Aix-Marseille Université
Faculté de Médecine de Marseille
Ecole Doctorale des Sciences de la Vie et de la Santé (EDSVS)

THESE DE DOCTORAT

présentée et soutenue **le 11 octobre 2013** par

Laura FANCELLO

En vue de l'obtention du grade de docteur de l'Université Aix-Marseille

Spécialité : **Pathologie humaine et Maladies Infectieuses**

A viral metagenomic approach to study taxonomic and functional diversity of viral communities from the environment to humans

---

Composition du jury :

M. le Professeur **Bernard LA SCOLA**	Président du Jury
M. le Professeur **Didier RAOULT**	Directeur de Thèse
Mme le Docteur **Christelle DESNUES**	Co-directrice de Thèse
M. le Professeur **Patrick FORTERRE**	Rapporteur
M. le Professeur **Renaud MAHIEUX**	Rapporteur
M. le Docteur **François ENAULT**	Examinateur

Unité de Recherche sur les Maladies Infectieuses Tropicales et Emergentes (URMITE), UM 63 CNRS 7278 IRD 198 INSERM 1095

To my family ... on Earth and in Heaven
Alla mia famiglia... in Cielo e sulla Terra

Preamble

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

Professeur Didier RAOULT

Abstract

Viruses are the most abundant and diverse organisms. They play an important ecological role and they represent a major issue for human health. However, little is known about their diversity and distribution in the environment and in the human body. Recently, viral metagenomics has allowed performing broad unselective exploration of uncultivated viral communities, bypassing the limits of classical viral detection tools. However, most viral metagenomes are generated from temperate regions and North America (for environmental studies) or from modern stool samples, sera/blood and naso-/oropharyngeal samples (for human-associated studies). Therefore, the initial purpose of my thesis is to study viral communities in the least investigated environments or human samples, using viral metagenomics.

The first part of my thesis is a review of the main computational tools for the analysis of viral metagenomes, with a focus on their application in human clinical research. The second part of my thesis presents the computational analysis of viral metagenomes from four perennial bodies of water in the Sahara desert. This is the first viral metagenomic study in the Sahara and the first high-throughput viral metagenomic study in the African continent. In the third part, I investigate human-associated viral communities in non-pathological and pathological conditions. In particular, I describe the first viral metagenome generated from a human coprolite, which is representative of viral flora from an individual human gut in the Middle Ages. Then, I present two human-associated viral metagenomic studies that have applications in clinical research. One study investigates viral communities in human pericardial fluids from idiopathic pericarditis cases. The other study is a functional-level investigation of previously described viral metagenomes from cystic fibrosis patient sputa that focuses on antimicrobial resistance genes carried by bacteriophages to better understand the emergence of multidrug-resistance bacteria in the airways of cystic fibrosis patients.

This thesis work provides original data on unexplored viral communities and shows the potential of viral metagenomics to give insights on viral diversity, reveal the presence of expected and unexpected viruses and decipher their role in the ecosystem.

Keywords : Metagenomics, virus, environment, human body, idiopathic diseases.

Résumé

Les virus constituent les entités biologiques les plus abondantes et les plus diversifiées de la biosphère. Outre leur influence fondamentale sur l'écologie de l'ensemble des écosystèmes sur Terre, ils constituent un problème majeur pour la santé humaine. Pourtant, leur diversité et leur distribution dans l'environnement ou dans le corps humain sont encore très peu connues. Récemment, la métagénomique virale a permis d'explorer les communautés virales sans a priori, s'affranchissant ainsi des limites des outils utilisés classiquement pour la détection virale. Néanmoins, la plupart des viromes environnementaux générés à ce jour proviennent de régions tempérées et d'Amérique du Nord et la plupart des viromes humains ont été générés à partir d'échantillons modernes de selles, sang ou de prélèvements oro-naso-pharyngés. Dans ce cadre, l'objectif de mon travail de thèse était d'apporter de nouvelles connaissances sur les communautés virales d'environnements et d'échantillons humains les moins étudiés en utilisant une approche de métagénomique virale basée sur des nouvelles techniques de séquençage haut-débit.

La première partie de cette thèse consiste en une revue de la littérature qui décrit les principaux outils d'analyse des métagénomes viraux, en portant une attention spécifique à leur application dans des études cliniques. La deuxième partie présente l'analyse de métagénomes viraux associés à quatre étendues d'eau pérennes situées dans le désert du Sahara. Il s'agit ici de la première étude de métagénomique virale au Sahara et du premier virome généré par séquençage à haut débit à partir d'échantillon provenant du continent africain. Dans la troisième partie de ma thèse, je présente une étude des communautés virales associées à l'Homme en condition pathologique ou non-pathologique. Je décris notamment le premier métagénome viral issu d'un coprolithe humain, représentant la flore virale associée à l'intestin d'un individu du Moyen Âge. Je décris ensuite deux études de métagénomique virale humaine avec une application dans la recherche clinique. La première de ces études examine les communautés virales de liquides péricardiques humains provenant de patients atteints d'une péricardite infectieuse d'origine inconnue. La seconde étude propose une analyse fonctionnelle de métagénomes viraux précédemment publiés et réalisés à partir d'expectorations de patients atteints de mucoviscidose. Cette étude fait le point sur les gènes de résistance aux antibiotiques portés par les bactériophages dans le but de mieux appréhender l'apparition de bactéries multi-résistantes dans les voies respiratoires de patients atteints de mucoviscidose.

Ce travail présente ainsi des données inédites sur certaines communautés virales peu étudiées et confirme le potentiel de la métagénomique virale pour étudier la diversité virale, révéler la présence de virus inattendus ou inconnus et comprendre leur rôle dans leur écosystème d'origine.

Mots clés : Métagénomique, virus, environnemental, humain, pathologies sans agent étiologique connu.

Contents

List of Figures

List of acronyms

ACLAME A CLAssification of Mobile genetic Elements

BLAST Basic Local Alignment Tool

bp base pair

CAMERA Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis

CF Cystic Fibrosis

Contig Contiguous sequence

CRISPR Clustered Regularly Interspaced Short Palindromic Repeat

DNA Deoxyribonucleic Acid

dsDNA double stranded DNA

Gb Gigabases or 1,000,000,000 nucleotides

h hours

Helicos SMS Helicos Single Molecule Sequencing

Kb kilobases or 1,000 nucleotides

LADS Linear amplification for Deep Sequencing

LASL Linker-Amplified Shotgun Library

Mb Megabases or 1,000,000 nucleotides

MERS-Cov Middle East Respiratory Syndrome coronavirus

MG-RAST metagenomics RAST server

NCBI National Center for Biotechnology Information

NCLDV Nucleocytoplasmic large DNA virus

NGS Next-generation sequencing

ORF Open Reading Frame

PacBio SMRT Pacific Bioscience Single Molecule Real Time

PCR Polymerase Chain Reaction

Read A sequenced DNA fragment

RNA Ribonucleic Acid

rRNA Ribosomal Ribonucleic Acid

Shotgun sequencing Sequencing of random fragments

SOLiD Sequencing by Oligo Ligation Detection

ssDNA single stranded DNA

tRNA Transfer Ribonucleic Acid

TTV Torque Teno virus

Virome Viral metagenome

WGA Whole Genome Amplification

Chapter 1

Introduction

I developed my thesis work in the field of viral metagenomics. The following section introduces the reader to viral metagenomic studies on environmental and human-associated samples and their application to human clinical research.

1.1 Virus detection and discovery

Viruses are the most abundant and diverse biological entities on earth [1, 2, 3]. It is estimated that in the oceans there are approximately ten times more prokaryotic viruses than their hosts and that these viruses play an important role in geochemical cycles and in shaping microbial communities [4]. Today, most viral diversity is unknown. Until recently, the available viral detection tools did not allow for the systematic exploration of the viral world. Most of these tools require the virus to be isolated and cultured, but viral hosts can be unknown or may be one of the 90%-99% of microbial species that are assumed to be difficult to grow in laboratory conditions [5]. Moreover, some viruses are overlooked as they do not exhibit characteristic cytopathic effects in culture [6]. Among classical detection tools, immunological assays do not require isolation. Still, they fail to identify unexpected or unknown viruses, which are usually too divergent to cross-react. Recently, molecular tools have been widely exploited to systematically describe prokaryotes phylogenetic diversity, using 16S rDNA. However, viruses lack a universally conserved genetic marker. Therefore, PCRs using different highly degenerate primers and PCRs targeting sequences conserved within viral groups have been used, which only identify close variants of known viruses or specific groups of viruses [7, 8, 9].

1.2 Viral Metagenomics: a culture-independent approach to the study of viral communities

Recently, a new approach, called viral metagenomics, has been developed. This approach is based on the sequence analysis of the entire collection of genomes directly isolated from a sample [10] and it allows a systematic comprehensive screening of “what is there”. Metagenomics does not require prior isolation and clonal culturing for species characterization, nor does it require previous assumptions about expected organisms or knowledge of the genomic sequences to be targeted. Common protocols to generate viral metagenomes involve an initial processing of the sample (homogenization, shaking or centrifugation) with or without further steps of viral purification (filtering, density gradient centrifugation). DNAse or RNAse treatments can be included to remove host DNA and RNA, prior to nucleic acid extraction. Different methods can also be adopted to increase the amount of nucleic acids and achieve the quantity which is required for sequencing. Low nucleic acid yields are a typical issue in viral metagenomics, as the majority of viral genomes are small (40-50 Kb). The first and most common amplification methods were the LASL (Linker-Amplified Shotgun Library) [11], which is based on cloning, and the whole genome amplification methods (WGA), such as the amplification via the ϕ29 polymerase [1]. These methods suffer from cloning bias and stochastic amplification biases, respectively. More recently, an optimized version of the LASL method has been introduced [12] as well as the linear amplification for deep sequencing method (LADS) [13] and the Nextera method [14] (Fig. 1.1). Detailed descriptions of the protocols to generate viral metagenomes are provided in the articles by Thurber et al. and by Duhaime et al. [15, 16].

The first viral metagenome was sequenced using the Sanger method (a first-generation sequencing technology), which required cloning [11]. Then, sequencing technologies bypassing the cloning step and yielding far higher throughputs at lower costs (second-generation sequencing technologies) were developed (Fig. 1.2). The first second-generation sequencing technology adopted (2005) in viral metagenomics is 454 pyrosequencing (Roche/454), which generated, at first, approximately 20 Mbp per run, with an average sequence length of 100 bp (Fig. 1.2). A few years later, new second-generation sequencing technologies, such as Illumina or SOLiD were implemented. These technologies generated even higher throughputs but had short read lengths, making it difficult to confidently identify and assemble reads. Today, newer versions of 454 pyrosequencing as well as of the Illumina and SOLiD platforms are continuously being developed with improving performances. At the same time, new sequencing technologies based on single molecule real-time sequencing and that do not require PCR amplification are being commercialized (third-generation sequencing technologies), such as the PacBio RS or Oxford Nanopore. One of the main advantages of these technologies is that they are not affected by amplification artifacts. However, the accuracy of the third-generation sequencing technologies remains rather low.

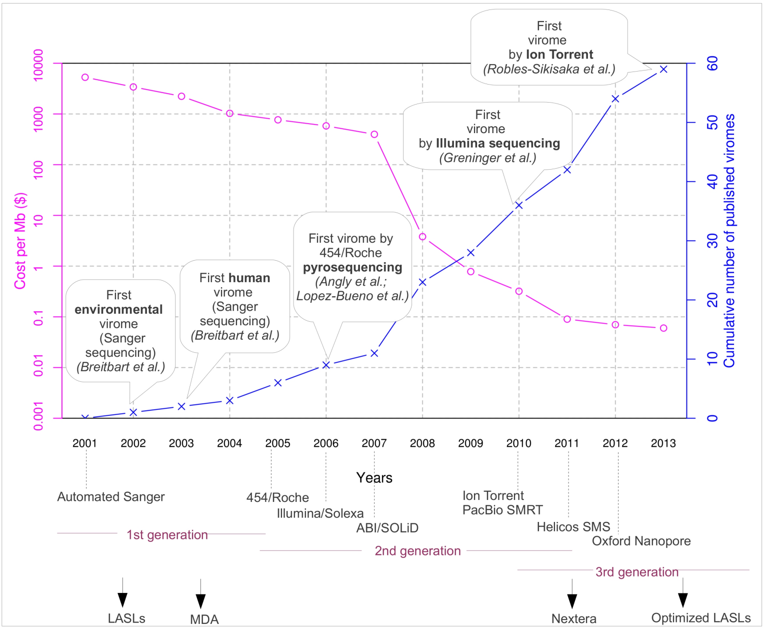


Figure 1.1: Development of sequencing technologies and viral metagenomic studies. The graph shows the decrease of sequencing costs (pink) and the cumulative number of viral metagenomes (blue) over time. Sequencing costs are estimated in dollars per megabase and their evolution is shown using a logarithmic scale (data from the NHGRI large-scale genome sequencing program: www.genome.gov/sequencingcosts). The cumulative number of viral metagenomes only considers published environmental and human-associated viral metagenomes. At the bottom, the development, in time, of different generations of sequencing technologies and of different methods of nucleic acid amplification to generate sufficient template amounts for sequencing are reported.

	ABI/ Life Technologies	454/Roche		Illumina		ABI/SOLiD		Ion Torrent/ Life Technologies	Pacific Biosciences
Method of sequencing	Dideoxy chain termination (fluorescence/optical detection)	Pyrosequencing (fluorescence/optical detection)		Reversible terminator sequencing by synthesis (fluorescence/optical detection)		Ligation-based sequencing (fluorescence/optical detection)		Sequencing by synthesis (detection of pH changes)	Single molecule real time sequencing (fluorescence/optical detection)
Platform (release date)	Sanger 3730xl (2002)	GS-20/FLX (2005)	GS FLX XL+ (2013)	Illumina/Solexa GA (2006)	Illumina HiSeq2500 (2013)	SOLiD (2007)	SOLiD v4 (2013)	Ion PGM 318TM Chip v2 (2012)	PacBio RS II (2013)
Generation	1st	2nd	2nd	2nd	2nd	2nd	2nd	2nd/3rd	3rd
Read length	400-900 bp	100-150 bp	700-1000 bp	35 bp	100 bp	35 bp	35-50 bp	400 bp	4-20 Kb
Output (per run)	1.9-84 Kb	20 Mb	0.7 Gb	1 Gb	540-600 Gb	3 Gb	80-120 Gb	1.2-2 Gb	217 Mb
Time (per run)	20 min-3 h	10 h	23 h	2-3 days	11 days	5-10 days	7 days	7.3 h	10 h

Figure 1.2: Principal next-generation sequencing technologies. The picture shows the principal next-generation sequencing technologies with their corresponding features: method of sequencing, generation of sequencing, read length, output per run and time per run. The data refer to the specified platform release. More releases are available for each cited technology. For the three main second-generation sequencing platforms both the first and the last available releases (as of July 2013) are reported, to illustrate the improving capacities of these technologies.

1.3 Environmental viral metagenomes

The first example of viral metagenomics was realized in 2002 by Breitbart et al. on marine water from two near-shore locations in California [11]. This study demonstrated that in approximately 200 liters of marine water more than 7,000 different viral genotypes were found and viral diversity had been largely underestimated (Shannon index: 7.56-7.99 nats). Approximately 65% of the generated sequences were unknown and the most prevalently identified were bacteriophages. Another study revealed that viral diversity in near-shore marine sediments was even more diverse (Shannon index: 9 nats) and more than 8,000 different viral genotypes were estimated to exist in 1 kilogram of sediment [17]. Once again most of this diversity was unknown. A comparative study between the viromes from four oceanic regions [1] demonstrated that the large majority of viral genotypes are shared between different locations, whereas their relative abundances differ. This implies that viral diversity would be locally high, as shown by viral metagenomic studies, whereas it would be relatively limited on a global scale [1, 18]. Since then, several other viral metagenomic studies have been performed to characterize viruses in marine environments [1, 19, 17, 9, 20, 21, 22]. Most of these studies describe the DNA fraction of viral communities, as oceans are assumed to be prevalently populated by bacteriophages and almost all bacteriophages are DNA viruses. Moreover, both ssDNA viruses and RNA viruses have been previously overlooked as they cannot be detected by direct count studies based on epifluorescence [23, 24, 25]. However, a recent study by Steward et al. revealed an astonishingly high proportion of RNA viruses in the sea [26] and, in particular, picorna-like viruses, which are known to infect eukaryotes. This reveals another large component of the unexplored viral diversity in the sea.

Viral communities from other aquatic environments have been investigated by metagenomics. Those associated with temperate natural freshwater lakes have been demonstrated to be distinct from other aquatic environments and have been shown to be rich in small ssDNA viruses and Caudovirales [27]. Viromes from hot springs presented diversity comparable to that of temperate freshwater viral communities, with a high proportion of unknown sequences, specific genome signatures and prevalent archeal and bacterial viruses [28, 29, 30, 31]. Solar saltern viral communities also have displayed high local diversity but are surprisingly similar between geographically distant locations [20, 32]. In another study, an antarctic lake virome displayed unexpected high diversity, considering its high latitude, and was rich in *Geminiviridae, Circoviridae, Nanoviridae, Microviridae* as well as *Phycodnaviridae*, with seasonal changes in the

Figure 1.3: Environmental viral metagenomes. The picture shows the geographical localization of environmental samples studied by viral metagenomics. The type of biome investigated is indicated by the color. Only published viral metagenomic studies available as of July 2013 are considered. The image is generated using Google Earth (www.google.com/earth).

taxonomic composition [33]. In addition, a hypersaline lake in Senegal was explored using Sanger sequencing [34]. Of the 983 sequences obtained, only 171 were known and mostly represented haloarcheal viruses. Most of remaining sequences had significant similarity to sequences from other viromes associated to high salinity solar salterns.

In addition to marine and freshwater, a few other environments have been explored by viral metagenomics, such as soil [35, 36], water from aquaculture farms [37], air [38] and microbialites [39] (Fig. 1.3). However, most viral metagenomes have been generated from the marine environment and fewer examples exist for other environments. Similarly, a geographical bias of sampling exists that favors temperate regions and the North American continent (Fig. 1.3).

1.4 Human viral metagenomes

Viruses have a major impact on human populations indirectly (*e.g.*, plant viruses affecting cultures or farm animals) or directly (*e.g.*, human pathogenic viruses). Today, several factors promote the emergence or re-emergence of viruses in human population: globalization favors their worldwide spread; immunosuppressive therapies create a favorable field for viruses that are not commonly pathogenic; deforestation and agriculture expansion, especially in the highly biodiverse tropical regions, create new contacts with wildlife and lead to the development of zoonoses [40]. New viruses are continuously being identified in the human population. A recent example is the pandemic H1N1 influenza A virus that emerged in 2009 as a result of the reassortment of genomic segments from swine, human and avian influenza strains [41, 42]. H1N1 influenza A virus is believed to have circulated undetected in the human population for months due to a lack of tools to specifically recognize it [43]. Coronaviruses are another example. Until the 1960s only 2 coronaviruses were known to infect humans [44, 45]. In 2002 and 2003 the SARS coronavirus was discovered as a human pathogen as well as two further new coronaviruses in 2004 and 2005 [46, 47, 48, 49]. Recently, in 2012, the Middle East Respiratory Syndrome coronavirus (MERS-CoV) was described in Saudi Arabia [50]. Thus, surveillance systems are needed to monitor viruses circulating in the human population and to detect even unexpected or unknown viruses. Viral metagenomics is a suitable approach for this task due to its capability to create a comprehensive catalog of the variety of viral genomes present in a sample without *a priori* targeting or knowledge of what is present [51, 52]. Human viral metagenomic studies aim at the characterization of the viral flora in both pathological and non-pathological conditions. Indeed, to establish any association between a disease and a detected viral pathogen, a reference baseline represented by human viral flora in a non-pathological state is required.

The very first viral metagenomic study on a human sample was performed by Breitbart et al., who investigated the DNA viral community associated with a stool sample from an adult healthy individual [53]. The majority of the studies on human viral flora have been performed using human stool samples because of the large availability of material without invasive collection methods. These studies indicated that human gut DNA viromes mainly consisted of unknown viral sequences (59%-98%) and bacteriophages were highly prevalent among the identifiable viruses [53, 54, 55, 56]. The bacteriophage community was relatively stable over time within the same individual at the taxonomic level [55, 56], suggesting that the bacteria-bacteriophages dynamics observed in the environment and described by the Lotka-Volterra or the kill-the-winner models [37] were not likely to occur in human gut. Large taxonomic variability was described between individuals, although more similar viral communities were observed for genetically related individuals and individuals following the same diet [55, 56]. Moreover, human gut bacteriophages were found to carry a strikingly high variety of functions such as CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats), antibiotic resistance, lysins, holins, bacteriocins, virulence factors, and restriction/modification systems [55]. These functions are likely to be laterally transferred to their bacterial hosts, as there is some evidence of the prevalence of lysogeny among these bacteriophages [56]. Fewer studies exist on the RNA portion of viral flora associated with the human gut. These studies demonstrate the prevalence of eukaryotic viruses and, in particular, plant viruses, which were likely to have been introduced by feeding. In addition, fewer unknown viral sequences are detected and the RNA viral community was observed to be dynamic, as it changes substantially in the same individual over time [57, 58].

The normal viral flora of only few other human body sites have been investigated by viral metagenomics up to now (Fig. 1.4). The DNA viromes associated with blood were generated from healthy blood donors and revealed the presence of sequences from a novel anellovirus [18]. Human oropharynx DNA viromes have been demonstrated to be dominated by bacteriophages and to be a reservoir of bacteriophage-encoded virulence genes [59]. DNA viromes associated with the human upper respiratory tract were also studied in a comparative study between healthy individuals and patients affected by cystic fibrosis [60]. These viromes were dominated by bacteriophages but also contained eukaryotic viruses such as geminiviruses, papillomaviruses, adenoviruses, herpesviruses or poxviruses. Human saliva DNA viral communities were first described by Pride et al., who demonstrated the high prevalence of bacteriophages and the stability of the taxonomic profile over time within each individual [61]. Large inter-individual variability was observed even between individuals sharing similar bacterial communities and more similarities were observed for individuals of the same household. This suggests a significant role of the environment in shaping saliva viral communities, as was further confirmed by a study comparing individuals from shared living environments to individuals from different environments [62]. The detected bacteriophages were observed to carry an abundance of virulence-associated genes and genes involved in lysogeny, which may imply a role of these bacteriophages in the pathogenicity of their hosts [61]. Finally, one study exists on the human skin DNA viral flora, in particular facial skin, which mainly contains eukaryotic viruses of the *Papillomaviridae, Polyomaviridae* and *Circoviridae* families [63]. High diversity and large interpersonal variability were observed.

Despite these first efforts toward the characterization of human viral flora, there are still only a few body sites that have been targeted and very little is known about the dynamics of the viral communities in most of them. Studies on larger cohorts of individuals differing in geographical origin, sex, age and lifestyle are needed to provide a more representative assessment of viral flora. These studies are of particular importance for establishing a reference baseline of the physiological human-associated viral flora against which viral communities from patients in clinical research can be compared.

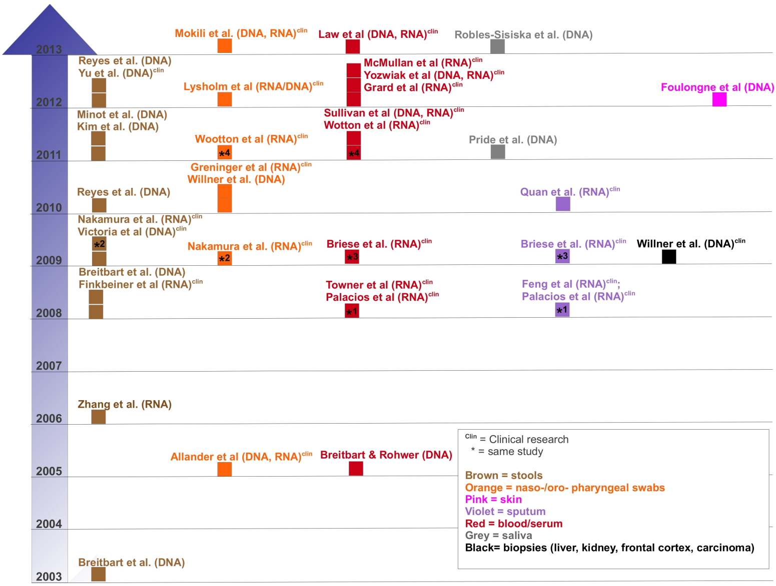


Figure 1.4: Human viral metagenomes. The picture shows the accumulation of viral metagenomic studies on human-associated samples over time. For each study, the type of investigated sample (color), the author and the type of viral genomes studied (DNA viruses or RNA viruses) are reported. Viral metagenomic studies applied to clinical research are also indicated ($^{\mathrm{Clin}}$).

1.5 Viral metagenomics for human clinical research

The principal application of human viral metagenomic studies is clinical investigation. Viral metagenomics has been successfully used to investigate both the origin of unexplained disease outbreaks or the etiology of common idiopathic diseases. The very first example of clinical research by a viral metagenomic approach was provided in 2008 by Palacios et al., who investigated a cluster of unexplained fatal transplant-associated diseases [64]. High-throughput sequencing performed on patient tissues, blood and cerebrospinal fluid allowed the identification of a new arenavirus transmitted through solid-organ transplantation and responsible for the patient deaths. After this first success, viral metagenomics proved to be efficient and fast for identifying and completely sequencing viral pathogens in viral disease outbreaks, even when almost nothing was known about these viruses. In South Africa, in 2008, an outbreak of unexplained hemorrhagic fever was reported and viral metagenomics allowed the detection of the pathogen, which was a novel arenavirus [65]. Similarly, in Uganda, no pathogens could be detected for an unusual outbreak of fever causing bleeding and death. As clinical tests against hemorrhagic fever viruses likely to cause these symptoms were all negative, a viral metagenomic approach was chosen to search for the pathogen without any assumption regarding its identity. An unsuspected yellow fever virus, untargeted by clinical tests, was found [66]. More recently, the same approach allowed for the detection of a new coronavirus and a new rhabdovirus in two outbreaks of hemorrhagic fever and respiratory distress syndrome, respectively [67, 68]. Viral metagenomics has also been successfully applied to the investigations of several diseases of unknown etiology for which a viral origin is suspected: acute flaccid paralysis [69], chronic fatigue syndrome [70], respiratory infections [71, 72, 73, 74, 58, 75] and diarrhea [76, 58, 75]. These studies allowed the discovery of unknown viruses belonging to new species [71, 67, 65, 73, 69] or new genera [72, 69]. In addition, they allowed the formulation of new hypotheses regarding the tropism of some known viruses that were detected for the first time in sites or hosts they were not known to populate. For example, a new astrovirus was observed to be associated with the central nervous system of a patient, whereas such virus is usually described in the gastrointestinal tract [77]. Moreover, plant viruses and circoviruses, known to infect insects and fish, were found in human stool samples and serum [76, 69, 78, 57]. Furthermore, the use of viral metgenomics in clinics can suggest new therapeutic approaches. For instance, a metagenomic study on lung-associated viral communities of patients affected by cystic fibrosis compared with healthy individuals revealed that differences between the diseased- and healthy-associated viral communities are defined by their functional rather than their taxonomic profile. This finding implies that patient treatment should change the respiratory environment rather than target the prevalent detected species [60, 59].

1.6 Computational analysis of human viral metagenomes

Today, the amount of sequences generated by second- and third-generation sequencing technologies (collectively referred to as next-generation sequencing technologies, NGS), in most cases, is likely to be high enough to represent all viruses present in a sample; therefore, sequencing output does not represent any more a limit in viral metagenomics for an exhaustive assessment of viral diversity. On the contrary, the new bottleneck of metagenomics is the computational analysis of the generated data. A virome computational analysis mainly consists of the quality processing of the raw reads, assembly, gene prediction, taxonomic and functional assignment, estimation of the community structure and diversity and comparison between viromes. The first issue in computational analysis is the dimension of generated datasets, which makes them difficult to handle and time-consuming to analyze. In addition, sequences generated by NGS are shorter than those generated by previous technologies, which complicates the task of taxonomic assignment and assembly. Finally, the taxonomic assignment of sequences by similarity searches may be especially challenging for viruses, given their high variability and genetic diversity and given the carriage of bacterial genes in bacteriophage genomes.

The very first computational tools for the analysis of metagenomic data, such as MG-RAST [79] or CAMERA [80], were initially developed to analyze microbial metagenomes. Following the development of viral metagenomics, new tools have been implemented to specifically analyze viral sequences and store viral metagenomes, such as MetaVir [81] or VIROME [82]. Moreover, new tools are continuously being developed to specifically address the new types of sequence data generated by NGS and to routinely assemble, compare and combine datasets from different sequencing technologies. For all these reasons, a wide variety of different tools have been developed and adopted, as well as in-house scripts. This has made the computational analysis of viromes fragmented and not comparable. No standardized protocols exist for the analysis of viral metagenomes, neither universal parameters for assembly and BLAST searches. The standardization and coordination of viral metagenomes analyses are required, as in the Human Microbial Project for bacterial communities. The main issues of computational analysis of human viral metagenomes are more widely addressed in the review reported in Chapter 2.

1.7 Objectives of my thesis work

My thesis work proceeds from the observation that, despite the recent advent of viral metagenomics, which represents an efficient tool for systematic characterization of viral communities, little is still known about viral diversity. New viruses are continuously being discovered and a high percentage of viral sequences generated by viral metagenomics results to be unknown. Although an increasing number of viral metagenomes have been generated since the first one by Breitbart et al. [11], viral metagenomics has not been applied to a wide variety of environments and/or samples. Environmental viral metagenomic studies mostly target marine water or freshwater and sampling is geographically biased towards temperate regions and North America (Fig. 1.3). Similarly, the sampling of human-associated samples in viral metagenomics is highly biased towards stool samples, sera/blood and naso-/oro- pharyngeal samples (Fig. 1.4). The availability of larger amounts of these materials and non-invasive methods of collection have made them the preferential targets. In addition, until now, all viral metagenomic studies have been performed on modern human samples and no systematic survey of viral diversity in ancient ones exists. Therefore, the initial purpose of my thesis work is to fill in the gap of knowledge on viral communities in the least investigated environments and human samples, in both pathological and non-pathological conditions, using viral metagenomics.

In the first part of my thesis, I review the main computational tools available for the analysis of viral metagenomes, with specific attention to their application in human clinical research (Chapter 2). The main steps of virome analysis are reported as well as relative issues. This review is not intended to be exhaustive of all the tools available in the literature but aims at providing a guide in this fragmented and rapidly evolving field.

In the second part of my thesis, I report my work on the computational analyses of environmental and human-associated viral metagenomes. This part is structured in two main chapters. The first chapter illustrates the investigation using viral metagenomics of four perennial bodies of water in the Sahara desert (Chapter 3). This is the first viral metagenomic study investigating viral diversity in the Sahara, the largest hot desert in the world. Moreover, it is the first viral metagenomic study in the African continent performed using high-throughput sequencing. In the following chapter, I present three viral metagenomic studies of human-associated viral communities in non-pathological and pathological conditions (Chapter 4). One study (Chapter 4.1) is the analysis of a 14^{th} century human coprolite from Belgium and it represents the first viral metagenomic study of an ancient specimen. This type of specimen is assumed to represent normal human gut viral flora in the Middle Ages. The human gut is an open system and the associated viral flora is thus supposed to be influenced by environmental viral communities and to be more diverse than in other body sites [83]. Therefore, this study may be considered to be at the interface between environmental and human-associated viral metagenomics. The two remaining human-associated studies involve clinical research, as they investigate human viral flora in disease states. In particular, one study (Chapter 4.2) focuses on human pericardial fluids with the purpose of describing associated viruses and discovering new or unexpected ones that could be responsible for some idiopathic pericarditis cases. The other study (Chapter 4.3) is a functional-level investigation of viral metagenomes from cystic fibrosis patients sputa, previously described by Willner et al [60, 84]. Its aim was to identify and analyze specific antimicrobial resistance genes carried by bacteriophages and better understand the emergence of multidrug-resistance bacteria in the airways of cystic fibrosis patients.

Chapter 2

Computational tools for viral metagenomics and their application in clinical research

2.1 Review. Computational tools for viral metagenomics and their application in clinical research

Computational tools for viral metagenomics and their application in clinical research.
Fancello Laura[1], Raoult Didier[1], Desnues Christelle[1*].

Published in Virology. 2012 Dec 20; 434(2):162-74.

[1] Aix Marseille University, URMITE, UM63, CNRS 7278, IRD 198, Inserm 1095, 13005 Marseille, France.
[*] Corresponding author. Email: christelle.desnues@univ-amu.fr

Virology 434 (2012) 162–174

Contents lists available at SciVerse ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Review

Computational tools for viral metagenomics and their application in clinical research

L. Fancello, D. Raoult, C. Desnues *

Aix Marseille University, URMITE, UM63, CNRS 7278, IRD 198, Inserm 1095, 13005 Marseille, France

A R T I C L E I N F O

Available online 11 October 2012

Keywords:
Virus
Metagenomics
Computational tools
Clinical research
Virome
Emerging disease

A B S T R A C T

There are 100 times more virions than eukaryotic cells in a healthy human body. The characterization of human-associated viral communities in a non-pathological state and the detection of viral pathogens in cases of infection are essential for medical care and epidemic surveillance. Viral metagenomics, the sequenced-based analysis of the complete collection of viral genomes directly isolated from an organism or an ecosystem, bypasses the "single-organism-level" point of view of clinical diagnostics and thus the need to isolate and culture the targeted organism. The first part of this review is dedicated to a presentation of past research in viral metagenomics with an emphasis on human-associated viral communities (eukaryotic viruses and bacteriophages). In the second part, we review more precisely the computational challenges posed by the analysis of viral metagenomes, and we illustrate the problem of sequences that do not have homologs in public databases and the possible approaches to characterize them.

Contents

* Corresponding author.
E-mail address: christelle.desnues@univ-amu.fr (C. Desnues).

0042-6822/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.virol.2012.09.025

Viral infections and the need for better viral discovery tools

Viral infections may become more prevalent in the future as multiple factors contribute to the emergence of new viral pathogens (Delwart, 2007; Wang, 2011). The expansion of the human population has led to the removal of barriers between animal and human communities, which favors the development of zoonoses. In addition, modern immunosuppressive therapies create favorable environments for the replication of viruses that are not commonly pathogenic. Furthermore, the spread of viruses worldwide is promoted by globalization and climate change, which extend the active ranges for some viral vectors, and there still exist several common pathologies, such as encephalitis and many respiratory syndromes, for which extensive classical diagnostic testing has failed to determine the etiology and which are thought to be of viral origin (Glaser et al., 2003; Quan et al., 2007).

Thus, an improved detection of newly emerging and re-emerging viruses and a systematic characterization of the full range of viruses that infect humans are needed (Anderson et al., 2003).

Classical methods of viral detection have several limitations. First, most of them are based on isolation and culture of the viral pathogen, but frequently the virus or its host cannot be cultivated under laboratory conditions, or the virus does not exhibit its characteristic cytopathic effects in culture (Specter, 1992). Moreover, these methods target known agents, and they are thus unsuitable for the detection of unexpected pathological agents or for the discovery of new ones. Immunological assays, for example, fail to identify unexpected or unknown viruses because such viruses are usually too divergent to cross-react. With respect to molecular tools, viruses lack a universally conserved genetic marker to target, and PCR assays directed towards conserved sequences within viral groups can only identify close variants of those groups (Staheli et al., 2011; Rose et al., 1998). Although the use of a wide set of different and highly degenerate primers has allowed the identification of numerous viruses (Culley et al., 2003), it does not allow a systematic and comprehensive screening to determine the identity of every virus that may be present.

Viral metagenomics and its first applications

Metagenomics, which is commonly defined as the sequenced-based analysis of the whole collection of genomes directly isolated from a sample (Handelsman et al., 1998), overcomes the principal limitations of the classical tools for viral detection. In fact, unlike traditional techniques for microbial and viral identification, metagenomics does not require prior isolation and clonal culturing for species characterization, nor does it rely on previous assumptions about what organisms are expected to be present or the genomic sequences that are to be targeted. Thus, it is particularly suitable to provide a global overview of the community diversity (species richness and distribution) and functional (metabolic) potential and to identify new species. In principle, it allows the identification of any organism, including those commonly not detected because they are difficult to isolate and grow under laboratory conditions. Such organisms are estimated to constitute between 90% and 99% of microbial species (Rappé and Giovannoni, 2003; Pace, 1997). Indeed the method of viral isolation, library preparation and sequencing affects the type of viruses which are retrieved. These issues have to be considered when analyzing the taxonomical profile of a metagenome and will be discussed later (see “General considerations on technical issues and potential biases in metagenome preparation”).

Metagenomics has a wide variety of applications from ecology and environmental sciences (Breitbart et al., 2002; Dinsdale et al., 2008) to the chemical industry (Lorenz and Eck, 2005) and human health (Turnbaugh et al., 2007; Ravel et al., 2010; Sullivan et al., 2011; Nakamura et al., 2009; Minot et al., 2011). Historically, it was first associated with the study of uncultured microbial organisms (bacteria and archaea) in environmental samples (Handelsman et al., 1998; Hugenholtz and Tyson, 2008). More recently, it has also been applied to the characterization of viral communities, a task that it is particularly suited for because the small size of viral genomes makes their coverage more comprehensive using the same number of metagenomic sequences. The first example of viral metagenomics was performed by Breitbart et al. in 2002. This study revealed that viral diversity had been widely underestimated because, in approximately 200 l of marine water, more than 7000 different viral genotypes were found. This high degree of viral genetic diversity has been confirmed by further metagenomic studies of marine water (Angly et al., 2006), marine sediments (Breitbart et al., 2004) and freshwater (Lopez-Bueno et al., 2009). Today, viruses are considered the most abundant and diverse living forms on earth (Culley et al., 2006; Suttle, 2005). Their diversity has been explored by metagenomics in a wide variety of environments: oceans (Williamson et al., 2008), stromatolites (Desnues et al., 2008), acidic hot springs (Rice et al., 2001), and subterranean and hypersaline environments (Dinsdale et al., 2008).

Identifying human-associated viral communities (the human virome)

A preliminary step in identifying viral agents that cause disease is the characterization of the viral microflora associated with humans in a non-pathological state. To date, only a few viral metagenomic studies have been performed on human samples. Moreover, due to the limited availability and size of human samples, most of these studies used fecal samples (Reyes et al., 2010; Breitbart et al., 2008,2003; Minot et al., 2011,2012; Zhang et al., 2006; Kim et al., 2011).

The first contribution to the assessment of the human virome by metagenomics was made in 2003 by Breitbart et al. who studied the DNA virus community that was associated with the human gut through partial shotgun sequencing of the feces of a healthy adult. Most of the sequences generated were unknown (59% according to a tblastx search against the Genbank non-redundant database with an E-value $< 1e-03$). Among the identifiable viral sequences, the majority were phages (Breitbart et al., 2003). The community was estimated to have a high richness (approximately 1200 different genotypes) and diversity as estimated by the Shannon–Wiener index ($H' = 6.4$ nats) which determines species diversity on the basis of both the number of species and the relative contribution of each of these species to the total number of individuals in a community. Breitbart et al. performed an analogous study in 2008 using the feces of a 1-week-old infant. Similarly to the 2003 study, an elevated percentage of unknown sequences (66%) and a significant abundance of phages were found. Similar observations were also reported by two recent studies on the DNA virome of the human gut (Reyes et al., 2010; Minot et al., 2011) in which the percentage of unknown sequences was 81% and 98%, respectively, and phages dominated the viral community. However, the richness and diversity of these viral communities were significantly lower in comparison with the results obtained by Breitbart in 2003 and in particular to the 1-week-old infant, whose virome richness was 8 genotypes and whose Shannon–Wiener index was only 1.63 nats. In addition to the DNA viruses, the RNA viruses of the human gut have also been studied (Zhang et al., 2006; Nakamura et al., 2009). In a study performed using stool samples from two healthy adults, Zhang et al. found that only 8.9% of the sequences were unknown (tblastx search with $E < 1e^{-03}$) and that among the identifiable viral sequences there was an insignificant number of phages. The majority of the identifiable viruses were plant viruses (91.5%). Among these viruses, they found viruses that infect consumable crops and fruits, which were most likely introduced through consumption of contaminated produces. They also observed that the viral community was dynamic and that it changed substantially in the same individual over time (Zhang et al., 2006).

Few other body sites have been targeted by viral metagenomics. In 2005, Breitbart and Rohwer analyzed the DNA virus communities associated with blood samples from healthy donors, and they were able to recover sequences from a novel anellovirus whose presence in the general population was then confirmed by specific PCR on a pool of 100 blood donors (Breitbart and Rohwer, 2005). In 2010, Willner et al. analyzed the DNA virus community of the human oral cavity using oropharyngeal swabs and showed that it was dominated by phages; the only eukaryotic virus detected was Epstein–Barr virus (Willner et al., 2010). A comparative study between patients affected by cystic fibrosis and healthy individuals showed that, in a non-disease state, the DNA virus community populating the sputum, which should be representative of the human respiratory tract, was again dominated by phages; among the eukaryotic viruses detected were adenoviruses, herpesviruses and poxviruses (Willner et al., 2009). Moreover, different individuals presented different viral communities, which likely were representative of a random sample of the inhaled organisms from the exterior environment; these viral particles are thought to establish transient infections that are rapidly cleared by the immune system or to be simply removed from the airway by mucociliary clearance. Interestingly, these communities were transient from a taxonomic point of view but constant with respect to the metabolic functions encoded. The estimated richness was 243 different genotypes, and the diversity, as measured by the Shannon–Wiener index, was as low as 4.83 nats.

A human salivary virome has also been described (Pride et al., 2011). Saliva samples from five healthy human subjects were studied over a 2- to 3-month period. The viral communities were dominated by bacteriophages, in contrast to the communities from human stool samples or the respiratory tract, and were likely the result of environmental influences. More than 122 thousands of homologs to genes involved in bacterial pathogenicity were identified in the salivary virome. This suggests that the bacteriophages contained in the saliva may serve as a reservoir of virulence-associated genes in the human oral environment.

Today, the assessment of the human virome in the non-disease state is still widely incomplete. Viral metagenomic studies characterizing the common “viral flora” associated with humans in the non-disease state need to be continued because they constitute a reference point in viral metagenomic clinical investigations. Indeed, they provide a baseline against which clinical samples can be compared to identify novel or divergent human viruses and assess which viruses are potentially responsible for idiopathic human diseases.

Bacteriophages in the human virome

Metagenomic studies aimed at characterizing the human virome have noted the prevalence and ubiquity of bacteriophages (viruses of bacteria) in humans. The vast majority of human viruses recovered by metagenomics were identified as viruses of bacteria, as shown in salivary (Pride et al., 2011), respiratory tract (Willner et al., 2009), gastrointestinal tract (Reyes et al., 2010) and oropharyngeal samples (Willner et al., 2010).

