UNIVERSITE D'AIX-MARSEILLE

ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

Unité de recherche : Microbes, Evolution, Phylogeny and Infection (MEPHI)

Thèse présentée pour obtenir le grade de DOCTEUR d'AIX-MARSEILLE UNIVERSITE

Discipline : Pathologie Humaine / Spécialité : Maladies Infectieuses

Par Mr GAUDIN Maxime

Human RNA bait library depletion for human (viral) pathogen discovery using shotgun metagenomic sequencing

---

Soutenue le 23 novembre 2018

Membres du jury de la thèse :

Mr le Professeur Philippe COLSON	Président du jury
Mr le Professeur Bruno POZZETTO	Rapporteur
Mr le Docteur Philippe ROUMAGNAC	Rapporteur
Mme le Docteur Christelle DESNUES	Directrice de thèse

IHU Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille

Résumé

Le développement des techniques de séquençage de nouvelle génération (NGS) a révolutionné la recherche et le diagnostic dans le domaine des maladies infectieuses humaines. En virologie clinique, la métagénomique virale qui repose sur le séquençage aléatoire de type shotgun de l'ensemble des génomes viraux d'un échantillon (le virome), est une approche prometteuse pour la détection et l'identification sans *a priori* de potentiels nouveaux pathogènes. Cependant, son utilisation reste encore marginale en raison de l'importante contamination des viromes par les séquences nucléiques de l'hôte qui masque le signal viral, limite la reconstruction de génomes viraux et requiert une profondeur importante de séquençage, générant ainsi un cout élevé. Ces dernières années, de nombreux protocoles reposant principalement sur des étapes de filtration/centrifugation et digestions enzymatiques, ont été développés pour diminuer cette contamination humaine avec un succès limité notamment dans le cas de biopsies cliniques. Dans ce contexte, ce travail de thèse avait pour objectif d'améliorer l'approche de métagénomique pour le diagnostic clinique de maladies infectieuses virales en augmentant le ratio de séquences pathogène/hôte par déplétion des acides nucléiques humains.

Le premier chapitre de cette thèse consiste en une synthèse bibliographique des approches de métagénomique virale en recherche clinique et des challenges à relever dans ce domaine. Cette synthèse bibliographique inclut également une revue sur les approches de capture/séquençage ciblées de certains pathogènes dans le domaine des maladies infectieuses humaines.

Le deuxième chapitre de cette thèse propose une mise au point méthodologique permettant d'enrichir les métagénomes en séquences non-humaines basée sur l'hybridation et la capture de l'ensemble des acides nucléiques de l'hôte après hybridation avec des sondes ARN humaines biotinylées. La déplétion des acides nucléiques humains a été optimisée et vérifiée sur un métagénome viral artificiel constitué de proportions variables d'acides nucléiques humains et viraux (Herpes simplex virus 1). Nous avons ensuite validé son application en démontrant une réduction de plus de 90% de la contamination humaine par PCR quantitative en temps réel. Les résultats après séquençage NGS confirment une déplétion en séquences humaines de 56,5 fois et un enrichissement de 64 fois en séquences virales.

Le troisième chapitre de cette thèse est divisé en deux sous-chapitres qui proposent l'application de ce protocole à la détection d'agents potentiellement impliqués (1) dans un cas fatal d'encéphalite et (2) dans un cas énigmatique d'endocardite infectieuse à hémoculture négative. Dans le premier cas, le génome complet d'un nouveau gemycircularvirus de 2134 pb a pu être reconstruit à partir d'un échantillon de biopsie cérébrale tandis que dans le second, nous avons pu identifier une nouvelle souche de *Moraxella osloensis* à partir d'un échantillon de valve mitrale et reconstruire un génome *quasi*-complet avec une couverture moyenne > 200X.

Enfin, dans un quatrième chapitre, l'approche méthodologique que nous avons développée est discutée et les résultats sont replacés dans un contexte élargi d'émergence des maladies infectieuses et de lien de causalité entre l'agent détecté et la pathologie observée.

Abstract

The development of Next Generation Sequencing (NGS) techniques has revolutionized research and diagnostic in the field of human infectious diseases. In clinical virology, viral metagenomics, which is based on the random shotgun sequencing of all viral genomes present in a sample, is a promising approach for blind detection and identification of potential new pathogens. Its use is however still marginal because of the large proportion of human nucleic sequences which masks the viral signal, limits the reconstruction of viral genomes and requires a ultra-deep sequencing, thus generating higher sequencing costs. In recent years, numerous protocols based on filtration/centrifugation and nuclease digestion steps have been developed to reduce human contamination with limited success particularly in the case of clinical biopsies. In this context, this thesis work aims at improving the metagenomic approach for the clinical diagnosis of viral infectious diseases by increasing the ratio of pathogen-to-host sequences trough depletion of human nucleic acids from the samples.

The first chapter of this thesis consists in a bibliographic synthesis of viral metagenomic approaches in clinical research and the challenges we faced in this field. This bibliographic overview also includes a review article on targeted-enrichment sequencing approaches for pathogen detection in the field of human infectious diseases.

The second chapter of this thesis proposes a methodological development allowing the enrichment of non-human sequences from metagenomes through hybridization and capture of human nucleic acids with biotinylated human RNA probes. Depletion of human nucleic acids was optimized and verified on a mock viral metagenome consisting of varying proportions of human and viral nucleic acids (Herpes simplex virus 1). We then validated its application by reducing human contamination by more than 90% as revealed by real-time quantitative PCR. The results after NGS sequencing confirm an average depletion of 56.5-fold for human sequences and an enrichment of 64-fold for viral sequences.

The third chapter of this thesis is divided into two sub-chapters that propose the application of this protocol to the detection of putative pathogens in (1) a fatal case of encephalitis and (2) an enigmatic case of blood-culture negative infectious endocarditis. In the first case, the 2,134 bp complete genome of a new gemycircularvirus was reconstructed from a cerebral biopsy sample while in the second, we identified a new strain of *Moraxella osloensis* from a mitral valve sample and reconstructed its nearly-complete genome with an average coverage >200X.

The methodological approach developed during this work is finally discussed in a fourth chapter, which also replaces the results obtained in the broader context of emerging infectious diseases and validation of the causal link between the agent detected and the observed pathology.

Remerciements

Dans un premier temps, je tenais à remercier vivement le docteur Christelle Desnues, directrice de l'équipe Pathovirome au sein de l'unité de recherche MEPHI et qui fut également ma directrice de master puis de thèse au cours de ces quatre dernières années. En travaillant à ces cotés j'ai pu m'épanouir et apprendre tout au long de mon projet car j'ai bénéficié à la fois d'un très bon accueil au sein de son équipe, de ses qualités humaines, d'un encadrement maniant à la fois rigueur scientifique et sens critique, de sa confiance ainsi qu'une très grande disponibilité pour mener à bien mon travail.

J'exprime également ma gratitude à la fois au Professeur Didier Raoult ainsi qu'au professeur Michel Drancourt de m'avoir accueilli et permis de réaliser ma thèse au sein de leurs laboratoires.

Je voudrais également exprimer ma plus grande gratitude aux Pr. Bruno Pozzetto et Philippe Roumagnac de m'avoir fait l'honneur d'évaluer mon travail et de participer à ce jury comme rapporteur, ainsi qu'au Pr. Philippe Colson d'avoir accepté d'être examinateur de ce travail.

Je tiens aussi à adresser mes remerciements à tous ceux qui m'ont aidé dans l'accomplissement de ce travail et plus particulièrement à tous les membres actuels ou anciens de mon équipe. A ce titre, je remercie vivement Sonia Monteil Bouchard, Priscilla Jardot, Sarah Temmam, Nicolas Rascovan, Sébastien Halary, Alexia Bordigoni, Stéphanie Gambut, Caroline Autréau et Laure Sauvat pour leur aide, leur soutien et leur bonne humeur communicative. Mention spéciale à ma petite Kelly Goldlust, étudiante en master au sein de mon équipe qui, plus qu'une collègue de travail est devenue une véritable amie si ce n'est ma meilleure (« On ne se quitte plus maintenant »).

De plus travailler au sein d'un laboratoire si riche tant sur le plan culturel que professionnel m'a permis d'établir un lien privilégié avec des personnes extraordinaires. Soyez en sûr que même si je ne sais pas où l'avenir me mènera, je n'oublierai jamais tous nos fous rires, nos débats et nos discussions que j'ai bien pu avoir avec vous. C'est pour cela que je ne vous ferai pas un adieu le jour de mon départ parce que je sais qu'on gardera contact et c'est dans cette optique que je tenais à remercier énormément Sarah Aherfy, Marielle Bedotto, Michelle Estel, Emeline Baptiste, Remi Barbieri, Estelle Menu, Eya Ben-Azzouz, Asma Boumaza, Lina Barassi, Enzo Parisi ainsi que Cheraz Riabi.

Merci aussi à l'ensemble du personnel de l'URMITE : ingénieurs, techniciens, secrétaires, étudiants, car je ne peux pas tous les citer mais sans qui tout aurait été plus compliqué.

Enfin, je ne pourrais pas finir ces remerciements sans penser à ma famille pour avoir toujours cru en moi et dont l'affection, l'amour, le soutien et l'encouragement constants m'ont été d'un grand réconfort et ont contribué à l'aboutissement de ce travail.

Table des matières

Liste des figures et tableaux

Cette liste ne tient pas compte des figures et des tableaux contenus dans les articles.

Figures introduction bibliographique et discussion

Tableaux introduction bibliographique et discussion

Tableau 1. Exemples de virus découverts par métagénomique virale et leur implication dans des pathologies.

Tableau 2. Méthylation durant la phase active de réplication ou la phase de latence de certains familles de virus à ADN.

Liste des abréviations

ADN : Acide désoxyribonucléique

ADNc : ADN complémentaire

AN : Acide nucléique

ARN : Acide ribonucléique

ARNr : ARN ribosomique

CMV : Cytomégalovirus

DOP-PCR : Degenerate Oligonucleotide Primed-PCR (PCR utilisant des oligonucléotides dégénérés)

dsDNA : ADN double brin

dsDNA-RT : ADN double brin avec une phase intermédiaire à ARN

EBV : Epstein-Barr Virus (Virus d'Epstein-Barr)

EEG : électroencéphalogramme

EI : Endocardite Infectieuse

EIHN : Endocardites Infectieuses à Hémocultures Négatives

FISH : Hybridation in situ en fluorescence

HERV : Human endogenous retrovirus (Rétrovirus endogène humain)

HSV-1 : Herpes Simplex virus type 1 (virus herpès simplex de type 1)

HTS : High-Throughput Sequencing (Séquençage à haut-débit)

ICTV : Comité International de la Taxonomie des Virus

IHC : Immunohistochemistry (Immunohistochimie)

IRM : Imagerie par Résonance Magnétique

ITS : Internal Transcribed Spacer (espaceurs internes transcrits)

MDA : Multiple Displacement Amplification (Amplification par déplacements multiples de brin)

NCLDV : NucleoCytoplasmic Large DNA Viruses (Grands virus nucléocytoplasmiques)

NGS : Next Generation Sequencing (Séquençage de nouvelle génération)

OMS : Organisation Mondiale de la Santé

PBS : Tampon phosphate salin

PCR : Polymerase Chain Reaction (Réaction en chaîne par polymérase)

RCA : Rolling Circle amplification (amplification en cercle roulant)

ssRNA-RT : ARN sens simple brin ayant une étape intermédiaire à ADN

SARS : Syndrome Aigu Respiratoire Sévère

SIA : Sequence-Independent Amplification (Amplification séquence-indépendante)

SIDA : Syndrome de l'ImmunoDéficience Acquise

SISPA : Sequence-Independent Single Primer Amplification

ssDNA : ADN simple brin

(-)-ssRNA : ARN anti-sens simple brin

(+)-ssRNA : ARN sens simple brin

VLPs : Virus-Like Particles (Particules virales circulantes)

VZV : Virus varicelle-zona (VZV)

WISC : Whole-genome In-Solution Capture

WGS : Whole Genome Sequencing (Séquençage de génome entier)

Chapitre I : Introduction bibliographique

I. Découverte des virus, classification et mécanismes d'infection

1. *Historique de la découverte des virus*

Les maladies infectieuses virales, telles que la variole ou la rage, sont connues depuis l'Antiquité. En effet, au XVIe siècle, des médecins chinois pratiquaient déjà une technique appelée « la variolisation ». Pour cela ils récupéraient le contenu des vésicules de patients malades qu'il laissait sécher à l'air libre, puis ils inoculaient cette forme atténuée à des personnes saines, afin de les immuniser [1]. Ce principe fut importé en occident au début du XVIIIe et mis au point par un médecin anglais, Edward Jenner, qui immunisa des patients sains en leur inoculant la variole bovine ou « cowpox ». Le principe de la vaccination a ainsi vu le jour [2].

Il faudra attendre 1892, pour que le premier grand pas vers la découverte des virus ait lieu. Le biologiste Dimitri Ivanosky, en travaillant sur la maladie de la mosaïque du tabac, démontra que le filtrat obtenu après extraction d'un broyat de feuilles malades et passage à travers un filtre de porcelaine (conçu pour arrêter les bactéries) était encore infectieux. N'ayant pas pris toute la mesure de sa propre découverte, il faudra attendre 1898 pour que le microbiologiste Martinus Willem Beijerinck reprenne ces travaux en démontrant que la sève de plants de tabac infectés était capable de retransmettre la maladie lorsqu'elle était inoculée à des plantes saines. De ces travaux a émergé la notion de virus décrit comme des agents infectieux invisibles, plus petits que les bactéries et capables de se multiplier dans les cellules vivantes [1].

Avec cette nouvelle notion de virus, la « virologie » a ainsi pu voir le jour. Cette nouvelle branche de la microbiologie se définit comme l'étude des virus et des maladies qu'ils engendrent. L'histoire de la virologie repose quant à elle sur trois dates clés :

- 1901 avec la découverte du premier virus humain et responsable de la fièvre jaune par Walter Reed, James Carroll and Jesse Lazear [1].
- 1915-1921 avec la découverte de virus capable d'infecter et de se répliquer à l'intérieur des bactéries par Frederick William Twort et Felix d'Hérelle. Ce dernier les nommera par la suite des « bactériophages » [1].
- 2008, et la découverte du premier virophage : un virus capable d'infecter un autre virus par l'équipe du professeur Didier Raoult [3].

2. *Définition et classification des virus*

Les virus sont des parasites obligatoires qui se répliquent uniquement à l'intérieur d'une cellule hôte contaminée. Ils présentent une grande variabilité en termes de taille, de structure de leur capside, de diversité de leur organisation génomique et de stratégie de réplication. Cependant, les virus partagent des caractères précis :

- Ils sont constitués d'un génome composé d'une ou plusieurs molécules d'acide nucléique (ADN ou ARN).
- Ils se comportent comme des parasites intracellulaires stricts car ils sont capables de se répliquer uniquement en détournant la machinerie cellulaire des cellules infectées.
- Les virus sont capables d'infecter chaque domaine du vivant : eucaryotes, bactéries et archées.
- Le matériel génétique est protégé par une structure protéique nommée la capside éventuellement entourée d'une enveloppe lipidique provenant des membranes des cellules hôtes.

Le terme virus définit la forme présente dans les cellules infectées tandis que le virion (ou particule virale) est la forme libre présente dans la circulation et capable d'infecter d'autres cellules cibles [4]. Historiquement, les virus étaient classés en fonction des maladies qu'ils provoquaient. L'utilisation de cette classification fut définitivement abandonnée avec l'avènement de la biologie moléculaire. Désormais, les virus sont le plus souvent classés suivant la classification de Baltimore basée en fonction du type de support de l'information génétique (ADN, ARN, simple ou double brin). Il y a sept groupes viraux dans cette classification : ADN double brin (dsDNA), ADN simple brin (ssDNA), ARN double brin (dsRNA), ARN anti-sens simple brin ((-)-ssRNA), ARN sens simple brin ((+)-ssRNA), ARN sens simple brin ayant une étape intermédiaire à ADN (ssRNA-RT), ADN double brin avec une phase intermédiaire à ARN (dsDNA-RT) *(https://viralzone.expasy.org/)*. Au niveau taxonomique le Comité International de la Taxonomie des Virus (ICTV) ne reconnaît que 5 niveaux taxonomiques : ordre, famille, sous-famille, genre et espèce. En 2017, l'ICTV a répertorié 8 ordres, 122 familles, 35 sous-familles, 735 genres et 4404 espèces virales existantes [5].

3. *Les mécanismes d'infection virale*

La majorité des infections virales sont aiguës (**Figure 1**). Certains exemples très connus [6] comme les grippes ou les gastro-entérites, aboutissent après plusieurs jours de manifestations cliniques (liées à la fois à la réplication virale et à la réponse immunitaire forte de l'hôte) à l'éradication de l'infection avec une immunité protectrice d'une durée variable contre (et en fonction) du type de virus impliqué. L'évolution des maladies virales aiguës dépend de la virulence du virus et de l'hôte et les réactions de défense sont différentes d'un sujet à l'autre. Enfin, de nombreuses infections virales aiguës sont asymptomatiques et seule la présence d'anticorps révèle la trace de l'infection comme c'est très souvent le cas pour le cytomégalovirus (CMV) ou le virus de la rubéole.

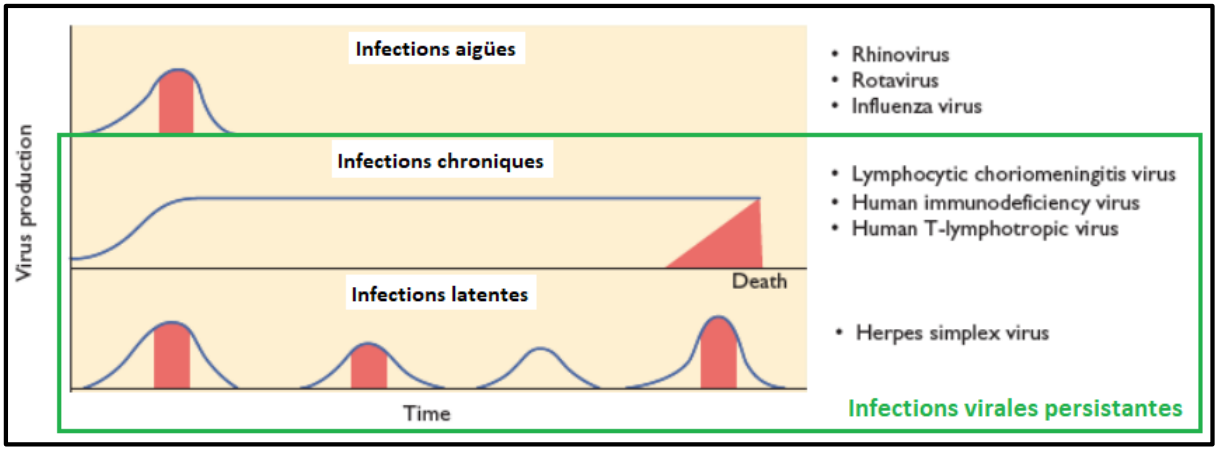


Figure 1 : Pathogénèse des infections virales chez l'homme.

La disparition des symptômes et la régression de la maladie ne s'accompagne pas forcement de l'élimination complète de l'agent infectieux et certains virus ont réussi à adopter des stratégies leur permettant d'établir des relations durables avec l'hôte infecté. Ces infections, dites « persistantes », résultent d'un équilibre entre les défenses de l'organisme et les moyens mis en jeu par le virus pour échapper au système immunitaire. Les infections virales persistantes sont soit latentes, soit chroniques (**Figure 1**). Dans les infections virales latentes, l'infection aigüe est suivie d'une période durant laquelle le virus rentre en dormance dans ces cellules cibles. Au cours de cette période dépourvue de signe clinique, aucune particule virale infectieuse ne peut être isolée. Lorsque cet équilibre est rompu, le virus rentre dans une phase de « réactivation » caractérisée par une reprise de la production de particules virales infectieuses et associée à un tableau parfois symptomatique caractéristique. Le virus

jusqu'alors silencieux peut ainsi déclencher une maladie potentiellement plusieurs années après l'atteinte initiale [7]. A l'inverse, une infection chronique est caractérisée par la présence en permanence de particules virales infectieuses, notamment dans le sang (**Figure 1**). Une personne ayant une infection chronique peut transmettre le virus à n'importe quel moment, tandis qu'une personne ayant une infection latente n'est contagieuse qu'au cours des phases de réactivation [7, 8]. La persistance chronique dans le sang et dans les organes cibles des virus sans déclencher de forte réponse immunitaire de l'hôte peut engendrer des dommages progressifs sur les cellules cibles pouvant conduire au développement de pathologie inflammatoires ou cancers (**Figure 2**) [7].

VIRUS	SIGNES DE LA PRIMO-INFECTION	SITE DE LA PERSISTANCE	FORME DE L'AGENT INFECTIEUX PERSISTANT	SIGNES DE LA RÉCURRENCE	CONSÉQUENCES POSSIBLES À LONG TERME
VIRUS DE L'HERPÈS	VÉSICULES OU AUCUN	NEURONES	ADN	VÉSICULES	AUCUNE
VIRUS DE LA VARICELLE	VARICELLE	NEURONES ET/OU CELLULES ENVIRONNANTES	ADN	ZONA	AUCUNE
VIRUS EPSTEIN-BARR	MONONUCLÉOSE INFECTIEUSE OU AUCUN	LYMPHOCYTES *B* CELLULES ÉPITHÉLIALES	ADN	AUCUN	LYMPHOME DE BURKITT CANCER NASOPHARYNGÉ
PAPILLOMAVIRUS	AUCUN	CELLULES ÉPITHÉLIALES	ADN ET VIRUS INFECTIEUX		CARCINOME
VIRUS DE LA ROUGEOLE	ROUGEOLE	NEURONES ET/OU CELLULES ENVIRONNANTES	PARTICULES VIRALES IMMATURES		PANENCÉPHALITE SUBAIGUË SCLÉROSANTE
HIV	AUCUN OU VARIABLES	LYMPHOCYTES *T* CD4 MACROPHAGES CELLULES MICROGLIALES	ADN ET VIRUS INFECTIEUX		SIDA
VIRUS DE L'HÉPATITE B	HÉPATITE OU AUCUN	CELLULES DU FOIE	VIRUS INFECTIEUX		HÉPATITE CHRONIQUE CIRRHOSE CANCER DU FOIE

Figure 2 : Exemple d'infections virales persistantes : virus latents au sens strict (jaune) et virus responsables d'infections chroniques (vert).

II. Outils historiques et modernes pour l'étude des virus humains

1. *Outils historiques dans la découverte des virus chez l'homme*

C'est grâce à l'avènement de techniques innovantes telles que la microscopie électronique, la culture cellulaire, la diffraction des rayons X, et la biologie moléculaire que les scientifiques ont pu progresser dans l'identification des virus, leur isolement, la compréhension de leur mécanisme de réplication et dans la santé humaine (diagnostic fiable et vaccination) [9].

L'utilisation de la culture cellulaire en virologie a débuté au début de XXe siècle mais était réservée à la production de particules virales à des fins de vaccinations [10]. Le premier virus humain cultivé sur des œufs de poule embryonnés était celui de la grippe entre 1931 et 1933. Avec l'avènement des antibiotiques en 1929 et leur utilisation comme adjuvant en culture cellulaire, de nombreux virus tel que les herpes virus ou le virus de la rubéole furent isolés [1]. C'est en 1983 que la culture cellulaire a réalisé l'une de ses plus grandes découvertes avec l'isolement d'un rétrovirus capable de se multiplier dans les lymphocytes et responsable du syndrome d'immunodéficience acquise (SIDA) [11]. L'apport de la microscopie électronique et de la sérologie n'en reste pas moins important car ils ont permis d'identifier les virus responsables de l'hépatite A et B au début des années 1970 [12, 13]. Par la suite, l'émergence de la biologie moléculaire comprenant les techniques de clonage, le séquençage de Sanger et par la suite l'utilisation de la technique de réaction de polymérisation en chaine (PCR) a permis de découvrir de nombreux virus comme par exemple celui de l'hépatite E [14] ou celui responsable du syndrome respiratoire aigüe sévère (SRAS) [15]. Par la suite, l'utilisation de l'ensemble de ces outils ont conduit à la découverte fortuite d'*Acanthamoeba polyphaga* mimivirus, un membre de la famille des grand virus nucléocytoplasmiques (en anglais ; nucleocytoplasmic large DNA viruses (NCLDV)), infectant l'amibe *Acanthamoeba polyphaga* et introduisant ainsi la notion du « quatrième domaine de vie » [16, 17].

Malgré l'existence, comme décrit précédemment, d'un panel d'outil utilisable pour identifier et caractériser de nouveaux virus certaines difficultés résident. Celles-ci sont principalement liées au fait que de nombreux virus ne sont pas capable de se multiplier *in-vitro* [18], qu'il est nécessaire d'avoir préalablement isolé et caractérisé le virus pour les techniques immunologiques et, qu'a l'inverse des bactéries avec leur gène codant la sous-unité 16S de l'ARN ribosomal (ARNr), il n'existe aucun gène conservé chez tous les virus, limitant ainsi leur découverte par PCR [19]. Les approches de métagénomique, qui consistent à séquencer de manière aléatoire l'ensemble des génomes d'un échantillon (un métagénome) en s'affranchissant de la culture et des besoins de connaissance sur les cibles potentielles, ont permis de faire un bond en avant dans ce domaine.

2. *La métagénomique et les NGS*

La première étude de métagénomique virale appliquée à l'homme s'est intéressée aux virus à ADN d'un échantillon de selles d'un donneur sain et a permis de révéler une diversité jusqu'alors insoupçonnée du virome intestinal humain [20]. Le virome humain est défini comme la collection de tous les virus eucaryotes, d'archées et de bactéries qui se trouvent à l'intérieur et à l'extérieur de notre corps en condition physiologique ou pathologique [21–23]. En effet, nous sommes tous des "méta-organismes" abritant, en termes d'abondance, plus de cellules bactériennes, d'archées et de particules virales que nos propres cellules humaines [24, 25]. Le virome humain est composé de particules virales libres (virus eucaryotes transitoires ou résidants susceptibles de provoquer une infection aiguë symptomatique ou non, bactériophages et virus d'archées) ainsi que d'éléments viraux intracellulaires (infections virales persistantes et anciens éléments dérivés de virus insérés dans nos chromosomes) (**Figure 3**).

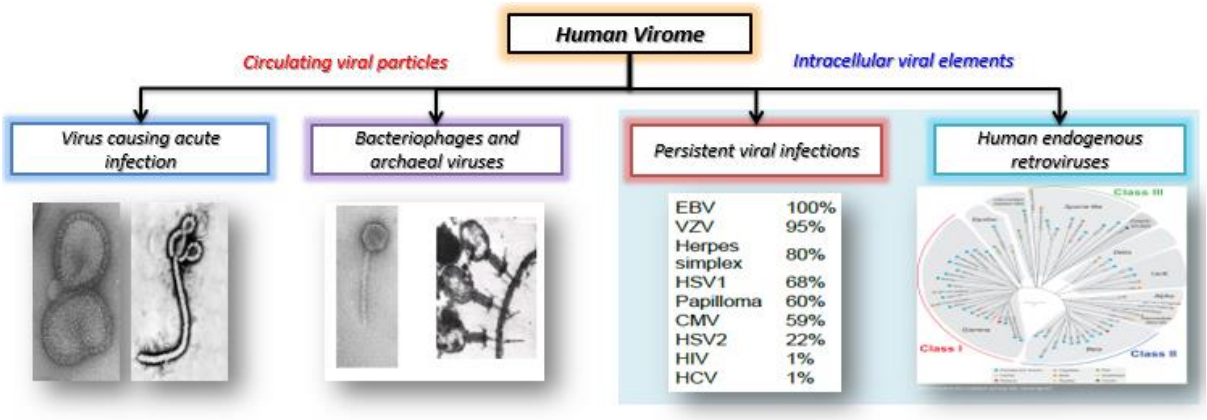


Figure 3 : Le virome humain : collection de tous les virus présents à l'intérieur ou à l'extérieur de notre organisme.

Les premiers viromes publiés utilisaient la technique de clonage et de séquençage Sanger qui ne permettait d'obtenir que quelques milliers de fragments [26]. L'apparition et l'évolution très rapide des techniques de séquençage haut-débit (HTS pour High-Throughput Sequencing ou NGS pour Next-Generation Sequencing) [27] ont profondément modifié les données obtenues et les analyses qui en découlent. Ainsi, du fait de sa capacité à générer rapidement des millions de séquences en une seule réaction pour un coût de plus en plus bas [26], cette approche offre donc des possibilités sans précédent pour la caractérisation du virome humain.

Contrairement à l'approche dite d'amplicon sequencing qui cible un gène ou une région conservé au sein d'une même espèce (comme le gène codant l'ARN ribosomal 16S chez les bactéries et archées [28], l'ARN ribosomal 18S chez les eucaryotes ou plus précisément la région comprenant les espaceurs internes transcrits (ITS) chez les champignons [29]), la métagénomique virale repose sur une approche de séquençage shotgun (séquençage aléatoire) qui consiste à extraire les acides nucléiques (AN) totaux d'un échantillon complexe (AN humains, bactériens, archées, viraux), réaliser une étape de reverse-transcription si on s'intéresse aux métagénomes ARN (i.e., générer les ADN complémentaires ou ADNc), et séquencer la totalité du contenu génomique par NGS sans amplification préalable (**Figure 4,** [30]).

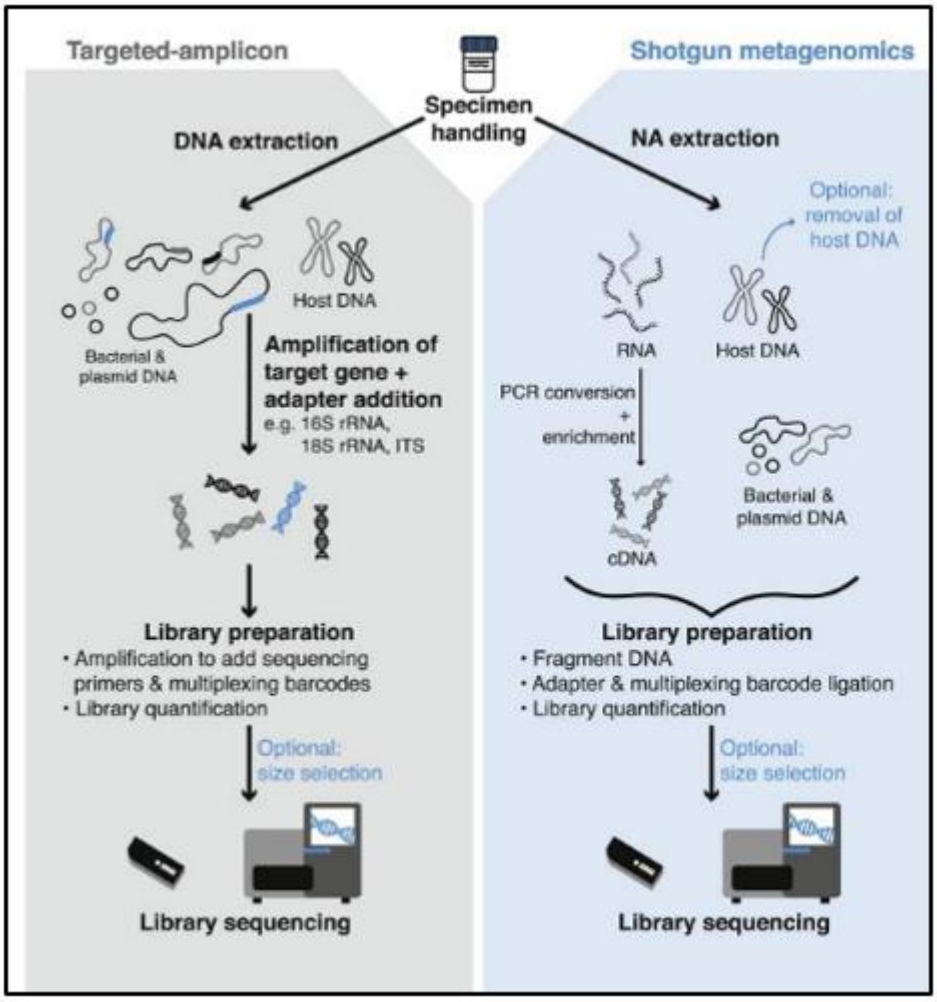


Figure 4 : Principales techniques de séquençage à haut débit utilisées en métagénomique clinique, d'après [30].

Plusieurs approches peuvent être utilisées pour réaliser un métagénome viral. Ces approches et leurs limites sont présentées dans la **partie IV** de cette introduction. Ainsi, la métagénomique virale a permis de révéler une diversité insoupçonnée du virome humain mais également permis un meilleur diagnostic et une meilleure gestion des maladies infectieuses virales chez l'homme.

III. La métagénomique virale pour la recherche et le diagnostic clinique

1. *Les maladies infectieuses humaines*

Les maladies infectieuses représentent de nos jours l'un des plus grands défis en santé publique. En effet, malgré les avancés thérapeutiques considérables, trois pathologies infectieuses étaient classées parmi les 10 causes majeures de mortalités dans le monde en 2016 : les infections respiratoires basses, les diarrhées et la tuberculose (**Figure 5**).

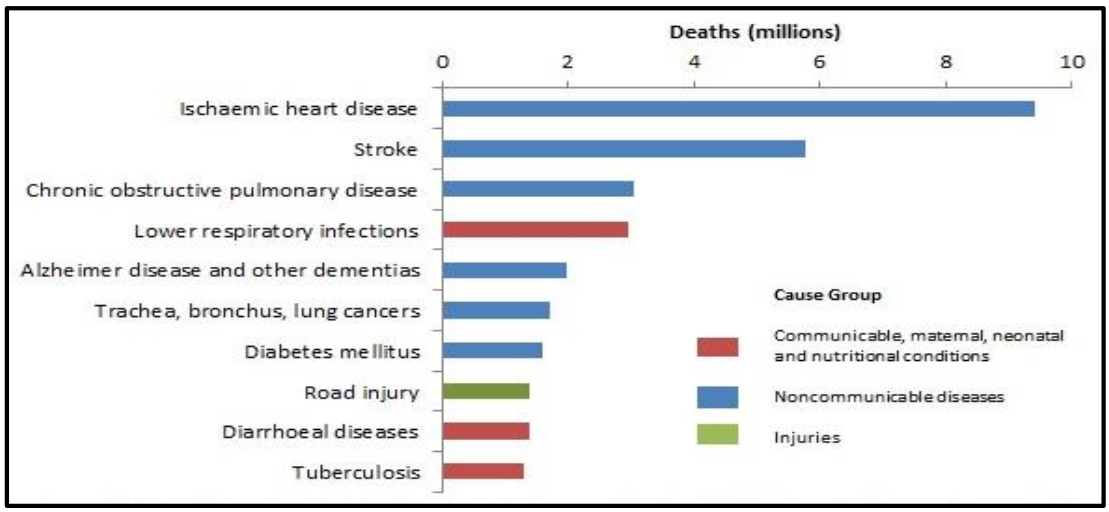


Figure 5 : Top 10 des causes de décès dans le monde. *Source: Global Health Estimates 2016. Deaths by cause, age, sex, country and by country and by region. World Health Organisation, 2018.*

Les maladies infectieuses transmissibles (notamment des voies respiratoires basses) représentent une la première cause de mortalité dans les pays à faibles revenus et la 6ème dans ceux à revenus élevés. A ce constat, il faut ajouter que, malgré l'arsenal diagnostic disponible dans les pays développés, le taux d'échec dans la détermination des causes étiologiques de certaines pathologies principalement virales peut s'avérer très élevé, variant par exemple de 30

à 85% pour les encéphalites [31], 40 à 80% pour les pneumonies d'origine communautaire [32] et 40% pour la gastro-entérite [33].

A l'échelle mondiale, une autre menace dans le domaine des maladies infectieuses réside dans l'émergence de nouvelles maladies transmissibles chez l'homme et la réémergence de maladies anciennes qui avaient disparues. C'est à la suite de l'épidémie de syndrome respiratoire aigu sévère lié à coronavirus (le SARS-CoV) en 2003 [34] que l'organisation mondiale de la santé (OMS) a créé son comité d'urgence en 2005. Depuis l'état d'urgence sanitaire internationale a été déclaré 4 fois et concernait la pandémie de grippe H1N1 en 2009, la résurgence de la poliomyélite en 2014, l'épidémie d'Ebola en Afrique de l'Ouest en 2014 et l'émergence du virus Zika en Amérique en 2016 [35]. Ces quatre alertes ont été déclenchées par des infections virales qui représentent la majorité des maladies infectieuses émergentes ou ré-émergentes comme l'illustre la **figure 6** [36] et dont une grande partie trouve leur origine dans des réservoirs animaux [37].

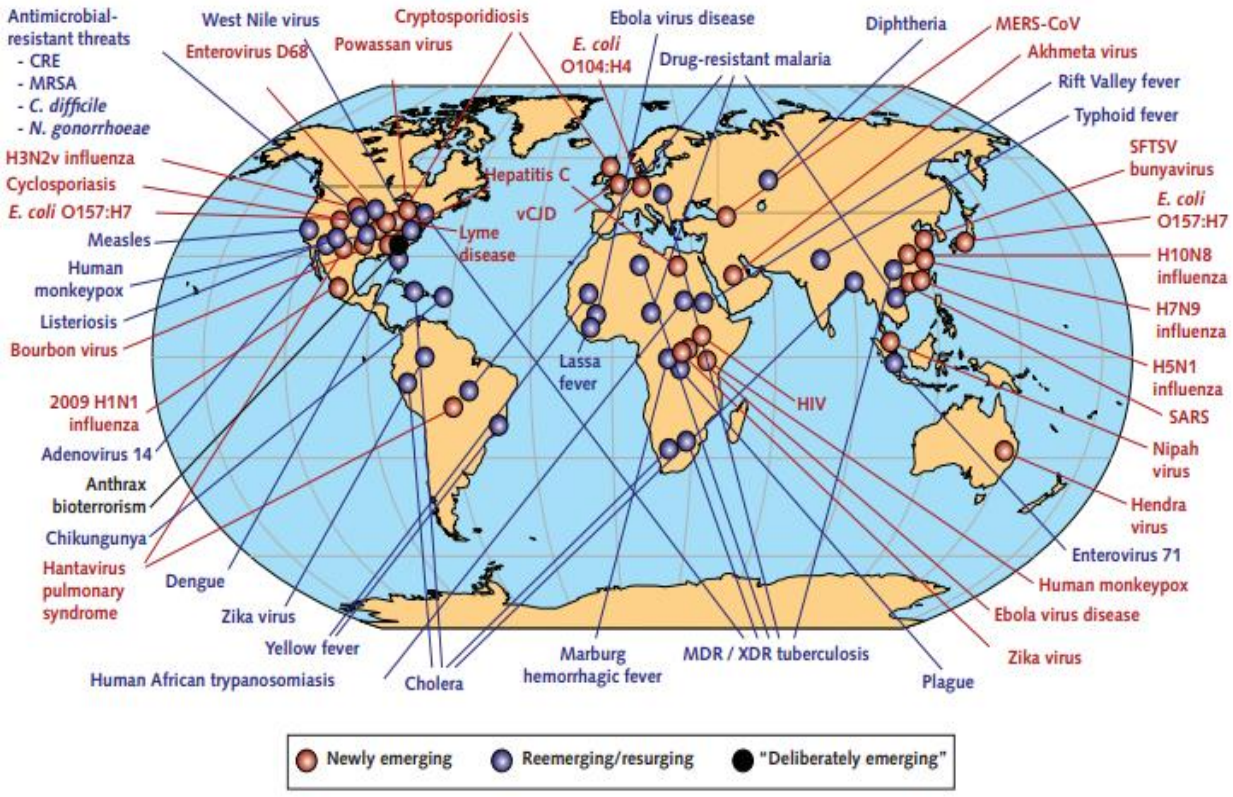


Figure 6 : Exemples de maladies infectieuses émergentes ou ré-émergentes de 1977 à 2007, d'après [36].

Les changements climatiques, la mondialisation des échanges et l'expansion rapide de la population humaine au détriment des forêts favorisent d'une part la propagation rapide d'agents pathogènes non identifiés capables de provoquer des épidémies humaines dévastatrices et d'autre part un contact plus étroit avec la faune sauvage favorisant l'émergence [38]. Les animaux domestiques et sauvages ainsi que les arthropodes hématophages tels que les moustiques et les tiques sont des réservoirs et/ou vecteurs reconnus de nouveaux agents pathogènes viraux [39]. Enfin, le lien entre les nouveaux virus et la maladie ne se limite pas aux maladies aiguës, mais peut également être observé dans des états chroniques, comme le montre la forte association entre l'infection par le nouveau polyomavirus à cellules de Merkel et une maladie tumorale très agressive chez les patients âgés [40].

Un modèle statistique des tendances temporelles de la découverte de virus humains développé en 2008 montrait que l'identification de nouveaux agents infectieux viraux était loin d'être complète et prédisait 10 à 40 nouveaux virus humains découverts d'ici 2020 et probablement des centaines à découvrir dans le futur [41]. On comprend donc ici toute l'importance de définir l'étendue de la diversité du virome humain et d'identifier des agents pathogènes viraux (en phase aigüe ou chronique) pour la santé humaine.

2. La métagénomique virale comme outil de diagnostic en maladie infectieuse

Avec sa capacité théorique à répertorier la totalité du contenu génétique de tous les organismes présents dans un échantillon clinique, la métagénomique se positionne comme un des outils indispensables pour la surveillance des virus circulants chez l'homme et dans la détection de virus encore inconnus ou inattendus dans certaines pathologies. Fin 2017, environ 65 études publiées ont utilisé la métagénomique pour la détection d'agents pathogènes (tous organismes confondus) majoritairement pour le diagnostic au cours d'infections respiratoires (29,2%), neurologiques (26,2%), cardiaques/sanguines (13,9%) et gastro-intestinales (10,8%) (**Figure 7**, [30]).

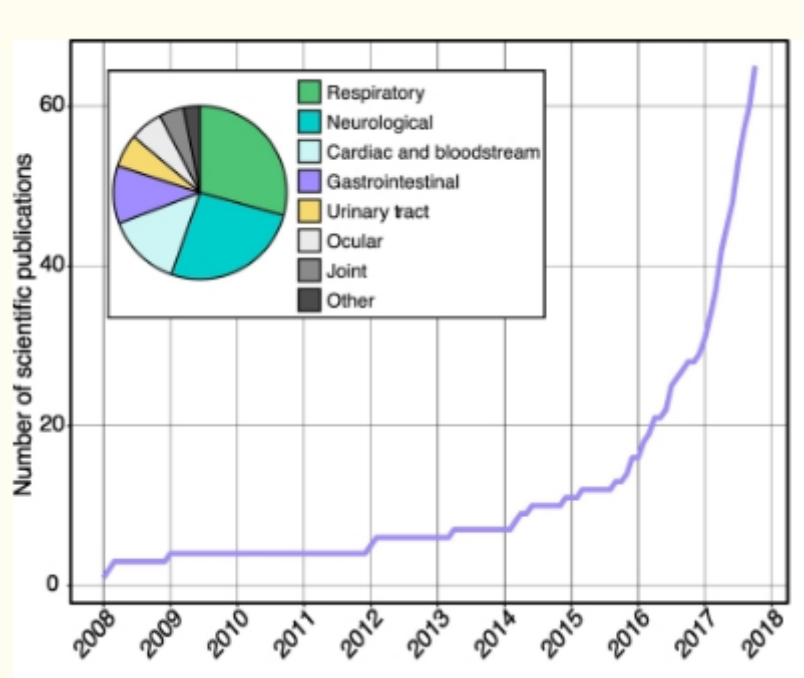


Figure 7 : Rythme du nombre de publications utilisant la métagénomique pour la détection de pathogènes en maladie infectieuse chez l'homme, d'après [30].

La métagénomique a ainsi démontré son utilité pour la découverte de nouveaux virus encore inconnus chez l'homme et pouvant être impliqué dans le phénotype d'une maladie idiopathique (**Tableau 1 - A**), la détection d'un virus déjà connu chez l'homme pour provoquer ou non une pathologie établie mais impliqué ici dans un contexte clinique inattendu (**Tableau 1 -B**) et enfin, le diagnostic d'un virus connu pour provoquer la maladie d'un patient, mais rarement testé du fait d'une probabilité faible d'en être l'agent (**Tableau 1 - C**).

Tableau 1. Exemples de virus découverts par métagénomique virale et leur implication dans des pathologies.

Virus	*Technologie*	*Maladie associée*	*Pathogénicité*	*Corrélation virus / pathologie*	*Référence*
A: Découverte de nouveaux virus infectant l'homme					
Merkel Cell polyomavirus	Sanger	Carcinome à cellules de Merkel	Chronique	++	[40]
Arenavirus	Roche 454	Maladie fatale après transplantation	Aiguë	++	[42]
African Swine Fever Virus	Roche 454	Affection fébrile aiguë	Aiguë	-	[43]
Lujo virus	Roche 454	Fièvres hémorragiques virales	Aiguë	++	[44, 45]
Bas-Congo virus	Roche 454		Aiguë	++	[46]
Variegated Squirrel Bornavirus 1	Illumina	Affection du système nerveux central (encéphalite, méningo-encéphalite, paraplégie...)	Aiguë	++	[47, 48]
Gemycircularvirus	Illumina		Aiguë	++	[49, 50]
Cyclovirus	Roche 454		Aiguë	+	[51, 52]
Severe fever with thrombocytopenia virus	Illumina	Fièvre et thrombocytopénie	Aiguë	++	[53, 54]
Giant Blood Marseillevirus	Roche 454	Bonne santé Adénopathie non fébrile Lymphome	Aigue or Chronique	+	[55, 56] [57] [58]
Astrovirus VA1	Roche 454	Gastro-entérite	Aiguë	+	[59]
B: Découverte de virus déjà connus pour infecter l'homme mais inattendus dans ces maladies					
Astrovirus	Roche 454	Affection du système nerveux central (encéphalite, méningo-encéphalite, paraplégie...)	Aiguë	++	[60–62]
Human Coronavirus OC43	Illumina		Aiguë	++	[63]
Hepatitis G virus	Illumina		Aiguë	+	[64]
Mumps vaccine virus	Illumina		Chronique	+	[65]
Virus respiratoire syncytial	Illumina		Aiguë	++	[66]
Hepatitis G virus	Roche 454	Fatigue chronique Sclérose en plaques	Chronique	-	[67, 68] [69]
Cyclovirus ChileNPA1	Illumina	Infection respiratoire	Aiguë	+	[70]
Human Rhinovirus B91	?	Pneumonie	Aiguë	++	[71]
Human polyomavirus-6	Illumina	Maladie de Kimura	Chronique	+	[72]
Trichodysplasia spinulosa-associated polyomavirus	Illumina	Myocardite	Aiguë	-	[73]
C: Découverte de virus connus pour être responsables de la maladie mais oubliés des tests diagnostics					
St. Louis Encephalitis Virus	Illumina	Affection du système nerveux central (encéphalite, méningo-encéphalite)	Aiguë	++	[74]
West Nile virus	Illumina		Aiguë	++	[75]
Enterovirus 71	Roche 454	Infection respiratoire	Aiguë	++	[76]
Rubella virus	Illumina	Uvéite chronique	Chronique	++	[77]
Hepatitis A virus	Illumina	Affection fébrile aiguë	Aiguë	++	[78]

Malgré son succès évident, il existe encore de nombreuses limites pour l'utilisation en routine de la métagénomique en diagnostic clinique dans le domaine des maladies infectieuses virales humaines. Outre les limites de coût de séquençage, de temps et de complexité d'analyse des viromes qui sont bien décrites [79, 80] et que nous n'aborderons pas ici, il existe également de nombreuses limites biologiques notamment de sensibilité en raison des faibles charges virales, de la taille généralement petite des génomes viraux et de l'importante contamination provenant des acides nucléiques de l'hôte ou d'autres organismes (bactéries, archées, micro-eucaryotes. Pour pallier à ces limites, différentes approches sont décrites dans la littérature pour générer un métagénome viral. L'objectif ultime de ces approches est d'enrichir en séquences virales les données issues du séquençage.

