UNIVERSITÉ D'AIX-MARSEILLE
FACULTÉ DE MÉDECINE DE MARSEILLE
ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

THÈSE

Présentée et publiquement soutenue devant

LA FACULTÉ DE MÉDECINE DE MARSEILLE

Le 05 Juillet 2018

Par Awa DIOP

Analyse des séquences des génomes bactériens en tant que source d'information taxonomique

Pour obtenir le grade de Docteur de l'Université d'AIX-MARSEILLE

Pathologie Humaine ; Spécialité Maladies Infectieuses

Membres du Jury de la Thèse :

Mme Christelle DESNUES	Présidente du jury
Mr Raymond RUIMY	Rapporteur
Mr Laurent BOYER	Rapporteur
Mr Pierre-Edouard FOURNIER	Directeur de thèse

Unité de Recherche Vecteurs-Infections Tropicales et Méditerranéennes
Aix-Marseille Université, IRD, SSA, AP-HM
Institut Hospitalo-Universitaire, Méditerranée Infection

Avant-propos

Le format de présentation de cette thèse correspond à une recommandation de la spécialité Maladies Infectieuses et Microbiologie, à l'intérieur du Master des Sciences de la Vie et de la Santé qui dépend de l'Ecole Doctorale des Sciences de la Vie de Marseille.

Le candidat est amené à respecter des règles qui lui sont imposées et qui comportent un format de thèse utilisé dans le Nord de l'Europe et qui permet un meilleur rangement que les thèses traditionnelles. Par ailleurs, la partie introduction et bibliographie est remplacée par une revue envoyée dans un journal afin de permettre une évaluation extérieure de la qualité de la revue et de permettre à l'étudiant de commencer le plus tôt possible une bibliographie exhaustive sur le domaine de cette thèse. Par ailleurs, la thèse est présentée sur article publié, accepté ou soumis associé d'un bref commentaire donnant le sens général du travail. Cette forme de présentation a paru plus en adéquation avec les exigences de la compétition internationale et permet de se concentrer sur des travaux qui bénéficieront d'une diffusion internationale.

Professeur Didier RAOULT

Remerciements

Je souhaite remercier toutes les personnes que j'ai cotoyées au cours de ma thèse et de mes études.

Tout d'abord, Je tiens à exprimer mes plus vifs remerciements et ma profonde gratitude au Professeur Pierre-Edouard FOURNIER de m'avoir accueillie dans son équipe et de m'avoir encadrée et guidée tout au long de cette thèse. J'ai pu bénéficier de sa patience et enthousiasme, de ses conseils, de son sens critique, de sa rigueur dans le travail, de ses compétences, et de ses grandes qualités pédagogiques qui ont été précieux pour moi et qui seront aussi l'excitation dans ma carriére future. Ses qualités humaines m'ont profondément touché. Ce fut un grand plaisir de passer ma thèse à vos côtés.

Je voudrais aussi remercier le Professeur Didier RAOULT de m'avoir accueillie dans son laboratoire et de m'avoir donné l'opportunité de réaliser cette thèse, pour m'avoir aussi pris en stage en master et de m'avoir ainsi donné la chance de mettre un pied dans le monde de la recherche.

Je tiens aussi à remercier les membres de mon jury de thése pour leurs regards critiques et pour avoir évalué soigneusement mes travaux de thése. Un grand merci au Docteur Christelle DESNUES d'avoir accepté de présider ce jury. Je remercie aussi le Professeur Raymond Ruimy et le Docteur Laurent Boyer qui ont accepté d'être rapporteurs de cette thèse.

Je voudrais aussi adresser un grand merci au Docteur Kahlid El-Karkouri de m'avoir initié aux expériences de génomique comparative et pangénomique surtout l'analyse d'évolution génomique et taxonomique des espèces du genre *Rickettsia,* au professeur Hervé Seligmann et Docteur Mathieu Million.

Je suis redevable d'exprimer mes remerciements au Professeur Florence Fenollar et Dr Oleg Mediannikov ainsi qu'à mes collègues Khoudia Diop, Amadou Hamidou Togo et El Hadji Seck pour le travail collaboré.

Merci à tous les membres de l'URMITE ayant participé de près ou de loin à ce travail incluant les techniciens, les ingénieurs plus particulièrement aux informaticiens Aurélia Caputo et Jeremy Delerce, à Frederic Cadoret et le personnel administratif et mes collègues étudiants. Et un grand merci à tous les membres de notre équipe Génomique bactérienne pour l'entraide, les conseils et les collaborations pendant ces trois années. Rita, Mamadou Beye…

Je remercie tous mes amis pour leur soutien et leur amitié. Je m'excuse de ne pas les mentionner individuellement.

Merci à mon mari Massaer GUEYE de m'avoir écoutée, soutenue et supportée au quotidien depuis qu'on s'est dit oui pour le meilleur et pour le pire.

Enfin, Je souhaite remercier toute ma famille plus particulièrement, à ma maman (Aby Gueye), à mon défunt père (Gora Diop) et à ma grande mère (Maty Djitté) pour leur amour inestimable, leurs sacrifices et pour tout ce que vous avez fait tout au long de mon éducation. A ma tante Aida FALL, à mon oncle Mamadou Mbeingue Gueye et à ma petite famille de Grenoble pour leur soutien, leurs sacrifices, et leurs encouragements tout au long de mes études.

Cette thése est aussi la vôtre je vous aime tous!

SOMMAIRE

Résumé

L'Identification rapide et la classification microbienne précise sont cruciales en microbiologie médicale pour la surveillance de la santé humaine et animale, établir un diagnostic clinique approprié et choisir des mesures thérapeutiques et de contrôle optimales. Initialement, la classification taxonomique des espèces bactériennes était basée sur des caractéristiques phénotypiques. Cependant, de nombreux outils génotypiques ont été mis au point pour compléter progressivement la définition des espèces bactériennes de façon plus fiable et precise dans une approche polyphasique intégrant les caractéristiques phénotypiques, l'analyse de la similarité et la phylogénie des séquences du gène de l'ARN ribosomique 16S (ARNr 16S), la teneur en G + C de l 'ADN (G+C%) ainsi que l'hybridation ADN-ADN (DDH). Même si ces outils sont largement utilisés, ils présentent plusieurs limites et inconvénients. En effet, les seuils universels de similarité de séquence de l'ARNr 16S (95% et 98,65% aux rangs du genre et de l'espèce, respectivement), de différence en G+C % (>5% entre deux espèces) et de DDH (<70% entre deux espèces) utilisés pour la définition des espèces ne sont pas applicables à de nombreux genres bactériens. C'est notamment le cas des espèces du genre *Rickettsia*, alpha-protéobactéries strictement intracellulaires qui expriment peu de caractéristiques phénotypiques. Ainsi, la définition des espèces au sein du genre *Rickettsia* a longtemps fait l'objet de débat. Mais en 2003, l'introduction d'un outil moléculaire basé sur l'analyse des séquences de cinq gènes a révolutionné la caractérisation et la classification taxonomique des rickettsies et constitue la base de leur classification à ce jour. En dépit de tous ces efforts, la taxonomie des membres du genre *Rickettsia* est restée un sujet de débat. Au cours des deux dernières décennies, les progrès remarquables de la technologie et de l'application du séquençage de l'ADN ont permis l'accès aux séquences génomiques complètes, permettant un accès sans précédant à des données précieuses pour une classification taxonomique plus précise des prokaryotes. Plusieurs outils

taxonomiques basés sur les séquences génomiques ont été développés. Compte tenu de la disponibilité des séquences génomiques de près de 100 génomes de *Rickettsia*, nous avons voulu évaluer une gamme de paramètres taxonomiques basés sur l'analyse des séquences génomiques afin de mettre au point des recommandations pour la classification des isolats au niveau de l'espèce et du genre. Nous avons également utilisé la genomique pour la caractérisation et la description des nouveaux isolats bactériens isolés par la méthode de "culturomique bactérienne" à partir de divers échantillons cliniques. En comparant le degré de similarité des séquences de 78 génomes de *Rickettsia* et 61 génomes de 3 genres étroitement apparentés (*Orientia*, 11 génomes, *Ehrlichia*, 22 génomes et *Anaplasma*, 28 génomes) en utilisant plusieurs paramètres génomiques (hybridation ADN-ADN, dDDH; l'identité nucléotidique moyenne par orthologie, OrthoANI et AGIOS; ou l'identité moyenne des séquences protéiques AAI, nous avons montré que les outils taxonomiques basés sur les séquences génomiques sont simples à utiliser et rapides, et permettent une classification taxonomique fiable et reproductible des isolats au sein des espèces du genre *Rickettsia*, avec des seuils spécifiques. Les résultats obtenus nous ont permis d'élaborer des lignes directrices pour la classification des isolats de rickettsies au niveau du genre et de l'espèce. À l'aide de la taxono-génomique, nous avons également pu décrire 17 nouvelles espèces bactériennes associées à l'homme sur la base d'une combinaison de l'analyse génomique et des propriétés phénotypiques. L'utilisation des outils génomiques est donc parfaitement adaptée à la classification taxonomique et peut changer radicalement notre vision de la taxonomie et de l'évolution bactérienne à l'avenir.

Mots clés: Génomique comparative, Génome bactérien, Taxonomie, Microbiologie, Definition d'espèce, *Rickettsia*

Abstract

Rapid identification and precise microbial classification are crucial in medical microbiology for human and animal health monitoring, appropriate clinical diagnosis and selection of optimal therapeutic and control measures. Initially, the taxonomic classification of bacterial species was based on phenotypic characteristics. However, many genotypic tools have been developed to progressively supplement the definition of bacterial species more reliably and accurately in a polyphasic approach incorporating phenotypic characteristics, analysis of similarity and phylogeny of sequences of the 16S ribosomal RNA gene (16S rRNA), the G + C content of DNA (G+C%), and DNA-DNA hybridization (DDH). Although these tools are widely used, they have several limitations and disadvantages. Indeed, the universal 16S rRNA sequence similarity thresholds (95% and 98.65% at the genus and species ranks, respectively), difference in G+C% (> 5% between two species) and DDH (< 70% between two species) used for the definition of species are not applicable to many bacterial genera. This is particularly true of species of the genus *Rickettsia* which are strictly intracellular alpha-proteobacteria that express few phenotypic characteristics. Thus, the definition of species within the genus *Rickettsia* has long been a matter of debate. But in 2003, the introduction of a molecular tool based on the analysis of five genes has revolutionized the characterization and taxonomic classification of rickettsiae and is the current basis for their classification. Despite these efforts, the taxonomy of members of the genus *Rickettsia* remained a subject of debate. Over the past two decades, the remarkable advances in DNA sequencing technologies have allowed access to complete genomic sequences, allowing unprecedented access to valuable data for a more accurate taxonomic

classification of prokaryotes. Several taxonomic tools based on genomic sequences have been developed. Given the availability of genomic sequences of nearly 100 rickettsial genomes, we wanted to evaluate a range of taxonomic parameters based on genomic sequence analysis, to develop guidelines for the classification of *Rickettsia* isolates at the genus and species levels. We have also used genomic sequences for the characterization and description of new bacterial isolates isolated by the "bacterial culturomics" method from various clinical specimens. By comparing the degree of similarity of the sequences of 78 genomes from *Rickettsia* species and 61 genomes from 3 closely related genera (*Orientia*, 11 genomes; *Ehrlichia*, 22 genomes; and *Anaplasma*, 28 genomes) using several genomic parameters (DNA-DNA hybridization, dDDH; the mean nucleotide identity by orthology, OrthoANI and AGIOS; or the mean identity of protein sequences AAI, we have shown that genome-based taxonomic tools are simple to use and fast, and allow for a reliable and reproducible taxonomic classification of isolates within species of the genus *Rickettsia*, with specific thresholds. The obtained results enabled us to develop guidelines for classifying rickettsial isolates at the genus and species levels. Using taxono-genomics, we have also been able to describe 17 new human-associated bacterial species on the basis of a combination of genomic analysis and phenotypic properties. The use of genomic tools is therefore perfectly adapted to taxonomic classification and can dramatically change our vision of taxonomy and bacterial evolution in the future.

Keywords: Comparative genomics, Bacterial genome, Taxonomy, Microbiology, Species definition, *Rickettsia*

INTRODUCTION

L'Identification rapide et la classification microbienne précise sont cruciales en microbiologie médicale pour la surveillance de la santé humaine et animale, établir un diagnostic clinique approprié et choisir des mesures thérapeutiques et de contrôle optimales des maladies infectieuses. Initialement, la classification taxonomique des espèces bactériennes était basée sur des caractéristiques phénotypiques [1, 2]. Cependant, de nombreux outils génotypiques ont été mis au point pour compléter progressivement la définition et la caracterisation des espèces bactériennes de façon plus fiable et plus précise dans une approche polyphasique [3–6]. Dans les années 1980 la taxonomie a connu un grand bouleversement provoqué par l'arrivée de méthodes de biologie moléculaire, notamment l'analyse de similarité de la sequence de l'ARN ribosomal 16S (ARNr 16S) [7–9]. Ainsi, l'approche polyphasique intégrant les caractéristiques phénotypiques, l'analyse de la similarité et la phylogénie des séquences de l'ARNr 16S, la teneur en G + C de l'ADN (G+C%) et l'hybridation ADN-ADN (DDH) est la stratégie de description taxonomique la plus largement acceptée des espèces bactériennes depuis 20 ans [10, 11]. Cependant, même si ces outils sont largement utilisés, ils présentaient plusieurs limites et inconvénients [12]. En effet, les seuils universels de similarité de séquence de l'ARNr 16S (95% et 98,65% aux rangs du genre et de l'espèce, respectivement), de différence en G+C % (>5% entre deux espèces) et de DDH (<70% entre deux espèces) utilisés pour la définition des espèces ne sont pas applicables à de nombreux genres bactériens [4, 13–

17]. C'est notamment le cas des espèces du genre *Rickettsia [13, 14, 12]*.

Les bactéries du genre *Rickettsia* sont des alpha-protéobactéries, bactéries strictement intracellulaires qui causent une gamme de maladies le plus souvent bénignes et d'évolution favorable, mais parfois aussi graves et mortelles [18, 19]. Elles sont transmises à l'homme et aux animaux dans le monde entier par divers vecteurs arthropodes (tiques, puces, poux, acariens). Les plus fréquentes des rickettsioses sont le SENLAT, la fièvre africaine à tiques (ATBF), le typhus murin, la fièvre boutonneuse méditerranéenne (MSF), la fièvre pourprée des montagnes Rocheuses (RMSF) et le typhus épidémique [20–22]. Les rickettsioses expriment peu de caractéristiques phénotypiques et présentant une faible hétérogénéité génétique [13, 14, 23]. Par conséquent, la définition des espèces au sein du genre *Rickettsia* a longtemps fait l'objet d'un débat et se basait uniquement sur des caractéristiques cliniques, épidémiologiques et des tests de serotypage chez la souris [13, 14, 24, 25]. En 2003, l'introduction d'un outil moléculaire basé sur l'analyse des séquences de cinq gènes: ARNr 16S, *gltA*, *ompA*, *ompB* et *sca4* a révolutionné la caractérisation et la classification taxonomique des rickettsies et constitue la base de leur classification à ce jour [13]. Cependant, en dépit de ces efforts, la taxonomie des membres du genre *Rickettsia* est restée un sujet de débat. A ce jour, il y a 30 espèces officiellement validées (www.bacterio.net/rickettsia.html) et de nombreux autres isolats de rickettsies qui n'ont pas encore été entièrement caractérisés, ou qui n'ont pas reçu de désignation d'espèce, ont également été récemment décrits.

En 1995, le séquençage complet du premier génome bactérien grâce à la méthodologie de Sanger, celui d'*Haemophilus influenzae* [26] a marqué le début de l'ère génomique. Ce fut un grand pas en avant en microbiologie en démontrant l'utilité de la génomique pour dévoiler le contenu génique complet d'une bactérie. Au cours des deux décennies suivantes, les progrès remarquables de la technologie et de l'application du séquençage de l'ADN à haut débit [27, 28] ont permis d'obtenir des séquences génomiques complètes (incluant plus de140 000 génomes bactériens dont plus de 100 genomes de *Rickettsia* à ce jour (Figure 1)), permettant l'accès sans précédent à des données précieuses pour une classification taxonomique plus précise des procaryotes. Par conséquent plusieurs outils taxonomiques basés sur les génomes ont été développés incluant l'hybridation ADN-ADN in silico (dDDH) [29–31], l'identité nucléotidique moyenne (ANI) [32–34], ou plus récemment l'identité nucléotidique moyenne par orthologie (OrthoANI) [35], l'identité moyenne des séquences protéiques (AAI) [17], l'indice maximal unique de l'ADN (MUMi) [36, 37], le pourcentage de protéines conservées (POCP) entre paires de genomes [38], la distance nucléotidique moyenne (FOA) [39] etc. Parmi ces méthodes, le DDH sert toujours de référence dans la classification taxonomique des procaryotes [10, 40]. Cependant, l'ANI constitue l'une des mesures les plus utilisées pour la délimitation des espèces dans l'ère génomique. Elle présente une forte corrélation avec les valeurs DDH, et a été proposée comme une alternative à DDH [41, 42]. Récemment, une approche légèrement différente de celle de la méthode ANI a été créée dans notre laboratoire pour calculer l'identité génomique entre paires de génomes [28, 40]. Le pipeline MAGi

(Marseille Average Genomic Identity) est un script perl qui permet à calculer l'identité génomique moyenne des séquences de gènes codant pour des protéines orthologues (AGIOS) entre deux génomes de souches bactériennes. Les paramètres AGIOS et ANI sont différents car pour ce dernier, les fragments orthologues sont identifiés en utilisant BLASTN, qui est moins sensible que BLASTP utilisé dans l'analyse AGIOS [40]. L'utilité des approches génomiques à des fins taxonomiques a été démontrée pour de nombreuses espèces bactériennes [41, 43–46]. Aujourd'hui, l'application de l'information génomique est recommandée pour la description taxonomique des espèces bactériennes [27]. Cependant, il n'existe aucune norme génomique spécifique pour la délimitation des espèces du genre *Rickettsia*. C'est dans cette optique que ce travail de doctorat s'inscrit avec comme objectif principal d'intégrer l'analyse des séquences génomiques en termes de contenu de gènes aussi bien que de similarité de séquence pour une meilleure delimitation des espèces, notamment par la mise au point de cutoffs génomiques entre genres et espèces. Dans un second temps, utiliser les outils génomiques pour la caractérisation et la description des nouveaux isolats bactériens isolés par la méthode de "culturomique bactérienne" à partir de divers échantillons cliniques.

Ce projet de thése est subdivisé en quatre sections présentées comme suit :

La première section (**Chapitre I**) a été consacrée à deux revues de la littérature scientifique sur les génomes des espèces de *Rickettsia.* La première revue soumise au journal Tick and Tick-borne diseases décrit l'évolution de la taille et du contenu

du génome des *Rickettsia*. Nous avons fait le point sur les différents mécanismes évolutifs qui façonnent le génome des rickettsies, à savoir une évolution convergente incluant une forte réduction génomique parallèlement à une expansion paradoxale de divers éléments génétiques. Et donc nous avons cherché à comprendre leur mode d'adaption dans un mode de vie strictement intracellaire. Ainsi la perte selective de gènes, la duplication de gènes, la prolifération d'éléments génétiques et le transfert horizontal de gènes ont tous façonné l'évolution des genomes des rickettsies **(Article 1)**. Dans la deuxième revue **(Article 2),** nous avons fait un lien entre l'évolution réductive du génome et l'augmentation de la virulence chez les rickettsies. Une conclusion frappante de l'étude genomique des rickettsies a été que les espèces les plus virulentes présentaient les génomes les plus réduits et les plus dégradés par rapport aux espèces les moins pathogènes ou non pathogènes étroitement proches qui en revanche, abritaient le plus grand nombre d'éléments génétiques mobiles. Par conséquent, l'évolution génomique réductrice contribue à l'émergence de la pathogénicité mais les mécanismes aboutissant à cet effet restent à élucider.

Dans la deuxième section (**Chapitre II**), nous proposons l'utilisation des données des séquences des genomes entiers pour la définition et la classification taxonomique des espèces du genre *Rickettsia*. Nous avons cherché à évaluer une gamme de paramètres génomiques basés sur l'analyse des séquences génomiques afin de mettre au point des recommandations pour la délimitation et la classification des isolats au niveau de l'espèce et du genre. Soixante-dix-huit génomes de souches de

Rickettsia disponibles dans GenBank ont été analysés et comparés.

La troisième section (**Chapitre III**) portant sur la taxono-génomique, a été introduite par une revue qui traite de l'impact de la culturomique sur la taxonomie en microbiologie clinique tout en tenant en compte de l'apport de la génomique. L'approche taxono-génomique consiste à incorporer les informations génomique notamment le séquençage du génome entier, la comparaison des caractéristiques génomiques associées aux données phénotypiques et protéomiques pour la caractérisation et la description des nouveaux isolats bactériens isolés par la méthode de "culturomique bactérienne" à partir de divers échantillons cliniques. Cette section contient des articles décrivant les 17 nouvelles espèces étudiées.

Dans la dernière section (**Chapitre IV**) contient deux articles décrivant le séquençage du génome entier d'espèces déjà connues et notamment l'analyse génomique de la souche type de l'espèce *Ezakiella peruensis* M6.X2 dont le premier génome séquencé et d'une nouvelle souche de *Megamonas funiformis* Marseille-P3344 isolée dans notre laboratoire.

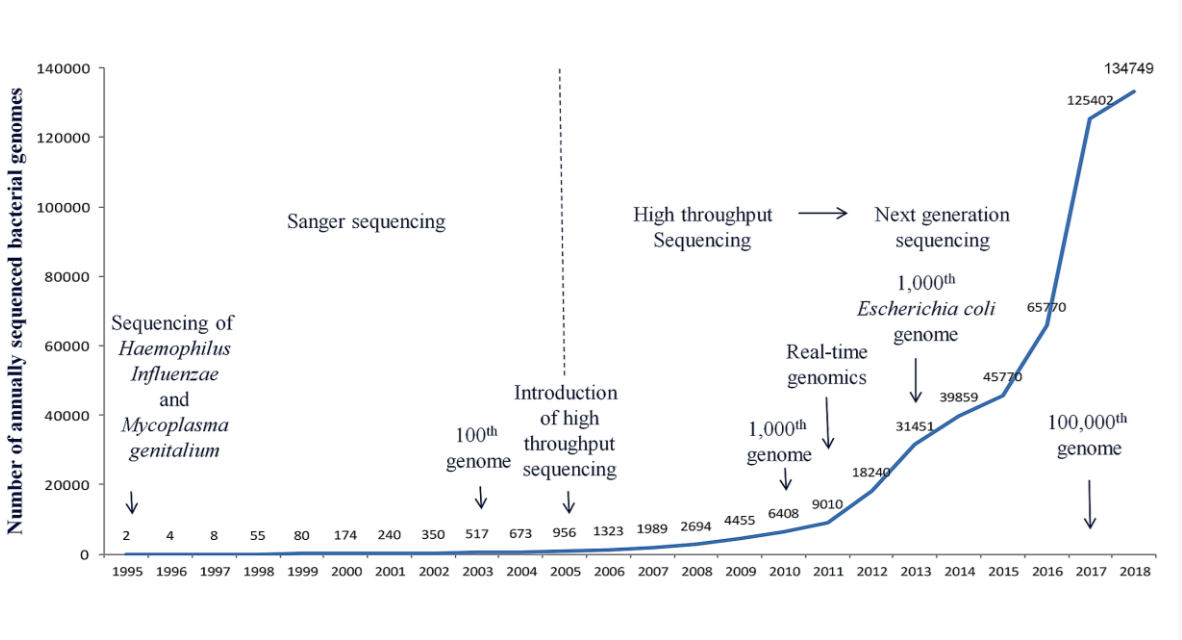


Figure 1: Nombre de séquences génomiques procaryotes publiées dans les bases de données publiques par an : Mars 2018

CHAPITRE I

Approche de l'évolution génomique des rickettsies

Article 1:

Paradoxical evolution of rickettsial genomes

Awa Diop, Didier Raoult, Pierre-Edouard Fournier

[Submitted in Ticks and Tick-borne Diseases]

Paradoxical evolution of rickettsial genomes

Awa Diop[1], Didier Raoult[2] and Pierre-Edouard Fournier[1*]

[1] UMR VITROME, Aix-Marseille University, IRD, Service de Santé des Armées, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Uuniversitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France Tel: +33 413 732 401, Fax: +33 413 732 402.

[2] UMR MEPHI, Aix-Marseille University, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Uuniversitaire Méditerranée Infection, Marseille, France

[*]Corresponding author: Pr Pierre-Edouard Fournier

Email: pierre-edouard.fournier@univ-amu.fr

Abstract:

Rickettsia species are strictly intracellular bacteria that evolved approximately 150 million years ago from a presumably free-living common ancestor of the order *Rickettsiales* that followed a transition to an obligate intracellular lifestyle. Rickettsiae are best known as human pathogen vectored by various arthropods causing a range of mild to severe human diseases. As part of their obligate intracellular lifestyle, rickettsial genomes have undergone a convergent evolution that includes a strong genomic reduction resulting from progressive gene degradation, genomic rearrangements as well as a paradoxical expansion of various genetic elements, notably short palindromic elements whose role remains unknown. This reductive evolutionary process is not unique to members of the *Rickettsia* genus but is common to several human pathogenic bacteria. Gene loss, gene duplication, DNA repeats duplication and horizontal gene transfer all have shaped rickettsial genome evolution. Gene loss mostly involved amino-acid, ATP, LPS and cell wall component biosynthesis and transcriptional regulators, but with a high preservation of toxin-antitoxin (TA) modules, recombination and DNA repair proteins. Surprisingly the most virulent *Rickettsia* species were shown to have the most drastically reduced and degraded genomes compared to closely related species of mild pathogenesis. In contrast, the less pathogenic species harbored the greatest number of mobile genetic elements. Thus, this distinct evolutionary process observed in *Rickettsia* species may be correlated with the differences in virulence and pathogenicity observed in these obligate intracellular bacteria. However, future investigations are needed to provide novel insights into the evolution of genome sizes and content, for that a better understanding of the balance between proliferation and elimination of genetic material in these intracellular bacteria is required.

Keywords: *Rickettsia*, genomics, evolution, virulence, genome rearrangement, non-coding DNA, gene loss, DNA repeats**.**

1 Introduction

The genus *Rickettsia* (order *Rickettsiales*, family *Rickettsiaceae*) comprises strictly intracellular α-proteobacteria mostly associated to diverse arthropod vectors around the world (Raoult and Roux, 1997; Stothard et al., 1994). These bacteria are also well known with infect mammalian hosts, mostly through arthropod bites or arthropod feces infecting scratching lesions. On the basis of their phenotypic properties and the diseases that they cause in humans, *Rickettsia* species were primarily phylogenetically classified into two major groups, namely the spotted fever group (SFG, Figure 1, Table 1) that contains species causing spotted fevers as well as numerous species of as-yet unknown pathogenicity. SFG rickettsiae are mostly associated with ticks, fleas and mites. The second major phylogenetic group, the typhus group (TG) is only made of *R. prowazekii* and *R. typhi* causing epidemic and murine typhus, and associated with human body lice and rat fleas, respectively. However, the SFG group was further divided into distinct phylogenetic subgroups on the basis of gene sequence comparisons (Gillespie et al., 2007; Merhej et al., 2014; Merhej and Raoult, 2011). In addition, two species, *R. bellii* and *R. canadensis*, associated with ticks but not causing any recognized human disease to date, diverged early from SFG and TG rickettsiae (Figure 1, Table 1). *Rickettsia* species evolved approximately 150 million years ago from a common ancestor of *Rickettsiales* that was presumably free-living, and progressively followed a transition to an obligate intracellular lifestyle that occurred 775–525 million years ago and then to primarily infecting arthropod lineages approximately 525–425 million years ago (El Karkouri et al., 2016; Merhej and Raoult, 2011; Weinert et al., 2009). *Rickettsia* species cause a range illnesses, from mild and self-limiting to severe and life-threatening diseases (Table 1). Currently, the most common rickettsioses are African tick-bite fever caused by *R. africae*, scalp eschar and neck lymphadenopathy (SENLAT) caused by *R. slovaca*, Mediterranean spotted fever (MSF) caused by *R. conorii*, Rocky Mountain spotted fever (RMSF) caused by

R. rickettsii and murine typhus caused by *R. typhi*. (El Karkouri et al., 2017; Parola et al., 2013; Sahni et al., 2013). *Rickettsia prowazekii*, the historical agent of epidemic typhus, is only rarely encountered currently but has a strong epidemic potential (Parola et al., 2013). Furthermore, recent studies have reported the association of other *Rickettsia* lineages with other reservoirs including protozoa, algae, leeches, plants or insects (Merhej and Raoult, 2011; Murray et al., 2016; Weinert et al., 2009).

In 1998, the first full *Rickettsia* genome and seventh bacterial genome to be sequenced was that of *R. prowazekii* strain Madrid E (Andersson et al., 1998). Subsequently, the genomes of most *Rickettsia* species have been fully sequenced, allowing a better knowledge of the molecular mechanisms involved in their pathogenicity (Balraj et al., 2009). Genome sequencing also appeared as a potential tool to revolutionize the phylogenetic and evolutionary investigations of prokaryotes, especially endosymbiotic bacteria. Hence, deciphering rickettsial genomes appeared as an efficient tool to understand the evolution of these obligate intracellular bacteria.

2 General features of rickettsial genomes

Rickettsia species have small genome sizes ranging from 1.1 to 2.3 Mbp. They are also AT-rich (G+C content from 28.9 to 33%, Table 2). SFG and TG rickettsiae exhibit genome sizes from 1.25 to 2.3 Mb and 1.11 Mb, with G+C contents ranging from 32.2 to 33.0% and 28.9 to 29%, respectively (Table 2). *Rickettsia* species have numbers of predicted protein-coding genes varying between 817 and 2,479 (Table 2) and many of them maintain a near perfect chromosomal colinearity (Ogata, 2001). This high degree of genomic synteny (Fig. 2) (Merhej and Raoult, 2011), enabled the identification of an ongoing and progressive genome degradation (Ogata, 2001). Rickettsial genomes contain many functional or unfunctional pseudogenes and possess a high percentage of non-coding DNA (Fig. 3) (Guillaume Blanc et

al., 2007; McLeod et al., 2004). *Rickettsia conorii* and *R. prowazekii* contain 19 and 24% of non-coding DNA respectively (Table 2). By comparison, *Chlamydia trachomatis*, another strictly intracellular bacterium, possesses only 10% non-coding DNA (Andersson et al., 1998; Holste et al., 2000; Rogozin et al., 2002).This pseudogenization progressively leads to a genome downsizing and results from a switch from a free-living to an obligate intracellular lifestyle. This progressive reductive evolution has allowed rickettsiae to purge unnecessary and redundant genes mainly involved in metabolisms supplied by eukaryotic host cells (Georgiades and Raoult, 2011; Merhej et al., 2009). Paradoxically to this ongoing genomic reduction, rickettsial genomes exhibit another marker of convergent evolution, *i. e.*, the expansion of genetic elements including plasmids, tandem repeats, short palindromic elements named rickettsia palindromic elements (RPEs) (Ogata et al., 2002), ankyrin and tetratricopeptide repeats and gene family duplication mainly ADP-ATP translocases, toxin-antitoxin modules and type IV secretion system (T4SS). Another unexpected property of rickettsial genomes is the presence of plasmids, the first described in obligate intracellular bacteria. The first plasmid was identified in *R. felis* (Ogata et al., 2005a). To date, at least 20 rickettsial plasmids have been described in 11 species. Their number varies from 1 to 4 per species/strain (Baldridge et al., 2007; G. Blanc et al., 2007; El Karkouri et al., 2016). These findings suggest possible exchanges of genetic material by conjugation, a mechanism that was thought to be absent in obligate intracellular and allopatric bacteria (Georgiades and Raoult, 2011; Merhej et al., 2009; Ogata et al., 2005a).

3 Rickettsial genome in an ongoing convergent evolution

3.1 Ongoing reductive evolution of rickettsial genomes

Following their adaptation from a free-living to an obligate intracellular lifestyle in eukaryotic cells, rickettsiae underwent genomic changes to fit their specific bottleneck

ecosystem, resulting not only in a reducing genome size but also in a specific genomic architecture (Keeling et al., 1994; Sicheritz-Pontén and Andersson, 1997). Comparative genomics revealed that rickettsiae, by taking advantage of host cell metabolites, underwent a genome reductive evolution (Georgiades and Raoult, 2011; Merhej et al., 2009) that occurred through a progressive pseudogenization (Fig. 3) and gene loss of selected biosynthetic pathway components (Andersson et al., 1998; Audia and Winkler, 2006; Fournier et al., 2009; Ogata, 2001; Sakharkar, 2004; Walker, 2005; Wolf and Koonin, 2013). In addition, genomic degradation was detrimental for the G+C content, as it led to an enrichment in A+T, in particular in the high proportion of non coding DNA (Sakharkar, 2004). However, a great variation in chromosome size, ranging from 1.1 to 2.3 Mb, is observed in rickettsiae (Table 2), indicating that some species are at a more advanced stage of reductive genomic evolution (TG rickettsiae) than others (SFG rickettsiae) (Ogata, 2001). An unexpected finding of rickettsial genomics was that the most virulent species had the most reduced genomes (Fournier et al., 2009). Such a finding is not an isolated phenomenon as in *Mycobacterium*, *Streptococcus* spp., *Corynebacterium* spp. and other genera, the highest degree of gene loss is observed in the most virulent species when compared to closely related and milder or nonpathogenic species (Guillaume Blanc et al., 2007; Merhej et al., 2013; Ogata, 2001).

Many of the genes required by free-living bacteria are absent in *Rickettsia* (Bechah et al., 2010) and degraded genes include mostly those coding for amino-acid, ATP, LPS and cell wall component biosynthesis (Blanc, 2005; Ogata, 2001; Renesto et al., 2005). Analysis of *R. conorii* and *R. prowazekii* genomes (Dunning Hotopp et al., 2006; Ogata, 2001) revealed that genes coding glycolytic enzymes and those required for nucleotide or cofactor biosynthesis are totally absent in *R. conorii* and *R. prowazekii* when compared to most genera in the order *Rickettsiales* that have complete glycolytic pathways. Nevertheless, rickettsiae must obtain glycerol-3-phosphate from the host via a glycerol-3-phosphate transporter (Dunning Hotopp

et al., 2006). This ATP production profile is similar for *Rickettsia* and mitochondria, as they possess a high number of ATP/ADP translocases, suggesting that they have both evolved from a common ancestor (Andersson et al., 1998; Renesto et al., 2005). In addition, the genome sequencing of *R. prowazekii* revealed a lack of amino acid metabolism such as those for glutamate metabolism (Andersson et al., 1998; Fuxelius et al., 2007). The enzymes involved in the aspartate and alanine metabolism pathways, and those playing a role in the biosynthesis of leucine, valine, isoleucine and aromatic amino acids (tryptophan, tyrosine, phenylalanine) are similarly missing in *Rickettsia* species (Renesto et al., 2005), suggesting the use of host-derived amino acids for their growth, survival and replication. Additionally, all *Rickettsia* species except *R. bellii* have a reduced set of folate biosynthesis genes (Fuxelius et al., 2007). In TG rickettsiae all five genes required for the de novo folate biosynthesis are lacking (Hunter et al., 2015). Furthermore, a limited set of genes for LPS and cell wall component biosynthesis, including lipid-A and peptidoglycan, respectively, were identified in *Rickettsia* species (Fuxelius et al., 2007). The rickettsial surface protein-coding genes *rickA* and *sca*2 are another example of genes that were degraded or eliminated by *Rickettsia* species during their specialization. The RickA protein participates in actin polymerization through the activation of Arp2/3 similar to that found in *Listeria monocytogenes* and *Shigella* spp. (Balraj et al., 2008b; Gouin et al., 2004, 1999). While lacking in the TG, *rickA* is present in all AG and SFG rickettsial genomes avalaible (Baldridge et al., 2005; Balraj et al., 2008a, 2008b; Heinzen et al., 1993; Jeng et al., 2004; McLeod et al., 2004; Ogata, 2001; Ogata et al., 2006, 2005a). The absence of *rickA* in *R. prowazekii* is not surprising if we consider its lack of actin motility. In contrast, *R. typhi* exhibits a unique and erratic actin-based motility despite having a nonfunctional RickA protein (McLeod et al., 2004; Reed et al., 2014). In addition, *R. canadensis* expresses RickA but does not exhibit actin-based motility (Heinzen et al., 1993). These data suggest the possible involvement of other actin polymerization mechanisms and

that RickA alone may not be sufficient or required for actin-based rickettsial motility. Nevertheless, it was proposed that RickA originated early in rickettsial evolution and may have been lost during the divergence of the TG. Recent research suggests that *Rickettsia* spp. use also Sca2 for actin-based motility with a distinct mechanism compared to RickA. Sca2 was found to be intact in *R. conorii*, absent in *R. prowazekii* and pseudogenized in *R. typhi* (McLeod et al., 2004). In *R. typhi*, Sca2 lacks the FH1 (formin homology 1) domain and contains only a proline-rich tract and a series of five WH2 domains (β-domains) in different locations with a divergence in sequences (Sears et al., 2012). The evolutionary process of genome degradation in rickettsiae led to loss of transcriptional regulator genes with a decreased translational capacity as observed in *R. prowazekii* (Andersson and Kurland, 1998), despite conserved gene sets coding for toxins, toxin-antitoxin (TA) modules and recombination and DNA repair proteins most likely needed for protection against host immune response (Moran, 2002).

The reductive evolution of rickettsial genomes is not only the consequence of gene degradation or loss, but it is also linked to a differential expression level of genes. Some genes under the influence of evolutionary forces are dormant or repressed while others under this effect are overexpressed. Recent research involving two virulent and two milder SFG rickettsiae demonstrated that the two virulent agents *R. conorii* (MSF) and *R. slovaca* (SENLAT) have the most reduced genome and displayed less up-regulated than down-regulated genes than the milder *R. massiliae* and *R. raoultii* causing MSF and SENLAT, respectively (El Karkouri et al., 2017), that have less reduced genomes. Consequently, to adapt to their specific intracellular environment, *Rickettsi*a species were shaped by distinct evolutionary processes. The most pathogenic species are characterized by a strong reductive genomic evolution, with a higher genome degradation rate and accumulation of non-coding DNA than less pathogenic species. These findings suggest that reductive genomic evolution,

resulting in protein structural variations, is associated to the emergence of virulence (El Karkouri et al., 2017). It was speculated that the loss of regulator genes, as observed in several intracellular pathogens, is a critical cause of virulence (Darby et al., 2007). This phenomenon was also observed in several human pathogens not genetically related to *Rickettsia* species such as *Treponema* spp., *Mycobacterium* spp. or *Yersinia* spp (Merhej et al., 2009; Walker, 2005; Wixon, 2001). As examples, *Mycobacterium leprae*, *Treponema pallidum* and *Yersinia pestis* have smaller genomes than closely related, but less virulent species, in their respective genera. Thus, genomic reduction with alteration of the regulation of invasion, replication and transmission processes, in addition to a differential level or degradation of expression of common proteins, may be correlated with an emergence of high pathogenicity. Overall, during the course of evolution, rickettsial genomes exhibit a trend toward gene loss rather than acquisition, but strong selective effects co-exist with functional duplication required for survival.

3.2 Gene order, recombination events and ‘‘junk DNA’’ in rickettsial genomes

A comparison of 8 rickettsial genomes (Fig. 3) demonstrated that they exhibit a highly conserved synteny and present few genomic rearrangements, except for *R. bellii* that exhibits little colinearity with other genomes, and *R. felis* that underwent several inversions. In addition, *R. typhi*, underwent a 35-kb inversion close to the replication terminus and a specific 124-kb inversion nearby the origin of replication when compared to *R. prowazekii* and *R. conorii* (McLeod et al., 2004). Inversions that occured in the origin of replication region are also found in *R. australis*, *R. helvetica* and *R. honei* (X. Dong et al., 2012; Xin Dong et al., 2012; Xin et al., 2012), indicating that this region constitutes a hot spot for genomic rearrangement. Homologous intra-chromosomal recombination, the principal mechanism for genomic rearrangement in rickettsiae, occured between repeated sequences or by site-specific recombination. Consequently, duplications, deletions and inversions arose through these

structures (Andersson and Kurland, 1998; Krawiec and Riley, 1990). Such events have been observed in Rickettsia spp., in the so-called super-ribosomal protein gene operon (Amiri, 2002). Highly conserved in a broad range of bacteria and archaea, this operon consists of about 40 genes located in seven operons in the same order (Sicheritz-Pontén and Andersson, 1997). Despite their conserved order in many bacteria including *E. coli* and *Bacillus subtilis*, genes in the ribosomal protein gene operon are scattered around the genomes of *Haemophilus influenzae*, *Mycoplasma genitalium* and *R. prowazeki* (Andersson and Kurland, 1998; Fraser et al., 1995). Ribosomal RNA genes in bacterial genomes are normally organized into an operon with a conserved order 16S-23S-5S, and tRNA genes are often found in the spacer between the 16S and the 23S rRNA genes (Krawiec and Riley, 1990). However, an unusual arrangement of rRNA genes has been observed in all avalaible *Rickettsia* genomes, as the 16S rRNA gene is separated from the 23S and 5S rRNA gene cluster (Andersson et al., 1999; Munson et al., 1993). The upstream spacer of the rearranged 23S rRNA gene in some *Rickettsia* species contains short repetitive sequences that have been eliminated in other related species, suggesting that the rearrangement of rRNA genes occurred by intra-chromosomal recombination prior to speciation in *Rickettsia* spp. Rickettsial genome analysis highlighted a second major genomic rearrangement in rickettsiae, the elongation factor proteins (tuf and fus) being present in more than one copy in *Rickettsia* genomes (Syvänen et al., 1996). These genes can serve as repeat sequences, and initiate a rapid gene loss through intra-chromosomal recombination (Krawiec and Riley, 1990). In addition, the degree and positions of deletions caused by intra-chromosomal recombination in *Rickettsia* is different among the species, which suggests that the homologous recombination is an ongoing process that may result in an ongoing genes loss under weak or no selection pressure (Amiri, 2002).

When compared to other bacterial genomes, rickettsial genomes have a high percentage of non-coding DNA sequences which also contains many DNA repeat sequences (Holste et al.,

2000; Rogozin et al., 2002). Non-coding DNA in rickettsial genomes is traditionally considered as "junk DNA" resulting from gene degradation. *R. prowazekii* and *R. typhi*, the most reduced rickettsial genomes, harbor high rates of non-coding DNA with 24.6 and 23.7%, respectively. However, *R. bellii* exhibits the lowest rickettsial level of non-coding DNA with 14.8% (Table 2).

3.3 Paradoxical genomic expansions

From a general point of view, rickettsial genomes are typical of those of symbiotic bacteria, in which the reductive trend is the dominant mode of evolution (Andersson and Andersson, 1999; Georgiades and Raoult, 2011; Merhej et al., 2009; Ogata, 2005). However, despite this reductive evolution, a paradoxical expansion of genetic elements can still occur in rickettsial genomes (Ogata et al., 2002). This genomic expansion may occur through proliferation of selfish DNA (small non coding RNAs (sRNAs), tandem repeats and rickettsia palindromic elements (RPEs)), gene duplications and horizontal gene transfer (Merhej and Raoult, 2011). Genome sequence analysis revealed that rickettsial genomes harbor many repetitive mobile elements, mainly sRNAs, tandem repeats and RPEs. Bacterial non-coding RNAs in intergenic regions were well documented in many bacterial taxa including *Enterobacteriaceae*, *Listeria monocytogenes*, *Clostridium perfringens*, *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Mycobacterium tuberculosis* (Papenfort and Vanderpool, 2015). Fifteen to 191 sRNAs were found in intergenic sequences, depending on species (Schroeder et al., 2015). These post-transcriptional regulators are assumed to influence virulence and adaptation depending on the host niche through transcriptomic regulation (Schroeder et al., 2015) . Their presence may explain why early comparative studies had identified highly conserved intergenic spacers (Ogata, 2001). A total of 1,785 sRNAs were predicted in 16 genomes of 13 species spanning all rickettsial groups (Schroeder

et al., 2015). *Rickettsia prowazekii* was shown to possess stem loop structures after homopolymeric poly(T) stretches in the termination sites (Woodard and Wood, 2011) where harbored mostly region encoding for sRNAs (Schroeder et al., 2015). Tandem repeats are generally distributed in intergenic regions (Fournier et al., 2004) and RPEs are present in both non-coding sequences and genes (Amiri et al., 2002; Ogata et al., 2000). These mobile genetic elements are found in most organisms (Ogata et al., 2000) and were considered an important factor in genome evolution. RPEs are more abundant in SFG than TG rickettsiae (Fig. 3). In the *R. conorii* genome, a total of 656 RPEs, classified into 8 families, were identified (RPE-1 to RPE-8) and represent 3.2% of the entire genome (Ogata et al., 2002). By comparison, only 10 of the 44 RPE-1 copies described in *R. conorii* were found in the *R. prowazekii* genome. Surprisingly, nine of these 10 RPE-1 copies that are present in *R. prowazekii* are inserted in protein-coding genes, *versus* 19/44 in *R. conorii*. In addition, the RPE-1s inserted into protein-coding genes have a position compatible with the 3-dimentional fold and function of proteins (Ogata et al., 2000). This process of genomic evolution by inserting RPEs within protein-coding genes was initially thought to be unique to *Rickettsia* species but is also encountered in the *Wolbachia* genus (Ogata et al., 2005b; Riegler et al., 2012). Bacteria may use this random strategy to adapt their genetic repertoire in response to selective environmental pressure. The presence of a mobile element inserted in many unrelated genes also suggests the potential role of selfish DNA in rickettsial genome for de novo creation of new protein sequences during the course of evolution, suggesting an implication in the dynamics of genome evolution (Claverie and Ogata, 2003). Moreover, genomic comparison also enabled the identification of several copies of Ankyrin and Tetratricopeptide (TPR)-repeats in rickettsiae. Such repeated elements are frequently found in endosymbionts and assumed to play a role in host-pathogen interaction (Caturegli et al., 2000; Felsheim et al., 2009; Seshadri et al., 2003; Wu et al., 2004). Twenty-two copies of ankyrin- and 11 copies of

TPR-repeats were found in *R. felis* (Ogata et al., 2005a). In both species, they were proposed to be linked to pathogenicity. In *Legionella pneumophila*, which exhibits 20 Ankyrin-repeat copies and numerous TPR-repeat copies, these elements are suspected to play a modulatory role in the interactions with the host cytoskeleton and in interferences with the host cell trafficking events, respectively (Cazalet et al., 2004).

In addition to DNA repeat sequences, gene family duplications are frequent in rickettsial genomes. Gene duplication was considered as an important source of bacterial adaptation to environmental changes in the host (Hooper, 2003). Following duplication, gene copies can evolve by conserving the same functions or undergoing mutations and becoming non-functional or assuming new functions, thus providing a putative new selective advantage in a new environment (Greub and Raoult, 2003; Walsh, 1995). *Rickettsia prowazekii*, the most reduced and degraded rickettsial genome that lacks the genes encoding the biosynthesis of purines and pyrimidines (Andersson et al., 1998), and *R. conorii* exhibit five copies of *tlc* genes. These genes encode ADP/ATP translocases responsible of energy exploitation from host cells (Greub and Raoult, 2003; Renesto et al., 2005). Similar sequences were found in *R. typhi*, *R. rickettsii* and *R. montanensis*. Four to 14 copies of *spoT* genes, involved in stringent response and the adaptation to intracellular environment, were also found in rickettsiae (Ogata et al., 2005a; Renesto et al., 2005; Rovery et al., 2005) . Other multicopy gene families present in *Rickettsia* genomes include Proline/Betaine transporters, toxin/antitoxin modules, Type IV secretion systems (T4SS), *sca* and *ampG*. All of these gene families are involved in rickettsial pathogenesis as previously described (Blanc, 2005; Georgiades and Raoult, 2011; Ogata, 2001; Renesto et al., 2005). The T4SS, a multiple component, membrane-spanning transporter system containing eight distinct classes such as the MPF-T class (P-T4SSs), is largely found in many rickettsial genomes. Rickettsiae possess an incomplete P-T4SS system (related to systems of the IncP group conjugative plasmid) that is characterized by the lack of

298 *virB*5 but the duplication of the *virB*4, *virB*6, *virB*8 and *virB*9 genes (Gillespie et al., 2016).
299 The *R. prowazekii* genome has six Vir components (*virB*4, *virB*8-*virB*11, *virD*4), and the
300 *virB*4 and *virB*9 were duplicated (Gillespie et al., 2009). Seventeen orthologous surface cell
301 antigen-coding genes (*sca*) were identified in rickettsial genomes (Blanc, 2005). SCA proteins
302 autotransporter proteins that were demonstrated to play roles in mammalian cell infection as
303 well as infection of their arthropod host cells, notably by promoting actin-based motility
304 (Sears et al., 2012). A set of conjugation genes (tra cluster, T4SS, ADP/ATP translocases and
305 patatin-encoding genes) found in Rickettsia spp. are phylogenetically close to those found in
306 many amoeba-associated bacteria, suggesting their acquisition by horizontal transfer events
307 between Rickettsia and non-rickettsial bacteria (Merhej and Raoult, 2011; Ogata et al., 2006).
308 Within amoebae, HGTs have given the *Rickettsia* ancestor the access to novel gene pools,
309 with possibility to acquire foreign DNA from other intracellular bacteria, thus, in capability of
310 adaptation environment (Ogata et al., 2006).

311 Finally, a large number of mobile genetic elements (MGEs) referred to as as mobilome
312 are found in rickettsiae despite their reduced genome size. This mobilome, mostly consisting
313 of plasmids, may ensure DNA movement within and between genomes. To date, at least 20
314 known rickettsial plasmids have been described in 11 species despite their allopatric lifestyle
315 (Table 2). Plasmids were most likely acquired vertically from *Orientia*/*Rickettsia*
316 chromosome ancestors (El Karkouri et al., 2016). Recent phylogenomic analysis revealed that
317 rickettsial plasmids are undergoing reductive evolutionary events similar to those affecting
318 their co-residing chromosomes (El Karkouri et al., 2016). Rickettsial plasmids were thus
319 shaped by a biphasic model of convergent evolution including a strong reductive evolution as
320 well as an increased complexity via horizontal gene transfer and gene duplication and genesis
321 (El Karkouri et al., 2016). The most reduced and virulent rickettsial genomes have probablely
322 lost plasmid(s) during their evolution when compared to the related milder or non pathogenic

species (Darby et al., 2007; El Karkouri et al., 2017; Ogata et al., 2005a). In addition, The genome from REIS, the largest rickettsial genome described to date, is characterized by a remarkable proliferation of mobile genetic elements (35% of the entire genome) including a RAGE module resulting from multiplied genomic invasion events, and was considered as a genetic exchange facilitator (Gillespie et al., 2014, 2012). The RAGE module was also described in *O. tsutsugamushi*, *R. massiliae* (G. Blanc et al., 2007), *R. bellii* (Ogata et al., 2006) and in the pLbaR plasmid of *R. felis* strain LSU-Lb (Gillespie et al., 2015).

4 Conclusions and Perspectives

Rickettsia species are strictly intracellular bacteria that are likely to have evolved approximately 150 million years ago from a common ancestor of *Rickettsiales* that was presumably free-living and followed a transition to an obligate intracellular lifestyle. To adapt to such a bottleneck lifestyle associated with genetic drift, *Rickettsia* species have been shaped by distinct evolutionary processes resulting not only in differences in genome size, but also in genomic architecture. Generally, rickettsial genomes are small and contain a high ratio of non-coding DNA, which suggests that the reductive trend is their dominant mode of evolution. Comparative sequence analysis has provided important clues on the mechanisms driving the genome-reduction process of *Rickettsia* spp. This phenomenon is marked by a selected loss of genes such as those associated with amino-acid, ATP, LPS and cell wall component biosynthesis with a loss of regulatory genes and a high preservation of toxin-associated proteins and toxin-antitoxin modules. Homologous intra-chromosomal recombination, principal mechanism for genomic rearrangement structures seems play a role in rapid gene loss. Consequently, rickettsiae have evolved under a distinct process including a strong reductive evolution as well as a paradoxical expansion of genetic elements acquired by horizontal gene transfer and gene duplication and genesis. Thus, during the course of

evolution, rickettsial genomes had a trend of gene loss rather than gene acquisition or duplication, but these strong selective effects co-exist with functional duplications required for survival. In order to understand the evolution of genome size and content, it is necessary to understand the balance between proliferation and elimination of genetic material in these intracellular bacteria.

5 References

Amiri, H., 2002. Patterns and Processes of Molecular Evolution in Rickettsia. DIVA.

Amiri, H., Alsmark, C., Andersson, S., 2002. Proliferation and Deterioration of Rickettsia Palindromic Elements.

Andersson, J.O., Andersson, S.G., 1999. Genome degradation is an ongoing process in Rickettsia. Mol. Biol. Evol. 16, 1178–1191. https://doi.org/10.1093/oxfordjournals.molbev.a026208

Andersson, S.G.., Kurland, C.G., 1998. Reductive evolution of resident genomes. Trends Microbiol. 6, 263–268. https://doi.org/10.1016/S0966-842X(98)01312-2

Andersson, S.G., Stothard, D.R., Fuerst, P., Kurland, C.G., 1999. Molecular phylogeny and rearrangement of rRNA genes in Rickettsia species. Mol. Biol. Evol. 16, 987–995. https://doi.org/10.1093/oxfordjournals.molbev.a026188

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C.M., Podowski, R.M., Näslund, A.K., Eriksson, A.-S., Winkler, H.H., Kurland, C.G., 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 396, 133–140.

Audia, J.P., Winkler, H.H., 2006. Study of the Five Rickettsia prowazekii Proteins Annotated as ATP/ADP Translocases (Tlc): Only Tlc1 Transports ATP/ADP, While Tlc4 and Tlc5 Transport Other Ribonucleotides. J. Bacteriol. 188, 6261–6268. https://doi.org/10.1128/JB.00371-06

Baldridge, G.D., Burkhardt, N., Herron, M.J., Kurtti, T.J., Munderloh, U.G., 2005. Analysis of Fluorescent Protein Expression in Transformants of Rickettsia monacensis, an Obligate Intracellular Tick Symbiont. Appl. Environ. Microbiol. 71, 2095–2105. https://doi.org/10.1128/AEM.71.4.2095-2105.2005

Baldridge, G.D., Burkhardt, N.Y., Felsheim, R.F., Kurtti, T.J., Munderloh, U.G., 2007. Transposon Insertion Reveals pRM, a Plasmid of Rickettsia monacensis. Appl. Environ. Microbiol. 73, 4984–4995. https://doi.org/10.1128/AEM.00988-07

Balraj, P., Karkouri, K.E., Vestris, G., Espinosa, L., Raoult, D., Renesto, P., 2008a. RickA Expression Is Not Sufficient to Promote Actin-Based Motility of Rickettsia raoultii. PLoS ONE 3, e2582. https://doi.org/10.1371/journal.pone.0002582

Balraj, P., Nappez, C., Raoult, D., Renesto, P., 2008b. Western-blot detection of RickA within spotted fever group rickettsiae using a specific monoclonal antibody. FEMS Microbiol. Lett. 286, 257–262. https://doi.org/10.1111/j.1574-6968.2008.01283.x

Balraj, P., Renesto, P., Raoult, D., 2009. Advances in Rickettsia Pathogenicity. Ann. N. Y. Acad. Sci. 1166, 94–105. https://doi.org/10.1111/j.1749-6632.2009.04517.x

Bechah, Y., El Karkouri, K., Mediannikov, O., Leroy, Q., Pelletier, N., Robert, C., Medigue, C., Mege, J.L., Raoult, D., 2010. Genomic, proteomic, and transcriptomic analysis of virulent and avirulent Rickettsia prowazekii reveals its adaptive mutation capabilities. Genome Res. 20, 655–663. https://doi.org/10.1101/gr.103564.109

Blanc, G., 2005. Molecular Evolution of Rickettsia Surface Antigens: Evidence of Positive Selection. Mol. Biol. Evol. 22, 2073–2083. https://doi.org/10.1093/molbev/msi199

Blanc, G., Ogata, H., Robert, C., Audic, S., Claverie, J.-M., Raoult, D., 2007. Lateral gene transfer between obligate intracellular bacteria: Evidence from the Rickettsia massiliae genome. Genome Res. 17, 1657–1664. https://doi.org/10.1101/gr.6742107

Blanc, G., Ogata, H., Robert, C., Audic, S., Suhre, K., Vestris, G., Claverie, J.-M., Raoult, D., 2007a. Reductive genome evolution from the mother of Rickettsia. PLoS Genet 3, e14.

Blanc, G., Ogata, H., Robert, C., Audic, S., Suhre, K., Vestris, G., Claverie, J.-M., Raoult, D., 2007b. Reductive Genome Evolution from the Mother of Rickettsia. PLoS Genet. 3, e14. https://doi.org/10.1371/journal.pgen.0030014

Caturegli, P., Asanovich, K.M., Walls, J.J., Bakken, J.S., Madigan, J.E., Popov, V.L., Dumler, J.S., 2000. ankA: an Ehrlichia phagocytophila group gene encoding a cytoplasmic protein antigen with ankyrin repeats. Infect. Immun. 68, 5277–5283.

Cazalet, C., Rusniok, C., Brüggemann, H., Zidane, N., Magnier, A., Ma, L., Tichit, M., Jarraud, S., Bouchier, C., Vandenesch, F., Kunst, F., Etienne, J., Glaser, P., Buchrieser, C., 2004. Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity. Nat. Genet. 36, 1165–1173. https://doi.org/10.1038/ng1447

Claverie, J.-M., Ogata, H., 2003. The insertion of palindromic repeats in the evolution of proteins. Trends Biochem. Sci. 28, 75–80. https://doi.org/10.1016/S0968-0004(02)00036-1

Darby, A.C., Cho, N.-H., Fuxelius, H.-H., Westberg, J., Andersson, S.G.E., 2007. Intracellular pathogens go extreme: genome evolution in the Rickettsiales. Trends Genet. 23, 511–520. https://doi.org/10.1016/j.tig.2007.08.002

Dong, X., El Karkouri, K., Robert, C., Gavory, F., Raoult, D., Fournier, P.-E., 2012. Genomic Comparison of Rickettsia helvetica and Other Rickettsia Species. J. Bacteriol. 194, 2751–2751. https://doi.org/10.1128/JB.00299-12

Dong, X., El Karkouri, K., Robert, C., Raoult, D., Fournier, P.-E., 2012. Genome Sequence of Rickettsia australis, the Agent of Queensland Tick Typhus. J. Bacteriol. 194, 5129. https://doi.org/10.1128/JB.01117-12

Dunning Hotopp, J.C., Lin, M., Madupu, R., Crabtree, J., Angiuoli, S.V., Eisen, J., Seshadri, R., Ren, Q., Wu, M., Utterback, T.R., Smith, S., Lewis, M., Khouri, H., Zhang, C., Niu, H., Lin, Q., Ohashi, N., Zhi, N., Nelson, W., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sundaram, J., Daugherty, S.C., Davidsen, T., Durkin, A.S., Gwinn, M., Haft, D.H., Selengut, J.D., Sullivan, S.A., Zafar, N., Zhou, L., Benahmed, F., Forberger, H., Halpin, R., Mulligan, S., Robinson, J., White, O., Rikihisa, Y., Tettelin, H., 2006. Comparative Genomics of Emerging Human Ehrlichiosis Agents. PLoS Genet. 2, e21. https://doi.org/10.1371/journal.pgen.0020021

El Karkouri, K., Kowalczewska, M., Armstrong, N., Azza, S., Fournier, P.-E., Raoult, D., 2017. Multi-omics Analysis Sheds Light on the Evolution and the Intracellular Lifestyle Strategies of Spotted Fever Group Rickettsia spp. Front. Microbiol. 8. https://doi.org/10.3389/fmicb.2017.01363

El Karkouri, K., Mediannikov, O., Robert, C., Raoult, D., Fournier, P.-E., 2016a. Genome Sequence of the Tick-Borne Pathogen *Rickettsia raoultii*. Genome Announc. 4, e00157–16. https://doi.org/10.1128/genomeA.00157-16

El Karkouri, K., Pontarotti, P., Raoult, D., Fournier, P.-E., 2016b. Origin and Evolution of Rickettsial Plasmids. PLOS ONE 11, e0147492. https://doi.org/10.1371/journal.pone.0147492

Felsheim, R.F., Kurtti, T.J., Munderloh, U.G., 2009. Genome Sequence of the Endosymbiont Rickettsia peacockii and Comparison with Virulent Rickettsia rickettsii: Identification of Virulence Factors. PLoS ONE 4, e8361. https://doi.org/10.1371/journal.pone.0008361

Fournier, P.-E., El Karkouri, K., Leroy, Q., Robert, C., Giumelli, B., Renesto, P., Socolovschi, C., Parola, P., Audic, S., Raoult, D., 2009. Analysis of the Rickettsia africae genome reveals that virulence acquisition in Rickettsia species may be explained by genome reduction. BMC Genomics 10, 166. https://doi.org/10.1186/1471-2164-10-166

Fournier, P.-E., Zhu, Y., Ogata, H., Raoult, D., 2004. Use of Highly Variable Intergenic Spacer Sequences for Multispacer Typing of Rickettsia conorii Strains. J. Clin. Microbiol. 42, 5757–5766. https://doi.org/10.1128/JCM.42.12.5757-5766.2004

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, R.D., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., Venter, J.C., 1995. The minimal gene complement of Mycoplasma genitalium. Science 270, 397–403.

Fuxelius, H.-H., Darby, A., Min, C.-K., Cho, N.-H., Andersson, S.G.E., 2007. The genomic and metabolic diversity of Rickettsia. Res. Microbiol. 158, 745–753. https://doi.org/10.1016/j.resmic.2007.09.008

Georgiades, K., Raoult, D., 2011. Genomes of the Most Dangerous Epidemic Bacteria Have a Virulence Repertoire Characterized by Fewer Genes but More Toxin-Antitoxin Modules. PLoS ONE 6, e17962. https://doi.org/10.1371/journal.pone.0017962

Gillespie, J.J., Ammerman, N.C., Dreher-Lesnick, S.M., Rahman, M.S., Worley, M.J., Setubal, J.C., Sobral, B.S., Azad, A.F., 2009. An Anomalous Type IV Secretion System in Rickettsia Is Evolutionarily Conserved. PLoS ONE 4, e4833. https://doi.org/10.1371/journal.pone.0004833

Gillespie, J.J., Beier, M.S., Rahman, M.S., Ammerman, N.C., Shallom, J.M., Purkayastha, A., Sobral, B.S., Azad, A.F., 2007. Plasmids and Rickettsial Evolution: Insight from Rickettsia felis. PLoS ONE 2, e266. https://doi.org/10.1371/journal.pone.0000266

Gillespie, J.J., Driscoll, T.P., Verhoeve, V.I., Utsuki, T., Husseneder, C., Chouljenko, V.N., Azad, A.F., Macaluso, K.R., 2015. Genomic Diversification in Strains of Rickettsia felis Isolated from Different Arthropods. Genome Biol. Evol. 7, 35–56. https://doi.org/10.1093/gbe/evu262

Gillespie, J.J., Joardar, V., Williams, K.P., Driscoll, T., Hostetler, J.B., Nordberg, E., Shukla, M., Walenz, B., Hill, C.A., Nene, V.M., Azad, A.F., Sobral, B.W., Caler, E., 2012. A Rickettsia Genome Overrun by Mobile Genetic Elements Provides Insight into the Acquisition of Genes Characteristic of an Obligate Intracellular Lifestyle. J. Bacteriol. 194, 376–394. https://doi.org/10.1128/JB.06244-11

Gillespie, J.J., Kaur, S.J., Rahman, M.S., Rennoll-Bankert, K., Sears, K.T., Beier-Sexton, M., Azad, A.F., 2014. Secretome of obligate intracellular *Rickettsia*. FEMS Microbiol. Rev. n/a–n/a. https://doi.org/10.1111/1574-6976.12084

Gillespie, J.J., Phan, I.Q.H., Driscoll, T.P., Guillotte, M.L., Lehman, S.S., Rennoll-Bankert, K.E., Subramanian, S., Beier-Sexton, M., Myler, P.J., Rahman, M.S., Azad, A.F., 2016. The *Rickettsia* type IV secretion system: unrealized complexity mired by gene family expansion. Pathog. Dis. 74, ftw058. https://doi.org/10.1093/femspd/ftw058

Gouin, E., Egile, C., Dehoux, P., Villiers, V., Adams, J., Gertler, F., Li, R., Cossart, P., 2004. The RickA protein of Rickettsia conorii activates the Arp2/3 complex. Nature 427, 457.

Gouin, E., Gantelet, H., Egile, C., Lasa, I., Ohayon, H., Villiers, V., Gounon, P., Sansonetti, P.J., Cossart, P., 1999. A comparative study of the actin-based motilities of the pathogenic bacteria Listeria monocytogenes, Shigella flexneri and Rickettsia conorii. J. Cell Sci. 112, 1697–1708.

Greub, G., Raoult, D., 2003. History of the ADP/ATP-Translocase-Encoding Gene, a Parasitism Gene Transferred from a Chlamydiales Ancestor to Plants 1 Billion Years Ago. Appl. Environ. Microbiol. 69, 5530–5535. https://doi.org/10.1128/AEM.69.9.5530-5535.2003

Heinzen, R.A., Hayes, S.F., Peacock, M.G., Hackstadt, T., 1993. Directional actin polymerization associated with spotted fever group Rickettsia infection of Vero cells. Infect. Immun. 61, 1926–1935.

Holste, D., Weiss, O., Grosse, I., Herzel, H., 2000. Are Noncoding Sequences of Rickettsia prowazekii Remnants of ``Neutralized'' Genes? J. Mol. Evol. 51, 353–362. https://doi.org/10.1007/s002390010097

Hooper, S.D., 2003. On the Nature of Gene Innovation: Duplication Patterns in Microbial Genomes. Mol. Biol. Evol. 20, 945–954. https://doi.org/10.1093/molbev/msg101

Hunter, D.J., Torkelson, J.L., Bodnar, J., Mortazavi, B., Laurent, T., Deason, J., Thephavongsa, K., Zhong, J., 2015. The Rickettsia endosymbiont of Ixodes pacificus contains all the genes of de novo folate biosynthesis. PloS One 10, e0144552.

494 Jeng, R.L., Goley, E.D., D'Alessio, J.A., Chaga, O.Y., Svitkina, T.M., Borisy, G.G., Heinzen, R.A., Welch,
495 M.D., 2004. A Rickettsia WASP-like protein activates the Arp2/3 complex and mediates actin-
496 based motility: Rickettsia RickA activates the Arp2/3 complex. Cell. Microbiol. 6, 761–769.
497 https://doi.org/10.1111/j.1462-5822.2004.00402.x
498 Keeling, P.J., Charlebois, R.L., Ford Doolittle, W., 1994. Archaebacterial genomes: eubacterial form
499 and eukaryotic content. Curr. Opin. Genet. Dev. 4, 816–822. https://doi.org/10.1016/0959-
500 437X(94)90065-5
501 Krawiec, S., Riley, M., 1990. Organization of the bacterial chromosome. Microbiol. Rev. 54, 502–539.
502 McLeod, M.P., Qin, X., Karpathy, S.E., Gioia, J., Highlander, S.K., Fox, G.E., McNeill, T.Z., Jiang, H.,
503 Muzny, D., Jacob, L.S., Hawes, A.C., Sodergren, E., Gill, R., Hume, J., Morgan, M., Fan, G.,
504 Amin, A.G., Gibbs, R.A., Hong, C., Yu, X. -j., Walker, D.H., Weinstock, G.M., 2004. Complete
505 Genome Sequence of Rickettsia typhi and Comparison with Sequences of Other Rickettsiae.
506 J. Bacteriol. 186, 5842–5855. https://doi.org/10.1128/JB.186.17.5842-5855.2004
507 Merhej, V., Angelakis, E., Socolovschi, C., Raoult, D., 2014. Genotyping, evolution and epidemiological
508 findings of Rickettsia species. Infect. Genet. Evol. 25, 122–137.
509 https://doi.org/10.1016/j.meegid.2014.03.014
510 Merhej, V., Georgiades, K., Raoult, D., 2013. Postgenomic analysis of bacterial pathogens repertoire
511 reveals genome reduction rather than virulence factors. Brief. Funct. Genomics 12, 291–304.
512 https://doi.org/10.1093/bfgp/elt015
513 Merhej, V., Raoult, D., 2011. Rickettsial evolution in the light of comparative genomics. Biol. Rev. 86,
514 379–405. https://doi.org/10.1111/j.1469-185X.2010.00151.x
515 Merhej, V., Royer-Carenzi, M., Pontarotti, P., Raoult, D., 2009. Massive comparative genomic analysis
516 reveals convergent evolution of specialized bacteria. Biol. Direct 4, 13.
517 https://doi.org/10.1186/1745-6150-4-13
518 Moran, N.A., 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell 108, 583–
519 586.
520 Munson, M.A., Baumann, L., Baumann, P., 1993. Buchnera aphidicola (a prokaryotic endosymbiont of
521 aphids) contains a putative 16S rRNA operon unlinked to the 23S rRNA-encoding gene:
522 sequence determination, and promoter and terminator analysis. Gene 137, 171–178.
523 https://doi.org/10.1016/0378-1119(93)90003-L
524 Murray, G.G.R., Weinert, L.A., Rhule, E.L., Welch, J.J., 2016. The Phylogeny of *Rickettsia* Using
525 Different Evolutionary Signatures: How Tree-Like is Bacterial Evolution? Syst. Biol. 65, 265–
526 279. https://doi.org/10.1093/sysbio/syv084
527 Ogata, H., 2005. Rickettsia felis, from Culture to Genome Sequencing. Ann. N. Y. Acad. Sci. 1063, 26–
528 34. https://doi.org/10.1196/annals.1355.004
529 Ogata, H., 2001. Mechanisms of Evolution in Rickettsia conorii and R. prowazekii. Science 293, 2093–
530 2098. https://doi.org/10.1126/science.1061471
531 Ogata, H., Audic, S., Abergel, C., Fournier, P.-E., Claverie, J.-M., 2002. Protein coding palindromes are
532 a unique but recurrent feature in Rickettsia. Genome Res. 12, 808–816.
533 Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P.-E., Raoult, D., M Claverie, J., 2000. Selfish
534 DNA in Protein-Coding Genes of Rickettsia.
535 Ogata, H., La Scola, B., Audic, S., Renesto, P., Blanc, G., Robert, C., Fournier, P.-E., Claverie, J.-M.,
536 Raoult, D., 2006. Genome Sequence of Rickettsia bellii Illuminates the Role of Amoebae in
537 Gene Exchanges between Intracellular Pathogens. PLoS Genet. 2, e76.
538 https://doi.org/10.1371/journal.pgen.0020076
539 Ogata, H., Renesto, P., Audic, S., Robert, C., Blanc, G., Fournier, P.-E., Parinello, H., Claverie, J.-M.,
540 Raoult, D., 2005a. The Genome Sequence of Rickettsia felis Identifies the First Putative
541 Conjugative Plasmid in an Obligate Intracellular Parasite. PLoS Biol. 3, e248.
542 https://doi.org/10.1371/journal.pbio.0030248
543 Ogata, H., Suhre, K., Claverie, J.-M., 2005b. Discovery of protein-coding palindromic repeats in
544 Wolbachia. Trends Microbiol. 13, 253–5. https://doi.org/10.1016/j.tim.2005.03.013

Papenfort, K., Vanderpool, C.K., 2015. Target activation by regulatory RNAs in bacteria. FEMS Microbiol. Rev. 39, 362–378. https://doi.org/10.1093/femsre/fuv016
Parola, P., Paddock, C.D., Socolovschi, C., Labruna, M.B., Mediannikov, O., Kernif, T., Abdad, M.Y., Stenos, J., Bitam, I., Fournier, P.-E., Raoult, D., 2013. Update on Tick-Borne Rickettsioses around the World: a Geographic Approach. Clin. Microbiol. Rev. 26, 657–702. https://doi.org/10.1128/CMR.00032-13
Raoult, D., Roux, V., 1997. Rickettsioses as paradigms of new or emerging infectious diseases. Clin. Microbiol. Rev. 10, 694–719.
Reed, S.C.O., Lamason, R.L., Risca, V.I., Abernathy, E., Welch, M.D., 2014. Rickettsia Actin-Based Motility Occurs in Distinct Phases Mediated by Different Actin Nucleators. Curr. Biol. 24, 98–103. https://doi.org/10.1016/j.cub.2013.11.025
Renesto, P., Ogata, H., Audic, S., Claverie, J.-M., Raoult, D., 2005. Some lessons from *Rickettsia* genomics. FEMS Microbiol. Rev. 29, 99–117. https://doi.org/10.1016/j.femsre.2004.09.002
Riegler, M., Iturbe-Ormaetxe, I., Woolfit, M., Miller, W.J., O'Neill, S.L., 2012. Tandem repeat markers as novel diagnostic tools for high resolution fingerprinting of Wolbachia. BMC Microbiol. 12, S12.
Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., Koonin, E.V., 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. 30, 4264–4271.
Rovery, C., Renesto, P., Crapoulet, N., Matsumoto, K., Parola, P., Ogata, H., Raoult, D., 2005. Transcriptional response of Rickettsia conorii exposed to temperature variation and stress starvation. Res. Microbiol. 156, 211–218. https://doi.org/10.1016/j.resmic.2004.09.002
Sahni, S.K., Narra, H.P., Sahni, A., Walker, D.H., 2013. Recent molecular insights into rickettsial pathogenesis and immunity. Future Microbiol. 8, 1265–1288. https://doi.org/10.2217/fmb.13.102
Sakharkar, K.R., 2004. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. Int. J. Syst. Evol. Microbiol. 54, 1937–1941. https://doi.org/10.1099/ijs.0.63090-0
Schroeder, C.L.C., Narra, H.P., Rojas, M., Sahni, A., Patel, J., Khanipov, K., Wood, T.G., Fofanov, Y., Sahni, S.K., 2015. Bacterial small RNAs in the Genus Rickettsia. BMC Genomics 16. https://doi.org/10.1186/s12864-015-2293-7
Sears, K.T., Ceraul, S.M., Gillespie, J.J., Allen, E.D., Popov, V.L., Ammerman, N.C., Rahman, M.S., Azad, A.F., 2012. Surface Proteome Analysis and Characterization of Surface Cell Antigen (Sca) or Autotransporter Family of Rickettsia typhi. PLoS Pathog. 8, e1002856. https://doi.org/10.1371/journal.ppat.1002856
Seshadri, R., Paulsen, I.T., Eisen, J.A., Read, T.D., Nelson, K.E., Nelson, W.C., Ward, N.L., Tettelin, H., Davidsen, T.M., Beanan, M.J., others, 2003. Complete genome sequence of the Q-fever pathogen Coxiella burnetii. Proc. Natl. Acad. Sci. 100, 5455–5460.
Sicheritz-Pontén, T., Andersson, S.G., 1997. GRS: a graphic tool for genome retrieval and segment analysis. Microb. Comp. Genomics 2, 123–139.
Stothard, D.R., Clark, J.B., Fuerst, P.A., 1994. Ancestral divergence of Rickettsia bellii from the spotted fever and typhus groups of Rickettsia and antiquity of the genus Rickettsia. Int. J. Syst. Evol. Microbiol. 44, 798–804.
Syvänen, A.-C., Amiri, H., Jamal, A., Andersson, S.G., Kurland, C.G., 1996. A chimeric disposition of the elongation factor genes in Rickettsia prowazekii. J. Bacteriol. 178, 6192–6199.
Walker, D.H., 2005. Progress in Rickettsial Genome Analysis from Pioneering of Rickettsia prowazekii to the Recent Rickettsia typhi. Ann. N. Y. Acad. Sci. 1063, 13–25. https://doi.org/10.1196/annals.1355.003
Walsh, J.B., 1995. How often do duplicated genes evolve new functions? Genetics 139, 421–428.
Weinert, L.A., Werren, J.H., Aebi, A., Stone, G.N., Jiggins, F.M., 2009. Evolution and diversity of Rickettsia bacteria. BMC Biol. 7, 6. https://doi.org/10.1186/1741-7007-7-6

Wixon, J., 2001. Featured organism: reductive evolution in bacteria: Buchnera sp., Rickettsia prowazekii and Mycobacterium leprae. Comp. Funct. Genomics 2, 44–48.
Wolf, Y.I., Koonin, E.V., 2013. Genome reduction as the dominant mode of evolution: Prospects & Overviews. BioEssays 35, 829–837. https://doi.org/10.1002/bies.201300037
Woodard, A., Wood, D.O., 2011. Analysis of Convergent Gene Transcripts in the Obligate Intracellular Bacterium Rickettsia prowazekii. PLoS ONE 6, e16537. https://doi.org/10.1371/journal.pone.0016537
Wu, M., Sun, L.V., Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J.C., McGraw, E.A., Martin, W., Esser, C., Ahmadinejad, N., Wiegand, C., Madupu, R., Beanan, M.J., Brinkac, L.M., Daugherty, S.C., Durkin, A.S., Kolonay, J.F., Nelson, W.C., Mohamoud, Y., Lee, P., Berry, K., Young, M.B., Utterback, T., Weidman, J., Nierman, W.C., Paulsen, I.T., Nelson, K.E., Tettelin, H., O'Neill, S.L., Eisen, J.A., 2004. Phylogenomics of the Reproductive Parasite Wolbachia pipientis wMel: A Streamlined Genome Overrun by Mobile Genetic Elements. PLoS Biol. 2, e69. https://doi.org/10.1371/journal.pbio.0020069
Xin, D., El Karkouri, K., Robert, C., Raoult, D., Fournier, P.-E., 2012. Genomic Comparison of Rickettsia honei Strain RBT and Other Rickettsia Species. J. Bacteriol. 194, 4145. https://doi.org/10.1128/JB.00802-12

614 **Table 1: Classification, diseases, vectors and geographic distribution of *Rickettsia* species with known pathogenicity for humans.**

Rickettsial group	Species	Rickettsiosis	Vector	Geographic distribution
Ancestral group	*R. bellii*	Unknown pathogenesis	*Dermacentor varabilis*	
	R. canadensis	Unknown pathogenesis	*Haemaphysalis leporis-palustris*	
Typhus group	*R. prowazekii*	Epidemic typhus; Brill-Zinsser disease	*Pediculus humanus corporis; flying squirrelsectoparasites*	Africa; Mexico; Central America; South America; Eastern Europe; India; China and Afghanistan
	R. typhi	Murine typhus; Endemic typhus	*Fleas : Xenopsylla cheopis; Ctenocephalides feli; Leptosylla segnis*	USA; Mediterranean area; Asia; Africa
Spotted fever group	*R. aeschlimannii*	Rickettsiosis	*Hyalomma m. sp.*	South Africa; Morocco; Mediterranean littoral
	R. africae	African tick-bite fever	*Amblyomma variegatum; A. hebraum*	Sud-Saharan Africa; West Indies

R. conorii	Mediterranean spotted fever;Israeli spotted fever; Astrakhan fever; Indian tick typhus	*Rhipicephalus sanguineus; R. pumilio*	North Caspian Region of Russia; Southern Europe; Africa; South Asia; South Europe and Middle East
R. heilongjiangensis	Far Eastern tick borne rickettsiosis	*Dermacentor silvarum*	Far East of Russia; Northern China; eastern Asia
R. honei	Flinders Island spotted fever;Thai tick typhus	*Aponomma hydrosauri; Ixodes granulatus*	Australia; Thailand
R. japonica	Japanese spotted fever or Oriental spotted fever	*Haemaphysalis sp.; Ixodes ovatus*	Japon
R. massiliae	Mediterranean spotted fever	*Rhipicephalus turanicus; R. sanguineus*	France; Grece, Spain; Portugal; Swizerland, Silicity; Central Africa and Mali
R. parkeri	Unnamed rickettsiosis	*Amblyomma maculatum*	North and South America
R. raoultii	scalp eschar and neck lymphadenopathy (SENLAT)	*Dermacentor sivarum*	France; Spain; Croatia; Russia and Kazakhstan

R. rickettsii	Rocky Mountain spotted fever	*Dermacentorandersoni; D. variabilis; Amblyomma cajennense; Rhipicephalus sanguineus*	North; Central and South America
R. sibirica	North Asian tick typhus; Siberian tick typhus; Lymphangitis-associated rickettsiosis	*Dermacentor nuttallii; D. sinicus; D. marginatus; D. silvatum; D. pictus; D. auratus; Hyalomma asiaticum; H. truncatum*	Siberia and Far East, Asiatic; Russia; South Africa; Southern France; Grece, Spain; Portugal; Egypt
R. slovaca	scalp eschar and neck lymphadenopathy (SENLAT)	*Dermacentor marginatu; D. reticulates*	Southern and eastern Europe; Asia
R. akari	Rickettsialpox	*Allodermanyssus sanguineus*	Countries of the former Sovet Union; South Africa; Korea; Turkey; Balkan countries; North and South America
R. australis	Queensland tick typhus	*Ixodes holocyclus*	Australia; Tasmania

R. felis	Flea-borne spotted fever	*Ctenocephalides felis; Liposcelis botrychopila*	Europe; North and South America; Africa; Asia
R. helvetica	Aneruptive fever/Unnamed rickettsiosis	*Ixodes ricinus*	Central and Northern Europe; Asia

615

616 **Table 2: Main characteristics of avalaible rickettsial genomes in GenBank**

Species	Strain	Genome size (Mb)	G+C content (%)	Protein-coding genes	Plasmids	% non-coding sequences	Chromosome accession number
R. aeschlimannii	MC16	1.31	32.2	1051	Plasmid 1, Plasmid 2	-	CCER01000000
R. africae	ESF-5	1.28	32.4	1219	pRaf	21.74	CP001612
R. akari	Hartford	1.23	32.3	1259		22.6	CP000847
R. amblyommatis	Ac37	1.46	32.4	1511	pRAMAC18 pRAMAC23	-	NZ_CP012420
R. amblyommatis	AcPa	1.44	32.4	1123		-	LANR01000001
R. amblyommatis	Darkwater	1.44	32.8	1060		-	LAOH01000001
R. amblyommatis	GAT-30V	1.48	32.4	1550	pMCE1 pMCE2 pMCE3	-	NC_017028
*R. argasii**	T170-B	1.44	32.3	1187		-	LAOQ01000006
R. asembonensis	NMRCii	1.36	32.3	1212	pRAS01	-	JWSW01000001

R. australis	Phillips	1.32	32.2	1099	pRau01	-	AKVZ01000001
R. australis	Cutlack	1.33	32.3	1136	pMC5_1	-	NC_017058
R. bellii	RMLAn4	1.54	31.6	1311		-	LAOI01000001
R. bellii	RMLMog	1.62	31.5	1336		-	LAOJ01000001
R. bellii	OSU 85-389	1.52	31.6	1476		-	NC_009883
R. bellii	RML369-C	1.52	31.7	1429		14.8	NC_007940
R. endosymbiont of	REIS	1.82	33.0	2309	pReis1 pReis2	-	CM000770
Ixodes scapularis					pReis3 pReis4		
R. canadensis	CA410	1.15	31.1	1016		-	NC_016929
R. canadensis	McKiel	1.16	31.1	902		24.8	NC_009879
R. conorii	Malish 7	1.27	32.4	1227		18.5	NC_003103
R. conorii	A-167	1.26	32.5	1210		-	AJUR01000001
R. conorii	ITTR	1.25	32.4	1157		-	AJHC01000001
R. conorii	ISTT CDC1	1.25	32.5	1200		-	AJVP01000001
R. endosymbiont of	Humboldt	1.56	32.2	1294		-	LAOP01000001
*Ixodes pacificus**							

R. felis	LSU	1.54	32.4	1970	pRF	-	JSEM01000001
R. felis	LSU lb	1.58	32.4	1691	pRF pLbaR	-	JSEL01000001
R. felis	Pedreira	1.49	32.5	1594		-	LANQ01000001
R. felis	URRWXCal2	1.49	32.5	1444	pRF pRFδ	16.4	NC_007109
R. gravesii	BWI-1	1.37	32.2	1158	pRgr	-	AWXL01000001
R. heilongjiangensis	*O54*	1.28	32.3	1140		-	CP002912
R. helvetica	*C9P9*	1.37	32.2	1114	pRhe	-	CM001467
R. honei	*RB*	1.27	32.4	1171		-	AJTT01000001
R. hoogstraalii	*Croatica*	1.48	32.4	1250		-	CCXM01000001
R. hoogstraalii	RCCE3	2.3	32.4	2479		-	LAOB01000001
R. japonica	*YH*	1.28	32.4	1142		-	NC_016050
R. massiliae	AZT80	1.28	32.5	1207	pRmaB	-	NC_016931
R. massiliae	MTU5	1.37	32.5	1152	pRma	-	NC_009900
*R. monacensis**	IrR/Munich	1.35	32.4	1447	pRM	-	NZ_LN794217
R. montanensis	*OSU 85-930*	1.28	32.6	1125		-	CP003340
R. parkeri	AT#24	1.3	32.4	1226		-	LAOL01000001

R. parkeri	GrandBay	1.31	32.4	1223		-	LAOK01000001
R. parkeri	Portsmouth	1.3	32.4	1228		-	NC_017044
R. parkeri	TatesHell	1.3	32.4	1227		-	LAOO01000001
R. peacockii	*Rustic*	1.29	32.6	927	pRpe	-	CP001227
*R. philipii**	364D	1.29	32.5	1218		-	CP003308
R. prowazekii	Breinl	1.11	29	842		-	NC_020993
R. prowazekii	BuV67-CWPP	1.11	29	843		-	NC_017056
R. prowazekii	Cairo3	1.11	29	842		-	APMO01000001
R. prowazekii	Chernikova	1.11	29	845		-	NC_017049
R. prowazekii	Dachau	1.11	29	839		-	NC_017051
R. prowazekii	GvV257	1.11	29	829		-	NC_017048
R. prowazekii	Katsinyian	1.11	29	844		-	NC_017050
R. prowazekii	Madrid E	1.11	29	834		24.6	NC_000963
R. prowazekii	NMRC Madrid E	1.11	29	830		-	NC_020992
R. prowazekii	Rp22	1.11	29	864		23.8	NC_017560

R. prowazekii	RpGvF24	1.11	29	870		-	NC_017057
R. raoultii	Khabarovsk	1.34	32.8	1334	pRa1 pRa2 pRa3 pRa4	-	CP010969
R. rhipicephali	3-7-female6-CWPP	1.31	32.4	1117	pRrh	-	NC_017042
R. rhipicephali	Ect	1.27	32.6	1067		-	LAOC01000001
R. rhipicephali	HJ#5	1.45	32.3	1200	pHJ51 pHJ52	-	NZ_CP013133
R. rickettsii	Arizona	1.27	32.4	1343		-	NC_016909
R. rickettsii	Brazil	1.25	32.4	1339		-	NC_016913
R. rickettsii	Colombia	1.27	32.4	1342		-	NC_016908
R. rickettsii	Hauke	1.27	32.4	1347		-	NC_016911
R. rickettsii	Hino	1.27	32.4	1346		-	NC_016914
R. rickettsii	Hlp#2	1.27	32.4	1339		-	NC_016915
R. rickettsii	Iowa	1.27	32.4	1384		-	NC_010263
R. rickettsii	Morgan	1.27	32.4	1343		-	NZ_CP006010
R. rickettsii	R	1.26	32.4	1334		-	NZ_CP006009

R. rickettsii	Sheila Smith	1.26	32.5	1345		21.5	NC_009882
R. sibirica	246	1.25	32.5	1227		22.2	AABW01000001
R. sibirica	HA-91	1.25	32.4	1175		-	AHZB01000001
R. sibirica	BJ-90	1.25	32.4	1217		-	AHIZ01000001
R. slovaca	D-CWPP	1.27	32.5	1261		-	NC_017065
R. slovaca	13-B	1.27	32.5	1260		-	NC_016639
R. tamurae	AT-1	1.44	32.4	1200	Plasmid 1 Plasmid 2	-	CCMG01000008
R. typhi	B9991CWPP	1.11	28.9	819		-	NC_017062
R. typhi	TH1527	1.11	28.9	819		-	NC_017066
R. typhi	Wilmington	1.11	28.9	817		23.7	NC_006142

[a]Species with as yet no standing in nomenclature are written with quotation marks, (-) = no avalaible data

Figure 1: Phylogenetic tree of 31 *Rickettsia* species based on 591 concatenated core proteins using the Maximum Likehood method with JTT and GAMMA models and display only topology. Values at the nodes are percentages. Numbers at the nodes represent the percentages of bootstrap values obtained by repeating the analysis 500 times to generate a majority consensus tree. Only values greater than 70 % were reported.

Figure 2: Genomic alignment showing the high degree of conserved synteny between *Rickettsia* spp. The figure was generated using the Mauve rearrangement viewer (Darling et al., 2004). It shows a linear representation of the genomes of *R. bellii* RML369-C, *R. felis* URRWXCal2, *R. africae* ESF-5, *R. conorii* Malish7, *R. massiliae* MTU5, *R. raoultii Khabarovsk*, *R. prowazekii* Madrid E, and *R. typhi* Wilmington. The size of the horizontal bars corresponds to genome size (Kb)

Figure 3: Phylogenomic tree based on 600 core proteins and, pathogenic and genomic features, of ten mild to highly pathogenic *Rickettsia* species. Bootstrap values greater than 90% are shown at the nodes. All data presented in the Figure were deduced from the following references (Andersson et al., 1998; G. Blanc et al., 2007; Guillaume Blanc et al., 2007; El Karkouri et al., 2017, 2016; Fournier et al., 2009; McLeod et al., 2004; Ogata, 2001; Ogata et al., 2006, 2005). NA = data not available.

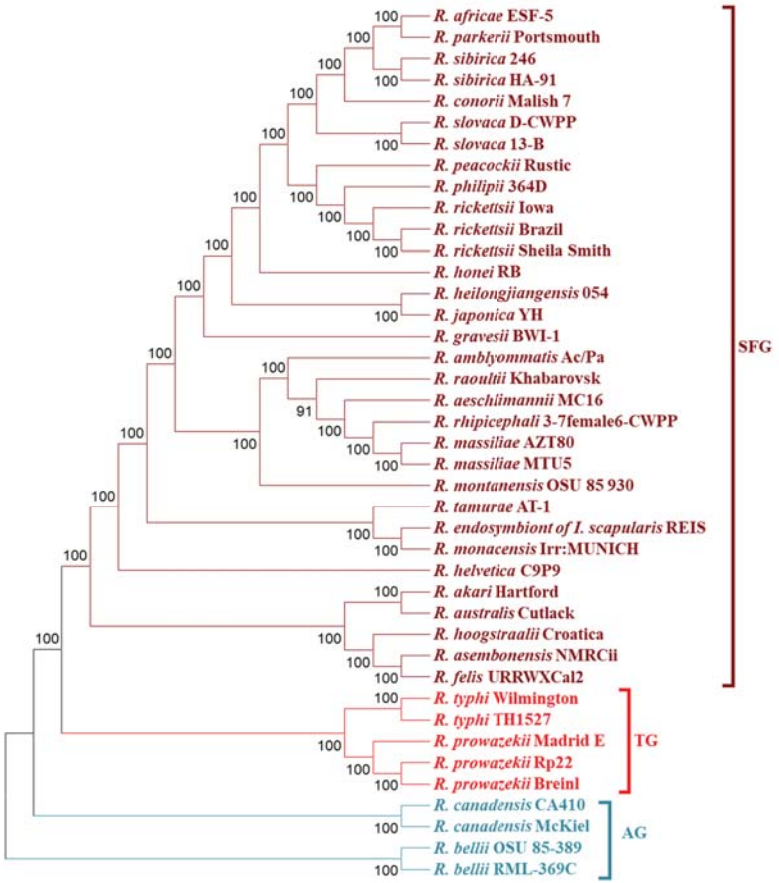


Figure 1: Phylogenetic tree of 31 *Rickettsia* species based on 591 concatenated core proteins using the Maximum Likehood method with JTT and GAMMA models and display only topology. Values at the nodes are percentages. Numbers at the nodes represent the percentages of bootstrap values obtained by repeating the analysis 500 times to generate a majority consensus tree. Only values greater than 70 % were reported.

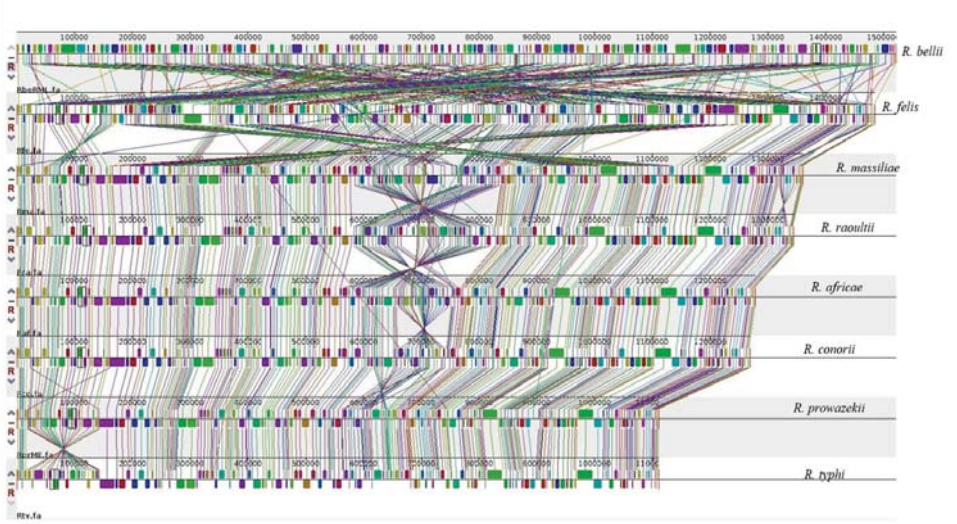
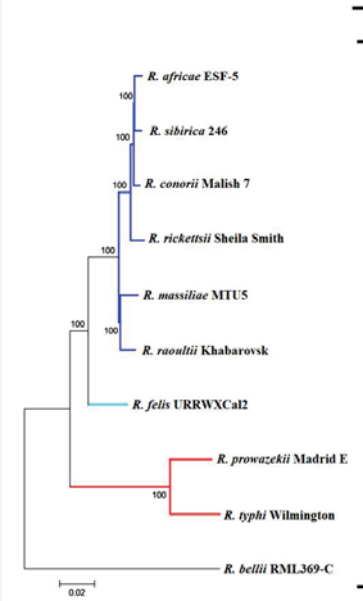


Figure 2: Genomic alignment showing the high degree of conserved synteny between *Rickettsia* spp. The figure was generated using the Mauve rearrangement viewer (Darling et al., 2004). It shows a linear representation of the genomes of *R. bellii* RML369-C, *R. felis* URRWXCal2, *R. africae* ESF-5, *R. conorii* Malish7, *R. massiliae* MTU5, *R. raoultii Khabarovsk*, *R. prowazekii* Madrid E, and *R. typhi* Wilmington. The size of the horizontal bars corresponds to genome size (Kb)



Virulence	Size (bp)	protein-coding genes	% coding sequences	RNAs	pseudogenes	RPEs
Mild	1,278,540 pRaf:12,377	1112	78.26	39	246	460
Virulent	1,250,021	1083	77.76	36	200	NA
Virulent	1,268,755	1374	81.5	39	252	559
Virulent	1,257,710	1345	NA	36	NA	NA
Mild	1,360,898 pRma:15,286	968	69	39	286	562
Mild	1,344,605 pRra1:20,840 pRra2:83,219 pRra3:34,583	1180	NA	39	339	NA
Mild	1,485,148 pRF:6,282 pRFδ:39,268	1444	83.8	39	130	726
Highly virulent	1,111,523	834	76.2	39	181	120
Virulent	1,111,496	838	76.3	39	185	121
Unknown	1,522,076	1429	82.5	40	100	526

Figure 3: Phylogenomic tree based on 600 core proteins and, pathogenic and genomic features, of ten mild to highly pathogenic *Rickettsia* species. Bootstrap values greater than 90% are shown at the nodes. All data presented in the Figure were deduced from the following references (Andersson et al., 1998; G. Blanc et al., 2007; Guillaume Blanc et al., 2007; El Karkouri et al., 2017, 2016; Fournier et al., 2009; McLeod et al., 2004; Ogata, 2001; Ogata et al., 2006, 2005). NA = data not available.

Article 2:

Rickettsial genomics and the paradigm of genome reduction associated with increased virulence

Awa Diop, Didier Raoult, Pierre-Edouard Fournier

[Published in Microbes and Infections]

ELSEVIER
Institut Pasteur
Microbes and Infection xx (2017) 1−9
www.elsevier.com/locate/micinf

Rickettsial genomics and the paradigm of genome reduction associated with increased virulence

Awa Diop [a], Didier Raoult [a,b], Pierre-Edouard Fournier [a,*]

[a] *Aix-Marseille Université, Institut Hospitalo-Universitaire Méditerranée-Infection, URMITE, UM63, CNRS 7278, IRD 198, Inserm U1095, Assistance Publique-Hôpitaux de Marseille, 19-21 Boulevard Jean Moulin, 13005 Marseille, France*
[b] *Campus International UCAD-IRD, Dakar, Senegal*

Received 22 August 2017; accepted 15 November 2017
Available online ■ ■ ■

Abstract

Rickettsia species are arthropod endosymbiotic α-proteobacteria that can infect mammalian hosts during their obligate intracellular lifecycle, and cause a range of mild to severe diseases in humans. Paradoxically, during their adaptation to a bottleneck lifestyle, rickettsial genomes have undergone an evolution marked by a progressive chromosomic and plasmidic degradation resulting in a genome reduction from 1.5 to 1.1 Mb, with a coding capacity of 69−84%. A striking finding of rickettsial genomics has been that the most virulent species had genomes that were drastically reduced and degraded when compared to closely related less virulent or nonpathogenic species. This paradoxical evolution, which is not unique to members of the genus *Rickettsia* but has been identified as a convergent evolution of several major human pathogenic bacteria, parallels a selected loss of genes associated with transcriptional regulators, but with a high preservation of toxin-antitoxin (TA) modules and recombination and DNA repair proteins. In addition, these bacteria have undergone a proliferation of genetic elements, notably short palindromic elements, whose role remains unknown. Recent proteomic and transcriptomics analyses have revealed a differential level or degradation of gene expression that may, at least partially, explain differences in virulence among *Rickettsia* species. However, future investigations are mandatory to provide novel insights into the mechanisms by which genomic reductive evolution contributes to an emergence of pathogenesis.

Keywords: Rickettsial genomics; Reductive evolution; Virulence; Pathogenesis

1. Introduction

The genus *Rickettsia* (order *Rickettsiales*, family *Rickettsiaceae*) is currently made of obligate intracellular α-proteobacteria mostly associated to arthropods worldwide [1,2]. These bacteria can also infect mammalian hosts, mostly through arthropod bites or feces, and were initially phylogenetically classified into two major groups on the basis of their pathogenicity for humans. These groups include the spotted fever group (SFG) that currently contains 16 pathogenic agents (Table 1) causing spotted fevers, as well as numerous species of as-yet unknown pathogenicity, associated with ticks, fleas and mites; and the typhus group (TG) that is made of *Rickettsia prowazekii* and *Rickettsia typhi* causing typhus and associated with human body lice and rat fleas, respectively. However, the SFG group was later demonstrated to be divided into distinct phylogenetic subgroups on the basis of gene sequence comparisons [3−5]. *Rickettsia* species cause a range of mild to severe diseases, the most common being scalp eschar and neck lymphadenopathy (SENLAT), also named tick-borne lyphadenopathy (TIBOLA) or Dermacentor-borne necrosis, erythema and lymphadenopathy (DEBONEL), murine typhus, Mediterranean spotted fever (MSF), Rocky Mountain spotted fever (RMSF), and epidemic typhus [6−8]. In addition to spotted fever and typhus group rickettsiae, two species, *Rickettsia bellii* and *Rickettsia canadensis*, associated with ticks but not causing to date any recognized human

* Corresponding author. Fax: +33 413 732 402.
E-mail address: pierre-edouard.fournier@univ-amu.fr (P.-E. Fournier).

https://doi.org/10.1016/j.micinf.2017.11.009
1286-4579/

Table 1
Classification, vectors, reservoirs and diseases of *Rickettsia* species with known pathogenicity to humans.

Antigenic group	Species	Strain_name	Rickettsiosis	Vector	Reservoirs
Spotted fever group	*R. aeschlimannii*	MC16	Rickettsiosis	Ticks: *Hyalomma m.* sp.	Unknown
	R. africae	ESF-5	African tick-bite fever	Ticks: *Amblyomma variegatum*	Ruminants
	R. akari	Hartford	Rickettsialpox	Mites: *Allodermanyssus sanguineus*	Mice, rodents
	R. australis	Cutlack	Queensland tick typhus	Ticks: *Ixodes* sp.	Rodents
	R. conorii	Malish 7	Mediterranean spotted fever	Ticks: *Rhipicephalus* sp.	Dogs, rodents
	R. felis	URRWXCal2	Flea spotted fever	Flea: *Ctenocephalides felis*	Cats, rodents, opossums
	R. heilongjiangensis	*O54*	Far Eastern tick borne rickettsiosis	Ticks: *Dermacentor silvarum*	Rodents
	R. helvetica	*C9P9*	Aneruptive fever/Unnamed rickettsiosis	Ticks: *Ixodes ricinus*	Rodents
	R. honei	*RB*	Flinders Island spotted fever, Thai tick typhus	Ticks: *Aponomma hydrosauri*	Rodents, reptiles
	R. japonica	*YH*	Japanese spotted fever or Oriental spotted fever	Ticks: *Haemaphysalis* sp.	Rodents
	R. massiliae	MTU5	Mediterranean spotted fever	Tck: *Rhipicephalus turanicus*	Unknown
	R. parkeri	Portsmouth	Unnamed rickettsiosis	Ticks	Rodents
	R. raoultii	Khabarovsk	scalp eschar and neck lymphadenopathy (SENLAT)	Ticks: *Dermacentor sivarum*	Unknown
	R. rickettsii	Sheila Smith	Rocky Mountain spotted fever	Ticks: *Dermacentor* sp.	Rodents
	R. sibirica	246	North Asian tick typhus, Siberian tick typhus	Ticks: *Dermacentor* sp	Rodents
	R. sibirica	HA-91	Lymphangitis-associated rickettsiosis	Ticks: *Dermacentor* sp.	Rodents
	R. slovaca	13-B	scalp eschar and neck lymphadenopathy (SENLAT)	Ticks: *Dermacentor* sp.	Lagomorphes, rodents
Typhus group	*R. prowazekii*	Breinl	Epidemic typhus, Brill-Zinsser disease	Louse: *Pediculus humanus*	Humans, flying squirrels
	R. prowazekii	Rp22	Epidemic typhus	Louse: *Pediculus humanus*	Humans, flying squirrels
	R. typhi	Wilmington	Murine typhus	Fleas: *Xenopsylla cheopis*	Rodents

disease, diverged early from these two groups. Furthermore, recent studies have reported the association of other *Rickettsia* lineages with other reservoirs including protozoa, algae, leeches plants or insects [4,9,10].

In 1995, the complete genome sequencing of *Haemophilus influenzae* (the first sequenced genome) [11] marked the beginning of the genomic era. Over the past two decades, the completion of the genome sequences of most *Rickettsia* species, starting with that of *R. prowazekii*, allowed better knowledge about the molecular mechanisms involved in their pathogenicity [12] (see Fig. 1).

2. Characteristics and genome architecture of *Rickettsia* species

Rickettsia species have genome sizes ranging from 1.1 to 2.3 Mbp and exhibit a G + C content of 29–33% (Table 2). *Rickettsia hoogstraalii* and *Rickettsia* endosymbiont of *Ixodes scapularis* [13] have the largest genomes sequenced to date but exhibit no known pathogenic effects. Rickettsial genomes are also characterized by a high degree of synteny (Fig. 2) [4] despite the presence of numerous pseudogenes and a large fraction of non-coding DNA, reaching 24% in *R. prowazekii* [14,15]. This genomic degradation likely results from their endosymbiotic lifestyle that has allowed them to discard genes involved in metabolisms supplied by their eukaryotic host cells [16,17]. This genomic downsizing has occurred through a progressive gene degradation, from complete functional genes to functional pseudogenes to non functional pseudogenes to gene remnants to discarded genes [18–21]. Generally, rickettsial genomes are typical of those of symbiotic bacteria, which are obligate intracellular and are characterized by a reduced genome, relatively small, made of a single circular chromosome, evolving slowly, and maintaining a near perfect colinearity between species [22]. However, in parallel to this reduction phenomenon, rickettsial genomes exhibit a paradoxical expansion of genetic elements, including plasmids, short palindromic elements named rickettsia palindromic elements (RPEs) [23], ankyrin and tetratricopeptide repeats, toxin-antitoxin modules, ADP-ATP translocases, type IV secretion system (T4SS), as well as *sca*, *spoT*, *proP* and *ampG* genes. Moreover, the presence of plasmids in *Rickettsia* genomes was first detected in *Rickettsia felis*, demonstrating that these bacteria were able to exchange genetic material by conjugation, a mechanism that was thought to be absent from obligate intracellular and allopatric bacteria [16,17,24]. To date, 20 plasmids have been identified in 11 species, some species having 1 to 4 distinct plasmids [25–27].

Table 2
Main characteristics of rickettsial genomes available in Genbank.

Species	Strain	Genome size (Mbp)	G + C content (%)	Presence of plasmid (s)	Protein-coding genes	% coding sequences	Rickettsia palindromic elements	Chromosome accession number
R. aeschlimannii	MC16	1.31	32.2	Plasmid 1, Plasmid 2	1051	–	–	CCER01000000
R. africae	ESF-5	1.28	32.4	pRaf	1219	78.26	–	CP001612
R. akari	Hartford	1.23	32.3		1259	77.4	–	CP000847
R. amblyommatis	Ac37	1.46	32.4	pRAMAC18, pRAMAC23	1511	–	–	NZ_CP012420
R. amblyommatis	AcPa	1.44	32.4		1123	–	–	LANR01000001
R. amblyommatis	Darkwater	1.44	32.8		1060	–	–	LAOH01000001
R. amblyommatis	GAT-30V	1.48	32.4	pMCE1, pMCE2, pMCE3	1550	–	–	NC_017028
"*R. argasii*"	T170-B	1.44	32.3		1187	–	–	LAOQ01000006
R. asembonensis	NMRCii	1.36	32.3	pRAS01	1212	–	–	JWSW01000001
R. australis	Phillips	1.32	32.2	pRau01	1099	–	–	AKVZ01000001
R. australis	Cutlack	1.33	32.3	pMC5_1	1136	–	–	NC_017058
R. bellii	RMLAn4	1.54	31.6		1311	–	–	LAOI01000001
R. bellii	RMLMog	1.62	31.5		1336	–	–	LAOJ01000001
R. bellii	OSU 85-389	1.52	31.6		1476	–	–	NC_009883
R. bellii	RML369-C	1.52	31.7		1429	85.2%	525	NC_007940
R. endosymbiont of Ixodes scapularis	REIS	1.82	33.0	pReis1, pReis2, pReis3, pReis4	2309	–	–	CM000770
R. canadensis	CA410	1.15	31.1		1016	–	–	NC_016929
R. canadensis	McKiel	1.16	31.1		902	75.2%	–	NC_009879
R. conorii	Malish 7	1.27	32.4		1227	81.5	559	NC_003103
R. conorii	A-167	1.26	32.5		1210	–	–	AJUR01000001
R. conorii	ITTR	1.25	32.4		1157	–	–	AJHC01000001
R. conorii	ISTT CDC1	1.25	32.5		1200	–	–	AJVP01000001
R. endosymbiont of *Ixodes pacificus*	Humboldt	1.56	32.2		1294	–	–	LAOP01000001
"*R. felis*"	LSU	1.54	32.4	pRF	1970	–	–	JSEM01000001
"*R. felis*"	LSU lb	1.58	32.4	pRF, pLbaR	1691	–	–	JSEL01000001
"*R. felis*"	Pedreira	1.49	32.5		1594	–	–	LANQ01000001
"*R. felis*"	URRWXCal2	1.49	32.5	pRF, pRFδ	1444	83.6%	726	NC_007109
R. gravesii	BWI-1	1.37	32.2	pRgr	1158	–	–	AWXL01000001
R. heilongjiangensis	*O54*	1.28	32.3		1140	–	–	CP002912
R. helvetica	*C9P9*	1.37	32.2	pRhe	1114	–	–	CM001467
R. honei	*RB*	1.27	32.4		1171	–	–	AJTT01000001
R. hoogstraalii	*Croatica*	1.48	32.4		1250	–	–	CCXM01000001
R. hoogstraalii	RCCE3	2.3	32.4		2479	–	–	LAOB01000001
R. japonica	*YH*	1.28	32.4		1142	–	–	NC_016050
R. massiliae	AZT80	1.28	32.5	pRmaB	1207	–	–	NC_016931
R. massiliae	MTU5	1.37	32.5	pRma	1152	–	565	NC_009900
"*R. monacensis*"	IrR/Munich	1.35	32.4	pRM	1447	–	–	NZ_LN794217
R. montanensis	*OSU 85-930*	1.28	32.6		1125	–	–	CP003340
R. parkeri	AT#24	1.3	32.4		1226	–	–	LAOL01000001
R. parkeri	GrandBay	1.31	32.4		1223	–	–	LAOK01000001
R. parkeri	Portsmouth	1.3	32.4		1228	–	–	NC_017044
R. parkeri	TatesHell	1.3	32.4		1227	–	–	LAOO01000001
R. peacockii	*Rustic*	1.29	32.6	pRpe	927	–	–	CP001227
"*R. philipii*"	364D	1.29	32.5		1218	–	–	CP003308
R. prowazekii	Breinl	1.11	29		842	–	–	NC_020993
R. prowazekii	BuV67-CWPP	1.11	29		843	–	–	NC_017056
R. prowazekii	Cairo3	1.11	29		842	–	–	APMO01000001
R. prowazekii	Chernikova	1.11	29		845	–	–	NC_017049
R. prowazekii	Dachau	1.11	29		839	–	–	NC_017051
R. prowazekii	GvV257	1.11	29		829	–	–	NC_017048
R. prowazekii	Katsinyian	1.11	29		844	–	–	NC_017050
R. prowazekii	Madrid E	1.11	29		834	75.4%	120	NC_000963
R. prowazekii	NMRC Madrid E	1.11	29		830	–	–	NC_020992
R. prowazekii	Rp22	1.11	29		864	76.2%	–	NC_017560
R. prowazekii	RpGvF24	1.11	29		870	–	–	NC_017057
R. raoultii	Khabarovsk	1.34	32.8	pRa1, pRa2, pRa3, pRa4	1334	–	–	CP010969
R. rhipicephali	3-7-female 6-CWPP	1.31	32.4	pRrh	1117	–	–	NC_017042
R. rhipicephali	Ect	1.27	32.6		1067	–	–	LAOC01000001
R. rhipicephali	HJ#5	1.45	32.3	pHJ51, pHJ52	1200	–	–	NZ_CP013133

(*continued on next page*)

Table 2 (*continued*)

Species	Strain	Genome size (Mbp)	G + C content (%)	Presence of plasmid (s)	Protein-coding genes	% coding sequences	Rickettsia palindromic elements	Chromosome accession number
R. rickettsii	Arizona	1.27	32.4		1343	–	–	NC_016909
R. rickettsii	Brazil	1.25	32.4		1339	–	–	NC_016913
R. rickettsii	Colombia	1.27	32.4		1342	–	–	NC_016908
R. rickettsii	Hauke	1.27	32.4		1347	–	–	NC_016911
R. rickettsii	Hino	1.27	32.4		1346	–	–	NC_016914
R. rickettsii	Hlp#2	1.27	32.4		1339	–	–	NC_016915
R. rickettsii	Iowa	1.27	32.4		1384	–	–	NC_010263
R. rickettsii	Morgan	1.27	32.4		1343	–	–	NZ_CP006010
R. rickettsii	R	1.26	32.4		1334	–	–	NZ_CP006009
R. rickettsii	Sheila Smith	1.26	32.5		1345	78.5%	–	NC_009882
R. sibirica	246	1.25	32.5		1227	77.8%	–	AABW01000001
R. sibirica	HA-91	1.25	32.4		1175	–	–	AHZB01000001
R. sibirica	BJ-90	1.25	32.4		1217	–	–	AHIZ01000001
R. slovaca	D-CWPP	1.27	32.5		1261	–	–	NC_017065
R. slovaca	13-B	1.27	32.5		1260	–	–	NC_016639
R. tamurae	AT-1	1.44	32.4	Plasmid 1, Plasmid 2	1200	–	–	CCMG01000008
R. typhi	B9991CWPP	1.11	28.9		819	–	–	NC_017062
R. typhi	TH1527	1.11	28.9		819	–	–	NC_017066
R. typhi	Wilmington	1.11	28.9		817	76.3%	121	NC_006142

Species with as yet no standing in nomenclature are written with quotation marks (–) = no available data.

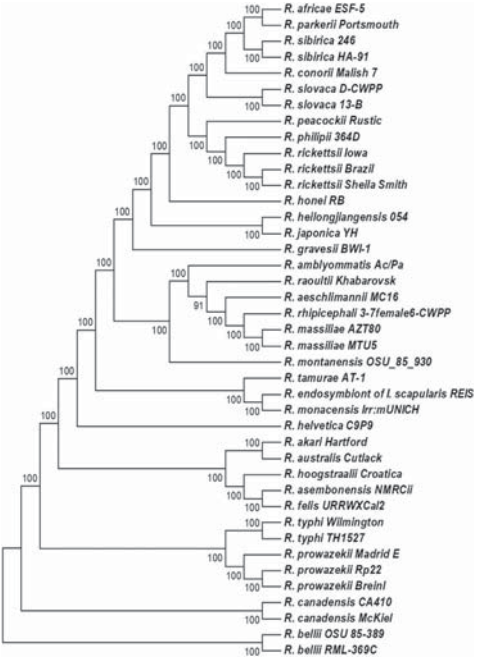


Fig. 1. Phylogenetic tree of 31 *Rickettsia* species with validly published names based on the alignment of 450 concatenated core proteins using the Maximum Likelihood method with JTT and GAMMA models and display only topology. Values at the nodes represent the percentages of bootstrap values obtained by repeating analysis 500 times to generate a majority consensus tree. Only values greater than 70% were indicated.

3. Comparative analysis of rickettsial genomes

The first genomic comparison of *Rickettsia* species was that of the first two sequenced genomes from *Rickettsia conorii* and *R. prowazekii* [22]. This study showed a near perfect colinearity between both species (Fig. 2) but the latter species had a smaller genome and a higher proportion of non coding DNA, including many pseudogenes. Further comparisons confirmed this trend in genomic reduction (1.5–1.1 Mb, coding capacity 69–84%) through progressive gene degradation until complete disappearance [28]. Degraded genes include mostly those coding for amino-acid, ATP, LPS and cell wall component biosynthesis [14,22,29].

Comparative genomic analysis of *Rickettsia* species revealed variations in chromosome size and plasmid number and size (Table 2), despite a common ongoing reductive evolution [30] by progressive gene loss and concomitant gene gain by gene duplication, proliferation of RPEs and horizontal gene transfer [4]. Gene family duplication is frequent in rickettsial genomes and is thought to enable adaptation to environmental changes in the host. The two most duplicated genes encode ADP/ATP translocases, often found in several copies and enabling energy exploitation produced by host cells [29,31], and *spoT* genes found in 4–14 copies and involved in the microbial response to environmental stress [24,29,32]. Other duplicated gene families include proline/betaine transporters, toxin/antitoxin modules, T4SS, *sca* and *ampG* involved in rickettsial pathogenesis as previously described [17,22,29,33]. Rickettsiae possess an incomplete P-T4SS system that is characterized by the lack of *virB5* but the duplication of the *virB4*, *virB6*, *virB8* and *virB9* genes [34]. Surface cell antigen (Sca) proteins are a family of 17 orthologous autotransporters diversely detected in all rickettsial genomes [33]. They were demonstrated to be localized at the

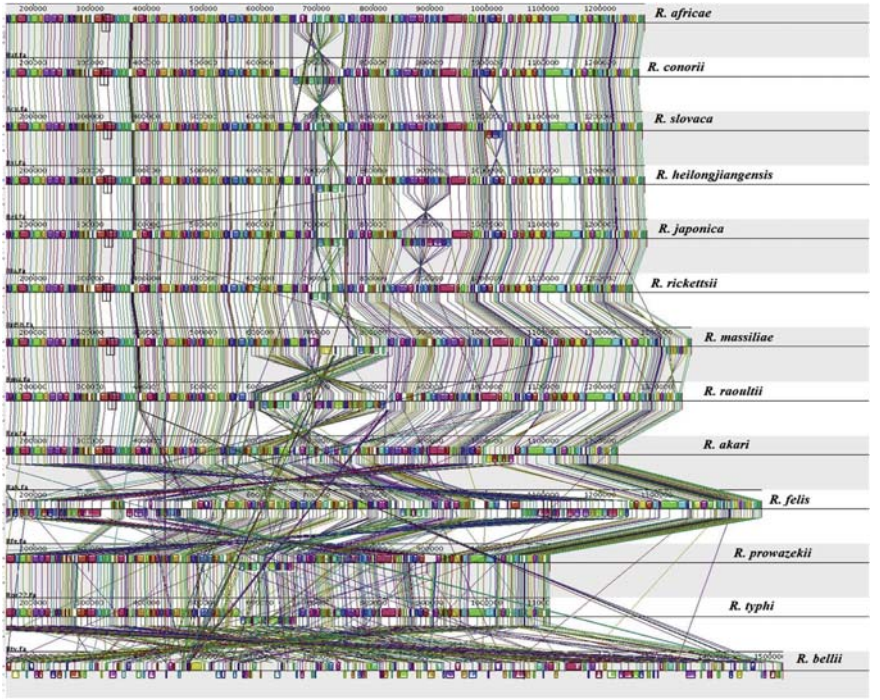


Fig. 2. Genomic alignment showing the high degree of conserved genomic synteny between Rickettsia species.

surface of bacteria and play roles in mammalian cell infection as well as infecting their arthropod hosts' cells, notably by promoting actin-based motility [35], In addition, many repetitive elements are distributed in intergenic regions (tandem repeats) [36] or both intergenic and coding regions (RPEs) [37,38]. RPEs are at least five times more numerous in SFG rickettsia than in TG rickettsia (Table 2). They are assumed to play a role in the evolution of rickettsial genomes by promoting the emergence of new proteins [39]. Twenty-two copies of ankyrin and 11 copies of tetratricopeptide repeats (TPR-repeat), frequently found in endosymbionts [40−43] are found in *R. felis* [44]. Finally, plasmids are less abundant in virulent than less virulent species [8,24,45]. They were most likely acquired vertically from *Orientia*/*Rickettsia* chromosome ancestors [27]. The genome from REIS, the largest rickettsial genome to date, is characterized by a remarkable proliferation of mobile genetic elements (35% of the entire genome) including a RAGE module considered as a genetic exchange facilitators [46] and resulting from multiplied genomic invasion events [13]. It was also described in *Orientia tsutsugamushi*, *Rickettsia massiliae* [25], *R. bellii* [47] and in the pLbaR plasmid of *R. felis* strain LSU-Lb [48]. Several genes including tra cluster, T4SS, ADP/ATP translocases and patatin-encoding genes found in *Rickettsia* spp. are phylogenetically close to those found in many amoeba-associated bacteria, suggesting their acquisition by horizontal transfer events between *Rickettsia* and non-rickettsial bacteria [4,47].

Other lessons from rickettsial genome comparison are the identification of 15−191 small non-coding RNAs (sRNAs) in intergenic sequences, depending on species [49]. These post-transcriptional regulators are assumed to influence virulence and adaptation depending on the host niche through transcriptomic regulation [49]. Their presence may explain why early comparative studies had identified highly conserved intergenic spacers [22]. A total of 1785 sRNAs were detected from 13 species spanning all rickettsial groups, and the expression of sRNAs was demonstrated in *R. prowazekii* [49]. In addition, all five genes required for the de novo folate biosynthesis were demonstrated to be present in 15 *Rickettsia* species, including both human pathogens and non pathogens but excluding the typhus group [50].

Finally, comparative genomics at the intraspecies level enabled identification of variable situations [51]. In *Rickettsia japonica*, 31 strains from the three major lineages exhibited only 112 single nucleotide polymorphisms (SNPs) and 44 InDels, thus suggesting a long generation time in nature or a

recent clonal expansion [51]. In *R. prowazekii*, similar findings were identified, with 81 SNPs observed among 3 strains [51]. In *Rickettsia rickettsii*, the comparison of 4 strains, two eastern an two western strains showed geographic divergences but an overall high genetic homology with few differences in coding regions [52]. This study also demonstrated that the avirulent strain Iowa only diverged from virulent strains by 29 SNPs in addition to a 891-bp insertion in the *ompA* gene [52]. In contrast, the comparison of 3 *R. felis* strains, including two from cat fleas and one from book lice, demonstrated that not only was the book louse strain divergent, with a unique plasmid and SNPs occurring in intergenic regions, RPEs and conserved *Rickettsia* genes, but also were both cat flea strains which exhibited SNPs in genes associated to the *Rickettsia* mobilome [48]. These data suggest that the observed difference may result from spatial isolation for cat flea strains and host specialization in the case of the book louse strain.

4. Paradigm of genome reduction associated with increased virulence

For long, it was believed that bacteria gain virulence by the acquisition of foreign genetic material. However, the comparison of the *R. prowazekii* and *R. conorii* genomes demonstrated that the former species, which is the most virulent, has a drastically degradated genome [19]. Further studies demonstrated that, in *Rickettsia* spp., some speculated virulence factors were found in both pathogenic and nonpathogenic bacteria, and genomes from the most pathogenic species were found to have few or no additional genes when compared to closely related but lesser pathogens. In addition, no association was found between virulence and the presence of plasmids or gene acquisition [45]. *R. prowazekii*, the most pathogenic *Rickettsia* species and agent of epidemic typhus has the smallest genome and an inverse correlation exists between genome size and degree of pathogenicity [21]. These findings suggested a new paradigm in rickettsial pathogenicity that linked increased virulence to genome reductive evolution rather than virulence gene acquisition. Comparative genomics showed a loss of nonessential genes including genes coding for the amino acid synthesis and biosynthetic pathway components during reductive evolution [53]. The most virulent *R. prowazekii* has lost transcriptional regulator genes with a decreased translational capacity [54], but conserved genes coding for toxins, toxin-antitoxin (TA) modules and recombination and DNA repair proteins most likely needed for protection against host immune response [55]. In addition, recent multi-omics data showed a link between reductive evolution and differential gene expression between two virulent and two less virulent SFG rickettsiae. The two virulent *R. conorii* (MSF) and *Rickettsia slovaca* (SENLAT) agents exhibit less up-regulated than down-regulated genes and than the less virulent *R. massiliae* (MSF) and *Rickettsia raoultii* (SENLAT) agents [8]. The former two species have more reduced genomes with plasmid loss than the latter two, suggesting that reductive genomic evolution associated with increased virulence may not be only a question of presence or lack of a specific protein but may also result from differential level or degradation of expression of common proteins [8]. It was speculated that loss of regulator genes, as observed in several intracellular pathogens, is a critical cause of virulence [45].

This phenomenon was also observed in other human pathogens not genetically related to *Rickettsia* species such as *Treponema* spp., *Mycobacterium* spp. or *Yersinia* spp. [16,20,56]. As examples, *Mycobacterium leprae*, *Treponama pallidum* and *Yersinia pestis* have smaller genomes than closely related but less virulent species in their respective genera. Thus, genomic reductive evolution with alteration of the regulation of invasion, replication and transmission processes, in addition to a differential level or degradation of expression of common proteins may result in an emergence of high pathogenicity.

5. Identified virulence factors in rickettsial genomes

Predicting virulence factors from genome sequences has been among the first objectives of genomics, especially for intracellular bacteria expressing few phenotypic characters. Therefore, several studies were conducted to compare rickettsial species or strains exhibiting diverse virulence phenotypes in order to identify pathogenesis factors. Surprisingly, no association was found between pathogenesis and the acquision of novel virulence genes [17,21,45]. In contrast, outer membrane proteins, notably Sca2 in *R. rickettsii*, and ankyrin repeat-coding genes were demonstrated to be essential virulence determinants [43,57]. However, RelA/SpoT responsible for the synthesis and hydrolysis of (p)ppGpp [58] and RickA, involved in actin-based bacterial motility [22] were found in both avirulent and virulent *R. rickettsii* strains and were thus ruled out as essential pathogenesis determinants [57]. In *R. prowazekii*, three virulence markers were identified through genome comparison, including *recO*, involved in DNA repair, *metK* and *adr1* encoding a S-adenosyl-methonine synthase and an adhesin, respectively, which are mutated in avirulent strains [53]. In addition, the RalF protein, a T4SS effector coded by genes conserved in all species, was demonstrated to play a role in host cell invasion in *R. typhi*, in contrast with SFG species in which it is pseudogenized [59].

6. Role of rickettsial plasmids in virulence

The presence of plasmids in *Rickettsia* genomes was first detected in that of *R. felis* [24]. To date, plasmids have been detected in 11 *Rickettsia* species [27]. Rickettsial plasmids result from vertical inheritance, mainly from *Orientia/Rickettsia* chromosome ancestors [27]. However, plasmids vary in number within and between species [27,60,61]. A variable plasmid content was observed in strains of *Rickettsia africae*, *R. bellii*, *Rickettsia akari*, *Rickettsia amblyommatis* and *R. felis* [21,60,61]. In addition, plasmid loss was demonstrated in cell culture [61]. As plasmids were present in several pathogenic species and contained protein-encoding genes necessary for recognition, invasion and pathogenicity, their role in rickettsial

Please cite this article in press as: Diop A, et al., Rickettsial genomics and the paradigm of genome reduction associated with increased virulence, Microbes and Infection (2017), https://doi.org/10.1016/j.micinf.2017.11.009

virulence was questioned [27]. However, the unstable plasmid content of *R. africae* did not support a role of plasmid in virulence in this species [21]. Furthermore, a strong correlation was observed between plasmid and genome sizes, with a paralleling decrease existing between plasmid size, number, and chromosome size. As examples, several species causing mild or no disease, such as the SFG *Rickettsia helvetica*, *R. felis*, and *Rickettsia peacockii*, possess one or more plasmids [21,24,43] whereas the most virulent species *R. prowazekii* and *R. typhi* that exhibit the most reduced genomes are plasmidless [15,52,53]. Furthermore, a recent multi-omics-study that compared four SFG rickettsiae showed that *R. conorii* and *R. slovaca*, the agents of MSF and SENLAT, respectively, were plasmidless but *R. massiliae* and *R. raoultii*, two less virulent agents of these diseases, harbor one and three plasmids, respectively [8]. Moreover, plasmids were also shown to undergo reductive evolutionary events similar to those affecting rickettsial chromosomes [27]. These findings support the absence of association between the presence of plasmids and difference in virulence in *Rickettsia* species.

7. Transcriptomic and proteomic investigation of rickettsial virulence

To date, several studies have demonstrated that transcriptomic and proteomic results are complementary to genomic analyses for analyzing bacterial virulence. A proteomic analysis of *Rickettsia parkeri* revealed that 91 proteins, including mostly virulence-related surface proteins (OmpA, OmpB, β-peptide, RickA), were differentially expressed during human infection [62]. Proteomic profile comparison of *R. prowazekii* grown in different cell lines, revealed an up-regulation of stress-related proteins in L929 murine fibroblasts [63]. In addition, proteins involved in protein synthesis, especially enoyl-(acyl carrier protein) reductase, a protein involved in fatty acid biosynthesis, were highly expressed when grown in *I. scapularis* ISE6 cells, suggesting that this rickettsia has the ability to regulate differentially its proteome according to the host [63]. Using transcriptomic and proteomic analyses of virulent and avirulent *R. prowazekii* strains, we identified four phenotypes that differed in virulence depending on the regulation of anti-apoptotic genes or the interferon I pathway in host cells [53]. Furthermore, *R. prowazekii* protein methylation (overproduced in virulent strains) and surface protein expression (Adr1 altered in avirulent Madrid E) varied with virulence, supporting the assumption that methylation of surface-exposed protein plays a role in the virulence of *R. prowazekii* [53]. In addition, in a recent proteomic and transcriptomic study, we compared two virulent agents, *R. conorii* and *R. slovaca*, causing MSF and SENLAT diseases, respectively, to two less virulent agents of the same diseases (*R. massiliae* and *R. raoultii*, respectively) [8]. Virulent species differed from less virulent ones by exhibiting mainly less up-regulated (8) than down-regulated (61) proteins. These included proteins associated mainly with translation, ribosomal structure and biogenesis, post-translational modification, protein turnover, chaperones, energy production and conversion [8]. In addition, virulent agents had rarely specifically expressed proteins [8]. This provides novel insights into the pathogenesis of *Rickettsia* species and suggests that virulence may not only be a question of presence or lack of a specific protein but may also result from a differential level or degradation of expression of a common protein.

8. Conclusion and perspective

Rickettsia spp., living mainly intracellularly in various arthropods, have undergone a particular paradoxical evolution marked by an evolutive chromosomic and plasmidic degradation resulting in a progressive genome reduction from 1.5 to 1.1 Mb with a coding capacity of 69−84%. This reductive evolution is marked by a selected loss of genes such as those associated with ATP, amino-acid and LPS metabolism or with synthesis of cell wall molecular components. In addition, a loss of regulatory genes and a high preservation of toxin-associated proteins and toxin-antitoxin modules are correlated to a rise in pathogenicity. However, paradoxically, these bacteria have undergone a proliferation of genetic elements whose role remains to be determined. As proteomic and transcriptomic analyses have just started to unveil the molecular mechanisms explaining the differences in virulence among *Rickettsia* species, and because the phenomenon of genome reduction associated with increased virulence seems to occur in other major human pathogens, these being examples of convergent evolution, *i. e.* natural selection leading to a similar biological outcome occurring independently in more than one unrelated biological group, future studies should identify which of the differences in rickettsial genomes account for their phenotypes.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This study was funded by the Mediterranee Infection Foundation and the French Agence Nationale de la Recherche under reference Investissements d'avenir Méditerranée Infection 10-IAHU-03.

References

[1] Stothard DR, Clark JB, Fuerst PA. Ancestral divergence of *Rickettsia bellii* from the spotted fever and typhus groups of *Rickettsia* and antiquity of the genus *Rickettsia*. Int J Syst Evol Microbiol 1994;44:798−804.
[2] Raoult D, Roux V. Rickettsioses as paradigms of new or emerging infectious diseases. Clin Microbiol Rev 1997;10:694−719.
[3] Gillespie JJ, Beier MS, Rahman MS, Ammerman NC, Shallom JM, Purkayastha A, et al. Plasmids and rickettsial evolution: insight from *Rickettsia felis*. PLos One 2007;2:e266.
[4] Merhej V, Raoult D. Rickettsial evolution in the light of comparative genomics. Biol Rev 2011;86:379−405.
[5] Merhej V, Angelakis E, Socolovschi C, Raoult D. Genotyping, evolution and epidemiological findings of *Rickettsia* species. Infect Genet Evol 2014;25:122−37.

[6] Parola P, Paddock CD, Socolovschi C, Labruna MB, Mediannikov O, Kernif T, et al. Update on tick-borne Rickettsioses around the world: a geographic approach. Clin Microbiol Rev 2013;26:657–702.
[7] Sahni SK, Narra HP, Sahni A, Walker DH. Recent molecular insights into rickettsial pathogenesis and immunity. Future Microbiol 2013;8:1265–88.
[8] El Karkouri K, Kowalczewska M, Armstrong N, Azza S, Fournier P-E, Raoult D. Multi-omics analysis sheds light on the evolution and the intracellular lifestyle strategies of spotted fever group *Rickettsia* spp. Front Microbiol 2017;8.
[9] Weinert LA, Werren JH, Aebi A, Stone GN, Jiggins FM. Evolution and diversity of *Rickettsia* bacteria. BMC Biol 2009;7:6.
[10] Murray GGR, Weinert LA, Rhule EL, Welch JJ. The phylogeny of *Rickettsia* using different evolutionary signatures: how tree-like is bacterial evolution? Syst Biol 2016;65:265–79.
[11] Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995;269:496–512.
[12] Balraj P, Renesto P, Raoult D. Advances in *Rickettsia* pathogenicity. Ann N Y Acad Sci 2009;1166:94–105.
[13] Gillespie JJ, Joardar V, Williams KP, Driscoll T, Hostetler JB, Nordberg E, et al. A Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. J Bacteriol 2012;194:376–94.
[14] Blanc G, Ogata H, Robert C, Audic S, Suhre K, Vestris G, et al. Reductive genome evolution from the mother of *Rickettsia*. PLoS Genet 2007;3:e14.
[15] McLeod MP, Qin X, Karpathy SE, Gioia J, Highlander SK, Fox GE, et al. Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. J Bacteriol 2004;186:5842–55.
[16] Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct 2009;4:13.
[17] Georgiades K, Raoult D. Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more toxin-antitoxin modules. PLos One 2011;6:e17962.
[18] Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UCM, Podowski RM, et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 1998;396:133–40.
[19] Sakharkar KR. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. Int J Syst Evol Microbiol 2004;54:1937–41.
[20] Walker DH. Progress in rickettsial genome analysis from pioneering of *Rickettsia prowazekii* to the recent *Rickettsia typhi*. Ann N Y Acad Sci 2005;1063:13–25.
[21] Fournier P-E, El Karkouri K, Leroy Q, Robert C, Giumelli B, Renesto P, et al. Analysis of the *Rickettsia africae* genome reveals that virulence acquisition in *Rickettsia* species may be explained by genome reduction. BMC Genomics 2009;10:166.
[22] Ogata H. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 2001;293:2093–8.
[23] Ogata H, Audic S, Abergel C, Fournier P-E, Claverie J-M. Protein coding palindromes are a unique but recurrent feature in *Rickettsia*. Genome Res 2002;12:808–16.
[24] Ogata H, Renesto P, Audic S, Robert C, Blanc G, Fournier P-E, et al. The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite. PLoS Biol 2005;3:e248.
[25] Blanc G, Ogata H, Robert C, Audic S, Claverie J-M, Raoult D. Lateral gene transfer between obligate intracellular bacteria: evidence from the *Rickettsia massiliae* genome. Genome Res 2007;17:1657–64.
[26] Baldridge GD, Burkhardt NY, Felsheim RF, Kurtti TJ, Munderloh UG. Transposon insertion reveals pRM, a plasmid of *Rickettsia monacensis*. Appl Environ Microbiol 2007;73:4984–95.
[27] El Karkouri K, Pontarotti P, Raoult D, Fournier P-E. Origin and evolution of rickettsial plasmids. PLos One 2016;11:e0147492.
[28] Merhej V, Georgiades K, Raoult D. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. Brief Funct Genomics 2013;12:291–304.
[29] Renesto P, Ogata H, Audic S, Claverie J-M, Raoult D. Some lessons from *Rickettsia* genomics. FEMS Microbiol Rev 2005;29:99–117.
[30] Andersson JO, Andersson SG. Genome degradation is an ongoing process in *Rickettsia*. Mol Biol Evol 1999;16:1178–91.
[31] Greub G, Raoult D. History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. Appl Environ Microbiol 2003;69:5530–5.
[32] Rovery C, Renesto P, Crapoulet N, Matsumoto K, Parola P, Ogata H, et al. Transcriptional response of *Rickettsia conorii* exposed to temperature variation and stress starvation. Res Microbiol 2005;156:211–8.
[33] Blanc G. Molecular evolution of *Rickettsia* surface antigens: evidence of positive selection. Mol Biol Evol 2005;22:2073–83.
[34] Gillespie JJ, Phan IQH, Driscoll TP, Guillotte ML, Lehman SS, Rennoll-Bankert KE, et al. The *Rickettsia* type IV secretion system: unrealized complexity mired by gene family expansion. Pathol Discov 2016;74:ftw058.
[35] Sears KT, Ceraul SM, Gillespie JJ, Allen ED, Popov VL, Ammerman NC, et al. Surface proteome analysis and characterization of surface cell antigen (Sca) or autotransporter family of *Rickettsia typhi*. PLoS Pathog 2012;8:e1002856.
[36] Fournier P-E, Zhu Y, Ogata H, Raoult D. Use of highly variable intergenic spacer sequences for multispacer typing of *Rickettsia conorii* strains. J Clin Microbiol 2004;42:5757–66.
[37] Amiri H, Alsmark CM, Andersson SG. Proliferation and deterioration of *Rickettsia* palindromic elements. Mol Biol Evol 2002;19:1234–43.
[38] Ogata H, Audic S, Barbe V, Artiguenave F, Fournier PE, Raoult D, et al. Selfish DNA in protein-coding genes of *Rickettsia*. Science 2000;290:347–50.
[39] Claverie J-M, Ogata H. The insertion of palindromic repeats in the evolution of proteins. Trends Biochem Sci 2003;28:75–80.
[40] Seshadri R, Paulsen IT, Eisen JA, Read TD, Nelson KE, Nelson WC, et al. Complete genome sequence of the Q-fever pathogen *Coxiella burnetii*. Proc Natl Acad Sci U S A 2003;100:5455–60.
[41] Caturegli P, Asanovich KM, Walls JJ, Bakken JS, Madigan JE, Popov VL, et al. ankA: an *Ehrlichia phagocytophila* group gene encoding a cytoplasmic protein antigen with ankyrin repeats. Infect Immun 2000;68:5277–83.
[42] Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, et al. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. PLoS Biol 2004;2:e69.
[43] Felsheim RF, Kurtti TJ, Munderloh UG. Genome sequence of the endosymbiont *Rickettsia peacockii* and comparison with virulent *Rickettsia rickettsii*: identification of virulence factors. PLos One 2009;4:e8361.
[44] Ogata H. *Rickettsia felis*, from culture to genome sequencing. Ann N Y Acad Sci 2005;1063:26–34.
[45] Darby AC, Cho N-H, Fuxelius H-H, Westberg J, Andersson SGE. Intracellular pathogens go extreme: genome evolution in the *Rickettsiales*. Trends Genet 2007;23:511–20.
[46] Gillespie JJ, Kaur SJ, Rahman MS, Rennoll-Bankert K, Sears KT, Beier-Sexton M, et al. Secretome of obligate intracellular Rickettsia. FEMS Microbiol Rev 2015;39:47–80.
[47] Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C, et al. Genome sequence of *Rickettsia bellii* illuminates the role of *Amoebae* in gene exchanges between intracellular pathogens. PLoS Genet 2006;2:e76.
[48] Gillespie JJ, Driscoll TP, Verhoeve VI, Utsuki T, Husseneder C, Chouljenko VN, et al. Genomic diversification in strains of *Rickettsia felis* isolated from different arthropods. Genome Biol Evol 2015;7:35–56.
[49] Schroeder CLC, Narra HP, Rojas M, Sahni A, Patel J, Khanipov K, et al. Bacterial small RNAs in the genus Rickettsia. BMC Genomics 2015;16.
[50] Hunter DJ, Torkelson JL, Bodnar J, Mortazavi B, Laurent T, Deason J, et al. The *Rickettsia endosymbiont of Ixodes pacificus* contains all the genes of de novo folate biosynthesis. PLos One 2015;10:e0144552.
[51] Akter A, Ooka T, Gotoh Y, Yamamoto S, Fujita H, Terasoma F, et al. Extremely low genomic diversity of *Rickettsia japonica* distributed in Japan. Genome Biol Evol 2017:evw304.

[52] Clark TR, Noriea NF, Bublitz DC, Ellison DW, Martens C, Lutter EI, et al. Comparative genome sequencing of *Rickettsia rickettsii* strains that differ in virulence. Infect Immun 2015;83:1568−76.