It is estimated that approximately 10^{13} to 10^{15} bacteriophages populate the human body (Haynes and Rohwer, 2011). These bacteriophages may have a substantial role in shaping and regulating human bacterial communities through lysis and horizontal gene transfer; a similar role has already been shown in environmental bacterial communities (Letarov and Kulikov, 2009; Weinbauer, 2004; Breitbart et al., 2004). Thus, they are also thought to be able to influence healthy and disease states in humans by, for example, eradicating certain bacteria or by conferring on bacteria a new pathogenic phenotype (Breitbart and Rohwer, 2005). Metagenomic analysis of viral communities populating the human oropharynx has suggested that bacteriophages are important reservoirs of virulence genes, such as the platelet-binding factors *pblA* and *pblB*, for oropharyngeal bacteria. Moreover, considerable differences were observed in the human respiratory tract between the bacteriophage communities associated with healthy subjects and the communities of cystic fibrosis patients (Willner et al., 2009). Antibiotic resistance genes were also found in bacteriophages colonizing cystic fibrosis patients, which could be passed through horizontal gene transfer to other bacterial communities and make those bacteria resistant. This phenomenon may represent a potential new therapeutic target to prevent the emergence of multidrug-resistant bacteria, which is a major problem in the treatment of cystic fibrosis patients (Fancello et al., 2011).

Clinical applications: discovery of human pathogens

The first application of viral metagenomics to human clinical research was in 2008 when Palacios et al. used the 454/Roche pyrosequencing platform to detect the pathogen responsible for a cluster of fatal transplant-associated diseases and identified a new arenavirus that was transmitted through solid-organ transplantation (Palacios et al., 2008). Since that initial study, viral metagenomics has led to the discovery of other previously unknown and potentially pathogenic viruses in stool samples (Victoria et al., 2009; Sullivan et al., 2011; Finkbeiner et al., 2008; Holtz et al., 2008), nasopharyngeal aspirates (Allander et al., 2005), serum/blood samples (Sullivan et al., 2011; Briese et al., 2009; McMullan et al., 2012) and a frontal lobe biopsy (Quan, 2010) collected from patients affected by idiopathic diseases. An overview of viral metagenomics studies on human clinical samples is provided in Table 1.

The interest in applying viral metagenomics to human patients comes not only from its capacity to identify new viruses that could potentially be implicated in a targeted disease but also from its capacity to confirm the presence of known pathogenic viruses even at concentrations lower than the levels detectable by PCR (Nakamura et al., 2009). Moreover, metagenomics can also highlight unexpected tropisms of known viruses and the potential pathogenicity of known viruses that are not suspected in the studied disease and thus are not targeted by standard diagnostic tests. An example is the implication of yellow fever virus in the hemorrhagic fever outbreak in October, 2010, in Uganda (McMullan et al., 2012). Also in 2010, Greninger et al. demonstrated that metagenomics was an efficient approach to rapidly identify and characterize the full genome of a flu virus without a priori information (Greninger et al., 2010). Clinical applications of viral metagenomics can also give important clues about which therapeutic measures to develop. For example, the metagenomic study of the viral communities populating human lungs in cystic fibrosis patients and healthy controls revealed that the diseased and non-diseased states are defined by their metabolic, rather than phylogenetic, profiles. Thus, therapeutic measures may be more effective if directed at changing the respiratory environment rather than targeting the dominant taxa (Willner et al., 2009, 2010).

General considerations on technical issues and potential biases in metagenome preparation

The way a viral metagenome is generated can widely affect the type of viruses retrieved and it should be taken into consideration for downstream analyses. Most of the biases related to

Table 1
Viral metagenomic studies on human samples for clinical application. Targeted disease, nucleic acids type (DNA or RNA viral genomes), sample type, eventual discovery of new viruses and sequencing technology are reported, as well as the method of viral particles isolation and the computational tools used for assembly and annotation.

Targeted disease	Nucleic acid	Samples	New virus discovered	Sequencing method	Viral particles isolation	Assembly	Annotation	Reference
Lower respiratory tract infection	DNA RNA	Nasopharyngeal aspirates	Parvovirus, coronavirus	Sanger	Ultracentrifugation; 0.22 μm filtering	Not performed	BLAST	(Allander et al., 2005)
Human merkel cell carcinoma	RNA	Cell carcinoma tissues (biopsies)	Polyomavirus	454/Roche	(Direct nucleic acids extraction)	Not performed	BLAST	(Feng et al., 2008)
Diarrhea	RNA	Stool	Astrovirus, torque teno virus, norovirus, picobirnavirus, enterovirus, nodavirus	Sanger	Centrifugation; 0.45 μm filtering	Not performed	BLAST	(Finkbeiner et al., 2008)
Acute respiratory infections and diarrhea	RNA	Nasopharyngeal aspirates, stool	–	454/Roche	Centrifugation	Not performed	BLAST, SSEARCH	(Nakamura et al., 2009)
Fatal transplant-associated disease	RNA	Brain, cerebrospinal fluid, serum, kidney, liver	Arenavirus	454/Roche	(Direct nucleic acids extraction)	CAP3 (Huang and Madan. 1999)	BLAST	(Palacios et al., 2008)
Hemorragic fever	RNA	Liver biopsies, serum	Arenavirus	454/Roche	(direct Nucleic acids extraction)	GCG Package (Accelrys, San Diego, CA, USA)	CLC RNA Workbench (CLC bio, Århus, Denmark)	(Briese et al., 2009)
Acute flaccid paralysis	DNA	Stool	Bocavirus, picornaviruses, circovirus, nodavirus, dicistroviruses	454/Roche, Sanger	Centrifugation; 0.45μm filtering	Sequencher (Gene Codes Corporation, Ann Arbor, MI USA)	BLAST	(Victoria et al., 2009)
Cystic fibrosis	DNA	Sputum	–	454/Roche	0.45 μm filtering; CsCl gradient	PHRAP (www.phrap.org)	BLAST, MG-RAST	(Willner et al., 2009)
Upper respiratory tract infection	RNA	Nasopharyngeal aspirates	–	Illumina	(Direct nucleic acids extraction)	Geneious (http://www.geneious.com)	BLAST	(Greinger et al., 2010)
Encephalitis	RNA	Frontal cortex (biopsy)	Astrovirus	454/Roche	(Direct nucleic acids extraction)	GreenPortal website (http://tako.cpmc.columbia.edu/Tools)	BLAST	(Quan, 2010)
Chronic fatigue syndrome	DNA RNA	Serum	–	454/Roche, Sanger	0.22 μm/0.45 μm filtering; ultracentrifugation	miraEST (Chevreux et al., 2004)	BLAST	(Sullivan et al., 2011)
Acute exacerbation of idiopathic pulmonary fibrosis	RNA	Bronchoalveolar lavage and serum	–	Illumina	(Direct nucleic acids extraction)	Not performed	MegaBLAST, BLAST	(Wootton et al., 2011)
Lower respiratory tract infections	DNA & RNA	Nasopharyngeal aspirates	Rhinovirus C	454/Roche	0.22 μm/0.45 μm filtering; ultracentrifugation	miraEST (Chevreux et al., 2004)	MegaBLAST, BLAST	(Lysholm et al., 2012)
Hemorragic fever	RNA	Serum	–	454/Roche	(Direct nucleic acids extraction)	Newbler (Roche); CLC (CLC bio, Aarhus, Denmark)	BLAST, MEGAN	(McMullan et al., 2012)
Cystic fibrosis	DNA	Lung tissue (biopsies)	–	454/Roche	0.45 μm filtering; CsCl gradient	CAP3 (Huang and Madan, 1999)	BLAST	(Willner et al., 2011)
Tropical febrile illness	DNA RNA	Serum	Circovirus	Illumina	(Direct nucleic acids extraction)	Not performed	BLAST	(Yozwiak et al., 2012)

metagenome preparation have already been discussed elsewhere (Morgan et al., 2010; Thomas et al., 2012). Here, we will briefly resume potential biased related to viral particles isolation, nucleic acid amplification and the sequencing technology used.

Viral particle isolation is usually performed by a combination of filtration and/or (ultra)centrifugation. Viral particles can be further purified onto a cesium chloride density gradient (Thurber et al., 2009). Sample filtering is often necessary to eliminate contamination by host cells and other non-viral cells. Because viral genomes generally are shorter than those of their eukaryotic or prokaryotic hosts, a minimal contamination would result in the preferential sequencing of those longer genomes which would "mask" viral sequences. However, most environmental metagenomic studies filter samples at 0.2 μm, which does not allow recovering large viruses and thus introduces a bias in the resulting metagenome taxonomic composition as already pointed out elsewhere (Thurber et al., 2009).

Another issue in metagenomes preparation is the need of a nucleic acids amplification step before sequencing as a result of the small amount of nucleic acids extracted from isolated viral particles. This is particularly critical for human-associated viral metagenomes as the volume of available sample may be more limited than in environmental studies. Nucleic acids may be amplified using the LASL (Linker Amplified Shotgun Library) method where the viral DNA (or the cDNA obtained from viral RNA genomes) is fragmented, ligated with an adapter and PCR amplified with a single primer specific to the adapter (Breitbart et al., 2002). Because the adapter ligation is only possible for dsDNA fragments, ssDNA viral genomes are not amplified and cannot be recovered in the metagenome (Kim and Bae, 2011).

Another common technique is the multiple displacement amplification (MDA), i.e. the isothermal amplification of the DNA (or the cDNA obtained from viral RNA genomes) by using random hexamers and the phi29 DNA polymerase. MDA is known to amplify more efficiently small circular DNA than linear DNA and preferentially ssDNA rather than dsDNA (Kim and Bae, 2011; Kim et al., 2011). It may also generate chimeras (Lasken and Stockwell, 2007) and introduce quantitative biases (Yilmaz et al., 2010). As different protocols can give different views on the diversity of the viral community studied, the biases introduced in the metagenome preparation have to be considered in downstream analyses and further comparative metagenomics.

Computational tools and algorithms in clinical viral metagenomics

One of the hardest challenges in metagenomic studies is sequence analysis, particularly because there is a large amount of data. For this reason, bioinformatics is essential to extract meaningful information from metagenomes. Computational analysis of metagenomes is particularly challenging in the case of viral community surveys. Viruses have an extremely high mutation rate, and they can be highly divergent, which hampers the identification of known homologs using similarity searches. In addition, viruses may exist in a proviral form, which complicates the task of distinguishing viral genomic sequences from host sequences. In the workflow for the analysis of a viral metagenome, the principal steps, aside from quality processing of raw reads, address the taxonomical and functional characterization of metagenomes, the gene prediction, the (partial) assembly of the genomes, the characterization of the community structure and diversity and comparisons of metagenomes. Due to the earlier and wider expansion of bacterial metagenomics over viral metagenomics, the first tools developed in this field were designed for the analysis of bacterial communities (Kunin et al., 2008; Wooley and Ye, 2010; Raes et al., 2007; Wooley et al., 2010) and may be unsuitable for the analysis of viral communities (see Fig. 1). The following sections present the computational tools and algorithms commonly used in viral metagenomics, with specific attention paid to clinical research.

Pre-processing and quality control

A typical metagenomic data workflow begins with quality control and the pre-processing of the raw reads produced by high-throughput sequencing technologies. The main goal is to create a high-quality metagenomic dataset that is faithfully representative of the genotypes present in the sample and of their relative abundances. Quality control includes the investigation of length, GC content, quality score, number of ambiguous bases "N" and the sequence complexity distribution of the reads. The criteria and methods for quality control are highly dependent on the sequencing technology used. These are general issues for all kinds of studies using data from high-throughput sequencing technologies and therefore are not the object of this review. Instead, we will treat here another pre-processing issue which is specific to metagenomics and in particular viral metagenomics:

		GENOMICS	(NON-VIRAL) METAGENOMICS	VIRAL METAGENOMICS
Pre-processing	Quality control	*Lucy (Li and Chou 2004); Phred (Ewing et al. 1998); Mothur (Schloss et al. 2009); TagDust (Lassmann et al. 2009); PIQA (Martínez-Alcántara et al. 2009); SolexaQA (Cox et al. 2010); PRINSEQ (Schmieder & Edwards 2011)*		
	Duplicates removal (artifacts from 454/Roche sequencing)	-	• Artificial duplicates may bias species relative abundances and the community structure *CD-HIT-454 (Li and Godzik 2006)*	
	Filtering	-		• Contamination from human genomic material is a major issue in human viral metagenomes. • Removal of bacterial-like reads must be carefully evaluated (part of these might come from genes of bacterial origin transferred to their phages or from excised prophages mistakenly annotated as bacteria).
			Deconseq (Schmieder and Edwards 2011)	
Annotation		-		• Annotation is challenging in viral metagenomics because of the high divergence of viral sequences and the limited number of reference genomes for similarity searches. • Annotation of bacterial-annotated reads must be carefully evaluated (part of these might come from genes of bacterial origin transferred to their phages or from excised prophages mistakenly annotated as bacteria).
			BLAST (Altschul et al. 1990), MEGAN (Huson et al. 2007); MG-RAST (Meyer et al. 2008);	
			TETRA (Teeling et al. 2004b); SPHINX (Mohammed et al. 2011); Phymm/PhymmBL (Brady and Salzberg 2011); PhyloPythia (McHardy et al. 2006)	*ProVide (Ghosh et al. 2011); MGTAXA (http://mgtaxa.jcvi.org); MetaVir (Roux et al. 2011); VIROME (Bhavsar et al. in preparation); VMGAP (Lorenzi et al. 2011)*
Assembly		*Euler (Pevzner et al. 2001) SOAPdenovo; (Li et al. 2009); AbySS (Simpson et al. 2009); ALLPATHs (Butler et al. 2008); Newbler (Roche); CLC (CLC bio, Aarhus, Denmark) Velvet (Zerbino et al. 2008); CAP3 (Huang et al. 1999)*	• Only partial assembly can be made, due to the complexity of a metagenomic sample and the limited depth of sequencing. We may reconstruct partial or complete genomes of dominant species in case of low community diversity. • The presence of conserved genomic regions between distantly related species may lead to the generation of chimeric contigs. • Assembly is complicated by the different coverage across species due to uneven species frequency in the sample. *Genovo (Laserson et al. 2011); MetaORFA (Ye and Tang 2009); MetaIDB (Peng et al. 2011); MetaVelvet (Namiki et al. 2012); Bambus 2 (Koren et al. 2011); MAP (Lai et al. 2012)*	
Estimation of genotype abundances, community structure and diversity		-	*GAAS(Angly et al. 2009); GRAMMy (Xia et al. 2011); GASiC (Lindner and Renard, 2012); Circonspect* (Angly et al. 2006); *PHACCS (Angly et al. 2005)*	

Fig. 1. Overview of the main issues and tools for computational analysis in genomics, metagenomics and viral metagenomics. For each step of the computational analysis, we reported specific issues, if any, relative to (non viral) metagenomic and viral metagenomics. Corresponding computational tools are reported in italic.

the presence of contaminating sequences in raw metagenomes. Filtering should be performed to obtain a metagenome that only contains sequences of interest (i.e., viral sequences). Filtering step limits misassemblies, and the resulting reduced size of the dataset speeds up the downstream analysis. There are two main sources of contamination: (i) primers and their eventual concatenations that are produced when metagenomes are generated by pre-amplification with primer-based methods (e.g., RNA virus communities generated by a Whole Transcriptome Amplification approach); and (ii) genomic material from organisms present in the sample that are not the targets of the metagenomic survey (e.g., host eukaryotic cells or prokaryotic material when the viral community is being studied). To eliminate contaminating primers, TagCleaner (Schmieder et al., 2010) and TagDust (Lassmann et al., 2009) can be used on 454/Roche- and Illumina-generated sequences, respectively. Contamination from genomic material can be removed after a BLAST search of all the reads that match with the genomes of the contaminating organisms; this task is automated by DeconSeq (Schmieder and Edwards, 2011). Recent studies have shown that viral metagenomes generated from human samples may contain over 90% host-derived sequences when nucleic acids are isolated without prior elimination of host or bacterial cells (Nakamura et al., 2009). Contamination from host genomic material can still represent a serious concern even in protocols that have been optimized to remove host and bacterial cells. For example, in a study by Willner et al., the percentage of human-derived sequences could be as high as 34% (Willner et al., 2009), although their protocol included a filtration step at 0.45 μm and a viral particle purification step using a cesium chloride gradient.

Human viral metagenomes are frequently dominated by sequences annotated as bacteria (Edwards and Rohwer, 2005; Rosario and Breitbart, 2011). Annotation and removal of bacterial-annotated reads must be carefully evaluated, as part of these might come from genes of bacterial origin transferred to their phages (Beumer and Robinson, 2005; Ghosh et al., 2008) or from excised prophages mistakenly annotated as bacteria. Recently, it has been proposed that the extensive presence of bacterial-like genes in viral metagenomes could be due to the presence of Gene-Transfer Agents (GTA) (Kristensen et al., 2010). These are phage-like particles found in a wide range of prokaryotes which are able to mediate gene transfers (Lang et al., 2012). Although similar to transducing bacteriophages, their production by a cell does not result from a phage infection, the amount of DNA packaged in GTAs is insufficient to encode the protein components of the particle itself and it contains a random piece of the genome of the producing cell. So far, the proportion of GTAs in viral metagenomes is unknown and the reason for such a large number of bacterial sequences retrieved from viral metagenomes is not clear (Lang et al., 2012).

Annotation, assembly and estimation of the community diversity and structure

Taxonomic identification, i.e., the assignment of each sequence to the genome from which it was generated, is one of the main goals of metagenomic studies. Indeed, it is a difficult task, especially for reads produced by high-throughput sequencing technologies that are only 50–500 nucleotides. Because of their short lengths, these reads are less informative and can be difficult to classify. An assembly step introduced prior the taxonomic classification could thus be very helpful by providing a better accuracy and sensitivity in the sequence assignments. At the same time, assembly itself constitutes a challenge in metagenomic studies which may be simplified by previous binning of sequences according to their putative taxonomic assignment (García Martín et al., 2006; Woyke et al., 2006). Taxonomic assignment and assembly, although described separately in the following sections, are deeply intertwined.

Taxonomic classification

Taxonomic classification is currently one of the most active fields in metagenomics. Several approaches have been developed and can be principally classified as either "similarity-based" methods or "composition-based" methods.

Similarity-based methods are most frequently used to describe the taxonomic profile of viral metagenomes. They are usually based on BLAST searches (Altschul et al., 1990), although other useful algorithms exist, including FAAST, which uses pyrosequencing flowpeak information to improve the alignment accuracy (Lysholm et al., 2011), or BLAT (Kent, 2002). Because most metagenomic sequences belong to unknown organisms, searches based on stringent *E*-values can yield too few classifiable sequences. In contrast, less stringent *E*-values can result in a high number of incorrect assignments. Thus, a few similarity-based taxonomic classifiers have been developed to evaluate taxonomic assignments that are based on alignment parameters. One of the most frequently used is MEGAN (Huson et al., 2007), a rank-flexible taxonomic classifier, i.e., a classifier that attempts to assign reads to the most appropriate taxonomic level when lacking sufficient phylogenetic information without forcing them to a particular rank to avoid misclassification of ambiguous reads. Although MEGAN has been adopted for viral metagenomic analysis (Kim et al., 2011; Yang et al., 2011), it was not specifically developed for this task. Conversely, ProViDE (Program for Viral Diversity Estimation) is a software tool based on a set of alignment parameter thresholds that are specific for viral metagenomic analysis (Ghosh et al., 2011). These thresholds take into account the patterns of sequence divergence and the non-uniform taxonomic hierarchies observed within/across viral taxonomic groups to increase the percentage of correct taxonomic assignments. Several biases affect the performance of similarity-based taxonomic classification methods. First, the content of public sequence databases is incomplete and only poorly reflects the existing biological diversity (McHardy and Rigoutsos, 2007). This is especially true in the viral world, which is mostly unknown; the majority of sequences obtained from viral metagenome projects has no homology to previously described sequences stored in public databases (Edwards and Rohwer, 2005) and cannot be classified by similarity searches. Moreover, viruses have high genetic diversity and divergence, which limits the probability of finding remote similarities between unknown and known viruses. Indeed, BLASTx, rather than BLASTn, searches are suggested for the classification of metagenomic sequences (Kunin et al., 2008). Because synonymous mutations are bypassed in the translation step, this method is more sensitive for recovering remote similarities. Additionally, the short lengths of metagenomic sequences can make reaching statistical significance in similarity searches difficult; prior assembly into longer sequences (called contigs) can thus be helpful in the taxonomic analysis. Finally, another drawback of these methods is that they are extremely time consuming.

Composition-based methods are taxonomic classification methods that are based on nucleotide composition. They are computationally faster than similarity-based methods, and they are useful for the classification of sequences that are highly divergent from the sequences in public databases. However, they depend on read length and have lower accuracy than similarity-based methods. They start from the assumption that the genome sequence composition varies among different organisms. Indeed, sequence composition is driven by taxonomy-related forces, such as the translational selection exerted on the synonymous codon usage of coding sequences, the polymerase nucleotide incorporation biases, the context-dependent mutation pressures and the optimal growth temperature of the organism (Karlin et al., 1997,1994; Perry and Beiko, 2010; Deschavanne et al., 1999). Genomic sequence composition has been shown to be sufficiently organism-specific to allow discrimination among several species (Kariin and Burge., 1995; Karlin et al., 1997) and thus to be employed for taxonomic classification. In addition, in the study by Teeling et al., the GC content and tetranucleotide signatures were adapted for the taxonomic classification of sequences from bacterial soil metagenomes (Teeling et al., 2004a). One of the first composition-based taxonomic methods, the TETRA software, is based on the computation of tetranucleotide usage patterns and performs comparisons with pre-computed patterns from organisms in a reference dataset (Teeling et al., 2004b). Unfortunately, this reference dataset does not contain viral genomes, and comparisons are not yet possible for viral metagenomes. More recently, programs based on the oligonucleotide composition of variable-length genome fragments have also been developed to achieve higher accuracy and sensitivity, including PhyloPythia (McHardy et al., 2007) and Phymm (Brady and Salzberg, 2011); other programs have been specifically developed to work correctly with metagenomes that exhibit both even and highly uneven species abundance distributions, e.g., Metacluster 3.0 (Leung et al., 2011) and Metacluster 4.0 (Wang et al., 2012). Finally, there are hybrid methods that combine similarity-based and composition-based approaches, including SPHINX (Mohammed et al., 2011) and PhymmBL (Brady and Salzberg, 2011). However, all of these methods are not suitable for viral metagenomes analysis because they are not trained or benchmarked on viral genomes. To our knowledge, the only composition-based tool specifically suited to predict the taxonomy of viral metagenomic sequences is MGTAXA (http://mgtaxa.jcvi.org), which was developed at the J. Craig Venter Institute and is freely available on the galaxy platform (http://galaxyproject.org). Based on Phymm, it is trained on viral genomes as well. Although composition-based methods have mostly been used for bacterial metagenomes, this approach has already been successfully tested on viral sequence classifications (Trifonov and Rabadan, 2010; Willner et al., 2009). Moreover, nucleotide composition analysis can also be used to infer the potential hosts of uncharacterized viral sequences. Indeed, the genome nucleotide composition of a virus is influenced by its host because it depends on the host for its replication (Kapoor et al., 2010). However, the compositional similarity between bacteriophage genomes and their hosts' genomes can be a confounding factor in the classification task. Therefore, the application of composition-based classification methods to viral metagenomes is a promising field of research, but further efforts in this area are needed.

Assembly

Assembly of metagenomic data is a complicated task due to the following factors: (i) the presence of several different genomes; (ii) non-species-specific contigs; (iii) conserved genomic regions that are shared between distantly related species; iv) the high frequency of polymorphisms and genome variation even at the subspecies level; (v) repeated regions; and (vi) the different coverages across species due to uneven species frequencies in the sample. The extreme richness and complexity of an environmental metagenomic sample and the limited depth of sequencing make virtually impossible to assemble all the individual genomes of a metagenomic project. However, it can be possible to reconstruct the genome(s) of the dominant species in the case of a highly uneven community. This is particularly true for viruses due to their shorter genome lengths. Such scenarios are of particular interest in metagenomics that is applied to clinical research because viral infection is expected to produce high viral loads of one dominant viral genotype over other residual viruses. Other interests of assembly are an improved length of assembled contigs compared to unassembled reads, which facilitates the taxonomic assignment and increases its accuracy in case of ambiguous reads. Moreover assembly may provide full-length coding sequences for subsequent analyses. Finally, assembly reduces the volume of the dataset and therefore the processing requirements.

So far, most studies have used *de novo* assemblers developed for single genome sequencing. The choice of assemblers depends on the average read length of the dataset, thus on the sequencing technology used. Phrap (http://www.phrap.org), Arachne (Batzoglou et al., 2002) and JAZZ (Aparicio et al., 2002) were for example used for Sanger-generated reads. Following the development of next-generation sequencing technologies and their application to metagenomic studies new versions of these *de novo* assembly tools and completely new algorithms were implemented to deal with the high throughput short reads generated by these technologies. Most of the new algorithms were based on the "de Bruijn graph" approach. Euler (Pevzner et al., 2001), ALLPATH (Butler et al., 2008), Velvet (Zerbino and Birney, 2008), SOAPdenovo (Li et al., 2009) and AbySS (Simpson et al., 2009) were initially developed for very short reads (<100 bp). The commercial assembler Newbler was implemented by Roche to specifically assemble 454-generated reads. For more information about these and further single genome NGS assemblers we address the reader to a specific review on this subject (Miller et al., 2010).

Still these assemblers were not specifically designed for metagenomes assembly. Some strategies had been adopted to make classic assemblers suitable for the analysis of metagenomic data, including the use of reference sequences (Rusch et al., 2007) and the pre-binning of reads on the basis of their sequence composition, which should be suggestive of their taxonomic classification (García Martín et al., 2006; Woyke et al., 2006). These methods may be affected by errors and may produce fragmented assemblies, hampering downstream analysis. These limits have been highlighted on simulated metagenomes (Pignatelli and Moya, 2011; Mavromatis et al., 2007).

More recently, new assembly algorithms have been implemented that specifically address the metagenome assembly problems. Genovo, for example, is an assembler based on the construction of a Bayesian probabilistic model of read generation from metagenomic samples, and it functions by discovering likely sequence reconstructions under this model (Laserson et al., 2011). Another approach is the assembly of translated ORFs rather than raw reads. This method, implemented by MetaORFA (Ye and Tang, 2009), simplifies the assembly task because it eliminates repeated regions (which are much more frequent in non-coding DNA than in ORFs) and thus avoids chimeric contigs. The assembly of sequences with synonymous mutations can also be easier because these mutations do not appear at the amino acid level, i.e., in translated ORFs. A further advantage is that downstream homology searches on longer peptide sequences assembled from ORFs are more sensitive and specific than searches using raw reads or single ORFs identified in an individual read. Another metagenome-specific assembler is Meta-IDB, which is not only capable of reconstructing longer contigs but also provides multiple alignments of similar contigs from different subspecies (variants) of the same species (Peng et al., 2011). Longer contigs can be produced because of two of the program's strengths: (i) its efficiency in eliminating genomic regions that are common to multiple species, thus isolating species that are different from each other; and (ii) its capacity to produce a unique consensus for different variants of the same subspecies instead of different contigs. Variations of this consensus are then represented by a multiple sequence alignment. Similarly to Meta-IDB (Peng et al., 2011), MetaVelvet (Namiki et al., 2012) and Bambus2 (Koren et al., 2011) focus on the detection of genomic repeats, which can generate chimeric sequences, and on the detection of polymorphisms, which can fragment the assembly into multiple contigs that represent different variants of the same subspecies (Koren et al., 2011). Moreover, Bambus2 is capable of using mate-paired data for metagenome scaffolding (i.e., the process through which read pairing information is used to order and orient the contig along a chromosome). Bambus 2 is used for the scaffolding step of the assembly process and is compatible with the output of most modern assemblers. Finally, among de novo assemblers specifically implemented for metagenome assembly, we can cite MAP (Metagenomic Assembly Program) which is developed for Sanger and 454/Roche generated reads (Lai et al., 2012). It uses mate pairs information to construct contigs when repeats confound the assembly.

Genotype abundances, community diversity and structure

An application of taxonomic classification and assembly is the characterization of the community's diversity and structure, which relies on estimating the number of different genotypes in the sample (richness) and defining their relative abundances and distribution (evenness) among the metagenomes. Simple read counts are often erroneously used to indicate relative abundances of different genotypes or different protein families within a metagenome. Indeed, metagenomic sequences only are a subset of the genomic sequences present in the sample and are obtained in a stochastic manner through high-throughput sequencing. Thus, longer genomes have a higher probability of being sequenced. Moreover, metagenomes usually contain high percentages of unknown sequences, which are usually not accounted for in the results of similarity-based taxonomic classification methods and which, conversely, should be considered in diversity estimates. The problem of the accurate estimation of species' relative abundances has been addressed by the GAAS tool. GAAS (Genome relative Abundance and Average Size) is a freely available tool fundamentally based on the assumption that the probability that a genome will be sequenced in a metagenomic study is directly proportional to its length (Angly et al., 2009). Thus, it performs sequence similarity searches and normalizes the number of reads recovered for a specific genome to the length of that genome, thus achieving more precise estimates. The accuracy of GAAS depends on the frequency of the ambiguous taxonomic assignment of reads (i.e., reads that cannot be reliably assigned to a unique genome) as it weights hits only by *E*-value (Xia et al., 2011; Lindner and Renard, 2012). The more recent GRAMMy tool (Genome Relative Abundance estimates based on Mixture Model theory) filters hits by *E*-value, alignment length and identity rate, and it manages ambiguous read assignments in a probabilistic way (Xia et al., 2011). It performs taxonomic assignment and computes the probability that each read is assigned to one of the reference genomes. Estimates of relative abundances as well as log-likelihood and standard error are then computed by maximum likelihood method. A different approach is implemented by GASiC (Genome Abundance Similarity Correction) (Lindner and Renard, 2012). This tool assumes that similarities among reference genomes are one of the major sources of ambiguities in reads assignments. Thus it computes abundances on the basis of reads alignments to reference genomes and then it directly uses observations on reference genomes similarities to correct the observed abundances. The community structure and diversity of viral communities can be estimated from metagenomic data using the Circonspect (Angly et al., 2006) and PHACCS tools (Angly et al., 2005). Circonspect uses an external assembly program and a bootstrap technique to automate the generation of the contig spectrum, which is the count of the number of contigs of each different size in an assembly. It relies on the assumption that the larger the contigs in the contig spectrum are for one genotype, the higher is the number of copies and the more abundant is this genotype. Thus, a highly diverse metagenome is supposed to produce a high number of small contigs and vice versa for a less diverse one. The contig spectrum is used as an input by PHACCS (PHAge Communities from Contig Spectrum) along with the average genome size estimated by GAAS to mathematically model the structure of viral communities and make predictions about diversity. Indeed, because not all sequences are entirely sequenced in a metagenomic survey, it predicts diversity by constructing models of species' relative abundances from available data and then extrapolating the diversity expected at an infinite sampling effort. In this way, it gives estimates of community richness, evenness and diversity. Interestingly, the method uses all of the available information, i.e., both known and unknown sequences. Indeed, it is based on the contig spectrum, which is computed using the whole set of metagenomic sequences.

Statistical tools for the analysis of clinical metagenomic samples

Statistical considerations are essential for the correct interpretation of metagenomic data in a wide range of cases, such as accurately estimating species' relative abundances or the community diversity. Metagenome comparisons also require statistical tests to assess the significance of observed differences or normalization procedures to account for the different sizes of the compared metagenomes. Most tools in comparative metagenomics were specifically developed for phylogenetic comparisons and, in particular, for 16S rRNA gene metagenomic surveys. Other tools were then developed for random sequencing of high-throughput data, such as ShotgunFunctionalizeR (Kristiansson et al., 2009) for functional comparisons of metagenomes. This tool focuses on the abundance of gene families, i.e., sets of functionally similar genes. Changes in gene family abundances between metagenomes can be linked to functional differences based on their corresponding annotations. XIPE-TOTEC (Rodriguez-Brito et al., 2006) is a rapid and user-friendly non-parametric statistical test that is designed for pairwise comparisons. However, a common issue with these tools is their inability to address multiple comparisons. This is an essential task in viral metagenomics when applied to clinical research because it relies on the comparison of two populations (patients and controls), each comprising multiple samples. Furthermore, it is of vital interest to precisely identify what is the statistically significant differential feature between the two populations studied (patients and controls) when we aim to detect, for example, those viruses whose presence or absence contributes to human disease.

Recently, Metastats (White et al., 2009) and STAMP (Parks and Beiko, 2010) have been developed to identify differentially abundant features between metagenomes. Metastats has been specifically implemented for clinical metagenomic sample analyses, and it provides a robust statistical framework. Metastats normalizes data to account for differences in metagenome sizes, can be confidently applied to non-normally distributed data, applies multiple comparison corrections and handles sparse counts using Fisher's exact test. STAMP is another valuable tool that uses confidence intervals and effect size statistics (i.e., the magnitude of the observed difference). Confidence intervals are more informative than the more commonly used p-value. Effect size statistics are used to assess whether a differentially abundant feature is not only statistically significant (as indicated by the *p*-value) but also biologically relevant; arbitrarily small effects can have statistically significant *p*-values when the sample sizes are sufficiently large.

These methods are of paramount interest for the detection of differentially abundant features in clinical samples compared with healthy controls. However, the assessment of an observed correlation between a specific feature and the disease state is a much more complicated task. Disease-association studies are complicated by the wide range of different viral genotypes observed in many viral groups in which each genotype can be associated or not to different symptoms. In addition, many viral infections seem to cause symptoms only in a subset of individuals, and co-infections can further complicate the interpretation of the results. The efficacy and informativeness of the described types of comparative analyses depend on the depths to which the functional and/or taxonomical annotations of viral metagenomes are performed. Although metagenome comparisons have yielded useful information to researchers about the differences, for example, between the viral communities associated with the sputa of healthy individuals and cystic fibrosis patients (Willner et al., 2009), they are still based on partial views of the sampled communities. Indeed, they do not take into consideration the unknown metagenomic sequences, which constitute a significant proportion of viral metagenomes. Conversely, Maxiphi (Angly et al., 2006) allows comparison of metagenomes at the sequence level rather than at the annotation level so that all of the reads are informative. Briefly, this tool assembles a random subset of sequences that equally represents each metagenome and analyzes the amount of overlap between sequences from different metagenomes, i.e., how many sequences from one metagenome overlap with sequences from another metagenome. The amount of this overlap indicates the degree of similarity between the two metagenomes. Then, it performs Monte Carlo simulations to estimate whether the differences are due to changes in the relative abundances of the viruses in the two metagenomes or to the presence of fundamentally different viruses. The output is the estimation of the "beta-diversity", which is based on the percentages of species that are shared between the metagenomes and the percentages of the permuted abundances of these species. However, we lack tools that precisely identify the statistically significant differential features between two metagenomes while considering unknown sequences in the comparison. Thus, further efforts should be applied to this area to improve metagenome annotation and decrease the percentage of unknown sequences.

Characterization of the "unknown"

The first metagenomic surveys performed on environmental viral communities showed that more than 60% of the sequences had no significant similarity to sequences stored in public databases (Edwards and Rohwer, 2005). A high percentage of unidentifiable sequences, classified as "unknown," are also found in metagenomic studies on viral communities that are associated with humans. The taxonomic identification and functional annotation of metagenomic sequences is a major problem, and until now it has been addressed mostly through BLAST searches. However, it is estimated that the use of existing BLAST-based approaches for taxonomic classification results in 10% to 90% of sequences being returned as unknown (Huson et al., 2007). Several factors contribute to the limited recovery rate of these approaches: (i) the short read lengths produced by high-throughput sequencing technologies; (ii) the incompleteness of public sequence databases; and (iii) sequencing errors. It has been proposed that integrating BLAST scores with information about gene adjacency will increase the efficacy of these similarity searches (Weng et al., 2010). In this approach, unclassified contigs or individual reads are blasted using less stringent E-values, and all of the top 250 hits are selected and compared in a pairwise fashion. Adjacent hits that are not consistent with the genomic arrangement of their reference genome are discarded, and between the remaining pairs the ones with the minimum *E*-value products are selected and used for taxonal classification of the sequence. However, this approach is based on the evolutionary conservation of gene order, which has been shown to be an important feature in prokaryotes but not in viruses (Tamames et al., 1997; Tamames, 2001).

Another approach to characterize unknown sequences by similarity-based methods derives from research on conserved protein domains, which are evolutionarily more conserved than the primary sequence and which can identify more remote similarities. Several databases of conserved protein domains exist, including Pfam, CDD, SMART and TIGRFAM (Punta et al., 2011; Marchler-Bauer et al., 2011; Letunic et al., 2011; Haft, 2003). These databases are commonly explored using BLAST or HMM-based alignments. The HMM-based alignment method has a high sensitivity for detecting remote homologs (Karplus et al., 1998). However, it cannot optimally classify sequences with frameshift errors. Thus, sequencing errors, such as those produced by high-throughput sequencing in metagenomic projects can hamper the identification of such domains. Recently, a new method of domain classification has been implemented that corrects for frameshift translations and is more suitable to metagenomic data analysis: HMM-FRAME (Zhang and Sun, 2011).

Another similarity-based approach for tentative sequence identification is phylogenetic analysis. This approach is based on the assumption that unknown genes, which are true remote homologs of known genes, should group with them in a phylogenetic tree. The construction of a phylogenetic tree for each unidentifiable sequence is rather inaccessible and time consuming for biologists without bioinformatics expertise. Thus, a user-friendly automated pipeline has been developed for the construction of multiple phylogenetic trees: Phylogena (Hanekamp et al., 2007). This tool allows automatic phylogenetic annotation of unknown sequences through an automated BLAST search of homologous sequences followed by the choice of a representative subset, computation of multiple alignments and construction of the phylogenetic tree. Still, this approach relies on the presence of (remote) homologs of the sequence in public databases and cannot be applied to highly divergent sequences. A radically different approach, independent from sequence similarity, is the use of composition-based methods for taxonomic classification, already cited in this review, which does not depend on the presence of homologs in public databases.

No more specific *in silico* methods are available, to our knowledge, for the characterization of unknown sequences. Some wet-lab experiments can be performed at this point, such as the cloning and expression of the unknown putative coding sequences followed by the characterization of the encoded protein's three-dimensional structure. Alternatively, it could be useful to study the metabolic function of the sequence by expressing it in *Escherichia coli* and observing the bacteria's growth in a chemostat culture. Recently, the cloning of sequences from a human gut microbiome and gulls metagenomes completed by an antibiotic resistance screening of the clones has allowed identifying several uncharacterized genes as antibiotic-resistance genes (Sommer et al., 2009; Martiny et al., 2011). However, given the large amount of unknown putative encoding sequences, the wet-lab approach is not an economical approach for characterizing all of them. Further *in silico* tools are thus needed to perform this task.

Next-generation sequencing technologies and the need for a common standardized pipeline analysis

The metagenomic field evolves in parallel with the development of sequencing technologies. The first metagenomic studies were based on Sanger sequencing, which yielded reads of approximately 800 bp. Later, the so-called "next-generation" sequencing (NGS) technologies were developed, which are currently capable of a much higher throughput, providing a more complete picture of the community and allowing discrimination between different sub-populations within the same sample. The first and still most used NGS platform is Roche/454 sequencing (Margulies et al., 2005). Recently, NGS such as ABI/SOLiD (Applied Biosystems by Life Technologies), the SMRT sequencing (Pacific Biosciences) and Illumina/Solexa (Bennett, 2004) which have even higher throughputs in comparison to Roche/454, have appeared. The SOLiD technology generates reads as short as 50 bp; thus, at the current state of the art, it is not used for metagenomic studies but only for whole genome re-sequencing (where deep sequencing allows correction of sequencing errors and detection of subpopulations) or RNA-sequencing projects.

The single-molecule real-time (SMRT) sequencing technology was developed by Pacific Biosciences in 2009 (Eid et al., 2009). In principle, it should allow to reach average read lengths as high as 3000 bp with instances of over 10,000 bp. However, accuracy of single reads is only at 85% which, up to now, makes the technology unusable in its current form for metagenomic applications. Illumina/Solexa technology, instead has already been successfully employed both in 16S rRNA metagenomic surveys on bacterial communities and in viral metagenomic projects (Greninger et al., 2010). It generates reads of about 100–150 bp and an output of up to 600 Gb per run. Its capacity to identify known and unknown viruses in biological samples has been compared to that of the Roche/454 platform in a blind metagenomic study on samples artificially spiked with viruses (Cheval et al., 2011). The results showed higher sensitivity for the detection of known viruses for the Illumina technology, which is most likely due to its considerably higher output compared to Roche/454. Conversely, Roche/454 sequencing performed better at the identification of unknown viruses because it generates longer reads, which allow easier assembly of de novo contigs of sufficient size to suggest the presence of a new virus.

The development of adapted bioinformatics tools still constitutes a bottleneck for the spread of the Illumina technology in the field of viral metagenomics. Most bioinformatics tools for metagenomic analyses were optimized for pyrosequencing-generated sequences and are not suitable for Illumina-generated reads whose shorter lengths complicate the taxonomic assignment of the reads and the assembly task. Moreover, we still need additional tools to routinely assemble or compare and combine data sets from different kinds of sequencing technologies, such as the recent Segminator II (Archer et al., 2012) and ngs_backbone softwares (Blanca et al., 2011).

Conclusions

The field of bioinformatics for metagenomics is very dynamic and new programs are continuously being created to manage the new NGS-generated data. Initial metagenomic studies used several tools previously developed for single genomic projects. However, it has become evident that metagenomics brings specific issues which have to be addressed by specific or adapted algorithms. Analyses that may be common with single genomics projects still present some specific issues when performed on metagenomic data. For instance, the assembly of a metagenome may be challenged by the presence of sequences from different organisms that share some genomic regions, further leading to the in silico generation of chimeric contigs. New assemblers have thus been specifically developed for metagenomic studies. In addition, some issues are specific to the nature of studied community (viruses, bacteria…). Hence, the computational tools developed initially for bacterial metagenomics may not be applicable to viral metagenomics and this is particularly true in the annotation field. Fig. 1 reports examples of tools developed for each step of a metagenomic analysis and the specific issues (if any) which have to be addressed in metagenomics (with emphasis on viral metagenomics). Indeed, a variety of different programs can be adapted for metagenomic analyses and frequently small in-house scripts are required. Presently, no common strategies have been established for the analysis of viral metagenomes and no universal standard parameters exist for assembly, BLAST searches or the quality trimming of reads. All of these factors make viral metagenomic analyses difficult to compare and difficult to reproduce. Standardization and coordination of efforts to analyze viral communities that are associated with humans are needed, which have already been undertaken in the Human Microbial Project for bacterial communities. In this view, although no completely exhaustive databases exist for viral metagenome submission and analysis, some platforms have been developed that allow for storage, public access and analysis of metagenomes, such as MetaVir (Roux et al., 2011) and VIROME (Wommack et al., 2012) and VMGAP (Viral MetaGenome Annotation Pipeline) for functional annotation (Lorenzi et al., 2011). Such initiatives constitute valuable first efforts towards data sharing and analysis standardization.