IV. Les challenges de la métagénomique virale en recherche et diagnostic clinique

Différents protocoles ont donc été utilisés pour préparer un métagénome viral. Dans les protocoles directs, les acides nucléiques totaux sont directement extraits de l'échantillon et séquencés sans aucun traitement préalable. Ces protocoles introduisent peu de biais (hormis celui de la méthode d'extraction) et permettent d'accéder théoriquement à la totalité du virome humain (i.e., particules virales circulantes et les éléments viraux intracellulaires). Ces protocoles se sont ainsi avéré performant pour la détection de virus pathogènes ([42], [43], [59], [63], [65], [73], [74], [75], [77], [78], [81]), pour la détection de variant viraux rares [82] et dans la reconstruction de génomes complets [83]. Cependant, ils entrainent une contamination très importante avec des acides nucléiques non-viraux (notamment humains) ([62], [73], [81], [84], [85], [86], [87], [88]) qui nécessite une profondeur de séquençage très élevée entrainant un coût associé non négligeable. Malgré cette profondeur très élevée, le séquençage direct de certains échantillons humains testés positifs pour des virus particuliers par biologie moléculaire n'a permis d'identifier qu'entre zéro et moins d'une dizaine de séquences de ces virus sur 25 millions que contenaient les données de séquençage [89]. Ainsi, d'autres approches, dites indirectes, ont été développées afin d'enrichir en séquences virales les données de séquençages. Ces protocoles introduisent des traitements (mécaniques, chimiques, enzymatiques, moléculaires) appliqués sur l'échantillon clinique avant ou après l'extraction des acides nucléiques (**Figure 8**).

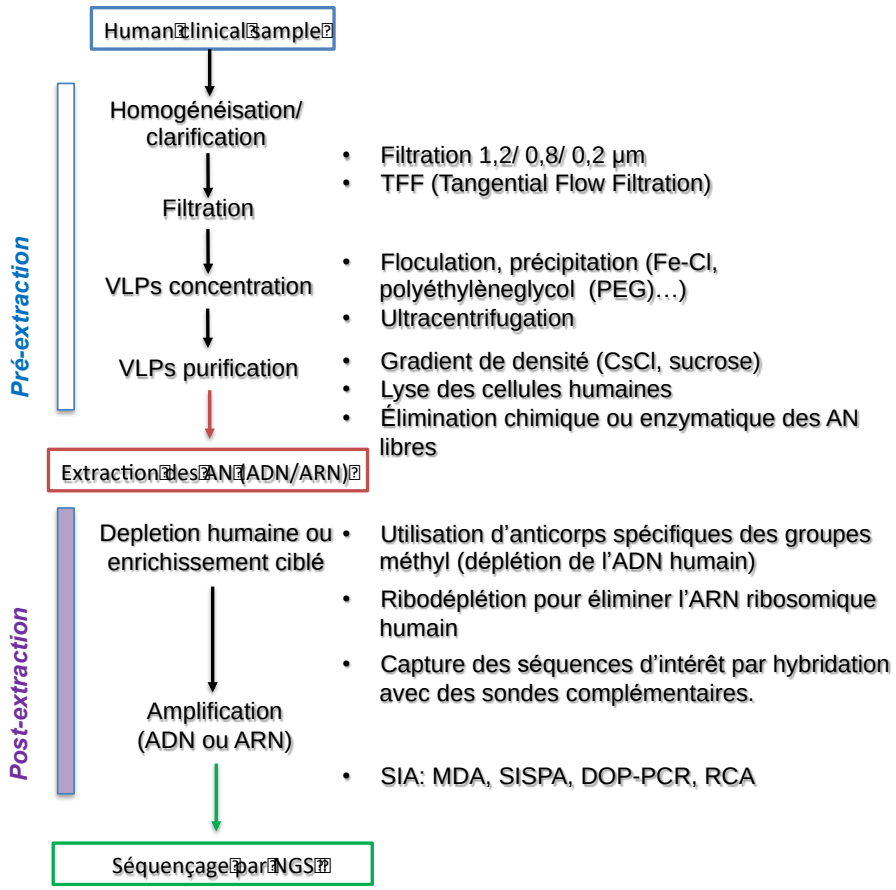


Figure 8 : Les différentes étapes pré- et post-extraction pouvant être utilisées pour la préparation des viromes dans les protocole indirects.

1. *Traitements pré-extraction*

Ainsi, avant l'extraction des acides nucléiques, des procédures d'enrichissement des particules virales circulantes peuvent être introduites **(Figure 8)**. Elles sont généralement composées d'étapes d'homogénéisation mécanique (dans le cas de tissus), de clarification par centrifugation à basse vitesse pour éliminer les cellules eucaryotes et les débris, de filtrations à différentes porosités, de concentration et purification des particules virales par précipitation, floculation, filtration tangentielle ou encore ultracentrifugation sur gradients de densité [89]. Des traitements chimiques et enzymatiques supplémentaires peuvent être introduits pour éliminer les acides nucléiques contaminants qui ne sont pas protégés par une capside [90, 91]. Cependant, ces procédures peuvent entraîner une perte significative d'informations sur les infections virales persistantes, les rétrovirus endogènes et les virus géants qui peuvent être retenus sur les filtres [92]. Bien que l'ensemble de ces étapes ait déjà permis de générer des

échantillons fortement enrichis en virus [93], une importante contamination en séquence humaine peut persister en fonction de la nature de l'échantillon clinique [94–97].

2. *Traitements post-extraction*

Ainsi, des approches complémentaires situées après l'extraction des acides nucléiques existent afin de diminuer la contamination en ADN/ARN de l'hôte. Celles-ci incluent l'utilisation d'anticorps spécifiques des groupes méthyl et de traitements nucléases pour éliminer sélectivement l'ADN de l'hôte [98] et des traitements utilisant des sondes spécifiques ou des endonucléases pour soustraire les ARN ribosomaux humains, majeurs contaminants présents dans les échantillons cliniques [99, 100]. Cependant, la digestion de l'ADN méthylé n'est pas préconisé quand on s'intéresse à l'étude des viromes et la découverte de nouveaux virus car la présence de méthylation au sein des génomes de virus varie largement en fonction des familles virales (**Tableau 2**) [101, 102].

Tableau 2 : Méthylation durant la phase active de réplication ou la phase de latence de certains familles de virus à ADN, d'après [101].

Virus	Genome size[a] (kb)	GC frequency[a]	CpG content[a] (pCPG)	Methylation status during active replication		Methylation status during latency			Host species	Effect on host methylation
				Replicating	Reference[b]	Integrated	Episomal	Reference[b]		
Large dsDNA viruses										
Adenoviridae	28–45	0.3–0.65	0.5–1.13	Un/hypomethylated	(121)	Methylated	–	(121)	Mammal, bird	DNMT upregulation
Alpha-herpesvirinae	130–150	0.4–0.71	0.9–1.17	Un/hypomethylated	(67)	–	Un/hypo-mehtylated	(89)	Mammal, bird	DNMT upregulation
Beta-herpesvirinae	140–240	0.4–0.67	1.0–1.25	Unknown	N/A	–	Unknown	N/A	Mammal	Unknown
Gamma-herpesvirinae (122)	110–185	0.3–0.61	0.3–0.66	Un/hypomethylated	(66)	–	Methylated	(65)	Mammal	DNMT upregulation
Ranid herpesvirus[c]	220–230	0.5–0.55	0.8–0.95	Methylated	(70)	–	Unknown	N/A	Amphibian	Viral 5-cytosine methyltransferases?
Iridoviridae	140–383	0.2–0.56	0.5–0.84	Methylated	(74)	–	–	–	Amphibian, fish	Viral 5-cytosine methyltransferases?
Poxviridae	130–375	0.2–0.64	0.8–1.23	Unknown	N/A	–	–	–	Mammal bird invertebrate	Unknown
Small dsDNA viruses										
Papilloma-viridae	7–8	0.4–0.54	0.1–0.57	Partially methylated	(60)	Methylated	–	(123)	Mammal	DNMT upregulation
Polyoma-viridae	5	0.4–0.48	0.05–0.78	Un/hypomethylated	(51)	Methylated	–	(51)	Mammal, bird	DNMT upregulation
Small ssDNA viruses										
Autonomous Parvoviridae	4–6	0.3–0.5	0.3–0.71	Unknown	N/A	–	–	–	Mammal	Unknown
Dependo-virinae	4–6	0.4–0.58	0.6–1.03	Unknown	N/A	–	–	–	Mammal	Unknown
Circoviridae	2	0.5–0.57	0.4–0.87	Unknown	N/A	–	–	–	Mammal, bird	Unknown
Anellovirus	4	0.5	0.67	Unknown	N/A	–	–	–	Human	Unknown

De plus, bien que les techniques de ribodéplétion se soient avérées efficaces dans l'étude des virus à ARN dans des échantillons comme le sérum et le liquide céphalo-rachidien [103, 104], cette soustraction d'ARN ribosomique n'est pas prometteuse lorsqu'on travaille sur des échantillons ou des tissus qui contiennent une forte abondance en ADN génomique et transcrits [105]. Une troisième approche post-extraction consiste en un enrichissement ciblé après hybridation et capture. Cette approche sera développée plus en détail dans la partie suivante.

Enfin, chaque traitement va diminuer de manière significative la quantité d'ADN/ARN utilisable de telle sorte que la concentration en matériel génétique obtenue à la fin ne soit plus suffisante pour procéder directement au séquençage. Ainsi pour la plupart des viromes générés par des protocoles indirects, une étape d'amplification est réalisée. Cette amplification est dite « séquence-indépendante » (SIA). Il existe différentes techniques de SIA [106] comme l'amplification par déplacement de brin (MDA) *via* la polymérase phi29 [107], l'amplification séquence indépendante utilisant une amorce unique (SISPA) ou la PCR utilisant des oligonucléotides dégénérés (DOP-PCR). Ces techniques présentent néanmoins différents biais d'amplification [108–112], résultant en une difficulté à assigner une abondance en génomes viraux dans l'échantillon originel et une amplification préférentielle de certaines familles virales [108–112].

L'utilisation de protocoles indirects permet donc de diminuer la quantité d'acides nucléiques contaminant les viromes, réduisant ainsi la profondeur nécessaire et le coût de séquençage mais entrainant un coût en matériels, réactifs et temps plus élevé que dans les protocoles directs [104, 113, 114]. Malgré tout, la contamination en séquence humaine peut s'avérer encore importante en fonction du type d'échantillon cliniques traités [94–97].

3. *La capture ciblée d'acides nucléiques viraux.*

La capture (et l'enrichissement) ciblée d'acides nucléiques d'un échantillon complexe est une approche post-extraction appliquée récemment à l'étude du virome humain et à la détection de pathogènes viraux [115–118]. Cette approche s'affranchit entièrement des étapes d'enrichissements pré-extraction et permet une hybridation des sondes directement sur des acides nucléiques extraits. Elle est donc particulièrement adaptée aux laboratoires de diagnostic en maladies infectieuses. Cette approche permet de reconstruire des génomes entiers ou quasi entiers de virus connus ou de variants de virus connus avec une couverture de séquençage très élevée facilitant des études ultérieures de phylogénie, d'évolution, de génotypage ou de virulence.

La revue ci-après propose une présentation rapide du principe de la technique d'enrichissement ciblé et de ses premières applications en génétique humaine puis s'intéresse ensuite à son application récente en recherche et diagnostic des maladies infectieuses virales, bactériennes, fongiques et parasitaires.

Article 1: Hybrid capture-based next generation sequencing and its application to human infectious diseases

Maxime Gaudin[1] and Christelle Desnues[*1]

[1]Aix-Marseille Université, IRD 198, CNRS FRE2013, Assistance-Publique des Hôpitaux de Marseille, UMR Microbes, Evolution, Phylogeny and Infections (MEPHI), IHU Méditerranée Infection, Marseille France.

*Corresponding author

Christelle Desnues, Ph.D.,

MEPHI, IHU Méditerranée Infection,

19-21 Boulevard Jean Moulin, 13005 Marseille, France

Fax: (+33) 4 13 73 24 24

Email: christelle.desnues@univ-amu.fr

- **Statut : Accepté dans Frontiers in Microbiology (Submission number 426578)**

Abstract

This review describes target-enrichment approaches followed by next generation sequencing and their recent application to the research and diagnostic field of modern and past infectious human diseases caused by viruses, bacteria, parasites and fungi.

Introduction

The development of next-generation sequencing (NGS) approaches has revolutionized human clinical research because of its ability to rapidly generate large volumes of sequencing data per run, with a concomitant decrease of sequencing costs (Shendure and Ji, 2008). Unbiased ultra-deep sequencing of complex samples is now accessible, although bioinformatics analyses may still be long and tedious. This issue is particularly problematic in the field of infectious disease diagnostic for which the rapid identification and functional characterization of a particular pathogen is critical for the clinical management of infected patients. So far, polymerase chain reaction (PCR) has been the gold standard method for the clinical diagnosis of infectious diseases (Edwards and Gibbs, 1994). This approach, which is based on the amplification of a generally short and conserved genomic region, can provide information on the presence/absence and abundance of a targeted microbial pathogen. PCR has numerous advantages, such as low cost, rapid processing and results acquisition, automation, sensitivity and specificity. However, and precisely because of its high specificity, PCR may not detect microorganisms whose sequences are too divergent from those targeted by the primers and probes designed. In addition, PCR will provide only partial information on the genetic diversity, genotype, functional potential, nutritional requirements as well as virulence or resistance to antimicrobial. Such information, that could only be retrieved from whole genome sequencing (WGS), usually requires culture of the pathogen, which can be unsuccessful in the majority of cases (and particularly for viruses and intracellular pathogens

that need host cells), can take several weeks for fastidious microorganisms or can be prevented by early administration of antimicrobial drugs. The power of NGS might thus be of particular interest in those cases for reconstructing full genomes of pathogens directly from nucleic acids extracted from clinical samples. However, due to the low pathogen/nucleic acid ratio in these complex biological samples, NGS may fail to detect/reconstruct genomes from pathogens present in low copy numbers in the sample. To overcome these limitations, capture methods, such as hybridization capture followed by NGS sequencing (also called hybrid-capture sequencing or target-enrichment sequencing) applied directly on human clinical samples have been developed (Mamanova et al., 2010). These approaches allow retrieving large genomic fragments to complete genomes with high sequencing coverage, which facilitate downstream investigations, such as phylogenetics, evolution, epidemiology and drug resistance. In this review, we will briefly describe the different principles of hybridization capture coupled with NGS, its early developments on human genetic studies and applications in the recent years to the study of present and past human infectious diseases (Gasc et al., 2016) directly from biological samples.

Overview of the experimental procedure

NGS hybridization-based capture is an approach directly applied after nucleic acid extraction and library preparation (**Figure 1**). Fragmented shotgun libraries are denatured by heating and subjected to hybridization with DNA or RNA single-stranded oligonucleotides (called also ‘probes’ or ‘baits’) specific to the region of interest (Kozarewa et al., 2015). RNA baits are preferable, because RNA:DNA duplexes are better in term of hybridization efficiency and stability, compared to DNA:DNA hybrids (Lesnik and Freier, 1995). Nonspecific unbound molecules are washed away, and the enriched DNA is eluted for NGS (Kozarewa et al., 2015). The hybridization between DNA libraries and baits can be carried out in solution or on a solid

support. In “solid-phase”, DNA probes are bound to a solid support, such as a glass microarray slide (Albert et al., 2007; Okou et al., 2007), on which in “solution-capture”, free DNA or RNA probes are biotinylated, allowing them to isolate the targeted fragment-probe heteroduplexes using magnetic streptavidin beads (Gnirke et al., 2009).

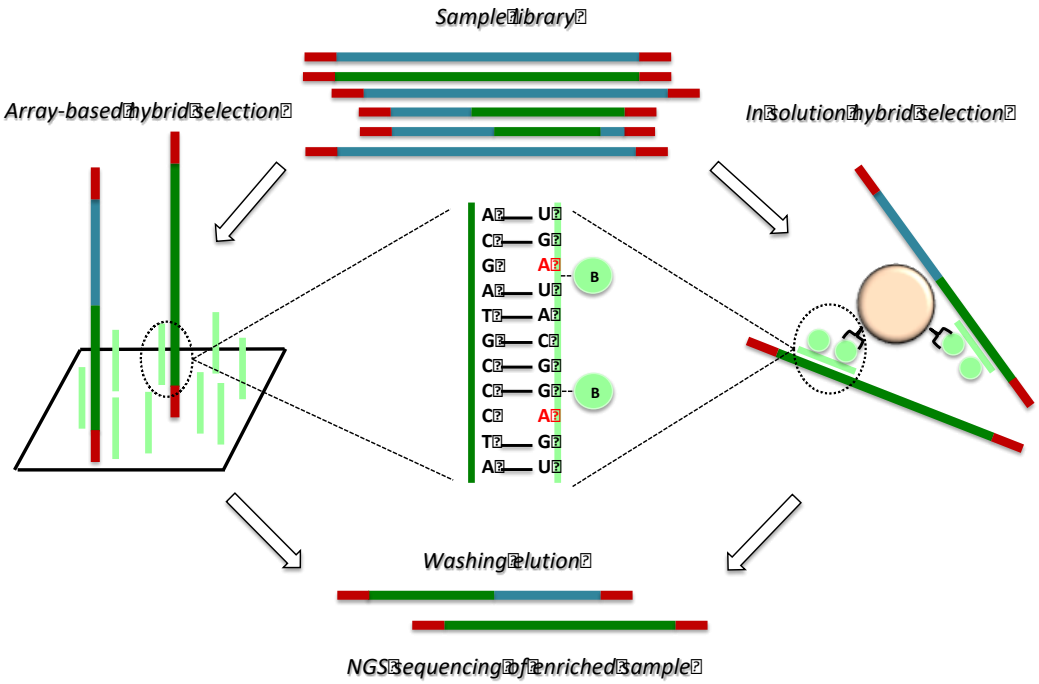


Figure 1: Principle of target-enrichment sequencing

Fragmented shotgun library with adapters (red) is hybridized against a set of specific probes. Hybridization can be performed either in solution (in solution hybrid selection) with biotinylated probes captured by streptavidin-coated magnetic beads, or on a solid support (array-based hybrid selection) on which probes are spotted. After hybridization, the nonspecific unbound molecules are washed out and the enriched DNA is eluted for NGS.

Early developments of hybrid-capture strategies: human genetic studies on modern and ancient samples

Target-enrichment strategy using hybrid capture was originally developed for human genomic studies for which it was used to capture and sequence the entire human exome. This genomic technique, also called exome sequencing (or whole exome sequencing) was first applied by using an array-based hybrid capture method in 2007 (Hodges et al., 2007). In this study, the authors developed six customized Nimblegen arrays to capture about 180,000 coding exons with overlapping 60-90-nt probes allowing an average enrichment of exon DNA sequences of 323 folds. Whole exome sequencing using capture arrays has proven its usefulness in identifying rare variants and mutations causing disease (Choi et al., 2009; Ng et al., 2009). The limitations of this technique include the need to design an array and a relatively large amount of DNA. To overcome some of the weaknesses of the previous method, Gnirke et al. have developed an in-solution hybrid capture method for human whole exome sequencing (Gnirke et al., 2009). To do so, biotinylated RNA baits of 170 bases in length were constructed, targeting 5,565 human protein-coding exons. In this study, authors have demonstrated the possibility to perform hybrid selection in solution. Following this, many targeted human exome in-solution enrichment methods for NGS have been developed, including those commercialized by Illumina (www.illumina.com) and Agilent Genomics (www.genomics.agilent.com) (Chen et al., 2015b, 2015a). In-solution capture for exome sequencing turned out to be an effective approach applied to discover the causal mutation of rare Mendelian disorders (Martignetti et al., 2013; Nectoux et al., 2015; Rousseau-Nepton et al., 2015; Shearer et al., 2010), of complex disorders (Griesi-Oliveira et al., 2015; Guipponi et al., 2014; Pérez-Serra et al., 2015; Poultney et al., 2013), mitochondrial disorders (Calvo et al., 2012; Gai et al., 2013) and more recently to the screening of potential genetic mutations of patients suffering from cancer (Clark et al., 2018b; Drilon et al., 2015; Rozenblum et al., 2017; Schrock et al., 2018; Sikkema-Raddatz et al., 2013; Xie et al., 2016; Xu et al., 2017).

The power of hybridization capture has been also successfully used to study human ancient DNA (aDNA) preserved in ancient human remains. Indeed, in ancient human samples, DNA is highly fragmented and dominated by a large contamination of environmental and bacterial DNA, which poses a limitation in shotgun aDNA sequencing experiment (Knapp and Hofreiter, 2010). Due to its higher copy number in the cell than nuclear DNA, mitochondrial DNA (mtDNA) was the first genetic marker to be analyzed in human paleogenetic studies. Probe hybridization assays use biotinylated DNA or RNA probes targeting the two hypervariable segments of the mtDNA control region (CR) (Briggs et al., 2009; Eduardoff et al., 2017; Enk et al., 2013; Kihana et al., 2013; Krause et al., 2010; Loreille et al., 2018; Maricic et al., 2010; Templeton et al., 2013). Another uniparental marker, the Y-chromosome DNA (Y-DNA), was also used to study aDNA. As each cell possesses only one copy of the Y chromosome, the hybridization capture was carried out to enrich specific genomic regions of the Y chromosome both on solid support (Fu et al., 2013) or in solution (Cruz Dávalos et al., 2017). However, targeting mitochondrial DNA or Y chromosome involves discarding a large proportion of potentially informative sequences present in autosomal DNA. For this reason, in 2013, Carpenter and colleagues reported a new capture-based method, called whole-genome in-solution capture (WISC), using modern DNA as bait covering the entire human genome. This method was applied to 12 ancient human DNA libraries and showed an enrichment of 6 to 159 folds of the sequence mapping to the human genome with enrichments of 2 to 13 folds for unique fragments (Carpenter et al., 2013). As for modern human genetic studies, commercial kits targeting mitochondrial DNA, custom loci, or entire nuclear genomes, such as those developed by Arbor Bioscience (myBaits® - http://www.arborbiosci.com) are now employed in the genetic sequencing of ancient DNA (Enk et al., 2014; Lindo et al., 2017).

Applications of target-enrichment sequencing to human infectious diseases (Table 1)

Parasites and fungi

The first application of hybrid selection method for infectious diseases was in the field of human parasitology research (Melnikov et al., 2011). To overcome the low proportion of *Plasmodium falciparum* sequences relative to that of their human host, authors have proposed to adapt in-solution NGS hybrid capture method to enrich this pathogen. This protocol has been tested in both mock mixtures composed of 99% human DNA and 1% Plasmodium but also in *P. falciparum* clinical samples. For this purpose, synthetic 140 bp oligos labeled with biotin were designed to capture exonic regions of the *P. falciparum* genome, whereas 250 bp oligos were constructed to target the entire genome. Processed and unprocessed samples were then sequenced with an Illumina technology. In the mock metagenome, sequencing of the hybrid-selected samples yielded between 37 to 44-fold enrichment of the parasite DNA. In the human clinical samples, Illumina sequencing showed that at least 5.9% of reads mapped to *Plasmodium*, but no data was provided regarding the percentage of *Plasmodium* reads obtained without hybrid capture (Melnikov et al., 2011). However, this first study highlighted the good performance of NGS hybrid capture to sequence parasite genome from human clinical samples. In 2012, other studies confirmed the good performance of in-solution hybrid capture to enrich *P. falciparum* (Smith et al., 2012) and *P. vivax* (Bright et al., 2012) sequences.

Fungi are also a major cause of human diseases that can be particularly serious in immunocompromized patients or in patients hospitalized for serious diseases (Pfaller and Diekema, 2010). For example, systemic infections with *Candida albicans* in immunocompromized patients result in mortality rates of about 50% (Pfaller and Diekema, 2010). The prevention, diagnosis and therapy of fungi infections remain very difficult and comprehension of transcriptional regulation between fungal pathogens and host is an important

step to identify potential novel targets for drug development (Pfaller and Diekema, 2010). Again, the limitations of host and pathogen transcriptome analysis lie in the low proportion of fungal RNA present in the total extracted RNA. The use of specific enrichment procedures before RNASeq analysis has then been proposed as an alternative method to overcome the problem of low fungus/host RNA ratio. For this purpose, Amorim-Vaz et al. have designed a set of 55,342 biotinylated 120 bp-RNA probes covering 6,094 *C. albicans* ORFs (Amorim-Vaz et al., 2015). cDNA libraries were established using SureSelect (Agilent) after extraction of RNA from mice kidney or *Galleria mellonella* larvae infected with *C. albicans*, and were subjected to capture with biotinylated probes before Illumina HiSeq sequencing. Results showed up to a 1670-fold enrichment of *C. albicans* reads in a given biological and a detection of more than 86% of its genes. Many genes that have been regulated in in vivo infection experiments have functions that have not yet been characterized and will require further research to understand their role during infection (Amorim-Vaz et al., 2015).

Bacteria

In bacteriological research and diagnostic, targeted capture strategies prior to sequencing could be a powerful tool in the management and therapeutics of patients with infectious disease. Indeed, the rapid identification of antimicrobial resistance is essential for a rapid and effective treatment. Regarding *Mycobacterium tuberculosis*, current methods of screening for antimicrobial resistance, which are based on the culture of the organism from sputum samples before sequencing, can take up to several weeks. To overcome these limitations, Brown et al. have proposed to use oligonucleotide enrichment technology to capture *M. tuberculosis* genome sequences directly from positive smear sputum samples (Brown et al., 2015). Whole genome baits (120-mer RNA baits) were designed to span the entire positive strand of the *H37Rv M. tuberculosis* reference genome and synthesized by Agilent Technologies. The authors demonstrated the reliability of targeted sequencing to

recover and sequence, in less than 96 hours, nearly complete genomes directly from 81% (21/26) smear positive sputa but also its robustness to identify the genotype and resistance determinants of all samples that were previously tested positive. This study emphasizes the use of hybrid selection target enrichment that could allow personalized antimicrobial treatment in multidrug-resistant tuberculosis (Brown et al., 2015). Other studies have used biotinylated baits spanning entire genomes for high-resolution strain genotyping directly from clinical samples. Indeed, discrimination of *Chlamydia trachomatis* serovars from genital samples would facilitate the study of population structures and modes of transmission (Christiansen et al., 2014) while genomic data from uncultured *Neisseria meningitidis* not grown in the case of invasive meningococcal would allow increased surveillance of vaccine antigens and studies on possible vaccine deficiencies (Clark et al., 2018a).

Viruses

In viral research and diagnostic laboratories, viral WGS is also essential for the detection of drug resistance and the development of novel treatments and vaccines (Houldcroft et al., 2017). In this domain, the first study that demonstrated the effectiveness of target capture technology for reconstructing full herpesvirus genomes from complex biological samples was proposed by Depledge et al. in 2011 (Depledge et al., 2011). In this study, 120-mer RNA baits generating a 2x coverage for Varicella-Zoster Virus (VZV), a 5x coverage for Epstein-Barr virus (EBV) and Kaposi's sarcoma-associated Herpesvirus (KSHV), were synthetized and hybridized with DNA extracted from a range of clinical samples including blood, saliva, vesicle fluid, cerebrospinal fluid and tumor cell lines. Full-length herpes virus genomes were reconstructed at high read depth for the 13 samples tested and generated further studies on the structure and diversity of the viral population (Depledge et al., 2011). Following this study, the capture of whole genomic hybrids made it possible to study the genomic diversity of 8 new complete EBV genomes isolated from biopsy specimens of primary nasopharyngeal

carcinomas (Kwok et al., 2014), 37 Zika virus genomes (ZIKV) samples out of 66 attempts (Metsky et al., 2017), 453 complete genomes (with >90% genome coverage and >100-fold read depth) of different norovirus genotypes from 509 stool samples (Brown et al., 2016) and to achieve sufficient coverage for de novo genome assembly and detection of single nucleotide variants of Lassa virus (LASV) from ultra-low input samples (Matranga et al., 2014). This approach has been also used to characterize other clinically relevant viruses, such as hepatitis C virus (HCV) (Thomson et al., 2016), varicella zoster virus (Depledge et al., 2014), human herpesvirus 7 (HHV-7) (Donaldson et al., 2013, 7) and the herpes simplex virus 1 and 2 (HSV-1 and HSV-2) (Greninger et al., 2018). Hybrid capture associated with shotgun sequencing could also be performed using a combination of several viral species used as baits. Indeed, in 2015, Wylie et al. developed ViroCap, a panel of probes designed to enrich nucleic acid from 34 families of DNA and RNA viruses (190 viral genera and 337 species) that infect vertebrate hosts, except human endogenous retrovirus (Wylie et al., 2015). These probes were tested both on a pool of 14 clinical samples, which tested positive for a viral infection, and on eight samples from young children with fever, also positive for one or more viruses. Libraries were sequenced before capture (pre-capture) and following capture using ViroCap (post-capture). Combining results from both experiments, 32 viruses were detected (11 additional in the post-captured samples), including diverse DNA and RNA viruses (with genomes ranging from 5-161 kb) with genomic coverage >80% for 16 of the 32 genomes. Several complete genomes were reconstructed and belonged to human bocavirus 1, human parvovirus B19, human adenovirus B (type 3), human adenovirus C (type 1), KI polyomavirus, sapovirus and human astrovirus 1. Finally, although ViroCap cannot detect viral sequences that are completely novel, its design, which includes neighbor genomes of reference sequences, allows variants with nucleotide sequence identity as low as 58% to be identified (Wylie et al., 2015). In 2018, a similar approach called ViroFind was designed to target 535 DNA and RNA viruses, which are known

to infect humans or cause zoonoses. This in-solution target enrichment was applied to the brain biopsy samples of 5 patients with progressive multifocal leukoencephalopathy (PML) (Chalkias et al., 2018). It allowed the description of highly complex polyoma virus JC populations as well as the detection of large genetic divergence among variants, with some of these mutations conferring viral fitness advantages (Chalkias et al., 2018). Lastly, other applications of target-enrichment sequencing have been described, such as the study of viral genome integrations within the human genome. This approach was powerful and efficient to identify Merkel Cell Polyomavirus (MCPyV) insertion sites on DNA extracted from formalin-fixed and paraffin-embedded tissue from Merkel cell carcinoma (Duncavage et al., 2011). It also allowed to analyze retroviral genomes integrated within host genomic DNA in case of human T-cell leukemia virus type-1 (HTLV-1) and human immunodeficiency virus type-1 (HIV-1) infections (Miyazato et al., 2016).

Paleomicrobiology

Paleomicrobiology is an emerging research field dedicated to the detection, identification and characterization of microorganisms (bacteria, viruses and parasites) in ancient specimens. Elucidating past infectious diseases can lead to a better understanding of the temporal and geographical distributions of infected individuals, the introduction of microorganisms into human populations, the host-pathogen relationships but also the genetic evolution of the microorganisms (Drancourt and Raoult, 2005). The main limitations of palaeomicrobiological studies concern the degradation of ancient DNA (aDNA) and the risk of contamination by modern DNA (Riviera-Perez et al., 2016). Target-enrichment prior to sequencing is therefore a particularly relevant tool in this context for genomes study. The first two studies using targeted enrichment in paleomicrobiology have investigated genetic changes and virulence factor of *Yersinia pestis*, the causal agent of the second plague pandemics (Black Death, 14-17th centuries) (Bos et al., 2011; Schuenemann et al., 2011). To this end, an array-

based enrichment using probe targeting either the full *Yersinia pestis* chromosome or *pestis*-specific virulence plasmids was applied directly after the DNA extraction from ancient bones (Schuenemann et al., 2011, 1) and/or teeth (Bos et al., 2011; Schuenemann et al., 2011). Using targeted DNA capture approach combined with high-throughput sequencing, the authors reconstructed 99% of the pPCP1 plasmid sequence (Schuenemann et al., 2011) and a draft genome of *Y. pestis* (Bos et al., 2011) with the molecular damages typically associated with aDNA. Comparisons with modern genomes did not identify any significant genetic variation that could explain the differences between the ancient and modern forms of the disease (Schuenemann et al., 2011). More recently, 3 other draft genomes of *Y. pestis* have been recovered from individuals who died during the first plague pandemics (the Plague of Justinian, 6-8th centuries) in 2 different rural sites in southern Germany (Feldman et al., 2016; Wagner et al., 2014). Genetic characterization showed that these 3 drafts derived from a single Justinianic strain which is unique and harbors novel substitutions and structural polymorphism (Feldman et al., 2016; Wagner et al., 2014). Finally, target enrichment sequencing also allowed the reconstruction of new *Y. pestis* strains from Bronze Age individuals (~3800BP) (Spyrou et al., 2018) providing further datas into the early stages of *Y. pestis* genome evolution including on genomic characteristics supporting flea-borne transmission in rodents or humans (Spyrou et al., 2018). Finally, target-enrichment sequencing approaches in the paleomicrobiological research field have not been exclusively applied to the study of ancient plague pandemics, but have also allowed genomic investigation of ancient *Mycobacterium tuberculosis* (Bos et al., 2014), *Mycobacterium leprae* (Schuenemann et al., 2013), *Variola virus* (Duggan et al., 2016), *Plasmodium falciparum* (Marciniak et al., 2016) and *Treponema pallidum* (Schuenemann et al., 2018) in human remains.

Table 1. Example of studies that used target-enrichment sequencing for parasitic, fungal, bacterial or viral diseases in modern and ancient samples.

Targeted organism	Probe design	Infectious Disease	Sample tested	Methods	Nucleic acids	Sequencing	Date	Reference
Parasitic and fungal disease								
Plasmodium falciparum 3D7	Whole genome	Malaria	Mock sample Whole blood	In-solution	DNA	Illumina GAIIx Illumina HiSeq	2011	*(Melnikov et al., 2011)*
Plasmodium falciparum 3D7	Whole genome	Malaria	Mock sample Whole blood	In-solution	DNA	Illumina GAIIx	2012	*(Smith et al., 2012)*
Plasmodium vivax	Whole genome	Malaria	Whole blood	In-solution	DNA	Illumina Hi-Seq2000	2012	*(Bright et al., 2012)*
Candida albicans	6094 ORFS	Systemic candidiasis	Mouse kidneys	In-solution	RNA (cDNA)	Illumina HiSeq	2015	*(Amorim-Vaz et al., 2015)*
Bacterial disease								
Chlamydia trachomatis	74 references whole genome	Trachoma	Vaginal swabs Urine	In-solution	DNA	Illumina MiSeq	2014	*(Christiansen et al., 2014)*
M. tuberculosis H37Rv	Whole genome	Multidrug-resistant tuberculosis	Sputum	In-solution	DNA	Illumina MiSeq	2015	*(Brown et al., 2015)*
Neisseria meningitidis	77 complete reference genome and 2898 drafts	Meningococcal	CSF Whole blood	In-solution	DNA	Illumina MiSeq	2018	*(Clark et al., 2018a)*
Viral disease								
Varicella-zoster virus Epstein–Barr virus Kaposi's Sarcoma-Associated Herpesvirus	Whole genome	Zoster-vaccine rash, wild-type zoster, encephalitis	Clinical and cultured samples	In-solution	DNA	Illumina GAIIx	2011	*(Depledge et al., 2011)*
Merkel cell polyomavirus (MCPyV)	23 overlapping PCR products that tiled across the MCPyV genome	Merkel cell carcinoma	Formalin-Fixed, Paraffin-Embedded Tissue	In-solution	DNA	Illumina GAIIx	2011	*(Duncavage et al., 2011)*
Human Herpesvirus 7 (HHV7)	Whole genome	x	x	In-solution	DNA	Illumina MiSeq	2013	*(Donaldson et al., 2013)*
Epstein–Barr Virus (EBV)	Whole genome	Nasopharyngeal carcinoma (NPC)	NPC tumor biopsy	In-solution	DNA	Illumina MiSeq	2014	*(Kwok et al., 2014)*
Zika virus (ZIKV)	Whole genome	ZIKV infection	Blood, urine, cerebrospinal fluid, and saliva	In-solution	RNA (cDNA)	Illumina HiSeq	2014	*(Metsky et al., 2017)*
Lassa virus	2 references complete genome	Hemorrhagic fevers	Plasma or serum	In-solution	RNA (cDNA)	Illumina HiSeq	2014	*(Matranga et al., 2014)*
Varicella Zoster Virus (VZV)	Whole genome	Vaccine-associated rashes	Vesicular fluid	In-solution	DNA	Illumina (GAIIx, HiSeq, and MiSeq)	2014	*(Depledge et al., 2014)*
Virocap: *337 DNA and RNA viral species that are known to infect human and animals*	1456 RefSeq + genome neighbor sequences (185,835 fasta total)	?	Panel of human samples	?	DNA and RNA (cDNA)	Illumina HiSeq	2015	*(Wylie et al., 2015)*
Norovirus	662 references complete or partial genome	Acute gastroenteritis	Stool suspensions	In-solution	RNA (cDNA)	Illumina MiSeq	2016	*(Brown et al., 2016)*

Hepatitis C Virus (HCV)	953 references complete genome	x	Plasma Mock samples RNA transcripts	In-solution	RNA (cDNA)	Illumina MiSeq	2016	*(Thomson et al., 2016)*
Exogenous retroviruses: Immunodeficiency virus type-1 (HIV-1) and the human T-cell leukemia virus type-1 (HTLV-1)	Whole genome	x	Infected cells	In-solution	DNA	Illumina MiSeq	2016	*(Miyazato et al., 2016)*
Human Herpesvirus 1 and 2 (HHV1 / HHV2)	Whole genome	Swab on the lesion	Skin rash: herpetic lesions	In-solution	DNA	Illumina MiSeq	2018	*(Greninger et al., 2018)*
Virofind *535 DNA and RNA viral species that are known to infect humans*	Whole genome	Progressive multifocal leukoencephalopathy (PML)	¨Post mortem PML brain samples	In-solution	DNA and RNA (cDNA)	Illumina MiSeq	2018	*(Chalkias et al., 2018)*
Infectious diseases in paleomicrobiology								
Yersinia pestis strain CO92	Whole genome or pCD1 and pMT1 plasmids	Plague (Black Death) 1347–1351	Teeth	Microarray	DNA	Illumina GAIIx	2011	*(Bos et al., 2011)*
Yersinia pestis strain CO92	Portion of the Y. pestis pPCP1 plasmid	Plague (Black Death) 1347–1351	Bones and Teeth	In-solution	DNA	Illumina GAIIx	2011	*(Schuenemann et al., 2011, 1)*
Yersinia pestis strain CO92	Core genome, plasmids and 155 other genes	Plague (Justinian) (6–8th centuries)	Teeth	In-solution	DNA	Illumina HiSeq	2014	*(Wagner et al., 2014)*
Mycobacterium tuberculosis	rpoB, gyrA, gyrB, katG, and mpt40 genes	Turberculosis (1000 year old)	Skeletal samples	In-solution	DNA	Illumina HiSeq	2014	*(Bos et al., 2014)*
Mycobacterium leprae	3 genomic loci	Leprosy (medieval period)	Bones and teeth	In-solution	DNA	x	2013	*(Schuenemann et al., 2013)*
Variola virus	Whole genome	Smallpox (1643 and 1665)	Bones of a child mummy	In-solution	DNA	x	2016	*(Duggan et al., 2016)*
Plasmodium falciparum	mitochondrial genomes of Plasmodium spp	Malaria (1st–5th century)	Teeth	In-solution	DNA	x	2016	*(Marciniak et al., 2016)*
Treponema pallidum	Whole genome	Syphilis (1681 to 1861)	Bones	Microarray	DNA	Illumina Hiseq	2018	*(Schuenemann et al., 2018)*

Conclusion

Target-enrichment sequencing is an efficient approach that allows large fragments and even entire sequences of the genome of targeted microorganisms to be reconstructed directly from complex biological samples containing a low pathogen/host nucleic acid ratio. The information provided by the genome can be used to explore the genetic diversity, epidemiology, evolution, transmission networks or antimicrobial resistance of the target pathogen or its variants. The main current limitations of democratizing target-enrichment sequencing in clinical diagnostic laboratories are its high cost, the expertise required for library preparation and the time required to generate biotinylated probes from reference genomes, which hampers a rapid response to an emerging pathogen. Above all, it is not suitable for the detection and characterization of completely novel microorganisms, including viruses whose emergence may represent one of the main threats to human health in the near future.

References

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., et al. (2007). Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4, 903–905. doi:10.1038/nmeth1111.

Amorim-Vaz, S., Tran, V. D. T., Pradervand, S., Pagni, M., Coste, A. T., and Sanglard, D. (2015). RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens In Vivo Applied to the Fungus Candida albicans. mBio 6, e00942-15. doi:10.1128/mBio.00942-15.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514, 494–497. doi:10.1038/nature13591.

Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., et al. (2011). A draft genome of Yersinia pestis from victims of the Black Death. Nature 478, 506–510. doi:10.1038/nature10549.

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., et al. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325, 318–321. doi:10.1126/science.1174462.

Bright, A. T., Tewhey, R., Abeles, S., Chuquiyauri, R., Llanos-Cuentas, A., Ferreira, M. U., et al. (2012). Whole genome sequencing analysis of Plasmodium vivax using whole genome capture. BMC Genomics 13, 262. doi:10.1186/1471-2164-13-262.

Brown, A. C., Bryant, J. M., Einer-Jensen, K., Holdstock, J., Houniet, D. T., Chan, J. Z. M., et al. (2015). Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J. Clin. Microbiol. 53, 2230–2237. doi:10.1128/JCM.00486-15.

Brown, J. R., Roy, S., Ruis, C., Yara Romero, E., Shah, D., Williams, R., et al. (2016). Norovirus Whole-Genome Sequencing by SureSelect Target Enrichment: a Robust and Sensitive Method. J. Clin. Microbiol. 54, 2530–2537. doi:10.1128/JCM.01052-16.

Calvo, S. E., Compton, A. G., Hershman, S. G., Lim, S. C., Lieber, D. S., Tucker, E. J., et al. (2012). Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. Sci. Transl. Med. 4, 118ra10. doi:10.1126/scitranslmed.3003310.

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., et al. (2013). Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am. J. Hum. Genet. 93, 852–864. doi:10.1016/j.ajhg.2013.10.002.

Chalkias, S., Gorham, J. M., Mazaika, E., Parfenov, M., Dang, X., DePalma, S., et al. (2018). ViroFind: A novel target-enrichment deep-sequencing platform reveals a complex JC virus population in the brain of PML patients. PLoS ONE 13. doi:10.1371/journal.pone.0186945.

Chen, R., Im, H., and Snyder, M. (2015a). Whole-Exome Enrichment with the Agilent SureSelect Human All Exon Platform. Cold Spring Harb. Protoc. 2015, pdb.prot083659. doi:10.1101/pdb.prot083659.

Chen, R., Im, H., and Snyder, M. (2015b). Whole-Exome Enrichment with the Illumina TruSeq Exome Enrichment Platform. Cold Spring Harb. Protoc. 2015, pdb.prot084863. doi:10.1101/pdb.prot084863.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. U. S. A. 106, 19096–19101. doi:10.1073/pnas.0910672106.

Christiansen, M. T., Brown, A. C., Kundu, S., Tutill, H. J., Williams, R., Brown, J. R., et al. (2014). Whole-genome enrichment and sequencing of Chlamydia trachomatis directly from clinical samples. BMC Infect. Dis. 14, 591. doi:10.1186/s12879-014-0591-3.

Clark, S. A., Doyle, R., Lucidarme, J., Borrow, R., and Breuer, J. (2018a). Targeted DNA enrichment and whole genome sequencing of Neisseria meningitidis directly from clinical specimens. Int. J. Med. Microbiol. IJMM 308, 256–262. doi:10.1016/j.ijmm.2017.11.004.

Clark, T. A., Chung, J. H., Kennedy, M., Hughes, J. D., Chennagiri, N., Lieber, D. S., et al. (2018b). Analytical validation of a hybrid capture-based next-generation sequencing clinical assay for genomic profiling of cell-free circulating tumor DNA. J. Mol. Diagn. JMD. doi:10.1016/j.jmoldx.2018.05.004.

Cruz Dávalos, D. I., Nieves-Colón, M. A., Sockell, A., Poznik, D. G., Schroeder, H., Stone, A. C., et al. (2017). In-solution Y-chromosome capture-enrichment on ancient DNA libraries. doi:10.1101/223214.

Depledge, D. P., Kundu, S., Jensen, N. J., Gray, E. R., Jones, M., Steinberg, S., et al. (2014). Deep Sequencing of Viral Genomes Provides Insight into the Evolution and Pathogenesis of Varicella Zoster Virus and Its Vaccine in Humans. Mol. Biol. Evol. 31, 397–409. doi:10.1093/molbev/mst210.