[53] Bechah Y, El Karkouri K, Mediannikov O, Leroy Q, Pelletier N, Robert C, et al. Genomic, proteomic, and transcriptomic analysis of virulent and avirulent *Rickettsia prowazekii* reveals its adaptive mutation capabilities. Genome Res 2010;20:655−63.

[54] Andersson SG, Kurland CG. Reductive evolution of resident genomes. Trends Microbiol 1998;6:263−8.

[55] Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. Cell 2002;108:583−6.

[56] Wixon J. Featured organism: reductive evolution in bacteria: *Buchnera* sp., *Rickettsia prowazekii* and *Mycobacterium leprae*. Comp Funct Genomics 2001;2:44−8.

[57] Ellison DW, Clark TR, Sturdevant DE, Virtaneva K, Porcella SF, Hackstadt T. Genomic comparison of virulent *Rickettsia rickettsii* Sheila Smith and avirulent *Rickettsia rickettsii* Iowa. Infect Immun 2008;76:542−50.

[58] Clark TR, Ellison DW, Kleba B, Hackstadt T. Complementation of *Rickettsia rickettsii* RelA/SpoT restores a nonlytic plaque phenotype. Infect Immun 2011;79:1631−7.

[59] Rennoll-Bankert KE, Rahman MS, Gillespie JJ, Guillotte ML, Kaur SJ, Lehman SS, et al. Which way in? The RalF Arf-GEF orchestrates *Rickettsia* host cell invasion. PLoS Pathog 2015;11:e1005115.

[60] Baldridge GD, Burkhardt NY, Felsheim RF, Kurtti TJ, Munderloh UG. Plasmids of the pRM/pRF family occur in diverse *Rickettsia* species. Appl Environ Microbiol 2008;74:645−52.

[61] Fournier P-E, Belghazi L, Robert C, Elkarkouri K, Richards AL, Greub G, et al. Variations of plasmid content in *Rickettsia felis*. PLos One 2008;3:e2289.

[62] Pornwiroon W, Bourchookarn A, Paddock CD, Macaluso KR. Proteomic analysis of *Rickettsia parkeri* strain Portsmouth. Infect Immun 2009;77:5262−71.

[63] Tucker AM, Driskell LO, Pannell LK, Wood DO. Differential proteomic analysis of *Rickettsia prowazekii* propagated in diverse host backgrounds. Appl Environ Microbiol 2011;77:4712−8.

Please cite this article in press as: Diop A, et al., Rickettsial genomics and the paradigm of genome reduction associated with increased virulence, Microbes and Infection (2017), https://doi.org/10.1016/j.micinf.2017.11.009

CHAPITRE II

Classification taxonomique des espèces du genre *Rickettsia* sur la base des données des séquences génomiques

Avant-propos

Actuellement, l'information génomique est de plus en plus utilisée pour la définition et la classification des espèces procaryotes grâce à l'accessibilité sans précédent à des données génomiques adéquates couplée à la disponibilité d'outils génomiques innovants, objectifs and reproductibles pour une classification taxonomique plus précise. Cependant, les critères génomiques usuels les plus largement acceptés pour la définition des espèces bactériennes ne sont pas applicables à de nombreux genres bactériens. Ainsi le statut taxonomique de plusieures espèces bactériennes reste encore un sujet de débat. C'est notamment le cas des espèces du genre *Rickettsia.*

Les rickettsies sont des alpha-protéobactéries strictement intracellulaires possédant de petits génomes avec un taux de G+C% faible (29-33%) et qui expriment peu de caractéristiques phénotypiques. A ce jour, il y a 30 espèces officiellement validées (www.bacterio.net/rickettsia.html) avec près de 100 génomes de *Rickettsia* disponibles et de nombreux autres isolats de rickettsies qui n'ont pas encore été entièrement caractérisés, ou qui n'ont pas reçu de désignation d'espèce, ont également été récemment décrits sur la base de la caractérisation moléculaire des rickettsies basée sur les séquences de plusieurs gènes.

Dans cette partie de nos travaux de thèse, notre objectif était d'évaluer une gamme de paramètres taxonomiques basés sur l'analyse des séquences génomiques afin de mettre au point des recommandations pour la classification des isolats au niveau de l'espèce et du genre. Ainsi, En comparant le degré de similarité des séquences de 78 génomes de *Rickettsia* et 61 génomes de 3

genres étroitement apparentés (*Orientia*, 11 *génomes*, *Ehrlichia*, 22 génomes et *Anaplasma phagocytophilum*, 28 génomes) utilisés comme outgroup, en utilisant plusieurs paramètres génomiques basés sur la taxonomie: hybridation ADN-ADN in silico (dDDH); Identité nucléotidique moyenne par orthologie (OrthoANI) et identité génomique moyenne des séquences de gènes orthologues (AGIOS), nos résultats montrent que les outils AGIOS et OrthoANI sont les meilleures méthodes permettant de definir qu'un isolat bactérien appartient bien au genre *Rickettsia* avec une specifité de 100%. Au sein de l'ordre des *Rickettsiales*, les rangs de genres et espèces ne présentaient aucun chevauchement en termes de valeurs d'OrthoANI. Toutes les souches des 28 espèces valides étudiées, étaient correctement classées dans le genre *Rickettsia* avec des seuils définis ≥ 80,5 et ≥ 80,5% pour les valeurs OrthoANI et/ou AGIOS, respectivement. D'après les résultats des tests de corrélations obtenus, ces deux cut-offs correspondaient exactement aux seuils de 98.1% et 86.5% de similarité de la séquence du gène de l'ARNr 16S et du gène *gltA* établis pour définir la limite au niveau du genre chez les espèces de *Rickettsia*. Donc pour qu'un isolat soit classé comme un membre du genre *Rickettsia,* il doit présenter des valeurs d'OrthoANI et/ou AGIOS avec l'une des espèces de *Rickettsia* reconnues supérieures ou égales à ces seuils. En revanche, le dDDH était le meilleur outil pour definir si un isolat bactérien était une nouvelle espèce ou appartenait à une espèce de *Rickettsia* connue avec un seuil ≥ 92.3%. Ce seuil correspondait parfaitement au seuil de 99.8% de similarité de la séquence du gène de l'ARNr 16S recommandé pour definir les espèces. Cependant les outils AGIOS et OrthoANI peuvent également être utilisés comme méthodes complémentaires, mais

pas pour les espèces étroitement apparentées à *R. conorii*. Ainsi pour être classé comme une nouvelle espèce de *Rickettsia*, un isolat bactérien ne devrait pas présenter plus d'une des valeurs de similarité génomique suivantes avec les espèces validées les plus proches: ≥ 92,3, ≥ 99,2 et ≥ 98,6% pour le dDDH, OrthoANI et AGIOS, respectivement. Nous avons montré que les outils taxono-génomiques sont des méthodes relativement simples d'utilisation en laboratoire et permettent une classification taxonomique fiable, rapide et facile pour les espèces de *Rickettsia* avec des seuils spécifiques. Les résultats obtenus nous ont permis ainsi d'élaborer des lignes directrices pour la classification des isolats de rickettsies au niveau du genre et de l'espèce.

Dans ce travail, nous avons également fait la caractérisation et la description d'une nouvelle espèce de *Rickettsia* nommée *Rickettsia fournieri* souche AUS118, qui a été inclue dans cette précédente étude.

Article 3:

Genome sequence-based criteria for species demarcation and definition: insight from the genus *Rickettsia*

Awa Diop, EL Karkouri Khalid, Didier Raoult
and Pierre-Edouard Fournier

[Submitted in International Journal of Systematic and Evolutionary Microbiology]

Genome sequence-based criteria for species demarcation and definition : Insight from the genus *Rickettsia*

Awa Diop[1], Khalid El Karkouri[1], Didier Raoult[2] and Pierre-Edouard Fournier[1*]

[1] UMR VITROME, Aix-Marseille University, IRD, Service de Santé des Armées, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France Tel: +33 413 732 401, Fax: +33 413 732 402.

[2] UMR MEPHI, Aix-Marseille University, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Uuniversitaire Méditerranée Infection, Marseille, France.

[*]Corresponding author: Pr Pierre-Edouard Fournier

Email: pierre-edouard.fournier@univ-amu.fr

Abstract :

Over recent years, genomic information has increasingly been used for prokaryotic species definition and classification. Genome sequence-based alternatives to the gold standard DNA-DNA hybridization (DDH) relatedness have been developed, notably the average nucleotide identity (ANI) that is one of the most useful measurements for species delineation in the genomic era. However, that strictly intracellar lifestyle, the few measurable phenotypic properties and the low level of genetic heterogeneity made the current standard genomic criteria for bacterial species definition inapplicable to *Rickettsia* species.We attempted to evaluate a range of genome-based taxonomic parameters, to develop guidelines for the classification of *Rickettsia* isolates at the genus and species levels. By comparing the degree of similarity of the sequences of 78 genomes from *Rickettsia* species and 61 genomes from 3 closely related genera (*Orientia*, 11 genomes; *Ehrlichia*, 22 genomes; and *Anaplasma*, 28 genomes) using digital DDH (dDDH), ANI by orthology (OrthoANI) and average genomic identity of orthologous genes sequences (AGIOS), we demonstrated that genome-based taxonomic tools are easy-to-use and fast and can serve as a robust genomic index for establishing *Rickettsia* genus and species boundaries. Within the order *Rickettsiales*, genus and family ranks showed no overlap in terms of OrthoANI values. Basically, to be classified as a member of the genus *Rickettsia*, an isolate should exhibit OrthoANI and AGIOS values between any of the *Rickettsia* species with standing in nomenclature of ≥ 80.5. To be classified as a new *Rickettsia* species, an isolate should not exhibit more than one of the following degrees of genomic relatedness levels with the most closely related species: ≥ 92.3, ≥ 99.2 and ≥ 98.6% for the dDDH, OrthoANI, and AGIOS values, respectively. Thus, we propose that whole-genome data can be used to efficiently delimitate *Rickettsia* species.

Keywords: Whole-genome data, Genome-based taxonomy, *Rickettsia*, dDDH, AGIOS, OrthoANI, Species definition.

1 Introduction

The genus *Rickettsia* was first proposed by da Rocha-Lima in 1916 (1) after Howard Taylor Ricketts and Stanislav von Prowazek laid the foundation of modern rickettsiology and eventually the recognition of new species and rickettsial infections (2). In 1980, the genus was listed in *Bergey's Manual of Systematic Bacteriology* (3). The term rickettsiae has once been used to describe (2), any strictly intracellular bacterium (4). In the early 1980s, the order *Rickettsiales* consisted of the families *Rickettsiaceae*, *Bartonellaceae*, and *Anaplasmataceae* (5). The use of 16S rRNA gene (rrs) sequences in the 1990s, deeply changed the classification of rickettsiae (6,7). *Eperythrozoon* spp. and *Haemobartonella* spp. were reclassified within the family *Mycoplasmataceae* (7,8), *Coxiella burnetii* and *Rickettsiella grylli* within the *Legionellaceae* family (6,7,9). Likewise, *Wolbachia melophagi*, *Rochalimaea* sp., *Grahamella* sp., and *Bartonella* sp., were reclassified within the family *Bartonellaceae* (7,10,11) and removed from the order *Rickettsiales*. As a consequence, the order *Rickettsiales* is only made of two families: *Rickettsiaceae* that includes the genera *Rickettsia* and *Orientia* and *Anaplasmataceae* with the genera *Ehrlichia*, W*olbachia*, *Anaplasma* and *Neorickettsia* (7,12,13).

Within the *Rickettsia* genus, species were classified in three groups: the typhus group (TG), the spotted fever group (SFG) and the scrub typhus group, on the basis of their phenotypic characteristics including ecological and epidemiological characteristics, pathogenicity and clinical data as well as results from the mouse serotyping test (7,14,15). In 1995, after analyzing its 16S rRNA gene sequence, *Rickettsia tsutsugamushi* was reclassified into a new genus, *Orientia* (12). To date, there are 30 *Rickettsia* species with standing in nomenclature within the genus, species were classified into three groups based on clinical, genotypic and phenotypic features: the ancestral group (AG) that contains *R. bellii* and *R. canadensis* associated with ticks and not pathogenic, the spotted fever group (SFG) that contains pathogenic agents causing spotted fevers as well as numerous species of as-yet unknown pathogenicity, are mostly associated with ticks, motile into the nuclei of host cells and cross-react with *Proteus vulgaris* OX-2 and have an optimal growth temperature of

32°C and the typhus group (TG) that includes *R. prowazekii* and *R. typhi* which cause typhus and are associated with human body lice and rat fleas respectively, not motile and cross-react with *Proteus vulgaris* OX-19 and have an optimal temperature of growth of 35°C. In addition to the 30 recognized species, numerous other rickettsial isolates which have not yet been fully characterized, have also been recently described based on molecular characterization (15).

The mouse serotyping test, developed in 1978, has long considered as the reference method for rickettsial identification (16). However, mouse serotyping method has many drawbacks including a lack of reproducibility, and is labour intensive to compare each new isolate to all previously described species. The use of the 16S ribosomal RNA gene sequence similarity (16S rRNA), the G + C content of DNA (G+C%), the DNA-DNA hybridization (DDH) relatedness and the description of phenotypic characteristics in a polyphasic classification strategy are the basis for the most widely accepted description of bacterial species (17,18). However, their strictly intracellar lifestyle, Their few phenotypic properties and their low level of genetic heterogeneity, making the universal 16S rRNA sequence similarity thresholds (95% and 98.65-98.7% at the genus and species ranks, respectively) and or divergence (3%), difference in G+C% (> 5% between two species) and DDH (< 70% between two species) used for the definition of species are not applicable to *Rickettsia* species (7,15,19). Thus, the definition of species within the genus *Rickettsia* has long been a matter of debate particularly in regarding their taxonomy due to the lack of official rules (7). But in 2003, the introduction of a molecular tool based on the analysis of five genes sequences: 16S rRNA, *gltA*, *ompA*, *ompB* and *sca4* genes has revolutionized the characterization and taxonomic classification of rickettsiae and is the current basis for their classification (15) with reliable phylogenetic estimation based on three or four concatenated MLST genes than with single gene (20). Despite these efforts, the taxonomy of members of the genus *Rickettsia* remained a subject of debate.

Over the past two decades, the remarkable advances in DNA sequencing technologies have allowed access to complete genomic sequences, allowing unprecedented access to valuable data for a more accurate taxonomic classification of prokaryotes (21–23). Therefore, whole-genome

sequencing has delivered several taxonomic tools based on genomic sequences coined as the overall genome related index (OGRI) (24) such as digital DNA-DNA hybridization (dDDH) (25–27), the average nucleotide identity (ANI) (27–29) or most recently the average nucleotide identity by orthology (OrthoANI) (30), average amino acid identity (AAI) (31) and average genomic identity of orthologous genes sequences (AGIOS) (23,32). Nowadays, genomic information is increasingly applied to prokaryotic species definition and classification. Despite, DDH relatedness still serves as the gold standard in prokaryotic taxonomy (21,22), the ANI (OrthoANI) (95~96% between two species) become one of the most useful measurements for species delineation in the genomic era and exhibited a strong correlation with DDH values (22,27). Over the past 10 years, the emergency of rickettsial genomics proved its usefulness in a variety of applications (7). In addition, phylogenomic treeing based on core gene sets of rickettsial genomes was demonstrated to provide more precise phylogenetic relationship supported by elevated boostrap values (7,33,34). Furthermore, the use of minimum number of genes to be 31 house-keeping, which is higher than that used in the traditional multilocus sequence analysis (MLSA) for phylogenomic study, was recommended by Chun et al., in 2018 (21).

Given the availability of genomic sequences of nearly 100 rickettsial genomes, we wanted to evaluate a range of taxonomic parameters based on genomic sequence analysis, to develop guidelines for the classification of *Rickettsia* isolates at the genus and species levels. In pursuit of this aim, we analyzed and compared the published whole-genome sequences from validated and unvalidated *Rickettsia* species available in Genbank.

2 Materials and Methods

2.1 Data set.

All analyzed genomes were downloaded from GenBank (ftp://ftp.ncbi.nih.gov/Genome/). These include the genomes from 78 *Rickettsia* strains (48 “complete” and 30 “incomplete genome sequences (WGS)”), 11 *Orientia tsutsugamushi* (2 “complete” and 9 “incomplete genome

sequences"), 22 *Ehrlichia* strains (13 "complete" and 9 "incomplete genome sequences ") and 28 *Anaplasma phagocytophilum* genomes (5 "complete" and 23 "incomplete genome sequences"). For *Rickettsia* species, we studied genomes from 28 species with standing in nomenclature (http://www.bacterio.net/) and 6 *Rickettsia* isolates from as yet unofficial species (Table 1). Genome sequences of members of the closely genera *Orientia*, *Ehrlichia* and *Anaplasma* were used as outgroup for the present study. The list of the 139 studied genomes is presented in Table 1. Three genome similarity parameters (dDDH, OrthoANI and AGIOS) were used. In addition, the complete sequences of the five genes: 16S rRNA, *gltA*, *ompA*, *ompB* and *sca4* extracted directly from each genome were included in the present study for statistical correlation tests.

2.2 Digital DNA-DNA Hybridization (dDDH) relatedness prediction

The dDDH relatedness values between genome pairs were predicted using the GGDC 2.1 web server (35) available at (http://ggdc.dsmz.de/distcalc2.php) (36).

2.3 Determination of average nucleotide identity by Orthology (OrthoANI)

The ANI (OrthoANI) values between two genome sequences were calculated using the OrthoANI algorithm version v0.91 as described by Lee *et al.* (30). The TMev software (http://sigenae.org/index.php?id=88) was used to visualize the results as a heatmap. For ANI values below 75%, the average amino acid identity (AAI) (37) was calculated on the basis of the overall similarity between two genomic datasets of predicted proteins using the web server available at http://enve-omics.ce.gatech.edu/aai/index.

2.4 Determination of the Average genomic identity of orthologous gene sequences (AGIOS)

For the calculation of AGIOS values, the degrees of genomic sequence similarity among compared genomes were estimated using the MAGI (Marseille Average Genomic Identity) home-made pipeline (38). The first step is to determine non ambiguous orthologous genes shared by the genomes using ProteinOrtho (39) that allows to detect orthologous genes group in pairwise genomic comparisons.

2.5 Calculation of cutoff values at the genus and species levels.

The cutoff values at the genus level for each genomic method used was calculated as previously described (15). Briefly, the mean dDDH, OrthoANI and AGIOS values between *Rickettsia* species were first calculated. Second, the standard deviation (SD) was calculated at the genus level. Subsequently, the cutoff was defined as the mean less 3 SDs. Thus, a strain with a degree of genomic sequence similarity of at least 3 SDs lower than the mean genomic sequence divergence between each species pair within the genus *Rickettsia* would be likely (with more than 99% probability) not to belong to this genus.

In order to validate each threshold, we applied to the pairwise genomic sequence similarity rates between all species used to establish the threshold as well as species of the three genera used as outgroups (*Orientia*, *Ehrlichia* and *Anaplasma*).

The sensitivity and specificity of a threshold for a given group (species or genus level) were also determined as previously described (15).

To calculate thresholds at the species level, we first evaluated the minimum dDDH, OrthoANI and AGIOS values at the intra-species level for each *Rickettsia* species with at least 2 strains. Second, we evaluated the cutoff value for each method according to the highest degree of similarity of genomic sequences in pairs observed among all validated species. Subsequently, to validate the obtained cutoffs, they were applied to 72 genomes of the officially species used to calculated them.

Finally, to evaluate the usefulness of our genomic criteria thresholds, they were applied to six previously classified member of the genus *Rickettsia*, namely: "*R. monacensis* strain IrR/Munich" (40), "*R. endosymbiont of ixodes pacificus* strain Humboldt, "*R. endosymbiont of ixodes scapularis*" (41) , all three of which were phylogenetically closely related to *R. tamurae* on the basis of genotypic and phenotypic criteria, "*R. fournieri*" a new isolate from our laboratory, closely related to *R. japonica* and *R. heilongjiangensis*, but considered as a distinct species on the basis of genotypic criteria, "*R. argasii*" strain T170-B very close to R. *helongjiangensis* and "*R. philipii*" strain 364D very close to *R. rickettsii* but considered as a distinct species on the basis of

epidemiological characteristics and serotyping tests (42,43).

2.6 Core genome phylogenetic analysis

Phylogenetic relationships between *Rickettsia* species was not well established with the use of a single gene, and concatenated MLST genes (16S rRNA, *gltA*, *sca4*, *ompA* and or *ompB* genes) were used to infer efficiently the phylogenetic relationships of these bacteria. In this aim, we attempted to reconstruct a phylogeny based on more comprehensive gene set precisely the core genome of the 78 *Rickettsia* strains. For each genome, gene prediction was done using the Prokka software (44) in order to generate sets of gene (orfeome file) and protein sequences (proteome file). The core genome was identified using the ProteinOrtho software (39). To compare the taxonomic discrimination power from our genomic criterion to those deduced from phylogenomic analysis based to conserved genes between all strains, the amino acid sequences of these 591 proteins were concatenated for each genome and multiple alignment was performed using the Mafft software (45). Gapped positions were removed. The phylogenetic inferences were obtained using Maximum Likelihood method within the MEGA software (Molecular Evolutionary Genetics Analysis), version 6 (46). Branching support was evaluated using the bootstrap method with 500 replications.

2.7 Statistical analysis

Statistical analysis was performed using the GraphPad Prism version 5.04 (GraphPad Software Inc, 2012, La Jolla, CA, www.graphpad.com/prism). The Pearson's correlation coefficient was used for the correlation analysis with linear regression. Values were considered statistically significant at a 95% confidence level when $P < 0.05$. We evaluated the correlation between dDDH, OrthoANI, AGIOS data and the pairwise nucleotide sequence similarity generated by 16S rRNA, *gltA*, *ompA*, *ompB* and *sca4* individually using the linear regression model (Table S4).

3 Results

3.1 Defining *Rickettsia* species on the basis of whole-genome sequence analysis

The complete nucleotide sequences of 16S rRNA (1484-1509 bp) and *gltA* (1305-1335 bp) and the partial sequence size used by Fournier et al., 2003 (15) of *ompA* (1-590 bp), *ompB* (296-5141 bp) and *sca4* (33-2979 bp) genes of 72 strains of the 28 valid *Rickettsia* species were studied. When the widely used species boundary for dDDH >70% and ANI values > 95~96, respectively were applied to our dataset, we were able to classify 32 of the 78 strains into eleven previously named species included *R. canadensis* and *R. bellii* (AG), *R. typhi* and *R. prowazekii* (TG), *R. akari*, *R. australis*, *R. felis*, *R. helvetica*, *R. hoogstraalii*, *R. asembonensis* and *R. tamurae* (TRG or SFG). All other spotted fever group species were classified within a single species (Fig. 1; Fig. 2). This result confirmed that thresholds used for other genera were not adequate for *Rickettsia* species, highlighting the need define specific genomic thresholds for Rickettsia species delineation based on genomic tools.

3.2 Evaluation of genome similarity of the genus level

dDDH analysis: among *Rickettsia* species, dDDH values ranged from 23.2% between *R. bellii* and *R. typhi* to 92.3% between *R. sibirica* and *R. parkeri* ((Fig. 3; Table S1). The mean dDDH level less 3 SDs among the 28 species studied was thus 12.02%. When this value was applied to the 28 *Rickettsia* species, it was validated for 4826 of 4826 similarity rates (sensitivity, 100%) (Table S1). All species from the three genera used as outgroup (61 strains) exhibited dDDH values with any tested *Rickettsia* species greater than 12.02% (specificity, 0%) (Table 2)

OrthoANI and AAI measurements of relatedness: Within the genus *Rickettsia*, OrthoANI values ranged from 79.6% between R. *bellii* and *R. prowazekii* to 99.2% between *R. sibirica* and *R. parkeri* (Fig 3; Table S2). The mean level of genomic sequence similarity less 3 SDs among the 28 species was 80.5%. When this value was applied to the 28 *Rickettsia* species, it exhibited a sensitivity of 4770 of 4826 (97.7%) (Table S2). OrthoANI values between outgroup and *Rickettsia* species were all lower than 75% (range from 62.8 to 67.0%) (Table 2). AAI values between outgroup and

Rickettsia species ranged from 40.9 to 49.5% (Table 2). Therefore, none of the three outgroup genera (61 genomes) fulfilled this criterion with any strains of the 28 *Rickettsia* species (specificity, 100%) (Fig 3; Table 2).

AGIOS measurement of relatedness: AGIOS values among *Rickettsia* species ranged from 78.5% between *R. canadensis* and *R. felis* to 98.6% between *R. sibirica* and *R. parkeri* (Fig. 3; Table S3). The mean AGIOS values less 3 SDs among the 28 species was 80.5%. When this value was applied to the 28 *Rickettsia* species and species of the three outgroup genera, it had a sensitivity of 4544 of 4826 similarities rates (94.2%) and none of the three outgroup genera (61 strains) fulfilled this criterion with any of the 28 *Rickettsia* species (specificity, 100%) (Table S3; Table 2).

3.3 Application of the genus criteria to Rickettsia species.

Due to its poor specificity, dDDH was not suitable to delineate *Rickettsia* species at the genus level, in contrast to OrthoANI and AGIOS values.

3.4 Use of genome-based criteria at the species level

We also evaluated the pairwise genomic sequence similarity for each of the three methods among strains within of the 14 *Rickettsia* species for which at least two strains were available (Table 1). Our results showed that dDDH is more variable from one species to another when compared to OrthoANI and GAIOS. It ranged from 88.8% between the *R. canadensis* strains to 99.9 between the *R. australis* strains (Table S1). In addition the dDDH criterion among strains within each strains of the fourteen studied species had a specifity of 100%. In addition, OrthoANI and AGIOS values ranged from 98.8 between *R. canadensis* strains to 99.9 between *R. australis* strains or *R. slovaca* strains and from 97.4 between *R. typhi* strains to 99.5 between *R. parkeri* strains respectively (Table S2 ; S3). These criteria had a specifity of 100% for all 14 studied *Rickettsia* species with the exception of *R. conorii* for which the specificity was 89.1 and 96.4% for the OrthoANI and AGIOS parameters, respectively. Thus, at the intra-species level, the dDDH method was more specific than OrthoANI and AGIOS making dDDH the best tool to define if a bacterial isolate was a new *Rickettsia* species or an isolate belonging to a previously known *Rickettsia* species. Nevertheless,

AGIOS and OrthoANI thresholds can also be used as complementary methods, but not for species closely related to *R. conorii*. In addition, the highest pairwise genomic sequence similarity rates among the 28 validated species were 92.3, 99.2 and 98.6% for the dDDH, OrthoANI, and AGIOS values, respectively (Fig. 3; Table S1; S2; S3). When these criteria were applied to all 72 strains of the 28 *Rickettsia* species, almost of these strains were correctly classified in their corresponding previously named species with 100% of specificity and exhibited levels of genomic sequence similarity to other strains of their respective species higher than these criteria excepted *R. canadensis* (88.0 and 98.8% for dDDH and OrthoANI values respectively) *R. conorii* (91.5, 99.0 and 98.1% for dDDH, OrthoANI and AGIOS values respectively), *R. massiliae* (90.5, 99.0 and 97.9% for dDDH, OrthoANI and AGIOS values respectively) and *R. felis* (97.6% for AGIOS values) (Table S1; S2; S3).

3.5 Application of genome similarity threshold to Rickettsia species of uncertain taxonomic status

By using the above-described genome-based taxonomic criteria (Fig. 3), all six unvalidated species belonged to the genus *Rickettsia* (Table 2). *R. monacensis, R. Endosymbiont of Ixodes scapularis, R. Endosymbiont of Ixodes pacificus* closely related to *R. tamurae* and *R. fournierii* closely related to *R. japonica* and *R. heilongjiangensis* fulfilled the three genomic cutoffs (Fig. 3) and were classified as new distinct species. In addition, *R. argasii* and *R. philipii*, phylogenetically closely related to *R. heilongjiangensis* et *R. rickettsii*, respectively, and previously proposed as new species, did not validate the genomic criteria (Fig. 3) for considering them as new species, but belonged to the *R. heilongjiangensis* and *R. rickettsii* species respectively.

3.6 Comparison of genomic similarity parameters and MLST

We found a strong positive and significant linear correlation among all genomic parameters tested ($P < 0.0001$, Fig. 4). The highest correlation was obtained between OrthoANI and AGIOS parameters ($r^2 = 0.9872$), and the lowest correlation between dDDH and AGIOS ($r^2 = 0.8623$) (Fig. 4F; Fig. 4H, respectively). In adition, we found a strong positive and significant linear correlation

between genome-based taxonomic parameters and the reference gene sequences tested ($P < 0.0001$ for all tests). Among these, *ompB* gene showed the highest correlation ($r^2 = 0.9836$) to OrthoANI, *sca4* gene showed the highest correlation to dDDH ($r^2 = 0.9196$) and *gltA* showed the highest correlation to AGIOS ($r^2 = 0.9653$) (Fig. 4G; 4E). The 16S rRNA gene showed the lowest correlation ($r^2 = 0.6850$, 0.5510 and 0.5101, respectively) to dDDH, OrthoANI and AGIOS (Fig. 4A; 4B; 4C, respectively). While *ompA* showed significantly lower correlation to dDDH and AGIOS ($r^2 = 0.8800$ and 0.8751, respectively) than *ompB* ($r^2 = 0.9159$ and 0.9633, respectively) and to OrthoANI ($r^2 = 0.9013$) than *gltA* and *sca4* ($r^2 = 0.9698$ and 0.9664 respectively) (Fig. 4G; 4D). In addition, the 80.5% threshold for OrthoANI and AGIOS corresponded well to the 98.1% and 86.5% 16S rRNA and gltA thresholds respectively, used to define *Rickettsia* boundary at the genus levels (Fig. 4B; 4C; 4D; 4E). Moreover, the 92.3% threshold for dDDH corresponded well to the 99.8% 16S rRNA threshold gene sequence similarity established to define *Rickettsia* boundary at the species level (Fig. 4A). Furthermore, the cutoff point of 80.5% of OrthoANI corresponded well to the 80.5% of AGIOS determined (Fig. 4F).

3.7 Phylogenomic analysis

Most of the widely used phylogenetic methods have been developed to infer the phylogeny of a gene, but not the entire genome sequence. Many genes have undergone horizontal transfer events, making difficult to elucidate precise phylogenetic relationships between genomes. We built a phylogenomic tree based on 591 common genes, that supported the monophyletic status of previously named species within the genus by elevated boostrap values and was similar to the classical classification of rickettsiae within three main clusters (Fig. 5). The first group included *R. bellii* and *R. canadensis*, the most outlying rickettsiae. The second cluster grouped the typhus group rickettsiae made of *R. typhi* and *R. prowazekii*. The last cluster grouped the largest number of rickettsiae (spotted fever group). Taxonomic classification error was discovered for one of the 78 studied strains. This strains named *R. rhipicephali* strain Ect was previously classified as a *R. rhipicephali* strain, but phylogenetically, clustered with the two *R. massiliae* strains with 100%

boostrap value (Fig. 5). In addition genome and gene sequence-based criteria confirmed this finding.

4 Discussion

We propose genome-based criteria as an alternative method to the traditional genotypic tools for the taxonomic classification of rickettsial isolates at the genus and species levels. The definition of species within the genus *Rickettsia* has long been a matter of debate because of their strict intracellular lifestyle, making it difficult to define the species boundaries among these bacteria (6,7). Moreover, the phenotypic criteria used for extracellular bacterial species definition are not applicable since few are expressed by these bacteria (7,15,19). Thus, various methods have been used for rickettsial species identification but failed to provide easily reproducible identification tools. Among these, cross-immunity and vaccine protection tests with guinea pigs (15,47), complement fixation tests (15,48), mouse toxin neutralization tests (15,49), mouse serotyping assays (15,16), sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and pulsed-field gelelectrophoresis (PFGE) (7,15) have all proven to be useful for differentiating rickettsiae but all suffered limitations and disadvantages such a lack of reproducibility and the difficulty to standardize for monoclonal antibody tests (7,15,50,51), changes in molecular weights of rOmpA and rOmpB among species limiting SDS-PAGE result values or the absence of database allowing the comparison of profiles PFGE (15). In 2003, the introduction of MLST scheme based on the analysis of five genes (16S rRNA, *gltA*, *ompA*, *ompB*, and *sca4*) has facilitated the characterization and taxonomic classification of rickettsial isolates and is the current basis for their classification. This was the first method allowing to define rickettsial species boundary with an accepted standard panel for all known isolates (7,15,52). However, over the past two decades, the remarkable advances in DNA sequencing technologies have allowed access to complete genomic sequences, within a short time and for an affordable budget allowing unprecedented access to valuable data for a more accurate taxonomic classification of prokaryotes. Several genome-based tools have been developed including ANI, AAI, digital DDH, that provide a numerical standard

threshold and has been shown to be applicable to a diverse group of bacteria but not to all (7,22,31,32,53,54). The usefulness of whole-genomic approaches for taxonomic purposes was demonstrated for many bacterial species definition (22,55–57). However, genome-based taxonomic tools have not been evaluated for *Rickettsia* species delineation. With the availability of genomic sequences of nearly 100 rickettsial genomes, we evaluated a range of genome-based taxonomic parameters, and proposed guidelines for the classification of new rickettsial isolates (Fig. 3). Our results showed that the AGIOS and OrthoANI parameters were the best tools to classify that rickettsia-like organism into the genus *Rickettsia*, supported by elevated sensitivities and specificities. Although the ANI parameter has been proposed to provide a high degree of resolution at the species and sub-species levels (22,31,54), within the order *Rickettsiales*, at the genus and species levels, OrthoANI values did not overlap, allowing us to use this parameter to define boundaries at the genus level. The AGIOS parameter, a tool created in our laboratory, has been used for taxonomic description of various new bacterial species and demonstrated a high sensitivity and specificity for Rickettsia species. When applied to the 28 studied species, we determined thresholds values of $\geq$ 80.5 and $\geq$ 80.5% at the genus level for the OrthoANI and AGIOS parameters, respectively. Hence, a rickettsia-like organism can be classified as a member of the *Rickettsia* genus, if it exhibits an OrthoANI and/or AGIOS values with one of the recognized *Rickettsia* species greater than or equal to 80.5%. Both the OrthoANI and the AGIOS cutoffs were validated by comparison with 3 closely related genera (61 species). In addition, we demonstrated that AGIOS and OrthoANI exhibit a high degree of correlation well between them and with 16S rRNA and the *gltA* gene sequences similarity analyses. The 80.5% threshold corresponded well to the 98.1% and 86.5% 16S rRNA and *gltA* threshold respectively, at the genus levels (15). In contrast, dDDH was the best to the three tested tools to define whether a rickettsia-like organism was a new species or belonged to a known *Rickettsia* species with a predicted cutoff value of 92.3%. A strong correlation was observed between dDDH values and the 16S rRNA gene sequence similarities and this threshold corresponded well to the 99.8% of the 16S rRNA gene sequence similarity threshold

established to define *Rickettsia* boundary at the species level (15). However, the AGIOS and OrthoANI tools can also be used as complementary methods to define *Rickettsia* boundaries at the species level but not for species closely related to *R. conorii*. To be classified as a new species an isolate should not exhibit more than one of the following degrees of dDDH, OrthoANI and AGIOS values with at least 1 of the 28 validated *Rickettsia* species: ≥ 92.3, ≥ 99.2 and ≥ 98.6% respectively. When our genomo-taxonomic scheme was applied to six rickettsial strains not previously officially classified, all of them were correctly classified into the genus *Rickettsia*. Our results also confirmed the previous tentative taxonomic classification of four strains whose taxonomic status is not yet established. On the basis of phenotypic and genotypic analysis these four strains were previously proposed to be new Rickettsia species. Our data confirm that these rickettsiae belongs to 4 new separated distinct species. In contrast, *R. argasii* and *R. philipii*, previously proposed as new species, belong to *R. heilongjiangensis* and *R. rickettsii* respectively. On the basis of genomic and phylogenomic analysis, we also identified a taxonomic classification error of *R. rhipicephali* strain Ect that rather belongs to *R. massiliae* rather than *R. rhipicephali*. This finding is congruent with the results of gene sequence-based analysis. Our study has shown that genome-based taxonomic tools are well suited, reliable and reproducible for the delineation of *Rickettsia* species, using specific thresholds. In addition, we demonstrated a high correlation between MLST, the reference method for the classification of rickettsial isolates, and genome-based tools. The dDDH, OrthoANI and AGIOS can serve as genomic standards for *Rickettsia* species demarcation and would provide valuable information for future reclassification. The obtained results enabled us to develop guidelines for classifying rickettsial isolates at the genus and species levels. The use of genomic tools is therefore perfectly adapted to the taxonomic classification of rickettsial isolates. We thus recommend that any description of a new rickettsial species should include complete genome sequencing.

CONFLICT OF INTEREST

The authors declare no competing interest in relation to this research.

ACKNOWLEDGEMENTS

This study was funded by the Méditerranée-Infection foundation and the French Agence Nationale de la Recherche under reference Investissements d'Avenir Méditerranée Infection 10-IAHU-03.

5 Reference

1. DA ROCHA-LIMA H. Zur Aetiologie des Fleckfeibers. Berl Klin Wochenschr. 1916;53(0):567–9.

2. Ngwamidiba M, Raoult D, Fournier PE. Rickettsia: history and current position. Antibiotiques. 2006 May 1;8(2):117–31.

3. SKERMAN VBD, McGowan V, Sneath PHA. Approved lists of bacterial names. Int J Syst Evol Microbiol. 1980;30(1):225–420.

4. Bergey DH, Krieg NR, Holt JG. Order I. Rickettsiales Gieszczkiewicz 1939. Baltimore, MD: Williams & Wilkins; 1984. 687-703 p. (Bergey's manual of systematic bacteriology).

5. Raoult D, Roux V. Rickettsioses as paradigms of new or emerging infectious diseases. Clin Microbiol Rev. 1997;10(4):694–719.

6. Weisburg WG, Dobson ME, Samuel JE, Dasch GA, Mallavia LP, Baca O, et al. Phylogenetic diversity of the Rickettsiae. J Bacteriol. 1989;171(8):4202–6.

7. Fournier P-E, Raoult D. Current Knowledge on Phylogeny and Taxonomy of Rickettsia spp. Ann N Y Acad Sci. 2009 May;1166(1):1–11.

8. Neimark H, Johansson KE, Rikihisa Y, Tully JG. Proposal to transfer some members of the genera Haemobartonella and Eperythrozoon to the genus Mycoplasma with descriptions of "Candidatus Mycoplasma haemofelis", "Candidatus Mycoplasma haemomuris", "Candidatus Mycoplasma haemosuis" and "Candidatus Mycoplasma wenyonii." Int J Syst Evol Microbiol. 2001 May;51(Pt 3):891–9.

9. Roux V, Bergoin M, Lamaze N, Raoult D. Reassessment of the taxonomic position of Rickettsiella grylli. Int J Syst Bacteriol. 1997 Oct;47(4):1255–7.

10. Birtles RJ, Harrison TG, Saunders NA, Molyneux DH. Proposals to unify the genera Grahamella and Bartonella, with descriptions of Bartonella talpae comb. nov., Bartonella peromysci comb. nov., and three new species, Bartonella grahamii sp. nov., Bartonella taylorii sp. nov., and Bartonella doshiae sp. nov. Int J Syst Bacteriol. 1995 Jan;45(1):1–8.

11. Brenner DJ, O'Connor SP, Winkler HH, Steigerwalt AG. Proposals to unify the genera Bartonella and Rochalimaea, with descriptions of Bartonella quintana comb. nov., Bartonella vinsonii comb. nov., Bartonella henselae comb. nov., and Bartonella elizabethae comb. nov., and to remove the family Bartonellaceae from the order Rickettsiales. Int J Syst Bacteriol. 1993 Oct;43(4):777–86.

12. Tamura A, Ohashi N, Urakami H, Miyamura S. Classification of Rickettsia tsutsugamushi in a new genus, Orientia gen. nov., as Orientia tsutsugamushi comb. nov. Int J Syst Bacteriol. 1995 Jul;45(3):589–91.

13. Dumler JS, Barbet AF, Bekker CP, Dasch GA, Palmer GH, Ray SC, et al. Reorganization of genera in the families Rickettsiaceae and Anaplasmataceae in the order Rickettsiales: unification of some species of Ehrlichia with Anaplasma, Cowdria with Ehrlichia and Ehrlichia with Neorickettsia, descriptions of six new species combinations and designation of Ehrlichia equi and "HGE agent" as subjective synonyms of Ehrlichia phagocytophila. Int J Syst Evol Microbiol. 2001 Nov;51(Pt 6):2145–65.

14. Drancourt M, Raoult D. Taxonomic position of the rickettsiae: current knowledge. FEMS Microbiol Rev. 1994 Jan;13(1):13–24.

15. Fournier P-E, Dumler JS, Greub G, Zhang J, Wu Y, Raoult D. Gene Sequence-Based Criteria for Identification of New Rickettsia Isolates and Description of Rickettsia heilongjiangensis sp. nov. J Clin Microbiol. 2003 Dec 1;41(12):5456–65.

16. Philip RN, Casper EA, Burgdorfer W, Gerloff RK, Hughes LE, Bell EJ. Serologic typing of rickettsiae of the spotted fever group by microimmunofluorescence. J Immunol Baltim Md 1950. 1978 Nov;121(5):1961–8.

17. Grimont PA. Use of DNA reassociation in bacterial classification. Can J Microbiol. 1988 Apr;34(4):541–6.

18. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, et al. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Evol Microbiol. 1987;37(4):463–4.

19. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. 2014 Feb 1;64(Pt 2):346–51.

20. Fournier P-E, Raoult D. Current Knowledge on Phylogeny and Taxonomy of Rickettsia spp. Ann N Y Acad Sci. 2009 May;1166(1):1–11.

21. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol. 2018 Jan 1;68(1):461–6.

22. Chan JZ, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ. Defining bacterial species in the genomic era: insights from the genus Acinetobacter. BMC Microbiol. 2012;12(1):302.

23. Padmanabhan R, Mishra AK, Raoult D, Fournier P-E. Genomics and metagenomics in medical microbiology. J Microbiol Methods. 2013 Dec;95(3):415–24.

24. Chun J, Rainey FA. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int J Syst Evol Microbiol. 2014 Feb 1;64(Pt 2):316–24.

25. Klenk H-P, Meier-Kolthoff JP, Göker M. Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. Int J Syst Evol Microbiol. 2014 Feb 1;64(2):352–6.

26. Meier-Kolthoff JP, G?ker M, Spr?er C, Klenk H-P. When should a DDH experiment be mandatory in microbial taxonomy? Arch Microbiol. 2013 Jun;195(6):413–8.

27. Klappenbach JA, Goris J, Vandamme P, Coenye T, Konstantinidis KT, Tiedje JM. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007 Jan 1;57(1):81–91.

28. Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics. 2016 Mar 15;32(6):929–31.

29. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci. 2009;106(45):19126–31.

30. Lee I, Ouk Kim Y, Park S-C, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol. 2016 Feb 1;66(2):1100–3.

31. Konstantinidis KT, Tiedje JM. Towards a Genome-Based Taxonomy for Prokaryotes. J Bacteriol. 2005 Sep 15;187(18):6258–64.

32. Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol. 2014 Feb 1;64(Pt 2):384–91.

33. Fournier P-E, Belghazi L, Robert C, Elkarkouri K, Richards AL, Greub G, et al. Variations of Plasmid Content in Rickettsia felis. Herman C, editor. PLoS ONE. 2008 May 28;3(5):e2289.

34. Gillespie JJ, Beier MS, Rahman MS, Ammerman NC, Shallom JM, Purkayastha A, et al. Plasmids and Rickettsial Evolution: Insight from Rickettsia felis. Snel B, editor. PLoS ONE. 2007 Mar 7;2(3):e266.

35. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013;14(1):60.

36. Auch AF, von Jan M, Klenk H-P, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci. 2010 Jan 28;2(1):117–34.

37. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.

38. Rodriguez-R LM, Konstantinidis KT. Bypassing cultivation to identify bacterial species. Microbe. 2014;9(3):111–8.

39. Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol. 2014 Feb 1;64(Pt 2):384–91.

40. Lechner M, Findei\s s S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC Bioinformatics. 2011;12(1):124.

41. Simser JA, Palmer AT, Fingerle V, Wilske B, Kurtti TJ, Munderloh UG. Rickettsia monacensis sp. nov., a Spotted Fever Group Rickettsia, from Ticks (Ixodes ricinus) Collected in a European City Park. Appl Environ Microbiol. 2002 Sep 1;68(9):4559–66.

42. Gillespie JJ, Joardar V, Williams KP, Driscoll T, Hostetler JB, Nordberg E, et al. A Rickettsia Genome Overrun by Mobile Genetic Elements Provides Insight into the Acquisition of Genes Characteristic of an Obligate Intracellular Lifestyle. J Bacteriol. 2012 Jan 15;194(2):376–94.

43. serotypic.pdf.

44. Padgett KA, Bonilla D, Eremeeva ME, Glaser C, Lane RS, Porse CC, et al. The Eco-epidemiology of Pacific Coast Tick Fever in California. Lopez JE, editor. PLoS Negl Trop Dis. 2016 Oct 5;10(10):e0005020.

45. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

454 46. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
455 Improvements in Performance and Usability. Mol Biol Evol. 2013 Apr 1;30(4):772–80.

456 47. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary
457 Genetics Analysis Version 6.0. Mol Biol Evol. 2013 Dec;30(12):2725–9.

458 48. Davis GE, Parker RR. Comparative Experiments on Spotted Fever and Boutonneuse Fever (I).
459 Public Health Rep 1896-1970. 1934;49(13):423–8.

460 49. Pickens EG, Bell EJ, Lackman DB, Burgdorfer W. Use of Mouse Serum in Identification and
461 Serologic Classification of Rickettsia Akari and Rickettsia Australis. J Immunol. 1965 Jun
462 1;94(6):883–9.

463 50. Lackman DB, Bell EJ, Stoenner HG, Pickens EG. THE ROCKY MOUNTAIN SPOTTED
464 FEVER GROUP OF RICKETTSIAS. Health Lab Sci. 1965 Jul;2:135–41.

465 51. Walker DH, Liu QH, Yu XJ, Li H, Taylor C, Feng HM. Antigenic diversity of Rickettsia
466 conorii. Am J Trop Med Hyg. 1992 Jul;47(1):78–86.

467 52. Xu W, Raoult D. Taxonomic relationships among spotted fever group rickettsiae as revealed
468 by antigenic analysis with monoclonal antibodies. J Clin Microbiol. 1998 Apr;36(4):887–96.

469 53. Merhej V, Raoult D. Rickettsial evolution in the light of comparative genomics. Biol Rev.
470 2011 May;86(2):379–405.

471 54. Garrity GM. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There
472 Yet? Kraft CS, editor. J Clin Microbiol. 2016 Aug;54(8):1956–63.

473 55. Qin Q-L, Xie B-B, Zhang X-Y, Chen X-L, Zhou B-C, Zhou J, et al. A Proposed Genus
474 Boundary for the Prokaryotes Based on Genomic Insights. J Bacteriol. 2014 Jun
475 15;196(12):2210–5.

476 56. Gupta A, Sharma VK. Using the taxon-specific genes for the taxonomic classification of
477 bacterial genomes. BMC Genomics. 2015 May 20;16:396.

478 57. Thompson CC, Vicente A, Souza RC, Vasconcelos A, Vesth T, Alves N, et al. Genomic
479 taxonomy of vibrios. BMC Evol Biol. 2009;9(1):258.

480 58. Thompson CC, Vieira NM, Vicente ACP, Thompson FL. Towards a genome based taxonomy
481 of Mycoplasmas. Infect Genet Evol. 2011 Oct 1;11(7):1798–804.

482

483 **Table 1: List of 139 genomes used in this study**

Species	Strain	Status	Genome size (Mb)	Accession no.
***Rickettsia* species with standing in nomenclature**				
Rickettsia aeschlimannii	MC16	WGS	1.31	CCER00000000
Rickettsia africae	ESF-5	Complete	1.28	CP001612
Rickettsia akari	Hartford	Complete	1.23	CP000847
Rickettsia amblyommatis	Ac37	Complete	1.46	NZ_CP012420
Rickettsia amblyommatis	AcPa	WGS	1.44	LANR00000000
Rickettsia amblyommatis	Darkwater	WGS	1.44	LAOH00000000
Rickettsia amblyommatis	GAT-30V	Complete	1.48	NC_017028
Rickettsia asembonensis	NMRCii	WGS	1.36	JWSW00000000
Rickettsia australis	Phillips	WGS	1.32	AKVZ00000000
Rickettsia australis	Cutlack	Complete	1.33	NC_017058
Rickettsia bellii	RML An4	WGS	1.54	LAOI00000000
Rickettsia bellii	RML Mog	WGS	1.62	LAOJ00000000
Rickettsia bellii	OSU 85-389	Complete	1.52	NC_009883
Rickettsia bellii	RML369-C	Complete	1.52	NC_007940
Rickettsia canadensis	CA410	Complete	1.15	NC_016929
Rickettsia canadensis	McKiel	Complete	1.16	NC_009879
Rickettsia conorii	Malish 7	Complete	1.27	NC_003103
Rickettsia conorii	A-167	WGS	1.26	AJUR00000000
Rickettsia conorii	ITTR	WGS	1.25	AJHC00000000
Rickettsia conorii	ISTT CDC1	WGS	1.25	AJVP00000000
Rickettsia felis	LSU	WGS	1.54	JSEM00000000
Rickettsia felis	LSU lb	WGS	1.58	JSEL00000000
Rickettsia felis	Pedreira	WGS	1.49	LANQ00000000
Rickettsia felis	URRWXCal2	Complete	1.49	NC_007109
Rickettsia gravesii	BWI-1	WGS	1.35	AWXL00000000
Rickettsia heilongjiangensis	O54	Complete	1.28	CP002912
Rickettsia helvetica	C9P9	WGS	1.37	CM001467
Rickettsia honei	RB	WGS	1.27	AJTT00000000
Rickettsia hoogstraalii	Croatica	WGS	1.48	CCXM00000000
Rickettsia japonica	YH	Complete	1.28	NC_016050
Rickettsia massiliae	AZT80	Complete	1.28	NC_016931
Rickettsia massiliae	MTU5	Complete	1.37	NC_009900
*Rickettsia rhipicephali**	Ect	WGS	1.27	LAOC00000000
Rickettsia montanensis	OSU 85-930	Complete	1.28	CP003340
Rickettsia parkeri	AT#24	WGS	1.3	LAOL00000000
Rickettsia parkeri	GrandBay	WGS	1.31	LAOK00000000
Rickettsia parkeri	Portsmouth	Complete	1.3	NC_017044
Rickettsia parkeri	TatesHell	WGS	1.3	LAOO00000000
Rickettsia peacockii	Rustic	Complete	1.29	CP001227
Rickettsia prowazekii	Breinl	Complete	1.11	NC_020993
Rickettsia prowazekii	BuV67-CWPP	Complete	1.11	NC_017056
Rickettsia prowazekii	Cairo3	WGS	1.11	APMO00000000
Rickettsia prowazekii	Chernikova	Complete	1.11	NC_017049
Rickettsia prowazekii	Dachau	Complete	1.11	CP003394
Rickettsia prowazekii	GvV257	Complete	1.11	NC_017048
Rickettsia prowazekii	Katsinyian	Complete	1.11	NC_017050
Rickettsia prowazekii	Madrid E	Complete	1.11	NC_000963

Rickettsia prowazekii	NMRC Madrid E	Complete	1.11	NC_020992
Rickettsia prowazekii	Rp22	Complete	1.11	NC_017560
Rickettsia prowazekii	RpGvF24	Complete	1.11	NC_017057
Rickettsia raoultii	Khabarovsk	Complete	1.34	CP010969
Rickettsia rhipicephali	3-7-female6-CWPP	Complete	1.31	NC_017042
Rickettsia rhipicephali	HJ#5	Complete	1.45	NZ_CP013133
Rickettsia rickettsii	Arizona	Complete	1.27	NC_016909
Rickettsia rickettsii	Brazil	Complete	1.25	NC_016913
Rickettsia rickettsii	Colombia	Complete	1.27	NC_016908
Rickettsia rickettsii	Hauke	Complete	1.27	NC_016911
Rickettsia rickettsii	Hino	Complete	1.27	NC_016914
Rickettsia rickettsii	Hlp#2	Complete	1.27	NC_016915
Rickettsia rickettsii	Iowa	Complete	1.27	NC_010263
Rickettsia rickettsii	Morgan	Complete	1.27	NZ_CP006010
Rickettsia rickettsii	R	Complete	1.26	NZ_CP006009
Rickettsia rickettsii	Sheila Smith	Complete	1.26	NC_009882
Rickettsia sibirica	246	WGS	1.25	AABW00000000
Rickettsia sibirica	HA-91	WGS	1.25	AHZB00000000
Rickettsia sibirica	BJ-90	WGS	1.25	AHIZ00000000
Rickettsia slovaca	D-CWPP	Complete	1.27	NC_017065
Rickettsia slovaca	13-B	Complete	1.27	NC_016639
Rickettsia tamurae	AT-1	WGS	1.45	CCMG00000000
Rickettsia typhi	B9991CWPP	Complete	1.11	NC_017062
Rickettsia typhi	TH1527	Complete	1.11	NC_017066
Rickettsia typhi	Wilmington	Complete	1.11	NC_006142
Rickettsial strains from as yet unvalidated species				
Rickettsia argasii	T170-B	WGS	1.44	LAOQ00000000
Rickettsia endosymbiont of Ixodes scapularis		WGS	1.82	CM000770
Rickettsia endosymbiont of Ixodes pacificus	Humboldt	WGS	1.56	LAOP00000000
Rickettsia fournieri	AUS118	WGS	1.45	OFAL00000000
Rickettsia monacensis	IrR/Munich	Complete	1.35	NZ_LN794217
Rickettsia philipii	364D	Complete	1.29	CP003308
Species from closely related genera				
Anaplasma phagocytophilum	BOV-10_179	WGS	1.37	CCXQ00000000
Anaplasma phagocytophilum	Annie	WGS	1.52	LAON00000000
Anaplasma phagocytophilum	ApMUC09	WGS	1.52	LANV00000000
Anaplasma phagocytophilum	ApNP	WGS	1.52	LANW00000000
Anaplasma phagocytophilum	ApNYW	WGS	1.50	LAOG00000000
Anaplasma phagocytophilum	ApWI1	WGS	1.50	LAOF00000000
Anaplasma phagocytophilum	C1	WGS	1.68	FLLR00000000
Anaplasma phagocytophilum	C2	WGS	1.64	FLMA00000000
Anaplasma phagocytophilum	C3	WGS	1.56	FLMB00000000
Anaplasma phagocytophilum	C4	WGS	1.60	FLLZ00000000
Anaplasma phagocytophilum	C5	WGS	1.72	FLMD00000000
Anaplasma phagocytophilum	CR1007	WGS	1.50	LASO00000000
Anaplasma phagocytophilum	CRT35	WGS	1.45	JFBI00000000
Anaplasma phagocytophilum	CRT38	WGS	1.51	APHI00000000
Anaplasma phagocytophilum	CRT53	WGS	1.57	LAOD00000000
Anaplasma phagocytophilum	Dog2	Complete	1.47	NC_021881
Anaplasma phagocytophilum	H1	WGS	1.17	FLMF00000000

Anaplasma phagocytophilum	HGE1	WGS	1.47	APHH00000000
Anaplasma phagocytophilum	HGE1mutant	WGS	1.49	LASP00000000
Anaplasma phagocytophilum	HGE2	WGS	1.48	LAOE00000000
Anaplasma phagocytophilum	HZ	Complete	1.47	NC_007797
Anaplasma phagocytophilum	HZ2	Complete	1.48	NC_021879
Anaplasma phagocytophilum	JM	Complete	1.48	NC_021880
Anaplasma phagocytophilum	MRK	WGS	1.48	JFBH00000000
Anaplasma phagocytophilum	NCH-1	WGS	1.50	LANT00000000
Anaplasma phagocytophilum	Norway variant2	Complete	1.55	NZ_CP015376
Anaplasma phagocytophilum	RD1	WGS	1.59	FLME00000000
Anaplasma phagocytophilum	Webster	WGS	1.48	LANS00000000
Ehrlichia canis	Jake	Complete	1.32	NC_007354
Ehrlichia chaffeensis	Arkansas	Complete	1.18	NC_007799
Ehrlichia chaffeensis	Heartland	Complete	1.17	NZ_CP007473
Ehrlichia chaffeensis	Jax	Complete	1.18	NZ_CP007475
Ehrlichia chaffeensis	Liberty	Complete	1.18	NZ_CP007476
Ehrlichia chaffeensis	Osceola	Complete	1.18	NZ_CP007477
Ehrlichia chaffeensis	Sapulpa	WGS	1.01	AAIF00000000
Ehrlichia chaffeensis	Saint Vincent	Complete	1.17	NZ_CP007478
Ehrlichia chaffeensis	Wakulla	Complete	1.17	NZ_CP007479
Ehrlichia chaffeensis	WestPaces	Complete	1.17	NZ_CP007480
Ehrlichia muris	AS145	Complete	1.20	NC_023063
Ehrlichia muris	EmCRT	WGS	1.15	LANU00000000
Ehrlichia ruminantium	Crystal Springs	WGS	1.48	BDDK00000000
Ehrlichia ruminantium	Gardel	Complete	1.50	NC_006831
Ehrlichia ruminantium	Kerr Seringe	WGS	1.45	BDDL00000000
Ehrlichia ruminantium	Palm River	WGS	1.49	LUFS00000000
Ehrlichia ruminantium	Pokoase	WGS	1.47	BDDM00000000
Ehrlichia ruminantium	Sankat430	WGS	1.46	BDDN00000000
Ehrlichia ruminantium	Senegal virulent	WGS	1.45	MQUJ00000000
Ehrlichia ruminantium	Senegalp63	WGS	1.45	MRDC00000000
Ehrlichia ruminantium	Welgevonden	Complete	1.52	NC_005295
Ehrlichia ruminantium	Welgevonden	Complete	1.51	NC_006832
Orientia tsutsugamushi	AFSC4	WGS	1.30	LYMT00000000
Orientia tsutsugamushi	AFSC7	WGS	1.44	LYMB00000000
Orientia tsutsugamushi	Gilliam	WGS	2.00	LANO00000000
Orientia tsutsugamushi	Karp	WGS	1.45	LANM00000000
Orientia tsutsugamushi	Karp	WGS	2.03	LYMA00000000
Orientia tsutsugamushi	Kato	WGS	1.48	LANN00000000
Orientia tsutsugamushi	Sido	WGS	7.13	LAOM00000000
Orientia tsutsugamushi	UT144	WGS	1.69	LAOR00000000
Orientia tsutsugamushi	UT716	WGS	2.22	LAOA00000000
Orientia tsutsugamushi	Boyond	Complete	2.12	NC_009488
Orientia tsutsugamushi	Ikead	Complete	2.01	NC_010793

484

Table 2: Range of dDDH, OrthoANI and AGIOS values of the unvalidated *Rickettsia* isolates (6 genomes) and species (61 genomes) of the genera *Orientia*, *Ehrlichia* and *Anaplasma* with the 28 validated *Rickettsia* species (72 genomes) used to establish the taxono-genomic criteria.

Species name	Strain name	Range of Pairwise comparison (%)			
		dDDH	OrthoANI	AAI	AGIOS
Unvalidated *Rickettsia* isolates					
R.argasii	T170-B	25.8 - 94.7	81.22 - 99.22	/	80.64 - 98.97
R.endosymbiont of Ixodes scapularis	-	29 - 75.3	82.07 - 97.68	/	80.87 - 97.65
R.endosymbiont of Ixodes pacificus	Humboldt	25.2 - 81.3	81.22 - 98.09	/	80.36 - 98.25
R.fournieri	AUS118	26 - 90.2	81.37 - 98.98	/	80.93 - 98.55
R.monacensis	IrR/Munich	25.5 - 81.3	81.54 - 98.02	/	80.60 - 98.14
R.philipii	364D	25.9 - 94.9	81.06 - 99.47	/	80.74 - 98.92
Inter-genera					
A.phagocytophilum	BOV-10_179	25.10 - 26.00	63.00 - 64.28	41.3 - 42.2	56.3 - 58.8
	Annie	23.40 - 25.90	63.08 - 64.09	41.3 - 42.2	56.3 - 58.8
	ApMUC09	23.40 - 25.90	63.07 - 64.23	41.3 - 42.2	56.2 - 58.5
	ApNP	25.20 - 26.00	62.94 - 64.13	41.3 - 42.2	55.9 - 58.2
	ApNYW	23.30 - 25.90	63.20 - 64.10	41.3 - 42.2	56.3 - 58.7
	ApWI1	23.40 - 25.90	63.22 - 64.41	41.3 - 42.2	56.3 - 58.7
	C1	25.10 - 26.00	63.01 - 64.03	41.3 - 42.2	56.3 - 58.8
	C2	25.10 - 26.00	62.96 - 64.05	41.1 - 42.2	56.3 - 58.7
	C3	25.10 - 26.00	63.18 - 64.43	41.3 - 42.2	56.3 - 58.7
	C4	25.10 - 26.00	62.87 - 64.12	41.2 - 42.2	56.2 - 58.7
	C5	25.10 - 29.00	63.00 - 64.02	40.9 - 42.2	56.3 - 58.8
	CR1007	23.40 - 25.90	62.86 - 64.04	41.4 - 42.4	56.3 - 58.6
	CRT35	23.40 - 26.00	62.97 - 63.86	41.4 - 42.4	56.4 - 58.7
	CRT38	23.80 - 26.30	62.99 - 64.13	41.4 - 42.4	56.2 - 58.7
	CRT53	22.50 - 26.10	62.93 - 64.07	41.4 - 42.4	56.4 - 58.7
	Dog2	23.40 - 25.90	62.77 - 64.34	41.4 - 42.4	56.3 - 58.7
	H1	23.30 - 25.80	63.03 - 64.36	41.4 - 42.4	56.3 - 58.7
	HGE1	23.40 - 25.90	63.13 - 64.07	41.4 - 42.4	56.3 - 58.7
	HGE1mutant	23.40 - 25.90	63.16 - 64.15	41.4 - 42.4	56.3 - 58.7
	HGE2	23.40 - 25.90	63.10 - 64.18	41.4 - 42.4	56.3 - 58.7
	HZ	23.30 - 25.90	63.02 - 63.99	41.4 - 42.4	56.3 - 58.7
	HZ2	23.30 - 25.90	63.02 - 64.09	41.4 - 42.4	56.3 - 58.7
	JM	23.30 - 25.90	62.96 - 64.13	41.4 - 42.4	56.5 - 58.9
	MRK	23.40 - 26.00	62.82 - 64.13	41.4 - 42.4	56.4 - 58.9
	NCH-1	23.30 - 25.80	62.77 - 63.71	41.4 - 42.4	56.4 - 58.8
	Norway variant2	25.10 - 26.00	63.15 - 64.38	41.4 - 42.4	56.4 - 59.0
	RD1	23.20 - 29.80	63.15 - 64.67	41.4 - 42.4	56.5 - 59.0
	Webster	23.30 - 25.90	62.88 - 64.17	41.4 - 42.4	56.5 - 58.9
E. canis	Jake	23.10 - 26.00	64.46 - 65.49	42.7 - 43.5	59.8 - 63.1
E. chaffeensis	Arkansas	25.50 - 26.10	63.92 - 65.34	42.7 - 43.6	59.8 - 63.0
	Heartland	25.50 - 26.10	64.07 - 65.32	42.7 - 43.6	59.7 - 62.9
	Jax	25.50 - 26.10	64.02 - 65.40	42.7 - 43.6	59.7 - 62.8

	Liberty	25.50 - 26.10	64.18 - 65.33	42.7 - 43.6	59.7 - 63.0
	Osceola	25.50 - 26.10	64.24 - 65.44	42.7 - 43.6	59.8 - 63.0
	Sapulpa	25.50 - 26.30	64.46 - 65.68	42.7 - 43.6	59.8 - 62.7
	Saint Vincent	25.50 - 26.10	63.97 - 65.27	42.7 - 43.6	59.7 - 63.0
	Wakulla	25.50 - 26.10	64.07 - 65.44	42.7 - 43.6	59.7 - 62.9
	WestPaces	25.50 - 26.10	63.82 - 64.95	42.7 - 43.6	59.7 - 62.9
E. muris	AS145	24.80 - 25.80	64.36 - 65.50	42.6 - 43.9	59.6 - 63.0
	EmCRT	24.80 - 25.80	64.17 - 65.50	42.6 - 43.9	59.6 - 63.0
E. ruminantium	Crystal Springs	25.80 - 26.70	63.97 - 65.13	42.4 - 43.4	59.7 - 63.0
	Gardel	25.80 - 26.70	64.04 - 65.11	42.4 - 43.4	59.7 - 63.0
	Kerr Seringe	25.70 - 26.60	64.13 - 65.40	42.4 - 43.4	59.7 - 63.0
	Palm River	25.70 - 26.70	64.09 - 65.35	42.4 - 43.4	59.7 - 63.0
	Pokoase	25.70 - 26.60	64.11 - 65.22	42.4 - 43.4	59.7 - 62.7
	Sankat430	25.70 - 26.60	63.78 - 65.12	42.4 - 43.4	59.8 - 63.0
	Senegal virulent	25.70 - 26.60	63.97 - 65.16	42.4 - 43.4	59.8 - 63.0
	Senegalp63	25.70 - 26.60	63.97 - 65.19	42.4 - 43.4	59.8 - 63.0
	Welgevonden	25.80 - 26.70	64.04 - 65.31	42.4 - 43.4	59.9 - 63.0
	Welgevonden	25.70 - 26.60	63.98 - 65.25	42.4 - 43.4	60.0 - 63.0
O. tsutsugamushi	AFSC4	24.10 - 35.40	65.36 - 66.50	48.3 - 49.5	62.6 - 65.7
	AFSC7	26.10 - 35.70	65.43 - 66.49	48.3 - 49.5	62.5 - 65.7
	Gilliam	23.90 - 39.20	65.51 - 66.54	47.3 - 49.5	61.5 - 64.7
	Karp	23.20 - 35.60	65.28 - 66.28	48.3 - 49.5	62.6 - 65.7
	Karp	19.70 - 35.60	65.40 - 66.61	48.3 - 49.5	62.5 - 65.7
	Kato	23.30 - 36.10	65.38 - 66.56	48.3 - 49.5	62.5 - 65.7
	Sido	28.50 - 41.50	65.32 - 66.98	47.1 - 49.5	61.7 - 65.2
	UT144	25.00 - 36.90	65.14 - 66.08	47.8 - 49.5	62.2 - 65.3
	UT716	21.50 - 38.20	65.48 - 66.51	48.0 - 49.5	62.4 - 65.6
	Boyond	24.60 - 36.00	65.37 - 66.35	48.2 - 49.5	62.0 - 65.5
	Ikead	24.30 - 36.20	65.46 - 66.8	48.3 - 49.5	62.1 - 65.6

488

Legends figures :

Figure 1 : Clusters obtained from pairwise similarity analysis of 72 genomes of 28 validated *Rickettsia* species based on digital DDH with recommended cutoff 70% for species demarcation.

Figure 2 : Clusters obtained from pairwise similarity analysis of 72 genomes of 28 validated *Rickettsia* species based on OrthoANI with recommended cutoff 95~96 **for species demarcation.**

Figure 3 : Proposal genomic scheme for classification of the rickettsiae at the genus and species levels.

Figure 4 : Relationships between dDDH, OrthoANI, AGIOS values and 16S rRNA, *gltA*, *sca4*, *ompA* and *ompB* gene sequence similarity for pairs of genomes among the 28 *Rickettsia* species (72 genomes). Each filled circle represents one hand the value for 16S rRNA gene identity between two strains (y-axis), plotted against the dDDH values between the strains (A), the OrthoANI values between the strains (B) and the AGIOS values between the strains (C). On the other hand the *gltA* gene identity between two strains (y-axis), plotted against the OrthoANI values between the strains (D) and the AGIOS values between the strains (E) and finally, the OrthoANI values between two strains (y-axis), plotted against the AGIOS values between the strains (F). The relationships of OrthoANI, AGIOS and dDDH to *sca4*, *ompA and ompB* genes (G). The relationships of OrthoANI, AGIOS and *gltA* gene to dDDH (H). A linear trend line is shown. The horizontal broken lines denote the 98.1, 99.8, 86.5% 16S rRNA and *gltA* genes identities recommendation for *Rickettsia* species delineation, while the vertical broken lines denote the corresponding dDDH (A), OrthoANI (B; D), and AGIOS (C; E) values for linear regression.

Figure 5 : Phylogenomic tree constructed with 591 concatenated core protein sequences from 78 *Rickettsia* genomes (in bold as well as their group affiliation). Sequences were aligned using mafft alignment algorithm. Phylogenetic inference was obtained by Maximum Likehood method with JTT and GAMMA models within the MEGA software and display only topology. Numbers at

the nodes represent the percentages of bootstrap values obtained by repeating analysis 500 times to generate a majority consensus tree. The scale bar represents a 2 % nucleotide sequence divergence.

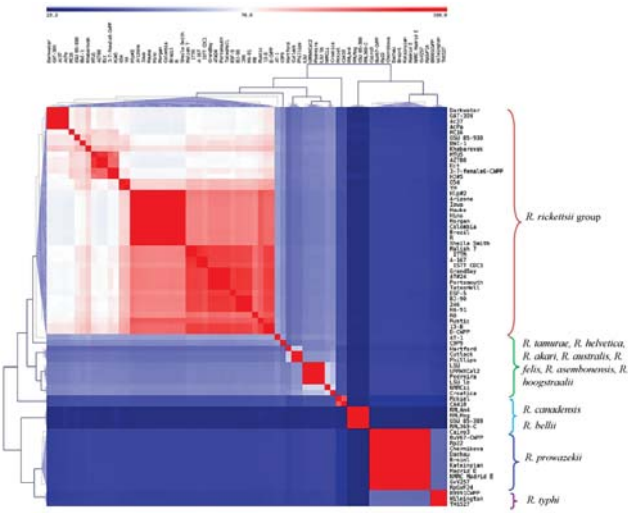


Figure 1 : Clusters obtained from pairwise similarity analysis of 72 genomes of 28 validated *Rickettsia* species based on digital DDH with recommended cutoff 70% for species demarcation.

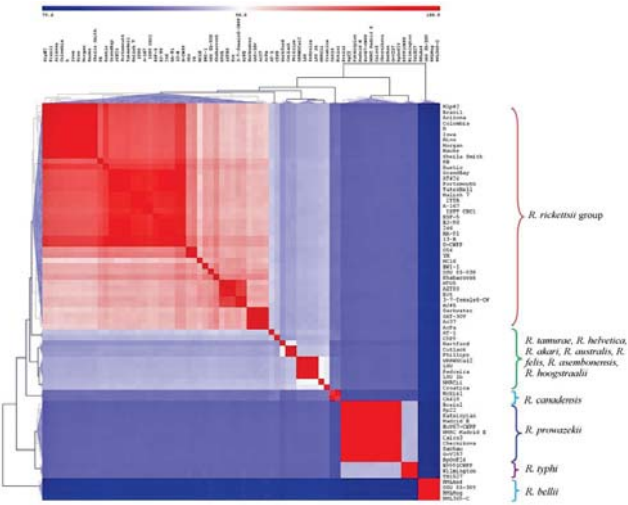
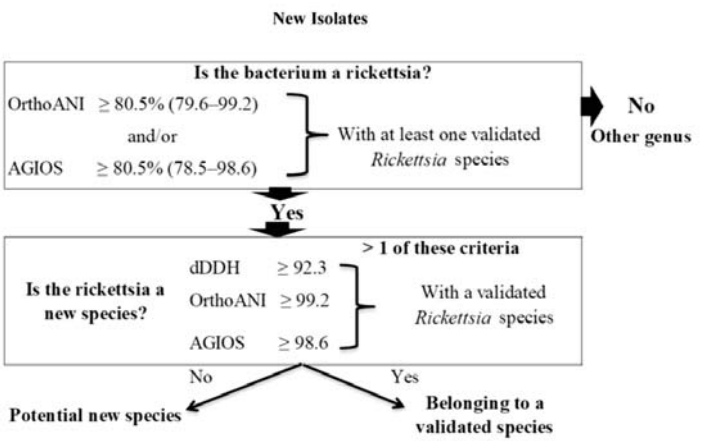


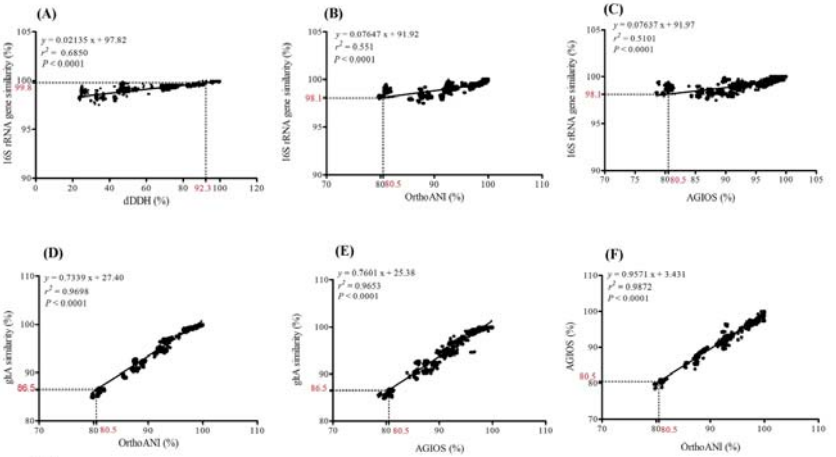
Figure 2 : Clusters obtained from pairwise similarity analysis of 72 genomes of 28 validated *Rickettsia* species based on OrthoANI with recommended cutoff 95~96 **for species demarcation.**



525

526 **Figure 3 : Proposal genomic scheme for classification of the rickettsiae at the genus and**

527 **species levels.**



(G)

genome-based parameters	Genes (r^2)		
	sca4	*ompA*	*ompB*
dDDH	0.9196	0.8800	0.9159
OrthoANI	0.9664	0.9013	0.9836
AGIOS	0.9579	0.8751	0.9633

(H)

Parameters/genes	dDDH (r^2)
OrthoANI	0.8776
AGIOS	0.8623
gltA	0.8172

Figure 4 : Relationships between dDDH, OrthoANI, AGIOS values and 16S rRNA, *gltA*, *sca4*, *ompA* and *ompB* gene sequence similarity for pairs of genomes among the 28 *Rickettsia* species (72 genomes). Each filled circle represents one hand the value for 16S rRNA gene identity between two strains (y-axis), plotted against the dDDH values between the strains (A), the OrthoANI values between the strains (B) and the AGIOS values between the strains (C). On the other hand the *gltA* gene identity between two strains (y-axis), plotted against the OrthoANI values between the strains (D) and the AGIOS values between the strains (E) and finally, the OrthoANI values between two strains (y-axis), plotted against the AGIOS values between the strains (F). The relationships of OrthoANI, AGIOS and dDDH to *sca4*, *ompA and ompB* genes (G). The relationships of OrthoANI, AGIOS and *gltA* gene to dDDH (H). A linear trend line is shown. The horizontal broken lines denote the 98.1, 99.8, 86.5% 16S rRNA and *gltA* genes identities recommendation for *Rickettsia* species delineation, while the vertical broken lines denote the corresponding dDDH (A), OrthoANI (B; D), and AGIOS (C; E) values for linear regression.

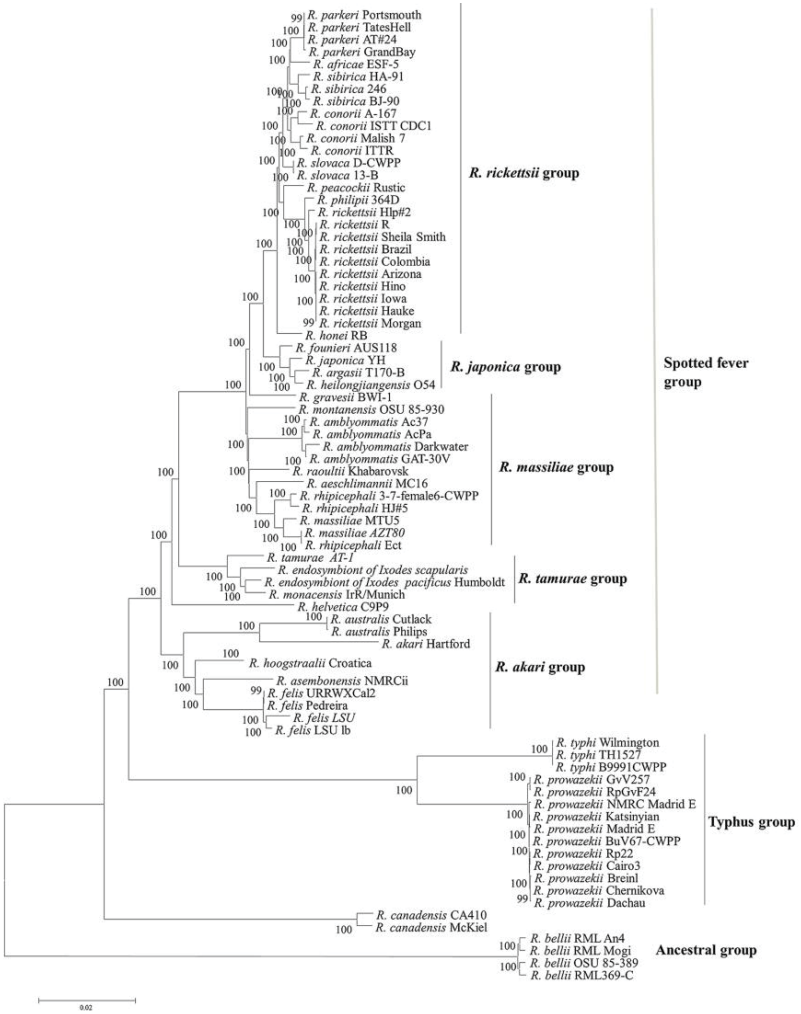


Figure 5 : Phylogenomic tree constructed with 591 concatenated core protein sequences from 78 *Rickettsia* genomes (in bold as well as their group affiliation). Sequences were aligned using mafft alignment algorithm. Phylogenetic inference was obtained by Maximum Likehood method with JTT and GAMMA models within the MEGA software and display only topology. Numbers at the nodes represent the percentages of bootstrap values obtained by repeating analysis 500 times to generate a majority consensus tree. The scale bar represents a 2 % nucleotide sequence divergence.

Article 4:

Rickettsia fournieri sp. nov. strain AUS118^{T}, a novel spotted fever group rickettsia first isolated from *Argas lagenoplastis* ticks in Australia.

Awa Diop, Stephen C. Barker, Mey Eberhard, Barker Dayana, Thi Tien Nguyen, Fabrizio Di Pinto, Didier Raoult, Oleg Mediannikov

[Submitted in International Journal of Systematic and Evolutionary Microbiology]

Rickettsia fournieri sp. nov. strain AUS118^T, a novel spotted fever group rickettsia from *Argas lagenoplastis* ticks in Australia.

Awa Diop[1], Stephen C. Barker[2], Eberhard Mey [2], Dayana Campelo[2], Thi Tien Nguyen[1], Fabrizio di Pinto[3], Didier Raoult[3], Oleg Mediannikov[3,*]

[1]UMR VITROME, Aix-Marseille University, IRD, Service de Santé des Armées, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France Tel: +33 413 732 401, Fax: +33 413 732 402.

[2]Department of Parasitology, School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane QLD 4072, Queensland, Australia Tel: +61 33 65 33 03.

[3]UMR MEPHI, Aix-Marseille University, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France

[*]Corresponding author: Dr Oleg Mediannikov

[3]UMR MEPHI, Aix-Marseille University, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France Tel: +33 413 732 401, Fax: +33 413 732 402.

Email: olegusss1@gmail.com

Running title: *Rickettsia fournieri* sp. nov.

Abstract words count: 186

Text word count: 2651

Keywords: *Rickettsia fournieri;* new species*; Argas lagenoplastis; Ixodida; Australia.*

Abstract:

A novel spotted fever group *Rickettsia* was found in bird-associated ticks, *Argas lagenoplastis,* collected from the nests of *Petrochelidon ariel* (fairy martin) in Australia in 2013. Based on the presence of this *Rickettsia* (strain AUS118^T) in tick tissues and cell cultures, confirmed by transmission electron microscopy, and analysis of its phylogenetic, genotypic and phenotypic relationships with type strains *Rickettsia* type strains, strain AUS118^T was sufficiently divergent to be classified within a novel species. Multigene sequences and the core proteins analyses, showed that strain AUS118^T was most closely related to *Rickettsia japonica* and *R. heilongjiangensis* of the spotted fever group. Furthermore, strain AUS118^T has levels of sequence similarity with its both closely related species respectively of 99.79, 99.52, 98.94, 97.12 and 98.71% and of 99.72, 99.60, 98.99, 97.80 and 98.6%, for the 16S rRNA, citrate synthase, *sca4*, *ompA*, and *ompB* genes, respectively. This supported also the new species status of this strain. Regarding its specific genotypic and phenotypic characteristics, we propose the creation of a novel species named *Rickettsia fournieri* sp. nov. Strain AUS118^T (DSM 28985 and CSUR R501) is the type strain of *Rickettsia fournieri* sp.nov.

Introduction

Rickettsia species are obligate Gram-negative intracellular α-proteobacteria associated with arthropod-vectors worldwide [1, 2]; many species of which can also infect mammalian hosts, mostly through arthropod bites or feces. Currently, there are at least 30 officially validated species in the genus (http://www.bacterio.net/rickettsia.html). However, numerous other putative species have also been recently proposed from molecular characterization of rickettsiae at three or four gene loci. On the basis of clinical, genotypic and phenotypic features, *Rickettsia* species were initially classified into two groups: (i) the spotted fever group (SFG) that currently contains pathogenic agents that cause spotted fevers as well as species of as-yet unknown pathogenicity associated with ticks, fleas and mites; and (ii) the typhus group (TG) that cause epidemic and murine typhus and associated with human body lice and rat fleas correspondingly. The SFG group has been further divided into phylogenetic subgroups on the basis of gene sequence comparisons [3–5]. *Rickettsia* species cause mild to severe diseases, the most common being scalp eschar and neck lymphadenopathy (SENLAT), Mediterranean spotted fever (MSF), Far Eastern spotted fever, Rocky Mountain spotted fever (RMSF), and African tick-bite fever [6–8]. In addition to spotted fever and typhus group rickettsiae, two species, *R. bellii* and *R. canadensis*, are associated with ticks and insects but do not cause any recognized human disease to date.

The majority of SFG rickettsiae are associated with ticks that serve as a vectors and often reservoir [9, 10]. The SFG rickettsiae species known to occur in Australia are *R. australis*, the aetiological agent of Queensland tick typhus (QTT) transmitted by *Ixodes holocyclus* and *Ixodes tasmani*, [11–15] *R. honei*, the agent of Flinders Island spotted fever transmitted by *Bothriocroton hydrosauri* and *R. honei* subsp. *marmionii*, the agent of Australian spotted fever with *Haemaphysalis novaeguineae* as tick-vector [11, 13, 16, 17]. Further rickettsial DNA sequences from *R. felis* were detected in fleas *Ctenocephalides felis* from cats and dogs

in Western Australia (WA) but as yet, no human infections caused by these rickettsiae have been reported in Australia [18, 19]. In addition to these rickettsial pathogens, the existence of a novel spotted fever group (SFG) *Rickettsia*, *R. gravesii* was demonstrated recently in *Amblyomma triguttatum triguttatum* ticks from Barrow Island, Western Australia but no human pathogenicity was described [20].

In the present study, we began to explore *Rickettsia* spp. in Australian soft ticks. Fourteen species of soft ticks (*Argasidae*) are known in Australia [21]: none of these has been examined for *Rickettsia* before the present work. A novel SFG rickettsia has been detected by molecular methods in bird-associated ticks, *Argas lagenoplastis*, collected in Australia and then five rickettsial strains including AUS118^{T} were isolated from these ticks. The creation of the novel species *Rickettsia fournieri* sp. nov. is proposed that includes strain AUS118^{T} as type strain.

In 2013, two hundred and twenty five ticks (one hundred and sixty five live ticks and sixty dead ticks) were collected from abandoned nests of *Petrochelidon ariel*, the fairy martin, in Queensland, Australia (-28.1022694 S, 144.1605377 E, Lake Bindegally, Qld). These were preserved in 70% ethanol for PCR screening or kept alive in sterile conditions for subsequent rickettsial isolation. The ticks were identified as *Argas lagenoplastis* by SCB and DB using standard taxonomic keys [22, 23] Twenty ticks were homogenized and blindly inoculated into a cell culture (XTC-2). DNA from the cell culture suspension supernatant and from homogenized ethanol-preserved ticks was extracted using an EZ-1 automate (Qiagen) and screened for the presence of rickettsiae by previously described quantitative real-time PCR (qPCR) [24]. In total, one hundred and thirty seven of the two hundred and twenty five *Argas lagenoplastis* ticks (60.1%) were PCR-positive for *Rickettsia* spp. DNA. Five randomly chosen ticks were subjected to *Rickettsia*-specific standard PCR assays using primer pairs RpCS.409d and RpCS.1258r (Bioprobe Systems, France) that target a 770-nucleotide region of the citrate synthase-encoding gene (*gltA*) [25]. BLAST searche of the 728 nucleotide obtained sequence, exhibited 99.58% sequence similarity with *R. japonica* strain YH^T (NC_016050); the most closely related species with a validly published name.

Isolation of rickettsial strains from ticks was attempted in XTC-2 cells line using the shell-vial technique [26]. XTC-2 cells were grown in L15 medium (Leibovitz medium) supplemented with 5% (w/v) foetal calf serum (FCS), 5% tryptose phosphate and 2 mmol/l L-glutamine in the atmosphere containing 5% (v/v) CO_2 at 28°C. Cultures were observed weekly under light microscopy. The scraped XTC-2 cells were applied to a microscope slide and the presence of rickettsiae in culture was detected by Giemenez staining [27] and confirmed by *gltA* qPCR as described above. Growth was also tested in L929 cells at 32°C in minimal essential medium supplemented with 2% heat-inactivated fetal calf serum. For electron microscopy analysis (TEM), a 3.5 µL drop of bacterial suspension was applied for

60s to the top of a formvar carbon 400 mesh nickel grid (FCF400-Ni, EMS) which was previously glow discharged. After drying on filter paper, bacteria were immediately stained with 1 % ammonium molybdate (Thermofisher, geel, Belgium) for 1s. Electron micrographs were taken with a Tecnai G20 transmission electron microscope (FEI) operated at 200 Kev. We succeeded in isolating the isolate named strain AUS118^T after seven days of incubation in the entire body of *Argas lagenoplastis* tick subcultured in XTC-2 cell. Growth was observed similarly in L929 cells. No cytopathic effect was observed. Staining by the Gimenez method revealed small, purple-coloured intracellular, rod-shaped bacteria, observed both in the cytoplasm and the nucleus of XTC-2 cells (Fig. 1A). Cells measured a mean size of 1.5µm in length and 0.3µm in width under electron microscopy using a Tecnai G20 operating at 200 keV (Fig. 1B).

Rickettsia species express few phenotypic properties. DNA sequences are highly conserved between different rickettsial species, making the thresholds of 16S rRNA sequence similarity, G + C content and DNA-DNA hybridization relatedness used to define bacterial species [28], inapplicable to the *Rickettsia* species delimitation. Thus, in 2003, a molecular scheme for the taxonomic classification of rickettsial species using a multi-locus sequence typing (MLST) approach based on the 16S rRNA, *gltA*, *sca4*, *sca0* (*ompA*) and *sca5* (*ompB*) genes was proposed [29]. Using this MLST classification scheme, a novel SFG to be confirmed as a new species should not exhibit more than one of the following degrees of nucleotide similarity with of the most homologous established rickettsial species: 99.8, 99.9, 98.8, 99.2 and 99.3% for the above-listed genes, respectively.

The sequences from 16S rRNA, *gltA*, *sca4*, *ompA* and *ompB* genes for strain AUS118^T were amplified and sequenced using the previously described primers and methods [30, 31]. These sequences were compared respectively with those of 27 validated *Rickettsia* species (The Genbank accession numbers of the genome from which the gene sequences were

extracted are indicated in Table 1), by pairwise nucleotide sequence similarity analysis, in order to estimate the genetic differences between *Rickettsia* sp. strain AUS118^T and its closest phylogenetically related species. Pairwise sequence similarities were calculated using the method recommended by Meier-Kolthoff et al. [32] available via the GGDC web server (http://ggdc.dsmz.de/) [33] available at (http://ggdc.dsmz.de/). The nucleotide sequences of the 16S rRNA, *gltA*, *ompA*, *ompB*, and *sca4* genes of *R. fournieri* sp. nov. have been deposited in the EMBL-EBI under accession numbers KF666475, KF666471, KF666477, KF666469, and KF666473, respectively. For the 16S rRNA gene, the level of similarity ranged from 98.10% with *R. akarii* to 99.79 % with *R. japonica* (99.72 % for *R. heilongjiangensis*). For *gltA* and *sca4*, the levels of similarity ranged from 87.17 % with *R. bellii* to 99.60 % with *R. heilongjiangensis* (99.52% for *R. japonica*) and from 82.22 % *R.prowazekii* to 99.00 % with *R. slovaca* (98.99 % for *R. heilongjiangensis*, 98.94 % for *R. japonica*), respectively. For *ompA* and *ompB*, the levels of similarity ranged from 82.40 % with *R. canadensis* to 97.80 % with *R. heilongjiangensis* (97.12 % for *R. japonica*) and from 83.52 % with *R. prowazekii* to 98.71 % with *R. japonica* (98.6% *R. heilongjiangensis*), respectively (Table 1). These values were lower than the cut-offs proposed for *Rickettsia* species definition cited above [29]. Therefore, on the basis of genotypic criteria, *Rickettsia* sp. strain AUS118^T demonstrated enough diversity to be classified as a new *Rickettsia* species.

The phylogenetic relationships of strain AUS118^T with 27 *Rickettsia* species with validly published names were estimated first by aligning sequences from the concatenated 16S rRNA, *gltA*, *sca4*, *ompB* and *ompA* genes using CLUSTALW 2.0 alignment algorithm [34] and second, by aligning sequences from 633 concatenated core proteins using the Mafft alignment algorithm [35]. The phylogenetic trees were inferred by the Maximum Likehood method with the Kimura 2-parameter model for the multigene sequences based tree and with JTT and GAMMA models for core proteome based tree within the MEGA software, version 6 [36]. In

addition a third phylogenetic tree among diverse *Rickettsia* species, inferred from sequence analysis of the 16S rRNA gene only was conducted in the same way as the first one. The position of stain AUS118^T was also established when phylogenetic analysis was inferred from the five concatenated multi-loci gene sequences comparisons (Fig. 2). A similar phylogenetic profile was obtained with the phylogenetic analysis from the concatenated core proteome sequence comparisons among the 28 *Rickettsia* species (Fig. 3). Based on these comparisons, strain AUS118^T was most closely related to the *R. japonica* group (including *R. japonica* and *R. heilongjiangensis)* (Fig. 2; Fig. 3; Fig. S1). Phylogenetic analyses on the basis of the 16S rRNA gene sequence only (Fig. S1) and of the concatenated MLST genes sequences (Fig. 2) revealed that *Rickettsia* spp. are associated with an extremely diverse host range including vertebrates, arthropods, leeches, insects (Fig. 2; Fig. S1). Furthermore, the *R. felis* group (*R. felis*, *R. akari*, *R. australis*, *R. hoogstraalii*, *R. asembonensis*) was placed between the typhus group and the ancestral group but not within the spotted fever group (Fig. S1).

Genomic DNA of *R. fournieri* sp. nov. strain AUS118^T was sequenced using a MiSeq sequencer with the mate pair strategy (Illumina Inc., San Diego, CA, USA). DNA was quantified by a Qubit assay with the high sensitivity kit (Life Technologies, Carlsbad, CA, USA) at 78 ng/µl and was barcoded in order to be mixed with 11 other projects with the Nextera Mate Pair sample prep kit (Illumina Inc., San Diego, CA, USA). For the mate pair library preparation, DNA was then diluted to obtain 1.5µg of genomic DNA as input. The tagmentation step fragmented the gDNA into a range from 1.5 kb up to 11kb with an optimal size at 5.63 kb inserts and tagged with a mate pair junction adapter. The fragmentation pattern was validated on an Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA 7500 labchip. The normalized libraries at 2nM were pooled for sequencing on the MiSeq. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated

cluster generation and sequencing run were performed in a single 39-hours run in a 2x251-bp. The 2,002,666 high-quality paired-end reads were trimmed and then assembled using the Spades assembler program [37].

Genome sequence comparison

The genome of strain AUS118^T (OFAL00000000) is 1,447,739 bp-long with a G+C content of 32.4 mol% organized into five scaffolds (composed to 6 contigs). The chromosome contains 1638 predicted protein-encoding genes and, like other *Rickettsia* species, 3 noncontiguous rRNAs (5S, 16S and 23S rRNA), 33 tRNAs, and 4 other RNAs) (Fig. 4). No plasmid was identified. Strain AUS118^T exhibited 46.03% genes associated to mobilome, and 28.57% duplicate genes. Furthermore, many genes (5.12%) associated to toxine / antitoxine modules were predicted.

When compared to the genomes of 27 valid *Rickettsia* species, strain AUS118^T had a smaller genome than *R. tamurae*, *R. amblyommatis*, *R. hoogstraalii*, *R. felis*, and *R. bellii*, but bigger than the other 22 species. The distribution of genes into COGs functional categories is presented in Fig. 4 and in Fig. S2. All compared genomes had similar COGs profiles, with the absence of genes encoding RNA processing and modification; Chromatin structure and dynamics; Extracellular structures; Nuclear structure and Cytoskeleton function.

In order to estimate the degrees of nucleotide sequence similarity at the genome level between *Rickettsia* sp. strain AUS118^T and other *Rickettsia* species, we first determined the Average Genomic Identity of orthologous gene Sequences (AGIOS) between pair genomes using the MAGI software [38]. Second, digital DNA–DNA hybridization (dDDH) relatedness values were predicted using the genome to genome distance calculator [39] via the GGDC 2.1 server (http://ggdc.dsmz.de/distcalc2.php). Finally, the average nucleotide identity by orthology analysis based on the overall similarity between pairs of genome sequences was

estimated using the OrthoANI algorithm version v0.91 [40]. Over all, among all compared genomes, AGIOS values ranged from 69.58 % between *R. bellii* and *R. felis* to 98.22 % between *R. sibirica* and *R. parkeri*. Strain AUS118^T shared a number of orthologous genes ranging from 747 (45.60 %) with *R. typhi* to 1062 (64.83 %) with *R. japonica* (979 (59.76 %) with *R. heilongjiangensis*), and exhibited AGIOS values ranging from 77.02 % with *R. felis* to 98.55 % with *R. heilongjiangensis* (98.27 % with *R. japonica*) (Table S1, available in the online Supplementary Material). In addition, dDDH values among *Rickettsia* species ranged from 23.2 % between *R. bellii* and *R. typhi* to 91.8 % between *R. sibirica* and *R. parkeri*. Strain AUS118^T exhibited dDDH values ranging from 26.0 % with *R. felis* to 90.2 % with *R. japonica* (89.4 % with *R. heilongjiangensis*) (Table S2). Moreover, OrthoANI values among species ranged from 79.74 % between *R. bellii* and *R. prowazekii* to 99.17 % between *R. sibirica* and *R. parkeri*. Strain AUS118^T exhibited OrthoANI values ranging from 81.37 % with *R. bellii* to 98.98 % with *R. japonica* (98.91 % with *R. heilongjiangensis*) (Table S3). On the basis of the results described above, we proposed that strain AUS118^T should be classified within a distinct spotted fever group species.

Description of *Rickettsia fournieri* sp. nov.

Rickettsia fournieri sp. nov. (four.ni.e'ri. N.L. masc. gen. n. *fournieri* of Fournier, named after the French clinical microbiologist Pierre-Edouard Fournier for his contribution to the taxono-genomic description of rickettsiae).

Obligately intracellular, Gram-negative, rod-shaped bacterium. Growth obtained in XTC-2 cells at 28° C in L-15 medium (Leibovitz medium) supplemented with 5% (w/v) foetal calf serum (FCS), 5% tryptose phosphate and 2 mmol/l L-glutamine and also in L929 cells at 32° C in minimal essential medium supplemented with 2% heat-inactivated fetal calf serum and 2mM L-glutamine. Detected by Gimenez staining and observed both in the cytoplasm and the

nucleus of XTC-2 cells. Bacterial cells measured a mean size of 1.5µm in length and 0.3µm in width by TEM. Strain AUS118^T is most closely related to the *R. japonica* group. G+C content is 32.4 mol%. No cytopathic effect was observed and pathogenicity of *R. fournieri* sp. nov. for vertebrate hosts is as yet unknown.

The type strain of *R. fournieri* sp. nov. is strain AUS118^T (=DSM 28985^T = CSUR R501^T). It was first, isolated in the entire body from an *Argas lagenoplastis* tick from Australia, in 2013 on XTC-2 cells at 28°C in L-15 medium (Leibovitz medium) supplemented with 5% (w/v) foetal calf serum (FCS), 5% tryptose phosphate and 2 mmol/l L-glutamine. Strain AUS118^T was deposited in the Deutsche Sammlung von Mikroorganismen un Zellkulturen (DSMZ) and the stands for Collection de Souches de l'Unité des Rickettsies (CSUR) under references DSM 28985^T and CSUR R501^T, respectively. The genome sequence of *R. fournieri* sp. nov. strain AUS118^T is deposited in EMBL-EBI under accession number OFAL00000000.

Funding information

This study was supported by the Fondation Méditerranée Infection and the French National Research Agency under the program "investissements d'avenir", reference ANR-10-IAHU-03.

Acknowledgements

We are also grateful to Sophie Edouard for PCR screening and Nathalie Duclos for her technical help with cell culture.

Conflicts of interest

The authors declare that they have no competing interest in relation to this research.

Reference

1. **Stothard DR, Clark JB, Fuerst PA**. Ancestral divergence of Rickettsia bellii from the spotted fever and typhus groups of Rickettsia and antiquity of the genus Rickettsia. *Int J Syst Evol Microbiol* 1994;44:798–804.

2. **Raoult D, Roux V**. Rickettsioses as paradigms of new or emerging infectious diseases. *Clin Microbiol Rev* 1997;10:694–719.

3. **Gillespie JJ, Beier MS, Rahman MS, Ammerman NC, Shallom JM, *et al.*** Plasmids and Rickettsial Evolution: Insight from Rickettsia felis. *PLoS ONE* 2007;2:e266.

4. **Merhej V, Raoult D**. Rickettsial evolution in the light of comparative genomics. *Biol Rev* 2011;86:379–405.

5. **Merhej V, Angelakis E, Socolovschi C, Raoult D**. Genotyping, evolution and epidemiological findings of Rickettsia species. *Infect Genet Evol* 2014;25:122–137.

6. **Parola P, Paddock CD, Socolovschi C, Labruna MB, Mediannikov O, *et al.*** Update on Tick-Borne Rickettsioses around the World: a Geographic Approach. *Clin Microbiol Rev* 2013;26:657–702.

7. **Sahni SK, Narra HP, Sahni A, Walker DH**. Recent molecular insights into rickettsial pathogenesis and immunity. *Future Microbiol* 2013;8:1265–1288.

8. **El Karkouri K, Kowalczewska M, Armstrong N, Azza S, Fournier P-E, *et al.*** Multi-omics Analysis Sheds Light on the Evolution and the Intracellular Lifestyle Strategies of Spotted Fever Group Rickettsia spp. *Front Microbiol*;8. Epub ahead of print 20 July 2017. DOI: 10.3389/fmicb.2017.01363.

9. **Fournier P-E, Raoult D**. Current Knowledge on Phylogeny and Taxonomy of Rickettsia spp. *Ann N Y Acad Sci* 2009;1166:1–11.

10. **Merhej V, Raoult D**. Rickettsial evolution in the light of comparative genomics. *Biol Rev* 2011;86:379–405.

11. **Stewart A, Armstrong M, Graves S, Hajkowicz K**. Clinical Manifestations and Outcomes of Rickettsia australis Infection: A 15-Year Retrospective Study of Hospitalized Patients. *Trop Med Infect Dis* 2017;2:19.

12. **McBride WJ, Hanson JP, Miller R, Wenck D**. Severe spotted fever group rickettsiosis, Australia. *Emerg Infect Dis* 2007;13:1742.

13. **Graves SR, Stewart L, Stenos J, Stewart RS, Schmidt E, *et al.*** Spotted fever group rickettsial infection in south-eastern Australia: isolation of rickettsiae. *Comp Immunol Microbiol Infect Dis* 1993;16:223–233.

14. **Sexton DJ, Dwyer B, Kemp R, Graves S**. Spotted fever group rickettsial infections in Australia. *Rev Infect Dis* 1991;13:876–886.

15. **Barker SC, Walker AR**. Ticks of Australia. The species that infest domestic animals and humans. *Zootaxa* 2014;1–144.

16. **Unsworth NB, Stenos J, McGregor AR, Dyer JR, Graves SR**. Not only 'Flinders Island' spotted fever. *Pathology (Phila)* 2005;37:242–245.

17. **Graham RMA, Donohue S, McMahon J, Jennison AV**. Detection of Spotted Fever Group Rickettsia DNA by Deep Sequencing. *Emerg Infect Dis* 2017;23:1911–1913.

18. **Williams M, Izzard L, R Graves S, Stenos J, J Kelly J**. *First probable Australian cases of human infection with*. 2011.

19. **Teoh YT, Hii SF, Graves S, Rees R, Stenos J, *et al.*** Evidence of exposure to Rickettsia felis in Australian patients. *One Health* 2016;2:95–98.

20. **Abdad MY, Abdallah RA, Karkouri KE, Beye M, Stenos J, *et al.*** Rickettsia gravesii sp. nov.: a novel spotted fever group rickettsia in Western Australian Amblyomma triguttatum triguttatum ticks. *Int J Syst Evol Microbiol* 2017;67:3156–3161.

21. **Barker SC, Walker AR, Campelo D**. A list of the 70 species of Australian ticks; diagnostic guides to and species accounts of Ixodes holocyclus (paralysis tick), Ixodes cornuatus (southern paralysis tick) and Rhipicephalus australis (Australian cattle tick); and consideration of the place of Australia in the evolution of ticks with comments on four controversial ideas. *Int J Parasitol* 2014;44:941–953.

22. **Hoogstraal H, Kohls GM**. Observation on the subgenus Argas (Ixodoidea: Argasidae: Argas). 6. Redescription and biological notes on A. lagenoplastis Froggat, 1906 of Austraulian fairy martins, Hylochelidon ariel (Gould). *Ann Entomol Soc Am* 1963;56:577–582.

23. **Roberts FHS**. *Australian Ticks*. In: Melbourne, Vic: CSIRO. 1970; 267.

24. **Sokhna C, Mediannikov O, Fenollar F, Bassene H, Diatta G, *et al.*** Point-of-Care Laboratory of Pathogen Diagnosis in Rural Senegal. *PLoS Negl Trop Dis* 2013;7:e1999.

25. **Regnery RL, Spruill CL, Plikaytis BD**. Genotypic identification of rickettsiae and estimation of intraspecies sequence divergence for portions of two rickettsial genes. *J Bacteriol* 1991;173:1576–1589.

26. **Sekeyová Z, Mediannikov O, Subramanian G, Kowalczewska M, Quevedo-Diaz M, *et al.*** Isolation of Rickettsia helvetica from ticks in Slovakia. *Acta Virol* 2012;56:247–252.

27. **Gimenez DF**. Staining rickettsiae in yolk-sac cultures. *Stain Technol* 1964;39:135–140.

28. **Kim M, Oh H-S, Park S-C, Chun J**. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.

29. **Fournier P-E, Dumler JS, Greub G, Zhang J, Wu Y, *et al.*** Gene Sequence-Based Criteria for Identification of New Rickettsia Isolates and Description of Rickettsia heilongjiangensis sp. nov. *J Clin Microbiol* 2003;41:5456–5465.

30. **Roux V, Raoult D**. Phylogenetic analysis of members of the genus Rickettsia using the gene encoding the outer-membrane protein rOmpB (ompB). *Int J Syst Evol Microbiol* 2000;50:1449–1455.

31. **Sekeyova Z, Roux V, Raoult D**. Phylogeny of Rickettsia spp. inferred by comparing sequences of'gene D', which encodes an intracytoplasmic protein. *Int J Syst Evol Microbiol* 2001;51:1353–1360.

32. **Meier-Kolthoff JP, G?ker M, Spr?er C, Klenk H-P**. When should a DDH experiment be mandatory in microbial taxonomy? *Arch Microbiol* 2013;195:413–418.

33. **Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M**. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:60.

34. **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, *et al.*** Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23:2947–2948.

35. **Katoh K, Standley DM**. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013;30:772–780.

36. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S**. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* 2013;30:2725–2729.

37. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.*** SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 2012;19:455–477.

38. **Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, *et al.*** A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* 2014;64:384–391.

39. **Klenk H-P, Meier-Kolthoff JP, G?ker M**. Taxonomic use of DNA G+C content and DNA?DNA hybridization in the genomic age. *Int J Syst Evol Microbiol* 2014;64:352–356.

40. **Ouk Kim Y, Chun J, Lee I, Park S-C**. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol* 2016;66:1100–1103.

Table 1: Gene sequence similarity (%) of five genes of *R. fournieri* sp. nov. strain AUS118^T, compared with sequences of 27 validated species of the genus *Rickettsia*.

The Genbank accession numbers indicated corresponding of those of the genome from which the gene sequences were extracted. Pairwise sequence similarities were calculated using the method recommended by Meier-Kolthoff et al. [32] via the GGDC web server (http://ggdc.dsmz.de/). NA, Nucleotide sequences were either not applicable in the analysis; NE, do not exist in TG rickettsiae.

Strain species	*Rickettsia fournieri* sp.nov. strain AUS118					Genome accession number
	rrs (KF666475)	gltA (KF666471)	Sca4 (KF666473)	ompA (KF666477)	ompB (KF666469)	
R. aeschlimannii MC16^T	99.23	98.88	97.88	95.21	96.47	CCER01000000
R. africae ESF-5^T	99.44	99.12	98.38	96.77	96.98	CP001612
R. akari Hartford	98.10	94.08	87.47	84.11	88.89	CP000847
R. amblyommatis Ac/Pa	99.23	98.56	97.97	95.57	96.57	LANR01000001
R. asembonensis NMRCii^T	99.09	94.48	91.45	84.48	92.26	JWSW01000001
R. australis Cutlack	98.94	95.04	88.22	86.28	90.83	NC_017058
R. bellii RML369-C^T	99.09	87.17	NA	NA	NA	NC_007940
R. canadensis Mckiel^T	98.45	92.31	84.53	82.40	85.50	NC_009879
R. conorii Malish 7^T	99.51	99.12	98.38	95.25	97.35	NC_003103
R. heilongjiangensis O54^T	99.72	99.60	98.99	97.80	98.60	CP002912
R. felis URRWXCal2	99.30	94.56	89.81	NA	92.05	NC_007109
R. helvetica C9P9	99.09	96.80	92.37	NA	90.57	CM001467
R. honei RB^T	99.44	99.04	98.51	96.26	97.02	AJTT01000001
R. hoogstraalii Croatica^T	99.09	94.32	87.89	86.38	88.49	CCXM01000001
R. japonica YH^T	99.79	99.52	98.94	97.12	98.71	NC_016050
R. massiliae MTU5	99.51	98.80	98.25	95.39	96.46	NC_009900
R. montanensis OSU 85-930^T	99.16	98.96	98.01	94.55	95.81	CP003340
R. parkeri Portsmouth	99.44	99.20	98.25	94.92	97.05	NC_017044
R. peacockii Rustic	99.51	99.20	98.64	93.81	97.24	CP001227
R. prowazekii Breinl^T	98.17	92.71	82.22	NE	83.52	NC_020993
R. raoultii Khabarovsk^T	99.58	99.04	98.24	95.91	96.69	CP010969
R. rhipicephali 3-7-female6-CWPP^T	99.44	98.72	98.12	95.39	96.72	NC_017042
R. rickettsii Sheila Smith^T	99.51	99.12	98.29	95.58	96.98	NC_009882
R. sibirica 246^T	99.51	99.28	98.24	96.26	97.05	AABW01000001
R. slovaca 13-B	99.58	99.36	99.00	97.11	97.16	NC_016639
R. tamurae AT-1^T	99.09	96.72	95.50	89.10	93.02	CCMG01000008
R. typhi Wilmington^T	98.31	92.71	82.24	NE	83.70	NC_006142

Figure 1 A: Gimenez staining of XTC-2 cells infected with *Rickettsia fournieri* sp. nov. strain AUS118^T, seventh day post-inoculation. B: Transmission electron microscopy of *Rickettsia fournieri* sp. nov. strain AUS118^T using a Tecnai G20, operating at 200 keV.

Figure 2: Phylogenetic tree highlighting the position of *Rickettsia fournieri* strain AUS118^T relative to other closely related rickettsia type strains. The sequences of the 16S rRNA (1421 bp), *gltA* (1250 bp), *sca4* (2289 bp), *ompB* (2716 bp) and *ompA* (590 bp) genes were concatenated, and then aligned using CLUSTALW, with default parameters. Phylogenetic inference was obtained by the Maximum Likehood method with the Kimura 2-parameter model within the MEGA6 software. The Genbank accession numbers of the genome from which the gene sequences were extracted are in Table 1. Numbers at the nodes represent the percentages of bootstrap values obtained by repeating analysis 500 times to generate a majority consensus tree. Only values higher than 95 % are shown. The scale bar represents a 5 % nucleotide sequence divergence.

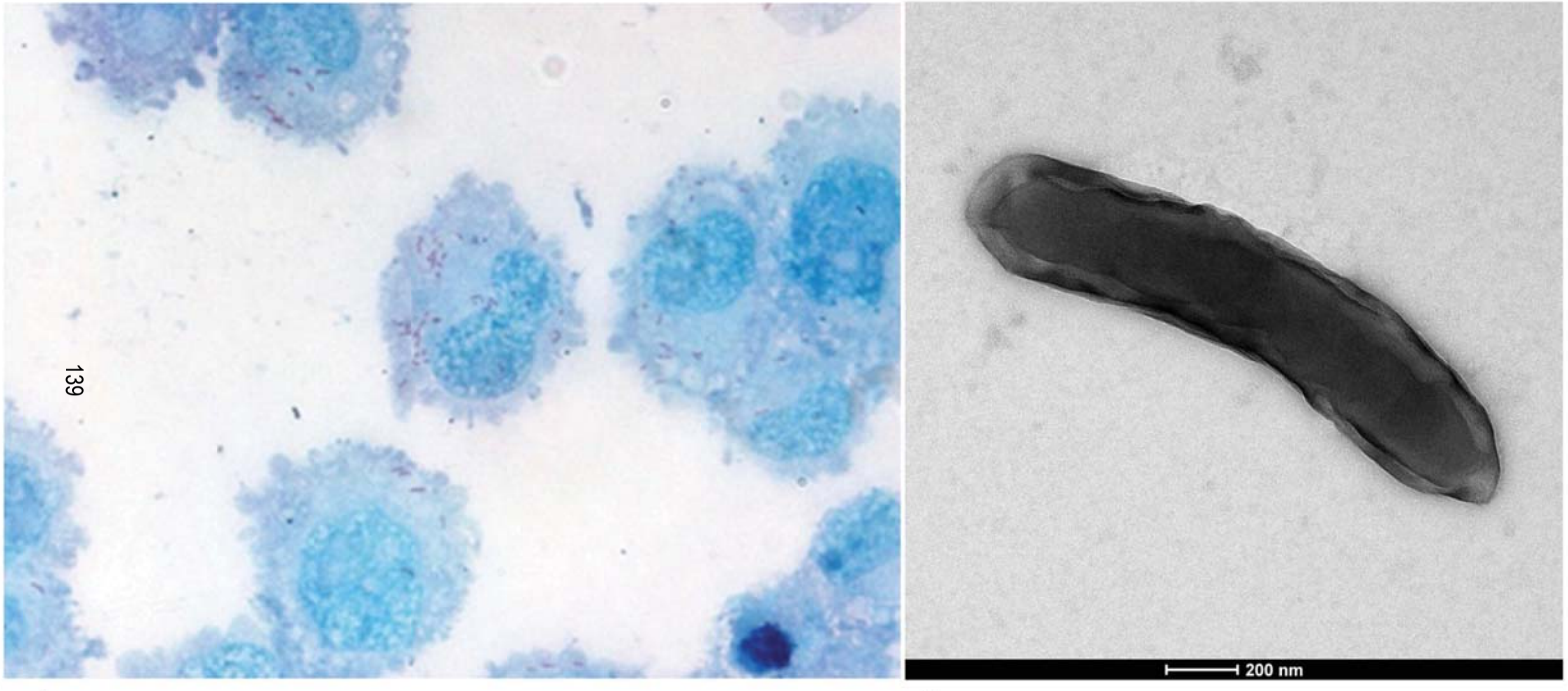
Figure 3: Phylogenetic tree of 28 valid *Rickettsia* species based on 633 concatenated core proteins. Sequences were aligned using mafft alignment algorithm. Phylogenetic inference was obtained by Maximum Likehood method with JTT and GAMMA models within the MEGA software and display only topology. Numbers at the nodes represent the percentages of bootstrap values obtained by repeating analysis 500 times to generate a majority consensus tree. The scale bar represents a 2 % nucleotide sequence divergence.

Figure 4: Graphical circular map of the chromosome of *Rickettsia fournieri* sp. nov. strain AUS118^{T}.

From the outside to the center: Genes on the forward strand colored by Clusters of Orthologous Groups of proteins (COG) categories (only genes assigned to COG), genes on the reverse strand colored by COG categories (only gene assigned to COG), RNA genes (tRNAs green, rRNAs red), GC content and GC skew.

Figure

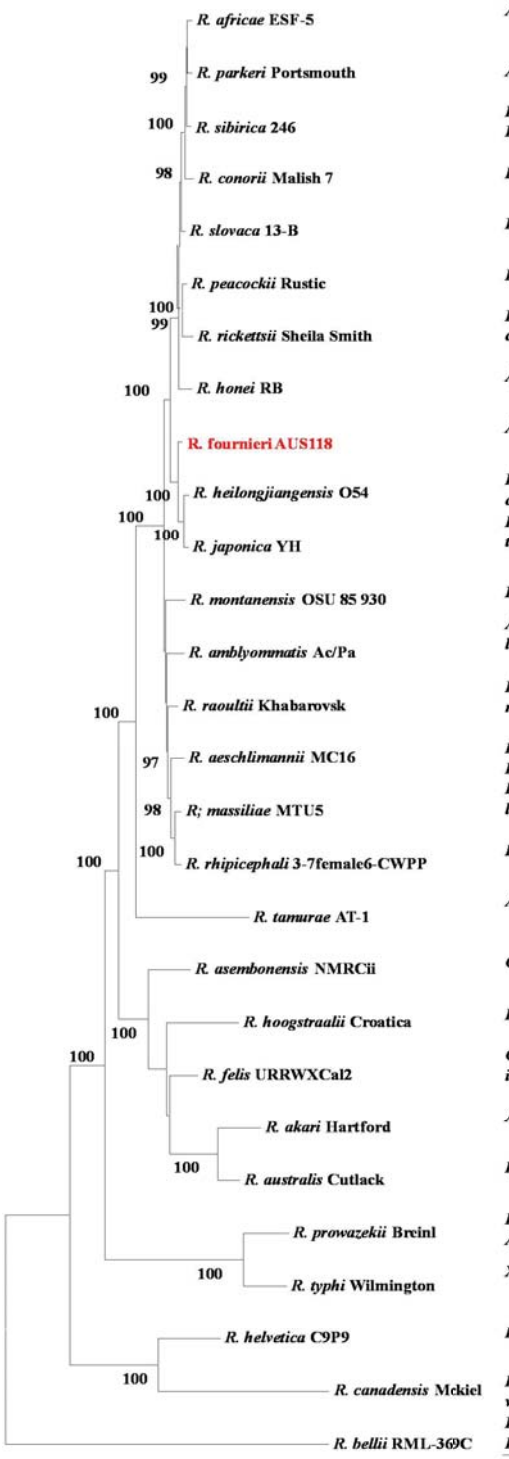
Click here to download Figure Figure 1.pdf



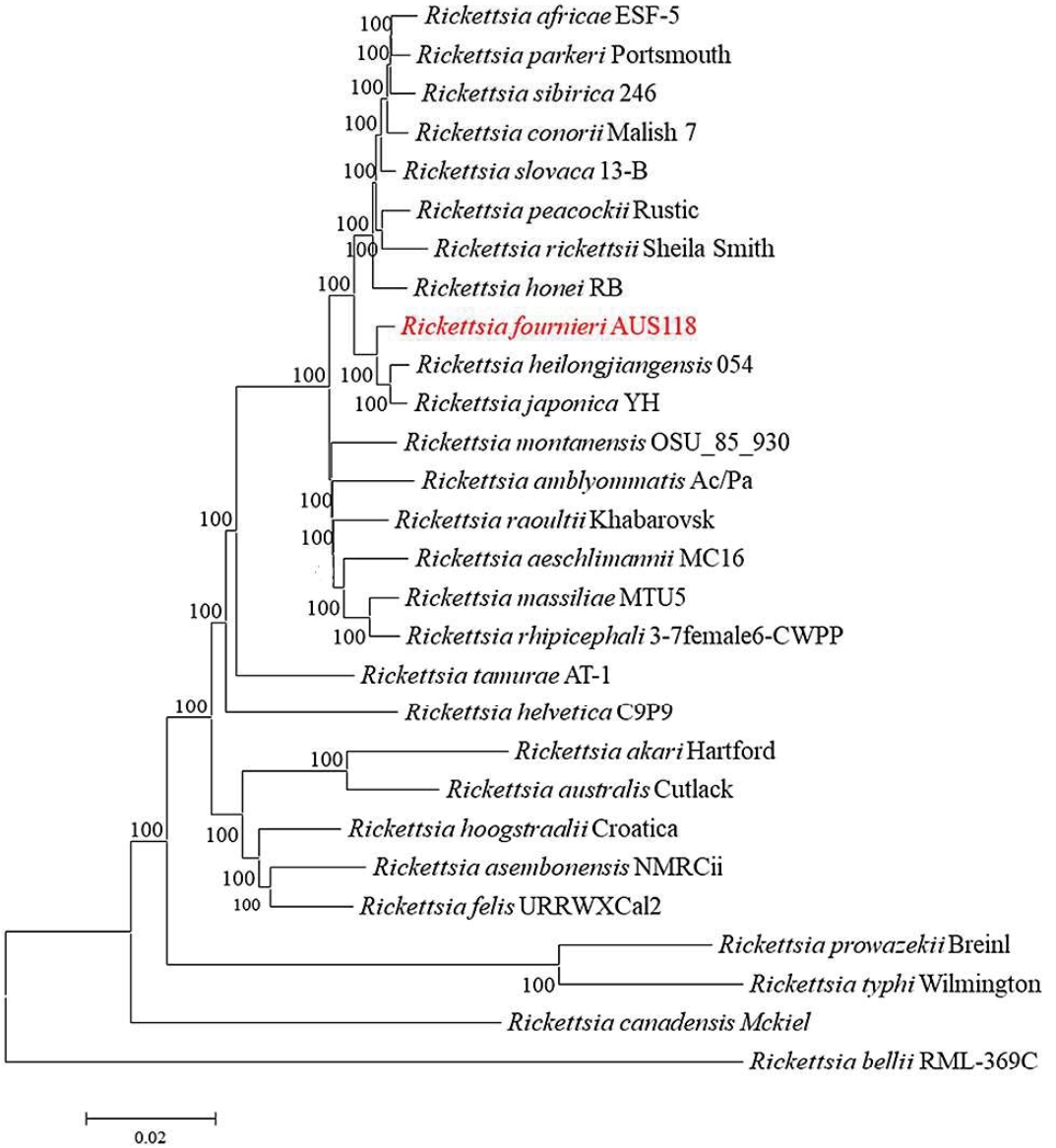
A

B

Figure

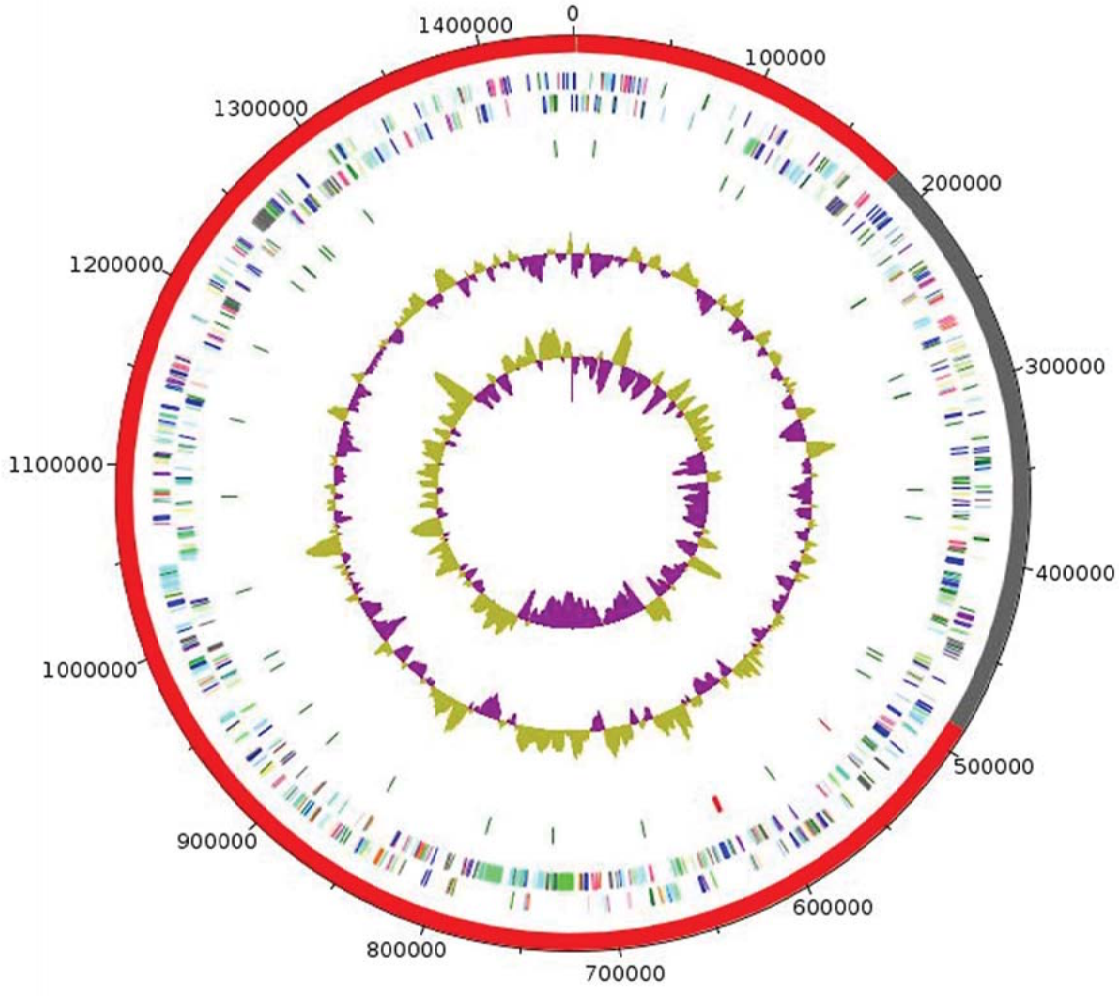


Vectors	Chromosome size (Mb)	Presence of plasmids
Amblyomma variegatum; A. hebraum	1.28	pRaf
Amblyomma maculatum; A. americanum	1.30	-
Dermacentor nuttallii; D. sinicus; D. marginatus;D. silvatum; D. pictus; D. auratus	1.25	-
Rhipicephalus sanguineus; R. pumilio	1.27	-
Dermacentor marginatus; D. reticulates	1.27	-
Dermacentor andersoni	1.29	pRpe
Dermacentor andersoni; D. variabilis; Amblyomma cajennense; Rhipicephalus sanguineus	1.26	-
Aponomma hydrosauri; Ixodes granulatus	1.27	-
Argas lagenoplastis	1.45	-
Dermacentor silvarum; Haemophysalis japonica douglasi; H. concinna	1.28	-
Haemaphysalis fava; H. longicornis; Dermacentor taiwanensis; Ixodes ovatus	1.28	-
Dermacentor variabilis; D. andersoni	1.28	-
Amblyomma americanum; A. cajennense;A. coelebs; A. longirostre	1.44	-
Dermacentor silvarum; D. nutallii;D. reticulatus; D. marginatus; D. niveus	1.34	pRa1; pRa2;
Hyalomma m. marginatum; H. m.rufipes; R. appendiculatus; H. punctata	1.31	Plasmid 1; Plasmid 2
Rhipicephalus turanicus; R. sanguineus; R. mushamae; R. lunulatus; R. sulcatus	1.37	pRma
Rhipicephalus sanguineus	1.31	pRrh
Amblyomma testudinarium	1.44	Plasmid 1; Plasmid 2
Ctenocephalides felis	1.36	Pras01
Haemaphysalis sulcata	1.48	-
Ctenocephalides felis; C. canis; Archeopsylla erinacei;Pulex irritans; Xenopsylla cheopis; Anomiopsyllus nudat	1.49	pRF; pRFδ
Allodermanyssus sanguineus; Liponyssoides sanguineus	1.23	-
Ixodes holocyclus; I. tasmani; I. cornuatus	1.33	pMC5_1
Pediculus humanus corporis; Orchopeas howardii; Amblyomma cajennense	1.11	-
Xenopsylla cheopis; Ctenocephalides felis; Leptosylla segnis	1.11	-
Ixodes ricinus; I. ovatus; I. persulcatus; I. monospinosus	1.37	pRhe
Haemaphysalis leporispalustris; Dermacentor andersoni; D. variabilis; Amblyomma americanum	1.16	-
Dermacentor sp.; Amblyomma sp.; Ixodes loricatus; Haemaphysalis sp.; Ornithodoros concanensis; Argas cooleyi	1.52	-



Figure

Click here to download Figure Figure 4.pdf



Rickettsia fournieri sp. nov. strain AUS118^{T}, a novel spotted fever group rickettsia from *Argas lagenoplastis* ticks in Australia

SUPPLEMENTARY DATA

Awa Diop[1], Stephen C. Barker[2], Eberhard Mey [2], Dayana Campelo[2], Thi Tien Nguyen[1], Fabrizio di Pinto[3], Didier Raoult[3], Oleg Mediannikov[3,*]

[1]UMR VITROME, Aix-Marseille University, IRD, Service de Santé des Armées, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France Tel: +33 413 732 401, Fax: +33 413 732 402

[2]Department of Parasitology, School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane QLD 4072, Queensland, Australia

[3]UMR MEPHI, Aix-Marseille University, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France

Email: olegusss1@gmail.com

SUPPLEMENTARY TABLES

Table S1: Numbers of orthologous proteins shared between genomes (upper right), and AGIOS values (%) (lower left). Bold numbers indicate the numbers of protein per genome

	R. aeschlimannii	*R. africae*	*R. akari*	*R. amblyommatis*	*R. asembonensis*	*Rickettsa australis*	*R. bellii*	*R. canadensis*	*R. conorii*	*R. heilongjiangensis*	*R. felis*	*R. fournieri*	*R. helvetica*	*R. honei*	*R. hoogstraalii*	*R. japonica*	*R. massiliae*	*R. montanensis*	*R. parkeri*	*R. peacockii*	*R. prowazekii*	*R. raoultii*	*R. rhipicephali*	*R. rickettsii*	*R. sibirica*	*R. slovaca*	*R. tamurae*	*R. typhi*
R. aeschlimannii	**1,536**	80.4	77.2	92.3	79.0	81.7	78.1	76.3	89.7	85.8	81.7	89.5	77.9	85.5	85.6	90.1	91.9	88.6	92.1	90.6	75.4	83.7	95.1	95.4	90.8	92.2	81.5	74.5
R. africae	96.28	**1,444**	80.1	94.4	81.0	84.8	80.7	76.7	1,196	90.6	84.4	98.4	80.6	97.6	87.2	94.7	92.9	92.7	1,27	1,054	76.4	85.7	95.4	116.0	123.0	119.4	85.8	75.8
R. akari	90.96	90.9	**1,314**	79.9	77.9	86.9	78.0	75.0	81.6	79.2	79.1	78.7	73.7	78.6	81.8	80.5	80.6	80.7	81.0	79.4	76.5	77.8	81.5	82.5	80.6	81.6	78.2	75.6
R. amblyommatis	96.19	95.63	90.62	**1,596**	82.6	83.1	80.6	77.4	93.5	92.4	86.3	94.7	83.5	90.2	87.3	93.4	96.2	96.2	96.2	94.8	75.5	97.0	95.4	93.8	93.6	96.1	87.4	75.1
R. asembonensis	92.3	91.91	92.35	91.85	**1,589**	82.6	81.4	73.8	81.5	81.3	92.1	81.4	78.9	79.0	87.1	81.1	80.9	82.4	80.7	78.6	74.5	80.1	81.4	80.3	94.7	99.2	79.6	71.4
Rickettsa australis	91.39	91.34	95.37	91.09	92.59	**1,456**	79.0	74.8	83.7	83.8	84.1	83.1	78.5	82.4	84.5	83.1	81.2	82.3	84.2	82.5	74.3	82.3	81.5	83.7	83.5	84.0	82.9	73.9
R. bellii	80.84	80.7	80.35	80.78	80.6	80.5	**1,516**	73.7	80.4	80.5	90.2	81.4	79.5	78.9	85.7	80.5	82.4	82.5	79.9	79.4	75.2	82.2	81.0	80.6	79.5	81.3	84.3	74.6
R. canadensis	78.62	78.94	78.25	78.28	78.35	77.95	72.75	**1,066**	75.8	76.1	76.1	75.8	73.5	75.2	75.9	76.5	77.4	76.6	76.7	75.2	73.3	76.2	76.6	77.0	76.2	76.6	75.8	73.6
R. conorii	96.04	97.65	90.99	95.55	91.76	91.12	80.49	78.63	**1,435**	90.2	84.2	97.1	82.4	1.040	84.8	98.2	92.3	94.1	1,173	1,068	75.9	94.3	96.0	109.0	116.6	122.4	85.3	75.4
R. heilongjiangensis	96.99	97.26	91.31	96.44	92.33	91.21	80.82	78.67	96.5	**1,449**	87.2	97.9	83.8	96.2	87.0	1.057	91.5	95.2	1,023	975	75.1	92.1	95.1	102.2	100.1	103.0	80.3	74.9
R. felis	76.9	76.66	77.36	76.49	77.37	76.68	69.58	72.00	76.07	76.34	**1,517**	85.7	85.2	83.6	97.4	86.4	84.8	87.0	85.4	86.3	77.6	89.8	87.1	85.6	84.8	86.3	89.3	77.6
R. fournieri	96.71	96.86	90.92	96.14	92.38	91.22	81.17	79.03	96.72	98.55	77.02	**1,638**	81.8	94.3	86.2	1.062	91.6	95.3	1034	96.4	75.2	95.4	95.0	103.1	99.4	103.1	86.1	74.7
R. helvetica	92.75	92.27	90.98	92.02	91.94	91.4	80.57	79.63	92.2	92.44	76.28	92.59	**1,537**	79.1	83.6	82.5	82.0	82.6	83.0	80.8	73.7	81.4	83.8	84.1	83.6	84.5	80.1	72.9
R. honei	96.83	97.71	90.78	96.28	92.41	91.33	80.89	78.78	97.02	97.13	76.95	97.62	92.45	**1,472**	83.0	96.2	91.2	92.6	1,095	1,045	74.7	86.0	90.7	108.6	106.2	109.2	83.0	74.1
R. hoogstraalii	92.53	92.29	92.88	91.98	94.9	93.34	80.69	79.01	92.44	92.32	77.16	92.57	92.1	92.59	**1,537**	85.7	85.3	87.1	85.9	84.3	76.1	85.7	85.0	85.9	85.7	86.9	87.7	75.8
R. japonica	96.32	97.32	91.12	96.14	92.03	91.42	80.6	78.89	96.6	98.49	76.69	98.27	92.51	97.00	92.54	**1,453**	91.5	93.7	1,025	988	75.2	89.2	93.1	102.1	100.6	102.6	85.7	74.9
R. massiliae	96.82	96.61	91.07	96.16	92.12	91.71	80.79	78.72	96.26	96.46	76.89	96.77	92.57	96.68	92.46	96.73	**1,558**	92.8	93.1	90.7	75.2	95.2	110.4	92.5	91.8	93.0	83.5	74.7
R. montanensis	96.67	96.78	91.28	96.18	91.97	91.7	81.14	79.04	96.26	96.45	76.42	96.61	92.62	96.46	92.26	96.8	97.05	**1,416**	97.8	92.3	75.0	91.9	96.3	94.2	92.6	94.5	85.2	74.8
R. parkeri	96.29	98.62	91.1	95.99	92.1	91.29	80.59	78.83	97.75	96.55	76.56	96.82	92.35	97.33	92.45	96.86	96.76	96.46	**1,462**	112.8	74.9	95.0	97.0	113.3	122.6	119.3	86.5	74.8
R. peacockii	96.34	98.16	91.26	95.88	92.18	91.19	80.8	78.94	97.28	96.76	76.52	97.34	92.86	97.56	92.13	97.00	96.67	96.59	97.86	**1,401**	73.8	92.0	92.1	119.5	110.6	114.2	84.6	73.7
R. prowazekii	87.77	87.61	86.99	87.69	87.55	87.57	79.71	76.3	87.52	87.8	74.8	87.92	87.41	87.87	87.93	87.68	87.69	87.74	87.65	87.63	**980**	75.6	75.5	76.0	75.7	76.0	75.2	80.2
R. raoultii	97.51	97.6	91.91	96.38	92.53	91.81	80.5	78.96	96.56	96.71	76.52	96.86	93.11	97.21	92.41	97.38	97.15	97.44	96.89	97.12	87.66	**1,527**	96.8	94.6	94.2	95.0	83.1	75.1
R. rhipicephali	96.57	96.45	91.21	96.00	92.21	92.00	80.46	78.99	95.75	96.17	76.47	96.6	92.53	96.55	92.49	96.75	97.64	96.84	96.42	96.63	87.71	97.14	**1,493**	95.0	94.4	94.7	86.6	74.9
R. rickettsii	95.71	97.19	90.84	95.8	92.07	91.33	80.52	78.88	96.79	96.12	76.64	96.53	92.13	96.87	92.19	96.78	96.61	96.48	97.41	97.85	87.55	96.78	96.31	**1,414**	110.1	113.4	86.5	74.9
R. sibirica	96.13	98.34	91.06	96.09	96.59	91.25	80.71	78.8	97.42	96.85	76.45	97.31	92.16	97.7	92.51	96.53	96.45	96.44	98.58	97.61	87.76	96.75	96.12	97.06	**1,431**	117.0	85.1	74.6
R. slovaca	95.94	97.7	90.73	95.32	95.87	91.11	80.31	78.56	97.81	96.15	76.09	96.46	91.87	96.87	92.02	96.52	96.31	96.27	97.61	97.34	87.41	96.58	96.2	96.83	97.2	**1,467**	86.7	75.0
R. tamurae	94.06	93.34	91.67	92.79	93.32	91.72	80.56	78.82	93.32	94.19	76.6	93.68	93.51	93.76	92.82	93.15	93.46	93.56	93.08	93.27	87.82	93.78	93.14	92.96	93.3	92.8	**1,654**	74.0
R. typhi	86.09	86.19	85.55	85.86	85.71	86.01	78.49	75.1	86.03	85.99	72.78	86.11	86.05	86.04	86.04	86.17	86.16	86.27	86.15	86.11	92.25	86.22	86.27	86.09	86.00	85.9	86.1	**868**

Table S2: dDDH values (%) obtained by pairwise comparisons of studied genomes (upper right)

	R. aeschlimannii	*R. africae*	*R. akari*	*R. amblyommatis*	*R. asembonensis*	*Rickettsia australis*	*R. bellii*	*R. canadensis*	*R. conorii*	*R. heilongjiangensis*	*R. felis*	*R. fournieri*	*R. helvetica*	*R. honei*	*R. hoogstraalii*	*R. japonica*	*R. massiliae*	*R. montanensis*	*R. parkeri*	*R. peacockii*	*R. prowazekii*	*R. raoultii*	*R. rhipicephali*	*R. rickettsii*	*R. sibirica*	*R. slovaca*	*R. tamurae*	*R. typhi*
R. aeschlimannii	100.00	69.2±2.9	42.6±2.6	70.2±2.9	49.2±2.6	45.2±2.6	24.9±2.4	36.8±2.5	70±2.9	70.2±2.9	47.2±2.6	72.2±2.9	50.2±2.7	70.2±2.9	48±2.6	70.2±2.9	74.5±2.9	72.2±2.9	70.2±2.9	72±2.9	33.2±2.5	74.2±2.9	74.2±2.9	67.2±2.9	69.2±2.9	72.2±2.9	54.2±2.7	32.2±2.5
R. africae		100.00	42.6±2.5	68.8±2.9	48.6±2.7	44.5±2.6	24.9±2.4	36.4±2.5	89.8±2.2	76±2.8	47±2.6	78.1±2.8	48.6±2.6	84.1±2.6	47.7±2.6	76.6±2.8	70±2.9	71.1±2.9	91.2±2	86.6±2.5	32.9±2.5	73.7±2.9	70.6±2.9	82.1±2.7	90.7±2.1	90.1±2.1	53.2±2.7	32.3±2.5
R. akari			100.00	42.9±2.6	47.8±2.6	61±2.8	24.2±2.4	34.2±2.5	42.8±2.6	42.7±2.6	47.3±2.6	43.5±2.6	42.8±2.6	42.9±2.6	49.4±2.6	42.6±2.5	43.1±2.6	43.4±2.6	42.8±2.6	43±2.6	32.1±2.5	43.5±2.6	43.2±2.6	42.4±2.5	42.6±2.6	43.2±2.6	44±2.6	31.7±2.5
R. amblyommatis				100.00	48.6±2.6	44.6±2.6	25.5±2.4	36.2±2.5	70±2.9	70.4±2.9	47.1±2.6	70.1±2.9	49.1±2.6	70.5±2.9	48.1±2.6	70.8±2.9	70.1±2.9	72.3±2.9	69.3±2.9	71±2.9	33.2±2.5	74.8±2.9	71.7±2.9	68.1±2.9	69.8±2.9	72.2±2.9	52.7±2.7	32.5±2.5
R. asembonensis					100.00	51.1±2.7	25.5±2.4	36.8±2.5	48.9±2.6	49±2.6	63.1±2.8	49.5±2.6	48.8±2.6	48.9±2.7	63.4±2.8	48.7±2.6	49.1±2.6	49.3±2.6	48.7±2.6	49.1±2.7	33.7±2.5	49.8±2.6	49.2±2.6	48.1±2.6	48.7±2.7	49.4±2.6	50.9±2.7	33±2.5
Rickettsia australis						100.00	24.4±2.4	35±2.5	44.7±2.6	45±2.6	50.5±2.7	46.1±2.6	44.8±2.6	44.7±2.6	53.1±2.7	44.9±2.6	45.5±2.6	45.5±2.6	44.5±2.6	45.2±2.6	32.8±2.5	45.8±2.6	45.7±2.6	44.1±2.6	44.5±2.6	45.1±2.6	46.4±2.6	32.4±2.5
R. bellii							100.00	24.7±2.4	25±2.4	24.7±2.4	25.1±2.4	26±2.4	25±2.4	25±2.4	25.5±2.4	24.7±2.4	25.3±2.4	26.9±2.4	25±2.4	25.3±2.4	24.1±2.4	25±2.4	25±2.4	24.8±2.4	24.9±2.4	25±2.4	25.6±2.4	23.4±2.4
R. canadensis								100.00	36.4±2.5	36.6±2.5	36.5±2.5	36.9±2.5	37.9±2.5	36.7±2.5	37.1±2.5	36.6±2.5	36.6±2.5	36.6±2.5	36.3±2.5	36.6±2.5	29.9±2.3	37±2.5	36.9±2.5	36.1±2.5	36.3±2.5	36.6±2.5	37.5±2.5	29.7±2.5
R. conorii									100.00	76.7±2.9	47.1±2.6	79.8±2.8	49.4±2.6	85±2.6	48.2±2.6	77.1±2.8	71±2.9	72.1±2.9	90.9±2	87.6±2.4	33.1±2.5	74.7±2.9	71.6±2.9	82.9±2.7	90±2.2	90.7±2.1	54±2.7	32.4±2.5
R. heilongjiangensis										100.00	47.1±2.6	89.4±2.2	49.7±2.6	76.9±2.8	48.7±2.7	92.4±1.9	71.1±2.9	72.6±2.9	76.9±2.9	78.1±2.8	33±2.5	75±2.9	72±2.9	74.1±2.9	76.1±2.9	78.7±2.8	53.2±2.7	32.4±2.5
R. felis											100.00	48.6±2.6	47.2±2.6	47.2±2.6	58.7±2.7	47.2±2.6	47.6±2.6	48.4±2.6	47±2.6	47.7±2.6	33.5±2.5	48.5±2.6	47.5±2.6	46.1±2.6	47.1±2.6	47.5±2.6	49.6±2.6	32.6±2.5
R. fournieri												100.00	49.9±2.6	80.3±2.8	49.6±2.6	90.2±2.1	71.5±2.9	74±2.9	77.4±2.8	80.2±2.7	33±2.5	77.2±2.8	74.3±2.9	77.3±2.8	79.2±2.8	82.3±2.7	54±2.7	32.5±2.5
R. helvetica													100.00	49.6±2.6	48.4±2.6	49.7±2.6	49.5±2.6	50.8±2.7	48.7±2.7	50±2.7	33.4±2.5	51±2.7	50.1±2.7	48.5±2.6	49.2±2.6	56.2±2.7	53±2.7	32.7±2.5
R. honei														100.00	48.6±2.7	77.1±2.9	71.3±2.9	72.6±2.9	85.4±2.5	86.3±2.5	33.2±2.5	74.9±2.9	71.7±2.9	82±2.7	84.7±2.6	87.7±2.3	53.5±2.7	32.3±2.5
R. hoogstraalii															100.00	48.8±2.7	49.1±2.7	49.8±2.7	47.8±2.6	48.8±2.6	33.5±2.5	49.9±2.7	49.3±2.7	47.4±2.6	48±2.6	48.8±2.6	50.2±2.6	33±2.5
R. japonica																100.00	71.5±2.9	73.1±2.9	77.4±2.8	78.6±2.8	33±2.5	75.3±2.9	72.5±2.9	74.4±2.9	76.8±2.8	79.6±2.8	53.4±2.7	32.4±2.5
R. massiliae																	100.00	73.5±2.9	70.4±2.9	71.6±2.9	33.3±2.5	76.4±2.8	86.1±2.5	69±2.9	70.8±2.9	73.4±2.9	53.9±2.7	32.7±2.5
R. montanensis																		100.00	71.5±2.9	73.4±2.9	33.2±2.5	77.2±2.8	73.9±2.9	70.3±2.9	72±2.9	74.2±2.9	55.1±2.7	32.4±2.5
R. parkeri																			100.00	87.1±2.4	33.1±2.5	74.3±2.9	71.8±2.9	83.2±2.7	91.8±1.9	91±2	53.1±2.7	32.4±2.5
R. peacockii																				100.00	33.3±2.5	75.7±2.9	73±2.9	87.5±2.4	87.2±2.4	90.6±2.1	54.3±2.7	32.4±2.5
R. prowazekii																					100.00	33.6±2.5	33.4±2.5	33.2±2.5	33.1±2.5	33.3±2.5	33.6±2.5	42.7±2.6
R. raoultii																						100.00	77.2±2.8	72.2±2.9	74.4±2.9	76.7±2.9	55.8±2.7	32.6±2.5
R. rhipicephali																							100.00	69.5±2.9	71.5±2.9	73.7±2.9	54.4±2.7	32.8±2.5
R. rickettsii																								100.00	82.7±2.7	85.5±2.5	52.6±2.7	32.3±2.5
R. sibirica																									100.00	90.5±2.1	53.4±2.7	32.3±2.5
R. slovaca																										100.00	54.5±2.7	32.6±2.5
R. tamurae																											100.00	33.1±2.5
R. typhi																												100.00

Table S3: OrthoANI values (%) obtained by pairwise comparisons of studied genomes (upper right)

	R. aeschlimannii	*R. africae*	*R. akari*	*R. amblyommatis*	*R. asembonensis*	*Rickettsa australis*	*R. bellii*	*R. canadensis*	*R. conorii*	*R. heilongjiangensis*	*R. felis*	*R. fournieri*	*R. helvetica*	*R. honei*	*R. hoogstraalii*	*R. japonica*	*R. massiliae*	*R. montanensis*	*R. parkeri*	*R. peacockii*	*R. prowazekii*	*R. raoultii*	*R. rhipicephali*	*R. rickettsii*	*R. sibirica*	*R. slovaca*	*R. tamurae*	*R. typhi*
R. aeschlimannii	100.00	96.56	91.38	96.77	92.98	92.01	80.92	88.79	96.71	96.72	92.56	**96.91**	92.88	96.73	95.11	96.86	97.17	97.00	96.51	96.77	87.69	97.25	97.31	96.41	96.74	96.95	93.94	87.29
R. africae		100.00	91.28	96.7	92.93	91.93	80.9	88.79	98.91	97.48	92.64	**97.86**	92.98	98.37	92.87	97.58	96.77	96.98	98.97	98.63	87.46	97.28	96.88	98.2	99.00	98.97	93.89	87.4
R. akari			100.00	91.32	92.75	95.53	80.56	88.02	91.39	91.53	92.66	**91.6**	91.13	91.47	93.17	91.26	91.65	91.6	91.39	91.48	87.00	91.59	91.65	91.32	91.3	91.52	91.62	86.98
R. amblyommatis				100.00	92.92	91.98	81.04	88.78	96.82	96.87	96.85	**96.72**	92.88	96.83	93.13	96.88	96.81	97.15	96.83	96.86	87.63	97.47	97.12	96.63	96.82	97.14	93.69	87.41
R. asembonensis					100.00	93.55	81.5	89.12	92.9	92.85	95.82	**93.17**	92.81	92.75	95.63	92.83	93.07	93.18	92.91	93.03	87.69	93.2	93.13	92.63	92.76	93.11	93.44	87.41
Rickettsa australis						100.00	80.9	88.36	92.01	91.87	93.29	**92.36**	91.75	92.11	93.99	91.99	92.02	92.12	91.94	92.1	87.5	92.37	92.24	91.72	91.88	91.96	92.42	87.32
R. bellii							100.00	80.71	80.77	80.96	80.89	**81.37**	81.26	80.84	81.45	80.95	81.15	81.55	80.91	81.07	79.74	81.1	80.93	80.84	80.94	80.81	81.35	79.74
R. canadensis								100.00	88.86	89.06	88.84	**89.00**	89.52	88.83	89.15	88.86	88.85	89.18	88.78	88.82	85.79	88.94	88.96	88.97	88.89	88.91	89.17	85.32
R. conorii									100.00	97.53	92.61	**97.9**	93.00	98.44	95.11	97.65	97.00	97.12	98.97	98.68	87.68	97.36	97.05	98.17	98.95	99.01	93.89	87.51
R. heilongjiangensis										100.00	92.49	**98.91**	93.1	97.6	93.05	99.19	96.96	97.11	97.6	97.71	87.73	97.38	97.04	97.3	97.5	97.8	93.94	87.59
R. felis											100.00	**92.91**	92.55	92.6	95.35	92.64	92.7	92.94	92.55	92.69	87.51	92.89	92.56	92.38	92.58	92.58	93.03	87.32
R. fournieri												**100.00**	**93.19**	**97.98**	**93.39**	**98.98**	**97.22**	**97.38**	**97.8**	**98.03**	**87.85**	**97.71**	**97.35**	**97.72**	**97.89**	**98.26**	**94.06**	**87.42**
R. helvetica													100.00	93.07	92.93	93.03	93.1	93.25	92.81	92.91	87.74	93.44	93.09	93.00	93.14	93.13	93.88	87.43
R. honei														100.00	93.11	97.7	96.99	97.14	98.48	98.52	87.74	97.35	97.04	98.15	98.4	98.74	94.05	87.45
R. hoogstraalii															100.00	93.1	93.22	93.2	93.15	93.16	87.86	93.27	93.12	92.84	92.91	93.12	93.43	87.51
R. japonica																100.00	97.08	97.2	97.73	97.84	87.62	97.48	97.11	97.45	97.63	97.91	94.07	87.31
R. massiliae																	100.00	97.21	96.82	97.03	87.7	97.52	98.58	96.71	96.97	97.14	93.99	87.46
R. montanensis																		100.00	97.02	97.22	87.65	97.74	97.36	96.88	97.07	97.32	94.37	87.3
R. parkeri																			100.00	98.65	87.51	97.31	97.03	98.3	99.17	99.03	93.79	87.43
R. peacockii																				100.00	87.7	97.47	97.19	98.65	98.64	98.99	94.1	87.31
R. prowazekii																					100.00	87.76	87.49	87.53	87.66	87.62	87.71	92.39
R. raoultii																						100.00	97.65	97.11	97.34	97.62	94.4	87.57
R. rhipicephali																							100.00	96.73	96.98	97.23	94.14	87.43
R. rickettsii																								100.00	98.23	98.51	93.59	87.64
R. sibirica																									100.00	99.00	93.84	87.4
R. slovaca																										100.00	94.18	87.43
R. tamurae																											100.00	87.67
R. typhi																												100.00

SUPPLEMENTARY FIGURE LEGENDS

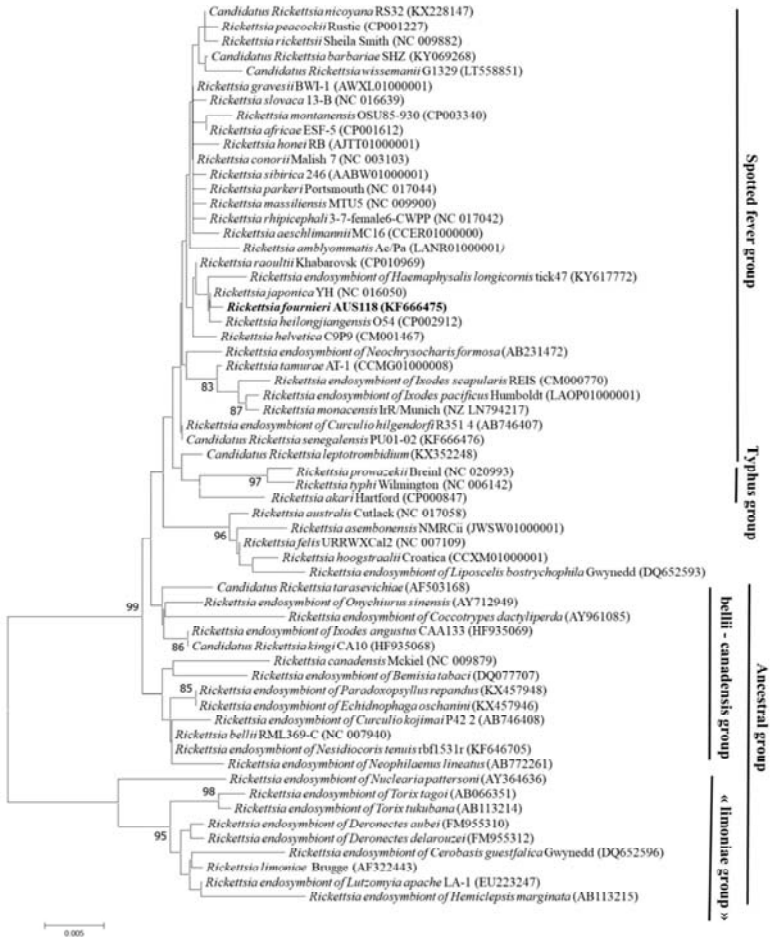


Figure S1: Phylogenetic tree highlighting the position of *Rickettsia fournieri* strain AUS118^{T} relative to other closely related *Rickettsia* species based on the 16S rRNA gene sequence analysis. Sequences were aligned using CLUSTALW, with default parameters. Phylogenetic analysis was inferred by Maximum Likehood method with the Kimura 2-parameter model within the MEGA6 software. Numbers at the nodes represent the percentages of bootstrap values obtained by repeating analysis 500 times to generate a majority consensus tree. Only values higher than 95 % are shown. The scale bar represents a 2 % nucleotide sequence divergence.