Acknowledgments

This work was funded by a Starting Grant number 242729 from the European Research Council to CD.

References

Allander, T., Tammi, M.T., Eriksson, M., Bjerkner, A., Tiveljung-Lindell, A., Andersson, B., 2005. Cloning of a human parvovirus by molecular screening of respiratory tract samples. Proc. Natl. Acad. Sci. USA 102, 12891–12896.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Anderson, N.G., Gerin, J.L., Anderson, N.L., 2003. Global screening for human viral pathogens. Emerging Infect. Dis. 9, 768–774.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J., Rohwer, F., 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6, 41.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C.A., Rohwer, F., 2006. The marine viromes of four oceanic regions. PLoS Biol. 4, e368.

Angly, F.E., Willner, D., Prieto-Davó, A., Edwards, R.A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D.A., Barott, K., Cottrell, M.T., Desnues, C., Dinsdale, E.A., Furlan, M., Haynes, M., Henn, M.R., Hu, Y., Kirchman, D.L., McDole, T., McPherson, J.D., Meyer, F., Miller, R.M., Mundt, E., Naviaux, R.K., Rodriguez-Mueller, B., Stevens, R., Wegley, L., Zhang, L., Zhu, B., Rohwer, F., 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput. Biol. 5, e1000593.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301–1310.

Archer, J., Baillie, G., Watson, S.J., Kellam, P., Rambaut, A., Robertson, D.L., 2012. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. BMC Bioinformatics 13, 47.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., Lander, E.S., 2002. ARACHNE: a whole-genome shotgun assembler. Genome Res. 12, 177–189.

Bennett, S., 2004. Solexa Ltd. Pharmacogenomics 5, 433–438.

Beumer, A., Robinson, J.B., 2005. A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. Appl. Environ. Microbiol. 71, 8301–8304.

Blanca, J.M., Pascual, L., Ziarsolo, P., Nuez, F., Cañizares, J., 2011. ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using Next Generation Sequence. BMC Genomics 12, 285.

Brady, A., Salzberg, S., 2011. PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat. Methods 8. (367–367).
Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., Rohwer, F., 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. USA 99, 14250–14255.
Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P., Rohwer, F., 2003. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. 185, 6220–6223.
Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., Rohwer, F., 2004. Diversity and population structure of a near-shore marine-sediment viral community. Proc. Biol. Sci. 271, 565–574.
Breitbart, M., Rohwer, F., 2005. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. BioTechniques 39, 729–736.
Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R.A., Felts, B., Mahaffy, J.M., Mueller, J., Nulton, J., Rayhawk, S., 2008. Viral diversity and dynamics in an infant gut. Res. Microbiol. 159, 367–373.
Briese, T., Paweska, J.T., McMullan, L.K., Hutchison, S.K., Street, C., Palacios, G., Khristova, M.L., Weyer, J., Swanepoel, R., Egholm, M., Nichol, S.T., Lipkin, W.I., 2009. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathogens 5, e1000455.
Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., Jaffe, D.B., 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 18, 810–820.
Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F., Berthet, N., Brisse, S., Moszer, I., Bourhy, H., Manuguerra, C.J., Lecuit, M., Burguiere, A., Caro, V., Eloit, M., 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J. Clin. Microbiol. 49, 3268–3275.
Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E.G., Wetter, T., Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14, 1147–1159.
Culley, A.I., Lang, A.S., Suttle, C.A., 2003. High diversity of unknown picorna-like viruses in the sea. Nature 424, 1054–1057.
Culley, A.I., Lang, A.S., Suttle, C.A., 2006. Metagenomic analysis of coastal RNA virus communities. Science 312, 1795–1798.
Delwart, E.L., 2007. Viral metagenomics. Rev. Med. Virol. 17, 115–131.
Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.
Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., Liu, H., Furlan, M., Wegley, L., Chau, B., Ruan, Y., Hall, D., Angly, F.E., Edwards, R.A., Li, L., Thurber, R.V., Reid, R.P., Siefert, J., Souza, V., Valentine, D.L., Swan, B.K., Breitbart, M., Rohwer, F., 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452, 340–343.
Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M.A., Nelson, K.E., Nilsson, C., Olson, R., Paul, J., Brito, B.R., Ruan, Y., Swan, B.K., Stevens, R., Valentine, D.L., Thurber, R.V., Wegley, L., White, B.A., Rohwer, F., 2008. Functional metagenomic profiling of nine biomes. Nature 452, 629–632.
Edwards, R.A., Rohwer, F., 2005. Opinion: viral metagenomics. Nat. Rev. Microbiol. 3, 504–510.
Eid, J., Fehr, A., Gray, J., Lyong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., de Winter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Hong, X., Kuse, R., Lacroix, Y., Lin, S., Lunquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.
Fancello, L., Desnues, C., Raoult, D., Rolain, J.M., 2011. Bacteriophages and diffusion of genes encoding antimicrobial resistance in cystic fibrosis sputum microbiota. J. Antimicrob. Chemother. 66, 2448–2454.
Feng, H., Shuda, M., Chang, Y., Moore, P.S., 2008. Clonal integration of a polyomavirus in human merkel cell carcinoma. Science 319, 1096–1100.
Finkbeiner, S.R., Allred, A.F., Tarr, P.I., Klein, E.J., Kirkwood, C.D., Wang, D., 2008. Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathogens 4, e1000011.
García Martín, H., Ivanova, N., Kunin, V., Warnecke, F., Barry, K.W., McHardy, A.C., Yeates, C., He, S., Salamov, A.A., Szeto, E., Dalin, E., Putnam, N.H., Shapiro, H.J., Pangilinan, J.L., Rigoutsos, I., Kyrpides, N.C., Blackall, L.L., McMahon, K.D., Hugenholtz, P., 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat. Biotechnol. 24, 1263–1269.
Ghosh, D., Roy, K., Williamson, K.E., White, D.C., Wommack, K.E., Sublette, K.L., Radosevich, M., 2008. Prevalence of lysogeny among soil bacteria and presence of 16S rRNA and trzN genes in viral-community DNA. Appl. Environ. Microbiol. 74, 495–502.
Ghosh, T.S., Mohammed, M.H., Komanduri, D., Mande, S.S., 2011. ProViDE: a software tool for accurate estimation of viral diversity in metagenomic samples. Bioinformation 6, 91–94.
Glaser, C.A., Gilliam, S., Schnurr, D., Forghani, B., Honarmand, S., Khetsuriani, N., Fischer, M., Cossen, C.K., Anderson, L.J., 2003. In search of encephalitis etiologies: diagnostic challenges in the California Encephalitis Project, 1998–2000. Clin. Infect. Dis. 36, 731–742.
Greninger, A.L., Chen, E.C., Sittler, T., Scheinerman, A., Roubinian, N., Yu, G., Kim, E., Pillai, D.R., Guyard, C., Mazzulli, T., Isa, P., Arias, C.F., Hackett, J., Schochetman, G., Miller, S., Tang, P., Chiu, C.Y., 2010. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS ONE 5, e13381.
Haft, D.H., 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31, 371–373.
Handelsman, J., et al., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5 (10), R245–R249.
Hanekamp, K., Bohnebeck, U., Beszteri, B., Valentin, K., 2007. PhyloGena—a user-friendly system for automated phylogenetic annotation of unknown sequences. Bioinformatics 23, 793–801.
Haynes, M., Rohwer, F., 2011. The Human Virome. In: Nelson, K.E. (Ed.), Metagenomics of the Human Body. Springer, New York, NY, pp. 63–77.
Holtz, L.R., Finkbeiner, S.R., Kirkwood, C.D., Wang, D., 2008. Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. Virol. J. 5, 159.
Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.
Hugenholtz, P., Tyson, G.W., 2008. Microbiology: metagenomics. Nature 455, 481–483.
Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Res. 17, 377–386.
Kapoor, A., Simmonds, P., Lipkin, W.I., Zaidi, S., Delwart, E., 2010. Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. J. Virol. 84, 10322–10328.
Kariin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290.
Karlin, S., Ladunga, I., Blaisdell, B.E., 1994. Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. USA 91, 12837–12841.
Karlin, S., Mrázek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179, 3899–3913.
Karplus, K., Barrett, C., Hughey, R., 1998. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846–856.
Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.
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.
Kim, M.S., Park, E.-J., Roh, S.W., Bae, J.-W., 2011. Diversity and abundance of single-stranded DNA viruses in human feces. Appl. Environ. Microbiol. 77, 8062–8070.
Koren, S., Treangen, T.J., Pop, M., 2011. Bambus 2: scaffolding metagenomes. Bioinformatics 27, 2964–2971.
Kristensen, D.M., Mushegian, A.R., Dolja, V.V., Koonin, E.V., 2010. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. 18, 11–19.
Kristiansson, E., Hugenholtz, P., Dalevi, D., 2009. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics 25, 2737–2738.
Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., Hugenholtz, P., 2008. A bioinformatician's guide to metagenomics. Microbiol. Mol. Biol. Rev. 72, 557–578.
Lai, B., Ding, R., Li, Y., Duan, L., Zhu, H., 2012. A de novo metagenomic assembly program for shotgun DNA reads. Bioinformatics 28, 1455–1462.
Lang, A.S., Zhaxybayeva, O., Beatty, J.T., 2012. Gene transfer agents: phage-like elements of genetic exchange. Nat. Rev. Microbiol. 10, 472–482.
Laserson, J., Jojic, V., Koller, D., 2011. Genovo: de novo assembly for metagenomes. J. Comput. Biol. 18, 429–443.
Lasken, R.S., Stockwell, T.B., 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol. 7, 19.
Lassmann, T., Hayashizaki, Y., Daub, C.O., 2009. TagDust—a program to eliminate artifacts from next generation sequencing data. Bioinformatics 25, 2839–2840.
Letarov, A., Kulikov, E., 2009. The bacteriophages in human- and animal body-associated microbial communities. J. Appl. Microbiol. 107, 1–13.
Letunic, I., Doerks, T., Bork, P., 2011. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 40, D302–D305.
Leung, H.C.M., Yiu, S.M., Yang, B., Peng, Y., Wang, Y., Liu, Z., Chen, J., Qin, J., Li, R., Chin, F.Y.L., 2011. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics 27, 1489–1495.
Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Yang, H., Wang, J., 2009. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20, 265–272.
Lindner, M.S., Renard, B.Y., 2012. Metagenomic abundance estimation and diagnostic testing on species level. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/gks803.
Lopez-Bueno, A., Tamames, J., Velazquez, D., Moya, A., Quesada, A., Alcami, A., 2009. High diversity of the viral community from an Antarctic Lake. Science 326, 858–861.
Lorenz, P., Eck, J., 2005. Outlook: metagenomics and industrial applications. Nat. Rev. Microbiol. 3, 510–516.
Lorenzi, H.A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L., Williamson, S.J., 2011. TheViral MetaGenome Annotation Pipeline(VMGAP):an automated tool

for the functional annotation of viral Metagenomic shotgun sequencing data. Stand. Genomic Sci. 4, 418–429.
Lysholm, F., Andersson, B., Persson, B., 2011. FAAST: Flow-space Assisted Alignment Search Tool. BMC Bioinformatics 12, 293.
Lysholm, F., Wetterbom, A., Lindau, C., Darban, H., Bjerkner, A., Fahlander, K., Lindberg, A.M., Persson, B., Allander, T., Andersson, B., 2012. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. PLoS ONE 7, e30875.
Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Zheng, C., Bryant, S.H., 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39, D225–D229.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.
Martiny, A.C., Martiny, J.B., Weihe, C., Field, A., Ellis, J.C., 2011. Functional metagenomics reveals previously unrecognized diversity of antibiotic resistance genes in gulls. Front. Microbiol. 2, 238.
Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., Kyrpides, N.C., 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat. Methods 4, 495–500.
McHardy, A.C., Rigoutsos, I., 2007. What's in the mix: phylogenetic classification of metagenome sequence samples. Curr. Opin. Microbiol. 10, 499–503.
McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I., 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods 4, 63–72.
McMullan, L.K., Frace, M., Sammons, S.A., Shoemaker, T., Balinandi, S., Wamala, J.F., Lutwama, J.J., Downing, R.G., Stroeher, U., MacNeil, A., Nichol, S.T., 2012. Using next generation sequencing to identify yellow fever virus in Uganda. Virology 422, 1–5.
Miller, J.R., Koren, S., Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics 96, 315–327.
Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., Bushman, F.D., 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. 21, 1616–1625.
Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D., Bushman, F.D., 2012. Hypervariable loci in the human gut virome. Proc. Natl. Acad. Sci. USA 109, 3962–3966.
Mohammed, M.H., Ghosh, T.S., Singh, N.K., Mande, S.S., 2011. SPHINX—an algorithm for taxonomic binning of metagenomic sequences. Bioinformatics 27, 22–30.
Morgan, J.L., Darling, A.E., Eisen, J.A., 2010. Metagenomic sequencing of an in vitro-simulated microbial community. PLoS ONE 5, e10209.
Nakamura, S., Yang, C.-S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., Iida, T., Nakaya, T., 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS ONE 4, e4219.
Namiki, T., Hachiya, T., Tanaka, H., Sakakibara, Y., 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/gks678.
Pace, N.R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734–740.
Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., Briese, T., Conlan, S., Quan, P.-L., Hui, J., Marshall, J., Simons, J.F., Egholm, M., Paddock, C.D., Shieh, W.-J., Goldsmith, C.S., Zaki, S.R., Catton, M., Lipkin, W.I., 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. 358, 991–998.
Parks, D.H., Beiko, R.G., 2010. Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26, 715–721.
Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2011. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27, i94–i101.
Perry, S.C., Beiko, R.G., 2010. Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives. Genome Biol. Evol. 2, 117–131.
Pevzner, P.A., Tang, H., Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA 98, 9748–9753.
Pignatelli, M., Moya, A., 2011. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. pLoS ONE 6, e19984.
Pride, D.T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R.A., Loomer, P., Armitage, G.C., Relman, D.A., 2011. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. ISME J. 6, 915–926.
Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R.D., 2011. The Pfam protein families database. Nucleic Acids Res. 40, D290–D301.
Quan, P.L., Palacios, G., Jabado, O.J., Conlan, S., Hirschberg, D.L., Pozo, F., Jack, P.J.M., Cisterna, D., Renwick, N., Hui, J., Drysdale, A., Amos-Ritchie, R., Baumeister, E., Savy, V., Lager, K.M., Richt, J.A., Boyle, D.B., García-Sastre, A., Casas, I., Perez-Breña, P., Briese, T., Lipkin, W.I., 2007. Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. J. Clin. Microbiol. 45, 2359–2364.
Quan, P.L., 2010. Astrovirus encephalitis in boy with X-linked agammaglobulinemia. Emerging Infect. Dis. 16, 918–925.
Raes, J., Foerstner, K.U., Bork, P., 2007. Get the most out of your metagenome: computational analysis of environmental sequence data. Curr. Opin. Microbiol. 10, 490–498.
Rappé, M.S., Giovannoni, S.J., 2003. The uncultured microbial majority. Annu. Rev. Microbiol. 57, 369–394.
Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., Brotman, R.M., Davis, C.C., Ault, K., Peralta, L., Forney, L.J., 2010. Colloquium Paper: vaginal microbiome of reproductive-age women. Proc. Natl. Acad. Sci. USA 108, 4680–4687.
Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., Gordon, J.I., 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466, 334–338.
Rice, G., Stedman, K., Snyder, J., Wiedenheft, B., Willits, D., Brumfield, S., McDermott, T., Young, M.J., 2001. Viruses from extreme thermal environments. Proc. Natl. Acad. Sci. USA 98, 13341–13345.
Rodriguez-Brito, B., Rohwer, F., Edwards, R.A., 2006. An application of statistics to comparative metagenomics. BMC Bioinformatics 7, 162.
Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 289–297.
Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M., Henikoff, S., 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res. 26, 1628–1635.
Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., Enault, F., 2011. Metavir: a web server dedicated to virome analysis. Bioinformatics 27, 3074–3075.
Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.-H., Falcón, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C., 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5, e77.
Schmieder, R., Lim, Y., Rohwer, F., Edwards, R., 2010. TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics 11, 341.
Schmieder, R., Edwards, R., 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS ONE 6, e17288.
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.
Sommer, M.O., Dantas, G., Church, G.M., 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 325, 1128–1131.
Specter, S, 1992. Clinical virology manual, 2nd ed. Elsevier, New York.
Staheli, J.P., Boyce, R., Kovarik, D., Rose, T.M., 2011. CODEHOP PCR and CODEHOP PCR primer design. Methods Mol. Biol. 687, 57–73.
Sullivan, P.F., Allander, T., Lysholm, F., Goh, S., Persson, B., Jacks, A., Evengård, B., Pedersen, N.L., Andersson, B., 2011. An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. BMC Microbiol. 11, 2.
Suttle, C.A., 2005. Viruses in the sea. Nature 437, 356–361.
Tamames, J., Casari, G., Ouzounis, C., Valencia, A., 1997. Conserved clusters of functionally related genes in two bacterial genomes. J. Mol. Evol. 44, 66–73.
Tamames, J., 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. 2, research0020-research0020.11.
Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., Glockner, F.O., 2004a. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol. 6, 938–947.
Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glöckner, F.O., 2004b. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5, 163.
Thomas, T., Gilbert, J., Meyer, F., 2012. Metagenomics—a guide from sampling to data analysis. Microb. Inf. Exp. 2, 3.
Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., Rohwer, F., 2009. Laboratory procedures to generate viral metagenomes. Nat Protocols 4, 83–470.
Trifonov, V., Rabadan, R., 2010. Frequency analysis techniques for identification of viral genetic data. MBio 1. (e00156-10–e00156-17).
Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I., 2007. The Human Microbiome Project. Nature 449, 804–810.
Victoria, J.G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S., Delwart, E., 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. 83, 4642–4651.
Wang, L.F., 2011. Discovering novel zoonotic viruses. N. South Wales Public Health Bull. 22, 113.
Wang, Y., Leung, H.C., Yiu, S.M., Chin, F.Y., 2012. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. J. Comput. Biol. 19, 241–249.
Weinbauer, M.G., 2004. Ecology of prokaryotic viruses. FEMS Microbiol. Rev. 28, 127–181.
Weng, F.C., Su, C.-H., Hsu, M.-T., Wang, T.-Y., Tsai, H.-K., Wang, D., 2010. Reanalyze unassigned reads in Sanger based metagenomic data using conserved gene adjacency. BMC Bioinformatics 11, 565.
White, J.R., Nagarajan, N., Pop, M., 2009. statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. 5, e1000352.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C.S., Sutton, G., Frazier, M., Venter, J.C., 2008. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3, e1456.
Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., Rohwer, F., 2009. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS ONE 4, e7370.
Willner, D., Furlan, M., Schmieder, R., Grasis, J.A., Pride, D.T., Relman, D.A., Angly, F.E., McDole, T., Mariella, R.P., Rohwer, F., Haynes, M., 2010. Colloquium Paper: metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc. Natl. Acad. Sci. USA 108, 4547–4553.
Willner, D., Haynes, M.R., Furlan, M., Hanson, N., Kirby, B., Lim, Y.W., Rainey, P.B., Schmieder, R., Youle, M., Conrad, D., Rohwer, F., 2011. Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. Am. J. Respir. Cell Mol. Biol. 46, 127–131.
Wooley, J.C., Godzik, A., Friedberg, I., 2010. A primer on metagenomics. PLoS Comput. Biol. 6, e1000667.
Wooley, J.C., Ye, Y., 2010. Metagenomics: facts and artifacts, and computational challenges. J. Comput. Sci. Technol. 25, 71–81.
Wommack, K.E., Bhavsar, J., Polson, S.W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S., Nasko, D.J, 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand. Genomic. Sci. 6, 421–433.
Wootton, S.C., Kim, D.S., Kondoh, Y., Chen, E., Lee, J.S., Song, J.W., Huh, J.W., Taniguchi, H., Chiu, C., Boushey, H., Lancaster, L.H., Wolters, P.J., DeRisi, J., Ganem, H.R., Collard, H.R., 2011. Viral infection in acute exacerbation of idiopathic pulmonary fibrosis. Am. J. Respir. Crit. Care Med. 183, 1698–1702.
Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O., Boffelli, D., Anderson, I.J., Barry, K.W., Shapiro, H.J., Szeto, E., Kyrpides, N.C., Mussmann, M., Amann, R., Bergin, C., Ruehland, C., Rubin, E.M., Dubilier, N., 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443, 950–955.
Xia, L.C., Cram, J.A., Chen, T., Fuhrman, J.A., Sun, F., 2011. Accurate genome relative abundance estimation based on shotgun metagenomic reads. PLoS ONE 6, e27992.
Yang, J., et al., 2011. Unbiased parallel detection of viral pathogens in clinical samples using a metagenomic approach. J. Clin. Microbiol. 49 (10), 3463–3469.
Ye, Y., Tang, H., 2009. An ORFome assembly approach to metagenomics sequences analysis. J. Bioinformatics Comput. Biol. 7, 455–471.
Yilmaz, S., Allgaier, M., Hugenholtz, P., 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat. Methods 7, 943–944.
Yozwiak, N.L., Skewes-Cox, P., Stenglein, M.D., Balmaseda, A., Harris, E., DeRisi, J.L., 2012. Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Neglected Trop. Dis. 6, e1485.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.
Zhang, T., Breitbart, M., Lee, W.H., Run, J.-Q., Wei, C.L., Soh, S.W.L., Hibberd, M.L., Liu, E.T., Rohwer, F., Ruan, Y., 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 4, e3.
Zhang, Y., Sun, Y., 2011. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. BMC Bioinformatics 12, 198.

Chapter 3

Environmental viral metagenomics

3.1 Article 1. Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara

Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara.

Fancello Laura[1], Trape Sebatien[1], Robert Catherine[1], Boyer Michael[1], Popgeorgiev Nikolay[1], Raoult Didier[1], Desnues Christelle[1*]

Published in ISME Journal. 2013 Feb; 7(2):359-69.

[1] Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE CNRS-IRD UMR 7278, Aix-Marseille Université, Faculté de médecine, 27, Bd Jean Moulin, 13385 MARSEILLE France.

[*] Corresponding author. Email: christelle.desnues@univ-amu.fr

Preamble to article 1

Little is known about viral diversity and ubiquity in arid regions, such as hot deserts, where unfavorable conditions limit the development and the activity of eukaryotic life, especially in Africa. In 2007, a viral metagenome associated with the sand from a hot desert in Kansas was generated by Sanger sequencing [35]. The data revealed high local diversity as well as significant differences with the viral communities from other environments. Another viral metagenome associated with an ice-covered lake from the cold desert of Antarctica displayed unexpected high richness with eukaryotic ssDNA viruses in spring and dsDNA bacteriophages as well as phycodnaviruses and mimiviruses in the summer [33]. In Africa, the only available viral metagenome was generated by Sanger sequencing from a hypersaline lake in Senegal and mostly contained haloarcheal viruses or sequences previously retrieved in viral metagenomes from solar salterns [34].

Here, we present the first metagenomic study on viral communities associated with freshwater perennial bodies of water (ponds) in the hot desert of Sahara, in Africa. The Sahara was rich in aquatic environments in the early Holocene period, of which a few traces persist today [85]. Previous electron microscopy and pulsed field gel electrophoresis studies on Sahara sands have revealed the presence of bacteriophages and 45-350 kb viral genomes, respectively [86, 87]. In this work, we investigated the ubiquity and diversity of viruses in four ponds. Three of them (Ilij, Molomhar and Hamdoun) belong to the same hydrologic basin, whereas the fourth (El Berbera) belongs to a different basin. Viromes were generated by filtering the water through a large pore size filter (0.45 μm) followed by the purification of viral particles on a cesium chloride gradient. Viral DNA was extracted, amplified using the ϕ29 polymerase and pyrosequenced. The majority of sequences were unknown, whereas among the identifiable viral sequences most were related to tailed bacteriophages, in particular *Prochlorococcus* phages and *Synechococcus* cyanophages. Large nucleocytoplasmic DNA viruses were also detected, in particular Mimivirus-like viruses. Both the presence of tailed bacteriophages and of a giant virus were confirmed by electron microscopy. Moreover, we observed lower diversity for the viral community of El Berbera, a guelta with sustained human activity, compared with the pristine Ilij and Molomhar, and we retrieved sequences related to viruses infecting crop pests, a likely consequence of the agricultural use of the soil. However, the El Berbera viral community shared the common global characteristics of the pristine ponds, such as the prevalence of lytic phages, especially *Myoviridae* infecting photosynthetic cyanobacteria. In contrast, the Hamdoun viral community was characterized by a larger proportion of temperate phages (*Siphoviridae* and phages infecting heterotrophic terrestrial bacteria). We hypothesized that the high proportion of induced lysogens in the Hamdoun viral community reflects stress from the critically low water level of the pond.

Open

The ISME Journal (2013) 7, 359–369

www.nature.com/ismej

ORIGINAL ARTICLE

Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara

Laura Fancello[1], Sébatien Trape[2], Catherine Robert[1], Mickaël Boyer[1,3], Nikolay Popgeorgiev[1], Didier Raoult[1] and Christelle Desnues[1]

[1]*Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE UM63, CNRS 7278, IRD 198, Inserm 1095, Aix-Marseille Université, Faculté de médecine, Marseille, France and* [2]*IRD, UR CoReUs2, promenade Roger-Laroque, Nouméa cedex (New Caledonia), France*

Here, we present the first metagenomic study of viral communities from four perennial ponds (gueltas) located in the central Sahara (Mauritania). Three of the four gueltas (Ilij, Molomhar and Hamdoun) are located at the source of three different wadis belonging to the same hydrologic basin, whereas the fourth (El Berbera) belongs to a different basin. Overall, sequences belonging to tailed bacteriophages were the most abundant in all four metagenomes although electron microscopy and sequencing confirmed the presence of other viral groups, such as large DNA viruses. We observed a decrease in the local viral biodiversity in El Berbera, a guelta with sustained human activities, compared with the pristine Ilij and Molomhar, and sequences related to viruses infecting crop pests were also detected as a probable consequence of the agricultural use of the soil. However, the structure of the El Berbera viral community shared the common global characteristics of the pristine gueltas, that is, it was dominated by *Myoviridae* and, more particularly, by virulent phages infecting photosynthetic cyanobacteria, such as *Prochlorococcus* and *Synechococcus* spp. In contrast, the Hamdoun viral community was characterized by a larger proportion of phages with the potential for a temperate lifestyle and by dominant species related to phages infecting heterotrophic bacteria commonly found in terrestrial environments. We hypothesized that the differences observed in the structural and functional composition of the Hamdoun viral community resulted from the critically low water level experienced by the guelta.
The ISME Journal (2013) **7,** 359–369; doi:10.1038/ismej.2012.101; published online 4 October 2012
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: viral metagenomics; giant virus; Sahara desert; perennial water pond; Mauritania; Adrar plateau

Introduction

Viruses can colonize virtually all ecosystems on Earth and are found wherever cellular life exists (Le Romancer *et al.*, 2007). In the ocean, viruses (the majority of which are bacteriophages) represent the most abundant biological component of the ecosystem and influence horizontal gene transfer, microbial diversity and biogeochemical cycling (Fuhrman, 1999; Suttle, 2005, 2007). Metagenomics (the sequence-based analysis of the collective genomes contained in an environmental sample) was first applied to environmental viral communities in marine waters 10 years ago (Breitbart *et al.*, 2002). This study demonstrated that the viral fraction represents a vast reservoir of unexplored biodiversity. Since then, viral diversity has been investigated using metagenomics in a wide range of environments, including marine waters (Angly *et al.*, 2006; Culley *et al.*, 2006), freshwaters (Dinsdale *et al.*, 2008a; Djikeng *et al.*, 2009), hot springs (Schoenfeld *et al.*, 2008), soils (Williamson *et al.*, 2005; Fierer *et al.*, 2007), stromatolites and thrombolites (Desnues *et al.*, 2008), and animal-associated biomes (Breitbart *et al.*, 2003; Zhang *et al.*, 2006; Vega Thurber *et al.*, 2008; Ng *et al.*, 2011).

To date, studies on viral diversity have mainly focused on marine waters from temperate regions, and data exploring the extent of viral diversity and ubiquity in arid regions are largely scarce. Unfavorable conditions found in cold or hot deserts limit the development and the activity of eukaryotic life. In such environments, the study of viral assemblages is of particular interest because microbial communities are mainly regulated by viral lysis (Weinbauer, 2004; Laybourn-Parry, 2009). A recent

Correspondence: C Desnues, Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE CNRS-IRD UMR 7278, Aix-Marseille Université, Faculté de médecine, 27 Bd Jean Moulin, Marseille 13385, France.
E-mail: christelle.desnues@univ-amu.fr
[3]Current address: Danone Research, 92190 Meudon, France
Received 10 February 2012; revised 18 June 2012; accepted 16 July 2012; published online 4 October 2012

study in the cold desert of Antarctica used a 0.45-µm size selective metagenomic analysis to show that the viral community of an ice-covered lake presented an unexpectedly high genetic richness, distinct from that of other aquatic viral metagenomes, and was dominated by small single-stranded DNA viruses infecting eukaryotes in the spring and by large double-stranded DNA (dsDNA) viruses (mostly Phycodnaviruses and Mimiviruses) and dsDNA bacteriophages in summer (Lopez-Bueno *et al.*, 2009).

The Sahara is the largest non-polar desert on Earth. In the early Holocene period, the Sahara experienced humid episodes (Kuper and Kröpelin, 2006) that sustained the development of numerous lakes and wetlands; remnants of these aquatic environments still persist today. The Mauritanian Adrar is one of the mountainous massifs of the central Sahara and contains >20 perennial and semi-perennial freshwater bodies. Among them, rocky pools (gueltas) are found in the higher reaches of gorge-like watercourses and are alimented by subterranean seep and rainfall. Some of these gueltas are sites of permanent human settlements and are used for agricultural and animal-farming purposes.

Previous electron microscopy and pulse-field gel electrophoresis studies from bacteriophage particles induced from Namib and Sahara sands have revealed the presence of different morphotypes with genome sizes varying from 45 to 350 kb, suggesting the existence of large viral particles (Prigent *et al.*, 2005; Prestel *et al.*, 2008). The objective of this study was to fill the gaps in our knowledge of the ubiquity and diversity of viruses in hot desert environments. Using metagenomic approaches, we investigated the composition, taxonomy and functional diversity of the viral communities from four gueltas located in the Mauritanian Sahara.

Materials and methods

Geographic location of the sampling sites

During the dry season of 2009 (June), water samples were collected from four different gueltas: Ilij (20°38′046 N, 13°08′490 W), Molomhar (20°35′229 N, 13°08′794 W), Hamdoun (20°19′380 N, 13°08′550 W) and El Berbera (19°59′181 N, 12°49′3744 W) located in the Adrar plateau of Mauritania (Figure 1). Ilij, Molomhar and Hamdoun belong to the same hydrographic network (the Seguellîl wadi basin), whereas El Berbera belongs to a different network (the Timinit wadi basin). The local human population occasionally and permanently occupies the Hamdoun and El Berbera gueltas, respectively. Hamdoun is located close to the main Atar-Nouakchott road and may serve as a temporary open well, whereas El Berbera hosts a permanent human settlement and is a site of intensive date fruit production. During the driest periods of the year, the residual water volume of the Hamdoum guelta can drop to $<2\,m^3$, whereas the volumes of the other gueltas remain between 200 and $500\,m^3$.

Sampling procedure, virus purification, transmission electron microscopy, nucleic acid extraction and sequencing

During a 2-day mission, one liter of water was collected from each guelta and filtered through a 0.45-µm pore filter. Virus-size particles contained in the filtrate were precipitated on site using PEG (10%)

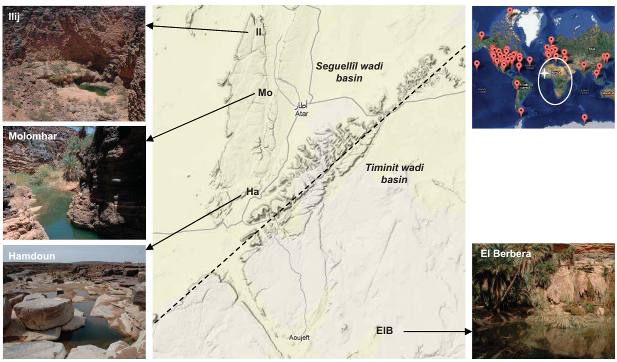


Figure 1 Geographic localization and pictures of the samp ling sites. Central: A Google Earth map of the Adrar region showing mountains (gray) and sandstones (yellow). Il: Ilij guelta, Mo: Molomhar guelta, Ha: Hamdoun guelta, ElB: El Berbera guelta. The dashed line on the map indicates the separation between the two hydrologic systems: the Seguellîl wadi basin and the Timinit wadi basin. Images of the sampling sites are provided on the left and on the right of the map. Upper right: Localization of the 205 environmental metagenomic projects recorded in the Genome On Line Database (GOLD) as of 2012-01-10. Africa is outlined by a circle, and the Adrar plateau of Mauritania is indicated by a star.

and NaCl (1 M final) in bottles maintained at 4 °C in a cool-box. Samples (kept at 4 °C) were brought to the laboratory immediately (within the 48 h) for further processing. Precipitated viral particles were purified using CsCl density gradient ultracentrifugation and DNAse treated as previously described (Thurber *et al.*, 2009). Purified viral particles were stained with 3.5% uranyl acetate and lead citrate and then examined by transmission electron microscopy (TEM) (Philips Morgagni 268D, FEI Co., Eindhoven, The Netherlands). Nucleic acids were extracted using the formamide procedure (Thurber *et al.*, 2009) and amplified using the illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Freiburg, Germany). Because phi29 DNA polymerase has been shown to preferentially amplify circular DNA and genomes from single-stranded DNA viruses (Kim *et al.*, 2008; Kim and Bae, 2011), duplicate reactions were performed to minimize this bias, as previously suggested (Thurber *et al.*, 2009). Amplification products were then pooled, ethanol purified and pyrosequenced on a Roche Applied Sciences (454 Life Sciences, Basel, Switzerland) GS20 platform. Metagenomes are freely accessible on the MG-RAST annotation server with the following accession numbers: El Berbera 2 (4446033.3), Molomhar Guelta (4445718.3), Ilij Guelta (4445716.3) and Hamdoun Guelta (4445715.3).

Taxonomic and functional annotations
Metagenomes were annotated using MG-RAST version 2 (Meyer *et al.*, 2008) with an *E*-value cutoff of 10^{-5}. The MG-RAST server produces automated taxonomic assignments using Blastx searches against the SEED non-redundant database and other accessory databases (rRNA, chloroplast and mitochondrial databases) and also produces metabolic profiles of metagenomes by Blastx comparisons using the SEED-Subsystem data set. Pairwise comparisons of the metabolic profiles were performed using XIPE-TOTEC (Rodriguez-Brito *et al.*, 2006), a non-parametric pairwise bootstrap statistical test that was specifically developed for metagenomic functional comparisons and is based on median difference analysis. This test locates statistically significant differences and identifies subsystems that are overrepresented in each comparison. The confidence level chosen for the test was 98%.

Sequence analysis
The GC content of the four metagenomes was analyzed using the geecee function of EMBOSS. The average GC fraction was computed for each metagenome as a whole or separately for subsets of bacterial- and viral-annotated reads.

Assembly and phylogeny
The assembly of each metagenome was performed using the Genome Sequencer (GS) *De Novo* Assembler version 2.0.01 (Roche Diagnostics, Meylan, France), an application especially suited to the analysis of GS-FLX data. We chose a minimum overlap length of 20 bp and a minimum overlap identity of 95%. We only kept contigs longer than 300 bp for subsequent analyses because the average read length was 251–258 bp. Open Reading Frames (ORFs) were searched on large contigs (>1500 bp) by Prodigal (Hyatt *et al.*, 2010) and MetaGeneMark (Zhu *et al.*, 2010). Phylogenetic trees were constructed for ORFs with at least 10 homologs, according to a Blastx search against the NCBI non-redundant database (*E*-value $<1e^{-10}$). An ORF and its homologs were aligned using MUSCLE (Edgar, 2004), and the alignment was curated using Gblocks (Castresana, 2000). Phylogenetic trees were constructed using PhyML (Guindon *et al.*, 2010), with 100 bootstrap replicates, and visualized using MEGA v5 software (Tamura *et al.*, 2011). A specific research of phages and prophages has also been performed on assembled contigs by a Blastn search (*E*-value $<1e^{-05}$) against the Aclame database (Leplae *et al.*, 2004).

Mapping
For each of the most abundant organisms found in the MG-RAST analysis, metagenomic reads were mapped (that is, aligned against a reference sequence) on the genome of that organism. Mapping was performed using the GS Reference Mapper version 2.0.01 (Roche).

Population modeling
Information about community structure and diversity was obtained for each metagenome using the following workflow: (i) computation of the community contig spectrum using the free Circonspect software, (ii) evaluation of average genome sizes using GAAS free software (Angly *et al.*, 2009) with an *E*-value cutoff of 10^{-3} and (iii) mathematical modeling of the community structure and diversity by PHACCS (Angly *et al.*, 2005).

Phylogenetic tests
Several phylogenetic tests were performed using the FastUniFrac tool (Hamady *et al.*, 2009) to find statistically significant differences among the four metagenomes. The analyses were performed on the subset of viral sequences of each metagenome. For each metagenome, FastUniFrac uses phylogenetic information to assemble metagenomic sequences into a tree. P-test and UniFrac metric capture significant diversity between the trees associated with the different metagenomes, and they account only for tree topology or for both tree topology and branch length, respectively. The two tests can be used for multiple or pairwise comparisons or to compare one particular tree with all others.

Principal Component Analysis (PCA) and hierarchical clustering were also performed using FastUniFrac. The robustness of the clustering results to the sampling effort and evenness was determined using the Jackknife Environment Clusters analysis option.

Comparative metagenomics of viral communities from freshwater environments
A multiple comparison of the phylogenetic profiles of different natural and non-natural freshwater viral communities was performed using the MG-RAST server (Meyer *et al*., 2008). Ten viral metagenomes were compared; the four analyzed in this work, two from two different temperate freshwater lakes (Roux *et al.,* 2012), one from an Antarctic lake sampled in the spring and summer seasons (Lopez-Bueno *et al.* 2009), and two from an aquaculture system (Dinsdale *et al*., 2008a). The phylogenetic profile was based on the sequence taxonomic assignment according to a Blastx search (*E*-value $<1e^{-05}$) against the NCBI GenBank non-redundant database. Multiple comparisons were performed by PCA on the MG-RAST server using normalized values and a Bray-Curtis distance matrix. *P*-values were computed on the MG-RAST server (Meyer *et al*., 2008).

Results

Taxonomic composition of the viral metagenomes
A total of 82 814 818 bp of sequence was generated from the four samples (Ilij ~17 Mbp, Molomhar ~25 Mbp, Hamdoun ~15 Mbp, El Berbera ~24 Mbp), corresponding to 324 603 sequences with an average length of 250 bp. Annotation of the sequence fragments by MG-RAST using an *E*-value cutoff of $1e^{-05}$ indicated that 70.50–83.21% of these fragments had no significant hits to known sequences stored in the SEED non-redundant database or other accessory databases (Figure 2). According to the MG-RAST annotation, 8.06–34.42% of the known reads were classified as viruses. The majority of viral reads belonged to dsDNA viruses (Table 1) and, among these, >92% matched with Caudovirales. Sequences belonging to the *Myoviridae* were the most abundant in all metagenomes followed by *Podoviridae* and *Siphoviridae*. The presence of tailed phages was confirmed by TEM (Figures 3a and b). Viral morphotypes were usually between 50 and 200 nm in diameter, but some viral particles with diameters <50 nm (Figures 3d–f, arrows) and >200 nm (Figure 3c) were also observed. Among the dsDNA viruses, sequences belonging to eukaryotic viruses were also found (Tables 1 and 2). Four out of the seven families of nucleocytoplasmic large DNA viruses group were represented (Table 1, in bold). Viruses from the *Phycodnaviridae* (infecting algae) and *Mimiviridae* (infecting amoebas and algae) were more abundant in Hamdoun (4.25%) and El Berbera (3.18%). *Poxviridae* sequences were also more frequently found in the El Berbera metagenome. Only a few reads were associated with single-stranded DNA viruses. The majority of these reads were found in the Molomhar metagenome and were related to *Microviridae* (2.92%). To analyze the taxonomic composition more accurately, we used GAAS that normalizes the number of hits for the genome size and then provides a more realistic description of species abundances. According to the GAAS analysis, the most represented viral genotype was that of *Prochlorococcus* phage, found in three out of the four gueltas (Ilij, Molomhar and El Berbera) with relative abundances ranging between 31.54 and 55.24% (Table 2). In contrast, Hamdoun

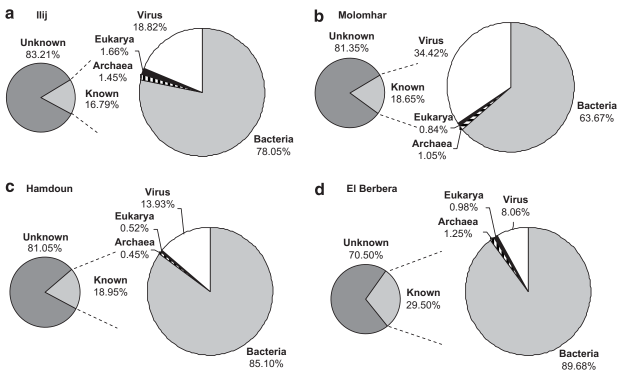


Figure 2 Reads classification according to their Best Blast Hit (*E*-value $<10^{-5}$) in the MG-RAST analysis. (**a**) Ilij guelta, (**b**) Molomhar guelta, (**c**) Hamdoun guelta and (**d**) El Berbera guelta.