Depledge, D. P., Palser, A. L., Watson, S. J., Lai, I. Y.-C., Gray, E. R., Grant, P., et al. (2011). Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLOS ONE 6, e27805. doi:10.1371/journal.pone.0027805.

Donaldson, C. D., Clark, D. A., Kidd, I. M., Breuer, J., and Depledge, D. D. (2013). Genome Sequence of Human Herpesvirus 7 Strain UCL-1. Genome Announc. 1. doi:10.1128/genomeA.00830-13.

Drancourt, M., and Raoult, D. (2005). Palaeomicrobiology: current issues and perspectives. Nat. Rev. Microbiol. 3, 23–35. doi:10.1038/nrmicro1063.

Drilon, A., Wang, L., Arcila, M. E., Balasubramanian, S., Greenbowe, J. R., Ross, J. S., et al. (2015). Broad, hybrid capture-based next-generation sequencing identifies actionable genomic alterations in "driver-negative" lung adenocarcinomas. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 21, 3631–3639. doi:10.1158/1078-0432.CCR-14-2683.

Duggan, A. T., Perdomo, M. F., Piombino-Mascali, D., Marciniak, S., Poinar, D., Emery, M. V., et al. (2016). 17th Century Variola Virus Reveals the Recent History of Smallpox. Curr. Biol. CB 26, 3407–3412. doi:10.1016/j.cub.2016.10.061.

Duncavage, E. J., Magrini, V., Becker, N., Armstrong, J. R., Demeter, R. T., Wylie, T., et al. (2011). Hybrid Capture and Next-Generation Sequencing Identify Viral Integration Sites from Formalin-Fixed, Paraffin-Embedded Tissue. J. Mol. Diagn. JMD 13, 325–333. doi:10.1016/j.jmoldx.2011.01.006.

Eduardoff, M., Xavier, C., Strobl, C., Casas-Vargas, A., and Parson, W. (2017). Optimized mtDNA Control Region Primer Extension Capture Analysis for Forensically Relevant Samples and Highly Compromised mtDNA of Different Age and Origin. Genes 8. doi:10.3390/genes8100237.

Edwards, M. C., and Gibbs, R. A. (1994). Multiplex PCR: advantages, development, and applications. PCR Methods Appl. 3, S65-75.

Enk, J. M., Devault, A. M., Kuch, M., Murgha, Y. E., Rouillard, J.-M., and Poinar, H. N. (2014). Ancient whole genome enrichment using baits built from modern DNA. Mol. Biol. Evol. 31, 1292–1294. doi:10.1093/molbev/msu074.

Enk, J., Rouillard, J.-M., and Poinar, H. (2013). Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. BioTechniques 55, 300–309. doi:10.2144/000114114.

Feldman, M., Harbeck, M., Keller, M., Spyrou, M. A., Rott, A., Trautmann, B., et al. (2016). A High-Coverage Yersinia pestis Genome from a Sixth-Century Justinianic Plague Victim. Mol. Biol. Evol. 33, 2911–2923. doi:10.1093/molbev/msw170.

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. A., Kelso, J., et al. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. Proc. Natl. Acad. Sci. U. S. A. 110, 2223–2227. doi:10.1073/pnas.1221359110.

Gai, X., Ghezzi, D., Johnson, M. A., Biagosch, C. A., Shamseldin, H. E., Haack, T. B., et al. (2013). Mutations in FBXL4, encoding a mitochondrial protein, cause early-onset mitochondrial encephalomyopathy. Am. J. Hum. Genet. 93, 482–495. doi:10.1016/j.ajhg.2013.07.016.

Gasc, C., Peyretaillade, E., and Peyret, P. (2016). Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. Nucleic Acids Res. 44, 4504–4518. doi:10.1093/nar/gkw309.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189. doi:10.1038/nbt.1523.

Greninger, A. L., Roychoudhury, P., Xie, H., Casto, A., Cent, A., Pepper, G., et al. (2018). Ultrasensitive Capture of Human Herpes Simplex Virus Genomes Directly from Clinical Samples Reveals Extraordinarily Limited Evolution in Cell Culture. mSphere 3. doi:10.1128/mSphereDirect.00283-18.

Griesi-Oliveira, K., Acab, A., Gupta, A. R., Sunaga, D. Y., Chailangkarn, T., Nicol, X., et al. (2015). Modeling non-syndromic autism and the impact of TRPC6 disruption in human neurons. Mol. Psychiatry 20, 1350–1365. doi:10.1038/mp.2014.141.

Guipponi, M., Santoni, F. A., Setola, V., Gehrig, C., Rotharmel, M., Cuenca, M., et al. (2014). Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. PloS One 9, e112745. doi:10.1371/journal.pone.0112745.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., et al. (2007). Genome-wide in situ exon capture for selective resequencing. Nat. Genet. 39, 1522–1527. doi:10.1038/ng.2007.42.

Houldcroft, C. J., Beale, M. A., and Breuer, J. (2017). Clinical and biological insights from viral genome sequencing. Nat. Rev. Microbiol. 15, 183–192. doi:10.1038/nrmicro.2016.182.

Kihana, M., Mizuno, F., Sawafuji, R., Wang, L., and Ueda, S. (2013). Emulsion PCR-coupled target enrichment: an effective fishing method for high-throughput sequencing of poorly preserved ancient DNA. Gene 528, 347–351. doi:10.1016/j.gene.2013.07.040.

Knapp, M., and Hofreiter, M. (2010). Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. Genes 1, 227–243. doi:10.3390/genes1020227.

Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E., and Hendrickson, C. L. (2015). Overview of Target Enrichment Strategies. Curr. Protoc. Mol. Biol. 112, 7.21.1-23. doi:10.1002/0471142727.mb0721s112.

Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., et al. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464, 894–897. doi:10.1038/nature08976.

Kwok, H., Wu, C. W., Palser, A. L., Kellam, P., Sham, P. C., Kwong, D. L. W., et al. (2014). Genomic Diversity of Epstein-Barr Virus Genomes Isolated from Primary Nasopharyngeal Carcinoma Biopsy Samples. J. Virol. 88, 10662–10672. doi:10.1128/JVI.01665-14.

Lesnik, E. A., and Freier, S. M. (1995). Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. Biochemistry 34, 10807–10815.

Lindo, J., Achilli, A., Perego, U. A., Archer, D., Valdiosera, C., Petzelt, B., et al. (2017). Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. Proc. Natl. Acad. Sci. U. S. A. 114, 4093–4098. doi:10.1073/pnas.1620410114.

Loreille, O., Ratnayake, S., Bazinet, A. L., Stockwell, T. B., Sommer, D. D., Rohland, N., et al. (2018). Biological Sexing of a 4000-Year-Old Egyptian Mummy Head to Assess the Potential of Nuclear DNA Recovery from the Most Damaged and Limited Forensic Specimens. Genes 9. doi:10.3390/genes9030135.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., et al. (2010). Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118. doi:10.1038/nmeth.1419.

Marciniak, S., Prowse, T. L., Herring, D. A., Klunk, J., Kuch, M., Duggan, A. T., et al. (2016). Plasmodium falciparum malaria in 1st-2nd century CE southern Italy. Curr. Biol. CB 26, R1220–R1222. doi:10.1016/j.cub.2016.10.016.

Maricic, T., Whitten, M., and Pääbo, S. (2010). Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLOS ONE 5, e14004. doi:10.1371/journal.pone.0014004.

Martignetti, J. A., Tian, L., Li, D., Ramirez, M. C. M., Camacho-Vanegas, O., Camacho, S. C., et al. (2013). Mutations in PDGFRB Cause Autosomal-Dominant Infantile Myofibromatosis. Am. J. Hum. Genet. 92, 1001–1007. doi:10.1016/j.ajhg.2013.04.024.

Matranga, C. B., Andersen, K. G., Winnicki, S., Busby, M., Gladden, A. D., Tewhey, R., et al. (2014). Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol. 15, 519. doi:10.1186/PREACCEPT-1698056557139770.

Melnikov, A., Galinsky, K., Rogov, P., Fennell, T., Van Tyne, D., Russ, C., et al. (2011). Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biol. 12, R73. doi:10.1186/gb-2011-12-8-r73.

Metsky, H. C., Matranga, C. B., Wohl, S., Schaffner, S. F., Freije, C. A., Winnicki, S. M., et al. (2017). Zika virus evolution and spread in the Americas. Nature 546, 411–415. doi:10.1038/nature22402.

Miyazato, P., Katsuya, H., Fukuda, A., Uchiyama, Y., Matsuo, M., Tokunaga, M., et al. (2016). Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. Sci. Rep. 6. doi:10.1038/srep28324.

Nectoux, J., de Cid, R., Baulande, S., Leturcq, F., Urtizberea, J. A., Penisson-Besnier, I., et al. (2015). Detection of TRIM32 deletions in LGMD patients analyzed by a combined strategy of CGH array and massively parallel sequencing. Eur. J. Hum. Genet. 23, 929–934. doi:10.1038/ejhg.2014.223.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276. doi:10.1038/nature08250.

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., and Zwick, M. E. (2007). Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4, 907–909. doi:10.1038/nmeth1109.

Pérez-Serra, A., Toro, R., Campuzano, O., Sarquella-Brugada, G., Berne, P., Iglesias, A., et al. (2015). A novel mutation in lamin a/c causing familial dilated cardiomyopathy associated with sudden cardiac death. J. Card. Fail. 21, 217–225. doi:10.1016/j.cardfail.2014.12.003.

Pfaller, M. A., and Diekema, D. J. (2010). Epidemiology of invasive mycoses in North America. Crit. Rev. Microbiol. 36, 1–53. doi:10.3109/10408410903241444.

Poultney, C. S., Goldberg, A. P., Drapeau, E., Kou, Y., Harony-Nicolas, H., Kajiwara, Y., et al. (2013). Identification of Small Exonic CNV from Whole-Exome Sequence Data and Application to Autism Spectrum Disorder. Am. J. Hum. Genet. 93, 607–619. doi:10.1016/j.ajhg.2013.09.001.

Riviera-Perez, J. I., Santiago-Rodriguez, T. M., and Toranzos, G. A. (2016). Paleomicrobiology: a Snapshot of Ancient Microbes and Approaches to Forensic Microbiology. Microbiol. Spectr. 4. doi:10.1128/microbiolspec.EMF-0006-2015.

Rousseau-Nepton, I., Okubo, M., Grabs, R., Mitchell, J., Polychronakos, C., and Rodd, C. (2015). A founder AGL mutation causing glycogen storage disease type IIIa in Inuit identified through whole-exome sequencing: a case series. CMAJ Can. Med. Assoc. J. 187, E68–E73. doi:10.1503/cmaj.140840.

Rozenblum, A. B., Ilouze, M., Dudnik, E., Dvir, A., Soussan-Gutman, L., Geva, S., et al. (2017). Clinical Impact of Hybrid Capture-Based Next-Generation Sequencing on Changes in Treatment Decisions in Lung Cancer. J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer 12, 258–268. doi:10.1016/j.jtho.2016.10.021.

Schrock, A. B., Pavlick, D., Klempner, S. J., Chung, J. H., Forcier, B., Welsh, A., et al. (2018). Hybrid Capture–Based Genomic Profiling of Circulating Tumor DNA from Patients with Advanced Cancers of the Gastrointestinal Tract or Anus. Clin. Cancer Res. 24, 1881–1890. doi:10.1158/1078-0432.CCR-17-3103.

Schuenemann, V. J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., et al. (2011). Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. Proc. Natl. Acad. Sci. 108, E746–E752. doi:10.1073/pnas.1105107108.

Schuenemann, V. J., Lankapalli, A. K., Barquera, R., Nelson, E. A., Hernández, D. I., Alonzo, V. A., et al. (2018). Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. PLoS Negl. Trop. Dis. 12, e0006447. doi:10.1371/journal.pntd.0006447.

Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jäger, G., Bos, K. I., et al. (2013). Genome-wide comparison of medieval and modern Mycobacterium leprae. Science 341, 179–183. doi:10.1126/science.1238286.

Shearer, A. E., DeLuca, A. P., Hildebrand, M. S., Taylor, K. R., Gurrola, J., Scherer, S., et al. (2010). Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. Proc. Natl. Acad. Sci. U. S. A. 107, 21104–21109. doi:10.1073/pnas.1012989107.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145. doi:10.1038/nbt1486.

Sikkema-Raddatz, B., Johansson, L. F., de Boer, E. N., Almomani, R., Boven, L. G., van den Berg, M. P., et al. (2013). Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum. Mutat. 34, 1035–1042. doi:10.1002/humu.22332.

Smith, M., Campino, S., Gu, Y., Clark, T. G., Otto, T. D., Maslen, G., et al. (2012). An In-Solution Hybridisation Method for the Isolation of Pathogen DNA from Human DNA-rich Clinical Samples for Analysis by NGS. Open Genomics J. 5. doi:10.2174/1875693X01205010018.

Spyrou, M. A., Tukhbatova, R. I., Wang, C.-C., Valtueña, A. A., Lankapalli, A. K., Kondrashin, V. V., et al. (2018). Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. Nat. Commun. 9, 2234. doi:10.1038/s41467-018-04550-9.

Templeton, J. E. L., Brotherton, P. M., Llamas, B., Soubrier, J., Haak, W., Cooper, A., et al. (2013). DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. Investig. Genet. 4, 26. doi:10.1186/2041-2223-4-26.

Thomson, E., Ip, C. L. C., Badhan, A., Christiansen, M. T., Adamson, W., Ansari, M. A., et al. (2016). Comparison of next generation sequencing technologies for the comprehensive assessment of full-length hepatitis C viral genomes. J. Clin. Microbiol., JCM.00330-16. doi:10.1128/JCM.00330-16.

Wagner, D. M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J. W., et al. (2014). Yersinia pestis and the plague of Justinian 541-543 AD: a genomic analysis. Lancet Infect. Dis. 14, 319–326. doi:10.1016/S1473-3099(13)70323-2.

Wylie, T. N., Wylie, K. M., Herter, B. N., and Storch, G. A. (2015). Enhanced virome sequencing using targeted sequence capture. Genome Res. 25, 1910–1920. doi:10.1101/gr.191049.115.

Xie, J., Lu, X., Wu, X., Lin, X., Zhang, C., Huang, X., et al. (2016). Capture-based next-generation sequencing reveals multiple actionable mutations in cancer patients failed in traditional testing. Mol. Genet. Genomic Med. 4, 262–272. doi:10.1002/mgg3.201.

Xu, M.-D., Liu, S.-L., Feng, Y.-Z., Liu, Q., Shen, M., Zhi, Q., et al. (2017). Genomic characteristics of pancreatic squamous cell carcinoma, an investigation by using high throughput sequencing after in-solution hybrid capture. Oncotarget 8, 14620–14635. doi:10.18632/oncotarget.14678.

V. Objectifs et présentation de la thèse

Dans ce contexte (**Chapitre I**), l'objectif de mon travail de doctorat visait à améliorer l'approche de métagénomique virale pour la détection de pathogènes et d'en faire un outil à la fois performant et sensible pour la détection de l'origine infectieuse de certaines pathologies idiopathiques.

J'ai ainsi proposé et développé un protocole expérimental applicable directement après l'extraction des acides nucléiques qui permettait de diminuer très sensiblement la contamination humaine présente dans des échantillons biologiques complexes (**Chapitre II**). J'ai ensuite appliqué ce protocole à deux cas cliniques d'infections idiopathiques que nous avons eu à élucider dans notre laboratoire (**Chapitre III**). La première étude a porté sur l'identification de la cause virale associée à un cas d'encéphalite fatale chez une enfant tandis que la seconde, collaborative, s'est intéressée à un cas énigmatique d'endocardite infectieuse pour laquelle la culture, la sérologie et la PCR à *Coxiella burnetii* était négative alors que l'immunohistochimie (IHC) était positive.

Pour finir (**Chapitre IV**), j'ai discuté les possibilités et les freins à l'intégration de cette approche dans les laboratoires de diagnostic cliniques, les limites du diagnostic par NGS et les développements possibles pour confirmer le lien de causalité entre un agent détecté et une pathologie observée.

Chapitre II : Mise au point d'un protocole de déplétion de la contamination humaine en métagénomique virale

Préambule à l'article 2 "Application of an inverted human Whole-genome In-Solution Capture (inv-WISC) to viral metagenomics"

Un des principaux challenges de la métagénomique virale appliquée à des prélèvements cliniques consiste à limiter la contamination des métagénomes par les séquences nucléiques de l'hôte. Dans ce chapitre, j'ai mis au point une technique applicable après l'extraction des acides nucléiques et donc indépendante de la nature de l'échantillon (biopsie ou fluide biologique) et de sa conservation (frais ou congelé), permettant de diminuer significativement la quantité d'acides nucléiques humains de l'échantillon et ainsi la quantité de séquences humaines après séquençage NGS. Pour ce faire, j'ai détourné le principe de l'approche dite d'enrichissement ciblé utilisant des sondes ARN qui vont servir d'appât pour capturer (après hybridation) les séquences d'une cible, dans notre cas le génome humain. Contrairement au protocole original, les séquences hybridées (hybride ARN:ADN) ne sont pas conservées tandis que la fraction restante non hybridée (non-humain) est purifiée et séquencée par NGS. Pour construire les sondes ARN, je me suis basé sur une étude publiée par Carpenter et al. en 2013 dans laquelle les auteurs ont mis au point une technique appelé WISC (Whole-genome In-Solution Capture) permettant d'enrichir des échantillons biologiques anciens en ADN endogène humain en utilisant des sondes ARN biotinylées couvrant la totalité du génome humain [119].

J'ai dans un premier temps mis au point et évalué l'efficacité de cette approche en fonction de la quantité d'ADN humain engagé (100 ou 10 ng), de la durée d'hybridation (16, 24 ou 66h) et du protocole de purification des acides nucléiques non hybridés. Une fois ces paramètres définis, j'ai testé cette approche sur un métagénome viral artificiel constitué d'ADN humain enrichi avec différentes concentrations en acides nucléiques viraux (virus de l'herpès 1 / HSV-1) et quantifié par PCR quantitative avant et après capture, l'efficacité de la déplétion et son effet sur la quantité de cibles virales. Une fois le protocole mis au point, je l'ai validé en séquençant le métagénome contenant le plus faible nombre de copie de génome viral. Les résultats ont montré que la capture a permis de passer de 99,6% à 56,3% de séquences humaines dans les métagénomes, soit une déplétion de 56,5 fois et d'enrichir de 64 fois le nombre de séquences virales (HSV-1) détecté. Enfin, j'ai appliqué ce protocole à 7 échantillons cliniques de plasma sanguin et montré une diminution significative de la contamination humaine des métagénomes après capture et séquençage NGS. Contrairement aux méthodes classiques qui

additionnent plusieurs traitements dans la préparation d'un métagénome, l'utilisation de cette approche limite les biais tout en réduisant de façon significative la quantité de séquences d'hôte.

Article 2: Application of an inverted human Whole-genome In-Solution Capture (inv-WISC) to viral metagenomics

Maxime Gaudin[1], Sonia Monteil-Bouchard[1], Caroline Michelle[1], Catherine Robert[1], Didier Raoult[1] and Christelle Desnues[*1]

[1]Aix-Marseille Université, IRD 198, CNRS FRE2013, Assistance-Publique des Hôpitaux de Marseille, Microbes, Evolution, Phylogeny and Infections (MEPHI), IHU Méditerranée Infection, Marseille France.

Corresponding author

Christelle Desnues, Ph.D.,

MEPHI, IHU Méditerranée Infection,

19-21 Boulevard Jean Moulin, 13005 Marseille, France

Fax: (+33) 4 13 73 24 24

Email: christelle.desnues@univ-amu.fr

- **Statut : En cours de révision dans Journal of Clinical Microbiology (submission number JCM01576-18)**

ABSTRACT

Shotgun Next Generation Sequencing (NGS) is a promising tool for the untargeted identification of viral pathogens in diagnostic microbiology. However, virus-derived sequences typically represent a very limited proportion of metagenomic samples compared to host-derived sequences. Recently, a procedure called WISC (for Whole-genome In-Solution capture), which captures and enriches human ancient DNA from complex biological sample, has been described. This approach relies on the mechanical capture of human nucleic acids after hybridization on a human RNA bait library. In this study, we tested an inverted WISC (inv-WISC) approach for depleting human sequences in viral metagenomes. Depletion of human nucleic acids was optimized and verified and on a mock human/viral metagenome consisting of human DNA extracted from PBMC and spiked with different concentrations of viral nucleic targets (Herpes Simplex Virus 1). After defining the best conditions, we validated this protocol by showing a reduction of more than 90% of human contamination by real-time quantitative PCR. NGS results confirmed a 56.5-fold depletion of the human reads and 64-fold enrichment of the HSV-1 reads after capture, further supporting the inv-WISC method for the detection of low-copy viral pathogens. Finally, NGS results confirmed the efficiency of human depletion on nucleic acids extracted from 7 human clinical specimens.

IMPORTANCE

Hybridization and capture of DNA directly after extraction by using human RNA bait libraries result in a significant reduction of background human DNA in shotgun metagenomes. It improves the sensitivity of next-generation sequencing approaches for viral metagenomics on human samples and is particularly adapted for clinical diagnostic laboratories.

INTRODUCTION

The shotgun next generation sequencing (NGS) of viral nucleic acids in a particular sample, aka viral metagenomics, has been largely applied in the recent years, both for viral diversity studies and pathogen discovery (1-3). One of the challenges of viral metagenomics applied to human clinical samples is that viruses typically represent a very small fraction of the sequencing data compared to host-derived sequences (4). Since the genome size of viruses is significantly smaller than that of their host, a minimal contamination with human cells can flood this sample with non-viral information (5). For instance, Wylie et al. experienced fewer than 10 viral reads per 25 million reads sequenced in human-associated samples that were positively tested by molecular biology (6, 7). To overcome these limitations, several procedures have been used to enrich virions in a biological sample and most of them rely on the properties that viral genomes are protected by protein capsids and sometimes lipid envelope. Thus, common virome preparation is based on initial centrifugation steps to clarify the samples followed by filtration, nuclease digestion and/or ultracentrifugation on density gradients (8). A recent study also showed that preprocessing of clinical specimens with detergents prior to nuclease digestion resulted in a significant reduction of background human DNA for NGS analysis (9). Thus, treated samples theoretically leave only enriched encapsidated viral DNA and RNA before extraction (8). However, a high proportion of host contaminants may remain, particularly in the case of human clinical samples (10-14). Moreover these downstream processes may lead to a significant loss of information about persistent viral infections, endogenous retroviruses and large viruses that can be retained on filters (15).

Other approaches, targeting post-extraction nucleic acids, have been recently developed to improve NGS sensitivity for viral sequence detection. These include either the use of methylation-specific DNases to digest host/bacteria genome (16, 17), capture based hybridization assay with specific probes to deplete host/bacterial material (18, 19) or capture

of a specific or a set of viral DNA/RNA targets (7, 20-23). Viral target isolation by hybridization and subsequent enrichment has been shown to be effective in the diagnostic (7, 20), the detection of viral variant (7, 20), the analysis of DNA mutation (21), the study about dynamism of human virome (20, 22) and the characterization of integrated retroviruses (23). However, these post-extraction methods are not appropriated in the context of viral pathogen discovery since they require some knowledge of the sequences of the target.

In 2013, Carpenter *et al.,* developed a procedure called WISC (for Whole-genome In-Solution capture) that captures and enriches human ancient DNA from complex biological sample (24). This approach relies on the mechanical capture of human nucleic acids after hybridization on a human RNA bait library. Here, we tested this approach and used an inverted capture-based hybridization assay with the creation of human genomic RNA "bait" libraries derived from a modern reference individual spanning the entire human genome. Depletion of the human nucleic acids was verified on mock viral metagenomes consisting of human DNA extracted from PBMC and spiked with different concentrations of viral nucleic targets and on clinical samples.

RESULTS

Preparation of human RNA bait libraries

Five whole-genome fragment libraries were created after shearing human DNA. Fragments of size comprised between 100-120 bp on average were generated (**Figure 1A**). After end-repair and dA-tailing, a double strand T7 promoter was ligated at both ends of the fragmented DNA. Non-ligated adapters were removed after a gel size selection with purification of bands at 160–260 bp (inserts 120–240 bp) for 4 libraries (**Figure 1B**, libraries 2, 3, 4 and 5) as the library 1 gave a very week signal after purification and was not used for the downstream experiment. To increase the yield of each library, 5 µl of purified ligation

product were amplified by PCR using T7 primers in two separate reactions. All 8 reactions were pooled, concentrated and quantified to obtain 100 ng of human libraries per µl. 500 ng of these amplified DNAs (5 µl) was transcripted into biotinylated RNA with T7 RNA polymerase using biotin-16-UTP. After DNase treatment to digest non-transcript DNAs, the RNA was purified, quantified at 1660 ng/µl (~50 µg total) and the size of the RNA was verified on a RNA6000 Nano chip. A 216 bp peak, as expected, was measured (**Figure 1C**).

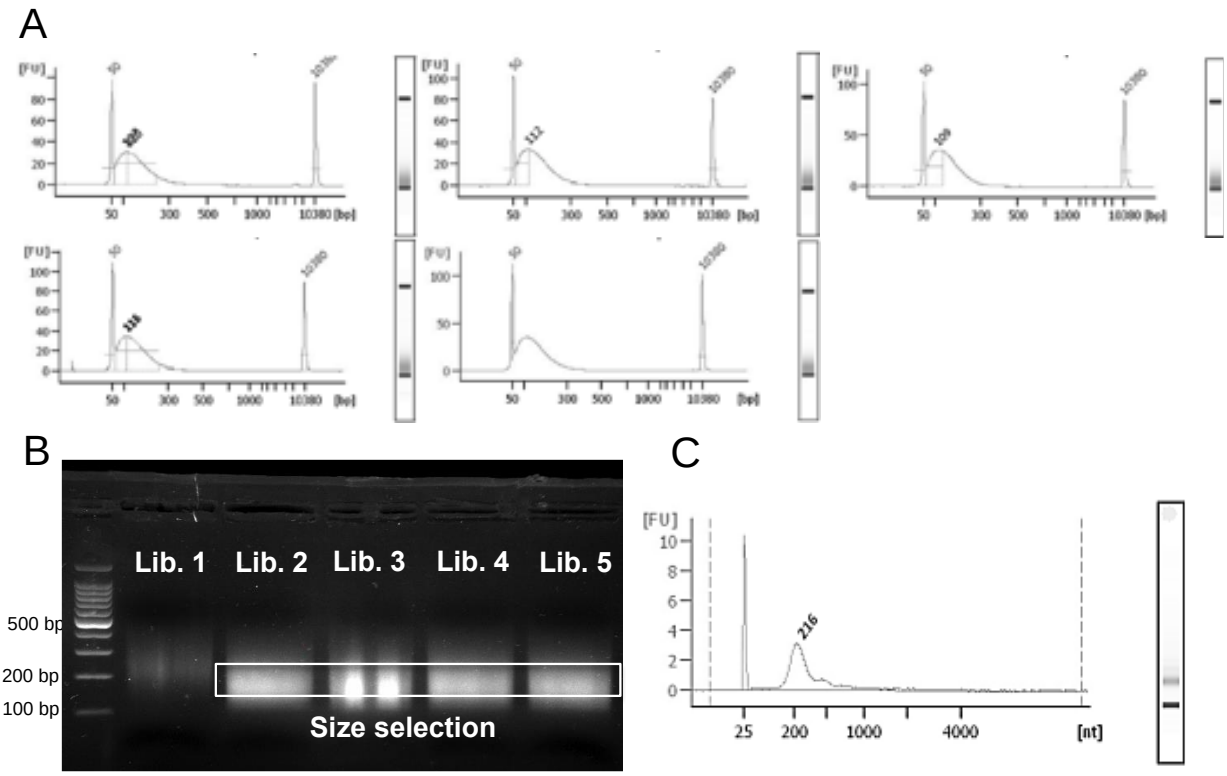
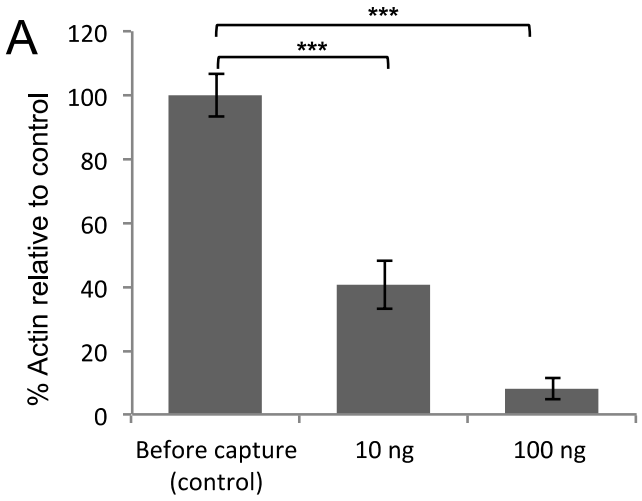


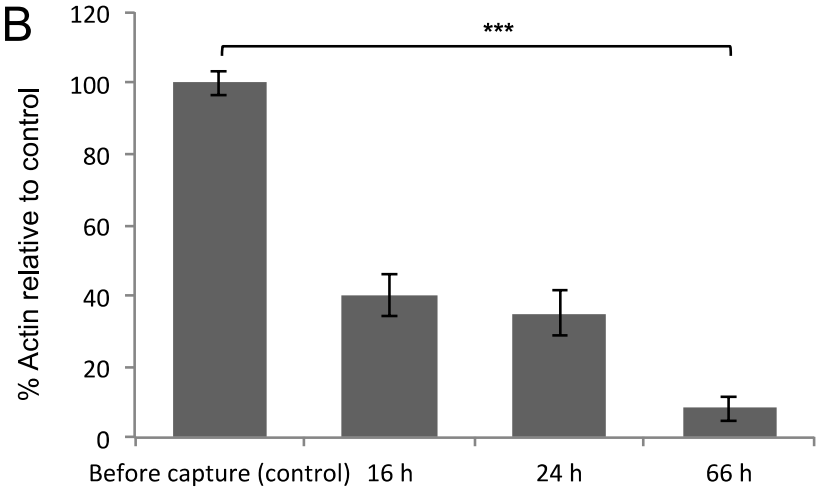
Figure 1. Preparation of human RNA baits libraries

(A) Verification of size fractionation. Five sheared human DNA libraries were run on an Agilent chip (DNA 7500 ladder). Size-fractionation of DNA was observed between a 100–120 bp range. **(B)** Agarose gel electrophoresis after ligation of T7 adaptators. All T7 ligated products were run on a 2% agarose gel and bands at 160-260 bp were purified for libraries 2 to 5 (highlighted by the box). **(C)** Verification of the size after human RNA transcription. Human biotinylated RNA library was run on an Agilent chip (RNA 6000 ladder).

Capture efficiency according to the quantity of human DNA enrolled and the hybridization time.

Hybridization and capture was first tested on 100 ng (1.14E+06 actin copies) and 10 ng (1.16E+05 actin copies) of DNAs extracted from human PBMCs (**Figure 2**). The depletion protocol was performed using 50 µl of Dynabeads MyOne Streptavidin C1 beads for the capture. After 66 hours of hybridization, the number of copies of actin was measured by qPCR in the depleted fraction before any purification and compared with that measured before the capture. As compared with the undepleted samples, the proportion of remaining human DNA was evaluated at 40.4 ± 7.6% and 8.2 ± 3.5% for 10 ng and 100 ng of human DNA involved in the capture reaction. These values corresponded to a host depletion yield of 59.6 and 91.8 for 10 ng and 100 ng, respectively (**Figure 2A**). The yield of host depletion was not improved using 100 µl Dynabeads MyOne Streptavidin C1 beads either with 10 and 100 ng involved (data not shown). Human depletion was also verified according to the hybridization time (16 h and 24 h), with 100 ng of DNA enrolled and 50 µl of streptavidin beads. The efficiency of capture decreases with shorter capture times at 59.8 ± 5.9% and 64.9 ± 6.6% of depletion after 16 hours and 24 hours, respectively (**Figure 2B**), compared to the 91.8 ± 3.5% obtained with 66 hours of hybridization. Finally, we tested 3 different methods to clean up the depleted fraction (**Figure 2C**) and quantified the recovery of human DNA. We reported better recovery performances with the 1.8x AMPure XP beads (87.3 ± 10.8%), with no significant differences in the quantity of DNA recovered after cleaning (P=0.143), in contrast to the QIAquick PCR and MinElute PCR purification kits.


A
***
***
120
100
80
60
40
20
0
% Actin relative to control
Before capture (control)
10 ng
100 ng


B
***
120
100
80
60
40
20
0
% Actin relative to control
Before capture (control)
16 h
24 h
66 h

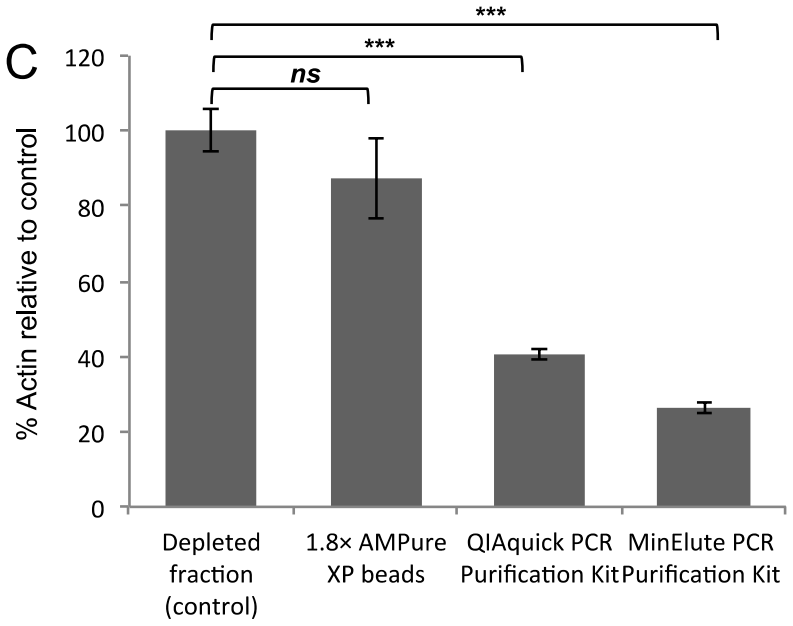

C
***
***
ns
120
100
80
60
40
20
0
% Actin relative to control
Depleted fraction (control)
1.8× AMPure XP beads
QIAquick PCR Purification Kit
MinElute PCR Purification Kit

Figure 2. Capture yield according to the amount of DNA enrolled (A) the hybridization time (B) and the purification method (C).

Depletion efficiency, expressed as the percentage of actin relative to the undepleted (or unpurified) controls was quantified for (A) 100 ng and 10 ng of the human DNA engaged, **(B)** three hybridization times (16 h, 24 h and 66h) and (C) three different purification methods. All data corresponded to the absolute numbers of actin copies (mean ± SD) measured by a minimum of 3 replicates of real-time quantitative PCR assays (except for 16h and 24h were duplicates have been done). Statistical significant differences between conditions ($P<0.001$ (***)) were calculated, ns=non significant ($P>0.05$).

Evaluation of the specificity of the capture-based method on a mock human/viral metagenomes

To assess the specificity of the capture, mock human/viral metagenomes were prepared. DNAs extracted from human PBMCs and Herpes Simplex Virus 1 (HSV-1) were mixed to reach a theoretical ratio of 1E+06 copies of actin with 1E+05 copies of HSV-1 (metagenome A) or 1E+06 copies of actin with 1E+04 copies of HSV-1 (metagenome B, reflecting low viral loads). The mock human/viral metagenomes were then sheared at 500 bp and the number of copies of actin and HSV-1 was determined**.** The reliability of the depletion protocol to capture human DNA, without affecting HSV-1 quantity was tested on these two mock human/viral metagenomes. To do so, 4 µl of mixtures previously sheared at 500 bp were enrolled for hybridization/capture process using 50 µl of beads. After depletion (depleted fraction), quantification results confirm the specificity of the method to capture only human DNA without significantly reducing the quantity of HSV nucleic acids measured (HSV in depleted fraction, **Figures** 3A and 3B), for both the metagenomes A and B. However, the purification process using AMPure XP beads significantly affects the yield of HSV recovery as about half of the HSV DNA was lost for both the metagenomes A and B (**Figures 3A and 3B**).

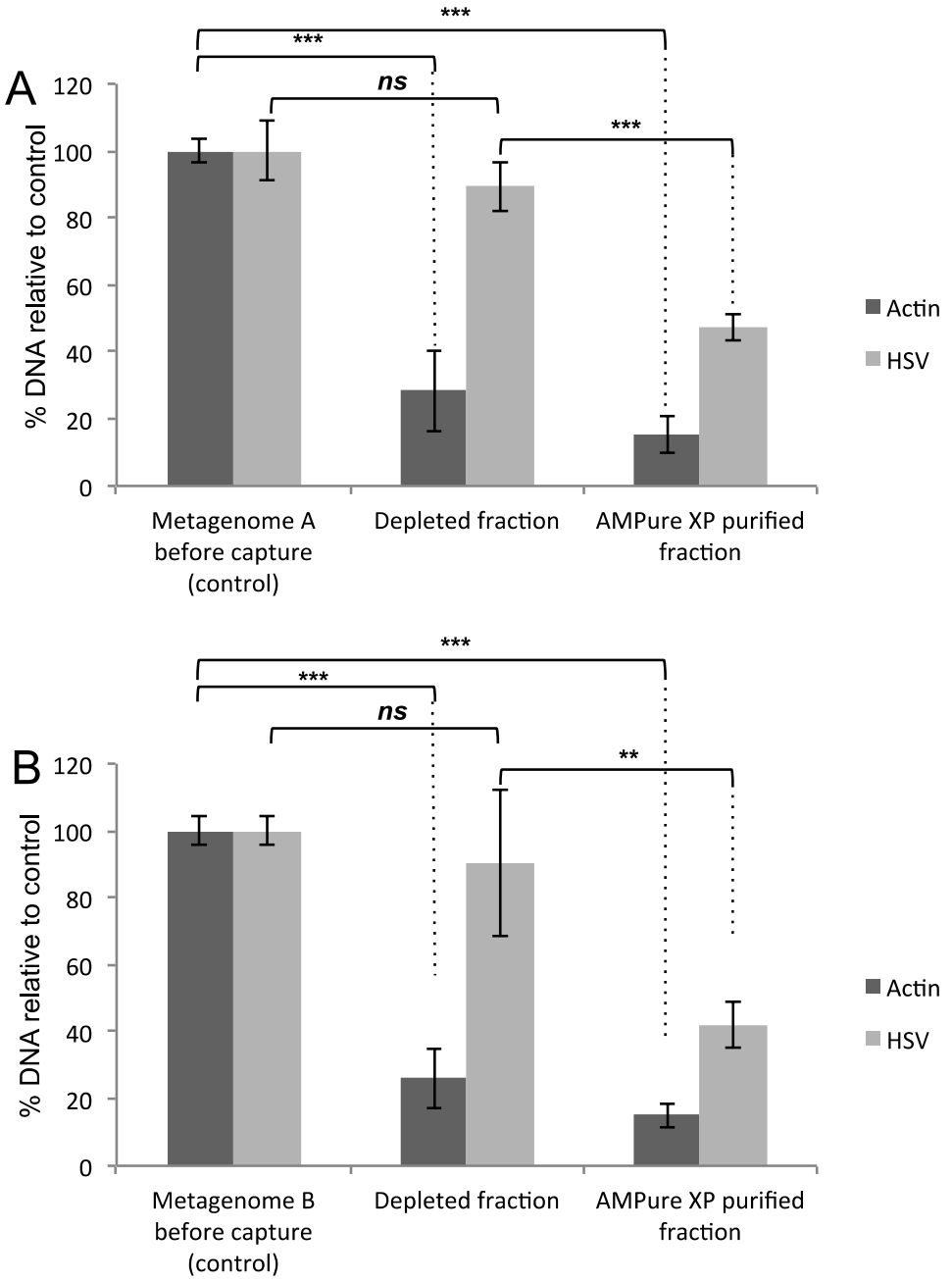


Figure 3. Specificity of the capture-based method on mock human/viral metagenomes

Hybridization and capture were tested on mock human/viral metagenomes containing (A) 1E+06 copies of actin with 1E+05 copies of HSV-1 (metagenome A), and (B) 1E+06 copies of actin with 1E+04 copies of HSV-1 (metagenome B). Data corresponded to the mean ± SD of beta-actin and HSV-1 polymerase absolute copy numbers measured by real-time quantitative PCR at least in triplicates in the undepleted sample (control), the depleted and purified fractions. Results are expressed as the % of DNA quantified relative to the control and statistical significant differences between conditions (P<0.01 (**) or P<0.001 (***)) were calculated, ns=non significant (P>0.05).

Capture yield according to RNA-bait quantities and fragmentation size

The capture efficiency was measured using twice as much RNA baits. Using the same method, 4 µl of artificial metagenome B (1E+06 copies of actin /1E+04 copies of HSV-1) previously sheared at 500 bp, was depleted using 1 ug of RNA baits. Quantification of human and viral DNA was performed on the original sample and after purification of the depleted fraction (**Figure 4**). No significant improvement (*P*=0.870) of human DNA depletion yield was observed when adding 1 µg of RNA bait libraries (84.2 ± 10.0%) compared to 500 ng (85.0 ± 3.7%) (**Figure 4**). Since NGS sequencing requires large fragments for library construction, we also examined whether the size of DNA fragmentation prior to hybridization could influence the effectiveness of depletion. Metagenome B was thus sheared at 1,500 bp and enrolled for depletion (**Figure 4**). Human/HSV quantification exhibited non-significant differences (*P*=0.250) of capture efficiency when the nucleic acids are sheared at 1,500 bp (89.1 ± 5.5%) compared to 500 bp. Finally, the different conditions tested (1 µg of RNA bait libraries or 1,500 bp fragmentation) did not significantly affect the yield of HSV recovery (**Figure 4**). Based on the yield of human DNA depletion and HSV recovery, we defined that the best parameters for the capture protocol were: about 100 ng of nucleic acids that are fragmented at 1,500 bp and hybridized during 66 hours with 500 ng of biotinylated RNA baits. After capture, the unbound fraction is cleaned up with 1.8X AMPure beads.

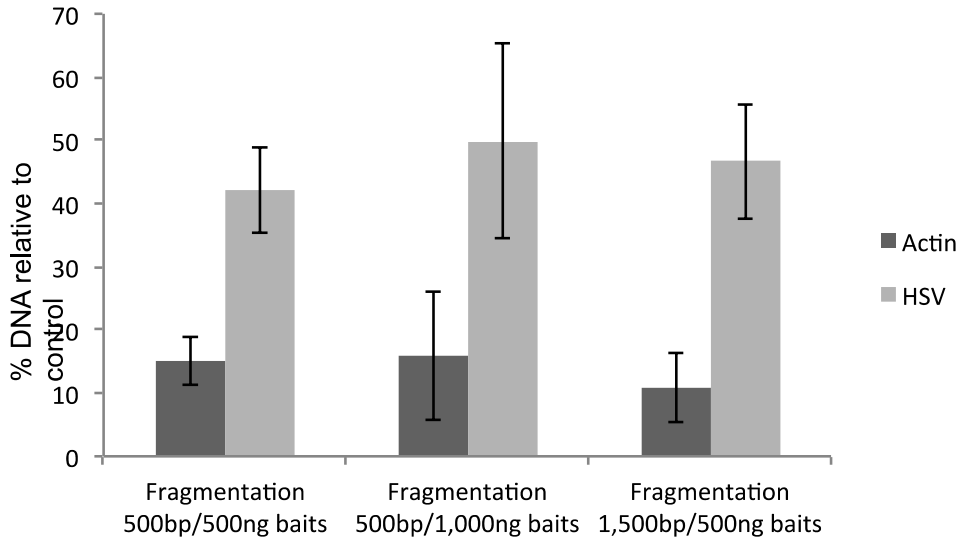


Figure 4. Capture efficiency according to RNA-bait quantities and fragmentation size

The metagenome B was enrolled for hybridization and capture with 1,000 ng of human RNA bait library. The same quantity of metagenome B was also sheared at 1,500 bp and depleted using 500 ng of human RNA bait library. Data corresponded to the mean ± SD of beta-actin and HSV-1 polymerase absolute copy numbers measured by real-time quantitative PCR at least in triplicates in the undepleted sample (control) and the AMPure XP purified fractions. Results are expressed as the % of DNA relative to the control and statistical significant differences between conditions were calculated using the two-tailed Student's *t*-test.

NGS before and after capture on the mock human/viral and clinical metagenomes

We therefore used these optimized parameters on the simulated human/viral metagenome B and the Illumina libraries were built before and after capture for sequencing on a MiSeq platform. After sequencing, the paired reads were imported, trimmed and the number of human reads was estimated after mapping on the human genome (**Table 1**). The number of HSV-1 reads was then detected by mapping the non-human reads on the HSV-1 genome. Prior to capture, the proportion of human paired reads was of 99.6% and it decreased to 56.3% after capture, which corresponded to a depletion of 56.5-fold, whereas an enrichment of 64-fold was obtained for the HSV-1 reads.

Table 1. Number of Illumina paired-reads obtained before and after capture of host nucleic acids

Pre- or post-capture	Number of paired reads	Number of trimmed-paired reads	Number of human reads	Ratio of human reads (%)	Fold depletion	Reads mapped on HSV-1	Ratio of HSV-1	Fold Enrichment
Pre	4,288,522	4,250,232	4,231,942	99.6	56.5	3	0.00007 %	64
Post	1,046,896	1,023,635	576,113	56.3		47	0.0045 %	

In a second time, we applied the same protocol to nucleic acids extracted from 7 human plasma samples and sequenced the pre- and post-capture samples on a MiSeQ platform. The average number of trimmed reads was of 1,108,870 and 1,037,028 before and after capture, respectively. These reads were mapped on the human genome and the results indicated that, although the ratio of human reads remains above 95% for 2 samples after capture (**Figure 5**), the average ratio of human reads significantly decreased (P<0,006) in the depleted metagenomes compared to the non-depleted ones **(Figure 5**).