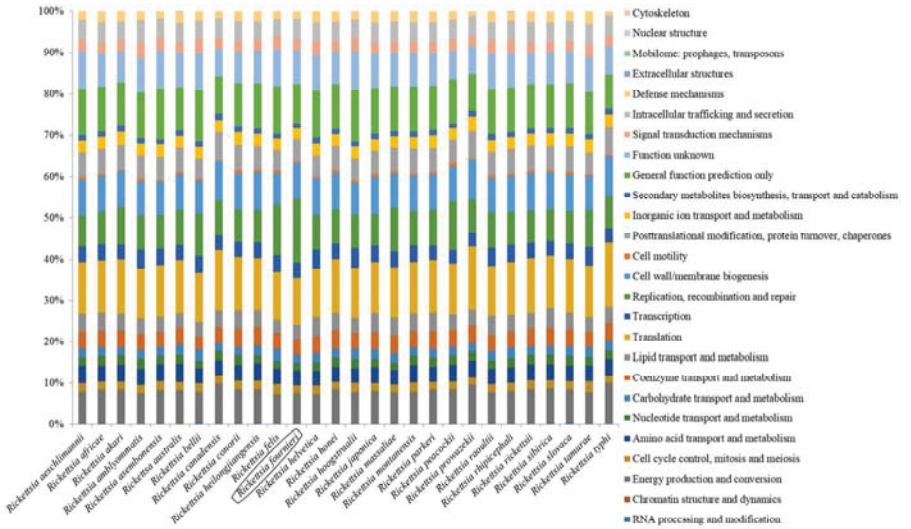


Figure S2: Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of *Rickettsia fournieri* strain AUS118^{T} among other *Rickettsia* species.

CHAPITRE III

Taxono-génomique: Utilisation des données génomiques pour la description taxonomique des nouveaux isolats bactériens issues du projet « culturomique »

Avant-propos

Cette partie a été consacrée à la caractérisation et à la description taxonomique de 17 nouvelles espèces bactériennes isolées à partir de divers échantillons cliniques par méthode « culturomique bactérienne», exceptée une qui a été isolée à partir de nourriture salée. Depuis 2009, un nouveau concept « microbial culturomics » a été développé au sein de notre laboratoire. Il s'agit d'un concept révolutionnaire de culture microbienne qui repose sur la variation des paramètres physico-chimiques des conditions de culture, dans le but d'explorer au maximum la diversité microbienne. Il s'appuie sur une méthode d'identification rapide des isolats par spectrométrie de masse MALDI-TOF complétée par le séquençage de l'ARNr 16S lorsque cela est nécessaire. La culturomique a permis l'isolement de plus de 1000 espèces bactériennes distinctes associées à l'homme depuis 2012, y compris environ 400 nouvelles espèces. La description taxonomique des nouvelles espèces bactériennes a évolué au cours du temps en fonction des nouveaux outils disponibles. Initialement basée sur des caractéristiques phénotypiques y compris la morphologie et les tests biochimiques, l'hybridation ADN-ADN, la teneur en G+C% et l'analyse de la similarité et la phylogénie des séquences de l'ARNr 16S ont été intégrées dans la description

des nouveaux taxons dans une approche polyphasique. Cette approche polyphasique est à la base de la classification taxonomique la plus largement acceptée des procaryotes. Cependant, le progrès remarquable des technologies de séquençage à haut débit, de plus en plus performantes et de moins en moins chères a permis l'accès sans précédent à des données du genome entier. Ainsi, l'intégration des informations génomique notamment les données de séquençage du génome entier et la comparaison des caractéristiques génomiques a été recommandée pour la description taxonomique des nouvelles espèces. En 2014, une méthode innovante appelée "taxono-genomics" a été développée dans notre laboratoire pour la caractérisation et la description des nouvelles espèces bactériennes. Ce concept « taxono-genomics » est une approche polyphasique qui intègre les informations génomiques à savoir les données de séquençage et de l'analyse fonctionnelle et les données de l'analyse comparative de similarité des séquences génomiques, les informations protéomiques obtenues par spectrométrie de masse (MALDI-TOF MS), en plus de la description phénotypique. Cette approche polyphasique surmonte les limites des méthodes conventionnelles basées sur les caractéristiques génotypiques, phénotypiques et chimiotaxonomiques pour la description de nouvelle espèce.

Dans cette partie nous présentons d'abord une revue qui examine le développement des cultures et de la génomique dans le domaine de la microbiologie clinique et leur impact sur la taxonomie bactérienne en tenant compte de l'apport de la génomique (**Article 6**).

Ensuite, nous présentons la caractérisation génomique et la description de 17 nouvelles espèces incluant 3 espèces halophiles dont 2 isolées dans la nourriture (*Gracilibacillus massiliensis et Bacillus salis*) et une isolée dans le tube digestif humain (*Gracilibacillus timonensis*), 8 isolées dans le vagin de patients souffrant de vaginose bactérienne (*Khoudiadiopia massiliensis, Olegusella massiliensis, Murdochiella vaginalis, Prevotella lascolaii, Collinsella vaginalis, Peptoniphilus vaginalis, Peptoniphilus raoultii, Peptoniphilus pacaensis*), 3 isolées à partir d'échantillon de selles de patients obéses (*Butyricimonas phoceensis, Eisenbergiella massiliensis, Mediterraneibacter phoceensis*), une nouvelle espèce isolée à partir d'échantillons fécaux d'un Bedouin sain à l'Arabie saoudite (*Raoultibacter massiliensis*), une nouvelle espèce isolée à partir des excréments d'un pygmée femelle vivant au Congo (*Raoultibacter timonensis*) et une nouvelle espèce de *Bartonella* isolée chez des rongeurs *Mastomys erythroleucus* (*Bartonella mastomydis*).

Article 5:

The impact of culturomics on taxonomy in clinical microbiology

Rita Abou Abdallah, Mamadou Beye, Awa Diop, Sofiane Bakour, Didier Raoult, Pierre-Edouard Fournier

[Published in Antonie van Leeuwenhoek]

Antonie van Leeuwenhoek
DOI 10.1007/s10482-017-0871-1

ORIGINAL PAPER

The impact of culturomics on taxonomy in clinical microbiology

Rita Abou Abdallah · Mamadou Beye · Awa Diop · Sofiane Bakour · Didier Raoult · Pierre-Edouard Fournier

Received: 19 January 2017 / Accepted: 4 April 2017

Abstract Over the past decade, new culture methods coupled to genome and metagenome sequencing have enabled the number of isolated bacterial species with standing in nomenclature to rise to more than 15,000 whereas it was only 1791 in 1980. 'Culturomics', a new approach based on the diversification of culture conditions, has enabled the isolation of more than 1000 distinct human-associated bacterial species since 2012, including 247 new species. This strategy was demonstrated to be complementary to metagenome sequencing for the exhaustive study of the human microbiota and its roles in health and diseases. However, by identifying a large number of new bacterial species in a short time, culturomics has highlighted a need for taxonomic approaches adapted to clinical microbiology that would include the use of modern and reproducible tools, including high throughput genomic and proteomic analyses. Herein, we review the development of culturomics and genomics in the clinical microbiology field and their impact on bacterial taxonomy.

Keywords Culturomics · Bacteria · Human microbiota · Taxonomy · Genome

Introduction

The isolation and description of microorganisms are essential for understanding their relationships with other living organisms. Over the past two decades, several important technical advances have marked clinical microbiology, including genome sequencing, the development of new culture strategies and identification of clinical isolates using MALDI-TOF mass spectrometry (MS) (Fournier et al. 2015). In addition, the emergence of high throughput metagenomics (Marchesi and Ravel 2015) has enabled the deciphering of the human microbiota and demonstrated that diseases may not exclusively result from the presence of a pathogen but also from an imbalance among members of the physiological microbiota, a phenomenon also referred to as dysbiosis (Karlsson et al. 2013). This made the scientific community neglect classical culture techniques for being fastidious and unable to isolate new microorganisms. However, metagenomics exhibits a number of drawbacks, notably the ignorance of minor populations, present at a concentration lower than 10^5 CFU/ml and the

Electronic supplementary material The online version of this article (doi:10.1007/s10482-017-0871-1) contains supplementary material, which is available to authorized users.

R. A. Abdallah · M. Beye · A. Diop · S. Bakour · D. Raoult · P.-E. Fournier (✉)
Unité de recherche sur les maladies infectieuses et tropicales émergentes (URMITE), UM 63CNRS 7278IRD 198Inserm 1095IHU Méditerranée Infection, Faculté de Médecine, Aix-Marseille Université, 27 Bd Jean Moulin, 13385 Marseille Cedex5, France
e-mail: pierre-edouard.fournier@univ-amu.fr

Published online: 07 April 2017

unreliable taxonomic characterisation of microbiota members at the species level (Lagier et al. 2012; Sankar et al. 2015). These disadvantages and the need to fully characterise bacteria motivated some researchers to express more interest in culture by developing new techniques aiming at growing previously uncultured bacteria (Overmann and Garcia-Pichel 2013; Overmann 2015). Among these methods, 'culturomics', first developed in 2012 and based on the diversification of culture conditions to mimic as closely as possible the natural environments in which bacteria live, has enabled the isolation of more than 1000 bacterial species from the human gut over the past five years (Lagier et al. 2012, 2016).

One of the most significant effects of the above-mentioned technical progresses on microbiology has been the rapid increase in the number of bacterial species with validly published names. Currently, more than 15,000 species have standing in nomenclature (www.bacterio.net) whereas this number was 1791 in the first list of validated prokaryotes published in 1980 (Tindall et al. 2010). Moreover, the number of available bacterial genome sequences has exploded in the past decade, following the introduction of high throughput sequencing methods (HTS) and is currently more than 60,000. Concurrent with these changes, taxonomy has also evolved over the years in order to fit the needs of the scientific community and to take advantages of the available data (Fournier et al. 2015). In this review we revisit the importance of culture in the clinical microbiology field, and we emphasise the outcomes of the culturomics revolution along with its impact on taxonomy and the evolution of the latter.

Evolution of culturing approaches

In medical microbiology, the isolation and growth of microorganisms is crucial for diagnostic purposes and the establishment of an effective treatment. Bacterial culture also has a key role in the evaluation of antibiotic susceptibility and virulence, and enables genomic studies (Singh et al. 2013; Lagier et al. 2015a). To cope with the variety of bacterial growth requirements, microbiologists may vary four essential parameters that include nutrient choice, atmosphere, temperature and incubation time (Lagier et al. 2015b).

The first culture media used in bacterial culture were mainly composed of cooking ingredients or the extracts of environmental elements. Indeed, the initial culture substrates included meat infusions, heart or brain extracts, vegetables and yeast extracts that remain among the major components of many media. In addition to these nutritional elements, peptones, casein, soy and gelatin were, and still are, often used as additives in culture media (Lagier et al. 2015b). By using solidifying components such as gelatin, agar or coagulated eggs, microbiologists were later able to observe bacterial colonies on solid culture media. This microbiological advance also allowed the description of bacterial species (Lagier et al. 2015b). However, although the nutrients cited above are used to compose the most common culture media for prokaryotes, microbiologists soon noticed that these media do not make provision for the growth of all bacteria, especially those that are fastidious. In order to facilitate the growth of these fastidious bacteria, media were enriched with a number of additives, notably blood (Drancourt et al. 2003; Drancourt and Raoult 2007). Then, selective culture media were developed to isolate specific pathogenic microorganisms from complex microbial communities. These differential media contained various substrates inhibiting the growth of undesired species. An example is given by the Chapman agar (culture medium enriched in NaCl) for the isolation of *Staphylococcus* species. Several antibiotics and antiseptics such as bromocresol purple are also used in culture media to inhibit the growth of some bacterial genera or species and select others (LeChevallier et al. 1983; Subramanyam et al. 2012).

Temperature is one of the most relevant factors influencing bacterial growth (Guijarro et al. 2015), ranging from ice surfaces (Antony et al. 2012) to hot springs (Liu et al. 2016), and the optimal growth temperatures of bacteria are species-dependent. In medical microbiology, most human-associated species, pathogenic or not, are mesophilic, growing at temperatures ranging between 25 and 45 °C (Lagier et al. 2015a).

In addition to the temperature, the atmosphere is also essential for the isolation and identification processes. Indeed, a primary characteristic is whether an organism grows aerobically, anaerobically, or microaerobically (Lagier et al. 2015a).

Finally, bacterial growth is also dependent on the incubation time. Most clinical pathogens grow easily

within 24–48 h of incubation (Lagier et al. 2015a), but several bacteria require a much longer incubation time, up to several days, as observed for *Helicobacter* species (Jiang and Doyle 2002) or weeks as is the case for some *Mycobacterium* species or *Tropheryma whipplei* (Simner et al. 2016).

The culturomics approach, a powerful tool to study the human microbiota

Studying complex microbiotas, notably those associated with humans, and their roles in health and diseases, has long been a challenge (Turnbaugh et al. 2007). The first microbiota studies were mainly based on culture (Finegold et al. 1974). However, the introduction of molecular biology methods in microbiology led to a progressive disinterest in culture based approaches, notably for the study of complex microbial communities. In particular, metagenomic studies dramatically expanded the known diversity of the human microbiome (Andersson et al. 2008; Turnbaugh et al. 2010; Claesson et al. 2010) and demonstrated that a majority of human-associated bacteria were not cultivable using standard techniques (Schmeisser et al. 2003; Turnbaugh et al. 2007). In the past few years, the number of publications on the human microbiome has massively expanded (Hiergeist et al. 2015) and clear links between the microbiota composition and many disorders such as obesity (Armougom et al. 2009), diabetes (Larsen et al. 2010), Crohn's disease, necrotizing enterocolitis, colo-rectal cancer (De Hertogh et al. 2006; Siggers et al. 2008), immune response variation (Kau et al. 2011), depression, anxiety and autism (Wang and Kasper 2014) have been presented (Hugon et al. 2016). However, metagenomic and other molecular biology techniques have several drawbacks, including the fact that a large fraction of obtained sequences have not been assigned to a known microorganism (Raoult 2016), that the primers used may not amplify all bacteria, that the DNA may not be homogenously extracted depending on the species and that bacteria present at a concentration lower than 10^5 CFU/mL may not be detected, even if they are clinically relevant (Lagier et al. 2012).

In addition to these limitations of metagenomics, the need to study the pathogenicity, antibiotic susceptibility, metabolic pathways and other phenotypic characteristics, as well as to elaborate new diagnostic tools (Singh et al. 2013), prompted many researchers over the past two decades to design new culture strategies and media for the isolation of uncultured bacteria (Goodman et al. 2011; Bomar et al. 2011). Many studies were conducted to isolate a maximum of previously uncultured bacteria, especially from the human gut. In 2011, Kim et al. used three culture media: brain heart infusion broth, and high- and low-carbohydrate medium with different growth supplements to study the human gut microbiota (Kim et al. 2011). In the same year, the concept of culture-enriched molecular profiling was launched and was used to study the airways microbiota of cystic fibrosis patients (Sibley et al. 2011) and then for the study of the human gut microbiota (Lau et al. 2016). In 2012, Lagier et al. launched the concept of culturomics (Lagier et al. 2012). This approach is based on the diversification of culture conditions to mimic as closely as possible the natural environments in which bacteria live, coupled to the use of MALDI-TOF MS and, when necessary, 16S rRNA gene amplification and sequencing, to identify bacterial colonies. In this article, we mainly focus on studies that were conducted on the human gut microbiota. In their first study, by testing 212 different culture conditions on three stool samples, Lagier et al. screened 32,500 colonies, representing 340 bacterial species including 31 putative new species (Lagier et al. 2012). The term culturomics was coined by analogy with other—OMICS strategies (genomics, metagenomics, proteomics, metabolomics…) for a method allowing an extensive assessment of the microbial composition by high-throughput culturing (Greub 2016).

The comparison of metagenomics and culturomics for the study of the human gut microbiota showed that the overlap in detected genera and species between both methods was less than 10%, each strategy identifying specific taxa (Lagier et al. 2012). More specifically, in this early study, culturomics was less efficient than metagenomics for the detection of anaerobic bacteria despite a high workload that consisted in cultivating the samples in 212 different culture conditions (Lagier et al. 2012). In order to overcome these initial weaknesses, several changes were made. A careful analysis showing that all the identified bacterial species could be isolated using only 70 of the 212 culture conditions led to a reduction of these conditions to 70 (Lagier et al. 2012). In 2014, this number was once more reduced, to the 12 culture

conditions enabling the greatest number and diversity of cultures. This decision was based on the identification of three essential steps to isolate the maximal number of microorganisms: (i) a pre-incubation in a blood culture bottle (56% of the new species isolated); (ii) the addition of filter-sterilised rumen fluid for this pre-incubation (40% of the new species isolated); and (iii) the addition of 5% sheep blood (25% of the new species isolated) (Lagier et al. 2015a). This refinement resulted in reducing the workload and extending the stool testing capacity. Another improvement was the systematic detection of micro-colonies grown on agar (Lagier et al. 2016). These bacterial colonies, exhibiting diameters ranging from 100 to 300 µm, are barely visible to the naked eye. Magnifying glasses were used to visualise the micro-colonies. Finally, the culture of halophilic bacteria was implemented using culture media supplemented with salt (Lagier et al. 2016).

Performance of cuturomics

Following the first two published studies (Lagier et al. 2012, 2015a), several other culturomics projects were conducted, including the analyses of the gut microbiotas from premature infants with necrotizing enterocolitis, pilgrims returning from the Hajj and patients before or after bariatric surgery (Lagier et al. 2016). In another study, 28 fresh stool samples were inoculated in order to overcome the impact of storage and processing delays, especially for anaerobic bacteria. Then studies focused on the isolation of proteobacteria, microaerophilic bacteria, halophilic prokaryotes and microcolonies. Finally, differences in bacterial composition of duodenal, small bowel intestine and colonic samples were evaluated (Lagier et al. 2016).

Briefly, the culture of around 1000 stool samples using culturomics has enabled the isolation of 1170 out of the 1525 currently known human gut prokaryotes (Lagier et al. 2016). These numbers show the high throughput capacity of culturomics and they are detailed in Table 1. The bacterial species identified using culturomics belong to ten different phyla (Fig. 1), including 630 within the phylum *Firmicutes* with the most represented genera being *Clostridium*, *Paenibacillus*, *Staphylococcus* and *Streptococcus*; 225 are classified in the phylum *Actinobacteria* (mostly in the genus *Corynebacterium* with 36 species); 187 belong to the phylum *Proteobacteria* (28 of them are *Pseudomonas* species); and 102 are classified in the phylum *Bacteroidetes*. In addition, ten, four, one, one and one species are classified in the phlya *Fusobacteria*, *Synergistetes*, *Deinococcus-Thermus*, *Lentispharae* and *Verrucomicrobium*, respectively. In addition, we isolated eight archaea, including one NS, five NH, one H and one HGUT (as defined in Table 1). On the other hand, laboratories studying the human gut using methods other than culturomics identified only 477 species belonging to 11 phyla. Among these, the *Synergistetes* and *Deinococcus-Thermus* are not represented, while the members of the phyla *Chlamydiae, Spirochetae* and *Tenericutes* phyla were identified in these studies but missing from the culturomics project.

Table 1 Culturomics results

Categorie	Isolated bacterial species count
Total	1170
NS	247
NH	269
H	250
HGUT	404

NS new species, *NH* prokaryotes isolated for the first time in humans, *H* prokaryotes already known in humans but isolated for the first time in the gut, *H*(*GUT*) prokaryotes known in the human gut but newly isolated by culturomics

Among the bacterial species identified using culturomics, the 247 new species belong to 6 distinct phyla, including 159 that were classified in the phylum *Firmicutes*. Within this phylum, the most represented genera were *Clostridium, Paenibacillus* and *Peptoniphilus*, which contain anaerobic bacteria, and *Bacillus* that includes facultative aerobes. Forty-two new species belong to the phylum *Actinobacteria*, the most represented genera being *Actinomyces* and *Corynebacterium* which are respectively facultative anaerobic and aerobic bacteria; thirty-one species were classified as belonging to the phylum *Bacteroidetes*, with *Alistipes* and *Bacteroides* being the most represented genera (both include anaerobic bacteria); twelve species belong to the *Proteobacteria* phylum; and the *Fusobacteria* and *Synergistetes* phyla each contain a new anaerobic species. In conclusion, the culturomics approach has doubled the number of known human gut bacteria, including microorganisms that had previously been detected using metagenomics but had remained unassigned due to the lack of an

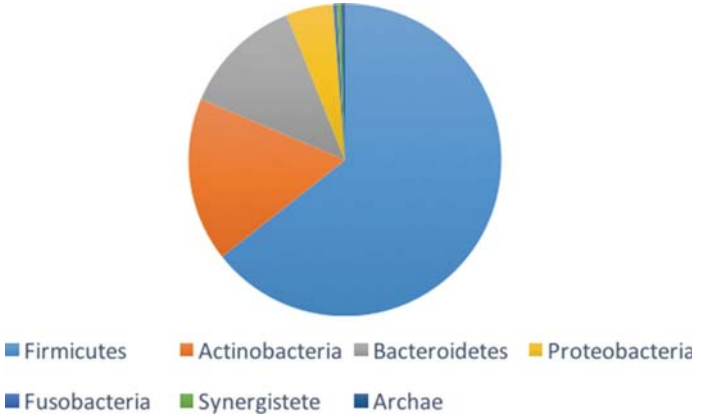


Fig. 1 Distribution of the new species isolated using culturomics in bacterial phyla

isolate to complete their characterisation. Therefore, a large panel of new species, mostly anaerobic, have been obtained in a short period of time, resulting in a need for modern tools enabling their proper characterisation and taxonomic classification.

The evolution of bacterial taxonomy

In 1872 Cohn compiled the first taxonomic description by characterising six genera of bacteria, including *Micrococcus luteus,* on the basis of their morphology (Schleifer 2009). At the beginning of the 20th century, more and more physiological and biochemical properties were used, and bacterial taxonomy relied on a combination of phenotypic characteristics such as colony size and colour, staining properties using Ziehl-Neelsen and Gram staining, motility, morphology and growth requirements, in addition to ultrastructure and chemical composition of the cell wall and outer membrane, metabolic pathways and protein composition (Collins 2004; Schleifer 2009).

Between 1960 and 1980, new parameters were added, notably chemotaxonomy (Minnikin et al. 1975), genomic DNA-DNA hybridization, G+C content and numerical taxonomy (Johnson 1973; Brenner et al. 1969; Johnson 1991) (Fig. 2). In the 1980s, the advent of DNA amplification and sequencing techniques, in particular of the 16S rRNA gene, constituted a major progress in bacterial taxonomy by enabling reclassification of many strains, leading to the creation of many new species (Vandamme and Coenye 2004; Goris et al. 2007). In 1980, the first Approved List of bacterial names was created and the number of bacterial species was reduced from 30,000 to 1800 (Skerman et al. 1989).

Currently, prokaryotic taxonomy relies on a 'polyphasic' combination of available phenotypic and genotypic data introduced in 1996 by Vandamme et al. (Vandamme and Coenye 2004; Vandamme et al. 1996). This was refined by Tindall et al. (2010) who proposed using 16S rRNA gene sequence similarity and phylogeny, followed by genomic DNA G+C content, DNA-DNA hybridization (DDH), cell morphology and Gram-staining properties, as well as phenotypic and chemotaxonomic criteria (Tindall et al. 2010) (Fig. 2).

Among the genotypic criteria, DNA–DNA hybridization (DDH) is used to estimate the genetic relatedness between microorganisms. A DDH value $\leq$70% indicates that the tested bacteria belong to distinct species (Wayne et al. 1987). The DNA G+C content of prokaryotes may also be used to classify prokaryotes (Ramasamy et al. 2014; Kim et al. 2015), a difference higher than 1–5% reflecting distinct species and a difference higher than 10% reflecting distinct genera. However, it is not applicable to all genera (Wayne et al. 1987) and errors in laboratory methods are evident (Kim et al. 2015). Regarding the 16S rRNA sequence identity and phylogenetic analysis (Fox et al. 1992; Hugenholtz et al. 1998; Ludwig and Klenk 2001), in 1994,

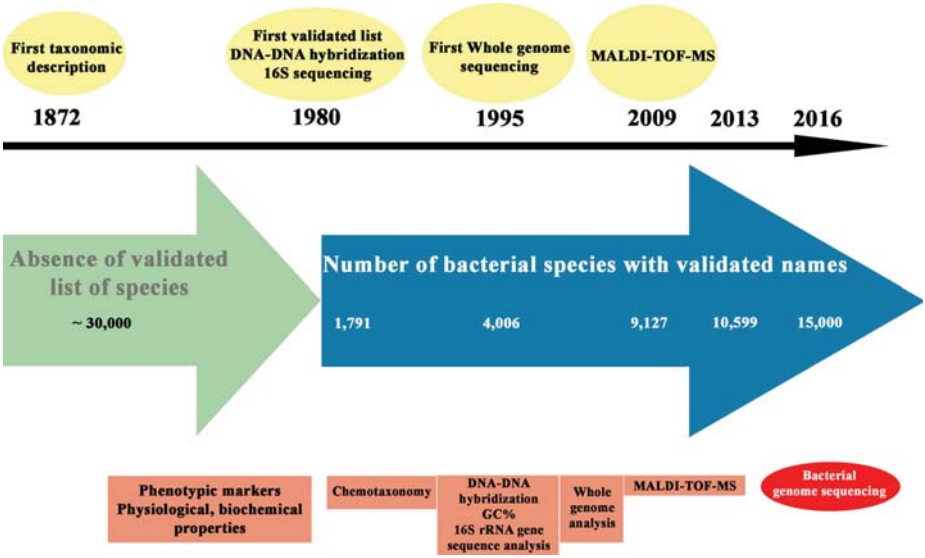


Fig. 2 Evolution of bacterial taxonomy. The most important changes in bacterial taxonomic tools over the years, as well as the number of species with standing in nomenclature

scientists considered that two bacteria belonged to a distinct genus if they shared a 16S rRNA sequence similarity lower than 95%, and to two different species if this value was between 95 and 97% (Stackebrandt and Goebel 1994). In 2006, the latter threshold value was re-evaluated at 98.7% (Stackebrandt and Ebers 2006) and then 98.65% (Kim et al. 2014).

Limitations of the traditional taxonomic tools

Currently, there is no universal strategy for the classification of prokaryotes, which thus remains a matter of debate. The most widely used methods present several inconsistencies and limitations. First, the 16S rRNA gene sequence similarity thresholds are not applicable to multiple genera (Rossi-Tamisier et al. 2015), the multiple rRNA operons in a single genome may exhibit nucleotide variations (Ramasamy et al. 2014), and some of the 16S rRNA gene copies may be acquired by horizontal gene transfer which may distort taxa relationships in phylogenetic trees (Zhi et al. 2012). Second, despite the fact that DDH has been considered as a gold standard for the taxonomic classification of prokaryotes (Ramasamy et al. 2014), the 70% threshold is not applicable to all bacterial genera (Sentausa and Fournier 2013), the method lacks reproducibility between and within laboratories, and the DDH experiments are labour-intensive (Azevedo et al. 2015).

Use of genome sequences in taxonomy

The sequencing of the first bacterial genome, that of *Haemophilus influenzae*, marked the beginning of the genomic era (Fleischmann et al. 1995). It was a major step forward in microbiology, by giving access to the full genetic content of a bacterial strain. This led many researchers to propose using genomic sequences as a source of taxonomic parameters such as the presence or absence of genes within genomes, chromosomal gene order, comparison of orthologous genes and the presence of indels or single nucleotide polymorphisms (Snel et al. 1999; Huson and Steel 2004). However, genome sequencing remained labour and money

consuming (Ramasamy et al. 2014) until the development of high-throughput sequencing technologies that resulted in a progressive decrease in genomic sequencing costs. Subsequently, the genomic sequences of thousands of bacteria have become increasingly available. To date, several genome based taxonomic tools have been proposed as alternatives to DDH. Multilocus sequence analysis (MLSA) is based on sequence analysis of multiple protein coding genes and uses these sequences to create phylogenetic trees and delineate species within a genus (Glaeser and Kämpfer 2015). It is derived from multilocus sequence typing (MLST) that was first introduced by Maiden et al. in 1998 for strain genotyping (Maiden et al. 1998). However, although it was proposed that a 3% sequence divergence of concatenated gene sequences was equivalent to the 70% DDH threshold (Vanlaere et al. 2009), this value is not a universal cutoff and does not apply to many genera (Glaeser and Kämpfer 2015). In addition, various overall genome relatedness indices (OGRI) were proposed. The Average Nucleotide Identity (ANI) (Goris et al. 2007), calculated from two genome sequences using BLASTn, was demonstrated to be a valid alternative to DDH, with an ANI value of 95–96% corresponding to a 70% DDH. Several new species were described using this method, such as *Dehalococcoides mccartyi* (Löffler et al. 2013) and *Streptococcus dentisani* (Camelo-Castillo et al. 2014). However, since the concept of ANI derives from DDH, it presents the same drawback, which is the inequality of two reciprocal values and should not be used as a single tool for prokaryotic classification (Tindall et al. 2010). In order to overcome this drawback, Lee et al. developed orthoANI, in which genomic fragments are reciprocally searched using BLASTn (Lee et al. 2016). The maximum unique matches index (MUMi), based on DNA conversation of the core genome as well as the proportion of shared DNA by two genomes, is well correlated with DDH and ANI, but is not applicable to draft genomes (Richter and Rosselló-Móra 2009). The GGDC online software (http://ggdc.dsmz.de/distcalc2.php) allows the genome to genome comparison and the study of genetic relatedness degree among bacterial isolates by determination of digital DDH (dDDH). Ramasamy et al. developed the AGIOS parameter obtained by identifying orthologous genes using BLASTP and then determining the mean percentage of nucleotide sequence identity using the Marseille Average Genomic Identy (MAGi) pipeline (Ramasamy et al. 2014). This approach does not use a universal cutoff and is always combined with phenotypic criteria for taxonomic purposes. However despite the decreasing cost of sequencing and the growing number of microbiologists supporting the incorporation of genome sequence analysis into taxonomy (Vandamme and Peeters 2014), the whole genome sequence information of prokaryotic strains has only been accepted recently by taxonomists.

An example of integrating genome analysis in prokaryotic taxonomy: 'taxono-genomics'

Coming from the need to characterise and classify the large number of new bacteria isolated by culturomics, a strategy named taxono-genomics was proposed and adopted recently in our laboratory for the description and classification of new bacterial species (Ramasamy et al. 2014). Taxono-genomics is a polyphasic approach that systematically combines genomic and MALDI-TOF MS data with other phenotypic and genotypic criteria for the taxonomic circumscription of bacterial species. Briefly, this approach includes several steps summarised as follows: a putative new species is suspected when exhibiting a MALDI-TOF MS score <2 and a 16S rRNA sequence similarity with the closest related species with standing in nomenclature is <98.7%. Then, its complete genome sequence is compared to those of phylogenetically close species or genera in terms of size, DNA G+C content, percentage of coding sequences, gene content, numbers of RNA genes, gene distribution in COG categories (Tatusov et al. 2001), presence of mobile genetic elements, signal peptides and transmembrane helices. The degree of genetic relatedness between the compared bacterial isolates is also evaluated by determination of the digital DDH using the GGDC online software (http://ggdc.dsmz.de/distcalc2.php) and of the average of genomic identity of orthologous gene sequences (AGIOS) using the MAGI software. To date, this taxono-genomics strategy has been used to describe more than 80 novel species and genera including *Gracilibacillus massiliensis* (Diop et al. 2016), *Anaerococcus rubiinfantis* (Tidjani Alou et al. 2016) or *Senegalimassilia anaerobia* (Lagier et al. 2013) (Supplementary Table 1). Therefore, genomic and MALDI-TOF MS data may be used as efficient

alternatives to chemotaxonomy for the description of bacteria (Fournier and Drancourt 2015).

Conclusion

Over the past few years, culturomics has stimulated the field of microbiology by enabling the isolation of many human-associated bacteria and thereby has helped precipitate a taxonomic challenge. Several initiatives and new publication formats have been proposed to simplify and accelerate the publication of new bacterial species. These include the Digital protologue and New Species Announcement article formats (Rossello-Mora et al. 2017; Fournier et al. 2016). Coordination of these new intiatives (and reconciliation with the requirements of the International Code of Nomenclature of Prokaryotes) is likely to be of importance in the next few years.

As culturomics will be carried out at larger scales on different types of microbiotas, neglecting genome sequences, which give access to the full genetic information of prokaryotes for an acceptable cost, does not seem justifiable for their taxonomic classification (Sutcliffe 2015). In addition, as the number of genomes from species with standing in nomenclature is continuously increasing, obtaining taxonomic information from genomic comparisons will soon be achievable by most scientists. Therefore, genomic data represent today a valid alternative, in combination to phenotypic criteria, to chemotaxonomic approaches for the taxonomic description of new bacterial species.

Compliance with ethical standards

Conflict of interest The authors declares that they do not have conflict of interest.

References

Andersson AF, Lindberg M, Jakobsson H et al (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS ONE 3:e2836. doi:10.1371/journal.pone.0002836

Antony R, Krishnan KP, Laluraj CM et al (2012) Diversity and physiology of culturable bacteria associated with a coastal Antarctic ice core. Microbiol Res 167:372–380. doi:10.1016/j.micres.2012.03.003

Armougom F, Henry M, Vialettes B et al (2009) Monitoring bacterial community of human gut microbiota reveals an increase in lactobacillus in obese patients and methanogens in anorexic patients. PLoS ONE 4:e7125. doi:10.1371/journal.pone.0007125

Azevedo H, Lopes F, Silla P, Hungria M (2015) A database for the taxonomic and phylogenetic identification of the genus Bradyrhizobium using multilocus sequence analysis. BMC Genom 16(Suppl 5):S10. doi:10.1186/1471-2164-16-S5-S10

Bomar L, Maltz M, Colston S, Graf J (2011) Directed culturing of microorganisms using metatranscriptomics. mBio 2:e00012–11. doi:10.1128/mBio.00012-11

Brenner DJ, Fanning GR, Rake AV, Johnson KE (1969) Batch procedure for thermal elution of DNA from hydroxyapatite. Anal Biochem 28:447–459. doi:10.1016/0003-2697(69)90199-7

Camelo-Castillo A, Benítez-Páez A, Belda-Ferre P et al (2014) *Streptococcus dentisani* sp. nov., a novel member of the mitis group. Int J Syst Evol Microbiol 64:60–65. doi:10.1099/ijs.0.054098-0

Claesson MJ, Wang Q, O'Sullivan O et al (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res 38:gkq873–e200. doi:10.1093/nar/gkq873

Collins MD (2004) *Corynebacterium caspium* sp. nov., from a Caspian seal (Phoca caspica). Int J Syst Evol Microbiol 54:925–928. doi:10.1099/ijs.0.02950-0

De Hertogh G, Aerssens J, De Hoogt R et al (2006) Validation of 16S rDNA sequencing in microdissected bowel biopsies from Crohn's disease patients to assess bacterial flora diversity. J Pathol 209:532–539. doi:10.1002/path.2006

Diop A, Khelaifia S, Armstrong N et al (2016) Microbial culturomics unravels the halophilic microbiota repertoire of table salt: description of *Gracilibacillus massiliensis* sp. nov. Microb Ecol Health Dis 27:32049. doi:10.3402/mehd.v27.32049

Drancourt M, Raoult D (2007) Cost-effectiveness of blood agar for isolation of mycobacteria. PLOS Negl Trop Dis 1:e83. doi:10.1371/journal.pntd.0000083

Drancourt M, Carrieri P, Gévaudan MJ, Raoult D (2003) Blood agar and mycobacterium tuberculosis: the end of a dogma. J Clin Microbiol 41:1710–1711. doi:10.1128/JCM.41.4.1710-1711.2003

Finegold SM, Attebery HR, Sutter VL (1974) Effect of diet on human fecal flora: comparison of Japanese and American diets1′2. Am J Clin Nutr 27(12):1456–1469

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fournier PE, Drancourt M (2015) New microbes new infections promotes modern prokaryotic taxonomy: a new section "TaxonoGenomics: new genomes of microorganisms in humans". New Microbes New Infect 7:48–49. doi:10.1016/j.nmni.2015.06.001

Fournier PE, Lagier J-C, Dubourg G, Raoult D (2015) From culturomics to taxonomogenomics: a need to change the taxonomy of prokaryotes in clinical microbiology. Anaerobe 36:73–78. doi:10.1016/j.anaerobe.2015.10.011

Fournier PE, Raoult D, Dancourt M (2016) New species announcements: a new format to prompt the description of new human microbial species. New Microbes New Infect 15:136–137. doi:10.1016/j.nmni.2016.04.006

Fox GE, Wisotzkey JD, Jurtshuk P Jr (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Evol Microbiol 42:166–170. doi:10.1099/00207713-42-1-166

Glaeser SP, Kämpfer P (2015) Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol 38:237–245. doi:10.1016/j.syapm.2015.03.007

Goodman AL, Kallstrom G, Faith JJ et al (2011) Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proc Natl Acad Sci USA 108:6252–6257. doi:10.1073/pnas.1102938108

Goris J, Konstantinidis KT, Klappenbach JA et al (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. doi:10.1099/ijs.0.64483-0

Greub G (2016) Culturomics: a new approach to study the human microbiome. Clin Microbiol Infect 18:1157–1159. doi:10.1111/1469-0691.12032

Guijarro JA, Cascales D, García-Torrico AI et al (2015) Temperature-dependent expression of virulence genes in fish-pathogenic bacteria. Front Microbiol 6:700. doi:10.3389/fmicb.2015.00700

Hiergeist A, Gläsner J, Reischl U, Gessner A (2015) Analyses of intestinal microbiota: culture versus sequencing. ILAR J 56:228–240. doi:10.1093/ilar/ilv017

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765–4774. doi:10.1007/BF00039173

Hugon P, Lagier J-C, Colson P et al (2016) Repertoire of human gut microbes. Microb Pathog. doi:10.1016/j.micpath.2016.06.020

Huson DH, Steel M (2004) Phylogenetic trees based on gene content. Bioinformatics 20:2044–2049. doi:10.1093/bioinformatics/bth198

Jiang X, Doyle MP (2002) Optimizing enrichment culture conditions for detecting helicobacter pylori in foods. J Food Prot 65(12):1949–1954

Johnson JL (1973) Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. Int J Syst Evolutionary Microbiol 23:308–315

Johnson JL (1991) DNA reassociation experiments. In: Stackebrandt E, Goodfellow M (eds) Nucleic acid techniques in bacterial systematics. Wiley, Chichester, pp 21–44

Karlsson F, Tremaroli V, Nielsen J, Bäckhed F (2013) Assessing the human gut microbiota in metabolic diseases. Diabetes 62:3341–3349. doi:10.2337/db13-0844

Kau AL, Ahern PP, Griffin NW et al (2011) Human nutrition, the gut microbiome and the immune system. Nature 474:327–336. doi:10.1038/nature10213

Kim BS, Kim JN, Cerniglia CE (2011) In vitro culture conditions for maintaining a complex population of human gastrointestinal tract microbiota. J Biomed Biotechnol. doi:10.1155/2011/838040

Kim M, Oh H-S, Park S-C, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351. doi:10.1099/ijs.0.059774-0

Kim M, Park SC, Baek I et al (2015) Large-scale evaluation of experimentally determined DNA G+C contents with whole genome sequences of prokaryotes. Syst Appl Microbiol 38:79–83. doi:10.1016/j.syapm.2014.11.008

Lagier JC, Armougom F, Million M et al (2012) Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 18:1185–1193. doi:10.1111/1469-0691.12023

Lagier J-C, Elkarkouri K, Rivet R et al (2013) Non contiguous-finished genome sequence and description of *Senegalemassilia anaerobia* gen. nov., sp. nov. Stand Genom Sci 7:343–356. doi:10.4056/sigs.3246665

Lagier J-C, Hugon P, Khelaifia S et al (2015a) The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. Clin Microbiol Rev 28:237–264. doi:10.1128/CMR.00014-14

Lagier J-C, Edouard S, Pagnier I et al (2015b) Current and past strategies for bacterial culture in clinical microbiology. Clin Microbiol Rev 28:208–236. doi:10.1128/CMR.00110-14

Lagier J-C, Khelaifia S, Alou MT et al (2016) Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol 1:16203. doi:10.1038/nmicrobiol.2016.203

Larsen N, Vogensen FK, van den Berg FWJ et al (2010) Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS ONE 5:e9085. doi:10.1371/journal.pone.0009085

Lau JT, Whelan FJ, Herath I et al (2016) Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. Genome Med 8:1635. doi:10.1186/s13073-016-0327-7

LeChevallier MW, Cameron SC, McFeters GA (1983) New medium for improved recovery of coliform bacteria from drinking water. Appl Environ Microbiol 45:484–492

Lee I, Ouk Kim Y, Chun J, Park S-C (2016) OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103. doi:10.1099/ijsem.0.000760

Liu L, Salam N, Jiao J-Y et al (2016) Diversity of culturable thermophilic actinobacteria in hot springs in tengchong, China and studies of their biosynthetic gene profiles. Microb Ecol 72:150–162. doi:10.1007/s00248-016-0756-2

Löffler FE, Yan J, Ritalahti KM et al (2013) *Dehalococcoides mccartyi* gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, *Dehalococcoidia classis* nov., order *Dehalococcoidales* ord. nov. and family *Dehalococcoidaceae* fam. nov., within the phylum Chloroflexi. Int J Syst Evol Microbiol 63:625–635. doi:10.1099/ijs.0.034926-0

Ludwig W, Klenk H-P (2001) Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. Bergey's manual® of systematic bacteriology. Springer, New York, pp 49–65

Maiden M, Bygraves JA, Feil E et al (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145. doi:10.1073/pnas.95.6.3140

Marchesi JR, Ravel J (2015) The vocabulary of microbiome research: a proposal. Microbiome. doi:10.1186/s40168-015-0094-5

Minnikin DE, Alshamaony L, GOODFELLOW M (1975) Differentiation of mycobacterium, nocardia, and related taxa by thin-layer chromatographic analysis of whole-organism methanolysates. Microbiology 88:200–204. doi:10.1099/00221287-88-1-200

Overmann J (2015) Green sulfur bacteria. Wiley, Chichester

Overmann J, Garcia-Pichel F (2013) The Phototrophic Way of Life. The Prokaryotes. Springer, Heidelberg, pp 203–257

Ramasamy D, Mishra AK, Lagier J-C et al (2014) A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 64:384–391. doi:10.1099/ijs.0.057091-0

Raoult D (2016) Human gut microbiota: repertoire and variations. Front Cell Infect Microbiol. doi:10.3389/fcimb.2012.00136/abstract

Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA 106:19126–19131. doi:10.1073/pnas.0906412106

Rossello-Mora R, Trujillo ME, Sutcliffe IC (2017) Introducing a digital protologue: a timely move towards a database-driven systematic of archaea and bacteria. Antonie Van Leeuwenhoek 110:455–456. doi:10.1007/s10482-017-0841-7

Rossi-Tamisier M, Benamar S, Raoult D, Fournier PE (2015) Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. Int J Syst Evol Microbiol 65:1929–1934. doi:10.1099/ijs.0.000161

Sankar SA, Lagier J-C, Pontarotti P et al (2015) The human gut microbiome, a taxonomic conundrum. Syst Appl Microbiol 38:276–286. doi:10.1016/j.syapm.2015.03.004

Schleifer KH (2009) Classification of bacteria and archaea: past, present and future. Syst Appl Microbiol 32:533–542. doi:10.1016/j.syapm.2009.09.002

Schmeisser C, Stöckigt C, Raasch C et al (2003) Metagenome survey of biofilms in drinking-water networks. Appl Environ Microbiol 69:7298–7309. doi:10.1128/AEM.69.12.7298-7309.2003

Sentausa E, Fournier PE (2013) Advantages and limitations of genomics in prokaryotic taxonomy. Clin Microbiol Infect 19:790–795. doi:10.1111/1469-0691.12181

Sibley CD, Grinwis ME, Field TR et al (2011) Culture enriched molecular profiling of the cystic fibrosis airway microbiome. PLoS ONE 6:e22702. doi:10.1371/journal.pone.0022702

Siggers RH, Siggers J, Boye M et al (2008) Early administration of probiotics alters bacterial colonization and limits diet-induced gut dysfunction and severity of necrotizing enterocolitis in preterm pigs. J Nutr 138:1437–1444

Simner PJ, Doerr KA, Steinmetz LK, Wengenack NL (2016) Mycobacterium and aerobic actinomycete culture: are two medium types and extended incubation times necessary? J Clin Microbiol 54:1089–1093. doi:10.1128/JCM.02838-15

Singh S, Eldin C, Kowalczewska M, Raoult D (2013) Axenic culture of fastidious and intracellular bacteria. Trends Microbiol 21:92–99. doi:10.1016/j.tim.2012.10.007

Skerman VBD, McGowan V, Sneath PHA (eds) (1989) Approved lists of bacterial names (Amended). ASM Press, Washington (DC)

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21:108–110. doi:10.1038/5052

Stackebrandt E, Ebers J (2006) Taxonomic parameters revisited: tarnished gold standards. Microbiol Today 33:152

Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Evol Microbiol 44:846–849. doi:10.1099/00207713-44-4-846

Subramanyam B, Sivaramakrishnan GN, Dusthackeer A et al (2012) Phage lysin as a substitute for antibiotics to detect Mycobacterium tuberculosis from sputum samples with the BACTEC MGIT 960 system. Clin Microbiol Infect 18:497–501. doi:10.1111/j.1469-0691.2011.03601.x

Sutcliffe IC (2015) Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: rip it up and start again. Front Genet 6:218. doi:10.3389/fgene.2015.00218

Tatusov RL, Natale DA, Garkavtsev IV et al (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28. doi:10.1093/nar/29.1.22

Tidjani Alou M, Khelaifia S, Michelle C et al (2016) *Anaerococcus rubiinfantis* sp. nov., isolated from the gut microbiota of a Senegalese infant with severe acute malnutrition. Anaerobe 40:85–94. doi:10.1016/j.anaerobe.2016.06.007

Tindall BJ, Rosselló-Móra R, Busse HJ et al (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 60:249–266. doi:10.1099/ijs.0.016949-0

Turnbaugh PJ, Ley RE, Hamady M et al (2007) The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449:804–810. doi:10.1038/nature06244

Turnbaugh PJ, Quince C, Faith JJ et al (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc Natl Acad Sci USA 107:7503–7508. doi:10.1073/pnas.1002355107

Vandamme P, Coenye T (2004) Taxonomy of the genus cupriavidus: a tale of lost and found. Int J Syst Evol Microbiol 54:2285–2289. doi:10.1099/ijs.0.63247-0

Vandamme P, Peeters C (2014) Time to revisit polyphasic taxonomy. Antonie Van Leeuwenhoek 106:57–65. doi:10.1007/s10482-014-0148x

Vandamme P, Pot B, Gillis M et al (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407–438

Vanlaere E, Baldwin A, Gevers D et al (2009) Taxon K, a complex within the Burkholderia cepacia complex, comprises at least two novel species, *Burkholderia contaminans* sp. nov. and *Burkholderia lata* sp. nov. Int J Syst Evol Microbiol 59:102–111. doi:10.1099/ijs.0.001123-0

Wang Y, Kasper LH (2014) The role of microbiome in central nervous system disorders. Brain Behav Immun 38:1–12. doi:10.1016/j.bbi.2013.12.015

Wayne LG, Brenner DJ, Colwell RR et al (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37:463–464

Zhi X-Y, Zhao W, Li W-J, Zhao G-P (2012) Prokaryotic systematics in the genomics era. Antonie Van Leeuwenhoek 101:21–34. doi:10.1007/s10482-011-9667-x

Description des nouvelles espèces halophiles isolées à partir de la nourriture et du tube digestif humain

Article 6:

Microbial culturomics unravels the halophilic microbiota repertoire of table salt: description of *Gracilibacillus massiliensis* sp. nov.

Diop A, Khelaifia S, Armstrong N, Labas N, Fournier PE, Raoult D, Million M

[Published in Microbial Ecology in Health and Disease]

MICROBIAL ECOLOGY in Health and Disease

co action PUBLISHING

ORIGINAL ARTICLE

Microbial culturomics unravels the halophilic microbiota repertoire of table salt: description of *Gracilibacillus massiliensis* sp. nov.

Awa Diop[1], Saber Khelaifia[1], Nicholas Armstrong[1], Noémie Labas[1], Pierre-Edouard Fournier[1], Didier Raoult[1,2] and Matthieu Million[1]*

[1]Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, AMU UM 63, CNRS UMR7278, IRD 198, INSERM U1095, Institut Hospitalo-Universitaire Méditerranée-Infection, Faculté de médecine, Aix-Marseille Université, Marseille, France; [2]Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

Background: Microbial culturomics represents an ongoing revolution in the characterization of environmental and human microbiome.
Methods: By using three media containing high salt concentration (100, 150, and 200 g/L), the halophilic microbial culturome of a commercial table salt was determined.
Results: Eighteen species belonging to the Terrabacteria group were isolated including eight moderate halophilic and 10 halotolerant bacteria. *Gracilibacillus massiliensis* sp. nov., type strain Awa-1^T (=CSUR P1441=DSM 29726), is a moderately halophilic gram-positive, non-spore-forming rod, and is motile by using a flagellum. Strain Awa-1^T shows catalase activity but no oxidase activity. It is not only an aerobic bacterium but also able to grow in anaerobic and microaerophilic atmospheres. The draft genome of *G. massiliensis* is 4,207,226 bp long, composed of 13 scaffolds with 36.05% of G+C content. It contains 3,908 genes (3,839 protein-coding and 69 RNA genes). At least 1,983 (52%) orthologous proteins were not shared with the closest phylogenetic species. Hundred twenty-six genes (3.3%) were identified as ORFans.
Conclusions: Microbial culturomics can dramatically improve the characterization of the food and environmental microbiota repertoire, deciphering new bacterial species and new genes. Further studies will clarify the geographic specificity and the putative role of these new microbes and their related functional genetic content in environment, health, and disease.

Keywords: Gracilibacillus massiliensis; *taxono-genomics*; *culturomics*; *microbial community*; *salt*; *halophile*

*Correspondence to: Matthieu Million, URMITE, CNRS UMR7278, IRD 198, INSERM U1095, AMU UM63, Faculté de Médecine, Aix-Marseille Université, 27 Boulevard Jean Moulin, FR-13385 Marseille Cedex 5, France, Email: matthieumillion@gmail.com

Received: 26 April 2016; Accepted: 22 September 2016; Published: 18 October 2016

Salt (sodium chloride) is the main mineral constituent of sea water, the oldest and most ubiquitous of food seasonings and an important method of food preservation. Salt was considered hostile to most forms of life; however, it favored the emergence and growth of halophilic bacteria in salty foods (1). Therefore, study on the diversity of hypersaline environmental microorganisms brings important information in the field of environmental microbiology. Recent studies have reported the isolation of new species from salty and/or fermented food (2, 3).

As part of the ongoing microbial culturomics revolution in our laboratory (4), we performed the 'microbial culturome' of a table salt isolating a new moderately halophilic bacterial species belonging to the genus *Gracilibacillus*. First described by Wainø et al. in 1999 (5), the genus *Gracilibacillus* includes, moderately halophilic or halotolerant, mobile, gram-positive bacteria, most of them forming endospores or filaments containing menaquinone-7 (MK-7) as predominant respiratory quinone (6). This genus includes 12 species (www.bacterio.net) described with valid published names (7). Members of the genus *Gracilibacillus* are salty environmental bacteria isolated most often from soil (8), food (9), lakes and salty sea water (10, 11).

To extend the halophilic environmental repertoire, we report here the characterization of a new halophilic species

Citation: Microbial Ecology in Health & Disease 2016, **27**: 32049 - http://dx.doi.org/10.3402/mehd.v27.32049

using the taxono-genomics strategy. Taxono-genomics integrate proteomic information obtained by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) and genomic tests to describe new bacterial species (12, 13). This polyphasic approach overcomes limitations of conventional methods based on genetic, phenotypic, and chemotaxonomic characteristics for new species description (14, 15).

Our new bacterial species *Gracilibacillus* Awa-1^T (=CSUR P1441=DSM 29726, CSUR stands for 'Collection de Souches de l'Unité des Rickettsies' and DSM stands for 'Deutsche Sammlung von Mikroorganismen'), type strain of *Gracilibacillus massiliensis* sp. nov., was isolated from a sample of commercial table salt, a hand-harvested 'fleur de sel', salt from the Camargue natural region. Naturally white, it contains 67% (w/v) NaCl. Fleur de sel is a hand-harvested sea salt collected by workers who scrape only the top layer of salt before it sinks to the bottom of large salt pans. It was harvested in the Saline of Aigues-Mortes in southern France, in a wild, unusual, and unexplored biodiversity habitat. The microbial culturome of this table salt sample and the phenotypic, phylogenetic, and genomic characteristics of the new species isolated in this culturomics approach are reported here.

Materials and methods

Strain isolation

The Camargue sea salt 'Fleur de Sel de Camargue' sample was bought in a supermarket. The sample was transported to our laboratory in the same conditions as at the point of sale, at room temperature. The salinity of the sample was measured using a digital refractometer (Fisher Scientific, Illkirch, France) and its pH was measured using a pH-meter (Eutech Instruments, Strasbourg, France). For the cultivation of halophilic microorganisms, we created media containing high salt concentrations (100, 150, and 200 g/L) (16). *Gracilibacillus* strain Awa-1^T was isolated in September 2014 by cultivation under aerobic conditions, on a homemade halophilic culture medium consisting of a Columbia agar (42 g/L) culture medium (Sigma-Aldrich, Saint-Louis, MO, USA) supplemented by the addition of (per liter) $MgCl_2\,6H_2O$, 10 g; $MgSO_4\,7H_2O$, 10 g; KCl, 4 g; $CaCl_2\,2H_2O$, 1 g; $NaHCO_3$, 0.5 g; glucose, 2 g; 100–150 g/L of NaCl and 5 g of yeast extract (Becton Dickinson, Le-Pont-de-Claix, France). The pH was adjusted to 7.5 with 10 M NaOH before autoclaving at 120°C.

Strain identification by MALDI-TOF MS

MALDI-TOF MS protein analysis was performed using a Microflex spectrometer (Bruker Daltonics, Leipzig, Germany), as previously reported (17). Each separate colony selected was deposited in duplicate on a MALDI-TOF target to be analyzed. A matrix solution of 1.5 µL (saturated solution of α-cyano-4-hydroxycinnamic acid diluted in 50% acetonitrile, 2.5% of trifluoroacetic acid, completed with HPLC water) was deposed on each spot. After reading of the plate, the obtained protein spectra were compared with those of the Bruker database (continuously updated with our recent data) in order to obtain a score, which enables, or not, identification of the strain.

Strain identification by 16S rRNA gene sequencing

The colonies unidentified by the MALDI-TOF after three tests were suspended in 200 µL of distilled water for deoxyribonucleic acid (DNA) extraction by EZ1 DNA Tissue Kit (Qiagen, Courtaboeuf, France). The amplification of the 16S rRNA gene was done by standard polymerase chain reaction (PCR), with the use of universal primers pair FD1 and rp2. The amplified DNA was revealed by electrophoresis on 1.5% agarose gel. Once validated, the PCR product was purified and sequenced using the Big Dye Terminator Sequencing Kit and the following internal primers: 536F, 536R, 800F, 800R, 1050F, 1050R, 357F, and 357R, as previously described (4).

Description of a new species by taxono-genomics

Phylogenetic analysis

We performed a phylogenetic analysis based on 16S rRNA of our isolate to identify its phylogenetic affiliations with other isolates of the genus *Gracilibacillus*. Sequences were aligned using Muscle software (18) and phylogenetic inferences were obtained using the approximately maximum likelihood method within the FastTree software (19). Numbers at the nodes are support local values computed through the Shimodaira–Hasegawa test (20).

Microscopy, sporulation, and motility assays

To observe *G. massiliensis* strain Awa-1^T morphology, transmission electron microscopy was performed after negative staining, using a Tecnai G20 (FEI Company, Limeil-Brevannes, France) at an operating voltage of 60 KV. The gram staining was performed and observed using a photonic microscope Leica DM2500 (Leica Microsystems, Nanterre, France) with a 100X oil-immersion objective. Motility testing was performed by observation of a fresh colony between the blades and slats using DM1000 photonic microscope (Leica Microsystems) at 40x. For the sporulation test, our strain was grown on Chapman agar (Oxoid, Dardilly, France) for 1 week, followed by gram staining and observation for the presence or absence of spores on colonies under the microscope.

Antimicrobial susceptibility and biochemical and atmospheric tests

Sensitivity to antibiotics was determined on a Mueller–Hinton agar in a petri dish (BioMerieux, Marcy-l'Etoile, France). The following antibiotics were tested using Sirscan discs (i2a, Perols, France): doxycycline, rifampicin, vancomycin, amoxicillin, erythromycin, ceftriaxone,

Table 1. Description of the table salt microbiota

	Species	Halophile	Salt concentration in the medium[a]
MALDI-TOF identification			
	Bacillus firmus	Halotolerant	75–150 g/L
	Bacillus licheniformis	Halotolerant	75–150 g/L
	Gracilibacillus dipsosauri	Moderate halophile	75–150 g/L
	Halobacillus trueperi	Moderate halophile	75–150 g/L
	Micrococcus luteus	Halotolerant	75–150 g/L
	Oceanobacillus picturae	Moderate halophile	75–150 g/L
	Planococcus rifietoensis	Halotolerant	75–150 g/L
	Staphylococcus capitis	Halotolerant	75–150 g/L
	Staphylococcus cohnii	Halotolerant	75–150 g/L
	Staphylococcus haemolyticus	Halotolerant	75–150 g/L
	Staphylococcus hominis	Halotolerant	75–150 g/L
	Staphylococcus epidermis	Halotolerant	75–150 g/L
	Staphylococcus warneri	Halotolerant	75–150 g/L
16S identification			
	Alkalibacillus halophilus	Moderate halophile	75–150 g/L
	Paraliobacillus quinghaiensis	Moderate halophile	75–150 g/L
	Thalassobacillus devorans	Moderate halophile	75–150 g/L
	Virgibacillus picturae	Moderate halophile	75–150 g/L
	Gracilibacillus massiliensis sp.nov	Moderate halophile	75–150 g/L

[a]No colonies grew on the medium with 200 g/L of salt.

ciprofloxacin, gentamicin, penicillin, trimethoprim/sulfamethoxazole, imipenem, and metronidazole. Scan 1200 was used to interpret the results (Interscience, Saint Nom la Bretêche, France).

The commercially available API ZYM, API 50CH, and API 20 NE strips (BioMerieux, Marcy-l'Etoile, France) were used for biochemical tests according to the manufacturer's instructions. The time of incubation was 4 h for API ZYM and 48 h for the others.

Growth of the strain Awa-1^T was tested with different growth temperatures (25°C, 30°C, 37°C, 45°C) under aerobic conditions and also in anaerobic and microaerophilic atmospheres, created using AnaeroGen^{TM} (Atmosphere Generation Systems, Dardily, France) and anaerobic jars (Mitsubishi) with GENbag microaer system (BioMerieux), respectively.

Cellular fatty acid analysis

Fatty acid methyl ester (FAME) analysis was performed by Gaz chromatography/mass spectrometry (GC/MS). Two samples were prepared with approximately 40 mg of bacterial biomass, each harvested from several culture plates. FAMEs were prepared as described by Sasser (21). GC/MS analyses were carried out as described before (22). Briefly, FAMEs were separated using an Elite 5-MS column and monitored by mass spectrometry (Clarus 500 – SQ 8 S, Perkin Elmer, Courtaboeuf, France). A spectral database search was performed using MS Search 2.0,

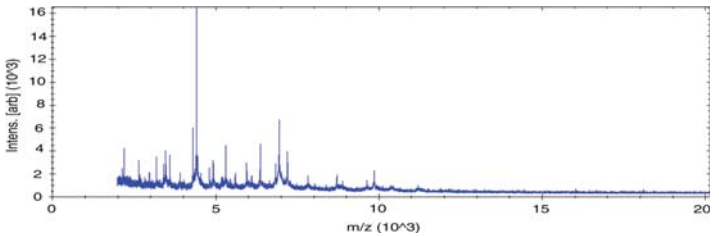


Fig. 1. Reference mass spectrum from *Gracilibacillus massiliensis* strain Awa-1^T spectra.

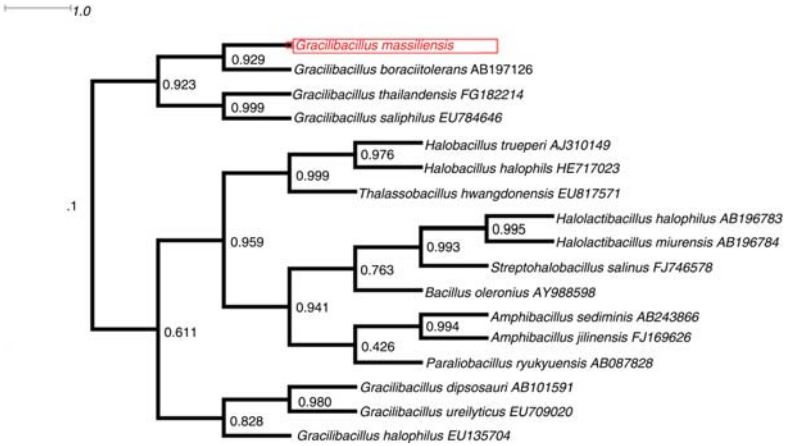


Fig. 2. Phylogenetic tree highlighting the phylogenetic position of *Gracilibacillus massiliensis* strain Awa-1^T relative to other species. GenBank accession numbers are indicated after the name. Sequences were aligned using Muscle software, and phylogenetic inferences were obtained by using the approximately maximum likelihood method within the FastTree software. Numbers at the nodes are support local values computed through the Shimodaira–Hasegawa test.

operated with the Standard Reference Database 1A (NIST, Gaithersburg, MD, USA) and the FAMEs mass spectral database (Wiley, Chichester, UK).

Genomic DNA preparation

After 48 h of growth of the strain Awa-1^T in four petri dishes using our homemade halophilic culture medium, bacteria were resuspended in sterile water and centrifuged at 4°C at 2,000 ×g for 20 min. Cell pellets were resuspended in 1 mL Tris/EDTA/NaCl (10 mM Tris/HCl (pH7.0), 10 mM EDTA (pH8.0), and 300 mM NaCl) and recentrifuged under the same conditions. The pellets were then resuspended in 200 µL Tris-EDTA buffer (TE buffer) and Proteinase K and kept overnight at 37°C for cell lysis. DNA was purified with phenol/chloroform/isoamylalcohol (25:24:1), followed by a precipitation with ethanol at −20°C. The DNA was resuspended in TE buffer and quantified by Qubit fluorometer using the

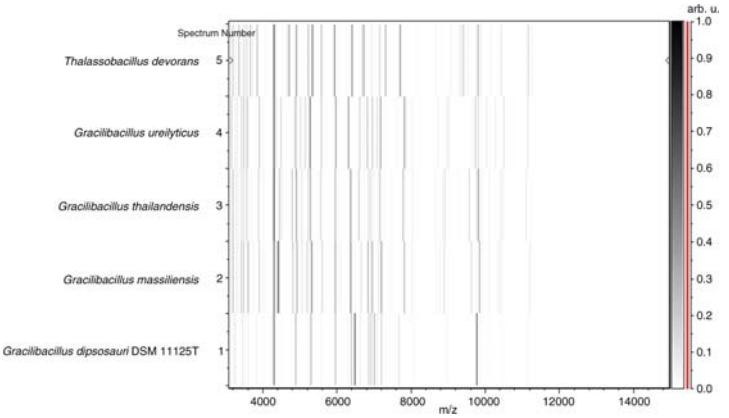


Fig. 3. Gel view comparing *Gracilibacillus massiliensis* strain Awa-1^T to other species within the genera *Gracilibacillus* and *Thalassobacillus.*

Citation: Microbial Ecology in Health & Disease 2016, **27**: 32049 - http://dx.doi.org/10.3402/mehd.v27.32049

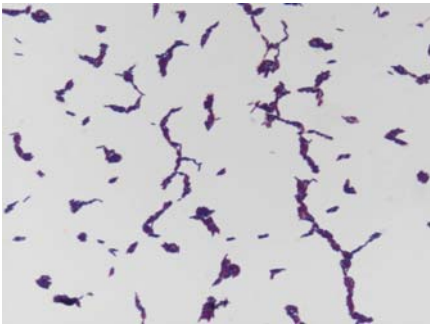


Fig. 4. Gram staining of *Gracilibacillus massiliensis* strain Awa-1^T.

high-sensitivity kit (Life Technologies, Carlsbad, CA, USA) to 112.7 ng/μL.

Genome sequencing and assembly

Genomic DNA (gDNA) of *G. massiliensis* was sequenced on the MiSeq Technology (Illumina Inc, San Diego, CA, USA) with the mate pair strategy. The gDNA was barcoded in order to be mixed with 11 other projects with the Nextera Mate Pair sample prep kit (Illumina). The mate pair library was prepared with 1.5 μg of gDNA using the Nextera mate pair Illumina guide. The gDNA sample was simultaneously fragmented and tagged with a mate pair junction adapter. The pattern of the fragmentation was validated on an Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA

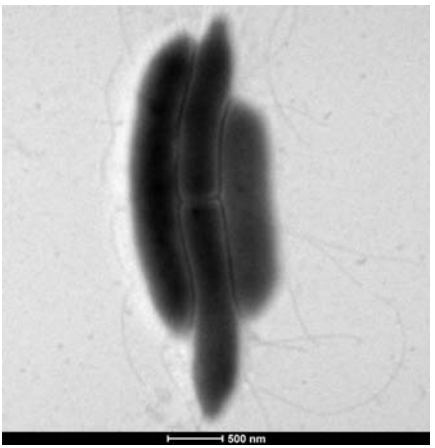


Fig. 5. Transmission electron microscopy of *Gracilibacillus massiliensis* strain Awa-1^T.

7500 labchip. The DNA fragments ranged in size from 1.5 up to 11 kb with an optimal size at 6.641 kb. No size selection was performed and 600 ng of tagmented fragments were circularized. The circularized DNA was mechanically sheared to small fragments with an optimal at 1,309 bp on the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The library profile was visualized on a high-sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) and the final concentration library was measured at 47.82 nmol/L. The libraries were normalized at 4 nM and pooled. After a denaturation step and dilution, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. The automated cluster generation

Table 2. Classification and general features of *Gracilibacillus massiliensis* strain Awa-1^T according to the MIGS recommendations (23)

MIGS ID	Property classification	Term	Evidence code[a]
		Domain: Bacteria	TAS (36)
		Phylum: Firmicutes	TAS (37)
		Class: Bacilli	TAS (36)
		Order: Bacillales	TAS (36)
		Family: Bacillaceae	TAS (36)
		Genus: *Gracilibacillus*	TAS (5)
		Species: *Gracilibacillus massiliensis*	IDA
		Type strain: Awa-1^T	IDA
	Gram strain	Positive	IDA
	Cell shape	Rods	IDA
	Motility	Motile	IDA
	Sporulation	No sporulating	IDA
	Temperature (°C)	Mesophile (25–45)	IDA
	Optimum temperature	37°C	IDA
	pH range: optimum	6.0–9.0: 7.0–8.0	IDA
	Carbon source	Unknown	IDA
MIGS-6	Habitat	Salt environment	IDA
MIGS-6.3	NaCl range: optimum	75–150:75 g/L	IDA
MIGS-22	Oxygen requirement	Aerobic	IDA
MIGS-15	Biotic relationship	Free-living	IDA
MIGS-14	Pathogenicity	Unknown	IDA

[a]Evidence codes – IDA, inferred from direct assay; TAS, traceable author statement (i.e. a direct report exists in the literature). These evidence codes are from the Gene Ontology project (38).

Table 3. Differential characteristics of *Gracilibacillus massiliensis* compared to other close bacteria of the genus *Gracilibacillus*

Properties	*G. massiliensis*	*G. thailandensis*	*G. saliphilus*	*G. orientalis*	*G. ureilyticus*	*G. halophilus*	*G. boraciitolerans*	*G. kekensis*	*G. halotolerans*	*G. alcaliphilus*
Cell diameter (μm)	0.3–1.8	0.3–0.4	0.7–0.9	0.7–0.9	0.7–1	0.3–0.5	0.5–0.9	0.2–1.05	0.4–0.6	0.5–0.7
Pigmentation	White	White	Creamy white	Creamy	Creamy	White	Dirty white	Creamy white	Creamy white	Creamy white
Oxygen requirement	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic
Gram stain	+	+	+	+	+	+	+	+	+	+
Salt requirement	+	+	+	+	+	+	+	+	+	+
Motility	+	+	+	+	+	+	+	+	+	+
Sporulation	−	+	+	+	+	+	+	+	+	+
Indole	−	−	−	−	−	−	−	−	−	−
Production of										
Alkaline phosphate	−	−	+	NA	+	+	+	NA	+	−
Catalase	+	+	+	+	+	+	+	NA	+	+
Oxidase	−	+	+	−	+	+	+	−	+	−
Nitrate reductase	−	+	+	−	+	+	−	−	+	+
Urease	+	−	+	−	+	−	−	−	+	+
Arginine dihydrolase	NA	−	−	−	+	−	−	NA	−	−
β-galactosidase	−	NA	+	NA	+	+	+	NA	+	NA
α-galactosidase	+	NA	−	NA	+	−	+	NA	NA	−
N-acetyl-glucosamine	−	NA	+	NA	NA	−	NA	NA	NA	+
Acid from										
L-Arabinose	−	+	+	+	+	−	+	+	+	+
Ribose	−	+	+	NA	NA	+	+	+	+	+
D-mannose	−	+	+	−	+	−	+	+	−	−
D-mannitol	−	+	+	+	+	+	+	+	+	+
D-sucrose	NA	+	+	+	+	+	NA	+	−	+
D-glucose	−	+	+	+	+	+	+	+	+	+
D-fructose	−	+	+	+	NA	+	+	+	+	+
D-maltose	−	+	+	+	+	−	+	+	−	+
D-lactose	−	−	+	+	+	−	+	+	−	+
DNA G+C content (mol%)	36.05	37.6	40.1	37.1	35.3	42.3	35.8	35.8	38	41.3
Habitat	Cooking salt	Fermented fish	Salt lake	Salt lake	Saline-alkaline soil	Salt soil	Soil	Salt lake	Saline soil	Fermentation liquor for dyeing

G. massiliensis Awa-1^T; *G. thailandensis* TP2-8^T(9); *G. orientalis* XH-63^T(39); *G. ureilyticus* MF38^T (6); *G. halophilus* YIM-C55.5^T(8); *G. boraciitolerans* T-16X^T(40); *G. saliphilus* YIM91119^T(41); *G. kekensis* K170^T(11); *G. halotolerans* NN^T(5); *G. alcaliphilus* SG103^T(7). NA = not available.

Table 4. Total cellular fatty acid composition of *Gracilibacillus massiliensis* strain Awa-1^T

Fatty acids	IUPAC name	Mean relative (%)[a]
15:0 anteiso	12-methyl-tetradecanoic acid	45.6±0.3
15:0 iso	13-methyl-tetradecanoic acid	21.2±0.3
17:0 anteiso	14-methyl-hexadecanoic acid	7.9±0.2
16:0	Hexadecanoic acid	5.7±0.1
15:0	Pentadecanoic acid	5.4±0.1
16:0 iso	14-methyl-pentadecanoic acid	3.4±0.02
14:0 iso	12-methyl-tridecanoic acid	3.0±0.2
16:1n9	7-hexadecenoic acid	2.5±0.2
14:0	Tetradecanoic acid	1.4±0.1
16:1n6 iso	14-methylpentadec-9-enoic acid	1.2±0.1
5:0 anteiso	2-methyl-butanoic acid	TR
16:1n7	9-hexadecenoic acid	TR
17:1n7 anteiso	14-methylhexadec-9-enoic acid	TR
17:0 iso	15-methyl-hexadecanoic acid	TR
17:0	Heptadecanoic acid	TR
18:0	Octadecanoic acid	TR

[a]Mean peak area percentage calculated from the analysis of FAMEs in two sample preparations ± standard deviation ($n=3$); TR = trace amounts <1%.

and sequencing run were performed in a single 2 × 251-bp run.

Total information of 7.9 Gb was obtained from an 816 K/mm^2 cluster density with cluster passing quality control filters of 91.7% (15,550,000 passing filter paired reads). Within this run, the index representation for *G. massiliensis* was determined to be 5.41%. The 841,255 paired reads were trimmed then assembled to 13 scaffolds.

Genome annotation and comparison

Prodigal was used for open reading frames (ORFs) prediction (23) with default parameters. Predicted ORFs spanning a sequencing gap region (containing N) were excluded. Bacterial protein sequences were predicted using BLASTP (*E*-value 1e^{-03}, coverage 0.7 and identity percent 30%) against the clusters of orthologous groups (COG) database. If no hit was found, a search against the non redundant (NR) database (24) was performed using BLASTP with *E*-value of 1e^{-03} coverage 0.7 and an identity percent of 30%. If sequence lengths were smaller than 80 amino acids, we used an *E*-value of 1e^{-05}. PFAM-conserved domains (PFAM-A and PFAM-B domains) were searched on each protein with the hhmscan tools analysis. RNAmmer (25) was used to find ribosomal RNAs genes, whereas tRNA genes were found using the tRNAScanSE tool (26). We predicted the lipoprotein signal peptides and the number of transmembrane helices

Table 5. Nucleotide content and gene count levels of the genome

Attribute	Value	% of total[a]
Size (bp)	4,207,226	100
G+C content (bp)	1,516,759	36.05
Coding region (bp)	3,579,496	85.07
Total genes	3,908	100
RNA genes	69	1.76
Protein-coding genes	3,839	98.23
Genes with function prediction	2,647	68.95
Genes assigned to COGs	2,455	63.94
Genes with peptide signals	430	11.20
Genes with transmembrane helices	1,063	27.68

[a]The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

using Phobius (27). ORFans were identified if all the BLASTP performed had negative results (*E*-value smaller than 1e^{-03} for ORFs with sequence size greater than 80 aa or *E*-value smaller than 1e^{-05} for ORFs with sequence length smaller than 80 aa). Artemis (28) and DNA Plotter (29) were used for data management and for visualization of genomic features, respectively. We used the MAGI homemade software to estimate the mean level of nucleotide sequence similarity at the genome level. It calculated the average genomic identity of gene sequences (AGIOS) among compared genomes (30). This software combines the Proteinortho software (31) for detecting orthologous proteins in pairwise genomic comparisons, then retrieves the corresponding genes and determines the mean percentage of nucleotide sequence identity among orthologous ORFs using the Needleman–Wunsch global alignment algorithm. Genomes from the genus *Gracilibacillus* and closely related genera were used for the calculation of AGIOS values. The genome of *G. massiliensis* strain Awa-1^T (EMBL-EBI accession number CZRP00000000) was compared with that of *Halobacillus halophilus* type strain DSM 2266 (HE717023), *Amphibacillus jilinensis* strain Y1 (AMWI00000000), *Halobacillus trueperi* strain HT-01 (CCDJ000000000), *Gracilibacillus halophilus* strain YIM-C55.5 (APML00000000), and *Gracilibacillus boraciitolerans* strain JCM 21714 (BAVS00000000). Annotation and comparison processes were performed in the Multi-Agent software system DAGOBAH (32), which include Figenix (33) libraries that provide pipeline analysis. We also performed genome-to-genome distance calculator (GGDC) analysis using the GGDC web server as previously reported (34).

Accession numbers

The 16S rRNA and genome sequences are deposited in EMBL-EBI under accession numbers LN626645 and CZRP00000000, respectively.

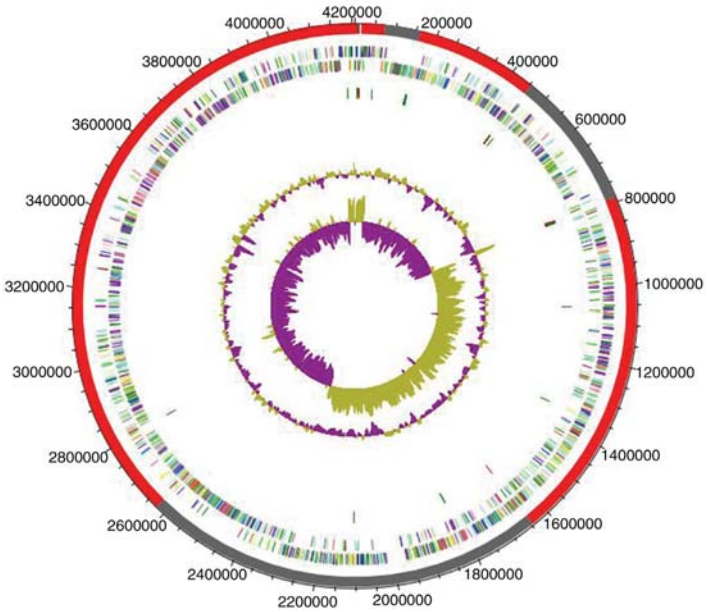


Fig. 6. Graphical circular map of the chromosome. From outside to the center: Genes on the forward strand colored by clusters of orthologous groups of proteins (COG) categories (only genes assigned to COG), genes on the reverse strand colored by COG categories (only gene assigned to COG), RNA genes (tRNAs green, rRNAs red), GC content, and GC skew.

Results

Description of the table salt microbiota community

The cultivable halophilic and halotolerant bacterial consortia isolated from the fleur de sel of Camargue included 18 bacterial species (Table 1) from 4,303 colonies. MALDI-TOF MS identified 13 species, whereas 16S rRNA gene sequencing identified five other species including a new species (*G. massiliensis* sp. nov.). Among the four culture conditions used, only three conditions yielded colonies. All colonies were isolated from media with a concentration of 75 g/L (standard Chapman medium), 100 g/L and 150 g/L NaCl (house-made media). Conversely, in the culture medium containing 200 g/L NaCl, no bacterial colonies were isolated. Among the 18 cultured species, 10 were halotolerant and 8 were halophilic species (Table 1).

Identification and phylogenetic analysis of the new species

MALDI-TOF score obtained for strain Awa-1^T against our database (Bruker database constantly incremented with new data) suggests that our isolate was not a member of a known species. We added the spectrum from strain Awa-1^T to our database (Fig. 1).

PCR-based identification of the 16S rRNA of *G. massiliensis* (EMBL-EBI accession number LN626645) yielded 96.9% 16S rRNA gene sequence similarity with the reference *Gracilibacillus thailandensis* (GenBank accession number NR116568), the phylogenetically closest validated *Gracilibacillus* species (Fig. 2). This value was lower than the 98.7% 16S rRNA gene sequence threshold advised by Meier-Kolthoff et al. (35) to delineate a new species without carrying out DNA–DNA hybridization. The gel view demonstrated the spectral differences with other members of the genus *Gracilibacillus* (Fig. 3).

Physiological and biochemical characteristics

G. massiliensis is a gram-positive (Fig. 4) thin, long rod, with a mean diameter of 0.3 µm and a length of 1.8 µm measured through electron microscopy (Fig. 5). This strain is non-spore-forming, peritrichous, and motile. It grew under aerobic conditions but was also able to grow in anaerobic (at 29°C) and microaerophilic (at 29°C −37°C) atmospheres. The colonies are convex, creamy white, circular, and measured 0.2–0.3 mm in diameter after 2–4 days of growth in our homemade culture

Citation: Microbial Ecology in Health & Disease 2016, **27**: 32049 - http://dx.doi.org/10.3402/mehd.v27.32049

medium. Classification and general features are summarized in Table 2.

The strain was catalase test positive and oxidase negative. Using API ZYM, API 20NE, and API 50CH identification strips, positive reactions were observed for esterase, lipase, α-galactosidase, β-glucuronidase, β-glucosidase, N-acetyl-β-glucosaminidase, urease, and 4-nitrophenyl-βD-galactopyranoside. Acid was not produced from D-glucose, D-mannitol, D-saccharose, D-maltose, D-lactose, L-arabinose, glycerol, D-mannose, D-fructose or D-ribose. Esculin was hydrolyzed, but nitrate was not reduced and indole was negative. Phenotypic characteristics were compared to those of other members of the genus *Gracilibacillus* (Table 3). Antimicrobial susceptibility tests demonstrated that the isolate was susceptible to doxycycline, rifampicin, vancomycin, erythromycin, ciprofloxacin, gentamicin, trimethoprim/sulfamethoxazole, and imipenem, but resistant to metronidazole, amoxicillin, ceftriaxone, and penicillin G.

Analysis of the total cellular fatty acid composition of *G. massiliensis* demonstrated that the fatty acids detected are mainly saturated. The most abundant species (15:0 anteiso, 15:0 iso, and 17:0 anteiso) are branched fatty acids. A few unsaturated fatty acids were detected at low abundances (Table 4).

Genome properties

The draft genome of *G. massiliensis* strain Awa-1^T is 4,207,226 bp long with 36.05% G+C content (Table 5 and Fig. 6). It is composed of 13 scaffolds with 13 contigs. Of the 3,908 predicted genes, 3,839 were protein-coding genes, and 69 were RNAs (7 genes are 5S rRNA, 1 gene is 16S rRNA, 1 gene is 23S rRNA, and 60 genes are tRNA genes). A total of 2,647 genes (68.95%) were assigned as putative functions (by COGs or by NR blast). A total of 126 genes (3.28%) were identified as ORFans. The remaining genes were annotated as hypothetical proteins (875 genes = 22.79%). Genome statistics are summarized in Table 5 and the distribution of the genes into COGs functional categories is presented in Table 6.

Table 6. Number of genes associated with the 25 general COG functional categories

Code	Value	% value	Description
J	206	5.36	Translation
A	0	0	RNA processing and modification
K	205	5.33	Transcription
L	90	2.34	Replication, recombination, and repair
B	1	0.026	Chromatin structure and dynamics
D	51	1.32	Cell cycle control, mitosis, and meiosis
Y	0	0	Nuclear structure
V	65	1.69	Defense mechanisms
T	140	3.64	Signal transduction mechanisms
M	125	3.25	Cell wall/membrane biogenesis
N	53	1.38	Cell motility
Z	0	0	Cytoskeleton
W	9	0.23	Extracellular structures
U	32	0.83	Intracellular trafficking and secretion
O	105	2.73	Posttranslational modification, protein turnover, and chaperones
X	46	1.19	Mobilome: prophages and transposons
C	138	3.59	Energy production and conversion
G	328	8.54	Carbohydrate transport and metabolism
E	208	5.41	Amino acid transport and metabolism
F	87	2.26	Nucleotide transport and metabolism
H	148	3.85	Coenzyme transport and metabolism
I	97	2.52	Lipid transport and metabolism
P	144	3.75	Inorganic ion transport and metabolism
Q	70	1.82	Secondary metabolites biosynthesis, transport, and catabolism
R	244	6.35	General function prediction only
S	191	4.97	Function unknown
–	1,384	36.05	Not in COGs

Genome comparison

The G+C content of *G. massiliensis* strain Awa-1^T (36.05%) is smaller than that of *H. trueperi, H. halophilus, A. jilinensis,* and *G. halophilus* (41.66, 41.82, 37.27, and 37.92%, respectively) but larger than that of *G. boraciitolerans* (35.83%). The gene content of *G. massiliensis* (3,839) is smaller than that of *H. trueperi, H. halophilus,* and *G. boraciitolerans* (4,000, 4,135, and 4,450, respectively) but larger than that of *A. jilinensis* and *G. halophilus* (3,594 and 2,968, respectively). However, the distribution of genes into COG categories was similar among all compared genomes (Fig. 7). In addition, *G. massiliensis* shared 1,856 orthologous genes with the most closely related species (*G. halophilus*): 1,780, 1,614, 1,781, and 1,611 orthologous genes with *H. halophilus, A. jilinensis, H. trueperi,* and *G. boraciitolerans*, respectively (Table 7). The average percentage of nucleotide sequence identity ranged from 72.17 to 78.29% at the intraspecies level between *G. massiliensis* and the two *Gracilibacillus* species, but it ranged from 52.49 to 68.02% at interspecies level between *G. massiliensis* and other species. Similar results were obtained for the analysis of the digital DNA–DNA hybridization (dDDH) using GGDC software (Table 8).

The Awa-1^T strain, moderate halophilic bacterium, was isolated from a sample of cooking salt (Sel de Camargue) when studying salt-tolerant bacteria in salty food in the context of the culturomics project. On the basis of the phenotypic characteristics, phylogenetic and genomic analysis, Awa-1^T strain is proposed to represent a novel species named *G. massiliensis* sp. nov.

Description of Gracilibacillus massiliensis sp. nov.

G. massiliensis (mas.si.li.en'sis. L. adj. *massiliensis* relating to Massilia, the ancient Roman name of Marseille, France, where the type strain was isolated and characterized, like

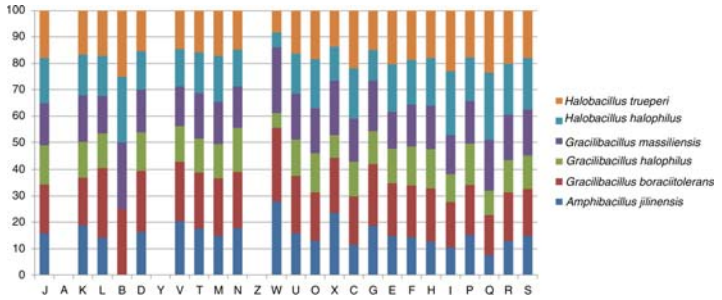


Fig. 7. Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of *Gracilibacillus massiliensis* strain Awa-1^T among other species.

many other species). This bacterium is motile through the use of its peritrichous flagella. It is a moderately halophilic, gram-positive, non-spore-forming rod, with a mean diameter of 0.3 µm and a length of 1.8 µm. The colonies are convex, creamy white, circular and measuring 0.2–0.3 mm in diameter after 2–4 days of growth on our home-made culture medium. Strain Awa-1^T is not only aerobic but also able to grow in anaerobic (at 29°C) and microaerophilic (at 29–37°C) atmospheres. Its optimal conditions for growth are 37°C at pH 7.0–8.0 with 75 g/L of NaCl.

Using API identification strips, catalase, urease, esterase, lipase, α-galactosidase, β-glucuronidase, β-glucosidase, N-acetyl-β-glucosaminidase, and 4-nitrophenyl-βD-galactopyranoside activities are found positive. Oxidase, nitrate reductase, and indole tests are negative. The isolate is susceptible to doxycyclin, rifampicin, vancomycin, erythromycin, ciprofloxacin, gentamicin, trimethoprim/sulfamethoxazole, and imipenem, but resistant to metronidazole, amoxicillin, ceftriaxone, and penicillin G.

The G+C% content of the genome is 36.05%. The 16S rRNA and genome sequences are deposited in EMBL-EBI under accession numbers LN626645 and CZR P00000000, respectively. The type strain of *G. massiliensis* is strain Awa-1^T (= CSUR P1441 = DSM 29726) and was isolated from Salt specimen (Salt of Camargue).

Discussion

Because of the concept of 'microbial culturomics', which is based on the variation of physicochemical parameters of the culture conditions to explore microbial diversity (4), many new bacterial species have been discovered. As mentioned in our seminal work (4), microbial culturomics provides culture conditions simulating, reproducing, or mimicking the entirety of selective constraints that have shaped natural microbiota for millions of years. Here, the use of hypersaline conditions led to the comprehensive description of the hitherto unknown halophilic repertoire of table salt including a new *Gracilibacillus* species. All correspond to the Terrabacteria taxonomic group, evidencing the terrestrial adaptation of such microbes with very high resistance to desiccation by salt. The members of *Gracilibacillus* genus are all gram-positive bacteria, aerobic, motile and peritrichous, moderately halophile, white, and endospore-forming at the terminal position in general. Our strain Awa-1^T does not form spores, the first differentiating characteristic compared to other species. It was selected for sequencing based on its phenotypic differences, phylogenetic position, and 16S rRNA sequence similarity with other members of the genus *Gracilibacillus.* The G+C content of the genomic DNA varies from 35.3 to 42.3 mol% (7). According to the fact that the G+C content deviation within species is at most

Table 7. Numbers of orthologous proteins shared between genomes (upper right) and AGIOS values obtained (lower left)

	GM	HH	AJ	HT	GH	GB
GM	**3,839**	1,780	1,614	1,781	1,856	1,611
HH	52.49%	**4,135**	1,446	1,813	1,551	1,316
AJ	68.02%	52.84%	**3,594**	1,448	1,430	1,193
HT	66.14%	53.12%	65.43%	**4,000**	1,560	1,316
GH	72.17%	52.66%	67.75%	65.98%	**2,968**	1,403
GB	78.29%	52.63%	67.13%	65.30%	70.63%	**4,450**

The numbers of proteins per genome are indicated in bold. GM, *Gracilibacillus massiliensis* Awa-1^T; HH, *Halobacillus halophilus* DSM 2266; AJ, *Amphibacillus jilinensis* Y1; HT, *Halobacillus trueperi* HT-01; GH, *Gracilibacillus halophilus* YIM-C55.5^T; GB, *Gracilibacillus boraciitolerans* JCM 21714.

Citation: Microbial Ecology in Health & Disease 2016, **27**: 32049 - http://dx.doi.org/10.3402/mehd.v27.32049

Table 8. dDDH values obtained by comparison of all studied genomes

	HH	AJ	HT	GH	GB
GM	24.4% ± 0.17	20.7% ± 0.21	27.0% ± 0.16	19.0% ± 0.23	22.2% ± 0.19
HH		21.9% ± 0.20	21.6% ± 0.20	26.2% ± 0.16	22.7% ± 0.19
AJ			24.2% ± 0.18	18.6% ± 0.23	24.6% ± 0.17
HT				33.2% ± 0.12	28.7% ± 0.14
GH					17.4% ± 0.25

dDDH, digital DNA-DNA hybridization. GM, *Gracilibacillus massiliensis* Awa-1^T; HH, *Halobacillus halophilus* DSM 2266; AJ, *Amphibacillus jilinensis* Y1; HT, *Halobacillus trueperi* HT-01; GH, *Gracilibacillus halophilus* YIM-C55.5^T; GB, *Gracilibacillus boraciitolerans* JCM 21714.

1%, these values confirm the classification of strain Awa-1^T in a distinct species (42). Furthermore, the values of the AGIOS and dDDH of *G. massiliensis* compared to all other known species confirm its new species status. Microbial culturomics significantly extend the halophilic repertoire of salty food and/or salt table. This will improve the understanding of the possible involvement of table salt microbiota in human health and disease, with significant contributions to food and environmental microbiology.

Authors' contributions

AD performed the bacterium phenotypic characterization and the genomic analyses and drafted the manuscript. SK participated in its design and helped draft the manuscript. NA performed the cellular fatty acids analysis and helped draft the manuscript. NL performed the genomic sequencing and helped draft the manuscript. PEF and DR conceived the study and helped draft the manuscript. MM conceived the study, participated in its design and coordination, and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Xegen Company (www.xegen.fr) for automating the genomic annotation process. They also thank Karolina Griffiths for English reviewing and Claudia Andrieu for administrative assistance.

Conflict of interest and funding

The authors declare that they have no competing interests. This work was supported by the 'Fondation Méditerranée Infection'.

References

1. Cantrell SA, Dianese JC, Fell J, Gunde-Cimerman N, Zalar P. Unusual fungal niches. Mycologia 2011; 103: 1161–74.
2. Hong SW, Kwon SW, Kim SJ, Kim SY, Kim JJ, Lee JS, et al. *Bacillus oryzaecorticis* sp. nov., a moderately halophilic bacterium isolated from rice husks. Int J Syst Evol Microbiol 2014; 64: 2786–91.
3. Lo N, Lee SH, Jin HM, Jung JY, Schumann P, Jeon CO. *Garicola koreensis* gen. nov., sp. nov., isolated from saeu-jeot, traditional Korean fermented shrimp. Int J Syst Evol Microbiol 2015; 65: 1015–21.
4. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, et al. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 2012; 18: 1185–93.
5. Wainø M, Tindall BJ, Schumann P, Ingvorsen K. *Gracilibacillus* gen. nov., with description of *Gracilibacillus halotolerans* gen. nov., sp. nov.; transfer of *Bacillus dipsosauri* to *Gracilibacillus dipsosauri* comb. nov., and *Bacillus salexigens* to the genus *Salibacillus* gen. nov., as *Salibacillus salexigens* comb. nov. Int J Syst Bacteriol 1999; 49: 821–31.
6. Huo YY, Xu XW, Cui HL, Wu M. *Gracilibacillus ureilyticus* sp. nov., a halotolerant bacterium from a saline–alkaline soil. Int J Syst Evol Microbiol 2010; 60: 1383–6.
7. Hirota K, Hanaoka Y, Nodasaka Y, Yumoto I. *Gracilibacillus alcaliphilus* sp nov, a facultative alkaliphile isolated from indigo fermentation liquor for dyeing. Int J Syst Evol Microbiol 2014; 64: 3174–80.
8. Chen YG, Cui XL, Zhang YQ, Li WJ, Wang YX, Xu LH, et al. *Gracilibacillus halophilus* sp. nov., a moderately halophilic bacterium isolated from saline soil. Int J Syst Evol Microbiol 2008; 58: 2403–8.
9. Chamroensaksri N, Tanasupawat S, Akaracharanya A, Visessanguan W, Kudo T, Itoh T. *Gracilibacillus thailandensis* sp. nov., from fermented fish (pla-ra). Int J Syst Evol Microbiol 2010; 60: 944–8.
10. Jeon CO, Lim JM, Jang HH, Park DJ, Xu LH, Jiang CL, et al. *Gracilibacillus lacisalsi* sp. nov., a halophilic Gram-positive bacterium from a salt lake in China. Int J Syst Evol Microbiol 2008; 58: 2282–6.
11. Gao M, Liu ZZ, Zhou YG, Liu HC, Ma YC, Wang L, et al. *Gracilibacillus kekensis* sp. nov., a moderate halophile isolated from Keke Salt Lake. Int J Syst Evol Microbiol 2012; 62: 1032–6.
12. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2012; 40: D571–9.
13. Sentausa E, Fournier PE. Advantages and limitations of genomics in prokaryotic taxonomy. Clin Microbiol Infect 2013; 19: 790–5.
14. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 1996; 60: 407–38.
15. Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. Microbiol Today 2006; 33: 152–5.
16. Lagier JC, Hugon P, Khelaifia S, Fournier PE, La Scola B, Raoult D. The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. Clin Microbiol Rev 2015; 28: 237–64.

17. Seng P, Drancourt M, Gouriet F, La Scola B, Fournier PE, Rolain JM, et al. Ongoing revolution in bacteriology: routine identification of bacteria by matrix assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 2009; 49: 543–51.
18. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32: 1792–7.
19. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 2010; 5: e9490.
20. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 1999; 16: 1114–6.
21. Sasser, M. Bacterial identification by gas chromatographic analysis of fatty acids methyl esters (GC-FAME). Technical Note 101. Newark, DE: MIDI Inc; 2006.
22. Dione N, Sankar SA, Lagier JC, Khelaifia S, Michele C, Armstrong N, et al. Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. New Microbes New Infect 2016; 10: 66–76.
23. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010; 11: 119.
24. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2015; 43: D30–5.
25. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007; 35: 3100–8.
26. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997; 25: 955–64.
27. Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004; 338: 1027–36.
28. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics 2000; 16: 944–5.
29. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 2009; 25: 119–20.
30. Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 2014; 64: 384–91.
31. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC Bioinformatics 2011; 12: 124.
32. Gouret P, Paganini J, Dainat J, Louati D, Darbo E, Pontarotti P, et al. Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the multi-agent software system DAGOBAH. In: Pontarotti P, ed. Evolutionary biology – concepts, biodiversity, macroevolution and genome evolution. Berlin: Springer-Verlag; 2011, pp. 71–87.
33. Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG. FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinformatics 2005; 6: 198.
34. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 2013; 14: 60.
35. Meier-Kolthoff JP, Göker M, Spröer C, Klenk HP. When should a DDH experiment be mandatory in microbial taxonomy? Arch Microbiol 2013; 195: 413–8.
36. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. Proc Natl Acad Sci USA 1990; 87: 4576–9.
37. Murray RGE. The higher taxa, or, a place for everything? In: Holt JG, ed. Bergey's manual of systematic bacteriology. 1st ed. Vol. 1. Baltimore, MD: The Williams and Wilkins; 1984, pp. 31–4.
38. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology The Gene Ontology Consortium. Nat Genet 2000; 25: 25–9.
39. Carrasco IJ, Márquez MC, Yanfen X, Ma Y, Cowan DA, Jones BE, et al. *Gracilibacillus orientalis* sp. nov., a novel moderately halophilic bacterium isolated from a salt lake in Inner Mongolia, China. Int J Syst Evol Microbiol 2006; 56: 599–604.
40. Ahmed I, Yokota A, Fujiwara T. *Gracilibacillus boraciitolerans* sp. nov., a highly boron-tolerant and moderately halotolerant bacterium isolated from soil. Int J Syst Evol Microbiol 2007; 57: 796–802.
41. Tang SK, Wang Y, Lou K, Mao PH, Jin X, Jiang CL, et al. *Gracilibacillus saliphilus* sp. nov., a moderately halophilic bacterium isolated from a salt lake. Int J Syst Evol Microbiol 2009; 59: 1620–4.
42. Meier-Kolthoff JP, Klenk HP, Göker M. Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. Int J Syst Evol Microbiol 2014; 64: 352–6.

Article 7:

Genome sequence and description of *Gracilibacillus timonensis* sp. nov. strain Marseille-P2481^{T}, a moderate halophilic bacterium isolated from the human gut microflora

Diop A, Seck EH, Dubourg G, Armstrong N, Michelle C, Raoult D, Fournier PE

[Published in MicrobiologyOpen journal]

Received: 22 November 2017 | Revised: 31 January 2018 | Accepted: 9 March 2018

DOI: 10.1002/mbo3.638

ORIGINAL ARTICLE

WILEY MicrobiologyOpen Open Access

Genome sequence and description of *Gracilibacillus timonensis* sp. nov. strain Marseille-P2481^T, a moderate halophilic bacterium isolated from the human gut microflora

Awa Diop[1] | **El hadji Seck**[1] | **Gregory Dubourg**[1] | **Nicholas Armstrong**[1] | **Caroline Blanc-Tailleur**[1] | **Didier Raoult**[1,2] | **Pierre-Edouard Fournier**[1]

[1]URMITE, UM63, CNRS 7278, IRD 198, Inserm U1095, Aix-Marseille Université, Institut hospitalo-universitaire Mediterranee-infection, Marseille, France

[2]Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

Correspondence
Pierre-Edouard Fournier,
URMITE, UM63, CNRS 7278, IRD 198, Inserm U1095, Aix-Marseille Université, Institut hospitalo-universitaire Mediterranee-infection, Marseille, France.
Email: pierre-edouard.fournier@univ-amu.fr

Funding information
Mediterranee-Infection foundation; French Agence Nationale de la Recherche

Abstract

Microbial culturomics represents an ongoing revolution in the characterization of the human gut microbiota. By using three culture media containing high salt concentrations (10, 15, and 20% [w/v] NaCl), we attempted an exhaustive exploration of the halophilic microbial diversity of the human gut and isolated strain Marseille-P2481 (= CSUR P2481 = DSM 103076), a new moderately halophilic bacterium. This bacterium is a Gram-positive, strictly aerobic, spore-forming rod that is motile by use of a flagellum and exhibits catalase, but not oxidase activity. Strain Marseille-P2481 was cultivated in media containing up to 20% (w/v) NaCl, with optimal growth being obtained at 37°C, pH 7.0–8.0, and 7.5% [w/v] NaCl). The major fatty acids were 12-methyl-tetradecanoic acid and hexadecanoic acid. Its draft genome is 4,548,390 bp long, composed of 11 scaffolds, with a G+C content of 39.8%. It contains 4,335 predicted genes (4,266 protein coding including 89 pseudogenes and 69 RNA genes). Strain Marseille-P2481 showed 96.57% 16S rRNA sequence similarity with *Gracilibacillus alcaliphilus* strain SG103^T, the phylogenetically closest species with standing in nomenclature. On the basis of its specific features, strain Marseille-P2481^T was classified as type strain of a new species within the genus *Gracilibacillus* for which the name *Gracilibacillus timonensis* sp. nov. is formally proposed.

KEYWORDS

Gracilibacillus timonensis, halophilic, human gut flora, microbial culturomics, taxonogenomics

1 | INTRODUCTION

One of the most important methods of food preservation in history has been the use of salt (NaCl). Salt has also become an indispensable ingredient of any kitchen. Considered previously as hostile to most forms of life by limiting the growth of certain bacteria, it was demonstrated to favor the emergence and growth of others, mainly halophilic bacteria (Cantrell, Dianese, Fell, Gunde-Cimerman, & Zalar, 2011). Several recent studies have reported the isolation of new halophilic species from the human gut microflora (Khelaifia et al., 2016; Lagier, Khelaifia, et al., 2015). Therefore, exploring the diversity of halophilic microorganisms in the human gut flora may provide important insights into our understanding of their presence, interactions with the human digestive environment, and their influence on health.

In order to explore the human gut halophilic microbiota, and as part of the ongoing microbial culturomics study in our laboratory

(Lagier et al., 2012, 2016), we used high salt-containing culture media, which enabled us to isolate a new moderately halophilic bacterial strain, Marseille-P2481, that belongs to the genus *Gracilibacillus* (Senghor et al., 2017). First proposed by Wainø et al. in 1999 (Wainø, Tindall, Schumann, & Ingvorsen, 1999), the genus *Gracilibacillus* currently includes 13 species (http://www.bacterio.net/gracilibacillus.html) with validly published names (Parte, 2014). These are Gram stain-positive, aerobic, moderately halophilic or halotolerant, motile bacteria. In most species, cells are motile due to peritrichous flagella and form endospores and white colonies (Wainø et al., 1999). *Gracilibacillus* species were isolated from diverse salty environmental samples, including sea water, salty lakes (Gao et al., 2012; Jeon et al., 2008), soil (Chen et al., 2008; Huo, Xu, Cui, & Wu, 2010), and/or food (Chamroensaksri et al., 2010; Diop et al., 2016).

Using the taxonogenomics approach that includes phenotypic features, proteomic information obtained by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), and analysis of the complete genome sequence (Pagani et al., 2012; Ramasamy et al., 2014; Sentausa & Fournier, 2013), we present here the characterization of a new halophilic species for which we formally propose the name *Gracilibacillus timonensis* sp. nov. Strain Marseille-P2481^T (= CSUR P2481 = DSM 103076) is the type strain of *Gracilibacillus timonensis* sp. nov.

2 | MATERIALS AND METHODS

2.1 | Sample collection and culture conditions

A stool sample was collected from a 10-year-old healthy young Senegalese boy living in N'diop (a rural village in the Guinean-Sudanian zone of Senegal). The patient's parents gave an informed consent, and the study was approved by the National Ethics Committee of Senegal (N° 00.87 MSP/DS/CNERS) and by the local ethics committee of the IFR48 (Marseille, France) under agreement 09-022. The stool sample was collected immediately after defecation into a sterile plastic container, preserved at −80°C and transported to Marseille until further analysis.

The salinity of the sample was measured using a digital refractometer (Fisher scientific, Illkirch, France) and its pH measured using a pH-meter (Eutech Instruments, Strasbourg, France).

Strain Marseille-P2481 was isolated in aerobic conditions, on a home-made culture medium consisting of Columbia agar enriched with 10% (w/v) NaCl (Sigma-Aldrich, Saint-Louis, MO, USA), as previously described (Diop et al., 2016). Briefly, 1 g of stool sample was inoculated into 100 ml of our home-made liquid medium and incubated aerobically at 37°C. Subcultures were conducted after 1, 3, 7, 10, 15, 20, and 30 days of incubation. Serial dilutions of 10^{-1} to 10^{-10} were then performed in the home-made liquid culture medium and plated on Columbia and Chapman agar plates (Oxoid, Dardilly, France). After 2 days of incubation at 37°C, all apparent colonies were picked and subcultured several times to obtain pure cultures.

2.2 | MALDI-TOF MS strain identification

Briefly, one isolated bacterial colony was picked from chapman culture plate using a pipette tip and spread it as a thin film on a MTP 96 MALDI-TOF target plate for identification with a Microflex MALDI-TOF MS spectrometer (Bruker Daltonics, Leipzig, Germany). In total, 12 distinct deposits for strain Marseille-P2481were done from 12 individual colonies in duplicate. After air-drying, 2-µl matrix solution was applied per spot, as previously reported (Lagier, Khelaifia, et al., 2015). All spectra were recorded in positive linear mode for the mass range of 2,000–20,000 Da (parameter settings: ion source 1 (ISI), 20 kV; IS2, 18.5 kV; lens, 7 kV). The obtained protein spectra were compared with those of 2,480 spectra in the Bruker database enriched with our own database (Lagier, Hugon, et al., 2015). The strain was identified at the species level if the MALDI-TOF MS score was greater than 1.9. If the score was lower than this threshold, the identification was not considered as reliable and the 16S rRNA gene was sequenced.

2.3 | 16S rRNA gene sequencing identification

The 16S rRNA gene was amplified using the broad-range primer pair FD1 and rp2 (Drancourt et al., 2000). The primers were obtained from Eurogentec (Seraing, France). The obtained amplicon was sequenced using the Big Dye Terminator Sequencing kit and the following internal primers: 536f, 536r, 800f, 800r, 1050f, 1050r, 357f, 357r, as previously described (Drancourt, Bollet, & Raoult, 1997; Drancourt et al., 2000). The sequence was then compared with the NCBI database using the BLASTn algorithm (https://blast.ncbi.nlm.nih.gov/). If the 16S rRNA gene sequence similarity value was greater than 95% and lower 98.65% with the most closely related species with standing in nomenclature, as previously proposed (Kim, Oh, Park, & Chun, 2014; Stackebrandt & Ebers, 2006), the strain was proposed to belong to a new species (Konstantinidis, Ramette, & Tiedje, 2006).

2.4 | Phylogenetic analysis

The 16S sequences from the type strains of the species with a validly published name that exhibited the highest BLAST score with our new strain were downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/Genome/). Sequences were aligned using the CLUSTALW 2.0 software (Larkin et al., 2007), and phylogenetic inferences were obtained using the neighbor-joining method and the maximum likelihood method within the MEGA software, version 6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013). The evolutionary distances were computed based on the Kimura 2-parameter model (Kimura, 1980) with 95% of deletion, and bootstraping analysis was performed with 500 replications.

2.5 | Morphological observation

To observe the cell morphology, transmission electron microscopy of the strain was performed using a Tecnai G20 Cryo (FEI

company, Limeil-Brevannes, France) at an operating voltage of 60 Kv after negative staining. Gram staining was performed and observed using a photonic microscope Leica DM2500 (Leica Microsystems, Nanterre, France) with a 100X oil-immersion objective (Atlas & Snyder, 2011). The motility of the strain was assessed by the Hanging Drop method. The slide was examined using a DM1000 photonic microscope (Leica Microsystems) at 40×. Sporulation was tested following a thermic shock at 80°C during 20 min, and the endospore formation was visualized using a Tecnai G20 Cryo transmission electron microscope (FEI company, Limeil-Brevannes, France) at an operating voltage of 60 Kv after negative staining.

2.6 | Atmospheric tests, biochemical, and antimicrobial susceptibility

In order to evaluate the optimal culture conditions, strain Marseille-P2481 was cultivated on Chapman agar at different temperatures (25, 28, 37, 45 and 56°C) under aerobic conditions, and in anaerobic and microaerophilic atmospheres using GENbag Anaer and GENbag microaer systems (bioMérieux), respectively. The pH (pH 5, 6, 6.5, 7, and 8.5) and salinity (5–20% [w/v] NaCl) conditions were also tested.

Biochemical tests were performed using the API ZYM, API 50 CH, and API 20 NE strips (bioMerieux, Marcy-l'Etoile, France), according to the manufacturer's instructions. The API ZYM was incubated for 4 hr and the other two strips for 48 hr.

The antibiotic susceptibility of strain Marseille-P2481 was determined using the disk diffusion method as previously described (Diop et al., 2016). The following antibiotics were tested: penicillin G (10 µg), amoxicillin (25 µg), ceftriaxone (30 µg), imipenem (10 µg), rifampicin (30 µg), erythromycin (15 µg), gentamicin (500 µg), and metronidazole (4 µg). The results were interpreted using the Scan 1,200 automate (Interscience, Saint Nom la Bretêche, France).

2.7 | Fatty acid methyl ester (FAME) analysis by GC/MS

For the FAME analysis, strain Marseille-P2481 was cultivated on Chapman agar (7.5% NaCl) (Oxoid, Dardilly, France) at 37°C under aerobic atmosphere for 2 days. Cellular fatty acid methyl ester (FAME) analysis was performed by gas chromatography/mass spectrometry (GC/MS). Two samples were prepared with approximately 70 mg of bacterial biomass per tube harvested from several culture plates. FAMEs were prepared as described by Sasser (Sasser, 1990). GC/MS analyses were carried out as previously described (Dione et al., 2016). Briefly, FAMEs were separated using an Elite 5-MS column and monitored by mass spectrometry (Clarus 500 - SQ 8 S, Perkin Elmer, Courtaboeuf, France). Spectral database search was performed using the MS Search 2.0 software operated with the Standard Reference Database 1A (NIST, Gaithersburg, USA) and the FAMEs mass spectral database (Wiley, Chichester, UK).

2.8 | Extraction and genome sequencing

After a pretreatment by lysozyme incubation at 37°C for 2 hr, the DNA of strain Marseille-P2481 was extracted on the EZ1 biorobot (Qiagen) with EZ1 DNA Tissue kit. The elution volume was 50 µl. The gDNA was quantified by a Qubit assay with the high sensitivity kit (Life Technologies, Carlsbad, CA, USA) to 185 ng/µl.

A MiSeq sequencer and the mate-pair strategy (Illumina Inc, San Diego, CA, USA) were used to sequence the gDNA. The gDNA was barcoded in order to be mixed with 11 other projects with the Nextera Mate-Pair sample prep kit (Illumina). The mate-pair library was prepared with 1.5 µg of gDNA using the Nextera mate-pair guide. The genomic DNA sample was simultaneously fragmented and tagged with a mate-pair junction adapter. The pattern of the fragmentation was validated on an Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA 7500 labchip. The DNA fragments ranged in size from 1.5 to 11 kb with an optimal size at 5.314 kb. No size selection was performed and 600 ng of tagmented fragments was circularized. The circularized DNA was mechanically sheared with an optimal size at 939 bp on the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The library profile was visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA), and the final concentration library was measured at 8.38 nmol/L. The libraries were normalized at 2 nmol/L and pooled. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded. Automated cluster generation and sequencing run were performed in a single 39-hr run in a 2 × 251 bp.

A total sequencing output of 6.52 Gb was obtained from a 696 K/mm^2 cluster density with a cluster passing quality control filters of 95.6% (12,863,388 passing filter paired reads). Within this run, the index representation for strain Marseille-P2481 was determined to be 9.39%. The 1,207,306 paired reads were trimmed and then assembled.

2.9 | Genome annotation and comparison

Prodigal was used for open reading frame (ORF) prediction (Hyatt et al., 2010) with default parameters. Predicted ORFs spanning a sequencing gap region were excluded. Bacterial protein sequences were predicted using BLASTP (E-value 1e^{-03}, coverage 0.7 and identity percent 30%) against the Clusters of Orthologous Groups (COG) database. If no hit was found, a search against the nr database (Benson et al., 2015) was performed using BLASTP with E-value of 1e^{-03}, a coverage of 0.7 and an identity percent of 30%. If sequence lengths were smaller than 80 amino acids, we used an E-value of 1e^{-05}. Pfam conserved domains (PFAM-A an PFAM-B domains) were searched on each protein with the HHMscan tool (Finn et al., 2015). RNAmmer (Lagesen et al., 2007) and tRNAScanSE (Lowe & Eddy, 1997) were used to identify ribosomal RNAs and tRNAs, respectively. We predicted lipoprotein signal peptides and the number of transmembrane helices using Phobius (Käll, Krogh, & Sonnhammer, 2004). ORFans were identified if the BLASTP search was negative (E-value smaller than 1e^{-03} for ORFs with a

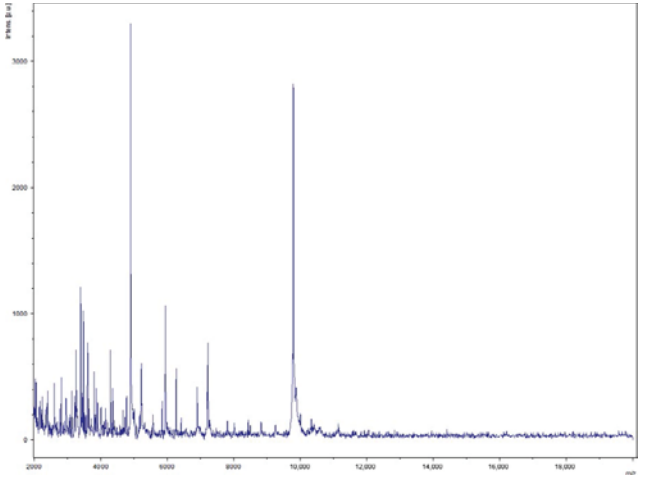


FIGURE 1 Reference mass spectrum from *Gracilibacillus timonensis* strain Marseille-P2481^T

sequence size larger than 80 aas or E-value smaller than 1e^{-05} for ORFs with sequence length smaller than 80 aas). Artemis (Carver, Harris, Berriman, Parkhill, & McQuillan, 2012) and DNA Plotter (Carver, Thomson, Bleasby, Berriman, & Parkhill, 2009) were used for data management and for visualization of genomic features, respectively. Genomes from members of the genus *Gracilibacillus* and closely related genera were used for the calculation of AGIOS values. The genome of strain Marseille-P2481 (EMBL-EBI accession number FLKH00000000) was compared with those of *Gracilibacillus halophilus* strain YIM-C55.5^T (APML00000000), *G. boraciitolerans* strain JCM 21714^T (BAVS00000000), *G. lacisalsi* strain DSM 19029 ^{T} (ARIY00000000), *G. massiliensis* strain Awa-1^T (CZRP00000000), *G. kekensis* strain K170 ^{T} (FRCZ01000001), *G. orientalis* strain XH-63 ^{T} (FOTR01000001), *G. ureilyticus* strain MF38 ^{T} (FOGL01000001), *B. clausii* strain KSM-K16^T (AP006627), and *B. alcalophilus* strain ATCC 27647^T (ALPT00000000). Annotation and comparison processes were performed using the multi-agent software system DAGOBAH (Gouret et al., 2011), which includes Figenix (Gouret et al., 2005) libraries that provide pipeline analysis. We also estimated the degrees of genomic sequence similarity among compared genomes using the following tools: first, we used the MAGI home-made software (Padmanabhan, Mishra, Raoult, & Fournier, 2013) This software calculates the average genomic identity of orthologous gene sequences (AGIOS) among compared genomes (Ramasamy et al., 2014). It combines the Proteinortho software (Lechner et al., 2011) for detecting orthologous proteins in pairwise genomic comparisons, then retrieves the corresponding genes and determines the mean percentage of nucleotide sequence identity among orthologous ORFs using the Needleman-Wunsch global alignment algorithm. Second,

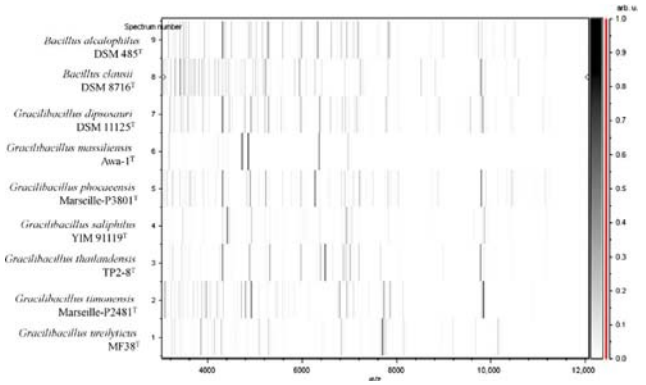


FIGURE 2 Gel view comparing *Gracilibacillus timonensis* strain Marseille-P2481^T with other species within the genera *Gracilibacillus* and *Bacillus*

the digital DNA–DNA hybridization was performed using the GGDC (Genome-to-Genome Distance Calculator) analysis via the GGDC web server as previously reported (Klenk, Meier-Kolthoff, & Göker, 2014). Finally, the average amino acid identity (AAI) was calculated, based on the overall similarity between two genomic datasets of proteins (Rodriguez-R & Konstantinidis, 2014) available at (http://enve-omics.ce.gatech.edu/aai/index).

3 | RESULTS

3.1 | Strain identification and phylogenetic analysis

A MALDI-TOF MS score of 1.4 was obtained for strain Marseille-P2481 against our database (Bruker database), suggesting that our isolate was not in the database. The MALDI-TOF MS spectrum from strain Marseille-P2481 (Figure 1) was added to our database and a gel view showed the spectral differences between our isolate and other closely related species (Figure 2). The 16S rDNA-based identification of strain Marseille-P2481 (EMBL-EBI accession number LT223702) yielded a 96.57% 16S rRNA gene sequence identity with *Gracilibacillus alcaliphilus* strain SG103^T (GenBank accession number NR_126185), the phylogenetically closest species with a validly published name (Figure 3). As this value was lower than the 98.65% 16S rRNA sequence identity threshold recommended to define a new species without carrying out DNA–DNA hybridization (Kim et al., 2014), strain Marseille-P2481 was considered as representative of a potential new species within the *Gracilibacillus* genus.

3.2 | Physiological and biochemical characteristics

Isolated for the first time in our home-made halophilic medium with 10% (w/v) NaCl, strain Marseille-P2481 was able to grow in media containing up to 20% (w/v) NaCl under aerobic conditions with a minimal concentration of growth at 7.5% NaCl, but was also able to grow in anaerobic and microaerophilic atmospheres (at 37°C). After 2 days of growth at 37°C, colonies were creamy orange and circular, with a mean diameter of 0.2 µm. Cells were Gram stain-positive (Figure 4a), endospore-forming (Figure 4b), and motile rods with a peritrichous flagellum. Cells were also slightly curved, with mean diameter and length of 0.5 and 1.9 µm, respectively (Figure 4b). Strain Marseille-P2481 exhibited positive catalase but no oxidase activity. General features and classification of *Gracilibacillus timonensis* strain Marseille-P2481^T are summarized in Table 1. Using an API ZYM strip, positive results were obtained for esterase, esterase lipase, acid phosphatase, naphtol-AS-BI-phosphohydrolase β-galactosidase, β-glucosidase, and α-glucosidase activities but no reaction was observed for alkaline phosphatase, lipase, Leucine arylamidase, Valine arylamidase, Cystine arylamidase, α-galactosidase, β-glucuronidase, trypsin, α-chymotrypsin, α-mannosidase, α-fucosidase, and N-acetyl-β-glucosaminidase. The API 50CH strip revealed that strain Marseille-P2481 exhibited esculin hydrolysis, but negative reactions were obtained for D-arabitol, L-arabitol, D-glucose, D-fructose, D-fucose, D-galactose, D-lactose, D-maltose, D-ribose, D-saccharose, D-lyxose, D-mannose L-sorbose, D-tagatose,

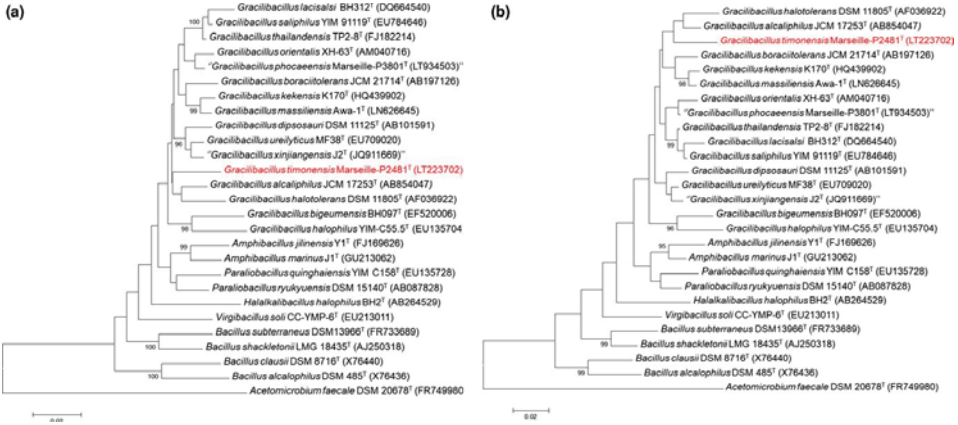


FIGURE 3 Phylogenetic tree highlighting the position of *Gracilibacillus timonensis* strain Marseille-P2481^T relative to other closely related species. GenBank accession numbers of each 16S rRNA are indicated after each species name. Sequences were aligned using CLUSTALW, and the evolutionary history was inferred using the Neighbor-Joining method (a) and the maximum likelihood method (b) with the Kimura 2-parameter method within MEGA6 software. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) is shown next to the branches. The analysis involved 24 nucleotide sequences. All positions with less than 95% site coverage were eliminated. There were a total of 1,404 positions in the final dataset. The scale bar represents a 2% nucleotide sequence divergence

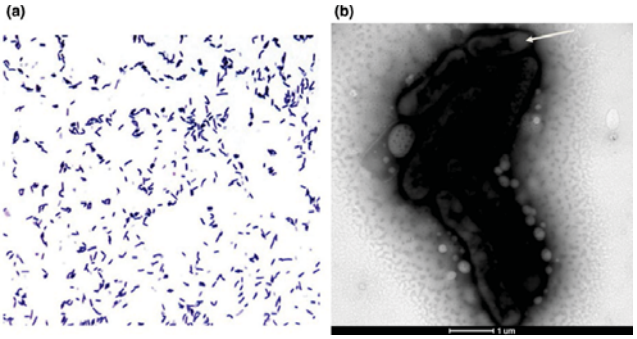


FIGURE 4 Bacterial morphology. (a) Gram staining of *Gracilibacillus timonensis* strain Marseille-P2481^T. (b) Transmission electron microscopy showing an endospore in terminal position (arrow). The scale bar represents 1 µm

TABLE 1 Classification and general features of *Gracilibacillus timonensis* strain Marseille-P2481^T according to the MIGS recommendations [23]

MIGS ID	Property	Term	Evidence code[a]
		Domain: *Bacteria*	TAS (Woese, Kandler, & Wheelis, 1990)
		Phylum: *Firmicutes*	TAS (Skerman & Sneath 1980, Murray, 1984, Gibbons and Murray, 1978, Garrity and Holt, 2001)
		Class: *Bacilli*	TAS (Ludwig, Schleifer, & Whitman 2009)
		Order: *Bacillales*	TAS (Skerman & Sneath 1980, Prevot, 1953)
		Family: *Bacillaceae*	TAS (Skerman & Sneath 1980, Fischer, 1985)
		Genus: *Gracilibacillus*	TAS (Wainø et al., 1999)
		Species: *Gracilibacillus timonensis*	IDA
		Type strain: Marseille-P2481^T	IDA
	Gram stain	Positive	IDA
	Cell shape	Rods	IDA
	Motility	Motile	IDA
	Sporulation	Spore-forming	IDA
	Temperature (°C)	Mesophile (25-45)	IDA
	Optimum temperature	37°C	IDA
	pH range: Optimal pH	6.0–9.0 7.0–8.0	IDA
	Carbon source	Unknown	IDA
MIGS-6	Habitat	Human gut	IDA
MIGS-6.3	NaCl range: Optimum NaCl	7.5–20% 7.5%	IDA
MIGS-22	Oxygen requirement	Aerobic	IDA
MIGS-15	Biotic relationship	Free living	IDA
MIGS-14	Pathogenicity	Unknown	IDA

[a]Evidence codes: IDA, Inferred from Direct Assay; TAS, Traceable Author Statement (i.e., a direct report exists in the literature). These evidence codes are from the Gene Ontology project (Ashburner et al. 2000).

TABLE 2 Differential characteristics of *Gracilibacillus timonensis* strain Marseille-P2481^T and other closely related members of the genus *Gracilibacillus*

Properties	*G. timonensis*	*G. saliphilus*	*G. bigeumensis*	*G. halophilus*	*G. boraciitolerans*	*G. kekensis*	*G. halotolerans*	*G. alcaliphilus*
Cell diameter (μm)	0.5–0.8	0.7–0. 9	0.3–0.5	0.3–0.5	0.5–0. 9	0.2–1.05	0.4–0.6	0.5–0.7
Pigmentation	Creamy orange	Creamy white	Creamy	White	Dirty white	Creamy white	Creamy white	Creamy white
Oxygen requirement	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic
Gram stain	+	+	+	+	+	+	+	+
Salt requirement	+	+	+	+	+	+	+	+
Motility	+	+	+	+	+	+	+	+
Sporulation	+	+	+	+	+	+	+	+
Indole	−	−	−	−	−	−	−	−
Production of								
Alkaline phosphate	−	+	+	+	+	NA	+	−
Catalase	+	+	+	+	+	NA	+	+
Oxidase	−	+	+	+	+	−	+	−
Nitrate reductase	−	+	−	+	−	−	+	+
Urease	+	+	−	−	−	−	+	+
β-galactosidase	+	+	+	+	+	NA	−	NA
α-galactosidase	−	−	−	−	+	NA	+	−
N-acetyl-glucosamine	−	+	−	−	NA	NA	NA	+
L-arabinose	−	+	+	−	+	+	+	+
Ribose	−	+	−	+	+	+	+	+
D-mannose	−	+	+	−	+	+	−	−
D-mannitol	−	+	+	+	+	+	+	+
D-glucose	+	+	+	+	+	+	+	+
D-fructose	−	+	+	+	+	+	+	+
D-maltose	−	+	+	−	+	+	−	+
D-lactose	−	+	+	−	+	+	−	+
DNA G+C content (mol %)	39.8	40.1	37.9	42.3	35.8	35.8	38	41.3
Habitat	Human gut	Salt lake	Solar saltern soil	Salty soil	Soil	Salty lake	Saline soil	Fermentation liquor for dyeing

NA, no data available.

G. timonensis strain Marseille-P2481^T, *G. Gracilibacillus bigeumensis* strain BH097^T (Kim et al., 2012), *G. halophilus* strain YIM-C55.5^T (Chen et al., 2008), *G. boraciitolerans* strain T-16X^T (Ahmed et al., 2007), *G. saliphilus* strain YIM91119^T (Tang et al., 2009), *G. kekensis* strain K170^T (Gao et al., 2012), *G. halotolerans* strain NN^T (Wainø et al., 1999), *G. alcaliphilus* strain SG103^T (Hirota, Hanaoka, Nodasaka, & Yumoto, 2014).