Table 1 Classification of reads hitting viral sequences

Group	*Order*	*Family*	*Ilij (%)*	*Molomhar (%)*	*Hamdoun (%)*	*El Berbera (%)*
		Unclassified	5.78	2.86	5.16	5.92
	Caudovirales	*Myoviridae*	63.33	73.04	48.03	56.43
		Podoviridae	11.75	8.45	22.61	18.92
		Siphoviridae	12.50	8.20	16.39	11.93
	Herpesvirales	*Herpesviridae*	0.05	0.02	0.00	0.00
dsDNA	–	*Tectiviridae*	0.05	0.06	0.00	0.00
	–	***Iridoviridae***	0.67	1.11	0.83	0.54
	–	***Phycodnaviridae***	2.08	1.20	4.25	2.25
	–	***Poxviridae***	0.28	0.02	0.38	0.54
	–	***Mimiviridae***	1.90	0.84	0.99	3.18
	–	*Baculoviridae*	0.00	0.00	0.00	0.15
ssDNA	–	*Circoviridae*	0.00	0.14	0.00	0.05
	–	*Microviridae*	0.00	2.92	0.15	0.09
	–	*Geminiviridae*	0.00	0.02	0.08	0.00
	–	*Nanoviridae*	0.05	0.14	0.08	0.00
	–	*Parvoviridae*	0.00	0.00	0.08	0.00
Retroviridae	–		0.00	0.00	0.08	0.00
Unclassified phages/viruses	–		1.56	0.98	0.91	0.00

Abbreviations: dsDNA, double-stranded DNA; ssDNA, single-stranded DNA.
Assignment was made according to the best Blastx hit (E-value $< 10^{-5}$) in the MG-RAST analysis.

was dominated by viruses that infect members of the genus *Microbacterium* (*Microbacterium* phage Min1), which represented >44% of the total viral genotype abundance (Table 2).

Although no bacterial cells could be detected via electron microscopy, 63.67–89.68% of the reads were classified as bacterial in the viral metagenomes (Figure 2; Supplementary Table 1). Using GAAS, all bacteria-annotated reads were dominated by sequences related to the *Acinetobacter* genus (Supplementary Table 2). Between 3 and 20 sequences matching bacterial 16S rRNA genes were also found for each metagenome (Supplementary Table 3).

Functional annotation and metabolic analysis
The metabolic profile of the four metagenomes was explored using MG-RAST, which assigns sequences to metabolic categories based on their Best Blastx Hit against the SEED database (E-value $< 1e^{-05}$). Only 6.22–17.43% of the sequences could be functionally classified in this way. The most represented categories were related to the metabolism of carbohydrates, amino acids, proteins, cofactors, vitamins, DNA, and nucleosides/nucleotides (Figure 4). We compared these data with the metabolic profile derived from the combined analysis of 42 viral metagenomes (subterranean, hypersaline, marine, aquaculture freshwater, coral, microbialites, fish, terrestrial animals and mosquito) described in a previous study (Dinsdale *et al.*, 2008a). The guelta metagenomes were depleted in virulence subsystems compared with the average value found for the other 42 viral metagenomes. Metabolic profile comparisons using XIPE-TOTEC showed that respiration, regulation and cell signaling, and motility and chemotaxis subsystems were overrepresented in the Hamdoun metagenome compared with the other gueltas (P-value < 0.02). Deeper in the respiration subsystem hierarchical levels, the electron donating reaction of the Hamdoun metagenome was dominated by the respiratory dehydrogenase I subsystem, which was mainly represented by the proline dehydrogenase. In contrast, the other three gueltas were dominated by the NAD(P)H dehydrogenase complex, which is classified as a respiratory complex I subsystem. An overrepresentation of RNA metabolism was also evidenced by the XIPE-TOTEC analyses of the El Berbera metagenome, whereas the Molomhar metagenome displayed a statistically significantly higher number of sequences related to photosynthesis and nucleoside/nucleotide metabolism (P-value < 0.02).

Assembly, contig analysis and mapping
Contigs were assembled using the GS *De Novo* Assembler, and only contigs longer than 300 bp were kept (Supplementary Table 4). The average contig length was from 742 to 990 bp, and large contigs were also obtained (for example, 55 kbp in El Berbera and 52 kbp in Hamdoun). Overall, 29.90–37.06% of the contigs had similarities to phage and prophage sequences in the Aclame database (Supplementary Table 5). Viral assemblies were dominated by phage genomes. However, mapping to fully sequenced genomes of the most abundant phages in the metagenomes resulted in low coverages (<5%; Supplementary Figure 1). Low coverage was also found for plasmids; the maximum plasmid coverage was 14.24% for the *Acinetobacter venetianus* pAV2 genome.

We were able to assemble 30.33–68.72% of all reads (Supplementary Table 4). Interestingly, between 34.98 and 93.03% of the 'unknown' reads

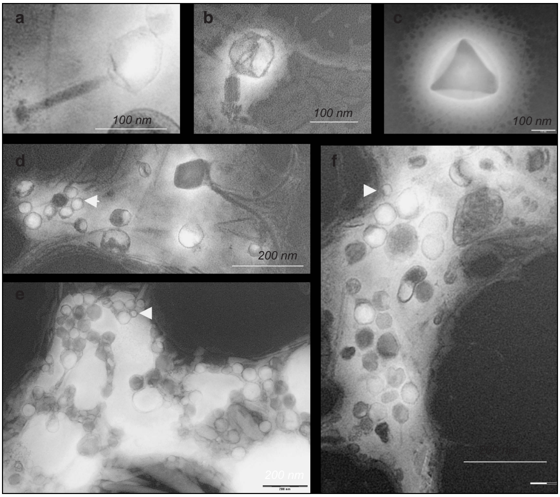


Figure 3 Viral morphotypes observed under TEM in the gueltas of the Adrar plateau. Example of tailed phages belonging to the Myoviridae family (**a**, **b**). Viral morphotypes were usually between 50 and 200 nm in diameter (**d–f**), but some small viral particles with diameters <50 nm (**d–f**, arrows) and large Mimivirus-like particles (**c**) were also observed. Images (**a**) and (**d**) are from Ilij, (**b**) and (**f**) are from Hamdoun, and (**c**) and (**e**) are from Molomhar.

Table 2 Most represented viral genotypes among the viral hits according to GAAS analysis

Metagenome	*Viral species*	*Relative abundance (%)*	*Host*
Ilij	***Prochlorococcus* phages**	**55.2449**	B
	Burkholderia phages	18.9451	B
	***Synechococcus*phages**	**15.9518**	B
	Roseobacter phage SIO1	5.8638	B
	Acanthocystis turfacea Chlorella virus 1	2.4366	E
	Aeromonas phages	1.3597	B
Molomhar	***Prochlorococcus* phages**	**46.0733**	B
	***Synechococcus* phages**	**35.6391**	B
	Mycobacterium phages	10.2428	B
	Bordetella phages	2.0430	B
	Acyrthosiphon pisum secondary endosymbiont phage	1.1061	E
	Acanthocystis turfacea Chlorella virus 1	1.0322	E
Hamdoun	*Microbacterium* phage Min1	44.3371	B
	***Synechococcus* phages**	**29.9623**	B
	***Prochlorococcus* phages**	**11.5155**	B
	Acanthocystis turfacea Chlorella virus 1	10.7050	E
	Acanthamoeba polyphaga mimivirus	3.4801	E
El Berbera	***Prochlorococcus* phages**	**31.5358**	B
	***Synechococcus* phages**	**26.8707**	B
	Mycobacterium phages	14.3964	B
	Lactobacillus phage phiJL-1	9.7423	B
	Burkholderia phage phi644-2	7.3403	B
	Spodoptera litura NPV	2.7741	E
	Musca domestica salivary gland hypertrophy virus	2.6775	E
	Acanthocystis turfacea Chlorella virus 1	2.4808	E
	Ostreococcus virus OsV5	1.7077	E

Abbreviations: B, Bacteria; E, Eukaryotes.
Only viral genotypes with a relative abundance superior to 1% are indicated. Viral species infecting cyanobacteria are shown in bold.

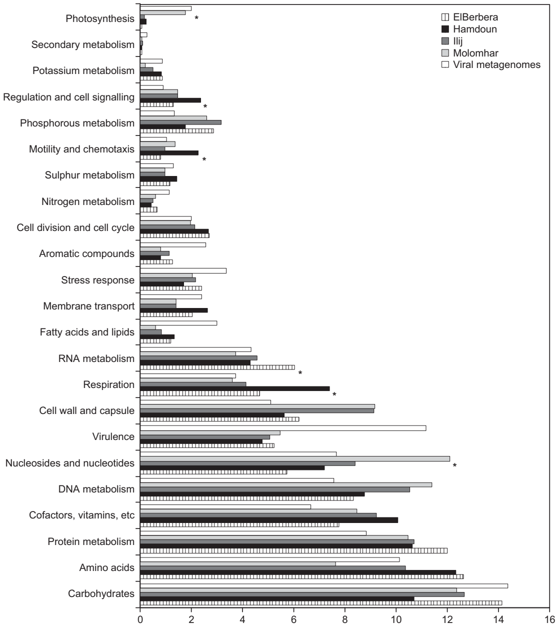


Figure 4 Relative abundances of sequences assigned to each metabolic subsystem by MG-RAST. The metabolic categorization is based on the sequences best Blast hits in the SEED database curated subsystems (E-value $<1e^{-05}$). Asterisks: metabolic subsystems for which pairwise comparisons were performed by XIPE-TOTEC to identify statistically significant differences ($P<0.05$) between the four guelta metagenomes.

identified by the MG-RAST annotation system were assembled into contigs (Supplementary Table 4). The assembly of these unknown reads into contigs could facilitate their taxonomic assignment. Indeed, short sequences are less likely than long sequences to retrieve statistically significant similarities in Blast searches and sequence assembly into longer contigs is helpful to overcome this difficulty (Wommack *et al.*, 2008). Moreover, long contigs can contain unknown reads and reads with far homologies to known sequences, which are suggestive of the putative phylogenetic origin of the whole contig. The largest contigs (>1500 bp) were annotated by ORF prediction and Blast search (Supplementary Table 6). When possible, phylogenetic analysis was performed to confirm the origin of the predicted ORFs (Supplementary Figure 2). A few large contigs contained a relevant proportion of predicted ORFs with similarities to phage sequences and coding for some specific conserved phage proteins, that is, terminases, structural proteins (mainly related to Caudovirales tail structures) and phage DNA polymerases (Supplementary Table 6). It has been previously shown that viral genomes contain more ORFans (that is, ORFs without homologs in the databases) than do bacteria (~30 and ~10% for viral and bacterial genomes, respectively) (Yin and Fischer, 2006; Boyer *et al.*, 2010) and that viral (meta)genomes tend to be more AT rich than those of their hosts (Rocha and Danchin, 2002; Willner *et al.*, 2009). Thus, contigs with >50% ORFans, low GC%, and for which conserved viral protein-encoding genes have been identified, can confidently be considered as of viral origin (for example, contigs ElBerbera_882 or Hamdoun_439 in Supplementary Table 6).

Community phylogenetic structure and diversity across sampling sites

Viral community structure and diversity estimations were performed using the PHACCS analysis system on each metagenome (Supplementary Figure 3). Briefly, contig spectra were generated using the Circonspect tool and the average genome size was estimated by GAAS. These parameters were passed to PHACCS for alpha-diversity analysis. Computed community structures (defined by richness (R), evenness (E) and diversity (H')) are graphically represented as rank-abundance curves in Supplementary Figure 3. Based on the obtained results, the samples with the highest viral diversity index were the pristine Ilij and Molomhar gueltas (H' = 4.83 and H' = 4.33, respectively), followed by El Berbera (H' = 4.19) and Hamdoun (H' = 2.21). The phylogenetic composition of the viral communities of the four metagenomes was then considered, and comparisons were computed using the FastUniFrac tool on the subset of viral annotated metagenomic sequences. Statistically significant differences were measured between two samples (P-value < 0.05) using the 'UniFrac significance analysis' test and further confirmed using the P-test. PCA, which was used to visualize multiple comparisons between samples (Supplementary Figure 4), showed that Hamdoun is isolated from the other samples. To evaluate the robustness of this clustering pattern, we performed a Jackknife environment cluster analysis. The results of this bootstrap procedure confirmed the confidence in the Hamdoun cluster node.

Comparisons with viral communities from other freshwater environments

The phylogenetic profile of the four gueltas was compared with that of other natural or non-natural freshwater environments, two temperate freshwater lakes (Roux *et al.*, 2012), an Antarctic lake sampled in the spring and summer seasons (Lopez-Bueno *et al.*, 2009) and two freshwater samples from a human-controlled aquaculture system (Dinsdale *et al.*, 2008a; Figure 5). As a phylogenetic profile representation, we used the metagenomic sequences classification according to their best Blast hit in a Blastx search against the NCBI GenBank non-redundant database (E-value < 1e^{-05}). Multiple comparisons were performed and visualized by PCA on the MG-RAST server. The results showed a geographic clustering pattern with the Mauritanian gueltas clustering together, separate from the others (and Hamdoun again separated from the other three Mauritanian gueltas) (Figure 5). Similarly, metagenomes from temperate natural and artificial freshwaters group together and are separate from the metagenomes of other environments. The two metagenomes from the Antarctic lake did not group together, which is consistent with the phylogenetic profile differences observed between the spring and summer communities from this lake (Lopez-Bueno *et al.*, 2009). The metagenome-clustering pattern showed statistically significant differences in the viral domain (P-value = 0.006).

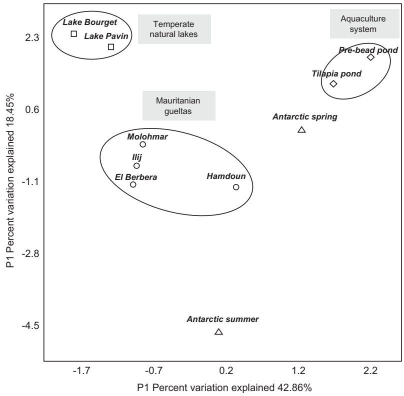


Figure 5 First two principal coordinates from the principal coordinate analysis of the viral communities in freshwater samples from different environments. The PCA was run in MG-RAST to visualize the overall patterns of variation between the samples.

Discussion

With the advent of metagenomics, an increasing number of studies describing viral and bacterial diversity have been conducted. Currently, only a few investigations have focused on viral assemblages in freshwaters, and most of them concern freshwaters from non-natural or polluted ecosystems. For example, viral communities have been described from aquaculture ponds (Dinsdale *et al.*, 2008a; Rodriguez-Brito *et al.*, 2010), a cattle farm pond (Rooks *et al.*), reclaimed and potable waters (Rosario *et al.*, 2009), hydrocarbon-polluted groundwater (Abbai *et al.*, 2012), and a man-made recreational lake in MD, USA (Bench *et al.*, 2007). Viral communities in natural freshwater systems have only been described in an ice-covered lake in Antarctica (Lopez-Bueno *et al.*, 2009) and, more recently from two temperate freshwater lakes in France (Roux *et al.*, 2012). By combining electron microscopy and metagenomics, we provide here the first comprehensive analysis of viral communities from freshwater ponds in the Sahara desert of Mauritania.

Most studies focusing on viral diversity in the environment use a 0.2-μm filtration step to separate viruses and bacteria on size criteria. One drawback of this method is that it may fail to recover large viral particles, which are supposed to be common in aquatic ecosystems (Claverie, 2005). However, using a 0.45-micron pore size filter, Lopez-Bueno *et al.* (2009) were able to identify sequences associated with large dsDNA viruses (mainly from the

Phycodnaviridae and *Mimiviridae* families) from an Antarctic freshwater lake in summer. In this study, up to 6.51% of the sequences matched large dsDNA viruses, and the presence of large viral particles (>200 nm) with Mimivirus-like morphologies was confirmed by electron microscopy (Figure 3c). These results further support that large DNA viruses are common in the environment (Ghedin and Claverie, 2005; Monier *et al.*, 2008a, b) and that the 0.2-μm filtration step currently used to prepare environmental viral metagenomes most likely leads to an underestimation of their genetic diversity.

No bacterial cells were observed under electron microscopy, and the number of reads annotated as bacteria (Figure 2) was similar to those reported for other environmental viral metagenomes (Edwards and Rohwer, 2005), indicating a low bacterial contamination of the metagenomes. In addition, the high proportion of unassigned sequences and relatively low number of 16S rRNA matching sequences (Supplementary Table 3) supported a viral origin for the bacterially annotated reads. Because bacterial genes can be packaged into generalized transducing phage particles (Beumer and Robinson, 2005; Ghosh *et al.*, 2008; Del Casale *et al.*, 2011), the bacterial-like sequences in the guelta metagenomes might come from excised prophages mistakenly annotated as bacterial and/or from genes of bacterial origins that were transferred to their phages.

Blast searches performed on the viral metagenomes showed that >70% of the sequences before assembly did not have homologs in current sequence databases (Figure 2). This result is consistent with results of previously published viral metagenomic projects (Breitbart *et al.*, 2002; Desnues *et al.*, 2008; Lopez-Bueno *et al.*, 2009) and again emphasizes that most of the biological diversity in the viral world is still unknown. In this case, sequence assembly using low stringency parameters (20 bp coverage and 95% identity) was of particular interest in classifying the unknown sequences. For example, the Hamdoun metagenome contained >90% unknown reads that could be assembled into contigs (Supplementary Table 4). The downstream identification of structurally conserved viral genes in these contigs (Supplementary Table 6) has provided information about the putative viral origin of these ORFans.

The guelta viral metagenomes were largely dominated by Caudovirales reads, and TEM confirmed the presence of tailed phages (Figure 3) along with other viral morphotypes. Caudovirales are common in the environment and are the dominant viral type recovered from metagenomic analyses in marine environments (Breitbart *et al.*, 2002; Suttle, 2005). Myoviruses, Siphoviruses and Podoviruses were also the most frequently observed viral particles in samples of Namib and Sahara desert sands after the mitomycin C induction of prophages and sonication to release pseudo-lysogens (Prigent *et al.*, 2005; Prestel *et al.*, 2008). The Molomhar metagenome presented the largest number of reads (73% of all reads) that were related to Myoviruses and only 8.2% of reads represented Siphoviruses. At a deeper taxonomic level, viruses infecting photosynthetic bacteria (for example, *Prochlorococcus* and *Synechococcus* phages; in bold, Table 2) were the most abundant in both absolute percentage and rank in the El Berbera, Ilij and Molomhar metagenomes.

Despite being a site for a permanent human settlement, the El Berbera viral community presented a viral community structure similar to those of the Ilij and Molomhar pristine gueltas. It has been previously shown that human activities can affect the diversity and the composition of microbial communities and, thus, their viral predators (reviewed in Horner-Devine *et al.*, 2004). For instance, microbial and viral communities from four coral atolls in the Pacific Ocean dramatically changed along a gradient of human disturbance; the most human-impacted atoll was dominated by heterotrophic microbes, including a large percentage of potential human pathogens (Dinsdale *et al.*, 2008b). In this study, the index of viral biodiversity was inversely correlated with human presence (Supplementary Figure 3), with El Berbera displaying a lower diversity index than the pristine Ilij and Molomhar gueltas. In addition, reads matching *Spodoptera litura* NPV, a baculovirus infecting *S. litura* (Lepidoptera: *Noctuidae*) a crop pest in tropical regions (Rao *et al.*, 1993), were found in the El Berbera metagenome; its presence is most likely linked to the agricultural activity (date production) that developed around the guelta. However, no significant change in the taxonomic structure of the viral communities was observed between the El Berbera human-populated and the Ilij and Molomhar non-populated gueltas. This stability reflects either that the magnitude of human disturbance is weak enough to be absorbed by the system or that the sequencing depth is not sufficient to statistically support finer differences in the phylogenetic profiles between the El Berbera, Ilij and Molomhar metagenomes.

In comparison with the Ilij, Molomhar and El Berbera metagenomes, the Hamdoun metagenome contains dramatically fewer reads matching Myoviruses but more reads related to Siphoviruses (Table 1). This result shows that as it is the case for terrestrial and sediment phage communities (Breitbart *et al.*, 2004; Williamson *et al.*, 2007), viruses with the potential for temperate lifestyles were common in this environment. This is also confirmed by the viral community taxonomic profile, which is dominated by the *Microbacterium* phage Min1 (Akimkina *et al.*, 2007) a potentially temperate phage that infects *Microbacterium* sp., a versatile heterotrophic bacteria that is frequently isolated from the rhizosphere and soils (Takeuchi and Hatano, 1998). The presence of viruses with the ability for being temperate in the free-viral fraction may be related to high nutrient availability, high bacterial density or to environmental stress that leads to virus induction (McDaniel *et al*., 2002; Breitbart *et al*., 2004). For decades, the Sahara has experienced a dramatic rainfall deficit, and a recent study has stressed that, with only 1.7 m^3 of remaining water in July 2007, Hamdoun is one of the most endangered gueltas of the Adrar plateau (Trape, 2009). We hypothesize that the high proportion of induced lysogens in the Hamdoun viral community compared with the other gueltas reflects stress associated with the unprecedentedly low water content of the pond. Further studies will be required (for example, after the rainy season) to confirm this hypothesis and to determine whether the current structure of the viral community is maintained over time.

Acknowledgements

We thank Florent Angly for his help with GAAS and Jean-François Trape for the valuable assistance during the field surveys. This work was funded by the Centre National de la Recherche Scientifique (crédits récurrents).

References

Abbai N, Govender A, Shaik R, Pillay B. (2012). Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater. *Mol Biotechnol* **50**: 39–48.

Akimkina T, Venien-Bryan C, Hodgkin J. (2007). Isolation, characterization and complete nucleotide sequence of a novel temperate bacteriophage Min1, isolated from the nematode pathogen Microbacterium nematophilum. *Res Microbiol* **158**: 582–590.

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P *et al.* (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**: 41.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R *et al.* (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5**: e1000593.

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K *et al.* (2007). Metagenomic Characterization of Chesapeake Bay Virioplankton. *Appl Environ Microbiol* **73**: 7629–7641.

Beumer A, Robinson JB. (2005). A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. *Appl Environ Microbiol* **71**: 8301–8304.

Boyer M, Gimenez G, Suzan-Monti M, Raoult D. (2010). Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. *Intervirology* **53**: 310–320.

Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P *et al.* (2004). Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**: 565–574.

Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P *et al.* (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220–6223.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D *et al.* (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99**: 14250–14255.

Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Claverie J-M. (2005). Giant viruses in the oceans: the 4th Algal Virus Workshop. *Virol J* **2**: 52.

Culley AI, Lang AS, Suttle CA. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**: 1795–1798.

Del Casale A, Flanagan PV, Larkin MJ, Allen CCR, Kulakov LA. (2011). Extent and variation of phage-borne bacterial 16S rRNA gene sequences in wastewater environments. *Appl Environ Microbiol* **77**: 5529–5532.

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M *et al.* (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**: 340–343.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al.* (2008a). Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L *et al.* (2008b). Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* **3**: e1584.

Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ. (2009). Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**: e7264.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Edwards RA, Rohwer F. (2005). Viral metagenomics. *Nat Rev Microbiol* **3**: 504–510.

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R *et al.* (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol* **73**: 7059–7066.

Fuhrman JA. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.

Ghedin E, Claverie JM. (2005). Mimivirus relatives in the Sargasso sea. *Virol J* **2**: 62.

Ghosh D, Roy K, Williamson KE, White DC, Wommack KE, Sublette KL *et al.* (2008). Prevalence of lysogeny among soil bacteria and presence of 16S rRNA and trzN genes in viral-community DNA. *Appl Environ Microbiol* **74**: 495–502.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Hamady M, Lozupone C, Knight R. (2009). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* **4**: 17–27.

Horner-Devine MC, Carney KM, Bohannan BJ. (2004). An ecological perspective on bacterial biodiversity. *Proc Biol Sci* **271**: 113–122.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

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.

Kim KH, Chang HW, Nam YD, Roh SW, Kim MS, Sung Y *et al.* (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol* **74**: 5975–5985.

Kuper R, Kröpelin S. (2006). Climate-Controlled Holocene Occupation in the Sahara: Motor of Africa's Evolution. *Science* **313**: 803–807.

Laybourn-Parry J. (2009). Microbiology. No place too cold. *Science* **324**: 1521–1522.

Le Romancer M, Gaillard M, Geslin C, Prieur D. (2007). Viruses in extreme environments. *Rev Environ Sci Biotechnol* **6**: 17–31.

Leplae R, Hebrant A, Wodak SJ, Toussaint A. (2004). ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res* **32**: D45–D49.

Lopez-Bueno A, Tamames J, Velazquez D, Moya A, Quesada A, Alcami A. (2009). High diversity of the viral community from an Antarctic lake. *Science* **326**: 858–861.

McDaniel L, Houchin LA, Williamson SJ, Paul JH. (2002). Lysogeny in marine Synechococcus. *Nature* **415**: 496.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Monier A, Claverie JM, Ogata H. (2008a). Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* **9**: R106.

Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H. (2008b). Marine mimivirus relatives are probably large algal viruses. *Virol J* **5**: 12.

Ng TFF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C *et al.* (2011). Broad surveys of DNA Viral diversity obtained through viral metagenomics of mosquitoes. *PLoS One* **6**: e20579.

Prestel E, Salamitou S, DuBow MS. (2008). An examination of the bacteriophages and bacteria of the Namib desert. *J Microbiol* **46**: 364–372.

Prigent M, Leroy M, Confalonieri F, Dutertre M, DuBow MS. (2005). A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles* **9**: 289–296.

Rao GVR, Wightman JA, Rao DVR. (1993). World review of the natural enemies and diseases of Spodoptera litura (F.) (Lepidoptera: Noctuiidae). *Insect Sci Appl* **14**: 273–284.

Rocha EP, Danchin A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet* **18**: 291–294.

Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M *et al.* (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J* **4**: 739–751.

Rodriguez-Brito B, Rohwer F, Edwards RA. (2006). An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**: 162.

Rooks DJ, Smith DL, McDonald JE, Woodward MJ, McCarthy AJ, Allison HE454-Pyrosequencing: a molecular Battiscope for freshwater viral ecology. *Genes* **1**: 210–226.

Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. (2009). Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol* **11**: 2806–2820.

Roux S, Enault F, Robin As, Ravet V, Personnic Sb, Theil Sb *et al.* (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**: e33641.

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D. (2008). Assembly of Viral Metagenomes from Yellowstone Hot Springs. *Appl Environ Microbiol* **74**: 4164–4174.

Suttle CA. (2005). Viruses in the sea. *Nature* **437**: 356–361.

Suttle CA. (2007). Marine viruses—major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.

Takeuchi M, Hatano K. (1998). Union of the genera Microbacterium Orla-Jensen and Aureobacterium Collins *et al.* in a redefined genus Microbacterium. *Int J Syst Bacteriol* **48**: 739–747.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731–2739.

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. (2009). Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4**: 470–483.

Trape S. (2009). Impact of climate change on the relict tropical fish fauna of central Sahara: threat for the survival of Adrar mountains fishes, Mauritania. *PLoS One* **4**: e4400.

Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C *et al.* (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. *Proc Natl Acad Sci USA* **105**: 18413–18418.

Weinbauer MG. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**: 127–181.

Williamson KE, Radosevich M, Smith DW, Wommack KE. (2007). Incidence of lysogeny within temperate and extreme soil environments. *Environ Microbiol* **9**: 2563–2574.

Williamson KE, Radosevich M, Wommack KE. (2005). Abundance and diversity of viruses in six Delaware soils. *Appl Environ Microbiol* **71**: 3119–3125.

Willner D, Thurber RV, Rohwer F. (2009). Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol* **11**: 1752–1766.

Wommack KE, Bhavsar J, Ravel J. (2008). Metagenomics: read length matters. *Appl Environ Microbiol* **74**: 1453–1463.

Yin Y, Fischer D. (2006). On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. *BMC Evol Biol* **6**: 63.

Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW *et al.* (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**: e3.

Zhu W, Lomsadze A, Borodovsky M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38**: e132.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

Chapter 4

Human viral metagenomics

4.1 Article 2. Viruses in a 14th century coprolite

Viruses in a 14th century coprolite

Appelt Sandra[1,*], Fancello Laura[1,*], Le Bailly Matthieu[2], Raoult Didier[1], Drancourt Michel[1], Desnues Christelle[†,1]

Submitted to Applied and Environmental Microbiology.

[1] Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE CNRS-IRD UMR 7278, Aix-Marseille Université, Faculté de médecine, 27, Bd Jean Moulin, 13385 MARSEILLE France.
[2] Franche-Comté University, CNRS UMR 6249 Chrono-Environment, 25 030 Besançon, France.

[*] These authors have contributed equally to this work.
[†] Corresponding author. Email: christelle.desnues@univ-amu.fr

Preamble to article 2

Coprolites are fossilized fecal material that can reveal information about ancient intestinal and environmental microbiota. A few paleomicrobiological studies have already been successfully performed on coprolites. However, today most studies performed on coprolites and, in general, ancient specimens target human and bacterial ancient DNA [88, 89, 90]. Little is known, instead, about viral particles persistence and viral DNA detectability in ancient specimens. One study indicated that viral particles in a 400-year old specimen could be observed but their viability was lost [91]. Moreover, according to PCR amplifications on ancient specimens, such as mummified soft tissues, bones and teeth, viral DNA can be detected for at least 1,500 years [92, 93, 94, 95]. All these studies were limited to the detection or observation of a few targeted viruses: no large-spectrum studies to characterize viral diversity of an ancient sample exist.

In this work, we analyzed the viral community of a 14^{th} century coprolite collected from a closed barrel in a Middle Age site in Namur (Belgium) using electron microscopy and, for the first time, viral metagenomics. We believe that this approach is particularly suitable to paleomicrobiological studies, as little is known about viruses in ancient specimens and viral metagenomics does not require previous knowledge or assumptions about which viruses are present in a sample. Viral particles were isolated from the coprolite sample by filtration and cesium chloride purification and a DNA viral metagenome was generated by pyrosequencing. The majority of generated sequences were of unknown origin and the known fraction of the DNA viral community was dominated by bacteriophages. Most bacteriophages were tailed bacteriophages, which were also observed and morphologically identified by transmission electron microscopy. *In silico* analyses revealed the presence of bacteriophages commonly observed in modern stool samples and soil. Some of these bacteriophages also infect bacteria that belong to genera including mammalian pathogens. Eukaryotic and archaeal viral sequences were only detected in low abundances and their presence was supported by contig recovery or confirmed by a suicide *ad hoc* PCR amplification on giant viruses. Taxonomic and functional comparisons of the coprolite viral community with modern stool viral communities revealed differences at the taxonomic level, even when functional profiles were similar. This finding is consistent with those obtained in a recent study, which demonstrated that, despite taxonomic inter-individual variability, the functional profile was significantly conserved within viromes from the same ecological niche [60]. Moreover, we detected and confidently annotated a sequence coding for the chloramphenicol O-acetyltransferase, which confers chloramphenicol resistance. The presence of antibiotic resistance genes has already been reported in viral metagenomes associated with modern human stool samples [55]. In addition, it has already been shown that the evolution and dissemination of resistance genes started well before the recent use of antibiotics [96]. In particular, direct evidence for the presence of antibiotic resistance genes in pre-antibiotic specimens was provided by PCR amplifications using DNA extracted from 30,000-year-old permafrost sediments in Canada [97].

Overall, this study furthers our understanding of past viral diversity and promotes the exploration of ancient viral communities using metagenomics.

Title: Viruses in a 14th century coprolite

Running title: Viruses in a 14th century coprolite

Sandra Appelt[1,*], Laura Fancello[1,*], Matthieu Le Bailly[2], Didier Raoult[1], Michel Drancourt[1], Christelle Desnues[†,1]

[1] Aix Marseille Université, URMITE, UM63, CNRS 7278, IRD 198, Inserm 1095, 13385 Marseille, France.

[2] Franche-Comté University, CNRS UMR 6249 Chrono-Environment, 25 030 Besançon, France.

[*] These authors have contributed equally to this work

† Corresponding author:

Christelle Desnues, Unité de recherche sur les maladies infectieuses et tropicales émergentes (URMITE), UM63, CNRS 7278, IRD 198, Inserm 1095, Faculté de médecine, Aix Marseille Université, 27 Bd Jean Moulin, 13385 Marseille, France. Tel: (+33) 4 91 38 46 30, Fax: (+33) 4 91 38 77 72.

Email: christelle.desnues@univ-amu.fr

Number of words in Abstract: 133 words

Number of words in Main Text: 1,765 words

Number of words in Methods: 773 words

Figures: 4, Supplementary Figures: 3

Tables: 0, Supplementary Tables: 4

Keywords: coprolite, paleomicrobiology, metagenomics, bacteriophages, viruses, ancient DNA

Abstract

Coprolites are fossilized fecal material that can reveal information about ancient intestinal and environmental microbiota. Viral metagenomics has allowed systematic characterization of viral diversity in environmental and human-associated specimens, but little is known about the viral diversity in fossil remains. Here, we analyzed the viral community of a 14^{th} century coprolite from a closed barrel in a Middle Age site in Belgium using electron microscopy and metagenomics. Viruses that infect eukaryotes, bacteria and archaea were detected, and we confirmed the presence of some of them by *ad hoc* suicide PCR. The coprolite DNA viral community was dominated by bacteriophages commonly found in modern stools and soil. Although their phylogenetic compositions differed, the metabolic functions of the viral communities have remained conserved across centuries. Antibiotic resistance was one of the reconstructed metabolic functions detected.

Introduction

Viral metagenomics is a sequencing-based analysis of all of viral genomes isolated from a sample. It has promoted the characterization of viral community diversity. Viral metagenomics has already been successfully applied to the exploration of modern environmental specimens sampled from marine water, freshwater, stromatolites and thrombolites and soil (1-4) and to modern human-associated specimens collected from the liver, blood, nasopharyngeal aspirates and stool (5-9). The DNA viromes generated from modern stools have been demonstrated to be dominated by bacteriophages (10, 11) and to be less diverse than environmental samples (8, 12).

Viral metagenomics does not require culturing viruses or *a priori* knowledge of the sequences that will be targeted, which allows for the identification of new, unknown or unexpected viruses and for the global assessment of the virome. Viral metagenomics is thus particularly suitable for paleomicrobiological studies, as little is known about which viruses are characteristic of ancient specimens. Indeed, the majority of ancient DNA (aDNA) studies are based on the analysis of human and bacterial aDNA (13-15), and viral persistence and its detectability in ancient specimens remains unclear. Electron microscopy has previously revealed that viral particles can persist for over 400 years, but their viability was lost (16). Moreover, PCR amplifications yielded positive results for viral aDNA in ancient specimens such as mummified soft tissues, bones and teeth. The amplification products varied between 100 and 570 bp in size, which indicated that viral aDNA can be detected for at least 1,500 years (17-20).

Here, we used electron microscopy and, for the first time, viral metagenomics to characterize the viral community of an ancient stool specimen. A viral DNA metagenome was generated from a 14th century coprolite sample that was recovered from a Middle Age site in Namur (Belgium).

Material and Methods

Virus-like particle isolation, Transmission Electron Microscopy (TEM) and DNA extraction. Four grams of the interior region of the coprolite were aseptically removed and solubilized over night at 4°C under continuous rotation in 50 mL of phosphate saline buffer (PBS), pH 7.4 (bioMérieux, Marcy-l'Etoile, France), which was previously filtered at 0.02 µm. The coprolite solution was centrifuged for 10 min at 550 g, the upper layer was removed and filtered in stages using sterile Whatman filters (pore sizes: 0.8 µm, 0.45 µm, and 0.2 µm, (Whatman Part of GE Healthcare, Dassel, Germany)). Twenty-five milliliters of the coprolite filtrate were used to precipitate and purify viral particles onto a cesium chloride density gradient using ultracentrifugation and DNAse treatment (Rhower et al. 2009). A 40-µl aliquot of the purified viral particles was stained by 1.5% ammonium molybdate (Euromedex) and observed by transmission electron microscopy using a Philips Morgagni 268D electron microscope (FEI Co., Eindhoven, Netherlands). To isolate the nucleic acids from the purified viral particles, the formamide procedure previously described by (21) was used.

Viral metagenomic library preparation and sequencing. Nucleic acids were amplified using the illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Freiburg, Germany). The phi29 DNA polymerase preferentially amplifies circular DNA, such as genomes from single-stranded viruses (4). To minimize this tendency, duplicate reactions were performed (21). Amplification products were pooled and ethanol purified.

A shotgun strategy was chosen for the high-throughput pyrosequencing on a 454 Life Sciences Genome FLX sequencer using titanium chemistry (Genome Sequencer RLX, Roche). Sequencing was performed using 1/16 of the picotiter plate.

Preprocessing of sequencing data. The reads were screened for quality using mother (22). Only reads longer than 50 bp and with an average quality score of greater than 21 were kept. Reads with more than two ambiguous base calls and/or reads with homopolymers longer than ten bases were eliminated. Identical sequences artificially generated by the pyrosequencing technology were also excluded using the "unique.seqs" mothur command.

Annotation of Reads. A BLASTN search against the non-redundant NCBI database (E-value<$1e^{-05}$) was performed. Reads with no significant similarity to sequences stored in the NCBI database were classified as "unknown reads". The relative abundances of the viral genotypes were estimated using GAAS (23), which is based on a BLASTX search against the RefSeq Viral Genomes database (E-value<$1e^{-05}$) and normalizes the number of reads matching each viral genotype by the genome length of that viral genotype.

Functional annotation was performed on the MG-RAST server (24) using the non-redundant SEED database (E-value<$1e^{-05}$).

Assembly and contig annotation. The reads were assembled into contigs using the Newbler de novo assembler (Roche) with at least a 98% identity and 35 bp of overlap. Only contigs longer than 400 bp were used in subsequent analyses.

Known and unknown contigs were identified on the basis of the BLASTN search against the non-redundant NCBI database (E-value<$1e^{-05}$). The taxonomical and functional contig classification was based on a BLASTX search against the non-redundant NCBI database (E-value<1^{-05}). A specific search for contigs encoding antibiotic resistance genes was also performed using BLAST on the ARDB (Antibiotic Resistance genes Database) with an E-value of <$1e^{-05}$ (25). Significant hits were manually verified.

Phylogenetic trees. When possible, phylogenetic trees of the contigs encoding antibiotic resistance genes were built. Prodigal was used to search for open reading frames (ORFs) in these contigs (26). Regions homologous to the translated ORFs were searched using BLASTP against the non-redundant NCBI database. A multiple alignment was constructed using MUSCLE (27) and curated by Gblocks (28). The phylogenetic tree was then built using the PhyML algorithm (29) with a bootstrap of 100. These tasks were all performed using the pipeline freely available on www.phylogenie.fr (30). The trees were visualized using MEGA v.4 (31).

Comparative metagenomics. The coprolite-associated virome was taxonomically and functionally compared to 21 published viromes of modern stools from healthy adult humans (12, 32). All viromes were taxonomically annotated (Genbank database, E-value<$1e^{-05}$) and functionally annotated (SEED database, E-value<1e-05) using the same method. The taxonomic and functional virome profiles were compared using principal component analysis on the MG-RAST server (normalized data, Bray-curtis measure of distance). Species richness estimations were obtained from the MG-Rast server. Functional diversity (measured by the Shannon-Wiener index) was calculated using the "estimateDiversity" function of the ShotGunFunctionalizeR package on the SEED-based functional metagenome annotations (E<$1e^{-05}$) (33).

Specific PCR amplifications and sequencing. Suicide PCR amplification (34) was performed to validate the high-throughput pyrosequencing results and perform target-orientated virus searches. The primer pair F8F (5'-TGGCCTGCGATCTATGTTCT-3') and F8R (5'-AGTCGTAACAAGGTAGCCGT-3') was used to amplify 133 bp of the genomic region of a *beta-Herpesviridae*. For *Giant virus* detection, we amplified the beta-unit of the polymerase (A10F01: 5'-AAGGGGACAAGGAGTTAAAATAT-3'; A10R01: 5'-TAGATATACGTTTGGTTTTGGAGTGA-3').

Results

The specimen was excavated in 1996 and collected from the interior of a closed barrel, which was commonly used during this period as a pit or latrine (35). The barrel was buried at a depth of 3.80 m. The 121.4-g coprolite specimen was dark brown and well preserved under anaerobic taphonomic conditions. Extensive precautions were undertaken to avoid contaminating the coprolite specimen in our laboratory environment, no positive control was used (15), and suicide PCR protocols were applied (34). All negative controls, used in a 1:4 control: specimen ratio, were consistent with current recommendations for paleomicrobiological and paleoparasitological studies (13, 15, 36, 37) and remained negative. Virus-like particles (VLPs) purified from the internal region of the coprolite, after the external layer was removed, were morphologically diverse and varied in size and shape. Oval particles of different lengths (up to 200 nm) and diameters (up to 100 nm), as well as rod-shaped structures (up to 250 nm in length), were observed (Fig. 1A). We identified a VLP with a dense core with a diameter of approximately 150 nm, which was surrounded by an envelope (Fig. 1B). Moreover, viral particles exhibiting typical characteristics (icosahedral head, long tail) of the *Siphoviridae* bacteriophage family were observed (Figs. 1C-1E).

High-throughput sequencing generated 30,654 reads corresponding to approximately 10.8 million bp. After quality trimming and duplicate removal, 29,811 reads remained (Supplementary Tables S1). The preprocessed read lengths ranged between 77 bp and 574 bp and displayed an average GC content of 47% (Supplementary Fig. S1). Finally, 41.93% of the reads were assembled into 1,464 contigs that ranged from 421 to 12,500 bp (Supplementary Tables S1). In total, 22.15% of all reads and 17.28% of all contigs were significantly similar to known sequences from public databases (Fig. 2A and Supplementary Fig. S2). The viral DNA community was dominated by double-stranded DNA viruses (85.21%) and single-stranded DNA viruses represented 0.81% of the community (Fig. 2B). The most abundant double-stranded DNA viral families were *Siphoviridae* (58.89%), *Myoviridae* (8.79%) and *Podoviridae* (5.95%). We identified viral families that can infect bacteria (*Siphoviridae*, *Myoviridae* and *Podoviridae*), eukaryotes (*Ascoviridae*, *Poxviridae*, *Iridoviridae*, *Adenoviridae*, *Mimiviridae*, *Herpesviridae*, *Baculoviridae*, *Polydnaviridae* and *Phycodnaviridae*) and archaea (*Lipothrixiviridae*, *Tectiviridae* and *Bicaudaviridae*) (Fig. 2B).

Eukaryotic viruses were present at a low abundance, wherein *Phycodnaviridae* was the most abundant family (0.81%) (Fig. 2B). Contig reconstruction and annotation enabled us to identify one contig encoding a tail fiber protein of the *Emiliania huxleyi* virus PS401. This virus is an unclassified dsDNA virus that infects photosynthetic plankton. We identified another contig that encoded a hypothetical protein of the Invertebrate Iridescent virus 3 (IIV-3). IIV-3 is a member of the *Iridoviridae* family, genus *Chloriridovirus*, with a large particle size (180 nm) and infects mosquitoes (Supplementary Table S2). Metagenomic results were confirmed by *ad hoc* suicide PCRs(34). In the presence of negative controls, a 167 bp fragment of a Mimiviridae-like polymerase beta subunit was amplified and sequenced, revealing 84% identity to the *Moumouvirus* of the *Mimiviridae* family (GenBank Accession No. GU265560.1) (Supplementary Table S3). The presence of *Herpesviridae* was also confirmed by PCR amplification. A 133 bp amplicon was 98% identical to the *Stealth Virus 1 clone 3B43* (GenBank Accession No. AF191073.1), a *Beta-Herpesviridae* derived from African green monkeys (38), which we have never studied in our laboratory (Supplementary Table S3 Viral families infecting Archeae were also identified at a low abundances. These corresponded to *Lipothrixiviridae* (0.04%), *Tectiviridae* (0.11%) and *Bicaudaviridae* (0.02%) (Fig. 2B). One contig identified an *Environmental Halophage eHP-6*, an unclassified bacteriophage that infects *Haloarchaea* (Supplementary Table S2).