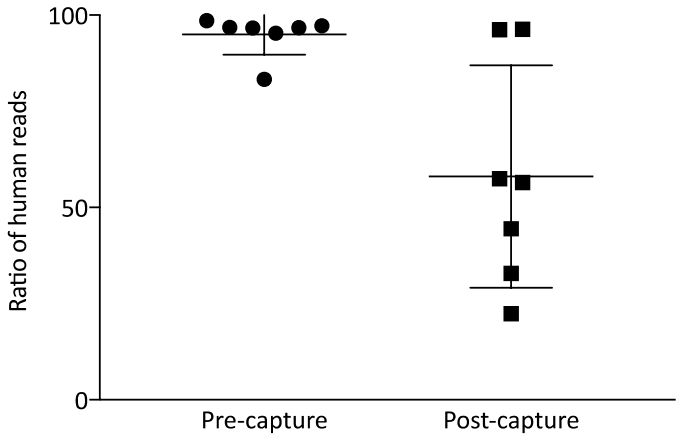


Figure 5. Ratio of human reads before and after capture on human plasma samples.

The ratio was calculated as the number of reads that were mapped onto the human genome compared to the total number of trimmed reads for each of the 7 datasets, before and after capture. Statistical significance was tested using the two-tailed Student's *t*-test.

DISCUSSION

One limit of shotgun metagenomics applied to clinical samples is the presence of human nucleic acids that may hamper the detection of pathogen-associated DNA or RNA. Several studies have highlighted different efficacy degrees of pre-extraction methods using commercial kits or various detergents on reducing the level of human contamination in clinical samples (9, 25-27), as measured by qPCR and/or sequencing. For instance, the MolYsis (Molzym GmbH & Co KG, Bremen Germany) is efficient to remove >90% of human DNA from oral samples (27), but was unsuccessful to recover *S. pneumoniae* DNA or influenza A RNA spiked in cerebrospinal fluids (CSF), or nasopharyngeal aspirates (NPA) (9). Recently, the use of osmotic lysis and treatment with propidium monoazide (lyPMA) on saliva samples has shown promising results in depleting human DNA, but its effectiveness in recovering viral DNA/RNA has not been tested (26). Most of these pre-extraction methods rely on the selective digestion/disruption of human cells, followed by a nuclease treatment (DNAse and/or RNAse) of the released human nucleic acids, and may conduct to a loss of information for latent DNA viruses, bacteria that lack cell walls, and eukaryotic pathogens. In addition, to reduce costs and facilitate transport, freezing samples is relatively common in clinical laboratories. However, several freezing and thawing cycles may disrupt both human cells and potential pathogens, and further limit the use of pre-extraction treatments.

One alternative resides in the post-extraction approaches, among whose capture of specific nucleic acids has gained a growing interest in the recent years (7, 28, 29). Such approach is usually dedicated to target specific viruses or a set of viruses (7, 20, 22, 23) with a certain capacity to detect variants, but it may yet miss uncommon, emerging or completely unknown viruses. Another alternative, developed in the present work, is to capture and remove human DNA directly from the extracted nucleic acids. Several in-solution hybridization methods to capture human genetic material with specific targets, such as the mitochondrial

genome, single nucleotide polymorphism (SNP) markers, or the exome have been published (18, 30, 31). However, generating baits by focusing only on a subset of human genome is clearly not appropriate in viral metagenomics where minimal remaining human contamination may alter sequencing results. Therefore, bait-libraries derived from the whole human genome, as the ones developed in the protocol proposed by Carpenter et al., 2013 (24) are required. Such libraries are also commercially available with the myBaits(R) manufacturer (pre-designed Human WGE kit). Both are indented to capture and enrich the endogenous human DNA from ancient sample, and as a consequence, some part of the protocol (e.g., throwing the supernatant after capture) and reagents (e.g., adding sequences of Human-*Cot1*, a blocker of repetitive sequences, and salmon sperm) are not adapted when sequencing the unbound fraction. In this study, we tested, optimized and applied an inverted whole genome in-solution capture protocol (inv-WISC) to deplete human nucleic acids before NGS sequencing of the unbound nucleic acids from a mock human/viral metagenome. The protocol we developed has reduced the level of contamination of human DNA by up to 90% compared to the native sample as measured by qPCR. After sequencing, a 56.5-fold depletion of the human sequences was observed. In addition, although the clean up by AMPure XP beads reduced by half the amount of viral DNA quantified by qPCR, an enrichment of 64-fold of the HSV-1 reads was still detected in the NGS data, further validating the inv-WISC method for low-copy viral pathogens in human sample. As it avoids any pre-extraction process and can operate on nucleic acids already extracted for other purposes, this approach is particularly suitable for diagnostic laboratories. Indeed, we observed a significant decrease in the average ratio of human reads after capture of human nucleic acids extracted from plasma samples. For two samples however, the depletion did not work and a second round of capture might be necessary. Although very promising, some issues still need to be optimized for routine diagnostic, including the turnaround time that has to be

reduced, the initial cost, the technical complexity and the standardization of the method for an application to other types of sample, such as saliva, sputum or tissue biopsy.

When working with human biological samples, the quantity of nucleic acids obtained after the extraction may be limited and even not measurable after capture. In that case, post-capture amplification may be required. We chose to use the commercial Illustra™ GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Freiburg, Germany) to generate sufficient material from the plasma samples. However, GenomiPhi amplification has been associated with the production of quantitative biases (32) and a preferential amplification of ssDNA viruses and small circular DNA viruses (33). To avoid these biases, the capture protocol can be applied on Illumina libraries that have been constructed right after nucleic acid extractions. In these cases, amplification with a limited number of cycles using primers targeting the Illumina adapters could be a good alternative (34). Such post-capture amplification step is commonly used when processing ancient DNA or forensic samples (24, 31).

Although not tested in this work, several perspectives of applications of the inv-WISC approach could be developed in the future. For instance, it could be adapted and experimentally tested for viruses with RNA genomes by hybridization and capture on retro-transcribed RNAs (cDNAs). Moreover, as this approach is untargeted and can virtually enrich in any non-human nucleic acids, it may be suitable not only for viruses, but also for the detection of sequences of any other extra or intracellular pathogens. Finally, RNA bait libraries could be constructed using genomic DNA from other species than humans, upon extraction of genetic material in good quality and quantity.

MATERIALS AND METHODS

Propagation of HSV-1 viral strain

Herpes Simplex Virus 1 strain Marseille/ (HSV-1) was propagated in Vero cells. After 4 passages, viral supernatant was precipitated with 10% PEG 8000 (Sigma-Aldrich, Saint-Quentin Fallavier, France) and 300 mM NaCl (Sigma-Aldrich, Saint-Quentin Fallavier, France) overnight at +4°C. After centrifugation at 12,000 g for 30 minutes at +4°C, the pellet was resuspended in 15 ml of Phosphate Buffer Saline solution (PBS), aliquoted and stored at -80°C until further use.

PBMC isolation from human blood sample

PBMC were isolated from 4 ml of EDTA-blood samples obtained from the Etablissement Français du Sang (Marseille, France). After centrifugation at 600 g at RT for 25 minutes, PBMC were separated by Ficoll density gradient (Eurobio). Cells were washed with RPMI Medium 1640 (GIBCO) and counted. Viability was estimated by Trypan blue dye exclusion. Freshly isolated PBMC were divided into 200 µl of RPMI medium aliquots of 2E+06 cells.

Clinical samples

A total of 7 plasma samples, collected for routine diagnostic purposes at the Mediteranean Infection Institute, were included in this study. Informed consent forms were obtained for each patient. Plasma samples have been kept frozen at -80°C before processing. Plasmas (300 to 500 µl depending on the sample) were centrifuged at 1,500xg for 10 minutes and the supernatants were digested with a cocktail of nucleases for 1 hour at 37°C as previously described (35).

Nucleic acid extraction and quantification

DNA extraction was performed on 100 µl of HSV-1 culture supernatant and 200 µl of PBMC (2E+06 cells) with the High Pure Viral Nucleic Acid kit (Roche Diagnostics, Meylan, France) Nucleic acids of plasma samples were extracted with the High Pure Viral Nucleic Acid Large Volume Kit (Roche Diagnostics, Meylan, France) according to the manufacturer's protocols. Nucleic acids were eluted in 50 µl of elution buffer, aliquoted and stored at -20°C. The DNA concentration was estimated with the QuantiFluor® dsDNA System (Promega, Charbonnières-les-Bains, France), according to the manufacturer's recommendation, and fluorescence was quantified with the Tecan GENios fluorometer.

Quantification of human beta-actin and DNA polymerase of HSV-1

Amplification by PCR of the human beta-actin and viral DNA polymerase genes was performed in a PCR reaction mix consisting in 0.25 U HostarTaq DNA polymerase (QIAGEN), 1X PCR buffer, 0.2 µM each primer couples (**Table 2**), 200 µM of dNTPs and 5 µl of template DNA. After an initial denaturation of 15 minutes for 95°C, 40 amplification cycles have been made under the following conditions: denaturation (30 sec, 94°C), hybridization (30 sec, 60°C) and elongation (30 sec, 72°C).

Table 2. Primers used in this study

Oligonucleotides used for the detection of HSV-1/2 polymerase and human beta-actin genes by real-time quantitative PCR along with the sequence of the primers, the amplicon length and the references.

Organism	Target gene	Name	Sequence (5' → 3')	Amplicon length (bp)	Ref
Human	Beta-actin	Actin-F	CATGCCATCCTGCGTCTGGA	172	*(36)*
		Actin R	CCGTGGCCATCTCTTGCTCG		
HSV-1/2	Polymerase	HSV1-2_DNApol-F	CATCACCGACCCGGAGAGGGAC	92	*(37)*
		HSV1-2_DNApol-R	GGGCCAGGCGCTTGTTGGTGTA		

Five microliters of the PCR product were verified by electrophoresis on a 2% agarose gel and the remaining volume was purified with the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's protocol. Purified PCR products were cloned into a TOPO TA vector (Invitrogen) by incubating 2 µl of purified PCR product, 1 µl of salt solution, and 1 µl TOPO vector at room temperature for 1 hour. Transformation was performed with TOP10 Competent cells, spread onto 400 µL pre-warmed LB Agar plates containing 50 mg/ml kanamycin, and then incubated at 37°C for 15-16 hours. Transformants were analyzed by PCR using M13 primers and positive clones were placed in 8 ml of LB broth containing 50 mg/ml kanamycin, and horizontally shaken for 15–16 hours. Plasmids were extracted using the QIAprep Spin Miniprep Kit (Qiagen) and verified by PCR and Sanger sequencing. Purified plasmids were linearized with BamHI digestion (Invitrogen) according to the manufacturer's recommendations. The restriction products were run on a 1% agarose gel and the bands at the proper size were purified using the QIAquick Gel Extraction kit (QIAGEN). The concentration of plasmid was quantified with the QuantiFluor® dsDNA System (Promega, Charbonnières-les-Bains, France) and was calculated as genome equivalents (copies/µl) based on the plasmid molecular weight. Serial dilutions of positive control quantification standard plasmid were aliquoted and stored at -20°C until further use.

Quantitative SYBR Green real-time polymerase chain reaction

One to five microliters of template DNA was added to a final volume of 20 µl containing 10 µl of QuantiTect SYBR Green Master Mix (QIAGEN) and 0.5 µM of each couple of primer listed in the **Table 2**. After an initial denaturation of 15 minutes at 95°C, 40 amplification cycles have been made under the following conditions: denaturation (30 sec at 94°C), hybridization (30 sec at 60°C) and elongation (30 sec at 72°C). Amplification and emission of each sample was recorded in real time with CFX 96 Real Time (Bio-Rad) to

generate cycle threshold (Ct) values. A standard curve was then computed from the Ct values of diluted standard (R^2>0.99 and E>92%) and absolute quantities were calculated based on their Ct values. Unless notified, all PCRs have been performedat least in triplicates and results are expressed as mean values ± standard deviation (SD). Statistical significance was tested using the two-tailed Student's *t*-test (P<0.05).

Construction of a human biotinylated RNA-bait library

Five whole-genome fragment libraries were constructed as described in the protocol developed by Carpenter et al., 2013 with slight modifications (24). Briefly, for each library, 5 µg of human DNA (HapMap individual NA21732) was sheared with a Covaris S2 sonicator and subjected to end-repair and dA-tailing using a KAPA library preparation kit (KAPA) according to the manufacturer's protocol. Ligation was performed using the custom adapters as described previously (24). The ligation products were extracted from a 2% agarose gel using the QIAquick Gel Extraction kit (QIAGEN) for 4 libraries (**Figure 1**, libraries 2 to 5). Each purified library was then PCR amplified in two separate reactions as previously described (24). All 8 reactions were pooled and purified with the MinElute PCR Purification Kit (Qiagen), eluted twice in 11 µl of elution buffer pre-warmed at 50°C to obtained 100 ng of human DNA libraries per µl. Transcription of the human DNA into biotinylated RNA was performed from 500 ng of human DNA as described previously (24). The RNA size has been verified on a RNA6000 Nano chip (Agilent Technologies, Les Ulis, France). For long-term storage, 1.5 µl of SUPERase-In was added, and the RNA was stored at −80°C.

Hybridization and capture

Hybridization was performed using 100 ng (~1E+06 copies of actin) or 10 ng of human DNA (~1E+05 copies of actin) or the mock human/viral metagenomes A (1E+06 copies of

actin with 1E+05 copies of HSV-1 DNA) or metagenome B (1E+06 copies of actin with 1E+04 copies of HSV-1 DNA) previously sheared at 500 bp or 1,500 bp. The sheared DNA was heated in a thermal cycler to 95°C for 5 minutes (denaturation), followed by 65°C for 5 minutes. In a separate tube, a RNA bait library consisting in 500 ng or 1 µg of biotinylated human RNA bait library with 3 µl SUPERase-In was prepared. When the DNA had been at 65°C for 2.5 min, the RNA bait mix was heated to 65°C for 2.5 min in a heat block. After denaturation, 10 µl of prewarmed (65°C) hybridization buffer 2X (10× SSPE, 10× Denhardt's, 10 mM EDTA, 0.2% SDS, and 0.01% Tween 20) was added, followed by the RNA bait mix. The hybridization reaction (~20 µl) was mixed by pipetting, and then incubated at 65°C for 16 hours, 24 hours or 66 hours. For the capture, 50 µl or 100 µl of Dynabeads MyOne Streptavidin C1 beads (Life Technologies) was mixed with 200 µl bead wash buffer (1 M NaCl, 10 mM Tris-HCl [pH 7.5], 1 mM EDTA, and 0.01% Tween 20), and washed as previously described (24). The beads were then resuspended into 80 µl of bead wash buffer and the DNA/RNA hybridization solution was added. The solution was vortexed for 10 s and incubated at room temperature for 1 hour under occasional vortexing. The mixture was then placed on a magnet particle collector to separate the beads, and the supernatant (unbound fraction) was kept. The unbound fraction was then concentrated and cleaned using 1.8x AMPure XP beads (Beckman Coulter) After capture, the depleted fraction was concentrated and cleaned by testing different protocols: 1.8x AMPure XP beads (Beckman Coulter), QIAquick PCR Purification kit (Qiagen) and MinElute PCR Purification kit (Qiagen) according to the manufacturer's protocol. Final elution step was conducted in 40 µl TE buffer (10 mM Tris, 1 mM EDTA, pH 8) for the 1.8x AMPure XP beads, and 40 µl or 11 µl of elution buffer for the QIAquick and MinElute PCR Purification kits, respectively. The quantification of the total number of copies of the beta-actin and HSV-1 polymerase genes was performed by real-time quantitative PCR on the sheared fraction (before capture), the depleted fraction (after capture) and the purified fraction (after purification).

Regarding the clinical samples, 10 µl of extracted DNAs (~100-170 ng depending on the sample) were sheared at 1,500 bp, hybridized during 66 hours with 500 ng of human RNA bait libraries, captured with 50 µl of Dynabeads MyOne Streptavidin C1 beads and the unbound fraction was purified with 1.8x AMPure XP beads. Non-depleted and depleted nucleic acids were amplified in duplicate with the Illustra™ GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Freiburg, Germany) and the resulting amplified DNAs were purified on silica columns (Qiagen).

Illumina sequencing of the mock viral metagenomes and clinical metagenomes before and after capture

Illumina libraries were generated using the Nextera XT library kit on nucleic acids before and after capture, and sequenced on a MiSeq platform using a paired-end strategy in a 2 × 250 bp format (Illumina Inc., San Diego, CA, USA).

Reads pre-processing, annotation and analysis

Reads were imported into the CLC Genomics Workbench 6.0.1 program (CLC Bio, Aarhus, Denmark) with default parameters. Raw Illumina reads were first trimmed according to their quality score (Illumina pipeline 1.8 and later) and their length (reads < 50 nt long were discarded). The amount of human DNA contamination was estimated by mapping the trimmed reads onto the GRCh38 human reference genome with Deconseq (38) (http://deconseq.sourceforge.net). As for the mock human/viral metagenomes, non-human reads were mapped onto the Herpes Simplex Virus 1 (NCBI: txid10298) reference genome using CLC Genomics with a minimal length fraction of 0.5 and a minimal similarity of 0.8 as mapping parameters. The efficiency of the human depletion process was determined by comparing the number of reads mapping to the human and HSV-1 genome relative to the total number of reads obtained. The ratio (%) was calculated as the (number of human or virus-

specific reads / total number of reads) ×100, whereas the depletion or the enrichment folds were calculated as the ratio of human or virus-specific reads in depleted fraction / ratio in the non-depleted sample.

ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche (reference: ANR-13-JSV6-0004), by the IHU Méditerranée Infection, Marseille, France, by the French Government under the «Investissements d'avenir» program (reference: Méditerranée Infection 10-IAHU-03), by the Région Provence-Alpes-Côte d'Azur and by the European funding FEDER PRIMI.

AUTHOR CONTRIBUTIONS

MG and C.D. designed the experiments. MG., C.M., C.R. and S.M.B performed the experiments. M.G., D.R. and C.D. analysed the results. M.G. and C.D. wrote the paper. All authors have approved the final version of the document.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests

REFERENCES

1. **Bexfield N, Kellam P.** 2011. Metagenomics and the molecular identification of novel viruses. Vet J **190:**191-198.
2. **Edwards RA, Rohwer F.** 2005. Viral metagenomics. Nat Rev Microbiol **3:**504-510.
3. **Tang P, Chiu C.** 2010. Metagenomics for the discovery of novel human viruses. Future Microbiol **5:**177-189.
4. **Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI.** 2012. Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol **10:**607-617.
5. **Auburn S, Campino S, Clark TG, Djimde AA, Zongo I, Pinches R, Manske M, Mangano V, Alcock D, Anastasi E, Maslen G, Macinnis B, Rockett K, Modiano D, Newbold CI, Doumbo OK,**

Ouedraogo JB, Kwiatkowski DP. 2011. An effective method to purify Plasmodium falciparum DNA directly from clinical blood samples for whole genome high-throughput sequencing. PLoS One **6:**e22213.

6. **Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA.** 2012. Sequence analysis of the human virome in febrile and afebrile children. PLoS One **7:**e27735.
7. **Wylie TN, Wylie KM, Herter BN, Storch GA.** 2015. Enhanced virome sequencing using targeted sequence capture. Genome Research **25:**1910-1920.
8. **Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F.** 2009. Laboratory procedures to generate viral metagenomes. Nat Protoc **4:**470-483.
9. **Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, Tilley P.** 2016. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. J Clin Microbiol **54:**919-927.
10. **Zoll J, Rahamat-Langendoen J, Ahout I, de Jonge MI, Jans J, Huijnen MA, Ferwerda G, Warris A, Melchers WJ.** 2015. Direct multiplexed whole genome sequencing of respiratory tract samples reveals full viral genomic information. J Clin Virol **66:**6-11.
11. **Sullivan PF, Allander T, Lysholm F, Goh S, Persson B, Jacks A, Evengard B, Pedersen NL, Andersson B.** 2011. An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. BMC Microbiol **11:**2.
12. **Kawada J, Okuno Y, Torii Y, Okada R, Hayano S, Ando S, Kamiya Y, Kojima S, Ito Y.** 2016. Identification of Viruses in Cases of Pediatric Acute Encephalitis and Encephalopathy Using Next-Generation Sequencing. Scientific Reports **6**.
13. **Perlejewski K, Popiel M, Laskus T, Nakamura S, Motooka D, Stokowy T, Lipowski D, Pollak A, Lechowicz U, Caraballo Cortes K, Stepien A, Radkowski M, Bukowska-Osko I.** 2015. Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: Unexpected detection of human herpesvirus 1 while searching for RNA pathogens. J Virol Methods **226:**1-6.
14. **Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY.** 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med **370:**2408-2417.
15. **Van Etten JL, Lane LC, Dunigan DD.** 2010. DNA Viruses: The Really Big Ones (Giruses). Annual Review of Microbiology, Vol 64, 2010 **64:**83-99.
16. **Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, Dimalanta ET, Amaral-Zettler LA, Davis T, Quail MA, Pradhan S.** 2013. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. PLoS One **8:**e76096.
17. **Matranga CB, Gladden-Young A, Qu J, Winnicki S, Nosamiefan D, Levin JZ, Sabeti PC.** 2016. Unbiased Deep Sequencing of RNA Viruses from Clinical Samples. J Vis Exp doi:10.3791/54117.
18. **Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C.** 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol **27:**182-189.
19. **Morlan JD, Qu K, Sinicropi DV.** 2012. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS One **7:**e42882.
20. **Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, Lipkin WI.** 2015. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. MBio **6:**e01491-01415.
21. **Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, Wylie T, Abel HJ, Pfeifer JD.** 2011. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J Mol Diagn **13:**325-333.

22. **Depledge DP, Palser AL, Watson SJ, Lai IY, Gray ER, Grant P, Kanda RK, Leproust E, Kellam P, Breuer J.** 2011. Specific capture and whole-genome sequencing of viruses from clinical samples. PLoS One **6:**e27805.
23. **Miyazato P, Katsuya H, Fukuda A, Uchiyama Y, Matsuo M, Tokunaga M, Hino S, Nakao M, Satou Y.** 2016. Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. Scientific Reports **6**.
24. **Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, Rasmussen M, Gravel S, Guillen S, Nekhrizov G, Leshtakov K, Dimitrova D, Theodossiev N, Pettener D, Luiselli D, Sandoval K, Moreno-Estrada A, Li YR, Wang J, Gilbert MTP, Willerslev E, Greenleaf WJ, Bustamante CD.** 2013. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. American Journal of Human Genetics **93:**852-864.
25. **Horz HP, Scheer S, Vianna ME, Conrads G.** 2010. New methods for selective isolation of bacterial DNA from human clinical specimens. Anaerobe **16:**47-53.
26. **Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K.** 2018. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome **6**.
27. **Horz HP, Scheer S, Huenger F, Vianna ME, Conrads G.** 2008. Selective isolation of bacterial DNA from human clinical specimens. Journal of Microbiological Methods **72:**98-102.
28. **Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ.** 2010. Target-enrichment strategies for next-generation sequencing. Nature Methods **7:**111-118.
29. **Lovett M, Kere J, Hinton LM.** 1991. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. Proc Natl Acad Sci U S A **88:**9628-9632.
30. **Tewhey R, Nakano M, Wang X, Pabon-Pena C, Novak B, Giuffre A, Lin E, Happe S, Roberts DN, LeProust EM, Topol EJ, Harismendy O, Frazer KA.** 2009. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol **10:**R116.
31. **Shih SY, Bose N, Goncalves ABR, Erlich HA, Calloway CD.** 2018. Applications of Probe Capture Enrichment Next Generation Sequencing for Whole Mitochondrial Genome and 426 Nuclear SNPs for Forensically Challenging Samples. Genes (Basel) **9**.
32. **Yilmaz S, Allgaier M, Hugenholtz P.** 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Methods **7:**943-944.
33. **Kim KH, Bae JW.** 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl Environ Microbiol **77:**7663-7668.
34. **Meyer M, Kircher M.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc **2010:**pdb prot5448.
35. **Temmam S, Monteil-Bouchard S, Robert C, Pascalis H, Michelle C, Jardot P, Charrel R, Raoult D, Desnues C.** 2015. Host-Associated Metagenomics: A Guide to Generating Infectious RNA Viromes. PLoS One **10:**e0139810.
36. **Angelakis E, Richet H, Rolain JM, La Scola B, Raoult D.** 2012. Comparison of real-time quantitative PCR and culture for the diagnosis of emerging Rickettsioses. PLoS Negl Trop Dis **6:**e1540.
37. **Kessler HH, Muhlbauer G, Rinner B, Stelzl E, Berger A, Dorr HW, Santner B, Marth E, Rabenau H.** 2000. Detection of Herpes simplex virus DNA by real-time PCR. J Clin Microbiol **38:**2638-2642.
38. **Schmieder R, Edwards R.** 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One **6:**e17288.

Chapitre III : Application à la détection d'agents potentiellement pathogènes dans des échantillons cliniques

Préambule à l'article 3 "Identification of a novel gemycircularvirus associated with a fatal case of child encephalitis"

L'encéphalite est une est une inflammation de l'encéphale associée à une dysfonction neurologique. Elle peut être d'origine auto-immune, oncologique, toxique, métabolique, vasculaire ou infectieuse [120]. Les encéphalites infectieuses correspondent quant à elle à une atteinte directe par un agent infectieux (encéphalite aiguë) ou par une réaction immunologique suite à une infection (phénomène post-infectieux). Leur diagnostic est généralement réalisé par une combinaison de résultats cliniques, de laboratoire, d'imagerie cérébrale et d'examens électrophysiologiques. Pour distinguer l'encéphalite infectieuse des autres causes, les principales manifestations possibles comprennent la présence de fièvre, une pléiocytose dans le liquide céphalo-rachidien, des anomalies de l'IRM ou de l'électroencéphalogramme (EEG) accompagnées par des troubles neurologiques [121]. Bien que l'incidence rapportée du nombre de cas d'encéphalite aiguë varie dans le monde entier, elle est généralement compris entre 1,5 et 7,4 cas pour 100 000 habitants par an [122, 123].

Le diagnostic des encéphalites reste encore difficile [124]. Outre les symptômes peu spécifiques et souvent trompeurs de nombreux agents pathogènes peuvent être impliqués et l'étiologie reste encore inconnue dans environ la moitié des cas. Dans le cas où l'étiologie est déterminée, l'encéphalite infectieuse est principalement causée par des virus, le virus de l'herpès simplex (HSV) étant l'entité la plus fréquemment isolée. D'autres virus tels que le virus varicelle-zona (VZV) et les entérovirus sont également fréquemment retrouvés comme agents principaux dans les diagnostics cliniques. Enfin, des pathogènes rares ou inconnus ont récemment fait leur apparition ou réapparition dans les encéphalites comme par exemple le virus du Nil Occidental, le virus Nipah, le virus Powassan et le virus Hendra. Certains de ces pathogènes entrainent des infections potentiellement graves, possiblement mortelles, avec un risque de séquelles à court et à long terme non négligeable, ce qui fait de l'encéphalite un véritable problème de santé publique.

Dans un récent article de revue de la littérature portant sur 25 articles décrivant 44 cas cliniques, l'utilisation du séquençage haut débit a permis d'identifier un agent potentiellement responsable chez 44 patients atteints d'encéphalite infectieuse. Dans 21 des 44 cas, les virus et bactéries retrouvés étaient déjà connus (causes fréquentes ou rares) pour être responsable d'encéphalites et auraient pu être identifiés par des méthodes classiques de dépistage. Dans les 23 cas restants, de nouveaux virus (arenavirus, variegated squirrel bornavirus, astrovirus

VA1/HMO-C, cyclovirus, gemycircularvirus, densovirus), ou des virus déjà connus pour infecter l'homme mais inattendus dans ce type de pathologie (parvovirus humain 4, coronavirus humain OC-43, astrovirus MLB1 et le virus présent dans le vaccin de la rougeole) ont été détectés [125]. Dans tous ces cas cliniques, les données issues du séquençage ont été obtenues après extraction des acides nucléiques à partir de liquide céphalorachidien (LCR) ou de biopsie cérébrale. Dans un peu moins de la moitié de ces études, les échantillons cliniques ont été traités suivant un protocole métagénomique direct, mais 91 à 99,8% des séquences générés s'alignaient contre le génome humain représentant un frein pour l'analyse des données [60, 85, 126, 127]. Certains auteurs ont inclus des étapes de traitement (digestion avec des nucléases, filtration ou ribodéplétion) des échantillons avant ou après l'extraction des acides nucléiques. La contamination en séquence de l'hôte s'est avéré encore très élevée dans cinq études (82 à 99% des séquences totales) [60, 63, 65, 94, 128] tandis que dans une étude les auteurs ont même observé une diminution de trois fois de la proportion en séquence virale malgré la faible teneur en séquences humaines présentes (44%) [85]. Ces études soulignent ainsi l'intérêt de la métagénomique pour le diagnostic des encéphalites idiopathiques mais mettent en évidence la nécessité d'améliorer le protocole de préparation des métagénomes pour ce type d'échantillons.

Dans ce contexte, nous avons appliqué le protocole mise au point dans le **chapitre II** de cette thèse afin de réaliser par séquençage haut débit une analyse métagénomique du microbiome d'une biopsie cérébrale prélevé chez une petite fille décédée d'une encéphalite fulgurante. L'analyse du virome ADN a permis de détecter des séquences d'un nouveau gemycircularvirus et de reconstruire son génome complet. Les analyses bioinformatiques et phylogénétiques placent ce virus proche d'un gemycircularvirus aviaire, ce qui soulève l'hypothèse d'une transmission zoonotique directe ou par l'intermédiaire d'un vecteur.

Article n°3: Identification of a novel gemycircularvirus associated with a fatal case of child encephalitis

Maxime Gaudin[1a], Sebastien Halary[1a], Ilia Stavroula[2], Kelly Goldlust[1], Sonia Monteil-Bouchard[1], Caroline Michelle[1], Catherine Robert[1], Didier Raoult[1], George Briassoulis[2], Emmanouil Angelakis[*1,3] and Christelle Desnues[*1]

[1]Aix-Marseille Université, IRD 198, CNRS FRE2013, Assistance-Publique des Hôpitaux de Marseille, Microbes, Evolution, Phylogeny and Infections (MEPHI), IHU Méditerranée Infection, Marseille France.

[2]Department of Pediatrics, Laboratory of Molecular Medicine and Human Genetics, and Department of Internal Medicine, University of Crete, Heraklion, Greece.

[3]Laboratory of Medical Microbiology, Hellenic Pasteur Institute, Athens, Greece.

[a] These authors have contributed equally to the work

*Corresponding authors

Christelle Desnues, Ph.D. and Emmanouil Angelakis, MD, Ph.D

MEPHI, IHU Méditerranée Infection,

19-21 Boulevard Jean Moulin, 13005 Marseille, France

Phone: (+33) 4 13 73 24 24/ (+33) 4 13 73

Email: christelle.desnues@univ-amu.fr

e.angelakis@hotmail.com

- **Statut : soumission prévue dans Virology**

ABSTRACT

Viruses with small circular rep-encoding ssDNA (CRESS-DNA) genomes were discovered from wide range of eukaryotic organisms ranging from fungi to mammals. The genomes of a novel CRESS-DNA virus, belonging to the gemycircularvirus genus (human brain-associated gemycircularvirus, HBa-GmV) was characterized by metagenomics in a post-mortem cerebral biopsy of a $3^{1/2}$ years-old child who died from an unexplained acute encephalitis. We obtained the reliable and complete circular genome of 2,134 nucleotides encoding three major proteins, including a capsid protein (CAP) and two replication-associated initiation proteins (REP1 and REP2). Phylogenetic analysis based on the amino acid sequence of concatenated REP proteins showed that HBa-GmV is closely related to a gemycircularvirus characterized from *Poecile atricapillus* feces, suggesting a transmission from bird to human. Further studies are required to confirm that gemycircularvirus infect humans and could be responsible of severe encephalitis.

HIGHLIGHTS

- A new gemycircularvirus was detected in a fatal case of encephalitis in a child.
- Gemycircularviruses might represent emerging human pathogens, possibly from an animal reservoir.
- Origin, cellular target and mechanisms of infection remain to be explored.

INTRODUCTION

Acute encephalitis (AE) is defined as an inflammation of the brain parenchyma with consecutive altered level of consciousness and/or epileptic convulsions [1]. In 2008, the global burden of encephalitis has been estimated to 500,000 people worldwide with a high mortality rate (up to 30%) [2] and morbidity as survivors usually suffer from cognitive or physical disability [3]. Despite hundreds of pathogens already associated with encephalitis (the majority of these being viruses) [4], specific etiologies are only characterized in <50% of cases [5, 6]. Traditional pathogen detection methods for suspected infectious cases have fundamental limitations because they do not permit the identification of new, or unexpected pathogens [7]. Recently, with the ability to sequence and catalogue all nucleic acids present in a sample, metagenomics have proved to be a powerful tool for the detection of viruses in clinical sample [8–11] including cerebrospinal fluid (CSF) or brain tissue [12, 13]. However, these methods may suffer from limitations due to the high abundance of sequences that align to the human genome after processing brain or CSF clinical specimens [14–17].

Circular replication initiation protein (Rep)–encoding single-stranded DNA (ssDNA) (CRESS-DNA) viruses possess small non-enveloped icosahedral capsid and are characterized by the presence in their genome of conserved Rep-encoding genes [18]. In 2018, the International Committee on Taxonomy of Viruses (ICTV) have classified the CRESS-DNA viruses into six families, namely *Circoviridae*, *Genomoviridae*, *Geminiviridae*, *Nanoviridae*, *Bacilladnaviridae* and *Smacoviridae* [19]. They represent the smallest genomes of autonomously replicating eukaryotic viruses and infect a wide variety of invertebrates and vertebrates in different natural and human-made environments. *Genomoviridae*, a recently described family of CRESS-DNA viruses [20] contains only one representative isolate; the *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1 (SsHADV-1), which infects

fungi [21]. All the other members are genomic sequences described from metagenomic datasets derived from a wide range of natural (air, sewage, seawater) and host-associated environments such as arthropods or animal feces [22, 23]. Genomovirus genomes have also been detected in human samples, both from healthy individuals [24, 25] and patients suffering from various pathologies including multiple sclerosis (in brain and serum samples), encephalitis (in cerebrospinal fluid), pericarditis (in pericardial fluid), and HIV (in blood samples) [26–28].

Here we report the detection of a novel CRESS-DNA virus of the *Genomoviridae* family and *gemycircularvirus* species (HBa-GmV) in the brain tissue of a young patient who died of encephalitis. For this, we have used an innovative strategy to increase the ratio of genomic viral-to-host signal from a brain specimen. Host contamination was depleted by in-solution capture of human nucleic acids using biotinylated RNA-baits [29] and processed as well as unprocessed samples were then submitted to DNA shotgun next-generation-sequencing (NGS).

CLINICAL CASE

On April 9, 2016, a female toddler, $3^{1/2}$ years old, was admitted to the Pediatric Intensive care Unit (PICU) of the University Hospital of Heraklion (Crete, Greece) with a clinical presentation compatible with severe encephalitis. According to her parents, the current illness started 2 days ago, with fever, truncal maculopapular rash and nasal congestion, while 12 hours prior to admission the child presented multiple episodes of seizures and gradual loss of consciousness. Due to Status Epilepticus she was intubated and transferred to PICU. A wide spectrum intravenous anti-infectious treatment against viral and bacterial agents was induced including acyclovir, oseltamivir, vibramycin, ceftriaxone and azithromycin. An urgent brain magnetic resonance imaging (MRI) was performed, which revealed global cerebral edema. Patient undergone emergency cranial decompression, and put on sedation, with no improvement of intracranial hypertension. A detailed panel of blood and urine laboratory tests for viruses, bacterial infections, metabolic diseases, immunosuppression, autoimmune diseases and toxicology tests did not manage to confirm a diagnosis for her critical condition. The only positive results were positive adenovirus antibodies (IgA and IgG) in blood and positive PCR for adenovirus in a pharyngeal sample. The child was on continuous monitoring with no improvement despite the aggressive treatment and finally died on the 2^{nd} day of her hospitalization in the PICU. One hour after death, a lumbar puncture was performed and cerebrospinal fluid was sent for microbiological and virology laboratory testing. All of them returned negative. Finally, a small part of brain tissue from biopsy was kept in -80°C.

METHODS

Samples collection and storage

On May 24, 2016, the post-mortem cerebral biopsy was sent in dry ice to the Mediterranean Infection Institute, Marseille France. This sample (*sample ID: BCE-01*) was kept at -80°C before processing. Informed consent for pursuing investigations on this case has been obtained from legal guardians.

Samples processing

One ml of phosphate-buffered saline (PBS; 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na2HPO4, and 1.4mM KH2PO4 [pH 7.3]) was added to the cerebral biopsy (*ID: BCE-01)* and homogenized using TissueLyser (Qiagen), with 3mm tungsten beads (Qiagen) according to the manufacturer instructions (3 cycles at 30m/s for 60s). Debris was pelleted by two successive centrifugations for 10 min at 1,500g and then 15 min at 10,000g. Pellet (*sample ID: pelBCE-01*) was frozen at -80°C whereas supernatant was divided into two 200 µl aliquots for further processing.

Nucleic acid extraction and quantification

DNA extraction was performed for both supernatant aliquots (*sample ID: sntBCE-01*) and 40mg of the cerebral biopsy pellet (*sample ID: pelBCE-01*) with the High Pure Viral Nucleic Acid kit (Roche Diagnostics, Meylan, France), according to the manufacturer's protocol. Nucleic acids were eluted in 50µl of elution buffer, aliquoted and stored at -20°C. DNA concentration was estimated with the QuantiFluor® dsDNA System (Promega, Charbonnières-les-Bains, France) according to the manufacturer's recommendations, and fluorescence was quantified with the Tecan GENios fluorometer.

Capture of human nucleic acids using human RNA-baits

200ng of DNA extracted from brain biopsy sample (*ID: sntBCE-01)* was sheared with Covaris S2 to generate fragments of approximately 1,500 bp in size. Depletion of host nucleic acid was performed by hybridization with a biotinylated human RNA-bait library and capture with Dynabeads MyOne Streptavidin C1 beads as previously described [29]. The unbound fraction (supernatant) corresponding to the depleted fraction was concentrated and cleaned using 1.8× AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol with elution into 40 µl of 1X TE buffer (10 mM Tris-Cl, pH 7.5. 1 mM EDTA).

DNA amplification and sequencing

Nucleic acids recovered from the human-depleted fractions (post-capture) or directly after nucleic acid extraction (pre-capture) were amplified with GenomiPhi (GE Healthcare) in duplicate to generate sufficient material for Illumina library preparations. The resulting amplified nucleic acids were purified with silica columns (Qiagen). DNA from these two metagenome samples (*sample ID: preBCE-01* and *postBCE-01*) was sequenced on a MiSeq platform using a paired-end strategy according to a Nextera XT library kit in a 2 × 250 bp format (Illumina Inc., San Diego, CA, USA).

Reads pre-processing, annotation and analysis

Reads were quality trimmed, filtered and assembled as described in [28]. The amount of human DNA contamination was estimated by mapping the trimmed reads onto the human genome with Deconseq (http://deconseq.sourceforge.net). Non-human assembled sequences (contigs) were aligned against the NCBI non-redundant protein database (http://www.ncbi.nlm.nih.gov/refseq/) using BlastX (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with E-value threshold <1E-05.

Genome finishing and diagnostic of a novel gemycricularvirus

To obtain and confirm the complete nucleotide sequence of the novel gemycircularvirus, primers were designed using Primer3 [30] based on the sequence obtained by sequencing (**Table 1**).

Table 1: Primers and pairs used in this study for HBa-GmV detection and genome finishing

Primer name	Sequence (5′ → 3′)	Primer pairs	Amplicon size	Used for
GcV-F1	GTCAAAACCACGCGTCACTC	GcV-F1	200 bp	Detection
GcV-R1	CGTCGACTGGGACGATACAG	GcV-R1		
GcV-F2	GACTGCCAAGAGAAGGACCC	GcV-F1	1663 bp	Genome finishing
GcV-R2	TTCTGGACTTGGATTGGGGC	GcV-R2		
GcV-F3	ACATTCGGGTGGTAACCGTC	GcV-F2	686 bp	Genome finishing
GcV-R3	CCCGAATTGGACCCCTACAG	GcV-R1		
GcV-F4	TCATCCCCAACAGTTCCACG	GcV-F3, GcV-R3, GcV-F4	In combination for genome finishing	

PCR were performed on the *postBCE-01* sample using the following specific primers: GcV-F1/GcV-R2 and GcV-F2/GcV-R1. For this, 2 µL of DNA template were added to a final volume of 25 µl containing 12.5 µl of AmpliTaq Gold® 360 Master Mix (Applied Biosystems) and 0.2 µM of each couple of primer describe previously. Amplification started with an initial denaturation step at 95°C for 10 min, followed by 40 cycles at 95°C for 30 seconds, at 60°C for 30 seconds, and at 72°C for 1 min/kb. Sequencing reactions were performed with the reagents of the ABI Prism dye terminator cycle sequencing ready reaction kit (Perkin Elmer Applied Biosystems, Foster City,CA) according to the manufacturer's instructions.

Genome annotation/Phylogeny

Complete genome was obtained by assembling Sanger reads using ChromasPro 2 (Technelysium.com.au). Genome sequence is deposited in the GenBank database under the accession number XXXX. Gene numbers and positions were determined by ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/), and the translated sequences were annotated by alignment against the non-redundant (nr) protein database with BlastP (https://blast.ncbi.nlm.nih.gov/Blast.cgi), using an E-value threshold < 1E-05. Gemycircularvirus characteristic hairpin structure and its associated nanonucleotide were characterized using MFold software (http://unafold.rna.albany.edu). Finally, genome-wide pairwise identities including all 121 available genomes described in [31] were calculated using SDT v1.2 [32].

Phylogeny

Concatenated REP protein sequences were compared to *nr* to grab the 50 most similar REP cluster using BlastP. All these sequences were aligned together using MAFFT (with local alignment option), and conserved sub-sequences found using GBlocks [33] were used for further analyses. The best substitution model (substitution matrix and rates), as well as phylogenetic inference were then calculated using PhyML [34, 35]. Phylogenetic tree visualization was performed using iTol (https://itol.embl.de/).

RESULTS

Metagenomic analysis and identification of a novel gemycircularvirus

Between 0.6 and 1.9 million single-reads were recovered after Illumina MiSeq sequencing of the whole DNA from the cerebral biopsy, with between 0.59 and 1.8 million

reads passing through the trimming and quality filtering steps. Before capture, the proportion of non-human reads was low (1.3%) whereas after capture, it reached 21%, which corresponded to non-human sequences enrichment of 16.1-fold (**Table 2**).

Table 2. Summary of the sequencing data before and after capture

Sample	pre- or post-capture	Number raw reads	Number trimmed reads (R1)	Number non-human reads	% of assigned non-human reads
Brain biopsy *(BCE-01)*	pre	1,850,978	1,773,603	23,076	1.3
	post	596,908	587,154	125,000	21

Non-human reads were assembled into contigs and annotated by BlastX against the GenBank nr database. Two CRESS-DNA virus contigs that were absent in the precapture dataset were identified in the post-capture dataset. The first one, a contig of 1,112 bp with an average coverage of 245X (recruiting 1,145 reads) shared 99% identity (100% coverage) with the capsid gene of a circovirus detected in marine isopods [36]. The second was a contig of 2,115 bp with an average coverage of 1,099X (recruiting 18,365 read). This contig displayed identities with replication initiation protein (82% identity) and capsid protein (39% identity) sequences of gemycircularviruses and was then named accordingly human brain-associated gemycircularvirus or HBa-GmV. A reliable and complete circular genome of 2,134bp was obtained after amplification using the GcV-F1/GcV-R2 and the GcV-F2/GcV-R1 primer pairs (**Table 1**) on the post capture metagenome sample (*postBCE-01*) and Sanger sequencing of amplicons. The HBa-GmV genome size and architecture are similar to those of already known *Genomoviridae* genomes, *i.e.* a 2,134bp circular DNA molecule carrying a capsid gene (*cap*) on 1 strand and 2 genes on the opposite strand, *rep1* and *rep2*, respectively involved in replication initiation and termination (**Figure 1**).

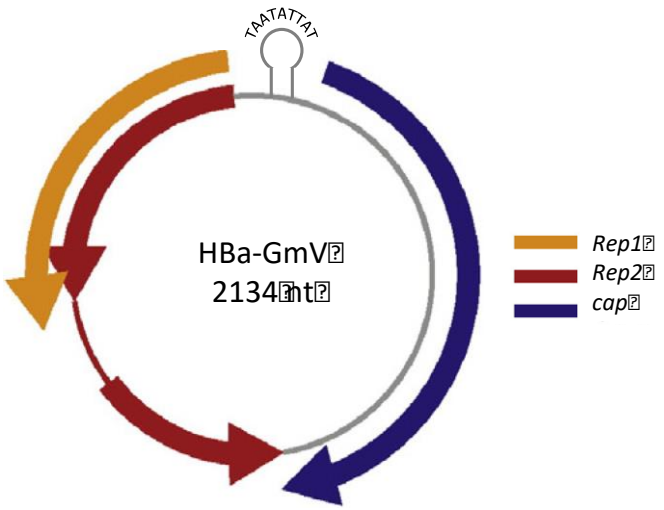


Figure 1: Genomic features of novel gemycircularvirus (HBa-GmV) including hairpin structure and predicted open reading frames. Cap, capsid; Rep, replication initiation protein.

Protein sequence similarity analysis showed a strong identity between each protein sequence and gemycirculavirus proteins from different viruses: Capsid (39% with mongoose feces-associated gemycircularvirus d), REP1 (88% with *Poecile atricapillus* GI tract-associated gemycircularvirus) and REP2 (80% with Pacific flying fox faeces associated gemycircularvirus-10), highlighting the high divergence and the high mosaicism within this virus family. The non-coding sequence between Rep1 and Cap displayed the hairpin structure harboring the gemycircularvirus typical nanonucleotide motif, TAATGTTAT (**Figure 1**). Phylogenetic inference calculated from concatenated Rep proteins confirmed that this virus belongs to the gemycircularviruses genus, with a virus characterized in bird *Poecile atricapillus* feces as closest neighbour (**Figure 2**).

Figure 2: Maximum likelihood phylogenetic inference of gemycircularvirus replication initiation proteins.

Genome-wide comparisons against 121 complete genomes indicate that this genome shares less than 64% identity with other members of the *Genomoviridae* family (**Figure 3**). HBa-GmV is thus a new species within the gemycircularvirus genus according to the recent <78% identity threshold defined for genomovirus species demarcation [31].