TABLE 3 Total cellular fatty acid composition of *Gracilibacillus timonensis* strain Marseille-P2481^T

Fatty acids	IUPAC name	Mean relative %[a]
15:0 anteiso	12-methyl-tetradecanoic acid	45.4 ± 1.5
16:0	Hexadecanoic acid	15.6 ± 1.1
17:0 anteiso	14-methyl-Hexadecanoic acid	13.9 ± 0.6
15:0 iso	13-methyl-tetradecanoic acid	10.3 ± 0.6
17:0 iso	15-methyl-Hexadecanoic acid	5.8 ± 1.0
16:0 iso	13-methyl-Pentadecanoic acid	3.4 ± 0.4
18:0	Octadecanoic acid	1.2 ± 0.1
15:0	Pentadecanoic acid	1.1 ± 0.2
14:0 iso	12-methyl-Tridecanoic acid	1.1 ± 0.1
17:0	Heptadecanoic acid	1.1 ± 0.1
14:0	Tetradecanoic acid	TR
10:0	Decanoic acid	TR
12:0	Dodecanoic acid	TR
13:0 anteiso	10-methyl-Dodecanoic acid	TR
13:0 iso	11-methyl-Dodecanoic acid	TR

[a]Mean peak area percentage calculated from the analysis of FAMEs in 2 sample preparations ± standard deviation (n = 3); TR= trace amounts < 1%.

TABLE 4 Nucleotide content and gene count of the genome

Attribute	Value	% of total[a]
Size (bp)	4,548,390	100%
G+C content (bp)	1,808,751	39.8%
Coding region (bp)	3,844,022	85.07%
Total genes	4,395	100%
RNA genes	63	1.76%
Protein-coding genes	4,332	98.23%
Genes with function prediction	3,043	68.95%
Genes assigned to COGs	2,797	63.94%
Genes with peptide signals	474	11.20%
Genes with transmembrane helices	1,191	27.68%

[a]The total is based on either the size of the genome in base pairs or the total number of protein-coding genes in the annotated genome.

TABLE 5 Number of genes associated with the 25 general COG functional categories

Code	Value	% value	Description
J	212	4.89	Translation
A	0	0	RNA processing and modification
K	266	6.14	Transcription
L	103	2.37	Replication, recombination, and repair
B	1	0.02	Chromatin structure and dynamics
D	52	1.20	Cell cycle control, mitosis, and meiosis
Y	0	0	Nuclear structure
V	98	2.26	Defense mechanisms
T	154	3.46	Signal transduction mechanisms
M	147	3.39	Cell wall/membrane biogenesis
N	49	1.13	Cell motility
Z	0	0	Cytoskeleton
W	3	0.06	Extracellular structures
U	30	0.69	Intracellular trafficking and secretion
O	107	2.46	Posttranslational modification, protein turnover, chaperones
X	57	1.31	Mobilome: prophages, transposons
C	113	2.60	Energy production and conversion
G	478	11.03	Carbohydrate transport and metabolism
E	201	4.63	Amino acid transport and metabolism
F	100	2.30	Nucleotide transport and metabolism
H	138	3.18	Coenzyme transport and metabolism
I	94	2.16	Lipid transport and metabolism
P	192	4.43	Inorganic ion transport and metabolism
Q	66	1.52	Secondary metabolites biosynthesis, transport, and catabolism
R	288	6.64	General function prediction only
S	212	4.89	Function unknown
-	1,535	35.43	Not in COGs

D-turanose, D-xylose, L-xylose, D-arabinose, L-arabinose, D-sorbitol, D-cellobiose, D-melezitose, D-melibiose, D-trehalose, D-raffinose, L-rhamnose, D-adonitol, D-mannitol, L-fucose, amygdalin, arbutin, erythritol, dulcitol, gentiobiose, glycerol, glycogen, inositol, inulin, salicin, starch, xylitol, αD-glucopyranoside, methyl-βD-xylopyranoside, methyl-αD-mannopyranoside, potassium gluconate, potassium-2-ketogluconate potassium-5-ketogluconate, N-acetylglucosamine. Using an API 20NE strip, fermentation of glucose, urease activity, and metabolism of L-arginine, esculin and 4-nitrophenyl-βD-galactopyrasinoside were positive. In contrast, nitrate and indole production, gelatinase activity and metabolism of D-glucose, L-arabinose, D-mannose, D-maltose, D-mannitol, N-acetyl-glucosamine, potassium gluconate, capric acid, malic acid, trisodium citrate, and phenylacetic acid were negative. Strain Marseille-P2481 differed

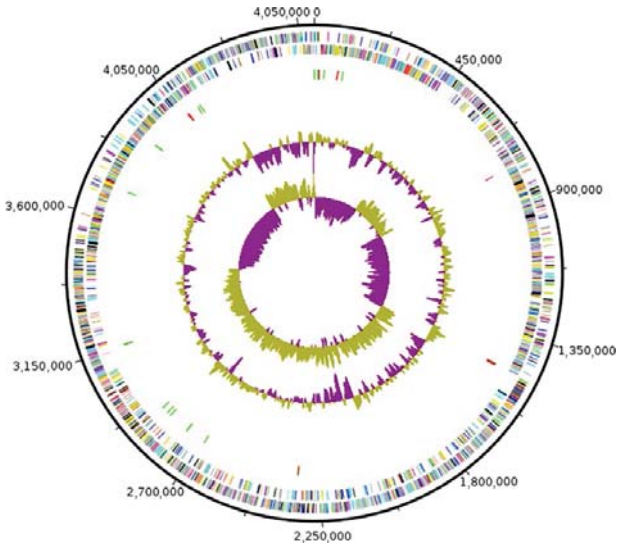


FIGURE 5 Graphical circular map of the chromosome. From the outside to the center: Genes on the forward strand colored by Clusters of Orthologous Groups of proteins (COG) categories (only genes assigned to COG), genes on the reverse strand colored by COG categories (only gene assigned to COG), RNA genes (tRNAs green, rRNAs red), GC content, and GC skew

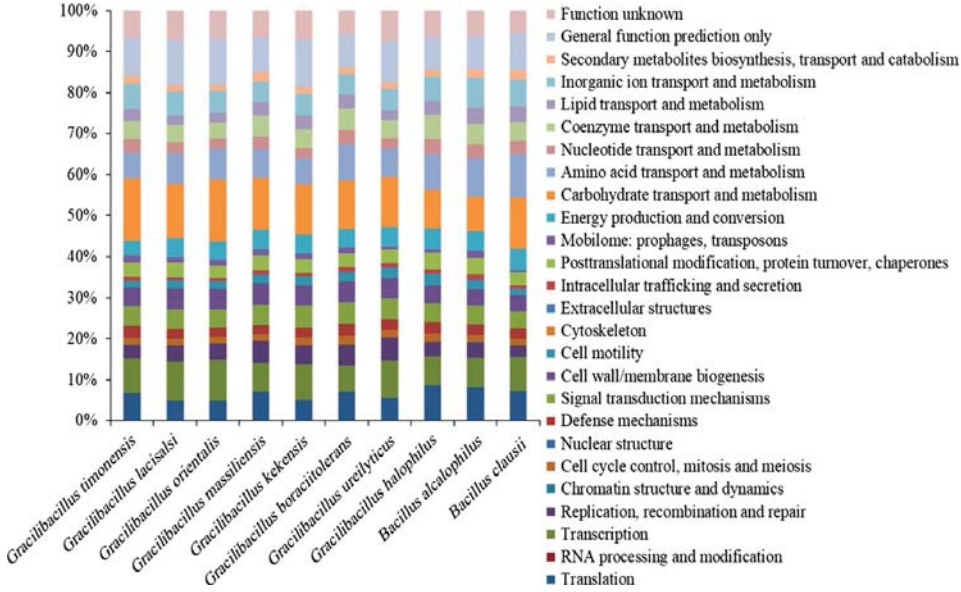


FIGURE 6 Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of *Gracilibacillus timonensis* strain Marseille-P2481^T and other compared species

TABLE 6 Numbers of orthologous proteins (upper right) and AGIOS values (lower left, %) obtained between compared genomes. The numbers of proteins per genome are indicated in bold

	GT	GL	GO	GM	GK	GB	GU	GH	BA	BC
GT	**4,333**	2,103	2,112	2,004	2,027	1,461	1,982	1,695	1,539	1,578
GL	72.3	**4,268**	2,654	2,405	2,467	1,693	2,374	1,995	1,654	1,703
GO	72.1	85.2	**4,313**	2,370	2,412	1,686	2,318	1,940	1,656	1,710
GM	72.0	77.0	77.0	**3,839**	2,559	1,724	2,346	1,892	1,569	1,567
GK	71.8	76.6	76.7	88.7	**3,730**	1,724	2,345	1,907	1,596	1,594
GB	71.0	75.2	75.2	78.1	77.9	**4,587**	1,612	1,408	1,166	1,151
GU	70.0	72.6	72.6	72.6	72.6	71.5	**4,001**	1,880	1,605	1,599
GH	69.8	71.8	71.9	71.9	71.7	70.7	70.6	**3,156**	1,348	1,363
BA	64.5	65.6	65.4	65.6	65.6	64.7	65.1	65.1	**4,269**	1,532
BC	62.9	63.0	62.8	62.8	62.7	62.1	62.9	63.1	66.6	**4,449**

GT: *Gracilibacillus timonensis* Marseille-P2481; GL: *Gracilibacillus lacisalsi* DSM 19029; GO: *Gracilibacillus orientalis* XH-63; GM: *Gracilibacillus massiliensis* Awa-1; GK: *Gracilibacillus kekensis* K170; GB: *Gracilibacillus boraciitolerans* JCM 21714; GU: *Gracilibacillus ureilyticus* MF38; GH: *Gracilibacillus halophilus* YIM-C55.5; BA: *Bacillus alcalophilus* ATCC 27647; BC: *Bacillus clausii* KSM-K16.

from all other studied members of the genus *Gracilibacillus* in a combination of negative alkaline phosphatase and nitrate reductase activities but the acidification of D-fructose (Table 2). The cellular fatty acids from strain Marseille-P2481 are mainly saturated and the most abundant were 12-methyl-tetradecanoic acid, hexadecanoic acid, and 14 methyl-hexadecanoic acid (45%, 16%, and 14%, respectively). No unsaturated fatty acid was detected (Table 3). Cells are resistant to Penicillin G, amoxicillin, ceftriaxone, and metronidazole, but susceptible to imipenem, rifampicin, gentamicin, and erythromycin.

3.3 | Genome properties

The genome is 4,548,390 bp long with a 39.8% G+C content. It is composed of 11 scaffolds (composed of 12 contigs). Of the 4,335 predicted genes, 4,266 were protein-coding genes and 69 were RNAs (4 complete 16S rRNA, 6 complete 5S rRNA gene, 2 complete and 2 partiel 23S rRNA, and 51 tRNA genes, as well as additional 4 other rRNAs). A total of 3,043 genes (70.24%) were assigned a putative function (by COGs or BLAST against nr). A total of 214 genes were identified as ORFans (6.94%). The remaining genes were annotated as hypothetical proteins (861 genes => 19.92%). The genome statistics are presented in Table 4, and the distribution of genes into COGs functional categories is summarized in Table 5.

3.4 | Comparative genomics

The draft genome sequence structure of strain Marseille-P2481 is summarized in Figure 5. It is smaller than those of *G. orientalis* (4.54 and 4.61 Mb, respectively), but larger than those of G. *halophilus*, *G. boraciitolerans*, *G. kekensis*, *G. ureilyticus*, *G. massiliensis*, *B. alcalophilus*, *G. lacisalsi*, and *B. clausii* (3.03, 3.65, 3.93, 4.07, 4.21, 4.37, 4.41 and 4.52 Mb, respectively). The G+C content of strain Marseille-P2481 is smaller than those of *B. clausii* (39.8 and 44.75%, respectively), but larger than those of *G. boraciitolerans*, *G. kekensis*, *G. massiliensis*, *G. orientalis*, *G. lacisalsi*, *B. alcalophilus*, *G. ureilyticus*, and G. *halophilus* (35.8, 36.0, 36.1, 36.3, 36.8, 37.4, 37.5, and 37.9%, respectively). The gene content of strain Marseille-P2481 is smaller than those of *G. orientalis*, *B. clausii*, and *G. boraciitolerans* (4,335, 4,350, 4,441, and 4,510 genes, respectively), but larger than those of G. *halophilus*, *G. kekensis*, *G. massiliensis*, *B. alcalophilus*, *G. ureilyticus*, and *G. lacisalsi*, (2,999, 3,842, 3,887, 3,973, 4,066, and 4,290 genes, respectively). The gene distribution into COG categories was similar among all compared genomes (Figure 6). In addition, the AGIOS analysis showed that strain Marseille-P2481 shared 2,103, 2,112, 2,004, 2,027, 1,461, 1,982, 1,695, 1,539, and 1,578 orthologous proteins with *G. lacisalsi*, *G. orientalis*, *G. massiliensis*, *G. kekensis*, *G. boraciitolerans*, *G. ureilyticus*, G. *halophilus*, *B. alcalophilus*, and *B. clausii*, respectively (Table 6). When comparing strain Marseille-P2481 to other species, AGIOS values were 69.8, 70.0, 71.0, 71.8, 72.0, 72.1, and 72.3% with G. *halophilus*, *G. ureilyticus*, *G. boraciitolerans*, *G. kekensis*, *G. massiliensis*, *G. orientalis*, and *G. lacisalsi*, respectively (Table 6), but ranged from 62.9% to 64.5% with *B. clausii* and *B. alcalophilus*, respectively (Table 6). In addition, dDDH values relatedness of strain Marseille-P2481 and the compared closest species varied between 19.1 and 28.67% and were 20.5, 19.8, 21.6, 20.1, 19.1, 21.4, 19.3, 23.6, and 28.67% for *G. lacisalsi*, *G. orientalis*, *G. massiliensis*, *G. kekensis*, *G. boraciitolerans*, *G. ureilyticus*, G. *halophilus*, *B. alcalophilus*, and *B. clausii*, respectively (Table 7). Finally, AAI values relatedness between strain Marseille-P2481, *G. lacisalsi*, *G. orientalis*, *G. massiliensis*, *G. kekensis*, *G. boraciitolerans*, *G. ureilyticus*, and G. *halophilus* were 68.72, 68.19, 68.18, 67.90, 68.08, 64.69, and 64.37%, respectively, but were lower when compared with *B. alcalophilus* and *B. clausii*, with 51.72 and 50.73%, respectively (Table 8). These dDDH and AAI values were less than the 70% and 95–96% threshold values for species demarcation, respectively (Chun et al., 2018; Klappenbach et al., 2007; Meier-Kolthoff, Auch,

TABLE 7 dDDH values obtained by comparison of all studied genomes

	GL	GO	GM	GK	GB	GU	GH	BA	BC
GT	20.5% ± 2.35	19.8% ± 2.3	21.6% ± 2.35	20.1% ± 2.3	19.1% ± 2.3	21.4% ± 2.35	19.3% ± 2.3	23.6% ± 2.4	28.67% ± 2.4
GL		29.1% ± 2.4	21.0% ± 2.35	20.9% ± 2.35	20.2% ± 2.3	19.4% ± 2.3	18.7% ± 2.25	18.1% ± 2.25	24.4% ± 2.35
GO			21.0% ± 2.35	20.9% ± 2.35	19.9% ± 2.3	19.4% ± 2.25	18.2% ± 2.25	18.4% ± 2.25	25.2% ± 2.4
GM				35.4% ± 2.45	22.2% ± 2.35	19.4% ± 2.3	19.1% ± 2.3	19.9% ± 2.3	31.2% ± 2.5
GK					21.8% ± 2.35	19.7% ± 2.3	19.2% ± 2.3	18.4% ± 2.25	29.5% ± 2.45
GB						18.5% ± 2.25	17.4% ± 2.2	18.2% ± 2.25	33.9% ± 2.5
GU							16.9% ± 2.2	20.9% ± 2.3	24.6% ± 2.4
GH								27.2% ± 2.4	29.8% ± 2.45
BA									27.4% ± 2.45

GT: *Gracilibacillus timonensis* Marseille-P2481; GL: *Gracilibacillus lacisalsi* DSM 19029; GO: *Gracilibacillus orientalis* XH-63; GM: *Gracilibacillus massiliensis* Awa-1; GK: *Gracilibacillus kekensis* K170; GB: *Gracilibacillus boracitolerans* JCM 21714; GU: *Gracilibacillus ureilyticus* MF38; GH: *Gracilibacillus halophilus* YIM-C55.5; BA: *Bacillus alcalophilus* ATCC 27647; BC: *Bacillus clausii* KSM-K16.

Klenk, & Göker, 2013; Richter & Rosselló-Móra, 2009; Rodriguez-R & Konstantinidis, 2014).

4 | DISCUSSION

Due to the concept of microbial culturomics, aiming at exploring the diversity of the human microbiota as exhaustively as possible, many new bacterial species have been discovered over the past 5 years (Lagier et al., 2016). This concept is based on the diversification of physicochemical parameters of culture conditions (Lagier et al., 2012, 2016; Lagier, Hugon, et al.,2015) to mimick as closely as possible the entirety of selective constraints that have shaped the human flora. To date, 329 new species have been characterized (Lagier et al., 2017). These new species include 52 species belonging to the order *Bacillales*, which is one of the most represented bacterial orders (Lagier et al., 2016). Using hypersaline conditions, many hitherto unknown bacteria extremely and or moderately halophilic have been identified in humans, including strain Marseille-P2481. To the best of our knowledge, this is the first *Gracilibacillus* species described in the human gut. Whether it is a resident species of the human gut or a transitory species brought by food is as yet unknown. Its phenotypic, phylogenetic, and genomic characteristics suggested that it represents a new species within the genus *Gracilibacillus*. Members of this genus are generally Gram-positive bacteria, aerobic, motile, moderately halophile and produce white colonies although *G. boraciitolerans* forms pink to red colonies (Ahmed, Yokota, & Fujiwara, 2007), and endospore-forming. However, *Gracilibacillus timonensis* sp. nov. differs from other *Gracilibacillus* species in colony color and metabolism of β-galactosidase, L-arabinose, and D-mannitol. In addition, its genomic DNA G + C content differed from those of other *Gracilibacillus* species, and the dDDH, AAI, and AGIOS values comforted its new species status.

5 | CONCLUSION

The moderately halophilic strain Marseille-P2481 was isolated from a stool sample of a 10-year-old healthy Senegalese boy as part of a study of halophilic bacteria from the human gut. Based on its phenotypic, phylogenetic, and genomic characteristics, this strain is proposed to represent a novel species in the genus *Gracilibacillus*, for which the name *Gracilibacillus timonensis* sp. nov. is proposed. Strain Marseille-P2481^T is the type strain of *Gracilibacillus timonensis* sp. nov.

5.1 | Description of *Gracilibacillus timonensis* sp. nov

Gracilibacillus timonensis (*ti.mo.nen'sis*, N. L adj. masc., *timonensis* of Timone, the name of the main hospital of Marseille, France, where the type strain was first isolated).

The bacterium is preferentially aerobic but is able to grow in anaerobic and microaerophilic atmospheres at 37°C. Strain

TABLE 8 Average amino acid identity (AAI) values (%) between *Gracilibacillus timonensis* strain Marseille-P2481T and other closely related species

	GL	GO	GM	GK	GB	GU	GH	BA	BC
GT	68.72	68.19	68.18	67.90	68.08	64.69	64.37	51.72	50.73
GL		85.64	77.21	76.84	75.47	70.41	68.82	52.40	51.31
GO			76.88	76.74	75.23	70.21	68.17	51.95	50.76
GM				90.32	79.78	70.72	68.09	52.02	50.74
GK					80.04	70.55	68.19	52.31	50.83
^{2}GB						69.60	67.34	51.99	50.92
GU							67.03	52.53	51.16
GH								51.53	50.77
BA									57.85

GT: *Gracilibacillus timonensis* Marseille-P2481; GL: *Gracilibacillus lacisalsi* DSM 19029; GO: *Gracilibacillus orientalis* XH-63; GM: *Gracilibacillus massiliensis* Awa-1; GK: *Gracilibacillus kekensis* K170; GB: *Gracilibacillus boraciitolerans* JCM 21714; GU: *Gracilibacillus ureilyticus* MF38; GH: *Gracilibacillus halophilus* YIM-C55.5; BA: *Bacillus alcalophilus* ATCC 27647; BC: *Bacillus clausii* KSM-K16.

Marseille-P2481^{T} is able to grow in media containing up to 20% (w/v) NaCl, but no growth occurs in the absence of NaCl. The optimal culture conditions are 37°C, pH 7.0-8.0, and 7.5% (w/v) NaCl. After 48 hr of incubation at 37°C on our home-made culture medium (7.5% [w/v] NaCl), colonies are creamy orange and circular and have a mean diameter of 0.2 µm. Cells are Gram-positive, motile rods (with peritrichous flagella) that form endospores rods and are slightly curved, with mean diameter and length of 0.5 and 1.9 µm, respectively.

Using an API ZYM strip, positive results were obtained for esterase, esterase lipase, acid phosphatase, naphtol-AS-BI-phosphohydrolase β-galactosidase, β-glucosidase, and α-glucosidase activities, but no reaction was observed for alkaline phosphatase, lipase, Leucine arylamidase, Valine arylamidase, Cystine arylamidase, α-galactosidase, β-glucuronidase, trypsin, α-chymotrypsin, α-mannosidase, α-fucosidase, and N-acetyl-β-glucosaminidase. The API 50CH strip revealed that strain Marseille-P2481 exhibited esculin hydrolysis, but negative reactions were obtained for D-arabitol, L-arabitol, D-glucose, D-fructose, D-fucose, D-galactose, D-lactose, D-maltose, D-ribose, D-saccharose, D-lyxose, D-mannose L-sorbose, D-tagatose, D-turanose, D-xylose, L-xylose, D-arabinose, L-arabinose, D-sorbitol, D-cellobiose, D-melezitose, D-melibiose, D-trehalose, D-raffinose, L-rhamnose, D-adonitol, D-mannitol, L-fucose, amygdalin, arbutin, erythritol, dulcitol, gentiobiose, glycerol, glycogen, inositol, inulin, salicin, starch, xylitol, αD-glucopyranoside, methyl-βD-xylopyranoside, methyl-αD-mannopyranoside, potassium gluconate, potassium-2-ketogluconate potassium-5-ketogluconate, N-acetylglucosamine. Using an API 20NE strip, fermentation of glucose, urease activity, and metabolism of L-arginine, esculin and 4-nitrophenyl-βD-galactopyrasinoside were positive. In contrast, nitrate and indole production, gelatinase activity and metabolism of D-glucose, L-arabinose, D-mannose, D-maltose, D-mannitol, N-acetyl-glucosamine, potassium gluconate, capric acid, malic acid, trisodium citrate, and phenylacetic acid were negative. Cell membrane fatty acids are mainly saturated structures, with 12-methyl-tetradecanoic acid (45%) and hexadecanoic acid (16%) being the most abundant. No unsaturated structure was found. The genomic DNA G+C content is 39.8 mol%. The 16S rRNA and genome sequences are deposited in EMBL-EBI under accession numbers LT223702 and FLKH00000000, respectively. The type strain of *Gracilibacillus timonensis* is strain Marseille-P2481^{T} (= CSUR P2481 = DSM 103076).

ACKNOWLEDGMENTS

This study was funded by the Méditerranée-Infection foundation and the French Agence Nationale de la Recherche under reference Investissements d'Avenir Méditerranée Infection 10-IAHU-03.

CONFLICT OF INTEREST

The authors declare no competing interest in relation to this research.

ORCID

Nicholas Armstrong http://orcid.org/0000-0002-8365-2244
Pierre-Edouard Fournier http://orcid.org/0000-0001-8463-8885

REFERENCES

Ahmed, I., Yokota, A., & Fujiwara, T. (2007). Gracilibacillus boraciitolerans sp. nov., a highly boron-tolerant and moderately halotolerant bacterium isolated from soil. *International Journal of Systematic and Evolutionary Microbiology*, *57*, 796–802. https://doi.org/10.1099/ijs.0.64284-0

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*, 25–29. https://doi.org/10.1038/75556

Atlas, R. M., & Snyder, J. W. (2011). Reagents, Stains, and Media: Bacteriology. In J. Versalovic, K. Carroll, G. Funke, J. Jorgensen, M. Landry & D. W. Warnock (Eds.), *Manual of Clinical Microbiology 10th Ed* (pp. 272–303). Washington, DC: ASM Press Wash. https://doi.org/10.1128/9781555816728.ch17

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. *Nucleic Acids Research*, *43*, D30–D35. https://doi.org/10.1093/nar/gku1216

Cantrell, S. A., Dianese, J. C., Fell, J., Gunde-Cimerman, N., & Zalar, P. (2011). Unusual fungal niches. *Mycologia*, *103*, 1161–1174. https://doi.org/10.3852/11-108

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, *28*, 464–469. https://doi.org/10.1093/bioinformatics/btr703

Carver, T., Thomson, N., Bleasby, A., Berriman, M., & Parkhill, J. (2009). DNAPlotter: Circular and linear interactive genome visualization. *Bioinformatics*, *25*, 119–120. https://doi.org/10.1093/bioinformatics/btn578

Chamroensaksri, N., Tanasupawat, S., Akaracharanya, A., Visessanguan, W., Kudo, T., & Itoh, T. (2010). *Gracilibacillus thailandensis* sp. nov., from fermented fish (pla-ra). *International Journal of Systematic and Evolutionary Microbiology*, *60*, 944–948. https://doi.org/10.1099/ijs.0.011981-0

Chen, Y.-G., Cui, X.-L., Zhang, Y.-Q., Li, W.-J., Wang, Y.-X., Xu, L.-H., … Jiang, C.-L. (2008). Gracilibacillus halophilus sp. nov., a moderately halophilic bacterium isolated from saline soil. *International Journal of Systematic and Evolutionary Microbiology*, *58*, 2403–2408. https://doi.org/10.1099/ijs.0.65698-0

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., … Trujillo, M. E. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *68*, 461–466. https://doi.org/10.1099/ijsem.0.002516

Dione, N., Sankar, S. A., Lagier, J.-C., Khelaifia, S., Michele, C., Armstrong, N., … Fournier, P.-E. (2016). Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. *New Microbes New Infect*, *10*, 66–76. https://doi.org/10.1016/j.nmni.2016.01.002

Diop, A., Khelaifia, S., Armstrong, N., Labas, N., Fournier, P.-E., Raoult, D., & Million, M. (2016). Microbial culturomics unravels the halophilic microbiota repertoire of table salt: Description of *Gracilibacillus massiliensis* sp. nov. *Microbial Ecology in Health and Disease*, *27*, https://doi.org/10.3402/mehd.v27.32049

Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J.-P., & Raoult, D. (2000). 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *Journal of Clinical Microbiology*, *38*, 3623–3630.

Drancourt, M., Bollet, C., & Raoult, D. (1997). *Stenotrophomonas africana* sp. nov., an opportunistic human pathogen in Africa. *International Journal of Systematic and Evolutionary Microbiology*, *47*, 160–163. https://doi.org/10.1099/00207713-47-1-160

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., … Eddy, S. R. (2015). HMMER web server: 2015 update. *Nucleic Acids Research*, *43*, W30–W38. https://doi.org/10.1093/nar/gkv397

Fischer, A. (1895). Untersuchungen über bakterien. *Jahrbücher für Wissenschaftliche Botanik*, *27*, 1–163.

Gao, M., Liu, Z.-Z., Zhou, Y.-G., Liu, H.-C., Ma, Y.-C., Wang, L., … Ji, X.-C. (2012). Gracilibacillus kekensis sp. nov., a moderate halophile isolated from Keke Salt Lake. *International Journal of Systematic and Evolutionary Microbiology*, *62*, 1032–1036. https://doi.org/10.1099/ijs.0.030858-0

Garrity, G. M., & Holt, J. (2001). The road map to the manual. In G. M. Garrity, D. R. Boone, & R. W. Castenholz (Eds.), *Bergey's manual of systematic bacteriology* (Vol. *1*, 2nd ed., pp. 119–169). New York: Springer-Verlag.

Gibbons, N. E., & Murray, R. G. E. (1978). Proposals concerning the higher taxa of *Bacteria*. *International Journal of Systematic and Evolutionary Microbiology*, *28*, 1–6. https://doi.org/10.1099/00207713-28-1-1

Gouret, P., Paganini, J., Dainat, J., Louati, D., Darbo, E., Pontarotti, P., & Levasseur, A. (2011). Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: The multi-agent software system DAGOBAH. In P. Pontarotti, (ed.) *Evolutionary biology – concepts, biodiversity, macroevolution and genome evolution* (pp. 71–87). Berlin, Heidelberg: Springer-Verlag. https://doi.org/doi-org.insb.bib.cnrs.fr/10.1007/978-3-642-20763-1_5

Gouret, P., Vitiello, V., Balandraud, N., Gilles, A., Pontarotti, P., & Danchin, E. G. (2005). FIGENIX: Intelligent automation of genomic annotation: Expertise integration in a new software platform. *BMC Bioinformatics*, *6*, 198. https://doi.org/10.1186/1471-2105-6-198

Hirota, K., Hanaoka, Y., Nodasaka, Y., & Yumoto, I. (2014). Gracilibacillus alcaliphilus sp. nov., a facultative alkaliphile isolated from indigo fermentation liquor for dyeing. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 3174–3180. https://doi.org/10.1099/ijs.0.060871-0

Huo, Y.-Y., Xu, X.-W., Cui, H.-L., & Wu, M. (2010). *Gracilibacillus ureilyticus* sp. nov., a halotolerant bacterium from a saline-alkaline soil. *International Journal of Systematic and Evolutionary Microbiology*, *60*, 1383–1386. https://doi.org/10.1099/ijs.0.016808-0

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Jeon, C. O., Lim, J.-M., Jang, H. H., Park, D.-J., Xu, L.-H., Jiang, C.-L., & Kim, C.-J. (2008). Gracilibacillus lacisalsi sp. nov., a halophilic Gram-positive bacterium from a salt lake in China. *International Journal of Systematic and Evolutionary Microbiology*, *58*, 2282–2286. https://doi.org/10.1099/ijs.0.65369-0

Käll, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, *338*, 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Khelaifia, S., Lagier, J.-C., Bibi, F., Azhar, E. I., Croce, O., Padmanabhan, R., … Raoult, D. (2016). Microbial culturomics to map halophilic bacterium in human gut: Genome sequence and description of *Oceanobacillus jeddahense* sp. nov. *Omics: A Journal of Integrative Biology*, *20*, 248–258. https://doi.org/10.1089/omi.2016.0004

Kim, P., Lee, J.-C., Park, D.-J., Shin, K.-S., Kim, J.-Y., & Kim, C.-J. (2012). Gracilibacillus bigeumensis sp. nov., a moderately halophilic bacterium from solar saltern soil. *International Journal of Systematic and Evolutionary Microbiology*, *62*, 1857–1863. https://doi.org/10.1099/ijs.0.034264-0

Kim, M., Oh, H.-S., Park, S.-C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 346–351. https://doi.org/10.1099/ijs.0.059774-0

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*, 111–120. https://doi.org/10.1007/BF01731581

Klappenbach, J. A., Goris, J., Vandamme, P., Coenye, T., Konstantinidis, K. T., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*, 81–91. https://doi.org/10.1099/ijs.0.64483-0

Klenk, H.-P., Meier-Kolthoff, J. P., & Göker, M. (2014). Taxonomic use of DNA G+C content and DNA?DNA hybridization in the genomic age. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 352–356. https://doi.org/10.1099/ijs.0.056994-0

Konstantinidis, K. T., Ramette, A., & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*, 1929–1940. https://doi.org/10.1098/rstb.2006.1920

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation

of ribosomal RNA genes. *Nucleic Acids Research*, *35*, 3100–3108. https://doi.org/10.1093/nar/gkm160

Lagier, J.-C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., … Raoult, D. (2012). Microbial culturomics: Paradigm shift in the human gut microbiome study. *Clinical Microbiology & Infection*, *18*, 1185–1193. https://doi.org/10.1111/1469-0691.12023

Lagier, J. C., Drancourt, M., Charrel, R., Bittar, F., La Scola, B., Ranque, S., & Raoult, D. (2017). Many more microbes in humans: Enlarging the microbiome repertoire. *Clinical Infectious Diseases*, *65*, S20–S29. https://https://doi.org/10.1093/cid/cix404

Lagier, J.-C., Hugon, P., Khelaifia, S., Fournier, P.-E., La Scola, B., & Raoult, D. (2015). The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. *Clinical Microbiology Reviews*, *28*, 237–264. https://doi.org/10.1128/CMR.00014-14

Lagier, J.-C., Khelaifia, S., Alou, M. T., Ndongo, S., Dione, N., Hugon, P., … Raoult, D. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature Microbiology*, *1*, 16203.https://doi.org/10.1038/nmicrobiol.2016.203

Lagier, J.-C., Khelaifia, S., Azhar, E. I., Croce, O., Bibi, F., Jiman-Fatani, A. A., … Raoult, D. (2015). Genome sequence of *Oceanobacillus picturae* strain S1, an halophilic bacterium first isolated in human gut. *Standards in Genomic Sciences*, *10*, https://doi.org/10.1186/s40793-015-0081-2

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (co-) orthologs in large-scale analysis. *BMC Bioinformatics*, *12*, 124.https://doi.org/10.1186/1471-2105-12-124

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*, 955–964. https://doi.org/10.1093/nar/25.5.0955

Ludwig, W., Schleifer, K. H., & Whitman, W. B. (2009). Class I. Bacilli class nov. In: P. De Vos, G. Garrity, D. Jones, N. R. Krieg, W. Ludwig, F. A. Rainey, K. H. Schleifer & W. B. Whitman (Eds.), *Bergey's manual of systematic bacteriology* (Vol. *3*, 2nd ed., pp. 19–20). New York: Springer-Verlag.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, *14*, 60. https://doi.org/10.1186/1471-2105-14-60

Murray, R. G. E. (1984). The higher taxa, or, a place for everything…? In: J. G. Holt (ed.), *Bergey's manual of systematic bacteriology* (Vol. *1*, 1st ed., pp. 31–34). Baltimore: The Williams and Wilkins Co.

Padmanabhan, R., Mishra, A. K., Raoult, D., & Fournier, P.-E. (2013). Genomics and metagenomics in medical microbiology. *Journal of Microbiol Methods*, *95*, 415–424. https://doi.org/10.1016/j.mimet.2013.10.006

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., … Kyrpides, N. C. (2012). The Genomes OnLine Database (GOLD) v. 4: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, *40*, D571–D579. https://doi.org/10.1093/nar/gkr1100

Parte, A. C. (2014). LPSN—list of prokaryotic names with standing in nomenclature. *Nucleic Acids Research*, *42*, D613–D616. https://doi.org/10.1093/nar/gkt1111

Ramasamy, D., Mishra, A. K., Lagier, J.-C., Padhmanabhan, R., Rossi, M., Sentausa, E., … Fournier, P.-E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 384–391. https://doi.org/10.1099/ijs.0.057091-0

Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*, *106*, 19126–19131. https://doi.org/10.1073/pnas.0906412106

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). Bypassing cultivation to identify bacterial species. *Microbe*, *9*, 111–118.

Sasser, M. (1990). Identification of bacteria by gas chromatography of cellular fatty acids. http://natasha.eng.usf.edu/gilbert/courses/Biotransport%20Phenomena/pdf/bacteria_gc_1.pdf (Accessed June 13, 2017)

Senghor, B., Seck, E. H., Khelaifia, S., Bassène, H., Sokhna, C., Fournier, P.-E., … Lagier, J.-C. (2017). Description of *"Bacillus dakarensis"* sp. nov., *"Bacillus sinesaloumensis"* sp. nov., *"Gracilibacillus timonensis"* sp. nov., *"Halobacillus massiliensis"* sp. nov., *"Lentibacillus massiliensis"* sp. nov., *"Oceanobacillus senegalensis"* sp. nov., *"Oceanobacillus timonensis"* sp. nov., *"Virgibacillus dakarensis"* sp. nov. and *"Virgibacillus marseillensis"* sp. nov., nine halophilic new species isolated from human stool. *New Microbes and new Infections*, *17*, 45–51. https://doi.org/10.1016/j.nmni.2017.01.010

Sentausa, E., & Fournier, P.-E. (2013). Advantages and limitations of genomics in prokaryotic taxonomy. *Clinical Microbiology & Infection*, *19*, 790–795. https://doi.org/10.1111/1469-0691.12181

Skerman, V. B. D., & Sneath, P. H. A. (1980). Approved list of bacterial names. *International Journal of Systematic Bacteriology*, *30*, 225–420. https://doi.org/10.1099/00207713-30-1-225

Stackebrandt, E., & Ebers, J. (2006). Taxonomic parameters revisited: Tarnished gold standards. *Microbiol Today.*, *33*, 152–155.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*, 2725–2729. https://doi.org/10.1093/molbev/mst197

Tang, S.-K., Wang, Y., Lou, K., Mao, P.-H., Jin, X., Jiang, C.-L., … Li, W.-J. (2009). Gracilibacillus saliphilus sp. nov., a moderately halophilic bacterium isolated from a salt lake. *International Journal of Systematic and Evolutionary Microbiology*, *59*, 1620–1624. https://doi.org/10.1099/ijs.0.006569-0

Wainø, M., Tindall, B. J., Schumann, P., & Ingvorsen, K. (1999). Gracilibacillus gen. nov., with description of *Gracilibacillus halotolerans* gen. nov., sp. nov.; transfer of *Bacillus dipsosauri* to *Gracilibacillus dipsosauri* comb. nov., and *Bacillus salexigens* to the genus Salibacillus gen. nov., as *Salibacillus salexigens* comb. nov. *International Journal of Systematic and Evolutionary Microbiology*, *49*, 821–831. https://doi.org/10.1099/00207713-49-2-821

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: Proposal for the domains *Archaea*, *Bacteria*, and *Eukarya*. *Proceedings of the National Academy of Sciences U.S.A.*, *87*, 4576–4579. https://doi.org/10.1073/pnas.87.12.4576

How to cite this article: Diop A, Seck EH, Dubourg G, et al. Genome sequence and description of *Gracilibacillus timonensis* sp. nov. strain Marseille-P2481^T, a moderate halophilic bacterium isolated from the human gut microflora. *MicrobiologyOpen*. 2018;e638. https://doi.org/10.1002/mbo3.638

Article 8:

Microbial culturomics to isolate halophilic bacteria from table salt: Genome sequence and description of the moderately halophilic bacterium *Bacillus salis* sp. nov.

Seck EH, Diop A, Dubourg G, Armstrong N, Delerce J, Fournier PE, Raoult D, Khelaifia S.

[Published in New Microbes New Infections]

Microbial culturomics to isolate halophilic bacteria from table salt: genome sequence and description of the moderately halophilic bacterium *Bacillus salis* sp. nov.

E. H. Seck[1], A. Diop[1], N. Armstrong[1], J. Delerce[1], P.-E. Fournier[1], D. Raoult[1,2] and S. Khelaifia[1]

1) URMITE, UM 63, CNRS 7278, IRD 198, Inserm 1095, Institut Hospitalo-Universitaire Méditerranée-Infection, Faculté de Médecine, Aix-Marseille Université, Marseille, France and 2) Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

Abstract

Bacillus salis strain ES3^T (= CSUR P1478 = DSM 100598) is the type strain of *B. salis* sp. nov. It is an aerobic, Gram-positive, moderately halophilic, motile and spore-forming bacterium. It was isolated from commercial table salt as part of a broad culturomics study aiming to maximize the culture conditions for the in-depth exploration of halophilic bacteria in salty food. Here we describe the phenotypic characteristics of this isolate, its complete genome sequence and annotation, together with a comparison with closely related bacteria. Phylogenetic analysis based on 16S rRNA gene sequences indicated 97.5% similarity with *Bacillus aquimaris,* the closest species. The 8 329 771 bp long genome (one chromosome, no plasmids) exhibits a G+C content of 39.19%. It is composed of 18 scaffolds with 29 contigs. Of the 8303 predicted genes, 8109 were protein-coding genes and 194 were RNAs. A total of 5778 genes (71.25%) were assigned a putative function.

Keywords: *Bacillus salis*, culturomics, genome, halophilic bacteria, human gut, taxonogenomics
Original Submission: 31 October 2017; **Revised Submission:** 13 December 2017; **Accepted:** 20 December 2017
Article published online: 10 January 2018

Corresponding author. S. Khelaifia, URMITE, UM63, UMR CNRS 7278, IRD198, INSERM U1095, Faculté de Médecine, Aix-Marseille Université, IHU-Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13385 Marseille Cedex 5, France.
E-mail: khelaifia_saber@yahoo.fr

Introduction

Halophiles are considered as microorganisms living in hypersaline environments which often require a high salt concentration for growth. They are involved in centuries-old processes, such as production of salt and fermentation of food consumed by humans [1,2]. Today, with the emergence of new biologic technologies, these organisms have been isolated and described from many traditional foods [2] such as salt [3].

Despite recent technologic advances in molecular biology, pure culture is the only way to characterize the physiologic properties of bacteria and to evaluate their potential virulence [4]. Therefore, we tried to investigate the population of halophilic prokaryotes in the human gut and salty food by using a culturomics approach. This approach allowed us to isolate a new member of the *Bacillus* genus. This bacterium is Gram negative, strictly aerobic, moderately halophilic and motile. It was isolated from commercial table salt. This isolation was part of a culturomics study using high-salt culture conditions in order to cultivate halophilic bacteria from human faeces and environmental samples [5]. This isolate is described using a new and innovative method that we have implemented [6]. The old methods, based on 16S rRNA sequencing, phylogeny, G + C content and DNA-DNA hybridization (DDH), are fastidious and include many limitations [6,7].

The emergence of new tools for DNA sequencing and technology, such as matrix-assisted desorption ionization–time of flight mass spectrometry (MALDI-TOF MS), has allowed an increase in available genomic and proteomic data over the last few years [8,9]. These technologic advances have allowed us to develop a new way of describing bacterial species that takes into account genomic and protonic information [10].

Here we present a summary classification and a set of features for *B. salis* strain ES3^T (= CSUR P1478 = DSM 100598), together with the description of its complete genomic sequence and its annotation.

Materials and methods

Strain isolation and identification

Culture condition. Culture was realized in an aerobic atmosphere on a homemade culture medium consisting of a Columbia agar culture (Sigma-Aldrich, Saint-Quentin Fallavier, France) modified by adding (per liter): $MgCl_2$ $6H_2O$, 5 g; $MgSO_4$ $7H_2O$, 5 g; KCl, 2 g; $CaCl_2$ $2H_2O$, 1 g; NaBr, 0.5 g; $NaHCO_3$, 0.5 g, glucose, 2 g and 100 g/L of NaCl. The pH was adjusted to 7.5 with 10 M NaOH before autoclaving [3].

New Microbe and New Infect 2018; **23:** 28–38

https://doi.org/10.1016/j.nmni.2017.12.006

MALDI-TOF MS identification. The identification of our strain was carried out by a MALDI-TOF MS analysis with a Microflex spectrometer (Bruker Daltonics, Leipzig, Germany) as previously described [11]. Obtained spectra were then compared by using MALDI Biotyper 3.0 software (Bruker) as well as the Unité des Maladies Infectieuses et Tropicales Emergentes's (URMITE) database, which is constantly updated. If no identification was possible at the genus or species level (score <1.7), sequencing of the 16S rRNA gene was performed to achieve a correct identification [12,13].

Sequencing of 16S rRNA gene. DNA extraction was performed using the EZ1 DNA Tissue Kit and BioRobot EZ1 Advanced XL (Qiagen, Courtaboeuf, France). The 16S rRNA gene was amplified using PCR technology and universal primers fD1 and rP2 [12] (Eurogentec, Angers, France). The amplifications and sequencing of the amplified products were performed as previously described [14]. Then 16S rRNA gene sequences were assembled and corrected using Codoncode Aligner software (http://www.codoncode.com/) and compared with those available in GenBank (http://www.ncbi.nlm.nih.gov/genbank/). Identification at the species level was defined by a 16S rRNA gene sequence similarity of ≥99% with the sequence of the type strain in GenBank. When the percentage of identity was <98.7%, the studied strain was considered as a new species [15].

Phylogenetic classification

Phylogenetic analysis based on 16S rRNA of our isolate was performed to identify its phylogenetic affiliations with other close isolates, including other members of the genus *Bacillus*. MEGA 6 (Molecular Evolutionary Genetics Analysis) software allowed us to construct a phylogenetic tree [16]. Sequence alignment of the different species was performed using CLUSTAL W [17], and evolutionary distance matrices for the neighbour-joining method were calculated using the algorithm of the Kimura two-parameter model [18].

Physiologic and phenotypic characteristics

Phenotypic tests. The phenotypic characteristics of this strain were studied by testing different parameters. Regarding temperature, we studied growth at 25, 30, 37, 45 and 56°C. Growth at various NaCl concentrations (0.5, 5, 7.5, 10, 15, 200 and 250%) was also investigated. The optimal pH for growth was determined by testing different pHs: 5, 6, 6.5, 7, 7.5, 8, 9 and 10. Growth of strain ES3^T was tested under aerobic atmosphere, in the presence of 5% CO_2 and also under anaerobic and microaerophilic atmospheres, created using AnaeroGen (Thermo Fisher Scientific, Saint Aubin, France) and CampyGen (Thermo Fisher Scientific) respectively.

Microscopy. Gram staining and motility were observed with a DM1000 light microscope (Leica Microsystems, Nanterre, France). Cell morphology was studied using a Tecnai G^{20} Cryo (FEI Company, Limeil-Brévannes, France) transmission electron microscope operated at 200 keV after negative staining of bacteria. Cells were first fixed with 2.5% glutaraldehyde in 0.1 M cacodylate buffer for at least 1 hour at 4°C. A drop of cell suspension was deposited for approximately 5 minutes on glow-discharged formvar carbon film on 400 mesh nickel grids (FCF400-Ni; Electron Microscopy Sciences (EMS), Hatfield, PA, USA). The grids were dried on blotting paper, and cells were negatively stained for 10 seconds with 1% ammonium molybdate solution in filtered water at room temperature. Formation of spores was determined after thermal shock and observed under a microscope.

Biochemical test. Acid production from carbohydrates was determined by using the API 50CHB system (bioMérieux, Marcy l'Etoile, France). Other physiologic tests were performed with the API 20NE system (bioMérieux) and API ZYM (bioMérieux), according to the manufacturer's instructions.

Antibiotic susceptibility test. Antibiotic susceptibility was determined on Mueller-Hinton agar in a petri dish using the disc diffusion method according to European Committee on Antimicrobial Susceptibility Testing recommendations (bioMérieux) [19]. The following antibiotics were tested: doxycycline, rifampicin, vancomycin, nitrofurantoin, amoxicillin, erythromycin, ampicillin, ceftriaxone, ciprofloxacin, gentamicin, penicillin, trimethoprim/sulfamethoxazole, imipenem and metronidazole.

Fatty acid analysis. Cellular fatty acid methyl ester (FAME) analysis was performed by gas chromatography/mass spectrometry (GC/MS). Two samples were prepared with approximately 85 mg of bacterial biomass per tube collected from several culture plates. FAMEs were prepared as described by Sasser [20]. GC/MS analyses were carried out as previously described [21]. Briefly, FAMEs were separated using an Elite 5-MS column and monitored by mass spectrometry (Clarus 500−SQ 8 S; Perkin Elmer, Courtaboeuf, France). Spectral database search was performed using MS Search 2.0 operated with the Standard Reference Database 1A (National Institute of Standards and Technology, Gaithersburg, MD, USA) and the FAME mass spectral database (Wiley, Chichester, UK).

Genome sequencing

Genomic DNA (gDNA) of *Bacillus salis* was extracted in two steps. A mechanical treatment was first performed by acid-washed glass beads (G4649-500g; Sigma-Aldrich, St. Louis, MO, USA) using a FastPrep BIO 101 instrument (Qbiogene,

TABLE 1. Classification and general features of *Bacillus salis* strain ES3^T

Property	Term
Current classification	Domain: *Bacteria* Phylum: *Firmicutes* Class: *Bacilli* Order: *Bacillales* Family: *Bacillaceae* Genus: *Bacillus* Species: *Bacillus salis* Type strain: ES3^T
Gram stain	Positive
Cell shape	Rod shaped
Motility	Motile
Sporulation	Endospore forming
Temperature range	Mesophile
Optimum temperature	37°C
Optimum pH	7.5
Salinity	5.0–200 g/L
Optimum salinity	100 g/L
Oxygen requirement	Aerobic

Strasbourg, France) at maximum speed (6.5 m/s) for 90 seconds. Then after a 2-hour lysozyme incubation at 37°C, DNA was extracted on the EZ1 biorobot (Qiagen) with an EZ1 DNA tissue kit. The elution volume was 50 µL. gDNA was quantified by a Qubit assay with the high-sensitivity kit (Life Technologies, Carlsbad, CA, USA) to 120 ng/µL.

gDNA was sequenced with MiSeq Technology (Illumina, San Diego, CA, USA) with the mate-pair strategy. The gDNA was barcoded to be mixed with 11 other projects with the Nextera Mate Pair sample prep kit (Illumina). The mate-pair library was prepared with 1.5 µg gDNA using the Nextera mate-pair Illumina guide. The gDNA sample was simultaneously fragmented and tagged with a mate-pair junction adapter. The pattern of the fragmentation was validated on an Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA) with a DNA 7500 labchip. The DNA fragments ranged in size from 1.5 to 11 kb, with an optimal size of 6.859 kb. No size selection was performed, and 600 ng of tagmented fragments were circularized. The circularized DNA was mechanically sheared to small fragments with an optimal at 921 bp on the Covaris S2 device in T6 tubes (Covaris, Woburn, MA, USA). The library profile was visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies), and the final concentration library was measured at 39.94 nmol/L. The libraries were normalized at 2 nM, and this library was added as two spots and all were pooled. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated

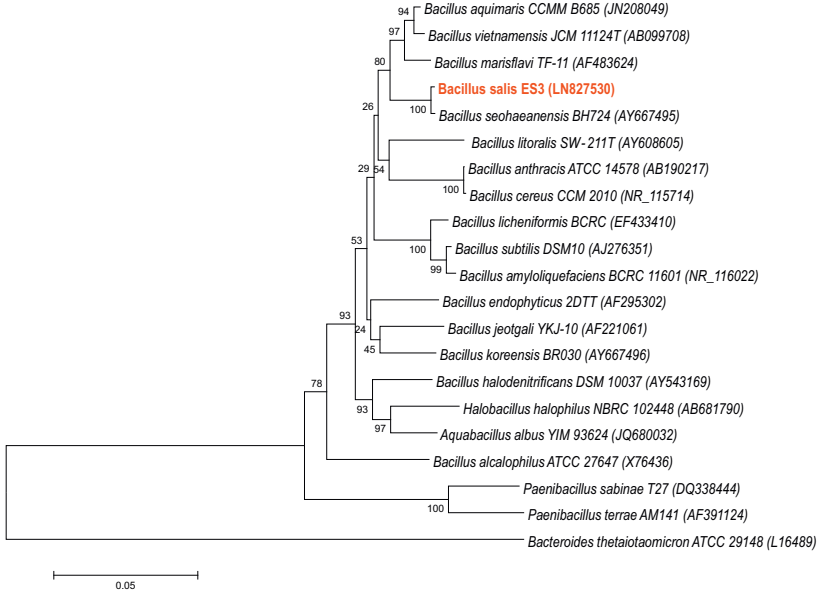


FIG. 1. Phylogenetic tree highlighting position of *Bacillus salis* strain ES3^T relative to other close species. GenBank accession numbers are indicated in parentheses. Sequences were aligned using CLUSTAL W, and phylogenetic inferences were obtained by Kimura two-parameter model within MEGA 6 software. *Bacteroides thetaiotaomicron* was used as outgroup. Scale bar represents 0.05% nucleotide sequence divergence.

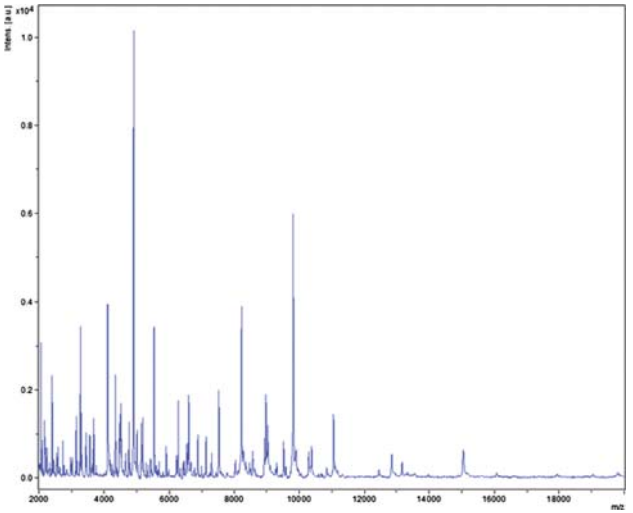


FIG. 2. Reference mass spectrum from *Bacillus salis* strain ES3^T. Spectra from 12 individual colonies were compared and reference spectrum generated.

cluster generation and a sequencing run were performed in a single 39-hour run with a 2 × 251 bp read length. Total information of 5.5 Gb was obtained from a 572K/mm^2 cluster density, with a cluster passing quality control filters of 96.33% (11 740 000 passing filter paired reads). Within this run, the index representation for *Bacillus salis* was determined to be 14.60%. The 1 662 573 paired reads were trimmed and then assembled.

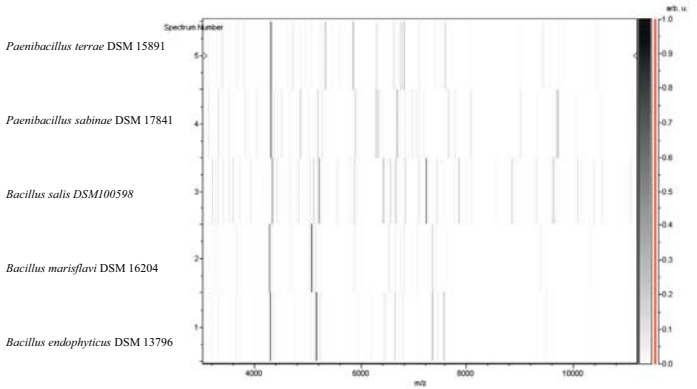


FIG. 3. Gel view comparing *Bacillus salis* strain ES3^T to members of genera *Bacillus* and *Paenibacillus*. Gel view displays raw spectra of all loaded spectrum files arranged in pseudo–gel-like look. *X*-axis records *m*/*z* value. Left *y*-axis displays running spectrum number originating from subsequent spectra loading. Peak intensity is expressed by greyscale scheme code. Colour bar and right *y*-axis indicate relation between colour peak; peak intensity is expressed in arbitrary units. Displayed species are indicated at left.

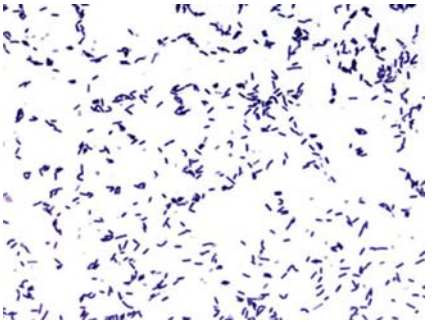


FIG. 4. Gram staining of *Bacillus salis* strain ES3^T.

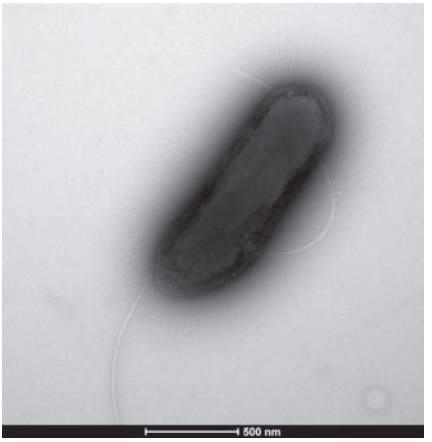


FIG. 5. Transmission electron microscopy of *Bacillus salis* strain ES3^T. Cells were observed with Tecnai G20 transmission electron microscope operated at 200 keV. Scale bar = 500 nm.

Genome annotation and comparison

The genome's assembly was performed with a pipeline that enabled us to create an assembly with different software (Velvet [22], Spades [23] and Soap Denovo [24]) on trimmed (MiSeq and Trimmomatic softwares) [25] or untrimmed data (only MiSeq software). For each of the six assemblies performed, GapCloser [24] was used to reduce gaps. Then contamination with Phage Phix was identified (BLASTn against Phage Phix174 DNA sequence) and eliminated. Finally, scaffolds of size <800 bp were removed, and scaffolds with a depth value of <25% of the mean depth were removed (identified as possible contaminants). The best assembly was selected by using different criteria (number of scaffolds, N50, number of N). For the studied strain, Spades gave the best assembly, with a depth coverage of 99×.

Open reading frames (ORFs) were predicted using Prodigal [26] with default parameters, but the predicted ORFs were excluded if they were spanning a sequencing gap region (contained N). The predicted bacterial protein sequences were searched against the Clusters of Orthologous Groups (COGs) database using BLASTP (*E* value 1e-03, coverage 0.7 and identity percentage 30%). If no hit was found, sequences were searched against the NR database using BLASTP with a *E* value of 1e-03, coverage 0.7 and identity percentage 30%. If the sequence length was smaller than 80 aa, we used an *E* value of 1e-05. The tRNAScanSE tool [27] was used to find transfer RNA genes, whereas ribosomal RNAs were found using RNAmmer [28]. Lipoprotein signal peptides and the number of transmembrane helices were predicted using Phobius [29]. ORFans were identified if the BLASTP performed did not give positive results (*E* value was lower than 1e-03 for ORFs with sequence size >80 aa; if alignment lengths were <80 aa, we used an *E* value of 1e-05). Such parameter thresholds have been used in previous work to define ORFans. The annotation process was performed in DAGOBAH [30], which includes Figenix [31] libraries that provided pipeline analysis.

Artemis was used for data management and DNAPlotter [32] for visualization of genomic features. The Mauve alignment tool (version 2.3.1) was used for multiple genomic sequence alignment [33]. To estimate the mean level of nucleotide sequence similarity at the genome level, we used MAGI homemade software to calculate the average genomic identity of orthologous gene sequences (AGIOS) among compared genomes. Briefly, this software is combined with the Proteinortho software [34] for detecting orthologous proteins in pairwise genomic comparisons; it then retrieves the corresponding genes and determines the mean percentage of nucleotide sequence identity among orthologous ORFs using the Needleman-Wunsch global alignment algorithm. Genomes from the genus *Bacillus* and closely related genera were used for the calculation of AGIOS values. The genomic similarity was evaluated among studied species close to the isolate by digital DNA-DNA hybridization (http://ggdc.dsmz.de/distcalc2.php).

Results and discussion

Strain identification and phylogenetic analyses

Strain ES3^T was first isolated in May 2014 (Table 1) after 30 days of preincubation in aerobic culture on our homemade culture medium at 37°C. No significant MALDI-TOF MS score was obtained for strain ES3^T against the Bruker and URMITE databases, suggesting that our isolate was not a member of a

TABLE 2. Differential characteristics of *Bacillus salis* strain ES3^T and *Bacillus marisflavi* strain TF-11^T [36], *Bacillus endophyticus* strain 2DT^T [37], *Halobacillus halophilus* strain SL-4^T [38], *Paenibacillus terrae* strain AM141^T [39] and *Paenibacillus sabinae* strain T27^T [40]

Characteristic	*B. salis*	*B. marisflavi*	*B. endophyticus*	*H. halophilus*	*P. terrae*	*P. sabinae*
Cell diameter (μm)	1.8	0.6–0.8	0.5–1.5	0.6–0.8	0.8–1.1	0.7–3.2
Oxygen requirement	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic	Aerobic
Gram stain	+	+ to v	+ to v	+	v	+
Motility	+	+	–	+	+	+
Endospore formation	+	+	–	+	+	+
Production of:						
Catalase	+	+	–	+	+	+
Oxidase	–	–	+	+	–	–
Nitrate reductase	+	NA	–	–	+	+
Urease	+	–	–	–	–	NA
β-Galactosidase	–	NA	NA	NA	–	NA
N-acetyl-β-glucosaminidase	–	NA	NA	NA	+	NA
Acid from:						
L-Arabinose	–	–	+	NA	–	–
D-Ribose	+	+	+	NA	–	+
D-Mannose	+	+	+	+	+	NA
D-Mannitol	–	+	–	–	+	NA
D-Sucrose	–	–	+	–	+	–
D-Glucose	+	+	+	–	+	+
D-Fructose	+	+	–	–	–	–
D-Maltose	–	–	–	NA	+	+
D-Lactose	–	–	–	NA	–	–
Starch	+	+	+	NA	NA	NA
Gelatin	+	+	+	NA	NA	NA
Habitat	Table salt	Seawater	Soil sediment	Soil	Soil	Salt lake

+, positive result; –, negative result; v, variable result; NA, data not available.

known species [9]. An almost complete 16S rRNA gene sequence of strain ES3^T (accession no. LN827530) comprising 1505 nt was analysed. Comparative 16S rRNA gene sequences analyses showed that strain ES3^T is phylogenetically affiliated with the *Bacillus* genus (Fig. 1). The phylogenetic distinctiveness (16S rRNA gene sequence similarity of <97%) confirms that strain ES3^T represents a distinct species from the recognized species belonging to *Bacillus* genus [35]. In fact, strain ES3^T exhibited 97.5% nucleotide sequence similarity with *Bacillus aquimaris*, the phylogenetically closest species with a validly published name [36]. The reference spectrum for strain ES3^T was thus incremented in our database (Fig. 2), then compared to other known species of the genus *Bacillus*. The differences exhibited are shown in Fig. 3 in the obtained gel view.

TABLE 3. Cellular fatty acid composition (%)

Fatty acid	IUPAC Name	Mean relative %[a]
15:0 anteiso	12-methyl-Tetradecanoic acid	59.6 ± 1.1
17:0 anteiso	14-methyl-Hexadecanoic acid	17.3 ± 1.0
15:0 iso	13-methyl-Tetradecanoic acid	10.1 ± 1.6
16:0	Hexadecanoic acid	3.7 ± 0.2
14:0	Tetradecanoic acid	2.7 ± 0.4
16:0 iso	14-methyl-Pentadecanoic acid	2.1 ± 0.3
17:0 iso	15-methyl-Hexadecanoic acid	1.5 ± 0.1
16:1n9	7-Hexadecenoic acid	TR
5:0 anteiso	2-methyl-Butanoic acid	TR
14:0iso	12-methyl-Tridecanoic acid	TR
13:0 anteiso	10-methyl-Dodecanoic acid	TR
17:1 iso	15-methyl-Hexadecenoic acid	TR
19:0 anteiso	16-methyl-Octadecanoic acid	TR
18:0	Octadecanoic acid	TR
16:1 iso	14-methyl-Pentadecenoic acid	TR
13:0 iso	11-methyl-Dodecanoic acid	TR
12:0	Dodecanoic acid	TR

IUPAC, International Union of Pure and Applied Chemistry; TR, trace amounts < 1%.
[a]Mean peak area percentage.

Phenotypic description

Strain ES3^T formed creamy, smooth, circular and slightly irregular colonies 5 to 8 mm in diameter after incubation at 37°C for 2 days on our halophilic medium under an aerobic atmosphere. Growth occurred between 25 and 40°C, but not at 55°C. No growth was observed without NaCl, and the strain grew at salt concentrations ranging from 1% to 25% (w/v) NaCl, with optimum growth occurring at 10% (w/v) NaCl. Growth occurred between pH 6 and 10, with an optimum at pH 7.5. Cells were motile and spore forming. Gram staining (Fig. 4) showed Gram-positive rods. Strain ES3^T exhibited catalase activity but no oxidase. Measured by electron microscopy, the rods had a mean diameter of 1.8 μm and a length of 5.9 μm (Fig. 5).

Biochemical test. Using API 50CH strip, positive reactions was observed for D-glucose, D-fructose, D-mannose, arbutin, esculin ferric citrate, salicin, D-maltose, D-saccharose, D-trehalose, melezitose, D-raffinose and amidon; and negative reactions were recorded for glycerol, erythritol, D-arabinose, L-arabinose, D-ribose, D-xylose, L-xylose, D-adonitol, methyl-βD-xylopyranoside, D-galactose, L-sorbose, L-rhamnose, dulcitol, inositol, D-mannitol, D-sorbitol, methyl-αD-mannopyranoside, methyl-αD-glucopyranoside, *N*-acetyl-glucosamine, D-

TABLE 4. Nucleotide content and gene count levels of genome

Attribute	Value	% of total[a]
Size (bp)	8 329 771	100
G+C content (bp)	3 263 777	39.18
Coding region (bp)	6 920 184	83.07
Total genes	8303	100
RNA genes	194	2.33
Protein-coding genes	8109	97.66
Genes with function prediction	5778	71.25
Genes assigned to COGs	5277	65.07
Genes with peptide signals	869	10.71
Genes with transmembrane helices	2032	25.05

COGs, Clusters of Orthologous Groups database.
[a]The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

TABLE 5. Number of genes associated with 25 general COGs functional categories

Code	Value	% value	Description
J	475	5.85	Translation
	0	0	RNA processing and modification
K	400	4.93	Transcription
L	215	2.65	Replication, recombination and repair
B	2	0.02	Chromatin structure and dynamics
D	102	1.25	Cell cycle control, mitosis and meiosis
Y	0	0	Nuclear structure
V	130	1.60	Defense mechanisms
T	288	3.55	Signal transduction mechanisms
M	260	3.20	Cell wall/membrane biogenesis
N	118	1.45	Cell motility
Z	0	0	Cytoskeleton
W	15	0.18	Extracellular structures
U	66	0.81	Intracellular trafficking and secretion
O	234	2.88	Posttranslational modification, protein turnover, chaperones
X	56	0.69	Mobilome: prophages, transposons
C	358	4.41	Energy production and conversion
G	431	5.31	Carbohydrate transport and metabolism
E	571	7.04	Amino acid transport and metabolism
F	208	2.56	Nucleotide transport and metabolism
H	318	3.92	Coenzyme transport and metabolism
I	333	4.10	Lipid transport and metabolism
P	323	3.98	Inorganic ion transport and metabolism
Q	176	2.17	Secondary metabolites biosynthesis, transport and catabolism
R	560	6.90	General function prediction only
S	403	4.96	Function unknown
—	2832	34.92	Not in COGs

COGs, Clusters of Orthologous Groups database.

cellobiose, inulin, glycogen, xylitol, gentiobiose, D-turanose, D-lyxose, D-tagatose, D-fucose, L-fucose, D-arabitol, L-arabitol, potassium gluconate, potassium 2-ketogluconate and potassium 5-ketogluconate.

Using API 20NE, positive reactions were obtained for esculin ferric citrate, potassium nitrate, L-tryptophane, D-glucose (fermentation), L-arginine and urea. Glucose was assimilated.

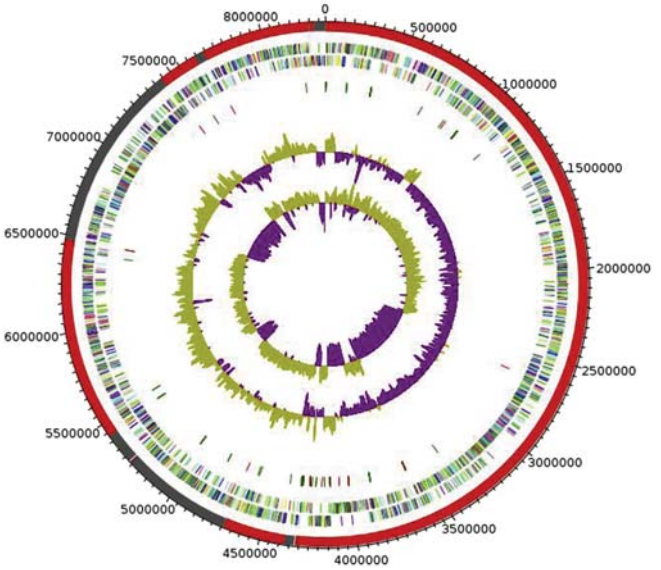


FIG. 6. Circular map of *Bacillus salis* strain ES3^T chromosome. From outside to centre: outer two circles show open reading frames oriented in forward (coloured by COGs categories) and reverse (coloured by COGs categories) directions, respectively. Third circle marks tRNA genes (green). Fourth circle shows G+C% content plot. Innermost circle shows GC skew, with purple indicating negative values and olive positive values. COGs, Clusters of Orthologous Groups database.

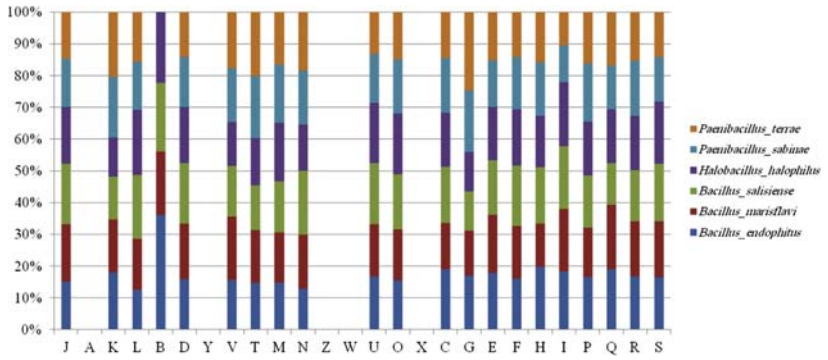


FIG. 7. Distribution of functional classes of predicted genes according to Clusters of Orthologous Groups of proteins.

Nitrophenyl-βD-galactopyranoside, L-arabinose, D-mannose, D-mannitol, *N*-acetyl-glucosamine, D-maltose, potassium gluconate, capric acid, adipic acid, malic acid, trisodium citrate and phenylacetic acid were not assimilated.

When assayed with the API ZYM system, alkaline phosphatase, esterase (C4), esterase lipase (C8), acid phosphatase and naphthol-AS-BI-phosphohydrolase had an enzymatic activity, but lipase (C14), leucine arylamidase, valine arylamidase, cystine arylamidase, trypsin, α-chymotrypsin, α-galactosidase, β-galactosidase, β-glucuronidase, α-glucosidase, β-glucosidase, *N*-acetyl-β-glucosaminidase, α-mannosidase and α-fucosidase had no activity. Table 2 compares these features with closely related species.

Antibiotic susceptibility test. Cells were resistant to metronidazole but susceptible to imipenem, doxycycline, rifampicin, vancomycin, amoxicillin, ceftriaxone, gentamicin, trimethoprim/sulfamethoxazole, erythromycin, ciprofloxacin, nitrofurantoin, ampicillin and penicillin.

Fatty acids analysis. The major fatty acids found for this strain were branched: 12-methyl-tetradecanoic acid (60%), 14-methyl-hexadecanoic acid (17%) and 13-methyl-tetradecanoic acid (10%). The most abundant fatty acids were saturated ones (99%) (Table 3).

Genome properties

The draft genome of strain ES3^T is 8 329 771 bp long with 39.19% G+C content (Table 4, Fig. 6). It is composed of 18 scaffolds with 29 contigs. Of the 8303 predicted genes, 8109 were protein-coding genes and 194 were RNAs (20 genes 5S rRNA, two genes 16S rRNA, two genes 23S rRNA and 170 genes tRNA). A total of 5778 genes (71.25%) were assigned a putative function (by COGs or by NR BLAST). A total of 180 genes (2.22%) were identified as ORFans. The remaining genes were annotated as hypothetical proteins (1748 genes, 21.569%). Table 4 summarizes the genome's properties. Table 5 presents the distribution of genes into COGs functional categories.

Genome comparison

We compared the genome sequence of strain ES3^T (accession no. FNMN00000000) with that of halophilic bacteria close to our strain: *Halobacillus halophilus* strain DSM 2266 (HE717023), *Bacillus endophyticus* Hbe603 (NZ_CP011974), *Bacillus marisflavi* JCM 11544 (LGUE00000000), *Paenibacillus sabinae* T27 (CP004078) and *Paenibacillus terrae* HPL-003 (CP003107). The draft genome of strain ES3^T (8.32 Mb) was larger than that of *B. endophyticus, B. marisflavi, H. halophilus, P. sabinae* and *P. terrae* (4.86, 4.31, 4.17, 5.27 and 6.08 Mb respectively). Its G+C content (39.19%) was smaller than that of *B. marisflavi, H. halophilus, P. sabinae* and *P. terrae* (48.60, 41.82, 52.6 and 46.80% respectively) but larger than that of *B. endophyticus* (36.60%). The gene content of strain ES3^T (8303) was larger than that of *B. endophyticus, B. marisflavi, H. halophilus, P. sabinae* and *P. terrae* (4816, 4319, 4857 and 5396 respectively). However, the distribution of genes into COGs categories was similar in all compared genomes (Fig. 7). In addition, strain ES3^T shared more orthologous genes with species belonging to the same genus (*B. endophyticus, B. marisflavi,* 1153 and 1151 genes respectively) than with other species belonging to other genus (*H. halophilus, P. sabinae* and *P. terrae* respectively shared 997, 701 and 725 orthologous genes) (Table 6). The average percentage of nucleotide sequence identity ranged from 65.34% to 65.84% at the intraspecies level between strain ES3^T and the

TABLE 6. Number of orthologous proteins shared between genomes (upper right) and AGIOS values obtained (lower left)

	BS	BE	BM	PS	PT	HH
BS	**8118**	1153	1151	701	725	997
BE	65.34%	**4846**	1036	657	717	818
BM	65.84%	62.01%	**4356**	639	678	822
PS	57.74%	57.64%	60.32%	**4866**	735	518
PT	60.05%	60.41%	60.35%	67.59%	**5446**	528
HH	66.03%	62.50%	61.65%	57.85%	59.29%	**4055**

The bold represents the total number of orthologous proteins for each species. AGIOS, average genomic identity of orthologous gene sequences; BE, *Bacillus endophyticus* strain Hbe603; BM, *Bacillus marisflavi* strain JCM 11544; BS, *Bacillus salis* strain ES3^T; HH, *Halobacillus halophilus* strain DSM 2266; PS, *Paenibacillus sabinae* strain T27^T; PT, *Paenibacillus terrae* strain HPL-003.

TABLE 7. Pairwise comparison of strain ES3^T with other species using GGDC, formula 2 (DDH estimates based on identities/HSP length)

	BE	BM	PS	PT	HH
BS	23.20 ± 2.38%	19.0 ± 2.30%	30.50 ± 2.45%	22.00 ± 2.39%	20.40 ± 2.32%
BE		26.50 ± 2.42%	29.20 ± 2.44%	28.50 ± 2.44%	29.80 ± 2.45%
BM			28.90 ± 2.44%	28.50 ± 2.44%	22.70 ± 2.37%
PS				26.00 ± 2.41%	29.40 ± 2.44%
PT					28.70 ± 2.44%

Confidence intervals indicate inherent uncertainty in estimating DDH values from intergenomic distances based on models derived from empirical test data sets (which are always limited in size). These results are in accordance with 16S rRNA (Fig. 1) and phylogenomic analyses as well as GGDC results.
BE, *Bacillus endophyticus* strain Hbe603; BM, *Bacillus marisflavi* strain JCM 11544; BS, *Bacillus salis* strain ES3^T; DDH, DNA-DNA hybridization; GGDC, Genome-to-Genome Distance Calculator; HH, *Halobacillus halophilus* strain DSM 2266; HSP, high-scoring segment pairs; PS, *Paenibacillus sabinae* strain T27; PT, *Paenibacillus terrae* strain HPL-003.

two *Bacillus* species, but it ranged from 57.74% to 60.05% between strain ES3^T and the two other *Paenibacillus* species. Similar results were obtained for the analysis of DDH using Genome-to-Genome Distance Calculator (GGDC) software (Table 7).

Conclusion

On the basis of the phenotypic properties (Table 2), phylogenetic tree (Fig. 1), MALDI-TOF MS analyses (Fig. 3), genomic comparison via taxonogenomics (Tables 6 and 7) and GGDC results, we propose the creation of *Bacillus salis* sp. nov., represented by the type strain ES3^T.

Description of *Bacillus salis* sp. nov.

***Bacillus salis* (sa'lis, L. gen. n., *salis*, from 'salt,' in which the strain was first identified)**

Colonies which grew after 48 hours' incubation at 37°C on our homemade culture medium were creamy, smooth, circular and slightly irregular, and measured 5 to 8 mm in diameter. Cells were Gram-positive rods and had a mean diameter of 1.8 µm and a length of 5.9 µm. The strain was able to form subterminal ellipsoidal spores and was motile with a single polar flagella. Growth occurred optimally at 37°C, pH 7.5 and 10% NaCl.

API 50CH strip testing showed positive reactions for D-glucose, D-fructose, D-mannose, arbutin, esculin ferric citrate, salicin, D-maltose, D-saccharose, D-trehalose, melezitose, D-raffinose and amidon. Negative reactions were recorded for glycerol, erythritol, D-arabinose, L-arabinose, D-ribose, D-xylose, L-xylose, D-adonitol, methyl-βD-xylopyranoside, D-galactose, L-sorbose, L-rhamnose, dulcitol, inositol, D-mannitol, D-sorbitol, methyl-αD-mannopyranoside, methyl-αD-glucopyranoside, *N*-acetyl-glucosamine, D-cellobiose, inulin, glycogen, xylitol, gentiobiose, D-turanose, D-lyxose, D-tagatose, D-fucose, L-fucose, D-arabitol, L-arabitol, potassium gluconate, potassium 2-ketogluconate and potassium 5-ketogluconate, potassium gluconate, potassium 2-ketogluconate and potassium 5-ketogluconate.

API 20NE testing showed positive reactions for esculin ferric citrate, potassium nitrate, L-tryptophane, D-glucose (fermentation), L-arginine and urea. Glucose was assimilated. Nitrophenyl-βD-galactopyranoside, L-arabinose, D-mannose, D-mannitol, *N*-acetyl-glucosamine, D-maltose, potassium gluconate, capric acid, adipic acid, malic acid, trisodium citrate and phenylacetic acid were not assimilated.

When assayed with the API ZYM system, alkaline phosphatase, esterase (C4), esterase lipase (C8), acid phosphatase and naphthol-AS-BI-phosphohydrolase had an enzymatic activity, but lipase (C14), leucine arylamidase, valine arylamidase, cystine arylamidase, trypsin, α-chymotrypsin, α-galactosidase, β-galactosidase, β-glucuronidase, α-glucosidase, β-glucosidase, *N*-acetyl-β-glucosaminidase, α-mannosidase and α-fucosidase had no activity.

The type strain was sensitive to imipenem, doxycycline, rifampicin, vancomycin, amoxicillin, ceftriaxone, gentamicin (500 µg), trimethoprim/sulfamethoxazole, erythromycin, ciprofloxacin, nitrofurantoin, ampicillin, penicillin and gentamicin (15 µg) but resistant to metronidazole (500 µg).

The major fatty acids found for this strain were branched: 12-methyl-tetradecanoic acid (60%), 14-methyl-hexadecanoic acid (17%) and 13-methyl-tetradecanoic acid (10%). The most abundant fatty acids were saturated ones (99%). The G+C content of the genome was 39.19%. The 16S rRNA gene sequence and whole-genome shotgun sequence of *B. salis* strain ES3^T were deposited in GenBank under accession numbers LN827530 and FNMN00000000, respectively. The type strain of *Bacillus salis* is strain ES3^T (= CSUR P1478 = DSM 100598) and was isolated from salt.

Acknowledgements

The authors thank the Xegen Company (www.xegen.fr) for automating the genomic annotation process, and M. Lardière for English-language editorial work. This study was funded by the Fondation Méditerranée Infection.

Conflict of interest

None declared.

References

[1] Kivistö AT, Karp MT. Halophilic anaerobic fermentative bacteria. J Biotechnol 2011;152:114–24.
[2] Lee HS. Diversity of halophilic archaea in fermented foods and human intestines and their application. J Microbiol Biotechnol 2013;23:1645–53.
[3] Diop A, Khelaifia S, Armstrong N, Labas N, Fournier P-E, Raoult D, et al. Microbial culturomics unravels the halophilic microbiota repertoire of table salt: description of *Gracilibacillus massiliensis* sp. nov. Microb Ecol Health Dis 2016;27:32049.
[4] Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of 'unculturable' bacteria. FEMS Microbiol Lett 2010;309:1–7.
[5] Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, et al. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 2012;18:1185–93.
[6] Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 2014;64:384–91.
[7] Auch AF, von Jan M, Klenk HP, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2010;2:117–34.
[8] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalog established by metagenomic sequencing. Nature 2010;464:59–65.
[9] Seng P, Abat C, Rolain JM, Colson P, Lagier JC, Gouriet F, et al. Identification of rare pathogenic bacteria in a clinical microbiology laboratory: impact of matrix-assisted laser desorption ionization–time of flight mass spectrometry. J Clin Microbiol 2013;51:2182–94.
[10] Bouvet P, Ferraris L, Dauphin B, Popoffa M-R, Butelb MJ, Julio Aires J. 16S rRNA gene sequencing, multilocus sequence analysis, and mass spectrometry identification of the proposed new species *'Clostridium neonatale'*. J Clin Microbiol 2014;52:4129–36.
[11] Lo CI, Fall B, Ba S, Diawara S, Gueye MW, Mediannikov O, et al. MALDI-TOF mass spectrometry: a powerful tool for clinical microbiology at Hôpital principal de Dakar, Senegal (West Africa). PLoS One 2015;10:e0145889.
[12] Weisburg WG, Barns SM, Pelletier DA, Lane DJ. 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 1991;173:697–703.
[13] Drancourt M, Bollet C, Carlioz A, Martelin R, Gayral JP, Raoult D, et al. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. J Clin Microbiol 2000;38:3623–30.
[14] Morel AS, Dubourg G, Prudent E, Edouard S, Gouriet F, Casalta J-P, et al. Complementarity between targeted real-time specific PCR and conventional broad-range 16S rDNA PCR in the syndrome-driven diagnosis of infectious diseases. Eur J Clin Microbiol Infect Dis 2015;34:561–70.
[15] Tindall BJ. The designated type strain of *Pseudomonas halophila* Fendrich 1989 is DSM 3051, the designated type strain of *Halovibrio variabilis* Fendrich 1989 is DSM 3050, the new name *Halomonas utahensis* (Fendrich 1989) Sorokin and Tindall 2006 is created for the species represented by DSM 3051 when treated as a member of the genus *Halomonas*, the combination *Halomonas variabilis* (Fendrich 1989) Dobson and Franzmann 1996 is rejected, and the combination *Halovibrio denitrificans* Sorokin et al. 2006 is validly published with an emendation of the description of the genus *Halovibrio* Fendrich 1989 emend. Sorokin et al. 2006. Opinion 93. Judicial Commission of the International Committee on Systematics of Prokaryotes. Int J Syst Evol Microbiol 2014;64:3588–9.
[16] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 2013;30:2725–9.
[17] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.
[18] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.
[19] Matuschek E, Brown DFJ, Kahlmeter G. Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. Clin Microbiol Infect 2014;20:O255–66.
[20] Sasser M. Bacterial identification by gas chromatographic analysis of fatty acids methyl esters (GC-FAME). Newark, NY: Microbial ID; 2006.
[21] Dione N, Sankar SA, Lagier JC, Khelaifia S, Michele C, Armstrong N, et al. Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. New Microbe New Infect 2016;11(10):66–76.
[22] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821–9.
[23] Bankevich A, Nurk S, Antipov D, Edouard S, Gouriet F, Casalta JP, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77.
[24] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012;1:18.
[25] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.
[26] Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 2010;11:1.
[27] Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997;25:955–64.
[28] Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35:3100–8.
[29] Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338:1027–36.
[30] Gouret P, Paganini J, Dainat J, Louati D, Darbo E, Pontarotti P, et al. Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the multi-agent software system DAGOBAH. In: Pontarotti P, editor. Evolutionary biology: concepts, biodiversity, macroevolution and genome evolution. Berlin: Springer Verlag; 2011. p. 71–87.
[31] Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG, et al. FIGENIX: intelligent automation of genomic annotation:

expertise integration in a new software platform. BMC Bioinform 2005;6:198.

[32] Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 2009;25:119–20.

[33] Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 2004;14:1394–403.

[34] Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ, et al. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinform 2011;12:124.

[35] Coorevits A, Logan NA, Dinsdale AE, Halket G, Scheldeman P, Heyndrickx M, et al. *Bacillus thermolactis* sp. nov., isolated from dairy farms, and emended description of *Bacillus thermoamylovorans*. Int J Syst Evol Microbiol 2011;61:1954–61.

[36] Yoon JH, Kim IG, Kang KH, Oh TK, Park YH. *Bacillus marisflavi* sp. nov. and *Bacillus aquimaris* sp. nov., isolated from sea water of a tidal flat of the Yellow Sea in Korea. Int J Syst Evol Microbiol 2003;53:1297–303.

[37] Reva ON, Smirnov VV, Pettersson B, Priest FG. *Bacillus endophyticus* sp. nov., isolated from the inner tissues of cotton plants (*Gossypium* sp.). Int J Syst Evol Microbiol 2002;52:101–7.

[38] Spring S, Lidwing W, Marquez MC, Ventosa A, Schleifer K-H. *Halobacillus* gen. nov., with descriptions of *Halobacillus litoralis* sp. nov., and *Halobacillus trueperi* sp. nov., and transfer of *Sporosarcina halophila* to *Halobacillus halophilus* comb. nov. Int J Syst Evol Microbiol 1996;46:492–6.

[39] Yoon JH, Oh HM, Yoon BD, Kang KH, Park YH. *Paenibacillus kribbensis* sp. nov. and *Paenibacillus terrae* sp. nov., bioflocculants for efficient harvesting of algal cells. Int J Syst Evol Microbiol 2003;53:295–301.

[40] Ma Y, Xia Z, Liu X, Chen S. *Paenibacillus sabinae* sp. nov., a nitrogen-fixing species isolated from the rhizosphere soils of shrubs. Int J Syst Evol Microbiol 2007;57:6–11.

Nouvelles espèces bactériennes du microbiome vaginal

Article 9:

Description of *Collinsella vaginalis* sp. nov. strain Marseille-P2666, a new member of the *Collinsella* genus isolated from genital tract of a patient suffering from bacterial vaginosis

Diop A, Diop Kh, Tomei E, Bretelle F, Raoult D, Fenollar F, Fournier PE

[Submitted in International Journal of Systematic and Evolutionary Microbiology]

Collinsella vaginalis sp. nov. strain Marseille-P2666^{T}, a new member of the Collinsella genus isolated from genital tract of a patient suffering from bacterial vaginosis.

Awa Diop[1], Khoudia Diop[1], Enora Tomei[1], Nicholas Armstrong[1], Florence Bretelle[1, 3], Didier Raoult[2,4], Florence Fenollar[1], Pierre-Edouard Fournier[1*]

[1]UMR VITROME, Aix-Marseille Université, IRD, Service de Santé des Armées, Assistance Publique-Hôpitaux de Marseille, Institut hospitalo-universitaire Méditerranée-infection, 19-21 Boulevard Jean Moulin 13005 Marseille, France Tel: +33 413 732 401, Fax: +33 413 732 402

[2]UMR MEPHI, Aix-Marseille University, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Uuniversitaire Méditerranée Infection, Marseille, France

[3]Department of Gynecology and Obstetrics, Gynépole, Marseille, Hôpital Nord, Assistance Publique-Hôpitaux de Marseille

[4]Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

*Corresponding author: Pr Pierre-Edouard Fournier

[1]UMR VITROME, Aix-Marseille Université, IRD, Service de Santé des Armées, Assistance Publique-Hôpitaux de Marseille, Institut hospitalo-universitaire Méditerranée-infection, 19-21 Boulevard Jean Moulin 13005 Marseille, France Tel: +33 413 732 401, Fax: +33 413 732 402

E-mail: pierre-edouard.fournier@univ-amu.fr

Keywords: *Collinsella vaginalis*; bacterial vaginosis; microbial culturomics; taxono-genomics; anaerobic bacteria; new species

ABSTRACT

A strictly anaerobic, Gram-stain-positive, non motile and non-spore-forming rod-shaped bacterium, strain Marseille-P2666, was isolated from a vaginal sample of a French patient suffering from bacterial vaginosis using the culturomics approach. Cells were saccharolytic and were negative for catalase, oxidase, urease, nitrate reduction, indole production, hydrolysis of aesculin and gelatin. Strain Marseille-P2666^{T} exhibited 97.04% 16S rRNA sequence similarity with *Collinsella tanakaei* type strain YIT 12063^{T}, the phylogenetically closest species with standing in nomenclature. The major fatty acids were $C_{18:1\omega9}$ (38%), $C_{16:0}$ (24%) and $C_{18:0}$ (19%). The G+C content of the genome sequence of strain Marseille-P2666 is 64.6 mol%. On the basis of its phenotypic, phylogenetic and genomic features, strain Marseille-P2666^{T} (= CSUR 2666^{T} = DSM103342^{T}) was classified as type strain of a novel species within the genus *Collinsella* for which the name *Collinsella vaginalis* sp. nov. is proposed.

Investigating the microbial diversity of bacterial vaginosis is part of the ongoing "Microbial Culturomics" project in our institute [1, 2], which consists in optimizing culture conditions to explore in depth the human microbiota. In 2015, we isolated a strictly anaerobic bacterial strain, strain Marseille-P2666^T, from a vaginal sample of a French woman patient suffering with Bacterial vaginosis (BV). Strain Marseille-P2666^T was classified as belonging to the genus *Collinsella*.

The genus C*ollinsella*, belonging to the family *Coriobacteriaceae* in the phylum Actinobacteria [3], was first described by Kageyama *et al.* in 1999 [4]. On the basis of 16S rRNA gene sequence and cell wall peptidoglycan divergence with other members of the genus *Eubacterium*, these authors reclassified *Eubacterium aerofaciens* into a the new genus *Collinsella*, with *Collinsella aerofaciens* being the type species [4]. Currently, five C*ollinsella* species have standing in nomenclature (www.bacterio.net), namely *C. aerofaciens* [4], *C. stercoris* [5], *C. intestinalis* [5], *C. tanakaei* [6] and *C. massiliensis* [7], all of which had been isolated from the gastro-intestinal tract of healthy humans. All five species are non spore-forming, non motile, rod-shaped cocci and contain an A4P-type peptidoglycan [4].

Thanks to the availability of genomic data from many bacterial species, we proposed since 2012 to include the complete genome sequence analysis in a polyphasic approach for the classification and description of new bacterial taxa, that we named named taxono-genomics [8]. On the basis of the analysis of phenotypic and phylogenetic characteristics, proteomic informations obtained by MALDI-TOF MS and genomics properties [8–10], we describe here a new *Collinsella* species for which we propose the name *Collinsella vaginalis* sp. nov.. Strain Marseille-P2666^T (= CSUR 2666^T = DSM103342^T) is the type strain of *C. vaginalis* sp. nov.

Strain Marseille-P2666 was isolated in May 2015 from a vaginal sample of a 26 year-old French woman diagnosed with bacterial vaginosis at the Nord hospital in Marseille, France. The sample was collected using a Sigma Transwab (Medical Wire, Corsham, United Kingdom) and then transported immediately to the microbiology laboratory of the Timone Hospital in Marseille. The patient was not treated with any antibiotic at the time of sampling. She gave an informed and signed consent and the study was validated by the ethics committee of the IFR48 (Marseille, France) under agreement 09-022. For strain isolation, the vaginal sample was first inoculated in an anaerobic blood culture bottle (Bactec Lytic/10 Anaerobic/F Culture Vials, Becton-Dickinson, Le Pont de Claix, Isère, France) supplemented with 4 mL filter-sterilized rumen fluid through a 0.2 µm pore filter (Thermo Fisher Scientific, Villebon-sur-Yvette, France) and 3 mL of sheep blood (bioMérieux, Marcy l'Etoile, France) and incubated at 37°C. After 72 hours of incubation, 50 µL of the supernatant was inoculated on 5% sheep blood-enriched CNA agar (Colistin and Naladixic Acid) (Becton-Dickinson) and incubated for 48 hours in anaerobic atmosphere (0% O2, 100% CO2 and 100% N2) at 37°C.

Isolated colonies were subcultured individually using the same conditions and each colony was deposited on a MTP 96 MALDI-TOF target plate (Bruker Daltonics, Leipzig, Germany) in duplicate for identification with a Microflex MALDI-TOF MS spectrometer (Bruker Daltonics, Leipzig, Germany), as described by Seng *et al*. [11]. The obtained protein spectra were compared with those of 8687 reference spectra in the Bruker database constantly enriched with our own database [12]. If the MALDI-TOF MS score was greater than 1.9 and 2.3, the bacterium was identified at the genus and species levels respectively. Conversely, if the score was lower than this threshold, the identification was not considered as reliable and the 16S rRNA gene was amplified and sequenced using the GeneAmp PCR System 2720 thermal cycler (Applied Bio systems, Bedford, MA, USA) and an ABI Prism 3130-XL capillary sequencer (Applied Biosciences, Saint Aubin, France), respectively, as previously

described [13]. The obtained sequence was corrected using the Chromas Pro 1.34 software (Technelysium Pty. Ltd., Tewantin, Australia) and then compared to the NCBI database using the BLASTn algorithm (https://blast.ncbi.nlm.nih.gov/) for taxonomic assignment. The 16S rRNA sequences of type strains from the species with a validly published name (http://www.bacterio.net/) exhibiting the closest phylogenetic relationship with strain Marseille-P2666 were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/Genome/). Sequences were aligned using MUSCLE [14]. Then, the degree of pairwise 16S rRNA sequence similarity between strain Marseille-P2666 and other closely related species were calculated using the GGDC web server [15] available at (http://ggdc.dsmz.de/) using the method proposed by Meier-Kolthoff [16]. Phylogenetic trees were inferred in the GGDC web server [15] using the DSMZ phylogenomics pipeline [17] adapted to single genes. Maximum likelihood (ML) and maximum parsimony (MP)-based trees were inferred from the alignment with RAxML [18] and TNT [19], respectively. For ML, rapid bootstrapping in conjunction with the autoMRE bootstopping criterion [20] and subsequent search for the best tree was used. The ML tree was inferred under the GTR+GAMMA model. For MP tree analysis, all sites with gaps were removed and 1000 bootstrapping replicates were used in conjunction with tree-bisection-and-reconnection branch swapping and ten random sequence addition replicates. The sequences were checked for a compositional bias using the X^2 test as implemented in PAUP* [21]. A supplementary phylogenetic tree using the Neighbor-joining method is presented in supplementary data. If the 16S rRNA sequence similarity value was lower than 95% or 98.65% with the most closely related species with standing in nomenclature, as proposed by Stackebrandt and Ebers [22], the strain was proposed to belong to a new genus or species, respectively [23].

In order to evaluate its ideal growth conditions, strain Marseille-P2666 was cultivated on 5% sheep blood-enriched Columbia agar (bioMérieux) at various temperatures (25, 28, 37,

45, 56°C) under aerobic conditions with or without 5% CO_2, and in anaerobic (0% O2, 100% CO2 and 100% N2) and microaerophilic atmospheres (5% O 2, 10% CO 2 and 85% N 2) using GENbag Anaer and GENbag microaer systems (bioMérieux) respectively. The tolerance to various NaCl concentrations (5 – 100 g/l NaCl) and pH values (pH 5, 6, 6.5, 7, 8.5) conditions was also tested. To observe the cell morphology, cells were fixed with 2.5% glutaraldehyde in a 0.1M cacodylate buffer at 4°C for at least an hour. One drop of cell suspension was deposited for approximately five minutes on glow-discharged formvar carbon film on 400 mesh nickel grids (FCF400-Ni, EMS). The grids were dried on blotting paper and the cells were negatively stained for 10 seconds with 1% ammonium molybdate solution in filtered water at RT. Electron micrographs were acquired using a Tecnai G20 Cryo (FEI company, Limeil-Brevannes, France) transmission electron microscope operated at 200 keV. Gram-stain, motility and sporulation were performed as previously described [24].

The biochemical properties of strain Marseille-P2666 were evaluated using API ZYM, API 20A, and API rapid ID 32A strips (bioMérieux) according to the manufacturer's instructions. The strips were incubated in anaerobic conditions (0% O2, 100% CO2 and 100% N2) at 37°C for 4, 24, and 4 hours respectively. Oxidase activity was tested using an oxidase reagent (Becton-Dickenson, Le Pont de Claix, and France) and catalase activity was assessed in 3% hydrogen peroxide solution (bioMérieux).

Amoxicillin (0.016-256 µg/mL), benzylpenicillin (0.002-32 µg/mL), ceftriaxone (0.016-256 µg/mL), vancomycin (0.016-256 µg/mL), metronidazole (0.016-256 µg/mL), rifampicin (0.002-32 µg/mL) and imipenem (0.002-32 µg/mL) were used to test the antibiotic susceptibility of strain Marseille-P2666. The minimal inhibitory concentrations (MICs) were then determined using E-test gradient strips (bioMérieux) according to the EUCAST recommendations [25, 26].

Cellular fatty acid methyl ester (FAME) analysis was performed using Gas Chromatography/Mass Spectrometry (GC/MS). Strain Marseille-P2666 was grown on 5% sheep blood-enriched Columbia agar (bioMérieux). Two samples were then prepared with approximately 16 mg of bacterial biomass per tube harvested from several culture plates. Fatty acid methyl esters were prepared as described by Sasser [27]. GC/MS analyses were carried out as described before [28]. Briefly, fatty acid methyl esters were separated using an Elite 5-MS column and monitored by mass spectrometry (Clarus 500 - SQ 8 S, Perkin Elmer, Courtaboeuf, France). Spectral database search was performed using MS Search 2.0 operated with the Standard Reference Database 1A (NIST, Gaithersburg, USA) and the FAMEs mass spectral database (Wiley, Chichester, UK).

The genomic DNA (gDNA) of the strain Marseille-P2666^{T} was sequenced using a MiSeq sequencer (Illumina Inc, San Diego, CA, USA) with the Mate Pair strategy. The gDNA was quantified by a Qubit assay with the high sensitivity kit (Life technologies, Carlsbad, CA, USA) to 68.1 ng/µl and a total of sequencing output of 5.1 Gb was obtained from a 542K/mm^2 cluster density with a cluster passing quality control filters of 95.7% (10,171,000 clusters). The 801,260 reads obtained by sequencing were trimmed, then assembled using the Spades assembler program [29]. A more detailed description of the sequencing methodology as well as the complete annotation of the genome is presented in the supplementary data section.

A MALDI-TOF-MS score of 1.3 was obtained for strain Marseille-P2666 against our database, suggesting that this isolate was not identified in the genus and species levels. The MALDI-TOF MS spectrum from strain Marseille-P2666 was added to our database to improve its content.

Using the Smith–Waterman algorithm [16], the 16S rDNA-based comparison of strain Marseille-P2666 (EMBL-EBI accession number LT598547) against GenBank yielded a

highest nucleotide sequence similarity of 97.04% with *C. tanakaei* strain YIT 12063^T (GenBank accession number AB490807), the phylogenetically-closest species with a validly published name. As this value was lower than the 98.65% 16S rRNA sequence identity threshold proposed to delineate a new species [22, 30], strain Marseille-P2666 was considered as a potential new species within the genus *Collinsella* in the family *Coriobacteriaceae*. The resulting combined ML/MP tree and the Neighbor-joining tree highlighting the position of *Collinsella vaginalis* strain Marseille-P2666 relative to other close strains with a validly published name is shown in Figure 1 and Figure 2.

For the phylogenetic inferences, the input nucleotide matrix comprised 21 operational taxonomic units and 1,572 characters, 500 of which were variable and 351 of which were parsimony-informative. The base-frequency check indicated a compositional bias ($p = 0.00$, $\alpha = 0.05$). ML analysis under the GTR+GAMMA model yielded a highest log likelihood of -8308.08, whereas the estimated alpha parameter was 0.20. The ML bootstrapping did not converge, hence 1,000 replicates were performed; the average support was 72.67%. MP analysis yielded a best score of 1315 (consistency index 0.57, retention index 0.66) and 6 best trees. The MP bootstrapping average support was 77.17%.

Colonies from strain Marseille-P2666 on CNA agar (Becton-Dickinson, Le pont de Claix, France) under anaerobic atmosphere are grey, opaque and circular with a diameter of 0.5-1.2 mm after 48 hours of growth at 37°C. The growth was obtained at temperatures ranging from 28 to 45 with optimal growth observed at 37°C in anaerobic atmosphere. No growth was obtained in neither aerobic nor microaerophilic atmospheres. Strain Marseille-P2666 needed a NaCl concentration below 5g/L and a pH ranging from 6.5 to 7.0 for its growth. Bacterial cells are rod-shaped Gram-stain-positive, non-motile and non spore-forming with a mean diameter of 0.4 µm and mean length of 1.8 µm and occur as single cells or in short chains. No oxidase or catalase activity was observed.

Using an API ZYM strip (bioMérieux), positive results were obtained for esterase (C4), esterase lipase (C8), alkaline phosphatase, leucine arylamidase, valine arylamidase, cystine arylamidase, acid phosphatase, naphtol-AS-BI-phosphohydrolase and N-acetyl-β-glucosaminidase but no reaction was observed for lipase (14), trypsin, α-chymotrypsin, α-galactosidase, β-galactosidase, β-glucuronidase, α-glucosidase, β-glucosidase, α-mannosidase and α-fucosidase. Using a Rapid ID32A strip (bioMérieux), positive reactions were obtained for N-Acetyl-ß-glucosaminidase, mannose fermentation, raffinose fermentation, alkaline phosphatase, arginine arylamidase, proline arylamidase, leucyl glycine arylamidase, leucine arylamidase, glycine arylamidase, histidine arylamidase and serine arylamidase. Cells showed no urease, arginine dihydrolase, α-galactosidase, β-galactosidase, 6-phospho-β-galactosidase, α-glucosidase, β-glucosidase, α-arabinosidase, β-glucuronidase, glutamic acid decarboxylase, α-fucosidase, reduction of nitrates, indole production, phenylalanine arylamidase, pyroglutamic acid arylamidase, tyrosine arylamidase and glutamyl-glutamic acid arylamidase activity. Using an API 20A strip (bioMérieux), strain Marseille-P2666 produced acid from D-glucose, D-lactose, D-saccharose, D-maltose, salicin, D-cellobiose, D-mannose and D-trehalose but not from D-mannitol, D-xylose, L-arabinose, gelatin, glycerol, D-melezitose, D-raffinose, sorbitol and D-rhamnose. Esculin ferric citrate was not hydrolyzed. Indole formation and urease activity were negative. Strain Marseille-P2666 differed from other members of the *Collinsella* genus [4–7] in esterase, esterase lipase and cystine arylamidase activities (Table 1). The most abundant cellular fatty acid found for strain Marseille-P2666 was the unsaturated acid $C_{18:1\omega9}$ (38%), followed by the saturated acids $C_{16:0}$ and $C_{18:0}$ (24 and 19%, respectively) (Table 2). Cells are susceptible to benzylpenicillin (MIC 0.38 µg/mL), amoxicillin (MIC 0.064 µg/mL), metronidazole (MIC 0.75 µg/mL), rifampicin (MIC 0.008 µg/mL), vancomycin (MIC 4 µg/mL) but resistant to ceftriaxone (MIC > 256 µg/mL) and imipenem (MIC > 32 µg/mL).

The draft genome of strain Marseille-P2666 is 2,162,909-bp long and has a G+C content of 64.6 mol% (Table S1, Figure 3). It is composed of 23 scaffolds composed of 63 contigs. Of the 1,907 predicted genes, 1,696 were protein-coding genes and 53 were RNAs (1 complete rRNA operon, 47 tRNA genes and 3 ncRNA genes). A total of 1,303 genes (76.8%) were assigned a putative function (by BLAST against the COGs or NR databases). A total of 121 genes were identified as ORFans (7.1%). The remaining 272 genes were annotated as hypothetical proteins (16.0%). Strain Marseille-P2666 has many genes related to virulence, including 13 bacteriocin-encoding genes (0.8%) and 50 toxin/ antitoxin modules (2.9%). By using PHAST and RAST, 691 genes (40.7%) were associated with mobile genetic elements. Genome statistics are summarized in Table S1 and the gene distribution into COGs functional categories is presented in Table S2.

The draft genome sequence structure of strain Marseille-P2666 is summarized in Figure S1. It is smaller than those of *C. aerofaciens, Collinsella tanakei* and *C. stercoris* (2.2, 2.4, 2.5 and 2.5 Mb, respectively), but larger than those of *C. intestinalis* (1.8 Mb). The G+C content of strain Marseille-P2666 (64.6 %) is greater than those of all compared *Collinsella* species (Table S3). The gene content of strain Marseille-P2666 (1,907) is smaller than those of *C. stercoris, Collinsella tanakei* and *C. aerofaciens* (2,119, 2,253 and 2437, respectively) but larger than those of *C. intestinalis* (1,630) (Table S3). The gene distribution into COG categories was similar among all compared genomes (Figure S2). However, *C. vaginalis* possessed fewer predicted genes of the "Mobilome: prophages, transposons" category than other compared *Collinsella* species (Figure S2). In addition, strain Marseille-P2666 exhibited digital DNA–DNA hybridization (dDDH) values of 22.4% with *C. aerofaciens* to 23.2% with *C. stercoris* (Table S4). Moreover, we observed AAI values of 64.7 to 66.9% between strain Marseille-P2666 and *C. aerofaciens* and *C.intestinalis* or *C. stercoris*, respectively, these

values obtained confirm the affiliation of the genus but also supported the status of new species of strain Marseille-P2666 (Table S5).

The obtained dDDH and AAI values were lower than the 70% and 95-96% threshold values for species demarcation, respectively [15, 31, 32]. Finally, strain Marseille-P2666 exhibited the genomic G+C content differences ranging from -1.3% when compared with *C. massiliensis* to +4.3% with *C. tanakaei*. As previously demonstrated, that the G + C content deviation within species does not exceed 1% [33].

By taking into consideration its phenotypic (Table 1), phylogenetic (Figure 1) and genomic characteristics (Supplementary data) when compared to *Collinsella* species with standing in nomenclature, strain Marseille-P2666 was considered as belonging to a new species within this genus, for which we propose the name *Collinsella vaginalis* sp. nov.

Description of *Collinsella vaginalis* sp. nov.

Collinsella vaginalis (va.gi.na'lis. L. n. fem. *vagina*, sheath, vagina; L. fem. gen. suff. – *alis*, suffix denoting pertaining to; N.L. fem. adj. *vaginalis*, pertaining to the vagina).

Strictly anaerobic, bacterial cells are rod-shaped, Gram-stain-positive, non-motile, non-sporforming, mesothermophilic, oxidase and catalase negative, with a mean diameter and length of 0.4 µm and 1.8 µm, respectively. Cells occur as single rods or in short chains. After two days of incubation at 37°C under anaerobic conditions, colonies on 5% sheep blood-enriched Columbia agar (BioMérieux), appear grey, opaque and circular with a diameter of 0.5-1.2 mm. Nitrate is not reduced; esculin ferric citrate, indole formation, gelatin hydrolysis and urease activities are not detected. Using an API 20A strip (BioMérieux), acid is produced from D-glucose, D-lactose, D-saccharose, D-maltose, salicin, D-cellobiose, D-mannose and D-trehalose but not from D-mannitol, D-xylose, L-arabinose, glycerol, D-melezitose, D-raffinose, sorbitol, D-rhamnose. By using API Rapid ID32A and API ZYM strips (BioMérieux), fermented reactions are observed for mannose and raffinose, N-acetyl-ß-

glucosaminidase, alkaline phosphatase, arginine arylamidase, proline arylamidase, leucyl-glycine arylamidase, leucine arylamidase, glycine arylamidase, histidine arylamidase, serine arylamidase, esterase (4), esterase lipase (8), leucine arylamidase, valine arylamidase, cystine arylamidase, acid phosphatase and naphtol-AS-BI-phosphohydrolase. Arginine dihydrolase, α-galactosidase, β-galactosidase, 6-phospho-β-galactosidase, α-glucosidase, β-glucosidase, α-arabinosidase, β-glucuronidase, glutamic acid decarboxylase, α-fucosidase, phenylalanine arylamidase, pyroglutamic acid arylamidase, tyrosine arylamidase, glutamyl glutamic acid arylamidase, lipase (14), trypsin, α-chymotrypsin and α-mannosidase activities were not detected. The most abundant fatty acids are 9-Octadecenoic acid ($C_{18:1\omega9}$) and Hexadecanoic acid ($C_{16:0}$). *C. vaginalis* was susceptible to benzylpenicillin, amoxicillin, metronidazole, rifampicin, and vancomycin and resistant to ceftriaxone and imipenem.

The type strain Marseille-P2666^T (= CSUR 2666 = DSM103342) was isolated from the vaginal sample of a French woman suffering from bacterial vaginosis. The genome of the type strain is 2,162,909-bp long and exhibits a G+C content of 64.6 mol%. The 16S rRNA and genome sequences are deposited in EMBL-EBI under accession numbers LT598547 and FWYK00000000, respectively.

FUNDING INFEORMATION

This study was funded by the Méditerranée-Infection foundation and the French Agence Nationale de la Recherche under reference Investissements d'Avenir Méditerranée Infection 10-IAHU-03.

CONFLICT OF INTEREST

The authors declare no competing interest in relation to this research.

ACKNOWLEDGEMENTS

Genome assembly was performed by the Xegen company.

References

1. **Lagier J-C, Armougom F, Million M, Hugon P, Pagnier I, *et al.*** Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin Microbiol Infect* 2012;18:1185–1193.

2. **Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, *et al.*** Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat Microbiol* 2016;1:16203.

3. **Stackebrandt E, Rainey FA, Ward-Rainey NL**. Proposal for a new hierarchic classification system, Actinobacteria classis nov. *Int J Syst Evol Microbiol* 1997;47:479–491.

4. **Kageyama A, Benno Y, Nakase T**. Phylogenetic and phenotypic evidence for the transfer of Eubacterium aerofaciens to the genus Collinsella as Collinsella aerofaciens gen. nov., comb. nov. *Int J Syst Evol Microbiol* 1999;49:557–565.

5. **Kageyama A, Benno Y**. Emendation of genus Collinsella and proposal of Collinsella stercoris sp. nov. and Collinsella intestinalis sp. nov. *Int J Syst Evol Microbiol* 2000;50:1767–1774.

6. **Nagai F, Watanabe Y, Morotomi M**. Slackia piriformis sp. nov. and Collinsella tanakaei sp. nov., new members of the family Coriobacteriaceae, isolated from human faeces. *Int J Syst Evol Microbiol* 2010;60:2639–2646.

7. **Padmanabhan R, Dubourg G, Lagier J-C, Nguyen T-T, Couderc C, *et al.*** Non-contiguous finished genome sequence and description of Collinsella massiliensis sp. nov. *Stand Genomic Sci* 2014;9:1144–1158.

8. **Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, *et al.*** A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* 2014;64:384–391.

9. **Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, *et al.*** The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2012;40:D571–D579.

10. **Sentausa E, Fournier P-E**. Advantages and limitations of genomics in prokaryotic taxonomy. *Clin Microbiol Infect* 2013;19:790–795.

11. **Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P, *et al.*** Ongoing Revolution in Bacteriology: Routine Identification of Bacteria by Matrix?Assisted Laser Desorption Ionization Time?of?Flight Mass Spectrometry. *Clin Infect Dis* 2009;49:543–551.

12. **Lagier J-C, Hugon P, Khelaifia S, Fournier P-E, La Scola B, *et al.*** The Rebirth of Culture in Microbiology through the Example of Culturomics To Study Human Gut Microbiota. *Clin Microbiol Rev* 2015;28:237–264.

13. **Drancourt M, Bollet C, Carlioz A, Martelin R, Gayral J-P, *et al.*** 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *J Clin Microbiol* 2000;38:3623–3630.

14. **Edgar RC**. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–1797.

15. **Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M**. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:60.

16. **Meier-Kolthoff JP, G?ker M, Spr?er C, Klenk H-P**. When should a DDH experiment be mandatory in microbial taxonomy? *Arch Microbiol* 2013;195:413–418.

17. **Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, *et al.*** Complete genome sequence of DSM 30083 T, the type strain (U5/41 T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genomic Sci* 2014;9:2.

18. **Stamatakis A**. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.

19. **Goloboff P, Farris J, C. Nixon K**. *TNT, a free program for phylogenetic analysis*. 2008.

20. **Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A**. How Many Bootstrap Replicates Are Necessary? In: Batzoglou S (editor). *Research in Computational Molecular Biology*. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 184–200.

21. **L. Swofford D**. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10*. Sinauer Associates. Sunderland; 2002.

22. **Stackebrandt E, Ebers J**. *Taxonomic parameters revisited: Tarnished gold standards*. 2006.

23. **Konstantinidis KT, Ramette A, Tiedje JM**. The bacterial species definition in the genomic era. *Philos Trans R Soc B Biol Sci* 2006;361:1929–1940.

24. **Diop A, Khelaifia S, Armstrong N, Labas N, Fournier P-E, *et al.*** Microbial culturomics unravels the halophilic microbiota repertoire of table salt: description of *Gracilibacillus massiliensis* sp. nov. *Microb Ecol Health Dis*;27. Epub ahead of print 18 October 2016. DOI: 10.3402/mehd.v27.32049.

25. **Citron DM, Ostovari MI, Karlsson A, Goldstein EJ**. Evaluation of the E test for susceptibility testing of anaerobic bacteria. *J Clin Microbiol* 1991;29:2197–2203.

26. **Matuschek E, Brown DFJ, Kahlmeter G**. Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. *Clin Microbiol Infect* 2014;20:O255–O266.

27. **Sasser M**. Identification of bacteria by gas chromatography of cellular fatty acids. http://natasha.eng.usf.edu/gilbert/courses/Biotransport%20Phenomena/pdf/bacteria_gc_1.pdf (1990, accessed 24 March 2016).

28. **Dione N, Sankar SA, Lagier J-C, Khelaifia S, Michele C, *et al.*** Genome sequence and description of Anaerosalibacter massiliensis sp. nov. *New Microbes New Infect* 2016;10:66–76.

29. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.*** SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 2012;19:455–477.

30. **Kim M, Oh H-S, Park S-C, Chun J**. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.

31. **Konstantinidis KT, Tiedje JM**. Towards a Genome-Based Taxonomy for Prokaryotes. *J Bacteriol* 2005;187:6258–6264.

32. **Rodriguez-R LM, Konstantinidis KT**. Bypassing cultivation to identify bacterial species. *Microbe* 2014;9:111–8.

33. **Klenk H-P, Meier-Kolthoff JP, G?ker M**. Taxonomic use of DNA G+C content and DNA?DNA hybridization in the genomic age. *Int J Syst Evol Microbiol* 2014;64:352–356.

377 **Table 1:** Compared characteristics of *Collinsella vaginalis* strain Marseille-P2666^T and other members of the genus *Collinsella*: *Collinsella*
378 *tanakaei* strain YIT 12063^T [6]; *C. stercoris* strain DSM 13279^T [5]; *C. intestinalis* strain DSM 13280^T [5]; *C. aerofaciens* strain ATCC
379 25986^T [4]; *C. massiliensis* strain GD3^T [7]. +: positive reaction; -: negative reaction; na: no available data.

Properties	*Collinsella vaginalis*	*Collinsella tanakei*	*Collinsella stercoris*	*Collinsella intestinalis*	*Collinsella aerofaciens*	*Collinsella massiliensis*
Cell diameter (µm)	0.3-0.5	0.5-1.0	0.3-0.5	0.3-0.5	0.3-0.7	0.57
Oxygen requirement	Anaerobic	Anaerobic	Anaerobic	Anaerobic	Anaerobic	Anaerobic
Gram stain	+	+	+	+	+	+
DNA G+C content (mol %)	64.6	60.2	63.2	62.5	60.6	65.8
Spore-forming	-	-	-	-	-	-
Motility	-	-	-	-	-	-
Production of						
Alkaline phosphatase	+	+	+	+	-	+
Acid phosphatase	+	+	+	+	-	+
α-galactosidase	-	-	-	-	+	+
β-galactosidase	-	-	+	-	+	+
α-glucosidase	-	-	-	-	+	+
Esterase lipase	+	-	-	-	-	-
N-acetyl-β-glucosaminidase	+	-	+	+	-	-
Cystine arylamidase	+	-	-	-	-	-
Acid form						
Mannose	+	+	+	+	+	-
Glucose	+	+	+	+	+	-
Salicin	+	+	+	-	+	-
Trehalose	+	+	+	-	-	-
Maltose	+	+	+	-	+	-
Lactose	+	+	+	-	+	-
Rhamnose	-	-	-	-	-	-
L-arabinose	-	-	-	-	-	-
Habitat	Human vagina	Human gut	Human gut	Human gut	Human gut	Human gut

380 **Table 2:** Cellular fatty acid composition (%).

Fatty acids	Name	Mean relative % (a)
18:1ω9	9-Octadecenoic acid	37.5 ± 1.0
16:00	Hexadecanoic acid	23.5 ± 0.5
18:00	Octadecanoic acid	18.5 ± 0.4
18:2ω6	9,12-Octadecadienoic acid	11.3 ± 0.3
14:00	Tetradecanoic acid	3.5 ± 0.3
18:1ω5	13-Octadecenoic acid	2.2 ± 0.3
10:00	Decanoic acid	TR
18:1ω7	11-Octadecenoic acid	TR
20:4ω6	5,8,11,14-Eicosatetraenoic acid	TR
17:00	Heptadecanoic acid	TR
17:0 anteiso	14-methyl-Hexadecanoic acid	TR
15:00	Pentadecanoic acid	TR
12:00	Dodecanoic acid	TR
15:0 anteiso	12-methyl-tetradecanoic acid	TR
17:0 iso	15-methyl-Hexadecanoic acid	TR

381 [a] Mean peak area percentage; TR = trace amounts

Figure legends

Figure 1. Maximum likelihood phylogenetic tree inferred under the GTR+GAMMA model and rooted by midpoint-rooting.

The branches are scaled in terms of the expected numbers of substitutions per site. The numbers above the branches are support values when larger than 60% from ML (left) and MP (right) bootstrapping.

Figure 2. Phylogenetic tree based on the 16S rRNA gene highlighting the position of *Collinsella vaginalis* strain Marseille-P2666^{T} relative to other close.

GenBank accession numbers of each 16S rRNA are noted in parenthesis. Sequences were aligned using CLUSTALW 2.0 software with default parameters and phylogenetic inferences were obtained using the neighbor-joining method with 500 bootstrap replicates, within MEGA6 software. The evolutionary distances were computed using the Kimura 2-parameter method and are in the units of the number of base substitutions per site. The scale bar represents a 2% nucleotide sequence divergence.

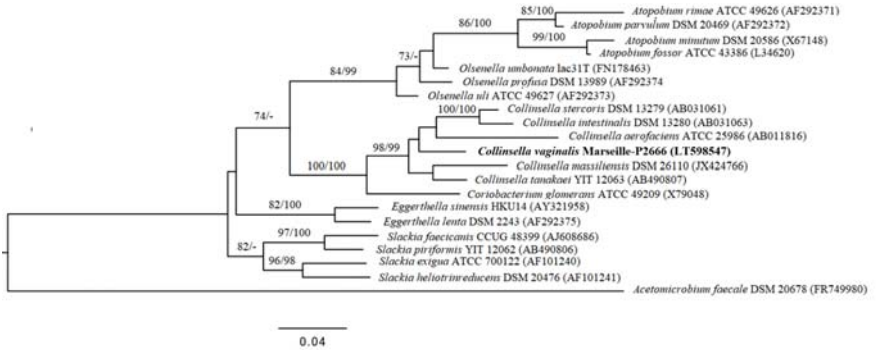


Figure 1. Maximum likelihood phylogenetic tree inferred under the GTR+GAMMA model and rooted by midpoint-rooting.

The branches are scaled in terms of the expected numbers of substitutions per site. The numbers above the branches are support values when larger than 60% from ML (left) and MP (right) bootstrapping.

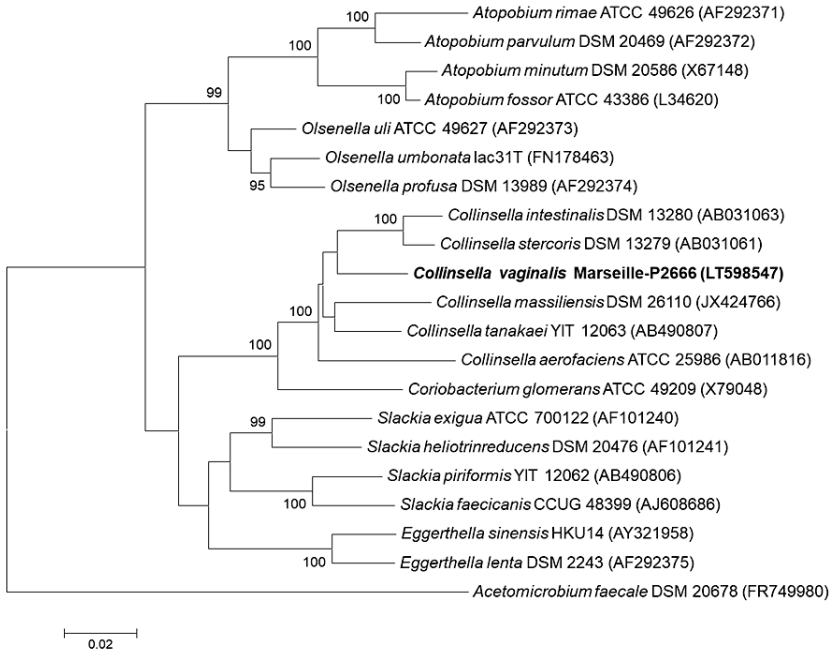


Figure 2. Phylogenetic tree based on the 16S rRNA gene highlighting the position of *Collinsella vaginalis* strain Marseille-P2666^{T} relative to other close.

GenBank accession numbers of each 16S rRNA are noted in parenthesis. Sequences were aligned using CLUSTALW 2.0 software with default parameters and phylogenetic inferences were obtained using the neighbor-joining method with 500 bootstrap replicates, within MEGA6 software. The evolutionary distances were computed using the Kimura 2-parameter method and are in the units of the number of base substitutions per site. The scale bar represents a 2% nucleotide sequence divergence.

SUPPLEMENTARY DATA

411 **Supplementary materials and methods**

412 **16S phylogenetic analysis using Neighbor-joining method.**

413 The 16S sequences of the type strains of the closest species to our new strain in the
414 BLAST search were downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/Genome/).
415 Sequences were aligned using CLUSTALW 2.0 software [1], with default parameters and
416 phylogenetic inferences were obtained using the neighbor-joining method within the MEGA
417 software, version 6 [2]. The evolutionary distances were computed using the Kimura 2-
418 parameter method [3] and the partial deletion option (95%) was used. The bootstraping analysis
419 was performed with 500 replications.

420

421 **DNA Extraction and genome sequencing**

422 After a pretreatment step by lysozyme incubation at 37°C for 2 hours, the Genomic DNA
423 (gDNA) of strain Marseille-P2666^{T} was extracted on the EZ1 biorobot (Qiagen, Hilden,
424 Germany) using the EZ1 DNA tissues kit. The elution volume was 50µL. gDNA was
425 quantified by a Qubit assay with the high sensitivity kit (Life technologies, Carlsbad, CA,
426 USA) to 68.1 ng/µl.

427 The gDNA was sequenced on the MiSeq sequencer (Illumina Inc, San Diego, CA, USA)
428 with the mate pair strategy. The gDNA was barcoded in order to be mixed with 11 other
429 projects using the Nextera Mate Pair sample prep kit (Illumina). The mate pair library was
430 prepared with 1.5 µg of gDNA using the Nextera mate pair Illumina guide. The genomic DNA
431 sample was simultaneously fragmented and tagged with a mate pair junction adapter. The
432 pattern of the fragmentation was validated on an Agilent 2100 BioAnalyzer (Agilent
433 Technologies Inc, Santa Clara, CA, USA) with a DNA 7500 labchip. The DNA fragments

ranged in size from 1.5 kb up to 11 kb with an optimal size at 9.088 kb. No size selection was performed and 600ng of tagmented fragments were circularized. The circularized DNA was mechanically sheared to small fragments with an optimal at 1325 bp on the Covaris device S2 in microtubes (Covaris, Woburn, MA, USA).The library profile was visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) and the final concentration library was measured at 11.99 nmol/l. The libraries were normalized at 2nM and pooled. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded. Automated cluster generation and sequencing run were performed in a single 39-hours run in a 2x151-bp.

A total of sequencing output of 5.1 Gb was obtained from a 542K/mm^2 cluster density with a cluster passing quality control filters of 95.7% (10,171,000 clusters). Within this run, the index representation for *Collinsella vaginalis* was determined to 7.88%. The 801,260 paired end reads were trimmed then assembled.

Genome annotation and comparison

Prodigal was used for Open Reading Frame (ORF) prediction [4] with default parameters. . Predicted ORFs spanning a sequencing gap region were excluded. Bacterial proteome was predicted using BLASTP (E-value 1e^{-03} coverage 0.7 and identity percent 30%) against the Clusters of Orthologous Groups (COG) database. If no hit was found, a search against the nr database [5] was performed using BLASTP with E-value of 1e^{-03}, a coverage of 0.7 and an identity percent of 30 %. If sequence lengths were smaller than 80 amino acids, we used an E-value of 1e^{-05}. Pfam conserved domains (PFAM-A an PFAM-B domains) were searched on each protein with the hhmscan tools analysis [6]. RNAmmer [7] and tRNAScanSE [8] were used to identify ribosomal RNAs and tRNAs, respectively. We predicted lipoprotein signal peptides and the number of transmembrane helices using Phobius [9]. ORFans were identified if the BLASTP search was negative (E-value smaller than 1e^{-03} for ORFs with a

459 sequence size larger than 80 aas or E-value smaller than $1e^{-05}$ for ORFs with sequence length
460 smaller than 80 aas). Artemis [10] and DNA Plotter [11] were used for data management and
461 for visualization of genomic features, respectively. Annotation and comparison processes were
462 performed using the multi-agent software system DAGOBAH [12], which include Figenix [13]
463 libraries that provide pipeline analysis. Genomes from members of the *Coriobacteriaceae*
464 family and closely related to our strain were used for the comparative genomics study.
465 Genomic informations from strain Marseille-P2666 and comparatively closest related species
466 are presented in Table 6. Finally, the average amino acid identity (AAI) was calculated, based
467 on the overall similarity between datasets of proteins of genome pairs belonging to the same
468 genus of *Collinsella* [17] available at (http://enve-omics.ce.gatech.edu/aai/index). We also
469 performed GGDC analysis using the GGDC web server, as previously reported [18].

SUPPLEMENTARY TABLES

Table S1. Nucleotide content and gene count levels of the genome of strain Marseille-P2666^{T}

Attribute	Value	% of total[a]
Size (bp)	2,162,909	100
G+C content (bp)	1,383,290	64.6
Coding region (bp)	1,624,759	75.1
Total genes	1,774	100
RNA genes	50	2.8
Protein-coding genes	1,724	100
Genes with function prediction	1,303	75.6
Genes assigned to COGs	1,191	69.1
Genes with peptide signals	141	8.2
Genes with transmembrane helices	389	22.6

a The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

474 **Table S2:** Number of genes associated with the 25 general COG functional categories of strain

475 Marseille-P2666^{T}

Code	Value	% of total	Description
[J]	137	8.0	Translation
[A]	0	0	RNA processing and modification
[K]	98	5.7	Transcription
[L]	49	2.8	Replication, recombination and repair
[B]	1	0.1	Chromatin structure and dynamics
[D]	15	0.9	Cell cycle control, mitosis and meiosis
[Y]	0	0	Nuclear structure
[V]	40	2.3	Defense mechanisms
[T]	51	3.0	Signal transduction mechanisms
[M]	65	3.8	Cell wall/membrane biogenesis
[N]	5	0.3	Cell motility
[Z]	0	0	Cytoskeleton
[W]	4	0.2	Extracellular structures
[U]	19	1.1	Intracellular trafficking and secretion
[O]	50	2.9	Post-translational modification, protein turnover, chaperones
[X]	6	0.3	Mobilome: prophages, transposons
[C]	77	4.5	Energy production and conversion
[G]	182	10.6	Carbohydrate transport and metabolism
[E]	115	6.7	Amino acid transport and metabolism
[F]	52	3.0	Nucleotide transport and metabolism
[H]	63	3.7	Coenzyme transport and metabolism
[I]	33	1.9	Lipid transport and metabolism
[P]	68	3.9	Inorganic ion transport and metabolism
[Q]	15	0.9	Secondary metabolites biosynthesis, transport and catabolism
[R]	104	6.0	General function prediction only
[S]	70	4.1	Function unknown
_	533	30.9	Not in COGs

476

477 **Table S3:** Genome comparison of closely related species to *Collinsella vaginalis* strain

478 Marseille P2666^T

Species	INSDC identifier[a]	Size (Mb)	G+C (mol %)	Gene Content
Collinsella vaginalis strain Marseille-P2666^T	FWYK00000000.1	2.2	64.6	1,907
Collinsella intestinalis DSM 13280	ABXH00000000.2	1.8	62.5	1,630
Collinsella aerofaciens ATCC 25986	AAVN00000000.2	2.4	60.5	2,437
Collinsella stercoris DSM 13279	ABXJ00000000.1	2.5	63.2	2,119
Collinsella tanakei YIT 12063	ADLS00000000.1	2.5	60.2	2,253
Coriobacterium glomerans ATCC 49209	CP002628.1	2.1	60.4	1,856
Olsenella profusa DSM 13989	AWEZ00000000.1	2.7	64.2	2,707
Olsenella uli ATCC 49627	CP002106.1	2.1	64.7	1,812

479 [a] INSDC: International Nucleotide Sequence Database Collaboration.

480 **Table S4:** dDDH values (%) obtained by comparison of all studied genomes

	CT	CS	CI	CA	CG	OP	OU
CV	22.6 ± 2.4	23.2 ± 2.4	23.0 ± 2.4	22.4 ± 2.4	20.4 ± 2.3	19.1 ± 2.8	19.7 ± 2.4
CT		25.0 ± 2.4	24.7 ± 2.4	22.5 ± 2.4	21.6 ± 2.4	20.0 ± 2.4	19.5 ± 2.3
CS			28.2 ± 2.5	23.9 ± 2.4	21.3 ± 2.3	19.1 ± 2.3	20.3 ± 2.3
CI				23.6 ± 2.4	21.2 ± 2.4	19.5 ± 2.3	20.4 ± 2.3
CA					21.0 ± 2.3	19.6 ± 2.3	20.0 ± 2.3
CG						20.0 ± 2.3	20.0 ± 2.3
OP							22.3 ± 2.4

481 dDDH: Digital DNA-DNA hybridization. CV: *Collinsella vaginalis* Marseille-P2666^T;
482 CT : *Collinsella tanakaei* YIT 12063^T*;* CS : *Collinsella stercoris* DSM 13279^T; CI : *Collinsella*
483 *intestinalis* DSM 13280^T; CA : *Collinsella aerofaciens* ATCC 25986^T; CG : *Coriobacterium*
484 *glomerans* ATCC 49209^T*;* OP : *Olsenella profusa* DSM 13989^T; OU : *Olsenella uli* ATCC
485 49627^T

486

487 **Table S5:** Average amino acid identity (AAI) values (%) between *Collinsella vaginalis*
488 strain Marseille P2666^T and other closely related *Collinsella* species.

	CT	CS	CI	CA
CV	66.9	66.1	66.9	64.7
CT		68.9	69.7	65.5
CS			79.5	66.3
CI				66.4

489 CV : *Collinsella vaginalis* Marseille-P2666^T; CT : *Collinsella tanakaei* YIT 12063^T*;* CS :
490 *Collinsella stercoris* DSM 13279^T; CI : *Collinsella intestinalis* DSM 13280^T; CA : *Collinsella*
491 *aerofaciens* ATCC 25986^T.

SUPPLEMENTARY FIGURE LEGENDS

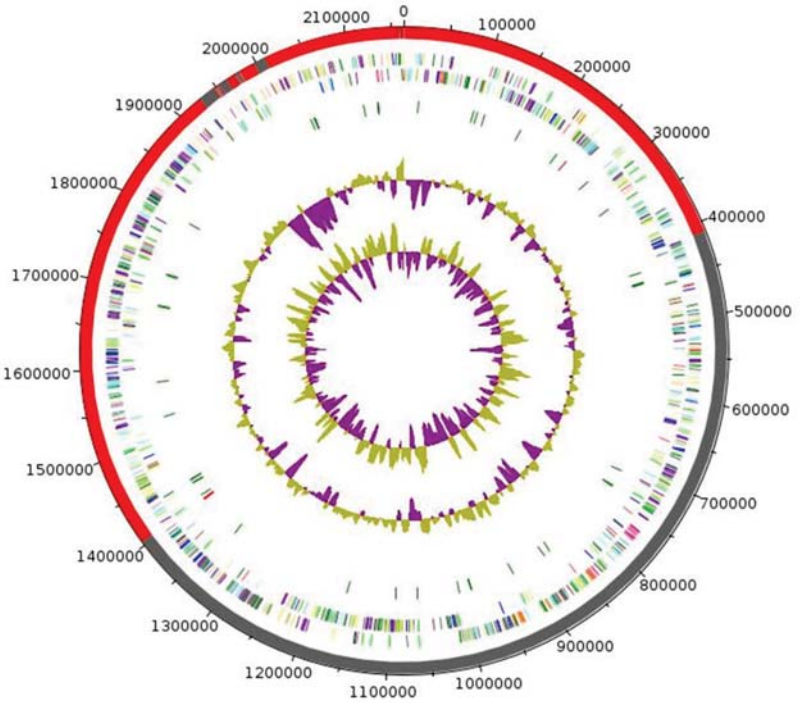
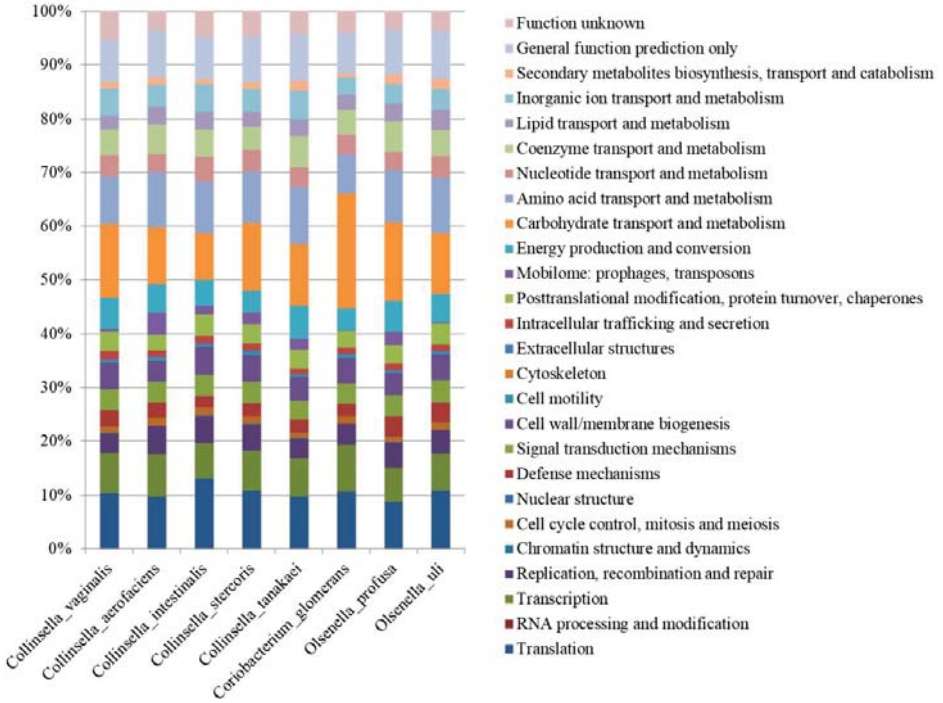


Figure S1. Graphical circular map of the genome. From the outside in: contigs (red/gray), COG category of genes on the forward strand (three circles), genes on the forward strand (blue circle), genes on the reverse strand (red circle), COG category on the reverse strand (three circles), G+C content.

Figure S3. Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of *Collinsella vaginalis* strain Marseille-P2666^{T} among other species.



References

1. **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, *et al.*** Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23:2947–2948.

2. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S**. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* 2013;30:2725–2729.

3. **Kimura M**. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980;16:111–120.

4. **Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, *et al.*** Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:1.

5. **Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, *et al.*** GenBank. *Nucleic Acids Res* 2015;43:D30–D35.

6. **Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, *et al.*** HMMER web server: 2015 update. *Nucleic Acids Res* 2015;43:W30–W38.

7. **Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, *et al.*** RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007;35:3100–3108.

8. **Lowe TM, Eddy SR**. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;25:955–964.

9. **Käll L, Krogh A, Sonnhammer EL.** A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J Mol Biol* 2004;338:1027–1036.

10. **Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA**. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 2012;28:464–469.

11. **Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J**. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 2009;25:119–120.

12. **Gouret P, Paganini J, Dainat J, Louati D, Darbo E, *et al.*** Integration of Evolutionary Biology Concepts for Functional Annotation and Automation of Complex Research in Evolution: The Multi-Agent Software System DAGOBAH. in: Pontarotti, P. (Ed.), Evolutionary Biology –Concepts, Biodiversity, Macroevolution and Genome Evolution. 581 Berlin, Heidelberg: *Springer Berlin Heidelberg*. 2011;pp. 71–87.

13. **Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, *et al.*** FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics* 2005;6:198.

14. **Padmanabhan R, Mishra AK, Raoult D, Fournier P-E**. Genomics and metagenomics in medical microbiology. *J Microbiol Methods* 2013;95:415–424.

15. **Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, *et al.*** A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* 2014;64:384–391.

16. **Lechner M, Findei\s s S, Steiner L, Marz M, Stadler PF, *et al.*** Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinformatics* 2011;12:124.

17. **Rodriguez-R LM, Konstantinidis KT**. Bypassing cultivation to identify bacterial species. *Microbe* 2014;9:111–8.

18. **Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M**. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:1.

Article 10:

Olegusella massiliensis gen nov, sp. nov., strain KHD7^{T}, a new bacterial genus isolated from the female genital tract of a patient with bacterial vaginosis

Diop Kh, Diop A, Bretelle F, Cadoret F, Michelle C, Richez M, Coccallemen JF, Raoult D, Fournier PE and Fenollar F

[Published in Anaerobe]

Anaerobe 44 (2017) 87–95

Contents lists available at ScienceDirect

Anaerobe

journal homepage: www.elsevier.com/locate/anaerobe

Anaerobes in the microbiome

Olegusella massiliensis gen. nov., sp. nov., strain KHD7^T, a new bacterial genus isolated from the female genital tract of a patient with bacterial vaginosis

Khoudia Diop [a], Awa Diop [a], Florence Bretelle [a, b], Frédéric Cadoret [a], Caroline Michelle [a], Magali Richez [a], Jean-François Cocallemen [b], Didier Raoult [a, c], Pierre-Edouard Fournier [a], Florence Fenollar [a, *]

[a] Aix Marseille Univ, Institut Hospitalo-Universitaire Méditerranée-Infection, URMITE, UM63, CNRS 7278, IRD 198, Inserm U1095, Faculté de médecine, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05, France
[b] Department of Gynecology and Obstetrics, Gynépole, Marseille, Pr Boubli et D'Ercole, Hôpital Nord, Assistance Publique-Hôpitaux de Marseille, AMU, Aix-Marseille Université, France
[c] Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 18 August 2016
Received in revised form
2 February 2017
Accepted 15 February 2017
Available online 20 February 2017

Handling Editor: Emma Allen-Vercoe

Keywords:
Olegusella massiliensis
Vaginal flora
Bacterial vaginosis
Culturomics
Taxono-genomics
Genome

ABSTRACT

Strain KHD7^T, a Gram-stain-positive rod-shaped, non-sporulating, strictly anaerobic bacterium, was isolated from the vaginal swab of a woman with bacterial vaginosis. We studied its phenotypic characteristics and sequenced its complete genome. The major fatty acids were C16:0 (44%), C18:2n6 (22%), and C18:1n9 (14%). The 1,806,744 bp long genome exhibited 49.24% G+C content; 1549 protein-coding and 51 RNA genes. Strain KHD7^T exhibited a 93.5% 16S rRNA similarity with *Olsenella uli*, the phylogenetically closest species in the family *Coriobacteriaceae*. Therefore, strain KHD7^T is sufficiently distinct to represent a new genus, for which we propose the name *Olegusella massiliensis* gen. nov., sp. nov. The type strain is KHD7^T.