In contrast, the majority of the sequences were related to bacteriophages, and the most abundant matching bacteriophages that can infect bacteria of the genus *Bacillus* (14.08%). The identified bacteriophages could infect as many as 37 different bacterial genera, including bacterial genera commonly associated with the human gut, such as *Enterobacteria* phages (11.54%), *Lactobacillus* phages (2.23%) and *Lactococcus* phages (2.14%) (Fig. 3). Furthermore, bacteriophages that infect typical soil-dwelling bacteria were also identified. These corresponded to *Geobacillus* phages (7.53%), *Streptomyces* phages (3.98%) and *Delftia* phages (0.11%). Several reads were related to bacteriophages whose bacterial hosts belong to genera also including human pathogens, such as *Mycobacterium* phages (7.89%), *Vibrio* phages (0.29%), *Pseudomonas* phages (4.01%), *Streptococcus* phages (5.06%), *Staphylococcus* phages (5.07%), *Listeria* phages (3.48%), *Burkholderia* phages (3.38%) and *Clostridium* phages (3.83%) (Fig. 3). The presence of some of these bacteriophages (*Bacillus*, *Clostridium*, *Mycobacterium* and *Burkholderia* phages) was further supported by contig reconstruction (Supplementary Table S2). Moreover, contigs also identified bacteriophages that likely infect hosts known to live in aquatic environments (*Cyanophage S-TIM5*, *Synechococcus* phage S-CBS3, *Celeribacter* phage P12053L, a prophage of *Planctomyces limnophilus DSM 3776* and one uncultured phage identified in a viral metagenomic study on marine water from the Mediterranean Sea). In addition, a 1,939 bp contig matched an unidentified phage previously described in a viral metagenomic study performed on modern human stools (32) (Supplementary Table S2). Only a scaffold is available for the unidentified phage, and the matched protein is annotated as a hypothetical protein. However, this hypothetical protein is predicted to harbor a conserved domain corresponding to an N-acetylmuramoyl-L-alanine amidase. This domain is characteristic of autolysins that degrade peptidoglycans and is typically observed in bacteriophage, prophage and bacterial genomes.

The coprolite-associated DNA virome was compared to the viromes of 21 modern human stool specimens (Fig. 4). Overall, the coprolite virome displayed a higher species richness (315.279) and was more functionally diverse (average Shannon-Wiener index of 4.8693) than modern stool viromes (average species richness of 77.824 and average Shannon-Wiener index of 4.1264) (Supplementary Table S4). At the taxonomic level, the coprolite virome did not group with modern stool viromes, whereas it was metabolically more similar to some of the modern stool samples (Fig. 4). A more extensive metabolic analysis revealed a contig encoding a chloramphenicol O-acetyltransferase gene that mediates chloramphenicol resistance. This gene was found to belong to *Chryseobacterium* sp. This BLAST-based annotation was further confirmed by a phylogenetic tree constructed from the ORF of this contig (Supplementary Fig. S3).

Discussion

We report the first metagenomic analysis of a human ancient DNA virome. The use of viral metagenomics allowed us to perform a systematic research of known and unknown viruses without a priori targeting of expected viruses.

Because minimizing contamination is vital in paleomicrobiology, extensive precautions, dictated by previously published recommended protocols, were implemented to avoid contaminating the coprolite specimen (13, 15, 36, 37). The coprolite here studied was recovered from a sealed barrel, which was still intact at the time it was found, which suggests that the coprolite was protected from contamination by environmental material for centuries. Only the internal region of the coprolite was used in our experiments. We ascertained the presence of viruses by three independent approaches, i.e., electron microscopy, metagenomics and suicide PCR. Accordingly, negative controls remained negative. The PCR amplification product sequences were original, i.e., they had not been previously observed in our laboratory.

The majority of generated metagenomic sequences were of unknown origin. The known sequences corresponded to viruses that infect eukaryotes, bacteria and archaea. Eukaryotic and archaeal viral sequences were only detected at low abundances, and their presence was supported by contig recovery or confirmed by suicide *ad hoc* PCR amplifications. The majority of the sequences recovered from the coprolite corresponded to bacteriophages, as previously shown for modern stools (10, 32). Apparently, the bacteriophages originated from two sources: the environment and the digestive tract microbiota. Some of the identified bacteriophages infect bacteria that belong to genera including mammal's pathogens. Modern human stool viromes did not group with the coprolite virome at the taxonomic level, whereas the metabolic functions were conserved. This finding is consistent with those obtained in a recent study, which demonstrated that despite taxonomic inter-individual variability, the metabolic profile was significantly conserved within viromes from the same ecological niche (39). This persistence of metabolic functionalities across centuries reinforces the crucial role of the viral community in the human gastrointestinal tract.

Furthermore, we detected a contig encoding a gene for chloramphenicol resistance (the chloramphenicol O-acetyltransferase), a broad-spectrum antibiotic that inhibits bacterial protein synthesis. The presence of antibiotic resistance genes in viral metagenomes has been reported in modern human stools (32). Indeed, bacteriophages themselves constitute a reservoir of resistance genes (40-42), and bacteriophage transduction represents one important way for the lateral transfer of resistance genes between bacterial species. Indeed, phylogenetic studies have demonstrated that the evolution and dissemination of resistance genes started well before the recent use of antibiotics(43-45). Accordingly, direct evidence for the presence of antibiotic resistance genes in pre-antibiotics specimens was provided by *ad hoc* PCR amplifications using DNA extracted from 30,000-year-old permafrost sediments in Canada (46). Here, we demonstrate that bacteriophages are an ancient reservoir of resistant genes associated with human samples that date back to as far as the Middle Ages.

Overall, the present study furthers our understanding of past viral diversity and distribution and promotes the further exploration of ancient viral communities using coprolite specimens.

Acknowledgments

We thank Sonia Monteil Bouchard and Catherina Robert for technical assistance. C.D. and L.F. were funded by a Starting Grant n°242729 from the European Research Council to CD.

The authors declare no competing interests.

Legends of figures and tables

Figure 1. Transmission electron microscopy of negative stained viral particles. (**A**) Overview of stained viral particles, which vary in size and shape, isolated from the Middle Age coprolite. (**B**) A representative virion and (**C-E**) viral-like particles with icosahedral nucleocapsids and a long filament tail characteristic of *Siphoviridae* bacteriophages.

Figure 2. (A) The proportion of known and unknown reads (in percent). Reads were defined as "unknown" if they lacked homology to the non-redundant NCBI database according to a BLASTN search (E-value<1e-05) and as "known" otherwise. **(B) The relative abundance of viral families.** The relative abundance of identified viral families was estimated using the GAAS software.

Figure 3. Relative bacteriophage abundance. The relative bacteriophage abundances were estimated using the GAAS software. The hosts of the bacteriophages that were also identified in a previous study on the bacterial community associated with this specimen (unpublished data (47)) are marked with a red point.

Figure 4. Comparison between the modern human stool viromes and the coprolite virome. Principal component analysis was used to compare the viral metagenomes associated with the coprolite (highlighted in red) to those associated with modern human stool samples (S1-S21) at the taxonomic (**A**) and functional (**B**) levels.

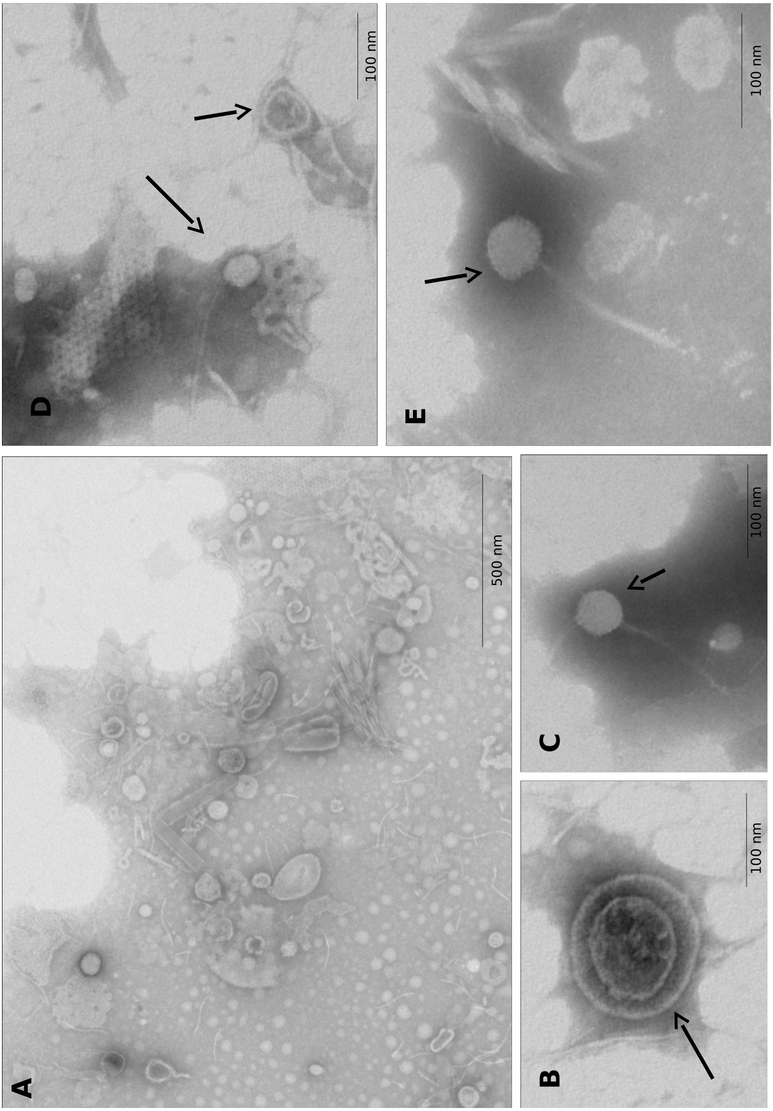


Figure 1.

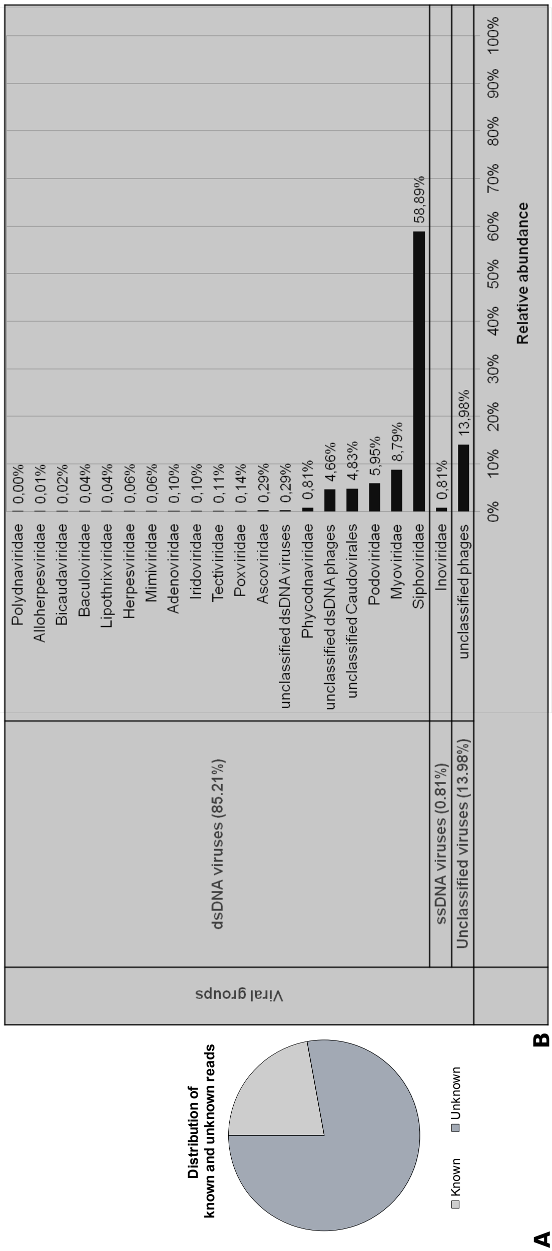

A
Distribution of known and unknown reads
Known
Unknown
B
Viral groups
dsDNA viruses (85.21%)
ssDNA viruses (0.81%)
Unclassified viruses (13.98%)
Polydnaviridae 0,00%
Alloherpesviridae 0,01%
Bicaudaviridae 0,02%
Baculoviridae 0,04%
Lipothrixviridae 0,04%
Herpesviridae 0,06%
Mimiviridae 0,06%
Adenoviridae 0,10%
Iridoviridae 0,10%
Tectiviridae 0,11%
Poxviridae 0,14%
Ascoviridae 0,29%
unclassified dsDNA viruses 0,29%
Phycodnaviridae 0,81%
unclassified dsDNA phages 4,66%
unclassified Caudovirales 4,83%
Podoviridae 5,95%
Myoviridae 8,79%
Siphoviridae 58,89%
Inoviridae 0,81%
unclassified phages 13,98%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Relative abundance

Figure 2.

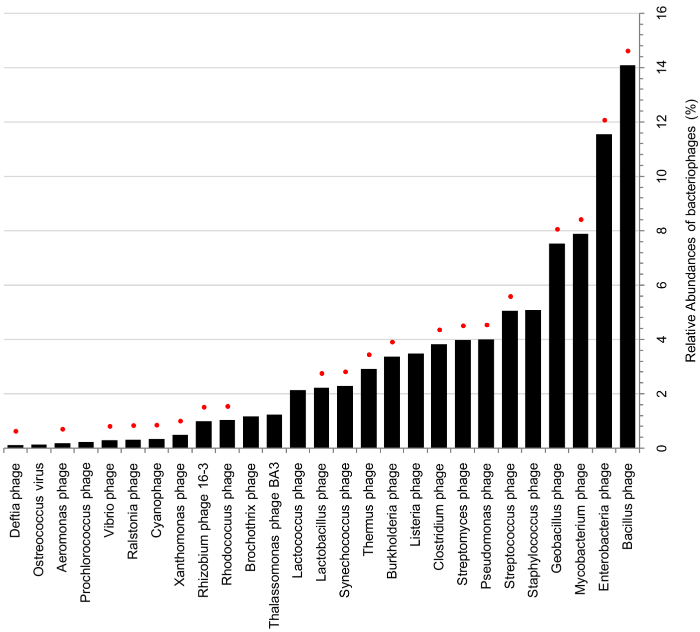

Relative Abundances of bacteriophages (%)
0
2
4
6
8
10
12
14
16
Deftia phage
Ostreococcus virus
Aeromonas phage
Prochlorococcus phage
Vibrio phage
Ralstonia phage
Cyanophage
Xanthomonas phage
Rhizobium phage 16-3
Rhodococcus phage
Brochothrix phage
Thalassomonas phage BA3
Lactococcus phage
Lactobacillus phage
Synechococcus phage
Thermus phage
Burkholderia phage
Listeria phage
Clostridium phage
Streptomyces phage
Pseudomonas phage
Streptococcus phage
Staphylococcus phage
Geobacillus phage
Mycobacterium phage
Enterobacteria phage
Bacillus phage

Figure 3.

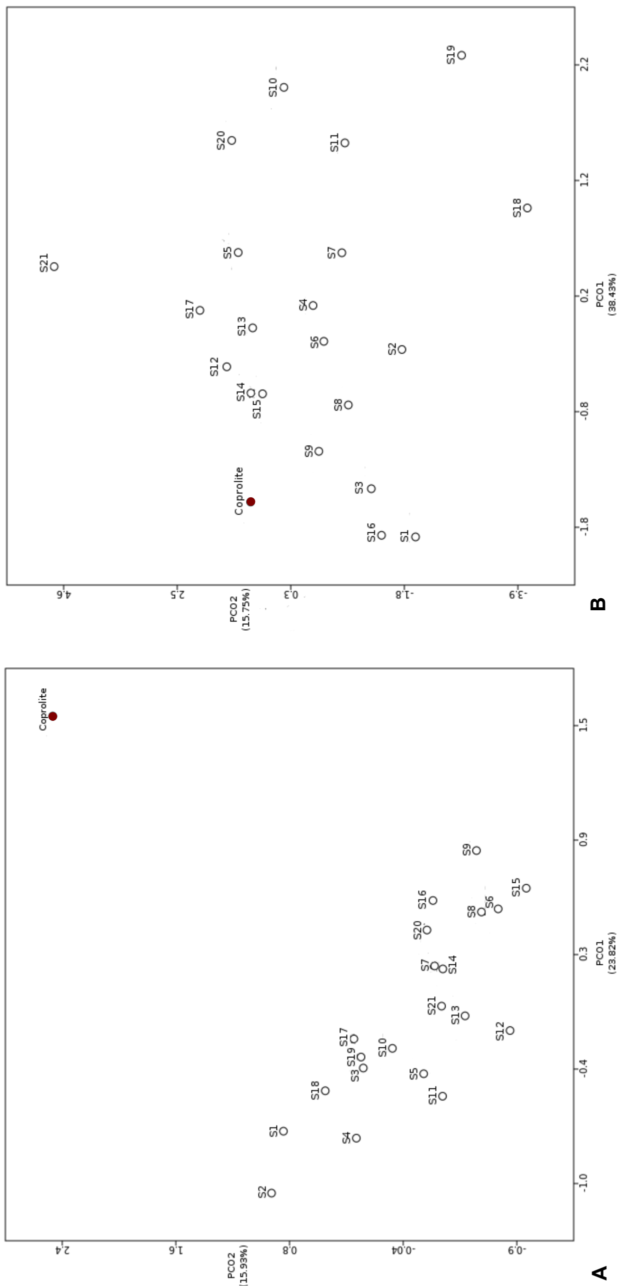

Coprolite
PCO1 (38.43%)
PCO2 (15.75%)
B
Coprolite
PCO1 (23.82%)
PCO2 (15.93%)
A

Figure 4.

References

1. **Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4:** e368.
2. **Lopez-Bueno A, Tamames J, Velazquez D, Moya A, Quesada A, Alcami A.** 2009. High diversity of the viral community from an Antarctic lake. Science. 326: 858-61.
3. **Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F.** 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature. **452:** 340-3.
4. **Kim KH, Chang HW, Nam YD, Roh SW, Kim MS, Sung Y, Jeon CO, Oh HM, Bae JW.** 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol. **74:** 5975-85.
5. **Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J, Swanepoel R, Egholm M, Nichol ST, Lipkin WI.** 2009. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog. **5:** e1000455.
6. **Anderson NG, Gerin JL, Anderson NL.** 2003. Global screening for human viral pathogens. Emerg Infect Dis. **9:** 768-74.
7. **Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, Mizutani T, Okamoto Y, Tagami M, Morita R, Maeda N, Kawai J, Hayashizaki Y, Nagai Y, Horii T, Iida T, Nakaya T.** 2009. Direct metagenomic detection of

viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS One. **4:** e4219.

8. **Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F.** 2003. Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol. **185:** 6220-3.
9. **Breitbart M, Rohwer F.** 2005. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques. **39:** 729-36.
10. **Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI.** 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature. **466:** 334-8.
11. **Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, Felts B, Mahaffy JM, Mueller J, Nulton J, Rayhawk S, Rodriguez-Brito B, Salamon P, Rohwer F.** 2008. Viral diversity and dynamics in an infant gut. Res Microbiol **159:** 367-73.
12. **Kim MS, Park EJ, Roh SW, Bae JW.** 2011. Diversity and abundance of single-stranded DNA viruses in human feces. Appl Environ Microbiol. **77:** 8062-70.
13. **Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M.** 2004. Genetic analyses from ancient DNA. Annu Rev Genet. **38:** 645-79.
14. **Willerslev E, Cooper A.** 2005. Ancient DNA. Proc Biol Sci **272:** 3-16.
15. **Drancourt M, Raoult D.** 2005. Palaeomicrobiology: current issues and perspectives. Nat Rev Microbiol. **3:** 23-35.
16. **Marennikova SS, Shelukhina EM, Zhukova OA, Yanova NN, Loparev VN.** 1990. Smallpox diagnosed 400 years later: results of skin lesions examination of 16th century Italian mummy. J Hyg Epidemiol Microbiol Immunol. **34:** 227-31.
17. **Bedarida S, Dutour O, Buzhilova AP, de Micco P, Biagini P.** 2011. Identification of viral DNA (*Anelloviridae*) in a 200-year-old dental pulp sample (Napoleon's Great Army,

Kaliningrad, 1812). Infect Genet Evol. **11:** 358-62.

18. **Biagini P, Theves C, Balaresque P, Geraut A, Cannet C, Keyser C, Nikolaeva D, Gerard P, Duchesne S, Orlando L, Willerslev E, Alekseev AN, de Micco P, Ludes B, Crubezy E.** 2012. Variola virus in a 300-year-old Siberian mummy. N Engl J Med. **367:** 2057-9.
19. **Li HC, Fujiyoshi T, Lou H, Yashiki S, Sonoda S, Cartier L, Nunez L, Munoz I, Horai S, Tajima K.** 1999. The presence of ancient human T-cell lymphotropic virus type I provirus DNA in an Andean mummy. Nat Med. **5:** 1428-32.
20. **Sonoda S, Li HC, Cartier L, Nunez L, Tajima K.** 2000. Ancient HTLV type 1 provirus DNA of Andean mummy. AIDS Res Hum Retroviruses. **16:** 1753-6.
21. **Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F.** 2009. Laboratory procedures to generate viral metagenomes. Nat Protoc. **4:** 470-83.
22. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. **75:** 7537-41.
23. **Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F.** 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol. **5:** e1000593.
24. **Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA.** 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC

Bioinformatics. **9:** 386.

25. **Liu B, Pop M.** 2009. ARDB--Antibiotic Resistance Genes Database. Nucleic Acids Res. **37:** D443-7.

26. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. **11:** 119.

27. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:** 1792-7.

28. **Talavera G, Castresana J.** 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. **56:** 564-77.

29. **Guindon S, Delsuc F, Dufayard JF, Gascuel O.** 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. **537:** 113-37.

30. **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:** W465-9.

31. **Tamura K, Dudley J, Nei M, Kumar S.** 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. **24:** 1596-9.

32. **Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD.** 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. **21:** 1616-25.

33. **Kristiansson E, Hugenholtz P, Dalevi D.** 2009. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics. **25:** 2737-8.

34. **Raoult D, Aboudharam G, Crubezy E, Larrouy G, Ludes B, Drancourt M.** 2000. Molecular identification by "suicide PCR" of *Yersinia pestis* as the agent of medieval black death. Proc Natl Acad Sci U S A. **97:** 12800-3.

35. **Rocha GCd.** 2003. Praça das Armas, Namur, Bélgica. Contribuição de um estudo paleoparasitológico. Escola Nacional de Saúde Pública-Fiocruz, Rio de Janeiro, Rio de Janeiro. 142 pp.

36. **Cooper A, Poinar HN.** 2000. Ancient DNA: do it right or not at all. Science. **289:** 1139.

37. **Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S.** 2001. Ancient DNA. Nat Rev Genet. **2:** 353-9.

38. **Martin WJ.** 1999. Stealth adaptation of an African green monkey simian cytomegalovirus. Exp Mol Pathol. **66:** 3-7.

39. **Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F.** 2009. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One. **4:** e7370.

40. **Muniesa M, Garcia A, Miro E, Mirelis B, Prats G, Jofre J, Navarro F.** 2004. Bacteriophages and diffusion of beta-lactamase genes. Emerg Infect Dis. **10:** 1134-7.

41. **Colomer-Lluch M, Jofre J, Muniesa M.** 2011. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. PLoS One. **6:** e17549.

42. **Mazaheri Nezhad Fard R, Barton MD, Heuzenroeder MW.** 2011. Bacteriophage-mediated transduction of antibiotic resistance in enterococci. Lett Appl Microbiol. **52:** 559-64.

43. **Hall BG, Barlow M.** 2004. Evolution of the serine beta-lactamases: past, present and future. Drug Resist Updat. **7:** 111-23.

44. **Garau G, Di Guilmi AM, Hall BG.** 2005. Structure-based phylogeny of the metallo-beta-lactamases. Antimicrob Agents Chemother. **49:** 2778-84.

45. **Aminov RI, Mackie RI.** 2007. Evolution and ecology of antibiotic resistance genes. FEMS Microbiol Lett. **271:** 147-61.

46. **D'Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R, Golding GB, Poinar HN, Wright GD.** 2011. Antibiotic resistance is

ancient. Nature. **477:** 457-61.

47. **Appelt S, Armougom F, Le Bailly LM, Robert C, Drancourt M, Raoult D.** 2013. Metagenomic Analysis of Middle Ages Coprolite. unpublished data

SUPPLEMENTAL MATERIAL

Supplementary Material Section 1. Prevention of contamination

After excavation, the coprolite was stored in a sterile forensic specimen bag. In 2006, the coprolite was sent to our laboratory, where it was handled only in a positive pressure room with isolated ventilation under strict aseptic conditions. Workbenches were stringently disinfected using absolute ethanol and UV-irradiated for at least 30 min. Non-disposable instruments were autoclaved. Reagents and chemicals were aliquoted from new stocks into sterile, single-use tubes and immediately discarded after use. The external portion of the coprolite was aseptically removed, and only the internal portion was used in this study. DNA extraction, PCR and post-PCR experiments were performed in separate rooms in isolated work areas. Positive controls were strictly avoided. Negative controls for DNA extraction and PCR were used in a 1:4 control:specimen ratio (1-5).

Supplementary Material Section 2. 454 sequencing.

The extracted DNA concentration (20.5ng/µL) was assessed using the Quant-iT PicoGreen kit (Invitrogen) and a Tecan GENios fluorometer. The DNA (500 ng) was nebulized, and a library was constructed according to the 454 Titanium shotgun protocol and the manufacturer's instructions. DNA fragmentation was visualized using the Agilent 2100 Bioanalyzer and a high-sensitivity labchip at an optimal size of 872 bp. The DNA stock concentration was measured using a TBS fluorometer at 2.47 E+09 molecules/µL, and the DNA was stored at -20°C. The library was clonally amplified with 2 cpb in a 1 emPCR reaction using the GS Titanium SV emPCR Kit (Lib-L) v2. The titration yield was 14.72 %. A total of 125,000 beads per project and per region were loaded onto the

GS Titanium PicoTiterPlate kit 70x75. The reaction was sequenced using the GS Titanium XLR70 Sequencing kit. The run was performed overnight and then analyzed on the ES Titanium computing cluster.

Supplementary Material Section 3. Recovery of total DNA from the coprolite.

Two different DNA extraction protocols were used, including the DNA extraction method recommended by Iniguez et al. (6) and the PowerSoilR DNA Isolation Kit (MoBio Laboratories, Inc., Carlsbad, SA) (7). One gram of the coprolite was solubilized over night at 4°C in 1mL TE buffer (Ethylenediamine Tetraacetic Acid; buffered solution, Tris HCl 10mM, EDTA 1mM, pH 8). A 500-µL aliquot of the solution was used for total DNA extraction as previously described (Iniguez et al. 2006), except that incubations into TE and digestion buffers were shortened to 1 day. TE buffer without coprolite was used as a negative control. For the DNA extraction using the PowerSoilR DNA Isolation Kit (MoBio Laboratories, Inc., Carlsbad, USA), 500 µL of solubilized coprolite specimen were also used. Incubation was extended to 24h/56°C in PowerSoilR bead tubes containing sodium dodecyl sulfate (SDS) and digestion buffer C1, under contentiously rotation followed by shaking in a Bio 101 FastPrep instrument (Qbiogene) at level 6.5 (full speed) for 95s. DNA extraction was performed according to the manufacturer's instructions. Extraction batches contained also negative controls composed of PowerSoilR bead tubes without coprolite. Total DNA extracts from both protocols above were pooled together in a 1:1 ratio.

Supplementary Material Section 4. Specific PCR amplifications and sequencing.

The PCRs were performed in a 50-µL final volume that included 1 x PCR buffer, 2 µL 25 mM $MgCl_2$, 200 µM of each dNTP, 1 µL of each 10 pM primer, 31.15 µL ddH_2O, 1 unit of

2

HotStar Taq Polymerase (Invitrogen, Villebon sur Yvette, France) and 57-112 ng of DNA extract. The total DNA recovered from the coprolite served as the DNA template in these reactions. The PCR steps comprised an initial incubation at 95°C for 15 min, 36-40 denaturation cycles at 95°C for 1 min, annealing for 30 sec at the corresponding primer annealing temperatures, elongation at 72°C for 90 sec and a final elongation at 72°C for 10 min; all of these steps were performed in a Gene Amp PCR System 2700 ABI Thermocycler (Applied Biosystems, Villbon sur Yvette, France). The PCR products were analyzed using a 2 % agarose gel (UltraPureTM agarose, Invitrogen, Villbon sur Yvette, France) and purified using the QIAquick PCR Purification Kit (QIAGEN, Courtaboeuf, France). All of the PCR products were sequenced in a final volume of 20 of µL (1 x sequencing buffer, 3.2 pM forward or reverse primer, 4 µL of BigDye Terminator V1.1 mix (Applied Biosystems), 7.4 µL ddH_2O and 4 µL of PCR product) after purification using the Sephadex Gel Filtration in the ABI PR ISM 3130xl genetic sequencer (Applied Biosystems, Villbon sur Yvette, France). The sequences were assembled using the Chromaspro software and compared with reference GenBank database sequences using NCBI BLAST searches.

Supplementary Table S1 | Reads before and after preprocessing and the read assembly process output.

Total number of raw reads	**30,654**
Total number of preprocessed reads	29,811
Average read length (bp)	358
Minimal read length (bp)	77
Maximal read length (bp)	574
Average GC content	47%
Total number of contigs	**1,464**
Total number of large contigs	104
Length of the longest contig (bp)	12,500
Number of assembled reads	13,682
Number of partially assembled reads	6,282
Number of singletons	9,481

Supplementary Table S2. Viral contigs[#].

Contig ID	Contig length (bp)	Hit description	E-value	Hit accession ID	Percent ID
contig00065	1,048	hypothetical protein MIV033L [Invertebrate iridescent virus 3]	2e-07	YP_654605	40.79
contig01018	421	tail fiber protein [Emiliania huxleyi virus PS401]	7e-06	AET73308	51.72
contig00021	543	gp160 [Mycobacterium phage Optimus]	5e-13	AEJ92216	45.78
contig00068	1,155	unnamed protein product [Synechococcus phage S-CBS3]	5e-55	YP_004421762	34.48
contig00101	1,005	putative prophage repressor protein [Burkholderia phage BcepC6B]	7e-08	YP_024964	48.39
contig00157	1,939	hypothetical protein 2200_scaffold2278_00035 [unidentified phage] stool sample from healthy person. Minot et al.. 2012	9e-25	AFB75852	52.44
contig00184	445	putative phage tail fiber protein [Celeribacter phage P12053L]gb|AFM54660.1|	2e-13	YP_006560940	66.15
contig00261	926	gp218 [Bacillus phage G]	3e-26	AEO93477	39.56
contig00396	689	hypothetical protein [uncultured phage MedDCM-OCT-S04-C136]	3e-06	ADD94244	35.71
contig00474	1,719	gp27 [Bacillus phage G]	7e-07	AEO93298	28.98
contig00611	488	phage terminase. large subunit. PBSX family [Thermoanaerobacterium phage THSA-485A]	2e-26	YP_006546303	49.64
contig00629	463	putative nuclease SbcCD D subunit [Bacillus phage BCP78]gb|AEW47679.1	6e-07	AEW47191	35.45
contig00656	402	gp128 [Bacillus phage G]	1e-22	AEO93390	68.42
contig00664	477	hypothetical protein OSG_eHP6_00230 [Environmental Halophage eHP-6]	1e-07	AFH21679	56.00
contig00702	554	conserved hypothetical protein [Clostridium phage D-1873]gb|EES90342.1|	2e-19	ZP_04863741	46.94
contig00926	473	hypothetical protein [Riemerella phage RAP44]	2e-15	AEB71650	44.19
contig01108	731	tail fiber protein [Cyanophage S-TIM5]	4e-07	AEZ65603	36.27
contig01456	734	DNA polymerase [Bacteriophage APSE-4]	6e-28	ACJ10154	34.80
contig00075	3,946	phage tape measure protein [Planctomyces limnophilus DSM 3776]	9e-62	YP_003632269	58.99
contig00078	8,733	phage major capsid protein. HK97 family [Clostridium cellulovorans 743B]	4e-100	YP_003842316	46.20
contig00073	436	phage-related protein-like protein [Acetivibrio cellulolyticus CD2]	1e-07	ZP_09464888	41.79
contig00076	2,218	phage tail tape measure protein. TP901 family [Thermoanaerobacter italicus Ab9]	2e-54	YP_003477199	64.49
contig01464	1,248	putative DNA-binding phage protein [Burkholderia cenocepacia J2315]	2e-18	YP_002234532	39.44
contig00055	639	phage-like protein [Thermoanaerobacterium saccharolyticum JW/SL-YS485]	1e-06	YP_006393008	41.54
contig00048	6,115	phage Gp37Gp68 [Mycobacterium massiliense 1S-151-0930]	3e-75	EIU63155	47.46
contig00230	7,394	TP901 family phage tail tape measure protein [Thermosipho melanesiensis BI429]	7e-10	YP_001306681	53.33
contig00143	1,495	Prophage LambdaBa04. site-specific recombinase. phage integrase [Bacillus cereus AH621]	3e-22	ZP_04293086	39.16
contig00181	3,804	phage-like protein [Bacillus licheniformis WX-02]	1e-66	EID48485	63.45

[#]The contig identifier, its length (bp) and the annotation according to the best BLAST hit (BLASTX versus the non-redundant NCBI database, E-value<1e-05) are summarized. The E-value, the hit accession identifier and the percent of identity are also provided.

Supplementary Table S3. Amplicons generated by ad hoc suicide PCR amplifications.

Species	Genomic Region	Size (bp)	Covera ge	E-value	Identity
Moumouvirus sp. GenBank Accession No. GU265560.1	Polymerase b	167	100%	2e-38	85%
Stealth Virus 1 clone 3B43 GenBank Accession No. AF191073.1	Genomic Sequence	133	96%	8e-56	98%

Supplementary Table S4. Viral metagenomes used for comparison, illustrated in Fig. 4.

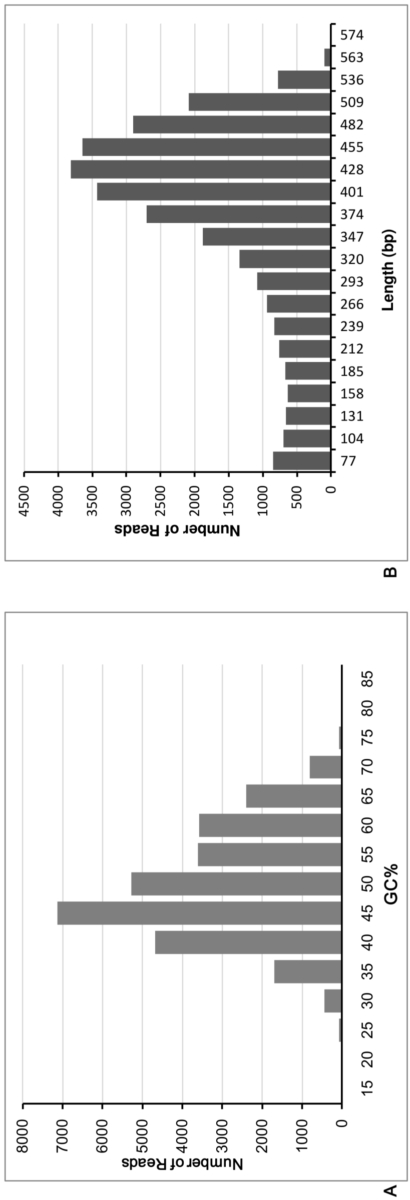
Identifier	Sample name	Sample type	Species richness (effective number of species)	Functional diversity (Shannon-Wiener index)	Reference
S1	F-A	Modern stool	138.760	4.6275	[9]
S2	F-B	Modern stool	141.912	4.1130	[9]
S3	F-C	Modern stool	127.776	4.7134	[9]
S4	F-D	Modern stool	20.232	3.9289	[9]
S5	F-E	Modern stool	42.160	4.1415	[9]
S6	X-1	Modern stool	112.061	4.4603	[10]
S7	L1-1	Modern stool	67.221	4.1169	[10]
S8	L1-8	Modern stool	99.298	4.4523	[10]
S9	L2-2	Modern stool	121.056	4.6202	[10]
S10	L2-7	Modern stool	65.893	3.4547	[10]
S11	L2-8	Modern stool	40.181	3.2639	[10]
S12	H1-1	Modern stool	130.359	4.0687	[10]
S13	H1-2	Modern stool	118.034	4.1585	[10]
S14	H1-7	Modern stool	60.617	4.4489	[10]
S15	H1-8	Modern stool	105.572	4.5599	[10]
S16	H2-1	Modern stool	56.515	4.7019	[10]
S17	H2-8	Modern stool	76.075	4.0324	[10]
S18	L3-1	Modern stool	32.353	4.2648	[10]
S19	L3-2	Modern stool	36.772	3.1021	[10]
S20	L3-7	Modern stool	21.855	3.6276	[10]
S21	L3-8	Modern stool	19.611	3.7963	[10]
Coprolite	Coprolite	Middle Age stool	315.279	4.8693	Present study

Supplementary Figures

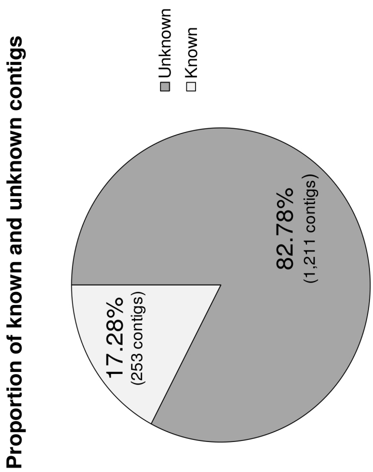
Supplementary Figure S1. GC content and read length distributions. (**A**) The GC content distribution of preprocessed metagenomic reads. The y-axis indicates the number of reads, whereas the x-axis indicates the (**B**) The read length distribution. The y-axis indicates the number of reads, whereas the x-axis indicates the length (in base pairs).

Supplementary Figure S2. Proportion of known and unknown contigs (in percent). Contigs were defined as "unknown" if they lacked homology to the non-redundant NCBI database according to a BLASTN search (E-value) and as "known" otherwise.

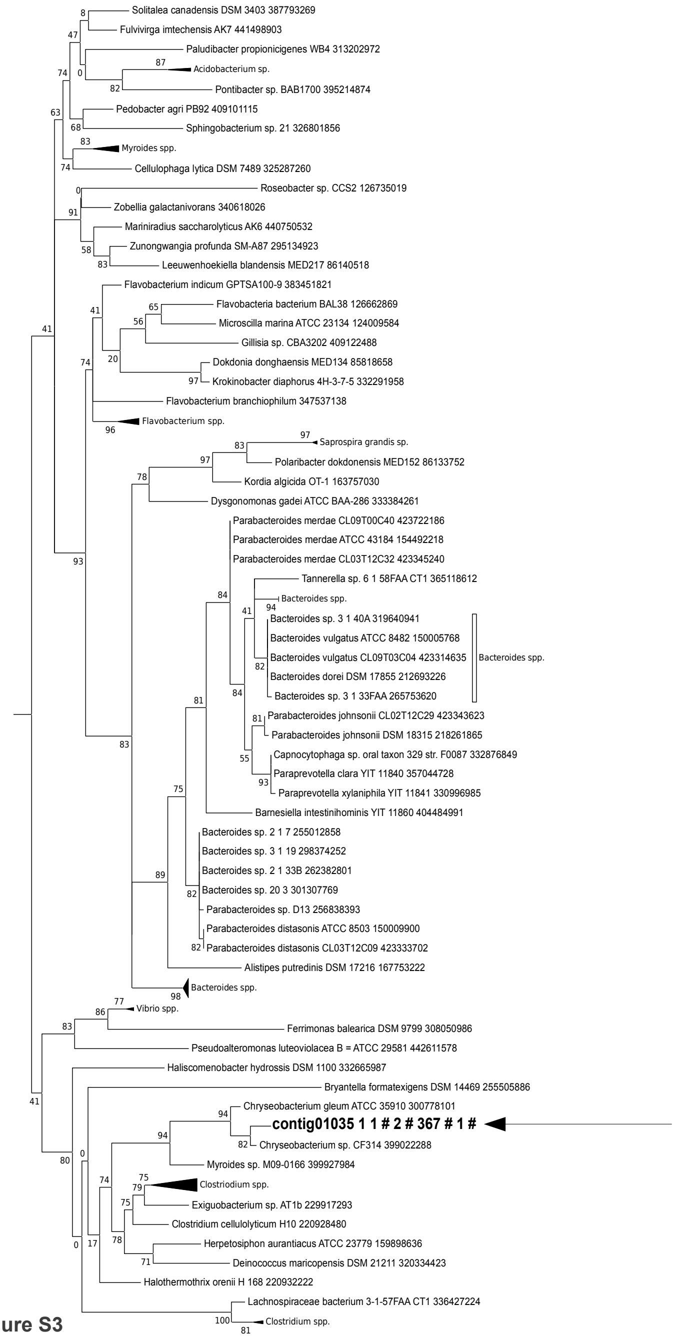
Supplementary Figure S3. Phylogenetic tree of a chloramphenicol O-acetyltransferase. A phylogenetic tree was generated from the translated open reading frame of a contig encoding a chloramphenicol O-acetyltransferase. The tree was constructed using the PhyML algorithm with a bootstrap of 100. The bootstrap support is reported for each branch.



Supplementary Figure S1



Supplementary Figure S2



Supplementary Figure S3

SUPPLEMENTARY REFERENCES

1. **Cooper A, Poinar HN.** 2000. Ancient DNA: do it right or not at all. Science. **289:** 1139.
2. **Willerslev E, Cooper A.** 2005. Ancient DNA. Proc Biol Sci. **272:** 3-16.
3. **Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S.** 2001. Ancient DNA. Nat Rev Genet. **2:** 353-9.
4. **Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M.** 2004. Genetic analyses from ancient DNA. Annu Rev Genet. **38:** 645-79.
5. **Drancourt M, Raoult D.** 2005. Palaeomicrobiology: current issues and perspectives. Nat Rev Microbiol. **3:** 23-35.
6. **Iniguez AM, Reinhard K, Carvalho Goncalves ML, Ferreira LF, Araujo A, Paulo Vicente AC.** 2006. SL1 RNA gene recovery from Enterobius vermicularis ancient DNA in pre-Columbian human coprolites. Int J Parasitol. **36:** 1419-25.
7. **Tito RY, Macmil S, Wiley G, Najar F, Cleeland L, Qu C, Wang P, Romagne F, Leonard S, Ruiz AJ, Reinhard K, Roe BA, Lewis CM, Jr.** 2008. Phylotyping and functional analysis of two ancient human microbiomes. PLoS One. **3:** e3703.