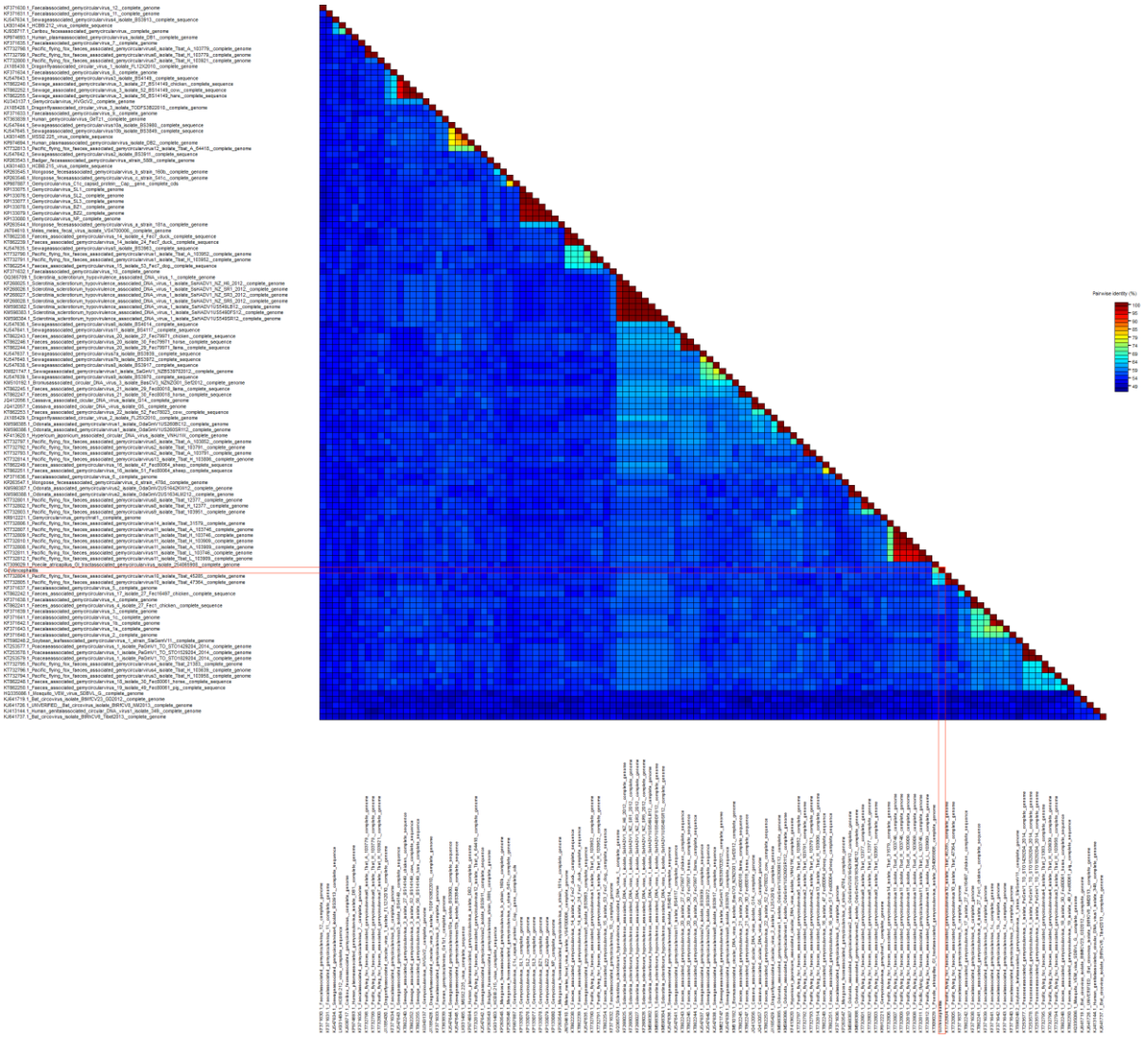


Figure 3: Genome-wide pairwise identity (122 taxa) of genomoviruses. HBa-GmV shares less than 78% of identy with other genomes and is thus a new gemicularvirus species (red frame).

PCR assay using primers GcV-F1/GcV-R1 amplified a 200-bp amplicon in metagenomic sample (*postBCE-01*, **Figure 4A**) and original samples (*sntBCE-01 and pelBCE-*

01, **Figure 4B and 4C**), whereas each PBS controls (PBS that underwent the same experimental process) and PCR controls remained negatives. The specificity of 200bp DNA band was confirmed by sequencing the PCR products, confirming the presence of this novel gemycircularvirus in the brain tissue of the patient.

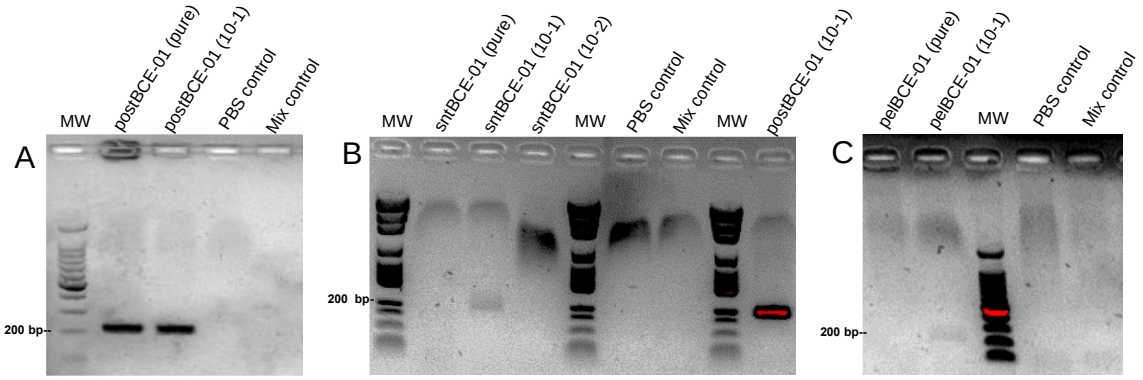


Figure 4: detection of HBa-GmV in the (A) post-capture sample (postBCE-01) that was NGS sequenced and, in the original extracted nucleic acids from (B) the supernatant (sntBCE-01) and (C) the pellet (pelBCE-01) after homogenization of the cerebral biopsy. PBS controls were subjected to the same experimental process. MW, molecular weight.

DISCUSSION

Genomoviridae family was first described in 2016 [20] Based on whole genome alignment of 121 NGS discovered genomoviruses and phylogenetic analysis using REP sequences, *Varsani et al*. have recently proposed 9 new genera within the *Genomoviridae* family: Gemycircularvirus (n = 73), Gemyduguivirus (n = 1), Gemygorvirus (n = 9), Gemykibivirus (n = 29), Gemykolovirus (n = 3), Gemykrogvirus (n = 3), Gemykroznavirus (n = 1), Gemytondvirus (n = 1), and Gemyvongvirus (n = 1) [31].

Here, we identified a novel gemycircularvirus in a cerebral biopsy from a pediatric case of unexplained fatal encephalitis. The HBa-GmV genome architecture exhibits the genomic

features of all *Genomoviridae* family members, with a small circular sequence of 2,134 bp, carrying two genes encoding Replication initiation (Rep1) and termination (Rep2) on the negative strand, and the gene coding for the capsid protein (Cap) on the positive strand, as well as a hairpin loop in the non-coding region. This structure exposes the classical gemycircularvirus nanonucleotide motif TAATATTAT, whose cleavage by REP1 is thought to initiate the genome replication [37].

Compared with the other known genomoviruses, HBa-GmV CAP shared 38 % sequence identity with Mongoose feces-associated gemycircularvirus based on the complete amino acid sequence of capsid. This strong divergence is common between *Genomiviridae* CAP sequences, and would be explained by selection pressures occurring on this gene to escape the host antibodies neutralization of virus infectivity and to promote interactions with plasma membrane components [38]. In contrast, the concatenated REP1-REP2 sequence was highly conserved, exhibiting 82% identity with the REP protein of a gemycircularvirus identified in gastrointestinal tract of *Poecile atricapillus* (Black-capped Chickadee). Phylogenetic analysis of predicted Rep protein confirmed these similarities and clustered HBa-GmV sequence in the same clade that these both REP proteins, with *Poecile atricapillus* feces gemycircularvirus as closest neighbour.

The characterization of a novel *Genomoviridae* genome in the human brain, a site expected to be sterile, supports the possibility of replication of these small viruses in human. Moreover, as no other human pathogen sequences were identified in this dataset, a potential role for this virus in the patient symptoms could be hypothesized. These findings are consistent with two other previous studies reporting the presence of novel gemycircularviruses in cerebrospinal fluid from patients with non-fatal unexplained cases of encephalitis [27, 39]. In the *Zhou et al.* study, a new gemycircularvirus (GeTz1) was detected in the CSF of a <6 years-old encephalitic child in China. Sequence analysis of the REP protein demonstrated a close

similarity with a replication-associated protein of a gemycircularvirus identified from bird feces [27]. These findings, in addition to our phylogenetic results and the fact that the patient lived in a rural environment and was in daily contact with backyard poultry and wild birds, raise the hypothesis of a direct or vector-borne zoonotic viral transmission from birds to human. However, gemycircularviruses have also been detected in blood samples of asymptomatic patients [24, 25] suggesting that they could also be part of the human commensal viral flora.

In this study, we identified a novel gemycircularvirus from a cerebral tissue biopsy of a child who died from severe acute encephalitis. This finding adds clues to the identification of potential new pathogen occurring in human encephalitis and underlies a possible zoonotic transmission from bird to human. However, reports confirming the association between gemycircularvirus and a human pathology are still lacking despite the growing number of *Genomoviridae* detected in human samples. Further studies aiming to explore the association between gemycircularviruses and diseases of humans and animals, such as Western blot experiment using recombinant viral proteins or a Fluorescent *In Situ* Hybridization assays (FISH) on human tissues, are needed.

BIBLIOGRAPHY

1. **Venkatesan A, Tunkel AR, Bloch KC, et al** (2013) Case Definitions, Diagnostic Algorithms, and Priorities in Encephalitis: Consensus Statement of the International Encephalitis Consortium. Clin Infect Dis Off Publ Infect Dis Soc Am 57:1114–1128 . doi: 10.1093/cid/cit458
2. **Jmor F, Emsley HCA, Fischer M, et al** (2008) The incidence of acute encephalitis syndrome in Western industrialised and tropical countries. Virol J 5:134 . doi: 10.1186/1743-422X-5-134
3. **Venkatesan A** (2015) Epidemiology and outcomes of acute encephalitis. Curr Opin Neurol 28:277–282 . doi: 10.1097/WCO.0000000000000199
4. **Granerod J, Tam CC, Crowcroft NS, et al** (2010) Challenge of the unknown. A systematic review of acute encephalitis in non-outbreak situations. Neurology 75:924–932 . doi: 10.1212/WNL.0b013e3181f11d65

5. **Khetsuriani N, Holman RC, Anderson LJ** (2002) Burden of encephalitis-associated hospitalizations in the United States, 1988-1997. Clin Infect Dis Off Publ Infect Dis Soc Am 35:175–182 . doi: 10.1086/341301
6. **Glaser CA, Honarmand S, Anderson LJ, et al** (2006) Beyond viruses: clinical profiles and etiologies associated with encephalitis. Clin Infect Dis Off Publ Infect Dis Soc Am 43:1565–1577 . doi: 10.1086/509330
7. **Miller RR, Montoya V, Gardy JL, et al** (2013) Metagenomics for pathogen detection in public health. Genome Med 5:81 . doi: 10.1186/gm485
8. **Finkbeiner SR, Li Y, Ruone S, et al** (2009) Identification of a Novel Astrovirus (Astrovirus VA1) Associated with an Outbreak of Acute Gastroenteritis. J Virol 83:10836–10839 . doi: 10.1128/JVI.00998-09
9. **Grard G, Fair JN, Lee D, et al** (2012) A Novel Rhabdovirus Associated with Acute Hemorrhagic Fever in Central Africa. PLOS Pathog 8:e1002924 . doi: 10.1371/journal.ppat.1002924
10. **Rascovan N, Monteil Bouchard S, Grob J-J, et al** (2016) Human Polyomavirus-6 Infecting Lymph Nodes of a Patient With an Angiolymphoid Hyperplasia With Eosinophilia or Kimura Disease. Clin Infect Dis 62:1419–1421 . doi: 10.1093/cid/ciw135
11. **Tsuzuki S, Fukumoto H, Mine S, et al** (2014) Detection of trichodysplasia spinulosa-associated polyomavirus in a fatal case of myocarditis in a seven-month-old girl. Int J Clin Exp Pathol 7:5308–5312
12. **Brown JR, Bharucha T, Breuer J** (2018) Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. J Infect 76:225–240 . doi: 10.1016/j.jinf.2017.12.014
13. **Piantadosi A, Kanjilal S, Ganesh V, et al** (2018) Rapid Detection of Powassan Virus in a Patient With Encephalitis by Metagenomic Sequencing. Clin Infect Dis Off Publ Infect Dis Soc Am 66:789–792 . doi: 10.1093/cid/cix792
14. **Jovel J, O'keefe S, Patterson J, et al** (2017) Cerebrospinal Fluid in a Small Cohort of Patients with Multiple Sclerosis Was Generally Free of Microbial DNA. Front Cell Infect Microbiol 6: . doi: 10.3389/fcimb.2016.00198
15. **Kawada J-I, Okuno Y, Torii Y, et al** (2016) Identification of Viruses in Cases of Pediatric Acute Encephalitis and Encephalopathy Using Next-Generation Sequencing. Sci Rep 6:33452 . doi: 10.1038/srep33452
16. **Brown JR, Morfopoulou S, Hubb J, et al** (2015) Astrovirus VA1/HMO-C: An Increasingly Recognized Neurotropic Pathogen in Immunocompromised Patients. Clin Infect Dis Off Publ Infect Dis Soc Am 60:881–888 . doi: 10.1093/cid/ciu940
17. **Morfopoulou S, Brown JR, Davies EG, et al** (2016) Human Coronavirus OC43 Associated with Fatal Encephalitis. N Engl J Med 375:497–498 . doi: 10.1056/NEJMc1509458
18. **Rosario K, Dayaram A, Marinov M, et al** (2012) Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). J Gen Virol 93:2668–2681 . doi: 10.1099/vir.0.045948-0
19. **Varsani A, Krupovic M** (2018) Smacoviridae: a new family of animal-associated single-stranded DNA viruses. Arch Virol 163:2005–2015 . doi: 10.1007/s00705-018-3820-z
20. **Krupovic M, Ghabrial SA, Jiang D, Varsani A** (2016) Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch Virol 161:2633–2643 . doi: 10.1007/s00705-016-2943-3
21. **Yu X, Li B, Fu Y, et al** (2010) A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc Natl Acad Sci U S A 107:8387–8392 . doi: 10.1073/pnas.0913535107
22. **Kraberger S, Hofstetter RW, Potter KA, et al** (2018) Genomoviruses associated with mountain and western pine beetles. Virus Res 256:17–20 . doi: 10.1016/j.virusres.2018.07.019
23. **Waits K, Edwards MJ, Cobb IN, et al** (2018) Identification of an anellovirus and genomoviruses in ixodid ticks. Virus Genes 54:155–159 . doi: 10.1007/s11262-017-1520-5

24. **Moustafa A, Xie C, Kirkness E, et al** (2017) The blood DNA virome in 8,000 humans. PLOS Pathog 13:e1006292 . doi: 10.1371/journal.ppat.1006292
25. **Zhang W, Li L, Deng X, et al** (2016) Viral nucleic acids in human plasma pools. Transfusion (Paris) 56:2248–2255 . doi: 10.1111/trf.13692
26. **Lamberto I, Gunst K, Müller H, et al** (2014) Mycovirus-Like DNA Virus Sequences from Cattle Serum and Human Brain and Serum Samples from Multiple Sclerosis Patients. Genome Announc 2: . doi: 10.1128/genomeA.00848-14
27. **Zhou C, Zhang S, Gong Q, Hao A** (2015) A novel gemycircularvirus in an unexplained case of child encephalitis. Virol J 12: . doi: 10.1186/s12985-015-0431-0
28. **Halary S, Duraisamy R, Fancello L, et al** (2016) Novel Single-Stranded DNA Circular Viruses in Pericardial Fluid of Patient with Recurrent Pericarditis. Emerg Infect Dis 22:1839–1841 . doi: 10.3201/eid2210.160052
29. **Gaudin M, Monteil-Bouchard S, Michelle C, et al** Application of an inverted human Whole-genome In-Solution Capture (inv-WISC) to viral metagenomics (submitted**)**
30. **Untergasser A, Cutcutache I, Koressaar T, et al** (2012) Primer3--new capabilities and interfaces. Nucleic Acids Res 40:e115 . doi: 10.1093/nar/gks596
31. **Varsani A, Krupovic M** (2017) Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evol 3: . doi: 10.1093/ve/vew037
32. **Muhire BM, Varsani A, Martin DP** (2014) SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PloS One 9:e108277 . doi: 10.1371/journal.pone.0108277
33. **Castresana J** (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552 . doi: 10.1093/oxfordjournals.molbev.a026334
34. **Guindon S, Dufayard J-F, Lefort V, et al** (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321 . doi: 10.1093/sysbio/syq010
35. **Lefort V, Longueville J-E, Gascuel O** (2017) SMS: Smart Model Selection in PhyML. Mol Biol Evol 34:2422–2424 . doi: 10.1093/molbev/msx149
36. **Bistolas KSI, Besemer RM, Rudstam LG, Hewson I** (2017) Distribution and Inferred Evolutionary Characteristics of a Chimeric ssDNA Virus Associated with Intertidal Marine Isopods. Viruses 9: . doi: 10.3390/v9120361
37. **Heyraud-Nitschke F, Schumacher S, Laufs J, et al** (1995) Determination of the origin cleavage and joining domain of geminivirus Rep proteins. Nucleic Acids Res 23:910–916
38. **Kolawole AO, Li M, Xia C, et al** (2014) Flexibility in Surface-Exposed Loops in a Virus Capsid Mediates Escape from Antibody Neutralization. J Virol 88:4543–4557 . doi: 10.1128/JVI.03685-13
39. **Phan TG, Mori D, Deng X, et al** (2015) Small viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. Virology 482:98–104 . doi: 10.1016/j.virol.2015.03.011

Résultats complémentaires au travail présenté

L'article précédent présente la détection et la caractérisation d'un nouveau gemycircularvirus à partir d'une biopsie cérébrale d'une patiente décédée d'une encéphalite infectieuse foudroyante. Pour confirmer le lien entre la présence de ce virus et la pathologie nous avons procédé à différentes analyses complémentaires comme la production de protéines recombinantes virales servant d'antigènes pour des tests de sérologie par Western blot, l'identification de la localisation cellulaire sur des coupes histologiques de tissu cérébral de la patiente par hybridation *in-situ* en fluorescence utilisant des sondes virales ou enfin des tentatives d'isolement sur lignées cellulaires permissives et par inoculation intracérébrale de souriceaux nouveau-nés. Les résultats préliminaires obtenus aux cours de ces tests sont présentés ci-dessous.

Protéines recombinantes virales et sérologies par Western blot:

Nous avons dans un premier temps produit des protéines virales recombinantes après transformation de la souche électrocompétente BL21 (DE3) d'*Escherichia coli* avec un plasmide d'expression pET-22b contenant des séquences partielles des gènes codant les protéines CAP (20kDa), REP1 (26kDa) ou REP2 (17kDa) du HBa-GmV. Après culture, induction, lyse et ultracentrifugation, l'efficacité de la production a été vérifiée par électrophorèse sur gel SDS 12,5% bis-acrylamide. Les résultats ont montré que les 3 protéines étaient insolubles et se retrouvaient dans le culot du lysat bactérien.

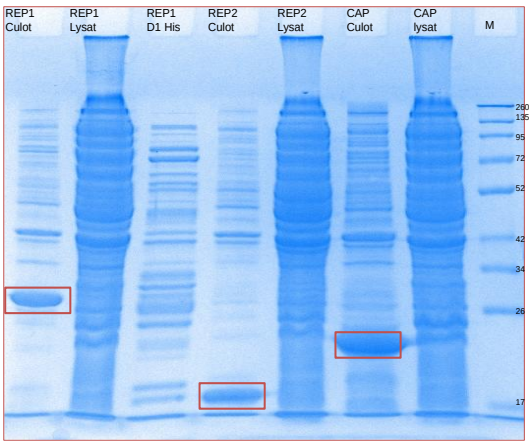


Figure : *SDS-Page (coloration au bleu de Comassie) des protéines isolées à partir de différentes fractions après production dans un hôte bactérien. Les protéines recombinantes d'intérêt CAP, REP1 et REP2 (encadré rouge) sont insolubles et sont identifiées par leur poids moléculaire dans le culot du lysat bactérien. Ces protéines ont été confirmées par MALDI-TOF.*

L'identification de ces protéines a été confirmée par MALDI-TOF (MALDI-TOF-MS Microflex LT, Bruker) et les protéines présentes dans le culot ont été solubilisées par sonication et centrifugation dans du tampon de lyse TS contenant 8M Urée, 2M thiourée et 4% CHAPS. Afin de détecter la présence d'anticorps anti-HBa-GmV dans le sérum de la patiente, une sérologie par western blot a été réalisé comme décrit précédemment [129]. Pour cela, 2 ou 6µg de protéines du culot bactérien contenant CAP, REP1 ou REP2 ont été déposé

sur gel et incubé avec une dilution au 1/1000ème du sérum de la patiente ou de trois nourissons (contrôles négatifs) comme anticorps primaire. Des anti-IgG ou anti-IgM humaines conjuguées à la péroxydase ont ensuite été utilisées comme anticorps secondaires. Les résultats ont montré un signal non spécifique des anticorps primaires et secondaires sur le lysat protéique. Une mise au point du protocole est en cours à travers une production des protéines virales à partir de la séquence génomique virale complète et une amélioration si nécessaire des étapes de solubilisation [130] et de capture des protéines virales d'intérêt.

Hybridation *in situ* en fluorescence (FISH) sur coupes histologiques de tissu cérébral

Nous avons ensuite tenter de visualiser la localisation et le tropisme cellulaire du virus sur des coupes histologiques de biopsie cérébrale par hybridation *in situ* en fluorescence (FISH). La sonde nucléique (1000 pb) a été obtenue après une amplification des acides nucléiques par PCR et un marquage à la biotine. Le protocole de FISH a été appliqué à des coupes de tissu cérébral de la patiente et de cinq autres biopsies cérébrales contrôles (contrôles négatifs sans pathologie infectieuse) comme décrit précédemment [55]. Un marquage cytosolique a été observé pour certaines cellules du tissu cérébral de la patiente mais également, dans une moindre mesure, chez 2 des 5 contrôles. Pour améliorer la spécificité, un marquage multiple utilisant un jeu de plusieurs sondes serait envisageable afin d'augmenter le nombre de régions cibles et d'observer d'éventuelles co-localisations. Une combinaison de FISH avec un marquage en immunofluorescence (IF-FISH) pourra être également envisagée, sous réserve d'obtention de protéines antigéniques en quantité et qualité suffisante pour une production d'anticorps chez la souris ou le lapin.

Tentatives d'isolement

Nous avons enfin tenté d'isoler ce virus sur différentes lignées cellulaires permissives (VERO, HeLa, HEK293) et parallèlement réalisé des inoculations intracérébrales de souriceaux nouveau-nés à partir de particules virales purifiées de la biopsie clinique. Les essais sur lignées cellulaires n'ont pas montré d'effet cytopathique tandis qu'aux 1er et 2ème passages (chacun à J15), aucun symptôme ni décès de souriceau n'a été constaté. Les qPCR réalisées sur les cerveaux, rates et sang des souriceaux n'ont pas permis de constater de réplication virale. Ces tests seront reproduits pour confirmer ces résultats.

Préambule à l'article 4 "An Enigmatic *Moraxella osloensis* Endocarditis Diagnosed by Laser-Capture Micro-Dissection and Human RNA Bait-Depletion"

L'endocardite est définie comme une inflammation de l'endocarde, le plus souvent d'origine infectieuse et plus rarement d'origine inflammatoire ou néoplasique. L'endocardite infectieuse (EI) résulte d'une inflammation de l'endocarde d'origine microbienne qui peut concerner : l'endothélium cardiaque valvulaire ou non, les prothèses valvulaires ainsi que tout autre matériel prosthétique intracardiaque [131]. C'est une maladie rare avec une incidence annuelle de 1.5 à 11.6 cas pour 100000 habitants [132] et grave avec une mortalité à un an de 20 à 30 %, voire 70% selon le micro-organisme responsable [133, 134]. Les staphylocoques (*Staphylococcus aureus* principalement), streptocoques et entérocoques sont responsables de 80% de tous les cas d'endocardites infectieuses. Cependant d'autres pathogènes ont été mis en évidence dans cette maladie incluant le groupe des bactéries Gram négatifs HACEK (*Haemophilus spp*., *Actinobacillus actinomycetemcomitans*, *Cardiobacterium hominis*, *Eikenella corrodens* and *Kingella kingae*), *Coxiella burnetii*, *Bartonella spp*, le genre Chlamydia mais également certains champignons et virus [134]. L'EI est une pathologie de diagnostic difficile, dont les signes d'appels peuvent être variés. Cependant, une fois le diagnostic évoqué, il devra être confirmé grâce à la réalisation d'examens complémentaires reposant sur la clinique, la biologie et les examens paracliniques (échographie, radiographie, électrocardiogramme, IRM cardiaque). Les hémocultures représentent le gold standard dans le diagnostic des EI avec généralement seulement 5 à 7% des hémocultures qui s'avèrent stériles (EIHN, endocardites infectieuses à hémocultures négatives) [134]. Cependant, cette proportion peut monter jusqu'à 70% dans certaines séries [135]. Pour s'affranchir de ces limitations, les médecins disposent d'autres techniques incluant la culture de tissus/valves, les examens histologiques, l'immunohistochimie (IHC), la microscopie électronique, la sérologie et la PCR [135]. Malgré le peu d'études sur le sujet, la métagénomique shotgun pourrait être utile pour compléter l'arsenal diagnostique des cas d'EIHN. Elle a ainsi déjà permis de détecter *Abiotrophia defectiva* et *Streptococcus sanguinis* comment agents pathogènes chez deux patients pour lesquelles les hémocultures étaient restées stériles [136, 137]. Dans ces études, les données issues du séquençage ont étés obtenues après extraction des acides nucléiques directement à partir de valves cardiaques. Cependant, les données issues du séquençage ont

souligné la forte contamination en séquence de l'hôte avec 95,9% à 99% des séquences générées qui s'alignaient contre le génome humain [137].

L'article suivant décrit un cas clinique d'EI pour lequel un signal anti-*Coxiella burnetii* a été mis en évidence par immunohistochimie sur valve mitrale tandis que la culture, la PCR et la sérologie étaient toutes négatives. Pour élucider ce mystère et définir s'il s'agissait ici du premier cas d'endocardite à *C. burnetii* avec une sérologie négative ou d'un faux positif, l'objectif consistait à identifier sans aucun *a priori* quel pathogène était à l'origine du signal positif détecté en IHC. A partir de cette valve, une microdissection laser a été réalisée afin d'isoler les amas de cellules infectées (zone avec un fort signal positif en IHC) des cellules non infectés (zone avec peu/pas de signal). Les acides nucléiques de ces deux zones ont été extraits et séquencés sur une plateforme Illumina MiSeq. L'analyse des données de séquençage n'a pas mis en évidence de séquence correspondant à *Coxiella burnetii* dans les deux échantillons. Ces résultats étaient en adéquation avec les résultats négatifs de la PCR. Le signal positif en IHC pourrait ainsi résulter d'une réaction croisée des anticorps anti *C. burnetii* avec un autre pathogène. Les résultats de la zone positive ont permis d'identifier qu'une bactérie du genre *Moraxella* constituait une piste prometteuse. Cependant, le faible nombre de séquences non humaines (1,5%) et plus particulièrement de celles correspondant au genre *Moraxella* (0,11%) limitait l'assignation au niveau de l'espèce. Le protocole d'hybridation et de capture en solution des acides nucléique humains que nous avons développé dans le **Chapitre II** a donc été appliqué sur les acides nucléiques extraits des deux zones (signal positif et négatif). Cela a permis d'améliorer considérablement proportion de séquences non humaines (72.3%) de l'échantillon positif et une identification sans équivoque d'une nouvelle souche de *Moraxella osloensis.* Un génome quasi complet de 2.4 Mbp avec une couverture moyenne de 234 fois a pu être reconstruit permettant une étude approfondie de son contenu génomique incluant des gènes de résistances aux antibiotiques. Les acides nucléiques de *M. osloensis* souche Marseille ont été détectés par PCR dans l'échantillon original tandis que la présence d'anticorps dirigés contre cet antigène a été mis en évidence dans le sérum de la patiente par Western blot. En conclusion la combinaison d'approches innovantes d'enrichissement a permis d'élucider un cas énigmatique d'endocardite infectieuse dû à une nouvelle souche de *Moraxella osloensis* chez cette patiente.

Article n°4: An Enigmatic *Moraxella osloensis* Endocarditis Diagnosed by Laser-Capture Micro-Dissection and Human RNA Bait-Depletion

Matthieu Million[1a], Maxime Gaudin[1a], Cléa Menelotte[1], Lionel Chasson[2], Messica Zeitoun[3], Elsa Prudent[1], Sophie Edouard[1], Bernard Amphoux[1], Elsa Prudent[1], Mathieu Fallet[2], Fred Cadoret[1], Stéphane Meresse[2], Julien Paganini[4], Hubert Lepidi[1], Caroline Michelle[1], Catherine Robert[5], Bernard LaScola[1], Philippe Naquet[2], Jean-Pierre Gorvel[2], Christelle Desnues[1] and Didier Raoult[1*]

[1]Aix-Marseille Université, IRD 198, CNRS FRE2013, Assistance-Publique des Hôpitaux de Marseille, UMR Microbes, Evolution, Phylogeny and Infections (MEPHI), IHU Méditerranée Infection, Marseille France.

[2]Centre d'Immunologie Marseille Luminy, 171 Avenue de Luminy, 13009 Marseille

[3] Department of Cardiology, Assistance Publique - Hôpitaux de Paris (AP-HP), Bichat Hospital, Paris, France

[4]Xegen, Gemenos, France

[5]Aix-Marseille Université, IRD 257, Service de Santé des Armées, Assistance-Publique des Hôpitaux de Marseille, UMR Vecteurs, Infections Tropicales et Méditerranéennes (VITROME), IHU Méditerranée Infection, Marseille France.

[a] These authors have contributed equally to the work

*Corresponding author

Didier Raoult, M.D., Ph.D.,

MEPHI, IHU Méditerranée Infection,

19-21 Boulevard Jean Moulin, 13005 Marseille, France

Phone: (+33) 4 13 73 24 24/ (+33) 4 13 73

Email: didier.raoult@gmail.com

- **Statut : Soumission prévue dans Plos Medicine**

INTRODUCTION

Etiological diagnosis of infectious diseases is one of the keys to therapeutic adaptation and improved prognosis, particularly for infections such as endocarditis. Endocarditis with negative culture can represent up to 70% of endocarditis cases depending on the series (1). In our center, we have developed a multimodal strategy that includes testing blood and valves when available using serology, molecular diagnostics with broad spectrum PCR, specific PCR and anatomopathology with immunohistochemistry (IHC) (2). This allowed us to find an etiology in 78% of cases (1). In our last study, we showed that molecular diagnosis improved diagnostic efficiency by 24% mainly by detecting enterococci and streptococci that had not been detected by other diagnostic methods (1). Thus, while these bacteria have the reputation of being easily cultivable, only molecular biology allows their identification in a certain number of cases.

Here, as an expert center in the diagnosis of blood culture negative endocarditis (BCNE) (3), we were confronted with a case of endocarditis where the anti-*Coxiella burnetii* IHC was positive (2) while our comprehensive syndromic approach was negative including culture, serology and PCR for *Coxiella burnetii*. As the French National Referral center for Q Fever with a 30-year experience, we recently reported a fiew cases of *Coxiella burnetii* endocarditis with low serological titers (4, 5). However, among 533 endocarditis (the largest series to date) (6), no definite case (positive culture and/or anatomopathology and/or PCR on the valve) of *Coxiella burnetii* endocarditis was diagnosed with a negative serology (6).

This enigmatic case prompts us to either confirm the possible first case of *Coxiella burnetii* endocarditis with negative serology or to identify an IHC false positive by searching which other microbe could be identified on the valve. For this, we developed an innovative strategy combining 2 different enrichment technologies to increase the ratio of bacterial-to-host signal of an IHC-positive mitral valve specimen. We first used laser micro-dissection (7)

on IHC positive and negative sections that were further submitted to DNA shotgun next-generation-sequencing (NGS) before and after in-solution capture of human nucleic acids using biotinylated RNA-baits. We were then able to diagnose a rare case of endocarditis due to a new strain of *Moraxella osloensis* for which we were able to reconstruct a nearly complete genome with a >200X average coverage. We subsequently confirmed the infection of the patient by Western-Blot and specific PCR. We also proved the cross-reactivity by immunostaining with the anti-*Coxiella burnetii* antibody produced in rabbit on a *Moraxella osloensis* strain from our collection.

RESULTS

CLINICAL CASE

A 37-year-old patient of Guinean origin, with a history of aortic valve biological prosthesis on rheumatic aortic insufficiency presented in May 2012 a degeneration of her prosthesis. She underwent surgery in May 2012 for a mechanical mitral and aortic valve replacement with tricuspid annuloplasty. She presented 2 blood culture negative endocarditis in December 2012 and June 2013. Despite broad spectrum antibiotic treatments, she required a new replacement of the aortic and mitral valves by a biological prosthesis and aortic homograft in July 2013. Surgical samples including mechanical aortic valve, aortic ring, mechanical mitral valve, mitral vegetation and electrode alongside sera were sent to our center as a worldwide expert center for BCNE (1, 3) (**Supplementary Table 1**). Non-specific histological examination of the mitral valve reported a highly inflammatory prosthetic valve tissue with neutrophils and macrophages. Macrophages were particularly visible within the valve tissue while the threads were rather localized on the surface with a small vegetation on the surface.

SYNDROMIC ETIOLOGICAL ASSESSEMENT

A comprehensive etiological assessment on surgical samples included PCR targeting the following pathogens (and specific methods): bacteria (real time short 16S rRNA gene and standard 16S rRNA gene PCRs), eukaryotes (18S rRNA gene, CU), *C. burnetii* (IS111, IS30a), *Coxiella* like, *Rickettsia, Bartonella, Staphylococcus aureus, Escherichia coli, Mycoplasma, Streptococcus oralis, Streptococcus gallolyticus, Enterococcus faecalis, Enterococcus faecium, Tropheryma whipplei, Mycobacterium*, and *Brucella*. Several culture conditions for intracellular bacteria (L929, HEL, MRC5 cells and amoeba) and mycobacteria were performed, also including inoculation on mice spleen. Histology was realized using Giemsa, Grocott, PAS, Warthin Starry and Ziehl staining. On sera, all serologies performed systematically for BCNE in our center were negative including that for *Coxiella burnetii, Rickettsia, Orientia tsutsugamushi* and *Bartonella*. IgG anticardiolipin and anti-pork antibodies were also assessed (8). All this comprehensive diagnostic strategy was negative (**Supplementary Table 1**).

IMMUNOHISTOCHEMLISTRY AND CONFOCAL SPECTRAL IMAGING

Unexpectedly, the immunohistochemistry on the valve against *Coxiella burnetii* antigens was positive, retested twice and confirmed by fluorescent *in situ* hybridization specific of *Coxiella burnetii* (**Figure 1**). Both techniques identified an intracellular cytoplasmic staining.

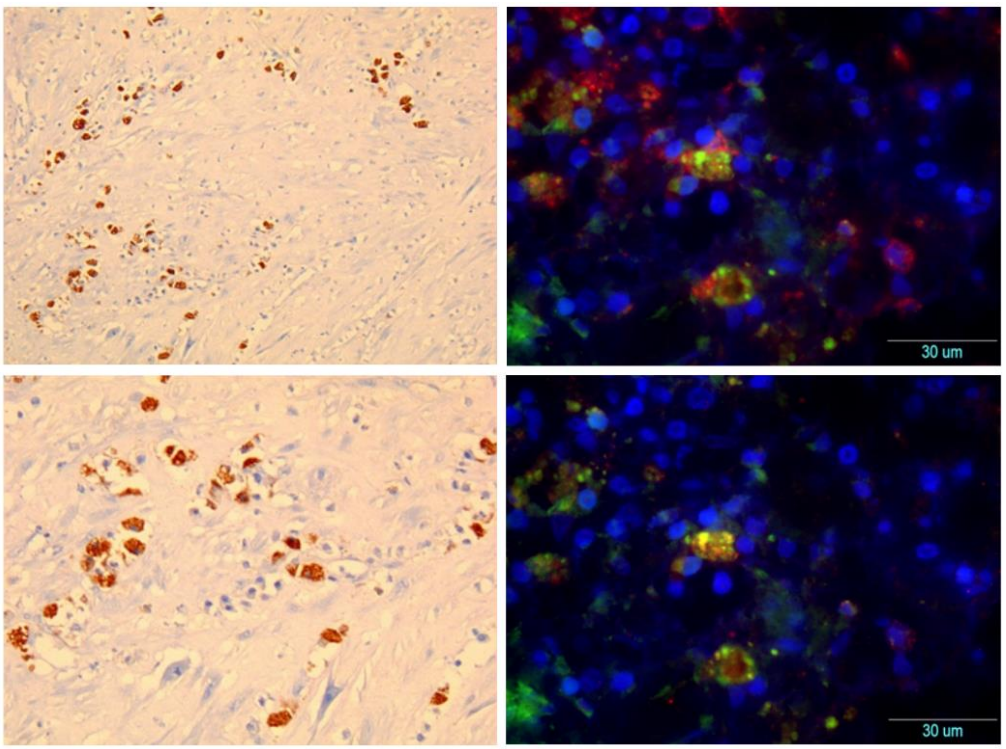


Figure 1. Positive immunohistochemistry and FISH against *Coxiella burnetii*

Left: immunohistochemistry against *Coxiella burnetii* (up: x100, bottom: x200). Right: fluorescent in situ hybridization with a probe designed to target *Coxiella burnetii*. Red: Eub338 probe, green: *C. burnetii* specific RNA probe (9), yellow: colocalization of the two signals. Note the intracellular cytoplasmic localization of the signals.

We improved the immunofluorescence signal by using confocal spectral microscopy that eliminates autofluorescence by determining autofluorescence spectra at 405 and 488nm (DAPI was used to stain nuclei, **Supplementary Figure 1**). Positive immunofluorescence was confirmed (**Figure 2**). Surprisingly, the bacteria were distributed very punctually, always isolated, without ever forming a dense zone of bacteria. This aspect contrasted with aspects usually reported for bacteria usually causing endocarditis where bacterial clusters are regularly observed (**Supplementary Figure 2**).

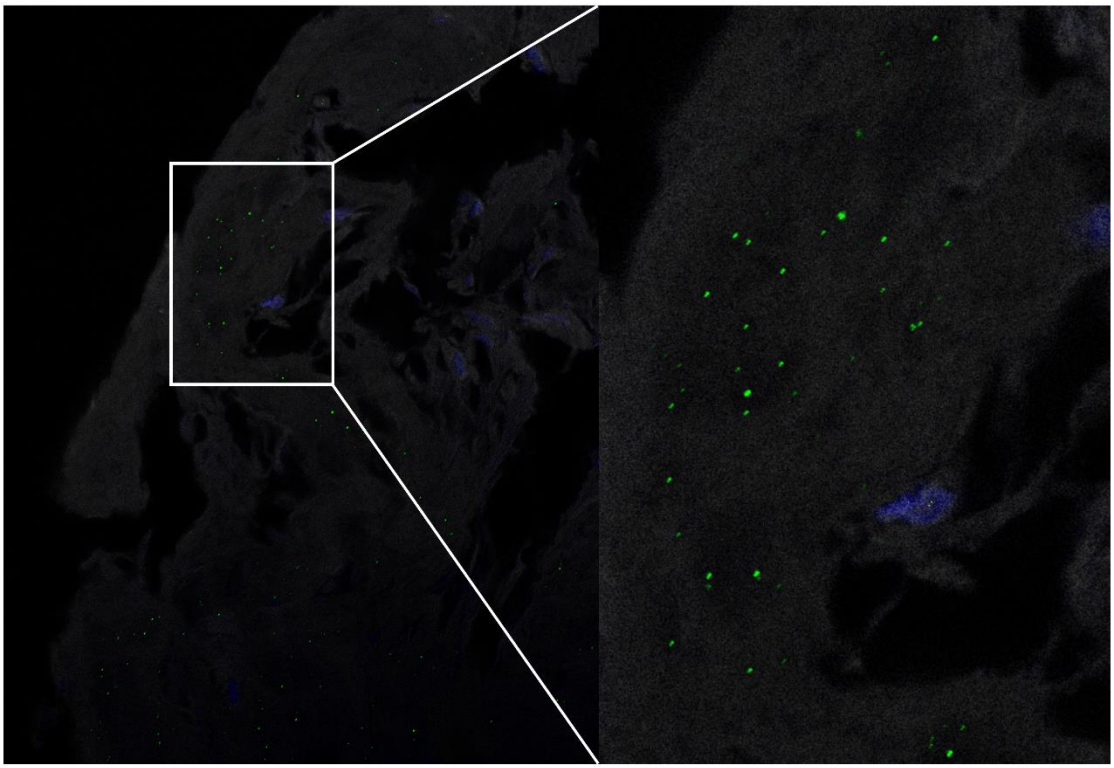


Figure 2. Confocal spectral imaging evidence bacteria distributed punctually

Green: Alexa Fluor 488nm, blue; DAPI. The autofluoresnce spectrum for 405 and 488nm have been subtracted (Supplementary Figure 1). Note the unusual punctuated distribution of bacteria without clusters, opposite to what can be observed for more usual bacteria causing endocarditis as Staphylococcus or Streptococcus (Supplementary Figure 2). Z-stack with orthogonal projection.

However, as a national reference center for *Coxiella burnetii* and with 30 years of experience of Q fever endocarditis (DR) (6, 10), we never diagnosed a definite *Coxiella burnetii* endocarditis in a patient with negative specific serology (among more than 533 Q fever endocarditis cases) (6). This precipitated us to obtain a definite diagnosis by an innovative and non-targeted diagnostic metagenomic approach after laser microdissection of immune-positive areas.

LASER CAPTURE MICRODISSECTION

Most of the 488nm signal corresponded to autofluorescence, so the positive areas were identified by confocal spectral immunofluorescence in order to exclude areas with non-specific autofluorescence signal (**Figure 3**). Several stained areas were cut out and placed in a positive (T+) tube. To control the major problem of contaminants in diagnostic metagenomics, we also cut two surface rounds in the unstained area (**Figure 3**) and placed in a negative tube (T-). These two tubes were transmitted for metagenomics.

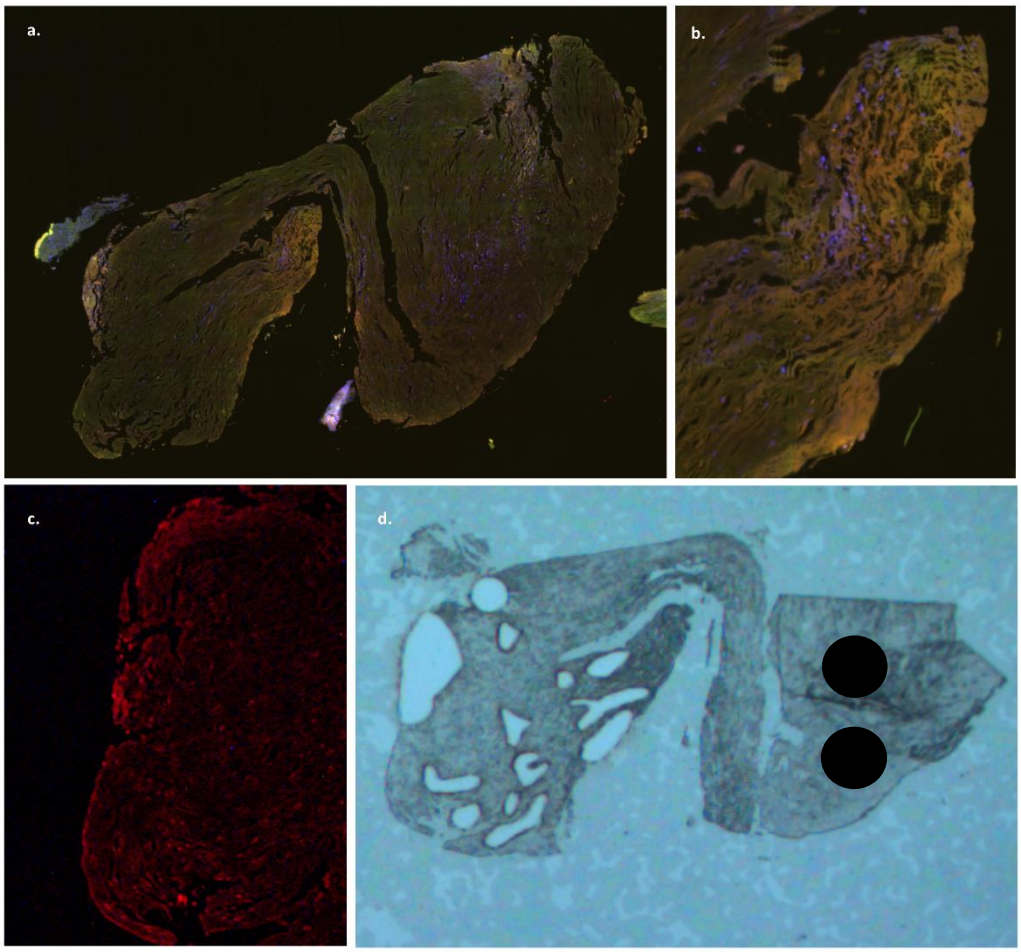


Figure 3. Laser capture microdissection

The areas stained with anti-*Coxiella burnetii* antibody were identified by immunofluorescence. a. x40 slide scanner of the whole mitral valve sample (vegetation), b. stained area corresponding to autofluorescence, no specific signal was found in this area by confocal spectral imaging, c. specific staining confirmed by spectral confocal imaging. d. sample after laser microdissection of the stained areas. Black rounds corresponded to negative stained area.