1. Introduction

The female genital tract is a complex ecosystem colonized by several types of microorganisms. Its composition was described for the first time in 1892 by Doderleïn and in 1901 by Beijerink, revealing that four species of *Lactobacillus* are predominant in healthy vaginal flora: *Lactobacillus crispatus*, *Lactobacillus gasseri*, *Lactobacillus jensenii*, and *Lactobacillus iners* [1,2]. The other bacteria include some anaerobic species such as *Bacteroides*, *Peptostreptococcus*, *Peptococcus*, *Corynebacterium*, and *Eubacterium* [3]. This mutualistic association maintains the stability of the vaginal environment, preventing infection by inhibiting the growth and expansion of pathogens through the production of antimicrobial molecules such as hydrogen peroxide, lactic acid, and bacteriocins [4,5].

This mutualism is disturbed in bacterial vaginosis (BV). The most common cause of vaginal discharge affecting women of childbearing age, BV is concurrently characterized by reduced *Lactobacillus* species and increased anaerobic bacteria including *Atopobium vaginae*, *Bacteroides* spp., *Mobiluncus* spp., *Prevotella* spp., *Peptoniphilus* spp., and *Anaerococcus* spp. [6–9]. The vaginal microbiota was first studied by conventional culture methods. These methods are limited because 80% of the bacterial microbiota is considered to be fastidious or not cultivable [10]. Advances in molecular techniques, with sequencing and phylogenetic analysis of the 16S rRNA gene, enhanced understanding of the human vaginal microbiota. These molecular methods allowed the detection of fastidious and uncultured bacteria, such as bacterial vaginosis-associated bacteria type 1 (BVAB1), BVAB2, and BVAB3 [11].

* Corresponding author.
E-mail address: florence.fenollar@univ-amu.fr (F. Fenollar).

http://dx.doi.org/10.1016/j.anaerobe.2017.02.012
1075-9964/

Abbreviations

AGIOS	Average of Genomic Identity of Orthologous gene Sequences
bp:	base pairs
COG	Clusters of Orthologous Groups
CSUR	Collection de souches de l'Unité des Rickettsies
DDH	DNA-DNA Hybridization
DSM	Deutsche Sammlung von Mikroorganismen
FAME	Fatty Acid Methyl Ester
GC/MS	Gas Chromatography/Mass Spectrometry
kb	kilobases
MALDI-TOF	Matrix-assisted laser-desorption/ionization time-of-flight
ORF	Open Reading Frame
TE buffer	Tris-EDTA buffer
URMITE	Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes

As part of a study on the diversity of the vaginal microbiota of patients with bacterial vaginosis using the culturomics approach, based on multiplication of culture conditions (variation of media, temperature, and atmosphere) with more rapid bacterial identification by MALDI-TOF mass spectrometry [12], we isolated a new member of the *Coriobacteriaceae* family. This family, created in 1997 by Stackebrandt, contains 35 species grouped in 13 validated genera [13,14].

Various parameters, including phenotypic and genotypic characteristics such as DNA-DNA hybridization, have been used to define a new species but they present certain limitations [15,16], so we introduced "taxono-genomics", a new approach that includes genomic analysis and proteomic information obtained by MALDI-TOF mass spectrometry analysis [17,18].

Here, we describe *Olegusella massiliensis* strain KHD7^T (= CSUR P2268 = DSM 101849), with its complete annotated genome, a new member of the *Coriobacteriaceae* family isolated in the vaginal flora of a patient with bacterial vaginosis.

2. Materials and methods

2.1. Sample collection

In October 2015, the vaginal sample of a French 33 year-old woman was collected at Hôpital Nord in Marseille (France). The patient was suffering from bacterial vaginosis, which was diagnosed as previously reported [19]. At the time of sample collection, she was not being treated with any antibiotics. She gave her written consent. This study was authorized by the local IFR48 ethics committee (Marseille, France) under agreement number 09-022. The sample was collected and transported using a Sigma Transwab (Medical Wire, Corsham, United Kingdom).

2.2. Strain identification by MALDI-TOF MS

After collection, the sample was first inoculated in a blood culture bottle (BD Diagnostics, Le Pont-de-Claix, France) supplemented with 4 mL of rumen that was filter-sterilized through a 0.2 μm pore filter (Thermo Fisher Scientific, Villebon-sur-Yvette, France), and 3 mL of sheep blood (bioMérieux, Marcy l'Etoile, France). The supernatant was then inoculated on 5% sheep blood-enriched CNA agar (BD Diagnostics) under anaerobic conditions at 37 °C. Isolated colonies were deposited in duplicate on a MTP 96 MALDI-TOF target plate (Bruker Daltonics, Leipzig, Germany) for identification with a microflex spectrometer (Bruker) [20]. Briefly, 1.5 μL of matrix solution, containing solution of α-cyano-4-hydroxycinnamic acid diluted in 500 μL acetonitrile, 250 μL 10% trifluoroacetic acid and 250 μL HPLC water was deposited on each spot for ionization and crystallization. All protein spectra obtained were compared with those in the MALDI-TOF database. If the score was greater than or equal to 1.9, the strain was considered identified. Otherwise, the identification failed.

2.3. Strain identification by 16S rRNA sequencing

For unidentified strains using MALDI-TOF MS, 16S rRNA sequencing was used to achieve identification [21]. As Stackebrandt and Ebers suggested, if the 16S rRNA sequence similarity value was lower than 98.7% or 95%, the strain was defined as a new species or genus, respectively [22–24].

2.4. Morphologic observation and growth conditions

Optimal strain growth was also tested at different temperatures (25, 28, 37, 45, and 56 °C) in an aerobic atmosphere with or without 5% CO_2, and in anaerobic and microaerophilic atmospheres using GENbag Anaer and GENbag miroaer systems (bioMérieux).

For electron microscopy, detection formvar-coated grids were dropped onto a 40 μL bacterial suspension before incubation at 37 °C for 30 min. Then, the grids were incubated on 1% ammonium molybdate for 10 s, dried on blotting paper and finally observed using a Tecnai G20 transmission electron microscope (FEI, Limeil-Brevannes, France) at an operating voltage of 60 Kv. Standard procedures were used to perform Gram-staining, motility, sporulation as well as oxidase and catalase tests [25].

2.5. Biochemical analysis and antibiotic susceptibility tests

Cellular fatty acid methyl ester (FAME) analysis was performed by GC/MS. Strain KHD7^T was grown on Columbia agar enriched with 5% sheep blood (bioMérieux). Then, two samples were prepared with approximately 30 mg of bacterial biomass per tube harvested from several culture plates. Fatty acid methyl esters were prepared as described by Sasser [26]. GC/MS analyses were realized by using a Clarus 500 gas chromatograph equipped with a SQ8S MS detector (Perkin Elmer, Courtaboeuf, France). 2 μL of FAME extracts were volatilized at 250 °C (split 20 mL/min) in a Focus liner with wool and separated on an Elite-5MS column (30 m, 0.25 mm i.d., 0.25 mm film thickness) using a linear temperature gradient (70–290 °C at 6 °C/min), allowing the detection of C4 to C24 fatty acid methyl esters. Helium flowing at 1.2 mL/min was used as carrier gas. The MS inlet line was set at 250 °C and EI source at 200 °C. Full scan monitoring was performed from 45 to 500 m/z. All data were collected and processed using Turbomass 6.1 (Perkin Elmer). FAMEs were identified by a spectral database search using MS Search 2.0 operated with the Standard Reference Database 1A (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA) and the FAMEs mass spectral database (Wiley, Chichester, UK). Retention time correlations with estimated nonpolar retention indexes from the NIST database were obtained using a 37-component FAME mix (Supelco; Sigma-Aldrich, Saint-Quentin Fallavier, France); FAME identifications were confirmed using this index).

API ZYM, API 20A, and API 50CH strips (bioMérieux) were used

to perform the biochemical test according to the manufacturer's instructions. The strips were incubated in anaerobic conditions and respectively for 4, 24, and 48 h. Antibiotic susceptibility was tested using the E-test gradient strip method (BioMerieux) to determine the minimal inhibitory concentration (MIC) of each tested antibiotic. Strain KHD7^T was grown on blood Colombia agar (BioMerieux) and a bacterial inoculum of turbidity 0.5 McFarland was prepared by suspending the culture in sterile saline (0.85% NaCl). Using cotton swabs, the inoculum was plated on 5% horse blood enriched Mueller Hinton Agar (BioMerieux) according to EUCAST recommendations [27,28]. E-test strips (amoxicillin, benzylpenicillin, imipenem, and vancomycin) were then deposited and the plates were incubated under anaerobic conditions for 48 h. Around the strip, Elliptic zones of inhibition were formed and the intersection with the strip indicates the MIC [28]. MICs were interpreted according to the EUCAST recommendations [29]. *Escherichia coli* strain DSM 1103 was used as a quality control strain.

2.6. Genomic DNA preparation

Strain KHD7^T was grown in anaerobic conditions at 37 °C using Columbia agar enriched with 5% sheep blood (bioMérieux) after 48 h on four Petri dishes. Bacteria were resuspended in 500 μL of TE buffer; 150 μL of this suspension was diluted in 350 μL 10× TE buffer, 25 μL proteinase K, and 50 μL sodium dodecyl sulfate for lysis treatment. This preparation was incubated overnight at 56 °C. DNA was purified using phenol/chloroform/isoamylalcohol successively for extraction and followed by ethanol precipitation at −20 °C of at least 2 h each. Following centrifugation, the DNA was suspended in 65 μL EB buffer. Genomic DNA concentration was measured at 46.06 ng/μL using the Qubit assay with the high-sensitivity kit (Life technologies, Carlsbad, CA, USA).

2.7. Genome sequencing and assembly

Genomic DNA of strain KHD7^T was sequenced on the MiSeq Technology (Illumina Inc., San Diego, CA, USA) with the mate pair strategy. The gDNA was barcoded with the Nextera Mate Pair sample prep kit (Illumina) in order to be mixed with 11 other projects.

gDNA was quantified by a Qubit assay with the high sensitivity kit (Life technologies, Carlsbad, CA, USA) to 26 ng/μL. The mate pair library was prepared with 1.5 μg of genomic DNA using the Nextera mate pair Illumina guide. The genomic DNA sample was simultaneously fragmented and tagged with a mate pair junction adapter. The pattern of the fragmentation was validated on an Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA 7500 labchip. The DNA fragments ranged in size from 1.5 kb up to 11 kb with an optimal size at 6.228 kb. No size selection was performed and 556 ng of tagmented fragments were circularized. The circularized DNA was mechanically sheared to small fragments with an optimal at 1275 bp on the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The library profile was visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) and the final concentration library was measured at 37.47 nmol/L.

The libraries were normalized at 2 nM and pooled. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated cluster generation and sequencing run were performed in a single 39-h run in a 2 × 151-bp. Total information of 6.5 Gb was obtained from 696 K/mm^2 cluster density with cluster passing quality control filters of 95.6% (12,863,000 passing filter paired reads). Within this run, the index representation for strain KHD7^T was determined at 6.26%. The 805,042 paired reads were trimmed then assembled in two scaffolds.

2.8. Genome annotation and analysis

Prodigal was used for Open Reading Frames (ORFs) prediction [30] with default parameters. We excluded predicted ORFs spanning a sequencing gap region (containing N). The bacterial proteome was predicted using BLASTP (E-value 1e^{-03} coverage 0.7 and identity percent 30) against the Clusters of Orthologous Groups (COG) database. If no hit was found we searched against the NR database [31] using BLASTP with E-value of 1e^{-03} coverage 0.7 and an identity percent of 30. An E-value of 1e^{-05} was used if sequence lengths were smaller than 80 amino acids. PFam conserved domains (PFAM-A an PFAM-B domains) were searched on each protein with the hhmscan tools analysis. RNAmmer [32] and tRNAScanSE tool [33] were used to find ribosomal RNAs genes and tRNA genes respectively. ORFans were identified if all the BLASTP performed had negative results (E-value smaller than 1e^{-03} for ORFs with sequence size above 80 aa or E-value smaller than 1e^{-05} for ORFs with sequence length below 80 aa). For data management and visualization of genomic features, Artemis [34] and DNA Plotter [35] were used, respectively. We used the MAGI in-house software to analyze the mean level of nucleotide sequence similarity at the genome level. It calculated the average genomic identity of gene sequences (AGIOS) among compared genomes [36]. This software combines the Proteinortho software [37] for detecting orthologous proteins in pairwise genomic comparisons. Then the corresponding genes were retrieved and the mean percentage of nucleotide sequence identity among orthologous ORFs was determined using the Needleman-Wunsch global alignment algorithm. Genomes from the genus *Atopobium*, *Olsenella*, and *Collinsella* were used for the calculation of AGIOS values. The genome of strain KHD7^T (FLLS00000000) was compared with that of *Olsenella uli* DSM 7084 (NC_014363); *Olsenella profusa* F0195 (AWEZ00000000); *Atopobium fossor* DSM 15642 (AXXR00000000); *Atopobium parvulum* DSM 20469 (NC_013203); *Atopobium rimae* ATCC 49626 (ACFE00000000); *Collinsella tanakaei* YIT 12063 (ADLS00000000). The Multi-Agent software system DAGOBAH [38] was used to perform annotation and comparison processes, which include Figenix [39] libraries that provide pipeline analysis. We also performed GGDC analysis using the GGDC web server as previously reported [40].

3. Results

3.1. Strain characterization

3.1.1. Strain identification by MALDI-TOF

Strain KHD7^T was first isolated in November 2015 after 10 days of pre-incubation in a blood culture bottle enriched with rumen and sheep blood under anaerobic conditions and sub-cultured on CNA agar with 5% sheep blood at 37 °C, also under anaerobic conditions. MALDI-TOF MS analysis of strain KHD7^T gave a low score (1.2), suggesting that our isolate was not in the database and could be a previously unknown species.

3.1.2. Strain identification by 16S rRNA sequencing gene

The 16S rRNA gene was then sequenced and the sequence obtained (accession number LN998058) shows 93.5% similarity with *Olsenella uli*, the phylogenetically closest bacterial species with a

validly published name (Fig. 1). As this value is lower than 95% threshold defined by Stackebrandt and Ebers for defining a new genus, we classified strain KHD7^T as the type strain of a new genus named *Olegusella* (Table 1). The reference spectrum was then added to our database (See Supplementary Table S1) and compared with those of the closest species (See Supplementary Table S2).

3.1.3. Phenotypic characteristics

Strain KHD7^T grew only in anaerobic conditions. Growth was observed at temperatures ranging from 25 to 42 °C, with optimal growth at 37 °C under anaerobic conditions after 48 h of incubation. The bacterium needed NaCl concentration below 0.5% and the pH for growth ranges from 6.5 to 7.0. On blood-enriched Columbia agar, colonies were pale white and translucent with a diameter of 1–1.2 mm. Gram-staining showed a rod-shaped Gram-positive bacterium (Fig. 2). On electron microscopy, individual cells appear with a mean diameter of 0.35 μm and a mean length of 0.42 μm (Fig. 3). Strain KHD7^T is non-motile and non-sporeforming.

The major fatty acid found for this strain was C16:0 acid (44%). Several unsaturated fatty acids were described including two abundant species: C18:2n6 (22%) and C18:1n9 (14%). Fatty acids with shorter aliphatic chains were also detected such as C8:0, C10:0, and C12:0 (Table 2).

Strain KHD7^T exhibited neither catalase nor oxidase activities.

Table 1
Classification and general features of *Olegusella massiliensis* strain khD7^T.

Properties	Terms
Taxonomy	**Kingdom**: *Bacteria* **Phylum**: *Acinetobacteria* **Class**: *Coriobacteriia* **Order**: *Coriobacteriales* **Family**: *Coriobacteriaceae* **Genus**: *Olegusella* **Species**: *Olegusella massiliensis*
Type strain	KhD7
Isolation site	Human vagina
Isolation country	France
Gram stain	Negative
Cell shape	Bacilli
Motility	No
Oxygen requirements	Anaerobic
Optimal temperature	37 °C
Temperature range	Mesophilic
Habitat	Host Associated
Biotic relationship	Free living
Host name	*Homo sapiens*
Sporulation	Nonsporulating
Metabolism	NA
Energy source	Chemoorganotrophic
Pathogenicity	Unknown
Biosafety level	2

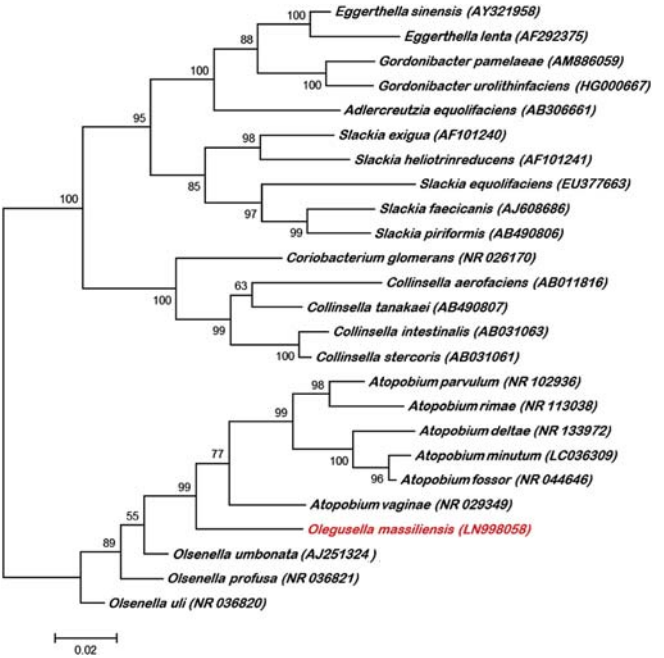


Fig. 1. Phylogenetic tree highlighting the position of *Olegusella massiliensis* strain KHD7^T relative to other close strains. GenBank accession numbers of each 16S rRNA is noted just after the name. Sequences were aligned using CLUSTALW, with default parameters and phylogenetic inferences were obtained using neighbor-joining method with 500 bootstrap replicates, within MEGA6 software. The scale bar represents a 2% nucleotide sequence divergence.

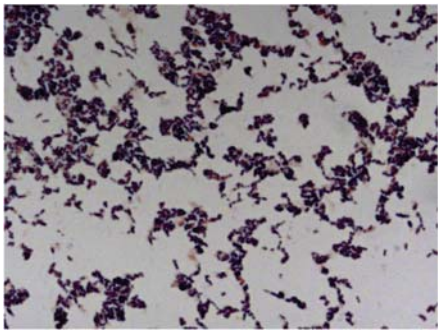


Fig. 2. Gram-staining of *Olegusella massiliensis* strain KHD7^T.

Fig. 3. Transmission electron microscopy of *Olegusella massiliensis* strain KHD7^T using a Tecnai G20 transmission electron microscope (FEI Company). The scale bar represents 200 nm.

Table 2
Cellular fatty acid composition (%).

Fatty acids	Name	Mean relative %[a]
16:0	Hexadecanoic acid	43.5 ± 0.7
18:2n6	9,12-Octadecadienoic acid	22.1 ± 0.4
18:1n9	9-Octadecenoic acid	13.8 ± 0.3
18:0	Octadecanoic acid	8.3 ± 0.1
14:0	Tetradecanoic acid	6.1 ± 0.4
10:0	Decanoic acid	1.6 ± 0.2
18:1n7	11-Octadecenoic acid	1.0 ± 0.1
18:1n6	12-Octadecenoic acid	TR
12:0	Dodecanoic acid	TR
15:0	Pentadecanoic acid	TR
16:1n7	9-Hexadecenoic acid	TR
15:0 anteiso	12-methyl-tetradecanoic acid	TR
15:0 iso	13-methyl-tetradecanoic acid	TR
8:0	Octanoic acid	TR

[a] Mean peak area percentage; TR = trace amounts < 1%.

Using API ZYM strip, positive reactions were detected for leucine arylamidase, acid phosphatase, naphthol phosphohydrolase, and *N*-acetyl-beta-glucosaminidase but no reaction was observed for alkaline phosphatase, lipases (C4, C8 and C14), valine and cysteine arylamidase, α-chymotrypsin, α-galactosidase, β-galactosidase, β-glucuronidase, α-glucosidase, β-glucosidase, α-mannosidase, and α-fucosidase. An API 50 CH strip revealed that strain KHD7^T metabolized D-glucose, D-mannose, *N*-acetylglucosamine, D-saccharose, and potassium 5-cetogluconate. This same strip show negative reactions for glycerol, erythritol, D-arabinose, arabinose (D and L), D-ribose, xylose, D-adonitol, methyl-βD-xylopyranoside, D-galactose, D-fructose, L-sorbose, L-rhamnose, dulcitol, inositol, D-mannitol, D-sorbitol, methyl-αD-mannopyranoside, methyl- αD-glucopyranoside, amygdaline, arbutine, esculin ferric citrate, salicine, D-cellobiose, D-maltose, D-lactose, D-melibiose, D-trehalose, inuline, D-melezitose, D-raffinose, starch, glycogene, xylitol, gentiobiose, D-turanose, D-lyxose, D-tagatose, fucose, arabitol, potassium gluconate, and potassium 2-cetogluconate. Based on API 20A strip, nitrate was not reduced, indole formation was negative. API 20A revealed also that esculin ferric citrate was hydrolyzed unlike gelatin.

Strain KHD7^T was susceptible to amoxicillin (MIC 0.38 μg/mL), benzylpenicillin (MIC 0.50 μg/mL), imipenem (MIC 1.25 μg/mL), and vancomycin (MIC 1 μg/mL). Phenotypic characteristics of strain KHD7^T compared with those of closely related species are shown in Table 3.

3.2. Genome properties

The final assembly identified two scaffolds (2 contigs) generating a genome size of 1,806,744 bp (1 chromosome, but no plasmid). The genome sequence was deposited in GenBank under accession number FLLS00000000. The G+C content was 49.24% (Table 4 and Fig. 4). Of the 1600 predicted genes, 1549 were protein-coding genes, and 51 were RNAs (two 5S rRNA, two 16S rRNA, two 23S rRNA, and 45 tRNA genes). A total of 1349 genes (87.08%) were assigned a putative function (by cogs or by NR blast): 54 genes were identified as ORFans (3.49%). The remaining genes were annotated as hypothetical proteins (224 genes, 14.46%). Genome statistics are summarized in Table 4. Genes are distributed according to COG functional categories in Table 5.

3.3. Genomic comparison

Compared to the genomes of other closed related species, the genome of strain KHD7^T (1.80 Mbp) is larger than those of *Atopobium fossor*, *Atopobium parvulum*, and *Atopobium rimae* (1.66; 1.54 and 1.63 Mbp respectively) but it is smaller than those of *Olsenella profusa*, *Olsenella uli*, and *Collinsella tanakaei* (2.72; 2.05; and 2.49 Mbp respectively). The G+C content of strain KHD7^T (49.24%) is smaller than those of *Olsenella uli*, *Olsenella profusa*, *Collinsella tanakaei*, and *Atopobium rimae* (64.70; 64.1; 60.2 and 49.30%, respectively) but larger than those of *Atopobium fossor* and *Atopobium parvulum* (45.4% and 45.70%, respectively). The gene content of strain KHD7^T (1,600) is smaller than those of *Olsenella uli*, *Olsenella profusa*, and *Collinsella tanakaei* (1,793, 2,474, and 2,150, respectively) but larger than those of *Atopobium fossor*, *Atopobium parvulum*, and *Atopobium rimae* (1,505, 1,406, and 1,511, respectively). However, the distribution of genes into COG categories was similar among all compared genomes (Fig. 5). In addition, strain KHD7^T shared on the one hand between 822 and 862 orthologous genes and on the other hand between 752 and 779 orthologous genes with the most closely related species belonging to the *Olsenella* and *Atopobium* genera (*O. uli*, *O. profusa* and *A. fossor*, *A. parvulum*, and *A. rimae*, respectively). Finally, it shared 745

Table 3
Differential characteristics of *Olegusella massiliensis* strain KHD7^T, *Olsenella uli* strain DSM 7084^T, *Olsenella umbonata* strain DSM 22620^T, *Olsenella profusa* strain DSM 13989^T, *Atopobium parvulum* strain ATCC 33793^T, *Atopobium rimae* strain ATCC 49626^T, *Atopobium fossor* strain NCTC 11919^T, *Atopobium deltae* strain CCUG 65171^T, and *Collinsella tanakaei* strain DSM 22478^T [40–46].

Properties	*Olegusella massiliensis*	*Olsenella uli*	*Olsenella umbonata*	*Olsenella profusa*	*Atopobium parvulum*	*Atopobium rimae*	*Atopobium fossor*	*Atopobium deltae*	*Collinsella tanakaei*
Cell diameter (μm)	0.3–0.4	na	0.3–0.6	0.6–0.8	0.3–0.6	na	0.5–0.9	1–1.2	0.5–1
Major fatty acid	$C_{16:0}$ (43.5%)	$C_{18:0}$ (31.7%)	$C_{18:0}$ (51%)	$C_{14:0}$-antesio (68.7%)	$C_{18:1}$ *cis*-9 FAME (38.2%)	$C_{18:1}$ *cis*-9 FAME (32.5%)	$C_{16:0}$ (33.3%)	C16:0 (33.3%)	C18:1 *cis*-9 FAME (44.91%)
DNA G+C content (mol%)	49.24	64.70	63	64.1	45.7	49.30	45.4	50.3	60.2
Production of									
Alkaline phosphatase	–	–	–	+	na	na	na	–	+
β-galactosidase		–	–	+	+	–	na	–	–
N-acetyl-glucosamine	+	–	–	+	na	na	na	–	–
Acid from									
Ribose	–	–	na	na	–	+	–	na	na
Mannitol	–	–	–	+	–	–	–	–	–
Sucrose	–	+	+	+	+	+	–	+	+
D-fructose	–	+	+	+	+	+	–	na	na
D-maltose	–	+	+	+	+	+	–	na	+
D-lactose	–	–	–	+	+	–	–	+	+
Habitat	Human vagina	Human gingival crevices	Sheep rumen	Human subgingival	Human gingival crevices	Human gingival crevices	Horse oropharyngeal	Human blood	Human faeces

+: positive reaction; –: negative reaction; na: not available data. Data are from literature except DNA G+C content which was calculated by EMBOSS software online (http://www.bioinformatics.nl/emboss-explorer/.

Table 4
Nucleotide content and gene count levels of the genome.

Attribute	Value	of total[a]
Size (bp)	1,806,744	100
G+C content (bp)	889,672	49.24
Coding region (bp)	1,610,188	89.12
Total genes	1600	100
RNA genes	51	3.18
Protein-coding genes	1549	96.81
Genes with function prediction	1349	87.08
Genes assigned to COGs	1219	78.69
Genes with peptide signals	125	8.06
Genes with transmembrane helices	371	23.95

[a] The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

orthologous genes with the most distant species belonging to the *Collinsella* genus (*C. tanakaei)* (Table 6). The same trend was observed when we analyzed the average percentage of nucleotide sequence identity, which ranged from 64.76% to 66.04% between *O. uli*, *O. profusa*, *A. parvulum*, *A. rimae*, and *A. fossor* species, but was 62.98% between strain KHD7^T and *C. tanakaei*. We obtained similar results for the analysis of the digital DNA-DNA hybridization (dDDH) using Genome-to-Genome Distance Calculator (GGDC) software (Table 7).

4. Discussion

Strain KHD7^T was isolated as part of a "culturomics" study of the vaginal flora aiming to isolate all bacterial species within the vagina. Strain KHD7^T was considered as a new genus on the basis of its unique MALDI-TOF MS spectrum, the genome comparison and its low 16S rRNA similarity level. The latter value was 93.5% with *O. uli*, which was lower than the recommended 95% threshold to define a new genus [22]. Strain KHD7^T is a member of the family *Coriobacteriaceae* belonging to the phylum *Actinobacteria.* This family comprises 35 species divided into 13 validated genera [13,14]. Most members of the *Coriobacteriaceae* are Gram-positive, non-motile, and non-sporulating bacteria. All these criteria are observed for *Olegusella massiliensis* strain KHD7^T. Bacterial species of the *Coriobacteriaceae* family have been detected in diverse habitats such as the intestinal tracts of humans and rodents, horse oropharynxes, human blood, and sheep rumen [41–46]. Furthermore, *Olsenella uli* was first isolated in the human gingival crevice; this bacterium is also associated with tissue destruction and periodontal inflammation [47].

A polyphasic taxono-genomics strategy [17,18], based on the combination of phenotypic and genomic analyses was used to characterize strain KHD7^T and the new genus from which it is the type strain. Phenotypically, strain KHD7^T exhibited a specific MALDI-TOF MS spectrum and differed from the other closed studied bacterial species in their fermentation of carbohydrate. Most often, the species of the *Coriobacteriaceae* family ferment glucose and mannose as observed for *Olegusella massiliensis*. Their differences lie on the fermentation of other carbohydrates such as ribose, mannitol, fructose, sucrose, lactose, and maltose. Unlike *O. uli*, *O. umbonata*, *O. profusa*, and *A. parvulum*, strain KHD7^T does not ferment sucrose, fructose, or maltose.

The G+C content of strain KHD7^T and its phylogenetically closest species varies from 45.4 to 64.70%. The genomic similarity of strain KHD7^T with species of *Coriobacteriaceae* family was evaluated by 2 parameters: DDH and AGIOS. The values found in DDH and AGIOS of *O. massiliensis* are in the range of those observed in the other genera of this family.

5. Conclusion

Based on the phenotypic analysis, phylogenetic and genomic results, strain KHD7^T may be a member of a new genus named *Olegusella* with *Olegusella massiliensis* as the type strain. It was isolated among the vaginal flora of a 33 year-old French woman suffering from bacterial vaginosis.

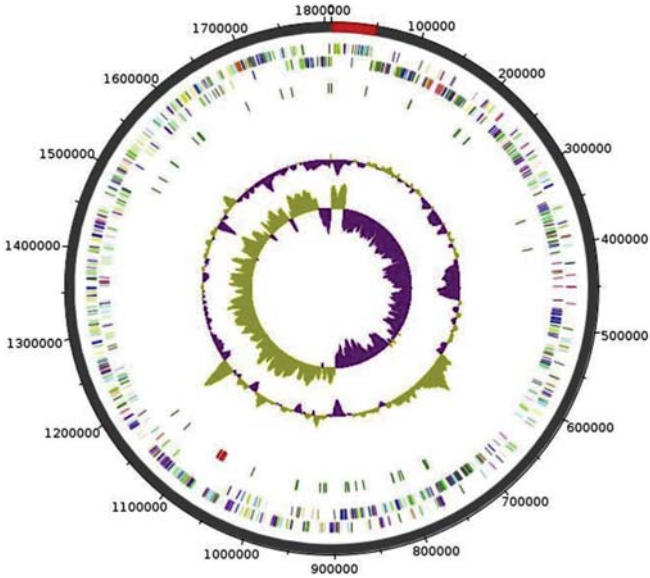


Fig. 4. Graphical circular map of the chromosome. From outside to the center: Genes on the forward strand colored by Clusters of Orthologous Groups of proteins (COG) categories (only genes assigned to COG). Genes on the reverse strand colored by COG categories (only gene assigned to COG). RNA genes (tRNAs green. rRNAs red). GC content and GC skew. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Number of genes associated with the 25 general COG functional categories.

Code	Value	% value	Description
J	125	10.25	Translation
A	0	0	RNA processing and modification
K	85	6.97	Transcription
L	74	6.07	Replication, recombination and repair
B	0	0	Chromatin structure and dynamics
D	17	1.39	Cell cycle control, mitosis and meiosis
Y	0	0	Nuclear structure
V	54	4.43	Defense mechanisms
T	39	3.20	Signal transduction mechanisms
M	85	6.97	Cell wall/membrane biogenesis
N	2	0.16	Cell motility
Z	0	0	Cytoskeleton
W	0	0	Extracellular structures
U	15	1.23	Intracellular trafficking and secretion
O	45	3.69	Posttranslational modification, protein turnover, chaperones
X	6	0.49	Mobilome: prophages, transposons
C	53	4.35	Energy production and conversion
G	111	9.11	Carbohydrate transport and metabolism
E	113	9.27	Amino acid transport and metabolism
F	51	4.18	Nucleotide transport and metabolism
H	34	2.79	Coenzyme transport and metabolism
I	26	2.13	Lipid transport and metabolism
P	49	4.02	Inorganic ion transport and metabolism
Q	9	0.74	Secondary metabolites biosynthesis, transport and catabolism
R	121	9.93	General function prediction only
S	105	8.61	Function unknown
_	330	21.30	Not in COGs

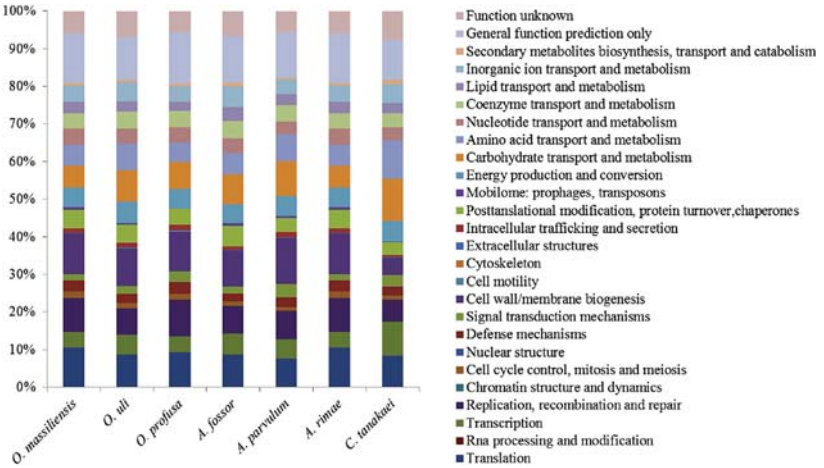


Fig. 5. Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of *Olegusella massiliensis* strain KHD7^T among other species.

Table 6
Numbers of orthologous proteins shared between genomes (upper right) and AGIOS values obtained (lower left). The numbers of proteins per genome are indicated in bold.

	OM	OU	OP	AF	AP	AR	CT
OM	**1550**	862	822	779	755	752	745
OU	64.76%	**1775**	928	836	816	837	814
OP	64.81%	75.26%	**2593**	790	817	821	811
AF	66.04%	62.79%	62.74%	**1487**	758	753	743
AP	65.77%	63.02%	62.91%	66.67%	**1363**	899	716
AR	65.37%	64.62%	64.56%	65.65%	72.13%	**1478**	**718**
CT	62.98%	62.98%	67.42%	62.46%	62.59%	63.35%	**2194**

OM *Olegusella massiliensis* KHD7^T; OU *Olsenella uli* DSM 7084; OP *Olsenella profusa* F0195; AF *Atopobium fossor* DSM 15642; AP *Atopobium parvulum* DSM 20469; AR *Atopobium rimae* ATCC 49626; CT *Collinsella tanakaei* YIT 12063.

5.1. Taxonomic and nomenclatural proposals

5.1.1. Description of Olegusella gen. nov.

Olegusella (O.le.gu.sel'la. M.L. dim. suffix usel'la; M.L. fem. n.) was chosen to honor Dr. Oleg Mediannikov for his contribution to medical microbiology. Gram-stain-positive rods. Strictly anaerobic. Mesophilic. Non-motile. Does not exhibit catalase, oxidase nor nitrate reduction. Positive for D-glucose, D-mannose, *N*-acetylglucosamine, D-saccharose, potassium 5-cetogluconate, leucine arylamidase, acid phosphatase, naphthol phosphohydrolase, and *N*-acetyl-beta-glucosaminidase. Habitat: human vaginal flora. Type species: *Olegusella massiliensis*.

5.1.2. Description of Olegusella massiliensis gen. nov., sp. nov.

Olegusella massiliensis (mas.il'ien'sis, L. gen. fem. n. massiliensis, of Massilia, the Latin name of Marseille where the Type strain was first isolated).

Gram-stain-positive rods. Strictly anaerobic. Mesophilic. Optimal growth at 37 °C. Non-motile and non-sporulating. Colonies are pale white and translucent with 1–1.2 mm diameter on blood-enriched Colombia agar. Cells are rod-shapped with diameter approximately 0.35 μm and length approximately 0.42 μm. Strain KHD7^T exhibited neither catalase nor oxidase activities. Nitrate reduction is absent. Positive reactions were observed for D-glucose, D-mannose, *N*-acetylglucosamine, D-saccharose, potassium 5-cetogluconate, leucine arylamidase, acid phosphatase, naphthol phosphohydrolase, and *N*-acetyl-beta-glucosaminidase. The major fatty acids are C16:0 acid (44%), C18:2n6 (22%) and C18:1n9 (14%). Strain KHD7^T is susceptible to penicillin, oxacillin, ceftriaxone, imipenem, ciprofloxacin, clindamycin, erythromycin, gentamicin, metronidazole, rifampicin, teicoplanin, and vancomycin but it is resistant to colistin, doxycycline, fosfomycin and trimethoprim-sulfamethoxazole.

The 16S rRNA and genome sequences are deposited in GenBank

Table 7
dDDH values obtained by comparison of all studied genomes.

	OM	OU	OP	AF	AP	AR	CT
OM	100	25.10 ± 2.4	22.00 ± 2.35	22.00 ± 2.35	23.00 ± 2.35	20.80 ± 2.35	22.50 ± 2.4
OU		100	22.30 ± 2.35	21.70 ± 2.35	25.00 ± 2.4	24.90 ± 2.4	19.50 ± 2.3
OP			100	19.80 ± 2.3	24.00 ± 2.4	21.60 ± 2.35	20.00 ± 2.35
AF				100	20.30 ± 2.35	21.00 ± 2.3	23.60 ± 2.4
AP					100	23.90 ± 2.4	20.80 ± 2.35
AR						100	22.00 ± 2.35
CT							100

dDDH: Digital DNA-DNA hybridization. OM *Olegusella massiliensis* KHD7^T; OU *Olsenella uli* DSM 7084; OP *Olsenella profusa* F0195; AF *Atopobium fossor* DSM 15642; AP *Atopobium parvulum* DSM 20469; AR *Atopobium rimae* ATCC 49626; CT *Collinsella tanakaei* YIT 12063.

under accession numbers LN998058 and FLLS00000000 respectively. The genome is 1,806,744 bp long with a G+C content of 49.24%. The type strain KHD7^T (= CSUR P2268^T = DSM 101849^T) was isolated from the vaginal flora of a patient with bacterial vaginosis.

Conflict of interest

The authors declare no conflict of interest.

Funding

This study was supported by the Institut hospitalo-universitaire Mediterranée-Infection.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.anaerobe.2017.02.012.

References

[1] J.P. Lepargneur, V. Rousseau, Protective role of the Doderleïn flora, J. Gynecol. Obstet. Biol. Reprod. 31 (2002) 485–494.
[2] B. Chaban, M.G. Links, T.P. Jayaprakash, et al., Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle, Microbiome 2 (2014) 23.
[3] K. Pal, S. Roy, B. Behera, et al., Bacterial vaginosis: etiology and modalities of treatment-A brief note, J. Pharm. Bioallied Sci. 3 (2011) 496.
[4] J.P. Lepargneur, V. Rousseau, Protective role of Doderleïn flora, J. Gynecol. Obstet. Biol. Reprod. 31 (2002) 485–494.
[5] S. Borges, J. Silva, P. Teixeira, The role of lactobacilli and probiotics in maintaining vaginal health, Arch. Gynecol. Obstet. 289 (2014) 479–489.
[6] J. Pépin, S. Deslandes, G. Giroux, et al., The complex vaginal flora of West African women with bacterial vaginosis, PLoS One 6 (2011) e25082.
[7] E. Shipitsyna, A. Roos, R. Datcu, et al., Composition of the vaginal microbiota in women of reproductive age – sensitive and specific molecular diagnosis of bacterial vaginosis is possible? PLoS One 8 (2013) e60670.
[8] C.S. Bradshaw, S.N. Tabrizi, C.K. Fairley, et al., The association of *Atopobium vaginae* and *Gardnerella vaginalis* with bacterial vaginosis and recurrence after oral metronidazole therapy, J. Infect. Dis. 194 (2006) 828–836.
[9] A.B. Onderdonk, M.L. Delaney, R.N. Fichorova, The human microbiome during bacterial vaginosis, Clin. Microbiol. Rev. 29 (2016) 223–238.
[10] R.F. Lamont, J.D. Sobel, R.A. Akins, et al., The vaginal microbiome: new information about genital tract flora using molecular based techniques, BJOG 118 (2011) 533–549.
[11] D.N. Fredricks, T.L. Fiedler, J.M. Marrazzo, Molecular identification of bacteria associated with bacterial vaginosis, N. Engl. J. Med. 353 (2005) 1899–1911.
[12] J.C. Lagier, F. Armougom, M. Million, et al., Microbial culturomics: paradigm shift in the human gut microbiome study, Clin. Microbiol. Infect. 18 (2012) 1185–1193.
[13] T. Clavel, C. Charrier, M. Wenning, et al., *Parvibacter caecicola* gen. nov., sp. nov., a bacterium of the family *Coriobacteriaceae* isolated from the caecum of a mouse, Int. J. Syst. Evol. Microbiol. 63 (2013) 2642–2648.
[14] Classification of bacteria, http://www.bacterio.net/-classifphyla.html#Actinobacteria, (Accessed: 06 March 2016).
[15] J.Z. Chan, M.R. Halachev, N.J. Loman, et al., Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*, BMC Microbiol. 12 (2012) 302.
[16] P. Vandamme, B. Pot, M. Gillis, et al., Polyphasic taxonomy, a consensus approach to bacterial systematics, Microbiol. Rev. 60 (1996) 407–438.
[17] P.E. Fournier, M. Drancourt, New Microbes New Infections promotes modern prokaryotic taxonomy: a new section TaxonoGenomics: new genomes of microorganisms in humans, New Microbes New Infect. 7 (2015) 48–49.
[18] P.E. Fournier, J.C. Lagier, G. Dubourg, D. Raoult, From culturomics to taxonomogenomics: a need to change the taxonomy of prokaryotes in clinical microbiology, Anaerobe 36 (2015) 73–78.
[19] J.P. Menard, F. Fenollar, M. Henry, et al., Molecular quantification of *Gardnerella vaginalis* and *Atopobium vaginae* loads to predict bacterial vaginosis, Clin. Infect. Dis. 47 (2008) 33–43.
[20] P. Seng, M. Drancourt, F. Gouriet, et al., Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry, Clin. Infect. Dis. 49 (2009) 543–551.
[21] M. Drancourt, C. Bollet, A. Carlioz, et al., 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates, J. Clin. Microbiol. 38 (2000) 3623–3630.
[22] E. Stackebrandt, J. Ebers, Taxonomic parameters revisited: tarnished gold standards, Microbiol. Today 33 (2006) 152–155.
[23] M. Kim, H.-S. Oh, S.-C. Park, J. Chun, Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes, Int. J. Syst. Evol. Microbiol. 64 (2014) 346–351.
[24] P. Yarza, P. Yilmaz, E. Pruesse, et al., Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences, Nat. Rev. Micro 12 (2014) 635–645.
[25] These standards outline the test procedures for identifying the organisms, https://www.gov.uk/government/collections/standards-for-microbiology-investigations-smi#test-procedures, (Accessed 06 March 2016).
[26] M. Sasser, Bacterial Identification by Gas Chromatographic Analysis of Fatty Acids Methyl Ester (GC-FAME), 1990.
[27] E. Matuschek, D.F.J. Brown, G. Kahlmeter, Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories, Clin. Microbiol. Infect. 20 (2014) 255–266.
[28] D.M. Citron, M.I. Ostovari, A. Karlsson, E.J. Goldstein, Evaluation of the E test for susceptibility testing of anaerobic bacteria, J. Clin. Microbiol. 29 (1991) 2197–2203.
[29] The European Committee on Antimicrobial Susceptibility Testing, Breakpoint Tables for Interpretation of MICs and Zone Diameters. http://www.eucast.org (Accessed 22 November 2016).
[30] D. Hyatt, G.L. Chen, P.F. Locascio, et al., Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinforma. 11 (2010) 119.
[31] K. Clark, I. Karsch-Mizrachi, D.J. Lipman, et al., GenBank, Nucleic Acids Res. 44 (2016) D67–D72.
[32] K. Lagesen, P. Hallin, E.A. Rødland, et al., RNAmmer: consistent and rapid annotation of ribosomal RNA genes, Nucleic Acids Res. 35 (2007) 3100–3108.
[33] T.M. Lowe, S.R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res. 25 (1997) 955–964.
[34] T. Carver, S.R. Harris, M. Berriman, et al., Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data, Bioinformatics 28 (2012) 464–469.
[35] T. Carver, N. Thomson, A. Bleasby, et al., DNAPlotter: circular and linear interactive genome visualization, Bioinformatics 25 (2009) 119–120.
[36] D. Ramasamy, A.K. Mishra, J.C. Lagier, et al., A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species, Int. J. Syst. Evol. Microbiol. 64 (2014) 384–391.
[37] M. Lechner, S. Findeiss, L. Steiner, et al., Proteinortho: detection of (Co-) orthologs in large-scale analysis, BMC Bioinforma. 12 (2011) 124.
[38] P. Gouret, J. Paganini, J. Dainat, et al., Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the Multi-Agent Software System DAGOBAH, in: P. Pontarotti (Ed.), Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution, Springer Berlin Heidelberg, 2011, pp. 71–87.
[39] P. Gouret, V. Vitiello, N. Balandraud, et al., FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform, BMC Bioinforma. 6 (2005) 198.
[40] J.P. Meier-Kolthoff, A.F. Auch, H.P. Klenk, et al., Genome sequence-based species delimitation with confidence intervals and improved distance functions, BMC Bioinforma. 14 (2013) 60.
[41] P. Cools, M. Oyaert, M. Vaneechoutte, et al., *Atopobium deltae* sp. nov., isolated from the blood of a patient with Fournier's gangrene, Int. J. Syst. Evol. Microbiol. 64 (2014) 3140–3145.
[42] F.E. Dewhirst, B.J. Paster, N. Tzellas, et al., Characterization of novel human oral isolates and cloned 16S rDNA sequences that fall in the family *Coriobacteriaceae*: description of *Olsenella* gen. nov., reclassification of *Lactobacillus uli* as *Olsenella uli* comb. nov. and description of *Olsenella profusa* sp. nov. Int. J. Syst. Evol. Microbiol. 51 (2001) 1797–1804.
[43] G.D. Bailey, D.N. Love, *Eubacterium fossor* sp. nov., an agar-corroding organism from normal pharynx and oral and respiratory tract lesions of horses, Int. J. Syst. Evol. Microbiol. 36 (1986) 383–387.
[44] I. Olsen, J.L. Johnson, L.V.H. Moore, et al., *Lactobacillus uli* sp. nov. and *Lactobacillus rimae* sp. nov. from the human gingival crevice and emended descriptions of *Lactobacillus minutus* and *Streptococcus parvulus*, Int. J. Syst. Evol. Microbiol. 41 (1991) 261–266.
[45] M. Kraatz, R.J. Wallace, L. Svensson, *Olsenella umbonata* sp. nov., a microaerotolerant anaerobic lactic acid bacterium from the sheep rumen and pig jejunum, and emended descriptions of *Olsenella*, *Olsenella uli* and *Olsenella profusa*, Int. J. Syst. Evol. Microbiol. 61 (2011) 795–803.
[46] F. Nagai, Y. Watanabe, M. Morotomi, *Slackia piriformis* sp. nov. and *Collinsella tanakaei* sp. nov., new members of the family *Coriobacteriaceae*, isolated from human faeces, Int. J. Syst. Evol. Microbiol. 60 (2010) 2639–2646.
[47] A.P. Vieira Colombo, C.B. Magalhães, F.A. Hartenbach, et al., Periodontal-disease-associated biofilm: a reservoir for pathogens of medical importance, Microb. Pathog. 94 (2015) 27–34.

Article 11:

Microbial Culturomics Broadens Human Vaginal Flora Diversity: Genome Sequence and Description of *Prevotella lascolaii* sp. nov., a new species isolated from the genital tract of a patient with bacterial vaginosis

Diop Kh, Diop A, Levasseur A, Mediannikov O, Robert C, Couderc C, Bretelle F, Raoult D, Fournier PE and Fenollar F

[Published in OMICS]

OMICS A Journal of Integrative Biology
Volume 22, Number 3, 2018

DOI: 10.1089/omi.2017.0151

Microbial Culturomics Broadens Human Vaginal Flora Diversity: Genome Sequence and Description of *Prevotella lascolaii* sp. nov. Isolated from a Patient with Bacterial Vaginosis

Khoudia Diop,[1] Awa Diop,[1] Anthony Levasseur,[1] Oleg Mediannikov,[1] Catherine Robert,[1] Nicholas Armstrong,[1] Carine Couderc,[1] Florence Bretelle,[2] Didier Raoult,[1,3] Pierre-Edouard Fournier,[1] and Florence Fenollar[1]

Abstract

Microbial culturomics is a new subfield of postgenomic medicine and omics biotechnology application that has broadened our awareness on bacterial diversity of the human microbiome, including the human vaginal flora bacterial diversity. Using culturomics, a new obligate anaerobic Gram-stain-negative rod-shaped bacterium designated strain khD1^T was isolated in the vagina of a patient with bacterial vaginosis and characterized using taxonogenomics. The most abundant cellular fatty acids were $C_{15:0}$ anteiso (36%), $C_{16:0}$ (19%), and $C_{15:0}$ iso (10%). Based on an analysis of the full-length 16S rRNA gene sequences, phylogenetic analysis showed that the strain khD1^T exhibited 90% sequence similarity with *Prevotella loescheii*, the phylogenetically closest validated *Prevotella* species. With 3,763,057 bp length, the genome of strain khD1^T contained (mol%) 48.7 G + C and 3248 predicted genes, including 3194 protein-coding and 54 RNA genes. Given the phenotypical and biochemical characteristic results as well as genome sequencing, strain khD1^T is considered to represent a novel species within the genus *Prevotella*, for which the name *Prevotella lascolaii* sp. nov. is proposed. The type strain is khD1^T (=CSUR P0109, =DSM 101754). These results show that microbial culturomics greatly improves the characterization of the human microbiome repertoire by isolating potential putative new species. Further studies will certainly clarify the microbial mechanisms of pathogenesis of these new microbes and their role in health and disease. Microbial culturomics is an important new addition to the diagnostic medicine toolbox and warrants attention in future medical, global health, and integrative biology postgraduate teaching curricula.

Keywords: culturomics, taxonogenomics, *Prevotella lascolaii*, bacterial vaginosis, microbiome science

Introduction

The symbiotic relationship between humans and their associated bacteria plays a crucial role in their health. Changes in the proportion of microbial species in the vagina predispose that person to dysbioses such as bacterial vaginosis (BV) (Narayankhedkar et al., 2015). First studies using traditional culture methods identified only 20% of bacteria present in the vagina (Lamont et al., 2011). The vaginal flora diversity has been revealed further using molecular methods, sequencing, and phylogenetic analysis of the 16S rRNA gene, which show the detection of fastidious and uncultured bacteria, such as bacterial vaginosis-associated bacteria type 1 (BVAB1), BVAB2, and BVAB3 (Fredricks et al., 2005).

Recently, a new approach named "Microbial Culturomics," involving high-throughput culture conditions and matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) for bacterial identification, was initiated and used to study the human microbiota (Dubourg et al., 2013; Lagier et al., 2012). Culturomics broadened our awareness about the bacterial diversity of the human microbiome by analyzing different samples (such as stool, small-bowel, and colonic samples) from healthy individuals and patients with various diseases (such as anorexia nervosa, obesity, malnutrition,

[1]Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, UM 63, CNRS UMR 7278, IRD 198, INSERM U1095, Institut Hospitalo-Universitaire Méditerranée-Infection, Faculté de Médecine, Aix-Marseille University, Marseille, France.
[2]Department of Gynecology and Obstetrics, Gynépole, Hôpital Nord, Assistance Publique-Hôpitaux de Marseille, Marseille, France.
[3]Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia.

and HIV) from different geographical origins (Europe, rural and urban Africa, Polynesia, India, …) (Lagier et al., 2016).

In addition to improving culture and bacterial identification, culturomics is used with a new classification and nomenclature concept called taxonogenomics to better characterize and describe bacterial species (Fournier and Drancourt, 2015; Fournier et al., 2015). Taxogenomics combines classic bacterial description and phenotypic/genotypic characteristics such as DNA-DNA hybridization with the proteomic information obtained by MALDI-TOF mass spectrometry (MS) and the description of the complete genome.

We isolated a new member of the genus *Prevotella* in a culturomics study of the vaginal flora, which aimed to map the vaginal microbiome in healthy women and patients with BV to identify bacteria involved in this dysbiosis. Amended in 2012 (Sakamoto and Ohkuma, 2012), the *Prevotella* genus was created in 1990 by reclassifying some *Bacteroides* species. It contains gram-negative rod, strict anaerobic, nonspore forming, and nonmotile bacteria with *Prevotella melaninogenica* as the type strain (Shah and Collins, 1990).

Here follows the description of *Prevotella lascolaii* strain khD1^T (=CSUR P0109, =DSM 101754) with its annotated whole genome, isolated in the vaginal flora of a patient suffering from BV.

Materials and Methods

Ethics and sample collection

The vaginal sample of a 33-year-old French woman was collected at Nord Hospital in Marseille (France) in October 2015 using a Sigma Transwab (Medical Wire, Corsham, United Kingdom). As previously described (Menard et al., 2008), the patient was suffering from abnormal vaginal discharge and diagnosed with BV. During the sample collection, she was not treated with any antibiotics and she signed a written consent. The local ethics committee of the IFR48 (Marseille, France) had also authorized this study under agreement number 09-022.

Strain identification by MALDI-TOF MS

Initially, the vaginal sample was inoculated in a blood culture bottle (BD Diagnostics, Le Pont-de-Claix, France) enriched with sheep's blood (bioMérieux, Marcy l'Etoile, France) and rumen filtered at 0.2 μm by a pore filter (Thermo Fisher Scientific, Villebon-sur-Yvette, France). Fifty microliters of the supernatant was plated onto Schaedler agar enriched with vitamin K and sheep's blood (BD Diagnostics). Then, after 4 days of incubation at 37°C in anaerobic conditions, purified colonies were deposited on an MALDI-TOF target plate (Bruker Daltonics, Leipzig, Germany) in duplicate and, as previously described, 1.5 μL of matrix (Seck et al., 2015) was added on each spot.

The identification was carried out using a Microflex spectrometer (Bruker) (Seng et al., 2009), which compares identified protein spectra to those on the MALDI-TOF database containing 7567 references (composed of the Bruker database incremented with our data). The reliability of bacteria identification was indicated by a score. If the score was greater than 1.9, the bacterium was considered identified. Conversely, if the score was less than 1.9 it was not in the database or identification failed.

Strain identification by 16S rRNA sequencing

To identify unidentified bacterium, the 16S rRNA gene was sequenced using fD1-rP2 primers (Eurogentec, Angers, France). The obtained sequence was corrected using ChromasPro 1.34 software (Technelysium Pty. Ltd., Tewantin, Australia) and matched against the NCBI database using the BLAST algorithm (Drancourt et al., 2000).

Phylogenetic tree

All species from the same genus of the new species were retrieved and 16S sequences were downloaded from NCBI. Sequences were aligned using CLUSTALW, with default parameters and phylogenetic inferences obtained using the neighbor-joining method with 500 bootstrap replicates, using MEGA6 software.

Growth conditions

To evaluate ideal growth, the strain khD1^T was cultivated on Columbia agar with 5% sheep's blood and incubated at different temperatures (25°C, 28°C, 37°C, 45°C, and 56°C) in an aerobic atmosphere with or without 5% CO_2 and also in anaerobic and microaerophilic atmospheres using GENbag anaer and GENbag miroaer systems (bioMérieux), respectively.

Morphology

To observe cell morphology, cells were fixed with 2.5% glutaraldehyde in a 0.1 M cacodylate buffer for at least an hour at 4°C. One drop of cell suspension was deposited for ~5 min on glow-discharged formvar carbon film on 400 mesh nickel grids (FCF400-Ni; EMS). The grids were dried on blotting paper and the cells were negatively stained for 10 sec with 1% ammonium molybdate solution in filtered water at RT. Electron micrographs were acquired using a Tecnai G20 Cryo (FEI) transmission electron microscope operated at 200 keV. Gram staining, motility, and sporulation were performed as previously conducted (Murray et al., 2007).

Biochemical analysis

The biochemical characteristics of strain khD1^T have been determined using the API ZYM, 20A, and 50CH strips (bioMérieux) according to the manufacturer's instructions. API ZYM was performed for the research of enzymatic activities. It allows the systematic and rapid study of 19 enzymatic reactions using very small sample quantities. While API 20A (20 cupules) was used for the biochemical identification of the isolate and 50CH API (50 cupules) to study carbohydrate metabolism.

Cellular fatty acid methyl ester (FAME) analysis was performed using gas chromatography/mass spectrometry (GC/MS). Two samples were prepared with ~35 mg of bacterial biomass per tube harvested from several culture plates. FAMEs were prepared as described by Sasser (Sasser, 2006). First, fatty acids were released from lipids with a saponification step at 100°C during 30 min in the presence of 1 mL NaOH 3.75 M in water/methanol (50% v:v). Then, free fatty acids were transformed to methyl esters at 80°C during 10 min after adding 2 mL of HCl 6 N/methanol (54/46% v:v). The resulting FAMEs were then extracted in 1 mL of hexane/MTBE (50% v:v). Organic extracts were finally washed with 3 mL of NaOH 0.3 M to

remove free acids. GC/MS analyses were carried out using a Clarus 500 gas chromatograph connected to a SQ8S single quadrupole MS detector (Perkin Elmer, Courtaboeuf, France).

Two microliters of both FAME extracts were volatized at 250°C (split 20 mL/min) in a Focus liner with wool. Compounds were separated on an Elite-5MS column (30 m, 0.25 mm i.d., 0.25 mm film thickness) using a linear temperature gradient (70–290°C at 6°C/min) enabling the detection of C4 to C24 FAMEs. Helium flowing at 1.2 mL/min was used as carrier gas. MS inlet line was set at 250°C and electron ionization source at 200°C. Full scan monitoring was performed from 45 to 500 m/z. All data were collected and processed using Turbomass 6.1 (Perkin Elmer).

FAMEs were identified using the identity spectrum search using the MS Search 2.0 software, operated with the Standard Reference Database 1A (NIST, Gaithersburg, USA) and the FAME mass spectral database (Wiley, Chichester, United Kingdom). A 37-component FAME mix (Supelco; Sigma-Aldrich, Saint-Quentin Fallavier, France) was used to calculate the correlation between chromatographic retention times and nonpolar retention indexes from the NIST database. MS Search identifications were therefore validated if reverse/forward search scores were above 750 and if nonpolar retention indexes were correlated to the chromatographic retention time.

Antibiotic susceptibility tests

Amoxicillin, benzylpenicillin, imipenem, metronidazole, and vancomycin were used to test antibiotic susceptibility of strain khD1^T. The minimal inhibitory concentrations (MICs) were then determined using E-test gradient strips (bioMérieux) according to the EUCAST recommendations (Citron et al., 1991; Matuschek et al., 2014).

Genomic DNA preparation

Strain khD1^T was cultured on 5% sheep's blood-enriched Columbia agar (bioMérieux) at 37°C anaerobically. Bacteria grown on three Petri dishes were resuspended in $4 \times 100\ \mu$L of Tris–EDTA (TE) buffer. Next, 200 μL of this suspension was diluted in 1 mL TE buffer for lysis treatment, which included a 30-min incubation with 2.5 μg/μL lysozyme at 37°C, followed by an overnight incubation with 20 μg/μL proteinase K at 37°C. Extracted DNA was then purified using

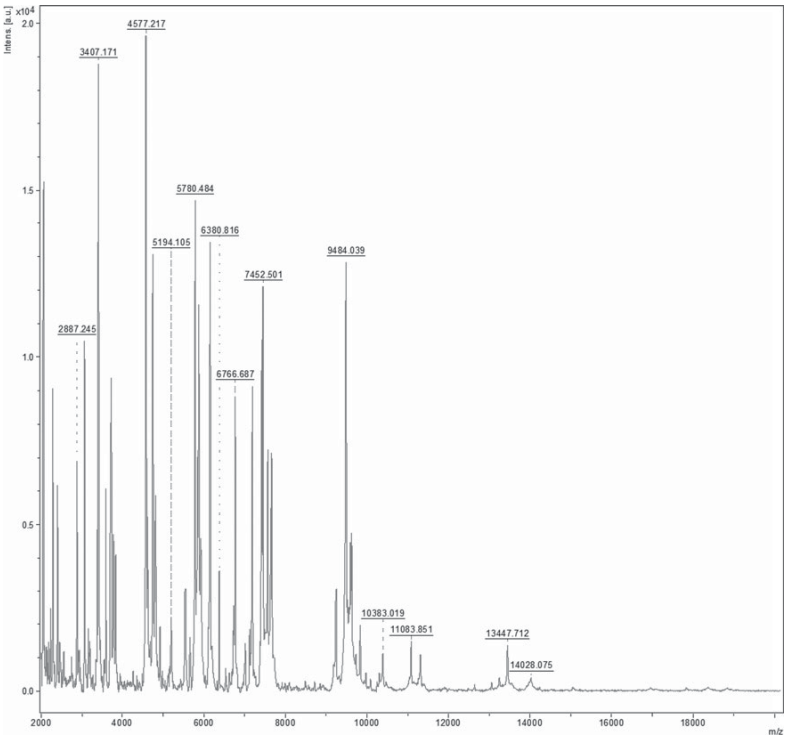


FIG. 1. Reference mass spectrum from the *Prevotella lascolaii* strain khD1^T.

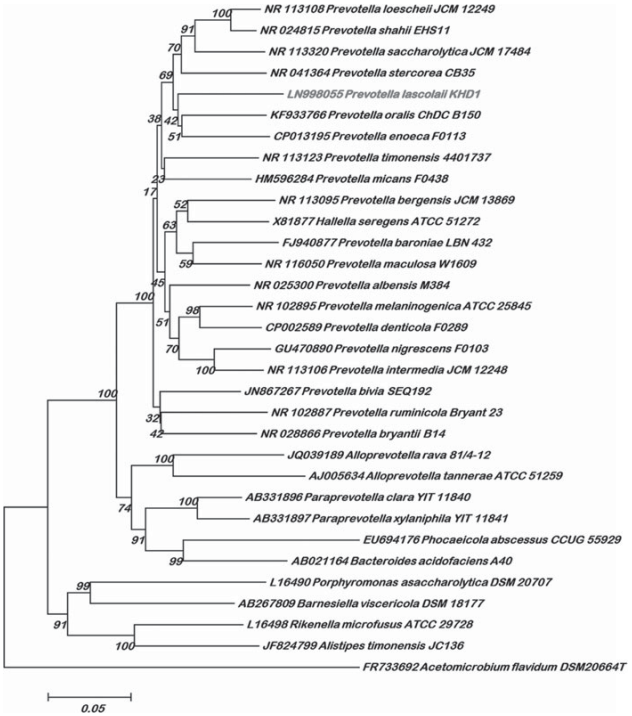


FIG. 2. Phylogenetic tree highlighting the position of *Prevotella lascolaii* strain khD1^T relative to other close strains. GenBank accession numbers of each 16S rRNA are noted before the name. Sequences were aligned using Muscle v3.8.31 with default parameters, and phylogenetic inferences were obtained using the neighbor-joining method with 500 bootstrap replicates, within MEGA6 software. The scale bar represents a 0.05% nucleotide sequence divergence.

three successive phenol–chloroform extractions and ethanol precipitations at −20°C overnight. After centrifugation, the DNA was resuspended in 160 μL TE buffer.

Genome sequencing and assembly

Genomic DNA (gDNA) of strain khD1^T was sequenced on the MiSeq Technology (Illumina, Inc., San Diego, CA, USA) with the mate-pair strategy. The gDNA was barcoded with the Nextera Mate-Pair sample prep kit (Illumina) to be mixed with 11 other projects.

gDNA was quantified by a Qubit assay with a high-sensitivity kit (Life technologies, Carlsbad, CA, USA) to 105.7 ng/μL. The mate-pair library was prepared with 1.5 μg of genomic DNA using the Nextera mate-pair Illumina guide. The genomic DNA sample was simultaneously fragmented and tagged with a mate-pair junction adapter. The pattern of fragmentation was validated on an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA, USA) with a

Table 1. Classification and General Features of *Prevotella lascolaii* Strain khD1^T

	Term
Current classification	Domain: *Bacteria* Phylum: *Bacteroidetes* Class: *Bacteroidia* Order: *Bacteroidales* Family: *Prevotellaceae* Genus: *Prevotella* Species: *Prevotella lascolaii* Type strain: khD1
Gram stain	Negative
Cell shape	Rod
Motility	Nonmotile
Sporulation	Nonsporulating
Temperature range	Anaerobic
Optimum temperature	37°C

DNA 7500 LabChip. The DNA fragments ranged in size from 1.5 to 11 kb with an optimal size at 5.203 kb. No size selection was performed and 440 ng of tagmented fragments were circularized.

The circularized DNA was mechanically sheared to small fragments with an optimal size of 985 bp on the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The library profile was viewed on a High-Sensitivity Bioanalyzer LabChip (Agilent Technologies, Inc., Santa Clara, CA, USA) and the final concentration library was measured at 4.17 nM.

The libraries were normalized at 2 nM and pooled. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated cluster generation and sequencing run were performed in a single 39-h run in a 2×151 bp.

Total information of 8.8 Gb was obtained from a 971 K/mm^2 cluster density with a cluster passing quality control filters of 93.1% (17,376,000 passing filter paired reads). Within this run, the index representation for strain khD1^T was determined to be 7.17%. The 1,246,384 paired reads were trimmed and then assembled in 27 scaffolds.

Genome annotation and analysis

Open reading frames (ORFs) were predicted using Prodigal software (Hyatt et al., 2010) with default parameters. Predicted ORFs spanning a sequencing gap region (containing N) were excluded. We predicted the bacterial proteome using BLASTP (E-value 1e-03 coverage 0.7 and identity percent 30) against the Clusters of Orthologous Groups (COGs) database. A search against the NR database (Clark et al., 2016) was performed if no hit was found, using BLASTP with E-value of 1e-03 coverage 0.7 and an identity percent of 30. An E-value of 1e-05 was used with sequence lengths smaller than 80 amino acids. The hhmscan tool analyses were used for searching PFam conserved domains (PFAM-A and PFAM-B domains) on each protein.

We used RNAmmer (Lagesen et al., 2007) and tRNAScanSE tools (Lowe and Eddy, 1997) to find ribosomal RNA genes and tRNA genes, respectively. Viewing and data managing genomic features were performed using Artemis (Carver et al., 2012) and DNA Plotter (Carver et al., 2009), respectively. For the mean level of nucleotide sequence similarity analysis at the genome level, we used the MAGI home-made software. It calculated the average genomic identity of gene sequences (AGIOS) among compared genomes (Ramasamy et al., 2014). The Proteinortho (Lechner et al., 2011) software was incorporated with the MAGI home-made software for detecting orthologous proteins in pair-wise genomic comparisons. Next, the corresponding genes were retrieved and the mean percentage of nucleotide sequence identity among orthologous ORFs was determined using the Needleman–Wunsch global alignment algorithm.

The Multi-Agent Software System DAGOBAH (Gouret et al., 2011) was used to perform annotation and comparison processes, which included Figenix (Gouret et al., 2005) libraries providing pipeline analysis. GGDC analysis was performed using the GGDC web server as previously reported (Meier-Kolthoff et al., 2013).

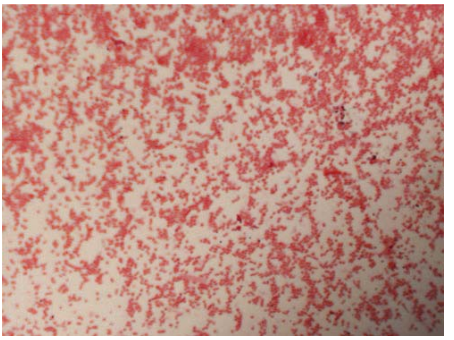


FIG. 3. Gram staining of *Prevotella lascolaii* strain khD1^T.

Results

Strain identification and phylogenetic analysis

P. lascolaii strain khD1^T was first isolated after 24 h preincubation of the vaginal sample in a blood culture bottle enriched with rumen, which was filter sterilized through a 0.2 μm pore filter (Thermo Fisher Scientific), and sheep's blood (bioMérieux) under anaerobic conditions at 37°C. Then, 50 μL of the supernatant was inoculated on Schaedler agar enriched with sheep's blood and vitamin K (BD Diagnostics) in the same conditions for 4 days. The MALDI-TOF identification gave us a score of 1.3. As the strain was not in the database, the reference spectrum (Fig. 1) was incremented in our database and the gene 16S rRNA was sequenced.

The sequence obtained (number accession LN998055) exhibited 90% similarity with *Prevotella loescheii*, the phylogenetically closest bacterial species with a validly published

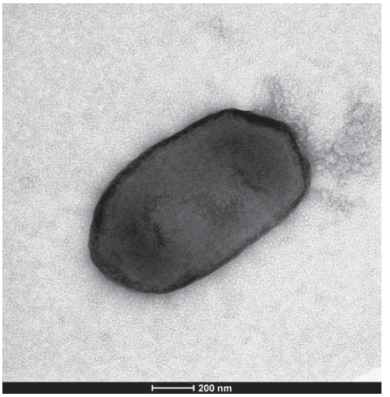


FIG. 4. Electron micrographs of *Prevotella lascolaii* strain khD1^T using a Tecnai G^{20} Cryo (FEI) transmission electron microscope operated at 200 keV. The scale bar represents 200 nm.

TABLE 2. PHENOTYPIC CHARACTERISTICS THAT DIFFERENTIATE *PREVOTELLA LASCOLAII* STRAIN KHD1^T SP. NOV. FROM RELATED *PREVOTELLA* SPECIES, *P. LOESCHEII*, *P. SHAHII*, *P. ORALIS*, *P. STERCOREA*, *P. ENOECA*, *P. TIMONENSIS*, AND *P. MICANS*

Characteristic	Prevotella lascolaii	Prevotella loescheii	Prevotella shahii	Prevotella oralis	Prevotella stercorea	Prevotella enoeca	Prevotella timonensis	Prevotella micans
Cell diameter (μm)	0.3–0.5	0.4–0.6	0.5–0.8	0.5–1	0.25–0.42	0.5	0.8–1.4	0.7
Endospore formation	−	−	−	−	−	−	−	na
Indole	−	−	−	−	−	−	na	+
Production of								
Alkaline phosphatase	+	na	+	na	+	na	+	+
Catalase	−	−	−	−	−	−	na	−
Nitrate reductase	−	−	−	−	na	−	na	−
Urease	−	na	−	−	−	na	na	na
β-galactosidase	+	na	+	na	+	na	+	+
N-acetyl-glucosamine	+	na	+	na	+	na	+	+
Production of								
L-arabinose	+	−	−	−	−	−	+	−
Ribose	+	−	na	na	na	−	+	−
Mannose	−	+	+	+	+	+	−	+
Sucrose	−	+	+	+	+	−	+	+
D-glucose	−	+	+	+	+	+	+	+
D-fructose	−	+	na	+	na	+	na	+
D-maltose	−	+	+	+	+	+	+	+
D-lactose	−	+	+	+	+	+	+	+
Major cellular fatty acids[a]	$C_{15:0}$ anteiso, $C_{16:0}$, $C_{15:0}$ iso	$C_{15:0}$ anteiso, $C_{18:1n9c}$, $C_{15:0}$ iso	$C_{18:1n9c}$, $C_{16:0}$, $C_{16:0}$ 3-OH	$C_{16:0}$, $C_{18:1n9c}$, $C_{16:0}$ 3-OH, $C_{15:0}$ anteiso	$C_{18:1n9c}$, $C_{15:0}$ iso, $C_{15:0}$ anteiso	$C_{15:0}$ anteiso, $C_{16:0}$, $C_{16:0}$ 3-OH, $C_{15:0}$ iso	$C_{14:0}$, $C_{16:0}$, $C_{18:2\ n6,9c}$/ $C_{18:0}$	na
G+C content (mol%)	48.7	46.9	44.3	43.1	48.2	47	40.50	46
Habitat	Human vagina	Human oral cavity	Human oral cavity	Human oral cavity	Human feces	Human gingiva	Breast abscess	Human oral cavity

The reference for the species data comes from descriptions of the original species. +, −, and na data.
[a]Major cellular fatty acids listed in order of predominance.
+, positive; −, negative; na, not available.

TABLE 3. CELLULAR FATTY ACID COMPOSITION (%) OF *PREVOTELLA LASCOLAII* STRAIN KHD1^T (DATA FROM THIS STUDY) COMPARED WITH CLOSEST SPECIES

Fatty acids	*Name*	Prevotella lascolaii	Prevotella loescheii	Prevotella shahii	Prevotella oralis	Prevotella stercorea	Prevotella enoeca	Prevotella timonensis
Saturated straight chain								
14:00	Tetradecanoic acid	1.5	1.1	10.9	2.1	0.8	4	**19.5**
15:00	Pentadecanoic acid	tr	3.8	1.0	tr	tr	na	na
16:00	Hexadecanoic acid	**18.8**	**12.5**	**16.9**	**19.2**	3.8	**17**	**15.3**
17:00	Heptadecanoic acid	tr	1.5	na	tr	na	na	na
18:00	Octadecanoic acid	tr	0.9	2.8	0.9	0.8	na	**16**
Unsaturated straight chain								
18:1n9	9-Octadecenoic acid	2.3	**15.0**	**18.7**	**18.6**	**14.7**	na	na
18:2n6	9,12-Octadecadienoic acid	4.0	2.0	na	na	2,2	na	**16**
20:4n6	5,8,11,14-Eicosatetraenoic acid	tr	na	na	na	na	na	na
Hydroxy acids								
16:0 3-OH	3-hydroxy-hexadecanoic acid	4.4	6.1	**16.3**	**10.4**	1	**10**	na
17:0 3-OH	3-hydroxy-heptadecanoic acid	7.7	na	na	na	na	na	na
18:0 3-OH	3-hydroxy-octadecanoic acid	tr	na	na	na	na	na	na
Saturated branched chain								
5:0 anteiso	2-methyl-butanoic acid	tr	na	na	na	na	na	na
14:0 iso	12-methyl-tridecanoic acid	1.5	2.1	4.4	3.0	2.7	3	**14**
15:0 iso	13-methyl-tetradecanoic acid	**9.9**	3.2	3.4	3.2	**23.7**	8	na
15:0 anteiso	12-methyl-tetradecanoic acid	**36.1**	**24.0**	6.8	**20.6**	**26.2**	**36**	na
16:0 iso	14-methyl-pentadecanoic acid	3.2	0.8	1.0	1.7	2.7	na	na
17:0 iso	15-methyl-hexadecanoic acid	4.8	1.1	na	tr	1.7	2	na
17:0 anteiso	14-methyl-hexadecanoic acid	4.3	1.7	na	1.5	1.3	na	na

Bold represents the majority fatty acid for this species; na, not available data; tr, trace amounts <1%. The reference for the species data comes from descriptions of the original species. *P. micans* was not listed because its complete fatty acid profile was not available.

Table 4. Nucleotide Content and Gene Count Levels of the Genome

Attribute	*Value*	*% of total*[a]
Size (bp)	3,763,057	100
G+C content (bp)	1,832,608	48.7
Coding region (bp)	3,186,418	84.67
Total genes	3248	100
RNA genes	54	1.60
Protein-coding genes	3194	98.33
Genes with function prediction	2034	63.68
Genes assigned to COGs	1691	52.9
Genes with peptide signals	643	20.13
Genes with transmembrane helices	2541	79.55

[a]The total is based on either the size of the genome in base pairs or the total number of protein-coding genes in the annotated genome.
COG, Clusters of Orthologous Group.

name (Fig. 2). Thus, as this value was under the threshold of 98.7%, established to delineate a new species (Kim et al., 2014; Stackebrandt and Ebers, 2006), strain khD1^T was classified as a new species within the *Prevotella* genus and named *P. lascolaii* (Table 1).

Phenotypic and biochemical characteristics

Cultivated on Columbia agar (bioMérieux) for 48 h in anaerobic conditions at 37°C, *P. lascolaii* strain khD1^T colonies were grayish-white, shiny, smooth, and circular with a diameter of 1.4 to 2 mm. Gram staining showed gram-negative short rod-shaped bacilli or coccobacilli (Fig. 3). Under electronic microscopy, individual cells had a mean diameter of 0.65 μm and mean length of 0.9 μm (Fig. 4). Nonmotile and nonspore-forming, *P. lascolaii* exhibited positive oxidase activity. Nevertheless, catalase activity was negative and nitrate was not reduced. Strictly anaerobic, strain khD1^T grows at temperatures between 25°C and 42°C, with optimal growth at 37°C after 48 h of incubation. Its growth also needs an NaCl concentration under 5 g/L and pH ranging from 6.5 to 8.5.

API ZYM strips revealed that strain khD1^T exhibited positive reactions for alkaline phosphatase, α-chymotrypsin, acid phosphatase, naphthol-AS-BI-phosphohydrolase, galactosidase (α and β), glucosidase (α and β), N-acetyl-β-glucosaminidase, and α-fucosidase enzymes. However, esterase, esterase lipase, lipase, leucine, cystine and valine arylamidase, trypsin, β-glucuronidase, and α-fucosidase were negative. API 50CH shows that strain khD1^T ferments arabinose, ribose, galactose, methyl-αD-mannopyranoside, β-galactosidase, melezitose, glycogen, turanose, tagose, and potassium 5-ketogluconate.

In contrast, arabinose, xylose, glucose, fructose, mannose, mannitol, cellobiose, maltose, lactose, sucrose, and starch were not metabolized. The same results were also observed using API 20A; ferric citrate esculin was hydrolyzed, but urease was not exhibited and carboxylates were not fermented. These

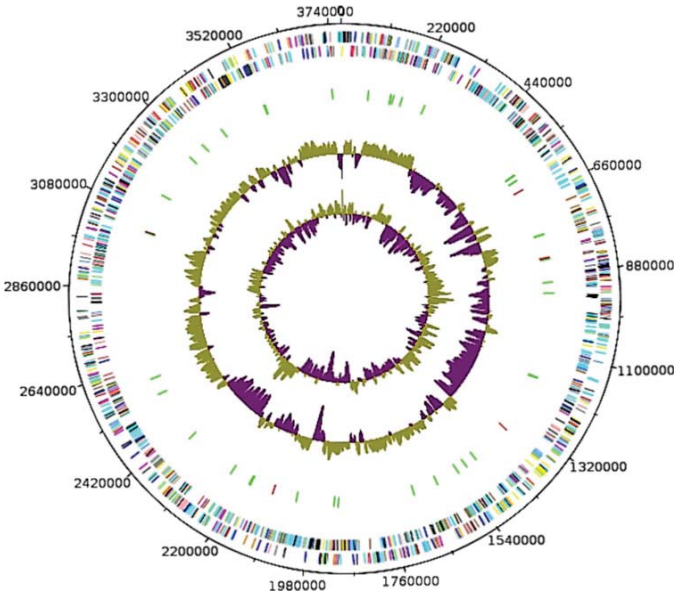


FIG. 5. Graphical circular map of the chromosome. From outside to the center: Contigs (*red/gray*), COG category of genes on the forward strand (three *circles*), genes on forward strand (*blue circle*), genes on the reverse strand (*red circle*), COG category on the reverse strand (three *circles*), GC content. COG, Clusters of Orthologous Group.

Table 5. Number of Genes Associated with the 25 General Clusters of Orthologous Group Functional Categories

Code	*Value*	*% value*	*Description*
J	133	7.9	Translation
A	0	0	RNA processing and modification
K	88	5.2	Transcription
L	159	9.4	Replication, recombination, and repair
B	0	0	Chromatin structure and dynamics
D	25	1.5	Cell cycle control, mitosis, and meiosis
Y	0	0	Nuclear structure
V	53	3.1	Defense mechanisms
T	49	2.9	Signal transduction mechanisms
M	169	10.0	Cell wall/membrane biogenesis
N	4	0.2	Cell motility
Z	0	0	Cytoskeleton
W	0	0	Extracellular structures
U	31	1.8	Intracellular trafficking and secretion
O	68	4.0	Posttranslational modification, protein turnover, chaperones
X	14	0.8	Mobilome: prophages, transposons
C	83	4.9	Energy production and conversion
G	131	7.7	Carbohydrate transport and metabolism
E	114	6.7	Amino acid transport and metabolism
F	59	3.5	Nucleotide transport and metabolism
H	69	4. 1	Coenzyme transport and metabolism
I	46	2.7	Lipid transport and metabolism
P	77	4. 6	Inorganic ion transport and metabolism
Q	8	0.5	Secondary metabolite biosynthesis, transport, and catabolism
R	202	11.9	General function prediction only
S	109	6.4	Function unknown
—	1504	47.1	Not in COGs

phenotypic characteristics of *P. lascolaii* strain khD1^T are summarized in Table 2.

The major fatty acids of strain khD1T were similar to those found in members of *Prevotella* genus (Table 3) with saturated structures: 12-methyl-tetradecanoic acid (36%), hexadecanoic acid (19%), and 13-methyl-tetradecanoic acid (10%). Several branched structures and characteristic 3-hydroxy fatty acids were also described.

P. lascolaii khD1^T is sensitive to imipenem (MIC 0.47 μg/mL) and metronidazole (MIC 0.19 μg/mL) but resistant to amoxicillin (MIC >256 μg/mL), benzylpenicillin (MIC >256 μg/mL), and vancomycin (MIC 24 μg/mL).

Genome properties

The draft genome of *P. lascolaii* khD1^T (accession number FKKG00000000) is 3,763,057 bp long with 48.7% G+C content (Table 4). It contains 27 scaffolds assembled in 42 contigs (Fig. 5). Of the 3248 predicted genes, 3194 were protein-coding genes and 54 were RNAs (4 genes were 5S rRNA, 1 gene was 16S rRNA, 1 gene was 23S rRNA, and 47 genes were tRNA genes). A total of 2034 genes (63.68%) were assigned as putative functions (by cogs or NR blast). Two hundred twelve genes were identified as ORFans (6.63%). The remaining genes were annotated as hypothetical proteins (897 genes = >27.52%). Genome statistics is summarized in Table 4 and the distribution of the genes in COG functional categories is presented in Table 5.

Genomic comparison

The genome comparison of *P. lascolaii* strain khD1^T with the closest related species of *Prevotella* genus (Table 6) shows that the draft genome sequence of our strain (3.76 Mbp) was bigger than those of *Prevotella enoeca* and *Prevotella micans* (2.86 and 2.43 Mbp, respectively) but smaller than those of *P. loescheii* (7.01 Mbp). The G+C content of strain khD1^T (48.7 mol%) is larger than those of all the compared *Prevotella* species except *P. stercorea* (49 mol%). However, gene distribution in COG categories was similar among all compared genomes (Fig. 6). In addition, the AGIOS analysis revealed that strain khD1^T shares 975 orthologous genes with *P. micans* and 1285 with *Prevotella oralis*, whereas the analysis of the average percentage of nucleotide sequence identity ranged from 65.38% to 70.94% with *P. micans* and *P. stercorea*, respectively (Table 7). Similar results were also observed in the analysis of the digital DNA-DNA hybridization (dDDH) (Table 8).

Description of P. lascolaii *strain khD1^T sp. nov.*

P. lascolaii (las.co.la'ii N.L. gen. masc. n. *lascolaii* of La Scola, the family name of the French microbiologist Bernard La Scola) is strictly anaerobic and is nonmotile and nonspore forming. It has positive oxidase activity. No production of

Table 6. Genome Comparison of Closely Related Species with the *Prevotella lascolaii* Strain khD1^T

Species	*INSDC identifier*	*Genome size (Mbp)*	*G+C percent*	*Protein-coding genes*
Prevotella lascolaii strain khD1	FKKG00000000	3.76	48.7	3194
Prevotella stercorea DSM 18206	AFZZ00000000	6.19	49	2677
Prevotella oralis ATCC 33269	AEPE00000000	5.67	44.5	2353
Prevotella loescheii JCM 12249	ARJO00000000	7.01	46.6	2828
Prevotella enoeca JCM 12259	BAIX00000000	2.86	46.5	2806
Prevotella micans DSM 21469	BAKH00000000	2.43	45.5	2828
Prevotella shahii DSM 15611	BAIZ00000000	3.49	44.4	3371
Prevotella timonensis 4401737	CBQQ000000000	6.34	42.5	2685

INSDC, International Nucleotide Sequence Database Collaboration.

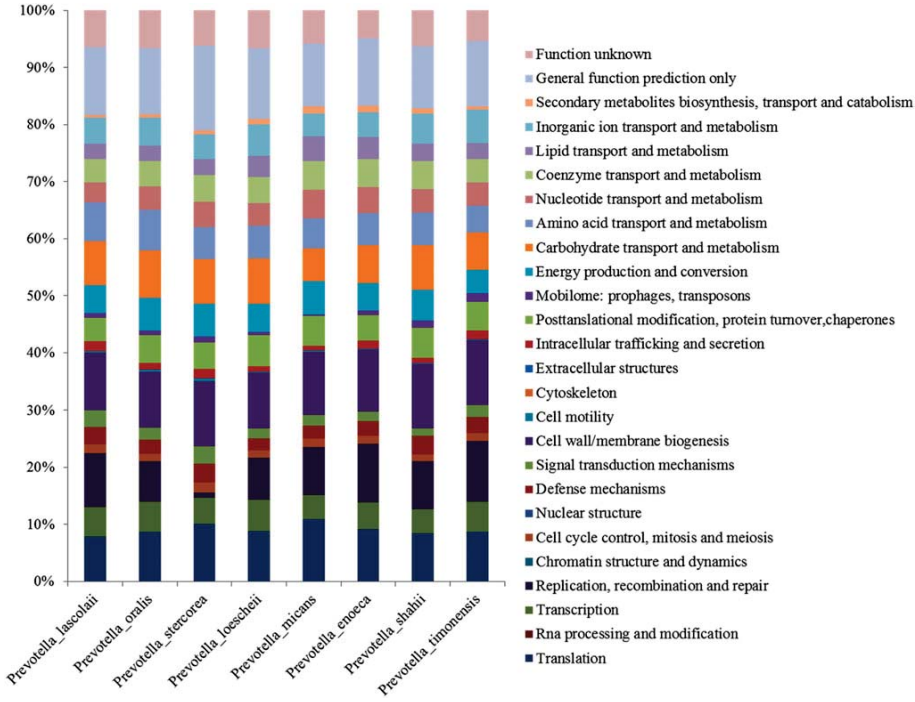


FIG. 6. Distribution of functional classes of predicted genes according to the COG of proteins of *Prevotella lascolaii* strain khD1^T among other species.

urease or catalase was observed. Cells are mesophilic, with optimal growth at 37°C, and are gram-negative bacilli with nearly 0.65 μm of diameter and 0.9 μm of length. On Columbia agar after 2 days of incubation at 37°C under anaerobic conditions, colonies appear grayish-white, shiny, smooth, and are circular with a diameter between 1.4 and 2 mm. It is moderately saccharolytic, and arabinose, ribose, galactose, melezitose are fermented while fructose, glucose, lactose, maltose, mannose, mannitol, raffinose, rhamnose, salicin, cellobiose, sucrose, trehalose, and xylose are not fermented. Ferric citrate esculin is hydrolyzed, but gelatin and urease are not hydrolyzed. Indole and catalase are not produced and nitrate is not reduced.

P. lascolaii exhibited positive enzymic reactions for alkaline phosphatase, α-chymotrypsin, acid phosphatase, naphthol-AS-BI-phosphohydrolase, α-galactosidase, β-galactosidase,

TABLE 7. NUMBERS OF ORTHOLOGOUS PROTEINS SHARED BETWEEN GENOMES (UPPER RIGHT) AND AVERAGE GENOMIC IDENTITY OF GENE SEQUENCE VALUES OBTAINED (LOWER LEFT)

	Prevotella lascolaii	Prevotella oralis	Prevotella stercorea	Prevotella loescheii	Prevotella micans	Prevotella enoeca	Prevotella shahii	Prevotella timonensis
P. lascolaii	**3194**	1285	1252	1255	975	1083	1099	1264
P. oralis	68.30%	**2353**	1226	1370	1038	1154	1185	1296
P. stercorea	70.94%	67.62%	**2677**	1228	968	1044	1057	1217
P. loescheii	67.21%	67.97%	67.78%	**2828**	1082	1215	1353	1311
P. micans	65.38%	66.46%	65.68%	65.70%	**2301**	935	956	983
P. enoeca	67.32%	69.23%	67.08%	68.27%	66.26%	**2806**	1055	1128
P. shahii	66.17%	67.21%	66.52%	81.03%	64.82%	67.33%	**3371**	1135
P. timonensis	66.97%	69.03%	66.71%	67.17%	65.66%	67.89%	67.02%	**2685**

The numbers of proteins per genome are indicated by bold numbers.

Table 8. Pairwise Comparison of *Prevotella lascolaii* with Other Species Using GGDC, Formula 2 (DDH Estimates Based on Identities/HSP length).[a]

	Prevotella lascolaii	Prevotella oralis	Prevotella stercorea	Prevotella loescheii	Prevotella micans	Prevotella enoeca	Prevotella shahii	Prevotella timonensis
P. lascolaii	100%	19.8% ± 2.3	31.6% ± 2.4	21.1% ± 2.3	20.0% ± 2.35	19.8% ± 2.3	22.4% ± 2.4	28.1% ± 2.4
P. oralis		100%	20.5% ± 2.3	19.7% ± 2.25	21.9% ± 2.35	20.0% ± 2.3	20.2% ± 2.35	21.0% ± 2.35
P. stercorea			100%	20.2% ± 2.3	21.4% ± 2.35	22.7% ± 2.35	21.5% ± 2.35	21.1% ± 2.4
P. loescheii				100%	24.0% ± 2.4	28.5% ± 2.45	24.9% ± 2.4	24.1% ± 2.4
P. micans					100%	29.4% ± 2.45	20.9% ± 2.3	25.2% ± 2.6
P. enoeca						100%	21.3% ± 2.3	24.0% ± 2.35
P. shahii							100%	25.7% ± 2.4
P. timonensis								100%

[a]The confidence intervals indicate the inherent uncertainty in estimating DDH values from intergenomic distances based on models derived from empirical test data sets (which are always limited in size). These results are in accordance with the 16S rRNA (Fig. 1) and phylogenomic analyses as well as the GGDC results.

DDH, DNA-DNA hybridization; HSP, high-scoring segment pairs.

α-glucosidase, β-glucosidase, N-acetyl-β-glucosaminidase, and α-fucosidase. The major fatty acids are $C_{15:0}$ anteiso (36%), $C_{16:0}$ (19%), and $C_{15:0}$ iso (10%).

P. lascolaii khD1^T is sensitive to imipenem and metronidazole but resistant to amoxicillin, benzylpenicillin, and vancomycin. Its genome contains 48.7%mol G+C and measured 3,763,057 bp long. The 16S rRNA and genome sequences are both deposited in GenBank under accession numbers LN998055 and FKKG00000000, respectively. The type strain khD1^T (=DSM 101754,=CSUR P0109) was isolated in the vaginal sample of a 33-year-old French woman afflicted with BV.

Discussion

Metagenomics has enhanced our knowledge of the relationships between human vaginal microbiome, health, and diseases, and also has shown the presence of a number of unknown and uncultured microorganisms such as BVAB1, BVAB2, and BVAB3 (Fredricks et al., 2005). In the postgenomic era, new technology and omics methodologies are being intensively developed. Culturomics is one of these new approaches dynamically describing new bacteria. Based on a multiplication of culture conditions combined with a rapid identification of bacteria, it was recently introduced and applied to samples from various body sites, including the human vagina.

First application of culturomics was to study the gut microbiota. Thus, microbial culturomics has expanded the diversity of the human microbiome to 1057 species, including 197 potential new bacterial species (Lagier et al., 2016). Recently, it has also enabled the culture and description of new bacterial species found in the vagina (Diop et al., 2016; 2017a; 2017b).

In this article, we described the isolation as well as the phenotypic and genomics characteristics of a new bacterial species *P. lascolaii* isolated from a vaginal sample of a 33-year-old French woman afflicted with BV. We described the sample using a polyphasic taxono-genomic strategy (Ramasamy et al., 2014) in sequencing its genome. The phylogenetic and genomic results agreed that *P. lascolaii* is indeed distinct from its phenotypically closest species and constitutes a new species.

After sampling under strict protocols, the sample was rapidly transported to the laboratory and cultured as soon as possible in aseptic conditions. This strictly anaerobic and nonmotile bacterium was also isolated in another vaginal specimen of a patient with BV and in stool samples, thus confirming that it is not a contamination but a member of the human microbiome (unpublished data). As suggested by several authors (Fenollar and Raoult, 2016), this also leads us to believe that BV results from fecal transplantation. To prove the authenticity of our isolate, a pure culture was deposed in two different microorganism collections: the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) and the Collection de Souches de l'Unité des Rickettsies (CSUR).

This work demonstrates the ability of culturomics and its taxonogenomics approach to, respectively, explore the human microbiome and describe new bacterial species. It should also be noted that this work does not attempt to describe the medical importance of this new bacterium in BV. Instead, it expands the human vaginal flora and by sequencing the genome of new species reduces the number of sequences not assigned to a known microorganism under metagenomics. To better understand the role of these species in vaginal health and vaginal dysbiosis, further laboratory experimentation will be needed to study their pathogenesis and virulence.

Conclusions

Phenotypic and phylogenetic analyses and genomic results mean we can propose strain khD1^T as the representative of a new species named *P. lascolaii* sp. nov. The type strain khD1^T was isolated from the vaginal sample of a patient suffering from BV. Using culturomics, which uses high-throughput culture conditions with a rapid bacterial identification by MALDI-TOF, several potential new bacterial species were found in the human vagina, thus suggesting that the vagina flora is a complex and still unknown ecosystem and its diversity should be explored as fully as possible. In sum, microbial culturomics is an important new addition to the diagnostic medicine toolbox and warrants attention in future medical, global health, and integrative biology postgraduate teaching curricula.

Acknowledgments

This study was supported by Méditerranée Infection and the National Research Agency under the program "Investissements d'avenir," reference ANR-10-IAHU-03.

The authors thank the Xegen Company (www.xegen.fr) for automating the genomic annotation process. They also thank TradOnline for reviewing the English.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

Carver T, Harris SR, Berriman M, Parkhill J, and McQuillan JA. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28, 464–469.

Carver T, Thomson N, Bleasby A, Berriman M, and Parkhill J. (2009). DNAPlotter: Circular and linear interactive genome visualization. Bioinformatics 25, 119–120.

Citron DM, Ostovari MI, Karlsson A, and Goldstein EJ. (1991). Evaluation of the E test for susceptibility testing of anaerobic bacteria. J Clin Microbiol 29, 2197–2203.

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Sayers EW. (2016). GenBank. Nucleic Acids Res 44, 67–72.

Diop K, Bretelle F, Michelle C, et al. (2017a). Taxonogenomics and description of *Vaginella massiliensis* gen. nov., sp. nov., strain Marseille P2517^T, a new bacterial genus isolated from the human vagina. New Microbes New Infect 15, 94–103.

Diop K, Diop A, Bretelle F, et al. (2017b). *Olegusella massiliensis* gen. nov., sp. nov., strain KHD7^T, a new bacterial genus isolated from the female genital tract of a patient with bacterial vaginosis. Anaerobe 44, 87–95.

Diop K, Raoult D, Bretelle F, and Fenollar F. (2016). ''*Murdochiella vaginalis*'' sp. nov., a new bacterial species cultivated from the vaginal flora of a woman with bacterial vaginosis. Hum Microbiome J 2, 15–16.

Drancourt M, Bollet C, Carlioz A, Martelin R, Gayral J-P, and Raoult D. (2000). 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. J Clin Microbiol 38, 3623–3630.

Dubourg G, Lagier JC, Armougom F, et al. (2013). The gut microbiota of a patient with resistant tuberculosis is more comprehensively studied by culturomics than by metagenomics. Eur J Clin Microbiol Infect Dis 32, 637–645.

Fenollar F, and Raoult D. (2016). Does bacterial vaginosis result from fecal transplantation? J Infect Dis 214, 1784–1784.

Fournier PE, and Drancourt M. (2015). New Microbes New Infections promotes modern prokaryotic taxonomy: A new section ''TaxonoGenomics: New genomes of microorganisms in humans.'' New Microbes New Infect 7, 48–49.

Fournier PE, Lagier JC, Dubourg G, and Raoult D. (2015). From culturomics to taxonomogenomics: A need to change the taxonomy of prokaryotes in clinical microbiology. Anaerobe 36, 73–78.

Fredricks DN, Fiedler TL, and Marrazzo JM. (2005). Molecular identification of bacteria associated with bacterial vaginosis. N Engl J Med 353, 1899–1911.

Gouret P, Paganini J, Dainat J, et al. (2011). Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: The multi-agent software system DAGOBAH. In: *Evolutionary Biology—Concepts, Biodiversity, Macroevolution and Genome Evolution*. Pontarotti P, ed. Heidelberg, Germany: Springer Berlin, 71–87.

Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, and Danchin EG. (2005). FIGENIX: Intelligent automation of genomic annotation: Expertise integration in a new software platform. BMC Bioinformatics 6, 198.

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, and Hauser LJ. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.

Kim M, Oh HS, Park SC, and Chun J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64, 346–351.

Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, and Ussery DW. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35, 3100–3108.

Lagier JC, Armougom F, Million M, et al. (2012). Microbial culturomics: Paradigm shift in the human gut microbiome study. Clin Microbiol Infect 18, 1185–1193.

Lagier JC, Khelaifia S, Alou MT, et al. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol 1, 16203.

Lamont R, Sobel J, Akins R, et al. (2011). The vaginal microbiome: New information about genital tract flora using molecular based techniques: Vaginal microbiome using molecular tools. BJOG Int J Obstet Gynaecol 118, 533–549.

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, and Prohaska SJ. (2011). Proteinortho: Detection of (Co-) orthologs in large-scale analysis. BMC Bioinformatics 12, 1.

Lowe TM, and Eddy SR. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955–964.

Matuschek E, Brown DFJ, and Kahlmeter G. (2014). Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. Clin Microbiol Infect 20, O255–O266.

Meier-Kolthoff JP, Auch AF, Klenk HP, and Göker M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14, 1.

Menard JP, Fenollar F, Henry M, Bretelle F, and Raoult D. (2008). Molecular quantification of *Gardnerella vaginalis* and *Atopobium vaginae* loads to predict bacterial vaginosis. Clin Infect Dis 47, 33–43.

Murray PR, Baron EJ, Jorgensen JH, Landry ML, and Pfaller MA. (2007). *Manual of Clinical Microbiology*, 9th ed. Washington, DC: ASM Press.

Narayankhedkar A, Hodiwala A, and Mane A. (2015). Clinicoetiological characterization of infectious vaginitis amongst women of reproductive age group from Navi Mumbai, India. J Sex Transm Dis 2015, 1–5.

Ramasamy D, Mishra AK, Lagier JC, et al. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 64, 384–391.

Sakamoto M, and Ohkuma M. (2012). Reclassification of *Xylanibacter oryzae* Ueki et al. 2006 as *Prevotella oryzae* comb. nov., with an emended description of the genus *Prevotella*. Int J Syst Evol Microbiol 62, 2637–2642.

Sasser M. (2006). Bacterial identification by gas chromatographic analysis of fatty acids methyl esters (GC-FAME). MIDI, Technical Note #101.

Seck E, Rathored J, Khelaifia S, et al. (2015). *Virgibacillus senegalensis* sp. nov., a new moderately halophilic bacterium isolated from human gut. New Microbes New Infect 8, 116–126.

Seng P, Drancourt M, Gouriet F, et al. (2009). Ongoing revolution in bacteriology: Routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 49, 543–551.

Shah HN, and Collins DM. (1990). NOTES: *Prevotella*, a new genus to include *Bacteroides melaninogenicus* and related species formerly classified in the genus *Bacteroides*. Int J Syst Evol Microbiol 40, 205–208.

Stackebrandt E, and Ebers J. (2006). Taxonomic parameters revisited: Tarnished gold standards. Microbiol Today 33, 152.

Address correspondence to:
Pr. Florence Fenollar, MD, PhD
URMITE, UM 63, CNRS UMR 7278, IRD
198, INSERM U1095
Aix-Marseille University
27 Bd Jean Moulin
Marseille 13005
France

E-mail: florence.fenollar@univ-amu.fr

Abbreviations Used

AGIOS = average genomic identity of gene sequences
BV = bacterial vaginosis
BVAB = bacterial vaginosis-associated bacteria
COG = Clusters of Orthologous Groups
CSUR = Collection de souches de l'Unité des Rickettsies
DSM = Deutsche Sammlung von Mikroorganismen
FAMEs = fatty acid methyl esters
GC/MS = gas chromatography/mass spectrometry
MALDI-TOF = matrix-assisted laser desorption/ionization–time of flight
MICs = minimal inhibitory concentrations
MTBE = methyl tert-butyl ether
ORFs = open reading frames
TE buffer = Tris–EDTA buffer

Article 12:

Characterization of a novel Gram-positive Anaerobic Coccus isolated from the female genital tract: Genome sequence and Description of *Murdochiella vaginalis* sp. nov.

Diop Kh, Diop A, Khelaifia S, Robert C, di pinto F, Delerce J, Raoult D, Fournier PE, Bretelle F, Fenollar F

[Published in MicrobiologyOpen]

Received: 29 August 2017 | Revised: 3 November 2017 | Accepted: 7 November 2017

DOI: 10.1002/mbo3.570

ORIGINAL RESEARCH

WILEY MicrobiologyOpen Open Access

Characterization of a novel Gram-stain-positive anaerobic coccus isolated from the female genital tract: Genome sequence and description of *Murdochiella vaginalis* sp. nov.

Khoudia Diop[1] | Awa Diop[1] | Saber Khelaifia[1] | Catherine Robert[1] | Fabrizio Di Pinto[1] | Jérémy Delerce[1] | Didier Raoult[1,2] | Pierre-Edouard Fournier[1] | Florence Bretelle[1,3] | Florence Fenollar[1]

[1]Aix-Marseille Univ, Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, UM 63, CNRS UMR 7278, IRD 198, INSERM U1095, Institut Hospitalo-Universitaire Méditerranée-Infection, Faculté de Médecine, Marseille, France

[2]Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

[3]Department of Gynecology and Obstetrics, Gynépole, Marseille, Hôpital Nord, Assistance Publique-Hôpitaux de Marseille, Marseille, France

Correspondence
Florence Fenollar, Aix-Marseille Univ, Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, UM 63, CNRS UMR 7278, IRD 198, INSERM U1095, Institut Hospitalo-Universitaire Méditerranée-Infection, Faculté de Médecine, Marseille, France.
Email: florence.fenollar@univ-amu.fr

Funding information
This study was supported by the Fondation Méditerranée Infection and the French National Research Agency under the "Investissements d'avenir" program, reference ANR-10-IAHU-03

Abstract

Strain Marseille-P2341^T, a nonmotile, nonspore-forming, Gram-stain-positive anaerobic coccus, was isolated in the vaginal specimen of a patient with bacterial vaginosis using culturomics. Its growth occurred at temperatures ranging from 25 to 42°C, with pH between 6.5 and 8.5, and at NaCl concentrations lower than 5%. The major fatty acids were $C_{18:1n9}$ (27.7%) and $C_{16:0}$ (24.4%). Its genome is 1,671,491 bp long with 49.48 mol% of G+C content. It is composed of 1,501 genes: 1,446 were protein-coding genes and 55 were RNAs. Strain Marseille-P2341^T shared 97.3% of 16S rRNA gene sequence similarity with *Murdochiella asaccharolytica*, the phylogenetically closest species. These results enabled the classification of strain Marseille-P2341^T as a new species of the genus *Murdochiella* for which we proposed the name *Murdochiella vaginalis* sp. nov. The type strain is strain Marseille-P2341^T (=DSM 102237, =CSUR P2341).

KEYWORDS
bacterial vaginosis, culturomics, genome, *Murdochiella vaginalis*, taxono-genomics, vaginal microbiota

1 | INTRODUCTION

Due to vaginal secretions and, sometimes, urine, the vagina is a humid biotope which constitutes a complex ecosystem colonized by several types of microorganisms (Pal et al., 2011). Its composition was described for the first time in 1892 by Döderlein, who revealed that the vaginal flora is homogeneous and composed of Gram-positive bacteria known as Döderlein bacilli (Lepargneur & Rousseau, 2002). Since then, many studies have been conducted, some of which suggest that this complex ecosystem is mostly dominated by the *Lactobacillus* genus (De Vos et al., 2009) with four main species: *Lactobacillus crispatus*, *Lactobacillus gasseri*, *Lactobacillus jensenii*, and *Lactobacillus vaginalis*. This constitutes the first line of defense against genital infections (Bohbot & Lepargneur, 2012; Turovskiy, Sutyak

Noll, & Chikindas, 2011). An imbalance in this flora is observed in bacterial vaginosis.

The vaginal microflora diversity of a patient suffering from bacterial vaginosis was first described by Schröder in 1921 (Pal et al., 2011). This dysbiosis is characterized by a progressive decrease or even a lack of normal *Lactobacillus* flora accompanied by an increased pH of the vaginal lumen and an abnormal proliferation of previously underrepresented bacteria and Gram-stain-negative anaerobic bacteria (*Gardnerella vaginalis*, *Atopobium vaginae*, *Mobiluncus curtisii*, etc.) (Pépin et al., 2011; Shipitsyna et al., 2013). The mechanism of bacterial vaginosis is unknown; its empirical treatment and relapse rate is estimated at 50% at 3 months (Bretelle et al., 2015). This disturbance is associated with some complications in pregnant women such as miscarriage, chorioamnionitis, and preterm birth (Bretelle et al., 2015; Svare, Schmidt, Hansen, & Lose, 2006).

Initially studied using conventional culture methods, the understanding of the human vaginal microbiota was enhanced through the use of molecular techniques involving sequencing and phylogenetic analysis of the 16S rRNA gene (Lamont et al., 2011). These molecular methods enabled the detection of fastidious and uncultured bacteria such as bacterial vaginosis-associated bacteria (BVAB): BVAB1 BVAB2, and BVAB3 (Fredricks, Fiedler, & Marrazzo, 2005). In order to identify all bacteria (uncultured and fastidious) present in the vagina and involved in this alteration, we studied normal vaginal flora and those from bacterial vaginosis using the concept of "microbial culturomics," based on the multiplication of culture conditions with variations in temperature, media, pH, and atmospheric conditions, and rapid bacterial identification using matrix-assisted laser-desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) (Lagier et al., 2012, 2015). This microbial culturomics approach enabled us to isolate a new member of the *Murdochiella* genus that did not correspond to other species of this genus. This strain is designated as Marseille-P2341^T. The *Murdochiella* genus was created in 2010, to include strain recovered from a human abdominal wall abscess and in a sacral pilonidal cyst aspirate (Ulger-Toprak, Liu, Summanen, & Finegold, 2010). This genus has only one valid species: *Murdochiella asaccharolytica*.

The description of new bacterial species is based on phenotypic and genotypic characteristics but has some limitations (Chan, Halachev, Loman, Constantinidou, & Pallen, 2012; Vandamme et al., 1996). In this manuscript we use taxonogenomics, a new approach combining classic characteristics with the proteomic information obtained from MALDI-TOF MS and the description of the annotated whole genome (Fournier & Drancourt, 2015; Fournier, Lagier, Dubourg, & Raoult, 2015), to describe *Murdochiella vaginalis* sp. nov. (=DSM 102237 = CSUR P2341).

2 | MATERIALS AND METHODS

2.1 | Sample ethics and strain isolation

Using a Sigma Transwab (Medical Wire, Corsham, United Kingdom), the vaginal specimen of a 33-year-old French woman was collected

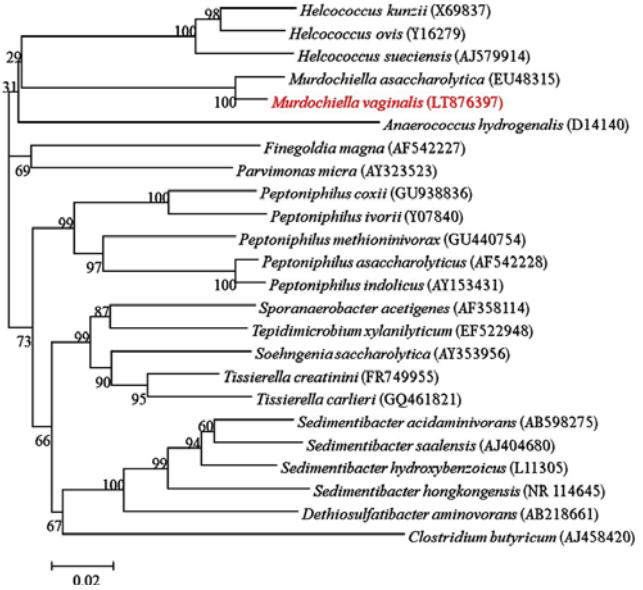


FIGURE 1 Phylogenetic tree highlighting the position of *Murdochiella vaginalis* strain Marseille-P2341^T relative to other close strains. GenBank accession numbers of each 16S rRNA are noted in parenthesis. Sequences were aligned using Muscle v3.8.31 with default parameters and phylogenetic inferences were obtained using the neighbor-joining method with 500 bootstrap replicates, within MEGA6 software. The scale bar represents a 2% nucleotide sequence divergence

TABLE 1 Classification and general features of *Murdochiella vaginalis* Marseille-P2341^T

Properties	Terms
Taxonomy	**Kingdom**: *Bacteria*
	Phylum: *Firmicutes*
	Class: *Clostridia*
	Order: *Clostridiales*
	Family: *Peptoniphiliaceae*
	Genus: *Murdochiella*
	Species: *M. vaginalis*
Type strain	Marseille-P2341^T
Isolation site	Human vagina
Isolation country	France
Gram stain	Positive
Cell shape	Coccus
Motility	No
Oxygen requirements	Anaerobic
Optimal temperature	37°C
Temperature range	Mesophilic

and transported to the La Timone hospital in Marseille (France). Diagnosed as previously reported (Menard, Fenollar, Henry, Bretelle, & Raoult, 2008), the patient was suffering from bacterial vaginosis. At the time the sample was collected, she was not being treated with any antibiotics. The study was authorized by the local IFR48 ethics committee (Marseille, France) under agreement number 09-022 and the patient also signed written consent. After sampling, the specimen was preincubated in a blood culture bottle (BD Diagnostics, Le Pont-de-Claix, France) enriched with 4 ml of rumen that was filter-sterilized through a 0.2 µm pore filter (Thermo Fisher Scientific, Villebon-sur-Yvette, France) and 3 ml of sheep's blood (bioMérieux, Marcy l'Etoile,

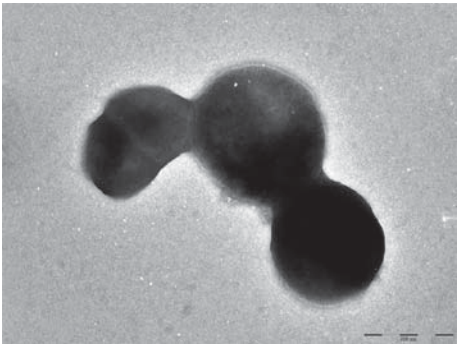


FIGURE 3 Transmission electron microscopy of *Murdochiella vaginalis* strain Marseille-P2341^T, using a Tecnai G20 transmission electron microscope (FEI Company). The scale bar represents 100 nm

France). After different preincubation periods (1, 3, 7, 10,15, 20, and 30 days), 50 µl of the supernatant was inoculated on Schaedler agar (BD Diagnostics) and then incubated for 7 days under anaerobic conditions at 37°C.

2.2 | Strain identification by MALDI-TOF MS and 16S rRNA gene sequencing

Isolated colonies were deposited in duplicate on a MTP 96 MALDI-TOF target plate (Bruker Daltonics, Leipzig, Germany) for identification with a microflex spectrometer (Bruker), as previously described (Seng et al., 2009). All obtained protein spectra were loaded into the MALDI Biotyper Software (Bruker Daltonics) and compared, as previously described (18), using the standard pattern-matching algorithm, which compared the acquired spectrum with those present

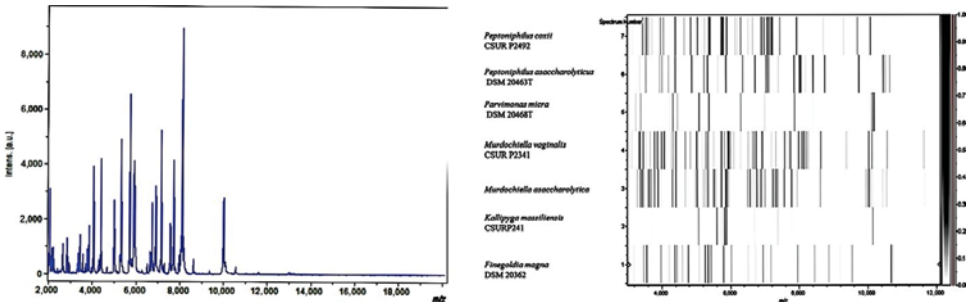


FIGURE 2 MALDI-TOF information. (**a**) Reference mass spectrum from *Murdochiella vaginalis* strain Marseille-P2341T spectra. (**b**) Gel view comparing *M. vaginalis* strain Marseille-P2341T to other species within Peptoniphilaceae family. The gel view displays the raw spectra of loaded spectrum files arranged with a pseudo-gel like appearance. The *x*-axis records the *m*/*z* value. The left *y*-axis displays the running spectrum number originating from subsequent spectra loading. The peak intensity is expressed by a gray scale scheme code. The right *y*-axis indicates the relation between the color of a peak and its intensity, in arbitrary units. Displayed species are indicated on the left

TABLE 2 Cellular fatty acid composition (%)

Fatty acids	Name	Mean relative % (a)
18:1n9	9-Octadecenoic acid	27.7 ± 6.6
16:0	Hexadecanoic acid	24.2 ± 4.1
18:2n6	9,12-Octadecadienoic acid	15.7 ± 4.4
18:0	Octadecanoic acid	13.4 ± 2.2
14:0	Tetradecanoic acid	5.9 ± 7.0
18:1n7	11-Octadecenoic acid	3.7 ± 0.6
15:0 iso	13-methyl-tetradecanoic acid	1.4 ± 1.7
17:0	Heptadecanoic acid	1.0 ± 0.1
14:0 3-OH	3-hydroxy-Tetradecanoic acid	TR
20:0	Eicosanoic acid	TR
18:0 9,10-methylene	2-octyl-Cyclopropaneoctanoic acid	TR
5:0 iso	3-methyl-butanoic acid	TR
20:4n6	5,8,11,14-Eicosatetraenoic acid	TR
15:0	Pentadecanoic acid	TR
16:1n5	11-Hexadecenoic acid	TR
17:0 anteiso	14-methyl-Hexadecanoic acid	TR
17:0 iso	15-methyl-Hexadecanoic acid	TR
20:1n9	11-Eicosenoic acid	TR
15:0 anteiso	12-methyl-tetradecanoic acid	TR
17:1n7	10-Heptadecenoic acid	TR
10:0	Decanoic acid	TR
20:2n6	11,14-Eicosadienoic acid	TR
12:0	Dodecanoic acid	TR
19:0	Nonadecanoic acid	TR
22:5n2	7,10,13,16,19-docosapentaenoic acid	TR
16:0 9,10-methylene	2-Hexyl-Cyclopropaneoctanoic acid	TR
13:0	Tridecanoic acid	TR
4:0	Butanoic acid	TR
22:6n3	4,7,10,13,16,19-Docosahexaenoic acid	TR

[a]Mean peak area percentage; TR = trace amounts <1%.

in the library (the Bruker database and our constantly updated database). If the score was greater than 1.9, the bacterium was considered to be identified at the species level. If not, identification failed and to achieve identification for unidentified colonies, the 16S rRNA gene was sequenced using fD1-rP2 primers (Eurogentec, Angers, France) and the obtained sequence was matched against the NCBI database using the BLAST algorithm (Drancourt et al., 2000). As suggested, if the 16S rRNA gene sequence similarity value was lower than 95% or 98.7%, the strain was defined as a new genus or species, respectively (Kim, Oh, Park, & Chun, 2014; Stackebrandt & Ebers, 2006).

2.3 | Phylogenetic analysis

All species from the same order of the new species were retrieved and 16S sequences were download from NCBI, by parsing NCBI eUtils results and the NCBI taxonomy page. Sequences were aligned using CLUSTALW, with default parameters and phylogenetic inferences obtained using the neighbor-joining method with 500 bootstrap replicates, within MEGA6 software.

2.4 | Growth conditions and morphological observation

To evaluate ideal growth, the strain Marseille-P2341^T was cultivated on Columbia agar with 5% sheep's blood (bioMérieux) and incubated at different temperatures (25, 28, 37, 45, and 56°C) in an aerobic atmosphere with or without 5% CO_2, and in anaerobic and microaerophilic atmospheres, using GENbag Anaer and GENbag microaer systems (bioMérieux). The salinity and pH conditions were also tested at different concentrations of NaCl (0%, 5%, 15%, and 45%) and different pH (5, 6, 6.5, 7, and 8.5).

Oxidase and catalase tests, Gram-stain, motility, and sporulation were performed using standard procedures (Murray, Baron, Jorgensen, Landry, & Pfaller, 2007). To observe cell morphology, they were fixed with 2.5% glutaraldehyde in 0.1 mol/L cacodylate buffer for at least 1 hr at 4°C. A drop of cell suspension was then deposited for approximately 5 min on glow-discharged formvar carbon film on 400 mesh nickel grids (FCF400-Ni, EMS). The grids were dried on blotting paper and cells were negatively stained for 10 s with 1% ammonium molybdate solution in filtered water at RT. Electron micrographs were acquired using a Tecnai G20 Cryo (FEI) transmission electron microscope operated at 200 keV.

2.5 | Biochemical and antibiotic susceptibility tests

Biochemical tests were performed using API ZYM, API 20A, and API 50CH strips (bioMérieux) according to the manufacturer's instructions. The strips were incubated for 4, 24, and 48 hr respectively.

Cellular fatty acid methyl ester (FAME) analysis was performed using Gas Chromatography/Mass Spectrometry (GC/MS). Strain Marseille-P2341^T was grown on Columbia agar enriched with 5% sheep's blood (bioMérieux). Two samples were then prepared with approximately 50 mg of bacterial biomass per tube harvested from several culture plates. Fatty acid methyl esters were prepared as described by Sasser (Sasser, 2006). GC/MS analyses were carried out as previously described (Dione et al., 2016). In brief, fatty acid methyl esters were separated using an Elite 5-MS column and monitored by mass spectrometry (Clarus 500—SQ 8 S, Perkin Elmer, Courtaboeuf, France). A spectral database search was performed using MS Search 2.0 operated with the Standard Reference Database 1A (NIST, Gaithersburg, USA) and the FAMEs mass spectral database (Wiley, Chichester, UK).

Antibiotic susceptibility was tested using the disc diffusion method (Le Page et al., 2015). The results were read using Scan 1200 (Interscience, Saint-Nom-la-Bretèche, France).

TABLE 3 Differential characteristics of *Murdochiella vaginalis* and the phylogenetically related species. *Murdochiella vaginalis* strain Marseille-P2341^T, *Murdochiella asaccharolytica* strain WAL 1855C^T, *Finegoldia magna* strain CCUG 17636^T, *Peptoniphilus indolicus* ATCC 29427^T, *Parvimonas micra* CCUG 46357^T, *Helcococcus sueciensis* CCUG 47334^T, and *Anaerococcus hydrogenalis* JCM 7635^T

Properties	*M. vaginalis*	*M. asaccharolytica*	*F. magna*	*P. indolicus*	*P. micra*	*H. sueciensis*	*A. hydrogenalis*
Cell diameter (μm)	0.6–0.8	0.5–0.6	0.8–1.6	0.7–1.6	0.3–0.7	na	0.7–1.8
Oxygen requirement	Anaerobic	Anaerobic	Anaerobic	Anaerobic	Anaerobic	Facultative anaerobic	Anaerobic
DNA G+C content (mol%)	49.5	na	na	31.69	28.65	29.5	29.64
Production of							
Alkaline phosphatase	−	−	Variable	+	+	+	−
Indole	−	−	−	+	−	−	+
Catalase	−	−	Variable	na	Variable	−	−
Nitrate reductase	−	−	−	+	−	−	−
Urease	−	−	−	−	−	−	Variable
β-galactosidase	+	−	−	−	−	−	-
N-acetyl-glucosamine	+	−	−	na	−	+	na
Acid from							
Mannose	+	−	−	−	−	−	+
Glucose	+	−	−	−	−	+	+
Lactose	−	−	−	−	−	+	+
Raffinose	−	−	−	−	−	−	+
Habitat	Vaginal discharges	Human wound	Human specimen	Summer mastitis of cattle	Human specimen	Human wound	Vaginal discharges

+, positive reaction; −, negative reaction; na, data not available.

2.6 | Genomic DNA preparation

Genomic DNA (gDNA) of strain Marseille-P2341^T was extracted in two steps: a mechanical treatment was first performed using acid-washed glass beads (G4649-500 g Sigma) and a FastPrep BIO 101 instrument (Qbiogene, Strasbourg, France) at maximum speed (6.5) for 3 × 30 s. Then after 2 hr of lysozyme incubation at 37°C, DNA was extracted on the EZ1 biorobot (Qiagen, Hilden, Germany) using the EZ1 DNA tissue kit. The elution volume was 50 µl. The gDNA was quantified by a Qubit assay using the high sensitivity kit (Life technologies, Carlsbad, CA, USA) to 103 ng/µl.

TABLE 4 Nucleotide content and gene count levels of the genome

Attribute	Value	% of total[a]
Size (bp)	1,671,491	100
G+C content (bp)	827,028	49.48
Coding region (bp)	1,511,436	90.42
Total genes	1,501	100
RNA genes	55	3.66
Protein-coding genes	1,446	100
Genes with function prediction	1,056	73.03
Genes assigned to COGs	965	66.74
Genes with peptide signals	160	11.06
Genes with transmembrane helices	369	25.52

[a]The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

2.7 | Genome sequencing and assembly

gDNA was sequenced on the MiSeq Technology (Illumina Inc, San Diego, CA, USA) using the mate pair strategy. The gDNA was barcoded using the Nextera Mate Pair sample prep kit (Illumina) to be mixed with 11 other projects. The mate pair library was prepared with 1.5 µg of genomic DNA using the Nextera mate pair Illumina guide. The genomic DNA sample was simultaneously fragmented and tagged with a mate pair junction adapter. The pattern of fragmentation was validated on an Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA 7500 labchip. The DNA fragments ranged in size from 1.5 kb to 11 kb with an optimal size at 3.716 kb. No size selection was performed and 652 ng of tagmented fragments were circularized. The circularized DNA was mechanically sheared to small fragments with a bi-modal pattern at 644 bp and 1,613 bp on the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The library profile was visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) and the final concentration library was measured at 53.40 nmol/L.

The libraries were normalized at 2 nmol/L and pooled. After a denaturation step and dilution at 15 pmol/L, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated cluster generation and sequencing run were performed in a single 39-hr run in a 2 × 251-bp.

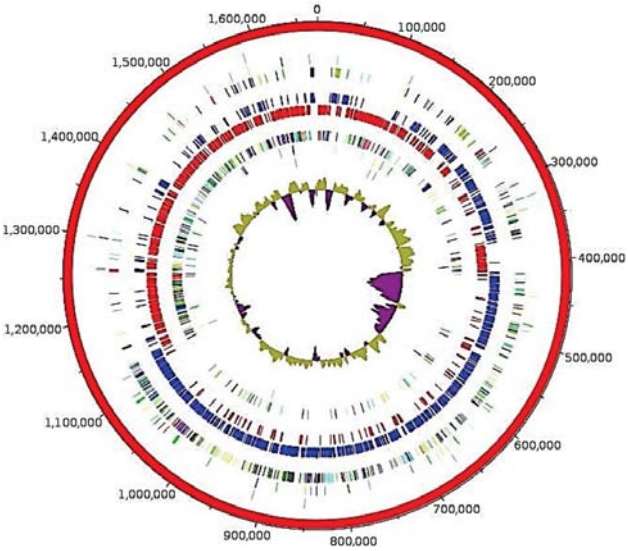


FIGURE 4 Graphical circular map of the genome. From outside to the center: Contigs (red/gray), COG category of genes on the forward strand (three circles), genes on forward strand (blue circle), genes on the reverse strand (red circle), COG category on the reverse strand (three circles), GC content

TABLE 5 Number of genes associated with the 25 general COG functional categories

Code	Value	% of total	Description
[J]	157	10.857538	Translation
[A]	0	0	RNA processing and modification
[K]	71	4.910097	Transcription
[L]	57	3.9419088	Replication, recombination and repair
[B]	0	0	Chromatin structure and dynamics
[D]	16	1.1065007	Cell cycle control, mitosis and meiosis
[Y]	0	0	Nuclear structure
[V]	45	3.1120331	Defense mechanisms
[T]	32	2.2130015	Signal transduction mechanisms
[M]	44	3.042877	Cell wall/membrane biogenesis
[N]	4	0.2766252	Cell motility
[Z]	0	0	Cytoskeleton
[W]	1	0.0691563	Extracellular structures
[U]	15	1.0373445	Intracellular trafficking and secretion
[O]	53	3.6652837	Post-translational modification, protein turnover, chaperones
[X]	8	0.5532504	Mobilome: prophages, transposons
[C]	60	4.149378	Energy production and conversion
[G]	81	5.60166	Carbohydrate transport and metabolism
[E]	80	5.5325036	Amino acid transport and metabolism
[F]	51	3.526971	Nucleotide transport and metabolism
[H]	52	3.5961275	Coenzyme transport and metabolism
[I]	34	2.351314	Lipid transport and metabolism
[P]	46	3.1811898	Inorganic ion transport and metabolism
[Q]	9	0.62240666	Secondary metabolites biosynthesis, transport and catabolism
[R]	92	6.3623796	General function prediction only
[S]	42	2.9045644	Function unknown
_	481	33.26418	Not in COGs

In total, 9.2 Gb of information was obtained from a 1042 K/mm^2 cluster density with a cluster passing quality control filters of 91.6% (18,078,000 passing filter paired reads). Within this run, the index representation for strain Marseille-P2341^T was determined to 13.14%. The 2,375,075 paired reads were trimmed then assembled in a scaffold.

2.8 | Genome annotation and analysis

Prodigal was used for open reading frame (ORF) prediction (Hyatt et al., 2010) with default parameters. We excluded predicted ORFs spanning a sequencing gap region (containing *N*). The bacterial proteome was predicted using BLASTP (E-value 1e^{-03} coverage 0.7 and identity percent 30) against the Clusters of Orthologous Groups (COG) database. If no hit was found, we searched against the NR database (Clark, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2016), using BLASTP with E-value of 1e^{-03} coverage 0.7 and an identity percent of 30. An E-value of 1e^{-05} was used if sequence lengths were shorter than 80 amino acids. PFam conserved domains (PFAM-A an PFAM-B domains) were searched on each protein with the hhmscan tools analysis. RNAmmer (Lagesen et al., 2007) and tRNAScanSE tools (Lowe & Eddy, 1997) were used to find ribosomal RNAs genes and tRNA genes, respectively. ORFans were identified if all the BLASTP performed had negative results (E-value smaller than 1e^{-03} for ORFs with sequence size above 80 aa or E-value smaller than 1e^{-05} for ORFs with sequence length below 80 aa). For data management and visualization of genomic features, Artemis (Carver, Harris, Berriman, Parkhill, & McQuillan, 2012) and DNA Plotter (Carver, Thomson, Bleasby, Berriman, & Parkhill, 2009) were used, respectively. We used the home-made MAGI software to analyze the mean level of nucleotide sequence similarity at the genome level. It calculated the average genomic identity of gene sequences (AGIOS) among compared genomes (Ramasamy et al., 2014). This software combines the Proteinrtho software (Lechner et al., 2011) for detecting orthologous proteins in pairwise genomic comparisons. The corresponding genes were then retrieved and the mean percentage of nucleotide sequence identity among orthologous ORFs was determined using the Needleman-Wunsch global alignment algorithm. The MultiAgent software system DAGOBAH (Gouret et al., 2011) was used to perform the annotation and comparison processes, which included Figenix (Gouret et al., 2005) libraries for pipeline analysis. We also performed GGDC analysis using the GGDC web server, as previously reported (Meier-Kolthoff, Auch, Klenk, & Göker, 2013).

3 | RESULTS

3.1 | Strain identification

Strain Marseille-P2341^T was first isolated after 15 days of preincubation of a vaginal sample in a blood culture bottle supplemented with rumen and sheep's blood under anaerobic conditions and then sub-cultured on Schaedler agar. A score of 1.3 was also obtained with MALDI-TOF MS identification, suggesting that this isolate was not in the database. The 16S rRNA gene sequence (accession number LT576397) of the strain exhibited 97.3% nucleotide sequence similarity with *M. asaccharolytica*, the phylogenetically-closest species with a validly published name (Figure 1). As this value was lower than 98.7%, the threshold recommended for delineating a new species (Kim et al., 2014; Stackebrandt & Ebers, 2006), strain Marseille-P2341^T was classified as a new species named *M. vaginalis* (Table 1). The reference

TABLE 6 Genome comparison of closely related species to *Murdochiella vaginalis* strain Marseille-P2341^T

Species	INSDC identifier	Size (Mb)	G+C (mol%)	Gene Content
M. vaginalis strain Marseille-P2341^T	LT632322	1.671	49.48	1,501
Anaerococcus hydrogenalis DSM 7454	ABXA00000000.1	1.89	29.64	2,069
Helcococcus kunzii NCFB 2900	AGEI00000000.1	2.10	29.35	1,882
Peptoniphilus indolicus ATCC 29427	AGBB00000000.1	2.24	31.69	2,269
Helcococcus sueciensis CCUG 47334	AUHK00000000.1	1.57	28.40	1,445
Peptoniphilus coxii RMA 16757	LSDG00000000.1	1.84	44.62	1,86
Parvimonas micra ATCC 33270	ABEE00000000.2	1.70	28.65	1,678

INSDC, International Nucleotide Sequence Database Collaboration.

spectrum of the strain Marseille-P2341^T (Figure 2a) was then added to our database and compared to other known species of the family *Peptoniphilaceae* (Johnson, Whitehead, Cotta, Rhoades, & Lawson, 2014). Their differences are shown in the gel view which was obtained (Figure 2b).

3.2 | Phenotypic characteristics

Only grown in anaerobic conditions, strain Marseille-P2341^T grows at temperatures between 25 to 42°C, with optimal growth at 37°C after 48 hr of incubation. It needs NaCl concentrations lower than 5 g/L and a pH ranging from 6.5 to 8.5. After 2 days of incubation at 37°C under anaerobic conditions on Columbia agar (bioMérieux), colonies are circular, white, and opaque with a diameter of 2–2.5 mm. Gram-staining shows a Gram-positive coccus. Individual cells show a diameter ranging from 0.6 to 0.8 μm under an electron microscope (Figure 3). Nonmotile and nonspore-forming, strain Marseille-P2341^T exhibited positive oxidase activity. Nevertheless, catalase activity was negative and nitrate was not reduced.

Using an API ZYM strip, positive reactions were observed for leucine arylamidase, Naphtol-AS-BI-phosphohydrolase, α and β-galactosidase, glucosidase (α and β), N-acetyl-β-glucosaminidase, α-mannosidase, and α-fucosidase. Alkaline phosphatase, lipases, and other reactions were negative. On an API 20A strip, we observed an acidification of glucose and an API 50CH strip revealed that only galactose, glucose, mannose, and potassium 5-ketogluconate were metabolized. All the other reactions were negative on both API strips. The most abundant fatty acids found were 9-Octadecenoic acid and Hexadecanoic acid (28% and 24%, respectively). Interesting minor fatty acids (<1%) are also described (Table 2). Cells were susceptible to oxacillin, penicillin, ceftriaxone, ciprofloxacin, clindamycin, doxycycline, erythromycin, fosfomycin, gentamycin, trimethoprim-sulfamethoxazole, rifampicin, and vancomycin but resistant to colistin. The phenotypic characteristics of strain Marseille-P2341^T were compared to those of closely related species and are summarized in Table 3 (Collins, 2004; Ezaki et al., 2001; Ezaki, Yamamoto, Ninomiya, Suzuki, & Yabuuchi, 1983; Murdoch & Shah, 1999; Tindall & Euzeby, 2006; Ulger-Toprak et al., 2010).

3.3 | Genome properties

The genome measures 1,671,491 bp long and has 49.48 mol% of G+C content (Table 4, Figure 4). It is composed of one scaffold composed of one contig. Of the 1,501 predicted genes, 1,446 were protein-coding genes and 55 were RNAs (two genes were 5S rRNA, two genes were 16S rRNA, two genes were 23S rRNA, 49 genes were tRNA genes). A total of 1,056 genes (73.03%) were assigned a putative function (by cogs or by NR blast). 56 genes were identified as ORFans (3.87%). The remaining 292 genes were annotated as hypothetical proteins (20.19%). Genome statistics are summarized in Table 4 and the distribution of the genes in COGs functional categories is presented in Table 5.

3.4 | Genomic comparison

The comparison of the genome of our species with the closest related species (Table 6) reveals that the genome sequence of strain Marseille-P2341^T (1.67 Mbp) is larger than that of *Helcococcus sueciensis* (1.57 Mbp), but smaller than those of *Parvimonas micra*, *Peptoniphilus coxii*, *Anaerococcus hydrogenalis*, *Helcococcus kunzii*, and *Peptoniphilus indolicus* (1.70, 1.84, 1.89, 2.10, and 2.24, respectively). The G+C content of strain Marseille-P2341 ^{T} (49.48 mol%) is greater than those of all compared species. The gene content of strain Marseille-P2341^T (1,446) is almost equal to that of *H. sueciensis* but is smaller than those of other compared genomes. However, in all the compared genomes, the distribution of genes in COG categories was similar. Nevertheless, there are fewer genes of *M. vaginalis* present in the COG categories X (Mobilome: prophages, transposons) and W (Extracellular structures) than other compared species (Figure 5). Moreover, the AGIOS analysis shows that strain Marseille-P2341^T shares between 509 and 542 orthologous genes with closely related species (Table 7) and analysis of the average percentage of nucleotide sequence identity ranged from 50.8% to 56.4% with *P. micra* and *H. sueciensis*, respectively (Table 7). In addition, the digital DNA-DNA hybridization (dDDH) of strain Marseille-P2341^T and its closest species varied between 22.40% to 36% with 22.40, 23.60, 23.70, 25.50, 25.90, and 36% for *H. kunzii*, *A. hydrogenalis*, *P. micra*, *P. coxii*, *H. sueciensis*, and *P. indolicus*, respectively. Unfortunately, *M. asaccharolytica* was not included in this comparison because its genome was not sequenced.

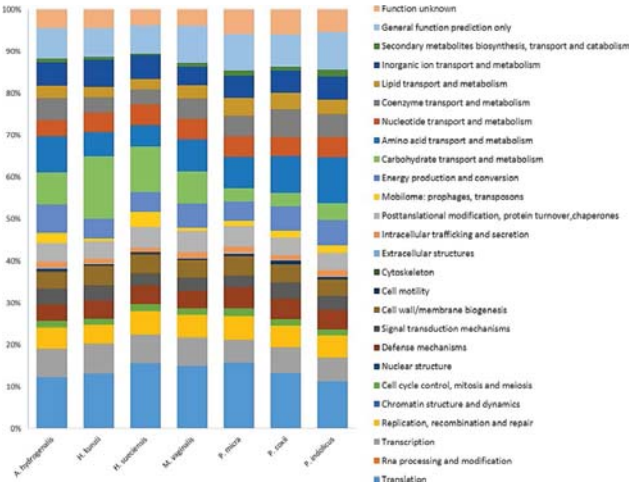


FIGURE 5 Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of *Murdochiella vaginalis* strain Marseille-P2341T among other species

4 | DISCUSSION

During the study of vaginal microbiota using culturomics, with the aim of exploring the vaginal flora as exhaustively as possible and identifying the bacteria involved in bacterial vaginosis in order to better manage this infection, strain Marseille-P2341^T was identified in the vaginal sample of a patient suffering from bacterial vaginosis. Its phenotypic characteristics, MALDI-TOF MS, 16S rRNA gene sequencing, and genome comparison with close phylogenic relatives enabled us to classify strain Marseille-P2341^T as a new species of the genus *Murdochiella*. The 16S rRNA gene sequence similarity was 97.3% with *M. asaccharolytica*, which was lower than the 98.7% threshold recommended for defining a new species (Kim et al., 2014; Stackebrandt & Ebers, 2006). Created in 2010, the genus *Murdochiella* contains Gram-positive staining anaerobic cocci bacteria which have been detected in human clinical samples (Ulger-Toprak et al., 2010). Members of this genus are nonmotile and nonsporulating, as observed for strain Marseille-P2341^T.

A polyphasic taxono-genomic strategy based on the combination of phenotypic and genomic analyses (Fournier & Drancourt, 2015; Fournier et al., 2015) was used to describe the new species whose strain Marseille-P2341^T is the type strain. Strain Marseille-P2341^T exhibited a specific MALDI-TOF MS spectrum and differed from the other studied closed bacterial species in their fermentation of carbohydrate. Bacteria in the *Murdochiella* genus are asaccharolytic and do not ferment carbohydrates. However, the *M. vaginalis* strain Marseille-P2341^T produces acid from glucose and mannose. This observation was confirmed by the annotation of the genome with the COGs database (Figure 5), showed that 7.7% of Marseille-P2341 predicted genes' were dedicated to carbohydrate transport and metabolism functions. These genes include carbohydrate enzymes such as glucose-6-phosphate isomerase, 6-phosphogluconolactonase, 6-phosphofructokinase, fructose-bisphosphate aldolase, triose-phosphate isomerase, glyceraldehyde-3-phosphate dehydrogenase, 3-phosphoglycerate kinase, phosphoglycerate mutase, enolase, pyruvate kinase, phosphomannomutase involved in carbohydrate metabolism, mainly in the process of glucose, fructose, and mannose metabolism.

TABLE 7 Numbers of orthologous proteins shared between genomes (upper right) and AGIOS values obtained (lower left)

	Murdochiella vaginalis	*Anaerococcus hydrogenalis*	*Helcococcus kunzii*	*Parvimonas micra*	*Helcococcus sueciensis*	*Peptoniphilus indolicus*	*Peptoniphilus coxii*
M. vaginalis	**1,446**	538	514	511	509	525	542
A. hydrogenalis	51.39	**2,069**	538	516	526	565	580
H. kunzii	51.12	57.33	**1,882**	541	653	511	534
P. micra	50.80	57.96	59.47	**1,678**	530	533	534
H. sueciensis	56.37	59.46	63.43	58.83	**1,445**	491	514
P. indolicus	52.45	58.27	56.33	58.43	59.21	**2,269**	614
P. coxii	52.67	53.15	52.95	53.78	50.25	52.93	**1,860**

The numbers of proteins per genome are indicated in bold.

The G+C content of strain Marseille-P2341^T and its phylogenetically-closest species ranges from 28.40 to 49.48 mol% and, as previously demonstrated, the difference in the G+C content is, at most, 1% in a species. Thus, overall, these values justify the strain Marseille-P2341^T being classified as a distinct species. The AGIOS and GGDC values also confirm it belongs to a new species (Klenk, Meier-Kolthoff, & Göker, 2014).

5 | TAXONOMIC AND NOMENCLATURE PROPOSAL

5.1 | Description of *Murdochiella vaginalis* sp. nov

Murdochiella vaginalis (va.gi.na'lis. L. n. *vagina*, sheath, vagina; L. fem. suff. *–alis*, suffix denoting pertaining to; N.L. fem. adj. *vaginalis*, pertaining to the vagina, of the vagina).