4.2 Article 3. Viral communities associated with human pericardial fluids in idiopathic pericarditis

Viral communities associated with human pericardial fluids in idiopathic pericarditis.

Fancello Laura[1], Monteil Sonia[1], Popgeorgiev Nikolay[1], Rivet Romain[1], Gouriet Frédérique[1], Fournier Pierre-Edouard[1], Raoult Didier[1] and Desnues Christelle[1,*]

Submitted to Plos One.

[1] Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE CNRS-IRD UMR 7278, Aix-Marseille Université, Faculté de médecine, 27, Bd Jean Moulin, 13385 MARSEILLE France.

[*] Corresponding author. Email: christelle.desnues@univ-amu.fr

Preamble to article 3

Since the publication of the first human-associated virome, a few viromes have been generated in human clinical studies to discover the etiology of unexplained epidemic outbreaks or to investigate some common idiopathic diseases, such as some acute diarrhea cases, respiratory infections, encephalitis or chronic fatigue syndrome [76, 68, 73, 66, 77, 70, 78]. Pericardits is a human disease defined by an inflammation of the pericardium and most commonly diagnosed as being of infectious origin. *Mycobacterium tuberculosis* is the most frequent bacterial cause of pericarditis, whereas echovirus, coxsackievirus, influenza, Epstein-Barr virus, cytomegalovirus, adenovirus, varicella, rubella, mumps, hepatitis B, hepatitis C, human immunodeficiency virus, parvovirus B19 and human herpesvirus 6 are the principal viral agents detected in pericarditis cases. However, the determination of pericarditis etiology is difficult and 40% to 85% of cases are still classified as idiopathic [98, 99]. Most of these cases are believed to be due to infection by undetected, unsuspected or unknown viruses.

The URMITE laboratory disposes of a very wide cohort of patients affected by pericarditis and the etiological diagnosis of pericarditis is one of its priorities [100, 98]. Due to its connection with Marseille's hospitals, hundreds of pericardial fluids samples are available. Here, I present a study on ten pericarditis cases investigating their etiological causes through a viral metagenomic approach. We generated for the first time the viromes associated to human pericardial fluids: seven were generated from the pericardial fluid of patients affected by pericarditis of unknown etiology, one from the pericardial fluid of a sudden infant death case, one from a patient affected by pericarditis caused by a human herpesvirus type 3 (positive control) and one from a pool of pericardial fluids of five different individuals affected by pericarditis of non-infectious origin (negative control). This work revealed that human pericardial fluids, although constituting a closed system, are not sterile. *Anelloviridae*, especially torque teno viruses, were observed for the first time in pericardial fluids from five patients whereas they were absent in the pool of negative controls. Moreover, co-infection by different genotypes of torque teno viruses was observed in these samples. Similarly, in the positive control, we found, as expected, the diagnosed human herpesvirus 3 but also previously undetected papillomaviruses. Bacteriophages were detected in three patients, in particular bacteriophages whose bacterial hosts belong to genera including human pathogens: *Staphylococcus, Enterobacteria, Streptococcus, Burkholderia* and *Pseudomonas*. However, except for one sample that came from a sudden infant death patient, the corresponding bacteria were not detected in the samples during previous hospital routine screenings. We hypothesize that bacteriophages are the trace of an earlier or current bacterial infection, that has not been detected by routine diagnostic tests as early antibiotic therapies and/or clearance by the host immune system made the bacterial DNA to be below the detection threshold of the diagnostic method used. As bacteriophages are estimated to outnumber their bacterial hosts by a ratio of approximately ten to one [2, 101], the detection of bacteriophages could thus be more sensitive than the detection of their bacterial hosts.

In conclusion, we provide the very first assessment of a DNA virome in human pericardial fluids as well as preliminary clues on potential unsuspected or undetected pathological agents involved in pericarditis cases of unknown etiology, for which further investigations are necessary.

VIRAL COMMUNITIES ASSOCIATED WITH HUMAN PERICARDIAL FLUIDS IN IDIOPATHIC PERICARDITIS

Fancello L., Monteil S., Popgeorgiev N., Rivet R., Gouriet F., Fournier PE., Raoult D., and Desnues C.

[1]Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE CNRS-IRD UMR 7278, Aix-Marseille Université, Faculté de médecine, 27, Bd Jean Moulin, 13385 MARSEILLE France

Keywords: virus, metagenomics, human pericardial fluids, idiopathic pericarditis, bacteriophages

*Corresponding author: christelle.desnues@univ-amu.fr

ABSTRACT

Pericarditis is a common human disease defined by inflammation of the pericardium. Currently, 40% to 85% of pericarditis cases have no identified etiology. Most of these cases are thought to be caused by an infection of undetected, unsuspected or unknown viruses.

In this work, we used a culture- and sequence-independent approach to investigate the viral DNA communities present in human pericardial fluids. Seven viral metagenomes were generated from the pericardial fluid of patients affected by pericarditis of unknown etiology. In addition, 1 metagenome was generated from the pericardial fluid of a sudden infant death case, 1 from that of a patient affected by pericarditis caused by herpesvirus type 3 (positive control) and 1 from a pool of pericardial fluids from 5 different individuals affected by pericarditis of non-infectious origin (negative control).

The results showed a significant presence of torque teno viruses in 5 patients, while herpesviruses and papillomaviruses were present in the positive control. Co-infections by different genotypes of the same viral type (torque teno viruses) or different viruses (herpesviruses and papillomaviruses) were observed. Sequences related to bacteriophages infecting *Staphylococcus, Enterobacteria, Streptococcus, Burkholderia* and *Pseudomonas* were also detected in three patients.

This study detected torque teno viruses and papillomaviruses, for the first time, in the pericardial fluids of patients with pericarditis. Bacteriophage sequences were hypothesized to be present due to trace infections by their corresponding bacterial hosts.

INTRODUCTION

Pericarditis is defined by inflammation of the pericardium, the sac-like membrane surrounding the heart. Pericarditis can be of non-infectious or infectious origin. Non-infectious pericarditis, which represents approximately 1/3 of cases, can be due to autoimmune or neoplastic diseases, metabolic disorders or traumas [1]. The infectious forms of pericarditis are mainly of viral or bacterial origin. *Mycobacterium tuberculosis* is the most frequently diagnosed bacterial agent and usually occurs in developing countries or immune-compromised hosts [1,2]. In Western countries, the large majority of acute pericarditis cases are viral, and echovirus, coxsackievirus, influenza, Epstein-Barr virus, cytomegalovirus, adenovirus, varicella, rubella, mumps, hepatitis B, hepatitis C, human immunodeficiency virus, parvovirus B19 and human herpesvirus 6 are the principal infectious agents [1–3].

Determination of pericarditis etiology is difficult, and a large number of cases remain unexplained. Indeed, idiopathic pericarditis represents between 40% and 85% of pericarditis cases, and most of them are suspected to be of viral origin [2,4]. Because serological tests are only suggestive and not diagnostic, the diagnosis of viral pericarditis requires invasive methods that evaluate pericardial effusion or tissue [3]. For example, in a study evaluating the etiology of 204 pericarditis cases, Levy et al. showed that no statistically significant difference between patients and controls could be observed in serological tests targeting adenovirus, influenza virus or cytomegalovirus, thus showing that the specificity of some of these tests is very low [5]. Moreover, classical diagnostic techniques do not allow the detection of unknown viruses or already-known viruses not suspected of involvement in the disease. Thus, an “a priori-free” approach needs to be applied to clarify the etiology of idiopathic pericarditis cases. Recently, the advent of next generation sequencing technologies has allowed the development of a metagenomic approach for a more complete and unbiased view of viral communities associated with a sample. This approach has already been successfully applied to human clinical studies and has led to the discovery of new potential pathogenic viruses as well as the identification of unsuspected viruses in some idiopathic diseases [6–13]. In the work presented here, we applied a metagenomic approach to investigate the DNA viral community colonizing human pericardial fluids and identify unsuspected or new, divergent viruses likely to be responsible for unexplained cases of pericarditis.

To this end, we generated viral DNA metagenomes of 7 pericardial fluid samples collected from patients affected by idiopathic pericarditis and from 1 patient as part of sudden infant death protocol and compared these metagenomes to a viral metagenome generated from a pool of pericardial fluids from 5 individuals affected by pericarditis of non-infectious origin.

MATERIALS AND METHODS

Samples collection

A total of 14 pericardial fluid samples were used in this study. All samples were collected between 2007 and 2010 in 5 French hospitals (the Hospital of Niort and Timone, Nord, Conception and Clairval hospitals in Marseille). Except for patient P4, samples P1 to P8 were collected from patients affected by pericarditis of unknown etiology (i.e., classical diagnostic tests performed at the hospital were all negative). Patient P4 was included as part of a sudden infant death protocol, although the culture showed a polymicrobial infection (*Escherichia coli* and *Pseudomonas* species). One sample (the positive control) came from a patient affected by a pericarditis caused by human herpesvirus 3. Five samples came from different individuals affected by pericarditis of non-infectious origin (*i.e.,* traumatic or post-surgery) and were pooled together (forming the negative control). Volumes of available pericardial fluids varied between approximately 130 µl and 2 ml. Samples were conserved at -80°C until processing. All samples are listed in Table 1 along with the age, gender, sample volume and hospital diagnosis.

Ethics Statement

All samples used in this study were collected from human subjects using a protocol approved by the local ethics committee IFR48 (Marseille, France). Written informed consent was obtained from the parents or legal guardians of all subjects.

Viral isolation and sequencing

Each sample was centrifuged at low speed to eliminate proteins and cellular debris. The resulting supernatant was collected and filtered through 0.45-μm filter pore. Virus-like particles were concentrated by ultracentrifugation at 55,000 g for 60 min. The resulting pellet was resuspended in a phosphate buffered saline solution (PBS) previously filtered at 0.02 μm. Purified VLPs were treated with DNAse and RNAse to remove any residual host and bacterial DNA as previously described [14]. Viral DNA was then extracted using the High Pure Viral Nucleic Acid Kit (Roche Applied Science, Inc, Branford, CT) following the manufacturer's recommendations. Extracted DNA was amplified using the commercial Illustra™ GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Freiburg, Germany) to generate sufficient material for shotgun 454 pyrosequencing library preparation. Amplified DNA was purified on silica columns (Qiagen Inc, Valencia, CA) to remove the enzyme, dNTPs and primers, and subsequently sequenced on a 454 Life Sciences Genome Sequencer FLX instrument using titanium chemistry (Roche Applied Science, Inc, Branford, CT).

Reads pre-processing and annotation

The generated sequences were screened to remove the exact and nearly identical duplicates using the CD HIT 454 program [15], available under the CAMERA 2.0 web portal [16]. All screened viromes are publicly available on the Metavir web-server (Table 1) [17]. A BLASTN search against the non-redundant NCBI database ($E<1e^{-05}$) was performed. Reads having no significant hits were classified as "unknown reads", whereas those with significant similarity to sequences stored in the NCBI database were classified as "known reads".

Estimation of viral genotype abundances

The GAAS (Genome relative Abundance and Average Size) program [18] was used to estimate the relative abundance of each viral genotype present in the metagenomes. Briefly, GAAS performs a BLASTX search against the Viral Refseq database and normalizes the number of reads matching one viral genotype to the genome length of that viral genotype. The BLASTX search was performed with an E-value of $1e^{-05}$.

Mapping

Reads were mapped onto reference genomes using the CLC Genomics Workbench version 4.9 (www.clcbio.com) with a minimal length fraction of 0.5 and a minimal similarity of 0.8 as mapping parameters.

Assembly and contig analysis

Read assemblies were performed for each sample using the GS *De Novo* Assembler (Roche Applied Science, Inc, Branford, CT), an application especially suited for the analysis of the 454 Life Sciences Genome Sequencer FLX-generated data. We chose a minimum overlap length of 35 bp and a minimum overlap identity of 98%. Only contigs longer than 400 bp were kept for further analyses. We classified as “large contigs” those spanning more than 1,500 bp. Contigs were classified as known if they had a significant hit in a BLASTN search against the non-redundant NCBI database (E-value$<1e^{-05}$), otherwise they were classified as unknown. Contigs were then annotated by a BLASTX search against the non-redundant NCBI database (E-value<1e^{-05}).

PCR for specific sequences

To confirm the presence of sequences found *in silico*, standard PCRs targeting specific viral genotypes recovered by the GAAS analysis were performed. All primers used are listed in Supplementary Table S1.

Torque teno viruses (TTV). Genomiphi-amplified DNA extracted from sample P7 was tested for torque teno virus presence by nested PCR using the primers previously described by Peng et al., which detect all genotypes and genetic groups currently identified for torque teno viruses [19]. These primers target the UTR universally conserved region of TTV genomes. Nested PCR was carried out with Phusion High-fidelity DNA Polymerase (Finnzymes Oy, Thermo Fisher Scientific) using 1 µl of template DNA. Three specific primer pairs were designed using Primer3 [20] to target the 3 longer TTV-like contigs assembled from sample P7. Standard PCRs with the designed primers were performed using Phusion High-fidelity DNA Polymerase (Finnzymes Oy, Thermo Fisher Scientific) on the remaining amplified viral DNA for sample P7; specific PCR products were purified and sequenced by Sanger technology for verification.

Enterobacteria, Staphylococcus, Pseudomonas, Burkholderia and Streptococcus bacteriophages. Primers from literature were used when available [21,22]. Otherwise, specific primers for the bacteriophages detected *in silico* were designed by Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/), PhiSigns [23] or Primer3, followed by BLAST verification of primers' specificity [20]. PCRs were performed on the Genomiphi-amplified DNA from each sample.

RESULTS

In this study, viral metagenomes from 10 human pericardial fluid samples were generated. Viral-like particles were purified and treated with DNAse and RNAse to minimize contamination by host and bacterial DNA as previously described [14]. Viral DNA was extracted, amplified in a sequence-independent manner and sequenced using the 454/Roche pyrosequencing technology. A total of 313,157 reads were generated. After removal of artificial duplicates, a total of 235,638 reads were left with an average length between 246 and 361 bp, depending on the sample (Supplementary Table S2). According to a BLASTN search (E-value<1e^{-05}) against the non-redundant NCBI database, between 0.95% and 23.18% of the reads had no significant similarity to known sequences and were thus classified as “unknown” (Fig. 1). GAAS [18] was used to more accurately estimate the relative abundance of each viral genotype (Fig. 2). Because longer genomes have higher probabilities of being sequenced in the shotgun approach, they are overrepresented in terms of number of sequences in the metagenome. To avoid this bias, GAAS normalizes the number of sequences associated with a viral genotype to its genome length. Relative abundances estimated by GAAS were synthesized and presented at the viral family level except for all bacteriophage-related sequences, which were grouped together. Reads assembly was performed on each metagenome and a total of 1,864 contigs were obtained, including 233 large contigs (longer than 1,500 bp). Depending on the sample, up to 76.66% of the reads were assembled into contigs, which were annotated on the basis of their best hit in a BLASTX search against the non-redundant NCBI database (E-value<$1e^{-05}$). The total number of contigs and large contigs generated for each sample as well as the percentage of assembled reads is reported in Supplementary Table S3.

Human herpesvirus 3 and papillomaviruses co-infection

As expected, the positive control showed high abundance of *Herpesviridae* (6,674 reads) and, in particular, reads matching human herpesvirus 3 (HHV3). Indeed, 5,990 reads related to HHV3 were recovered (BLASTX against the Viral Refseq Genomes Database, E-value<$1e^{-05}$). After mapping of the reads onto the reference HHV3 genome, 98% (122,502 bp) of the genome was reconstructed (Supplementary Figure S1) with an average coverage depth of 18.93 reads. In addition to HHV3, 173 reads matching viruses from the *Papillomaviridae* family were recovered and 11 contigs could be generated after *de novo* assembly (Table 2). Contig annotation retrieved 4 different types of human papillomaviruses (HPVs): HPV isolate 915 F 06 002 KN1, HPV type 50, HPV type 80, and HPV type 12 (Table 2). Major and minor capsid proteins as well as replication and regulatory proteins were encoded by the contigs (BLASTX against the non-redundant NCBI database, E-value<$1e^{-05}$). One contig (contig 37, 1196 bp long) was found to harbor two different ORFs matching the E2 and L2 genes of HPV type 50 (BLASTX against the non-redundant NCBI database, E-value<$1e^{-05}$). Both ORFs were incomplete, but their position on the contig was consistent with that of the E2 and L2 genes on the HPV type 50 genome (Figure 3). More than 50% of each of the four HPV genomes could be reconstructed by mapping (Supplementary Figure S2 and Supplementary Table S4), and the subsets of reads that mapped onto each reference genome were mutually exclusive, supporting the presence of multiple different genotypes of papillomaviruses. All recovered HPV genotypes were HPVs first isolated from human skin under healthy (HPV type 80, HPV isolate 915 F 06 002 KN1) or disease (HPV type 50, HPV type 12) conditions [24–27].

Human endogenous retroviruses

In contrast to the positive control, the negative control contained only one read that matched a herpesvirus (HHV3) and none related to papillomaviruses. The viral DNA community of the negative control was dominated by sequences from the *Retroviridae* family (90.85%) (Figure 2) and no contig could be reconstructed for any eukaryotic virus or bacteriophage. A similar profile was observed for the viral communities of patients P5 and P8. Indeed, viromes associated with these samples were dominated by *Retroviridae* (90.51% and 90.9% for P5 and P8, respectively) (Figure 2) and no viral contigs could be reconstructed. Analysis of the *Retroviridae*-related sequences (reads and contigs) identified them as human endogenous retroviruses, suggesting they may be derived from contamination by host DNA.

Variable *Anelloviridae* genotypes were detected in 5 patients

According to GAAS analysis, five metagenomes (P1, P2, P3, P6, P7) contained *Anelloviridae* (57.04%, 68.76%, 26.05%, 60.23%, and 98.59%, respectively) (Figure 2) and, in particular, torque teno viruses (TTVs) (49.6%, 66.97%, 26.05%, 60.23% and 98.59%, respectively). In total for *Anelloviridae* (torque teno viruses and others), we recovered 20 reads for P1, 108 for P2, 3 for P3, 56 for P6 and 1628 for P7 (BLASTX search against the non-redundant NCBI database, E-value<$1e^{-05}$, Supplementary Table S5). Among the recovered reads, there were reads matching unclassified *Anelloviridae* as well as reads matching Alfatorqueviruses, Betatorqueviruses and Gammatorqueviruses. More than 70% of several TTV genomes could be reconstructed by mapping reads from samples P2, P6 and P7. These genomes corresponded to the TTV clone Saf-09 in sample P2 and the TTV strain SIA109 and TTV 24 clone Saa-01 in sample P6 (Supplementary Figure S3 A and B). In addition, we were able to fully reconstruct the genome of three TTV isolates (torque teno virus TTV-HD14h, torque teno virus isolate TTVyon-LC011 and torque teno virus isolate JT41F), the genome of TTV 29 and that of Micro TTV isolate microTTV-HD14.2 in sample P7 (Supplementary Figure S3, C). Several contigs were also assembled *de novo* and annotated as TTVs (BLASTx against the non-redundant NCBI database, E-value<$1e^{-05}$) (Supplementary Table S6). The presence of contig 127, contig 129 and contig 64 was verified by standard PCR in sample P7, which, according to *in silico* analyses, had the greatest abundance of TTVs. PCRs targeting contig 127 and contig 64 showed non-specific amplification, while PCR targeting contig 129 yielded only a specific PCR product that was further verified by Sanger sequencing. This contig contained one ORF that matched the ORF1 of torque teno virus and one GC-rich region (85.7%), typical of TTV genome organization [19,28,29] (Figure 4). Torque teno virus presence in the same sample was further verified by a nested PCR targeting the TTV UTR region, which is conserved among all TTV genotypes and genetic groups. The PCR result was positive and specific. Overall, TTVs were detected in 5 out of 8 patients and in none of the 5 pooled negative controls.

Bacteriophages as traces of bacterial infections

GAAS analysis showed the presence of bacteriophages in 3 metagenomes (P1, P3, P4) with relative abundances of 4.74%, 33.69% and 49.32%, respectively. The total number of bacteriophage-related reads recovered for P1, P3 and P4 was 34, 91 and 455 sequences, respectively (BLASTX search against the Viral Genome database, E-value<1e^{-05}). *Enterobacteria* phages, *Pseudomonas* phages, *Staphylococcus* phages, *Streptococcus* phages, *Pseudomonas* phages and *Burkholderia* phages were among the most represented bacteriophage-related sequences (Table 3). Moreover 30 contigs matching the *Escherichia* phage ECML-117 were assembled from sample P4. Specific PCRs were performed on P1, P3 and P4 samples to verify the presence of the bacteriophage sequences recovered *in silico.* We used primers targeting the genomes of the bacteriophages of interest rather than primers designed specifically from the recovered metagenomic sequences. We performed a PCR targeting *Enterobacteria* phage lambda in sample P1 and targeting *Staphylococcus* phages of groups 3A-like and 11A-like and *Streptococcus* prophage EJ-1 in sample P3. For sample P4, we performed PCRs targeting Stx converting phage II, *Enterobacteria* phage BP-4795, *Pseudomonas* phage F8, *Pseudomonas* phage LMA2, *Pseudomonas* phage LBL3, *Burkholderia amphibaria* phage BcepF1 and *Enterobacteria* phage P1. Non-specific amplification was observed for *Enterobacteria* phage lambda in sample P1, for *Streptococcus* prophage EJ-1 in sample P3 and for Stx converting phage II, *Enterobacteria* phage BP-4795, *Pseudomonas* phage F8, *Pseudomonas* phage LMA2, *Pseudomonas* phage LBL3, *Burkholderia amphibaria* phage BcepF1 and *Enterobacteria* phage P1 in sample P4 (Supplementary Table S7). For sample P3, we obtained unique amplification products at the expected length (Supplementary Table S7) for *Staphylococcus* phages of groups 3A-like and 11A-like. The PCR products were sequenced and their correct identification was confirmed by Sanger technology.

DISCUSSION

In this study, we successfully generated viral metagenomes from human pericardial fluids. In the positive control patient, we reconstructed almost the entire genome of HHV3 (human herpesvirus 3), which had been previously detected by molecular diagnostic analyses. HHV3 infections are usually self-limiting, although pericarditis has been previously described as a rare complication [41,42]. Serendipitously, the virome generated for the positive control also revealed the presence of 4 different HPV genotypes. To our knowledge, HPV has never been associated with pericarditis. The presence of multiple concomitant viral species/genotypes may be explained by the fact that this patient suffered a systemic lupus erythematosus and immunosuppressive medication increased permissiveness to viral infections.

Two patients and the negative control viromes displayed a high relative abundance of *Retroviridae*, which may result from the residual presence of host DNA, although VLPs were DNAse and RNAse digested to eliminate peripheral contaminations. The presence of contaminating sequences is a well-known issue in the generation of host-associated viral metagenomes. In previous viral metagenomic studies, the proportion of human-related sequences represented 24%-36% of the virome [7]. These sequences represented up to 34% of generated sequences even when stringent protocols, such as filtration followed by cesium chloride density gradient purification, were adopted for the purification of viral particles [43]. In our work, abundant contamination from residual host DNA was detected in 3 patients, which may reflect the paucity of amplifiable VLPs in these samples.

Five metagenomes had sequences that matched *Anelloviridae*, particularly torque teno viruses. Torque teno viruses were originally identified in the serum of a patient with idiopathic hepatitis [44]. Since then, they have been demonstrated to be highly ubiquitous in the human population [45–47] and have been found in sera, blood and multiple tissues, as well as in several organs and cells, including liver and peripheral blood mononuclear cells [19,48,49]. Our virome analysis showed that up to several different genotypes of TTV could be retrieved concomitantly in an individual sample. Co-infection by different TTV species or isolates is common and has been hypothesized as necessary for productive infection [49–53]. In sample P7, the number of TTV sequences was high and comparable to that of HHV sequences in the positive control. In addition, the viral community composition of P7 was dominated by torque teno viruses and other *Anelloviridae.* These two observations suggest the occurrence of viremia by torque teno virus in P7. To our knowledge, this is the first time that the presence of TTVs has been described in pericardial fluids. No torque teno virus sequences could be detected in the virome associated with the negative control pool. However, no evidence of the TTV association with idiopathic pericarditis can be provided, nor has an association with this disease been described in literature. It has previously been demonstrated that TTVs are able to perform in vitro and in vivo intragenomic rearrangement [54]. In addition, the genetic variability within and between the genomes of torque teno virus strains has been proposed as a mechanism to evade the host immune response [52,55]. Thus, the pathogenic potential of these viruses could be related to specific genotypes, intragenomic rearrangements, co-infections or the viral load. Further experiments need to be performed to evaluate these hypotheses.

Finally, 3 patients showed the presence of sequences related to bacteriophages. The presence of bacteriophages in the human body was first discovered by Felix d'Herelle almost 90 years ago [56]. Metagenomic studies have revealed the prevalence of bacteriophages in viral communities associated with different human samples such as saliva [57], sputum [43], feces [30] and oropharyngeal samples [58]. In this study, we describe the occurrence of bacteriophages in pericardial fluids. The presence of bacteriophages in a closed system such as the pericardium indicates that phages may circulate in the human body. This hypothesis is supported a previous study showing that phages circulate in the blood of patients with septicemia [59]. We detected bacteriophages infecting *Enterobacteria, Pseudomonas, Staphylococci, Streptococci* and *Burkholderia*. The bacterial hosts of these bacteriophages can be pathogenic for humans and may play a role in the etiology of the pericarditis cases. However, except for sample P4, the corresponding bacteria were not detected in the samples during previous hospital routine screenings. Sample P4 came from a sudden infant death protocol and culture of the pericardial fluid showed a polymicrobial infection (*Escherichia* and *Pseudomonas* species) as a probable consequence of post-mortem bacterial invasion. In accordance with this result, the viral metagenomic analysis showed that the viral community was dominated by bacteriophage sequences related to these bacterial hosts. We were able to reconstruct 30 contigs that matched an *Escherichia coli* phage. Viral metagenomes from patients P1 and P3 also contained bacteriophage sequences, but the corresponding bacteria were not detected in the samples by culture. Early antibiotic therapy has been proposed as one factor responsible for the negative results of blood- and joint-cultures in the case of *S. aureus* endocarditis and osteoarticular infections, respectively [60,61]. We hypothesize that the presence of bacteriophages in these samples may either reflect an earlier bacterial infection or a current one in which the amount of bacterial DNA is below the detection threshold of the diagnostic method used due to antibiotic treatment and/or clearance by the host immune system. As bacteriophages are estimated to outnumber their bacterial hosts by a ratio of approximately 10 to 1 depending to the environmental niche [62,63], the detection of bacteriophages could thus be more sensitive than detection of their bacterial hosts.

In conclusion, the data provided in this study allow assessment of the DNA virome in human pericardial fluids and provide preliminary clues on potential unsuspected or undetected pathological agents involved in pericarditis cases of unknown etiology, for which further investigations are necessary.

Figures legends

Figure 1. Proportion of known and unknown reads. The chart shows the proportion of metagenomic reads classified as “known” or “unknown,” according to a BLASTN search against the non-redundant NCBI database (E-value<1e^{-05}). For each sample, known reads (in white) and unknown reads (in grey) are reported as both a percentage and absolute number (in parentheses).

Figure 2. Viral relative abundance in each virome. The estimated relative abundance for each viral group detected in a sample is shown. The data were generated by GAAS analysis. Eukaryotic viral genotypes are grouped together at the family level. All bacteriophages are grouped together in the “Bacteriophages” category.

Figure 3. Reconstructed contig matching a human papillomavirus. Contig 37 assembled from the positive control is shown and compared to the genome organization of human papillomavirus type 50. The arrows represent open reading frames (ORFs). Homologous ORFs between the contig and human papillomavirus type 50 are shown in black.

Figure 4. Contig 129 from sample P7. The contig was assembled *de novo* from the viral metagenome associated with sample P7 and matches a torque teno virus. A grey arrow indicates an identified ORF homologous to the ORF1 of torque teno virus. A star indicates a GC-rich region (85.7%). The dotted line under the contig shows the region that was amplified by PCR and sequenced.

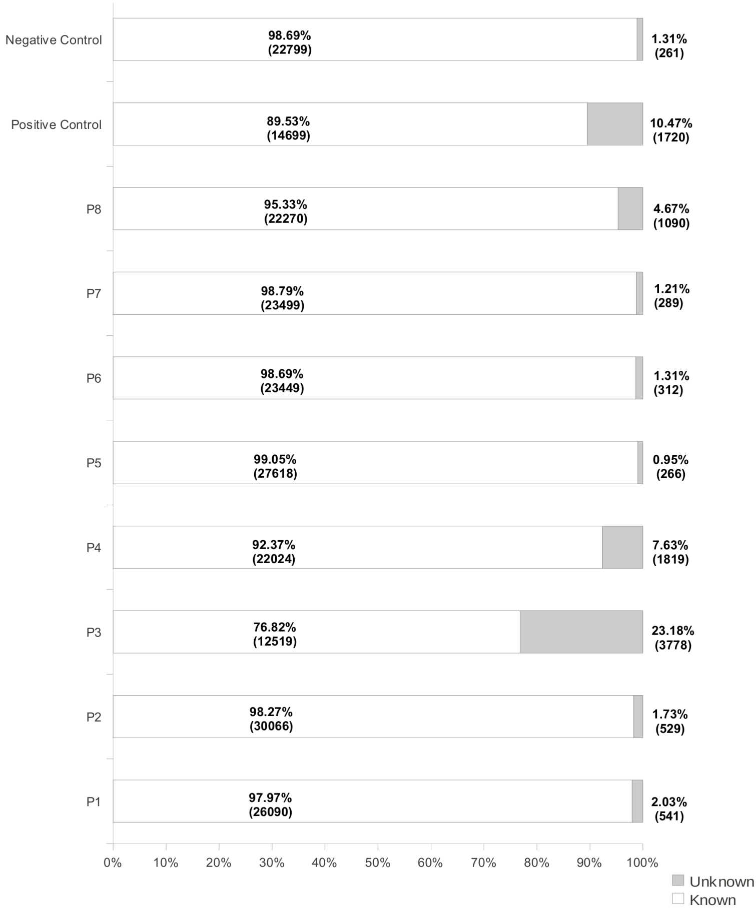

Negative Control
98.69%
(22799)
1.31%
(261)
Positive Control
89.53%
(14699)
10.47%
(1720)
P8
95.33%
(22270)
4.67%
(1090)
P7
98.79%
(23499)
1.21%
(289)
P6
98.69%
(23449)
1.31%
(312)
P5
99.05%
(27618)
0.95%
(266)
P4
92.37%
(22024)
7.63%
(1819)
P3
76.82%
(12519)
23.18%
(3778)
P2
98.27%
(30066)
1.73%
(529)
P1
97.97%
(26090)
2.03%
(541)
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Unknown
Known

Figure 1.

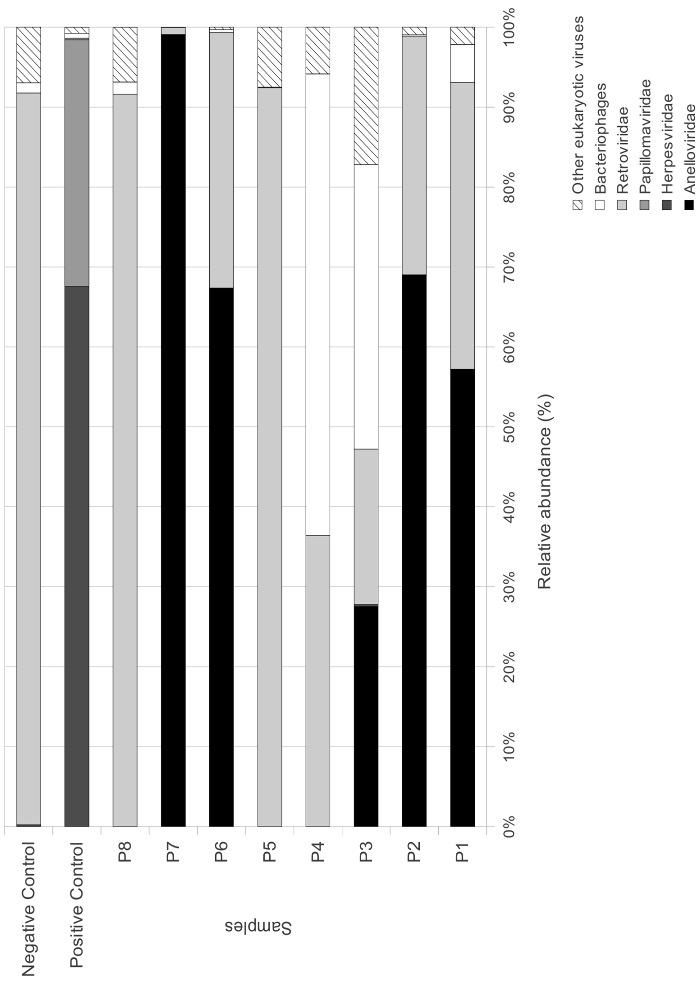

Samples
Negative Control
Positive Control
P8
P7
P6
P5
P4
P3
P2
P1
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Relative abundance (%)
Other eukaryotic viruses
Bacteriophages
Retroviridae
Papillomaviridae
Herpesviridae
Anelloviridae

Figure 2.

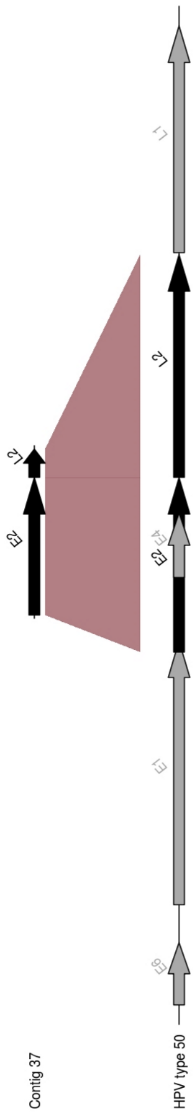


Figure 3.

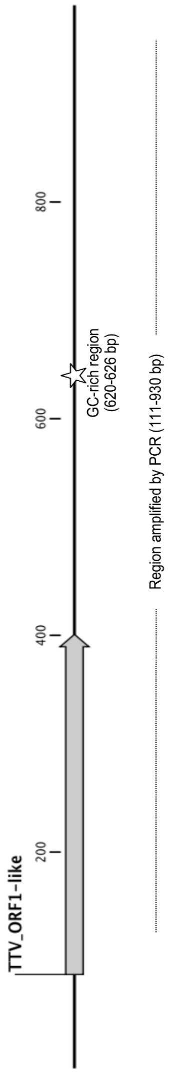

TTV_ORF1-like
200
400
600
800
GC-rich region
(620-626 bp)
Region amplified by PCR (111-930 bp)

Figure 4.

Table 1. Samples, diagnosis and metagenomic results.

Sample	Age[a]	Sex	Metavir identifier[b]	Diagnosis	Pathological agent (hospital diagnostic tests)	Most abundant viral type in viral metagenomes[c]
P1	34	M	LPC_P1	Idiopathic pericarditis	-	*Anelloviridae* (57.04%); *Retroviridae* (35.78%); Bacteriophages (4.74%)
P2	73	F	LPC_P2	Idiopathic pericarditis	-	*Anelloviridae* (68.76%); *Retroviridae* (29.68%)
P3	81	M	LPC_P3	Idiopathic pericarditis	-	Bacteriophages (33.69%); *Anelloviridae* (26.05%)
P4	6 m	M	LPC_P4	Sudden infant death	Polymicrobial infection	Bacteriophages (49.32%); *Retroviridae* (31.07%)
P5	66	M	LPC_P5	Idiopathic pericarditis	-	*Retroviridae* (90.51%)
P6	43	F	LPC_P6	Idiopathic pericarditis	-	*Anelloviridae* (60.23%); *Retroviridae* (28.58%)
P7	64	M	LPC_P7	Idiopathic pericarditis	-	*Anelloviridae* (98.59%)
P8	88	M	LPC_P8	Idiopathic pericarditis	-	*Retroviridae* (90.9%)
Positive control	19	F	LPC_PosContr	Viral pericarditis	Human herpesvirus 3	*Herpesviridae* (66.25%); *Papillomaviridae* (30.28%)
Negative control[d]	6	M	LPC_NegContr	Pericarditis of non infectious origin[d]	-	*Retroviridae* (90.85%)[d]
	87	F		Pericarditis of non infectious origin[d]	-	
	49	M		Pericarditis of non infectious origin[d]	-	
	7 m	F		Pericarditis of non infectious origin[d]	-	
	3	M		Pericarditis of non infectious origin[d]	-	

[a] In years, if not specified; in months if indicated (m).

[b] Identifier on the Metavir server (http://metavir-meb.univ-bpclermont.fr) for preprocessed viral metagenomes.

[c] The most abundant viral type in the associated viral metagenome according to the GAAS analysis.

[d] Negative control samples were pooled together and one unique viral metagenome was generated

Table 2. ***De novo*** **contigs matching** ***Papillomaviridae*** **from the viral metagenome associated with the positive control.**

Contig ID	Contig length (bp)	Number of reads[a]	Best Blast hit[b]	E-value	Percent identity	Alignment length (aa)
contig00017	650	8	major capsid protein L1 [Human papillomavirus type 50]	4.00E-106	98.39	186
contig00019	677	6	major capsid protein L1 [Human papillomavirus type 50]	5.00E-129	98.22	225
contig00027	487	7	replication protein E1 [Human papillomavirus type 50]	4.00E-065	87.12	132
contig00033	645	6	E1 protein [Human papillomavirus]	9.00E-111	96.91	162
contig00035	428	2	L2 [Human papillomavirus type 80]	3.00E-068	98.35	121
contig00037	1196	19	regulatory protein E2 [Human papillomavirus type 50]	3.00E-159	96.59	323
contig00049	787	5	Replication protein E1 [Human papillomavirus type 12]	1.00E-122	100	121
contig00062	1210	15	replication protein E1 [Human papillomavirus type 50]	1.00E-144	97.71	175
contig00080	665	4	E2 protein [Human papillomavirus]	4.00E-056	77.69	121
contig00082	771	10	minor capsid protein L2 [Human papillomavirus type 50]	7.00E-128	96.85	254
contig00089	461	4	L2 protein [Human papillomavirus]	5.00E-068	90.26	154

a Number of reads assembled in the contig

b Best BLAST hit according to a BLASTX search against the non-redundant NCBI database, $E<1e^{-05}$

Table 3. Detection of bacteriophage sequences in metagenomes P1, P3 and P4.

Sample	Bacteriophage[a]	Number of reads[b]	Number of contigs[c]
P1	*Enterobacteria* phages	19	0
P3	*Streptococcus* phages	18	0
	Staphylococcus phages	12	0
P4	*Pseudomonas* phages	170	0
	Burkholderia phages	76	0
	Enterobacteria phages	49	30

a Only the most abundant bacteriophage types (according to GAAS analysis) are reported. Bacteriophages are grouped according to the genus of their putative bacterial host.

b Number of reads matching the corresponding bacteriophage according to BLASTX against the RefSeq Viral Genomes database, E-value<1e-05.

c Number of contigs matching the corresponding bacteriophage according to BLASTX against the non-redundant NCBI database, E-value<1e-05.

BIBLIOGRAPHY

1. Imazio M, Spodick DH, Brucato A, Trinchero R, Markel G, et al. (2010) Diagnostic issues in the clinical management of pericarditis. International Journal of Clinical Practice 64: 1384–1392. Available: http://doi.wiley.com/10.1111/j.1742-1241.2009.02178.x. Accessed 31 May 2012.

2. Troughton RW, Asher CR, Klein AL (2004) Pericarditis. Lancet 363: 717–727. Available: http://www.ncbi.nlm.nih.gov/pubmed/15001332.

3. Levy P-Y, Habib G, Collart F, Lepidi H, Raoult D (2006) Etiological diagnosis of pericardial effusion. Future microbiology 1: 229–239. Available: http://www.ncbi.nlm.nih.gov/pubmed/17661668. Accessed 31 May 2012.

4. Knowlton KU, Savoia MC, Oxman MN (2000): Myocarditis and pericarditis. In: Mandell, Bennett D. Principles and Practice of Infectious Diseases. Philadelphia: Churchill-Livingstone. pp. 925–941.

5. Levy P-Y, Corey R, Berger P, Habib G, Bonnet J-L, et al. (2003) Etiologic diagnosis of 204 pericardial effusions. Medicine 82: 385–391. Available: http://www.ncbi.nlm.nih.gov/pubmed/14663288. Accessed 31 May 2012.

6. Palacios G, Druce J, Du L, Tran T, Birch C, et al. (2008) A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. New England Journal of Medicine 358: 991–998. Available: http://www.nejm.org/doi/abs/10.1056/NEJMoa073785. Accessed 13 September 2011.

7. Allander T (2005) From The Cover: Cloning of a human parvovirus by molecular screening of respiratory tract samples. Proceedings of the National Academy of Sciences 102: 12891–12896. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0504666102. Accessed 21 September 2011.

8. Greninger AL, Runckel C, Chiu CY, Haggerty T, Parsonnet J, et al. (2009) The complete genome of klassevirus - a novel picornavirus in pediatric stool. Virology Journal 6: 82. Available: http://www.virologyj.com/content/6/1/82. Accessed 30 April 2012.

9. Feng H, Shuda M, Chang Y, Moore PS (2008) Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma. Science 319: 1096–1100. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1152586. Accessed 27 April 2012.

10. Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, et al. (2012) Characterization of the Viral Microbiome in Patients with Severe Lower Respiratory Tract Infections, Using Metagenomic Sequencing. PLoS ONE 7: e30875. Available: http://dx.plos.org/10.1371/journal.pone.0030875. Accessed 27 April 2012.

11. McMullan LK, Frace M, Sammons SA, Shoemaker T, Balinandi S, et al. (2012) Using next generation sequencing to identify yellow fever virus in Uganda. Virology 422: 1–5. Available: http://www.ncbi.nlm.nih.gov/pubmed/21962764. Accessed 30 April 2012.

12. Sullivan PF, Allander T, Lysholm F, Goh S, Persson B, et al. (2011) An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. BMC Microbiology 11: 2. Available: http://www.biomedcentral.com/1471-2180/11/2. Accessed 13 September 2011.