HUMAN RNA BAIT-DEPLETION AND METAGENOMICS

Sequencing of the valve samples

Between 0.9 and 1.7 million single-reads were recovered after Illumina MiSeq sequencing of the IHC-positive and negative valve samples before and after human capture (**Table 1**). For the positive section before capture, the proportion of non-human reads that were taxonomically assigned was only 1.5%.

Identification of a putative specific species of the sample by z-score

We compared these 2 metagenomic datasets to those of 12 other heart valves processed identically but without capture (data not shown), Based on read annotation, we observed that the most abundant species in our sample (the fungus *Malassezia* and the bacterium *Cutibacterium acnes*, formerly called *Propionibacterium acnes*) were skin commensals (11, 12) and were also the dominant species in the previously analyzed valves (**supplementary Figure 3**). In order to control this metagenomic noise consisting mainly of contaminants, we devised an approach to identify if a species was specifically enriched in our sample. For this we calculated the z-score for each detected species (**supplementary Figure 4**). We were able to observe that *Moraxella caprae* clearly stood out from the 355 other species detected with at least 1 reads with a z-score of +82 standard deviations in the positive tube (**supplementary Figure 5**), whereas this was not the case in the negative tube (**supplementary Figure 6**). In addition, the first 3 species in descending order z-scores were *Moraxella*. These results encouraged us to more accurately identify whether we could identify a *Moraxella* species in a robust manner by using human DNA capture to further improve the signal-to-noise ratio.

Enrichment of non-human reads by human RNA-bait depletion

After capture, the proportion of non-human reads that were taxonomically assigned increased from 1.5% to 72.3% (**Table 1**). Among these assigned reads, 15.1 and 17.1%

belonged to the *Moraxella* genus before and after capture, respectively. For the negative section the proportion of taxonomically assigned non-human reads before capture was higher (27.8) and was not improved by the human capture process. Only 2.1 and 0.8% of these reads were assigned to the *Moraxella* genus before and after capture, respectively.

Table 1. Summary of the sequencing data before and after capture.

Sample	pre- or post-capture	Number trimmed reads (R1)	Number of assigned[a] non-human reads	% of assigned non-human reads	Number of reads assigned[a] at the *Moraxella* genus level	% reads assigned at the *Moraxella* genus level among all reads[b]
Positive section	pre	1,059,153	16,206	1.5	2,443	15.1
	post	1,742,732	1,260,424	72.3	215,528	17.1
Negative section	pre	967,349	268,996	27.8	5,571	2.1
	post	1,182,231	189,877	16.1	1,572	0.8

Number of R1-reads recovered after trimming and annotation by BLASTx with DIAMOND against GenBank nr and assignation using MEGAN for the positive and negative sections before and after capture of human nucleic acids

Identification of a new strain of *Moraxella osloensis* in the sample

We next focused on identifying the *Moraxella* strain detected in the positive sample after human capture. The paired-reads were trimmed and assembled into 10,076 contigs of size ranging from 89 bp to 38,393 bp (1,259 bp average). From these contigs, 3,061 ORFs were detected and among these ORFs, 2,558 had *M. osloensis* within the 10 BBH recovered. The paired-reads were then mapped onto the chromosome and 4 plasmids of *M. osloensis* KSH reference genome before and after capture of the host nucleic acids. Before capture, only 2,312 paired-reads (0.11%) mapped onto the *M. osloensis* genome and plamids whereas after capture, it reached 2,495,844 paired-reads (71.6% of the reads) (**Supplementary Table 2 and 3**). The capture process does not affect the ratio of single reads assigned to *Moraxella* (15.1 pre-capture vs. 17.1 post-capture) compared to the total number of assigned-reads but rather significantly increase its absolute number (2,443 pre-capture vs. 215,528 post-capture). A genome consensus sequence of 2.4 Mbp (2,348,787 bp, 187 contigs) with a GC content of 44.2% was obtained with an average coverage of 234X (**Supplementary Table 4**). Alignment with the *M.*

osloensis KSH reference genome showed syntheny conservation (**Supplementary Figure 7 & 8**) with gaps corresponding to the 4 copies of the 16-23S rRNA genes that generated conflicts in the mapping process (**Supplementary Figure 9**). 16S rRNA sequences were retrieved using the very sensitive mapping option implemented within Bowtie 2. A consensus sequence that shares 99 to 100% identity to the 4 copies of 16S rRNA gene of *Moraxella osloensis* KSH was replaced in a phylogenetic tree (**Figure 4**). This new strain of *Moraxella osloensis* was named *M. osloensis* strain Marseille (**Figure 4**). The partial genome of the *M. osloensis* strain Marseille was annotated using the RAST server and displayed 2,167 protein-encoding genes (PEGs) and 64 RNAs.

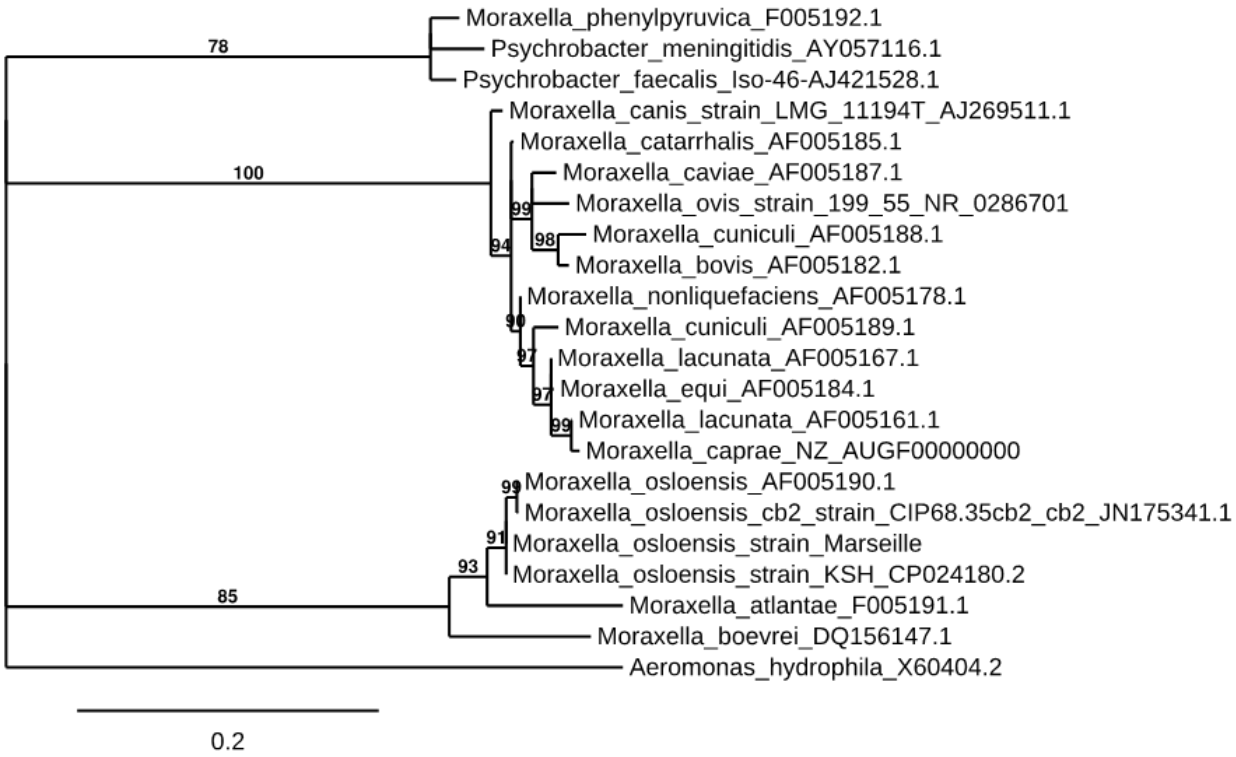


Figure 4. Phylogenetic tree of the *Moraxella* 16S rRNA genes

Sequences were aligned with MUSCLE (v3.8.31) configured for highest accuracy ambiguous regions (i.e. containing gaps and/or poorly aligned) were removed with Gblocks (v0.91b). The phylogenetic tree was reconstructed on 1348 residus using the maximum likelihood method implemented in the PhyML program (v3.1/3.0 aLRT). The HKY85 substitution model was selected assuming an estimated proportion of invariant sites (of 0.621) and 4 gamma-distributed rate categories to account for rate heterogeneity across sites. The gamma shape parameter was estimated directly from the data (gamma=0.375). Reliability for internal branch was assessed using the aLRT test (SH-Like). Branches with support values <50% were collapsed.

Antibiotic resistance genes (ARG) identified in the *Moraxella osloensis* strain Marseille genome

The 2,167 PEGs were compared to the ARG-ANNOT database (release May 2018) by BlastP (E-value < 1E-10). Among the 89 PEGs showing significant similarities with the selected cut-off, 10 hits presented very low E-values < 1E-50 and high bit-scores >200 (**Supplementary Table 5**). These hits belong to the fluoroquinolone (n=2), colistin (n=1), beta-lactamase (n=3), tetracyclin (n=1) and macrolide-lincosamide-streptogramin (n=3) ARG groups. PEG_1719 (1,087 aa) and PEG_1308 (1,053 aa) presented similarities with the OqxAB genes that codes for multidrug efflux pump found on the pOLA52 plasmid of *Escherichia coli* and the chromosome of *Klebsiella pneumoniae* clinical isolates. These 2 proteins were confirmed as efflux pumps of the resistance-nodulation-division (RND) AcrA/B, which confers resistance to multiple classes of antibiotics including the beta-lactams, quinolones, and aminoglycosides. Peg_1235 (559 aa) shares 65% identity with the mcr-6.1 of Moraxella sp. strains and the mcr-1/2 gene product of *Escherichia coli*. Mcr-1/2 gene product is responsible of colistin (polymyxin E) resistance and has been recently described as ICR-Mo in *Moraxella osloensis* (13). Phylogenetic analysis showed that MCR-1/2 and ICR-Mo protein sequences formed 2 separate clusters with the PEG_1235 of *M. osloensis* strain Marseille branching within the group of *M. osloensis* ICR-Mo (**Figure 5**). PEG-1036, PEG-253 and PEG-1986 were identified as penicillin-binding proteins in the ARG-ANNOT database with PEG-1036 (690aa) showing a highest identity score with OXA-372, a novel carbapenem-hydrolysing class D beta-lactamase recently described from *Citrobacter freundii* (14). However, BlastP against NCBI nr database did not confirm this result. Tetracycline resistance determinant (TetB) was also detected as previously described in *M. catarrhalis* (15). Finally, 3 PEGs presented similarities with the *Enterobacter faecalis* plasmid-borne ABC transporter gene optrA that confers resistance to oxazolidinones and phenicols (16).

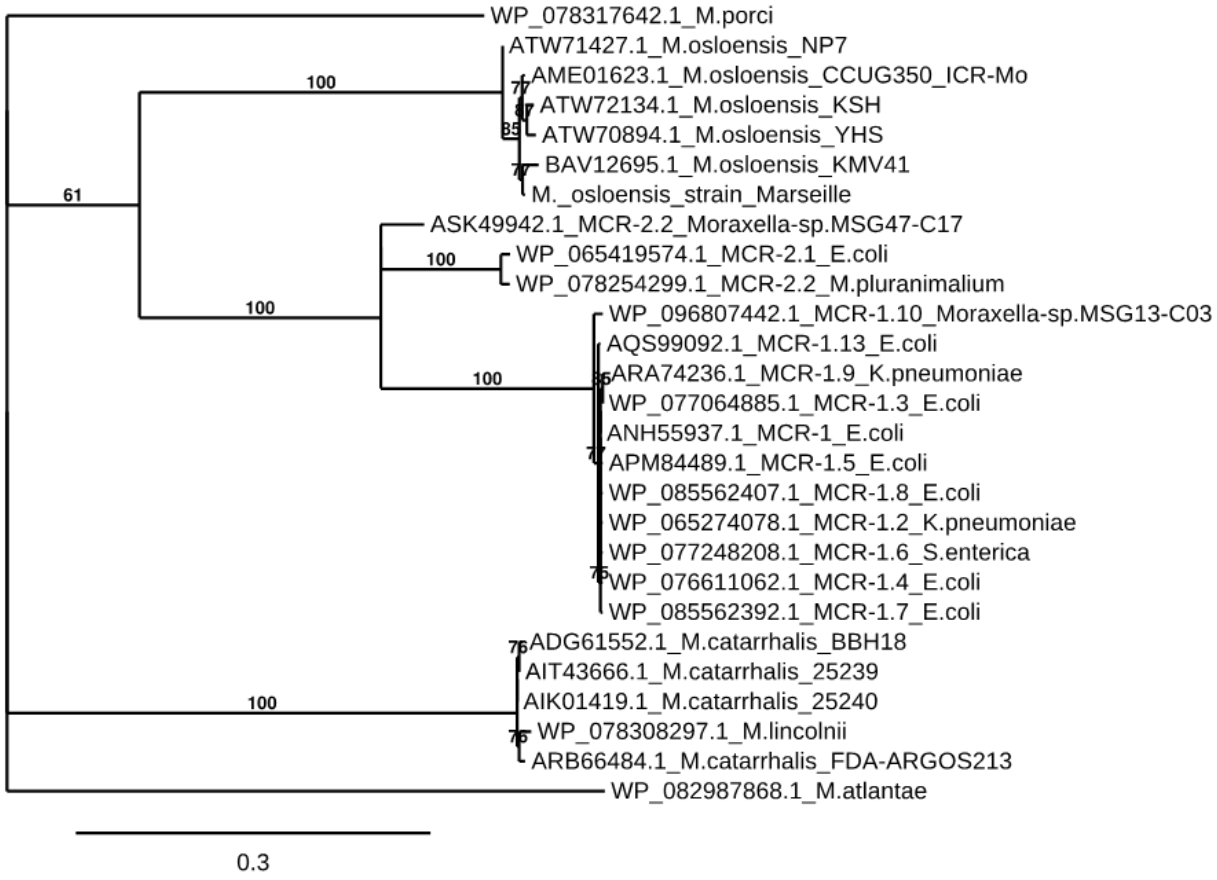


Figure 5. Phylogenetic tree of the MCR-1/2 genes.

Sequences were aligned with MUSCLE (v3.8.31) configured for highest accuracy. The phylogenetic tree was reconstructed using the maximum likelihood method implemented in the PhyML program (v3.1/3.0 aLRT). The WAG substitution model was selected assuming an estimated proportion of invariant sites (of 0.213) and 4 gamma-distributed rate categories to account for rate heterogeneity across sites. The gamma shape parameter was estimated directly from the data (gamma=1.732). Reliability for internal branch was assessed using the aLRT test (SH-Like). Branches with support values <50% were collapsed.

WESTERN BLOTS

In order to confirm the patient's infection by *Moraxella osloensis* at the time of cardiac surgery, we performed western blots with three sera of the patient and several *Moraxella* strains corresponding to closely related species; *Moraxella osloensis, M. bovoculi, M. caprae*, but also a *Coxiella burnetii* strain. No bands were observed with *Coxiella burnetii* antigens confirming false positive immunohistochemistry and FISH (**Figure 6**). On the other hand, at least one band was observed for all *Moraxella* strains and the western blot was the most positive for *Moraxella*

osloensis, especially on the June 2013 serum, i.e. the serum closer to the surgery corresponding to the mitral valve vegetation sample analyzed here (**Figure 6**).

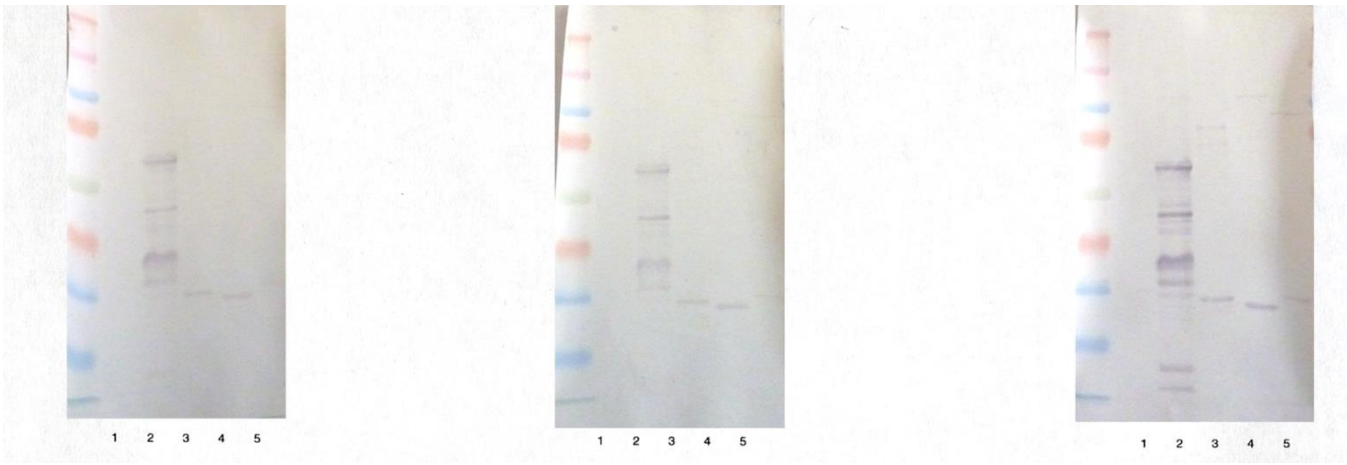


Figure 6. Western blot

Lane 1: *Coxiella burnetii*, lane 2 : *Moraxella osloensis*, lane 3 : *Moraxella bovoculi*, lane 4 : *Moraxella caprae*, lane 5 : *Moraxella lacunata*. Serum dilution 1/100. Left: plasma of the 27/11/2012, center: plasma of the 12/06/2013, right: serum of the 23/06/2013

IMMUNOFLUORESCENCE ON MORAXELLA OSLOENSIS STRAIN

To confirm the false positive of immunohistochemistry and immunofluorescence, we tested whether the primary antibody used produced in rabbits by *Coxiella burnetii* infection had a cross reaction with *Moraxella osloensis*. After culturing *Moraxella osloensis* strain CSURP3830, we were able to detect antibody staining used for anti-*Coxiella burnetii* immunohistochemistry and thus confirmed the cross reaction (**Figure 7**).

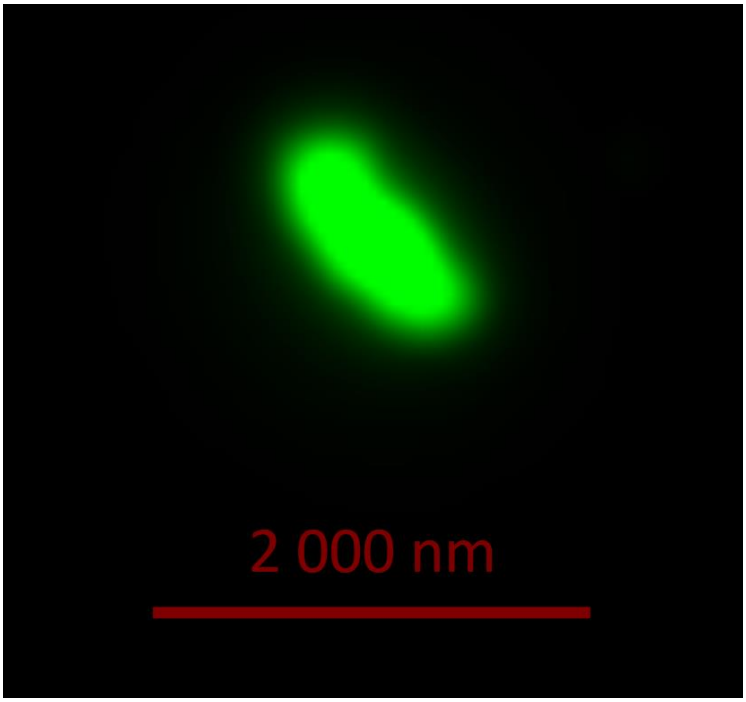


Figure 7. Immunostaining of *Moraxella osloensis* strain with anti-*Coxiella burnetii* antibody

Bacterial strain: *Moraxella osloensis* CSURP3830, primary antibody: antibody used for immunohistochemistry against *Coxiella burnetii*, secondary antibody: anti-rabbit AF488nm. A video is provided as supplementary data with the complete z-stack (Supplementary video 1).

PCR VALIDATION

For a highly sensitive and specific detection of *M. osloensis*, specific primers targeting a region that is conserved and repeated in the *M. osloensis* CCUG 350 reference genome were designed. These primers (M_oslo_transpF AAATGCGAGAACGCAGGTTG and M_oslo_transpR CCTTTCGGACTATTGGCGGT) amplify a 101 bp fragment of a gene that has 23 genomic copies and is coding for a transposase. A faint positive signal was detected in the original microdissected valve sample (**Figure 8A**) and confirmed as *M. osloensis* by Sanger sequencing. To exclude any contamination during the microdissection process, another tissue section, for which nucleic acids were directly extracted, was also tested and returned positive for *M. osloensis* (**Figure 8B**).

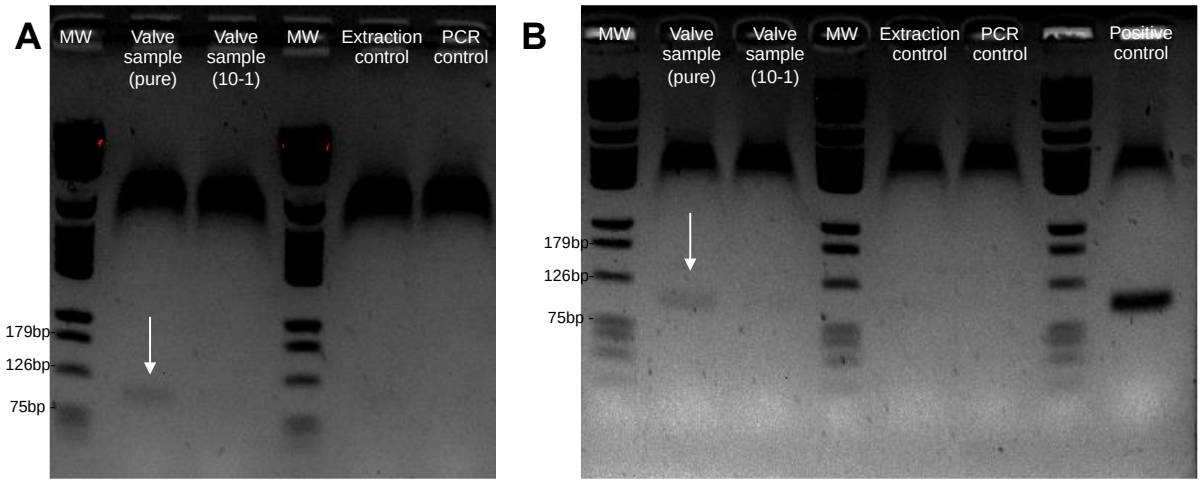


Figure 8. *M. osloensis* specific PCR on valve samples

(A) PCR on the nucleic acids (pure and 1/10 dilution) extracted from the microdissection positive area, an extraction negative control and a PCR negative control. (B) PCR on the nucleic acids (pure and 1/10 dilution) extracted from another valve section, a PBS extraction negative control, a PCR negative control and a PCR positive control (random-amplified nucleic acids extracted from the positive microdissection area). MW: molecular weight.

DISCUSSION

In this study, we combine laser micro-dissection and depletion of human genomic nucleic acids using biotinylated RNA-baits on a valve sample that was positive for *Coxiella burnetii* by immunohistochemistry. This was done to enhance the bacterial-to-human signal to confirm the presence of *C. burnetii* or to identify the aetiological agent of an enigmatic endocarditis case in a young woman. With this approach, we could not detect a single sequence related to *Coxiella burnetii*, thus confirming the negative PCR results, but instead we confidently identified a new strain of *Moraxella osloensis* that was further defined as strain Marseille. The presence of *M. osloensis* strain Marseille was further confirmed by specific PCR on the original valve sample. After capture of the human nucleic acids, a partial consensus genome sequence covering more than 94% of the *M. osloensis* KSH reference genome was reconstructed with a 234X average coverage. Indeed, the approach developed here led to a 593-fold enrichment of the ratio of paired-reads mapping on the *Moraxella osloensis* KSH genome

(66.63% vs 0.13%), although we used a medium-depth Illumina MiSeq shotgun approach and sequencing of a tissue sample, a material from which non-host signal is usually very low.

For the clinics, *M. osloensis* was first described in 1967 (17), and since then, only rare cases of infection, mostly meningitis and bacteriemia, have been reported in the literature (18). So far, only four cases of endocarditis due to *M. osloensis* endocarditis have been described (**Table 2**) both in immunocompromised and immunocompetent patients (19-21), further limiting diagnostic efforts in case of infective endocarditis. Including this study, 4 out of the 5 cases of *M. oslensis* endocarditis were described in the past 5 years reflecting either (or both) better diagnostic tools or the presence of an emerging pathogen.

Table 2. Documented cases of infective endocarditis due to *M. osloensis* since 1967

Country	Patient	Type of endocarditis	Concurrent condition/Clinical history	Detection	Treatment	Evolution	Reference
USA	66-years old male	Prosthetic aortic valve	renal failure	Blood culture	Penicillin, oxacillin, tobramycin	Died	(19)
France	75-years old male	Native aortic valve	B-cell chronic lymphocytic leukaemia	Histopathology, culture of native valve, MALDI-TOF, 16S sequencing	amoxicillin/ gentamicin	Cured	(20)
France	51-years old male	Aortic abcess, prosthetic mitral valve	IE on prothetic mitral valve, Hodgkin's lymphoma and a kidney transplant	Culture of Prosthetic valve, MALDI-TOF, 16S sequencing	Cefotaxime	Cured	(20)
Brazil	41-years old male	Native mitral valve	None	Blood culture	vancomycin, gentamicin, ampicillin, ceftriaxone	Cured	(21)
France	38-years old female	Mechanical aortic and mitral valve	stroke, renal failure	Histopathology	vancomycin, gentamicin, rifampicin, doxycycline	Cardiac transplant	This study

The *M. osloensis* strain Marseille harbors on its chromosome several protein-encoding genes that are related to resistance or reduced susceptibility to fluoroquinolone, tetracyclin,

penicillin, carbapenem and colistin. Most *M. osloensis* strains isolated so far are susceptible to penicillin and cephalosporins although penicillin and streptomycin resistant strains of *M. osloensis* have previously been described (22). In addition, several species from the *Moraxella* genus, such as *M. catarrhalis, M. lacunata* and *M. nonliquefasciens*, are known to produce BRO beta-lactamases (23). A BRO beta-lactamase-producing *M. lacunata* was also isolated from a 15-mont infant with endocarditis (24). In this work, BRO beta-lactamase could not be identified but genes potentially associated with penicillin and carbapenem resistance have been detected *in silico*. Unfortunately, antibiotic resistance spectrum of *M. osloensis* strain Marseille could not be experimentally evaluated, as all cultures (hemoculture and valve culture) returned sterile probably because of the on-going antibiotic treatments of the patient. Colistin (Polymixin E) is a final resort antibiotic against carbapenem-resistant bacteria, and this work identifies the presence of a chromosomally encoded mcr-1/2-like variant showing 99% identity with the one recently described in the genome of *Moraxella osloensis* strain CCUG 350 (ICR-Mo). ICR-Mo of *M. osloensis* strain CCUG350 confers resistance to colistin, although with lower levels than mcr-1, when expressed in *Escherichia coli* (13), which further reinforces the rising idea suggesting that *Moraxella* species represent a potential reservoir of MCR-1/2 (25, 26). Owing to the recent increased detection of *M. oslensis* strains in endocarditis cases along with other pathologies and the underlying risk of transferable resistance to other bacterial species, we propose the systematic search of *M. osloensis* at least in case of blood-negative endocarditis.

MATERIAL AND METHODS

CLINICAL SAMPLE

All clinical samples were obtained from the diagnostic laboratories of the Bichat hospital (Paris, France) and IHU Méditerranée Infection Institute (Marseille, France). The signed consent of the patient was obtained.

MICROBIAL CULTURE

All samples (heparinized blood, mechanical mitral valve, aortic valve ring, mitral valve vegetation) were inoculated for conventional axenic agar culture and cell culture. For axenic culture we inoculated columbia sheep blood agar plates (Biomerieux, Marcy L'etoile, France). Agar plates were incubated in a 5% CO2 incubator at 37°C for 20 days. Inoculated cells were human embryonic lung fibroblasts HEL and human endothelial cells ECV34. Cell culture were done by using the centrifugation-shell-vial technique (Sterilin-Felthan-England, 3.7ml) using 12-mm round coverslips seed with 1 ml of medium containing 50,000 cells and incubated in a 5% CO2 incubator at 37°C for 3 days to obtain confluent monolayer (27). HEL cells were maintained in minimal essential medium MEM (Gibco BRL kife technologies, Cergy pontoise, France) with 10% fetal calf serum and 2mM L-glutamine per liter. ECV34 cells were maintained in RPMI-1640 medium (Gibco) with 15% fetal calf serum and 2mM L-glutamine per liter. Each sample were inoculated onto three shell-vials for each cell support. After inoculation, sheel vials were centrifugated at 700xg/min for 1h at 22°C. Then supernatant of each shell vial was removed and replaced by fresh medium before incubation at 37°C. ECV34 cells were incubated for 5 weeks and HEL for 12 weeks. Detection of growing bacteria were assessed using Gimenez, Gram and Giemsa staining.

IMMUNOHISTOCHEMISTRY (IHC) AND FLUORESCENCE IN SITU HYBRIDIZATION (FISH)

Immunohistochemical detection using a *C. burnettii* monoclonal antibody coupled with an immunoperoxidase was performed as previously described (2). FISH experiments were done as previously described on 3-μm-thick formalin-fixed paraffin-embedded tissue sections (28).

CONFOCAL SPECTRAL IMAGING

Confocal spectral imaging was performed with a Zeiss LSM 780. Reference spectra of individual dyes (DAPI and AF488) and autofluorescence produced by each laser (405, 488) were acquired separately using the dichroic 405 and 488 and the Quasar detector composed of 32-channel PMT in the range of 410 to 695nm. Each channel simultaneously collects a discrete range of optical frequency (~8.9 nm) without the need for multiple emission filters. These references spectrum were then used on the Fingerprinting mode to unmix the autofluorescence contribution.

LASER CAPTURE MICRODISSECTION

The paraffin-embedded tissue section were cut at 3.5mm with the microtome Leica RM2245 and the blade was change for each tissue. For visualizing the areas of interest, the slide was used after labeling 1h with a primary antibody (rabbit anti-*Coxiella burnetii* dilution 1/800) and a secondary antibody (Donkey anti-rabbit antibody Alexa Fluor 488nm, (Jackson ImmunoResearch Laboratories Inc., West Grove, PA, USA) for 45min at room temperature. The slide was mount and scan with panoramic scanner (3DHISTECH Ltd., Budapest, Hungary) and analyzed by confocal (Zeiss LSM780).

For the laser microdissection, a second section was perform and mounted on PEN Membrane Glass Slide (Applied Biosystems/Life Technologies, Carlsbad, CA). The mounted tissue

sections were left to dry overnight at room temperature. The section was dewaxed by two bath of xylene 5 min each, rehydrated and dried at room temperature. One positive and one negative area were captured by laser capture microdissection (LCM) using an Arcturus® XT system (Applied Biosystems/Life Technologies, Carlsbad, CA according to the manufacturer's instructions. A CapSure™ LCM Caps (Applied Biosystems/Life Technologies, Carlsbad, CA) was placed over the target area. Laser pulsing through this cap caused a thermoplastic film to form a thin protrusion that bridged the membrane around the selected area. The membrane around the area was cut using a UV laser, and the cap was lifted to remove the target cell attached to it. The capture areas were directly collected into a sterile 1.5 ml microcentrifuge tubes containing 180 µl of Lysis Buffer T1 (Macherey-Nagel, Duren, Germany) and processed for extraction.

HUMAN RNA BAIT-DEPLETION

The two tissue sections (positive and negative) within the Lysis buffer T1 and a negative control (180 µl of Lysis buffer T1) were digested with proteinase K during 3h. DNA was extracted using the NucleoSpin® Tissue kit (Macherey-Nagel, Duren, Germany) and eluted with 60 µl of elution buffer. DNA concentration was estimated with the QuantiFluor® dsDNA System (Promega, Charbonnières-les-Bains, France) according to the manufacturer's recommendations, and fluorescence was quantified with the Tecan GENios fluorometer.

Capture of human nucleic acids using human RNA-baits

DNA extracted from the two microdissections was sheared with Covaris S2 to generate fragments of approximately 1,500 bp in size. Hybridization reaction with a biotinylated human RNA-bait library and capture with Dynabeads MyOne Streptavidin C1 beads using a magnetic particle collector was performed as previously described (29). The unbound fraction

(supernatant) corresponding to the depleted fraction was concentrated and cleaned using 1.8× AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol with elution into 40 µl of 1X TE buffer.

Random DNA amplification and sequencing

Nucleic acids recovered from the human-depleted fractions were amplified with GenomiPhi (GE Healthcare) in duplicate to generate sufficient material for Illumina library preparations. Nucleic acids were also amplified directly after nucleic acid extraction (before capture) with Genomiphi (GE Healthcare) for positive and negative sections. The resulting amplified nucleic acids were purified with silica columns (Qiagen). DNA was sequenced on a MiSeq platform using a paired-end strategy according to a Nextera XT library kit in a 2 × 300 bp format (Illumina Inc., San Diego, CA, USA).

Reads pre-processing, annotation and analysis

Single (R1) and paired-reads were imported using the CLC Genomics Workbench 7.5 software (CLC Bio, Aarhus, Denmark) with default parameters. Single and paired-reads were first trimmed according to their quality score (Illumina pipeline 1.8 and later) and their length (reads < 50 nt long were discarded). The amount of human DNA contamination was estimated on the single trimmed reads by mapping onto the human genome with Deconseq (30) (http://deconseq.sourceforge.net). Cleaned (non-human) reads were then BlastX (31) against the NCBI GenBank NR database using DIAMOND (32) and the resulting .daa file was imported in the MEGAN software (33) for visualization of the assigned reads. Paired-reads were assembled into contigs using the CLC Genomics Workbench 7.5 software (CLC Bio, Aarhus, Denmark) with default parameters. The resulting contigs were compared to the NCBI non-redundant protein database using the BlastX algorithm with the DIAMOND software and

the result were visualized with MEGAN. The open reading frames (ORFs) present in these contigs were determined by MetaGeneMark according to the default heuristic parameters (34). Translated ORFs were annotated by BlastP against the NCBI nr database and the results were parsed according to the taxonomic origin of the 10 Best Blast Hit (BBH). Trimmed paired-reads were mapped on the *Moraxella osloensis* KSH (GenBank assembly accession: GCA_002752795.2) reference genome using CLC Genomics with a minimal length fraction of 0.5 and a minimal similarity of 0.8 and a consensus sequence of 2,348,787 bp (187 contigs) was extracted. This consensus sequence was aligned against the *Moraxella osloensis* KSH (GenBank assembly accession: GCA_002752795.2) reference genome using progressive MAUVE (35) and annotated with the RAST server (36). Protein-encoding genes (PEGs) were compared to the ARG-ANNOT database (37) (release May 2018) using BlastP. R1 and R2 reads were also mapped using Bowtie 2 (38) on the *Moraxella osloensis* KSH reference genome (GenBank assembly accession: GCA_002752795.2) using the --very sensitive option. Consensus sequences of the 16SrRNA genes were retrieved from the mapping result. 16SrRNA genes from other *Moraxella* species and the MCR_1/2 protein sequences were retrieved from the GenBank database using BlastN and BlastP, respectively. Multiple sequence alignments were done using the MUSCLE aligner (v3.8.31) (39) configured for highest accuracy (MUSCLE with default settings) and the phylogenetic trees were reconstructed using the maximum likelihood method implemented in the PhyML program (v3.1/3.0) within the Phylogeny.fr platform (40). Graphical representation and edition of the phylogenetic tree were performed with TreeDyn (v198.3). Branches with support values <50% were collapsed.

PCR amplification of beta-actin gene

The quality of DNA extraction was verified by real-time quantitative PCR targeting the human beta-actin housekeeping gene (41). One microliters of template DNA (pure or diluted at 1/10) was added to a final volume of 20 µl containing 10 µl of QuantiTect SYBR Green Master Mix (QIAGEN) and 0.5 µM of each primer. After an initial denaturation of 15 min for 95°C, 40 amplification cycles have been made under the following conditions: denaturation (30 sec at 94°C), hybridization (30 sec at 60°C) and elongation (30 sec at 72°C). Amplification was performed with a CFX 96 Real Time (Bio-Rad) and fluorescence emission was real-time recorded to generate threshold Cycle (Ct) values, which are inversely proportional to the quantity of starting template.

WESTERN BLOTS

Western blotting procedures were performed as previously described (42); 1 µg of *Coxiella bunetii*, *Moraxella osloensis* CSURP3830, *Moraxella bovoculi* DSM2114, *Moraxella caprae* DSM19149 and *Moraxella lacunata* P1120 antigens was used per lane and incubated with 2 different patient plasma samples (collected in 11/2012 and 06/2013) or one serum sample (collected in 06/2013) at a 1/100 dilution as primary antibody. Immunoreactive bands were detected by chemiluminescence using a peroxidase conjugated goat anti-human antibody.

IMMUNOSTAINING OF MORAXELLA OSLOENSIS WITH ANTI-C. BURNETII

Moraxella osloensis P3830 cells were harvested, resuspended and washed three times in phosphate-buffered saline (PBS). After resuspension in 200 µl of PBS, cells were labeled for 1h with a primary antibody (rabbit anti-*Coxiella burnetii* dilution 1/800) and a secondary antibody (Donkey anti-rabbit antibody Alexa Fluor 488nm, (Jackson ImmunoResearch

Laboratories Inc., West Grove, PA, USA). After washing three times, the cells were resuspended in 500 µl of PBS and analyzed by fluorescence microscopy.

SPECIFIC PCR AND SEQUENCING

PCRs were carried out on the nucleic acids extracted from the positive microdissected sample (pure and 1/10 diluted), from a new valve tissue section (pure and 1/10 diluted), on a PBS negative control that was processed simultaneously and on a positive control (the random amplified nucleic acids from the positive microdissection area). Two microliters of DNA template were added to a final volume of 25 µl containing 12.5 µl of AmpliTaq Gold® 360 Master Mix (Applied Biosystems) and 0.2 µM of each primers (M_oslo_transpF AAATGCGAGAACGCAGGTTG and M_oslo_transpR CCTTTCGGACTATTGGCGGT) and 2 µl of GC enhancer. Amplification started with an initial denaturation step at 95°C for 10 min, followed by 40 cycles at 95°C for 30 seconds, at 56°C for 30 seconds, and at 72°C for 15 seconds. Amplification products were visualized on a 2% agarose gel. Positive PCR products were then purified and sequenced using the same primers with the BigDye version 1-1 Cycle Sequencing Ready Reaction Mix (Applied Biosystems, Foster City, CA) and an ABI 3100 automated sequencer (Applied Biosystems). The sequences were assembled and analyzed using the ChromasPro software (version 1.34) (Technelysium Pty. Ltd., Tewantin, Australia) and the BLAST website (http://blast.ncbi.nlm.nih.gov).

ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche (reference: ANR-13-JSV6-0004) to CD, by the IHU Méditerranée Infection, Marseille, France, by the French Government under the «Investissements d'avenir» program (reference: Méditerranée Infection 10-IAHU-03), by the Région Provence Alpes Côte d'Azur and by the European funding FEDER PRIMI.

ADDITIONAL INFORMATION

Competing financial interests: The authors declare no competing financial interests

The sequencing project has been deposited in the Short Read Archive under the number PRJEB26171 and contains the sequencing files ERR2528754 (positive before depletion), ERR2528753 (negative before depletion), ERR2528755 (positive after depletion) and ERR2528756 (negative after depletion).

An Enigmatic Moraxella osloensis Endocarditis Diagnosed by Laser-Capture Micro-Dissection and Human RNA Bait-Depletion

Matthieu Million[1a], Maxime Gaudin[1a], Cléa Menelotte[1], Lionel Chasson[2], Messica Zeitoun[3], Elsa Prudent[1], Sophie Edouard[1], Bernard Amphoux[1], Elsa Prudent[1], Mathieu Fallet[2], Fred Cadoret[1], Stéphane Meresse[2], Julien Paganini[4], Hubert Lepidi[1], Caroline Michelle[1], Catherine Robert[5], Bernard LaScola[1], Philippe Naquet[2], Jean-Pierre Gorvel[2], Christelle Desnues[1] and Didier Raoult[1*]

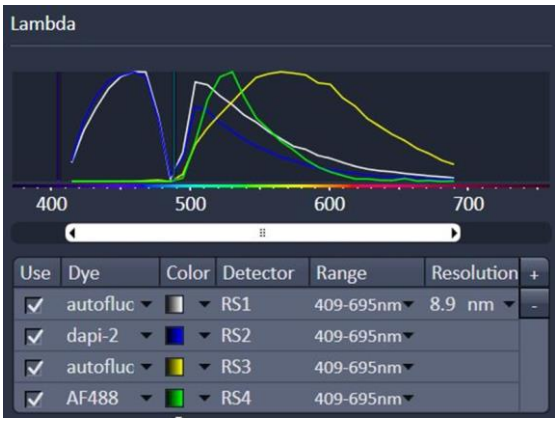
Supplementary data

Supplementary Table 1. Samples and techniques for the standard syndromic diagnostic approach and specific *Coxiella burnetii* tests

Date	Sample	Lab number analysis	C. burnetii serology	Sero Rickettsia	Sero Bartonella quintana	Serology Orientia tsutsugamushi	real time short 16S gene	PCR 16S gene	PCR 18S gene	PCR CU	PCR IS1111	PCR IS30a	PCR Coxiella like	PCR Rickettsia spp.	PCR Bartonella spp.	PCR Bartonella hensalae	PCR Staphylococcus aureus	PCR Escherichia coli	PCR Mycoplasma	PCR Streptococcus oralis	PCR Streptococcus gallolyticus	PCR Enterococcus faecalis	PCR Enterococcus faecium	PCR Tropheryma whippleii	PCR Mycobacterium tuberculosis	PCR Brucella spp.	DNA extraction control (cycle treshold)	Culture Coxiella burnetii	Ziehl staining	Culture of Mycobacterium	IgG anticardiolipin Ab	Antipork Ab
28/11/2012	Blood	1307945 & 1283122	Neg	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	ND	20	Neg	ND	ND	ND	ND
12/06/2013	Blood	1307969 & 1371981	Neg	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	ND	20	Neg	ND	ND	ND	ND
13/06/2013	Serum	1306485	Neg	Neg	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
02/07/2013	Cardiac valve	1372925 & 1372927	ND	ND	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	ND	21	ND	ND	ND	ND	ND
05/07/2013	Electrode	1373217	ND	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	ND	22	Neg	Neg	Neg	ND	ND
05/07/2013	Mechanical aortic valve	1373216	ND	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	23	Neg	Neg	Neg	ND	ND
05/07/2013	Mechanical aortic valve - reextraction - undiluted	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mechanical aortic valve - reextraction - dilution 1/10	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mechanical aortic valve - reextraction - dilution 1/100	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mechanical aortic valve - reextraction - dilution 1/1000	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mechanical aortic valve - reextraction - dilution 1/10000	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mechanical aortic valve - reextraction D-glycosylase - undiluted	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mechanical aortic valve - reextraction D-glycosylase - 1/10	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mechanical aortic valve - reextraction D-glycosylase - 1/100	1373216	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation	1373215	ND	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	18	Neg	Neg	Neg	ND	ND
05/07/2013	Mitral vegetation - reextraction - undiluted	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation - reextraction - dilution 1/10	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation - reextraction - dilution 1/100	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation - reextraction - dilution 1/1000	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation - reextraction - dilution 1/10000	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation - reextraction D-glycosylase - undiluted	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation - reextraction D-glycosylase - 1/10	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Mitral vegetation - reextraction D-glycosylase - 1/100	1373215	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring	1373214	ND	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	17	Neg	Neg	Neg	ND	ND
05/07/2013	Aortic ring - reextraction - undiluted	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring - reextraction - dilution 1/10	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring - reextraction - dilution 1/100	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring - reextraction - dilution 1/1000	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring - reextraction - dilution 1/10000	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring - reextraction D-glycosylase - undiluted	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring - reextraction D-glycosylase - 1/10	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Aortic ring - reextraction D-glycosylase - 1/100	1373214	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve	1373213	ND	ND	ND	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	ND	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	23	Neg	Neg	Neg	ND	ND
05/07/2013	mechanical mitral valve - reextraction - undiluted	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve - reextraction - dilution 1/10	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve - reextraction - dilution 1/100	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve - reextraction - dilution 1/1000	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve - reextraction - dilution 1/10000	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve - reextraction D-glycosylase - undiluted	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve - reextraction D-glycosylase - 1/10	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	mechanical mitral valve - reextraction D-glycosylase - 1/100	1373213	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
05/07/2013	Slide positive by immunohistochemistry	1374759	ND	ND	ND	ND	ND	ND	ND	ND	Neg	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	NA	ND	ND	ND	ND	ND
25/07/2013	serum & blood	1308243	Neg	ND	Neg	ND	Neg	ND	ND	ND	Neg	ND	Neg	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	23	Neg	ND	ND	7.4	<0.10

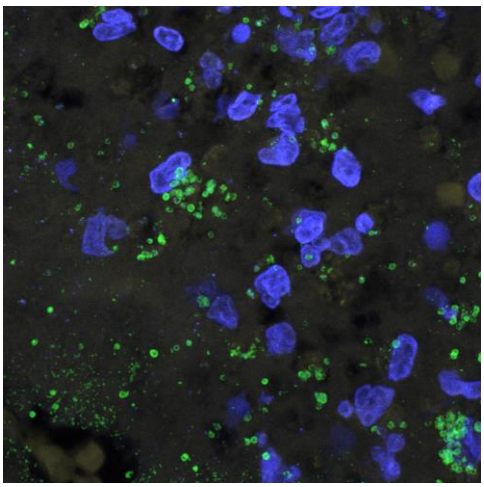
Neg: negative, ND: not done, NA: not applicable. Normal values for Ig G anticardiolipin Ab (N < 22 GPLU). Normal values for anti-pork Antibody (N < 0.10)

Supplementary Figure 1. Autofluorescence spectra for confocal spectral microscopy



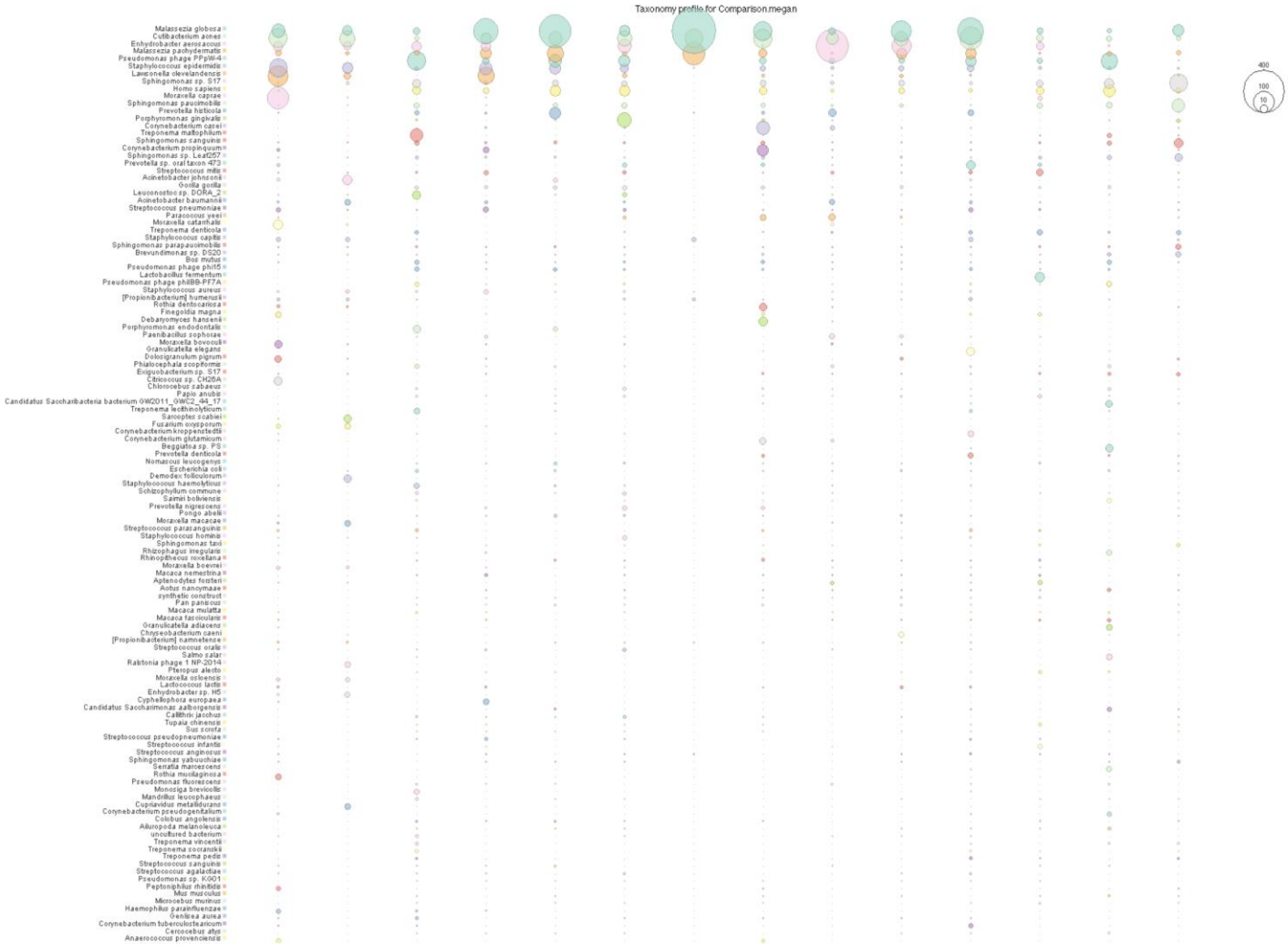
Grey: 405nm autofluorescence spectrum of the patient's cardiac valve, blue: DAPI spectrum, green: Alexa Fluor 488nm, Yellow: 488nm autofluorescence spectrum of the patient's cardiac valve.