Obligate anaerobic *M. vaginalis* cells are Gram-stain-positive and coccus-shaped. They are nearly 0.7 µm in diameter, nonmotile, nonspore-forming, mesophilic, and occur in pairs or short chains. After 2 days of incubation on Columbia agar with 5% sheep's blood (bioMérieux) at 37°C under anaerobic conditions, colonies appear circular, white, and opaque with a diameter of 2–2.5 mm. Nitrate is not reduced; catalase and urease are also negative. Weakly saccharolytic, acid is produced only from glucose, mannose, and galactose. Positive reactions are observed for leucine arylamidase, Naphtol-AS-BI-phosphohydrolase, α- galactosidase, β-galactosidase, α-glucosidase, β-glucosidase, N-acetyl-β-glucosaminidase, α-mannosidase, and α-fucosidase. The most abundant fatty acids are $C_{18:1n9}$ (27.7%) and $C_{16:0}$ (24.4%). The type strain is susceptible to oxacillin, penicillin, ceftriaxone, ciprofloxacin, clindamycin, doxycycline, erythromycin, fosfomycin, gentamycin, trimethoprim-sulfamethoxazole, vancomycin, and rifampicin but resistant to colistin.

Its genome contains 49.48 mol% of G+C content and measures 1,671,491 bp long. The 16S rRNA and whole-genome sequences are both deposited in EMBL-EBI under accession numbers LT576397 and LT632322 respectively. The type strain Marseille-P2341^T (=DSM 102237, =CSUR P2341) was isolated from the vaginal sample of a French woman suffering from bacterial vaginosis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors thank the Xegen Company (www.xegen.fr) for automating the genomic annotation process. We also thank TradOnline for English reviewing and Claudia Andrieu for administrative assistance.

ORCID

Khoudia Diop http://orcid.org/0000-0002-9296-563X
Saber Khelaifia http://orcid.org/0000-0002-9303-3893
Pierre-Edouard Fournier http://orcid.org/0000-0001-8463-8885

REFERENCES

Bohbot, J. M., & Lepargneur, J. P. (2012). La vaginose en 2011: Encore beaucoup d'interrogations. *Gynécologie Obstétrique & Fertilité*, *40*, 31–36. https://doi.org/10.1016/j.gyobfe.2011.10.013

Bretelle, F., Rozenberg, P., Pascal, A., Favre, R., Bohec, C., Loundou, A., … Groupe de Recherche en Obstetrique Gynecologie (2015). High *Atopobium vaginae* and *Gardnerella vaginalis* vaginal loads are associated with preterm birth. *Clinical Infectious Diseases*, *60*, 860–867. https://doi.org/10.1093/cid/ciu966

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, *28*, 464–469. https://doi.org/10.1093/bioinformatics/btr703

Carver, T., Thomson, N., Bleasby, A., Berriman, M., & Parkhill, J. (2009). DNAPlotter: Circular and linear interactive genome visualization. *Bioinformatics*, *25*, 119–120. https://doi.org/10.1093/bioinformatics/btn578

Chan, J. Z., Halachev, M. R., Loman, N. J., Constantinidou, C., & Pallen, M. J. (2012). Defining bacterial species in the genomic era: Insights from the genus *Acinetobacter*. *BMC Microbiology*, *12*, 302. https://doi.org/10.1186/1471-2180-12-302

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*, D67–D72. https://doi.org/10.1093/nar/gkv1276

Collins, M. D. (2004). *Helcococcus sueciensis* sp. nov., isolated from a human wound. *International Journal of Systematic and Evolutionary Microbiology*, *54*, 1557–1560. https://doi.org/10.1099/ijs.0.63077-0

De Vos, P., Garrity, G. M., Jones, D., Krieg, N. R., Ludwig, W., Rainey, F. A., … Whitman, W. B. (2009). The Firmicutes. In G. M. Garrity (Ed.), *Bergey's Manual of Systematic Bacteriology* (pp. 465–511). New York, NY: Springer.

Dione, N., Sankar, S. A., Lagier, J. C., Khelaifia, S., Michele, C., Armstrong, N., … Fournier, P. E. (2016). Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. *New Microbes and New Infections*, *10*, 66–76. https://doi.org/10.1016/j.nmni.2016.01.002

Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J.-P., & Raoult, D. (2000). 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *Journal of Clinical Microbiology*, *38*, 3623–3630.

Ezaki, T., Kawamura, Y., Li, N., Li, Z. Y., Zhao, L., & Shu, S. (2001). Proposal of the genera *Anaerococcus* gen. nov., *Peptoniphilus* gen. nov. and *Gallicola* gen. nov. for members of the genus Peptostreptococcus. *International Journal of Systematic and Evolutionary Microbiology*, *51*, 1521–1528. https://doi.org/10.1099/00207713-51-4-1521

Ezaki, T., Yamamoto, N., Ninomiya, K., Suzuki, S., & Yabuuchi, E. (1983). Transfer of *Peptococcus indolicus*, *Peptococcus asaccharolyticus*, *Peptococcus prevotii*, and *Peptococcus magnus* to the Genus *Peptostreptococcus* and Proposal of *Peptostreptococcus tetradius* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, *33*, 683–698. https://doi.org/10.1099/00207713-33-4-683

Fournier, P. E., & Drancourt, M. (2015). New Microbes New Infections promotes modern prokaryotic taxonomy: A new section "TaxonoGenomics: New genomes of microorganisms in humans". *New Microbes and New Infections*, *7*, 48–49. https://doi.org/10.1016/j.nmni.2015.06.001

Fournier, P. E., Lagier, J. C., Dubourg, G., & Raoult, D. (2015). From culturomics to taxonomogenomics: A need to change the taxonomy of prokaryotes in clinical microbiology. *Anaerobe*, *36*, 73–78. https://doi.org/10.1016/j.anaerobe.2015.10.011

Fredricks, D. N., Fiedler, T. L., & Marrazzo, J. M. (2005). Molecular identification of bacteria associated with bacterial vaginosis. *New England Journal of Medicine*, *353*, 1899–1911. https://doi.org/10.1056/NEJMoa043802

Gouret, P., Paganini, J., Dainat, J., Louati, D., Darbo, E., Pontarotti, P., & Levasseur, A. (2011). Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: The multi-agent software system DAGOBAH. In P. Pontarotti (Ed.), *Evolutionary biology – Concepts, biodiversity, macroevolution and genome*

evolution (pp. 71–87). New York, NY: Springer, Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20763-1

Gouret, P., Vitiello, V., Balandraud, N., Gilles, A., Pontarotti, P., & Danchin, E. G. (2005). FIGENIX: Intelligent automation of genomic annotation: Expertise integration in a new software platform. *BMC Bioinformatics*, *6*, 198. https://doi.org/10.1186/1471-2105-6-198

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Johnson, C. N., Whitehead, T. R., Cotta, M. A., Rhoades, R. E., & Lawson, P. A. (2014). *Peptoniphilus stercorisuis* sp. nov., isolated from a swine manure storage tank and description of *Peptoniphilaceae* fam. nov. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 3538–3545. https://doi.org/10.1099/ijs.0.058941-0

Kim, M., Oh, H. S., Park, S. C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 346–351. https://doi.org/10.1099/ijs.0.059774-0

Klenk, H. P., Meier-Kolthoff, J. P., & Göker, M. (2014). Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 352–356. https://doi.org/10.1099/ijs.0.056994-0

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*, 3100–3108. https://doi.org/10.1093/nar/gkm160

Lagier, J. C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., … Trape, J. F. (2012). Microbial culturomics: Paradigm shift in the human gut microbiome study. *Clinical Microbiology & Infection*, *18*, 1185–1193. https://doi.org/10.1111/1469-0691.12023

Lagier, J. C., Hugon, P., Khelaifia, S., Fournier, P. E., La Scola, B., & Raoult, D. (2015). The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. *Clinical Microbiology Reviews*, *28*, 237–264. https://doi.org/10.1128/CMR.00014-14

Lamont, R., Sobel, J., Akins, R., Hassan, S., Chaiworapongsa, T., Kusanovic, J., & Romero, R. (2011). The vaginal microbiome: New information about genital tract flora using molecular based techniques. *BJOG: An International Journal of Obstetrics and Gynaecology*, *118*, 533–549. https://doi.org/10.1111/j.1471-0528.2010.02840.x

Le Page, S., van Belkum, A., Fulchiron, C., Huguet, R., Raoult, D., & Rolain, J.-M. (2015). Evaluation of the PREVI® Isola automated seeder system compared to reference manual inoculation for antibiotic susceptibility testing by the disk diffusion method. *European Journal of Clinical Microbiology and Infectious Diseases*, *34*, 1859–1869. https://doi.org/10.1007/s10096-015-2424-8

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (Co-) orthologs in large-scale analysis. *BMC Bioinformatics*, *12*, 124. https://doi.org/10.1186/1471-2105-12-124

Lepargneur, J. P., & Rousseau, V. (2002). Protective role of the Doderleïn flora. *Journal de Gynecologie, Obstetrique et Biologie de la Reproduction*, *31*, 485–494.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*, 955–964. https://doi.org/10.1093/nar/25.5.0955

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, *14*, 60. https://doi.org/10.1186/1471-2105-14-60

Menard, J., Fenollar, F., Henry, M., Bretelle, F., & Raoult, D. (2008). Molecular Quantification of *Gardnerella vaginalis* and *Atopobium vaginae* Loads to Predict Bacterial Vaginosis. *Clinical Infectious Diseases*, *47*, 33–43. https://doi.org/10.1086/588661

Murdoch, D., & Shah, H. N. (1999). Reclassification of *Peptostreptococcus magnus* (Prevot 1933) Holdeman and Moore 1972 as *Finegoldia magna* comb. nov. and Peptostreptococcus micros (Prevot 1933) Smith 1957 as *Micromonas micros* comb. nov. *Anaerobe*, *5*, 555–559. https://doi.org/10.1006/anae.1999.0197

Murray, P. R., Baron, E. J., Jorgensen, J. H., Landry, M. L., & Pfaller, M. A. (2007). *Manual of clinical microbiology* (9th ed.). Washington, D.C.: ASM Press.

Pal, K., Roy, S., Behera, B., Kumar, N., Sagiri, S., & Ray, S. (2011). Bacterial vaginosis: Etiology and modalities of treatment-A brief note. *Journal of Pharmacy And Bioallied Sciences*, *3*, 496. https://doi.org/10.4103/0975-7406.90102

Pépin, J., Deslandes, S., Giroux, G., Sobéla, F., Khonde, N., Diakité, S., … Frost, E. (2011). The complex vaginal flora of west african women with bacterial vaginosis. *PLoS ONE*, *6*, e25082. https://doi.org/10.1371/journal.pone.0025082

Ramasamy, D., Mishra, A. K., Lagier, J. C., Padhmanabhan, R., Rossi, M., Sentausa, E., … Fournier, P. E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 384–391. https://doi.org/10.1099/ijs.0.057091-0

Sasser, M. (2006). Bacterial identification by gas chromatographic analysis of fatty acids methyl esters (GC-FAME). MIDI, Technical Note.

Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P., Rolain, J. M., & Raoult, D. (2009). Ongoing revolution in bacteriology: Routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Clinical Infectious Diseases*, *49*, 543–551. https://doi.org/10.1086/600885

Shipitsyna, E., Roos, A., Datcu, R., Hallén, A., Fredlund, H., Jensen, J. S., … Unemo, M. (2013). Composition of the vaginal microbiota in women of reproductive age – Sensitive and specific molecular diagnosis of bacterial vaginosis is possible? *PLoS ONE*, *8*, e60670. https://doi.org/10.1371/journal.pone.0060670

Stackebrandt, E., & Ebers, J. (2006). Taxonomic parameters revisited: Tarnished gold standards. *Microbiology Today*, *33*, 152.

Svare, J., Schmidt, H., Hansen, B., & Lose, G. (2006). Bacterial vaginosis in a cohort of Danish pregnant women: Prevalence and relationship with preterm delivery, low birthweight and perinatal infections. *BJOG: An International Journal of Obstetrics and Gynaecology*, *113*, 1419–1425. https://doi.org/10.1111/j.1471-0528.2006.01087.x

Tindall, B. J., & Euzeby, J. P. (2006). Proposal of *Parvimonas* gen. nov. and *Quatrionicoccus* gen. nov. as replacements for the illegitimate, prokaryotic, generic names *Micromonas* Murdoch and Shah 2000 and *Quadricoccus* Maszenan et al. 2002, respectively. *International Journal of Systematic and Evolutionary Microbiology*, *56*, 2711–2713. https://doi.org/10.1099/ijs.0.64338-0

Turovskiy, Y., Sutyak Noll, K., & Chikindas, M. L. (2011). The etiology of bacterial vaginosis. *Journal of Applied Microbiology*, *110*, 1105–1128. https://doi.org/10.1111/j.1365-2672.2011.04977.x

Ulger-Toprak, N., Liu, C., Summanen, P. H., & Finegold, S. M. (2010). *Murdochiella asaccharolytica* gen. nov., sp. nov., a Gram-stain-positive, anaerobic coccus isolated from human wound specimens. *International Journal of Systematic and Evolutionary Microbiology*, *60*, 1013–1016. https://doi.org/10.1099/ijs.0.015909-0

Vandamme, P., Pot, B., Gillis, M., De Vos, P., Kersters, K., & Swings, J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews*, *60*, 407–438.

How to cite this article: Diop K, Diop A, Khelaifia S, et al. Characterization of a novel Gram-stain-positive anaerobic coccus isolated from the female genital tract: Genome sequence and description of *Murdochiella vaginalis* sp. nov. *MicrobiologyOpen*. 2018;e570. https://doi.org/10.1002/mbo3.570

Article 13:

Description of three new species belonging to genus *Peptoniphilus* isolated from the vaginal fluid of a patient suffering with bacterial vaginosis: *Peptoniphilus vaginalis* sp. nov., *Peptoniphilus raoultii* sp. nov., and *Peptoniphilus pacaensis* sp. nov.

Diop Kh, Diop A, Cadoret F, Michelle C, Richez M, Rathored J, Raoult D, Bretelle F, Fournier PE and Fenollar F

[Published in MicrobiologyOpen]

Received: 1 February 2018 | Revised: 5 April 2018 | Accepted: 30 April 2018

DOI: 10.1002/mbo3.661

ORIGINAL ARTICLE

WILEY MicrobiologyOpen Open Access

Description of three new *Peptoniphilus* species cultured in the vaginal fluid of a woman diagnosed with bacterial vaginosis: *Peptoniphilus pacaensis* sp. nov., *Peptoniphilus raoultii* sp. nov., and *Peptoniphilus vaginalis* sp. nov.

Khoudia Diop[1] | Awa Diop[1] | Caroline Michelle[2] | Magali Richez[2] | Jaishriram Rathored[1] | Florence Bretelle[2,3] | Pierre-Edouard Fournier[1] | Florence Fenollar[1]

[1]Aix Marseille Univ, IRD, AP-HM, SSA, VITROME, IHU-Méditerranée Infection, Marseille, France

[2]Aix-Marseille Univ, IRD, AP-HM, MEPHI, IHU-Méditerranée Infection, Marseille, France

[3]Department of Gynecology and Obstetrics, Gynépole, Hôpital Nord, AP-HM, Marseille, France

Correspondence
Florence Fenollar
Institut Hospitalo-Universitaire Méditerranée-Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille Cedex 05, France.
Email: florence.fenollar@univ-amu.fr

Méditerranée Infection and the National Research Agency under the program "Investissements d'avenir", reference ANR-10-IAHU-03, supported this study.

Abstract

Three previously unidentified Gram-positive anaerobic coccoid bacteria, strains KhD-2^T, KHD4^T, and Kh-D5^T, isolated from a vaginal swab, were characterized using the taxonogenomics concept. The phylogenic analysis, phenotypic characteristics, and genotypic data presented in this report attest that these three bacteria are distinct from previously known bacterial species with standing in nomenclature and represent three new *Peptoniphilus* species. Strain KhD-2^T is most closely related to *Peptoniphilus* sp. DNF00840 and *Peptoniphilus harei* (99.7% and 98.2% identity, respectively); strain KHD4^T to *Peptoniphilus lacrimalis* (96%) and strain Kh-D5^T to *Peptoniphilus coxii* (97.2%). Strains KhD-2^T, KHD4^T, and Kh-D5^T DNA G+C contents are, respectively, 34.23%, 31.87%, and 49.38%; their major fatty acid was $C_{16:0}$ (41.6%, 32.0%, and 36.4%, respectively). We propose that strains KhD-2^T (=CSUR P0125 = DSM 101742), KHD4^T (=CSUR P0110 = CECT 9308), and Kh-D5^T (=CSUR P2271 = DSM 101839) be the type strains of the new species for which the names *Peptoniphilus vaginalis* sp. nov., *Peptoniphilus raoultii* sp. nov., and *Peptoniphilu pacaensis* sp. nov., are proposed, respectively.

KEYWORDS

bacterial vaginosis, culturomics, human microbiota, *Peptoniphilus pacaensis*, *Peptoniphilus raoultii*, *Peptoniphilus vaginalis*, taxogenomics

1 | INTRODUCTION

Since the 1800s, physicians and researchers investigate the vaginal bacterial community using both cultivation and culture-independent methods (Pandya et al., 2017; Srinivasan et al., 2016). To date, many species from the vaginal microbiota have been identified. The healthy vaginal flora is associated to a biotope rich in *Lactobacilli* species (Li, McCormick, Bocking, & Reid, 2012). The vaginal microbiota has a beneficial relationship with its host and can also impact women's health, that of their partners as well as their neonates (Lepargneur & Rousseau, 2002; Srinivasan & Fredricks, 2008). A depletion of vaginal *Lactobacilli* can lead to bacterial vaginosis (BV). This disease is a dysbiosis that may be associated to sexually transmitted infections as well as miscarriage and preterm birth in pregnant women (Afolabi, Moses, & Oduyebo, 2016; Martin & Marrazzo, 2016).

A microbial culturomics study exploring the bacterial community of the vaginal econiche flora in healthy women and patients suffering from bacterial vaginosis enabled the isolation of three Gram-positive-staining, anaerobic, and coccoid bacteria in the vaginal discharge of a woman with bacterial vaginosis (Lagier et al., 2015, 2016). These bacteria exhibited phylogenetic and phenotypic proximity to species of the *Peptoniphilus* genus. Created after the division of *Peptostreptococcus* genus into five genera (Ezaki et al., 2001), the *Peptoniphilus* genus belonging to the Peptoniphilaceae family that regroup members of the genera *Peptoniphilus*, *Parvimonas*, *Murdochiella*, *Helcococcus*, *Gallicola*, *Finegoldia*, *Ezakiella*, *Anaerosphaera*, and *Anaerococcus* (Johnson, Whitehead, Cotta, Rhoades, & Lawson, 2014; Patel et al., 2015). The *Peptoniphilus* genus is currently made of 16 valid published species (http://www.bacterio.net/peptoniphilus.html). These bacteria employ amino acids and peptone as a major energy sources (Ezaki et al., 2001). They are mainly cultivated from diverse human samples such as sacral ulcer, vaginal discharge, as well as ovarian, peritoneal, and lacrymal gland abscesses (Ezaki et al., 2001; Li et al., 1992; Ulger-Toprak, Lawson, Summanen, O'Neal, & Finegold, 2012).

Herein, we describe the isolation and taxonogenomic characterization (Fournier, Lagier, Dubourg, & Raoult, 2015) of strains KhD-2^T, KHD4^T, and Kh-D5^T as type strains of three new *Peptoniphilus* species for which the names *Peptoniphilus vaginalis* sp. nov. (=CSUR P0125, =DSM 101742), *Peptoniphilus raoultii* sp. nov. (=CSUR P0110, =CECT 9308), and *Peptoniphilus pacaensis* sp. nov. (=CSUR P2271, =DSM 101839), are proposed, respectively. All the three strains were cultivated from the vaginal swab of the same patient.

2 | MATERIALS AND METHODS

2.1 | Samples and ethics

The vaginal specimen from a French 33-year-old woman with bacterial vaginosis was sampled at Hospital Nord in Marseille (France) in October 2015 using a Sigma Transwab (Medical Wire, Corsham, United Kingdom). Bacterial vaginosis was diagnosed as previously described (Menard, Fenollar, Henry, Bretelle, & Raoult, 2008). The patient had not received any antibiotic for several months. The local IFR48 ethics committee in Marseille (France) authorized the study (agreement number: 09-022). In addition, the patient gave her signed informed consent.

2.2 | Bacterial strain isolation and identification

After sampling, the specimen was preincubated in a blood culture bottle (Becton-Dickinson Diagnostics, Le Pont-de-Claix, France). The blood culture bottle was enriched with 3 ml of sheep blood (bioMérieux, Marcy l'Etoile, France) and 4 ml of rumen fluid, filter-sterilized through a 0.2 µm pore filter (Thermo Fisher Scientific, Villebon-sur-Yvette, France). Various preincubation periods (1, 3, 7, 10, 15, 20, and 30 days) were tested. Then, 50 µl of the supernatant were inoculated on both Colistin-nalidixic acid (CNA) used for selective enrichment of Gram-positive bacteria and trypticase soy agar plates used for cultivation of nonfastidious and fastidious microorganisms (both BD Diagnostics), and then incubated for 4 days under anaerobic conditions at 37°C. Isolated colonies were purified and subsequently identified by matrix-assisted laser-desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry with a Microflex spectrometer (Bruker, Leipzig, Germany) that compared the new spectra with those present in the library (Bruker database and URMITE database, constantly updated), as previously reported (Seng et al., 2009). If the score was >1.99, the bacterium was considered as identified at the genus level (score between 2.0 and 2.299) or species level (score from 2.3 to 3.0). When the score was <1.7, no identification was considered reliable. The 16S rRNA sequence of unidentified isolates was obtained using an ABI Prism 3130xl Genetic Analyzer capillary sequencer (Applied Biosystems, Bedford, MA, USA), as previously described (Morel et al., 2015; Seng et al., 2009). Finally, the sequences were compared to the NCBI nr database using the BLAST algorithm (https://blast.ncbi.nlm.nih.gov/Blast.cgi). If the 16S rRNA sequence similarity value was lower than 98.7%, the isolate was considered as a putative new species (Kim, Oh, Park, & Chun, 2014; Stackebrandt & Ebers, 2006; Yarza et al., 2014).

2.3 | Phylogenetic analysis

The 16S rRNA sequences of isolates not identified using mass spectrometry and those of members of the family Peptoniphilaceae with standing in nomenclature (downloaded from the nr database) were aligned using CLUSTALW (Thompson, Higgins, & Gibson, 1994) with default setting. The phylogenetic inferences were performed using both the neighbor-joining and maximum-likelihood methods with the software MEGA version 6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013).

2.4 | Phenotypic characteristics

For each new isolate, cell morphology was visualized using optical and electron microscopy. Oxidase, catalase, motility, sporulation tests, as well as Gram stain were performed as already reported (Murray, Baron, Jorgensen, Landry, & Pfaller, 2007). Cells were fixed for electron microscopy for at least 1 hour at 4°C with 2.5% glutaraldehyde in a 0.1 mol L^{-1} cacodylate buffer. One drop of cell suspension was deposited for about 5 min on a glow-discharged formvar carbon film on 400-mesh nickel grids (FCF400-Ni, EMS). The grids were dried on a blotting paper. Then, the cells were negatively stained at room temperature for 10 s with a 1% ammonium molybdate solution in filtered water. Micrographs were obtained using a Tecnai G20 Cryo (FEI) transmission electron microscope operated at 200 keV.

In order to characterize the best growth conditions of each isolate, bacteria were inoculated on 5% sheep blood-enriched Columbia agar (bioMérieux) incubated at various atmospheres (aerobic, anaerobic, and microaerophilic) and temperatures (56, 42, 37, 28, and

25°C) (Mishra, Lagier, Nguyen, Raoult, & Fournier, 2013). Several salinity (NaCl concentrations of 0%, 5%, 15%, and 45%) and pH (5, 6, 6.5, 7, and 8.5) conditions were also tested.

Biochemical analyses were realized using various strips (API ZYM, API 20A, and API 50CH) according to the manufacturer's instructions (bioMérieux) (Avguštin, Wallace, & Flint, 1997; Durand et al., 2017). The tests were performed in anaerobic chamber. The strips were incubated there for 4, 24, and 48 hr, respectively.

For the analysis of cellular fatty acid methyl ester (FAME), gas chromatography/mass spectrometry (GC/MS) was achieved. All three isolates were grown anaerobically at 37°C on 5% sheep blood-enriched Columbia agar (bioMérieux). For each isolate, after 2 days of incubation, two aliquots with roughly 25–70 mg of bacterial biomass per tube were prepared. FAME preparation and GC/MS analyses were performed as already reported (Dione et al., 2016; Sasser, 2006). FAMEs were separated with an Elite 5-MS column and monitored by MS (Clarus 500-SQ 8 S, Perkin Elmer, Courtaboeuf, France). A spectral database search was done with MS Search 2.0 operated using the standard reference database 1A (NIST, Gaithersburg, USA) as well as the FAMEs mass spectral database (Wiley, Chichester, UK).

The susceptibility of all three isolates was tested for 11 antibiotics: amoxicillin (0.16–256 μg/ml), benzylpenicillin (0.002–32 μg/ml), ceftriaxone (0.002–32 μg/ml), ertapenem (0.002–32 μg/ml), imipenem (0.002–32 μg/ml), amikacin (0.16–256 μg/ml), erythromycin (0.16–256 μg/ml), metronidazole (0.16–256 μg/ml), ofloxacin (0.002–32 μg/ml), rifampicin (0.002–32 μg/ml), and vancomycin (0.16–256 μg/ml). Minimal inhibitory concentrations (MICs) were estimated using E-test strips (bioMérieux) and according to EUCAST recommendations (Citron, Ostovari, Karlsson, & Goldstein, 1991; Matuschek, Brown, & Kahlmeter, 2014).

2.5 | Genome sequencing and analyses

After a pretreatment of 2 hr at 37°C using lysozyme, the genomic DNAs (gDNAs) of strains KhD-2^T, KHD4^T, and Kh-D5^T were extracted using the EZ1 biorobot and EZ1 DNA Tissue kit (Qiagen). An elution volume of 50 μl was obtained for each sample. The gDNAs were quantified by a Qubit assay (Life technologies, Carlsbad, CA, USA) at 74.2, 22.4, and 16.4 ng/μl, respectively. Genomic sequencing of each strain was performed with a MiSeq sequencer (Illumina Inc, San Diego, CA, USA) and the Mate Pair strategy.

The Mate Pair library was prepared with the Nextera Mate Pair guide (Illumina) using 1.5 μg of gDNA. The gDNA samples were fragmented and tagged using a Mate Pair junction adapter (Illumina). Then, the fragmentation pattern was validated using a DNA 7500 labchip on an Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA). No size selection was done. Thus, 537, 600, and 480.7 ng of tagmented fragments were, respectively, circularized. Circularized DNAs were mechanically cut to smaller fragments using Optima on a bimodal curve at 507 and 1,244 bp for KhD-2^T, 975 and 1,514 bp for KHD4^T, and 609 and 999 bp for Kh-D5^T on the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The libraries profiles were visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) and the final concentrations libraries were determined. Then, the libraries were normalized at 2 nmol L^{-1}, pooled, denatured, diluted at 15 pmol L^{-1}, loaded onto the reagent cartridge, and onto the instrument. Sequencing was performed in a single 39-hr run in a 2 × 250-bp.

The genome assembly was performed with a pipeline that enabled to create an assembly with various software such as Velvet (Zerbino & Birney, 2008), Spades (Bankevich et al., 2012), and Soap Denovo (Luo et al., 2012), on trimmed data with MiSeq and Trimmomatic (Bolger, Lohse, & Usadel, 2014) software or untrimmed data with only MiSeq software. In order to reduce gaps, GapCloser was used (Luo et al., 2012). Phage contamination was searched (blastn against Phage Phix174 DNA sequence) and eliminated. Finally, scaffolds with sizes under 800 bp and scaffolds with a depth value lower than 25% of the mean depth were identified as possible contaminants and removed. The best assembly was considered by using several criteria including number of scaffolds, N50, and number of N. Spades gave the best assembly for the three studied strains with depth coverage of 518x.

Prodigal was used to predict open reading frames (ORFs) (Hyatt et al., 2010) using default parameters. However, the predicted ORFs were excluded if they spanned a sequencing gap region (containing Ns). The predicted bacterial protein sequences were analyzed as previously reported (Alou et al., 2017). tRNA genes were found using the tRNAScan-SE tool (Lowe & Eddy, 1997), while RNAmmer was used to find ribosomal RNAs (Lagesen et al., 2007). Phobius was used to predict lipoprotein signal peptides and the number of transmembrane helices (Käll, Krogh, & Sonnhammer, 2004). ORFans were identified when the BLASTP search failed to provide positive results (*E*-value smaller than $1e^{-03}$ for ORFs with a sequence size larger than 80 aa or an *E*-value smaller than $1e^{-05}$ for ORFs with a sequence length smaller than 80 aa), as previously reported (Alou et al., 2017). For genomic comparison, the closest species with validly published names in the 16S RNA phylogenetic tree were identified with the Phylopattern software (Gouret, Thompson, & Pontarotti, 2009). The complete genome, proteome, and ORFeome sequences were retrieved for each selected species in NCBI. An annotation of the entire proteome in order to define the distribution of functional classes of predicted genes according to the COG classification of their predicted protein products was performed as already reported (Alou et al., 2017). Annotation and comparison processes were done using the DAGOBAH software as previously described (Alou et al., 2017; Gouret et al., 2005, 2011). Finally, in order to evaluate the genomic similarity between the genomes, we determined two previously described parameters: average amino acid identity (AAI) based on the overall similarity between two genomic datasets of proteins available at (http://enve-omics.ce.gatech.edu/aai/index) and digital DNA–DNA hybridization (dDDH) (Auch, von Jan, Klenk, & Göker, 2010; Meier-Kolthoff, Auch, Klenk, & Göker, 2013; Alou et al., 2017; Rodriguez & Konstantinidis, 2014; Chun et al., 2018).

3 | RESULTS

3.1 | Strain identification and phylogenetic analysis

The MS identification of the three bacteria, secluded, respectively, after 24 hr (strains KhD-2^T and KHD4^T) and 15 days (Kh-D5^T) of preincubation, failed. This suggested that these isolates were not in the database and may be unknown species. Pairwise analysis of 16S rRNA sequences attested that strain KhD-2^T exhibited 92.8% and 87.4% sequence similarities with strains KHD4^T and Kh-D5^T, respectively, and strains KHD4^T and Kh-D5^T had an 88.7% identity. BLASTN sequence searches demonstrated that the three strains were related to the genus *Peptoniphilus*, suggesting that each strain represented a new species within this genus. Strain KhD-2^T exhibited a 16S rRNA similarity of 99.7% with *Peptoniphilus* sp. strain DNF00840 (GenBank KQ960236) over 1,842 bp and 98.2% with *Peptoniphilus harei* (GenBank NR_026358.1) over 1,488 bp. Strain KHD4^T exhibited a 16S rRNA similarity of 96% with *Peptoniphilus lacrimalis* (GenBank NR_041938.1) over 1,489 bp. Finally, strain Kh-D5^T exhibited a 16S rRNA similarity of 97.2% with *Peptoniphilus coxii* (GenBank NR_117556.1) over 1,491 bp (Figure 1). As these percentage similarities were under the threshold of 98.7% established to delineate new species (Kim et al., 2014; Stackebrandt & Ebers, 2006; Yarza et al., 2014), strains KhD-2^T, KHD4^T, and Kh-D5^T were considered as representative strains of putative new *Peptoniphilus* species. The names *P. vaginalis* sp. nov., *P. raoultii* sp. nov., and *P. pacaensis* sp. nov. are, respectively, proposed.

The reference MALDI-TOF MS spectra of our isolates were added in our database (http://www.mediterranee-infection.com/article.php?laref=256&titre=urms-database) and then compared to those of other *Peptoniphilus* spp. (Figure 2).

3.2 | Phenotypic features

Cells from all three novel strains (KhD-2^T, KHD4^T, and Kh-D5^T) were Gram- -positive cocci (mean diameter of 0.6–0.7 μm for each). After 4 days of incubation, colonies on blood agar were grey and circular, and all had a diameter ranging from 1 to 2 mm. For all the three strains, growth occurred only in anaerobic atmosphere. Besides, optimal growth occurred at 37°C, with a pH between 6.5 and 8.5, and a NaCl concentration lower than 5%. They exhibited no catalase, oxidase, and urease activities. Using API 20A strips, all tests including aesculin, arabinose, cellobiose, gelatin, glucose, glycerol, indole, lactose, maltose, mannitol, mannose, raffinose, rhamnose, saccharose, sorbitol, trehalose, urease, and xylose were negative for strains KHD4^T and Kh-D5^T, whereas for strain KhD-2^T, indole formation was positive, and gelatin was hydrolyzed. API ZYM strips showed that the three isolates exhibited positive reactions for acid phosphatase, esterase, and Naphthol-AS-BI-phosphohydrolase. In addition, strains KhD-2^T and KHD4^T had *N*-acetyl-β-glucosaminidase and leucine arylamidase activities. In contrast, an alkaline phosphatase activity was observed for strains KhD-2^T and Kh-D5^T. All other remaining tests including valine arylamidase, lipase, cystine arylamidase, trypsin, galactosidase, glucosidase, β-glucuronidase, α-mannosidase, and α-fucosidase were negative. Using API 50CH strips, all three isolates fermented ribose, tagatose, and potassium-5-ketogluconate. However, they did not ferment adonitol, aesculin, arabinose, arabitol, cellobiose, dulcitol, erythritol, fructose, fucose, galactose, glucose, glycerol, glycogen, inulin, lyxose, inositol, mannose, mannitol, maltose, melibiose, potassium gluconate, potassium-2-ketogluconate, salicine, saccharose, sorbitol, sorbose, trehalose, melezitose, raffinose, rhamnose, starch, turanose, xylitol, and xylose. Table 1 displayed the phenotypic differences between these bacteria and other *Peptoniphilus* spp.

The fatty acid composition of the three strains was as following: strain KhD-2^T contained saturated acid $C_{16:0}$ (41.6%) and $C_{14:0}$ (14.7%); unsaturated acids were also detected (Table 2); strains KHD4^T and Kh-D5^T contained $C_{16:0}$ (32% and 36%, respectively), $C_{18:2\omega6}$ (26% and 24%, respectively), and $C_{18:1\omega9}$ (26% and 21%, respectively) (Table 2). These fatty acid results were likened to those of related species in Table 2 (Johnson et al., 2014; Rooney, Swezey, Pukall, Schumann, & Spring, 2011). Strain KhD-2^T can be distinguished from its nearest neighbor *P. harei* by the production of $C_{14:0}$ (14.7% vs. 4.4%). Strain KHD4^T can be distinguished from its closest related species *P. lacrimalis* by the presence of fatty acids: $C_{14:0}$, $C_{17:0}$ iso 3-OH, and anteiso-$C_{17:0}$. Finally, strain Kh-D5^T showed a fairly similar profile with its neighbors *P. coxii* and *Peptoniphilus ivorii* with some differences such as the presence of antesio-$C_{5:0,}$ only in strain Kh-D5^T (4.5%), of iso-$C_{5:0}$ in *P. coxii* (5.5%), and $C_{17:0}$ iso 3-OH and antesio-$C_{17:0}$, solely in *P. ivorii* (7.7% and 3.8%, respectively). Besides, the three strains were sensitive to amoxicillin, benzylpenicillin, ceftriaxone, ertapenem, imipenem, metronidazole, rifampicin, and vancomycin, but resistant to amikacin, erythromycin, and ofloxacin (Table 3).

3.3 | Genome characteristics

Strains KhD-2^T, KHD4^T, and Kh-D5^T exhibited genomes sizes of 1,877,211, 1,623,601, and 1,851,572 bp long, respectively (Figure 3). The genome characteristics were detailed in Table 4. The repartition of genes into the 25 general COG categories was represented in Table 5 and Figure 4. When compared to other *Peptoniphilus* species, the three strains had genome sizes, G+C contents and total gene counts in the same range (Table 6, Figure 5). Although, base composition varies widely among bacterial species, the genes within a given genome are relatively similar in G+C content with the exception of recently acquired genes. As a matter of fact, DNA sequences acquired by horizontal transfer often bear unusual sequence characteristics and can be distinguished from ancestral DNA notably by a distinct G+C content (Lawrence & Ochman, 1997). The region between 100,000 and 600,000 bp of the chromosome from strain KhD-5^T showed a high variation in G+C content (Figure 3). Thus, 43 genes putatively acquired by horizontal gene transfer were identified in this region, including 25 genes specific for strain KhD-5^T and 18 genes shared with strain *Peptoniphilus urinimassiliensis*. Consequently, the presence of these genes may play a role in the

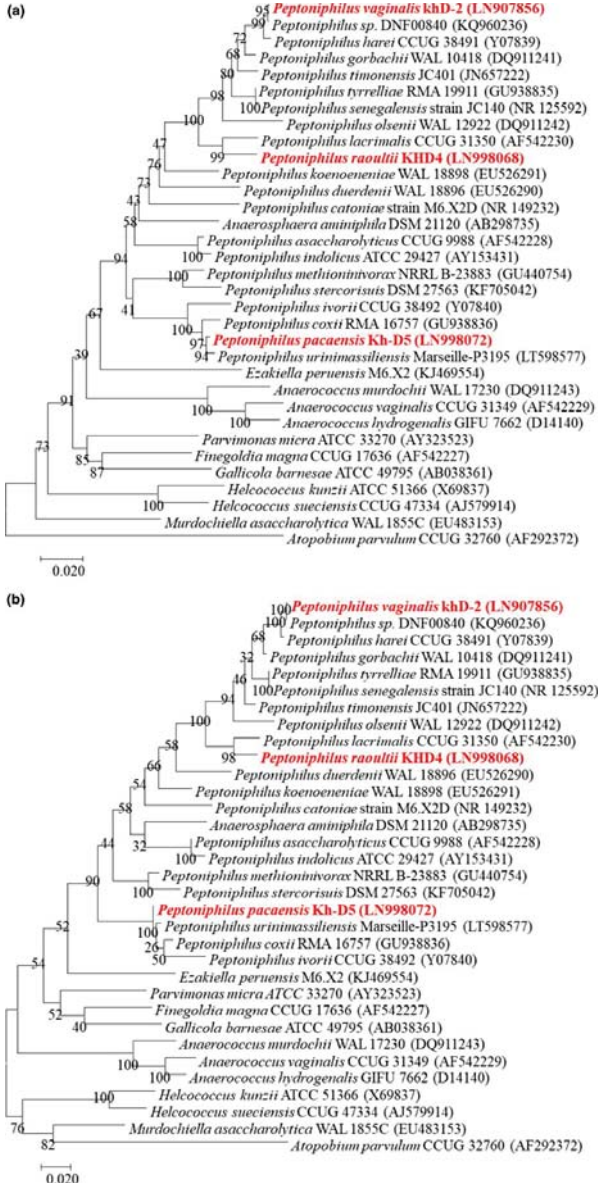


FIGURE 1 Phylogenetic analysis based on the 16S RNA gene sequence highlighting the position of *Peptoniphilus vaginalis* strain KhD-2^T, *Peptoniphilus raoultii* strain KHD4^T, and *Peptoniphilus pacaensis* strain Kh-D5^T relative to other closely related strains. GenBank accession numbers are indicated in parentheses. Sequences were aligned using Muscle v3.8.31 with default parameters and, phylogenetic inferences were performed using the neighbor-joining (a) and maximum-likelihood (b) methods with the software MEGA version 6. The scale bar represents a 2% nucleotide sequence divergence

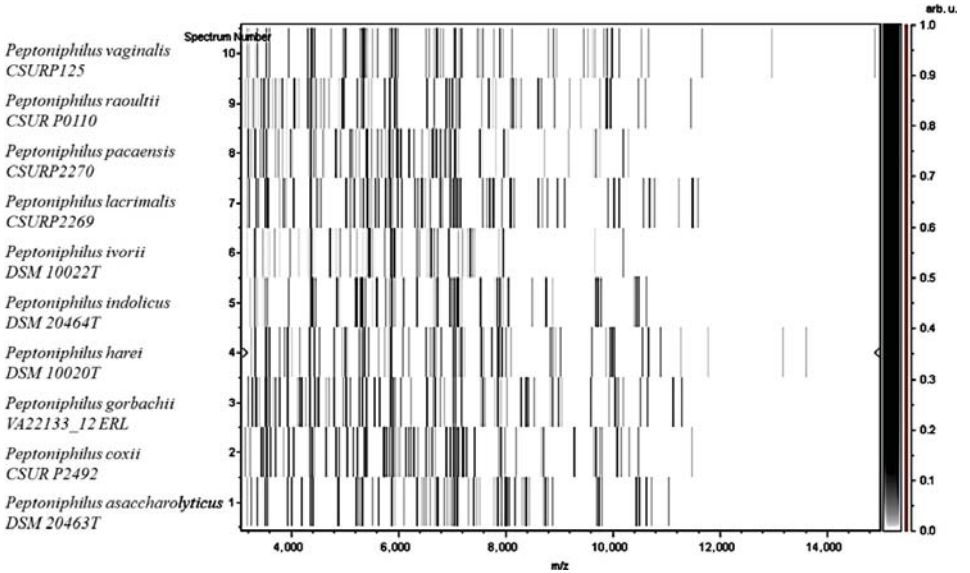


FIGURE 2 Gel view comparing strains KhD-2^T, KHD4^T, and Kh-D5^T to other species within the genus *Peptoniphilus*. The gel view displays the raw spectra of loaded spectrum files arranged in a pseudo-gel-like look. The *x*-axis records the *m/z* value. The left *y*-axis displays the running spectrum number originating from subsequent spectra loading. The peak intensity is expressed by a gray scale scheme code. The right *y*-axis indicates the relation between the color of a peak and its intensity, in arbitrary units. Displayed species are indicated on the left

significant difference in genomic G+C content observed between strain KhD-5^T and other compared *Peptoniphilus* species as well as the similar genomic G+C content observed between strain KhD-5^T and *P. urinimassiliensis*.

The dDDH values ranked from 20.1% ± 2.3% between *P. harei* and *P. duerdenii* to 56.4% ± 2.75% between *P. lacrimalis* and *P. urinimassiliensis* (Table 7). When comparing the three new strains to other *Peptoniphilus* species, strain KhD-2^T exhibited dDDH values ranging from 22.7% ± 2.4% with *Peptoniphilus indolicus* to 47.3% ± 2.55% with *P. coxii*; dDDH values from strain KHD4^T ranged from 19.0% ± 2.25% with *P. harei* to 44.3% ± 2.55% with *P. coxii*; and strain Kh-D5^T exhibited dDDH values ranging from 20.7% ± 2.35% with *P. coxii* to 45.0% ± 2.60% with *P. urinimassiliensis* (Table 7). Furthermore, the AAI values ranged from 51.3% between *P. coxii* and *P. indolicus* to 84.0% between *P. indolicus* and *Peptoniphilus asaccharolyticus* (Table 8). Comparing the three new isolates to their neighbors, strain KhD-2^T shared AAI values ranging from 51.5% with *P. urinimassiliensis* to 92.9% with *P. harei*, AAI values of strain KHD4^T ranging from 50.9% with *P. urinimassiliensis* to 70.6% with *P. lacrimalis*, and strain Kh-D5^T exhibited AAI values ranging from 50.2% with *P. asaccharolyticus* to 92.9% with *P. urinimassiliensis* (Table 8). According to the fact that the threshold of dDDH and AAI values for distinguishing different species are 70% and 95%–96%, respectively (Chun et al., 2018; Klappenbach et al., 2007; Meier-Kolthoff et al., 2013; Richter & Rosselló-Móra, 2009; Rodriguez-R & Konstantinidis, 2014), these data confirm the classification of strains KhD-2^T, KHD4^T, and Kh-D5^T in distinct species.

4 | DISCUSSION

The aim of this study was to investigate, using culturomics, the vaginal flora of a woman with bacterial vaginosis. Indeed, bacterial vaginosis is a gynecologic disorder marked by a perturbation of the vaginal microbiota equilibrium with a loss of commensal *Lactobacillus* spp. and their replacement with anaerobic bacteria including *Atopobium vaginae*, *Bacteroides* spp., *Mobiluncus* spp., *Prevotella* spp., and numerous Gram-positive anaerobic cocci (Bradshaw et al., 2006; Onderdonk, Delaney, & Fichorova, 2016; Shipitsyna et al., 2013). Gram-positive anaerobic cocci were associated to various infections (Murdoch, 1998). They represent about 24%–31% of anaerobic bacteria cultivated in clinical specimens (Murdoch, Mitchelmore, & Tabaqchali, 1994). In this present study, three novel Gram-positive-staining, anaerobic cocci (KhD-2^T, KHD4^T, and Kh-D5^T) were cultured in the vaginal discharge of a patient suffering from bacterial vaginosis. These bacteria exhibited sufficient MALDI-TOF MS profiles, 16S rRNA sequence,

TABLE 1 Compared phenotypic characteristics of *Peptoniphilus vaginalis* strain KhD-2^T, *Peptoniphilus raoultii* strain KHD4^T, *Peptoniphilus pacaensis* strain Kh-D5^T, and other closely related *Peptoniphilus* species. Data were obtained from the original descriptions of species

Properties	*P. vaginalis*	*P. raoultii*	*P. pacaensis*	*P. harei*	*P. lacrimalis*	*P. coxii*	*P. duerdenii*	*P. indolicus*	*P. asaccharolyticus*
Cell diameter (μm)	0.66	0.7	0.7	0.5–1.5	0.5–0.7	<0.7	≥0.7	0.7–1.6	0.5–1.6
% G+C	34.23	31.87	49.38	34.44	30.22	44.62	34.24	31.69	32.30
Major fatty acid (%)	$C_{16:00}$ (41.6)	$C_{16:00}$ (32)	$C_{16:00}$ (36.4)	$C_{16:00}$ (31.2)	$C_{16:00}$ (27.7)	$C_{16:00}$ (49.9)	$C_{16:00}$ (33)	$C_{16:00}$ (19.4)	C18:2ω6 (22.0)
Production of									
Alkaline phosphatase	+	−	+	−	−	−	−	+	+
Indole	+	−	−	+	−	−	+	+	−
Catalase	−	−	−	+	na	−	−	−	−
Urease	−	−	−	−	−	−	−	−	−
β-galactosidase	−	−	−	−	−	−	−	−	−
N-Acetyl-β-glucosaminidase	+	+	−	na	na	−	−	na	na
Acid from									
Ribose	+	+	+	−	−	−	−	−	−
D-fructose	+	−	−	−	−	−	−	−	−
Habitat	Human vagina	Human vagina	Human vagina	Human sacral ulcer	Human eyes	Human specimens	Human vagina	Summer mastitis of cattle	Human vagina

+, positive; −, negative; v, variable and na (not available) data.

TABLE 2 Cellular fatty acid profiles (%) of strains KhD-2^T, KHD4^T, and Kh-D5^T compared with other *Peptoniphilus* species

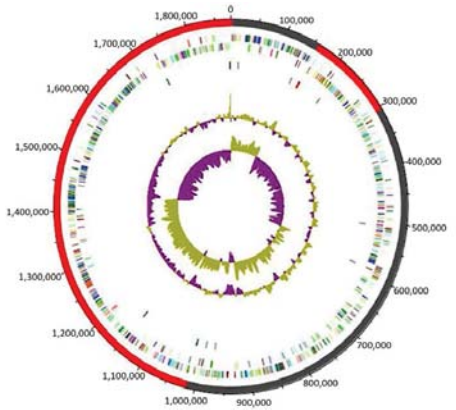
Fatty acids	Name	1	2	3	4	5	6	7	8	9	10
C4:00	Butanoic acid	TR	–	–	–	–	–	–	–	–	–
iso-C5:0	3-Methyl-butanoic acid	–	–	–	–	–	5.5	–	–	–	–
anteiso-C5:0	2-Methyl-butanoic acid	TR	–	4.5	–	–	–	–	–	–	–
C10:0	Decanoic acid	–	–	TR	TR	–	–	2.8	TR	–	–
C12:0	Dodecanoic acid	TR	–	TR	–	TR	TR	–	1.2	TR	2.3
C13:0	Tridecanoic acid	TR	–	–	–	–	–	–	–	–	–
C14:0	Tetradecanoic acid	**14.7**	TR	4.9	4.4	2.9	8.6	4.4	**12.6**	4.4	5.4
C14:1ω5	9-Tetradecenoic acid	TR	–	–	–	–	–	–	–	–	–
C15:0	Pentadecanoic acid	1.1	TR	TR	–	–	1.4	–	–	–	–
C16:0	Hexadecanoic acid	**41.6**	**32.0**	**36.4**	**32.1**	**27.7**	**49.9**	**33.0**	**19.4**	**29.5**	**14.4**
C16:0 9,10-methylene	2-Hexyl-cyclopropaneoctanoic acid	–	TR	–	–	–	–	–	–	–	–
C16:1ω5	11-Hexadecenoic acid	TR	–	–	–	–	–	–	–	–	–
C16:1ω7	9-Hexadecenoic acid	6.2	1.0	TR	1.0	3.2	–	–	–	1.0	3.9
C16:1ω9	7-Hexadecenoic acid	TR	–	–	–	–	–	–	3.6	–	–
C17:0	Heptadecanoic acid	TR	TR	TR	–	–	–	–	–	–	–
C17:0 iso 3-OH	3-Hydroxy-heptadecanoic acid	–	–	–	6.0	3.0	–	–	–	7.7	-
anteiso-C17:0	14-Methyl-hexadecanoic acid	TR	–	–	4.2	1.8	–	–	2.6	3.8	1.6
C17:1ω7	10-Heptadecenoic acid	TR	–	–	–	–	–	–	–	–	–
C18:0	Octadecanoic acid	3.9	8.8	3.6	7.2	**11.2**	**13.1**	**16.2**	2.5	4.8	9.4
C18:1ω7	11-Octadecenoic acid	4.8	3.7	2.0	1.9	3.5	–	–	3.5	2.6	–
C18:1ω9	9-Octadecenoic acid	**12.1**	**25.8**	**21.2**	**17.0**	**25.7**	**17.3**	**22.6**	6.2	**11.4**	**20.2**
C18:2ω6	9,12-Octadecadienoic acid	**12.0**	**26.4**	**24.4**	**17.0**	**13.6**	3.2	**21.1**	**13.0**	**24.0**	**22.0**

Strains: 1, *P. vaginalis* strain KhD-2^T; 2, *P. raoultii* strain KHD4^T; 3, *P. pacaensis* strain Kh-D5^T; 4, *Peptoniphilus harei* DSM 10020^T; 5, *P. lacrimalis* DSM 7455^T; 6, *P. coxii* CSUR 2492^T; 7, *P. uerdenii* WAL 18896^T; 8, *P. indolicus* DSM 20464^T, 9, *P. ivorii* CCUG 38492^T and 10, *P. asaccharolyticus* CCUG 9988^T. Strains 1, 2, 3, and 6 data are from this study and strains 4, 5, 7 to 9, data come from Rooney et al., 2011 and Johnson et al., 2014. Predominant products are shown in bold; TR, trace amounts < 1%; –, not detected.

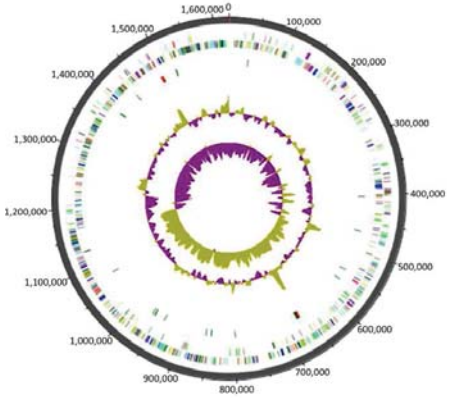
Antibiotics	Concentration (μg/ml)	*P. vaginalis* strain KhD-2^T	*P. raoultii* strain KHD4^T	*P. pacaensis* strain Kh-D5^T
Amoxicillin	0.016–256	0.032	0.016	0.016
Benzylpenicillin	0.002–32	0.094	0.002	0.002
Ceftriaxone	0.002–32	0.064	0.064	0.064
Ertapenem	0.002–32	0.002	0.003	0.002
Imipenem	0.002–32	0.004	0.002	0.002
Metronidazole	0.016–256	0.125	0.032	0.032
Rifampicin	0.002–32	0.002	0.002	0.002
Vancomycin	0.016–256	0.094	0.094	0.094
Amikacin	0.016–256	>256	>256	>256
Erythromycin	0.016–256	1	2	2
Ofloxacin	0.002–32	>256	>256	2

TABLE 3 Minimal inhibitory concentrations (MIC μg/μl) of antibiotics for *P. vaginalis* strain KhD-2^T, *P. raoultii* strain KHD4^T, and *P. pacaensis* strain Kh-D5^T

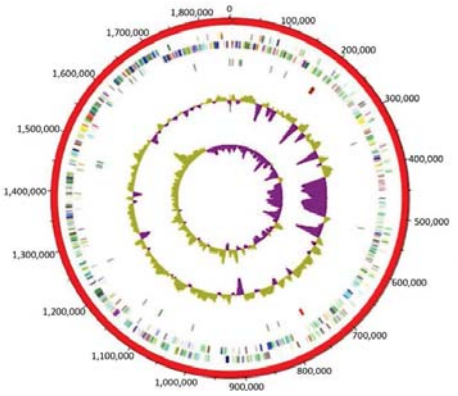
FIGURE 3 Graphical circular map of the three genomes. From outside to the center: Contigs (red/gray), COG category of genes on the forward strand (three circles), genes on forward strand (blue circle), genes on the reverse strand (red circle), COG category on the reverse strand (three circles), G+C content



Peptoniphilus vaginalis strain KhD-2^T



Peptoniphilus raoultii strain KHD4^T



Peptoniphilus pacaensis strain Kh-D5^T

TABLE 4 Nucleotide and gene count levels of the genomes

	P. raoultii		*P. vaginalis*		*P. vaginalis*	
Attribute	**Value**	**% of total[a]**	**Value**	**% of total[a]**	**Value**	**% of total[a]**
Size (bp)	1,623,601	100%	1,877,211	100%	1,851,572	100%
G+C content (bp)	517,506	31.87%	642,534	34.22%	914,357	49.38%
Coding region (bp)	1,467,557	90.39%	1,692,527	90.16	3,579,496	85.07%
Total genes	1,624	100%	1,780	100%	1,801	100%
RNA genes	42	2.59%	40	2.35%	54	3.00%
Protein-coding genes	1,520	93.60%	1,698	95.39%	1,699	94.34%
Genes with function prediction	1,222	75.25%	1,375	77.24%	1,323	73.45%
Genes assigned to COGs	1,048	65.53%	1,204	67.64%	1,175	65.24%
Genes with peptide signals	162	9.97%	169	9.49%	231	12.83%
Genes with transmembrane helices	349	21.49%	403	22.64%	414	22.98%

[a]The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

phenotypic, and genomic differences with *Peptoniphilus* species to be regarded as representative strains of three new species within this genus. Currently, this genus contains 16 species with validly published names. Most of them have been observed in human clinical specimens (Ezaki et al., 2001).

Data from phylogenetic analysis and genomic comparison exhibited the heterogeneity of this genus and revealed that strain KhD-2^T and *Peptoniphilus* sp. DNF00840^T share 99.79% 16S rRNA gene sequence similarity, an ANI value of 96.83% and 75.0% of dDDH. In fact, to differentiate bacterial species, thresholds lower than 98.7%, 94%, and 70% were delimited for 16S rRNA sequence identity, ANI, and dDDH values, respectively. Therefore, the obtained values suggest that the two strains (KhD-2^T and *Peptoniphilus* sp. DNF00840^T) belong to the same species. Unlike other *Peptoniphilus* spp., strains KhD-2^T, KHD4^T, and Kh-D5^T ferment ribose and tagatose. The study of their genomes revealed that strain Kh-D2^T had 75 genes associated to carbohydrate metabolism, including 4 genes (1 *rbsA* gene, 2 *rbsR* genes, and 1 *rpiB* gene) encoding proteins involved in fermentation of ribose; the genome from strain KHD4^T contained 61 genes associated to carbohydrate metabolism of which one *rpiB* gene is involved in fermentation of ribose; and strain KhD-5^T had 58 genes associated to carbohydrate metabolism with 3 genes implicated in ribose fermentation (2 *rpiB* genes and 1 *rbsK*) and 1 gene encoding a tagatose biphosphate aldolase enzyme involved in tagatose fermentation. In addition, the genomes of strains Kh-D2^T, KHD4^T, and KhD-5^T also had 25 genes (5 genes encoding proteins responsible for the degradation of histidine, 1 of lysine, 2 of threonine, 12 of methionine, and 5 of arginine), 20 genes (5 of histidine, 1 of lysine, 1 of threonine, 7 of methionine, and 6 of arginine), and 21 genes (14 which degraded methionine, 6 for arginine and 1 for lysine), associated to amino acid degradation, respectively.

Finally, we propose that strains KhD-2^T, KHD4^T, and KhD-5^T are type strains of *P. vaginalis* sp. nov., *P. raoultii* sp. nov., and *P. pacaensis* sp. nov., respectively.

4.1 | Description of *P. vaginalis* sp. nov

Peptoniphilus vaginalis (va.gi.na'lis. L. n. fem. gen. *vaginalis* from the feminine organ vagina; vaginalis pertaining to the vagina).

Gram-stain—positive. Coccus-shaped bacterium with a mean diameter of 0.66 µm. *Peptoniphilus vaginalis* sp. nov. is a mesophilic bacterium; its optimal growth occurs at temperature 37°C, a pH ranking from 6.5 to 8.5, and a NaCl concentration lower than 5%. Colonies are circular, translucent, gray, and have a diameter of 1–1.5 mm on Columbia agar. Cells are strictly anaerobic, not motile, and non-spore-forming. Catalase, oxidase, and urease activities are negative. Nitrate reduction is also negative nevertheless indole production is positive. *P. vaginalis* shows positive enzymatic activities for acid phosphatase, alkaline phosphatase, esterase, esterase lipase, leucine arylamidase, Naphthol-AS-BI-phosphohydrolase, and *N*-acetyl-β-glucosaminidase. *P. vaginalis* ferments fructose, potassium 5-ketogluconate, ribose, and tagatose. $C_{16:0}$, $C_{14:0}$, $C_{18:1\omega9}$, and $C_{18:2\omega6}$ are its main fatty acids. Strain KhD-2^T is sensitive to amoxicillin, benzylpenicillin, ceftriaxone, imipenem, ertapenem, metronidazole, rifampicin, and vancomycin but resistant to amikacin, erythromycin, and ofloxacin. Its 1,623,601-bp genome contains 34.23% G+C. In EMBL-EBI, the 16S rRNA gene sequence is deposited under accession number LN907856 and the draft genome sequence under accession number FXLP00000000. The type strain of *Peptoniphilus vaginalis* sp. nov. is strain KhD-2^T (=CSUR P0125 = DSM 101742), which was cultured from the vaginal discharge of a woman suffering from bacterial vaginosis.

4.2 | Description of *P. raoultii* sp. nov

Peptoniphilus raoultii (ra.oul'ti.i. N. L. masc. gen. n. *raoultii* of Raoult, to honor French scientist Professor Didier Raoult for his outstanding contribution to medical microbiology).

TABLE 5 Number of genes associated with the 25 general COG functional categories

	P. vaginalis		*P. raoultii*		*P. pacaensis*		
Code	Value	% value	Value	% value	Value	% value	Description
J	170	9.70	170	10.69	171	9.78	Translation
A	0	0	0	0	0	0	RNA processing and modification
K	75	4.28	63	3.96	78	4.46	Transcription
L	64	3.65	65	4.09	63	3.60	Replication, recombination, and repair
B	0	0	0	0	0	0	Chromatin structure and dynamics
D	20	1.14	18	1.13	23	1.31	Cell cycle control, mitosis, and meiosis
Y	0	0	0	0	0	0	Nuclear structure
V	61	3.48	40	2.51	60	2.97	Defense mechanisms
T	44	2.51	43	2.70	52	3.64	Signal transduction mechanisms
M	50	2.85	50	3.14	55	3.14	Cell wall/membrane biogenesis
N	7	0.39	7	0.44	8	0.45	Cell motility
Z	0	0	0	0	0	0	Cytoskeleton
W	3	0.17	3	0.18	2	0.11	Extracellular structures
U	15	0.85	16	1.00	15	0.85	Intracellular trafficking and secretion
O	58	3.31	51	3.20	54	3.08	Posttranslational modification, protein turnover, chaperones
X	68	3.88	22	1.38	44	2.51	Mobilome: prophages, transposons
C	83	4.74	66	4.15	75	4.29	Energy production and conversion
G	40	2.28	47	2.95	48	2.74	Carbohydrate transport and metabolism
E	115	6.56	105	6.60	112	6.40	Amino acid transport and metabolism
F	57	3.25	52	3.27	58	3.31	Nucleotide transport and metabolism
H	71	4.05	52	3.27	84	4.80	Coenzyme transport and metabolism
I	56	3.19	53	3.33	45	2.57	Lipid transport and metabolism
P	68	3.88	48	3.02	69	3.94	Inorganic ion transport and metabolism
Q	19	1.08	18	1.13	11	0.62	Secondary metabolites biosynthesis, transport, and catabolism
R	111	6.33	107	6.73	98	5.60	General function prediction only
S	62	3.54	51	3.20	71	4.06	Function unknown
-	547	31.23	541	34.04	573	32.78	Not in COGs

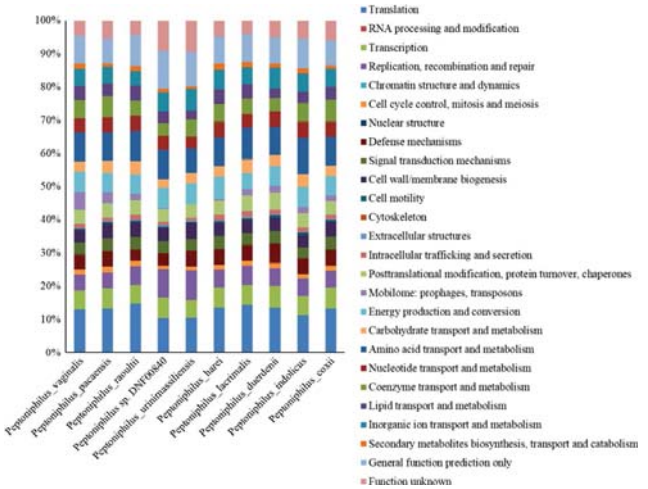


FIGURE 4 Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of *P. vaginalis* strain KhD-2^T, *P. raoultii* strain KHD4^T, and *P. pacaensis* strain Kh-D5^T among other species

TABLE 6 Genome comparison of closely related species to *P. vaginalis* strain KhD-2^T, *P. raoultii* strain KHD4^T, and *P. pacaensis* strain Kh-D5^T

Species	INSDC identifier[a]	Size (Mbp)	G+C Percent	Gene Content	Number of contigs	N50 Value
***P. vaginalis* KhD-2^T**	**FXLP00000000**	**1.88**	**34.2**	**1,791**	**5**	**707,77**
***P. raoultii* KHD4^T**	**FMWM00000000**	**1.62**	**31.9**	**1,631**	**2**	**1,62**
***P. pacaensis* Kh-D5^T**	**FLQT00000000**	**1.85**	**49.4**	**1,802**	**3**	**1,84**
Peptoniphilus sp. DNF00840	LSDH00000000	1.88	34.3	1,671	91	50,04
Peptoniphilus urinimassiliensis Marseille-P3195	FTPC00000000	1.82	49.7	1,770	5	563,37
Peptoniphilus harei ACS-146-V-Sch2b	AENP00000000	1.84	34.4	1,749	32	111,2
Peptoniphilus lacrimalis CCUG 31350	ARKX00000000	1.85	30.2	1,785	22	190,04
Peptoniphilus duerdenii WAL 18896	AEEH00000000	2.12	34.2	1,963	61	96,77
Peptoniphilus indolicus ATCC 29427	AGBB00000000	2.24	31.7	2,145	302	11,79
Peptoniphilus coxii RMA 16757	LSDG00000000	1.84	44.6	1,783	48	103,89
Peptoniphilus asaccharolyticus DSM 20463	FWWR00000000	2.23	32.3	2,054	17	1,358,172

[a]INSDC: International Nucleotide Sequence Database Collaboration. Text and values in bold have been used to highlight new species.

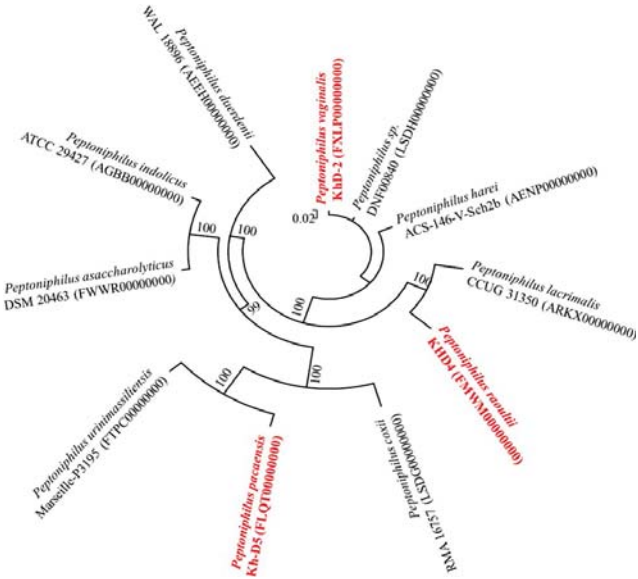


FIGURE 5 Phylogenetic tree based on whole genome sequence showing the position of *P. vaginalis* strain KhD-2^T, *P. raoultii* strain KHD4^T, and *P. pacaensis* strain Kh-D5^T relative to their nearest neighbors. GenBank accession numbers are indicated in parentheses. Sequences were aligned using Mugsy software, and phylogenetic inferences were performed using the maximum likelihood method with the software FastTree. The scale bar represents a 2% nucleotide sequence divergence

TABLE 7 dDDH values obtained by comparison of all studied genomes using GGDC, Formula 2 (DDH Estimates Based on Identities/HSP length)[a]

	P. vaginalis strain KhD-2^T	*P. raoultii* strain KHD4^T	*P. pacaensis* strain Kh-D5^T	*P. urini-massiliensis*	*P. harei*	*P. lacrimalis*	*P. duerdenii*	*P. indolicus*	*P. coxii*	*P. asaccharolyticus*
P. vaginalis	100 ± 00	22.9 ± 2.35	40.0 ± 2.50	35.3 ± 2.50	45.8 ± 2.60	25.6 ± 2.40	32.0 ± 2.45	22.7 ± 2.40	47.3 ± 2.55	33.20 ± 2.45
P. raoultii		100 ± 00	29.8 ± 2.45	40.5 ± 2.50	19.0 ± 2.25	20.4 ± 2.30	36.4 ± 2.55	22.2 ± 2.35	44.3 ± 2.55	28.40 ± 2.45
P. pacaensis			100 ± 00	45.0 ± 2.60	42.0 ± 2.55	41.9 ± 2.55	38.7 ± 2.50	27.3 ± 2.45	20.7 ± 2.35	29.30 ± 2.45
P. urinimassiliensis				100 ± 00	32.9 ± 2.50	56.4 ± 2.75	42.9 ± 2.50	33.0 ± 2.45	20.1 ± 2.30	32.30 ± 2.45
P. harei					100 ± 00	34.3 ± 2.50	39.2 ± 2.50	20.1 ± 2.30	36.2 ± 2.45	33.30 ± 2.45
P. lacrimalis						100 ± 00	39.3 ± 2.50	25.1 ± 2.40	40.6 ± 2.50	31.90 ± 2.45
P. duerdenii							100 ± 00	24.3 ± 2.35	38.2 ± 2.50	32.80 ± 2.50
P. indolicus								100 ± 00	44.0 ± 2.55	26.70 ± 2.45
P. coxii									100 ± 00	35.40 ± 2.45
P. asaccharolyticus										100 ± 00

[a]The confidence intervals indicate the inherent uncertainty in estimating DDH values from intergenomic distances based on models derived from empirical test data sets (which are always limited in size).

TABLE 8 AAI values obtained by comparison of all studied genomes

	P. raoultii strain KHD4^T	*P. pacaensis* strain Kh-D5^T	*P. urini-massiliensis*	*P. harei*	*P. lacrimalis*	*P. duerdenii*	*P. indolicus*	*P. coxii*	*P. asaccharolyticus*
P. vaginalis	62.7	51.2	51.5	92.9	61.5	57.0	55.9	53.2	57.9
P. raoultii		50.0	50.9	61.6	70.6	56.2	55.4	52.5	56.8
P. pacaensis			92.9	51.8	51.2	51.8	50.4	74.1	50.2
P. urinimassiliensis				52.0	52.7	52.2	51.4	73.4	51.3
P. harei					64.2	58.5	56.4	51.7	58.5
P. lacrimalis						58.0	55.9	51.8	57.1
P. duerdenii							54.7	53.1	57.0
P. indolicus								51.3	84.0
P. coxii									51.2

Gram-stain—positive. Coccus-shaped bacterium with a mean diameter of 0.7 µm. *Peptoniphilus raoultii* sp. nov. is a mesophilic bacterium; its optimal growth occurs at temperature 37°C, a pH ranking from 6.5 to 8.5, and a NaCl concentration lower than 5%. Colonies are circular, translucent, gray, and have a diameter of 1–1.5 mm on Columbia agar. Cells are strictly anaerobic, not motile, and non-spore-forming. Catalase, oxidase, urease, indole, and nitrate activities are negative. *P. raoultii* exhibits positive enzymatic activities for acid phosphatase, esterase, esterase lipase, leucine arylamidase, Naphthol-AS-BI-phosphohydrolase, and *N*-acetyl-β-glucosaminidase. *P. raoultii* ferments potassium 5-ketogluconate, ribose, and tagatose. $C_{16:0}$, $C_{18:2\omega6}$, and $C_{18:1\omega9}$ are its main fatty acids. Strain KHD4^T is sensitive to amoxicillin, benzylpenicillin, ceftriaxone, imipenem, ertapenem, metronidazole, rifampicin, and vancomycin but resistant to amikacin, erythromycin, and ofloxacin. The genome is 1,877,211 bp long and contains 31.87% G+C. In EMBL-EBI, the 16*S* rRNA gene sequence is deposited under accession number LN998068 and the draft genome sequence under accession number FMWM00000000. Strain KHD4^T (=CSUR P0110 = CECT 9308) is the type strain of *P. raoultii* sp. nov., which was cultured from the vaginal discharge of a woman suffering from bacterial vaginosis.

4.3 | Description of *P. pacaensis* sp. nov

Peptoniphilus pacaensis (pa.ca.en'sis N. L. gen. masc. n. *pacaensis*, from the acronym PACA, of Provence-Alpes-Côte d'Azur, the region where the type strain was first cultured and characterized).

Gram-stain—positive. Coccus-shaped bacterium with a mean diameter of 0.7 µm. *Peptoniphilus pacaensis* sp. nov. is a mesophilic bacterium; its optimal growth occurs at temperature 37°C, a pH ranking from 6.5 to 8.5, and a NaCl concentration lower than 5%. Colonies are circular, translucent, gray, and have a diameter of 1–1.5 mm on Columbia agar. Cells are strictly anaerobic, not motile, and non-spore-forming. Catalase, oxidase, urease, indole, and nitrate activities are negative. *P. pacaensis* shows positive enzymatic activities for alkaline phosphatase, acid phosphatase, esterase, esterase lipase, and Naphthol-AS-BI-phosphohydrolase. *P. pacaensis* ferments potassium 5-ketogluconate, ribose, and tagatose. $C_{16:0}$, $C_{18:2\omega6}$, and $C_{18:1\omega9}$ are its main fatty acids. Strain Kh-D5^T is sensitive to amoxicillin, benzylpenicillin, ceftriaxone, imipenem, ertapenem, metronidazole, rifampicin, and vancomycin but resistant to amikacin, erythromycin, and ofloxacin. Its genome is 1,851,572 bp long with a 49.38% G+C content. In EMBL-EBI, the 16*S* rRNA gene sequence is deposited under accession number LN998072 and the draft genome sequence under accession number FLQT00000000. The type strain of *P. pacaensis* sp. nov. is strain Kh-D5^T (=CSUR P2270 = DSM 101839), which was cultured from the vaginal discharge of a woman suffering from bacterial vaginosis.

ACKNOWLEDGEMENTS

The authors thank Frederic Cadoret for administrative assistance and the Xegen Company (www.xegen.fr) for automating the genomic annotation process.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Khoudia Diop http://orcid.org/0000-0002-9296-563X

Pierre-Edouard Fournier http://orcid.org/0000-0001-8463-8885

REFERENCES

Afolabi, B. B., Moses, O. E., & Oduyebo, O. O. (2016). Bacterial vaginosis and pregnancy outcome in Lagos, Nigeria. *Open Forum Infectious Diseases*, *3*, ofw030. https://doi.org/10.1093/ofid/ofw030

Alou, M. T., Rathored, J., Michelle, C., Dubourg, G., Andrieu, C., Armstrong, N., … Fournier, P. E. (2017). *Inediibacterium massiliense* gen. nov., sp. nov., a new bacterial species isolated from the gut microbiota of a severely malnourished infant. *Antonie van Leeuwenhoek*, *110*, 737–750. https://doi.org/10.1007/s10482-017-0843-5

Auch, A. F., von Jan, M., Klenk, H.-P., & Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Standards in Genomic Sciences*, *2*, 117–134. https://doi.org/10.4056/sigs.531120

Avguštin, G., Wallace, R. J., & Flint, H. J. (1997). Phenotypic diversity among ruminal isolates of *Prevotella ruminicola*: Proposal of *Prevotella brevis* sp. nov., *Prevotella bryantii* sp. nov., and *Prevotella albensis* sp. nov. and redefinition of *Prevotella ruminicola*. *International Journal of Systematic and Evolutionary Microbiology*, *47*, 284–288.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pyshkin, A. V. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bradshaw, C. S., Tabrizi, S. N., Fairley, C. K., Morton, A. N., Rudland, E., & Garland, S. M. (2006). The association of *Atopobium vaginae* and *Gardnerella vaginalis* with bacterial vaginosis and recurrence after oral metronidazole therapy. *Journal of Infectious Diseases*, *194*, 828–836. https://doi.org/10.1086/506621

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., … Trujillo, M. E. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *68*, 461–466. https://doi.org/10.1099/ijsem.0.002516

Citron, D. M., Ostovari, M. I., Karlsson, A., & Goldstein, E. J. (1991). Evaluation of the E test for susceptibility testing of anaerobic bacteria. *Journal of Clinical Microbiology*, *29*, 2197–2203.

Dione, N., Sankar, S. A., Lagier, J. C., Khelaifia, S., Michele, C., Armstrong, N., … Fournier, P. E. (2016). Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. *New Microbes and New Infections*, *10*, 66–76. https://doi.org/10.1016/j.nmni.2016.01.002

Durand, G. A., Pham, T., Ndongo, S., Traore, S. I., Dubourg, G., Lagier, J. C., … Million, M. (2017). *Blautia massiliensis* sp. nov., isolated from a fresh human fecal sample and emended description of the genus Blautia. *Anaerobe*, *43*, 47–55. https://doi.org/10.1016/j.anaerobe.2016.12.001

Ezaki, T., Kawamura, Y., Li, N., Li, Z.-Y., Zhao, L., & Shu, S. (2001). Proposal of the genera *Anaerococcus* gen. nov., *Peptoniphilus* gen. nov. and *Gallicola* gen. nov. for members of the genus P*eptostreptococcus*. *International Journal of Systematic and Evolutionary Microbiology*, *51*, 1521–1528. https://doi.org/10.1099/00207713-51-4-1521

Fournier, P. E., Lagier, J. C., Dubourg, G., & Raoult, D. (2015). From culturomics to taxonomogenomics: A need to change the taxonomy of prokaryotes in clinical microbiology. *Anaerobe*, *36*, 73–78. https://doi.org/10.1016/j.anaerobe.2015.10.011

Gouret, P., Paganini, J., Dainat, J., Louati, D., Darbo, E., Pontarotti, P., & Levasseur, A. (2011). Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: The multi-agent software system DAGOBAH. In P. Pontarotti (Ed.), *Evolutionary biology – concepts, biodiversity, macroevolution and genome evolution* (pp. 71–87). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-642-20763-1

Gouret, P., Thompson, J. D., & Pontarotti, P. (2009). PhyloPattern: Regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinformatics*, *10*, 298. https://doi.org/10.1186/1471-2105-10-298

Gouret, P., Vitiello, V., Balandraud, N., Gilles, A., Pontarotti, P., & Danchin, E. G. (2005). FIGENIX: Intelligent automation of genomic annotation: Expertise integration in a new software platform. *BMC Bioinformatics*, *6*, 1.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 1.

Johnson, C. N., Whitehead, T. R., Cotta, M. A., Rhoades, R. E., & Lawson, P. A. (2014). *Peptoniphilus stercorisuis* sp. nov., isolated from a swine manure storage tank and description of *Peptoniphilaceae* fam. nov. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 3538–3545. https://doi.org/10.1099/ijs.0.058941-0

Käll, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, *338*, 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Kim, M., Oh, H.-S., Park, S.-C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 346–351. https://doi.org/10.1099/ijs.0.059774-0

Klappenbach, J. A., Goris, J., Vandamme, P., Coenye, T., Konstantinidis, K. T., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*, 81–91.

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*, 3100–3108. https://doi.org/10.1093/nar/gkm160

Lagier, J. C., Hugon, P., Khelaifia, S., Fournier, P. E., La Scola, B., & Raoult, D. (2015). The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. *Clinical Microbiology Reviews*, *28*, 237–264. https://doi.org/10.1128/CMR.00014-14

Lagier, J. C., Khelaifia, S., Alou, M. T., Ndongo, S., Dione, N., Hugon, P., … Durand, G. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature Microbiology*, *12*, 16203. https://doi.org/10.1038/nmicrobiol.2016.203

Lawrence, J. G., & Ochman, H. (1997). Amelioration of bacterial genomes: Rates of change and exchange. *Journal of Molecular Evolution*, *44*, 383–397. https://doi.org/10.1007/PL00006158

Lepargneur, J. P., & Rousseau, V. (2002). Protective role of the Doderleïn flora. *Journal de Gynecologie, Obstetrique et Biologie de la Reproduction*, *31*, 485–494.

Li, N., Hashimoto, Y., Adnan, S., Miura, H., Yamamoto, H., & Ezaki, T. (1992). Three new species of the genus *Peptostreptococcus* isolated from humans: *Peptostreptococcus vaginalis* sp. nov., *Peptostreptococcus lacrimalis* sp. nov., and *Peptostreptococcus lactolyticus* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, *42*, 602–605.

Li, J., McCormick, J., Bocking, A., & Reid, G. (2012). Importance of vaginal microbes in reproductive health. *Reproductive Sciences*, *19*, 235–242. https://doi.org/10.1177/1933719111418379

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*, 955–964. https://doi.org/10.1093/nar/25.5.0955

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Tang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*, 18. https://doi.org/10.1186/2047-217X-1-18

Martin, D. H., & Marrazzo, J. M. (2016). The vaginal microbiome: Current understanding and future directions. *Journal of Infectious Diseases*, *214*, S36–S41. https://doi.org/10.1093/infdis/jiw184

Matuschek, E., Brown, D. F., & Kahlmeter, G. (2014). Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. *Clinical Microbiology & Infection*, *20*, O255–O266. https://doi.org/10.1111/1469-0691.12373

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, *14*, 1.

Menard, J. P., Fenollar, F., Henry, M., Bretelle, F., & Raoult, D. (2008). Molecular quantification of *Gardnerella vaginalis* and *Atopobium vaginae* loads to predict bacterial vaginosis. *Clinical Infectious Diseases*, *47*, 33–43. https://doi.org/10.1086/588661

Mishra, A. K., Lagier, J. C., Nguyen, T. T., Raoult, D., & Fournier, P.-E. (2013). Non contiguous-finished genome sequence and description of *Peptoniphilus senegalensis* sp. nov. *Standards in Genomic Sciences*, *7*, 370–381. https://doi.org/10.4056/sigs.3366764

Morel, A. S., Dubourg, G., Prudent, E., Edouard, S., Gouriet, F., Casalta, J. P., … Raoult, D. (2015). Complementarity between targeted real-time specific PCR and conventional broad-range 16S rDNA PCR in the syndrome-driven diagnosis of infectious diseases. *European Journal of Clinical Microbiology and Infectious Diseases*, *34*, 561–570. https://doi.org/10.1007/s10096-014-2263-z

Murdoch, D. A. (1998). Gram-positive anaerobic cocci. *Clinical Microbiology Reviews*, *11*, 81–120.

Murdoch, D. A., Mitchelmore, I. J., & Tabaqchali, S. (1994). The clinical importance of gram-positive anaerobic cocci isolated at St Bartholomew's Hospital, London, in 1987. *Journal of Medical Microbiology*,*41*,36–44.https://doi.org/10.1099/00222615-41-1-36

Murray, P. R., Baron, E. J., Jorgensen, J. H., Landry, M. L., & Pfaller, M. A. (2007). *Manual of clinical microbiology*, 9th ed. Washington, D.C: ASM Press.

Onderdonk, A. B., Delaney, M. L., & Fichorova, R. N. (2016). The human microbiome during bacterial vaginosis. *Clinical Microbiology Reviews*, *29*, 223–238. https://doi.org/10.1128/CMR.00075-15

Pandya, S., Ravi, K., Srinivas, V., Jadhav, S., Khan, A., Arun, A., … Madhivanan, P. (2017). Comparison of culture-dependent and culture-independent molecular methods for characterization of vaginal microflora. *Journal of Medical Microbiology*, *66*, 149–153.

Patel, N. B., Tito, R. Y., Obregón-Tito, A. J., O'Neal, L., Trujillo-Villaroel, O., Marin-Reyes, L., … Lewis, C. M. Jr (2015). *Ezakiella peruensis* gen. nov., sp. nov. isolated from human fecal sample from a coastal traditional community in Peru. *Anaerobe*, *32*, 43–48. https://doi.org/10.1016/j.anaerobe.2014.12.002

Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*, *106*, 19126–19131. https://doi.org/10.1073/pnas.0906412106

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). Bypassing cultivation to identify bacterial species. *Microbe*, *9*, 111–118.

Rooney, A. P., Swezey, J. L., Pukall, R., Schumann, P., & Spring, S. (2011). *Peptoniphilus methioninivorax* sp. nov., a Gram-positive anaerobic coccus isolated from retail ground beef. *International Journal of Systematic and Evolutionary Microbiology*, *61*, 1962–1967. https://doi.org/10.1099/ijs.0.024232-0

Sasser, M. (2006). *Bacterial identification by gas chromatographic analysis of fatty acids methyl esters (GC-FAME)*. New york, NY: MIDI, Technical Note.

Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P. E., Rolain, J. M., & Raoult, D. (2009). Ongoing revolution in bacteriology: Routine identification of bacteria by matrix-assisted laser desorption ionization time of flight mass spectrometry. *Clinical Infectious Diseases*, *49*, 543–551. https://doi.org/10.1086/600885

Shipitsyna, E., Roos, A., Datcu, R., Hallén, A., Fredlund, H., Jensen, J. S., … Unemo, M. (2013). Composition of the vaginal microbiota in women of reproductive age–sensitive and specific molecular diagnosis of bacterial vaginosis is possible? *PLoS ONE*, *8*(4), e60670. https://doi.org/10.1371/journal.pone.0060670

Srinivasan, S., & Fredricks, D. N. (2008). The human vaginal bacterial biota and bacterial vaginosis. *Interdisciplinary Perspectives on Infectious Diseases*, *2008*, 1–22. https://doi.org/10.1155/2008/750479

Srinivasan, S., Munch, M. M., Sizova, M. V., Fiedler, T. L., Kohler, C. M., Hoffman, N. G., … Fredricks, D. N. (2016). More easily cultivated than identified: Classical isolation with molecular identification of vaginal bacteria. *Journal of Infectious Diseases*, *214*(Suppl 1), S21–S28. https://doi.org/10.1093/infdis/jiw192

Stackebrandt, E., & Ebers, J. (2006). Taxonomic parameters revisited: Tarnished gold standards. *Microbiology Today*, *33*, 152.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*, 2725–2729. https://doi.org/10.1093/molbev/mst197

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*, 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Ulger-Toprak, N., Lawson, P. A., Summanen, P., O'Neal, L., & Finegold, S. M. (2012). *Peptoniphilus duerdenii* sp. nov. and *Peptoniphilus koenoeneniae* sp. nov., isolated from human clinical specimens. *International Journal of Systematic and Evolutionary Microbiology*, *62*, 2336–2341. https://doi.org/10.1099/ijs.0.031997-0

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., … Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*, 635–645. https://doi.org/10.1038/nrmicro3330

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*, 821–829. https://doi.org/10.1101/gr.074492.107

How to cite this article: Diop K, Diop A, Michelle C, et al. Description of three new *Peptoniphilus* species cultured in the vaginal fluid of a woman diagnosed with bacterial vaginosis: *Peptoniphilus pacaensis* sp. nov., *Peptoniphilus raoultii* sp. nov., and *Peptoniphilus vaginalis* sp. nov. *MicrobiologyOpen*. 2018;e661. https://doi.org/10.1002/mbo3.661

Article 14:

Khoudiadiopia massiliensis' gen. nov., sp. nov., strain Marseille-P2746^{T}, a new bacterial genus isolated from the female genital tract

Diop A, Raoult D, Fenollar F, Fournier PE

[Published in New Microbes New Infections]

'*Khoudiadiopia massiliensis*' gen. nov., sp. nov., strain Marseille-P2746T^T, a new bacterial genus isolated from the female genital tract

A. Diop[1], **D. Raoult**[1,2], **F. Fenollar**[1,2] **and P.-E. Fournier**[1]

1) *Aix-Marseille Université, Institut hospitalo-universitaire Méditerranée-infection, URMITE, UM63, CNRS 7278, IRD 198, Inserm U1095, Marseille, France and* 2) *Campus International UCAD-IRD, Dakar, Senegal*

Abstract

We report the main characteristics of '*Khoudiadiopia massiliensis*' gen. nov., sp. nov., strain Marseille-P2746^T (= CSUR P2746), a new member of the *Peptoniphilaceae* family isolated from a vaginal swab of a patient suffering from bacterial vaginosis.

Keywords: Culturomics, human microbiome, *Khoudiadiopia massiliensis*, taxono-genomics, vaginal microbiota
Original Submission: 14 April 2017; **Revised Submission:** 23 May 2017; **Accepted:** 2 June 2017
Article published online: 8 June 2017

Corresponding author: P.-E. Fournier, Aix-Marseille Université, Institut hospitalo-universitaire Méditerranée-infection, URMITE, UM63, CNRS 7278, IRD 198, Inserm U1095, 19-21 Boulevard Jean Moulin, 13005 Marseille, France
E-mail: pierre-edouard.fournier@univ-amu.fr

The study of the vaginal microbiota diversity from patients with bacterial vaginosis is part of the ongoing microbial culturomics revolution in our laboratory [1]. A new member from the new family *Peptoniphilaceae* was isolated during this study that could not be identified by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry screening on a Microflex spectrometer (Bruker Daltonics, Leipzig, Germany), from a vaginal sample of a 26-year-old French woman suffering from bacterial vaginosis in the hospital Nord in Marseille (France). The patient gave her informed and signed consent and the study was authorized by the local ethics committee of the IFR48 (Marseille, France) under agreement 09-022. Strain Marseille-P2746T was first cultivated in April 2016 after 48 h of incubation in an anaerobic atmosphere at 37°C on Schaedler agar and Trypticase soy agar (BD Diagnostics, Le Pont de Claix, France), after 4 days of pre-incubation in a blood culture bottle enriched with rumen and sheep blood. Colonies were bright grey. Bacterial cells were Gram-positive, non-motile and non-spore-forming with a mean diameter of 0.55 µm. Strain Marseille-P2746^T is a strictly anaerobic coccus and exhibits oxidase activity but no catalase activity. Using the universal primer pair FD1 and rp2 as previously described [2], and a 3130-XL sequencer (Applied Biosciences, Saint Aubin, France), the 16S rRNA gene was sequenced. Strain Marseille-P2746^T exhibited an 89.28% 16S rRNA gene sequence identity with *Murdochiella asaccharolytica* strain WAL 1855C^T (GenBank Accession number EU483153), the phylogenetically closest species with a validly published name (Fig. 1). This value was lower than the 95% 16S rRNA gene sequence threshold proposed by Stackebrandt and Ebers [3] to define a new genus without carrying out DNA–DNA hybridization and classifies it as a new genus within the *Peptoniphilaceae* family (phylum *Firmicutes*), first created in 2014 [4]. *Murdochiella asaccharolytica* is an obligate anaerobic species isolated from a sacro-pilonidal cyst aspirate from an immunocompetent patient. It is also Gram-stain-positive, non-motile, non-spore-forming, and also shows a negative catalase activity [5].

Strain Marseille-P2746^T has >10% 16S rRNA gene sequence divergence with its closest phylogenetic neighbour [6], so we propose the creation of a new genus named '*Khoudiadiopia*' gen. nov. (khou.dia.dio'pia, N.L. fem. n. khoudiadiopia from the contraction of the first and last names of the Senegalese microbiologist Khoudia Diop). Strain Marseille-P2746^T is the type strain of '*Khoudiadiopia massiliensis*' gen. nov., sp. nov., the type species of the new genus '*Khoudiadiopia*' gen. nov.

http://dx.doi.org/10.1016/j.nmni.2017.06.002

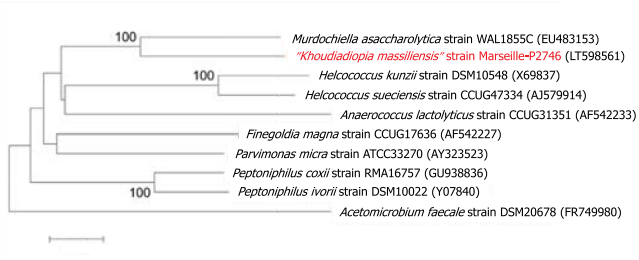


FIG. 1. Phylogenetic tree highlighting the phylogenetic position of '*Khoudiadiopia massiliensis*' gen. nov. strain Marseille-P2746T^{T} relative to other close species. GenBank accession numbers are indicated in parentheses. Sequences were aligned using CLUSTALW, and the tree was constructed with the Neighbour-joining method and 500 bootstrap replicates using the MEGA6 software. Numbers at the nodes are percentages of bootstrap values > 95%. The scale bar indicates a 2% nucleotide sequence divergence.

Nucleotide sequence accession number

The 16S rRNA gene sequence was deposited in EMBL-EBI under Accession number LT598561.

Deposit in a culture collection

'*Khoudiadiopia massiliensis*' gen. nov., sp. nov. was deposited in the 'Collection de Souches de l'Unité des Rickettsies' (CSUR, WDCM 875) under number CSUR P2746.

Acknowledgement

This research is funded by the Méditerranée-Infection Foundation.

Transparency declaration

No conflicts of interest declared.

References

[1] Lagier JC, Hugon P, Khelaifia S, Fournier PE, La Scola B, Raoult D. The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. Clin Microbiol Rev 2015;28:237–64.
[2] Drancourt M, Bollet C, Carlioz A, Martelin R, Gayral JP, Raoult D. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. J Clin Microbiol 2000;38:3623–30.
[3] Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. Microbiol Today 2006;33:152–5.
[4] Johnson CN, Whitehead TR, Cotta MA, Rhoades RE, Lawson PA. *Peptoniphilus stercorisuis* sp. nov., isolated from a swine manure storage tank and description of *Peptoniphilaceae* fam. nov. Int J Syst Evol Microbiol 2014;64:3538–45.
[5] Ulger-Toprak N, Liu C, Summanen PH, Finegold SM. Murdochiella asaccharolytica gen. nov., sp. nov., a Gram-stain-positive, anaerobic coccus isolated from human wound specimens. Int J Syst Evol Microbiol 2010;60:1013–6.
[6] Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, et al. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 2008;31:241–50.

Taxono-génomique des nouvelles espèces bactériennes du tube digestif de patients obèses

Article 15:

Butyricimonas phoceensis sp. nov., a new anaerobic species isolated from the human gut microbiota of a French morbidly obese patient

Togo AH, Diop A, Dubourg G, Nguyen TT, Andrieu C, Caputo A, Couderc C, Fournier PE, Maraninchi M, Valero R, Raoult D, Million M

[Published in New Microbes New Infections]

Butyricimonas phoceensis sp. nov., a new anaerobic species isolated from the human gut microbiota of a French morbidly obese patient

A. H. Togo[1], A. Diop[1], G. Dubourg[1], T. T. Nguyen[1], C. Andrieu[1], A. Caputo[1], C. Couderc[1], P.-E. Fournier[1], M. Maraninchi[2,3], R. Valero[2,3], D. Raoult[1,4] and M. Million[1]

1) *Aix Marseille Université, URMITE, Institut Hospitalier Universitaire Méditerranée-Infection, UM63, CNRS7278, IRD198, INSERM1095,* 2) *Aix Marseille Université, NORT "Nutrition, Obesity and Risk of Thrombosis", INSERM1062, INRA1260,* 3) *APHM, CHU Hôpital de la Conception, Service Nutrition, Maladies Métaboliques et Endocrinologie, F-13385 Marseille, France and* 4) *Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia*

Abstract

Butyricimonas phoceensis strain AT9 (= CSUR 2478 = DSM 100838) was isolated from a stool sample from a morbidly obese French patient living in Marseille using the culturomics approach. The genome of this Gram-negative-staining, anaerobic and non–spore forming rod bacillus is 4 736 949 bp long and contains 3947 protein-coding genes. Genomic analysis identified 173 genes as ORFans (4.5%) and 1650 orthologous proteins (42%) not shared with the closest phylogenetic species, *Butyricimonas virosa*. Its major fatty acid was the branched acid iso-C15:0 (62.3%).

Keywords: Butyrate, *Butyricimonas phoceensis*sp. nov., culturomics, genome, obesity, taxonogenomics
Original Submission: 18 May 2016; **Revised Submission:** 11 July 2016; **Accepted:** 25 July 2016
Article published online: 9 August 2016

Corresponding author: M. Million, Aix Marseille Université, URMITE, UM63, CNRS 7278, IRD 198, INSERM 1095, Marseille, France
E-mail: matthieumillion@gmail.com

Introduction

Butyricimonas phoceensis strain AT9 (= CSUR P2478 = DSM 100838 was isolated from the faeces of a 57-year-old French woman living in Marseille with class III morbid obesity (body mass index (BMI) 55.8 kg/m^2). This isolate is part of an exploratory study of the gut flora from obese patients before and after bariatric surgery. Bariatric surgery is the most effective treatment for morbid obesity for sustainable weight loss and leads to an enrichment of the gut flora [1]. The goal of our study was to compare microbial diversity of the gut flora in obese patients before and after bariatric surgery by culturomics. The aim of culturomics is to exhaustively explore the microbial ecosystem of gut flora by using different culture conditions followed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) identification [2].

The conventional approaches used in the delineation of bacterial species are 16S rRNA sequence identity and phylogeny [3], genomic (G+C content) diversity and DNA-DNA hybridization (DDH) [4,5]. However, these approaches present some difficulties, mainly as a result of their cutoff values, which change according to species or genera [6]. The accession of new technology tools, such as high-throughput sequencing, has made available nucleotide sequence libraries for many bacterial species [7]. We recently suggested incorporating genomic data in a polyphasic taxonogenomics approach to describe new bacteria. This procedure considers phenotypic characteristics, genomic analysis and the MALDI-TOF MS spectrum comparison [8,9].

Here we propose a classification and a set of characteristics for *Butyricimonas phoceensis* strain AT9, together with the description of complete genome sequencing, annotation and comparison as new species belonging to the genus *Butyricimonas*. The genus *Butyricimonas* was established in 2009 by

http://dx.doi.org/10.1016/j.nmni.2016.07.010

Sakamoto and encompasses four described species (*B. faecihominis, B. synergistica, B. paravirosa* and *B. virosa*). They were isolated from rat or human faeces and belong to the family *Porphyromonadaceae* [10,11]. The family *Porphyromonadaceae* contains 11 genera: *Porphyromonas* (type genus), *Barnesiella, Butyricimonas, Dysgonomonas, Macellibacteroides, Odoribacter, Paludibacter, Parabacteroides, Petrimonas, Proteiniphilum* and *Tannerella* [12]. *Butyricimonas virosa* bacteraemia has been described in patients with colon cancer [13,14] and in patients with posttraumatic chronic bone and joint infection [14].

Materials and Methods

Sample collection

A stool sample was collected from a 57-year-old obese French woman (BMI 55.8 kg/m^2; 150 kg, 1.64 m tall) in June 2012. Written informed consent was obtained from the patient at the nutrition, metabolic disease and endocrinology service at La Timone Hospital (Marseille, France). The study and assent procedure were approved by the local ethics committee (IFR 48, no. 09-022, 2010). The stool sample was stored at −80°C after collection.

Isolation and identification of strain

Strain isolation was performed in May 2015. Stool extract was preincubated in blood culture bottles enriched with lamb rumen juice and sheep's blood in anaerobic atmosphere as described elsewhere [2]. The culture was followed closely for 30 days. At different time points (days 1, 3, 7, 10, 15, 21 and 30), a seeding of the preincubated product was performed on sheep's blood–enriched Columbia agar (bioMérieux, Marcy l'Etoile, France) during 48 hours of incubation in an anaerobic atmosphere at 37°C. Colonies that emerged were cultivated in the same isolated conditions.

The colonies were then identified by MALDI-TOF MS as previously described [15]. Briefly, one isolated bacterial colony was picked up with a pipette tip from a culture agar plate and spread as a thin smear on a MTP 384 MALDI-TOF MS target plate (Bruker Daltonics, Leipzig, Germany). Measurement and identification were performed as previously described [16]. When a bacterium was unidentifiable, 16S rRNA gene amplification and sequencing were performed.

The 16S rRNA PCR coupled with sequencing were performed using GeneAmp PCR System 2720 thermal cyclers (Applied Biosystems, Bedford, MA, USA) and ABI Prism 3130xl Genetic Analyzer capillary sequencer (Applied Biosystems) respectively [17]. Chromas Pro 1.34 software (Technelysium, Tewantin, Australia) was used to correct sequences, and BLASTn searches were performed at the National Center for Biotechnology Information (NCBI) website (http://blast.ncbi.nlm.nih.gov.gate1.inist.fr/Blast.cgi).

Phylogenetic analysis

A custom Python script was used to automatically retrieve all species from the same family of the new species and download 16S sequences from NCBI by parsing NCBI results and NCBI taxonomy page. The scripts also remove species that are not found on the List of Prokaryotic Names With Standing in Nomenclature (LPSN) website (http://www.bacterio.net/). The script retains the most appropriate 16S sequence (the longest sequence with the smallest number of degenerate nucleotides) whilst also retaining one sequence from another genus as an outside group. It then aligns and trims the extremities of the sequences. Sequences were aligned using Muscle v3.8.31 with default parameters, and phylogenetic inferences were obtained using neighbour-joining method with 500 bootstrap replicates within MEGA6 software.

Phenotypic and biochemical characterization

Growth conditions. Different growth temperatures (28, 37, 45 and 55 °C) were tested on sheep's blood–enriched Columbia agar (bioMérieux). Growth of this strain was tested under anaerobic conditions using the GENbag anaer system (bioMérieux), microaerophilic conditions using the GENbag microaer system (bioMérieux) and under aerobic conditions with or without 5% CO_2. The tolerance to salt of this strain over a range salt concentrations (0–100 g/L) on Schaedler agar with 5% sheep's blood (bioMérieux) under anaerobic atmosphere was performed.

Microscopy. A heat shock at 80°C for 20 minutes was performed for the sporulation test. A fresh colony was observed between blades and slats using a photonic microscope Leica DM 1000 (Leica Microsystems, Nanterre, France) at 40× to assess the motility of the bacteria. Gram staining was performed and observed using a photonic microscope Leica DM 2500 with a 100× oil-immersion objective lens. Transmission electron microscopy using a Tecnai G20 device (FEI Company, Limeil-Brevannes, France) at an operating voltage of 60 kV was performed to observe strain AT9 after negative colouration.

Biochemical assays. Biochemical assays were performed using API Gallery systems (API ZYM, API 20A and API 50CH) according to the manufacturer's instructions (bioMérieux). Detection of catalase (bioMérieux) and oxidase (Becton Dickinson, Le Pont de Claix, France) was also performed according to the manufacturer's instructions.

Antibiotic susceptibility. The antibiotic susceptibility of the strain was tested using a disk diffusion method [18] for 21 antibiotics

including the following: amoxicillin 25 µg/mL, amoxicillin–clavulanic acid 30 µg/mL, ceftriaxone 30 µg, ciprofloxacin 5 µg, clindamycin (DA15), colistin (CT50), Dalacin 15 µg/mL, doripenem 10 µg/mL, doxycycline 30 IU, erythromycin 15 IU, fosfomycin 10 µg, gentamicin 500 µg, gentamicin 15 µg, imipenem 10 µg/mL, metronidazole 4 µg/mL, oxacillin 5 µg, penicillin G 10 IU, rifampicin 30 µg, sulfamethoxazole 23.75 µg, trimethoprim 1.25 µg, teicoplanin (TEC30) and vancomycin 30 µg (i2a, Montpellier, France). The 1200 scan was used for the interpretation of results (Interscience, Saint-Nom-La-Bretèche, France).

Fatty acid analysis. Fresh colonies from a plate of Columbia agar with 5% sheep's blood were collected after 48 hours' incubation at 37°C for fatty acid analysis. Cellular fatty acid analysis was performed by gas chromatography/mass spectrometry (GC/MS). Two samples were prepared with approximately 100 mg of bacterial biomass each collected from a culture plate. Cellular fatty acid methyl esters were prepared as described by Sasser [19]. GC/MS analyses were carried out as previously described [20]. Briefly, fatty acid methyl esters were separated using an Elite 5-MS column and monitored by a Clarus 500 gas chromatograph equipped with a SQ8S MS detector (PerkinElmer, Courtaboeuf, France). Fatty acid methyl esters were identified by using the spectral database search using MS Search 2.0 operated with the Standard Reference Database 1A (National Institute of Standards and Technology, Gaithersburg, MD, USA) and the FAMEs mass spectral database (Wiley, Chichester, UK).

Genome sequencing and assembly

Genomic DNA (gDNA) of strain AT9 was sequenced using MiSeq Technology (Illumina, San Diego, CA, USA) with the mate-pair strategy. The gDNA was barcoded so it could be mixed with 11 other projects with the Nextera Mate Pair sample prep kit (Illumina). The gDNA was quantified by a Qubit assay with the high-sensitivity kit (Thermo Fisher Scientific Life Sciences, Waltham, MA, USA) to 325 ng/µL. The mate-pair library was prepared with 1.5 µg of genomic DNA using the Nextera mate pair Illumina guide. The genomic DNA sample was simultaneously fragmented and tagged with a mate-pair junction adapter. The pattern of the fragmentation was validated on an Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA) with a DNA 7500 lab chip. The DNA fragments ranged in size from 1.5 to 11 kb with an optimal size at 4.8 kb. No size selection was performed, and 600 ng of tagmented fragments were circularized. The circularized DNA was mechanically sheared to small fragments with an optimal at 966 bp on the Covaris S2 device in T6 tubes (Covaris, Woburn, MA, USA). The library profile was visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies), and the final concentration library was measured at 24.3 nmol/L. The libraries were normalized at 2 nM and pooled. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated cluster generation and sequencing run were performed in a single 39-hour run at a 2 × 151 bp read length. Total information of 8.9 Gb was obtained from a 1009K/mm^2 cluster density, with a cluster passing quality control filters of 91.5% (17 486 000 passing filter-paired reads). Within this run, the index representation for strain AT9 was determined to be 8.38%. The 1 465 998 paired reads were trimmed then assembled in six scaffolds using Spades software [21].

Genome annotation and comparison

Open reading frames (ORFs) were predicted using Prodigal [22] with default parameters. Nevertheless, the predicted ORFs were excluded if they spanned a sequencing gap region (contains N). The predicted bacterial protein sequences were searched against the GenBank and Clusters of Orthologous Groups (COGs) databases using BLASTP (*E* value 1e-03 coverage). If no hit was found, it searched against the nr (nonredundant) database using BLASTP with an *E* value of 1e-03, coverage 70% and identity 30%. If the sequence length was smaller than 80 amino acids, we used an *E* value of 1e-05. The tRNAs and rRNAs were predicted using the tRNA Scan-SE and RNAmmer tools respectively [23,24]. Phobius was used to foresee the signal peptides and number of transmembrane helices respectively [25]. Mobile genetic elements were foretold using PHAST and RAST [26,27]. ORFans were identified if none of the BLASTP runs provided positive results (*E* value was lower than 1e-03 for an alignment length greater than 80 amino acids. If alignment lengths were smaller than 80 amino acids, we used an *E* value of 1e-05). Artemis and DNA Plotter were used for data management and visualization of genomic features respectively [28,29]. Genomes were automatically retrieved from the 16S rRNA tree using Xegen software (PhyloPattern) [30]. For each selected genome, complete genome sequence, proteome genome sequence and Orfeome genome sequence were retrieved from the NCBI FTP site. All proteomes were analysed with proteinOrtho [31]. Then for each couple of genomes, a similarity score was computed. This score is the mean value of nucleotide similarity between all couple of orthologous genes between the two genomes studied (average genomic identity of orthologous gene sequences (AGIOS)) [7]. For the genomic comparison of strain AT9, we used *Butyricimonas virosa* (type) strain JCM15149T (Genbank project number: JAEW00000000), *Odoribacter laneus* strain YIT12061 (ADMC00000000), *Bacteroides plebeius* strain DSM17135

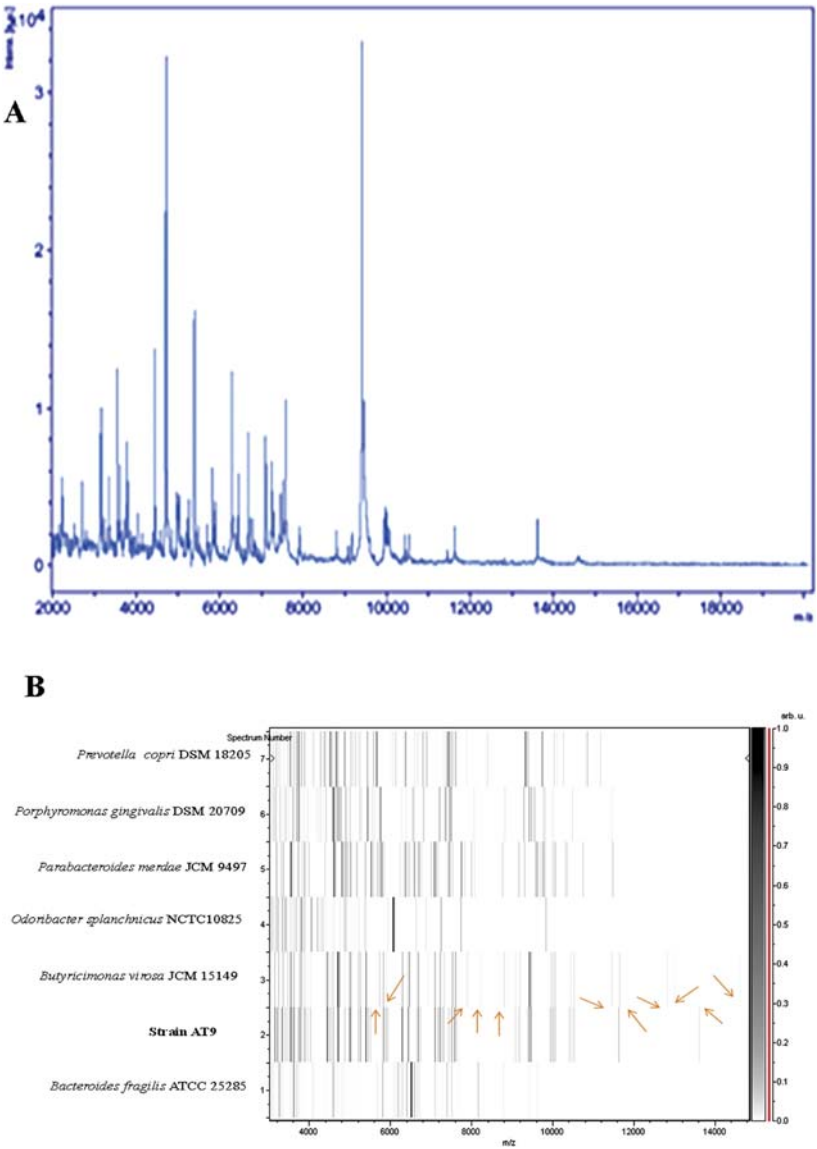


FIG. 1. MALDI-TOF MS analysis of *Butyricimonas phoceensis* strain AT9. (a) Reference mass spectrum from strain AT9. (b) Gel view comparing strain AT9 to other close species. Gel view displays raw spectra of loaded spectrum files arranged in pseudo–gel-like look. The *x*-axis records m/z value. Left *y*-axis displays running spectrum number originating from subsequent spectra loading. Peak intensity is expressed by greyscale scheme code. Colour bar and right *y*-axis indicate relation between colour peak is displayed with and peak intensity in arbitrary units. Displayed species are indicated at left. Arrows indicated discordant peaks between strain AT9 and its closest phylogenetic neighbour, *Butyricimonas virosa*.

(ABQC00000000), *Paraprevotella clara* strain YIT11840 (AFFY00000000), *Parabacteroides merdae* ATCC43184 (AAXE00000000), *Porphyromonas catoniae* ATCC 51270 (JDFF00000000) and *Odoribacter splanchnicus* strain DSM20712 (CP002544). An annotation of the entire proteome was performed to define the distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins (using the same method as for the genome annotation). The genome of strain AT9 was locally aligned pairwise using the BLAT algorithm [28,29] against each of the selected genomes previously cited, and DDH values were estimated from a generalized model [32]. Annotation and comparison processes were performed in the multiagent software system DAGOBAH [33], which includes Figenix [34] libraries that provide pipeline analysis.

Results

Phylogenetic analysis

The spectrum generated from clean strain AT9 spots did not match with those identified from the Bruker database even when two strains of *Butyricimonas virosa*, including the type strain (JCM15149T), were included in the database (Fig. 1a). The phylogenetic analysis, performed using 16S rRNA gene sequences, showed that our strain AT9 exhibited 98.3, 97.8, 97.5 and 94.2% similarity with *Butyricimonas virosa* JCM 15149T, *Butyricimonas faecihominis* JCM 18676T, *Butyricimonas paravirosa* JCM 18677T and *Butyricimonas synergistica* JCM 15148T respectively [9,10] (Table 1). However, this percentage remains lower than the 98.7% 16S rRNA gene sequence threshold recommended by Kim *et al.* [35] to delineate a new species. The neighbour-joining phylogenetic tree (Fig. 2), based on 16S rRNA gene sequences, shows the relationships between strain AT9 and some related taxa. The 16S rRNA sequence of strain AT9 was deposited in European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) under accession number LN881597. A gel view was performed in order to see the spectra differences of strain AT9 with other related bacteria. Eleven discordant peaks were found when we compared strain AT9 and the *B. virosa* JCM15149T profile (Fig. 1b).

Phenotypic and biochemical characterization

The growth of strain AT9 occurred between 28 to 37°C, but optimal growth was observed at 37°C after 48 hours' incubation in anaerobic atmosphere. It is an anaerobic bacillus, but it

TABLE 1. Percentage 16S rRNA gene similarity within *Butyricimonas* genus

	B. faecihominis JCM 18676T	*B. paravirosa* JCM 18677T	*B. synergistica* JCM 15148T	*B. virosa* JCM 15149T	*B. phoceensis* strain AT9
B. faecihominis JCM 18676T	100	97.30	94.07	96.84	97.77
B. paravirosa JCM 18677T		100	94.75	96.84	97.51
B. synergistica JCM 15148T			100	94.22	94.20
B. virosa JCM 15149T				100	98.38
B. phoceensis strain AT9					100

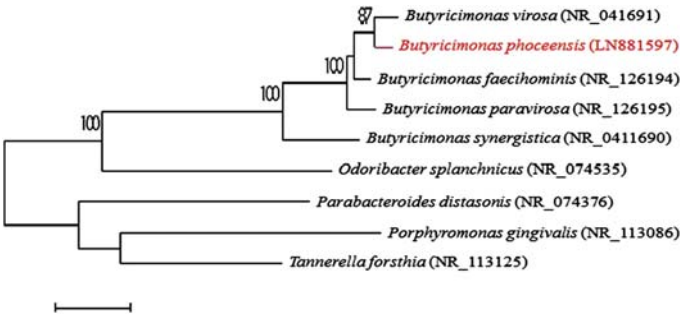


FIG. 2. Phylogenetic tree based on 16S rRNA highlighting position of *Butyricimonas phoceensis* strain AT9 relative to other close species. Corresponding GenBank accession numbers for 16S rRNA genes are indicated at right of strains in tree. Sequences were aligned using Muscle v3.8.31 with default parameters, and phylogenetic inferences were obtained using neighbour-joining method with 500 bootstrap replicates within MEGA6 software. Scale bar represents 2% nucleotide sequence divergence.

can also grow in microaerophilic atmospheres at 28°C. The colonies were ~1 to 2 mm in diameter and opalescent on 5% sheep's blood–enriched Columbia agar. Growth of this isolate was observed using 5 g of salt on Schaedler agar with 5% sheep's blood but not with 10 g/L of salt. This bacterium is not able to form spores. It is a Gram-negative stain (Fig. 3a); it is a motile rod-shaped bacterium that is catalase positive and oxidase negative. Cell diameter ranges 0.5 to 1.5 µm, with a mean diameter of 1 µm by electron microscopy (Fig. 3b). Table 2 summarizes the classification and main features of strain AT9.

Using the API ZYM strip, we observed that strain AT9 possesses alkaline phosphatase, esterase (C4), esterase lipase (C8), naphthol-AS-BI-phosphohydrolase, phosphatase acid and *N*-acetyl-β-glucosaminidase activities; there were no activities for the other enzymes tested. Using API 20A strip, positive reactions were obtained for indole, D-glucose, D-lactose, glycerol and D-mannose. Using the API 50 CH strip, positive reactions were observed only with esculin ferric citrate and potassium 2-ketogluconate. The differences of characteristics compared to other representatives of the genus *Butyricimonas* are detailed in Table 3.

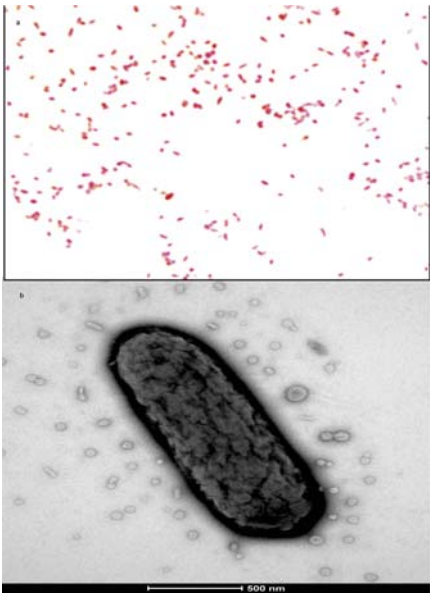


FIG. 3. Phenotypic features of *Butyricimonas phoceensis* strain AT9. (a) Gram stain. (b) Transmission electron microscopy using Tecnai G20 (FEI Company) at operating voltage of 60 kV. Scale bar = 500 nm.

TABLE 2. **Classification and general features of *Butyricimonas phoceensis* strain AT9**

Property	Term
Current classification	Domain: *Bacteria* Phylum: *Bacteroidetes* Class: *Bacteroidia* Order: *Bacteroidales* Family: *Porphyromonadaceae* Genus: *Butyricimonas* Species: *B. phoceensis* Type strain: AT9
Gram stain	Negative
Cell shape	Rod
Motility	Motile
Sporulation	Non–spore forming
Temperature range	Mesophile
Optimum temperature	37°C
Oxygen requirement	Anaerobic
Carbon source	Unknown
Energy source	Unknown
Habitat	Human gut
Biotic relationship	Free living
Pathogenicity	Unknown
Isolation	Human faeces

Of the 21 antibiotics tested, strain AT9 was susceptible to gentamicin 500 µg, vancomycin, doxycycline, trimethoprim–sulfamethoxazole, rifampicin, amoxicillin 25 µg/mL, metronidazole 4 µg/mL, amoxicillin–clavulanic acid 30 µg/mL, imipenem 10 µg/mL, penicillin G, teicoplanin and doripenem 10 µg/mL and was resistant to erythromycin, oxacillin, gentamicin 15 µg, colistin, ceftriaxone, ciprofloxacin, clindamycin, dalacin 15 µg/mL and fosfomycin. Analysis of the total cellular fatty acid composition demonstrated that the major fatty acid detected was the branched iso-C15:0 acid (62.3%). Hydroxy and cyclo fatty acids were also detected (Table 4).

Genome properties

The draft genome of strain AT9 (Fig. 4) (accession no. FBYB00000000) is 4 736 949 bp long with 42.51% G+C content (Table 5). It is composed of six scaffolds comprising seven contigs. Of the 4007 predicted genes, 3947 were protein-coding genes and 60 were RNAs (four genes 5S rRNA, one 16S rRNA, one 23S rRNA and 54 tRNA). A total of 2386 genes (60.45%) were assigned as putative functions (by COGs or by NR BLAST), 178 genes (4.51%) were identified as ORFans and ten genes were associated with polyketide synthase or non-ribosomal peptide synthetase [36]. Using ARG-ANNOT [37], three genes associated with resistance were found, including *TetQ, TetX* (which confers resistance to tetracycline) and *ErmF* (which confers resistance to erythromycin). This could represent the *in silico/in vitro* discordance for antibiotic resistance prediction, as strain AT9 was resistant to erythromycin but susceptible to doxycycline. The remaining 1316 genes (33.34%) were annotated as hypothetical proteins. Genome statistics are provided in Table 5. Table 6 lists the distribution of genes into COGs functional categories of strain AT9.

TABLE 3. Differential characteristics of strain *Butyricimonas phoceensis* strain AT9 with *Butyricimonas* species

Property	Strain AT9	*B. virosa*	*B. faecihominis*	*B. paravirosa*	*B. synergistica*
Cell diameter width/length (μm)	0.5/1.75	0.6–0.8/2.5–5	0.7–1/3–5	0.8–1/2–12.4	0.5–1/3–6
Oxygen requirement	–	–	–	–	–
Gram stain	–	–	–	–	–
Motility	+	–	–	–	–
Spore formation	–	–	–	–	–
Production of:					
Catalase	+	+	+	+	–
Oxidase	–	–	–	–	–
Urease	–	–	–	–	–
Indole	+	+	+	+	+
β-Galactosidase	+	+	+	+	+
N-acetyl-glucosaminidase	+	+	+	+	+
Utilization of:					
L-Arabinose	–	–	+	–	–
D-Mannose	+	–	+	+	+
D-Mannitol	–	–	–	–	–
D-Glucose	+	+	+	+	+
D-Maltose	–	–	+	–	+
Isolation source	Human faeces	Rat faeces	Human faeces	Rat faeces	Human faeces
DNA G+C content (mol%)	42.5	46.5	45.2	44.9	46.4

TABLE 4. Cellular fatty acid profiles of strain *Butyricimonas phoceensis* strain AT9 compared to other closely related *Butyricimonas* species

Fatty acid	Strain AT9	*B. faecihominis* JCM 18676T	*B. paravirosa* JCM 18677T	*B. synergistica* JCM 15148T	*B. virosa* JCM 15149T
C4:0	TR	NA	NA	NA	NA
C12: 0	NA	TR	TR	NA	NA
C14:0	TR	TR	1.8	NA	1.3
C15: 0	TR	TR	NA	NA	NA
C16:0	3.7	2.8	3.2	2.4	2.1
C18:0	TR	TR	TR	1.0	TR
iso-C5: 0	2.9	NA	NA	NA	NA
iso-C11: 0	NA	TR	TR	NA	NA
iso-C13: 0	NA	1.0	1.0	NA	TR
iso-C15: 0	62.3	64.6	57.6	61.8	68.6
anteiso-C15: 0	1.2	1.8	1.7	2.0	1.5
iso-C17: 0	NA	1.0	TR	NA	TR
C14:0 3-OH	TR	NA	NA	NA	NA
C16: 0 3-OH	4.8	1.7	6.3	1.6	5.2
C17: 0 3-OH	9.0	NA	NA	NA	NA
iso-C15: 0 3-OH	NA	TR	1.8	1.6	1.7
iso-C17: 0 3-OH	NA	5.3	10.6	14.9	10.4
C18: 2n6	2.9	NA	NA	NA	NA
C18:1n5	2.1	NA	NA	NA	NA
C16:1n7	TR	NA	NA	NA	NA
iso-C17: 0	TR	1.0	TR	NA	TR
iso-C15:1n5	TR	NA	NA	NA	NA
C18: 1ω9c	NA	8.3	9.5	12.6	6.0
C18: 2ω6, 9c	NA	1.4	1.5	2.3	1.2
C9, 10-methylene C16:0	7.0	NA	NA	NA	NA

Number are percentages. NA, not available; TR, trace amounts <1%.

Genome comparison

The draft genome (4.74 Mb) sequence of strain AT9 is smaller than those of *Butyricimonas synergistica* (4.77 Mb), but larger than those of *Butyricimonas virosa, Porphyromonas catoniae, Bacteroides plebeius, Paraprevotella clara, Odoribacter laneus, Parabacteroides merdae* and *Odoribacter splanchnicus* (4.72, 2.04, 3.27, 3.65, 4.43, 3.77 and 4.39 MB respectively).

The G+C content of strain AT9 (42.5%) is smaller than those of *Butyricimonas virosa, Odoribacter splanchnicus, Bacteroides plebeius, Parabacteroides merdae, Paraprevotella clara, Butyricimonas synergistica* and *Porphyromonas catoniae* (46.5, 43.4, 44.3, 44.8, 45.3, 48.1, 46.4 and 51.0% respectively) but larger than those of *Odoribacter laneus* (40.55). Fig. 5 shows that the distribution of genes into COGs categories was similar in all genomes compared. In addition, strain AT9 shared 2297, 1535, 742, 1720, 999, 1173, 2108 and 960 orthologous genes with *B. virosa*, *O. laneus*, *P. catoniae*, *O. splanchnicus*, *B. plebeius*, *P. merdae*, *B. synergistica* and *P. clara* respectively (Table 6). Accordingly, strain AT9 has 1650 (42%) of 3947 orthologous proteins not shared with its closest phylogenetic neighbour, *B. virosa*. The AGIOS values ranged from 53.3 to 76.2% among the compared closest species except strain AT9. When strain AT9 was compared to other close species, the AGIOS values ranged from 53.5% with *P. catoniae* to 97.7% with *B. virosa* (Table 7).

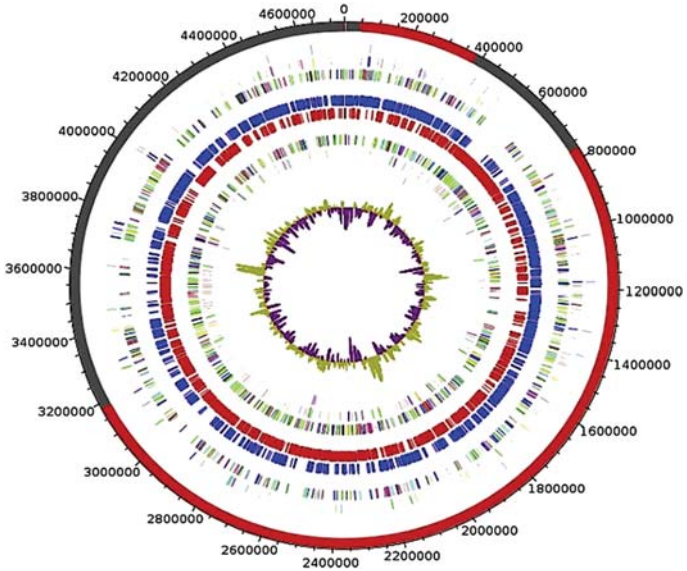


FIG. 4. Graphical circular map of genome of *Butyricimonas phoceensis* strain AT9. From outside to centre: contigs (red/grey), COGs category of genes on forward strand (three circles), genes on forward strand (blue circle), genes on reverse strand (red circle), COGs category on reverse strand (three circles), GC content.

The DDH value was 80.2% ± 2.7 with *B. virosa*, 17.7% ± 2.2 with *O. laneus*, 21.4% ± 2.3 with *B. plebeius*, 20.2% ± 2.3 with *P. clara*, 19.1% ± 2.2 with *P. merdae*, 18.3% ± 2.2 with *P. catoniae* and 17.3% ± 2.2 with *O. splanchnicus* (Table 8).

TABLE 5. Nucleotide content and gene count levels of the genome of *Butyricimonas phoceensis* strain AT9

	Genome (total)	
Attribute	**Value**	**% of total**
Size (bp)	4 736 949	100
G+C content (bp)	2 013 756	42.51
Coding region (bp)	4 330 163	91.40
Total genes	4007	100
RNA genes	60	1.50
Protein-coding genes	3947	98.50
Genes with function prediction	2386	60.45
Genes assigned to COGs	1880	47.63
Genes with peptide signals	1185	30.02
Gene associated to PKS or NRPS	10	0.25
Genes associated to ORFan	178	4.51
Genes associated to mobilome	1109	28.10
Genes associated to toxin/antitoxin	70	1.8
Genes associated to resistance genes	3	0.076
Genes with paralogues (*E* value 1e-10)	1449	36.71
Genes with paralogues (*E* value 1e-25)	1098	27.82
Gene associated to hypothetical protein	1316	33.34
Genes larger than 5000 nucleotides	5	0

COGs, Clusters of Orthologous Groups database; PKS, polyketide synthase; NRPS, nonribosomal peptide synthase.

Discussion

Strain AT9 is part of an exploratory culturomics study of the gut flora from obese patients before and after bariatric surgery. The aim of culturomics is to exhaustively explore the microbial ecosystem of gut flora by using different culture conditions followed by MALDI-TOF MS identification [2]. The phylogenetic analysis, performed using 16S rRNA sequences, showed that strain AT9 exhibited 98.3% similarity with *Butyricimonas virosa*. However, this percentage remains lower than the 98.7% 16S rRNA gene sequence threshold recommended to delineate a new species [3,38].

The genus *Butyricimonas* was established in 2009 by Sakamoto and includes four described species [9–11]. All the species of the genus *Butyricimonas* are anaerobic. These bacteria are isolated in human or rat faeces. To evaluate the genomic similarity with other closest species, we determined two parameters: DDH [39] and AGIOS [7]. Although the values of DDH (80.2%) and AGIOS (97.7%) were very high between strain AT9 and *Butyricimonas virosa* (type strain JCM15149T), we found several discrepancies justifying the description of a new species, including motility, D-mannose utilization (absent in

TABLE 6. Number of genes associated with the 25 general COGs functional categories of *Butyricimonas phoceensis* strain AT9

Code	Value	% value	Description
J	193	4.89	Translation
A	0	0	RNA processing and modification
K	192	4.87	Transcription
L	111	2.81	Replication, recombination and repair
B	0	0	Chromatin structure and dynamics
D	23	0.58	Cell cycle control, mitosis and meiosis
Y	0	0	Nuclear structure
V	85	2.15	Defence mechanisms
T	174	4.41	Signal transduction mechanisms
M	200	5.06	Cell wall/membrane biogenesis
N	20	0.51	Cell motility
Z	4	0.10	Cytoskeleton
W	3	0.07	Extracellular structures
U	28	0.71	Intracellular trafficking and secretion
O	91	2.30	Posttranslational modification, protein turnover, chaperones
X	32	0.81	Mobilome: prophages, transposons
C	122	3.09	Energy production and conversion
G	92	2.33	Carbohydrate transport and metabolism
E	120	3.04	Amino acid transport and metabolism
F	60	1.52	Nucleotide transport and metabolism
H	99	2.51	Coenzyme transport and metabolism
I	69	1.75	Lipid transport and metabolism
P	199	5.04	Inorganic ion transport and metabolism
Q	26	0.66	Secondary metabolites biosynthesis, transport and catabolism
R	150	3.80	General function prediction only
S	67	1.69	Function unknown
—	2067	52.36	Not in COGs

COGs, Clusters of Orthologous Groups database.

B. virosa but present in *B. paravirosa, B. synergistica* and *B. faecihominis*), MALDI-TOF MS spectrum (11 different peaks), different GC% (42.5 vs. 46.5% for *B. virosa*), high proportion of orthologous proteins not shared between the two species (1650/3947 (42%)) and different COGs repartition ((D) cell cycle control 110 vs. 124, (P) transport of inorganic ions 32 vs. 20 for strain AT9 and *B. virosa* respectively).

Conclusion

On the basis of phenotypic, chemotaxonomic, phylogenetic and genomic information, a novel species belonging to the genus *Butyricimonas* is proposed with the name *Butyricimonas phoceensis* sp. nov. The type strain is AT9. This bacterium was isolated from the faeces of a 57-year-old obese French woman living in Marseille after bariatric surgery. The isolation of this new species demonstrates that microbial culturomics extends the repertoire of human gut anaerobes, which are of critical importance to decipher the links among gut microbiota, health and disease, including obesity.

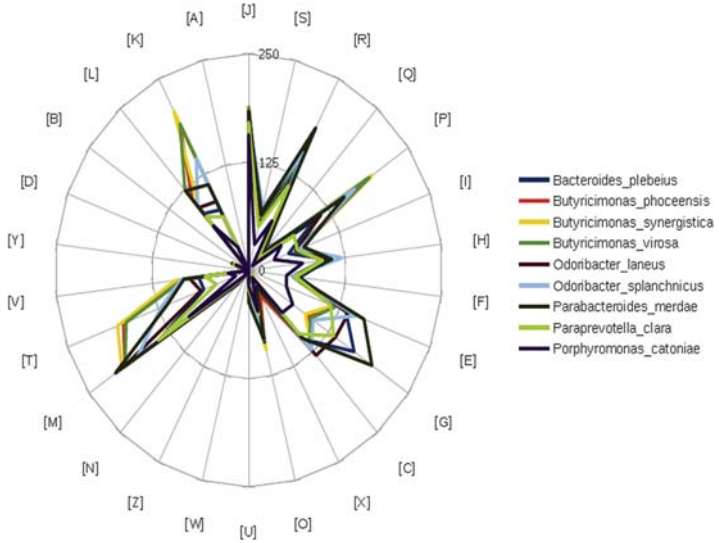


FIG. 5. Distribution of functional classes of predicted genes according to clusters of orthologous groups of proteins from *Butyricimonas phoceensis* strain AT9.

TABLE 7. Numbers of orthologous proteins shared between genomes (upper right), average percentage similarity of nucleotides corresponding to orthologous protein shared between genomes (lower left) and numbers of proteins per genome (bold)

	Odoribacter laneus	*Butyricimonas phoceensis* strain AT9	*Porphyromonas catoniae*	*Odoribacter splanchnicus*	*Bacteroides plebeius*	*Parabacteroides merdae*	*Butyricimonas virosa*	*Butyricimonas synergistica*	*Paraprevotella clara*
O. laneus	**3103**	1535	745	1472	1005	1187	1519	1480	964
B. phoceensis strain AT9	57.52	**3947**	742	1720	999	1173	2297	2108	960
P. catoniae	53.37	53.50	**1597**	737	726	826	729	725	746
O. splanchnicus	59.11	68.17	53.65	**3497**	977	1149	1702	1604	960
B. plebeius	55.52	62.17	55.11	62.84	**2643**	1175	986	963	1059
P. merdae	55.38	63.08	55.30	63.65	66.34	**4384**	1154	1130	1123
B. virosa	57.47	97.79	53.41	68.18	62.13	62.87	**3934**	2086	950
B. synergistica	57.22	76.18	53.75	68.24	62.15	62.93	76.24	**3874**	926
P. clara	54.31	61.92	54.84	62.34	68.02	65.40	61.87	62.09	**2847**

TABLE 8. Pairwise comparison of *Butyricimonas phoceensis* strain AT9 with other species using GGDC, formula 2 (DDH estimates based on identities/HSP length)[a]

	Strain AT9	*Odoribacter laneus*	*Bacteroides plebeius*	*Butyricimonas virosa*	*Paraprevotella clara*	*Parabacteroides merdae*	*Porphyromonas catoniae*	*Odoribacter splanchnicus*
Strain AT9	100% ± 0	17.7% ± 2.2	21.4% ± 2.3	80.2% ± 2.7	20.2% ± 2.3	19.1% ± 2.3	18.3% ± 2.3	17.3% ± 2.2
O. laneus		100% ± 0	19% ± 2.3	18.2% ± 2.3	20.5% ± 2.3	18.9% ± 2.3	19.6% ± 2.3	18.2% ± 2.3
B. plebeius			100% ± 0	19.9% ± 2.3	20.3% ± 2.3	21.5% ± 2.3	17.6% ± 2.2	18.4% ± 2.3
B. virosa				100% ± 0	20.3% ± 2.3	19.4% ± 2.3	19.0% ± 2.3	17.4% ± 2.2
P. clara					100% ± 0	18.9% ± 2.3	17.8% ± 2.2	17.7% ± 2;2
P. merdae						100% ± 0	17.6% ± 2.2	21.5% ± 2.3
P. catoniae							100% ± 0	18% ± 2.2
O. splanchnicus								100% ± 0

DDH, DNA-DNA hybridization; GGDC, Genome-to-Genome Distance Calculator; HSP, high-scoring segment pairs.
[a]Confidence intervals indicate inherent uncertainty in estimating DDH values from intergenomic distances based on models derived from empirical test data sets (which are always limited in size). These results are in accordance with 16S rRNA and phylogenomic analyses as well as GGDC results.

Taxonomic and nomenclatural proposals

Description of strain AT9 sp. nov.. Butyricimonas phoceensis (*pho.ce.en.sis*, N.L. gen. n. *phoceensis*, based on the acronym of the Phocean city where the type strain was isolated). Cells are Gram-negative-staining, non–spore forming, motile, rod-shaped bacilli, with a size of 0.5 to 1.5 µm in diameter. Colonies are opalescent with a diameter of 1 to 2 mm on 5% sheep's blood–enriched Columbia agar. The strain is oxidase negative and catalase positive. It has an optimum growth temperature of 37°C and is anaerobic, but it is able to grow in microaerophilic condition at 28°C. Using API Gallery systems, positive reactions were observed for alkaline phosphatase, esterase (C4), esterase lipase (C8), naphthol-AS-BI-phosphohydrolase, phosphatase acid, *N*-acetyl-β-glucosaminidase, indole, D-glucose, D-lactose, glycerol and D-mannose, esculin ferric citrate and potassium 2-ketogluconate. Cells are susceptible to gentamicin 500 µg, vancomycin, doxycycline, trimethoprim–sulfamethoxazole, rifampicin, penicillin G and teicoplanin. The major fatty acid detected was iso-C15:0. The length of the genome is 4 736 949 bp with 42.51% G+C content. The 16S rRNA gene sequence and whole-genome shotgun sequence of *B. phoceensis* strain AT9 were deposited in EMBL-EBI under accession numbers LN881597 and FBYB00000000, respectively. The type strain AT9 (= CSUR P2478 = DSM 100838) was isolated from the stool sample of a French obese woman. The habitat of this microorganism is the human digestive gut.

Acknowledgements

The authors thank the Xegen Company (http://www.xegen.fr/) for automating the genomic annotation process and K. Griffiths for English-language review. This study was funded by the Fondation Méditerranée Infection.

Conflict of Interest

None declared.

References

[1] Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, et al. Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci U S A 2009;106:2365–70.
[2] Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, et al. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 2012;18:1185–93.

[3] Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. Microbiol Today 2006;33:152–5.
[4] Garrity GM, Trüper HG, Whitman WB, Grimont PAD, Nesme X, Frederiksen W, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 2002;52:1043–7.
[5] Rosselló-Mora R. DNA-DNA reassociation methods applied to microbial taxonomy and their critical evaluation. In: Molecular identification, systematics, and population structure of prokaryotes. New York: Springer; 2006. p. 23–50.
[6] Welker M, Moore ERB. Applications of whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry in systematic microbiology. Syst Appl Microbiol 2011;34:2–11.
[7] Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 2014;64(Pt 2):384–91.
[8] Kokcha S, Mishra AK, Lagier JC, Million M, Leroy Q, Raoult D, et al. Non-contiguous-finished genome sequence and description of *Bacillus timonensis* sp. nov. Stand Genomic Sci 2012;6:346–55.
[9] Mishra AK, Lagier JC, Nguyen TT, Raoult D, Fournier PE. Non-contiguous-finished genome sequence and description of *Peptoniphilus senegalensis* sp. nov. Stand Genomic Sci 2013;7:370–81.
[10] Sakamoto M, Takagaki A, Matsumoto K, Kato Y, Goto K, Benno Y. *Butyricimonas synergistica* gen. nov., sp. nov. and *Butyricimonas virosa* sp. nov., butyric acid–producing bacteria in the family '*Porphyromonadaceae*' isolated from rat faeces. Int J Syst Evol Microbiol 2009;59(Pt 7):1748–53.
[11] Sakamoto M, Tanaka Y, Benno Y, Ohkuma M. *Butyricimonas faecihominis* sp. nov. and *Butyricimonas paravirosa* sp. nov., isolated from human faeces, and emended description of the genus *Butyricimonas*. Int J Syst Evol Microbiol 2014;64(Pt 9):2992–7.
[12] Sakamoto M. The Family *Porphyromonadaceae*. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, editors. The Prokaryotes—other major lineages of bacteria and the Archaea. Berlin: Springer; 2014. p. 811–24.
[13] Ulger Toprak N, Bozan T, Birkan Y, Isbir S, Soyletir G. *Butyricimonas virosa*: the first clinical case of bacteraemia. New Microbes New Infect 2015;4:7–8.
[14] Ferry T, Laurent F, Ragois P, Chidiac C, Lyon BJI Study Group. Post-traumatic chronic bone and joint infection caused by *Butyricimonas* spp., and treated with high doses of ertapenem administered subcutaneously in a 30-year-old obese man. BMJ Case Rep 2015;2015:212359.
[15] Seng P, Drancourt M, Gouriet F, La Scola B, Fournier PE, Rolain JM, et al. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 2009;49:543–51.
[16] Hugon P, Ramasamy D, Lagier JC, Rivet R, Couderc C, Raoult D, et al. Non-contiguous-finished genome sequence and description of *Alistipes obesi* sp. nov. Stand Genomic Sci 2013;7:427–39.
[17] Nkamga VD, Huynh HTT, Aboudharam G, Ruimy R, Drancourt M. Diversity of human-associated *Methanobrevibacter smithii* isolates revealed by multispacer sequence typing. Curr Microbiol 2015;70:810–5.
[18] Le Page S, van Belkum A, Fulchiron C, Huguet R, Raoult D, Rolain JM. Evaluation of the PREVI® Isola automated seeder system compared to reference manual inoculation for antibiotic susceptibility testing by the disk diffusion method. Eur J Clin Microbiol Infect Dis 2015;34:1859–69.
[19] Sasser M. Bacterial identification by gas chromatographic analysis of fatty acids methyl esters (GC-FAME). Technical note 101. Newark, DE: MIDI; 2006.
[20] Dione N, Sankar SA, Lagier JC, Khelaifia S, Michele C, Armstrong N, et al. Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. New Microbes New Infect 2016;10:66–76.
[21] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol 2012;19:455–77.
[22] Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119.
[23] Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997;25:955–64.
[24] Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35:3100–8.
[25] Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338:1027–36.
[26] Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res 2011;39(Web Server issue):W347–52.
[27] Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 2014;42(Database issue):D206–14.
[28] Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002;12:656–64.
[29] Auch AF, von Jan M, Klenk HP, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2010;2:117–34.
[30] Gouret P, Thompson JD, Pontarotti P. PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinformatics 2009;10:298.
[31] Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 2011;12:124.
[32] Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 2013;14:60.
[33] Gouret P, Paganini J, Dainat J, Louati D, Darbo E, et al. Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the multi-agent software system DAGOBAH. In: Springer-Verlag, editor. Evolutionary biology—concepts, biodiversity, macroevolution and genome evolution. Amsterdam: Springer-Verlag; 2011. p. 71–87.
[34] Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG. FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinformatics 2005;6:198.
[35] Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 2014;64(Pt 2):346–51.
[36] Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res 2013;41(Database issue):D402–7.
[37] Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014;58:212–20.
[38] Konstantinidis KT, Stackebrandt E. Defining taxonomic ranks. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E, editors. The Prokaryotes. New York: Springer; 2006.
[39] Auch AF, Klenk HP, Göker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci 2010;2:142–8.