13. Nakamura S, Yang C-S, Sakon N, Ueda M, Tougan T, et al. (2009) Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-

Throughput Sequencing Approach. PLoS ONE 4: e4219. Available: http://dx.plos.org/10.1371/journal.pone.0004219. Accessed 13 September 2011.

14. Thurber R V, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nature protocols 4: 470–483. Available: http://www.ncbi.nlm.nih.gov/pubmed/19300441. Accessed 24 May 2013.

15. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England) 22: 1658–1659. Available: http://www.ncbi.nlm.nih.gov/pubmed/16731699. Accessed 9 May 2012.

16. Sun S, Chen J, Li W, Altintas I, Lin A, et al. (2010) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Research 39: D546–D551.

17. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, et al. (2011) Metavir: a web server dedicated to virome analysis. Bioinformatics 27: 3074–3075. Available: http://bioinformatics.oxfordjournals.org/content/27/21/3074.abstract.

18. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, et al. (2009) The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. PLoS Computational Biology 5: e1000593. Available: http://dx.plos.org/10.1371/journal.pcbi.1000593. Accessed 13 September 2011.

19. Peng YH, Nishizawa T, Takahashi M, Ishikawa T, Yoshikawa A, et al. (2002) Analysis of the entire genomes of thirteen TT virus variants classifiable into the fourth and fifth genetic groups, isolated from viremic infants. Archives of virology 147: 21–41. Available: http://www.ncbi.nlm.nih.gov/pubmed/11855633. Accessed 7 May 2012.

20. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods in molecular biology (Clifton, NJ) 132: 365–386. Available: http://www.ncbi.nlm.nih.gov/pubmed/10547847. Accessed 7 May 2012.

21. Pantůcek R, Doskar J, Růzicková V, Kaspárek P, Orácová E, et al. (2004) Identification of bacteriophage types and their carriage in Staphylococcus aureus. Archives of virology 149: 1689–1703. Available: http://www.ncbi.nlm.nih.gov/pubmed/15593413. Accessed 9 May 2012.

22. Balding C, Bromley SA, Pickup RW, Saunders JR (2005) Diversity of phage integrases in Enterobacteriaceae: development of markers for environmental analysis of temperate phages. Environmental microbiology 7: 1558–1567. Available: http://www.ncbi.nlm.nih.gov/pubmed/16156729. Accessed 7 May 2012.

23. Dwivedi B, Schmieder R, Goldsmith DB, Edwards RA, Breitbart M (2012) PhiSiGns: an online tool to identify signature genes in phages and design PCR primers for examining phage diversity. BMC bioinformatics 13: 37. Available: http://www.ncbi.nlm.nih.gov/pubmed/22385976. Accessed 7 May 2012.

24. Favre M, Obalek S, Jablonska S, Orth G (1989) Human papillomavirus (HPV) type 50, a type associated with epidermodysplasia verruciformis (EV) and only weakly related to other EV-specific HPVs. Journal of Virology 63: 4910.

25. Li J, Cai H, Xu Z, Wang Q, Hang D, et al. (2012) Nine complete genome sequences of cutaneous human papillomavirus genotypes isolated from healthy skin of individuals living in rural He Nan province, China. Journal of virology 86: 11936.

26. Delius H HB (1994) Primer-directed sequencing of human papillomavirus types. Curr Top Microbiol Immunol 186: 13–31.

27. Delius H, Saegling B, Bergmann K, Shamanin V, De Villiers EM (1998) The genomes of three of four novel HPV types, defined by differences of their L1 genes, show high conservation of the E7 gene and the URR. Virology 240: 359–365. Available: http://www.ncbi.nlm.nih.gov/pubmed/9454709.

28. Hallett RL, Clewley JP, Bobet F, McKiernan PJ, Teo CG (2000) Characterization of a highly divergent TT virus genome. The Journal of general virology 81: 2273–2279.

29. Heller F, Zachoval R, Koelzer a, Nitschko H, Froesner GG (2001) Isolate KAV: a new genotype of the TT-virus family. Biochemical and biophysical research communications 289: 937–941. Available: http://www.ncbi.nlm.nih.gov/pubmed/11741280. Accessed 9 April 2013.

30. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334–338. Available: http://www.nature.com/doifinder/10.1038/nature09199. Accessed 26 September 2011.

31. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic Analyses of an Uncultured Viral Community from Human Feces. Journal of Bacteriology 185: 6220–6223. Available: http://jb.asm.org/cgi/doi/10.1128/JB.185.20.6220-6223.2003. Accessed 20 September 2011.

32. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, et al. (2008) Viral diversity and dynamics in an infant gut. Research in Microbiology 159: 367–373. Available: http://linkinghub.elsevier.com/retrieve/pii/S0923250808000648. Accessed 13 September 2011.

33. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. (2011) The human gut virome: Inter-individual variation and dynamic response to diet. Genome Research. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.122705.111. Accessed 28 September 2011.

34. Kim M-S, Park E-J, Roh SW, Bae J-W (2011) Diversity and Abundance of Single-Stranded DNA Viruses in Human Feces. Applied and Environmental Microbiology 77: 8062–8070. Available: http://aem.asm.org/cgi/doi/10.1128/AEM.06331-11. Accessed 3 May 2012.

35. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, et al. (2008) Metagenomic Analysis of Human Diarrhea: Viral Detection and Discovery. PLoS Pathogens 4: e1000011. Available: http://dx.plos.org/10.1371/journal.ppat.1000011. Accessed 13 September 2011.

36. Holtz LR, Finkbeiner SR, Kirkwood CD, Wang D (2008) Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. Virology Journal 5: 159. Available: http://www.virologyj.com/content/5/1/159. Accessed 27 April 2012.

37. Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, et al. (2009) Metagenomic Analyses of Viruses in Stool Samples from Children with Acute Flaccid Paralysis. Journal of Virology 83: 4642–4651. Available: http://jvi.asm.org/cgi/doi/10.1128/JVI.02301-08. Accessed 13 September 2011.

38. Breitbart M, Rohwer F (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. BioTechniques 39: 729–736. Available: http://www.ncbi.nlm.nih.gov/pubmed/16312220. Accessed 28 September 2011.

39. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, et al. (2008) Newly Discovered Ebola Virus Associated with Hemorrhagic Fever Outbreak in Uganda. PLoS

Pathogens 4: e1000212. Available: http://dx.plos.org/10.1371/journal.ppat.1000212. Accessed 27 April 2012.

40. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, et al. (2009) Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever–Associated Arenavirus from Southern Africa. PLoS Pathogens 5: e1000455.

41. Kao K-L, Yeh S-J, Chen C-C (2010) Myopericarditis associated with varicella zoster virus infection. Pediatric cardiology 31: 703–706.

42. Seddon DJ (1986) Pericarditis with pericardial effusion complicating chickenpox. Postgraduate medical journal 62: 1133–1134.

43. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. PLoS ONE 4: e7370.

44. Nishizawa T, Okamoto H, Konishi K, Yoshizawa H, Miyakawa Y, et al. (1997) A novel {DNA} virus ({TTV)} associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. Biochem Biophys Res Commun 241: 92–97. doi:10.1006/bbrc.1997.7765.

45. Hsieh SY, Wu YH, Ho YP, Tsao KC, Yeh CT, et al. (1999) High prevalence of {TT} virus infection in healthy children and adults and in patients with liver disease in Taiwan. J Clin Microbiol 37: 1829–1831.

46. Vasilyev E V, Trofimov DY, Tonevitsky AG, Ilinsky V V, Korostin DO, et al. (2009) Torque Teno Virus ({TTV)} distribution in healthy Russian population. 6: 134. Available: http://www.virologyj.com/content/6/1/134. Accessed 16 January 2013.

47. Abe K, Inami T, Asano K, Miyoshi C, Masaki N, et al. (1999) {TT} virus infection is widespread in the general populations from different geographic regions. J Clin Microbiol 37: 2703–2705.

48. Zhong S, Yeo W, Tang M, Liu C, Lin X, et al. (2002) Frequent detection of the replicative form of {TT} virus {DNA} in peripheral blood mononuclear cells and bone marrow cells in cancer patients. J Med Virol 66: 428–434.

49. Biagini P, Gallian P, Attoui H, Cantaloube JF, De Micco P, et al. (1999) Determination and phylogenetic analysis of partial sequences from TT virus isolates. The Journal of general virology 80 (Pt 2): 419–424. Available: http://www.ncbi.nlm.nih.gov/pubmed/10073702.

50. Ball JK, Curran R, Berridge S, Grabowska a M, Jameson CL, et al. (1999) TT virus sequence heterogeneity in vivo: evidence for co-infection with multiple genetic types. The Journal of general virology 80 (Pt 7): 1759–1768.

51. Irving WL, Ball JK, Berridge S, Curran R, Grabowska a M, et al. (1999) TT virus infection in patients with hepatitis C: frequency, persistence, and sequence heterogeneity. The Journal of infectious diseases 180: 27–34. Available: http://www.ncbi.nlm.nih.gov/pubmed/10353857.

52. Khudyakov YE, Cong ME, Nichols B, Reed D, Dou XG, et al. (2000) Sequence heterogeneity of {TT} virus and closely related viruses. J Virol 74: 2990–3000.

53. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. PLoS ONE 7: e30087. Available: http://dx.plos.org/10.1371/journal.pone.0030087. Accessed 24 April 2012.

54. Leppik L, Gunst K, Lehtinen M, Dillner J, Streker K, et al. (2007) In vivo and in vitro intragenomic rearrangement of {TT} viruses. J Virol 81: 9346–9356. doi:10.1128/JVI.00781-07.

55. Jelcic I, Hotz-wagenblatt A, Hunziker A, Hausen H, Villiers E De (2004) Isolation of Multiple TT Virus Genotypes from Spleen Biopsy Tissue from a Hodgkin ' s Disease Patient : Genome Reorganization and Diversity in the Hypervariable Region Isolation of Multiple TT Virus Genotypes from Spleen Biopsy Tissue from a Hodgkin ' s D. doi:10.1128/JVI.78.14.7498.

56. D'Herelle F (1922) The bacteriophage : its rôle in immunity. Baltimore [Md.]: Williams & Wilkins.

57. Pride DT, Salzman J, Haynes M, Rohwer F, Davis-Long C, et al. (2011) Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. The ISME Journal 6: 915–926.

58. Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, et al. (2010) Colloquium Paper: Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proceedings of the National Academy of Sciences 108: 4547–4553. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1000089107. Accessed 13 September 2011.

59. Gaidelyte A,Vaara M, Bamford DH (2007) Bacteria, phages and septicemia. PLoS ONE 2: e1145.59.

60. Levy P-Y, Fournier PE, Fenollar F, Raoult D (2013) Systematic PCR detection in culture negative osteoarticular infections. Am J Med. In press.

61. Fournier PE,Thuny F, Richet H, Lepidi H, Casalta J-P, Arzouni J-P, Maurin M, Célard M, Mainardi J-L, Caus T, Collart F, Habib G, Raoult D (2010) Comprehensive diagnostic strategy for blood culture-negative endocarditis: a prospective study of 819 new cases. Clin Infect Dis 51: 131-140.

62. Wommack KE, Ravel J, Hill RT, Colwell RR (1999) Hybridization Analysis of Chesapeake Bay Virioplankton. Appl Envir Microbiol 65: 241–250.

63. Bergh O, Børsheim KY, Bratbak G, Heldal M (1989) High abundance of viruses found in aquatic environments. Nature 340: 467–468.

Supplementary material.

Supplementary Table S1. Primers used in this study. Different PCR primers were used to confirm the presence of a selected viral species detected *in silico*. For each primer pair, the organism and sequence targeted, primer names, expected amplicon size, sequences and method used to design the primers are listed.

Organism	Targeted sequence	Primer name	Expected product size (bp)	Primer sense (5' → 3')	Primer antisense (5' → 3')	Primer design
TTV	UTR-region	Internal primers NG472, NG352;	91	GCGTCCCGWGGGCG GGTGCCG [W = A or T]	GAGCCTTGCCCATRGC CCGGCCAG [R = A or G]	Literature [18]
		External primers NG473, NG351	71	CGGGTGCCGDAGGTG AGTTTACAC [D = G, A or T]	CCCATRGCCCGGCCAG TCCCGAGC	
TTV	TTV-matching contig 127 (sample P7)	127L, 127R	819	CCCATGGAGCAGATG TCTTT	CAGGGCTAGACCAGCT CAAC	Primer3 [19]
TTV	TTV-matching contig 129 (sample P7)	129L, 129R	820	CCTGCTGCTCTGGTA TCCAT	TTACACACGACTGGGGC ATA	Primer3 [19]
TTV	TTV-matching contig 164 (sample P7)	164L, 164R	783	AGCGGAAGGAAGTCA CAAGA	GGACCCCCTCCTGTCAT AAT	Primer3 [19]
Staphylococcus phage 47, 42E, phi 12, tp310-2	tail fiber	SGA1, SGA2	744	TATCAGGCGAGAATTA AGGG	CTTTGACATGACATCCG CTTGAC	Literature [20]
Staphylococcus phage 96, phi ETA3	hypothetical tail protein	SGB1, SGB2	405	ACTTATCCAGGTGGY GTTATTG	TGTATTTAATTTCGCCG TTAGTG	Literature [20]
Enterobacteria phage P1	tyrosine recombinases	EphP1, EphP2	447	TGCTTATAACACCCTG TTACGTAT	CAGCCACCAGCTTGCAT GATC	Literature [21]
Enterobacteria phage lambda	minor capsid protein	EphL1, EphL2	378	TCATCGCCCGTGTGC GTGAC	GCCTGCGAAGCAGTGG CTGA	Primer-BLAST (www.ncbi.nlm.nih.gov/tools/primer-blast/)
Stx2 converting phage II, *Enterobacteria* phage BP-4795	NinG protein	RF41, RF42	453	CAAATCAGTGGTGGT GCT	TGGCGGTTATGGTTCG	PhiSigns [22]
Pseudomonas phage F8,LMA2, LBL3	phage protein	Pph1, Pph2	458	GTTGCTGCGGATTTG A	GTCTGCGGCGTTCTGG	PhiSigns [22]
Burkholderia ambifaria phage BcepF1 *Streptococcus* prophage EJ-1	hypothetical protein BcepF1.101	Bph1, Bph2	173	CGTAAATGCCTTGACC GTTT	CATCCAGTGGGTGAACA CAG	Primer3 [19]
Streptococcus prophage EJ-1	minor head protein	StrePh1, StrePh2	954	CTAAAACGCATTACGG GGCG	GCGGACTCTGTCCTTTT CCA	Primer-BLAST (www.ncbi.nlm.nih.gov/tools/primer-blast/)

Supplementary Table S2. High-throughput sequencing output. For each sample, the total number of reads generated, the number of reads remaining after duplicate elimination (preprocessing) and the average length of the preprocessed reads are reported.

Sample	Total number of reads	Number of reads after preprocessing	Average read length after preprocessing (bp)
P1	41711	26631	334.82
P2	34452	30595	311.79
P3	33909	16297	359.62
P4	35817	23843	341.57
P5	29614	27884	287.77
P6	26369	23761	246.22
P7	26220	23788	260.05
P8	28691	23360	280.66
Positive Control	29812	16419	361.8
Negative Control	26562	23060	336.84

Supplementary Table S3. Metagenomes assembly. For each sample, the total number of contigs (“Contigs”), the number of contigs spanning more than 1,500 bp (“Large contigs”) and the percentage of reads assembled into contigs (“Assembled reads”) are reported.

Sample	Contigs	Large contigs	Assembled reads (%)
P1	260	42	60.94
P2	415	95	43.9
P3	83	13	76.66
P4	282	19	44.71
P5	32	2	2.62
P6	226	8	16.93
P7	97	3	12.43
P8	233	17	35.11
Positive control	79	16	67.66
Negative control	157	18	22.42

Supplementary Table S4. Reconstruction of human papillomavirus genomes by mapping of the positive control virome. For each reference genome, we reported the length of the consensus sequence reconstructed by mapping, the number of reads mapped, the average depth coverage and the proportion of the reference genome that was reconstructed.

Reference genome	Consensus length (bp)	Number of reads mapped	Average depth coverage	% of reference covered
Human papillomavirus type 12	4413	20	0.93	57.51
Human papillomavirus type 50	6721	69	3.5	93.55
Human papillomavirus type 80	4975	23	1.04	66.98
Human papillomavirus isolate 915 F 06 002 KN1	5737	32	1.57	78

Supplementary Table S5. *Anelloviridae* detected in the viromes. For each virome, the absolute number of reads matching each identified species of *Anelloviridae* (BLASTX search against the non-redundant NCBI database, E-value<1e-05) is listed. Species are grouped according to the genus to which they belong.

Genus	Species	P1	P2	P3	P4	P5	P6	P7	P8	Positive Control	Negative Control
Alfatorquevirus	Torque teno virus 1	0	**4**	0	0	0	0	0	0	0	0
	Torque teno virus 10	0	0	0	0	0	0	**43**	0	0	0
	Torque teno virus 16	0	**7**	0	0	0	0	0	0	0	0
	Torque teno virus 19	0	**1**	0	0	0	0	0	0	0	0
	Torque teno virus 21	0	**13**	0	0	0	0	0	0	0	0
	Torque teno virus 23	0	**3**	0	0	0	**1**	0	0	0	0
	Torque teno virus 24	0	**1**	0	0	0	**12**	0	0	0	0
	Torque teno virus 27	0	0	0	0	0	0	**2**	0	0	0
	Torque teno virus 28	0	0	0	0	0	0	**1**	0	0	0
	Torque teno virus 29	0	0	0	0	0	0	**366**	0	0	0
Betatorquevirus	Torque teno mini virus 1	0	0	0	0	0	0	**1**	0	0	0
	Torque teno mini virus 2	0	**2**	0	0	0	0	0	0	0	0
	Torque teno mini virus 3	0	0	0	0	0	0	**1**	0	0	0
Gammatorquevirus	Torque teno midi virus 2	**1**	0	0	0	0	0	0	0	0	0
unclassified *Anelloviridae*	SEN virus	0	**2**	0	0	0	**2**	0	0	0	0
	Small anellovirus	**1**	0	0	0	0	0	0	0	0	0
	Torque teno midi virus	**18**	**4**	0	0	0	0	0	0	0	0
	Torque teno virus	0	**71**	0	0	0	**41**	**1214**	0	0	0
	TTV-like mini virus	0	0	**3**	0	0	0	0	0	0	0

Supplementary Table S6. Reconstructed contigs matching *Anelloviridae*. For each contig, we list the sample from which it was assembled, the length, the number of reads assembled in the contig and the GC content. The best BLAST hit (BLASTX against the non-redundant NCBI database, E-value<1e-05) is also shown, as well as the hit alignment parameters: E-value, percentage of identity, and alignment length.

Sample	Contig ID	Contig length (bp)	Number of reads	GC content (%)	Number of ORFs	Best BLAST hit	E-value	Percentage identity	Alignment length (aa)
P1	contig011	482	5	36.1	1	hypothetical protein ORF1-like [Torque teno midi virus]	3.00E-033	62.4	125
	contig088	455	7	38.9	1	hypothetical protein ORF1-like [Torque teno midi virus]	3.00E-024	39.62	159
	contig175	517	3	47	1	hypothetical protein ORF2-like [Torque teno midi virus]	5.00E-023	69.86	73
P2	contig198	3532	96	48.9	2	ORF1 [Torque teno virus]	0	69.46	681
P6	contig002	860	13	57.1	1	unnamed protein product ORF2-like [Torque teno virus 24]	4.00E-019	60	80
	contig008	1011	17	49.3	3	ORF1 [Torque teno virus]	2.00E-079	63.25	117
	contig188	1060	18	45.9	2	ORF1 [Torque teno virus]	2.00E-171	91.03	234
P7	contig001	526	4	51	1	hypothetical protein TTV10_gp4 ORF1-like [Torque teno virus 10]	6.00E-052	98.67	75
	contig064	2668	1358	56.5	4	ORF1 [Torque teno virus]	6.00E-142	87.4	262
	contig127	3572	276	52.2	3	ORF1 [Torque teno virus]	0	89.52	668
	contig129	981	739	49.4	1	ORF1 [Torque teno virus]	8.00E-176	88.27	324

Supplementary Table S7. Standard PCRs performed to confirm the presence of the bacteriophages detected *in silico.*

The sample, primers used, bacteriophage targeted and results of the amplification are listed.

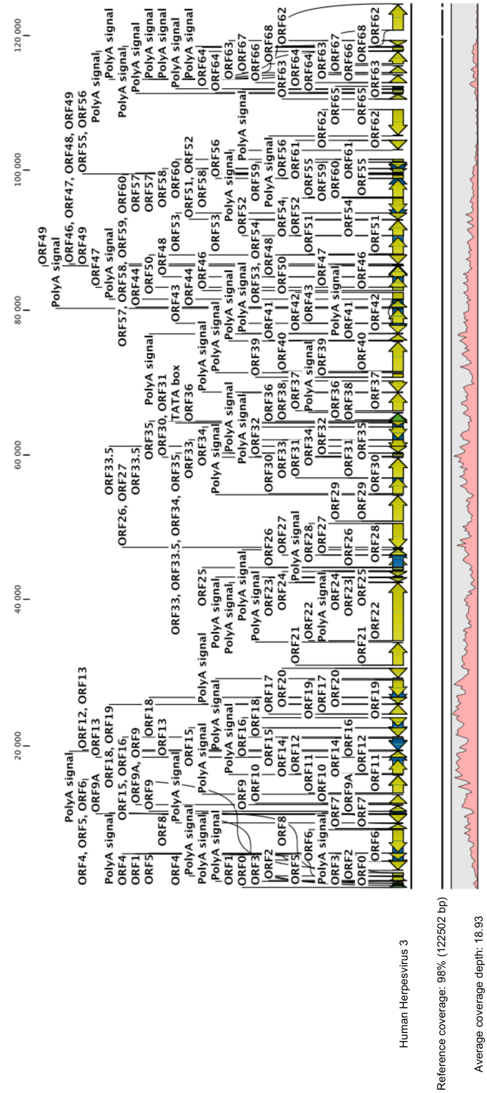
Sample	Primers	Targeted phage	PCR amplification
P1	EphL1, EphL2	*Enterobacteria* phage Lambda	Negative
P3	SGA1, SGA2	*Staphylococcus* phages 3A-like	Positive, specific
P3	SGB1, SGB2	*Staphylococcus* phages 11A-like	Positive, specific
P3	StrePh1, StrePh2	*Streptococcus* prophage EJ-1	Aspecific amplificates
P4	RF41, RF42	Stx converting phage II, *Enterobacteria* phage BP-4795	Aspecific amplificates
P4	Pph1, Pph2	*Pseudomonas* phage F8, LMA2, LBL3	Aspecific amplificates
P4	Bph1, Bph2	*Burkholderia amphibaria* phage BcepF1	Aspecific amplificates
P4	EphP1, Eph2	*Enterobacteria* phage P1	Aspecific amplificates

Supplementary Figures legends

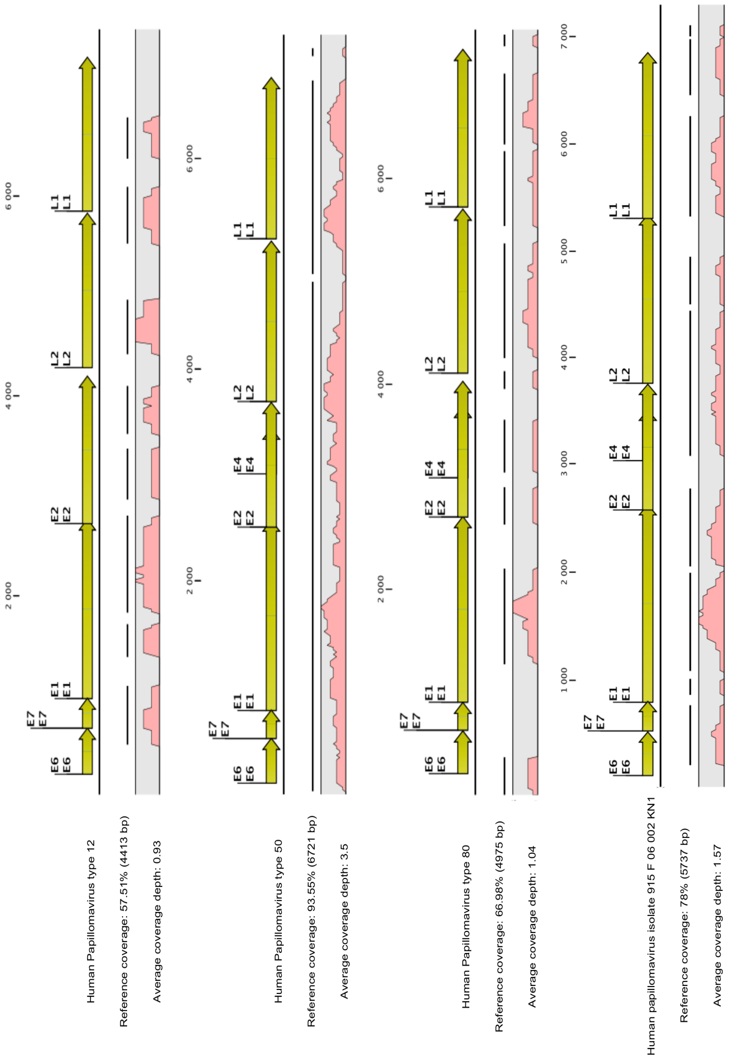
Supplementary Figure S1. Reconstruction of the human herpesvirus 3 genome from the positive control virome. The HHV3 reference genome was reconstructed by mapping the metagenomic reads generated from the positive control sample. Open Reading Frames of the reference genome (yellow arrows), reference coverage (black line) and coverage depth (pink shadow) are shown along the genome.

Supplementary Figure S2. Reconstruction of the human papillomavirus genomes detected in the positive control virome. The reference genomes of human papillomavirus type 12, type 50 and type 80 as well as that of human papillomavirus isolate 915 F 06 002 KN1 were reconstructed by mapping the metagenomic reads generated from the positive control sample. Open Reading Frames of the reference genome (yellow arrows), reference coverage (black line) and coverage depth (pink shadow) are shown.

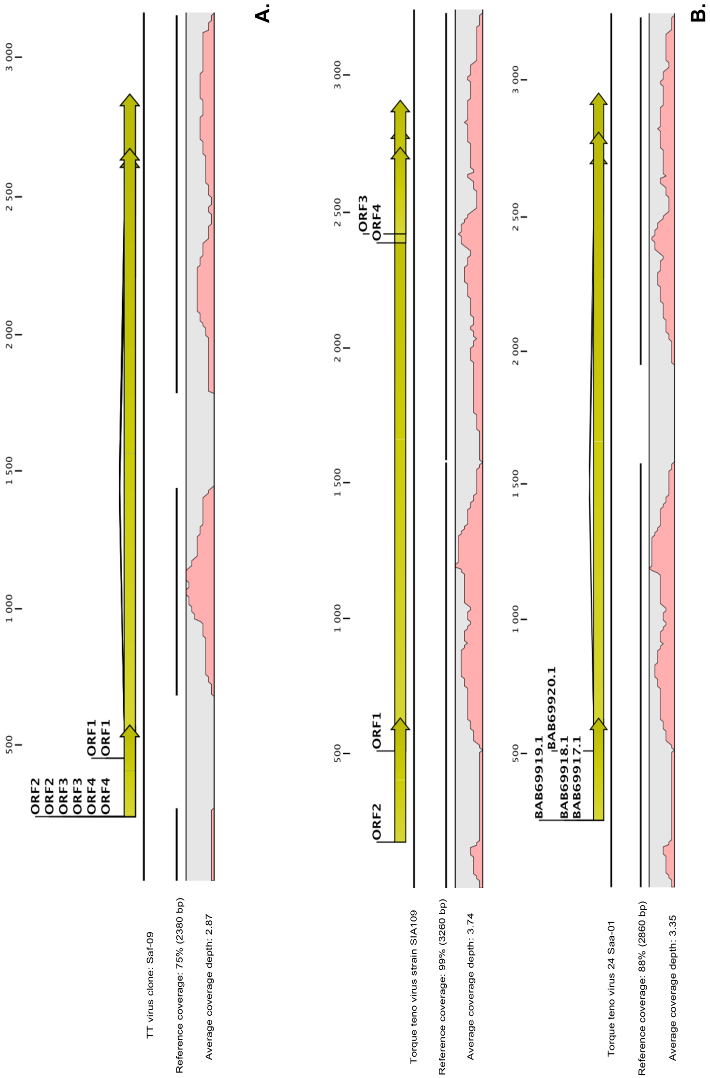
Supplementary Figure S3. Reconstruction of the torque teno virus genomes detected in patients' viromes. The reference genome of torque teno virus clone Saf-09 was reconstructed by mapping from sample P2 (A); the reference genomes of torque teno virus strain SIA 09 and torque teno virus 24 Saa-01 were reconstructed by mapping from sample P6 (B); the reference genomes of torque teno virus isolate TTVyon-LC011, torque teno virus 29, micro torque teno virus, isolate microTTV-HD14.2, torque teno virus TTV-HD14h and torque teno virus isolate JT41F were reconstructed by mapping from sample P7 (C). The open reading frames of the reference genome (yellow arrows), reference coverage (black line) and coverage depth (pink shadow) are shown.



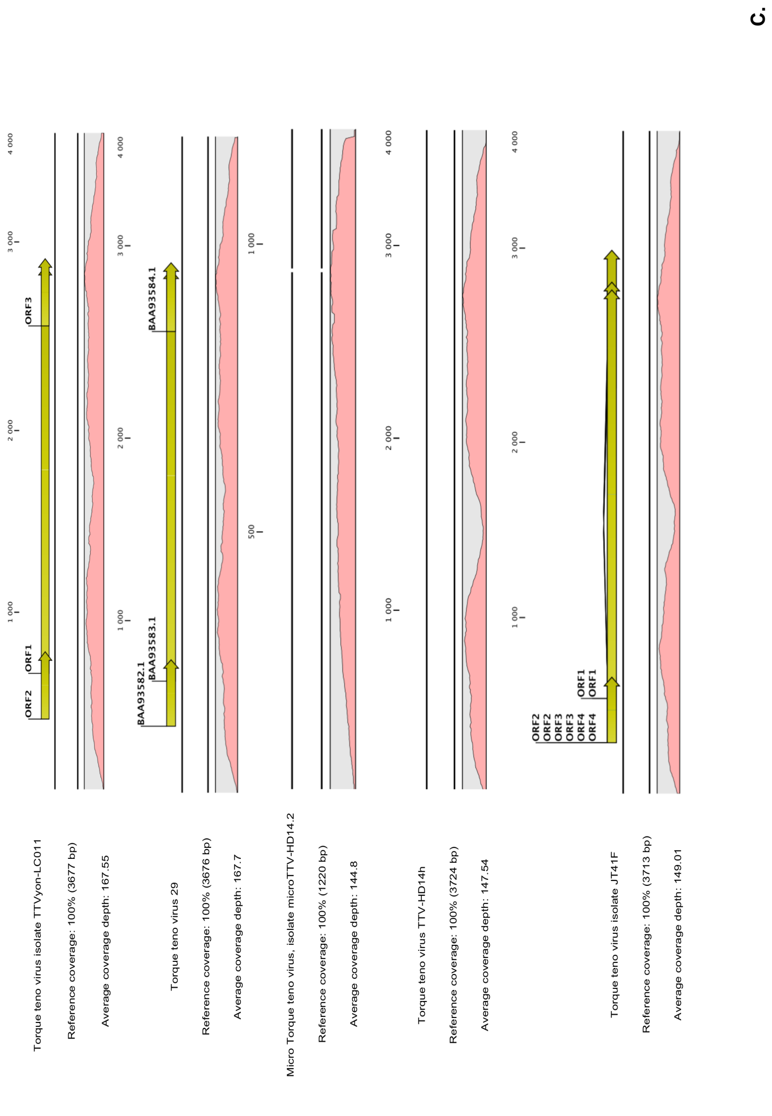
Supplementary Figure S1.



Supplementary Figure S2.



Supplementary Figure S3 A. and B.



Supplementary Figure S3 C.

4.3 Article 4. Bacteriophages and diffusion of genes encoding antimicrobial resistance in cystic fibrosis sputum microbiota

Bacteriophages and diffusion of genes encoding antimicrobial resistance in cystic fibrosis sputum microbiota.

Fancello Laura[1], Desnues Christelle[1], Raoult Didier[1], Rolain Jean Marc[1*].

Published in Journal of Antimicrobial Chemotherapy. 2011 Nov; 66(11):2448-54.

[1] URMITE UMR CNRS-IRD 6236, IFR48, Faculté de Médecine et de Pharmacie, Université de la Méditerranée, Marseille, France.

[*] Corresponding author. Email: jean-marc.rolain@univmed.fr

Preamble to article 4

Cystic fibrosis (CF) is a fatal genetic disease, which causes impaired mucociliary clearance and creates hypoxic microenvironments for bacterial proliferation [102]. It is characterized by chronic and persistent respiratory bacterial infections that are responsible for decreased lung functionality. Moreover, the frequent antibiotic treatments during bacterial infections lead to the selection of multidrug-resistant bacteria [103]. Several studies have characterized bacterial communities of CF lungs. However, little is known about the associated viral communities. A metagenomic study by Willner et al. revealed that viral communities associated with CF lungs prevalently consist of bacteriophages and these bacteriophages, regardless of their taxonomy, share a common core of metabolic functions in all CF patients studied [60]. This core of functions was significantly different from that of bacteriophages associated with the lung of individuals not affected by CF. Moreover, among the most represented functions, the researchers found those connected to bacterial virulence, including antibiotic resistance. Today, it is recognized that bacteriophages are an important reservoir of genes encoding metabolic functions which contribute to the adaptation of bacterial communities to the environment [104, 105]. Antibiotic resistance genes have been demonstrated to be carried by bacteriophages, especially in the environment, and were shown to be transmitted to pathogenic bacterial clinical strains [106, 107, 108, 109]. In addition, recent studies revealed that some antibiotics, frequently used in CF treatments, can mobilize bacteriophages, enhancing lateral gene transfer and thus the spread of antibiotic resistance genes [110, 111, 112].

In this work, we analyzed sequences related to the antibiotic resistance genes retrieved in the bacteriophages populating the airways of CF patients. We used the subset of sequences representing the core metabolism common and unique to CF patients, described by Willner et al. in a viral metagenomic study [84]. Overall, we found a significant difference between the number of sequences related to antibiotic and toxic compound resistance genes in CF viromes compared to non-CF viromes. Antibiotic resistance genes associated with CF viromes were analyzed and confidently annotated as efflux pump genes, fluoroquinolone resistance genes and beta-lactamase genes. The phylogeny of these sequences was also reconstructed revealing different origins. Beta-lactamases were observed to originate from different classes of Bacteroidetes (Flavobacteria, Cytophagia, Bacteroidia and Sphingobacteria) and, in one case, from Cyanobacteria; fluoroquinolones originated from Sphingobacteria and efflux pumps from Flavobacteria, Cytophagia, Bacteroidia and Sphingobacteria.

This work highlights the elevated number of genes involved in antibiotic and toxic compound resistance carried by bacteriophages and suggests the potential critical role of these bacteriophages in influencing the fitness and adaptation of bacteria to CF airways, especially the evolution of antibiotic resistant bacterial strains. Most human-associated viromes that have been studied up to now are prevalently composed of bacteriophages and have been mainly described taxonomically. As shown by this and other studies [59, 84] functional analyses are highly useful for elucidating the role of these bacteriophages in the host-associated environment.

J Antimicrob Chemother 2011; **66**: 2448–2454
doi:10.1093/jac/dkr315 Advance Access publication 4 August 2011

Journal of Antimicrobial Chemotherapy

Bacteriophages and diffusion of genes encoding antimicrobial resistance in cystic fibrosis sputum microbiota

Laura Fancello, Christelle Desnues, Didier Raoult and Jean Marc Rolain*

URMITE UMR CNRS-IRD 6236, IFR48, Faculté de Médecine et de Pharmacie, Université de la Méditerranée, Marseille, France

*Corresponding author. Tel: +33-4-91-32-43-75; Fax: +33-4-91-38-77-72; E-mail: jean-marc.rolain@univmed.fr

Received 1 July 2011; accepted 6 July 2011

Objectives: The cystic fibrosis (CF) airway is now considered the site of a complex microbiota, where cross-talking between microbes and lateral gene transfer are believed to contribute to the adaptation of bacteria to this specific environment and to the emergence of multidrug-resistant bacteria. The objective of this study was to retrieve and analyse specific sequences associated with antimicrobial resistance from the CF viromes database.

Methods: Specific sequences from CF metagenomic studies related to the 'antibiotic and toxic compound resistance' dataset were retrieved from the MG-RAST web site, assembled and functionally annotated for identification of the genes. Phylogenetic trees were constructed using a minimum parsimony starting tree topology search strategy.

Results: Overall, we found 1031 short sequences in the CF virome putatively encoding resistance to antimicrobials versus only 3 reads in the non-CF virome dataset (P=0.001). Among them, we could confidently identify 66 efflux pump genes, 15 fluoroquinolone resistance genes and 9 β-lactamase genes. Evolutionary relatedness determined using phylogenetic information demonstrates the different origins of these genes among the CF microbiota. Interestingly, among annotated sequences within CF viromes, we also found matching 16S rDNA sequences from *Escherichia*, Cyanobacteria and Bacteroidetes.

Conclusions: Our results suggest that phages in the CF sputum microbiota represent a reservoir of mobilizable genes associated with antimicrobial resistance that may spread in this specific niche. This phenomenon could explain the fantastic adaptation of CF strains to their niche and may represent a new potential therapeutic target to prevent the emergence of multidrug-resistant bacteria, which are responsible for most of the deaths in CF.

Keywords: mobilome, multidrug resistance, microbiota, transduction, phage induction

Introduction

Cystic fibrosis (CF) is an autosomal recessive disease caused by a mutation in the gene of the CF transmembrane conductance regulator and is characterized by recurrent respiratory infections.[1] Repeated bronchopulmonary infections are responsible for a progressive decrease in lung function, and the frequent use of antibiotics during acute exacerbations leads to the emergence and selection of multidrug-resistant (MDR) bacterial species.[2] Recent advances in molecular biology, especially full-genome sequence analysis, have shown that such MDR bacteria probably evolved during chronic colonization of the lungs in CF patients due to both intensive and recurrent antibiotic treatments as well as the vertical and horizontal transmission of antibiotic resistance determinants.[3–8] Prophages that play an important role in bacterial genomic evolution and diversity via horizontal gene transfer (HGT) were also found in these sequenced strains, including *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Burkholderia cenocepacia*.[3,4,6,9,10] Transduction is known to represent one of the most powerful phenomena for the dissemination of genes that allow bacteria to become more pathogenic and resistant to antimicrobials.[11,12] Interestingly, recent studies demonstrate that some antibiotics, especially those frequently used in CF patients (e.g. ciprofloxacin, tobramycin, co-trimoxazole and imipenem), can enhance phage mobility,[3,13,14] suggesting that frequent exchanges of genetic material and the spread of virulence and/or antibiotic resistance genes occurred among bacterial species in these communities (called microbiota). More recently, the metagenomic analysis of DNA viral communities in the respiratory tract of CF individuals (CF virome) as compared with non-CF individuals (non-CF virome) has shown a higher abundance of phage communities in CF patients, which was associated with airway pathology and specific metabolic profiles.[15] The core metabolism of CF-associated phages was further explored using a dataset containing sequences both common and specific to the CF viromes,

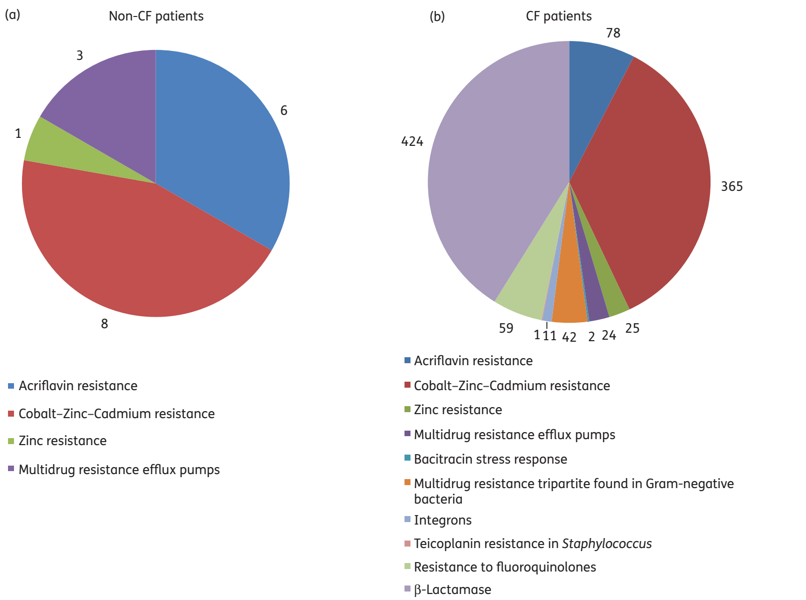


Figure 1. List and number of sequences encoding resistance to antimicrobials and toxic compounds in the virome of (a) non-CF and (b) CF patients. This figure appears in colour in the online version of *JAC* and in black and white in the print version of *JAC*.

Table 1. Number of contigs containing resistance genes for each CF patient

	Number of contigs					
Resistance gene	CF6	CF7	CF8	CF9	CF10	total
Multidrug efflux pumps	11	10	20	13	12	66
Resistance to fluoroquinolones	5	3	2	5	0	15
β-Lactamases	4	0	4	1	0	9
Arsenical resistance operon repressor	0	0	0	0	0	0

All contigs longer than 400 bp were considered. Contig annotation was performed by a blastx search (e-value < 1e-05) against the SEED subsystems database.

Table 2. Number of large contigs (>1500 bp) containing resistance genes for each CF patient

	Number of contigs					
Resistance gene	CF6	CF7	CF8	CF9	CF10	total
Multidrug efflux pumps	4	1	3	9	9	26
Resistance to fluoroquinolones	0	0	0	2	0	2
β-Lactamases	0	0	0	0	0	0
Arsenical resistance operon repressor	1	0	0	0	0	1

Only contigs longer than 1500 bp were considered. Contig annotation was performed by a blastx search (e-value < 1e-05) against the SEED subsystems database.

which included a unique set of 16059 sequences specific to the CF virome as compared with the non-CF virome.[16] These specific sequences are freely available in the MG-RAST web site, and the objective of our study was to retrieve and analyse specific sequences associated with antimicrobial resistance from the CF virome dataset.