Supplementary Figure 2. Confocal spectral imaging of a culture positive endocarditis



Note the bacterial clusters.

Supplementary Figure 3. Comparison of dominant species on our sample and 12 other cardiac valves

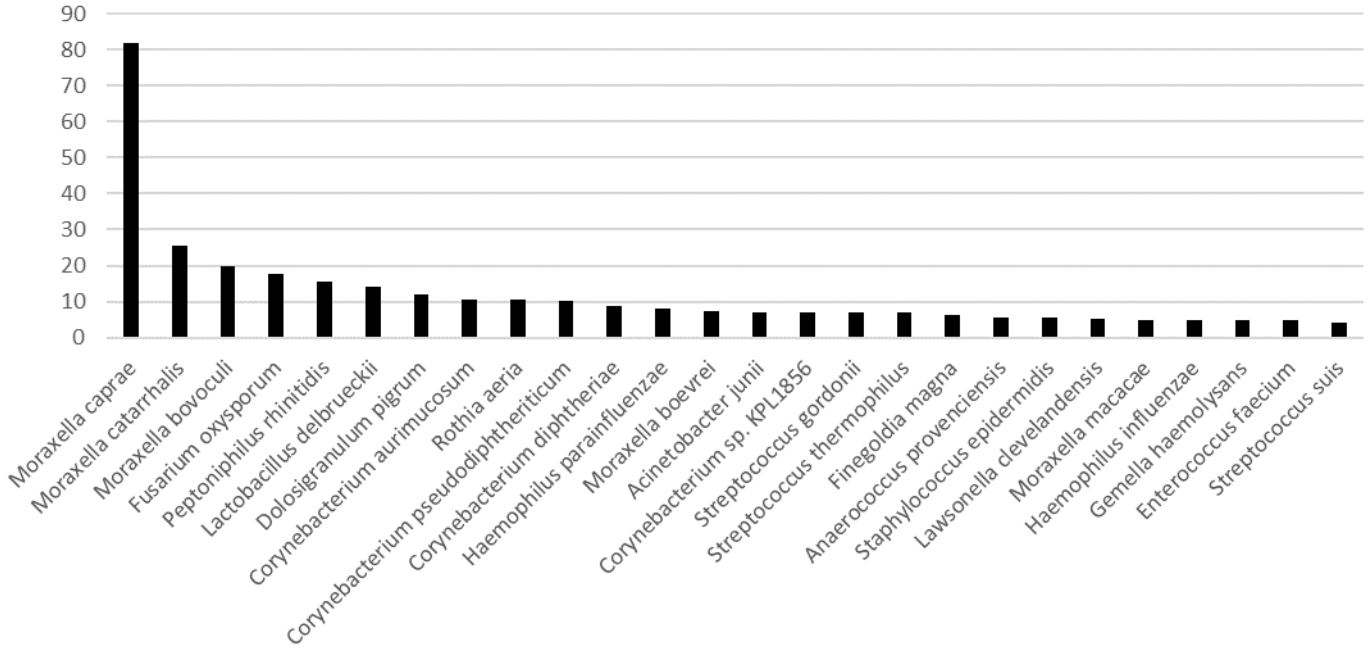


Reads were distributed according to their taxonomical annotation and the size of the circle is proportional to the number of reads. The two majority species are the fungus *Malassezia* and the bacterium *Cutibacterium acnes*, both are skin commensals. Metagenomes for the positive area is first (on the left), the negative area is at the second position and then metagenomes from 12 other valves samples are presented.

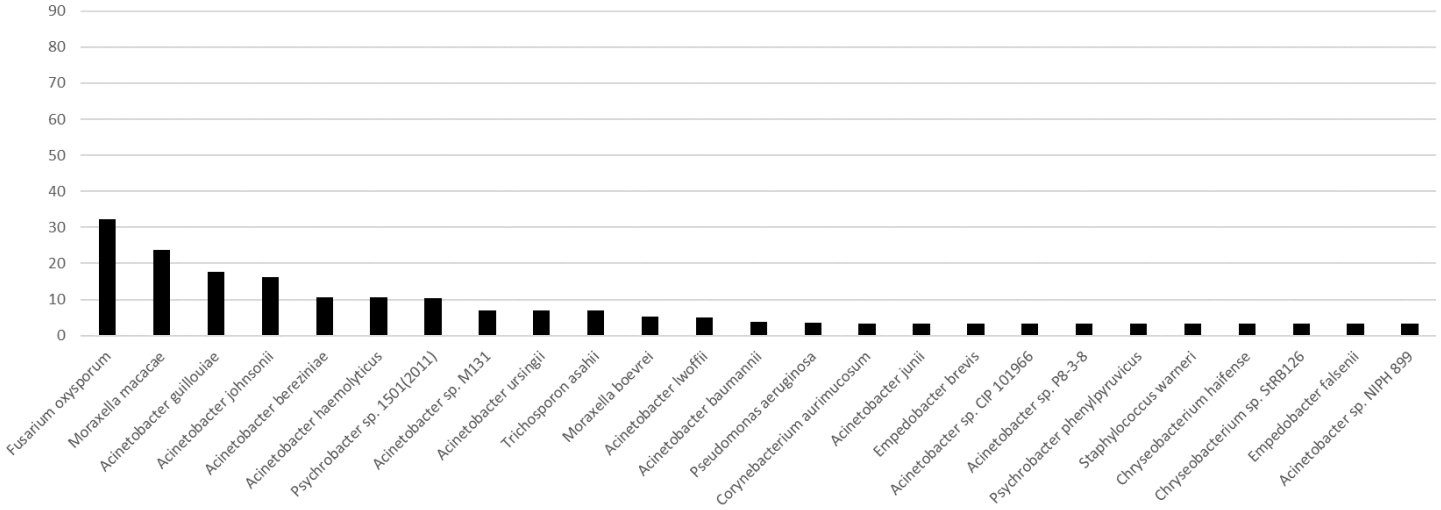
Supplementary Figure 4. Metagenomic z-score

$$Z - score\ for\ species\ X = \frac{(number\ of\ reads\ of\ species\ X\ in\ the\ tested\ sample\ - mean\ of\ reads\ of\ species\ X\ in\ all\ other\ samples)}{Standard\ deviation\ of\ reads\ of\ species\ X\ in\ all\ other\ samples}$$

Supplementary Figure 5. Z-scores of the top species of the IHC positive area



Supplementary Figure 6. Z-scores of the top species of the IHC negative area



Compared to the positive zone, no species clearly predominates.

Supplementary Table 2. Mapping on the Moraxella osloensis KSH reference genome (NZ_CP024180.2) and plasmids (NZ_CP024181.2 to NZ_CP024184.2) of the trimmed paired-reads before capture of the human nucleic acids

	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
References	5	-	545 042,80	2 725 214	-
Mapped reads	2 725	0,13%	229,32	624 895	0,13%
Not mapped reads	2 115 517	99,87%	230,25	487 106 102	99,87%
Reads in pairs	2 312	0,11%	327,78	523 263	0,11%
Broken paired reads	413	0,02%	246,08	101 632	0,02%
Total reads	2 118 242	100,00%	230,25	487 730 997	100,00%

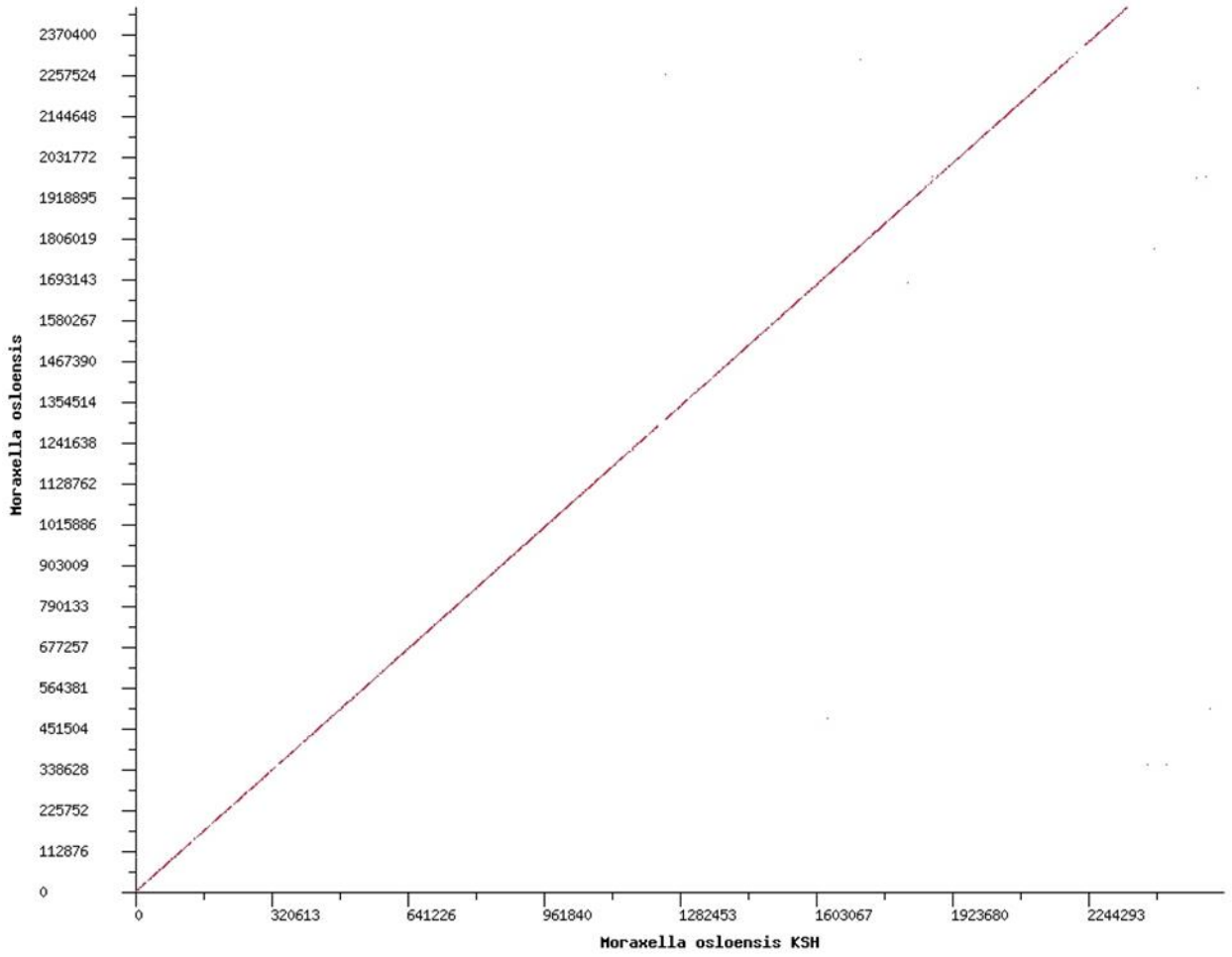
Supplementary Table 3. Mapping on the Moraxella osloensis KSH reference genome (NZ_CP024180.2) and plasmids (NZ_CP024181.2 to NZ_CP024184.2) of the trimmed paired-reads after capture of the human nucleic acids

	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
References	5	-	545 042,80	2 725 214	-
Mapped reads	2 705 641	77,63%	228,90	619 328 295	77,90%
Not mapped reads	779 721	22,37%	225,28	175 657 203	22,10%
Reads in pairs	2 496 844	71,64%	311,28	567 295 289	71,36%
Broken paired reads	208 797	5,99%	249,20	52 033 006	6,55%
Total reads	3 485 362	100,00%	228,09	794 985 498	100,00%

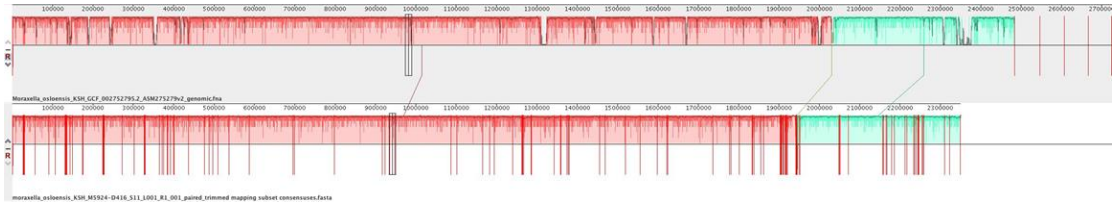
Supplementary Table 4. Length and coverage of the consensus sequences recovered after mapping on the *Moraxella osloensis* KSH reference genome (NZ_CP024180.2) and plasmids (NZ_CP024181.2 to NZ_CP024184.2) of the trimmed paired-reads after capture of the human nucleic acids

Name	Consensus length	Total read count	Average coverage	Reference sequence	Reference length
CP024180 mapping	2348973	2653891	234,21	CP024180	2483277
CP024181 mapping	28351	6781	22,10	CP024181	63733
CP024182 mapping	25000	5269	17,58	CP024182	60481
CP024183 mapping	46294	23166	83,76	CP024183	58635
CP024184 mapping	38748	16534	57,36	CP024184	59088

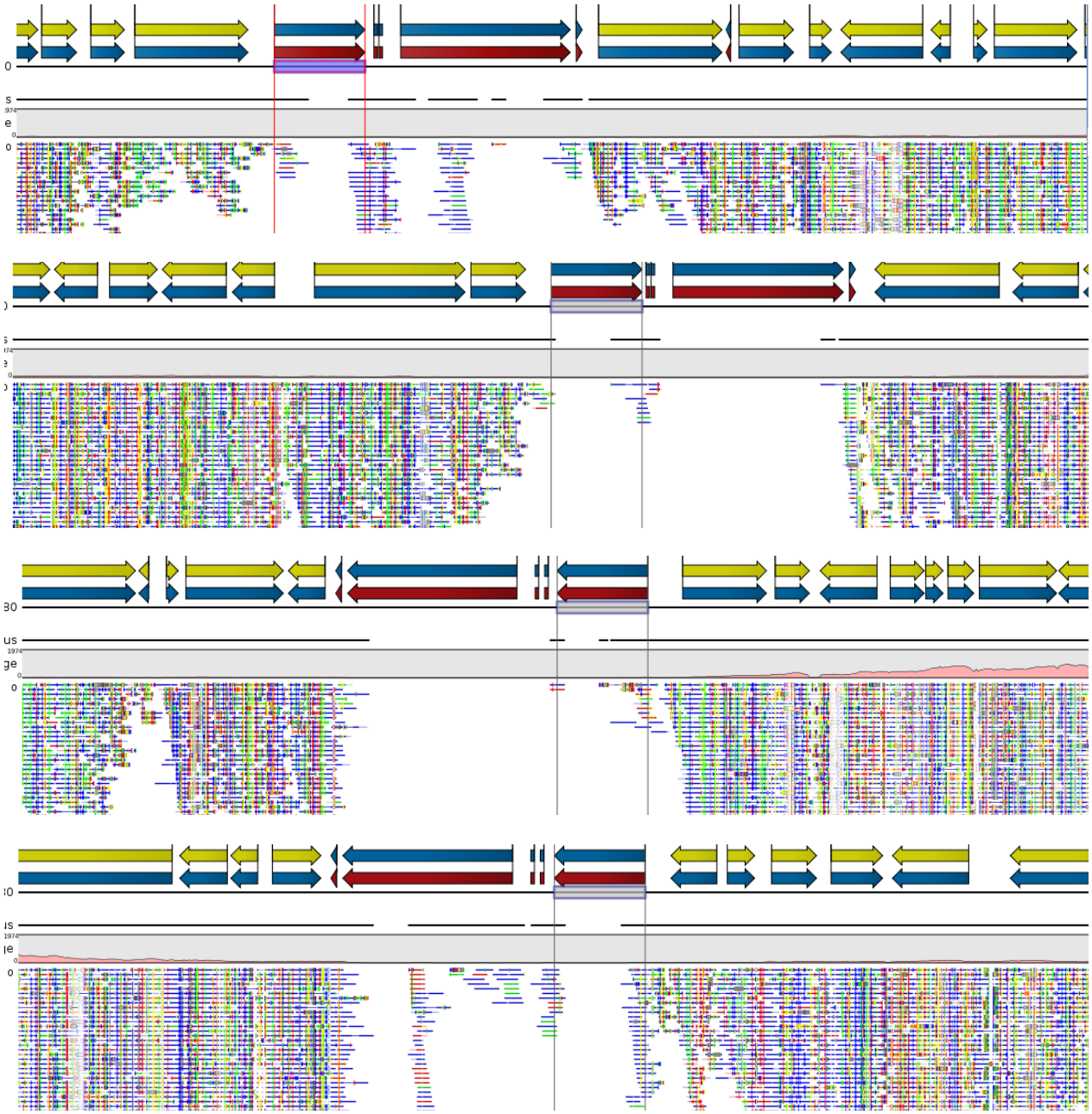
Supplementary Figure 7. Blast DotPlot of the 187 consensus genomic sequences recovered after mapping of the trimmed paired-reads after capture of the human nucleic acids on the *Moraxella osloensis* KSH reference genome (NZ_CP024180.2)



Supplementary Figure 8. MAUVE alignment of the 187 consensus genomic sequences recovered after mapping of the trimmed paired-reads after capture of the human nucleic acids on the *Moraxella osloensis* KSH reference genome (NZ_CP024180.2)



Supplementary Figure 9. Mapping of the paired-reads on the *Moraxella osloensis* KSH reference genome showing the 4 uncovered 16-23S rRNA regions



Supplementary Table 5. List of the 10 best blast hits (BBH) after BlastP of the 2167 protein-encoding genes of the *M. osloensis* strain Marseille against the ARG-ANNOT database. Lowest E-value as well as greatest identity, positive, hit length and bit score are given along with the corresponding accession.

Query	Number of hits	Lowest E-value	Accession (E-value)	Greatest identity %	Accession (identity %)	Greatest positive %	Accession (positive %)	Greatest hit length	Accession (hit length)	Greatest bit score	Accession (bit score)
peg.1719	1	0	(Flq)OqxBgb:EU370913:47851-51003:3153\	41,825	(Flq)OqxBgb:EU370913:47851-51003:3153\	61,122	(Flq)OqxBgb:EU370913:47851-51003:3153\	1041	(Flq)OqxBgb:EU370913:47851-51003:3153\	681,019	(Flq)OqxBgb:EU370913:47851-51003:3153\
peg.1235	10	0	(Col)mcr-6.1:MF176240:1-1617:1617\	65,666	(Col)mcr-6.1:MF176240:1-1617:1617\	78,987	(Col)mcr-6.1:MF176240:1-1617:1617\	531	(Col)mcr-1.5:KY283125:1-1626:1626\	737,643	(Col)mcr-6.1:MF176240:1-1617:1617\
peg.1036	10	3,1E-91	(Bla)penA:AB511945:1298-3049:1762\	35,556	(Bla)blaOXA-372:KJ746496:1-774:774\	50,909	(Bla)blaOXA-22:AF064820:952-1776:825\	607	(Bla)Penicillin_Binding_Protein_Ecoli:CP002291:664439-666340:1902	295,434	(Bla)penA:AB511945:1298-3049:1762\
peg.253	9	1,5E-75	(Bla)Penicillin_Binding_Protein_Ecoli:CP002291:664439-666340:1902	44,164	(Bla)Penicillin_Binding_Protein_Ecoli:CP002291:664439-666340:1902	60,568	(Bla)Penicillin_Binding_Protein_Ecoli:CP002291:664439-666340:1902	317	(Bla)Penicillin_Binding_Protein_Ecoli:CP002291:664439-666340:1902	243,432	(Bla)Penicillin_Binding_Protein_Ecoli:CP002291:664439-666340:1902
peg.1308	2	3,5E-75	(Flq)OqxBgb:EU370913:47851-51003:3153\	28,09	(Gly)vanS-D:AB242319:689-1834:1146\	48,148	(Flq)OqxBgb:EU370913:47851-51003:3153\	1003	(Flq)OqxBgb:EU370913:47851-51003:3153\	265,774	(Flq)OqxBgb:EU370913:47851-51003:3153\
peg.591	10	8,2E-74	(Tet)tetB-P:L20800:2309-4267:1959\	27,153	(Tet)tetW:AJ222769:3687-5606:1920\	46,2	(Tet)tetT:L42544:478-2433:1956\	636	(Tet)tet(44):FN594949:25245-27167:1923\	250,366	(Tet)tetB-P:L20800:2309-4267:1959\
peg.13	26	5,1E-73	(Oxzln)OptrA:KP399637:31477-33444:1968\	38,202	(MLS)msr(A):AY591760:274-1740:1467\	60,674	(MLS)msr(A):AY591760:274-1740:1467\	543	(MLS)car(A):M80346:411-2066:1656\	247,284	(Oxzln)OptrA:KP399637:31477-33444:1968\
peg.2091	30	3,5E-67	(Oxzln)OptrA:KP399637:31477-33444:1968\	42,424	(MLS)vga(E):FR772051:8741-10315:1575\	72,727	(MLS)vga(E):FR772051:8741-10315:1575\	524	(MLS)car(A):M80346:411-2066:1656\	231,106	(Oxzln)OptrA:KP399637:31477-33444:1968\
peg.480	23	8,6E-60	(Oxzln)OptrA:KP399637:31477-33444:1968\	45,714	(MLS)vga(E):FR772051:8741-10315:1575\	65,714	(MLS)vga(E):FR772051:8741-10315:1575\	529	(MLS)car(A):M80346:411-2066:1656\	208,379	(Oxzln)OptrA:KP399637:31477-33444:1968\
peg.1986	10	2,6E-56	(Bla)PBP1a:JN645776:1-2160:2160\	36,111	(Tmt)dfrA3:J03306:103-591:489\	50	(Bla)blaKPC-9:FJ624872:1-854:854\	585	(Bla)PBP1b:AF101781:1-2466:2466\	203,371	(Bla)PBP1a:JN645776:1-2160:2160\

BIBLIOGRAPHY

1. **Fournier PE, Gouriet F, Casalta JP, Lepidi H, Chaudet H, Thuny F, Collart F, Habib G, Raoult D.** 2017. Blood culture-negative endocarditis: Improving the diagnostic yield using new diagnostic tools. Medicine (Baltimore) **96:**e8392.
2. **Lepidi H, Houpikian P, Liang Z, Raoult D.** 2003. Cardiac valves in patients with Q fever endocarditis: microbiological, molecular, and histologic studies. J Infect Dis **187:**1097-1106.
3. **Brouqui P, Raoult D.** 2001. Endocarditis due to rare and fastidious bacteria. Clin Microbiol Rev **14:**177-207.
4. **Edouard S, Million M, Casalta JP, Collart F, Amphoux B, Raoult D.** 2017. Low antibodies titer and serological cross-reaction between Coxiella burnetii and Legionella pneumophila challenge the diagnosis of mediastinitis, an emerging Q fever clinical entity. Infection **45:**911-915.
5. **Edouard S, Million M, Lepidi H, Rolain JM, Fournier PE, La Scola B, Grisoli D, Raoult D.** 2013. Persistence of DNA in a cured patient and positive culture in cases with low antibody levels bring into question diagnosis of Q fever endocarditis. J Clin Microbiol **51:**3012-3017.
6. **Melenotte C, Protopopescu C, Million M, Edouard S, Carrieri P, Eldin C, Angelakis E, Djossou F, Bardin N, Fournier P-E, Mege J-L, Raoult D.** 2018. Clinical features and complications of 2434 Coxiella burnetii infections from the French National Resource Center for Q fever. JAMA Network Open **1:**15.
7. **Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA.** 1996. Laser capture microdissection. Science **274:**998-1001.
8. **Fournier PE, Thuny F, Grisoli D, Lepidi H, Vitte J, Casalta JP, Weiller PJ, Habib G, Raoult D.** 2011. A deadly aversion to pork. Lancet **377:**1542.
9. **Prudent E, Lepidi H, Angelakis E, Raoult D.** 2018. FISH and PNA FISH for the diagnosis of Q fever endocarditis and vascular infections. J Clin Microbiol doi:10.1128/JCM.00542-18.
10. **Million M, Thuny F, Richet H, Raoult D.** 2010. Long-term outcome of Q fever endocarditis: a 26-year personal survey. Lancet Infect Dis **10:**527-535.
11. **Bruggemann H, Henne A, Hoster F, Liesegang H, Wiezer A, Strittmatter A, Hujer S, Durre P, Gottschalk G.** 2004. The complete genome sequence of Propionibacterium acnes, a commensal of human skin. Science **305:**671-673.
12. **Gueho E, Boekhout T, Ashbee HR, Guillot J, Van Belkum A, Faergemann J.** 1998. The role of Malassezia species in the ecology of human skin and as pathogens. Med Mycol **36 Suppl 1:**220-229.
13. **Wei W, Srinivas S, Lin J, Tang Z, Wang S, Ullah S, Kota VG, Feng Y.** 2018. Defining ICR-Mo, an intrinsic colistin resistance determinant from Moraxella osloensis. PLoS Genet **14:**e1007389.
14. **Antonelli A, D'Andrea MM, Vaggelli G, Docquier JD, Rossolini GM.** 2015. OXA-372, a novel carbapenem-hydrolysing class D beta-lactamase from a Citrobacter freundii isolated from a hospital wastewater plant. J Antimicrob Chemother **70:**2749-2756.
15. **Roberts MC, Brown BA, Steingrube VA, Wallace RJ, Jr.** 1990. Genetic basis of tetracycline resistance in Moraxella (Branhamella) catarrhalis. Antimicrob Agents Chemother **34:**1816-1818.
16. **Wang Y, Lv Y, Cai J, Schwarz S, Cui L, Hu Z, Zhang R, Li J, Zhao Q, He T, Wang D, Wang Z, Shen Y, Li Y, Fessler AT, Wu C, Yu H, Deng X, Xia X, Shen J.** 2015. A novel gene, optrA, that confers transferable resistance to oxazolidinones and phenicols and its presence in Enterococcus faecalis and Enterococcus faecium of human and animal origin. J Antimicrob Chemother **70:**2182-2190.
17. **Bovre K, Henriksen SD.** 1967. A new Moraxella species, Moraxella osloensis, and a revised description of Moraxella nonliquefaciens. Int J Syst Bacteriol **17:**9.

18. **Shah SS, Ruth A, Coffin SE.** 2000. Infection due to Moraxella osloensis: case report and review of the literature. Clin Infect Dis **30:**179-181.
19. **Stryker TD, Stone WJ, Savage AM.** 1982. Renal failure secondary to Moraxella osloensis endocarditis. Johns Hopkins Med J **150:**217-219.
20. **Gagnard JC, Hidri N, Grillon A, Jesel L, Denes E.** 2015. Moraxella osloensis, an emerging pathogen of endocarditis in immunocompromised patients? Swiss Med Wkly **145:**w14185.
21. **Paiva PF, Paiva CF, Paiva EG, Fabri GMC, Fabri Junior J.** 2018. Endocarditis Caused by Gram-Negative Moraxella osloensis in an Immunocompetent Patient: First Case Report in Latin America. Case Rep Cardiol **2018:**4209094.
22. **Hansen W, Butzler JP, Fuglesang JE, Henriksen SD.** 1974. Isolation of penicillin and streptomycin resistant strains of Moraxella osloensis. Acta Pathol Microbiol Scand B Microbiol Immunol **82:**318-322.
23. **Wallace RJ, Jr., Steingrube VA, Nash DR, Hollis DG, Flanagan C, Brown BA, Labidi A, Weaver RE.** 1989. BRO beta-lactamases of Branhamella catarrhalis and Moraxella subgenus Moraxella, including evidence for chromosomal beta-lactamase transfer by conjugation in B. catarrhalis, M. nonliquefaciens, and M. lacunata. Antimicrob Agents Chemother **33:**1845-1854.
24. **Nagano N, Sato J, Cordevant C, Nagano Y, Taguchi F, Inoue M.** 2003. Presumed endocarditis caused by BRO beta-lactamase-producing Moraxella lacunata in an infant with Fallot's tetrad. J Clin Microbiol **41:**5310-5312.
25. **AbuOun M, Stubberfield EJ, Duggett NA, Kirchner M, Dormer L, Nunez-Garcia J, Randall LP, Lemma F, Crook DW, Teale C, Smith RP, Anjum MF.** 2017. mcr-1 and mcr-2 variant genes identified in Moraxella species isolated from pigs in Great Britain from 2014 to 2015. J Antimicrob Chemother **72:**2745-2749.
26. **Kieffer N, Nordmann P, Poirel L.** 2017. Moraxella Species as Potential Sources of MCR-Like Polymyxin Resistance Determinants. Antimicrob Agents Chemother **61**.
27. **Marrero M, Raoult D.** 1989. Centrifugation-shell vial technique for rapid detection of Mediterranean spotted fever rickettsia in blood culture. Am J Trop Med Hyg **40:**197-199.
28. **Prudent E, Lepidi H, Angelakis E, Raoult D.** 2018. Fluorescence In Situ Hybridization (FISH) and Peptide Nucleic Acid Probe-Based FISH for Diagnosis of Q Fever Endocarditis and Vascular Infections. J Clin Microbiol **56**.
29. **Gaudin M, Monteil-Bouchard S, Michelle C, Robert C, Raoult D, Desnues C.** 2018. Application of an inverted human Whole-genome In-Solution Capture (inv-WISC) to viral metagenomics. (submitted)
30. **Schmieder R, Edwards R.** 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One **6:**e17288.
31. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-410.
32. **Buchfink B, Xie C, Huson DH.** 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods **12:**59-60.
33. **Huson DH, Auch AF, Qi J, Schuster SC.** 2007. MEGAN analysis of metagenomic data. Genome Res **17:**377-386.
34. **Besemer J, Borodovsky M.** 1999. Heuristic approach to deriving models for gene finding. Nucleic Acids Res **27:**3911-3920.
35. **Darling AE, Mau B, Perna NT.** 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5:**e11147.
36. **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:**D206-214.
37. **Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain JM.** 2014. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother **58:**212-220.

38. **Langdon WB.** 2015. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Min **8:**1.
39. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792-1797.
40. **Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O.** 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res **36:**W465-469.
41. **Angelakis E, Richet H, Rolain JM, La Scola B, Raoult D.** 2012. Comparison of real-time quantitative PCR and culture for the diagnosis of emerging Rickettsioses. PLoS Negl Trop Dis **6:**e1540.
42. **Raoult D, Dasch GA.** 1989. Line blot and western blot immunoassays for diagnosis of Mediterranean spotted fever. J Clin Microbiol **27:**2073-2079.

Chapitre IV : Discussion générale - Conclusions

Dans le cadre de cette thèse, j'ai proposé et mis en place un projet de recherche afin d'améliorer la sensibilité de la métagénomique dans un but d'identification de pathogènes. Ce projet s'est axé exclusivement dans le domaine de la santé humaine et plus particulièrement dans un contexte de pathologies dont l'agent étiologique était inconnu. Bien que l'ensemble des résultats ait permis d'apporter des réponses à des questions d'ordre diagnostique, la métagénomique clinique se heurte encore à des défis d'ordres techniques et conceptuels pour qu'elle puisse être un jour être utilisée en première intention pour le dépistage d'agents infectieux dans des échantillons cliniques. Cette partie de discussion aborde donc différents aspects des enjeux et défis de la métagénomique clinique.

I. Implémentation de la métagénomique dans les laboratoires de diagnostic clinique

Les progrès des techniques de séquençage corrélés à la baisse spectaculaire de leur coût, permettent désormais d'envisager le NGS et la métagénomique pour une utilisation dans les laboratoires de diagnostic clinique. En 2017, une étude évaluait l'intérêt de la métagénomique comme outil de diagnostic de première intention dans le cadre d'infections chez des patients immunodéprimés. Les auteurs ont démontré que le séquençage haut débit permettait de détecter plus de virus et de bactéries cliniquement pertinents avec une meilleure valeur prédictive négative que les méthodes microbiologiques conventionnelles [138]. Cependant, cette population se caractérise par un risque plus élevé de voir se développer des infections impliquant des agents pathogènes nouveaux ou inattendus [139–142], responsables en grande partie de l'échec diagnostique des techniques classiques. Ainsi, avant de recourir à la métagénomique, de nombreux tests diagnostiques standard sont souvent utilisés en amont entraînant potentiellement des coûts inutiles et des retards de diagnostic. On comprend ici l'intérêt de mettre en place le séquençage haut débit comme méthode de choix dans le dépistage des infections dans ce groupe de patient. A l'inverse, lorsqu'un agent pathogène est connu, les approches actuelles de métagénomique présentent une sensibilité limitée par rapport aux techniques traditionnelles. Par exemple, dans le cas d'encéphalites aigüe, une origine virale est le plus souvent suspectée et la PCR en temps réel reste le gold standard pour en déterminer l'agent étiologique. Le délai entre le traitement de l'échantillon jusqu'à la réponse diagnostique varie entre 4 et 12h dans un laboratoire clinique [143] alors qu'il est compris entre 6h et 7 jours (moyenne de 48h) pour les techniques de métagénomique [144–148]. Ce retard dans le

diagnostic pourrait ainsi avoir des conséquences dramatiques dans le cas où la PCR aurait pu rapidement identifier le pathogène responsable [143]. En l'état actuel des choses, de par son coût et son délai de mise en œuvre, la métagénomique apparaît alors complémentaire aux approches moléculaires et réservée aux cas cliniques dont l'étiologie n'a pu être déterminée par PCR, culture et/ou sérologie.

En métagénomique virale, les données issues du séquençage sont généralement largement dominées par les séquences de l'hôte plutôt que par celles de pathogènes. Il en résulte un véritable gaspillage financier avec une perte considérable de sensibilité de la technique. De plus, pour atteindre cette faible proportion en séquences d'intérêts, il faut utiliser la profondeur maximale qu'offre la plateforme de séquençage tout en limitant le nombre d'échantillons séquencés en parallèles [143], générant de fait un coût important. Pour s'affranchir de ces limitations, l'élimination de l'ADN ou de l'ARN humain avant le séquençage est souvent nécessaire mais les méthodes actuellement existantes sont non standardisées et doivent être adaptées en fonction du type de pathogène ciblé dans l'échantillon : virus [90, 113], bactéries ou parasites [91, 98, 149–151]. Cette pléthore de méthodes induit différents types de biais et une difficulté décisionnelle pour déterminer quelle approche mettre en place dans le cas d'échantillons cliniques où le pathogène n'est pas connu. Ces étapes supplémentaires augmentent aussi la probabilité de contaminer les jeux de données par les microorganismes environnants qui s'additionnent aux contaminants déjà connus pour être présents dans les kits d'extraction et/ou d'amplification [152–154]. Ces séquences représentent un frein pour la découverte de nouveaux microbes, notamment ceux présents en faible quantité. Pour identifier ces contaminants, le séquençage en parallèle d'un contrôle négatif ou l'analyse exclusive des séquences les plus abondantes seraient des stratégies envisageables, au risque cependant de ne pas prendre en compte certains pathogènes rares. Dans les deux études du **chapitre III** de cette thèse, nous avons utilisé différents types de contrôles. Pour chaque échantillon, nous avons par exemple traité simultanément un contrôle négatif (tampon PBS) du début de la manipulation jusqu'à l'étape de séquençage. Pour des questions de coûts, ce contrôle n'est généralement pas séquencé, mais systématiquement testé par PCR pour vérifier que les agents d'intérêts détectés dans l'échantillon clinique ne proviennent pas de contaminations introduites lors du traitement des échantillons. Dans le cas de l'endocardite, la présence d'un signal en immunofluorescence nous a permis de définir des zones positives et négatives qui ont été isolées par microdissection pour séquençage shotgun NGS. Avant capture, les séquences les plus abondantes, présentes dans l'ensemble des échantillons, y compris, dans le contrôle négatif (zone non marquée par

immunofluorescence), correspondaient à des contaminants. Une analyse de Z-score sur chaque espèce détectée nous a cependant orientée sur *Moraxella* qui, bien que n'étant pas le genre présentant le plus de séquences dans l'échantillon avant capture, était le plus enrichi dans le métagénome positif. La déplétion a ensuite confirmé ce résultat et permis la reconstruction du génome *quasi* complet d'une nouvelle souche de *Moraxella osloensis* pour laquelle l'infection a été confirmée chez la patiente en sérologie par Western blot.

L'année dernière, des recommandations détaillées ont été émises au sujet de l'implémentation des NGS dans les laboratoires de diagnostic par le Collège des Pathologistes Américains [155] et au cours de la première Conférence Internationale sur la Métagénomique Clinique qui s'est tenue à Genève [156]. Ces recommandations portent sur les 4 principales étapes de la métagénomique qui sont présentées dans la **Figure 9.**

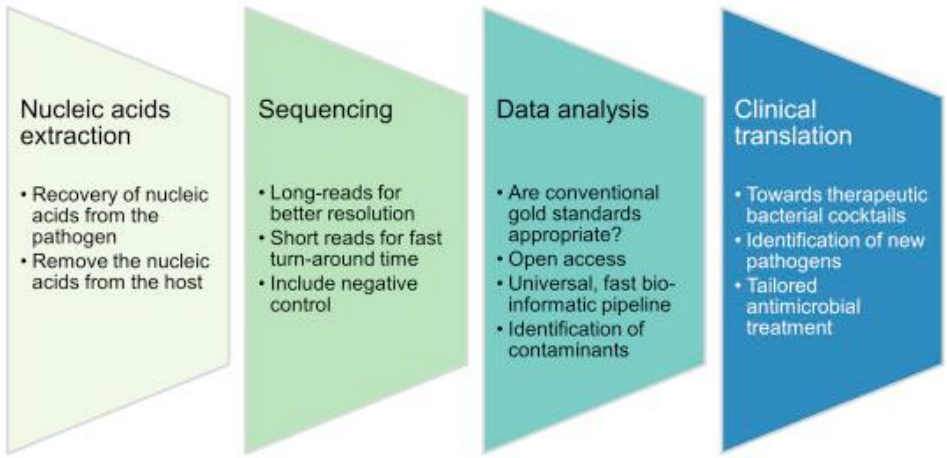


Figure 9 : les 4 grandes étapes de la métagénomique clinique et les limites qui ont été discutées lors de la première conférence Internationale sur la Métagénomique Clinique [156].

Ainsi, pour pouvoir prétendre à une utilisation comme outil diagnostique de première intention dans les maladies infectieuses ; virologues, bactériologistes, mycologues et parasitologues devront travailler ensemble afin de développer un protocole universel capable de s'appliquer à n'importe quel type d'échantillons (fluides biologiques et tissus) et dans n'importe quel contexte clinique [156]. Dans le **chapitre II** de cette thèse, j'ai présenté un protocole nommé inv-WISC répondant à une majorité des limites présentées dans la **Figure 9**. Initialement développé dans le cadre de la recherche de virus, ce protocole s'est également révélé performant pour la détection de pathogènes bactériens et son application en parasitologie et mycologie humaine est envisageable. Cependant, des améliorations techniques incluant (1)

sa rapidité d'exécution (2) son utilisation dans des études transcriptomiques et (3) sa performance et standardisation sur tous types d'échantillons biologiques humains, devront être considérées avant son utilisation en routine.

II. Le postulat de Koch à l'ère de la métagénomique

Bien qu'il ne soit pas exclu qu'elle puisse être présente un jour dans les laboratoires d'analyses, l'utilisation de la métagénomique dans le cadre du diagnostic est restreinte à l'étude descriptive de maladies pour lesquelles aucun agent causal n'est identifié. De plus, par sa nécessité d'employer des scientifiques qualifiés pour effectuer les expériences et analyser les données, à ce jour, elle est majoritairement exercée dans le domaine de la recherche universitaire plutôt que sur le front de la santé publique. Pour cela, des outils bioinformatiques sont disponibles afin de trier les séquences générées et de définir lesquelles seraient significatives dans la pathologie étudiée [157–160]. Si l'identification par NGS de séquences virales ou bactériennes constitue la première étape, elle n'est pas suffisante pour déterminer le spectre d'hôte ni si les micro-organismes détectés sont à l'origine des symptômes observés. Ainsi, de plus en plus d'études de métagénomique ne se contentent plus de la simple identification d'un pathogène dans un prélèvement clinique mais tentent d'apporter la preuve de la causalité des symptômes par cet agent causal. Pour ce faire, il faut souvent revenir au postulat de Koch [161] pour lequel un microorganisme est considéré comme responsable d'une maladie s'il répond aux critères suivants : (1) l'organisme suspecté doit être retrouvé chez tous les patients malades mais pas chez des personnes saines ; (2) l'organisme doit pouvoir être isolé de l'individu malade et cultivé en laboratoire ; (3) l'organisme cultivé doit pouvoir recréer des symptômes identiques chez un individu sain (par exemple dans un modèle animal) ; (4) le même organisme doit pouvoir être ré-isolé de l'organisme sain ainsi inoculé.

Dans le domaine de la virologie, ces quatre critères ont déjà été atteint dans le cas du coronavirus associé au SARS [162]. Cependant dans la majorité des infections virales, ce postulat rencontre assez vite des limites car (1) la grande majorité des virus ne sont pas cultivables ; (2) l'homme est le seul hôte pour le virus étudié (ex : VIH) et/ou le virus peut ne pas se répliquer en dehors de l'homme (ex : virus de l'hépatite B) ; (3) le virus peut être retrouvé dans les organes cibles de sujets non malades dans le cas d'infections asymptomatiques, soit parce que la période d'incubation est longue soit parce qu'il peut persister (ex : VIH, virus de l'hépatite C, West Nile virus, poliovirus, papillomavirus). Par

exemple, le papillomavirus est retrouvé dans la plupart des cancers du col de l'utérus mais également dans les tissues cervicaux normaux, ou encore que seul un très faible pourcentage d'individus infectés par le poliovirus seront paralysés [163, 164]. Depuis l'avènement des technologies de séquençage à haut débit, un pan de l'infectiologie ne repose plus sur l'idée qu'un pathogène donné soit responsable à lui seul d'une maladie rendant ainsi caduque la vision de Koch. En effet dans son concept de relation de cause à effet entre un microbe et une maladie, ces critères ne prennent pas en considération les infections multiples (co-infections) ou celles associés à un changement de la communauté microbienne (dysbiose microbienne) [165].

Dans le domaine médical, la co-infection est l'infection simultanée d'une cellule par au moins deux entités virales ou bactériennes différentes. Les co-infections sont divisées en trois catégories distinctes [166, 167] :

- Les infections synergiques : Un microorganisme génère une niche favorable pour l'infection et la colonisation par un autre agent souvent pathogène. Dans le cas des infections des voies respiratoires, la destruction de l'épithélium respiratoire par un virus induit une immunosuppression locale permettant la surinfection bactérienne.