Article 16:

Description of *Mediterraneibacter phoceensis*, gen. nov., sp. nov., a new species isolated from human stool sample from an obese patient before bariatric surgery and reclassification of *Ruminococcus faecis*, *Ruminococcus lactaris*, *Ruminococcus torques* and *Clostridium glycyrrhizinilyticum* as *Mediterraneibacter faecis* comb. nov., *Mediterraneibacter lactaris* comb. nov., *Mediterraneibacter torques* comb. nov. and *Mediterraneibacter glycyrrhizinilyticum* comb. nov.

Togo AH, Diop A, Bittar F, Maraninchi M, Valero R, Armstrong N, Dubourg G, Labas N, Richez M, Fournier PE, Raoult D, Million M

[Published in Antonie van Leeuwenhoek]

Antonie van Leeuwenhoek
https://doi.org/10.1007/s10482-018-1104-y

ORIGINAL PAPER

Description of *Mediterraneibacter massiliensis*, gen. nov., sp. nov., a new genus isolated from the gut microbiota of an obese patient and reclassification of *Ruminococcus faecis*, *Ruminococcus lactaris*, *Ruminococcus torques*, *Ruminococcus gnavus* and *Clostridium glycyrrhizinilyticum* as *Mediterraneibacter faecis* comb. nov., *Mediterraneibacter lactaris* comb. nov., *Mediterraneibacter torques* comb. nov., *Mediterraneibacter gnavus comb.* nov. and *Mediterraneibacter glycyrrhizinilyticus* comb. nov.

Amadou Hamidou Togo · Awa Diop · Fadi Bittar · Marie Maraninchi · René Valero · Nicholas Armstrong · Grégory Dubourg · Noémie Labas · Magali Richez · Jeremy Delerce · Anthony Levasseur · Pierre-Edouard Fournier · Didier Raoult · Matthieu Million

Received: 4 January 2018 / Accepted: 20 May 2018

Abstract An anaerobic isolate, strain AT7^T, was cultivated from a stool sample of a morbidly obese French woman using a microbial culturomics approach. The 16S rRNA gene sequence analysis showed that strain AT7^T exhibited 96% nucleotide sequence similarity with *Ruminococcus torques* strain JCM 6553^T (= ATCC 27756^T = VPI B2-51^T), currently the closest related species with a validly published name. The strain was observed to be a Gram-stain positive, non-motile, asporogenous and coccobacillary-shaped bacterium. It was found to be catalase positive and oxidase negative. Its major fatty acids were identified as $C_{16:0}$ (54%) and $C_{18:1n9}$ (30%). The draft genome of strain AT7^T is 3,069,882 bp long with 42.4% G+C content. 2925 genes were predicted, including 2867 protein-coding genes and 58 RNAs. Based on phenotypic, biochemical, phylogenetic and genomic evidence, we propose the creation of the new

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10482-018-1104-y) contains supplementary material, which is available to authorized users.

A. H. Togo · N. Armstrong · G. Dubourg · M. Richez · J. Delerce · A. Levasseur · D. Raoult · M. Million (✉)
Aix Marseille Univ, IRD, MEPHI, IHU-Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France
e-mail: matthieumillion@gmail.com

A. Diop · F. Bittar · N. Labas · P.-E. Fournier
Aix Marseille Univ, IRD, VITROME, IHU-Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France

M. Maraninchi · R. Valero
NORT "Nutrition, Obesity and Risk of Thrombosis", INSERM1062, INRA1260, Aix Marseille Université, 13385 Marseille, France

R. Valero
CHU Hôpital de la Conception, Service Nutrition, Maladies Métaboliques et Endocrinologie, APHM, 13385 Marseille, France

Published online: 31 May 2018

genus *Mediterraneibacter* and species, *Mediterraneibacter massiliensis*, that contains strain AT7^T (= CSUR P2086^T = DSM 100837^T), and the reclassification of *Ruminococcus faecis*, *Ruminococcus lactaris*, *Ruminococcus torques*, *Ruminococcus gnavus*, *Clostridium glycyrrhizinilyticum* as *Mediterraneibacter faecis* comb. nov., with type strain Eg2^T (= KCTC 5757^T = JCM15917^T), *Mediterraneibacter lactaris* comb. nov., with type strain ATCC 29176^T (= VPI X6-29^T), *Mediterraneibacter torques* comb. nov., with type strain ATCC 27756^T (= VPI B2-51^T), *Mediterraneibacter gnavus* comb. nov., with type strain ATCC 29149T (= VPI C7-9T) and *Mediterraneibacter glycyrrhizinilyticus* comb. nov., with type strain ZM35^T (= JCM 13368^T = DSM 17593^T), respectively.

Keywords *Mediterraneibacter massiliensis* · Taxonogenomics · Culturomics · Gut microbiota · Obesity

Abbreviations

AGIOS	Average of genomic identity of orthologous gene sequences
COG	Clusters of orthologous groups
CSUR	Collection de souches de l'Unité des Rickettsies
DDH	DNA–DNA hybridization
DSM	Deutsche Sammlung von Mikroorganismen
EUCAST	European Committee on antimicrobial susceptibility testing
FAME	Fatty acid methyl ester
GC/MS	Gas chromatography/mass spectrometry
GGDC	Genome-to-genome distance calculator
IUPAC	International Union of Pure and Applied Chemistry
ORF	Open reading frame
MALDI-TOF	Matrix-assisted laser-desorption/ionization time-of-flight

Introduction

Obesity is a major public health problem and the global obesity rate has doubled since 1980. In 2014, more than 1.9 billion adults were overweight and 600 million were obese (Ng et al. 2014). In France, the prevalence of obesity was 15.8% for men and 15.6% for women in 2016. Excess weight concerns nearly half of the French population (Matta et al. 2016). The treatment of obesity is a great challenge for health professionals. Bariatric surgery is currently the most effective treatment for morbid obesity. It is currently known that bariatric surgery leads to a lasting weight loss and reduces complications related to obesity. It has also been associated with an increase in the richness of the gut microbiota (Zhang et al. 2009; Kong et al. 2013). Bariatric surgery is a surgery that consists of gastric restriction (calibrated vertical gastroplasty, adjustable gastroplasty with adjustable rings and longitudinal gastrectomy) that reduces the amount of food to be ingested during a meal. It can be implemented in the form of a mixed system that combines gastric restriction with the bypass short-circuit (Roux-en-Y by-pass) to reduce the absorption of nutrients.

A new anaerobic bacterial species, strain AT7^T = CSUR P2086 = DSM 100837, was isolated by a'microbial culturomics' approach from the faeces of a morbidly obese patient before bariatric surgery. The goal of culturomics was to set up a collection of all human-associated microbes using different bacterial growth conditions to mimic natural conditions (Lagier et al. 2012, 2016). The conventional approaches for bacterial delineation have been based on phenotypic characteristics, the 16S RNA gene sequences similarity (Kim et al. 2014), phylogenetic relationship (Stackebrant and Ebers 2006), the G+C content of the genomic sequence and DNA-DNA hybridization (DDH) (Rosselló-Mora 2006; Meier-Kolthoff et al. 2014). However, these tools have some limitations. We proposed to include genomic and spectrometric data in a polyphasic approach to describe new bacterial taxa. This new method of delineation was named taxono-genomics (Ramasamy et al. 2014; Fournier et al. 2015). This approach combines the phenotypic, biochemical characteristics, the MALDI-TOF spectra, genomic analysis and phylogenetic comparison to delineate new bacterial taxa.

The bacterial strain isolated in this study clustered in phylogenetic analyses with some species of the genus *Ruminococcus*, which was first described in Antonie Van Leeuwenhoek with *Ruminococcus flavefaciens* as the type species (Sijpesteijn 1949). The genus is composed of Gram-positive bacteria and currently contains ten species as reported in the'List of

prokaryotic names with standing in nomenclature' (http://www.bacterio.net/ruminococcus.html). Of eight other species originally identified as belonging to the genus *Ruminococcus*, six species have been reclassified in the genus *Blautia* (Liu et al. 2008; Lawson and Finegold 2015) and two as *Trichococcus* (Liu 2002). However, the remaining members of the genus *Ruminococcus* form two distinct phylogenetic groups in two different families, as previously described (Rainey and Janssen 1995; Willems and Collins 1995; Rainey 2010; Lawson and Finegold 2015). the family *Ruminococcaceae* contains the *Ruminococcus* type species *Ruminococcus flavefaciens*, along with *Ruminococcus albus*, *Ruminococcus bromii*, *Ruminococcus callidus* and *Ruminococcus champanellensis* (*Ruminococcus* sensu stricto; Rainey 2010; Chassard et al. 2012), whereas *Ruminococcus faecis*, *Ruminococcus gnavus*, *Ruminococcus lactaris* and *Ruminococcus torques* cluster with members of the family *Lachnospiraceae*. This separation of members of the genus *Ruminococcus* into two distinct families suggested that taxonomy of the current *Ruminococcus* species should be clarified.

Here, we describe the main phenotypic, phylogenetic and genotypic features of strain AT7^T (= CSUR P2086 = DSM 100837) and propose the creation of a new genus, *Mediterraneibacter* gen. nov., that contains strain AT7^T as the type strain of *Mediterraneibacter massiliensis* sp. nov. Furthermore, creation of this new genus resolves most of the inconsistencies observed in the taxonomy of the genus *Ruminococcus*.

Materials and methods

Sample collection

Stool samples were collected for a study comparing the microbiota of subjects suffering from morbid obesity before and after surgery. The patients gave a written informed consent and the study was validated by the ethics committee of the Institut Federatif de Recherche IFR48 under agreement number 09-022, 2010. The stool sample containing the bacterium described here was collected from a 37-year-old obese French woman (BMI 44.75 kg/m^2; 116 kg, 1.61 m) in July 2012. The samples were aliquoted and stored at − 80 °C degrees before analysis.

Strain isolation and growth conditions

The strain was grown in May 2015. The stool sample of the patient was pre-incubated in blood culture bottles enriched with 10% filter-sterilised rumen fluid and 10% sheep blood, as described elsewhere (Lagier et al. 2016). The growth and monitoring procedures, colony identification and purification procedures were similar to those described elsewhere (Togo et al. 2017). The isolated colonies were then identified by MALDI-TOF-mass spectrometry, as previously described (Seng et al. 2009). The current Bruker and local "culturomics" database contains 8687 reference spectra of bacterial and fungal species.

Phenotypic and biochemical characterisation

Different growth temperatures (25, 28, 37, 45 and 55 °C) were tested on 5% sheep blood-enriched Columbia agar (bioMérieux, Marcy l'Etoile, France). Growth of strain AT7^T was tested under anaerobic atmosphere with the GENbag anaer system (bioMérieux), under microaerophilic atmosphere with the GENbag microaer system (bioMérieux) and under aerobic atmosphere, with or without 5% CO_2. Salt tolerance of the strain was tested using a 5–100 g/L NaCl concentration range on 5% sheep blood-enriched Schaedler agar (bioMérieux) under anaerobic atmosphere.

A fresh colony was observed between slides and slats using a Leica DM 1000 photonic microscope (Leica Microsystems, Nanterre, France) at 40× to assess bacterial motility. Transmission electron microscopy, using a Tecnai G20 microscope (FEI Company, Limeil-Brevannes, France) at an operating voltage of 60 kV was performed to observe strain AT7^T after negative coloration. Gram staining was performed using a Gram staining kit (bioMérieux) and observed using a photonic microscope Leica DM 2500 (Leica Microsystems, Nanterre, France) with a 100× oil-immersion objective lens. Thermal shock at 80 °C for 20 min was carried out to test for sporulation.

Biochemical assays were performed in triplicate using API Gallery systems: API® ZYM (bioMérieux), API® 20A (bioMérieux) and API®50 CH (bioMérieux) according to the manufacturer's instructions. Detection of catalase and oxidase activity (Becton, Dickenson and Company, Le Pont de Claix, France) was also performed.

The antibiotic susceptibility of strain AT7^T was tested following EUCAST recommendations (Citron et al. 1991; Matuschek et al. 2014). E-test strips for amikacin (0.016–256 µg/mL), vancomycin (0.016–256 µg/mL), imipenem (0.002–32 µg/mL), ceftriaxone (0.016–256 µg/mL), rifampicin (0.002–32 µg/mL), benzyl penicillin (0.002–32 µg/mL), amoxicillin (0.016–256 µg/mL), cefotaxime (0.002–32 µg/mL), metronidazole (0.016–256 µg/mL), minocycline (0.016–256 µg/mL), teicoplanin (0.016–256 µg/mL), erythromycin (0.016–256 µg/mL) and daptomycin (0.016–256 µg/mL) (bioMérieux) were deposited manually and the plates were incubated under anaerobic conditions for 48 h. Around the strip, elliptic zones of inhibition appeared and the intersection with the strip indicated the MIC (Citron et al. 1991). MICs were interpreted according to the EUCAST recommendations (http://www.eucast.org).

Fresh colonies were collected from 5% sheep blood-enriched Columbia agar (bioMérieux) after 48 h of incubation at 37 °C in an anaerobic atmosphere for cellular fatty acid methyl ester (FAME) analysis. The analysis was performed by Gas Chromatography/Mass Spectrometry (GC/MS), as described by Sasser (2006). GC/MS analyses were carried out as described by Dione et al. (2016). Metabolic end products were measured with a Clarus 500 chromatography system connected to a mass spectrometer (Perkin Elmer, Courtaboeuf, France), as detailed previously (Zhao et al. 2006), with some modifications. Acetic, propanoic, isobutanoic, butanoic, isopentanoic, pentanoic, isohexanoic, hexanoic and heptanoic acids were purchased from Sigma Aldrich (Lyon, France). A stock solution was prepared in water/methanol (50% v/v) at a final concentration of 50 mmol/L and then stored at − 20 °C. Calibration standards were freshly prepared in acidified water (pH 2–3 with 37% HCl) from the stock solution at the following concentrations: 0.5; 1; 5; 10 mmol/L. Short chain fatty acids were analysed from 3 independent culture bottles with BD Bactec™ Lytic/10 anaerobic/F culture vials media (Becton, Dickenson and Company); both blank and samples were analysed as described in previously (Togo et al. 2017).

Genomic characteristics

Sequencing and assembly

Genomic DNA (gDNA) of strain AT7^T was sequenced with the MiSeq technology (Illumina Inc, San Diego, CA, USA) using the mate pair strategy. It was barcoded in order to be mixed with 11 other projects using the nextera mate pair sample prep kit. Qubit assay with the high sensitivity kit (Thermo Fisher Scientific, Waltham, MA, USA) were used to quantify the gDNA of the strain at a concentration of 130 ng/µl. The nextera mate pair Illumina guide was used to prepare the mate pair library with 1.5 µg of gDNA. The sample was simultaneously fragmented and tagged with a mate pair junction adapter. The pattern of the fragmentation was validated on an Agilent 2100 bioanalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA 7500 labchip. The DNA fragments ranged from 1.5 kb up to 11 kb with an optimal size at 7.3 kb. No size selection was performed and 600 ng of tagmented fragments were circularised.

The circularised DNA was mechanically sheared to small fragments with an optimal size at 1336 bp on a Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The library profile was visualised on a high sensitivity bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) and the final concentration of the library was measured as 13.9 nmol/L. The libraries were normalised and pooled at 2 nM. After a denaturation step and dilution to 15 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated cluster generation and sequencing run were performed in a single 39-h run in a 2 × 151-bp. Total information of 8.9 Giga bases was obtained from a 1009 K/mm^2 cluster density with a cluster passing quality control filters of 91.5% (17,486,000 passing filter paired reads). Within this run, the index representation for strain AT7^T was determined to be of 8.4%. The 1,470,265 paired reads were trimmed and then assembled into 5 scaffolds using the SPAdes software (Bankevich et al. 2012).

Annotation and comparison

Open Reading Frames (ORFs) were predicted using Prodigal (Hyatt et al. 2010) with default parameters.

Nevertheless, the predicted ORFs were excluded if they spanned a sequencing gap region. The predicted bacterial protein sequences were searched against the GenBank and Clusters of Orthologous Groups (COG) databases using BLASTP (Evalue 1e-03, coverage 70% and identity percent 30%). The tRNAs and rRNAs were predicted using the tRNA Scan-SE and RNAmmer tools, respectively (Lowe and Eddy 1997). SignalP and TMHMM were used to identify signal peptides and the number of transmembrane helices, respectively (Krogh et al. 2001; Bendtsen et al. 2004). Mobile genetic elements were predicted using PHAST and RAST (Zhou et al. 2011; Overbeek et al. 2014). ORFans were identified if their BLASTP E-value was lower than 1e-03 for an alignment length greater than 80 amino acids. If alignment lengths were smaller than 80 amino acids, we used an E-value of 1e-05. Artemis and DNA Plotter were used for data management and visualisation of genomic features, respectively (Carver et al. 2009, 2012). Genomes were automatically retrieved from the 16S rRNA tree using XEGEN software (Phylopattern) (Gouret et al. 2009). For each selected genome, complete genome sequence, proteome genome sequence and orfeome genome sequence were retrieved from the FTP of NCBI. All proteomes were analysed with proteinOrtho (Lechner et al. 2011). Then, for each pair of genomes, a similarity score of the average genomic identity of orthologous gene sequences (AGIOS) was computed. This score is the mean value of nucleotide similarity between all pairs of orthologous proteins for the two genomes studied (Ramasamy et al. 2014). For the evaluation of genomic similarity, digital DDH (dDDH) values were estimated using GGDC formula 2 (Meier-Kolthoff et al. 2013b). The average amino acid identity (AAI) was also calculated, based on the overall similarity between two genomic datasets of proteins, (Konstantinidis and Tiedje 2005; Rodriguez-R and Konstantinidis 2014) and is available at http://enve-omics.ce.gatech.edu/aai/index.

For the genomic comparison of strain AT7^T, the genomes of *R. lactaris* strain ATCC 29176^T = VPI X6-29^T (ABOU00000000) (Moore et al. 1976), *R. torques* strain ATCC 27756^T = VPI B2-51^T (GCA0001153925) (Holdeman and Moore 1974), *R. faecis* strain Eg2^T = KCTC 5757^T = JCM15917^T (BBDW0100000) (Kim et al. 2011), *Clostridium glycyrrhizinilyticum* strain ZM35^T = JCM 13368^T = DSM 17593^T (BBAB0100000) (Sakuma et al. 2006), *R. gnavus* strain ATCC 29149^T = VPI C7-9^T (PUEL00000000) (Moore et al. 1976), *Ruminococcus gauvreauii* strain CCRI-16110^T = NML 060141^T = CCUG 54292^T = JCM 14987^T (AUDP00000000) (Domingo et al. 2008), *R. albus* strain 7^T = ATCC 27210^T = DSM 20455^T = JCM 14654^T (CP002403) (Hungate 1957), *R. bromii* strain V.P.I. 6883^T = ATCC 27255^T (FMUV00000000) (Moore et al. 1972), *R. callidus* strain ATCC 27760^T = VPI S7-31^T (AWVF00000000) (Holdeman and Moore 1974), *R. champanellensis* strain 18P13^T = DSM 18848^T = JCM 17042^T (FP929052) (Chassard et al. 2012) *Coprococcus comes* strain ATCC 27758^T = VPI C1-38^T (ABVR00000000) (Holdeman and Moore 1974) and *R. flavefaciens* strain C94^T = ATCC 19208^T (JAEF00000000) (Sijpesteijn 1949) were used.

An annotation of the entire proteome was performed to define the distribution of functional classes of predicted genes according to the Clusters of Orthologous Groups of proteins (by using the same method as for genome annotation). The genome of this AT7^T was locally aligned pairwise using the BLAST algorithm against each of the selected genomes (Kent 2002; Auch et al. 2010).

Phylogenetic analysis

To clarify the taxonomic inconsistencies among *Ruminococcus* species, we have achieved the most robust strategy to date based on a phylogenetic tree based on 271 orthologous genes from the genomes of 27 closely related species and 1 outgroup (*Escherichia coli*). All 28 genomes were downloaded from NCBI (www.ncbi.nlm.nih.gov). For orthologue detection, we applied Proteinortho with default values (Lechner et al. 2011). All orthologous genes were aligned using Muscle (Edgar 2004) and then concatenated. Phylogenetic reconstruction was performed using the maximum likelihood method with the Kimura 2 parameter model and bootstrap value of 100.

Results

MALDI-TOF analysis

The spectrum generated from strain AT7^T (Fig. 1) did not match with that of any reference strain in the

Bruker plus culturomics database. Accordingly, this strain was suspected to correspond to a new species so that phenotypic and chemotaxonomic characteristics were determined, and genome sequencing was performed.

Phenotypic and biochemical characterisation

Strain AT7^T was observed to be non-motile, asporogenous, coccobacillary -shaped, Gram-strain positive (Fig. S1) and anaerobic. The strain exhibits catalase activity but not oxidase activity. Growth was observed on 5% sheep blood Colombia agar plates between 28 and 45 °C, with optimal growth observed at 37 °C after 48 h of incubation under anaerobic atmosphere. The colonies were observed to be small (about 0.5–1 mm in diameter), translucent, punctiform and not haemolytic on 5% sheep blood Columbia agar. No growth of this bacterium was observed using 10–100 g/L of NaCl concentration on 5% sheep blood Schaedler agar plates. Strain AT7^T was observed to grow at pH ranging from 6.5 to 8.5, with optimal growth at 7.2. Cells were determined to be 0.2–0.4 wide and 1–1.4 µm long under electron microscopy (Fig. S2). The phenotypic characteristics of strain AT7^T were compared with those of its close phylogenetic neighbours, as shown in Table 1.

Using the API® ZYM test system, positive reactions were observed with trypsin, α-chymotrypsin, naphtol-AS-BI-phosphohydrolase and β-glucuronidase but negative reactions were observed with phosphatase alkaline, esterase, esterase lipase, lipase, leucine arylamidase, valine arylamidase, cystine arylamidase, phosphatase acid, α-galactosidase, β-galactosidase, α-glucosidase, β-glucosidase, *N*-acetyl-β-glucosaminidase, α-mannosidase and α-fructosidase. The API® 50 CH test system revealed that strain AT7^T exhibits positive reactions for aesculin, arbutine, D-arabinose, D-cellobiose, D-fructose, D-galactose, D-glucose, D-lactose, D-maltose, D-mannitol, D-mannose, D-melibiose, D-saccharose, D-trehalose, D-xylose, dulcitol, gentiobiose, inositol, L-arabinose, L-sorbose, L-xylose, methyl-α D-glucopyranoside, potassium 2-cetogluconate, salicin and xylitol. Negative reactions were obtained with adonitol, amygdalin, D-arabitol, D-fucose, D-lyxose, D-melezitose, D-raffinose, D-ribose, D-sorbitol, D-tagatose, D-turanose, erythritol, glycerol, inulin, L-rhamnose, methyl-αD-mannopyranoside, methyl-β D-xylopyranoside, *N*-acetyl-glucosamine, glycogen, L-arabitol, L-fucose, potassium gluconate, potassium 5-cetogluconate and starch. Using the API® 20A test system, positive reactions were observed with aesculin, D-cellobiose, D-glucose, D-lactose, D-maltose, D-mannitol, D-mannose, D-saccharose, D-xylose,

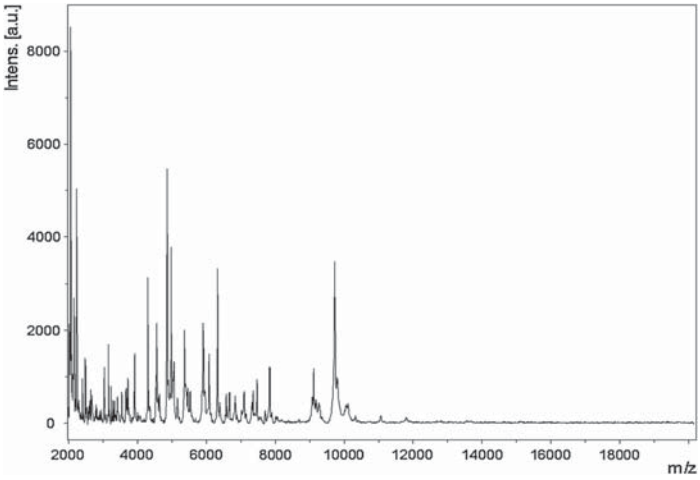


Fig. 1 Reference mass spectrum from strain AT7^T. Spectra from 12 individual colonies were compared and a reference spectrum was generated

Table 1 Differential characteristics of strain AT7^T compared to those of closely related species. (1) Strain AT7^T; (2) *M. faecis* strain Eg2^T (= KCTC 5757^T = JCM 15917^T); (3) *M. lactaris* strain ATCC 29176$^{T;}$ (= VPI X6-29^T); (4) *M. torques* strain ATCC 27756^T (= VPI B2-51^T); (5) *M. glycyrrhizinilyticus* strain ZM35^T (= JCM 13368^T = DSM 17593^T); (6) *M. gnavus* strain ATCC 29149^T (= VPI C7-9^T); (7) *Co. comes* strain ATCC 27758^T (= VPI C1-38^T), (8) *R. gauvreauii* strain CCRI-16110^T (= NML 060141^{T-} = CCUG 54292^T = JCM 14987^T); (9) *R. albus* strain 7^T (= ATCC 27210^T = DSM 20455^T = JCM 14654^T); (10) *R. bromii* strain V.P.I. 6883^T (= ATCC 27255^T); (11) *R. callidus* strain ATCC 27760^T (= VPI S7-31^T); (12) *R. champanellensis* strain 18P13^T (= DSM 18848^T = JCM 17042^T); (13) *R. flavefaciens* strain C94^T(ATCC 19208^T)

Properties	1	2	3	4	5	6	7	8	9	10	11	12	13
Catalase	+	+	+	+	−	+	−	−	+	−	+	−	−
Aesculin hydrolysis	+	+	−	+	−	+	V	−	−	−	+	+	−
Gelatine hydrolysis	+	−	+	+	−	+	+	−	Na	−	w	−	−
Acid production from													
Arabinose	+	−	−	−	+	+	−	−	−	−	−	−	−
Cellobiose	+	−	−	−	−	−	−	−	+	−	+	+	w
Erythritol	−	Na	−	−	Na	−	Na	−	Na	−	−	−	Na
Fructose	+	Na	+	+	−	+	+	+	−	−	w	−	−
Galactose	+	Na	Na	+	Na	+	+	+	−	−	−	−	−
Glucose	+	+	+	+	+	+	+	+	+	w	+	−	−
Lactose	+	+	+	+	+	−	+	−	−	−	+	−	−
Maltose	+	+	+	+	+	+	+	−	−	−	+	−	−
Mannitol	+	−	+	−	−	−	W	+	−	−	−	−	−
Mannose	+	−	w	−	−	−	W	−	+	−	w	−	−
Melibiose	+	Na	−	−	−	w	+	−	−	−	+	−	−
Raffinose	−	+	−	−	+	+	+	−	−	−	+	−	−
Rhamnose	−	−	−	−	+	+	Na	−	−	−	−	−	−
Ribose	−	Na	−	−	−	+	−	+	−	−	−	−	−
Saccharose	+	−	−	+	−	+	+	+	+	−	+	−	−
Salicin	+	−	−	w	−	+	W	−	−	−	−	−	−
Sorbitol	−	+	v	−	−	−	W	+	Na	−	−	−	Na
Starch	−	Na	−	−	Na	+	−	−	−	+	−	−	−
Trehalose	+	−	−	−	−	−	−	−	−	−	−	−	−
Xylose	+	−	−	−	+	+	+	−	−	−	w	−	−
Major end product of carbohydrate metabolism	A Ih	L A	F A L S	L A F	Na	F A L	L A B	A	A L S E	A F L P E	S A F	A S	A S F B L
G+C content (%)	42.4	43.4	45	42	45.7	43	40	47.6	44.2	39.1	43	53	43.2
Source	Human feces	Human feces	Human feces	Human feces	Human feces	Human feces	Human feces	Human feces	Rumen of cattle	Human feces	Human feces	Human feces	Human feces

A acetic acid, *F* formic acid, *L* lactic acid, *S* succinic acid, *E* ethanol, *P* pyruvic acid, B butyric acid, *Ih* isohexanoic acid, + positive reaction, − negative reaction, *Na* not available, *w* weakly reaction, *v* variable

D-trehalose, gelatine, L-arabinose and salicin. Reactions for D-raffinose, D-melezitose, D-sorbitol, glycerol, L-rhamnose, L-tryptophan and urea were found to be negative.

Strain AT7^T was found to be susceptible to vancomycin (2 μg/mL), imipenem (0.047 μg/mL), ceftriaxone (0.75 μg/mL), rifampicin (0.002 μg/mL), benzyl penicillin (0.094 μg/mL), amoxicillin (0.094 μg/mL), cefotaxime (2 μg/mL), metronidazole (0.19 μg/mL), minocycline (0.0125 μg/mL), teicoplanin (0.016 μg/mL), erythromycin (0.025 μg/mL) and daptomycin (1 μg/mL). However, the strain was found to be resistant to amikacin (> 256 μg/mL). The minimum inhibitory concentration for each antimicrobial used is in parenthesis.

Total cellular fatty acid composition analysis of strain AT7^T revealed that the most abundant fatty acids were $C_{16:0}$ (54%) and $C_{18:1n9}$ (30%). Minor amounts of other fatty acids ($C_{18:0}$, $C_{14:0}$, $C_{18:1n7}$, $C_{18:1n6}$, $C_{15:0}$, $C_{16:1n7}$, $C_{12:0}$, $C_{17:0}$, anteiso-$C_{15:0}$ and iso-$C_{15:0}$) were detected. The results of fatty acid analysis are summarised in Table 2.

Analysis of metabolic end products revealed that strain AT7^T produces (after 72 h) acetic acid (17.1 ± 0.5 mM), isohexanoic acid (6 ± 0.2 mM), isobutanoic acid (2.3 ± 0.1 mM), butanoic acid (1.3 ± 0.1 mM), isopentanoic acid (1.3 ± 0.1 mM) and propanoic acid (0.7 ± 0.1 mM), but also small quantities (< 0.5 mM) of pentanoic and hexanoic acid.

Genomic analysis

Genome properties

The draft genome of strain AT7^T has been deposited in EMBL-EBI under accession number FAVJ00000000 and is 3,069,882 bp long with 42.4% G+C content (Fig. 2). It is composed of five scaffolds and eight contigs. Among the 2925 predicted genes, 2867 are protein-coding genes and 58 are RNA genes (two 5S rRNA genes, one 16S rRNA gene, three 23S rRNA genes and fifty-two tRNA genes). A total of 2191 genes (76.4%) were assigned a putative function by COGs or NR blast. A total of 108 genes were identified as ORFans (4%). Using ARG-ANNOT (Gupta et al. 2014), no resistance genes were found, however, three genes (0.1%) were identified as PKS or NRPS (Conway and Boddy 2013). Using PHAST and RAST, 1136 genes (40%) were found to be associated with mobilome elements. The remaining 483 genes (17%) were annotated as hypothetical proteins.

Table 2 Cellular fatty acid profiles of strain AT7^T compared with those of closely related species; (1) Strain AT7^T; (2) *Ruminococcus faecis* strain Eg2^T (= KCTC 5757^T = JCM 15917^T); (3) *Ruminococcus gauvreauii* strain CCRI-16110^T (= NML 060141^T = CCUG 54292^T = JCM 14987^T) (4) *Ruminococcus champanellensis* strain 18P13^T (= DSM 18848^T = JCM 17042^T)

Fatty acids	1	2	3	4
anteiso-C15:0	< 1	ND	ND	**19.6**
anteiso-C17:0	0	ND	ND	2.8
C12:0	< 1	2.3	ND	ND
C13:1n12/C11:1 2-OH	0	1.9	ND	ND
C14:0	2.0 ± 0.2	**10**	**16.9**	ND
C15:0	< 1	ND	ND	ND
C15:2/C15:1n7	0	2.0	ND	ND
C16:0	**54.0 ± 4.2**	**27.7**	**19.9**	ND
C16:1n7	< 1	ND	ND	ND
C16:1n9	0	2.5	ND	ND
C17:0	< 1	ND	ND	0,4
C17:1n9/C17:2	0	2.7	ND	ND
C18: 1n11	0	ND	ND	ND
C18:0	9.0 ± 1.2	2.9	ND	0.7
C18:1c11/t9/t6	0	6.7	ND	ND
C18:1n11	0	ND	ND	ND
C18:1n6	2.0 ± 0.1	ND	ND	ND
C18:1n7	2.0 ± 1.2	ND	ND	ND
C18:1n9	**30.0 ± 2.3**	3.1	8.4	ND
C18:2n9, 12	0	3.3	ND	ND
iso-C13:03-OH	0	ND	ND	0.2
iso-C15:0	< 1	ND	ND	**26.6**
iso-C16:0	0	ND	ND	8.8
iso-C17:0	0	ND	ND	0.4

Date for 2–4 are taken from (Domingo et al. 2008; Kim et al. 2011; Chassard et al. 2012). *ND* not detected. Data were not available for *Ruminococcus torques* strain ATCC 27756^T, *Ruminococcus lactaris* strain ATCC 29176^T, *Clostridium glycyrrhizinilyticum* strain ZM35^T, *Coprococcus comes* strain ATCC 27758^T, *Ruminococcus gnavus* strain ATCC 29149^T, *Ruminococcus albus* strain 7^T, *Ruminococcus bromii* strain ATCC^T; *Ruminococcus callidus* strain ATCC 27760^T and *Ruminococcus flavefaciens* strain C94^T

Bold values indicate major cellular fatty acids of the strains

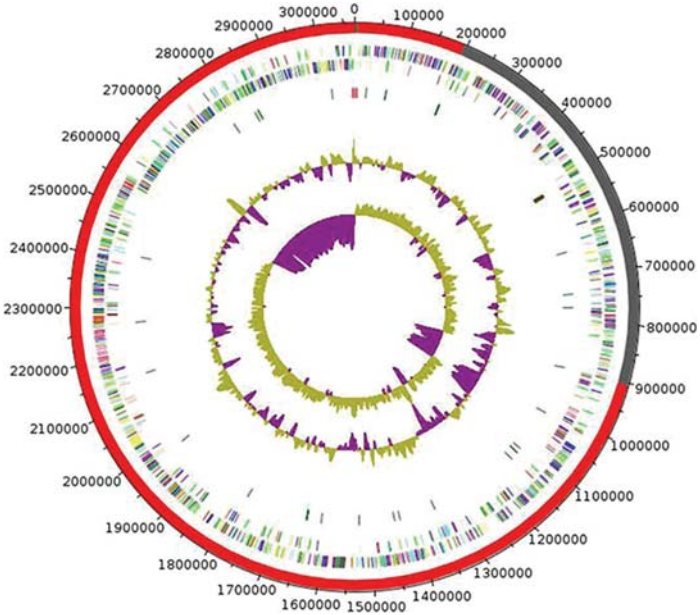


Fig. 2 Graphical circular map of the genome of strain AT7^T. From outside to the centre: Contigs (red/grey), COG category of genes on the forward strand (three circles), genes on forward strand (blue circle), genes on the reverse strand (red circle), COG category on the reverse strand (three circles), GC content

16S gene-based phylogenetic analysis

16S rRNA gene sequence similarity values lower than 98.7% or 95%, have been used to assign strain to novel species or genera, respectively (Stackebrant and Ebers 2006; Kim et al. 2014; Yarza et al. 2014). The 16S gene sequence of strain AT7^T exhibited a 95.2, 95.6, 95.6 and 95.9% nucleotide sequence similarity with *C. glycyrrhizinilyticum* strain ZM35^T (= JCM 13368^T = DSM 17593^T), *R. lactaris* strain ATCC 29176^T (= VPI X6-29^T), *R. faecis* strain Eg2^T (= KCTC 5757^T = JCM15917^T) and *R. torques* strain JCM 6553^T (= ATCC 27756^T = VPI B2-51^T), the closely related species with validly published names according to the phylogenetic analysis. The 16S rRNA gene sequence similarity values of strain AT7^T and other members of the genus *Ruminococcus* are displayed in Table 3. Supplementary figure 3 (Fig. S3) shows a 16S rRNA gene tree for all *Ruminococcus* type strains plus type strains of type species and other representative species of genera in the families *Lachnospiraceae* and *Ruminococcaceae*. The 16S rRNA gene sequence of strain AT7^T has been deposited in EMBL-EBI under accession number LN881607.

Genome comparison

The draft genome sequence of strain AT7^T (3.07 Mb) is smaller than those of *Co. comes*, *R. faecis*, *R. flavefaciens*, *R. gnavus*, *R. gauvreauii* and *R. albus* (3.24, 3.26, 3.44, 3.62, 3.73 and 3.84 Mb respectively), larger than those of *R. bromii*, *R. champanellensis*, *R. lactaris* and *R. torques* (2.28, 2.54, 2.73 and 2.74 Mb respectively) but similar to that of *R. callidus* (3.09 Mb). Its G+C content (42.4%) is similar to that of *Co. comes* (42.5), lower than those of *R. gnavus*, *R. callidus*, *R. faecis*, *R. flavefaciens*, *R. lactaris*, *C. glycirrhizinilyticum*, *R. albus* and *R. champanellensis*

Table 3 16S rRNA gene sequence similarity values of strain AT7^T obtained from comparisons with closely related species

RRNA sequences From	Strain AT7	*M. faecis*	*M. lactaris*	*M. torques*	*M. glycyrrhizinilyticus*	*M. gnavus*	*Co. comes*	*R. gauvreauii*	*R. albus*	*R. bromii*	*R. callidus*	*R. champanellensis*	*R. flavefaciens*
Similarity of 16S rRNA gene sequences													
Strain AT7^T (LN881607)													
M. faecis strain Eg2^T (FJ611794)	96												
M. lactaris strain ATCC 29176^T (L76602)	96	96											
M. torques strain VPI B2-51^T (L76604)	95	96	95										
M. glycyrrhizinilyticus strain ZM35^T (AB233029)	95	96	94	95									
M. gnavus strain ATCC 29149^T (X94967)	92	95	94	94	95								
Co. comes strain VPI C1-38^T (EF031542)	94	95	94	94	96	94							
R. gauvreauii strain CCRI-16110^T (EF529620)	91	93	92	92	92	93	93						
R. albus strain 7^T (L76598)	85	86	86	86	83	84	83	84					
R. bromii strain ATCC 27255^T (L76600)	82	83	82	82	82	93	82	83	89				
R. callidus strain ATCC 27760^T (L76596)	84	84	84	85	85	84	84	84	90	89			
R. champanellensis strain 18P13^T (AJ515913)	83	83	85	85	83	84	84	84	92	89	95		
R. flavefaciens strain C94^T (L76603)	84	83	83	86	84	84	82	83	91	89	93	94	

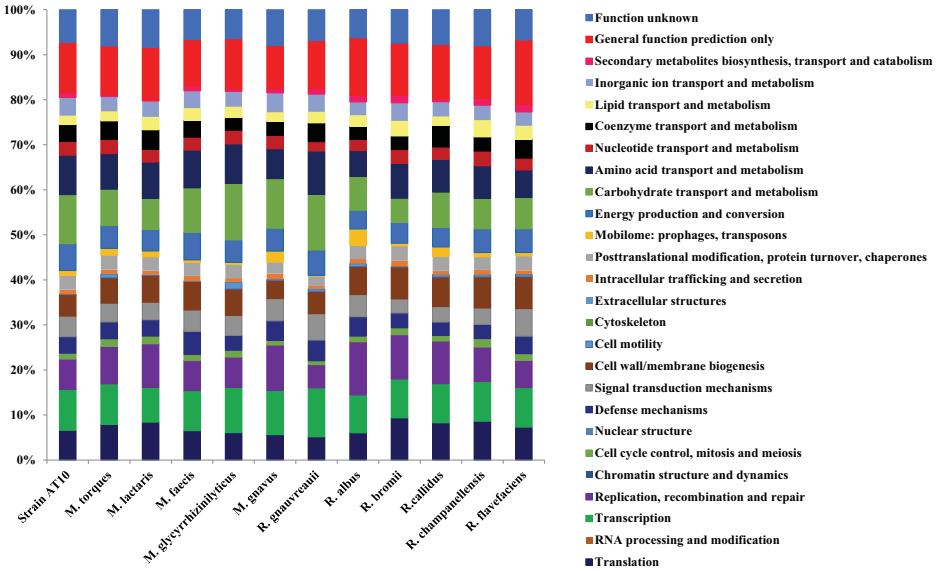


Fig. 3 Distribution of functional classes of predicted genes according to the clusters of orthologous groups of proteins of strain AT7^T compared with closely related species

(43, 43, 43, 43.4, 44, 45, 45. 3, 45.8 and 53% respectively), but higher than those of *R. bromii*, *R. gauvreauii* and *R. torques* (40, 40 and 42% respectively). Its gene content (2925) is lower than those of *R. faecis*, *R. albus*, *C. glycirrhizinilyticum*, *Co. comes* and *R. gnavus* (3220; 3335; 3359, 3529 and 3744 respectively), but higher than those of *R. flavefaciens*, *R. gauvreauii*, *R. champanellensis*, *R. torques*, *R. lactaris*, and *R. bromii* (1807; 2110; 2371; 2491; 2486 and 2852 respectively). Even so, the distribution of genes into COG was similar among all compared genomes (Fig. 3 and Table 4). AGIOS values (Table 5) among compared species, except for strain AT7^T, ranged from 59% between *R. torques* and *R. champanellensis* to 75.9% between *R. lactaris* and *R. faecis*. When strain AT7^T was compared to other species, this value ranged from 59.2% with *R. champanellensis* to 72.7% with *R. torques*. The dDDH values of strain AT7^T ranged from 17.7% with *R. gauvreauii* to 29.2% with *R. callidus* and are shown in Table 6. The average amino acid identity values between strain AT7^T and closely related species ranged from 60.98% between *Co. comes* and strain AT7 to 73.49% between *R. faecis* and *R. lactaris*. However, these values were lower when strain AT7^T and the group of closely related types strains were compared with *R. flavefaciens* and the species of the genus *Ruminococcus* sensu stricto as shown in Table 7.

Phylogenetic tree based on 271 concatenated orthologous genes from genomes of the 28 closest species

As *Ruminococcus* species can be separated into two different clusters belonging to two different taxonomic families (*Lachnospiraceae* and *Ruminococcaceae*), we decided to apply one of the best performing current taxonomic approaches based on genomic analysis using shared orthologous genes among closely related species (Fig. 4). Strain AT7^T was found to cluster with *R. faecis*, *R. lactaris*, *R. torques*, *R. gnavus* and *C. glycyrrhizinilyticum*, forming a homogeneous cluster within the family

Table 4 Number of genes associated with the 25 general COG functional categories of strain AT7^T compared to those of its closest species; (1) Strain AT7^T; (2) *M. faecis* strain Eg2^T(= KCTC 5757^T = JCM 15917^T); (3) *M. lactaris* strain ATCC 29176^T(= VPI X6-29^T); (4) *M. torques* strain ATCC 27756^T (= VPI B2-51^T); (5) *M. glycyrrhizinilyticus* strain ZM35^T (= JCM 13368^T = DSM 17593^T); (6) *M. gnavus* strain ATCC 29149^T (= VPI C7-9^T); (7) *Co. comes* strain ATCC 27758^T (= VPI C1-38^T); (8) *R. gauvreauii* strain CCRI-16110^T (= NML 060141^T = CCUG 54292^T = JCM 14987^T); (9) *R. albus* strain 7^T (= ATCC 27210^T = DSM 20455^T = JCM 14654^T); (10) *R. bromii* strain V.P.I. 6883^T (= ATCC 27255^T); (11) *R. callidus* strain ATCC 27760^T (= VPI S7-31^T); (12) *R. champanellensis* strain 18P13^T (= DSM 18848^T = JCM 17042^T); (13) *R. flavefaciens* strain C94^T (ATCC 19208^T)

Genes	1	2	3	4	5	6	7	8	9	10	11	12	13
Translation	143	174	149	140	150	144	162	153	145	148	145	134	151
RNA processing and modification	0	0	0	0	0	0	0	0	0	0	0	0	0
Transcription	198	238	137	160	248	252	216	323	202	137	153	138	181
Replication, recombination and repair	146	178	171	148	166	257	206	154	283	155	166	119	124
Chromatin structure and dynamics	0	0	0	0	0	0	0	0	0	0	0	0	0
Cell cycle control, mitosis and meiosis	28	39	30	30	37	28	37	25	31	24	22	29	31
Nuclear structure	0	0	0	0	0	0	0	0	0	0	0	0	0
Defence mechanisms	80	137	64	66	81	110	98	137	103	52	52	49	80
Signal transduction mechanisms	99	125	69	74	112	127	133	174	118	50	61	57	127
Cell wall/membrane biogenesis	104	169	105	100	142	102	140	148	150	112	115	107	149
Cell motility	2	1	3	15	38	6	3	16	15	2	9	8	9
Cytoskeleton	0	0	0	0	0	0	0	0	0	0	0	0	0
Extracellular structures	0	0	0	0	0	0	0	0	0	0	0	0	0
Intracellular trafficking and secretion	22	32	16	17	22	33	25	19	25	21	15	17	17
Posttranslational modification, protein turnover, chaperones	69	81	55	57	74	66	77	65	71	51	57	45	66
Mobilome: prophages, transposons	23	14	23	27	13	62	14	7	89	9	36	15	16
Energy production and conversion	129	165	85	91	121	130	128	165	99	73	77	82	108
Carbohydrate transport and metabolism	237	263	122	142	311	282	212	367	181	85	137	105	144
Amino acid transport and metabolism	190	224	142	140	215	170	229	287	138	122	128	113	126
Nucleotide transport and metabolism	66	77	51	56	75	75	61	62	59	49	48	51	54
Coenzyme transport and metabolism	82	100	76	73	70	80	86	124	70	48	84	49	87
Lipid transport and metabolism	43	73	51	38	59	52	55	73	60	53	36	58	61
Inorganic ion transport and metabolism	88	102	62	58	84	111	74	116	70	63	57	51	64
Secondary metabolites biosynthesis, transport and catabolism	23	28	4	6	14	24	16	39	35	26	11	24	34
General function prediction only	243	276	207	193	274	244	235	315	307	184	212	182	295
Function unknown	157	178	148	143	159	204	169	204	150	117	136	125	139

COGs Clusters of Orthologous Groups database

Table 5 Pairwise comparison of strain AT7^T with closely related species using the AGIOS parameter; (1) Strain AT7^T; (2) *M. faecis* strain Eg2^T (= KCTC 5757^T = JCM 15917^T); (3) *M. lactaris* strain ATCC 29176^T (= VPI X6-29^T); (4) *M. torques* strain ATCC 27756^T (= VPI B2-51^T); (5) *M. glycyrrhizinilyticus* strain ZM35^T (= JCM 13368^T = DSM 17593^T); (6) *M. gnavus* strain ATCC 29149^T (= VPI C7-9^T); (7) *Co. comes* strain ATCC 27758^T = VPI C1-38^T; (8) *R. gauvreauii* strain CCRI-16110^T (= NML 060141^T = CCUG 54292^T = JCM 14987^T); (9) *R. albus* strain 7^T (= ATCC 27210^T = DSM 20455^T = JCM 14654^T); (10) *R. bromii* strain V.P.I. 6883^T (= ATCC 27255^T); (11) *R. callidus* strain ATCC 27760^T (= VPI S7-31^T); (12) *R. champanellensis* strain 18P13^T (= DSM 18848^T = JCM 17042^T); (13) *R. flavefaciens* strain C94^T(ATCC 19208^T)

Species	1	2	3	4	5	6	7	8	9	10	11	12	13
Strain AT7	**2869**	1002	1122	1177	1185	1256	987	1060	690	646	646	661	703
M. faecis	71.07	**3921**	1018	914	945	1017	925	912	613	579	609	564	609
M. lactaris	72.32	75.92	**2479**	1118	1055	1142	1000	1024	724	673	701	669	719
M. torques	72.57	72.04	73.10	**2489**	1077	1174	901	996	675	661	666	638	698
M. glycyrrhizinilyticus	71.73	70.58	71.86	71.88	**3359**	1184	910	1004	661	635	636	618	656
M. gnavus	72.70	71.45	72.71	71.96	72.68	**3760**	989	1092	710	663	693	642	710
Co. comes	69.10	71.82	71.07	68.85	69.69	70.17	**3529**	936	629	575	618	571	619
R. gauvreauii	65.90	65.20	66.72	66.16	66.53	66.88	66.32	**3790**	749	696	706	703	764
R. albus	60.41	60.49	61.22	60.76	60.15	60.95	60.86	60.45	**4051**	724	841	883	948
R. bromii	60.67	61.16	61.25	61.40	60.33	61.06	61.16	60.17	62.61	**2485**	715	729	723
R. callidus	59.86	60.38	61.27	60.17	61.01	60.08	61.05	61.09	63.97	61.36	**2847**	886	941
R. champanellensis	59.23	58.76	60.05	58.99	60.34	60.23	59.71	60.54	63.99	60.47	68.44	**2356**	935
R. flavefaciens	60.30	60.86	61.20	60.77	60.02	60.87	60.77	60.47	66.96	63.39	65.43	65.32	**3089**

Upper right, numbers of orthologous proteins shared between genomes; lower left, average percentage similarity of nucleotides corresponding to orthologous proteins shared between genomes and in bold, number of proteins for each species genome

Table 6 Pairwise comparison of strain AT7^T with closely related species using the dDDH parameter; (1) Strain AT7^T; (2) *M. faecis* strain Eg2^T(= KCTC 5757^T = JCM 15917^T); (3) *M. lactaris* strain ATCC 29176^T (= VPI X6-29^T); (4) *M. torques* strain ATCC 27756^T (= VPI B2-51^T); (5) *M. glycyrrhizinilyticus* strain ZM35^T (= JCM 13368^T = DSM 17593^T); (6) *M. gnavus* strain ATCC 29149^T (= VPI C7-9^T); (7) *Co. comes* strain VPI C1-38^T (= ATCC 27758^T); (8) *R. gauvreauii* strain CCRI-16110^T (= NML 060141^T = CCUG 54292^T = JCM 14987^T); (9) *R. albus* strain 7^T (= ATCC 27210^T = DSM 20455^T = JC14654^T); (10) *R. bromii* strain V.P.I. 6883^T (= ATCC 27255^T); (11) *R. callidus* strain ATCC 27760^T (= VPI S7-31^T); (12) *R. champanellensis* strain 18P13^T (= DSM 18848^T = JCM 17042^T); (13) *R. flavefaciens* strain C94^T(= ATCC 19208^T)

Species	1	2	3	4	5	6	7	8	9	10	11	12	13
1		20.6% ± 2.3	19.4% ± 2.3	22.3% ± 2.3	18.9% ± 2.3	19.3% ± 2.3	24.1% ± 2.4	17.7% ± 2.2	26.7% ± 2.4	20.5% ± 2.3	29.2% ± 2.4	27.6% ± 2.4	27.1% ± 2.4
2			24.3% ± 2.3	23.2% ± 2.3	23.1% ± 2.4	25.1% ± 2.4	35.8% ± 2.5	24.4% ± 2.4	22.4% ± 2.3	16.7% ± 2.2	39.5% ± 2.5	20% ± 2.3	15.1% ± 2.1
3				24.6% ± 2.3	24.2% ± 2.4	21.3% ± 2.3	27.3% ± 2.5	21.9% ± 2.3	26.6% ± 2.4	19.5% ± 2.3	29.5% ± 2.4	23.5% ± 2.3	24.8% ± 2.4
4					24.5% ± 2.4	26.5% ± 2.4	27.2% ± 2.5	21.7% ± 2.3	25.8% ± 2.4	22.7% ± 2.4	38% ± 2.5	21.8% ± 2.3	26.5% ± 2.4
5						22.5% ± 2.4	24.2% ± 2.4	18.3% ± 2.3	28.4% ± 2.5	23.1% ± 2.4	23.6% ± 2.4	30.4% ± 2.5	40.6% ± 2.5
6							23.1% ± 2.3	19.6% ± 2.3	22.6% ± 2.4	21.7% ± 2.3	22.3% ± 2.3	26.8% ± 2.4	24.7% ± 2.4
7								23.9% ± 2.4	25.7% ± 2.4	21.8% ± 2.3	39.9% ± 2.5	28.8% ± 2.4	22.4% ± 2.4
8									18.3% ± 2.2	22.6% ± 2.4	19% ± 2.3	25.8% ± 2.4	18.8% ± 2.3
9										24.6% ± 2.4	24.4% ± 2.4	24.7% ± 2.4	18.8% ± 2.3
10											29.7% ± 2.4	19.3% ± 2.3	15.9% ± 2.2
11												20.4% ± 2.3	21.3% ± 2.3
12													17.7% ± 2.2
13													

Confidence intervals indicate inherent uncertainty in estimating DDH values from intergenomic distances based on models derived from empirical test data sets. These results are consistent with the 16S rRNA and phylogenomic analyses as well as the GGDC results: DDH, DNA-DNA hybridization and Genome-to-Genome Distance Calculator. HSP: high-scoring segment pairs

Table 7 The average amino acid identity values of strain AT7^T compared with those of its phylogenetically close neighbours; (1) Strain AT7; (2) *M. faecis* strain Eg2^T (= KCTC 5757^T = JCM 15917^T); (3) *M. lactaris* strain ATCC 29176^T (= VPI X6-29^T); (4) *M. torques* strain ATCC 27756^T (= VPI B2-51^T); (5) *M. glycyrrhizinilyticus* strain ZM35^T (= JCM 13368^T = DSM 17593^T); (6) *M. gnavus* strain ATCC 29149^T (= VPI C7-9^T); (7) *Co. comes* strain ATCC 27758^T (= VPI C1-38^T), (8) *R. gauvreauii* strain CCRI-16110^T (= NML 060141^T = CCUG 54292^T = JCM 14987^T); (9) *R. albus* strain 7^T (= ATCC 27210^T = DSM 20455^T = JCM 14654^T); (10) *R. bromii* strain V.P.I. 6883^T (= ATCC 27255^T); (11) *R. callidus* strain ATCC 27760^T (= VPI S7-31^T); (12) *R. champanellensis* strain 18P13^T (= DSM 18848^T = JCM 17042^T); (13) *R. flavefaciens* strain C94^T(ATCC 19208^T)

	1	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)	9 (%)	10 (%)	11 (%)	12 (%)	13 (%)
Strain AT7		65.8	66.8	69.2	67.9	68.2	60.9	53.8	43.3	44.0	43.3	43.5	43.7
M. faecis			73.4	67.0	64.9	65.0	67.1	53.9	44.0	44.6	46.1	44.1	43.7
M. lactaris				68.6	65.4	65.7	64.4	54.7	44.3	45.2	45.4	44.2	44.3
M. torques					67.3	66.0	60.6	54.7	44.0	44.6	44.3	44.0	44.3
M. glycyrrhizinilyticus						67.3	60.9	54.7	43.5	44.1	44.7	43.8	43.5
M. gnavus							60.3	54.3	43.3	43.9	44.2	43.1	43.3
Co. comes								54.2	44.0	44.5	44.9	43.8	43.8
R. gauvreauii									42.8	43.7	43.3	43.6	43.5
R. albus										46.2	50.0	50.9	53.6
R. bromii											47.0	47.5	46.9
R. callidus												55.4	54.7
R. champanellensis													54.7
R. flavefaciens													

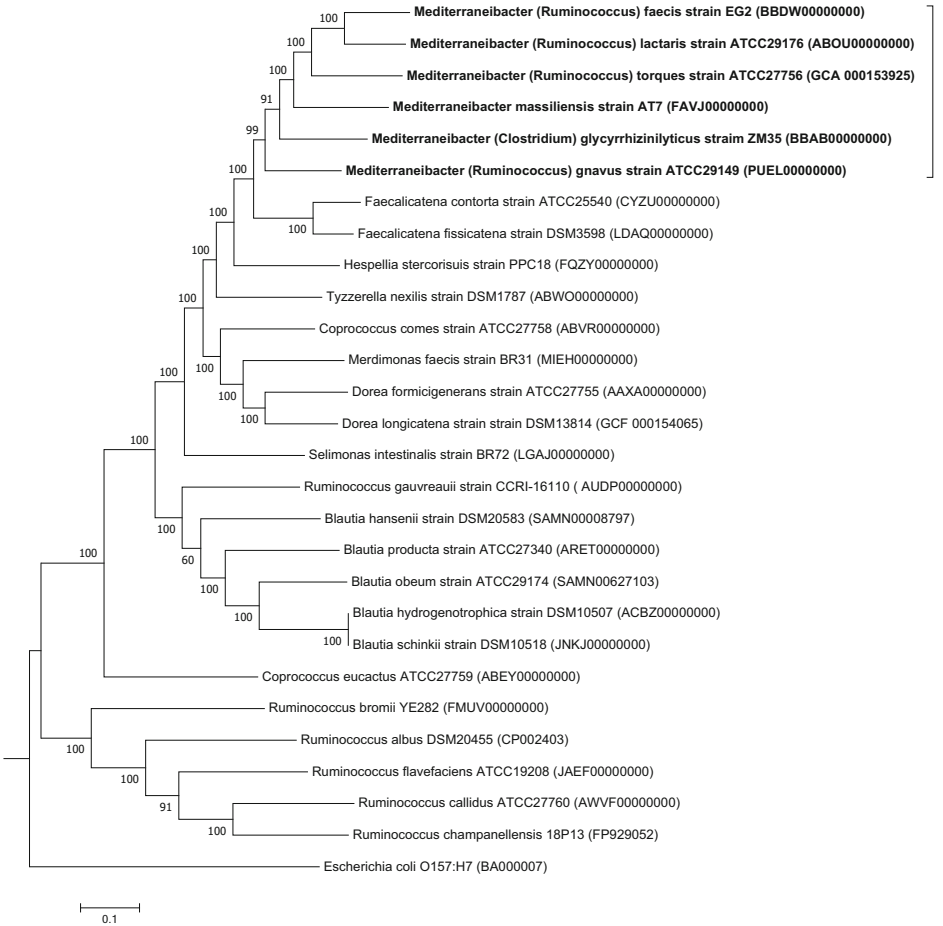


Fig. 4 Phylogenetic tree based on the 271 concatenated orthologous genes from the genomes of 28 related species. All 28 genomes were downloaded from NCBI (www.ncbi.nlm.nih.gov). For orthologous detection, we applied Proteinortho with default values (Lechner et al. 2011). All orthologous genes were aligned using Muscle (Edgar 2004) then concatenated. Phylogenetic reconstruction was performed using maximum likelihood method with the Kimura 2 parameter model and bootstrap value of 100

Lachnospiraceae. *R. gauvreauii* was also recovered as part of the family *Lachnospiraceae* but was not consistently related to the newly identified cluster (Fig. 4). In contrast, *R. bromii*, *R. albus*, *R. champanellensis*, *R. callidus* and *R. flavefaciens*, the type species of the genus *Ruminococcus*, formed a distinct cluster. Based on these observations, 16S gene similarities (Table 3), number of shared orthologous proteins (Table 5), average of genomic identity of orthologous gene sequences (AGIOS—Table 5), and average amino acid identity (AAI—Table 7), we propose a new genus, *Mediterraneibacter*, to include

a new species, *Mediterraneibacter massiliensis*, represented by the type strain AT7^T and to clarify the taxonomy of *Ruminococcus* species by reclassification of most of those species that do not cluster with the type species of the genus *Ruminococcus* in phylogenetic analyses. The phenotypic, chemotaxonomic, 16S similarities and genomic comparisons are shown in Tables 1, 2, 3, 4, 5, 6 and 7.

The 16S gene similarity between strain AT7^T and *R. gnavus* (92%) was lower than the usual threshold of 94% for delineating genera. However, recent findings suggest that using only the 16S rRNA gene similarity is not adequate and that genomic analysis based on shared orthologous genes is much more robust (Fox et al. 1992; Coenye et al. 2005; Konstantinidis and Tiedje 2005; Varghese et al. 2015). Indeed, the phylogenetic tree based on 271 concatenated shared orthologous genes (Fig. 4), the number of shared proteins (Table 5), AGIOS (Table 5) and AAI (Table 7) all confirm that *R. gnavus* should be included in the new genus.

The sequence of the 16S ribosomal RNA gene alone does not allow satisfactory discrimination of the species in the *Lachnospiraceae* family. This is illustrated by the very low bootstrap values (Figure S3). These values are all below 70% for nodes between species of the new genus (accordingly not shown in Fig. S3). In the phylogenetic tree based on 271 shared orthologous genes (Fig. 4), the bootstrap values of the nodes between the species of the new genus are between 91 and 100% and the bootstrap of the node that differentiates the new *Mediterraneibacter* genus and the closely related genus *Faecalicatena* is 100%. This means that the creation of the new genus is based on very robust results (concatenated phylogenetic tree based on 271 shared orthologous genes) whereas the analysis based on the 16S ribosomal gene alone was associated with a very high risk of phylogenetic error.

Based on these findings, we propose to reclassify these four *Ruminococcus* species, namely *R. faecis*, *R. lactaris*, *R. torques* and *R. gnavus* and *C. glycyrrhizinilyticum* within the new genus *Mediterraneibacter* as *Mediterraneibacter faecis* comb. nov., *Mediterraneibacter lactaris* comb. nov., *Mediterraneibacter torques* comb. nov., *Mediterraneibacter gnavus* comb. nov. and *Mediterraneibacter glycyrrhizinilyticus* comb. nov. In addition, we observed that *R. gauvreauii* should probably be reclassified in the *Blautia* genus but further analyses specifically focusing on this genus are necessary.

Discussion and conclusion

Strain AT7^T was considered to represent a new species of the new genus *Mediterraneibacter* based on its MALDI-TOF spectrum (Fig. 1), which could not be identified on our database that contains more than 8000 spectra, 16S rRNA similarity level and genomic characteristics. Comparison of this bacterial species with other closely related species (Table 1) showed that strain AT7^T can be differentiated by its metabolism of mannitol, mannose, salicin and trehalose. The dDDH (Table 6) values are very low when compared to closely related species, using threshold set at 70% according to Meier-Kolthoff et al. (2013a). The genomic comparisons (AGIOS and dDDH) reported in Tables 5 and 6 confirm that the similarities between strain AT7 and closely related species are in accordance with the proposition of a new species. Phenotypic differences, together with phylogenetic and genomic findings, allow us to propose strain AT7^T (= CSUR P2086^T = DSM 100837$^{T)}$ as the type strain of *Mediterraneibacter massiliensis* gen. nov., sp. nov.

This new bacterium is potentially important for human health because it has been isolated from a morbidly obesity patient. It is currently known that some species of the family *Lachnospiraceae* family, namely *R. gnavus*, *Blautia obeum* and *Coprococcus catus* are strongly associated with weight gain, both in humans and in experimental models (Sepp et al. 2013; Petriz et al. 2014; Zietak et al. 2016). More recently, *R. gnavus* was associated with adiposity in a microbiome-wide association study (MWAS) (Beaumont 2016). *R. gnavus* was also associated with obesity in another large-scale metagenomic study (Le Chatelier et al. 2013). This is particularly interesting because, to our knowledge, *R. gnavus* is one of the rare bacteria consistently associated with obesity and/or adiposity. However, based on our comprehensive phylogenetic analyses, *R. gnavus* should be classified in the genus *Mediterraneibacter*. Correcting this classification of this species is important as inaccurate nomenclature could lead researchers to draw erroneous conclusions about the role of the members of the genus *Ruminococcus* sensu stricto with regard to weight and adiposity regulation. Accordingly, the reclassification

of *R. gnavus* will help prevent confusion and will help studies analysing relationships between obesity and the gut microbiota.

In addition, we investigated the presence of 16S rRNA from strain AT7^T in the high throughput DNA and RNA sequence read archive (SRA) using an online open resource (Lagkouvardos et al. 2016). We found metagenomic sequences with a similarity greater than 97% with strain AT7^T in several gut metagenomes (human, bovine, chicken, mouse, rat, pig, primate and insect), skin (mouse, human) metagenomes, human oral metagenome, human lung metagenome, vaginal metagenome, food metagenome, as well as in environmental samples (wastewater, groundwater, seawater, marine sediment, bioreactor, hydrothermal vent, sludge, soil and insect). Metagenomic sequences corresponding to strain AT7^T were found in 7.9% (10844/135936) of all metagenomes and 30.7% (6191/20156) of the human gut metagenomes present in this database. Accordingly, the bacterium described here is found in the human mature anaerobic gut microbiota (HMAGM) (Million et al. 2017), consistent with its isolation from the stool sample of a 37-year-old French woman living in Marseille, who suffered from morbid obesity.

The Digital Protologue TaxoNumbers (http://imedea.uib-csic.es/dprotologue/index.php) of *M. massiliensis* gen. nov., sp. nov., *M. faecis* comb. nov., *M. lactaris* comb. nov., *M. torques* comb. nov., *M. gnavus* comb. nov. and *M. glycyrrhizinilyticus* comb. nov. are GA00061/TA00494, TA00495, TA00496, TA00497, TA00498 and TA00499, respectively.

Description of *Mediterraneibacter* gen. nov.

Mediterraneibacter (Me.di.ter.ra.ne.i.bac'ter. L. neut. n. *mediterraneum mare*, the Mediterranean sea; N.L. masc. n. *bacter* a rod; N.L. masc. n. *Mediterraneibacter* a rod from the Mediterranean Sea).

Gram-stain positive, asporogenous, non-motile, coccoid or coccobacillary-shaped, catalase positive and obligately anaerobic. The major end products of carbohydrate metabolism are acetic acid, formic acid and lactic acid. The DNA G+C content of the ranges from 42 to 45 mol %. The type species of the genus is *Mediterraneibacter massiliensis*, which was isolated from human faeces.

Description of *Mediterraneibacter massiliensis* sp. nov.

Mediterraneibacter massiliensis (mas.si.li.en'sis. L. masc. adj. *massiliensis*, of Massilia, the Latin name for Marseille).

In addition to the characteristics in the genus description, cells are coccobacillary-shaped, with a width ranging from 0.2 to 0.4 μm and a length ranging from 1 to 1.4 μm. Colonies are translucent with a diameter of 0.5–1 mm on 5% sheep blood Columbia agar. Oxidase negative. Optimum growth temperature is 37 °C under anaerobic conditions and pH tolerance ranges from 6.5 to 8.5. The major fatty acids are $C_{16:0}$ and $C_{18:1n9}$. The major end product of carbohydrate metabolism also include isohexanoic acid and isobutanoic acid. The draft genome of the type strain is 3,069,882 bp long with a DNA G+C content of 42.4%.

The type strain AT7^T has been deposited in the CSUR and DSM collections under numbers CSUR P2086 and DSM 100837, respectively. The type strain was isolated from the stool sample of a 37-year-old obese French woman. The draft genome and 16S rRNA sequences of the type strain have been deposited in EMBL-EBI under accession numbers FAVJ00000000 and LN881607, respectively.

Description of *Mediterraneibacter faecis* comb. nov.

Mediterraneibacter faecis (fae'cis. L. gen. n. *faecis*, of faeces, referring to its faecal origin).

Basonym: *Ruminococcus faecis* Kim et al. 2011.

The description of *Mediterraneibacter faecis* is the same as that given for *Ruminococcus faecis* (Kim et al. 2011). The type strain is Eg2^T (= KCTC 5757^T = JCM 15917^T).

Description of *Mediterraneibacter lactaris* comb. nov.

Mediterraneibacter lactaris (lac.ta'ris. L. masc. adj. *lactaris* milk-drinking [referring to its rapid fermentation of lactose and curding of milk]).

Basonym: *Ruminococcus lactaris* (Moore et al. 1976) Approved Lists 1980.

The description of *Mediterraneibacter lactaris* is the same as given for *Ruminococcus lactaris* (Moore et al. 1976). The type strain is ATCC 29176^T (= VPI X6-29^T).

Description of *Mediterraneibacter torques* comb. nov.

Mediterraneibacter torques (tor'ques. L. n. *torques* twisted necklace [referring to appearance of the chains from broth cultures]).

Basonym: *Ruminococcus torques* (Holdeman and Moore 1974) Approved Lists 1980.

The description of *Mediterraneibacter torques* is the same as given for *Ruminococcus torques* (Holdeman and Moore 1974). The type strain is ATCC 27756^T (= VPI B2-51^T).

Description of *Mediterraneibacter gnavus* comb. nov.

Mediterraneibacter gnavus (gna'vus. L. masc. adj. *gnavus* busy, active [referring to the active fermentative ability of this species]).

Basonym: *Ruminococcus gnavus* (Moore et al. 1976) Approved Lists 1980.

The description of *Mediterraneibacter gnavus* is the same as given for *Ruminococcus gnavus* (Moore et al. 1976). The type strain is ATCC 29149 (= VPI C7-9).

Description of *Mediterraneibacter glycyrrhizinilyticus* comb. nov.

Mediterraneibacter glycyrrhizinilyticus (gly.cy.rrhi.zi.ni.ly'ti.cus. N.L. neut. n. *glycyrrhizinum* glycyrrhizin [a sugar from the roots of Glycyrrhiza species], N.L. masc. adj. *lyticus* dissolving, able to dissolve, N.L. masc. adj. *glycyrrhizinilyticus* glycyrrhizin dissolving).

Basonym: *Clostridium glycyrrhizinilyticum* Sakuma et al. 2006.

The description of *Mediterraneibacter glycyrrhizinilyticus* is the same as given for *Clostridium glycyrrhizinilyticum* (Sakuma et al. 2006). The type strain is strain ZM35^T (= JCM 13368^T = DSM 17593^T).

Acknowledgements The authors thank the Xegen Company (www.xegen.fr) for automating the genomic annotation process and Magdalen LARDIERE for English correction.

Author contributions AHT isolated the bacterium, performed the phenotypic characterization, drafted the manuscript; AD performed the genomic analyses and drafted manuscript. FB and P-EF helped in data interpretation, drafted the manuscript and reference checking, MM and RV take care of the patient and provide samples; NA, GD, NL and MR performed genome sequencing and chemotaxonomic analysis; JD, AL performed comprehensive genomic analysis; DR designed and directed the project; MM drafted manuscript, checked the references and acted as corresponding author.

Funding This study was funded by the « Fondation Méditerranée Infection» and the French National Research Agency under the program "Investissements d'avenir" with the reference ANR-10-IAHU-03.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

References

Auch AF, von Jan M, Klenk H-P, Göker M (2010) Digital DNA–DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2:117–134

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol 19:455–477

Beaumont M (2016) Genetic and environmental factors affecting the human gut microbiom in obesity. Student thesis. Doctoral thesis, Doctor of Philosophy, King's College

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization. Bioinform Oxf Engl 25:119–120

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinform Oxf Engl 28:464–469

Chassard C, Delmas E, Robert C, Lawson PA, Bernalier-Donadille A (2012) *Ruminococcus champanellensis* sp. nov., a cellulose-degrading bacterium from human gut microbiota. Int J Syst Evol Microbiol 62:138–143

Citron DM, Ostovari MI, Karlsson A, Goldstein EJ (1991) Evaluation of the E test for susceptibility testing of anaerobic bacteria. J Clin Microbiol 29:2197–2203

Coenye T, Gevers D, Van de Peer Y, Vandamme P, Swings J (2005) Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev 29:147–167

Conway KR, Boddy CN (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res 41:D402–D407

Dione N, Sankar SA, Lagier J-C, Khelaifia S, Michele C, Armstrong N, Richez M, Abrahão J, Raoult D, Fournier P-E (2016) Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. New Microbes New Infect 10:66–76

Domingo M-C, Huletsky A, Boissinot M, Bernard KA, Picard FJ, Bergeron MG (2008) *Ruminococcus gauvreauii* sp. nov., a glycopeptide-resistant species isolated from a human faecal specimen. Int J Syst Evol Microbiol 58:1393–1397

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Fournier P-E, Lagier J-C, Dubourg G, Raoult D (2015) From culturomics to taxonomogenomics: a need to change the taxonomy of prokaryotes in clinical microbiology. Anaerobe 36:73–78

Fox GE, Wisotzkey JD, Jurtshuk P (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166–170

Gouret P, Thompson JD, Pontarotti P (2009) PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinform 10:298

Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain J-M (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 58:212–220

Holdeman LV, Moore WEC (1974) New genus, *Coprococcus*, twelve new species, and emended descriptions of four previously described species of bacteria from human feces. Int J Syst Evol Microbiol 24:260–277

Hungate RE (1957) Microorganisms in the rumen of cattle fed a constant ration. Can J Microbiol 3:289–311

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 11:119

Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664

Kim M-S, Roh SW, Bae J-W (2011) *Ruminococcus faecis* sp. nov., isolated from human faeces. J Microbiol Seoul Korea 49:487–491

Kim M, Oh H-S, Park S-C, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351

Kong L-C, Tap J, Aron-Wisnewsky J, Pelloux V, Basdevant A, Bouillot J-L, Zucker J-D, Dore J, Clement K (2013) Gut microbiota after gastric bypass in human obesity: increased richness and associations of bacterial genera with adipose tissue genes. Am J Clin Nutr 98:16–24

Konstantinidis KT, Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. J Bacteriol 187:6258–6264

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874

Lagier J-C, Armougom F, Million M, Hugon P, Pagnier I, Robert C, Bittar F, Fournous G, Gimenez G, Maraninchi M, Trape J-F, Koonin EV, La Scola B, Raoult D (2012) Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 18:1185–1193

Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, Caputo A, Cadoret F, Traore SI, Seck EH, Dubourg G, Durand G, Mourembou G, Guilhot E, Togo A, Bellali S, Bachar D, Cassir N, Bittar F, Delerce J, Mailhe M, Ricaboni D, Bilen M, Dangui Nieko NPM, Dia Badiane NM, Valles C, Mouelhi D, Diop K, Million M, Musso D, Abrahão J, Azhar EI, Bibi F, Yasir M, Diallo A, Sokhna C, Djossou F, Vitton V, Robert C, Rolain JM, La Scola B, Fournier P-E, Levasseur A, Raoult D (2016) Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol 1:16203

Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, Clavel T (2016) IMNGS: a comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci Rep 6:33721

Lawson PA, Finegold SM (2015) Reclassification of *Ruminococcus obeum* as *Blautia obeum* comb. nov. Int J Syst Evol Microbiol 65:789–793

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clément K, Doré J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker J-D, Raes J, Hansen T, Bork P, Wang J, Ehrlich SD, Pedersen O, Guedon E, Delorme C, Layec S, Khaci G, van de Guchte M, Vandemeulebrouck G, Jamet A, Dervyn R, Sanchez N, Maguin E, Haimet F, Winogradski Y, Cultrone A, Leclerc M, Juste C, Blottière H, Pelletier E, LePaslier D, Artiguenave F, Bruls T, Weissenbach J, Turner K, Parkhill J, Antolin M, Manichanh C, Casellas F, Boruel N, Varela E, Torrejon A, Guarner F, Denariaz G, Derrien M, van Hylckama Vlieg J E T, Veiga P, Oozeer R, Knol J, Rescigno M, Brechot C, M'Rini C, Mérieux A, Yamada T (2013) Richness of human gut microbiome correlates with metabolic markers. Nature 500:541–546

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ (2011) Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinform 12:124

Liu JR (2002) Emended description of the genus *Trichococcus*, description of *Trichococcus collinsii* sp. nov., and reclassification of *Lactosphaera pasteurii* as *Trichococcus pasteurii* comb. nov. and of *Ruminococcus palustris* as *Trichococcus palustris* comb. nov. in the low-G+C Gram-positive bacteria. Int J Syst Evol Microbiol 52:1113–1126

Liu C, Finegold SM, Song Y, Lawson PA (2008) Reclassification of *Clostridium coccoides*, *Ruminococcus hansenii*, *Ruminococcus hydrogenotrophicus*, *Ruminococcus luti*, *Ruminococcus productus* and *Ruminococcus schinkii* as *Blautia coccoides* gen. nov., comb. nov., *Blautia hansenii* comb. nov., *Blautia hydrogenotrophica* comb. nov., *Blautia luti* comb. nov., *Blautia producta* comb. nov., *Blautia schinkii* comb. nov. and description of *Blautia wexlerae* sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 58:1896–1902

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Matta J, Zins M, Feral-Pierssens AL, Carette C, Ozguler A, Golberg M, Czernichow S (2016) Prévalence du surpoids, de l'obésité et des facteurs de risque cardio-métaboliques dans la cohorte Constances. Bull Epidémiol Hebd 35–36:640–646

Matuschek E, Brown DFJ, Kahlmeter G (2014) Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. Clin Microbiol Infect 20:O255–O266

Meier-Kolthoff JP, Göker M, Spröer C, Klenk H-P (2013a) When should a DDH experiment be mandatory in microbial taxonomy? Arch Microbiol 195:413–418

Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M (2013b) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinform 14:60

Meier-Kolthoff JP, Klenk H-P, Göker M (2014) Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. Int J Syst Evol Microbiol 64:352–356

Million M, Diallo A, Raoult D (2017) Gut microbiota and malnutrition. Microb Pathog 106:127–138

Moore WEC, Cato EP, Holdeman LV (1972) *Ruminococcus bromii* sp. n. and emendation of the description of *Ruminococcus Sijpestein*. Int J Syst Bacteriol 22:78–80

Moore ERB, Johnson JL, Holdeman LV (1976) Emendation of *Bacteroidaceae* and *Butyrivibrio* and descriptions of *Desulfomonas* gen. nov. and ten new species in the genera *Desulfomonas*, *Butyrivibrio*, *Eubacterium*, *Clostridium*, and *Ruminococcus*. Int J Syst Evol Microbiol 26:238–252

Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, Mullany EC, Biryukov S, Abbafati C, Abera SF, Abraham JP, Abu-Rmeileh NME, Achoki T, AlBuhairan FS, Alemu ZA, Alfonso R, Ali MK, Ali R, Guzman NA, Ammar W, Anwari P, Banerjee A, Barquera S, Basu S, Bennett DA, Bhutta Z, Blore J, Cabral N, Nonato IC, Chang J-C, Chowdhury R, Courville KJ, Criqui MH, Cundiff DK, Dabhadkar KC, Dandona L, Davis A, Dayama A, Dharmaratne SD, Ding EL, Durrani AM, Esteghamati A, Farzadfar F, Fay DFJ, Feigin VL, Flaxman A, Forouzanfar MH, Goto A, Green MA, Gupta R, Hafezi-Nejad N, Hankey GJ, Harewood HC, Havmoeller R, Hay S, Hernandez L, Husseini A, Idrisov BT, Ikeda N, Islami F, Jahangir E, Jassal SK, Jee SH, Jeffreys M, Jonas JB, Kabagambe EK, Khalifa SEAH, Kengne AP, Khader YS, Khang Y-H, Kim D, Kimokoti RW, Kinge JM, Kokubo Y, Kosen S, Kwan G, Lai T, Leinsalu M, Li Y, Liang X, Liu S, Logroscino G, Lotufo PA, Lu Y, Ma J, Mainoo NK, Mensah GA, Merriman TR, Mokdad AH, Moschandreas J, Naghavi M, Naheed A, Nand D, Narayan KMV, Nelson EL, Neuhouser ML, Nisar MI, Ohkubo T, Oti SO, Pedroza A, Prabhakaran D, Roy N, Sampson U, Seo H, Sepanlou SG, Shibuya K, Shiri R, Shiue I, Singh GM, Singh JA, Skirbekk V, Stapelberg NJC, Sturua L, Sykes BL, Tobias M, Tran BX, Trasande L, Toyoshima H, van de Vijver S, Vasankari TJ, Veerman JL, Velasquez-Melendez G, Vlassov VV, Vollset SE, Vos T, Wang C, Wang X, Weiderpass E, Werdecker A, Wright JL, Yang YC, Yatsuya H, Yoon J, Yoon S-J, Zhao Y, Zhou M, Zhu S, Lopez AD, Murray CJL, Gakidou E (2014) Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 384:766–781

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 42:D206–D214

Petriz BA, Castro AP, Almeida JA, Gomes CP, Fernandes GR, Kruger RH, Pereira RW, Franco OL (2014) Exercise induction of gut microbiota modifications in obese, non-obese and hypertensive rats. BMC Genom 15:511

Rainey FA (2010) Family VIII. *Ruminococcaceae* fam. nov. In: De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB (eds) Bergey's manual of systematic bacteriology, 2nd edn. Springer, New York

Rainey FA, Janssen PH (1995) Phylogenetic analysis by 16S ribosomal DNA sequence comparison reveals two unrelated groups of species within the genus *Ruminococcus*. FEMS Microbiol Lett 129:69–73

Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, Sentausa E, Raoult D, Fournier P-E (2014) A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 64:384–391

Rodriguez-R LM, Konstantinidis KT (2014) Bypassing cultivation to identify bacterial species: culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. Microbe Mag 9:111–118

Rosselló-Mora R (2006) DNA-DNA reassociation methods applied to microbial taxonomy and their critical evaluation. In: Molecular identification, systematics, and population structure of prokaryotes. Springer, pp 23–50

Sakuma K, Kitahara M, Kibe R, Sakamoto M, Benno Y (2006) *Clostridium glycyrrhizinilyticum* sp. nov., a glycyrrhizin-hydrolysing bacterium isolated from human faeces. Microbiol Immunol 50:481–485

Sasser M (2006) Bacterial identification by gas chromatographic analysis of fatty acids methyl esters (GC-FAME)

Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P-E, Rolain JM, Raoult D (2009) Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 49:543–551

Sepp E, Lõivukene K, Julge K, Voor T, Mikelsaar M (2013) The association of gut microbiota with body weight and body

mass index in preschool children of Estonia. Microb Ecol Health Dis 24:19231
Sijpesteijn AK (1949) Cellulose-decomposing bacteria from the rumen of cattle. Antonie Van Leeuwenhoek 15:49–52
Stackebrant E, Ebers J (2006) Taxonomic parameters revisited: tarnished gold standards. Microbiol Today 33:152–155
Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci USA 101:11030–11035
Togo AH, Durand G, Khelaifia S, Armstrong N, Robert C, Cadoret F, Di Pinto F, Delerce J, Levasseur A, Raoult D, Million M (2017) *Fournierella massiliensis*, gen. nov., sp. nov., a new human-associated member of the family *Ruminococcaceae*. Int J Syst Evol Microbiol 67:1393–1399
Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A (2015) Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771
Willems A, Collins MD (1995) Phylogenetic analysis of *Ruminococcus flavefaciens*, the type species of the genus *Ruminococcus*, does not support the reclassification of *Streptococcus hansenii* and *Peptostreptococcus productus* as ruminococci. Int J Syst Bacteriol 45:572–575
Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635–645
Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, Parameswaran P, Crowell MD, Wing R, Rittmann BE, Krajmalnik-Brown R (2009) Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci USA 106:2365–2370
Zhao G, Nyman M, Jönsson JA (2006) Rapid determination of short-chain fatty acids in colonic contents and faeces of humans and rats by acidified water-extraction and direct-injection gas chromatography. Biomed Chromatogr BMC 20:674–682
Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS (2011) PHAST: a fast phage search tool. Nucleic Acids Res 39:347–352
Ziętak M, Kovatcheva-Datchary P, Markiewicz LH, Ståhlman M, Kozak LP, Bäckhed F (2016) Altered microbiota contributes to reduced diet-induced obesity upon cold exposure. Cell Metab 23:1216–1223

Article 17:

Draft genome and description of *Eisenbergiella massiliensis* strain AT11^{T}: a new species isolated from human faeces after bariatric surgery

Togo AH, Diop A, Million M, Maraninchi M, Lagier JC, Robert C, Di Pinto F, Raoult D, Fournier PE, Bittar F

[Published in Current Microbiology]

Current Microbiology
https://doi.org/10.1007/s00284-018-1520-2

Draft Genome and Description of *Eisenbergiella massiliensis* Strain AT11^T: A New Species Isolated from Human Feces After Bariatric Surgery

Amadou H. Togo[1] · Awa Diop[2] · Matthieu Million[1] · Marie Maraninchi[3] · Jean-Christophe Lagier[1] · Catherine Robert[2] · Fabrizio Di Pinto[1] · Didier Raoult[1] · Pierre-Edouard Fournier[2] · Fadi Bittar[1,4]

Received: 25 January 2018 / Accepted: 29 May 2018

Abstract
A novel strain of a Gram-stain negative, non-motile, non-spore forming rod-shaped, obligate anaerobic bacterium, designated AT11^T, was isolated from a stool sample of a morbidly obese woman living in Marseille, France. This bacterium was characterized using biochemical, chemotaxonomic, and phylogenetic methods. The 16S rRNA gene sequence analysis showed that strain AT11^T had a 97.8% nucleotide sequence similarity with *Eisenbergiella tayi* strain B086562^T, the closest species with standing in nomenclature. The major cellular fatty acids of the novel isolate were $C_{16:0}$ followed by saturated or unsaturated C_{18} fatty acids ($C_{18:1}$n9, $C_{18:1}$n5 and $C_{18:0}$). The draft genome of strain AT11^T is 7,114,554 bp long with 48% G+C content. 6176 genes were predicted, including 6114 protein-coding genes and 62 were RNAs (with 2 5S rRNA genes, two 16S rRNA genes, two 23S rRNA genes, and 56 tRNA genes). The digital DNA–DNA hybridization (dDDH) relatedness between the new isolate and *E. tayi* strain B086562^T was 23.1% ± 2.2. Based on the phenotypic, chemotaxonomic, genomic, and phylogenetic characteristics, *Eisenbergiella massiliensis* sp. nov., is proposed. The type strain is AT11^T (= DSM 100838^T = CSUR P2478^T).

Introduction

The number of people suffering from obesity has increased in recent decades [25]. It has been well established that the gut microbiota contributes to the development of human metabolic disorders such as obesity [18, 24]. Bariatric surgery is the most effective treatment for morbid obesity. It induces a sustainable weight loss, improves complications related to obesity, and increases the diversity of the gut flora [14, 34].

We conducted a study comparing the gut microbiota from obese patients before and after bariatric surgery using a new microbial high-throughput culture approach known as culturomics [16]. This new approach makes it possible to isolate and describe the living microbial diversity of any environmental and clinical sample. Using culturomics, we isolated a new anaerobic bacterium, strain AT11^T, from a stool sample harvested following bariatric surgery. The discovery of this bacterium has been previously reported as a new species announcement without a thorough description [31].

Herein, strain AT11^T was analyzed by a polyphasic approach in order to describe it as a new bacterial taxon. This combines phenotypic characteristics, the matrix laser desorption ionization-time of flight mass spectrometry (MALDI-TO MS) spectrum, and genomic properties known as taxono-genomics [27].

Here, we propose a classification and a set of phenotypic, chemical, and chemotaxonomic characteristics of a new bacterial species: strain AT11^T, which belongs to

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00284-018-1520-2) contains supplementary material, which is available to authorized users.

✉ Fadi Bittar
fadi.bittar@univ-amu.fr

[1] Aix Marseille Univ, IRD, APHM, MEPHI, IHU-Méditerranée Infection, Marseille, France

[2] Aix Marseille Univ, IRD, APHM, SSA, VITROME, IHU-Méditerranée Infection, Marseille, France

[3] Aix Marseille Univ, NORT "Nutrition, Obesity and Risk of Thrombosis", INSERM1062, INRA1260, 13385 Marseille, France

[4] IHU-Méditerranée Infection, 19-21 Bd Jean Moulin, 13005 Marseille, France

Published online: 01 June 2018

the genus *Eisenbergiella* [1], together with the description of the complete genome sequencing, annotation, and genomic comparison. To date, this genus includes only one species *Eisenbergiella tayi*, the type strain B086562^T (= LMG 27400^T = DSM 26961^T = ATCC BAA-2558^T) as reported in List of Prokaryotic Names with Standing in Nomenclature (http://www.bacterio.net/ruminococcus.html).

Materials and Methods

Ethics and Sample Collection

Once informed consent had been obtained, stool samples were collected before and after surgery. These samples were obtained from a 56-year-old obese French woman following bariatric surgery on April 27, 2011. All samples were stored at − 80 °C before culturing. The study and the assent procedure were approved by the local ethics committee of IFR 48, under ascent number 09-022, 2010.

Isolation and Identification of the Strain

Strain AT11^T was first grown on July 22, 2015. One gram of stool was pre-incubated in BD BACTEC™ Lytic/10 Anaerobic/F Culture Vials media culture bottles (Becton, Dickenson and Company, Le Pont de Claix, France) enriched with 4 ml of filtered rumen juice and 4 ml of sheep blood. The pre-incubated product was cultured on 5% sheep blood-enriched Columbia agar (bioMérieux, Marcy l'Etoile, France) as described elsewhere [32]. This strain was isolated 21 days after pre-incubation. The resulting colonies were then identified using MALDI-TOF mass spectrometry (Bruker Daltonics, Leipzig, Germany) as previously described [29]. When the spectra of a bacterium are not identified by MALDI-TOF MS screening, 16S rRNA gene amplification and sequencing is performed.

Phylogenetic Analysis

The 16S rRNA gene amplification PCR and sequencing were performed using GeneAmp PCR System 2720 thermal cyclers (Applied Bio systems, Bedford, MA, USA) and ABI Prism 3130xl Genetic Analyzer capillary sequencer (Applied Bio systems), respectively, as described by Drancourt et al. [6]. The CodonCode Aligner was used to correct sequences and BLASTn searches were performed on the NCBI (National Centre for Biotechnology Information) web server at http://blast.ncbi.nlm.nih.gov.gate1.inist.fr/Blast.cgi for the taxonomic assignation. Pairwise sequence similarities were calculated using the method recommended by Meier-Kolthoff et al. [23] and as described previously [33]. Sequences were aligned using ClustalW with default parameters and phylogenies were inferred using the GGDC web server available at http://ggdc.dsmz.de/ using the DSMZ phylogenomics pipeline.

Phenotypic, Biochemical, and Chemotaxonomic Characterization

Different growth temperatures (room temperature, 28, 37, 45, and 55 °C) were tested on sheep blood-enriched Columbia agar (bioMérieux) under anaerobic conditions using GENbag anaer system (bioMérieux), microaerophilic conditions using GENbag microaer system (bioMérieux), and aerobic conditions, with or without 5% CO_2.

Phenotypic and biochemical characteristics were performed as described elsewhere [32]. In addition to the three API gallery systems (API® ZYM, API® 20A, and API® 50 CH) usually used in our laboratory, API® Rapid ID 32A gallery system was added and the tests were done according to the manufacturer's instructions (bioMérieux).

E test strips for Amikacin 0.016–256 µg/ml, Vancomycin 0.016–256 µg/ml, Imipenem 0.002–32 µg/ml, Ceftriaxone 0.016–256 µg/ml, Rifampicin 0.002–32 µg/ml, Benzyl penicillin 0.002–32 µg/ml, Amoxicillin 0.016–256 µg/ml, Minocycline 0.016–256 µg/ml, Teicoplanin 0.016–256 µg/ml, Erythromycin 0.016–256 µg/ml, and Daptomycin 0.016–256 µg/ml (bioMérieux) were used for the antimicrobial agent susceptibility of strain AT11^T as recommended by EUCAST [4, 22]. Breakpoint tables for the interpretation of MICs and inhibition zone diameters, version 7.1, 2017, were used to interpret the results: these are available at http://www.eucast.org.

Cellular fatty acid methyl ester (FAME) analysis of this was then performed using gas chromatography/mass spectrometry (GC/MS) as described by Dione et al. [5].

Genome Sequencing and Assembling

The genomic DNA of strain AT11^T was sequenced and assembled as described in previous studies [33]. It was quantified by a Qubit assay using the high sensitivity kit (Life Technologies, Carlsbad, CA, USA) to 107.7 ng/µl and mechanically sheared with a circular shear to small fragments with an optimal length of 1401 bp using the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). A High Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) was used to visualize the library profile and the final concentration library was measured at 34.4 nmol/l. The libraries were then normalized

and pooled at 2 nM. After a denaturation step and dilution at 15 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated cluster generation and a sequencing run were performed in a single 2 × 251-bp run. A total of 5.6 Gb of information was obtained from the 589 K/mm^2 cluster density with a cluster passing quality control filters of 96.1% (11,444,000 passing filter paired reads). Within this run, the index representation for strain AT11^T was determined to 6.46%. The 697,439 paired reads were trimmed and assembled.

Genome Annotation and Comparison

Open reading frames (ORFs) were predicted using Prodigal [10] with default parameters but the predicted ORFs were excluded if they were spanning a sequencing gap region (contain N). The predicted bacterial protein sequences were searched again the Clusters of Orthologous Groups (COG) using BLASTP with an *E* value of 1e^{-03}, a coverage of 0.7, and a percent identity of 30%. If no hit was found, a search was conducted against the Nucleotide Redundant (NR) database using the same parameters. If the length of sequence was smaller than 80 amino acids, a 1e^{-05} *E* value was used. The tRNAScanSE tool [21] was used to find tRNA genes, while ribosomal RNAs were found by using RNAmmer [15]. Lipoprotein signal peptides and the number of transmembrane helices was predicted using Phobius [11]. ORFans were identified if all the BLASTP performed gave no positive results with an *E* value smaller than 1e^{-03} for ORFs with a sequence size larger than 80 amino acid or an *E* value smaller than 1e^{-05} for ORFs with a sequence length smaller than 80 amino acids. Paralog genes were defined by blasting each protein gene against all protein genes of this genome. For pseudogenes, the first step was to define the closed species genomes used for comparison analysis. Then, the potential missing genes in the genomes of interest were identified. All processes of annotation and comparison were performed using the Multi-Agent Software System DAGOBAH [8] that includes Figenix [9]. Genomic similarity was evaluated via digital DNA–DNA hybridization (dDDH) using the Genome to Genome Distance Calculator (GGDC) 2.1-DSM web service (http://ggdc.dsmz.de/ggdc.php).

The species used for genomic comparison were retrieved from the 16S rDNA gene tree. The following strains were selected: *Blautia producta* strain ATCC 27340^T (ARET00000000) [7, 19, 28], *Eisenbergiella tayi* strain B086562^T (MCGH00000000) [1, 2], *Anaerostipes hadrus* strain DSM 3319^T (AMEY00000000) [13], *Parasporobacterium paucivorans* strain DSM 15970^T [20], *Eubacterium ruminantium* strain ATCC 17233^T (GCA900167085) [3], *Clostridium bolteae* strain WAL 16351^T (AGYH00000000) [30], and *Clostridium clostridioforme* strain ATCC 25,537 (GCA900113155) [12]. For each selected strain, the complete genome sequence was retrieved from the FTP of NCBI (National Center for Biotechnology Information). The proteome was analyzed using proteinOrtho [17]. For each couple of genomes, a similarity score was then computed.

Accession Numbers

The 16S rRNA gene sequence and whole-genome shotgun sequence of strain AT11^T were deposited in EMBL-EBI under accession numbers LN881600 and OEZA00000000, respectively. The Digital Protologue database TaxonNumber for strain AT11^T is TA00401.

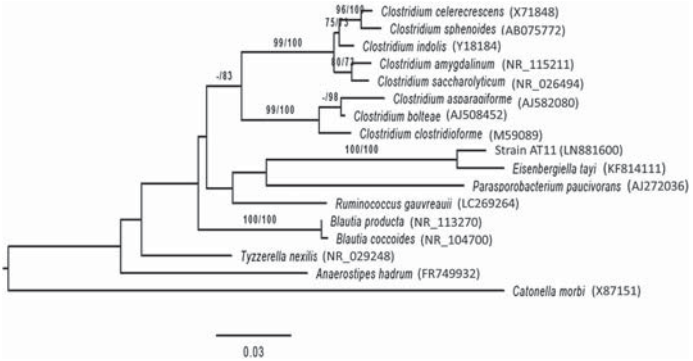


Fig. 1 Phylogenetic tree based on 16S rRNA sequence comparison highlighting the position of strain AT11^T against other most closely related type strains. The scale bar represents a 2% nucleotide sequence divergence

Results and Discussion

Phylogenetic Analysis

The spectrum generated from strain AT11^T spots did not match those of Bruker and our in-house database (Supplementary Fig. 1) available at http://www.mediterranee-infection.com/article.php?laref=933&titre=c-d-e. This new strain exhibited 97.76% nucleotide sequence similarity with *Eisenbergiella tayi*, the closest species with standing in nomenclature according to the 16S rDNA sequence analysis. Figure 1 presents the neighbor-joining phylogenetic tree (Fig. 1) based on 16S rRNA gene sequences and shows the relationships between strain AT11^T and some related taxa. This sequence of the strain was deposited in EMBL-EBI under accession number LN881600.

Phenotypic and Biochemical Characterization

Strain AT11^T is strictly anaerobic, its growth temperature was between 28 and 45 °C, and optimal growth was observed at 37 °C. Colonies appeared light gray in color and exhibited an irregular form with a diameter between 0.5 and 1.5 mm after 72 h of culture on Columbia agar with 5% sheep blood (bioMérieux). No growth was observed above 5 g/l (10–100 g/l) salt on Schaedler agar with 5% sheep blood (bioMérieux). Cells were Gram-negative, non-motile, non-spore-forming, catalase positive, and rod shaped, measuring 1–3 µm in length and 0.4–0.5 µm wide using electron microscopy (Supplementary Fig. 2). The negativity of Gram staining was confirmed by the positive KOH test, but the strain had a positive Gram structure in electron microscopy. The characteristics of strain AT11^T, according to API® gallery systems (50 CH, 20A, Zym and Rapid ID 32A), along with those of the closest species, *Eisenbergiella tayi* strain B086562^T, are listed in Supplementary Table 1 and the differences between these two species are presented in Table 1.

Hexadecanoic acid was the most abundant fatty acid (63%), followed by saturated and unsaturated C_{18} fatty acids representing approximately (33%) of total relative abundance. The fatty acid profiles of strain AT11^T and the closest strain *E. tayi* B086562^T are shown in Table 2.

Antimicrobial agent susceptibility was tested according to the EUCAST recommendations leading to the following MIC results: 32, 0.5, 0.125, 1.6, 0.064, 0.38, 0.5, 0.5, and 0.125 µg/ml, respectively, for Amikacin, Vancomycin, Imipenem, Ceftriaxone, Rifampicin, Benzyl penicillin, Amoxicillin, Minocycline, and Teicoplanin.

Table 1 Differential characteristic of strain AT11^T with *Eisenbergiella tayi* B086562^T

Properties	*Eisenbergiella massiliensis* AT11^T	*Eisenbergiella tayi* B086562^{Ta}
Indole production	V	−
Arabinose	+	−
Arbutin	+	−
Cellulose	+	−
Dulcitol	+	−
Gelatin	+	−
Glucose	+	−
Lactose	+	−
Maltose	V	−
Mannitol	V	−
Mannose	+	−
Raffinose	+	−
Rhamnose	+	−
Saccharose	+	−
Salicin	+	−
Sorbose	+	−
Tagatose	+	−
Trehalose	+	−
Trypsin	+	−
Xylose	V	−
Potassium 5-cetogluconate	+	−
Acid phosphatase	+	−
Alkaline phosphatase	V	+
Arginine hydrolase	+	−
Esterase	+	−
Esterase lipase	+	−
Naphthol-AS-BI-phosphohydrolase	+	−
Tyrosine arylamidase	−	+
α-Arabinosidase	−	+
α-Fructosidase	+	−
β-Glucuronidase	+	−
Isolated from	Human feces	Blood

+ Positive, − negative

v Variable

[a]Data for *E. tayi* were obtained from Amir et al. [25]

Genome Properties

The genome deposited in EMBL-EBI under accession number OEZA00000000 (Fig. 2) is 7,114,554 bp long with 48% GC content. It is composed of 19 contigs consisting of 17 scaffolds. Of the 6176 predicted genes, 6114 were protein-coding genes and 62 were RNAs (two 5S rRNA genes, two 16S rRNA genes, two 23S rRNA genes, 56 tRNA genes). A total of 4321 genes (70.67%) were assigned a putative

Table 2 Cellular fatty acid composition (%) of strain AT11^T compared to its closest neighbor *Eisenbergiella tayi* strain B086562^T

Fatty acid	Name	Strain AT11	*E. tayi*[a]
$C_{16:0}$	Hexadecanoic acid	62.7	45.4
$C_{18:1n9}$	9-Octadecenoic acid	10.3	14.8
$C_{18:1n5}$	13-Octadecenoic acid	9.2	ND
$C_{18:0}$	Octadecanoic acid	7.8	12.8
$C_{18:2n6}$	9,12-Octadecadienoic acid	4.4	1.3
$C_{18:1n7}$	11-Octadecenoic acid	1.6	3.2
$C_{17:0}$	Heptadecanoic acid	1.5	ND
$C_{15:0}$	Pentadecanoic acid	< 1	< 1
$C_{14:0}$	Tetradecanoic acid	< 1	6.3
$C_{16:1n7}$	9-Hexadecenoic acid	< 1	ND
$C_{20:4n6}$	5,8,11,14-Eicosatetraenoic acid	< 1	ND
9,10-Methylene-$C_{16:0}$	2-Hexyl-cyclopropaneoctanoic acid	< 1	ND
Iso-$C_{16:0}$	14-Methyl-pentadecanoic acid	< 1	ND
Iso-$C_{15:0}$	13-Methyl-tetradecanoic acid	< 1	ND
C16: 0 2-OH	2-Hydroxyhexadecanoic acid	ND	1.6
$C_{17:2}$	Heptadecadienoic acid	ND	11.6
$C_{13:1\ cis\ 12}$	12-Tridecanoic acid	ND	2.1
Anteiso-$C_{15:0}$	12-Methyl-tetradecanoic acid	ND	< 1

ND Not detected

[a]Data for *E. tayi* were obtained from Amir et al. [25]

function by COGs or by NR BLAST. 130 genes were identified as ORFans (2.13%). The remaining 1525 genes (24. 94%) were annotated as hypothetical proteins. Two genes associated with Vancomycin (Vancomycin B-type resistance protein, VanW) resistance and 20 genes associated with beta-lactamase resistance were found using the RAST web server [26]. The remaining 1525 genes (24.94%) were annotated as hypothetical proteins.

Genome Comparison

The draft genome sequence of strain AT11^T (7.11 MB) is larger in size than those of *C. bolteae, B. producta, C. clostridioforme, Eubacterium ruminantium*, and *A. hadrus* (6.38, 6.09, 5.46, 2.84, and 2.77 MB, respectively) but almost equal to that of *E. tayi* (7.15). Its G+C (48%) content is lower than that of *C. clostridioforme* and *C. bolteae* (49 and 49.6%, respectively), but higher than that of *E. tayi, B. producta, Eubacterium ruminantium*, and *A. hadrus* (46.3, 45.7, 37.2, and 37.2, respectively). Its gene content (6114) is higher than that of *C. bolteae*, *B. producta*, *C. clostridioforme*, *A. hadrus*, and *Eubacterium ruminantium* (5892, 5666, 5376, 2716, and 2533, respectively) but lower than that of *E. tayi* at 6156. The distribution of genes into COG categories was not entirely similar in all compared genomes (Fig. 3). The average genomic identity of orthologous gene sequences (AGIOS) values ranged from 61.7% between *C. bolteae* and *Eubacterium ruminantium* to 90.8% between *C. bolteae* and *C. clostridioforme* among compared species without strain AT11^T (Supplementary Table 2). When strain AT11^T was included in the comparison, these values ranged from 62.2% with *Eubacterium ruminantium* to 78.4% with *E. tayi* (Supplementary Table 2). The dDDH values for strain AT11^T ranged from 19.5% with *A. hadrus* to 34.4% with *C. clostridioforme* (Supplementary Table 3) with a probability of error of ± 2%. These values are very low and below the cutoff of 70%, thus also confirming that this strain is a new species.

Based on the phenotypic, chemotaxonomic, genomic, and phylogenetic characteristics, a novel bacterium isolated from the stool sample of a morbidly obese French woman, under the name *Eisenbergiella massiliensis* sp. nov., is proposed. The type strain is AT11^T = DSM 100838^T = CSUR P2478^T.

Description of *Eisenbergiella massiliensis* sp. nov

Eisenbergiella massiliensis (*mas.si.li.en'sis*. L. fem. adj. massiliensis, of Massilia, the Latin name for Marseille). It is a strictly anaerobic bacterium which grows at a mesothermal temperature of 37 °C. The colonies grown on Columbia agar with 5% enriched sheep blood are light gray, non-hemolytic, and irregular with a diameter of 0.5 mm.

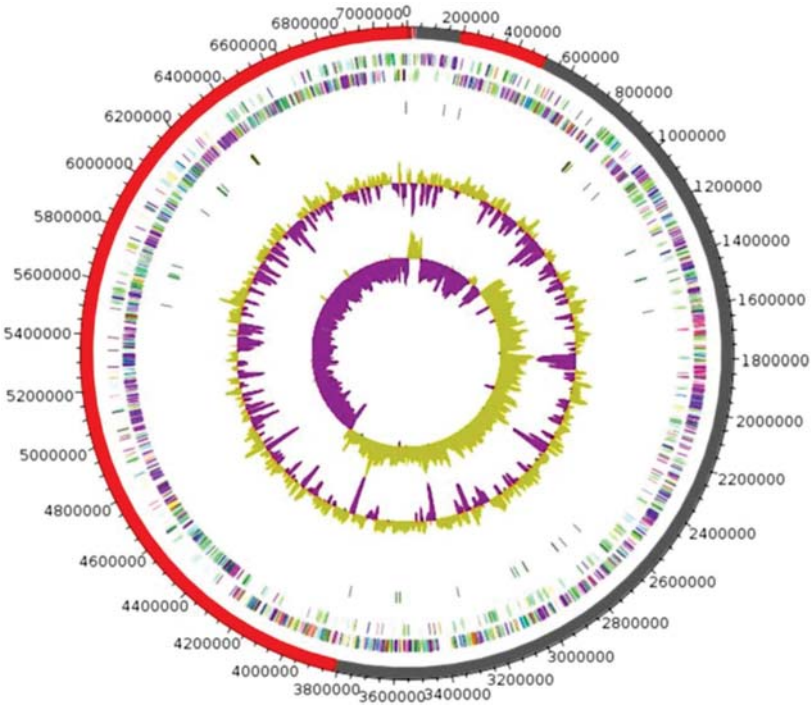


Fig. 2 Graphical circular map of the genome of strain AT11^T: From outside to the center: Contigs (red/gray), Clusters of Orthologous Groups (COGs) category of genes on the forward strand (three circles), genes on forward strand (blue circle), genes on the reverse strand (red circle), COGs category on the reverse strand (three circles), G+C content. (Color figure online)

Cells exhibit a negative Gram-stain, are non-spore-forming, non-motile, catalase positive, and rod-shaped bacilli, with a size of 0.4/2 µm. Using the API Gallery systems (API® ZYM API® 50CH API® 20A and API® rapid ID 32A) in anaerobic condition, positives reactions were observed for acid phosphatase, alkaline phosphatase, esterase, esterase lipase, naphthol-AS-BI-phosphohydrolase, *N*-acetyl-β-glucosaminidase, α-arabinosidase, α-fucosidase, α-galactosidase, β-galactosidase, α-glucosidase, β-glucosidase, β-glucuronidase, 6-phosphate-β-galactosidase, arbutin, D-cellobiose, D-glucose, D-lactose, D-lyxose, D-maltose, D-mannose, D-raffinose, D-saccharose, D-tagatose, D-trehalose, dulcitol, D-xylose, L-arabinose, L-rhamnose, L-sorbose, potassium 5-cetogluconate, and salicin. Urease and indole are not produced, gelatin was not liquefied and nitrate was not reduced, although esculin was hydrolyzed. The major cellular fatty acids detected were $C_{16:0}$ (62.7%) and $C_{18:1n9}$ (10.3%). Its genome, consisting of one chromosome, is 7,114,554 bp in length with 48% of G+C content. The type strain AT11^T = CSUR P2478^T = DSM 100838^T was isolated from the stool sample of a French morbidly obese woman following bariatric surgery.

Acknowledgements The authors thank the Xegen Company (http://www.xegen.fr) for automating the genomic annotation process.

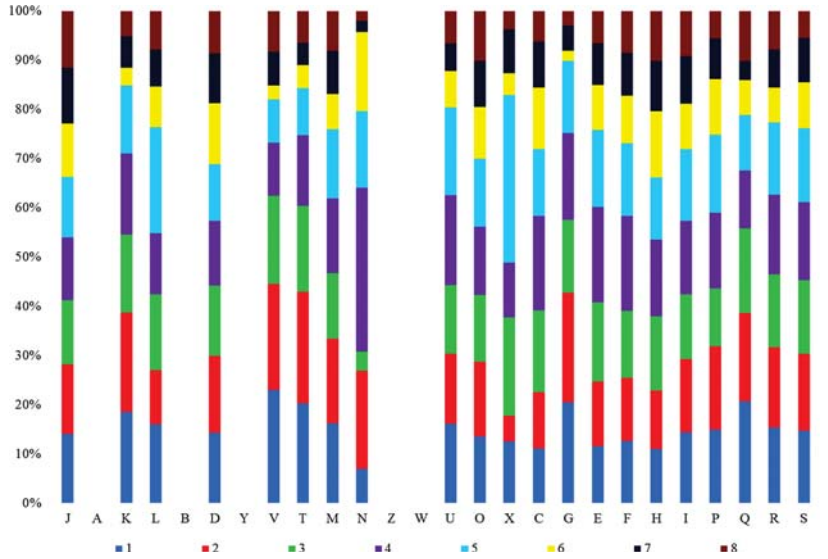


Fig. 3 Distribution of functional classes of predicted genes according to the Clusters of Orthologous Groups (COGs) of strain AT11^{T} with its closest species: 1, Strain AT11^{T}; 2, *Eisenbergiella tayi* strain DSM 26961^{T}; 3, *Blautia producta* strain ATCC 27340^{T}; 4, *Clostridium bolteae* strain WAL 16351^{T}; 5, *Clostridium clostridioforme* strain ATCC 25537^{T}; 6, *Parasporobacterium paucivorans* strain DSM 15970^{T}; 7, *Anaerostipes hadrus* strain ATCC 29173^{T}; and 8, *Eubacterium ruminantium* strain ATCC 17233^{T}. *Superscript T* Type strain, *ATCC* American Type Culture Collection, *DSM* Deutsche Sammlung von Mikroorganismen, *WAL* Wadsworth Anaerobe Laboratory

Funding This work was funded by Fondation Méditerranée Infection.

Compliance with Ethical Standards

Conflict of interest The author declares that they have no conflicts of interest.

References

1. Amir I, Bouvet P, Legeay C et al (2014) *Eisenbergiella tayi* gen. nov., sp. nov., isolated from human blood. Int J Syst Evol Microbiol 64:907–914
2. Bernard K, Burdz T, Wiebe D et al (2017) Characterization of isolates of *Eisenbergiella tayi*, a strictly anaerobic gram-stain variable bacillus recovered from human clinical materials in Canada. Anaerobe 44:128–132
3. Bryant MP (1959) Bacterial species of the rumen. Bacteriol Rev 23:125–153
4. Citron DM, Ostovari MI, Karlsson A, Goldstein EJ (1991) Evaluation of the E test for susceptibility testing of anaerobic bacteria. J Clin Microbiol 29:2197–2203
5. Dione N, Sankar SA, Lagier J-C et al (2016) Genome sequence and description of *Anaerosalibacter massiliensis* sp. nov. New Microbes New Infect 10:66–76
6. Drancourt M, Bollet C, Carlioz A et al (2000) 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. J Clin Microbiol 38:3623–3630
7. Ezaki T, Li N, Hashimoto Y et al (1994) 16S ribosomal DNA sequences of anaerobic cocci and proposal of *Ruminococcus hansenii* comb. nov. and *Ruminococcus productus* comb. nov. Int J Syst Bacteriol 44:130–136
8. Gouret P, Paganini J, Dainat J et al (2011) Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the multi-agent software system DAGOBAH. In: Pontarotti P (ed) Evolutionary biology—concepts, biodiversity, macroevolution and genome evolution. Springer, Berlin Heidelberg, pp 71–87
9. Gouret P, Vitiello V, Balandraud N et al (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinform 6:198
10. Hyatt D, Chen G-L, Locascio PF et al (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 11:119
11. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036
12. Kaneuchi C, Watanabe K, Terada A et al (1976) Taxonomic Study of *Bacteroides clostridiiformis* subsp. *clostridiiformis* (Burri and Ankersmit) Holdeman and Moore and of Related Organisms: Proposal of *Clostridium clostridiiformis* (Burri and Ankersmit) comb.

nov. and *Clostridium symbiosum* (Stevens) comb. nov. Int J Syst Bacteriol 26:195–204
13. Kant R, Rasinkangas P, Satokari R et al (2015) Genome sequence of the butyrate producing Anaerobic bacterium *Anaerostipes hadrus* PEL 85. Genome Announc 3:e00224-15
14. Kong L-C, Tap J, Aron-Wisnewsky J et al (2013) Gut microbiota after gastric bypass in human obesity: increased richness and associations of bacterial genera with adipose tissue genes. Am J Clin Nutr 98:16–24
15. Lagesen K, Hallin P, Rødland EA et al (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108
16. Lagier J-C, Hugon P, Khelaifia S et al (2015) The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. Clin Microbiol Rev 28:237–264
17. Lechner M, Findeiss S, Steiner L et al (2011) Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinform 12:124
18. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444:1022–1023
19. Liu C, Finegold SM, Song Y, Lawson PA (2008) Reclassification of *Clostridium coccoides, Ruminococcus hansenii. Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus* and *Ruminococcus schinkii* as *Blautia coccoides* gen. nov., comb. nov., *Blautia hansenii* comb. nov., *Blautia hydrogenotrophica* comb. nov., *Blautia luti* comb. nov., *Blautia producta* comb. nov., *Blautia schinkii* comb. nov. and description of *Blautia wexlerae* sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 58:1896–1902
20. Lomans BP, Leijdekkers P, Wesselink J-J et al (2001) Obligate sulfide-dependent degradation of methoxylated aromatic compounds and formation of methanethiol and dimethyl sulfide by a freshwater sediment isolate, *Parasporobacterium paucivorans* gen. nov., sp. nov. Appl Environ Microbiol 67:4017–4023
21. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964
22. Matuschek E, Brown DFJ, Kahlmeter G (2014) Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. Clin Microbiol Infect 20:O255–O266
23. Meier-Kolthoff JP, Göker M, Spröer C, Klenk H-P (2013) When should a DDH experiment be mandatory in microbial taxonomy? Arch Microbiol 195:413–418
24. Million M, Maraninchi M, Henry M et al (2012) Obesity-associated gut microbiota is enriched in *Lactobacillus reuteri* and depleted in *Bifidobacterium animalis* and *Methanobrevibacter smithii*. Int J Obes 36:817–825
25. Ng M, Fleming T, Robinson M et al (2014) Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet 384:766–781
26. Overbeek R, Olson R, Pusch GD et al (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 42:D206-214
27. Ramasamy D, Mishra AK, Lagier J-C et al (2014) A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 64:384–391
28. Rettedal EA, Gumpert H, Sommer MOA (2014) Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. Nat Commun 5:4714
29. Seng P, Drancourt M, Gouriet F et al (2009) Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 49:543–551
30. Song Y, Liu C, Molitoris DR et al (2003) *Clostridium bolteae* sp. nov., isolated from human sources. Syst Appl Microbiol 26:84–89
31. Togo AH, Khelaifia S, Bittar F et al (2016) '*Eisenbergiella massiliensis*', a new species isolated from human stool collected after bariatric surgery. New Microbes New Infect 13:15–16
32. Togo AH, Khelaifia S, Lagier J-C et al (2016) Noncontiguous finished genome sequence and description of *Paenibacillus ihumii* sp. nov. strain AT5. New Microbes New Infect 10:142–150
33. Togo AH, Durand G, Khelaifia S et al (2017) *Fournierella massiliensis*, gen. nov., sp. nov., a new human-associated member of the family *Ruminococcaceae*. Int J Syst Evol Microbiol 67:1393–1399
34. Zhang H, DiBaise JK, Zuccolo A et al (2009) Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci USA 106:2365–2370

Autres descriptions de nouvelles espèces bactériennes

Article 18:

Non-contiguous finished genome sequence and description of *Bartonella mastomydis* sp. nov.

M. Dahmani, G. Diatta, N. Labas, A. Diop, H. Bassene, D. Raoult, L. Granjon, F. Fenollar, O. Mediannikov

[Published in New Microbes New Infection]

Non-contiguous finished genome sequence and description of *Bartonella mastomydis* sp. nov.

M. Dahmani[1], G. Diatta[2], N. Labas[1], A. Diop[1], H. Bassene[2], D. Raoult[1], L. Granjon[3], F. Fenollar[1], O. Mediannikov[1,2]

[1] Aix Marseille Univ, CNRS, IRD, INSERM, AP-HM, URMITE, IHU - Méditerranée Infection, Marseille, France

[2] Research Unit of Emerging Infectious and Tropical Diseases (URMITE) UMR CNRS 7278 IRD 198, Institute of Research for Development, Dakar, Senegal

[3] CBGP, IRD, INRA, CIRAD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France

* Corresponding author
O. Mediannikov, URMITE, IHU - Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France

Tel.: +33 4 13 73 24 01
Fax: +33 4 13 73 24 02
Email: olegusss1@gmail.com

Non-contiguous finished genome sequence and description of *Bartonella mastomydis* sp. nov.

M. Dahmani[1], G. Diatta[2], N. Labas[1], A. Diop[1], H. Bassene[2], D. Raoult[1], L. Granjon[3], F. Fenollar[1], O. Mediannikov[1,2]

Word abstract count: 74

Word text count: 2,823

Abstract

Bartonella mastomydis sp. nov. strain 008 is the type strain of *B. mastomydis* sp. nov., a new species within the genus *Bartonella*. This strain was isolated from *Mastomys erythroleucus* rodents trapped in the Sine-Saloum region of Senegal. Here we describe the features of this organism, together with the complete genome sequence and its annotation. The 2,044,960 bp-long genomes with 38.44% GC content contains 1,674 protein-coding and 42 RNA genes, including three rRNA genes.

Key words: *Bartonella mastomydis* sp.nov, complete genome, *Mastomydis erythroleucus*

Introduction

Just over a century ago, the first historical record of the emerging *Bartonella* genus was made during World War I, when a million frontline troops were shown to be plagued by a disease later known as "trench fever". This was caused by the louse-borne bacterium now known as *Bartonella quintana* [1]. *Bartonella* are small facultative intracellular, vector-transmitted, Gram-negative, hemotropic bacilli, classified within the class of α-proteobacteria [2]. The genus was significantly expanded after Brenner *et al.* proposed the unification of genera *Bartonella* and *Rochalimaea* in 1993, and Birtles *et al.* unified *Bartonella* and *Grahamella* genera in 1995 [3]. The *Bartonellaceae* family (Gieszczykiewicz 1939) [4] contains 35 species and 3 sub-species [5] as of August 01, 2017 [6]. Bartonellae usually exists in two specific habitats: the gut of the obligately blood sucking arthropod vector and the bloodstream of the mammalian host [1]. Among the 38 recognized *Bartonella* species, seventeen species have been described as pathogenic for humans [7]. In humans, *Bartonella* bacteria are among the most described as being associated with endocarditis or cardiopathy. In animal hosts, a wide array of clinical syndromes from asymptomatic infection to endocarditis is described [7–9], although the infection is often asymptomatic.

New species and sub-species are constantly being proposed. Candidate species belonging to the genus *Bartonella* from a wide range of animal reservoirs have been described but not yet assigned new species designations [1]. Parasitism by Bartonellae is widespread among small mammals. Potentially new *Bartonella* species infecting bat communities were reported in Madagascar [10], Kenya [11], Puerto Rico [12], and French Guiana [13]. Rodents and insectivores were showed to maintain Bartonellae infections. Additionally, a large number of partially characterized *Bartonella* have been isolated from rodents in Southeast Asia [14], South Africa [15,16], Europe, North and South America [17], Nigeria [18], the Republic of Congo, and Tanzania [17]. In Senegal, West Africa, using the criteria proposed

by La Scola *et al.* based on the multilocus sequence analyses of four genes and the intergenic spacer as a tool to the description of Bartonellae [19], three new Bartonellae were isolated and described: *Bartonella senegalensis*, *Bartonella massiliensis* from soft ticks *Ornithodoros sonrai* [14], and *Bartonella davoustii* from cattle [20]. Our aim is to describe an additional *Bartonella* species isolated from small mammals in the region of Sine-Saloum, in western Senegal [21]. In this rural region, the biotype is favorable to the spread of commensal mammals harboring pathogenic microorganisms and often found in close contact with humans. This situation increases the risk of human and animal transmission of infectious disease from rodent-associated tick-borne pathogens. This work describes the genome sequence of the proposed candidate *Bartonella mastomydis* strain 008 isolated from *Mastomys erythroleucus* using a polyphasic approach combining matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry and genomic properties, as well as next-generation sequencing technology to complete description of a potentially new species [22]. Here we present the summary classification and a set of features for *B. mastomydis* sp. nov. strain 008 together with the description of the complete genomic sequences and annotation. These characteristics support the definition of the species *B. mastomydis*.

Samples and bacterial culture

In February 2013, rodents and insectivores were captured alive in two sites (Dielmo and Ndiop) using wire mesh traps baited with peanut butter or onions. Our aim was to investigate the presence of *Bartonella* spp. in commensal rodents in Sine-Saloum, Senegal. In this region, rodents and rodent-associated soft ticks are respectively the reservoirs and vectors of relapsing fever caused by *Borrelia crocidurae*. Trapped rodents and insectivores were anesthetized and autopsied in sterile conditions. Sampled blood was inoculated on home-made Columbia agar plates supplemented with 5% sheep blood. The results of this study were

reported elsewhere [21]. In total, within a 6-day period, 119 small mammals were captured: 116 rodents and three shrews (*Crocidura cf. olivieri*). Rodents were identified morphologically as follows: 5 *Arvicanthis niloticus*, 56 *Gerbilliscus gambianus*, 49 *Mastomys erythroleucus*, 5 *Mus musculus*, and 1 *Praomys daltoni*. Thirty isolates of *Bartonella* spp. were recovered from the rodent bloodstreams. None of those isolated belonged to previously described *Bartonella* species (Table 1).

Classification and features

The *gltA, rpoB*, 16S rRNA, *ftsZ* genes, and the intergenic spacer (ITS) have been amplified and sequenced from recovered *Bartonella* isolates [19,23–26]. *Bartonella mastomydis* (21 isolates) recovered only from *Mastomys erythroleucus* was obtained following the fifth to tenth incubations at 37°C in a 5% CO_2-enriched atmosphere on Columbia agar plates supplemented with 5% sheep blood. Other morphologically and genetically indistinguishable strains were isolated from *Mastomys erythroleucus.* The 21 isolates of *B. mastomydis* are almost genetically identical, however, strains type 008, 025, 086, 202 show different nucleotide identity. The identities between them are as follows: 100% for the *rrs* gene, 99% for the *rpoB* gene, and 98-99% for the *ftsZ* and *gltA* genes. The sequence of the intragenic spacer (ITS) of the strain 008 present 94-99% identity with the strain 025, 086, 202 presented by a 23 bp deletion and 4 bp insertion compared to the other strains. This study focused on the taxonomic description and identification of strains 008.

Strain 008 exhibits the following nucleotide sequence similarities for the *rrs* gene (KY555064): 99% with *Bartonella tribocorum* strain BM1374166 (HG969192), *Bartonella grahamii* as4aup (CP001562), *Bartonella vinsonii* subsp. *arupensis* strain OK 94-513 (NR_104902) and subsp. *berkhoffii* (CP003124), *Bartonella elizabethae* strain F9251 (NR_025889), *Bartonella henselae* strain Houston-1 (NR_074335), and finally *Bartonella quintana* str. Toulouse (BX897700). For the ITS (KY555067), 95% similarity was observed

with *B. elizabethae* (L35103). For the *gltA* gene (KY555066), 97% similarity was observed with *B. elizabethae* (Z70009), 94% with *B. tribocorum* strain BM1374166 (HG969192), *B. grahamii* as4aup (CP001562), and *Bartonella queenslandensis* strain AUST/NH12 (EU111798). For the *ftsZ* gene (KY555065), 98% of similarity was observed with *B. elizabethae* (AF467760), 96% with *B. tribocorum* strain BM1374166 (HG969192), *B. grahamii* as4aup (CP001562), and *B. queenslandensis* strain AUST/NH12 (EU111798). For *rpoB gene* (KY555068), 99% similarity was observed with multiple uncultured *Bartonella* amplified from small mammals from Ethiopia [27], Benin [28], Congo and Tanzania [17], and Nepal [29]. The closest recognized species was *B. elizabethae* (AF165992) at 98% homology (Figure 1).

MALDI-TOF mass spectrometry protein analysis was carried out as previously described [22]. Five isolated colonies of strain 008 were deposited as individual spots on the MALDI target plate. Each smear was overlaid with 2 µL of matrix solution (a saturated solution of alpha-cyano-4-hydroxycinnamic acid) in 50% acetonitrile/2.5% trifluoro acetic acid and allowed to dry for 5 minutes. Measurements were performed with a Microflex spectrometer (Bruker Daltonics, Leipzig, Germany). The five 008 spectra were imported into the MALDI BioTyper software (version 2.0, Bruker) and analyzed by standard pattern matching (with default parameter settings) against the main spectra of 4,613 bacteria in the BioTyper database and the 25 *Bartonella* species in our own database. The identification method included the m/z from 3,000 to 15,000 Da. For every spectrum, a maximum of 100 peaks was considered and compared with the spectra in the database. A score of below 1.7 meant identification was not possible. For strain 008, the scores obtained were always below 1.5, suggesting that our isolate was not a member of a known species. We added the spectrum from strain 008 to the database (Figure 2). A gel view comparing the spectrum of strain 008 with those of other *Bartonella* species is shown in (Figure 3).

Biochemical characterization and antibiotic susceptibility

Different growth temperatures (32, 37, 42°C) were tested. Growth occurred only at 37°C in 5% CO_2. Colonies were gray, opaque, and 0.5 mm to 1 mm in diameter on blood-enriched Columbia agar. A motility test was negative. Cells grown on agar were Gram-negative and have a mean length and width of 1369.8±423.8 nm and 530.9±105.8 nm, respectively, by electron microscopy (Figure 4). No flagella or pili were observed. Strain 008 exhibited neither catalase nor oxidase activity. Biochemical characteristics were assessed using API 50 CH (bioMérieux, Marcy l'Etoile, France), API ZYM (bioMérieux), and API Coryne (bioMérieux); none of the available biochemical tests were positive. Similar profiles were previously observed for *B. senegalensis* [30]. *Bartonella mastomydis* is sensitive to amoxicillin, amoxicillin-clavulanic acid, oxacillin, imipenem, rifampicin, nitrofurantoin, doxycyclin, linezolid, tobramycin, gentamycin, trimethoprim-sulfamethoxazole, fosfomycin, and ciprofloxacin. *Bartonella mastomydis* is resistant to metronidazole and colistin.

Genome sequencing information

Genome project history

The organism was selected for sequencing based on the similarity of its 16S rRNA, ITS, *ftsZ, gltA,* and *rpoB* to other members of the genus *Bartonella*. Nucleotide sequence similarities for these genes suggested that strain 008 represents a new species in the genus *Bartonella*. A summary of the project information is shown in Table 2. The GenBank accession number is GCA_900185775, and the entry consists of 12 scaffolds (>1,500 bp). Table 2 shows the project information and its association with MIGS version 2.0 compliance.

Genome sequencing and assembly

Bartonella mastomydis sp. nov. strain 008 (DSM 28002; CSUR B643) was grown on 5% sheep blood-enriched Columbia agar at 37°C in a 5% CO_2 atmosphere. gDNA of *B. mastomydis* sp. nov. strain 008 was extracted in two steps. A mechanical treatment was first

159 performed by acid-washed glass beads (G4649-500g Sigma) using a FastPrep BIO 101
160 instrument (Qbiogene, Strasbourg, France) at maximum speed (6.5 m/s) for 90 s. Then after a
161 2-hour lysozyme incubation at 37°C, DNA was extracted on the EZ1 biorobot (Qiagen,
162 Hilden, Germany) with the EZ1 DNA tissue kit. The elution volume was 50 µL. Genomic
163 DNA was quantified by a Qubit assay with the high sensitivity kit (Life technologies,
164 Carlsbad, CA, USA) to 66 ng/µL. Genomic DNA was sequenced on the MiSeq Technology
165 (Illumina Inc, San Diego, CA, USA) with the mate pair strategy. The gDNA was barcoded to
166 be mixed with 11 other projects with the Nextera Mate Pair sample prep kit (Illumina Inc).
167 The mate pair library was prepared with 1.5 µg of genomic DNA using the Nextera
168 mate pair Illumina guide. The genomic DNA sample was simultaneously fragmented and
169 tagged with a mate pair junction adapter. The profile of the fragmentation was validated on an
170 Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA
171 7500 labchip. The optimal size of obtained fragments was 7.77 kb. No size selection was
172 performed and 600 ng of tagmented fragments were circularized. The circularized DNA was
173 mechanically sheared to small fragments with optima on a bimodal curve at 593 and 1,377 bp
174 on the Covaris device S2 in T6 tubes (Covaris, Woburn, MA, USA). The library profile was
175 visualized on a High Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc) and the
176 final concentration library was measured at 49.16 nmol/L. The libraries were normalized at 2
177 nM, pooled with 11 other projects, denatured and diluted at 15 pM. Automated cluster
178 generation and 2x250-bp sequencing runs were performed in a 39-hour run.
179 Total information of 7.2 Gb was obtained from a 765 K/mm^2 cluster density with a
180 cluster passing quality control filters of 94.7% (14,162,000 passed filter clusters). Within this
181 run, the index representation for *B. mastomydis* was determined to 12.30%. The 1,742,441
182 paired end reads were filtered according to the read qualities.
183 ***Genome assembly***

The genome's assembly was performed with a pipeline that enabled creation of an assembly with different software programs (Velvet [31], Spades [32] and Soap Denovo [33]), on trimmed (MiSeq and Trimmomatic [34]) or untrimmed data (only MiSeq). For each of the six assemblies performed, GapCloser [33] was used to reduce gaps. Then contamination with Phage Phix was identified (BLASTN against Phage Phix174 DNA sequence) and eliminated. Finally, scaffolds under 800 bp were removed and scaffolds with a depth value lower than 25% of the mean depth were removed (identified as possible contaminants). The best assembly was selected by using different criteria (number of scaffolds, N50, number of N).

Genome annotation

Open Reading Frames (ORFs) were predicted using Prodigal [35] with default parameters but the predicted ORFs were excluded if they spanned a sequencing gap region (contained N). The predicted bacterial protein sequences were searched against the Clusters of Orthologous Groups (COG) database using BLASTP (E-value of 1e -03, coverage 0.7 and identity percent 30%). If no hit was found, it searches against the NR database using BLASTP (E-value of 1e-03, coverage 0.7 and identity percent of 30%). If the sequence length was smaller than 80 amino acids, we used an E-value of 1e-05. The tRNAScanSE [36] tool was used to find transfer RNA genes, whereas ribosomal RNA genes were found by using RNAmmer [37]. Lipoprotein signal peptides and the number of transmembrane helices were predicted using Phobius [38]. ORFans were identified if not all of the BLASTP performed gave positive results (E-value smaller than 1e-03 for ORFs with sequence size superior to 80 aa or E-value smaller than 1e-05 for ORFs with sequence length smaller than 80 aa). Such parameter thresholds have already been used in previous work to define ORFans.

Genome properties

The genome is 2,044,960 bp long with 38.44% GC content. It is composed of 12 scaffolds (composed of 14 contigs) (Figure 5). Of the 1,716 predicted genes, 1,674 were

protein-coding genes and 42 were RNAs (1 gene is 5S rRNA, 1 gene is 16S rRNA, 1 gene is 23S rRNA, 39 genes are tRNA genes). A total of 1,212 genes (72.4%) were assigned as putative function (by cogs or by NR blast). 56 genes were identified as ORFans (3.35%). The remaining 338 genes were annotated as hypothetical proteins (20.19%). The distribution of genes into COGs functional categories is presented in Table 3. The propriety and statistics of the genome are summarized in Tables 3 and 4. The most predicted functional genes are associated with translation (9.38%), followed by those involved in the basic biological functions, such as amino acid transport and metabolism (6.33%), energy production and conversion (4.42%), and carbohydrate transport and metabolism (3.35%) (Table 4).

Insights from the genome sequence

The draft genome sequence of *B. mastomydis* is smaller than those of *Bartonella rattaustraliani, Bartonella florencae, B. queenslandensis,* and *B. tribocorum* (2,045, 2,158, 2,054, 2,378, and 2,631 Mb, respectively), but larger than those of *B. elizabethae* and *B. vinsonii* subsp. *berkhoffii* (1,964 and 1,803 Mb, respectively). The G+C content of *B. mastomydis* is smaller than those of *B. rattaustraliani, B. vinsonii* subsp. *berkhoffii*, *B. florencae*, and *B. tribocorum* (38.44, 38.8, 38.83, 38.45, and 38.81%, respectively), but larger than those of *B. elizabethae* and *B. queenslandensis* (38.32 and 38.38%, respectively). The protein-coding gene content of *B. mastomydis* is smaller than those of *B. rattaustraliani, B. florencae, B. queenslandensis*, and *B. tribocorum* (1,674, 1,943, 1,886, 2,466, and 2,295, respectively), but larger than those of *B. elizabethae* and *B. vinsonii* subsp. *berkhoffii* (1,663 and 1,434, respectively). Similarly, the gene content of *B. mastomydis* (1,674) is smaller than those of *B. rattaustraliani, B. florencae, B. queenslandensis,* and *B. tribocorum* (1,943, 1,886, 2,466, and 2,295, respectively), but larger than those of *B. elizabethae* and *B. vinsonii* subsp. *berkhoffii* (1,663 and 1,434, respectively). The COG category gene distribution is not similar. *B. mastomydis* has fewer COG category genes belonging to transcription (58) than *B.*

tribocorum (73). *Bartonella mastomydis* has also fewer genes belonging to the replication, recombination and repair COG category (73) than *B. rattaustraliani* (108), *B. queenslandensis* (100), and *B. tribocorum* (95). Finally, *B. mastomydis* has also fewer genes belonging to mobilome: prophages, transposons COG category (25) than *B. tribocorum, B. rattaustraliani, B. queenslandensis, B. vinsonii* subsp. *berkhoffii*, and *B. florencae* (125, 56, 50, 45, and 43, respectively) (Figure 6). Among species with standing in nomenclature, AGIOS values ranged from 0.96 between *B. mastomydis* and *B. elizabethae* to 0.66 between *B. vinsonii* subsp *berkhoffii* and *B. rattaustraliani, B. queenslandensis, B. elizabethae, B. mastomydis, B. rattaustraliani, B. tribocorum, B. florencae,* and *B. tribocorum* (Table 5). To evaluate the genomic similarity among the strains, we determined two parameters, dDDH, which exhibits high correlation with DDH [39], and AGIOS [40], which was designed to be independent of DDH (Table 6).

Conclusion

Based on phenotypic, phylogenetic, and genomic analyses, we formally propose the creation of *Bartonella mastomydis* sp. nov. that contains the strain 008. This bacterial strain has been isolated from *Mastomys erythroleucus* blood samples trapped in the Sine-Saloum region of Senegal.

Description of *Bartonella mastomydis* sp. nov. strain 008

Bartonella mastomydis (mas.to′my.dis. N.L. gen. n. mastomydis of *Mastomys*, isolated from *Mastomys erythroleucus*) is a non-motile Gram-negative rod. Growth is only obtained at 37°C. Colonies are opaque, gray and 0.5 to 1 mm in diameter on blood-enriched Columbia agar. Cells are rod-shaped without flagella or pili. Length and width are 1369.8±423.8 nm and 530.9±105.8 nm, respectively. *Bartonella mastomydis* strain 008 exhibits neither biochemical nor enzymatic activities. The type strain 008 is sensitive to rifampicin, amoxicillin, amoxicillin-clavulanic acid, oxacillin, nitrofurantoin, doxycycline, linezolid, tobramycin,

gentamycin, imipenem, trimethoprim-sulfamethoxazole, fosfomycin and ciprofloxacin, and resistant to metronidazole and colistin. The G+C content of the genome is 38.44%. The 16S rRNA gene sequence and whole-genome shotgun sequence of strain 008 are deposited in GenBank under accession numbers (KY555064) and (GCA_900185775), respectively. The type strain 008 (CSUR B643, DSM2802) was isolated from the rodent *Mastomys erythroleucus* trapped in the region of Sine-Saloum, Senegal.

Legend

Figure 1. The evolutionary history of the sequenced samples was inferred using the maximum likelihood method implemented in MEGA7 [41] and based on concatenated *gltA*, *rpoB,* 16S RNA, and *ftsZ* (total length of 2,731 bp*)* sequences. The sequences of the *gltA*, *rpoB,*16S RNA, and *ftsZ* genes used for comparison were obtained from the GenBank database [42]. The sequences were aligned using BioEdit [43]. Firstly, for each gene individually, the sequences we used for comparison were first aligned using CLUSTAL W. All positions containing gaps and missing data were eliminated manually, then each alignment was concatenated, and a second alignment was performed. The evolutionary history was inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The initial tree for the heuristic search was obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 0.2144)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Statistical support for internal branches of the trees was evaluated by bootstrapping with 1000 iterations. The analysis involved 39 nucleotide sequences.

Figure 2. Reference mass spectrum from *Bartonella mastomydis* strain 008. Spectra from 12 individual colonies were compared and a reference spectrum was generated.

Figure 3. Gel view comparing *Bartonella mastomydis* strain 008 spectra with other members of the *Bartonella* genus. The gel view displays the raw spectra of loaded spectrum files arranged in a pseudo-gel like look. The x-axis records the m/z value. The left y-axis displays the running spectrum number originating from subsequent spectra loading. The peak intensity

290 is expressed by a Grayscale scheme code. The color bar and the right y-axis indicate the
291 relation between the color in which a peak is displayed and the peak intensity in arbitrary
292 units. Displayed species are indicated on the left.
293 **Figure 4.** Transmission electron microscopy of *Bartonella mastomydis* strain 008, using a
294 TECNAI G20 (FEI) at an operating voltage of 200 keV. The scale bar represents 200 nm.
295 **Figure 5.** Graphical circular map of the chromosome. From outside to the center: Genes on
296 the forward strand colored by COG categories (only genes assigned to COG), genes on the
297 reverse strand colored by COG categories (only gene assigned to COG), RNA genes (tRNAs
298 green, rRNAs red), GC content and GC skew.
299 **Figure 6.** Distribution of functional classes of predicted genes according to the clusters of
300 orthologous groups of proteins.
301
302 **Table 1.** Classification and general features of *Bartonella mastomydis* strain 008.
303 **Table 2.** Project information.
304 **Table 3.** Number of genes associated with the 25 general COG Functional categories.
305 **Table 4.** Nucleotide content and gene count levels of the genome.
306 **Table 5.** The numbers of orthologous protein shared between genomes (upper right)[a].
307 **Table 6.** Pairwise comparison of *Bartonella mastomydis* with six other species using GGDC,
308 formula 2 (DDH estimates based on identities / HSP length)[a].
309

310 **Conflict of interest statement**

311 None of the authors has any conflicts of interest related to this article.

312 **Funding**

313 This study was supported by IHU Méditerranée Infection and the French National
314 Research Agency under the program "Investissements d'avenir," reference ANR-10-IAHU-
315 03. The funders had no role in study design, data collection or analysis, decision to publish, or
316 manuscript preparation.

References

[1] Okaro U, Addisu A, Casanas B, Anderson B. *Bartonella* Species, an Emerging Cause of Blood-Culture-Negative Endocarditis. Clin Microbiol Rev 2017;30:709–46.