- Les infections additives : Deux ou plusieurs microorganismes faiblement pathogènes peuvent s'associer et augmenter leur virulence pour causer par exemple des bactériémies, des abcès abdominaux ou pulmonaires, des infections parodontales, des abcès cérébraux ou des otites [166, 167]. Un cas fatal dû à une infection simultanée par un adénovirus de type 7 associé à un bocavirus humain a été reporté alors qu'individuellement ces infections restent bégnines [168].

- L'interférence microbienne : C'est le cas lorsqu'un microorganisme supprime la virulence ou la colonisation d'un pathogène tel le virus de l'hépatite G supprimant la réplication du VIH *in vitro* et diminuant le risque de mortalité chez les patients infectés par le VIH [166, 167].

De plus, les co-infections peuvent parfois rendre le diagnostic plus difficile, l'un des pathogènes pouvant masquer ou modifier les symptômes de l'autre, et inversement [166, 167]. De nos jours, le domaine médical commence à reconnaitre l'importance des maladies polymicrobiennes et des principaux types d'interactions existant entre eux. De nombreuses thérapies commencent tout juste à prendre en compte les causes multi-infectieuses des maladies et l'importance de leur traitement et de leur prévention. L'existence de ce nouveau concept remet en cause la vision de Koch basé sur un concept réduit à l'association « un microbe, une maladie ». Ici, l'agent pathogène n'est plus étudié de manière isolé mais intégré au sein de son écosystème microbien (pas seulement des bactéries, mais aussi des protistes, champignons, virus et phages) avec les interactions qu'il y développe. Cette nouvelle approche en maladies infectieuses a fait l'objet d'une publication en 2014 définissant un nouveau concept nommé le « pathobiome » [169].

Dans le **chapitre III** de ce travail nous avons présenté un cas fatal d'encéphalite infectieuse pour lequel nous avons détecté la présence d'un nouveau gemycircularvirus. Cette étude n'est pas la première à rapporter la présence de gemycircularvirus dans des cas d'encéphalites, y compris d'encéphalite pédiatrique [49, 50]. Bien que la présence d'ADN ait été détectée spécifiquement dans une biopsie cérébrale de la patiente, en l'absence d'autres preuves de pathogénicité incluant une séroconversion (sérologies) ou la détection de séquence/particules virales dans le tissu cérébral (IF/IHC, FISH), le lien de causalité entre ce gemycircularvirus et la pathologie reste incertain. Par ailleurs, dans ce cas clinique, la fulgurance des symptômes et la sévérité de la pathologie soulèvent la possibilité d'une infection additive. En effet, outre le gemycircularvirus, des séquences d'un autre CRESS-DNA virus appartenant au genre des circovirus (famille des *Circoviridae*) ont été détectées. Des séquences de circovirus ont déjà été décrites dans des cas d'encéphalites chez des animaux [170, 171] , tandis que des cyclovirus, un autre genre de la famille *Circoviridae*, ont été retrouvés dans le liquide cérébro-spinal de personnes atteintes de paraplégies inexpliquées [51]. Enfin, quelques jours après le décès de l'enfant, un des personnels soignant de l'unité de réanimation pédiatrique a développé une fièvre et un rash maculo-papuleux spontanément résolutifs (Emmanouil Angelakis, communication personnelle). Bien que sans antécédents connus, d'autres facteurs génétiques ou immuns pourraient ainsi être associés à la sévérité de la pathologie chez cette enfant.

Ainsi le concept de maladies infectieuses doit maintenant tenir compte de leur origine multifactorielle dépendant à la fois des propriétés des microorganismes présents mais

également de nombreux déterminants de l'hôte (génétiques, immunologiques) ou de facteurs de l'environnement. L'expression d'une maladie et son intensité reposent donc sur des interactions entre un ou plusieurs facteurs de ces trois compartiments. Le microbiologiste travaille maintenant avec des millions de séquences génétiques traduisant la présence et la colonisation du corps humain par des millions de microorganismes. Pour pouvoir définir quels microorganismes pourraient être à l'origine des maladies observées, des analyses de corrélation sont souvent envisagées [169]. On comprend donc que pour l'utilisation des résultats issus des disciplines « omiques » en diagnostic, de nombreuses études proposant des modifications et actualisations du postulat de Koch ont été publiées [9, 27, 172, 173]. Cependant, quelle que soit la définition choisie (postulat de Koch originel ou récent), pour établir un lien solide entre le pathogène découvert par métagénomique et une maladie idiopathique des études complémentaires sur la population humaine (épidémiologie par PCR ou sérologie par exemple) sont requises. Enfin, il est indispensable de poursuivre les efforts d'isolement et de culture afin, par la suite, de définir le tropisme cellulaire, d'observer les effets cytopathiques et de produire en grande quantité le pathogène incriminé pour sa caractérisation (séquençage complet du génome et incrémentation des bases de données, détermination du protéome, etc.) ou la production d'anticorps pour des analyses sérologiques et d'immunofluorescence.

III. Conclusions

Au cours des dix dernières années, les techniques de séquençage haut-débit ont permis à la métagénomique de faire un bond en avant pour l'identification et la caractérisation de micro-organismes dans le domaine des maladies infectieuses. Elles ont notamment permis, et ce à plusieurs reprises, de démontrer le « proof of concept » de l'utilité de cette approche non biaisée pour l'identification de nouveaux pathogènes ou de variants, la détection de la transmission, l'analyse du profil de virulence ou de résistance aux antibiotiques, l'étude des co-infections et pathologies complexes, et l'identification de pathogènes présents même en faible nombre de copies (sous couvert d'une profondeur de séquençage suffisante). L'arrivée en 2011 des séquenceurs haut-débit de paillasse à des prix abordables a démocratisé cette technologie et la question maintenant n'est plus de savoir si mais plutôt quand et comment les NGS vont être utilisées dans les laboratoires de diagnostic microbiologique clinique. A ce jour, l'absence de protocoles standardisés, la multiplicité des technologies, la technicité requise et l'importante contamination par les acides nucléiques de l'hôte qui pose à la fois des problèmes de sensibilité et d'éthique (données génétiques confidentielles, base de données sécurisées) limitent leur développement comme outils de première intention. Au cours de ce travail, nous avons développé une approche innovante permettant de pallier certaines de ces limites. Validé pour la détection de cibles virales et bactérienne à partir d'échantillons complexes, ce protocole de inv-WISC est également envisageable dans le domaine du diagnostic en parasitologie et mycologie humaine. Des réductions significatives de coût, de temps d'expérimentation et d'analyse des données, seront nécessaires pour qu'il puisse s'intégrer dans un pipeline de routine en diagnostic microbiologique.

Chapitre V : Références bibliographiques

1. Berche P (2007) Une histoire des microbes. John Libbey Eurotext

2. Riedel S (2005) Edward Jenner and the history of smallpox and vaccination. Proc Bayl Univ Med Cent 18:21–25

3. La Scola B, Desnues C, Pagnier I, et al (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455:100–104 . doi: 10.1038/nature07218

4. Bândea CI (1983) A new theory on the origin and the nature of viruses. J Theor Biol 105:591–602 . doi: 10.1016/0022-5193(83)90221-7

5. Lefkowitz EJ, Dempsey DM, Hendrickson RC, et al (2018) Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res 46:D708–D717 . doi: 10.1093/nar/gkx932

6. Baron S, Fons M, Albrecht T (1996) Viral Pathogenesis. In: Baron S (ed) Medical Microbiology, 4th ed. University of Texas Medical Branch at Galveston, Galveston (TX)

7. Nicolas J-C, Maréchal V (1996) Les infections virales persistantes. 9

8. Virgin HW, Wherry EJ, Ahmed R (2009) Redefining Chronic Viral Infection. Cell 138:30–50 . doi: 10.1016/j.cell.2009.06.036

9. Sridhar S, To KKW, Chan JFW, et al (2015) A Systematic Approach to Novel Virus Discovery in Emerging Infectious Disease Outbreaks. J Mol Diagn 17:230–241 . doi: 10.1016/j.jmoldx.2014.12.002

10. Leland DS, Ginocchio CC (2007) Role of Cell Culture for Virus Detection in the Age of Technology. Clin Microbiol Rev 20:49–78 . doi: 10.1128/CMR.00002-06

11. Barré-Sinoussi F, Chermann JC, Rey F, et al (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868–871

12. Dane DS, Cameron CH, Briggs M (1970) Virus-like particles in serum of patients with Australia-antigen-associated hepatitis. The Lancet 295:695–698 . doi: 10.1016/S0140-6736(70)90926-8

13. Feinstone SM, Kapikian AZ, Purceli RH (1973) Hepatitis A: detection by immune electron microscopy of a viruslike antigen associated with acute illness. Science 182:1026–1028

14. Reyes GR, Purdy MA, Kim JP, et al (1990) Isolation of a cDNA from the virus responsible for enterically transmitted non-A, non-B hepatitis. Science 247:1335–1339

15. Peiris JSM, Lai ST, Poon LLM, et al (2003) Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet Lond Engl 361:1319–1325

16. Scola BL, Audic S, Robert C, et al (2003) A Giant Virus in Amoebae. Science 299:2033–2033 . doi: 10.1126/science.1081867

17. Colson P, de Lamballerie X, Fournous G, Raoult D (2012) Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. Intervirology 55:321–332 . doi: 10.1159/000336562

18. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394 . doi: 10.1146/annurev.micro.57.030502.090759

19. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3:504–510 . doi: 10.1038/nrmicro1163

20. Breitbart M, Hewson I, Felts B, et al (2003) Metagenomic Analyses of an Uncultured Viral Community from Human Feces. J Bacteriol 185:6220–6223 . doi: 10.1128/JB.185.20.6220-6223.2003

21. Wylie KM, Weinstock GM, Storch GA (2012) Emerging View of the Human Virome. Transl Res J Lab Clin Med 160:283–290 . doi: 10.1016/j.trsl.2012.03.006

22. Rascovan N, Duraisamy R, Desnues C (2016) Metagenomics and the Human Virome in Asymptomatic Individuals. Annu Rev Microbiol 70:125–141 . doi: 10.1146/annurev-micro-102215-095431

23. Virgin HW (2014) The Virome in Mammalian Physiology and Disease. Cell 157:142–150 . doi: 10.1016/j.cell.2014.02.032

24. Dethlefsen L, McFall-Ngai M, Relman DA (2007) An ecological and evolutionary perspective on human–microbe mutualism and disease. Nature 449:811–818 . doi: 10.1038/nature06245

25. Methé BA, Nelson KE, Pop M, et al (2012) A framework for human microbiome research. Nature 486:215–221 . doi: 10.1038/nature11209

26. Roux S, Debroas D, Enault F (2013) Application des approches métagénomiques à l'étude de la diversité virale environnementale. Virologie 17:229–242 . doi: 10.1684/vir.2013.0506

27. Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2:63–77 . doi: 10.1016/j.coviro.2011.12.004

28. Janda JM, Abbott SL (2007) 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J Clin Microbiol 45:2761–2764 . doi: 10.1128/JCM.01228-07

29. Lindahl BD, Nilsson RH, Tedersoo L, et al (2013) Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. New Phytol 199:288 . doi: 10.1111/nph.12243

30. Forbes JD, Knox NC, Peterson C-L, Reimer AR (2018) Highlighting Clinical Metagenomics for Enhanced Diagnostic Decision-making: A Step Towards Wider Implementation. Comput Struct Biotechnol J 16:108–120 . doi: 10.1016/j.csbj.2018.02.006

31. Granerod J, Crowcroft NS (2007) The epidemiology of acute encephalitis. Neuropsychol Rehabil 17:406–428 . doi: 10.1080/09602010600989620

32. Graf EH, Simmon KE, Tardif KD, et al (2016) Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: a Systematic Comparison to a Commercial PCR Panel. J Clin Microbiol 54:1000–1007 . doi: 10.1128/JCM.03060-15

33. Finkbeiner SR, Allred AF, Tarr PI, et al (2008) Metagenomic Analysis of Human Diarrhea: Viral Detection and Discovery. PLOS Pathog 4:e1000011 . doi: 10.1371/journal.ppat.1000011

34. Shaw K (2006) The 2003 SARS outbreak and its impact on infection control practices. Public Health 120:8–14 . doi: 10.1016/j.puhe.2005.10.002

35. Graham BS, Sullivan NJ (2018) Emerging viral diseases from a vaccinology perspective: preparing for the next pandemic. Nat Immunol 19:20–28 . doi: 10.1038/s41590-017-0007-9

36. Paules CI, Eisinger RW, Marston HD, Fauci AS (2017) What Recent History Has Taught Us About Responding to Emerging Infectious Disease Threats. Ann Intern Med 167:805 . doi: 10.7326/M17-2496

37. Chiu CY (2013) Viral pathogen discovery. Curr Opin Microbiol 16:468–478 . doi: 10.1016/j.mib.2013.05.001

38. Chomel BB, Belotto A, Meslin F-X (2007) Wildlife, Exotic Pets, and Emerging Zoonoses. Emerg Infect Dis 13:6–11 . doi: 10.3201/eid1301.060480

39. Temmam S, Davoust B, Berenger J-M, et al (2014) Viral metagenomics on animals as a tool for the detection of zoonoses prior to human infection? Int J Mol Sci 15:10377–10397 . doi: 10.3390/ijms150610377

40. Spurgeon ME, Lambert PF (2013) Merkel Cell Polyomavirus: A Newly Discovered Human Virus with Oncogenic Potential. Virology 435:118–130 . doi: 10.1016/j.virol.2012.09.029

41. Woolhouse MEJ, Howey R, Gaunt E, et al (2008) Temporal trends in the discovery of human viruses. Proc R Soc Lond B Biol Sci 275:2111–2115 . doi: 10.1098/rspb.2008.0294

42. Palacios G, Druce J, Du L, et al (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 358:991–998 . doi: 10.1056/NEJMoa073785

43. Loh J, Zhao G, Presti RM, et al (2009) Detection of Novel Sequences Related to African Swine Fever Virus in Human Serum and Sewage. J Virol 83:13019–13025 . doi: 10.1128/JVI.00638-09

44. Briese T, Paweska JT, McMullan LK, et al (2009) Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever–Associated Arenavirus from Southern Africa. PLOS Pathog 5:e1000455 . doi: 10.1371/journal.ppat.1000455

45. Moraz M-L, Kunz S (2011) Pathogenesis of arenavirus hemorrhagic fevers. Expert Rev Anti Infect Ther 9:49–59 . doi: 10.1586/eri.10.142

46. Grard G, Fair JN, Lee D, et al (2012) A Novel Rhabdovirus Associated with Acute Hemorrhagic Fever in Central Africa. PLOS Pathog 8:e1002924 . doi: 10.1371/journal.ppat.1002924

47. Hoffmann B, Tappe D, Höper D, et al (2015) A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. N Engl J Med 373:154–162 . doi: 10.1056/NEJMoa1415627

48. Tappe D, Schlottau K, Cadar D, et al (2018) Occupation-Associated Fatal Limbic Encephalitis Caused by Variegated Squirrel Bornavirus 1, Germany, 2013. Emerg Infect Dis 24:978–987 . doi: 10.3201/eid2406.172027

49. Zhou C, Zhang S, Gong Q, Hao A (2015) A novel gemycircularvirus in an unexplained case of child encephalitis. Virol J 12: . doi: 10.1186/s12985-015-0431-0

50. Phan TG, Mori D, Deng X, et al (2015) Small viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. Virology 482:98–104 . doi: 10.1016/j.virol.2015.03.011

51. Smits SL, Zijlstra EE, van Hellemond JJ, et al (2013) Novel Cyclovirus in Human Cerebrospinal Fluid, Malawi, 2010–2011. Emerg Infect Dis 19:1511–1513 . doi: 10.3201/eid1909.130404

52. Tan LV, Doorn HR van, Nghia HDT, et al (2013) Identification of a New Cyclovirus in Cerebrospinal Fluid of Patients with Acute Central Nervous System Infections. mBio 4:e00231-13 . doi: 10.1128/mBio.00231-13

53. Xu B, Liu L, Huang X, et al (2011) Metagenomic Analysis of Fever, Thrombocytopenia and Leukopenia Syndrome (FTLS) in Henan Province, China: Discovery of a New Bunyavirus. PLOS Pathog 7:e1002369 . doi: 10.1371/journal.ppat.1002369

54. Yu X-J, Liang M-F, Zhang S-Y, et al (2011) Fever with Thrombocytopenia Associated with a Novel Bunyavirus in China. N Engl J Med 364:1523–1532 . doi: 10.1056/NEJMoa1010095

55. Popgeorgiev N, Boyer M, Fancello L, et al (2013) Marseillevirus-like virus recovered from blood donated by asymptomatic humans. J Infect Dis 208:1042–1050 . doi: 10.1093/infdis/jit292

56. Popgeorgiev N, Colson P, Thuret I, et al (2013) Marseillevirus prevalence in multitransfused patients suggests blood transmission. J Clin Virol Off Publ Pan Am Soc Clin Virol 58:722–725 . doi: 10.1016/j.jcv.2013.10.001

57. Popgeorgiev N, Michel G, Lepidi H, et al (2013) Marseillevirus Adenitis in an 11-Month-Old Child. J Clin Microbiol 51:4102–4105 . doi: 10.1128/JCM.01918-13

58. Aherfi S, Colson P, Audoly G, et al (2016) Marseillevirus in lymphoma: a giant in the lymph node. Lancet Infect Dis 16:e225–e234 . doi: 10.1016/S1473-3099(16)30051-2

59. Finkbeiner SR, Li Y, Ruone S, et al (2009) Identification of a Novel Astrovirus (Astrovirus VA1) Associated with an Outbreak of Acute Gastroenteritis. J Virol 83:10836–10839 . doi: 10.1128/JVI.00998-09

60. Naccache SN, Peggs KS, Mattes FM, et al (2015) Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. Clin Infect Dis Off Publ Infect Dis Soc Am 60:919–923 . doi: 10.1093/cid/ciu912

61. Quan P-L, Wagner TA, Briese T, et al (2010) Astrovirus Encephalitis in Boy with X-linked Agammaglobulinemia. Emerg Infect Dis 16:918–925 . doi: 10.3201/eid1606.091536

62. Brown JR, Morfopoulou S, Hubb J, et al (2015) Astrovirus VA1/HMO-C: An Increasingly Recognized Neurotropic Pathogen in Immunocompromised Patients. Clin Infect Dis Off Publ Infect Dis Soc Am 60:881–888 . doi: 10.1093/cid/ciu940

63. Morfopoulou S, Brown JR, Davies EG, et al (2016) Human Coronavirus OC43 Associated with Fatal Encephalitis. N Engl J Med 375:497–498 . doi: 10.1056/NEJMc1509458

64. Fridholm H, Østergaard Sørensen L, Rosenstierne MW, et al (2016) Human pegivirus detected in a patient with severe encephalitis using a metagenomic pan-virus array. J Clin Virol Off Publ Pan Am Soc Clin Virol 77:5–8 . doi: 10.1016/j.jcv.2016.01.013

65. Morfopoulou S, Mee ET, Connaughton SM, et al (2017) Deep sequencing reveals persistence of cell-associated mumps vaccine virus in chronic encephalitis. Acta Neuropathol (Berl) 133:139–147 . doi: 10.1007/s00401-016-1629-y

66. Xu L, Gao H, Zeng J, et al (2018) A fatal case associated with respiratory syncytial virus infection in a young child. BMC Infect Dis 18:217 . doi: 10.1186/s12879-018-3123-8

67. Sullivan PF, Allander T, Lysholm F, et al (2011) An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. BMC Microbiol 11:2 . doi: 10.1186/1471-2180-11-2

68. Jones JF, Kulkarni PS, Butera ST, Reeves WC (2005) GB virus-C – a virus without a disease: We cannot give it chronic fatigue syndrome. BMC Infect Dis 5:78 . doi: 10.1186/1471-2334-5-78

69. Kriesel JD, Hobbs MR, Jones BB, et al (2012) Deep Sequencing for the Detection of Virus-Like Sequences in the Brains of Patients with Multiple Sclerosis: Detection of GBV-C in Human Brain. PLoS ONE 7: . doi: 10.1371/journal.pone.0031886

70. Phan TG, Luchsinger V, Avendaño LF, et al (2014) Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. J Gen Virol 95:922–927 . doi: 10.1099/vir.0.061143-0

71. Yan F, Xiao Y, Li M, et al (2017) Metagenomic Analysis Identified Human Rhinovirus B91 Infection in an Adult Suffering from Severe Pneumonia. Am J Respir Crit Care Med 195:1535–1536 . doi: 10.1164/rccm.201609-1908le

72. Rascovan N, Monteil Bouchard S, Grob J-J, et al (2016) Human Polyomavirus-6 Infecting Lymph Nodes of a Patient With an Angiolymphoid Hyperplasia With Eosinophilia or Kimura Disease. Clin Infect Dis 62:1419–1421 . doi: 10.1093/cid/ciw135

73. Tsuzuki S, Fukumoto H, Mine S, et al (2014) Detection of trichodysplasia spinulosa-associated polyomavirus in a fatal case of myocarditis in a seven-month-old girl. Int J Clin Exp Pathol 7:5308–5312

74. Al CYC et Diagnosis of Fatal Human Case of St. Louis Encephalitis Virus Infection by Metagenomic Sequencing, California, 2016 - Volume 23, Number 10—October 2017 - Emerging Infectious Diseases journal - CDC. doi: 10.3201/eid2310.161986

75. Wilson MR, Zimmermann LL, Crawford ED, et al (2017) Acute West Nile Virus Meningoencephalitis Diagnosed Via Metagenomic Deep Sequencing of Cerebrospinal Fluid in a Renal Transplant Patient. Am J Transplant 17:803–808 . doi: 10.1111/ajt.14058

76. Zhou Y, Fernandez S, Yoon I-K, et al (2016) Metagenomics Study of Viral Pathogens in Undiagnosed Respiratory Specimens and Identification of Human Enteroviruses at a Thailand Hospital. Am J Trop Med Hyg 95:663–669 . doi: 10.4269/ajtmh.16-0062

77. Doan T, Wilson MR, Crawford ED, et al (2016) Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. Genome Med 8:90 . doi: 10.1186/s13073-016-0344-6

78. Conteville LC, de Filippis AMB, Nogueira RMR, et al (2018) Metagenomic analysis reveals Hepatitis A virus in suspected yellow fever cases in Brazil. Mem Inst Oswaldo Cruz 113:66–67 . doi: 10.1590/0074-02760170260

79. Rose R, Constantinides B, Tapinos A, et al (2016) Challenges in the analysis of viral metagenomes. Virus Evol 2: . doi: 10.1093/ve/vew022

80. Fancello L, Raoult D, Desnues C (2012) Computational tools for viral metagenomics and their application in clinical research. Virology 434:162–174 . doi: 10.1016/j.virol.2012.09.025

81. Yang J, Yang F, Ren L, et al (2011) Unbiased Parallel Detection of Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach ▿ . J Clin Microbiol 49:3463–3469 . doi: 10.1128/JCM.00273-11

82. Lewandowska DW, Zagordi O, Zbinden A, et al (2015) Unbiased metagenomic sequencing complements specific routine diagnostic methods and increases chances to detect rare viral strains. Diagn Microbiol Infect Dis 83:133–138 . doi: 10.1016/j.diagmicrobio.2015.06.017

83. Li D, Li Z, Zhou Z, et al (2016) Direct next-generation sequencing of virus-human mixed samples without pretreatment is favorable to recover virus genome. Biol Direct 11: . doi: 10.1186/s13062-016-0105-x

84. Moustafa A, Xie C, Kirkness E, et al (2017) The blood DNA virome in 8,000 humans. PLOS Pathog 13:e1006292 . doi: 10.1371/journal.ppat.1006292

85. Perlejewski K, Popiel M, Laskus T, et al (2015) Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: Unexpected detection of human herpesvirus 1 while searching for RNA pathogens. J Virol Methods 226:1–6 . doi: 10.1016/j.jviromet.2015.09.010

86. Wootton SC, Kim DS, Kondoh Y, et al (2011) Viral Infection in Acute Exacerbation of Idiopathic Pulmonary Fibrosis. Am J Respir Crit Care Med 183:1698–1702 . doi: 10.1164/rccm.201010-1752OC

87. Lazarevic V, Whiteson K, Gaïa N, et al (2012) Analysis of the salivary microbiome using culture-independent techniques. J Clin Bioinforma 2:4 . doi: 10.1186/2043-9113-2-4

88. Salzberg SL, Breitwieser FP, Kumar A, et al (2016) Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. Neurol Neuroimmunol Neuroinflammation 3:e251 . doi: 10.1212/NXI.0000000000000251

89. Wylie KM, Mihindukulasuriya KA, Sodergren E, et al (2012) Sequence analysis of the human virome in febrile and afebrile children. PloS One 7:e27735 . doi: 10.1371/journal.pone.0027735

90. Thurber RV, Haynes M, Breitbart M, et al (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4:470–483 . doi: 10.1038/nprot.2009.10

91. Marotz CA, Sanders JG, Zuniga C, et al (2018) Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 6: . doi: 10.1186/s40168-018-0426-3

92. Van Etten JL, Lane LC, Dunigan DD (2010) DNA Viruses: The Really Big Ones (Giruses). Annu Rev Microbiol 64:83–99 . doi: 10.1146/annurev.micro.112408.134338

93. Thurber RV, Haynes M, Breitbart M, et al (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4:470–483 . doi: 10.1038/nprot.2009.10

94. Kawada J-I, Okuno Y, Torii Y, et al (2016) Identification of Viruses in Cases of Pediatric Acute Encephalitis and Encephalopathy Using Next-Generation Sequencing. Sci Rep 6:33452 . doi: 10.1038/srep33452

95. Somasekar S, Lee D, Rule J, et al (2017) Viral Surveillance in Serum Samples From Patients With Acute Liver Failure By Metagenomic Next-Generation Sequencing. Clin Infect Dis 65:1477–1485 . doi: 10.1093/cid/cix596

96. Greninger AL, Naccache SN, Messacar K, et al (2015) A novel outbreak enterovirus D68 strain associated with acute flaccid myelitis cases in the United States from 2012-2014: a retrospective cohort study. Lancet Infect Dis 15:671–682 . doi: 10.1016/S1473-3099(15)70093-9

97. Zoll J, Rahamat-Langendoen J, Ahout I, et al (2015) Direct multiplexed whole genome sequencing of respiratory tract samples reveals full viral genomic information. J Clin Virol Off Publ Pan Am Soc Clin Virol 66:6–11 . doi: 10.1016/j.jcv.2015.02.010

98. Feehery GR, Yigit E, Oyola SO, et al (2013) A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA. PLOS ONE 8:e76096 . doi: 10.1371/journal.pone.0076096

99. Gu W, Crawford ED, O'Donovan BD, et al (2016) Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol 17:41 . doi: 10.1186/s13059-016-0904-5

100. He S, Wurtzel O, Singh K, et al (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Methods 7:807–812 . doi: 10.1038/nmeth.1507

101. Ciuffi A (2016) Viral cell biology: HIV RNA gets methylated. Nat Microbiol 1:16037 . doi: 10.1038/nmicrobiol.2016.37

102. Hoelzer K, Shackelton LA, Parrish CR (2008) Presence and role of cytosine methylation in DNA viruses of animals. Nucleic Acids Res 36:2825–2837 . doi: 10.1093/nar/gkn121

103. Matranga CB, Andersen KG, Winnicki S, et al (2014) Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol 15:519 . doi: 10.1186/PREACCEPT-1698056557139770

104. Rosseel T, Ozhelvaci O, Freimanis G, Van Borm S (2015) Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. J Virol Methods 222:72–80 . doi: 10.1016/j.jviromet.2015.05.010

105. Li D, Li Z, Zhou Z, et al (2016) Direct next-generation sequencing of virus-human mixed samples without pretreatment is favorable to recover virus genome. Biol Direct 11: . doi: 10.1186/s13062-016-0105-x

106. Bexfield N, Kellam P (2011) Metagenomics and the molecular identification of novel viruses. Vet J Lond Engl 1997 190:191–198 . doi: 10.1016/j.tvjl.2010.10.014

107. Kumar G, Garnova E, Reagin M, Vidali A (2008) Improved multiple displacement amplification with phi29 DNA polymerase for genotyping of single human cells. BioTechniques 44:879–890 . doi: 10.2144/000112755

108. Yilmaz S, Allgaier M, Hugenholtz P (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Methods 7:943–944 . doi: 10.1038/nmeth1210-943

109. Kim K-H, Chang H-W, Nam Y-D, et al (2008) Amplification of Uncultured Single-Stranded DNA Viruses from Rice Paddy Soil. Appl Environ Microbiol 74:5975–5985 . doi: 10.1128/AEM.01275-08

110. Kim K-H, Bae J-W (2011) Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. Appl Environ Microbiol 77:7663–7668 . doi: 10.1128/AEM.00289-11

111. Marine R, McCarren C, Vorrasane V, et al (2014) Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. Microbiome 2:3 . doi: 10.1186/2049-2618-2-3

112. Rosseel T, Borm SV, Vandenbussche F, et al (2013) The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing. PLOS ONE 8:e76144 . doi: 10.1371/journal.pone.0076144

113. Kohl C, Brinkmann A, Dabrowski PW, et al (2015) Protocol for metagenomic virus detection in clinical specimens. Emerg Infect Dis 21:48–57 . doi: 10.3201/eid2101.140766

114. Conceição-Neto N, Zeller M, Lefrère H, et al (2015) Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci Rep 5:16532 . doi: 10.1038/srep16532

115. Chalkias S, Gorham JM, Mazaika E, et al (2018) ViroFind: A novel target-enrichment deep-sequencing platform reveals a complex JC virus population in the brain of PML patients. PLoS ONE 13: . doi: 10.1371/journal.pone.0186945

116. Wylie TN, Wylie KM, Herter BN, Storch GA (2015) Enhanced virome sequencing using targeted sequence capture. Genome Res 25:1910–1920 . doi: 10.1101/gr.191049.115

117. Kwok H, Wu CW, Palser AL, et al (2014) Genomic Diversity of Epstein-Barr Virus Genomes Isolated from Primary Nasopharyngeal Carcinoma Biopsy Samples. J Virol 88:10662–10672 . doi: 10.1128/JVI.01665-14

118. Miyazato P, Katsuya H, Fukuda A, et al (2016) Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. Sci Rep 6: . doi: 10.1038/srep28324

119. Carpenter ML, Buenrostro JD, Valdiosera C, et al (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am J Hum Genet 93:852–864 . doi: 10.1016/j.ajhg.2013.10.002

120. Kennedy PGE (2004) Viral Encephalitis: Causes, Differential Diagnosis, and Management. J Neurol Neurosurg Psychiatry 75:i10–i15 . doi: 10.1136/jnnp.2003.034280

121. Venkatesan A, Geocadin RG (2014) Diagnosis and management of acute encephalitis. Neurol Clin Pract 4:206–215 . doi: 10.1212/CPJ.0000000000000036

122. Granerod J, Crowcroft NS (2007) The epidemiology of acute encephalitis. Neuropsychol Rehabil 17:406–428 . doi: 10.1080/09602010600989620

123. Boucher A, Herrmann JL, Morand P, et al (2017) Epidemiology of infectious encephalitis causes in 2016. Médecine Mal Infect 47:221–235 . doi: 10.1016/j.medmal.2017.02.003

124. Venkatesan A, Tunkel AR, Bloch KC, et al (2013) Case Definitions, Diagnostic Algorithms, and Priorities in Encephalitis: Consensus Statement of the International Encephalitis Consortium. Clin Infect Dis Off Publ Infect Dis Soc Am 57:1114–1128 . doi: 10.1093/cid/cit458

125. Brown JR, Bharucha T, Breuer J (2018) Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. J Infect 76:225–240 . doi: 10.1016/j.jinf.2017.12.014

126. Wilson MR, Shanbhag NM, Reid MJ, et al (2015) Diagnosing Balamuthia mandrillaris Encephalitis With Metagenomic Deep Sequencing. Ann Neurol 78:722–730 . doi: 10.1002/ana.24499

127. Chan BK, Wilson T, Fischer KF, Kriesel JD (2014) Deep Sequencing to Identify the Causes of Viral Encephalitis. PLOS ONE 9:e93993 . doi: 10.1371/journal.pone.0093993

128. Brown JR, Morfopoulou S, Hubb J, et al (2015) Astrovirus VA1/HMO-C: An Increasingly Recognized Neurotropic Pathogen in Immunocompromised Patients. Clin Infect Dis Off Publ Infect Dis Soc Am 60:881–888 . doi: 10.1093/cid/ciu940

129. Raoult D, Dasch GA (1989) Line blot and western blot immunoassays for diagnosis of Mediterranean spotted fever. J Clin Microbiol 27:2073–2079

130. Lechtzier V, Hutoran M, Levy T, et al (2002) Sodium dodecyl sulphate-treated proteins as ligands in ELISA. J Immunol Methods 270:19–26

131. Moreillon P, Que Y-A (2004) Infective endocarditis. The Lancet 363:139–149 . doi: 10.1016/S0140-6736(03)15266-X

132. Bin Abdulhak AA, Baddour LM, Erwin PJ, et al (2014) Global and regional burden of infective endocarditis, 1990-2010: a systematic review of the literature. Glob Heart 9:131–143 . doi: 10.1016/j.gheart.2014.01.002

133. Cresti A, Chiavarelli M, Scalese M, et al (2017) Epidemiological and mortality trends in infective endocarditis, a 17-year population-based prospective study. Cardiovasc Diagn Ther 7:27–35 . doi: 10.21037/cdt.2016.08.09

134. Hill EE, Herijgers P, Herregods M-C, Peetermans WE (2006) Evolving trends in infective endocarditis. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis 12:5–12 . doi: 10.1111/j.1469-0691.2005.01289.x

135. Brouqui P, Raoult D (2001) Endocarditis Due to Rare and Fastidious Bacteria. Clin Microbiol Rev 14:177–207 . doi: 10.1128/CMR.14.1.177-207.2001

136. Fukui Y, Aoki K, Okuma S, et al (2015) Metagenomic analysis for detecting pathogens in culture-negative infective endocarditis. J Infect Chemother Off J Jpn Soc Chemother 21:882–884 . doi: 10.1016/j.jiac.2015.08.007

137. Imai A, Gotoh K, Asano Y, et al (2014) Comprehensive metagenomic approach for detecting causative microorganisms in culture-negative infective endocarditis. Int J Cardiol 172:e288–e289 . doi: 10.1016/j.ijcard.2013.12.197

138. Parize P, Muth E, Richaud C, et al (2017) Untargeted next-generation sequencing-based first-line diagnosis of infection in immunocompromised adults: a multicentre, blinded, prospective study. Clin Microbiol Infect 23:574.e1-574.e6 . doi: 10.1016/j.cmi.2017.02.006

139. Saylor D, Thakur K, Venkatesan A (2015) Acute encephalitis in the immunocompromised individual: Curr Opin Infect Dis 28:330–336 . doi: 10.1097/QCO.0000000000000175

140. Verykiou S, Goodhead C, Parry G, Meggitt S (2018) Legionella feeleii: an unusual organism associated with cutaneous infection in an immunocompromised patient. Clin Exp Dermatol 43:300–302 . doi: 10.1111/ced.13346

141. Qureshi S, Pandey A, Sirohi TR, et al (2014) Mixed pulmonary infection in an immunocompromised patient: a rare case report. Indian J Med Microbiol 32:79–81 . doi: 10.4103/0255-0857.124330

142. Waggoner JJ, Soda EA, Deresinski S (2013) Rare and Emerging Viral Infections in Transplant Recipients. Clin Infect Dis 57:1182–1188 . doi: 10.1093/cid/cit456

143. Brown JR, Bharucha T, Breuer J (2018) Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. J Infect 76:225–240 . doi: 10.1016/j.jinf.2017.12.014

144. Afshinnekoo E, Chou C, Alexander N, et al (2017) Precision Metagenomics: Rapid Metagenomic Analyses for Infectious Disease Diagnostics and Public Health Surveillance. J Biomol Tech JBT 28:40–45 . doi: 10.7171/jbt.17-2801-007

145. Wilson MR, Naccache SN, Samayoa E, et al (2014) Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. N Engl J Med 370:2408–2417 . doi: 10.1056/NEJMoa1401268

146. Pendleton KM, Erb-Downward JR, Bao Y, et al (2017) Rapid Pathogen Identification in Bacterial Pneumonia Using Real-Time Metagenomics. Am J Respir Crit Care Med 196:1610–1612 . doi: 10.1164/rccm.201703-0537LE

147. Greninger AL, Naccache SN, Federman S, et al (2015) Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med 7:99 . doi: 10.1186/s13073-015-0220-9

148. Kujiraoka M, Kuroda M, Asai K, et al (2017) Comprehensive Diagnosis of Bacterial Infection Associated with Acute Cholecystitis Using Metagenomic Approach. Front Microbiol 8:685 . doi: 10.3389/fmicb.2017.00685

149. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, et al (2016) Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods 127:141–145 . doi: 10.1016/j.mimet.2016.05.022

150. Noyes NR, Weinroth ME, Parker JK, et al (2017) Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. Microbiome 5: . doi: 10.1186/s40168-017-0361-8

151. Shih SY, Bose N, Gonçalves ABR, et al (2018) Applications of Probe Capture Enrichment Next Generation Sequencing for Whole Mitochondrial Genome and 426 Nuclear SNPs for Forensically Challenging Samples. Genes 9: . doi: 10.3390/genes9010049

152. Salter SJ, Cox MJ, Turek EM, et al (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12: . doi: 10.1186/s12915-014-0087-z

153. Laurence M, Hatzis C, Brash DE (2014) Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. PLoS ONE 9: . doi: 10.1371/journal.pone.0097876

154. Thoendel M, Jeraldo P, Greenwood-Quaintance KE, et al (2017) Impact of Contaminating DNA in Whole-Genome Amplification Kits Used for Metagenomic Shotgun Sequencing for Infection Diagnosis. J Clin Microbiol 55:1789–1801 . doi: 10.1128/JCM.02402-16

155. Schlaberg R, Chiu CY, Miller S, et al (2017) Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. Arch Pathol Lab Med 141:776–786 . doi: 10.5858/arpa.2016-0539-RA

156. Ruppé E, Greub G, Schrenzel J (2017) Messages from the first International Conference on Clinical Metagenomics (ICCMg). Microbes Infect 19:223–228 . doi: 10.1016/j.micinf.2017.01.005

157. Ahn T-H, Chai J, Pan C (2015) Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 31:170–177 . doi: 10.1093/bioinformatics/btu641

158. Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46 . doi: 10.1186/gb-2014-15-3-r46

159. Hong C, Manimaran S, Shen Y, et al (2014) PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2:33 . doi: 10.1186/2049-2618-2-33

160. Andrusch A, Dabrowski PW, Klenner J, et al (2018) PAIPline: pathogen identification in metagenomic and clinical next generation sequencing samples. Bioinformatics 34:i715–i721 . doi: 10.1093/bioinformatics/bty595

161. Koch R The etiology of anthrax, based on the life history of *Bacillus anthracis*.

162. Osterhaus ADME, Fouchier RAM, Kuiken T (2004) The aetiology of SARS: Koch's postulates fulfilled. Philos Trans R Soc B Biol Sci 359:1081–1082 . doi: 10.1098/rstb.2004.1489

163. Mammette A (2002) Virologie médicale. Presses Universitaires Lyon

164. Shors (2016) Understanding Viruses. Jones & Bartlett Publishers

165. Singh VP, Proctor SD, Willing BP (2016) Koch's postulates, microbial dysbiosis and inflammatory bowel disease. Clin Microbiol Infect 22:594–599 . doi: 10.1016/j.cmi.2016.04.018

166. Brogden KA, Guthmiller JM, Taylor CE (2005) Human polymicrobial infections. Lancet Lond Engl 365:253–255 . doi: 10.1016/S0140-6736(05)17745-9

167. Brogden KA, Guthmiller JM (2002) Polymicrobial Diseases: Current and Future Research. ASM Press

168. Heydari H, Mamishi S, Khotaei G-T, Moradi S (2011) Fatal type 7 adenovirus associated with human bocavirus infection in a healthy child. J Med Virol 83:1762–1763 . doi: 10.1002/jmv.22149

169. Vayssier-Taussat M, Albina E, Citti C, et al (2014) Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. Front Cell Infect Microbiol 4:29 . doi: 10.3389/fcimb.2014.00029

170. Bexton S, Wiersma LC, Getu S, et al (2015) Detection of Circovirus in Foxes with Meningoencephalitis, United Kingdom, 2009-2013. Emerg Infect Dis 21:1205–1208 . doi: 10.3201/eid2107.150228

171. Bukovsky C, Schmoll F, Revilla-Fernández S, Weissenböck H (2007) Studies on the aetiology of non-suppurative encephalitis in pigs. Vet Rec 161:552–558

172. Lipkin WI, Anthony SJ (2015) Virus hunting. Virology 479–480:194–199 . doi: 10.1016/j.virol.2015.02.006

173. Fredericks DN, Relman DA (1996) Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. Clin Microbiol Rev 9:18–33

Abstract

The development of Next Generation Sequencing (NGS) techniques has revolutionized research and diagnostic in the field of human infectious diseases. In clinical virology, viral metagenomics, which is based on the random shotgun sequencing of all viral genomes present in a sample, is a promising approach for blind detection and identification of potential new pathogens. Its use is however still marginal because of the large proportion of human nucleic sequences which masks the viral signal, limits the reconstruction of viral genomes and requires a ultra-deep sequencing, thus generating higher sequencing costs. In recent years, numerous protocols based on filtration/centrifugation and nuclease digestion steps have been developed to reduce human contamination with limited success particularly in the case of clinical biopsies. In this context, this thesis work aims at improving the metagenomic approach for the clinical diagnosis of viral infectious diseases by increasing the ratio of pathogen-to-host sequences trough depletion of human nucleic acids from the samples.

The first chapter of this thesis consists in a bibliographic synthesis of viral metagenomic approaches in clinical research and the challenges we faced in this field. This bibliographic overview also includes a review article on targeted-enrichment sequencing approaches for pathogen detection in the field of human infectious diseases.

The second chapter of this thesis proposes a methodological development allowing the enrichment of non-human sequences from metagenomes through hybridization and capture of human nucleic acids with biotinylated human RNA probes. Depletion of human nucleic acids was optimized and verified on a mock viral metagenome consisting of varying proportions of human and viral nucleic acids (Herpes simplex virus 1). We then validated its application by reducing human contamination by more than 90% as revealed by real-time quantitative PCR. The results after NGS sequencing confirm an average depletion of 56.5-fold for human sequences and an enrichment of 64-fold for viral sequences.

The third chapter of this thesis is divided into two sub-chapters that propose the application of this protocol to the detection of putative pathogens in (1) a fatal case of encephalitis and (2) an enigmatic case of blood-culture negative infectious endocarditis. In the first case, the 2,134 bp complete genome of a new gemycircularvirus was reconstructed from a cerebral biopsy sample while in the second, we identified a new strain of *Moraxella osloensis* from a mitral valve sample and reconstructed its nearly-complete genome with an average coverage >200X.

The methodological approach developed during this work is finally discussed in a fourth chapter, which also replaces the results obtained in the broader context of emerging infectious diseases and validation of the causal link between the agent detected and the observed pathology.

Résumé

Le développement des techniques de séquençage de nouvelle génération (NGS) a révolutionné la recherche et le diagnostic dans le domaine des maladies infectieuses humaines. En virologie clinique, la métagénomique virale qui repose sur le séquençage aléatoire de type shotgun de l'ensemble des génomes viraux d'un échantillon (le virome), est une approche prometteuse pour la détection et l'identification sans *a priori* de potentiels nouveaux pathogènes. Cependant, son utilisation reste encore marginale en raison de l'importante contamination des viromes par les séquences nucléiques de l'hôte qui masque le signal viral, limite la reconstruction de génomes viraux et requiert une profondeur importante de séquençage, générant ainsi un cout élevé. Ces dernières années, de nombreux protocoles reposant principalement sur des étapes de filtration/centrifugation et digestions enzymatiques, ont été développés pour diminuer cette contamination humaine avec un succès limité notamment dans le cas de biopsies cliniques. Dans ce contexte, ce travail de thèse avait pour objectif d'améliorer l'approche de métagénomique pour le diagnostic clinique de maladies infectieuses virales en augmentant le ratio de séquences pathogène/hôte par déplétion des acides nucléiques humains.

Le premier chapitre de cette thèse consiste en une synthèse bibliographique des approches de métagénomique virale en recherche clinique et des challenges à relever dans ce domaine. Cette synthèse bibliographique inclut également une revue sur les approches de capture/séquençage ciblées de certains pathogènes dans le domaine des maladies infectieuses humaines.

Le deuxième chapitre de cette thèse propose une mise au point méthodologique permettant d'enrichir les métagénomes en séquences non-humaines basée sur l'hybridation et la capture de l'ensemble des acides nucléiques de l'hôte après hybridation avec des sondes ARN humaines biotinylées. La déplétion des acides nucléiques humains a été optimisée et vérifiée sur un métagénome viral artificiel constitué de proportions variables d'acides nucléiques humains et viraux (Herpes simplex virus 1). Nous avons ensuite validé son application en démontrant une réduction de plus de 90% de la contamination humaine par PCR quantitative en temps réel. Les résultats après séquençage NGS confirment une déplétion en séquences humaines de 56,5 fois et un enrichissement de 64 fois en séquences virales.

Le troisième chapitre de cette thèse est divisé en deux sous-chapitres qui proposent l'application de ce protocole à la détection d'agents potentiellement impliqués (1) dans un cas fatal d'encéphalite et (2) dans un cas énigmatique d'endocardite infectieuse à hémoculture négative. Dans le premier cas, le génome complet d'un nouveau gemycircularvirus de 2134 pb a pu être reconstruit à partir d'un échantillon de biopsie cérébrale tandis que dans le second, nous avons pu identifier une nouvelle souche de *Moraxella osloensis* à partir d'un échantillon de valve mitrale et reconstruire un génome *quasi*-complet avec une couverture moyenne > 200X.

Enfin, dans un quatrième chapitre, l'approche méthodologique que nous avons développée est discutée et les résultats sont replacés dans un contexte élargi d'émergence des maladies infectieuses et de lien de causalité entre l'agent détecté et la pathologie observée.