[2] Tsai Y-L, Chang C-C, Chuang S-T, Chomel BB. *Bartonella* species and their ectoparasites: selective host adaptation or strain selection between the vector and the mammalian host? Comp Immunol Microbiol Infect Dis 2011;34:299–314. doi:10.1016/j.cimid.2011.04.005.

[3] Birtles RJ, Harrison TG, Saunders NA, Molyneux DH. Proposals to unify the genera *Grahamella* and *Bartonella*, with descriptions of *Bartonella talpae* comb. nov., *Bartonella peromysci* comb. nov., and three new species, *Bartonella grahamii* sp. nov., *Bartonella taylorii* sp. nov., and *Bartonella doshiae* sp. nov. Int J Syst Bacteriol 1995;45:1–8. doi:10.1099/00207713-45-1-1.

[4] Skerman VBD, McGowan V, Sneath PHA. Approved Lists of Bacterial Names. Int J Syst Evol Microbiol 1980;30:225–420. doi:10.1099/00207713-30-1-225.

[5] http://www.bacterio.net/ n.d.

[6] Guptill L. Bartonellosis. Vet Microbiol 2010;140:347–59. doi:10.1016/j.vetmic.2009.11.011.

[7] Angelakis E, Raoult D. Pathogenicity and treatment of *Bartonella* infections. Int J Antimicrob Agents 2014;44:16–25. doi:10.1016/j.ijantimicag.2014.04.006.

[8] Brouqui P, Raoult D. New insight into the diagnosis of fastidious bacterial endocarditis. FEMS Immunol Med Microbiol 2006;47:1–13. doi:10.1111/j.1574-695X.2006.00054.x.

[9] Chomel BB, Kasten RW, Williams C, Wey a C, Henn JB, Maggi R, et al. *Bartonella* endocarditis: a pathology shared by animal reservoirsand patients. Ann N Y Acad Sci 2009;1166:120–6. doi:10.1111/j.1749-6632.2009.04523.x.

[10] Brook CE, Bai Y, Dobson AP, Osikowicz LM, Ranaivoson C, Zhu Q, et al. *Bartonella* spp . in fruit bats and blood- feeding ectoparasites in Madagascar 2015:1–9. doi:10.1371/journal.pntd.0003532.

[11] Kosoy M, Bai Y, Lynch T, Kuzmin I V, Niezgoda M, Franka R, et al. *Bartonella* spp.

in bats, Kenya. Emerg Infect Dis 2010;16:1875–81. doi:10.3201/eid1612.100601.

[12] Olival KJ, Dittmar K, Bai Y, Rostal MK, Lei BR, Daszak P. Bartonella spp . in a Puerto Rican Bat Community 2015;51:274–8. doi:10.7589/2014-04-113.

[13] Davoust B, Marié J-L, Dahmani M, Berenger J-M, Bompar J-M, Blanchet D, et al. Evidence of *Bartonella* spp. in blood and ticks (*Ornithodoros hasei*) of bats, in French Guiana. Vector-Borne Zoonotic Dis 2016;16:516–9. doi:10.1089/vbz.2015.1918.

[14] Jiyipong T, Jittapalapong S, Morand S, Raoult D, Rolain J. Prevalence and genetic diversity of *Bartonella* spp . in small mammals from southeastern Asia 2012;78:8463–6. doi:10.1128/AEM.02008-12.

[15] Pretorius A-M, Beati L, Birtles RJ. Diversity of bartonellae associated with small mammals inhabiting Free State province, South Africa. Int J Syst Evol Microbiol 2004;54:1959–67. doi:10.1099/ijs.0.03033-0.

[16] Brettschneider H, Bennett NC, Chimimba CT, Bastos a DS. Bartonellae of the Namaqua rock mouse, *Micaelamys namaquensis* (*Rodentia: Muridae*) from South Africa. Vet Microbiol 2012;157:132–6. doi:10.1016/j.vetmic.2011.12.006.

[17] Gundi V a KB, Kosoy MY, Makundi RH, Laudisoit A. Identification of diverse *Bartonella* genotypes among small mammals from Democratic Republic of Congo and Tanzania. Am J Trop Med Hyg 2012;87:319–26. doi:10.4269/ajtmh.2012.11-0555.

[18] Kamani J, Morick D, Mumcuoglu KY, Harrus S. Prevalence and diversity of *Bartonella* species in commensal rodents and ectoparasites from Nigeria, West Africa. PLoS Negl Trop Dis 2013;7:e2246. doi:10.1371/journal.pntd.0002246.

[19] Scola B La, Zeaiter Z, Khamis A, Raoult D. Gene-sequence-based criteria for species definition in bacteriology: the *Bartonella* paradigm. Trends Microbiol 2003;11:318–21. doi:10.1016/S0966-842X(03)00143-4.

[20] Dahmani M, Sambou M, Scandola P, Raoult D, Fenollar F, Mediannikov O. *Bartonella bovis* and *Candidatus* Bartonella davousti in cattle from Senegal. Comp Immunol Microbiol Infect Dis 2017;50:63–9. doi:10.1016/j.cimid.2016.11.010.

[21] Mediannikov O, Aubadie M, Bassene H, Diatta G, Granjon L, Fenollar F. Three new *Bartonella* species from rodents in Senegal. Int J Infect Dis 2014;21:335.

doi:10.1016/j.ijid.2014.03.1112.

[22] Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P-E, Rolain JM, et al. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 2009;49:543–51. doi:10.1086/600885.

[23] Brenner DONJ, Connor SPO, Winkler HH, Steigerwalt AG. Proposals To Unify the Genera *Bartonella* and *Rochalimaea* , with descriptions of *Bartonella quintana* comb. nov., *Bartonella* vinsonii comb . nov. , *Bartonella henselae* comb. nov., and *Bartonella elizabethae* comb. nov., and to remove the family *Bartonellaceae*. Int J Syst BACTERIOLO 1993:777–86. doi:0020-7713/93/040777-10$02.00/0.

[24] Birtles RJ, Raoult D. Comparison of partial *Citrate Synthase* gene (*gltA*) sequences for phylogenetic analysis of *Bartonella* species. Int J Syst Bacteriol 1996;1147:33–891. doi:10.1099/00207713-46-4-891.

[25] Renesto P, Gouvernet J. Use of *rpoB* gene analysis for detection and identification of *Bartonella* species. J Clin Microbiol 2001;39:430–7. doi:10.1128/JCM.39.2.430.

[26] Zeaiter Z, Liang Z, Raoult D. Genetic Classification and differentiation of *Bartonella* species based on comparison of partial *ftsZ* gene sequences. J Clin Microbiol 2002;40:3641–7. doi:10.1128/JCM.40.10.3641.

[27] Meheretu Y, Leirs H, Welegerima K, Breno M, Tomas Z, Kidane D, et al. *Bartonella* prevalence and genetic diversity in small mammals from Ethiopia. Vector Borne Zoonotic Dis 2013;13:164–75. doi:10.1089/vbz.2012.1004.

[28] Martin-Alonso A, Houemenou G, Abreu-Yanes E, Valladares B, Feliu C, Foronda P. *Bartonella* spp. in small mammals, Benin. Vector-Borne Zoonotic Dis 2016;16:229–37. doi:10.1089/vbz.2015.1838.

[29] Gundi VAKB, Kosoy MY, Myint KSA, Shrestha SK, Shrestha MP, Pavlin JA, et al. Prevalence and genetic diversity of *Bartonella* species detected in different tissues of small mammals in Nepal. Appl Environ Microbiol 2010;76:8247–54. doi:10.1128/AEM.01180-10.

[30] Bakour S, Rathored J, Lo CI, Mediannikov O, Beye M, Ehounoud CB, et al. Non-contiguous finished genome sequence and description of *Bartonella senegalensis* sp.

nov. New Microbes New Infect 2016;11:93–102. doi:10.1016/j.nmni.2016.03.004.

[31] Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821–9. doi:10.1101/gr.074492.107.

[32] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77. doi:10.1089/cmb.2012.0021.

[33] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 2012;1:18. doi:10.1186/2047-217X-1-18.

[34] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.

[35] Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119. doi:10.1186/1471-2105-11-119.

[36] Lowe TM, Eddy SR. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1996;25:955–64. doi:10.1093/nar/25.5.0955.

[37] Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35:3100–8. doi:10.1093/nar/gkm160.

[38] Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338:1027–36. doi:10.1016/j.jmb.2004.03.016.

[39] Auch AF, von Jan M, Klenk H-P, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2010;2:117–34. doi:10.4056/sigs.531120.

[40] Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 2014;64:384–91. doi:10.1099/ijs.0.057091-

434 0.

435 [41] Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis
436 version 7.0 for bigger datasets. Mol Biol Evol 2016;33:msw054.
437 doi:10.1093/molbev/msw054.

438 [42] Home - Nucleotide - NCBI n.d.

439 [43] Hall TA. BioEdit: a user-frindly biological sequences alignement editors and analysis
440 program for Windows 95/98/NT 1999:95–8. doi:10.12691/ajmr-2-6-8.

441

442

Table 1: Classification and general features of *Bartonella massiliensis* strain 008.

MIGS ID	Property	Term	Evidence code[a]
		Domain *Bacteria*	TAS [44]
		Phylum *Proteobacteria*	TAS [45]
		Class *Alphaproteobacteria*	TAS [46]
	Current classification	Order *Rhizobiales*	TAS [47,48]
		Family *Bartonellaceae*	TAS [4,23]
		Genus *Bartonella*	TAS [3,4,23,49]
		Species *Bartonella mastomydis*	IDA
		Type strain 008	IDA
	Gram stain	Negative	IDA
	Cell shape	Rod	IDA
	Motility	Non-motile	IDA
	Sporulation	Non-sporulating	IDA
	Temperature range	Mesophilic	IDA
	Optimum temperature	37°C	IDA
MIGS-22	Oxygen requirement	Aerobic	IDA
	Carbon source	Unknown	IDA
	Energy source	Unknown	IDA
MIGS-6	Habitat	*Mastomys erythroleucus* bloodstream	IDA
MIGS-15	Biotic relationship	Facultative intracellular	IDA
	Pathogenicity	Unknown	
	Biosafety level	3	
MIGS-14	Isolation	*Mastomys erythroleucus*	IDA
MIGS-4	Geographic location	Senegal	IDA
MIGS-5	Sample collection	February 2013	IDA
MIGS-4.2	Latitude	14°03'N	IDA
MIGS-4.3	Longitude	15°31'W	IDA
MIGS-4.4	Altitude	8 m	IDA

[a]Evidence codes - IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (*i.e.*, a direct report exists in the literature); NAS: Non-traceable Author Statement (*i.e.*, not directly observed for the living, isolated sample but based on a generally accepted property for the species or anecdotal evidence). Evidence codes come from the Gene Ontology project [10]. If

the evidence is IDA, then the property was directly observed for a live isolate by one of the authors or an expert mentioned in the acknowledgements.

Table 2 : Project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	High-quality draft
MIGS-28	Libraries used	One paired-end 3-kb library
MIGS-29	Sequencing platforms	454 GS FLX Titanium
MIGS-31.2	Fold coverage	30×
MIGS-30	Assemblers	Newbler version 2.5.3
MIGS-12	Gene calling method	Prodigal
	Genbank ID	GCA_900185775
MIGS-13	Project relevance	Biodiversity of *Bartonella* spp. in rodents from Senegal

Table 3: Number of gene associated with the 25 general COG Functional categories.

Code	Value	% of total	Description
[J]	157	9.38	Translation
[A]	0	0	RNA processing and modification
[K]	58	3.46	Transcription
[L]	73	4.36	Replication, recombination and repair
[B]	0	0	Chromatin structure and dynamics
[D]	17	1.02	Cell cycle control, mitosis and meiosis
[Y]	0	0	Nuclear structure
[V]	21	1.25	Defense mechanisms
[T]	37	2.21	Signal transduction mechanisms
[M]	74	4.42	Cell wall/membrane biogenesis
[N]	4	0.24	Cell motility
[Z]	0	0	Cytoskeleton
[W]	0	0	Extracellular structures
[U]	42	2.51	Intracellular trafficking and secretion
[O]	74	4.42	Posttanslational modification, protein turnover, chaperones
[X]	25	1.49	Mobilome: prophages, transposons
[C]	74	4.42	Energy production and conversion
[G]	56	3.35	Carbohydrate transport and metabolism
[E]	106	6.33	Amino acid transport and metabolism
[F]	47	2.81	Nucleotide transport and metabolism
[H]	63	3.76	Coenzyme transport and metabolism
[I]	44	2.63	Lipid transport and metabolism
[P]	57	3.41	Inorganic ion transport and metabolism
[Q]	15	0.89	Secondary metabolites biosynthesis, transport and catabolism
[R]	74	4.42	General function prediction only
[S]	68	4.06	Function unknown
_	603	36.02	Not in COGs

Table 4: Nucleotide content and gene count levels of the genome.

Attribute	Genome (Total)	
	Value	**% of total[a]**
Size (bp)	2,044,960	100
G+C content (bp)	785,960	38.44
Coding region	1,555,569	76.07
Total gene	1,716	100
RNA genes	42	2.45
Protein-coding genes	1,674	100
Protein assigned to COGs	1,071	63.99
Protein with peptide signals	263	15.71
Genes with transmembrane helices	372	22.22

[a)] The total is based on either the size of the genome in base pairs of the total of protein coding genes in the annotated genome.

Table 5: The numbers of orthologous protein shared between genomes (upper right) [a]

	B. vinsonii* subsp *berkhoffii	***B. rattaustraliani***	***B. florencae***	***B. tribocorum***	***B. queenslandensis***	***B. elizabethae***	***B. mastomydis***
B. vinsonii subsp. berkhoffii	**1,434**	1,115	1,121	1,154	1,043	1,143	1,144
B. rattaustraliani	0.66	**1,943**	1,134	1,164	1,057	1,148	1,154
B. florencae	0.67	0.83	**1,886**	1,210	1,081	1,201	1,201
B. tribocorum	0.80	0.66	0.66	**2,295**	1,136	1,257	1,258
B. queenslandensis	0.66	0.82	0.83	0.70	**2,466**	1,114	1,115
B. elizabethae	0.66	0.82	0.84	0.70	0.90	**1,663**	1,264
B. mastomydis	0.66	0.82	0.84	0.70	0.90	0.96	**1,674**

396

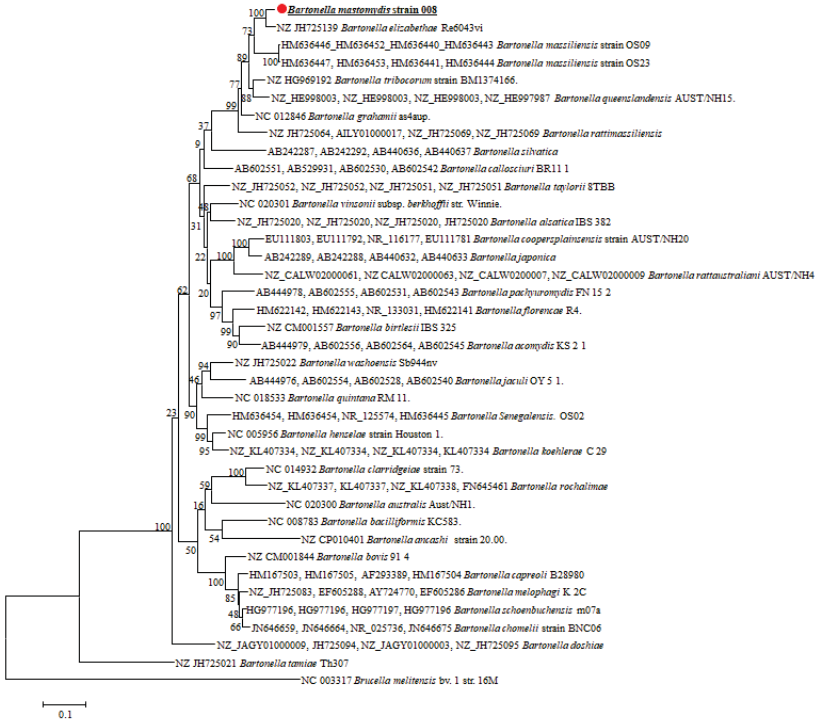
[a] Average percentage similarity of nucleotides corresponding to orthologous protein shared between genomes (lower left) and numbers of proteins per genome (bold).

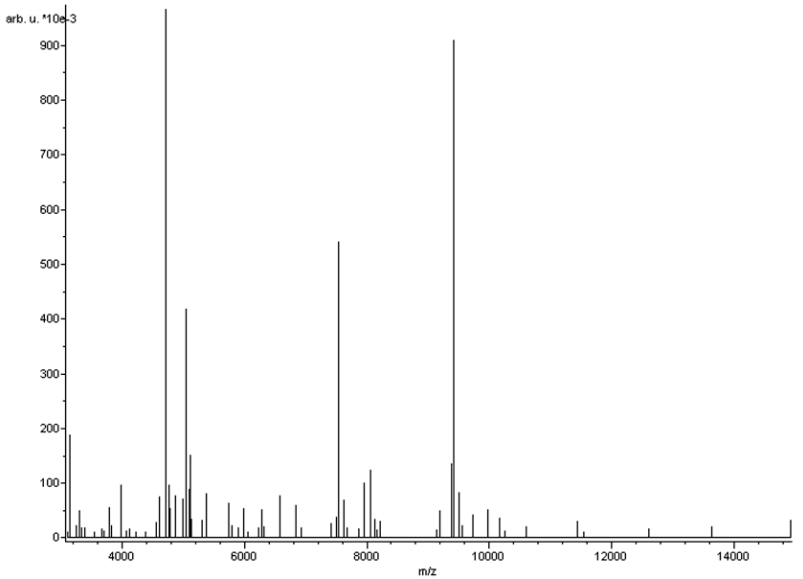
Table 6 : Pairwise comparison of *Bartonella mastomydis* with six other species using GGDC, formula 2 (DDH estimates based on identities / HSP length)[a]

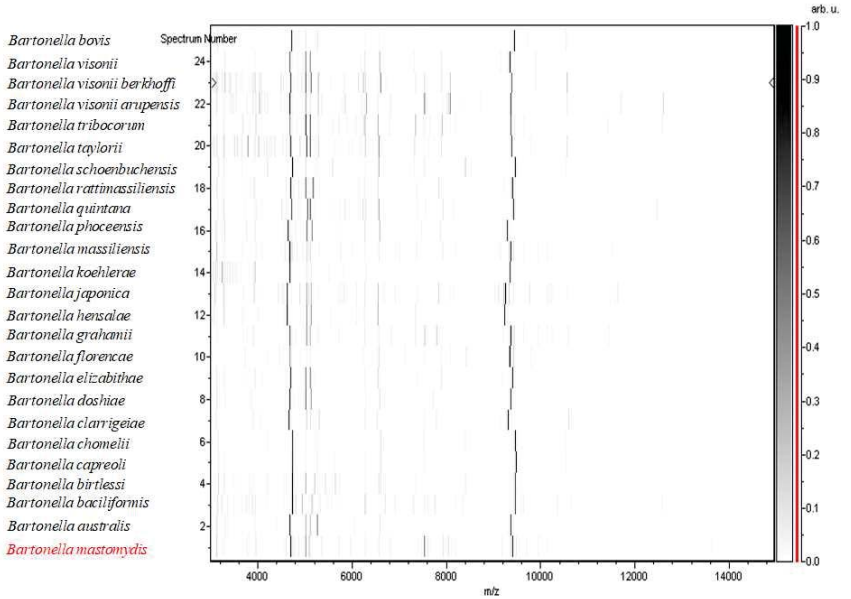
	B. vinsonii subsp berkhoffii	***B. rattaustraliani***	***B. florencae***	***B. tribocorum***	***B. queenslandensis***	***B. elizabethae***	***B. mastomydis***
B. vinsonii* subsp. *berkhoffii	100% ± 00	25.8% ± 2.45	27.1% ± 2.45	25.8% ± 2.4	25.9% ± 2.4	25.6% ± 2.4	25.5% ± 2.4
B. rattaustraliani		100% ± 00	25.5% ± 2.4	25.1% ± 2.4	27.5% ± 2.45	24.4% ± 2.4	24.2% ± 2.4
B. florencae			100% ± 00	26.7% ± 2.4	26.3% ± 2.45	26.8% ± 2.4	26.7% ± 2.4
B. tribocorum				100% ± 00	42% ± 2.55	37.3% ± 2.45	36.8% ± 2.5
B. queenslandensis					100% ± 00	37.6% ± 2.45	37.3% ± 2.5
B. elizabethae						100% ± 00	60.3% ± 2.8
B. mastomydis							100% ± 00

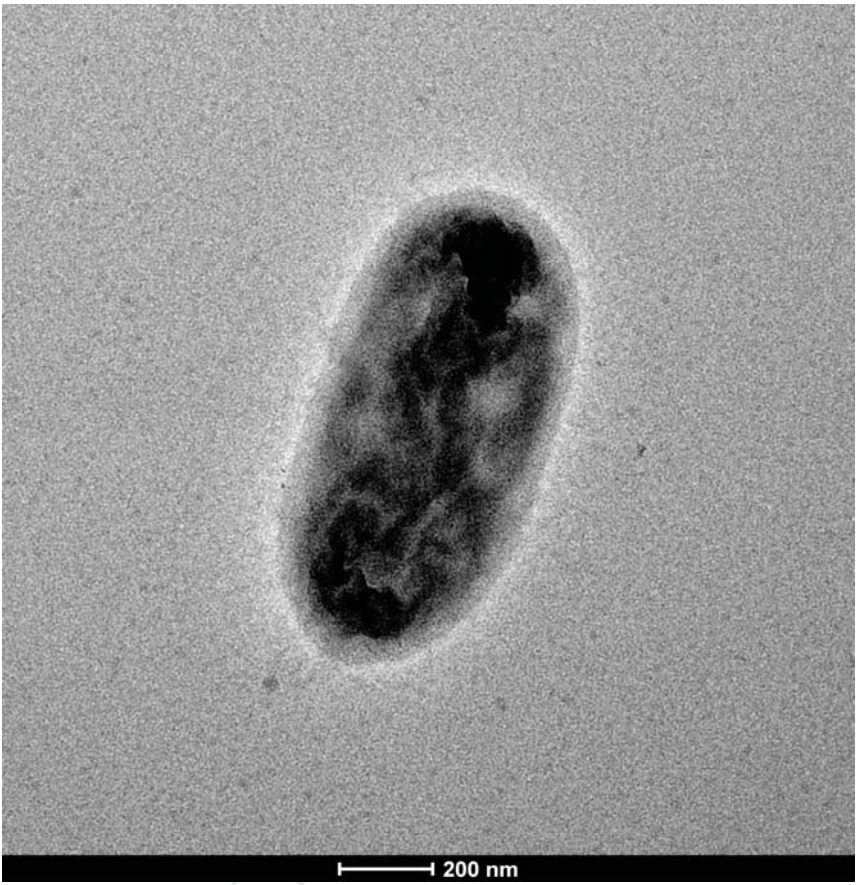
397

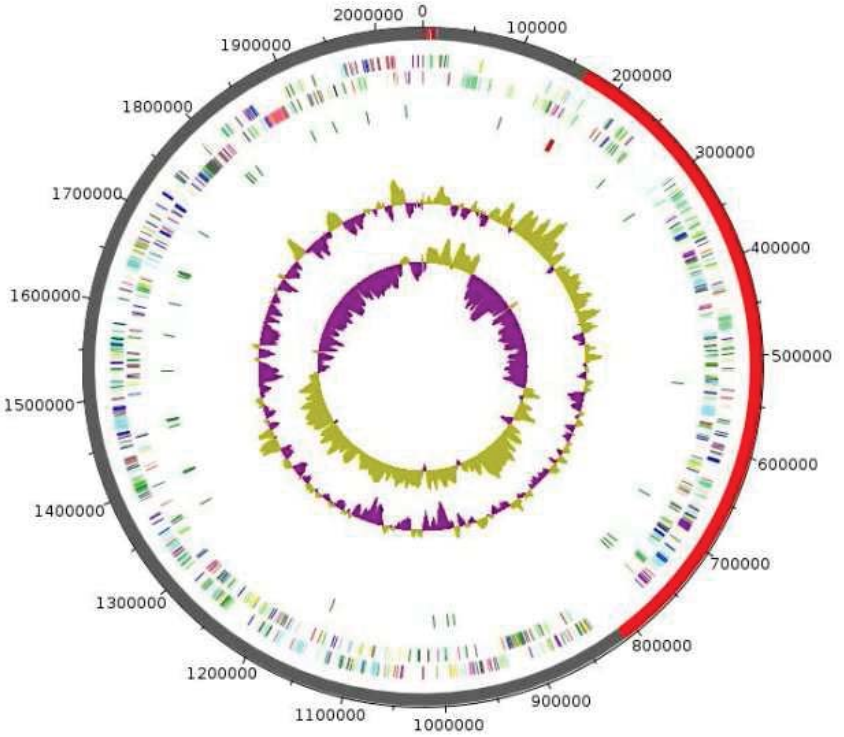
a) The confidence intervals indicate the inherent uncertainty in estimating DDH values from intergenomic distances based on models derived from empirical test data sets (which are always limited in size). These results are in accordance with phylogenomic analyses as well as the GGDC results.

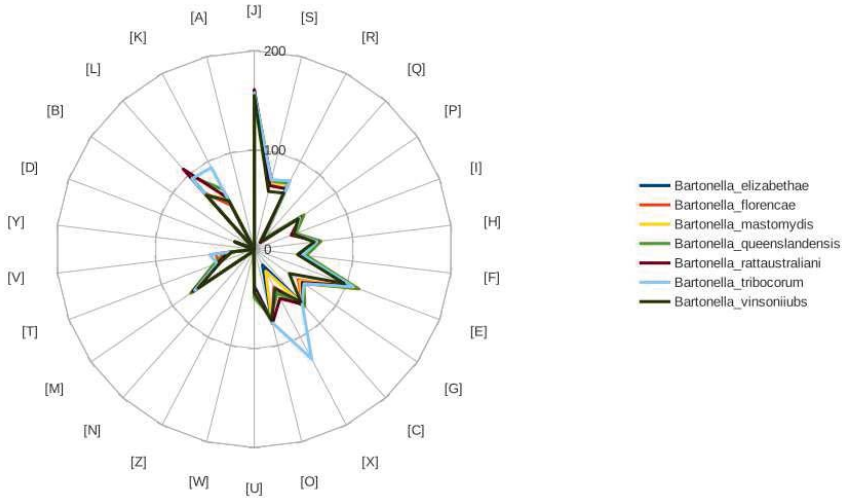

Bartonella mastomydis strain 008
NZ JH725139 Bartonella elizabethae Re6043vi
HM636446_HM636452_HM636440_HM636443 Bartonella massiliensis strain OS09
HM636447, HM636453, HM636441, HM636444 Bartonella massiliensis strain OS23
NZ HG969192 Bartonella tribocorum strain BM1374166.
NZ_HE998003, NZ_HE998003, NZ_HE998003, NZ_HE997987 Bartonella queenslandensis AUST/NH15.
NC 012846 Bartonella grahamii as4aup.
NZ JH725064, AILY01000017, NZ_JH725069, NZ_JH725069 Bartonella rattimassiliensis
AB242287, AB242292, AB440636, AB440637 Bartonella silvatica
AB602551, AB529931, AB602530, AB602542 Bartonella callosciuri BR11 1
NZ_JH725052, NZ_JH725052, NZ_JH725051, NZ_JH725051 Bartonella taylorii 8TBB
NC 020301 Bartonella vinsonii subsp. berkhoffii str. Winnie.
NZ_JH725020, NZ_JH725020, NZ_JH725020, JH725020 Bartonella alsatica IBS 382
EU111803, EU111792, NR_116177, EU111781 Bartonella coopersplainsensis strain AUST/NH20
AB242289, AB242288, AB440632, AB440633 Bartonella japonica
NZ_CALW02000061, NZ CALW02000063, NZ_CALW0200007, NZ_CALW02000009 Bartonella rattaustraliani AUST/NH4
AB444978, AB602555, AB602531, AB602543 Bartonella pachyuromydis FN 15 2
HM622142, HM622143, NR_133031, HM622141 Bartonella florencae R4.
NZ CM001557 Bartonella birtlesii IBS 325
AB444979, AB602556, AB602564, AB602545 Bartonella acomydis KS 2 1
NZ JH725022 Bartonella washoensis Sb944nv
AB444976, AB602554, AB602528, AB602540 Bartonella jaculi OY 5 1.
NC 018533 Bartonella quintana RM 11.
HM636454, HM636454, NR_125574, HM636445 Bartonella Senegalensis. OS02
NC 005956 Bartonella henselae strain Houston 1.
NZ_KL407334, NZ_KL407334, NZ_KL407334, KL407334 Bartonella koehlerae C 29
NC 014932 Bartonella clarridgeiae strain 73.
NZ_KL407337, KL407337, NZ_KL407338, FN645461 Bartonella rochalimae
NC 020300 Bartonella australis Aust/NH1.
NC 008783 Bartonella bacilliformis KC583.
NZ CP010401 Bartonella ancashi strain 20.00.
NZ CM001844 Bartonella bovis 91 4
HM167503, HM167505, AF293389, HM167504 Bartonella capreoli B28980
NZ_JH725083, EF605288, AY724770, EF605286 Bartonella melophagi K 2C
HG977196, HG977196, HG977197, HG977196 Bartonella schoenbuchensis m07a
JN646659, JN646664, NR_025736, JN646675 Bartonella chomelii strain BNC06
NZ_JAGY01000009, JH725094, NZ_JAGY01000003, NZ_JH725095 Bartonella dozhiae
NZ JH725021 Bartonella tamiae Th307
NC 003317 Brucella melitensis bv. 1 str. 16M
0.1








0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000



Article 19:

Non-contiguous finished genome sequence and description of *Raoultibacter massiliensis* gen. nov., sp. nov. and *Raoultibacter timonensis* sp. nov., two new bacterial species isolated from the human gut

Traore SI, Bilen M, Beye M, Diop A, Yasir M, I Azhar E, Fonkou Mbogning M, Tall ML, Michelle C, Bibi F, Bittar F, Jiman-Fatani AA, Daoud Z, Cadoret F, Fournier PE, Edouard S

[Submitted in MicrobiologyOpen]

Non-contiguous finished genome sequence and description of *Raoultibacter massiliensis* gen. nov., sp. nov. and *Raoultibacter timonensis* sp. nov, two new bacterial species isolated from the human gut

Running title: *Raoultibacter massiliensis* and *Raoultibacter timonensis* gen. nov., sp. nov.

Sory Ibrahima TRAORE[a★], Melhem BILEN[a,b★], Mamadou BEYE[c], Awa DIOP[c], Muhammad YASIR[d], Esam Ibraheem AZHAR[d,e], Maxime DESCARTES MBOGNING FONKOU[a], Mamadou Lamine TALL[a], Caroline MICHELLE[a], Fehmida BIBI[d], Fadi BITTAR[a], Asif Ahmad JIMAN-FATANI[f], Ziad DAOUD[f], Fréderic CADORET[a], Pierre-Edouard FOURNIER[c], Sophie EDOUARD[a*]

[a] Aix Marseille Univ, UMR MEPHI, Aix-Marseille Université, IRD, APHM, IHU Méditerranée-Infection, Marseille, France

[b] Clinical Microbiology Department, Faculty of Medicine and Medical sciences, University of Balamand, POBox:33, Amioun, Lebanon

[c] Aix Marseille Univ, UMR VITROME, IRD, Aix-Marseille Université, AP-HM, SSA, IHU Méditerranée-Infection, Marseille, France

[d] Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

[e] Medical Laboratory Technology Department, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

[f] Department of Medical Microbiology and Parasitology, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia

* Corresponding author. E-mail address: sophie.edouard@univ-amu.fr

★ The authors contributed equally to this article.

Keywords: Culturomics; taxonogenomics; *Raoultibacter massiliensis*; *Raoultibacter timonensis*; new bacterial species; human gut microbiota

Abstract

As part of the culturomics project aiming at describing the human microbiome, we report in this study the description of the new bacterial genus *Raoultibacter* gen. nov. that includes two new species, *i. e.*, *Raoultibacter massiliensis* sp. nov. and *R. timonensis* sp. nov. The *R. massiliensis* type strain Marseille-P2849^{T} was isolated from the fecal specimen of a healthy 19-year-old Saudi Bedouin while *R. timonensis* type strain Marseille-P3277^{T} was isolated from the feces of an 11-year-old pygmy female living in Congo. Strain Marseille-P2849^{T} exhibited 91.4% 16S rRNA sequence similarity with *Gordonibacter urolithinfaciens*, its phylogenetic closest neighbor with a validly published name. Strain Marseille-P3277^{T} exhibited 97.96% 16S rRNA similarity with strain Marseille-P2849^{T}. These novel Gram-negative, motile, non spore-forming coccobacilli form transparent micro-colonies on blood agar in both anaerobic and microaerophilic atmospheres and belong to the family *Eggerthellaceae*. The genome sizes of these strains were 3,657,161 bp and 4,000,215 bp, and their G+C contents were 59.02 and 59.9 mol%, respectively. Using a taxono-genomic approach combining the phenotypic, biochemical, phylogenetic and genomic characteristics, we propose the creation of the genus *Raoultibacter* gen. nov., which contains strains Marseille-P2849^{T} (= CSUR P2849^{T} = DSM 103407^{T}) and Marseille-P3277^{T} (=CCUG 70680, =CSUR P3277) as type strains of the species *Raoultibacte massiliensis* sp. nov and *R. timonensis* sp. nov., respectively.

1. INTRODUCTION

The human microbiota is a highly diverse consortium of microbes colonizing different regions of the human body. The role of the microbiota has generated an important interest in the scientific and medical communities as it was demonstrated to be involved in human health (Alegre et al. 2014;Glenwright et al. 2017;Honda and Littman 2016;Round and Mazmanian 2009). A dysbiosis of the microbiota has been proven to be implicated in a growing number of pathologies and its modulation can have benefic impacts on the host (Smits et al. 2013;Zak-Golab et al. 2014). Over the past decade, great advances have been achieved by the development of next-generation DNA sequencing technologies, which have allowed for considerable progress in the study of different ecosystems including the intestinal microbiota, which is the most studied human microbiota (Margulies et al. 2005). However, many drawbacks appear when using these molecular methods, such as the inability to distinguish between dead and living bacteria and the depth bias that neglects a minority but important bacterial species (Lagier et al. 2012). Consequently, a new approach named "culturomics" was developed in our laboratory in order to exhaustively explore the microbial ecosystems and to increase the chance of isolating previously uncultured bacteria (Lagier et al. 2015b;Lagier et al. 2016;Lagier and Raoult 2016). Culturomics relies on the multiplication of culture conditions (including the variation of temperature, media, atmosphere…) along with a rapid bacterial identification method by the means of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS). The latter proved its efficiency in describing the human gut microbiota by reporting a significant number of previously uncultured and novel bacterial species (Lagier et al 2016). Nevertheless, we are still far from understanding the human microbiome since only around 2,000 human bacterial species have been isolated, knowing that up to 10^{12} bacteria are estimated to be present in only 1g of stool (Hugon et al. 2015;Wu and Lewis 2013). In the present work, the two

understudied organisms, strains Marseille-P2849^T and Marseille-P3277^T, were isolated from the stool samples of a19-year-old healthy Saudi Bedouin and an 11-year-old Congolese pygmy female, respectively. These bacteria were not identified using MALDI-TOF-MS and the sequencing and phylogenetic analysis of their 16S rRNA genes classified them as members of a new genus within the family *Eggerthellaceae* (Gupta et al. 2013). This family contains the type genus *Eggerthella* and the genera *Adlercreutzia, Asaccharobacter, Cryptobacterium, Denitrobacterium, Enterorhabdus, Gordonibacter, Paraeggerthella* and *Slackia* (Gupta, Chen, Adeolu, & Chai 2013). Among its members, *Eggerthella lenta* is commonly found in humans, and has been associated with bacteremia in patients with intra-abdominal and gastrointestinal tract pathologies and bacteremia complicated by spondylodiscitis, psoas abscess, and meningitis (Gardiner et al. 2014;Gardiner et al. 2015;Wong et al. 2014). We herein describe the new genus *Raoultibacter* gen. nov. within the family *Eggerthellaceae* using the taxono-genomic approach including phenotypic, biochemical and genomic characteristics of studied strains (Fournier et al. 2015;Kokcha et al. 2012;Lagier et al. 2013;Seck et al. 2016). Strain Marseille-P2849^T (= CSUR P2849 = DSM 103407) is the type strain of the new species *Raoultibacter massiliensis* sp. nov and Marseille-P3277^T is the type strain of the species *Raoultibacter timonensis* sp. nov (=CCUG 70680, =CSUR P3277).

2. METHODS AND MATERIALS

2.1. Ethical requirements and sample collection

Strain Marseille-P2849 ^{T} was isolated in April 2016 from the stool sample of a 19-year-old healthy Bedouin male living in Saudi Arabia and strain Marseille-P3277^{T} was isolated in June 2016 from the stool specimen of an 11-year-old healthy Pygmy female living in Congo. The fecal specimens were preserved at 4°C after collection and were sent to Marseille, where they were stored frozen at -80°C until laboratory culture isolation. The donors gave a signed informed consent, and the study was validated by the ethics committee of the Institut Federatif de Recherche 48 under number 09-022.

2.2. Isolation of the strains

For the initial cultivation of the bacteria, stool samples were diluted with phosphate-buffered saline (Life Technologies, Carlsbad, CA, USA) and multiple culture conditions were performed as previously described (Lagier, et al 2012;Lagier et al. 2015a). We observed the first isolation of *R. massiliensis* when the sample collected from the Bedouin male was incubated in an anaerobic blood culture bottle (Becton-Dickinson, BACTEC Plus anaerobic/F Media, Le pont de Claix, France) supplemented with 5 mL filter-sterilized rumen for 7 days at 37°C. Then, we observe the first *R. timonensis* isolation when the sample collected from the Pygmy female was incubated in a similar blood culture bottle supplemented with 5ml sterile sheep blood and 5mL filtered rumen for 2 days at 37°C. Then, following the inoculation of each liquid culture on 5% sheep blood-enriched agar and incubation at 37°C under anaerobic condition using AnaeroGen (bioMérieux), the initial growth of strains Marseille-P2849^{T} and Marseille-P3277^{T} was detected after 4 and 2 days, respectively.

2.3. Strain identification by MALDI-TOF-MS and 16S rRNA gene sequencing

Identification of bacterial colonies was attempted using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) analysis as

previously described (Lagier et al. 2013). When MALDI-TOF MS failed to identify the new organisms (score <1.7), 16S rRNA gene sequencing was performed using the fD1 and rP2 primers as previously described (Drancourt et al. 2000), a GeneAmp PCR System 2720 thermal cycler (Applied Bio systems, Bedford, MA, USA) and an ABI Prism 3130-XL capillary sequencer (Applied Biosciences, Saint Aubin, France). Each 16S rRNA sequence was compared with the nr database of the National Center for Biotechnology Information using the BLAST software (https://blast.ncbi.nlm.nih.gov). Compared to its phylogenetically closest species with standing in nomenclature, a 95% similarity threshold was used to define a new genus and a 98.65% similarity threshold was used to define a new species (Meier-Kolthoff et al. 2013b;Tindall et al. 2010;Yarza et al. 2014). The mass spectrum and 16S rRNA sequence of the newly isolated species were submitted in the URMITE (http://www.mediterranee-infection.com/article.php?laref=256&titre=urms-database) and EMBL-EBI databases, respectively.

2.4 Phylogenetic tree

For phylogenetic analysis, sequences of the phylogenetically closest species were obtained after performing a BLASTn search within the 16S rRNA database of “The All-Species Living Tree" Project of Silva (The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks 2017). Alignment was performed using CLUSTALW (Thompson et al. 1994) and MEGA software (Kumar et al. 1994) was used for phylogenetic inferences generation using the maximum likelihood method.

2.5. Morphologic observation and growth conditions

Following Gram staining, bacterial cells were observed using a Leica DM 2500 photonic microscope (Leica Microsystems, Nanterre, France) with a 100X oil immersion lens. The motility of the bacterium was assessed using a Leica DM 1000 photonic microscope (Leica Microsystems) at a 100 X magnification. A Tecnai G20 (FEI company, Limeil-Brevannes,

France) electron microscope was used for bacterial cell imaging at an operating voltage of 60kV, as previously described (Elsawi et al. 2017).

Culture of strains Marseille P2849^T and Marseille P3277^T was attempted using several growth conditions in order to determine the optimal ones. Culture assays were performed on 5% sheep blood-enriched Columbia agar (bioMerieux) under anaerobic and microaerophilic conditions using GENbag Anaer and GENbag Microaer systems, respectively (BioMérieux, Marcy-l'Étoile, France), and under aerobic conditions, with or without 5% of CO_2. Different growth temperatures (25, 28, 37, 45, 55°C) and pH values (6-8.5) were also tested. Finally, NaCl tolerance was tested using a range of 5-100g/L NaCl concentrations on 5% sheep blood-enriched Schaedler agar (BioMérieux) in anaerobic conditions.

2.6. Biochemical analysis, Fatty acid methyl ester analysis and antibiotic susceptibility testing

Biochemical characteristics of the strains were investigated using API ZYM, 20A and 50CH strips (BioMérieux) according to the manufacturer's instructions. A 20-minute-thermic shock of fresh colonies at 80°C was done in order to test sporulation. Catalase (BioMérieux) activity was determined in 3% hydrogen peroxide solution and oxidase activity was assessed using an oxidase reagent (Becton-Dickinson).

Cellular fatty acid methyl ester (FAME) analysis was performed by gas chromatography/mass spectrometry (GC/MS). Two samples were prepared with approximately 17 mg of bacterial biomass per tube for strain Marseille-P2849^T and 5 mg per tube for strain Marseille-P3277^T. Briefly, fatty acid methyl esters were separated using an Elite 5-MS column and monitored by mass spectrometry (Clarus 500 - SQ 8 S, Perkin Elmer, Courtaboeuf, France) as previously described (Dione et al. 2016;Myron Sasser 2006). Spectral database search was performed using MS Search 2.0 operated with the Standard Reference Database 1A (NIST, Gaithersburg, USA) and the FAMEs mass spectral database (Wiley, Chichester, UK).

Antibiotic susceptibility was tested using the E-test gradient strip method (BioMerieux) to determine the minimal inhibitory concentration (MIC) of each tested antibiotic. Strains were grown on 5% sheep blood-enriched Columbia agar (bioMérieux) and a bacterial inoculum of turbidity 0.5 McFarland was prepared by suspending the culture in sterile saline solution (0.85% NaCl). Using cotton swabs, the inoculum was plated on 5% horse blood-enriched Mueller Hinton Agar (BioMerieux), E-test strips were deposited and the plates were incubated under anaerobic conditions for 48 hours (Citron et al. 1991;Matuschek et al. 2014). MICs were interpreted according to the 2017 EUCAST recommendations (Citron, Ostovari, Karlsson, & Goldstein 1991).

2.7. DNA extraction, genome sequencing and assembly

Genomic DNAs (gDNAs) of strains Marseille-P2849^{T} and Marseille-P3277^{T} were extracted in two steps. A mechanical treatment was first performed using acid-washed glass beads (G4649-500g Sigma) and a FastPrep BIO 101 instrument (Qbiogene, Strasbourg, France) at maximum speed (6.5) for 90s. Then after a 2-hour lysozyme incubation at 37°C, DNA was extracted on the EZ1 biorobot (Qiagen) with EZ1 DNA tissue kit according to the manufacturer's recommendations. Each gDNA was quantified by a Qubit assay with the high sensitivity kit (Life technologies, Carlsbad, CA, USA) to 69.9 and 107 ng/µl, respectively, and was sequenced using the MiSeq technology (Illumina Inc, San Diego, CA, USA) with the Mate-Pair strategy. Both gDNAs were barcoded in order to be mixed with 10 other projects with the Nextera Mate-Pair sample prep kit (Illumina).

Each Mate-Pair library was prepared with 1.5 µg of gDNA using the Nextera Mate-Pair Illumina guide. Both gDNAs were simultaneously fragmented and tagged with a Mate-Pair junction adapter. The fragmentation patterns were validated on an Agilent 2100 BioAnalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) with a DNA 7500 labchip. The DNA fragments ranged in size from 1.5 kb up to 11kb with optimal sizes at 8.345 and 6.291 kb,

192 respectively, for strains Marseille-P2849^{T} and Marseille-P3277^{T}, respectively. No size
193 selection was performed and 600ng of tagmented fragments were circularized for strain
194 Marseille-P2849^{T} and 404.1 ng for strain Marseille-P3277^{T}. The circularized DNAs were
195 mechanically sheared to small fragments with an optimal size at 960 bp on the Covaris device
196 S2 in T6 tubes (Covaris, Woburn, MA, USA).The library profiles were visualized on a High
197 Sensitivity Bioanalyzer LabChip (Agilent Technologies Inc, Santa Clara, CA, USA) and the
198 final concentration libraries were measured at 12.3 and 3.9 nmol/l for strains Marseille
199 P2849^{T} and Marseille P3277^{T}, respectively.
200 The libraries were normalized at 2nM and pooled. After a denaturation step and dilution at 15
201 pM, the pool of libraries was loaded onto the reagent cartridge and then onto the instrument
202 along with the flow cell. Automated cluster generation and sequencing run were performed in
203 a single 39-hour run in a 2x151-bp.
204 For strain Marseille P2849 ^{T}, total information of 4.5 Gb was obtained from a 477K/mm2
205 cluster density with a cluster passing quality control filters of 94.8 % (8,444,000 passing filter
206 paired reads). Within this run, the index representation for strain Marseille-P2849^{T} was
207 determined to be of 8.34 %. For strain Marseille-P3277^{T}, total information of 6.3 Gb was
208 obtained from a 673K/mm2 cluster density with a cluster passing quality control filters of
209 95.4% (12,453,000 clusters). Within this run, the index representation for this strain was
210 determined to be of 7.29%. The 769,472 and 907,611 paired reads of strains Marseille-P2849^{T}
211 and Marseille-P3277^{T}, respectively, were trimmed, assembled, annotated and analyzed using
212 the same pipeline adapted in our previous studies (Elsawi et al. 2017).
213 **2.8. Genome annotation and analysis**
214 Prodigal was used for Open Reading Frame (ORF) prediction (Hyatt et al. 2010) with default
215 parameters. We excluded predicted ORFs spanning a sequencing gap region (containing N).
216 The bacterial proteome was predicted using BLASTP (E-value of 1e 03, coverage of 0.7 and

identity percent of 30) against the Clusters of Orthologous Groups (COGs) database. If no hit was found we searched against the nr database (Clark et al. 2016) using BLASTP with an E-value of 1e03, coverage 0.7 and an identity percent of 30. An E-value of 1e05 was used if the length of sequences was smaller than 80 amino acids. Pfam conserved domains (PFAM-A and PFAM-B domains) were searched on each protein with the hhmscan tools analysis. RNAmmer (Lagesen et al. 2007) and tRNAScanSE tool (Lowe and Eddy 1997) were used to find ribosomal rRNAs genes and tRNA genes respectively. ORFans were identified if all the BLASTP performed had negative results (E-value inferior to 1e03 for ORFs with sequence size above 80 aa or E-value inferior to 1e05 for ORFs with sequence length smaller than 80 aa). For data management and visualization of genomic features, Artemis (Carver et al. 2012) and DNA Plotter (Carver et al. 2009) were used, respectively. We used the MAGI in-house software to analyze the mean level of nucleotide sequence similarity at the genome level. It calculated the average genomic identity of gene sequences (AGIOS) among compared genomes (Ramasamy et al. 2014). This software combines the Proteinortho software (Lechner et al. 2011) for detecting orthologous proteins in pairwise genomic comparisons. Then the corresponding genes were retrieved and the mean percentage of nucleotide sequence identity among orthologous ORFs was determined using the Needleman-Wunsch global alignment algorithm.

We also used the Genome-to-Genome Distance Calculator web service to calculate digital DNA:DNA hybridization estimates (dDDH) with confidence intervals under recommended settings (Formula 2, BLASTp) (Auch et al. 2010;Meier-Kolthoff et al. 2013a).

3. Results

3.1. Strain identification by MALDI-TOF-MS and 16S rRNA sequencing

MALDI-TOF-MS failed to identify strains Marseille-P2849^{T} and P3277 ^{T} at the genus and species levels (score <1.7). The spectra of strain Marseille-P2849^{T} and Marseille-P3277^{T}

were added to our URMS database. Close species, on the basis of 16S rRNA phylogenetic analysis and their presence in our MALDI-TOF MS spectrum database, were compared at the protein level with strains Marseille-P2849^T and Marseille-P3277 ^{T} and represented in a gel view (Figure 1). Mass spectrum of each organism was unique and did not match any other spectrum, confirming the novelty of both studied strains.

The16S rRNA gene from strain Marseille-P2849^T exhibited a 91.4% identity with *Gordonibacter urolithinfaciens* strain Marseille-AA00211^T (GenBank accession number LT223667), the phylogenetically closest species with standing in nomenclature (Figure 2). According to the criteria defined by Kim *et al*. (Kim et al. 2014), a new genus can be defined by a similarity level threshold lower than 95%, thus putatively classifying strain Marseille-P2849^T as a member of a new genus within the family *Eggerthellaceae*, for which we proposed the name *Raoultibacter*. Furthermore, two months later, when performing phylogenetic analyses for strain Marseille-P3277^T, we found that it exhibited a 97.96% sequence similarity with strain Marseille-P2849^T, enabling us to classify it as a putative new species within the *Raoultibacter* genus. The 16S rRNA sequences of strains Marseille-P2849^T and Marseille-P3277^T were deposited in EMBL-EBI under accession numbers LT576395 and LT623894, respectively.

3.2 Phenotypic characteristics and biochemical features

Strains Marseille-P2849^T and Marseille-P3277^T form translucent micro-colonies on 5% sheep blood-enriched Columbia agar (bioMérieux) with a mean diameter ranging from 0.1 to 0.4 mm. The growth of both strains was observed in anaerobic and microaerophilic atmospheres at 28, 37 and 45°C but optimal growth occurred under anaerobic conditions at 37°C after 48 hours of incubation. No growth was obtained at 55°C or in aerobic atmosphere. Bacterial cells were motile, Gram-negative (Figure 3a, 3b) and non spore-forming coccobacilli. Electron microscopy revealed that cells from strain Marseille-P2849^T ranged from 0.8 to 1.2-µm long

with a mean diameter ranging from 0.4 to 0.6µm (Figure 3c, 3d) while cells from strain Marseille-P3277^{T}were 1 to 2-µm long with a mean diameter ranging from 0.35 to 0.44µm. Strain Marseille-P2849^T was found to be catalase-positive and oxidase-negative but strain Marseille-P3277^T was both catalase-and oxidase-negative. Both strains tolerated pH levels ranging between 6 and 8.5 and could not sustain NaCl concentration > 5g/L. The classification and general features of strains Marseille-P2849^T and Marseille-P3277^T are summarized in Table 1.

Using an API® 50CH strip (bioMérieux), positive reactions were observed for both strains for glycerol, D-Ribose, D-Galactose, D-Glucose, D-Fructose, D-Mannose, D-Mannitol, D-Sorbitol, N-Acetylglucosamine, Amygdaline, Arbutine, Esculin ferric citrate, Salicine, D-Maltose, D-Lactose, D-Saccharose, D-Trehalose, D-Melezitose, Gentiobiose, D-Tagalose and potassium Gluconate. In addition, positive reactions were observed for strain Marseille-P2849^T with amidon and potassium 5-Cetogluconate, and for strain Marseille-P3277^T with methyl- αD-glucosamine, D-cellobiose and D-turanose (Table 2). Negative reactions were observed for both strains for Erythritol, D-Arabinose, L-Arabinose, D-Xylose, L-Xylose, D-Adonitol, Methyl-βD-Xylopyranoside, L-Sorbose, L-Rhamnose, Dulcitol, Inositol, Methyl-αD-Mannopyranoside, Methyl-αD-Glucopyranoside, D-Cellobiose, D-Melibiose, Inulin, D-Raffinose, Glycogen, Xylitol, D-Turanose, D-Xylose, D-Fucose, L-Fucose, D-Arabitol, L-Arabitol and Potassium 2-CetoGluconate.

Using an API® 20A strip (bioMérieux), both strains produced indole and positive reactions were observed for D-glucose, D-Mannitol, D-lactose, D-Saccharose, D-Maltose, Salicine, L-Arabinose, Gelatine, D-Mannose, Esculin ferric citrate, D-Cellobiose D-Melezitose, D-Rafinose, D-sorbitol and D-Trehalose. In addition, a positive reaction was observed for strain Marseille-P3277^T, but not Marseille-P2849^T, with L-Rhamnose. No reaction was obtained for urease and D-xylose for both strains.

Using an API® ZYM strip (bioMérieux), both strains exhibited esterase (C4), esterase lipase (C8), Lipase (C14), Leucine arylamidase, Valine arylamidase, Cystine arylamidase, phosphatase acid and naphtol phosphohydrolase activities but no phosphatase alkaline was observed. In addition, positive reactions were observed for strain Marseille-P3277^T with trypsin, α-chymotrypsin, α-galactosidase, β-galactosidase, β-glucuronidase, α-glucosidase, β-glucosidase, N-acetyl-β-glucosaminidase, α-mannosidase. An α-fucosidase activity was observed only for strain Marseille-P2849^T.

The major fatty acids identified for strains Marseille-P2849^T and Marseille-P3277^T were 9-Octadecenoic acid (l8:ln9, 36 % and 38%, respectively), Hexadecanoic acid (16:0, 18% and 25%) and Tetradecanoic acid (l4:0, 13% and 11%) (Table 3). Strain Marseille-P3277^T exhibited unusually long chain fatty acids (C20:4n6 and C20:5n3).

Among tested antibiotics, strains Marseille-P2849^T and Marseille-P3277^T were susceptible to amoxicillin (MIC 0.50 µg/mL and 1 µg/mL, respectively), imipenem (0.047 mg/mL and 0.047 µg/mL), metronidazole (0.023 µg/ml and 0.064 µg/ml), rifampicin (0.003 µg/ml and 0.008) and erythromycin (0.32 µg/ml and 0.016 µg/ml) but were resistant to daptomycin, minocycline, amikacin, vancomycin and cefotaxime.

3.3. Genomic properties

The draft genome of strain Marseille-P2849^T is 3,657,161-bp long with a G+C content of 59.02 % (Table 4; Figure 4a). It is composed of 9 scaffolds (35 contigs). Of the 3,073 predicted genes, 3,025 were protein-coding genes and 48 were RNAs (1 complete rRNA operon and 45 tRNA genes). A total of 2,365 proteins (76.86 %) were assigned to COGs and 253 genes were identified as ORFans (8.23%). Six genes were associated to polyketide synthases (PKS) or non ribosomal peptide synthetases (NRPS) (0.18%) and 470 genes were associated to virulence (15.29%). Regarding strain Marseille-P3277^T, the genome size was 4,000,215-bp long with a 59.9% G+C content (Figure 4b). It is composed of 21 scaffolds

(composed of 84 contigs). Of the 3,284 predicted genes, 3,232 were protein-coding genes and 52 were RNAs (1 complete rRNA operon and 49 tRNA genes). A total of 2,562 proteins (78.01%) were assigned to COGs and 323 genes were identified as ORFans (9.83%). The genome of strain Marseille-P3277^T contained 14 genes associated to PKS or NRPS (0.45%) and 481 genes associated to virulence (14.64%). The genome statistics are presented in Table 4 and the distribution of genes into COGs functional categories is summarized in Table 5.

3.3. Genomic comparison

The draft genome sequence structure of strains Marseille-P2849^T and Marseille-P3277^T are summarized in Figure 4. The draft genome sequence of strain Marseille-P2849^T is larger than that of *Atopobium fossor, Denitrobacterium detoxificans, Atopobium parvulum*, *Olsenella profusa, Olsenella uli, Eggerthella lenta* and *Gordonibacter pamelaeae* (1.66, 2.45, 1.54, 2.72, 2.05, 3.63 and 3.61 Mb, respectively) but smaller than that of strain Marseille-P3277^T (3.94 Mb, Table 6). The G+C content of strains Marseille-P2849 ^{T} and Marseille-P3277^T are larger than those of *A. fossor* and *A. parvulum* (59.02 and 59.9 versus 45.4 and 45.7, respectively), but smaller than those of *D. detoxificans, G. pamelaeae, E. lenta, O. profusa* and *O. uli* (59.5, 64.0, 64.2, 64.2 and 64.7%, respectively). The gene content of strain Marseille-P2849^T is smaller than that of strain Marseille-P3277^T (3,073 and 3,284 respectively), but larger than that of *A. fossor, G. pamelaeae, D. detoxificans, A. parvulum, O. profusa* and *E. lenta* (1,487, 2,027, 1,762, 1,353, 2,650 and 3,070, respectively). The distribution of functional classes of predicted genes of strains Marseille-P2849^T and Marseille-P3277^T according to the clusters of orthologous groups (COGs) of proteins is summarized in Figure 5.

Strain Marseille-P2849^T shared 1,542, 555, 571, 1,069, 693, 683, 1,084, 1,404 and 911 orthologous proteins with strain Marseille-P3277^T, *A. parvulum, A. fossor, A. equolifaciens, O. umbonata, O. profusa, G. pamelaeae, E. lenta* and *D. detoxificans,* respectively. The

AGIOS values among the 8 most closely related species ranged between 58.12% and 81.35%. When compared to these eight species, strain Marseille P2849^T AGIOS values ranging from 58.97% with *A. fossor* to 73.75% with *G. pamelaeae*. Similarly, strain Marseille P3277^T exhibited AGIOS values ranging from 58.95% with *A. fossor* to 74.19% with *G. pamelaeae* (Table 7). The AGIOS values obtained for strains Marseille P2849^T and Marseille P3277^T, between 58.12 and 81.35%, support their new species status.

In addition, dDDH values obtained between strain Marseille-P2849^T, strain Marseille-P3277^T, *A. parvulum, A. fossor, A. equolifaciens, O. umbonata, O. profusa, G. pamelaeae, E. lenta* and *D. detoxificans* were of 25.2% [22.9 -27.7], 28.1% [25.8-30.6], 30.7% [28.3-33.2], 20.3% [18.1-22.8%], 20.8% [18.6-23.3], 18.6% [16.5-21], 24.5% [22.2-27], 23.6% [21.3-26.1] and 19.1% [16.9-21.5], respectively (Table 8). These dDDH values were lower than the 70% value threshold for species demarcation, thus confirming that the two studied strains are representative of new species (Meier-Kolthoff et al. 2013c).

4. Discussion

Culturomics is a high-throughput culture approach that enabled the isolation of approximately 2,872 bacterial species including 247 new species from the human gut in our laboratory (Lagier et al. 2017). Along with the development of culturomics, a new polyphasic approach, taxonogenomics, was developed in order to describe novel bacterial species using their biochemical, proteomic and genomic properties (Fournier, Lagier, Dubourg, & Raoult 2015;Kokcha, Ramasamy, Lagier et al. 2012;Lagier et al. Fournier 2013;Seck et. 2016). This approach has the advantage of exhibiting a higher inter- and intra-laboratory reproducibility when compared to DNA-DNA hybridization and chemotaxonomic methods (Fournier, Lagier, Dubourg, & Raoult 2015). Based on MALDI-TOF MS analysis, 16S rRNA gene sequence comparison (< 95% similarity), genome comparison, AGIOS and dDDH values, we propose the creation of the new genus *Raoultibacter* gen. nov. within the family *Eggerthellaceae* that

belongs to the phylum Actinobacteria. Members of this family belong to the class *Coriobacteria*. Many revisions have been made to the classification of this group by using various molecular techniques and Gupta *et al.* proposed the taxonomic division of this class into two orders (*Coriobacteriales* and *Eggerthellales*) and three families including *Coriobacteriaceae*, *Atopobiaceae* and *Eggerthellaceae* (Gupta, Chen, Adeolu, & Chai 2013). Members of the latter family are predominantly anaerobic, non-spore forming, catalase-positive and Gram-positive rods or cocci. However, strains Marseille-P2849^{T} and Marseille-P3277^{T} are Gram-negative (Lau et al. 2004;Selma et al. 2014;Wurdemann et al. 2009). Most of the species closely related to the genus *Raoultibacter* gen. nov. were isolated from the human gut flora and, to date, exhibited a low pathogenicity (Gardiner, Korman, & Junckerstorff 2014;Lee et al. 2012).

Conclusion.

The biochemical, proteomic, genetic and genomic characteristics of strains Marseille-P2849^{T} and Marseille-P3277^{T} confirmed that they belong to two distinct species within a new genus in the family *Eggerthellaceae*, for which we propose the names *Raoultibacter* gen. nov., *Raoultibacter massiliensis* sp. nov. and *Raoultibacter timonensis* sp. nov. The type strain from *R. massiliensis* sp. nov., Marseille-P2849^{T}, was isolated from the feces of a 19-year-old healthy male Saudi Bedouin, whereas the type strain from *R. timonensis* sp. nov., Marseille-P3277^{T} was isolated from the feces of a healthy 11-year-old Pygmy female living in Congo.

5. Taxonomic and nomenclatural proposals

5.1 Description of *Raoultibacter* gen. nov.

Raoultibacter (ra.ou.l.ti.bac'ter. N.L. masc. n, "*Raoultibacter*", composed of *Raoult*, in honor of the French microbiologist Didier Raoult, founder of the IHU Mediterranée-Infection in Marseille and inventor of culturomics, the culture strategy that has enabled the discovery of more than 250 bacterial species, and *bacter*, for bacterium).

392 *Raoultibacter* forms transparent micro-colonies on blood agar with a mean diameter of 0.1-
393 0.3 mm. Cells are Gram-negative, non spore-forming, motile coccobacilli that grow in
394 microaerophilic and anaerobic atmospheres, with an optimal growth at 37°C after 48 hours of
395 incubation. The pH tolerance ranges from 6 to 8.5. The type species of the genus is
396 *Raoultibacter massiliensis* sp. nov. The type strain of the genus is strain Marseille-P2849^{T}.
397
398 **5.2 Description of *Raoultibacter massiliensis* sp. nov.**
399 *Raoultibacter massiliensis* (mas.si.li.en'sis. L. fem. adj. massiliensis, from Massilia, the Latin
400 name of Marseille, where the type strain was first isolated).
401 *Raoultibacter massiliensis* is a Gram-negative and motile coccobacillus whose individual
402 cells measure 0.8-1.2 µm in length and 0.4-0.6 µm in diameter. Transparent micro-colonies
403 obtained on 5% sheep blood-enriched Columbia agar exhibit a diameter of 0.1-0.3 mm. The
404 optimal growth is observed at 37°C after 48 hours of incubation. No oxidase activity, but
405 catalase activity is observed. Indole is produced. Using API strips, positive reactions are
406 observed with glycerol, D-Ribose, D-Galactose, D-Glucose, D-Fructose, D-Mannose, D-
407 Mannitol, N-Acetylglucosamine, Amygdaline, Arbutine, Esculin ferric citrate, Salicin, D-
408 Maltose, D-Lactose, D-Saccharose, D-Trehalose, D-Melezitose, Gentiobiose, D-Tagalose,
409 potassium Gluconate, L-Arabinose, Gelatine, D-Cellobiose, D-Melezitose, D-Rafinose, D-
410 sorbitol, amidon and potassium 5-Cetogluconate. Fucosidase, esterase (C4), esterase lipase
411 (C8), lipase (C14), Leucine arylamidase, Valine arylamidase, Cystine arylamidase, acid
412 phosphatase and naphtol phosphohydrolase activities are present but no reaction is obtained
413 for urease and alkaline phosphatase. The major fatty acids are 9-Octadecenoic acid (36 %),
414 Hexadecanoic acid (18 %) and Tetradecanoic acid (13 %). The genome is 3,657,161 bp long
415 with a DNA G+C content of 59.02mol%. The 16S rRNA and genome sequences were both
416 deposited in EMBL/EBI under accession numbers LT576395 and FZQX00000000,

417 respectively. The habitat of this bacterium is the human gut. The type strain Marseille-P2849^T
418 (= CSUR P2849 = DSM 103407) was isolated from a stool specimen of a healthy 19-year-old
419 male Bedouin living in Saudi Arabia.
420
421 **5.3 Description of *Raoultibacter timonensis* sp. nov.**
422 *Raoultibacter timonensis* (ti.mo.nen'sis, N.L. masc. adj., *timonensis* pertaining to La Timone,
423 the name of the university hospital in Marseille, France, where the strain was first isolated).
424 *Raoultibacter timonensis* is a Gram-negative and motile coccobacillus whose individual cells
425 measure 1-2 µm in length and 0.35-0. 44 µm in diameter. Transparent micro-colonies grown
426 on 5% sheep blood-enriched Columbia agar have a diameter of 0.1-0.4 mm with an optimal
427 growth at 37°C after a 48h incubation period in anaerobic conditions. No oxidase or catalase
428 activities were observed. Using API strips, positive reactions are observed with glycerol, D-
429 Ribose, D-Galactose, D-Glucose, D-Fructose, D-Mannose, D-Mannitol, N-
430 Acetylglucosamine, Amygdaline, Arbutine, Esculin ferric citrate, Salicin, D-Maltose, D-
431 Lactose, D-Saccharose, D-Trehalose, D-Melezitose, Gentiobiose, D-Tagalose, methyl- αD-
432 glucosamine, D-cellobiose, D-turanose, L-Rhamnose, glycerol, potassium gluconate, L-
433 Arabinose, gelatin, D-Cellobiose, D-Melezitose, D-Rafinose and D-sorbitol. Trypsin, α-
434 chymotrypsin, α-galactosidase, β-galactosidase, β-glucuronidase, α-glucosidase, β-
435 glucosidase, N-acetyl-β-glucosaminidase, α-mannosidase, exhibited esterase (C4), esterase
436 lipase (C8), Lipase (C14), Leucine arylamidase, Valine arylamidase, Cystine arylamidase,
437 acid phosphatase and naphtol phosphohydrolase activities are present. No reactions are
438 obtained for urease and phosphatase alkaline. The major fatty acids are 9-Octadecenoic acid
439 (38%), Hexadecanoic acid (25%) and Tetradecanoic acid (11%). Strain Marseille-P3277^T is
440 susceptible to amoxicillin, imipenem, metronidazole, rifampicin, erythromycin and resistant
441 to vancomycin, amikacin, Daptomycin, minocyclin and ceftriaxone. The genome is

4,000,215-bp-long with a DNA G+C content of 59.9 mol%. The 16S rRNA and genome sequences were deposited in EMBL/EBI under accession numbers LT623894 and OEPT00000000, respectively. The habitat of this microorganism is the human gut. The type strain Marseille- P3277^{T} (= CSUR P3277 = CCUG 70680) was isolated from the human stool of a 11-year-old healthy Pygmy female.

Funding.

This work was supported by the French Government under the « Investissements d'avenir » (Investments for the Future) program managed by the Agence Nationale de la Recherche (ANR, fr: National Agency for Research), (reference: Méditerranée Infection 10-IAHU-03) and by the National Plan for Science, Technology and Innovation (MAARIFAH) - King Abdulaziz City for Science and Technology - the Kingdom of Saudi Arabia - award number (12MED3108-03).

Acknowledgments

The authors thank the Xegen Company (http://www.xegen.fr/) for assisting in genomic analysis. The authors also acknowledge with thanks the Science and Technology Unit, King Abdulaziz University for their technical support.

Conflict of interest

The authors declare no conflict of interest

463

464 **References**

465 Alegre, M.L., Mannon, R.B., & Mannon, P.J. (2014). The microbiota, the immune system and
466 the allograft. *Am.J.Transplant.*, 14(6), 1236-1248.

467 Auch, A.F., Klenk, H.P., & Goker, M. (2010). Standard operating procedure for calculating
468 genome-to-genome distances based on high-scoring segment pairs. *Stand.Genomic.Sci.*, 2(1),
469 142-148.

470 Carver, T., Harris, S.R., Berriman, M., Parkhill, J., & McQuillan, J.A. (2012). Artemis: an
471 integrated platform for visualization and analysis of high-throughput sequence-based
472 experimental data. *Bioinformatics.*, 28(4), 464-469.

473 Carver, T., Thomson, N., Bleasby, A., Berriman, M., & Parkhill, J. (2009). DNAPlotter:
474 circular and linear interactive genome visualization. *Bioinformatics.*, 25(1), 119-120.

475 Citron, D.M., Ostovari, M.I., Karlsson, A., & Goldstein, E.J. (1991). Evaluation of the E test
476 for susceptibility testing of anaerobic bacteria. *J.Clin.Microbiol.*, 29(10), 2197-2203

477 Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., & Sayers, E.W. (2016). GenBank.
478 *Nucleic Acids Res.*, 44(D1), D67-D72.

479 Dione, N., Sankar, S.A., Lagier, J.C., Khelaifia, S., Michele, C., Armstrong, N., Richez, M.,
480 Abrahao, J., Raoult, D., & Fournier, P.E. (2016). Genome sequence and description of
481 Anaerosalibacter massiliensis sp. nov. *New Microbes.New Infect.*, 10, 66-76.

482 Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J.P., & Raoult, D. (2000). 16S
483 ribosomal DNA sequence analysis of a large collection of environmental and clinical
484 unidentifiable bacterial isolates. *J.Clin.Microbiol.*, 38(10), 3623-3630.

485 Elsawi, Z., Togo, A.H., Beye, M., Dubourg, G., Andrieu, C., Armsrtong, N., Richez, M., di,
486 P.F., Bittar, F., Labas, N., Fournier, P.E., Raoult, D., & Khelaifia, S. (2017). Hugonella
487 massiliensis gen. nov., sp. nov., genome sequence, and description of a new strictly anaerobic
488 bacterium isolated from the human gut. *Microbiologyopen.*, 6(4).

489 Fournier, P.E., Lagier, J.C., Dubourg, G., & Raoult, D. (2015). From culturomics to
490 taxonomogenomics: A need to change the taxonomy of prokaryotes in clinical microbiology.
491 *Anaerobe.*, 36, 73-78.

492 Gardiner, B.J., Korman, T.M., & Junckerstorff, R.K. (2014). Eggerthella lenta bacteremia
493 complicated by spondylodiscitis, psoas abscess, and meningitis. *J.Clin.Microbiol.*, 52(4)
494 1278-1280.

495 Gardiner, B.J., Tai, A.Y., Kotsanas, D., Francis, M.J., Roberts, S.A., Ballard, S.A.,
496 Junckerstorff, R.K., & Korman, T.M. (2015). Clinical and microbiological characteristics of
497 Eggerthella lenta bacteremia. *J.Clin.Microbiol.*, 53(2), 626-635.

498 Glenwright, A.J., Pothula, K.R., Bhamidimarri, S.P., Chorev, D.S., Basle, A., Firbank, S.J.,
499 Zheng, H., Robinson, C.V., Winterhalter, M., Kleinekathofer, U., Bolam, D.N., & van den
500 Berg, B. (2017). Structural basis for nutrient acquisition by dominant members of the human
501 gut microbiota. *Nature*, 541(7637), 407-411.

502 Gupta, R.S., Chen, W.J., Adeolu, M., & Chai, Y. (2013). Molecular signatures for the class
503 Coriobacteriia and its different clades; proposal for division of the class Coriobacteriia into
504 the emended order Coriobacteriales, containing the emended family Coriobacteriaceae and
505 Atopobiaceae fam. nov., and Eggerthellales ord. nov., containing the family Eggerthellaceae
506 fam. nov. *Int.J.Syst.Evol.Microbiol.*, 63(Pt 9), 3379-3397.

507 Honda, K. & Littman, D.R. (2016). The microbiota in adaptive immune homeostasis and

disease. *Nature*, 535(7610), 75-84.

Hugon, P., Dufour, J.C., Colson, P., Fournier, P.E., Sallah, K., & Raoult, D. (2015). A comprehensive repertoire of prokaryotic species identified in human beings. *Lancet Infect.Dis.*, 15(10), 1211-1219.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., & Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC.Bioinformatics.*, 11, 119.

Kim, M., Oh, H.S., Park, S.C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int.J.Syst.Evol.Microbiol.*, 64(Pt 2), 346-351.

Kokcha, S., Ramasamy, D., Lagier, J.C., Robert, C., Raoult, D., & Fournier, P.E. (2012). Non-contiguous finished genome sequence and description of Brevibacterium senegalense sp. nov. *Stand.Genomic.Sci.*, 7(2), 233-245.

Kumar, S., Tamura, K., & Nei, M. (1994). MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput.Appl.Biosci.*, 10(2), 189-191.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., & Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, 35(9), 3100-3108.

Lagier, J.C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., Trape, J.F., Koonin, E.V., La, S.B., & Raoult, D. (2012). Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin.Microbiol.Infect.*, 18(12), 1185-1193.

Lagier, J.C., Drancourt, M., Charrel, R., Bittar, F., La, S.B., Ranque, S., & Raoult, D. (2017). Many More Microbes in Humans: Enlarging the Microbiome Repertoire. *Clin.Infect.Dis.*, 65(suppl_1), S20-S29.

Lagier, J.C., Edouard, S., Pagnier, I., Mediannikov, O., Drancourt, M., & Raoult, D. (2015a). Current and past strategies for bacterial culture in clinical microbiology. *Clin.Microbiol Rev.*, 28(1), 208-236.

Lagier, J.C., Elkarkouri, K., Rivet, R., Couderc, C., Raoult, D., & Fournier, P.E. (2013). Non contiguous-finished genome sequence and description of Senegalemassilia anaerobia gen. nov., sp. nov. *Stand.Genomic.Sci.*, 7(3), 343-356.

Lagier, J.C., Hugon, P., Khelaifia, S., Fournier, P.E., La, S.B., & Raoult, D. (2015b). The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. *Clin.Microbiol.Rev.*, 28(1), 237-264.

Lagier, J.C., Khelaifia, S., Alou, M.T., Ndongo, S., Dione, N., Hugon, P., Caputo, A., Cadoret, F., Traore, S.I., Seck, E.H., Dubourg, G., Durand, G., Mourembou, G., Guilhot, E., Togo, A., Bellali, S., Bachar, D., Cassir, N., Bittar, F., Delerce, J., Mailhe, M., Ricaboni, D., Bilen, M., Dangui Nieko, N.P., Dia Badiane, N.M., Valles, C., Mouelhi, D., Diop, K., Million, M., Musso, D., Abrahao, J., Azhar, E.I., Bibi, F., Yasir, M., Diallo, A., Sokhna, C., Djossou, F., Vitton, V., Robert, C., Rolain, J.M., La, S.B., Fournier, P.E., Levasseur, A., & Raoult, D. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat.Microbiol.*, 1, 16203.

Lagier, J.C. & Raoult, D. (2016). [Culturomics: a method to study human gut microbiota]. *Med.Sci.(Paris)*, 32(11), 923-925.

Lau, S.K., Woo, P.C., Woo, G.K., Fung, A.M., Wong, M.K., Chan, K.M., Tam, D.M., & Yuen,

553 K.Y. (2004). Eggerthella hongkongensis sp. nov. and eggerthella sinensis sp. nov., two novel
554 Eggerthella species, account for half of the cases of Eggerthella bacteremia.
555 *Diagn.Microbiol.Infect.Dis.*, 49(4), 255-263.

556 Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P.F., & Prohaska, S.J. (2011).
557 Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC.Bioinformatics.*, 12,
558 124.

559 Lee, M.R., Huang, Y.T., Liao, C.H., Chuang, T.Y., Wang, W.J., Lee, S.W., Lee, L.N., &
560 Hsueh, P.R. (2012). Clinical and microbiological characteristics of bacteremia caused by
561 Eggerthella, Paraeggerthella, and Eubacterium species at a university hospital in Taiwan from
562 2001 to 2010. *J.Clin.Microbiol.*, 50(6), 2053-2055.

563 Lowe, T.M. & Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer
564 RNA genes in genomic sequence. *Nucleic Acids Res.*, 25(5), 955-964.

565 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J.,
566 Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V.,
567 Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.,
568 Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M.,
569 Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P.,
570 Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T.,
571 Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt,
572 K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., & Rothberg,
573 J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*,
574 437(7057), 376-380.

575 Matuschek, E., Brown, D.F., & Kahlmeter, G. (2014). Development of the EUCAST disk

diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories. *Clin.Microbiol.Infect.*, 20(4), O255-O266.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., & Goker, M. (2013a). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC.Bioinformatics.*, 14, 60.

Meier-Kolthoff, J.P., Goker, M., Sproer, C., & Klenk, H.P. (2013b). When should a DDH experiment be mandatory in microbial taxonomy? *Arch.Microbiol.*, 195(6), 413-418.

Meier-Kolthoff, J.P., Goker, M., Sproer, C., & Klenk, H.P. (2013c). When should a DDH experiment be mandatory in microbial taxonomy? *Arch.Microbiol.*, 195(6), 413-418.

Myron Sasser (2006). Bacterial Identification by Gas Chromatographic Analysis of Fatty Acids Methyl Esters (GC-FAME). *MIDI*

Ramasamy, D., Mishra, A.K., Lagier, J.C., Padhmanabhan, R., Rossi, M., Sentausa, E., Raoult, D., & Fournier, P.E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int.J.Syst.Evol.Microbiol.*, 64(Pt 2), 384-391.

Round, J.L. & Mazmanian, S.K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nat.Rev.Immunol.*, 9(5), 313-323.

Seck, E.H., Sankar, S.A., Khelaifia, S., Croce, O., Robert, C., Couderc, C., di, P.F., Sokhna, C., Fournier, P.E., Raoult, D., & Lagier, J.C. (2016). Noncontiguous finished genome sequence and description of Planococcus massiliensis sp. nov., a moderately halophilic bacterium isolated from the human gut. *New Microbes.New Infect.*, 10, 36-46.

Selma, M.V., Tomas-Barberan, F.A., Beltran, D., Garcia-Villalba, R., & Espin, J.C. (2014).

598 Gordonibacter urolithinfaciens sp. nov., a urolithin-producing bacterium isolated from the
599 human gut. *Int.J.Syst.Evol.Microbiol.*, 64(Pt 7), 2346-2352.

600 Smits, L.P., Bouter, K.E., de Vos, W.M., Borody, T.J., & Nieuwdorp, M. (2013). Therapeutic
601 potential of fecal microbiota transplantation. *Gastroenterology*, 145(5), 946-953.

602 The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks (2017).
603 Available at: https:, www.ncbi.nlm.nih.gov/pmc/articles/PMC3965112/. *(Accessed: 10th July*
604 *2017)*

605 Thompson, J.D., Higgins, D.G., & Gibson, T.J. (1994). CLUSTAL W: improving the
606 sensitivity of progressive multiple sequence alignment through sequence weighting, position-
607 specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22), 4673-4680.

608 Tindall, B.J., Rossello-Mora, R., Busse, H.J., Ludwig, W., & Kampfer, P. (2010). Notes on the
609 characterization of prokaryote strains for taxonomic purposes. *Int.J.Syst.Evol.Microbiol.*,
610 60(Pt 1), 249-266.

611 Wong, D., Aoki, F., & Rubinstein, E. (2014). Bacteremia caused by Eggerthella lenta in an
612 elderly man with a gastrointestinal malignancy: A case report.
613 *Can.J.Infect.Dis.Med.Microbiol.*, 25(5), e85-e86.

614 Wu, G.D. & Lewis, J.D. (2013). Analysis of the human gut microbiome and association with
615 disease. *Clin.Gastroenterol.Hepatol.*, 11(7), 774-777.

616 Wurdemann, D., Tindall, B.J., Pukall, R., Lunsdorf, H., Strompl, C., Namuth, T., Nahrstedt,
617 H., Wos-Oxley, M., Ott, S., Schreiber, S., Timmis, K.N., & Oxley, A.P. (2009). Gordonibacter
618 pamelaeae gen. nov., sp. nov., a new member of the Coriobacteriaceae isolated from a patient
619 with Crohn's disease, and reclassification of Eggerthella hongkongensis Lau et al. 2006 as

Paraeggerthella hongkongensis gen. nov., comb. nov. *Int.J.Syst.Evol.Microbiol.*, 59(Pt 6), 1405-1415.

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., & Rossello-Mora, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat.Rev.Microbiol.*, 12(9), 635-645.

Zak-Golab, A., Olszanecka-Glinianowicz, M., Kocelak, P., & Chudek, J. (2014). [The role of gut microbiota in the pathogenesis of obesity]. *Postepy Hig.Med.Dosw.(Online.)*, 68, 84-90.

630 **Table 1. Classification and general features of *Raoultibacter massiliensis* strain**

631 **Marseille-P2849^T and *Raoultibacter timonensis* strain Marseille-P3277^T**

Properties	**Term**	
Current classification	Domain: *Bacteria*	Domain: *Bacteria*
	Phylum: *Actinobacteria*	Phylum: *Actinobacteria*
	Class: *Coriobacteriia*	Class: *Coriobacteriia*
	Order: *Eggerthellales*	Order: *Eggerthellales*
	Family: *Eggerthellaceae*	Family: *Eggerthellaceae*
	Genus: *Raoultibacter*	Genus: *Raoultibacter*
	Species: *R. massiliensis*	Species: *R. timonensis*
	Type strain: Marseille-P2849^T	Type strain: Marseille-P3277^T
Gram-stain	Negative	Negative
Cell shape	coccobacilli	coccobacilli
Motility	Motile	Motile
Sporulation	Non-sporulating	Non-sporulating
Temperature range	25-45°C	25-45°C
Optimum temperature	37°C	37°C
Oxygen requirement	Anaerobic or microaerophilic	Anaerobic or microaerophilic
Biotic relationship	Free living	Free living
Isolation	Human feces	Human feces

632

633 **Table 2. Differential characteristics of *Raoultibacter massiliensis* strain Marseille-P2849^T, *Raoultibacter timonensis* strain Marseille-**

634 **P3277^T, *Gordonibacter pamelaeae* strain 7-10-1-b^T (Wurdemann D, *et al.*, 2009); *Gordonibacter urolithinfaciens* strain CEBAS 1/15P^T,**

635 **(Selma MV *et al.* 2014); *Eggerthella sinensis* HKU14 (Lau Susanna K. P *et al.*, 2004); *Paraeggerthella hongkongensis* strain HKU10^T**

636 **(Wurdemann D, *et al.*, 2009) and *Eggerthella lenta* JCM 997^T DSM 2243^T (Kageyama A, *et al.*, 1999).**

	Raoultibacter massiliensis	*Raoultibacter timonensis*	*Gordonibacter pamelaeae*	*Gordonibacter urolithinfaciens*	*Eggerthella sinensis*	*Paraeggerthella hongkongensis*	*Eggerthella lenta*
Cell lengh (µm)	0.8-1.2/ 0.4-0.6	0.8-1.2	1.2/ 0.5	1.57/ 0.61	NA	NA	0.2-0.4/ 0.2-2.0
Oxygen requirement	Anaerobe and micro aerophile	Anaerobe and micro aerophile	Strict anaerobe	Strict anaerobe	Strict anaerobe	Strict anaerobe	Strict anaerobe
Gram-stain	negative	negative	positive	positive	positive	positive	positive
Indole	+	+	NA	NA	-	-	-
Motility	+	+	+	+	-	-	-
Endospore formation	-	-	-	-	-	-	-
Production of							
Nitrate reductase	-	NA	-	-	-	-	+
Catalase	+	-	+	+	+	+	V
Urease	-	-	-	NA	-	-	-
Posphatase alkaline	-	-	-	-	-	-	-
Acid from							
L-Fucose	-	NA	-	+	-	-	-
D-Ribose	+	+	NA	NA	-	**NA**	+
L-arabinose	-	-	NA	-	-	-	+
D-Mannitol	+	+	NA	NA	NA	NA	NA
D-Mannose	+	+	-	-	-	-	-

Raffinose	+	+	-	-	-	-	-
L-Rhamnose	-	+	-	-	-	+	+
Trehalose	+	+	-	-	-	-	-
D-glucose	+	+	+	-	-	-	+
D-fructose	+	+	NA	+	NA	NA	NA
D-Maltose	+	+	NA	NA	NA	NA	NA
D-lactose	+	+	NA	NA	NA	NA	NA
DNA G+C content (mol%)	59.01	59.6	66.4	66.4	64.9, 65.6	61.1, 61.8	62.0, 61.8
Isolation source	Human feces	Human feces	human Colon	Human feces	Blood culture	Blood culture	Human feces

637 NA = data Not Available; v= variable

437

Table 3. Cellular fatty acid composition (%) of *Raoultibacter massiliensis* strain Marseille-P2849^T and *Raoultibacter timonensis* strain Marseille-P3277^T compared with other type strains of closely related species: 1, *R. massiliensis* strain Marseille-P2849^T; 2, *R. timonensis* strain Marseille-P3277^T 3, *Gordonibacter urolithinfaciens* strain CEBAS 1/15P^T; 4, *Gordonibacter pamelaeae* strain 7-10-1-b^T; 5, *Eggerthella hongkongensis DSM 16106*^{T}; 6, *Eggerthella lenta* DSM 2243^T; 7, *Eggerthella sinensis* DSM 16107^T .Values represent the percentage of total identified fatty acid methyl esters only (aldehydes, dimethyl acetals and unidentified "summed features" described previously were not included).

Fatty acids		1	2	3	4	5	6	7
18 :1n9	**9-Octadecenoic acid**	36.4	38.1	27.0	6.8	55.1	42.3	36.6
16 :0	**Hexadecanoic acid**	18.2	25.4	4.4	4.5	7.1	6.7	7.6
14 :0	**Tetradecanoic acid**	12.7	10.9	5.2	16.3	6.9	12.5	7.7
15 :0 anteiso	**12-methyl-tetradecanoic acid**	7.3	1.4	22.7	36.9	1.1	16.3	21.2
18 :2n6	**9,12-Octadecadienoic acid**	6.7	9	ND	ND	1.4	ND	ND
18 :0	**Octadecanoic acid**	3.4	5.7	5.6	1.5	4.7	1.4	1.5
18 :1n7	**11-Octadecenoic acid**	3.2	3.7	1.4	ND	4.3	2.6	2.3
15 :0 iso	**13-methyl-tetradecanoic acid**	2.8	2.8	3.6	5.5	0	1.1	0
12 :0	**Dodecanoic acid**	1.8	1.8	TR	5.0	7.7	2.9	1.1
13 :0 iso	**11-methyl-Dodecanoic acid**	1.5	ND	TR	2.0	ND	ND	ND
14 :0 iso	**12-methyl-Tridecanoic acid**	1.4	ND	13.4	18.3	0	7.5	17.1
15 :0	**Pentadecanoic acid**	1.2	1.1	ND	ND	ND	ND	ND
13 :0 anteiso	**10-methyl-Dodecanoic acid**	1.1	ND	ND	ND	ND	ND	1.0
20 :4n6	**5,8,11,14-Eicosatetraenoic acid**	TR	1.2	ND	ND	ND	ND	ND
20:5n3	**5,8,11,14,17-Eicosapentaenoic acid**	ND	TR	ND	ND	ND	ND	ND
5 :0 iso	**3-methyl-Butanoic acid**	TR	ND	ND	ND	ND	ND	ND
13 :0	**Tridecanoic acid**	TR	ND	ND	ND	ND	ND	ND
16 :1n7	**9-Hexadecenoic acid**	TR	ND	2.0	3.2	8.8	4.4	2.6

ND= Not detected

TR= trace amounts < 1 %

Table 4. Nucleotide content and gene count levels of the genome of strain *Raoultibacter massiliensis* Marseille-P2849^{T} and *Raoultibacter timonensis* strain Marseille-P3277^{T}.

	Raoultibacter massiliensis		*Raoultibacter timonensis*	
Size (bp)	3,657,161	100	4,000,215	100
Number of G+C	2,158,456	59	2,396,128	59.9
Number total of genes	3,073	100	3,284	100
Total number of protein-coding genes	3,025	98.4	3,232	99.33
Total number of RNA Genes	48	1.56	52	1.58
Total number of tRNA Genes	45	1.6	48	1.46
Total number of rRNA (5S, 16S, 23S) Genes	3	0.1	3	0.12
Coding sequence gene protein size	3,156,910	86.3	3,498,188	87.45
Number of proteins associated to COGs	2,365	77	2,562	78.01
Number of proteins associated to orfan	253	8,23	323	9.83
Number of proteins with peptide signal	385	12,5	512	15.59
Number of genes associated to PKS or NRPS	6	0.18	14	0.45
Number of genes associated to virulence	470	15.3	481	14.64
Number of proteins with TMH	855	27.8	940	28.62

The total is based on either the size of the genome in base pairs or the total number of protein- coding genes in the annotated genome

653 **Table 5. Number of genes associated with the 25 general COG functional categories.**

Code	*Raoultibacter massiliensis* Value	% of total	*Raoultibacter timonensis* Value	% of total	Description
[J]	134	4.43	142	4.39	Translation
[A]	0	0	0	0	RNA processing and modification
[K]	264	8.73	291	9.01	Transcription
[L]	102	3.37	95	2.94	Replication, recombination and repair
[B]	0	0	0	0	Chromatin structure and dynamics
[D]	23	0.76	16	0.5	Cell cycle control. mitosis and meiosis
[Y]	0	0	0	0	Nuclear structure
[V]	64	2.12	57	1.76	Defense mechanisms
[T]	181	5.98	214	6.62	Signal transduction mechanisms
[M]	121	4	115	3.56	Cell wall/membrane biogenesis
[N]	8	0.26	9	0.28	Cell motility
[Z]	0	0	0	0	Cytoskeleton
[W]	0	0	0	0	Extracellular structures
[U]	18	0.6	20	0.62	Intracellular trafficking and secretion
[O]	83	2.74	86	2.66	Posttranslational modification, protein turnover, chaperones
[X]	5	0.17	2	0.06	Mobilome: prophages, transposons
[C]	409	13.52	477	14.76	Energy production and conversion
[G]	118	3.9	132	4.08	Carbohydrate transport and metabolism
[E]	160	5.29	171	5.29	Amino acid transport and metabolism
[F]	55	1.82	58	1.79	Nucleotide transport and metabolism
[H]	65	2.15	69	2.13	Coenzyme transport and metabolism
[I]	49	1.61	55	1.7	Lipid transport and metabolism
[P]	120	3.97	139	4.3	Inorganic ion transport and metabolism
[Q]	18	0.6	21	0.65	Secondary metabolites biosynthesis, transport and catabolism
[R]	214	7.07	226	6.99	General function prediction only
[S]	154	5.09	167	5.18	Function unknown
-	660	21.82	670	20.73	Not in COGs

654 *The total is based on either the size of the genome in base pairs or the total number of

655 protein-coding genes in the annotated genome.

656 **Table 6. Genome comparison of species closely related to *Raoultibacter massiliensis***

657 **strain Marseille P2849^{T} and *Raoultibacter timonensis* strain Marseille P3277^{T} .**

Species	INSDC identifiera	Size (Mb)	G+C (mol %)	Gene Content
***Raoultibacter massiliensis* strain Marseille-P2849T**	FZQX00000000	3.65	59.01	3,021
***Raoultibacter timonensis* strain Marseille-P3277T**	OEPT00000000	3.94	59.6	3,277
***Eggerthella lenta* strain DSM 2243**	NC_013204.1	3.63	64.2	3,146
***Denitrobacterium detoxificans* strain NPOH1**	NZ_CP011402.1	2.45	59.5	2,023
***Gordonibacter pamelaeae* strain 7-10-1-b**	NC_021021.1	3.61	64.0	3,352
***Atopobium fossor* strain ATCC 43386 T**	AXXR00000000.1	1.66	45.4	1,505
***Atopobium parvulum* strain DSM 20469T**	NC_013203.1	1.54	45.7	1,406
***Olsenella profusa* DSM 13989**	AWEZ00000000.1	2.72	64.2	2,707
***Olsenella uli* ATCC 49627**	CP002106.1	2.05	64.7	1822

658 [a] INSDC: International Nucleotide Sequence Database Collaboration.

Table 7. Number of orthologous proteins shared between genomes (upper right) and AGIOS values (%) obtained (lower left). The number of proteins per genome is indicated in bold.

	Raoultibacter massiliensis	*Raoultibacter timonensis*	*Atopobium parvulum*	*Atopobium fossor*	*Adlercreutzia equolifaciens*	*Olsenella umbonata*	*Olsenella profusa*	*Gordonibacter pamelaeae*	*Eggerthella lenta*	*Denitrobacterium detoxificans*
Raoultibacter massiliensis	**3025**	1542	555	571	1069	693	683	1084	1404	911
Raoultibacter timonensis	81.25	**3232**	529	552	1029	647	643	1086	1373	863
Atopobium parvulum	59.35	59.27	**1363**	706	523	772	769	412	576	534
Atopobium fossor	58.97	58.95	66.76	**1487**	546	774	754	425	605	541
Adlercreutzia equolifaciens	69.69	70.09	58.3	58.12	**2278**	649	621	770	1094	861
Olsenella umbonata	64.29	64.82	63.57	62.66	66.2	**2059**	909	496	719	645
Olsenella profusa	63.81	64.37	62.95	62.73	65.97	74.21	**2593**	501	704	628
Gordonibacter pamelaeae	73.75	74.19	58.95	58.73	74.46	67.76	66.84	**3228**	1056	644
Eggerthella lenta	72.92	73.35	58.39	58.06	73.45	67	66.14	81.35	**3116**	921
Denitrobacterium detoxificans	68.46	68.75	60.29	60.14	68.84	64.956	64.84	70.75	69.92	1960

Table 8. Digital DNA-DNA hybridization values (%) obtained by comparison of *Raoultibacter massiliensis* strain Marseille-P2849^T and *Raoultibacter timonensis* strain Marseille P3277^T with other closely-related species using the GGDC formula 2 software (DDH estimates based on identities / HSP length)*, upper right.

	Raoultibacter massiliensis	*Raoultibacter timonensis*	*Atopobium parvulum*	*Atopobium fossor*	*Adlercreutzia equolifaciens*	*Olsenella umbonata*	*Olsenella profusa*	*Gordonibacter pamelaeae*	*Eggerthella lenta*	*Denitrobacterium detoxificans*
Raoultibacter massiliensis	**100**	25,2% ± 2.4	28,1% ± 2.4	30,7% ± 2.45	20,3% ± 2.35	20,8% ± 2.35	18,6% ± 2.25	24,5% ± 2.4	23,6% ± 2.4	19,1% ± 2.3
Raoultibacter timonensis		**100**	28% ± 2.4	30,1% ± 2.45	20,4% ± 2.35	21,5% ± 2.35	19% ± 2.3	22,9% ± 2.35	22% ± 2.35	19,1% ± 2.25
Atopobium parvulum			**100**	20,3% ± 2.35	22,6% ± 2.35	26,2% ± 2.4	24% ± 2.4	25,3% ± 2.4	25,8% ± 2.4	24,4% ± 2.4
Atopobium fossor				**100**	23,7% ± 2.4	21,3% ± 2.35	19,8% ± 2.3	26,8% ± 2.4	26,4% ± 2.45	25,2% ± 2.4
Adlercreutzia equolifaciens					**100**	18,2% ± 2.25	17,9% ± 2.25	22,4% ± 2.35	21,5% ± 2.35	19,5% ± 2.35
Olsenella umbonata						**100**	21,7% ± 2.35	18,2% ± 2.25	20,4% ± 2.35	33,7% ± 2.45
Olsenella profusa							**100**	18% ± 2.25	19,3% ± 2.3	22,3% ± 2.4
Gordonibacter pamelaeae								**100**	29,4% ± 2.45	19,7% ± 2.35
Eggerthella lenta									**100**	20,2% ± 2.35
Denitrobacterium detoxificans										**100**

*****The confidence intervals indicate the inherent uncertainty in estimating DDH values from intergenomic distances based on models derived from empirical test data sets (which are always limited in size).

Figure Legends.

Figure 1. Gel view comparing *Raoultibacter massiliensis* gen. nov., sp. nov. strain Marseille-P2849 ^{T} and strain *Raoultibacter timonensis* gen. nov., sp. nov. strain Marseille-P3277 ^{T} with other closely related species present in our MALDI-TOF-MS spectrum database. The gel view displays the raw spectra of loaded spectrum files arranged in a pseudo-gel like look. The x-axis records the m/z value. The left y-axis displays the running spectrum number originating from subsequent spectra loading. The peak intensity is expressed by a gray scale scheme code. The color bar and the right y-axis indicate the relation between the color of the peak and its intensity, in arbitrary units. Displayed species are indicated on the left.

Figure 2. Phylogenetic tree highlighting the position of *Raoultibacter massiliensis* strain gen. nov., sp. nov. strain Marseille-P2849^{T} and *Raoultibacter timonensis* gen. nov., sp. nov. strain Marseille-P3277^{T} relative to other closely related species. Strains and their GenBank accession numbers of 16S rRNA gene are indicated in brackets. Sequences were aligned using ClustalW, with default parameters and phylogenetic inferences obtained using the neighbor-joining method with 500 bootstrap replicates, within MEGA6 software. The scale bar represents a 2% nucleotide sequence divergence.

Figure 3. Gram-staining of (A) *Raoultibacter massiliensis* gen. nov., sp. nov. strain Marseille-P2849^{T} and (B) *Raoultibacter timonensis* gen. nov., sp. nov strain Marseille-P3277^{T}. Transmission electron microscopy images of *Raoultibacter massiliensis* gen. nov., sp. nov. strain Marseille-P2849T (C) and *Raoultibacter timonensis* gen. nov., sp. nov strain Marseille-P3277^{T} (D) using a Tecnai G20 transmission electron microscope (FEI Company). The scale bar represents 200 nm.

691 **Figure 4:** Graphical circular map of the genome of (A) *Raoultibacter massiliensis* gen. nov.,
692 sp. nov. strain Marseille-P2849^{T} and (B) strain *Raoultibacter timonensis* gen. nov., sp. nov.
693 strain Marseille-P3277^{T}. From the outside to the center, contigs (red / grey), COG category of
694 genes on the forward strand (three circles), genes on the forward strand (blue circle), genes on
695 the reverse strand (red circle), COG category of genes on the reverse strand (three circles),
696 G+C skew (purple indicates positive values and olive negative values).
697
698 **Figure 5.** Distribution of functional classes of predicted genes according to the clusters of
699 orthologous groups of proteins of *Raoultibacter massiliensis* gen. nov., sp. nov. strain
700 Marseille-P2849 and strain *Raoultibacter timonensis* gen. nov., sp. nov. strain Marseille-
701 P3277^{T} among other closely related species.

Figures:

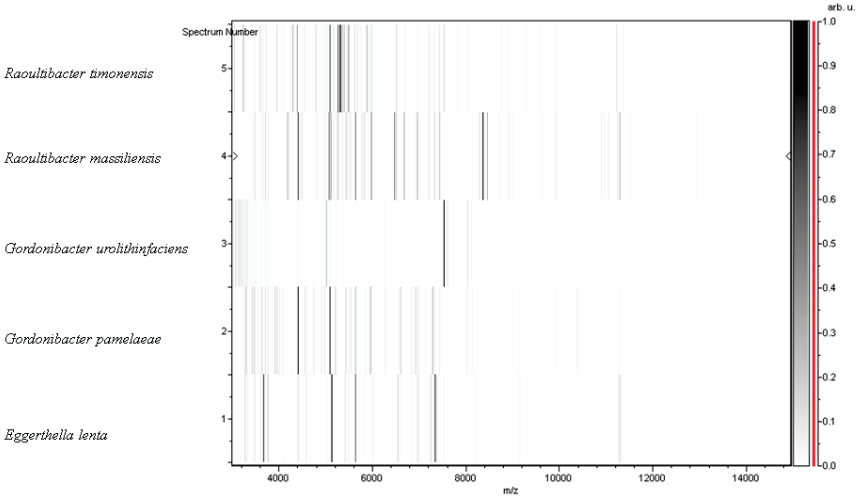


Figure 1. Gel view comparing *Raoultibacter massiliensis* gen. nov., sp. nov. strain Marseille-P2849 ^{T} and strain *Raoultibacter timonensis* gen. nov., sp. nov. strain Marseille-P3277 ^{T} with other closely related species present in our MALDI-TOF-MS spectrum database. The gel view displays the raw spectra of loaded spectrum files arranged in a pseudo-gel like look. The x-axis records the m/z value. The left y-axis displays the running spectrum number originating from subsequent spectra loading. The peak intensity is expressed by a gray scale scheme code. The color bar and the right y-axis indicate the relation between the color of the peak and its intensity, in arbitrary units. Displayed species are indicated on the left.

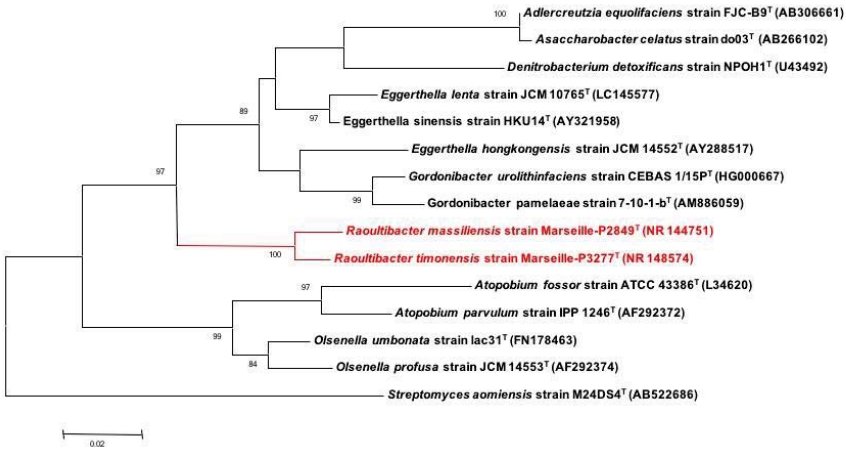
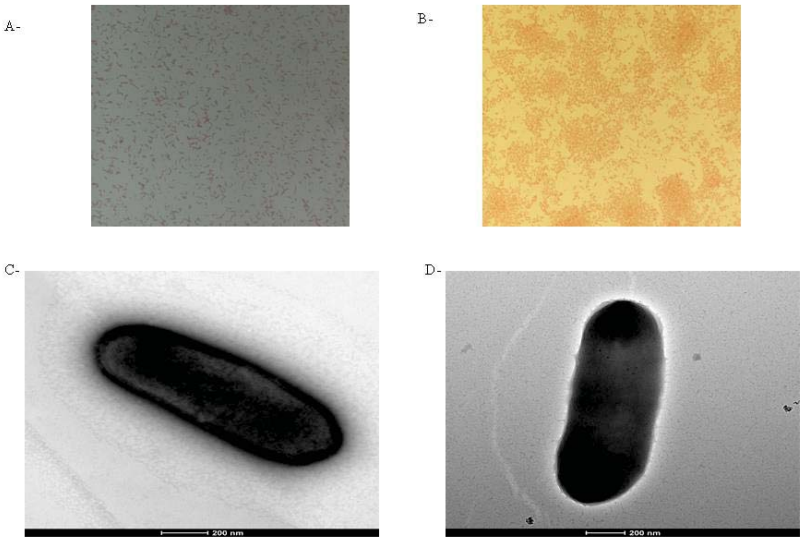


Figure 2. Phylogenetic tree highlighting the position of *Raoultibacter massiliensis* strain gen. nov., sp. nov. strain Marseille-P2849^{T} and *Raoultibacter timonensis* gen. nov., sp. nov. strain Marseille-P3277^{T} relative to other closely related species. Strains and their GenBank accession numbers of 16S rRNA gene are indicated in brackets. Sequences were aligned using ClustalW, with default parameters and phylogenetic inferences obtained using the neighbor-joining method with 500 bootstrap replicates, within MEGA6 software. The scale bar represents a 2% nucleotide sequence divergence.



720
721 **Figure 3.** Gram-staining of (A) *Raoultibacter massiliensis* gen. nov., sp. nov. strain Marseille-
722 P2849^{T} and (B) *Raoultibacter timonensis* gen. nov., sp. nov strain Marseille-P3277^{T}.
723 Transmission electron microscopy images of *Raoultibacter massiliensis* gen. nov., sp. nov.
724 strain Marseille-P2849T (C) and *Raoultibacter timonensis* gen. nov., sp. nov strain Marseille-
725 P3277^{T} (D) using a Tecnai G20 transmission electron microscope (FEI Company). The scale
726 bar represents 200 nm.

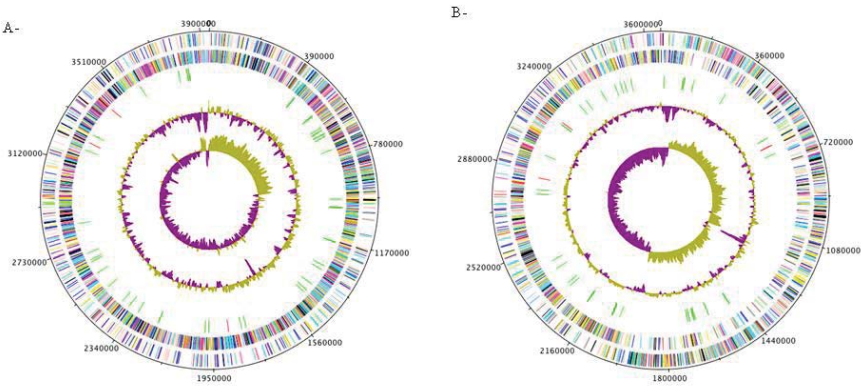
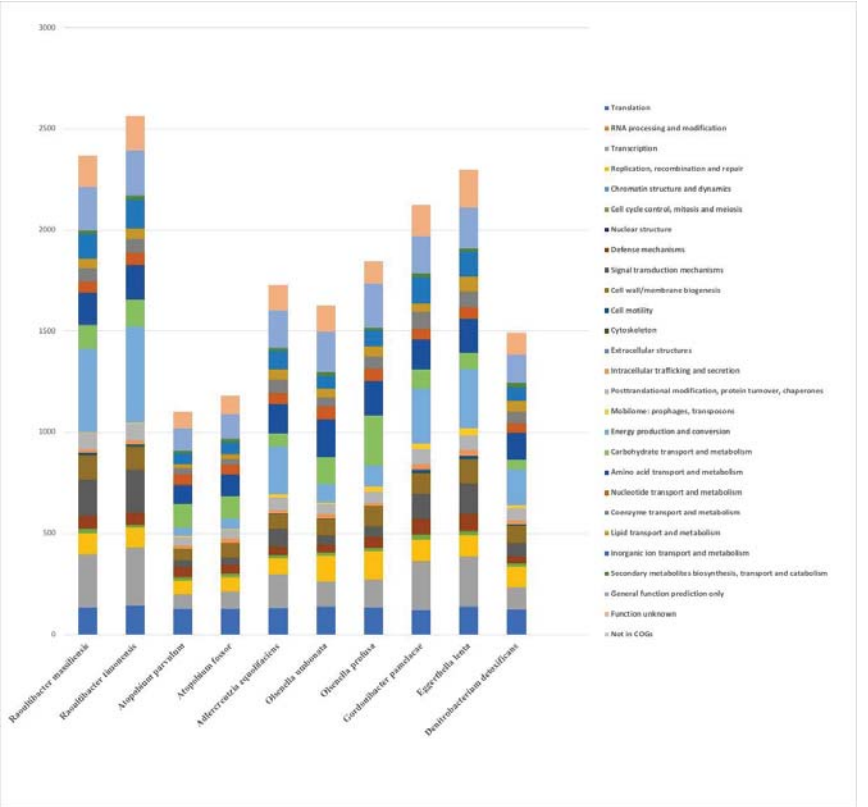


Figure 4: Graphical circular map of the genome of (A) *Raoultibacter massiliensis* gen. nov., sp. nov. strain Marseille-P2849^{T} and (B) strain *Raoultibacter timonensis* gen. nov., sp. nov. strain Marseille-P3277^{T}. From the outside to the center, contigs (red / grey), COG category of genes on the forward strand (three circles), genes on the forward strand (blue circle), genes on the reverse strand (red circle), COG category of genes on the reverse strand (three circles), G+C skew (purple indicates positive values and olive negative values).



734
735 **Figure 5.** Distribution of functional classes of predicted genes according to the clusters of
736 orthologous groups of proteins of *Raoultibacter massiliensis* gen. nov., sp. nov. strain
737 Marseille-P2849^{T} and strain *Raoultibacter timonensis* gen. nov., sp. nov. strain Marseille-
738 P3277^{T} among other closely related species.

CHAPITRE IV: (ANNEXES)

Microbio-génomique

Avant-propos

Cette dernière partie de mon travail doctoral contient deux articles décrivant le séquençage du génome entier d'espèces déjà connues notamment celui de la souche type de l'espèce *Ezakiella peruensis* M6.X2 (première séquence génomique de cette espèce) et celui de la souche *Megamonas funiformis* Marseille-P3344 nouvellement isolée dans le cadre du projet « culturomics » dans notre laboratoire. Ce séquençage du génome fait partie d'une étude «microbio-génomique» visant à séquencer et analyser les génomes d'espèces bactériennes pour lesquelles aucune séquence n'est disponible, ou les nouvelles souches bactériennes isolées dans notre laboratoire dans le but d'étendre les bases de données des génomes bactériens.

Ezakiella peruensis M6.X2^{T} est un coccus anaérobie à Gram positif isolé à partir d'un échantillon fécal d'un individu en bonne santé résidant dans une communauté traditionnelle côtière au Pérou. Le génome de la souche M6.X2, a une longueur de 1 672 788 pb et héberge 1 589 gènes codant pour des protéines, dont 26 gènes associées à la résistance aux antibiotiques avec 1 gène codant pour la résistance à la vancomycine. Le génome présente également une région CRISPR et 333 gènes acquis par transfert horizontal de gènes.

Le deuxième article décrit le draft génome de la souche *Megamonas funiformis* Marseille-P3344 isolée à partir d'un échantillon fécal d'un individu sain dans notre laboratoire. Il s'agit d'une bactérie à Gram négatif, strictement anaérobie. Le génome mesure 2 464 704 pb, avec 2 230 gènes codant pour des protéines et 76 gènes d'ARN. En outre, 46 gènes de virulence sont prédits incluant 30 gènes associés à la résistance aux antibiotiques, dont 3 bêta-lactamases.

Article 20:

Draft Genome Sequence of *Ezakiella peruensis* Strain M6X2^{T}, a human fecal Gram-stain positive anaerobic coccus

Awa Diop, Khoudia Diop, Enora Tomei, Didier Raoult, Florence Fenollar, Pierre-Edouard Fournier

[Published in Genome Announcements]

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

Draft Genome Sequence of *Ezakiella peruensis* Strain M6.X2, a Human Gut Gram-Positive Anaerobic Coccus

Awa Diop,[a] Khoudia Diop,[a] Enora Tomei,[a] Didier Raoult,[a,b] Florence Fenollar,[a] Pierre-Edouard Fournier[a]

[a]Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, Aix-Marseille Université, UM 63, CNRS UMR7278, IRD 198, INSERM U1095, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée-Infection, Faculté de Médecine, Marseille, France

[b]Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT We report here the draft genome sequence of *Ezakiella peruensis* strain M6.X2^T. The draft genome is 1,672,788 bp long and harbors 1,589 predicted protein-encoding genes, including 26 antibiotic resistance genes with 1 gene encoding vancomycin resistance. The genome also exhibits 1 clustered regularly interspaced short palindromic repeat region and 333 genes acquired by horizontal gene transfer.

Ezakiella peruensis is the type and only species of the genus *Ezakiella*, created in 2015 (1). *E. peruensis* occupies a unique position in an undefined family within the phylum *Firmicutes* (1). This microorganism is a Gram-positive anaerobic coccus. Gram-positive anaerobic cocci include many commensal species of humans and animals and also some human pathogens (2). The type strain M6.X2^T was isolated from a fecal sample of a healthy individual residing in a coastal traditional community in Peru (1). It is nonmotile and non-spore forming. Here, we present the annotated draft genome sequence of *E. peruensis* strain M6.X2^T (DSM 27367 = NBRC 109957 = CCUG 64571), which we obtained from the DSMZ collection.

Genomic DNA of *E. peruensis* strain M6.X2^T was sequenced using a MiSeq sequencer with the mate-pair strategy (Illumina, Inc., San Diego, CA, USA). DNA was quantified by a Qubit assay with a high-sensitivity kit (Life Technologies, Carlsbad, CA, USA) at 38.4 ng/μl. The 576,285 high-quality paired-end reads were trimmed and then assembled using the SPAdes assembler program (3). The draft genome sequence was annotated using Prokka software (4). Functional annotation was achieved using the BLASTp algorithm (5) against the Clusters of Orthologous Groups (COGs) database and the Rapid Annotations using Subsystems Technology (RAST) web server (6). Ribosomal RNAs (5S, 16S, and 23S rRNAs) were predicted using RNAmmer software (7).

The genome was 1,672,788-bp long, assembled in five scaffolds (seven contigs) with a G+C content of 36.9%. Overall, 1,589 protein-coding sequences were identified, including 1,165 (73.31%) protein-coding genes that had orthologs in the COGs database, 1,052 of which were assigned a putative function. A total of 46 tRNA loci and 1 rRNA operon (16S, 5S, and 23S rRNA) were identified in the genome. Strain M6.X2^T exhibited 26 genes associated with antibiotic resistance and toxic compounds, including one *vanW* gene encoding vancomycin resistance. No toxin/antitoxin module or bacteriocin-associated gene was identified. The genome of *E. peruensis* harbored 1 clustered regularly interspaced short palindromic repeat locus of 763 bp with 12 repeats (mean repeat length = 36 bp). We also detected 333 putative genes acquired by horizontal gene transfer, including 209 from bacteria within the order *Clostridiales*.

Accession number(s). The 16S rRNA and genome sequences from *Ezakiella peruensis* strain M6.X2^T are available in GenBank under accession numbers KJ469554 and OCSL00000000, respectively.

Received 28 November 2017 **Accepted** 6 February 2018 **Published** 1 March 2018

Citation Diop A, Diop K, Tomei E, Raoult D, Fenollar F, Fournier P-E. 2018. Draft genome sequence of *Ezakiella peruensis* strain M6.X2, a human gut Gram-positive anaerobic coccus. Genome Announc 6:e01487-17. https://doi.org/10.1128/genomeA.01487-17.

Address correspondence to Pierre-Edouard Fournier, pierre-edouard.fournier@univ-amu.fr.

ACKNOWLEDGMENTS

This study was funded by the Méditerranée-Infection Foundation and the French Agence National de la Recherche under reference Investissements d'Avenir Méditerranée-Infection 10-IAHU-03.

REFERENCES

1. Patel NB, Tito RY, Obregón-Tito AJ, O'Neal L, Trujillo-Villaroel O, Marin-Reyes L, Troncoso-Corzo L, Guija-Poma E, Hamada M, Uchino Y, Lewis CM, Lawson PA. 2015. *Ezakiella peruensis* gen. nov., sp. nov. isolated from human fecal sample from a coastal traditional community in Peru. Anaerobe 32:43–48. https://doi.org/10.1016/j.anaerobe.2014.12.002.
2. Ulger-Toprak N, Liu C, Summanen PH, Finegold SM. 2010. *Murdochiella asaccharolytica* gen. nov., sp. nov., a Gram-stain-positive, anaerobic coccus isolated from human wound specimens. Int J Syst Evol Microbiol 60:1013–1016. https://doi.org/10.1099/ijs.0.015909-0.
3. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
4. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
5. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
6. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: Rapid Annotations using Subsystems Technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75.
7. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. https://doi.org/10.1093/nar/gkm160.

Article 21:

Draft genome sequence of *Megamonas funiformis* strain Marseille-P3344 isolated from the human fecal microbiota

Mossaab Maaloum, Awa Diop, Sokhna Ndongo, Thi-Tien Nguyen, Frederic Cadoret, Didier Raoult, Pierre-Edouard Fournier

[Published in Genome Announcements]

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

Draft Genome Sequence of *Megamonas funiformis* Strain Marseille-P3344, Isolated from a Human Fecal Microbiota

Mossaab Maaloum,[a,b] Awa Diop,[a] Sokhna Ndongo,[a] Thi-Tien Nguyen,[a] Frederic Cadoret,[a] Didier Raoult,[a,c] Pierre-Edouard Fournier[a]

[a]URMITE, Institut Hospitalo-Universitaire Méditerranée-Infection, Aix-Marseille Université, UM63, CNRS 7278, IRD 198, Inserm U1095, Assistance Publique–Hôpitaux de Marseille, Marseille, France

[b]Faculty of Sciences Ben M'sik, Laboratory of Biology and Health, Hassan II University, Casablanca, Morocco

[c]Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT In this article, we present the draft genome sequence of *Megamonas funiformis* strain Marseille-P3344, isolated from a human fecal sample. The genome described here is composed of 2,464,704 nucleotides, with 2,230 protein-coding genes and 76 RNA genes.

Megamonas hypermegale was the first species of the *Megamonas* genus described. The bacterium was isolated from chicken cecum and first described in 1936 as *Bacteroides hypermegas* by Harrison and Hansen (1), and the original name was changed to *Megamonas hypermegale* by Euzéby in 1998 (2). This microorganism is strictly anaerobic and nonmotile. Its optimal growth temperature is 37°C. The species *Megamonas funiformis* was identified in human feces in 2008 in Japan by Sakon et al. (3). Cells from this bacterium are large Gram-negative rods, 5 to 10 μm in size. Some of the cells exhibit a central, subterminal, or terminal swelling of 2- to 4-μm diameter when grown in a broth medium supplemented with glucose.

In August 2016, as part of a microbial culturomics study, we cultivated strain Marseille-P3344 from a fecal sample of a healthy woman. This bacterium exhibited a 99.08% 16S rRNA sequence similarity with *M. funiformis* strain YIT 11815^T (=JCM 14723 =DSM 19343), its closest phylogenetic neighbor. Genomic DNA (gDNA) from *M. funiformis* strain Marseille-P3344, isolated from a human fecal specimen, was sequenced using a MiSeq sequencer and the mate pair strategy (Illumina, Inc., San Diego, CA, USA). The gDNA from *M. funiformis* strain Marseille-P3344 was barcoded in order to be mixed with 11 other projects with the Nextera mate pair sample prep kit (Illumina). The gDNA quantification by a Qubit assay with a high-sensitivity kit (Life Technologies, Inc., Carlsbad, CA, USA) was 148.7 ng/μL.

A total of 6.3 Gb was obtained from a 673,000/mm^2 cluster density with a cluster passing quality control filters of 95.4% (12,453,000 clusters). Within this run, the index representation for *M. funiformis* was 7.99%. The 995,543 mate pair reads were filtered according to the read quality.

The draft genome sequence of *M. funiformis* strain Marseille-P3344 is composed of 7 scaffolds for a total of 2,464,704 nucleotides (nt) and a G+C content of 31.4%. The coding capacity is 2,099,846 nt (85.1% of the total genome). Predicted genes include 2,230 protein-coding genes, of which 1,701 are assigned to clusters of orthologous groups and 76 (3.29%) are RNA genes (17 rRNAs and 59 tRNAs). A total of 228 genes (10.2%) have peptide signals, and 481 (21.5%) have transmembrane helices. In addition, 46 virulence genes are predicted, including 30 genes associated with antibiotic resistance, including 3 beta-lactamases. No toxin/antitoxin module or bacteriocin-associated gene could be found.

Received 22 November 2017 **Accepted** 29 November 2017 **Published** 11 January 2018

Citation Maaloum M, Diop A, Ndongo S, Nguyen T-T, Cadoret F, Raoult D, Fournier P-E. 2018. Draft genome sequence of *Megamonas funiformis* strain Marseille-P3344, isolated from a human fecal microbiota. Genome Announc 6:e01459-17. https://doi.org/10.1128/genomeA.01459-17.

Address correspondence to Pierre-Edouard Fournier, pierre-edouard.fournier@univ-amu.fr.

The genomes of *M. funiformis* strains Marseille-P3344 and YIT 11815^T (=JCM 14723 =DSM 19343) were compared using GGDC and OrthoANI softwares (4, 5). Digital DNA-DNA hybridization and OrthoANI values of 84.1% ± 2.6 (>70%) and 98.18% (>95.96%), respectively, were obtained, thus confirming that these strains belong to the same species.

Accession number(s). The 16S rRNA and whole-genome sequences reported here have been deposited in GenBank under accession numbers LT628480 and FQRY00000000, respectively.

ACKNOWLEDGMENT

This work was supported by the Méditerranée-Infection Foundation.

REFERENCES

1. Harrison AP, Hansen PA. 1963. *Bacteroides hypermegas* nov. spec. Antonie van Leeuwenhoek 29:22–28. https://doi.org/10.1007/BF02046035.
2. Euzéby JP. 1998. Taxonomic note: necessary correction of specific and subspecific epithets according to Rules 12c and 13b of the International Code of Nomenclature of Bacteria (1990 Revision) Int J Syst Bacteriol 48:1073–1075. https://doi.org/10.1099/00207713-48-3-1073.
3. Sakon H, Nagai F, Morotomi M, Tanaka R. 2008. *Sutterella parvirubra* sp. nov. and *Megamonas funiformis* sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 58:970–975. https://doi.org/10.1099/ijs.0.65456-0.
4. Auch AF, von Jan M, Klenk HP, Göker M. 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2:117–134. https://doi.org/10.4056/sigs.531120.
5. Lee I, Kim YO, Park S-C, Chun J. 2016. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103. https://doi.org/10.1099/ijsem.0.000760.

CONCLUSION ET PERSPECTIVES

L'approche polyphasique basique utilisée dans la taxonomie et la systématique des bactéries comprend l'utilisation de données phénotypiques, chimiotaxonomiques et génotypiques. Avec l'avènement des progrès remarquables de la technologie et de l'application du séquençage de « nouvelle génération » (NGS), en grande partie liée à la diminution des coûts de séquençage à une vitesse sans précédent, la systématique et la classification taxonomique des procaryotes est entrée dans l'ére génomique. Cela a permis l'accès à des séquences génomiques bactériennes complètes dont plus de 100 génomes d'espèces de *Rickettsia* officiellement validées et non officiellement reconnues. De plus, l'accés sans précédent aux séquences génomiques a non seulement permis l'utilisation de données précieuse pour une classification taxonomique plus fiable et précise des prokaryotes, mais aussi de déchiffrer le contenu génique complet d'une bactérie. De plus, le séquençage du génome fournit également une teneur précise en G + C du génome, ce qui a eu une grande valeur dans la taxonomie bactérienne. Ainsi, au travers de deux revues de la littérature sur les génomes des bactéries du genre *Rickettsia*, nous avons pu identifier les caractéristiques génomiques générales, les mécanismes évolutifs et les différences de pathogénicité en relation avec ces processus

évolutifs qui animent les génomes de *Rickettsia*. Les *Rickettsia* ont des génomes de petite taille et subissent une evolution convergente à la fois reductive avec dégradation ou perte selective de gènes parallèlement à une proliferation paradoxale d'éléments génétiques, duplication de gènes et ou transfert horizontal de gènes. Nous avons montré aussi que l'évolution réductive du génome contribue à l'émergence de la pathogénicité. Ainsi, des études futures seront nécessaires pour élucider notre compréhension sur les mécanismes par lesquels ce processus évolutif entraine une augmentation de la virulence. Ensuite, nous avons prouvé que l'utilisation de la génomique facilite la classification et l'identification des prokaryotes, notamment grâce à la disponibilité d'outils bioinformatiques assez simples d'utilisation. Nous proposons l'utilisation des données de séquence du génome entier pour la mise au point des recommandations pour la définition et la classification des isolats au niveau de l'espèce et du genre. En particulier, avec l'analyse de similarité des séquences génomiques de 78 souches de *Rickettsia* et de 61 souches de trois genres étroitement apparentés du genre *Rickettsia*, et en utilisant plusieurs paramètres génomiques basés sur la taxonomie: dDDH; OrthoANI et AGIOS, nous avons pu élaborer des recommandations pour la classification des isolats de *Rickettsia*

au niveau du genre et de l'espèce. Les outils AGIOS et OrthoANI sont les meilleures méthodes permettaient de definir qu'un isolat bactérien appartient bien au genre *Rickettsia.* En revanche, le dDDH est le meilleur outil pour definir si un isolat bactérien est une nouvelle espèce ou un isolat appartient à une espèce de *Rickettsia* connue. Néanmoins, les paramètres AGIOS et OrthoANI peuvent également être utilisés comme méthodes complémentaires, mais pas pour les espèces étroitement apparentées à *R. conorii.* Le paramètre AGIOS est légèrement différent de l'OrthoANI dans la mesure où ce dernier utilise BLASTN pour identifier les fragments orthologues qui est moins sensible que BLASTP utilisé par le premier. En plus l'outil AGIOS fournit en même temps le nombre de genes orthologues partagés entre deux génomes. Nous avons également trouvé une forte correlation positive entre nos données génomiques et les données dérivées des séquences de gènes. En outre nous avons montré que les outils taxono-génomiques sont des méthodes relativement simples d'utilisation en laboratoire et permettent une classification taxonomique fiable, rapide, facile et reproductible pour les especes de *Rickettsia* avec des seuils spécifiques. Avec le séquençage de plus en plus de souches bactériennes, nous prévoyons que l'outil AGIOS puisse être utilisé comme index

génomique pour la délimitation bactérienne dans un futur proche avec la détermination a posteriori de valeurs seuils standards ou spécifiques.

Par ailleurs, dans ce travail, nous avons utilisé la stratégie «taxono-genomics», intégrant les données de séquençage et de l'analyse génomique, le spectre protéique MALDI-TOF, en plus des propriétés phénotypiques et génotypiques, dans la description taxonomique de nouvelles espèces bactériennes. Nous avons analysé et décrire les génomes de 17 nouveaux isolats bactériens isolés par la méthode de "culturomique bactérienne" à partir de divers échantillons. En plus de cela, nous avons également analysé, caractérisé et décrit le premier génome séquencé de la souche type de l'espèce *Ezakiella peruensis* M6.X2^{T} et celui de la nouvelle souche *Megamonas funiformis* Marseille-P3344. Ceux-ci visent à étendre les bases de données des génomes bactériens. L'incorporation de la génomique dans la taxonomie et la systématique des bactéries couplée à la disponibilité d'outils bio-informatiques robustes augmentera la crédibilité de la taxonomie dans l'ère génomique. L'utilisation des outils génomiques est donc parfaitement adaptée à la classification taxonomique et peut changer radicalement notre vision de la taxonomie et de l'évolution bactérienne à l'avenir.

REFERENCES

1. **Karl Bernhard Lehmann RON**. *Atlas und Grundriss der Bakteriologie und Lehrbuch der speziellen bakteriologischen Diagnostik ...* München: Lehmann. http://archive.org/details/atlasundgrundri00neumgoog (1896).

2. **Schleifer KH**. Classification of Bacteria and Archaea: Past, present and future. *Syst Appl Microbiol* 2009;32:533–542.

3. **Stackebrandt E**. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* 2002;52:1043–1047.

4. **Stackebrandt E, Ebers J**. *Taxonomic parameters revisited: Tarnished gold standards*. 2006.

5. **Vandamme P, Pot B, Gillis M, De Vos P, Kersters K, *et al.*** Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* 1996;60:407–438.

6. **Tindall BJ, Rosselló-Móra R, Busse H-J, Ludwig W, Kämpfer P**. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* 2010;60:249–266.

7. **Coenye T, Vandamme P**. Use of the Genomic Signature in Bacterial Classification and Identification. *Syst Appl Microbiol* 2004;27:175–185.

8. **Konstantinidis KT, Ramette A, Tiedje JM**. The bacterial species definition in the genomic era. *Philos Trans R Soc B Biol Sci* 2006;361:1929–1940.

9. **Woese CR**. Bacterial evolution. *Microbiol Rev* 1987;51:221–271.

10. **Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, *et al.*** Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Evol Microbiol* 1987;37:463–464.

11. **Grimont PA**. Use of DNA reassociation in bacterial classification. *Can J Microbiol* 1988;34:541–546.

12. **Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, *et al.*** A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* 2014;64:384–391.

13. **Fournier P-E, Dumler JS, Greub G, Zhang J, Wu Y, *et al.*** Gene Sequence-Based Criteria for Identification of New Rickettsia Isolates and Description of Rickettsia heilongjiangensis sp. nov. *J Clin Microbiol* 2003;41:5456–5465.

14. **Fournier P-E, Raoult D**. Current Knowledge on Phylogeny and Taxonomy of Rickettsia spp. *Ann N Y Acad Sci* 2009;1166:1–11.

15. **Kim M, Oh H-S, Park S-C, Chun J**. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.

16. **Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M**. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:1.

17. **Konstantinidis KT, Tiedje JM**. Towards a Genome-Based Taxonomy for Prokaryotes. *J Bacteriol* 2005;187:6258–6264.

18. **Stothard DR, Clark JB, Fuerst PA**. Ancestral divergence of Rickettsia bellii from the spotted fever and typhus groups of Rickettsia and antiquity of the genus Rickettsia. *Int J Syst Evol Microbiol* 1994;44:798–804.

19. **Raoult D, Roux V**. Rickettsioses as paradigms of new or emerging infectious diseases. *Clin Microbiol Rev* 1997;10:694–719.

20. **Parola P, Paddock CD, Socolovschi C, Labruna MB, Mediannikov O, *et al.*** Update on Tick-Borne Rickettsioses around the World: a Geographic Approach. *Clin Microbiol Rev* 2013;26:657–702.

21. **Sahni SK, Narra HP, Sahni A, Walker DH**. Recent molecular insights into rickettsial pathogenesis and immunity. *Future Microbiol* 2013;8:1265–1288.

22. **El Karkouri K, Kowalczewska M, Armstrong N, Azza S, Fournier P-E, *et al.*** Multi-omics Analysis Sheds Light on the Evolution and the Intracellular Lifestyle Strategies of Spotted Fever Group Rickettsia spp. *Front Microbiol*;8. Epub ahead of print 20 July 2017. DOI: 10.3389/fmicb.2017.01363.

23. **Abdad MY, Abdallah RA, Karkouri KE, Beye M, Stenos J, *et al.*** Rickettsia gravesii sp. nov.: a novel spotted fever group rickettsia in Western Australian Amblyomma triguttatum triguttatum ticks. *Int J Syst Evol Microbiol* 2017;67:3156–3161.

24. **Drancourt M, Raoult D**. Taxonomic position of the rickettsiae: current knowledge. *FEMS Microbiol Rev* 1994;13:13–24.

25. **Philip RN, Casper EA, Burgdorfer W, Gerloff RK, Hughes LE, *et al.*** Serologic typing of rickettsiae of the spotted fever group by microimmunofluorescence. *J Immunol Baltim Md 1950* 1978;121:1961–1968.

26. **Fleischmann R, Adams M, White O, Clayton R, Kirkness E, *et al.*** Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 1995;269:496–512.

27. **Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, *et al.*** Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* 2018;68:461–466.

28. **Padmanabhan R, Mishra AK, Raoult D, Fournier P-E**. Genomics and metagenomics in medical microbiology. *J Microbiol Methods* 2013;95:415–424.

29. **Meier-Kolthoff JP, G?ker M, Spr?er C, Klenk H-P**. When should a DDH experiment be mandatory in microbial taxonomy? *Arch Microbiol* 2013;195:413–418.

30. **Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M**. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:60.

31. **Klenk H-P, Meier-Kolthoff JP, Göker M**. Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. *Int J Syst Evol Microbiol* 2014;64:352–356.

32. **Klappenbach JA, Goris J, Vandamme P, Coenye T, Konstantinidis KT, *et al.*** DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.

33. **Richter M, Rosselló-Móra R**. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci* 2009;106:19126–19131.

34. **Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J**. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 2016;32:929–931.

35. **Ouk Kim Y, Chun J, Lee I, Park S-C**. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol* 2016;66:1100–1103.

36. **Chun J, Rainey FA**. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int J Syst Evol Microbiol* 2014;64:316–324.

37. **Deloger M, El Karoui M, Petit M-A**. A Genomic Distance Based on MUM Indicates Discontinuity between Most Bacterial Species and Genera. *J Bacteriol* 2009;191:91–99.

38. **Qin Q-L, Xie B-B, Zhang X-Y, Chen X-L, Zhou B-C, *et al.*** A Proposed Genus Boundary for the Prokaryotes Based on Genomic Insights. *J Bacteriol* 2014;196:2210–2215.

39. **Shpynov S, Pozdnichenko N, Gumenuk A**. Approach for classification and taxonomy within family Rickettsiaceae based on the Formal Order Analysis. *Microbes Infect* 2015;17:839–844.

40. **Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, *et al.*** A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* 2014;64:384–391.

41. **Chan JZ, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ**. Defining bacterial species in the genomic era: insights from the genus Acinetobacter. *BMC Microbiol* 2012;12:302.

42. **Klappenbach JA, Goris J, Vandamme P, Coenye T, Konstantinidis KT, *et al.*** DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.

43. **Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, *et al.*** Complete genome sequence of DSM 30083 T, the type strain (U5/41 T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genomic Sci* 2014;9:2.

44. **Gupta A, Sharma VK**. Using the taxon-specific genes for the taxonomic classification of bacterial genomes. *BMC Genomics* 2015;16:396.

45. **Thompson CC, Vicente A, Souza RC, Vasconcelos A, Vesth T, *et al.*** Genomic taxonomy of vibrios. *BMC Evol Biol* 2009;9:258.

46. **Thompson CC, Vieira NM, Vicente ACP, Thompson FL**. Towards a genome based taxonomy of Mycoplasmas. *Infect Genet Evol* 2011;11:1798–1804.

Résumé

L'Identification rapide et la classification microbienne précise sont cruciales en microbiologie médicale pour la surveillance de la santé humaine et animale, établir un diagnostic clinique approprié et choisir des mesures thérapeutiques et de contrôle optimales. Initialement, la classification taxonomique des espèces bactériennes était basée sur des caractéristiques phénotypiques. Cependant, de nombreux outils génotypiques ont été mis au point pour compléter progressivement la définition des espèces bactériennes de façon plus fiable et precise dans une approche polyphasique intégrant les caractéristiques phénotypiques, l'analyse de la similarité et la phylogénie des séquences du gène de l'ARN ribosomique 16S (ARNr 16S), la teneur en G + C de l 'ADN (G+C%) ainsi que l'hybridation ADN-ADN (DDH). Même si ces outils sont largement utilisés, ils présentent plusieurs limites et inconvénients. En effet, les seuils universels de similarité de séquence de l'ARNr 16S (95% et 98,65% aux rangs du genre et de l'espèce, respectivement), de différence en G+C % (>5% entre deux espèces) et de DDH (<70% entre deux espèces) utilisés pour la définition des espèces ne sont pas applicables à de nombreux genres bactériens. C'est notamment le cas des espèces du genre *Rickettsia*, alpha-protéobactéries strictement intracellulaires qui expriment peu de caractéristiques phénotypiques. Ainsi, la définition des espèces au sein du genre *Rickettsia* a longtemps fait l'objet de débat. Mais en 2003, l'introduction d'un outil moléculaire basé sur l'analyse des séquences de cinq gènes a révolutionné la caractérisation et la classification taxonomique des rickettsies et constitue la base de leur classification à ce jour. En dépit de tous ces efforts, la taxonomie des membres du genre *Rickettsia* est restée un sujet de débat. Au cours des deux dernières décennies, les progrès remarquables de la technologie et de l'application du séquençage de l'ADN ont permis l'accès aux séquences génomiques complètes, permettant un accès sans précédant à des données précieuses pour une classification taxonomique plus précise des prokaryotes. Plusieurs outils taxonomiques basés sur les séquences génomiques ont été développés. Compte tenu de la disponibilité des séquences génomiques de près de 100 génomes de *Rickettsia*, nous avons voulu évaluer une gamme de paramètres taxonomiques basés sur l'analyse des séquences génomiques afin de mettre au point des recommandations pour la classification des isolats au niveau de l'espèce et du genre. Nous avons également utilisé la genomique pour la caractérisation et la description des nouveaux isolats bactériens isolés par la méthode de "culturomique bactérienne" à partir de divers échantillons cliniques. En comparant le degré de similarité des séquences de 78 génomes de *Rickettsia* et 61 génomes de 3 genres étroitement apparentés (*Orientia*, 11 génomes, *Ehrlichia*, 22 génomes et *Anaplasma*, 28 génomes) en utilisant plusieurs paramètres génomiques (hybridation ADN-ADN, dDDH; l'identité nucléotidique moyenne par orthologie, OrthoANI et AGIOS; ou l'identité moyenne des séquences protéiques AAI, nous avons montré que les outils taxonomiques basés sur les séquences génomiques sont simples à utiliser et rapides, et permettent une classification taxonomique fiable et reproductible des isolats au sein des espèces du genre *Rickettsia*, avec des seuils spécifiques. Les résultats obtenus nous ont permis d'élaborer des lignes directrices pour la classification des isolats de rickettsies au niveau du genre et de l'espèce. À l'aide de la taxono-génomique, nous avons également pu décrire 17 nouvelles espèces bactériennes associées à l'homme sur la base d'une combinaison de l'analyse génomique et des propriétés phénotypiques. L'utilisation des outils génomiques est donc parfaitement adaptée à la classification taxonomique et peut changer radicalement notre vision de la taxonomie et de l'évolution bactérienne à l'avenir.

Mots clés: Génomique comparative, Génome bactérien, Taxonomie, Microbiologie, Definition d'espèce, *Rickettsia*

Abstract

Rapid identification and precise microbial classification are crucial in medical microbiology for human and animal health monitoring, appropriate clinical diagnosis and selection of optimal therapeutic and control measures. Initially, the taxonomic classification of bacterial species was based on phenotypic characteristics. However, many genotypic tools have been developed to progressively supplement the definition of bacterial species more reliably and accurately in a polyphasic approach incorporating phenotypic characteristics, analysis of similarity and phylogeny of sequences of the 16S ribosomal RNA gene (16S rRNA), the G + C content of DNA (G+C%), and DNA-DNA hybridization (DDH). Although these tools are widely used, they have several limitations and disadvantages. Indeed, the universal 16S rRNA sequence similarity thresholds (95% and 98.65% at the genus and species ranks, respectively), difference in G+C% (> 5% between two species) and DDH (< 70% between two species) used for the definition of species are not applicable to many bacterial genera. This is particularly true of species of the genus *Rickettsia* which are strictly intracellular alpha-proteobacteria that express few phenotypic characteristics. Thus, the definition of species within the genus *Rickettsia* has long been a matter of debate. But in 2003, the introduction of a molecular tool based on the analysis of five genes has revolutionized the characterization and taxonomic classification of rickettsiae and is the current basis for their classification. Despite these efforts, the taxonomy of members of the genus *Rickettsia* remained a subject of debate. Over the past two decades, the remarkable advances in DNA sequencing technologies have allowed access to complete genomic sequences, allowing unprecedented access to valuable data for a more accurate taxonomic classification of prokaryotes. Several taxonomic tools based on genomic sequences have been developed. Given the availability of genomic sequences of nearly 100 rickettsial genomes, we wanted to evaluate a range of taxonomic parameters based on genomic sequence analysis, to develop guidelines for the classification of *Rickettsia* isolates at the genus and species levels. We have also used genomic sequences for the characterization and description of new bacterial isolates isolated by the "bacterial culturomics" method from various clinical specimens. By comparing the degree of similarity of the sequences of 78 genomes from *Rickettsia* species and 61 genomes from 3 closely related genera (*Orientia*, 11 genomes; *Ehrlichia*, 22 genomes; and *Anaplasma*, 28 genomes) using several genomic parameters (DNA-DNA hybridization, dDDH; the mean nucleotide identity by orthology, OrthoANI and AGIOS; or the mean identity of protein sequences AAI, we have shown that genome-based taxonomic tools are simple to use and fast, and allow for a reliable and reproducible taxonomic classification of isolates within species of the genus *Rickettsia*, with specific thresholds. The obtained results enabled us to develop guidelines for classifying rickettsial isolates at the genus and species levels. Using taxono-genomics, we have also been able to describe 17 new human-associated bacterial species on the basis of a combination of genomic analysis and phenotypic properties. The use of genomic tools is therefore perfectly adapted to taxonomic classification and can dramatically change our vision of taxonomy and bacterial evolution in the future.

Keywords: Comparative genomics, Bacterial genome, Taxonomy, Microbiology, Species definition, *Rickettsia*