Materials and methods

Dataset sequences

The viromes analysed were those recently published by Willner *et al.*,[15] representing the DNA viral community populations of the respiratory tract of five individuals with CF. The viral metagenomes were generated

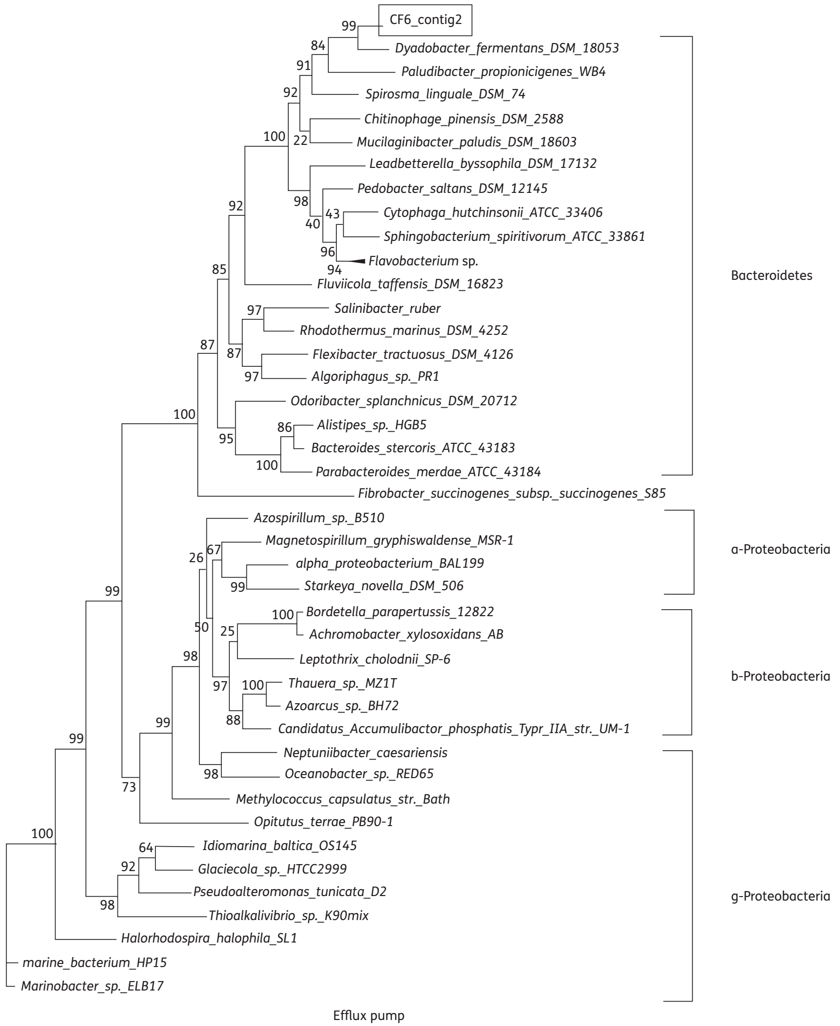


Figure 2. Phylogeny of an efflux pump-encoding gene on a contig assembled from the CF6 virome. Maximum likelihood phylogenetic reconstruction using Phyml. Values on branches represent maximum likelihood support values.

from the patients' sputum samples by GSFLX pyrosequencing and are available from NCBI (www.ncbi.nlm.gov) under the genome project ID 28441 (where NCBI stands for National Center for Biotechnology Information).

Assembly

For each patient we selected reads whose functional annotation was related to antibiotic and toxic compound resistance, according to the blastx search against the SEED database (e-value < 1e-05). These reads were assembled by the GS De Novo Assembler (Roche) using a minimal overlap length of 35 bp and 98% as the identity threshold parameter. Only contigs longer than 400 bp were considered for successive analyses. We classified as 'large' those contigs spanning >1500 bp. The obtained contigs were in turn functionally annotated on the basis of a blastx search against the SEED database (e-value < 1e-05). Larger contigs containing efflux pump genes and fluoroquinolone resistance genes, and contigs containing β-lactamase genes were kept for further analyses.

Open reading frames (ORFs)

On each selected contig we performed an analysis of the ORFs with the Prodigal algorithm.[17] ORFs were translated into amino acid sequences and a stringent blastp search (e-value < 1e-30) was realized for each of them. Only ORFs that had ≥10 hits in the blast non-redundant database were used for phylogenetic reconstruction.

Phylogenetic reconstruction

A phylogenetic tree was constructed for each ORF having enough potential homologues according to a blastp search against the GenBank non-redundant database, i.e. those ORFs presenting ≥10 statistically significant hits using an e-value of 1e-05. A meaningful subset of blast hits was selected to be included in the phylogenetic tree in order to reduce the number of sequences used, as phylogenetic trees comprising hundreds of sequences are computationally heavy to construct and visually difficult to interpret. The subset of sequences included representatives of the different genera and was selected from among the significant blast hits by using an in-house PERL script to create a non-redundant but exhaustive subset of homologues. Using the MUSCLE program,[18] we performed a multiple alignment on the obtained subset of sequences, which were then curated by Gblocks.[19] Tree reconstruction was implemented by the Phyml algorithm[20] using a minimum parsimony starting tree and optimizing tree topology to maximize the likelihood through a nearest neighbour interchange tree topology search strategy. When no phylogenetic information could be inferred from the obtained tree, a new tree was implemented by Phyml using different parameters, by using the subtree pruning and regrafting tree topology search strategy and 10 random starting trees. The constructed trees were visualized using Molecular Evolutionary Genetic Analysis software version 4.0.[21]

rRNA annotation

A search of 16S rRNA sequences was performed on each metagenome by a blastn search against the GreenGenes database using an e-value of 1e-05 and a minimum alignment length of 50 bp. The sequences that significantly matched a 16S rRNA were taxonomically assigned to bacterial phyla according to their best blast hit.

Results

Looking at these specific sequences in the MG-RAST web site (maximum e-value at 1e-05), we found 1031 short sequences (reads) in the CF virome (6.42% of the CF-specific annotated sequences) putatively encoding resistance to antimicrobials (n=563, including 66 encoding multidrug resistance, 59 encoding resistance to fluoroquinolones, 424 encoding β-lactamases and 1 encoding teicoplanin resistance) and toxic compounds or antiseptics (n=468) versus only 3/18 reads

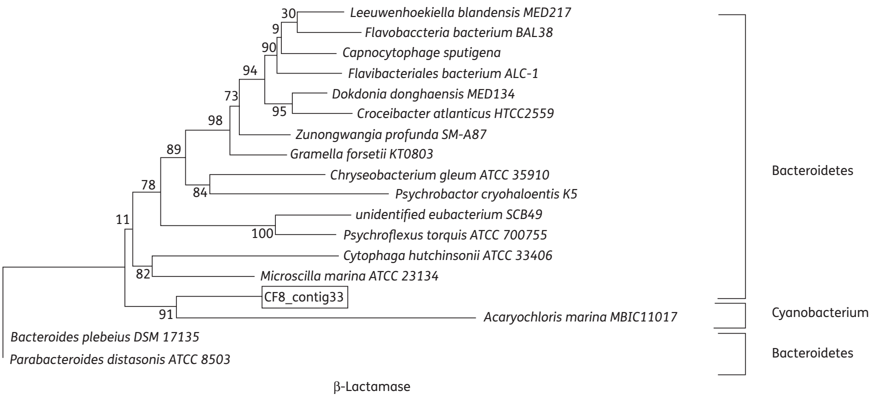


Figure 3. Phylogeny of a β-lactamase-encoding gene on a contig assembled from the CF8 virome. Maximum likelihood phylogenetic reconstruction using Phyml. Values on branches represent maximum likelihood support values.

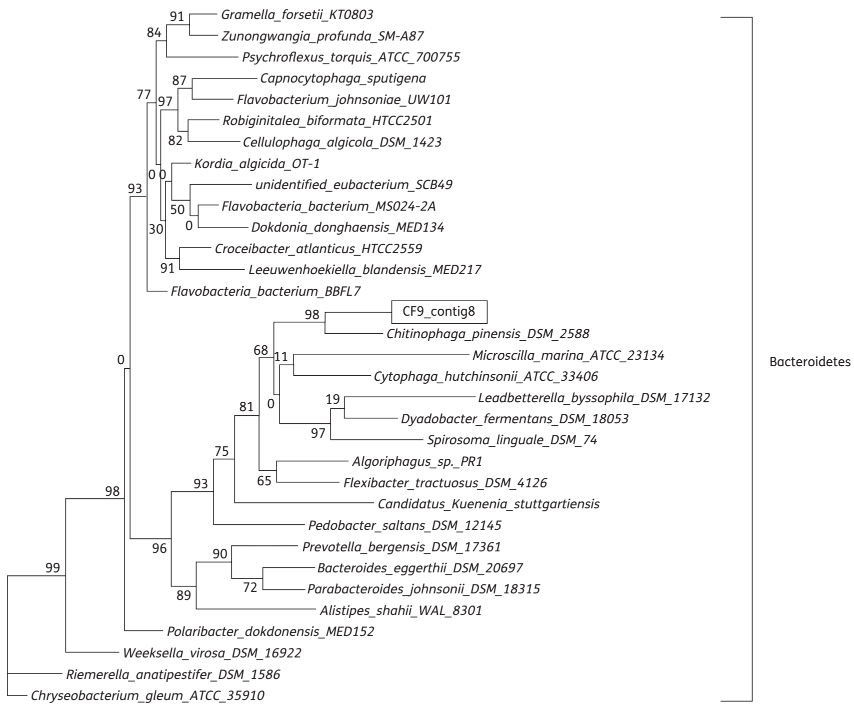


Figure 4. Phylogeny of a fluoroquinolone resistance gene on a contig assembled from the CF9 virome. Maximum likelihood phylogenetic reconstruction using Phyml. Values on branches represent maximum likelihood support values.

encoding antimicrobial resistance in the non-CF virome 'virulence' subsystem (P=0.001) (Figure 1). Starting from the reads annotated by MG-RAST in the 'antibiotic and toxic compound resistance' subsystem, we could assemble 186 reads in larger sequences called 'contigs'. Among them, according to a stringent and curated MG-RAST annotation, we could confidently identify 66 contigs containing different kinds of efflux pump genes and 24 contigs containing antimicrobial resistance genes (15 fluoroquinolone resistance genes and 9 β-lactamase genes; Tables 1 and 2). When possible, evolutionary relatedness was visualized by constructing phylogenetic trees for the ORFs found on the largest contigs. In order to identify the phylogenetic origin of the genes of antibiotic and toxic compound resistance found on these contigs, we searched for the ORFs on each contig and constructed their phylogenetic tree. According to blast results, only a few of these ORFs had enough potential homologues to be able to construct their phylogenetic tree. Moreover, among the constructed trees only a few carried confident and strong enough phylogenetic information to infer the origin of the ORF. However, we were able to extract phylogenetic evidence for nine efflux pump genes, six β-lactamase-encoding genes and six genes encoding fluoroquinolone resistance. According to our results, all these genes would have originated from different classes of Bacteroidetes except for one gene, encoding a β-lactamase, which seems to come from a member of the Cyanobacteria. More specifically, the fluoroquinolone resistance-encoding genes analysed would phylogenetically derive from the class Sphingobacteria of Bacteroidetes. Efflux pump-encoding ORFs would instead originate from four different classes of Bacteroidetes: Flavobacteria; Cytophagia; Bacteroidia;

Table 3. Number of bacterial 16S rRNA sequences found in each CF patient

	Number of 16S rRNA hits				
Phylum	CF6	CF7	CF8	CF9	CF10
Bacteroidetes	75	20	21	67	1
Firmicutes	3	1	2	13	3
Proteobacteria	8	0	0	96	202
Actinobacteria	0	0	1	0	2
Cyanobacteria	0	0	0	17	0
Chlorobi	0	0	0	0	1
Unclassified bacteria	0	0	0	0	1

Data are based on a blastn search against the GreenGenes database with an e-value < 1e-05 and 50 bp of minimum alignment overlap.

and Sphingobacteria. Similarly, β-lactamase-encoding ORFs would have originated from these four different classes of Bacteroidetes (Flavobacteria, Cytophagia, Bacteroidia and Sphingobacteria) and in one case from Cyanobacteria. Figures 2–4 show examples of phylogenetic trees obtained for an efflux pump gene, a β-lactamase-encoding gene and a fluoroquinolone resistance gene, respectively. A search for 16S rRNA sequences was performed on each CF patient's metagenome and a total of 534 16S rRNA matching reads were found. Among them, six bacterial phyla were detected and, interestingly, the Bacteroidetes phylum from which most of the previously analysed genes seem to have originated, was represented by 184 matching sequences (Table 3). More precisely, we found 71 and 56 reads matching with the 16S rRNA sequence from *Chitinophaga* in patient CF6 and patient CF9, respectively. In the same patients, on the basis of our phylogenetic analyses, *Chitinophaga* seems to be the origin of a β-lactamase-encoding gene in patient CF6 and the origin of an efflux pump-encoding gene as well as two fluoroquinolone resistance genes and a β-lactamase gene in CF9.

Discussion

Phages are known to encode bacterial exotoxins, to modify bacterial susceptibility to antibiotics via generalized transduction and to influence bacterial adhesion, colonization and invasion.[3,22,23] Most mobile antibiotic resistance genes are encoded on plasmids or transposons, but phages may play an important role, via transduction, in the mobility of these resistance genes among bacteria. As has been described for CF strains of *S. aureus*, *B. cenocepacia* and *P. aeruginosa*,[3,4,6] phages appear to be essential for the adaptation of the bacterial genomes to their niche, i.e. CF mucus and antimicrobial pressure. Metagenomic studies of DNA viruses in CF patients also showed that the viral community was highly adapted to the CF lung, as are the bacteria.[15,16] The phage taxonomy was common to all CF patients and different from that of healthy subjects, and these CF-associated phages presented with a specific metabolic profile reflecting adaptation to the particular nature of CF mucus.[15,16] This suggests that phage transduction is more frequent in CF, enhancing the essential role of phage transduction in the dramatic adaptation of CF strains to their particular environment, including antimicrobial resistance, as demonstrated in the present study. All CF strains whose genomes are sequenced present an increased genome size with many antimicrobial resistance-encoding genes, as compared with environmental strains, mainly due to a high number of prophages.[3,4,6] Therefore, the genomic evolution of CF strains has led them to fit perfectly into their environment, but on the other side, bacteria may become more resistant to antibiotics. Intensive antibiotic therapy has undeniably increased the life expectancy of CF patients.[1] However, the role of antibiotics in phage transduction is particularly dangerous in the context of CF. The 'collateral' effects of antibiotic treatments, i.e. the transduction of virulence, toxin or antimicrobial resistance genes, often appear at concentrations that are subinhibitory for bacterial growth.[24] If antibiotic treatment is adapted to the bacterium responsible for the disease, this situation may be considered unlikely from a clinical perspective. However, it is important to consider that it can easily occur in CF, where bronchial exacerbation is now considered polymicrobial.[25,26] Metagenomic studies of CF sputa revealed >50 bacterial species that had never been described before in this disease and that could be an incredible source for acquisition by transduction of new antimicrobial resistance determinants,[25,26] as suggested in the present work. Interestingly, we found that phages may also package 16S rRNA sequences from corresponding bacteria that are the source of antibiotic resistance determinants. Our work highlights the fact that our present way of considering CF is probably very incomplete, especially for antibiotic treatment. If phages that are known to be mobilized during chronic infection of the lungs of patients with CF,[14] especially broad-host-range phages, are induced among the microbiota spontaneously,[27] by specific environmental conditions in the CF respiratory tract or by antibiotics, there are many possibilities for HGT to create new bacterial species with specific repertoires adapted to these niches, including MDR bacteria that may spread in this community.

In conclusion, this work illustrates the potential critical role of phages in the adaptation of bacteria to the CF airway, especially involving resistance to antimicrobials.

Funding

This work was partly supported by CNRS, France and a Starting Grant (number 242729) to C. D. from the European Research Council.

Transparency declarations

None to declare.

References

1 Ratjen F, Doring G. Cystic fibrosis. *Lancet* 2003; **361**: 681–9.

2 Waters V, Ratjen F. Multidrug-resistant organisms in cystic fibrosis: management and infection-control issues. *Expert Rev Anti Infect Ther* 2006; **4**: 807–19.

3 Rolain JM, Francois P, Hernandez D *et al*. Genomic analysis of an emerging multiresistant *Staphylococcus aureus* strain rapidly spreading in cystic fibrosis patients revealed the presence of an antibiotic inducible bacteriophage. *Biol Direct* 2009; **4**: 1.

4 Winstanley C, Langille MG, Fothergill JL *et al*. Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res* 2009; **19**: 12–23.

5 Fothergill JL, Mowat E, Ledson MJ *et al*. Fluctuations in phenotypes and genotypes within populations of *Pseudomonas aeruginosa* in the cystic fibrosis lung during pulmonary exacerbations. *J Med Microbiol* 2010; **59**: 472–81.

6 Holden MT, Seth-Smith HM, Crossman LC *et al*. The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. *J Bacteriol* 2008; **191**: 261–77.

7 Smith EE, Buckley DG, Wu Z *et al*. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci USA* 2006; **103**: 8487–92.

8 Brockhurst MA, Buckling A, Rainey PB. The effect of a bacteriophage on diversification of the opportunistic bacterial pathogen, *Pseudomonas aeruginosa*. *Proc Biol Sci* 2005; **272**: 1385–91.

9 Finnan S, Morrissey JP, O'Gara F *et al*. Genome diversity of *Pseudomonas aeruginosa* isolates from cystic fibrosis patients and the hospital environment. *J Clin Microbiol* 2004; **42**: 5783–92.

10 Goerke C, Wirtz C, Fluckiger U *et al*. Extensive phage dynamics in *Staphylococcus aureus* contributes to adaptation to the human host during infection. *Mol Microbiol* 2006; **61**: 1673–85.

11 Colomer-Lluch M, Jofre J, Muniesa M. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS ONE* 2011; **6**: e17549.

12 Goodman AL, Kallstrom G, Faith JJ *et al*. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc Natl Acad Sci USA* 2011; **108**: 6252–7.

13 Goerke C, Koller J, Wolz C. Ciprofloxacin and trimethoprim cause phage induction and virulence modulation in *Staphylococcus aureus*. *Antimicrob Agents Chemother* 2006; **50**: 171–7.

14 Fothergill JL, Mowat E, Walshaw MJ *et al*. Effect of antibiotic treatment on bacteriophage production by a cystic fibrosis epidemic strain of *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother* 2011; **55**: 426–8.

15 Willner D, Furlan M, Haynes M *et al*. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 2009; **4**: e7370.

16 Willner D, Furlan M. Deciphering the role of phage in the cystic fibrosis airway. *Virulence* 2010; **1**: 309–13.

17 Hyatt D, Chen GL, Locascio PF *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 119.

18 Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**: 1792–7.

19 Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**: 540–52.

20 Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003; **52**: 696–704.

21 Tamura K, Dudley J, Nei M *et al*. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007; **24**: 1596–9.

22 Wagner PL, Waldor MK. Bacteriophage control of bacterial virulence. *Infect Immun* 2002; **70**: 3985–93.

23 Zhang X, McDaniel AD, Wolf LE *et al*. Quinolone antibiotics induce Shiga toxin-encoding bacteriophages, toxin production, and death in mice. *J Infect Dis* 2000; **181**: 664–70.

24 Matsushiro A, Sato K, Miyamoto H *et al*. Induction of prophages of enterohemorrhagic *Escherichia coli* O157:H7 with norfloxacin. *J Bacteriol* 1999; **181**: 2257–60.

25 Bittar F, Richet H, Dubus JC *et al*. Molecular detection of multiple emerging pathogens in sputa from cystic fibrosis patients. *PLoS ONE* 2008; **3**: e2908.

26 Harris JK, De Groote MA, Sagel SD *et al*. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. *Proc Natl Acad Sci USA* 2007; **104**: 20529–33.

27 Ghosh D, Roy K, Williamson KE *et al*. Acyl-homoserine lactones can induce virus production in lysogenic bacteria: an alternative paradigm for prophage induction. *Appl Environ Microbiol* 2009; **75**: 7142–52.

Chapter 5

Conclusions

5.1 Conclusions and perspectives

Investigation of unexplored environments

Despite the advent of viral metagenomics and the increasing number of generated viromes, the exploration of viral diversity is incomplete and biased towards the most abundant and easily available types of samples. This thesis contributes to the investigation, by viral metagenomics, of poorly characterized viral communities from environmental and human-associated samples. In particular, we generated the first viral metagenomic study of viral communities in the Sahara, which also represents the first high-throughput viral metagenomic study in hot deserts and in the African continent (chapter 3). Furthermore, we showed a pioneering application of viral metagenomics to the study of an ancient human specimen (chapter 4.1). Finally, we provided unique and original data on viral flora associated to human pericardial fluids (chapter 4.2).

Presence of giant viruses in environmental and human samples

The presence of giant viruses was detected in three different viral metagenomic studies that are contained in this thesis: in four perennial bodies of water in the Sahara desert (chapter 3), in a human 14^{th} century coprolite (chapter 4.1) and in the viral flora associated to human blood (appendix A). In viral metagenomics a 0.2-μm filtration step is currently used for viral particle purification due to an old and overcome definition of viruses as "ultrafilterable particles". The use of such stringent filtering prevents the detection of viruses larger than the filter pores and leads to an underestimation of giant virus diversity and abundance. In this thesis work, we were able to successfully recover giant viruses due to the adoption of a 0.45-μm filtration step, which enabled us to eliminate most bacterial and eukaryotic cells contamination without neglecting the detection of large viruses.

In the environment, the abundance and wide distribution of nucleocytoplasmic large DNA viruses (NCLDVs), including *Mimiviridae* and *Marseilleviridae*, has already been shown to be common [113, 114, 115]. Since the discovery of *Acanthamoeba polyphaga* Mimivirus, about ten years ago, mimiviruses and marseilleviruses have been isolated from freshwater, seawater, and soil samples [116]. Moreover, a previous viral metagenome generated with a 0.45-μm filtration step by Lopez-Bueno et al. also revealed their presence in a cold desert, an Antarctic freshwater lake [33]. In this work we detected *Mimiviridae* in another type of desert, the hot desert of the Sahara (chapter 3). Their presence was first detected *in silico* and then confirmed by electron microscopy observation of a large viral particle with Mimivirus-like morphology.

In contrast to the environment, little is known about the presence of giant viruses in the human body [117, 118]. In this thesis work we were able to recover sequences related to *Mimiviridae* in an ancient human stool sample (chapter 4.1) and to confirm its presence with molecular biology. In addition, following the recent discovery of a new giant virus in modern stool samples, we re-evaluated publicly available human viromes and microbiomes in search of further evidence of the presence of giant viruses in humans. Our results indicated that the presence of giant viruses in humans has been overlooked in some cases (appendix B). In another work that I collaborated on, a giant virus related to Marseillevirus was serendipitously discovered in the serum of a blood donor (appendix A). Although likely asymptomatic in the blood donor, its discovery as a circulating human virus has raised questions about its role in the human body and about the safety of blood transfusions [119]. This is particularly relevant considering the fortuitous discovery of high antibody titers to Marseillevirus in an 11-month-old child suffering from adenitis [120]. Marseillevirus presence was confirmed in the child's blood by PCR and in the lymph node by fluorescent in-situ hybridization and immuno-histochemistry. Finally, another work by Van Etten et al. revealed the identification of sequences related to *Phycodnaviridae, Asfaviridae* and *Mimiviridae* in the post-mortem hippocampus and frontal cortex of brains from individuals affected by mental disorders with significantly higher frequency than in healthy controls [121]. All of these data as well as our results should encourage the use of viral purification protocols that do not neglect giant viruses in both environmental and human-associated viral metagenomic surveys and stimulate a further investigation of the potential pathogenicity of giant viruses for humans.

Bacteriophages in human-related samples

Bacteriophages were found to be abundant in most of the DNA viral communities investigated in this thesis work. For instance, bacteriophages infecting cyanobacteria were observed to be prevalent in the viral communities associated to perennial bodies of water in the Sahara desert (chapter 2) and bacteriophages infecting bacteria commonly found in the gastrointestinal tract or the environment were detected in the viral community of the human coprolite presented in chapter 4.1. We also detected bacteriophage sequences in human pericardial fluids (chapter 4.2). Bacteriophages are known to be abundant in aquatic environments, as well as in some human body sites, such as the gastrointestinal tract or the oropharynx [55, 56, 61]. However, their presence in pericardial fluids was surprising as pericardial fluid is not in contact with the external environment. One of the samples harboring bacteriophages derived from a sudden infant death case and presented a polymicrobial infection, most likely as a consequence of a post-mortem bacterial invasion. Four additional samples were found to harbor bacteriophages infecting *Staphylococci, Enterobacteria, Pseudomonas, Streptococci* and *Burkholderia.* The bacterial hosts of these bacteriophages may be pathogenic for humans and may play a role in the etiology of the pericarditis cases. However, the corresponding bacteria were not detected in the samples during previous hospital routine screenings. We hypothesized that the detection of bacteriophages was more sensitive than the detection of their bacterial hosts. We further suggested that the presence of bacteriophages in these samples could either reflect a trace of an earlier bacterial infection or a current one in which the amount of bacterial DNA is below the detection threshold of the diagnostic method used due to antibiotic treatment [122, 123] and/or clearance by the host immune system. To explore this hypothesis, further studies have recently been conducted at the URMITE laboratory, focusing on clinical strains of *Staphylococcus aureus* and its bacteriophages [124]. The detection of *S. aureus* was performed using primers adopted in hospital diagnostics, whereas the detection of its bacteriophages was performed using primers from a multiplex PCR system targeting five serogroups of *S. aureus* bacteriophages [125]. Ten clinical strains and thirteen blood samples from patients affected by a *S. aureus* infection were tested by classical PCR and revealed the presence of one or more types of *Staphylococcus* bacteriophages. Qualitative and quantitative comparisons regarding the detection limits of classical PCRs targeting the bacterial host or its bacteriophages were also performed. PCR targeting bacteriophages resulted to be a specific method of detection for *S. aureus* infections and to be more sensible than the PCR targeting the bacterium. Furthermore, a qPCR detection system was developed for bacteriophages to simulate testing conditions in hospital screenings. The qPCR targeting bacteriophages was performed on 250 samples from patients affected by an infectious disease (47 patients positive for *S.aureus* and 203 patients negative for *S. aureus*). The results revealed comparable performance between the two systems and they were consistent with those from hospital diagnostic tests for the large majority of samples (232/250 samples). The specificity and sensibility demonstrated by the bacteriophage-based system to detect a *S. aureus* infection may suggest a diagnostic application complementary to already existing bacterium-based screenings. In addition, it may encourage the development of similar bacteriophage-based systems for other bacteria, especially fastidious ones.

The study of bacteriophages in human disease is not only interesting because they may represent a sensible system to detect bacterial infections but also because they play a significant role in shaping the infectious bacterial communities. Bacteriophages represent an important reservoir of metabolic functions for bacteria, which includes antibiotic resistance and virulence factors [55, 60, 59, 84]. Therefore, the study of their functional carriage may provide important clues about the potential emergence of antibiotic-resistant or virulent pathogenic bacteria infecting humans. For example, in this work, a sequence likely carried by a bacteriophage and coding for antibiotic-resistance was retrieved in a human coprolite-associated viral metagenome (chapter 4.1). This confirms previous studies that provide evidence for the dissemination and evolution of antibiotic-resistance genes well before the use of antibiotic treatments [96, 97]. In addition, in chapter 4.3, we analyzed the antibiotic resistance genes retrieved in viral metagenomes, mainly represented by bacteriophages, which were associated to human sputa from healthy individuals and cystic fibrosis patients [60]. A significantly more elevated abundance of these sequences was observed in patients than in controls. Moreover, phylogenetic analyses revealed that these sequences mosty likely originated from the environment. This study further supports the important role of bacteriophages in the emergence of multidrug-resistance bacteria in the airways of cystic fibrosis patients. More generally, it confirms the role of bacteriophages in bacterial adaptation to human-associated niches and it suggests the importance of metabolic analyses on bacteriophage communities to decode the mechanisms of interaction with their host.

Considerations regarding human clinical viral metagenomic studies and the assessment of viral pathogenicity

In the past, most viral studies were performed in a clinical context and new viruses were usually discovered from inquiries about a disease etiology, leading to the belief that all viruses would be pathogenic. Today, the observation of viruses in apparently asymptomatic individuals as well as their prevalence in the human population suggests that viruses can also populate different human body sites as part of the normal flora. Therefore, the simple presence of a virus in a patient does not imply its pathogenicity. Viral metagenomic studies have been increasingly applied to characterize human normal viral flora and to identify viral pathogens responsible for unexplained human diseases. New viruses have been successfully identified by viral metagenomics as responsible of disease outbreaks [65, 66] or as pathogen candidates for some human idiopathic diseases [71, 69, 73]. As discussed in chapter 4.2, co-infection by multiple torque teno virus genotypes has been detected in patients affected by idiopathic pericarditis. Therefore, the question of their potential role in the disease can be addressed. Since their discovery in 2002, TTVs have been observed to be associated with some human diseases. However, their role in disease causation has never been demonstrated and it is difficult to determine due to their prevalence in human population [126, 127, 128, 129]. In addition, TTVs present high genetic variability and their pathogenic potential may be due to a specific strain [130]. Epidemiological studies might be conducted to establish a potential association with idiopathic pericarditis and might be the starting point of further investigations. Recently, we generated sixty-two additional viral metagenomes from the pericardial fluids of patients with pericarditis of idiopathic, non-infectious or known viral origin (unpublished data). The analysis of these data will most likely help to clarify the role of TTVs in pericarditis.

When assessing the role and potential pathogenicity of a virus detected in human samples, it should be considered that Koch's postulates cannot always be satisfied and viral pathogenicity may be due to several factors depending on the host, on the virus itself and on the environment [131]. Indeed a new medical paradigm is emerging, which defines a disease as a disruption of the normal equilibrium state between the microbiome, the virome and the human host. Therefore, the complex web of interactions of a virus within the human virome, with the human microbiome and with the human host has to be considered. Interactions within the viral community are represented by concurrent infections by different viral species or different strains of the same virus. These co-infections may slow the progression of the disease caused by a virus [132], be essential for productive infection [133, 134, 135, 136, 137], or lead to pathology in cases where single infections would have been asymptomatic [138]. Interactions between the viral community and the microbial community populating the same body site may consist of the regulation of the bacterial community structure by bacteriophages or in the bacteriophage-mediated transfer to bacteria of virulence genes, antibiotic resistant genes or others [105, 104, 139]. Finally, concerning the interactions between the viral community and their host, acute infection, clearance of the infection, chronic infection, changes in the host immune and inflammatory response promoted by viral infection and even integration of viruses into the host genome may occur, as recently reviewed by Virgin et al. [140]. Considering this, it is clear that the definition of all of the acting characters of these interactions, beyond the human host and the single inquired virus, is necessary to understand the pathogenicity of a virus and a global approach, such as metagenomics, is suitable to provide it.

5.2 Future directions

Viral metagenomic analyses can be described as a flow, whose source is represented by the massive sequencing of viral communities (Fig. 5.1).

In this flow, three main steps can be distinguished:

1. sequence annotation, to describe “what is there”;
2. functional characterization and viromes comparison, to understand “what does it do”;
3. wet-lab experiments, to verify formulated hypotheses and provide new information.

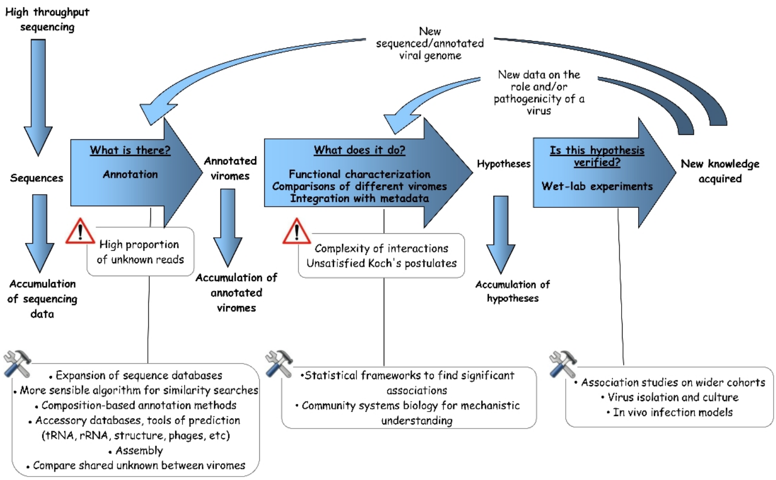


Figure 5.1: Work-flow of viral metagenomic analyses. The arrows represent the steps of the analysis and the scientific question they address. For each step the main issues are reported (warning sign pictogram) as well as the available tools and approaches to resolve them (hammer and wrench symbol).

The flow of a viral metagenomic analysis greatly benefits from close inter-connections between these three steps and from the coordinated use of both the computational approach and of classical biological tools.

Concerning sequence annotation (1.), one challenge is represented by the characterization of unknown sequences. Consistent with previous viral metagenomic studies, viromes presented in this work contain a significant fraction of unidentifiable sequences. Their characterization is one of the main perspectives of this work. The generation of unknown viral sequences represents one of the main advantages of viral metagenomics. However, these sequences have not been exploited or deeply investigated yet. Most computational tools for metagenomic sequences identification depend on BLASTX-based similarity searches, which are affected by the limited size of public sequence databases, the high genetic divergence of viruses and the short size of metagenomic reads. Moreover, in such searches, non-coding sequences are classified as unknown, leading to the over-estimation of the number of unknown reads and to the loss of relevant information for the metagenome description [141]. Due to the high computational cost of a metagenomic analysis, more specific searches using other biological databases and the use of additional tools of prediction, characterization or annotation (ACLAME, CRISPR, etc.) are usually not performed. Furthermore, a large assortment of genes, such as translational factors and aminoacyl-tRNA synthetases, have only recently been detected in viral genomes and computational tools for their prediction and annotation were, therefore, not adopted in viral metagenomics. Today, with the discovery of the wide assortment of genes in giant virus genomes and with the observation that bacteriophages may carry a variety of bacterial genes, it is clear that a viral metagenomic analysis may greatly benefit from these additional computational analyses. However, an increase of the available computational power is required for such analyses. Otherwise, strategies of data dimensions reduction may be used, such as assembly into contigs or the clustering of highly similar reads previous to annotation. Complementary to this approach, the use of wet-lab experiments may be performed but, given the large amount of unknown putative encoding sequences generated in viral metagenomics, an experimental approach is not applicable to all of them and new *in silico* tools are needed for this task. To our knowledge, no *in silico* tools specifically developed for unknown read characterization exist and no specific studies have been conducted to systematically investigate unknown sequences from different viral metagenomes. Nothing is known about the percentage of coding or non-coding sequences among unknown reads, their DNA composition, their average length or their prevalence in different metagenomes or assemblies. A preliminary broad survey of unknown sequences in public viromes would be useful for the design of an efficient strategy to characterize them. Such an investigation may compare unknown sequences from different environments, host-associated samples or host body sites and then cluster the subsets of shared unknown sequences between different types of samples. This would be essential for obtaining clues regarding potential tropism, host range or prevalence and distribution in different ecological niches of new viruses likely encoded by unknown sequences.

The basic annotation of viromes is essential to formulate hypotheses about "what do they do" (2.), which is to understand the role and interactions of the detected viruses in the environment or in the human host. This is a difficult task due to the complexity of the scenario. Statistical frameworks appropriate to viral metagenomics are required to establish significant associations and to provide unbiased comparisons between different viral communities. Significant differences or informative correlations (for example an association between the presence of a virus and a disease) are often difficult to decipher, despite the amount of data generated by next-generation sequencing technologies and despite the increasing number of available studies. Indeed, these data are difficult to interpret as they have an elevated signal to noise ratio. Moreover, a mechanistic understanding of interactions within the viral community or between the viral community and other organisms (including their host) is necessary. This mechanistic analysis may come from the application of a systems biology approach. Up to now, only a few examples of systems biology being applied to more than one model organism exist. A few studies described models for communities of single species organisms [142, 143], for the interaction between a microbe and its host [144] and for the interaction between two microbial species [145, 146]. Today, the development of a complex community systems biology approach using metagenomic data is both an opportunity and a great challenge [147].

In the human host, one of the major concerns with the "what do they do" is the potential involvement of detected viruses in pathology. Wet-lab experiments (3.) are essential to verify this hypothesis. To establish the disease-association of a virus recovered by viral metagenomics in unexplained clinical cases, serological or PCR-based association studies on large cohorts of samples are required. Moreover, virus isolation and culture may enable to describe the virus host range and tropism, to observe cytopathic effects and to produce high titers of the virus for complete genome sequencing or antibody production for *in vivo* infection models. In this work we presented a typical example of the integration of viral metagenomic discoveries with wet-lab experiments (appendix A). Indeed, following the *in silico* discovery of a new virus of the *Marseilleviridae* family in a human serum, morphological observation, detection by molecular biology and serology, isolation and complete sequencing of the virus genome were performed. A further study was then conducted to estimate the prevalence of the virus in asymptomatic blood donors and patients with thalassaemia [148]. The virus was detected by molecular biology in 4% of the donors and 9.1% of the thalassemic patients and serologic tests were positive for 12.6% of donors and 22.7% of thalassemic patients. Consequently, wet-lab experiments not only enable to verify formulated hypotheses regarding the role of a virus and its interactions (*e.g.*, their pathogenicity) and contribute to answering the question about "what does it do" (step 2.), but they also provide further data that are useful for annotation (step 1.), such as the sequencing of new viral genomes or the characterization of the virus host, prevalence and tropism.

In the metagenomic analysis flow illustrated above, each step is essential for the following one. Sufficient sequence data are required for a representative description of what is there and annotation is the fundamental basis for analyses regarding the role and interactions of viruses. These analyses, in turn, produce hypotheses to be verified and further investigated by wet-lab experiments. If any interruption occurs in the flow, an unfruitful accumulation of sequence data, viromes annotations or unverified hypotheses is produced. On the contrary, an uninterrupted flow establishes a virtuous cycle, where wet-lab experiments provide the characterization of sequenced viruses, which is useful for the annotation of new viral metagenomes and to understand the viruses' role and their interactions in human or environmental ecosystems. In such cases, viral metagenomics become a driving force in virology research, enabling us to gather new data on viral diversity, evolution, host range, tropism, role and interactions.

Bibliography

[1] F E Angly, B Felts, M Breitbart, P Salamon, R A Edwards, C Carlson, A M Chan, M Haynes, S Kelley, H Liu, J M Mahaffy, J E Mueller, J Nulton, R Olson, R Parsons, S Rayhawk, C A Suttle, and F Rohwer. The marine viromes of four oceanic regions. *PLoS Biology*, 4(11):e368, November 2006.

[2] O Bergh, K Y Bø rsheim, G Bratbak, and M Heldal. High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–8, August 1989.

[3] L M Proctor and J A Fuhrman. Viral mortality of marine bacteria and cyanobacteria. *Nature*, 343(6253):60–62, January 1990.

[4] C A Suttle. Viruses in the sea. *Nature*, 437(7057):356–361, September 2005.

[5] M S Rappé and S J Giovannoni. The uncultured microbial majority. *Annual Review of Microbiology*, 57:369–394, 2003. PMID: 14527284.

[6] Steven Specter. *Clinical virology manual.* Elsevier, New York, 2nd ed. edition, 1992.

[7] J P Staheli, R Boyce, D Kovarik, and T M Rose. CODE-

HOP PCR and CODEHOP PCR primer design. *Methods in Molecular Biology (Clifton, N.J.)*, 687:57–73, 2011.

[8] T M Rose, E R Schultz, J G Henikoff, S Pietrokovski, C M McCallum, and S Henikoff. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Research*, 26(7):1628–1635, April 1998.

[9] A I Culley, A S Lang, and C A Suttle. High diversity of unknown picorna-like viruses in the sea. *Nature*, 424(6952):1054–1057, August 2003.

[10] J Handelsman, M R Rondon, S F Brady, J Clardy, and R M Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):R245–249, October 1998.

[11] M Breitbart, P Salamon, B Andresen, J M Mahaffy, A M Segall, D Mead, F Azam, and F Rohwer. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14250–14255, October 2002. PMID: 12384570.

[12] M B Duhaime and M B Sullivan. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology*, 434(2):181–6, December 2012.

[13] W A M Hoeijmakers, R Bártfai, K-J Françcoijs, and H G Stunnenberg. Linear amplification for deep sequencing. *Nature protocols*, 6(7):1026–36, July 2011.

[14] R Marine, S W Polson, J Ravel, G Hatfull, D Russell, M Sullivan, F Syed, M Dumas, and K E Wommack. Evaluation of a transposase protocol for rapid generation of shotgun

high-throughput sequencing libraries from nanogram quantities of DNA. *Applied and environmental microbiology*, 77(22):8071–9, November 2011.

[15] M B Duhaime and M B Sullivan. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology*, 434(2):181–6, December 2012.

[16] R V Thurber, M Haynes, M Breitbart, L Wegley, and F Rohwer. Laboratory procedures to generate viral metagenomes. *Nature protocols*, 4(4):470–83, January 2009.

[17] M Breitbart, B Felts, S Kelley, J M Mahaffy, J Nulton, P Salamon, and F Rohwer. Diversity and population structure of a near-shore marine-sediment viral community. *Proceedings. Biological Sciences / The Royal Society*, 271(1539):565–574, March 2004.

[18] M Breitbart and F Rohwer. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *BioTechniques*, 39(5):729–736, November 2005. PMID: 16312220.

[19] S R Bench, T E Hanson, K E Williamson, D Ghosh, M Radosovich, K Wang, and K E Wommack. Metagenomic characterization of chesapeake bay virioplankton. *Applied and environmental microbiology*, 73(23):7629–41, December 2007.

[20] E Dinsdale, O Pantos, S Smriga, R Edwards, F Angly, L Wegley, M Hatay, D Hall, E Brown, M Haynes, L Krause, E Sala, S Sandin, R V Thurber, B L Willis, F Azam, N Knowlton, and F Rohwer. Microbial ecology of four coral atolls in the northern line islands. *PLoS ONE*, 3:e1584, February 2008.

[21] B L Hurwitz and M B Sullivan. The pacific ocean virome (POV): a marine viral metagenomic dataset and associated

protein clusters for quantitative viral ecology. *PloS one*, 8(2):e57355, January 2013.

[22] M Yoshida, Y Takaki, M Eitoku, T Nunoura, and K Takai. Metagenomic analysis of viral communities in (hado)pelagic sediments. *PloS one*, 8(2):e57271, January 2013.

[23] C P Brussaard, D Marie, and G Bratbak. Flow cytometric detection of viruses. *Journal of virological methods*, 85(1-2):175–82, March 2000.

[24] K Holmfeldt, D Odić, M B Sullivan, M Middelboe, and L Riemann. Cultivated single-stranded DNA phages that infect marine bacteroidetes prove difficult to detect with DNA-binding stains. *Applied and environmental microbiology*, 78(3):892–4, February 2012.

[25] Y Tomaru and K Nagasaki. Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. *Journal of Oceanography*, 63(2):215–221, April 2007.

[26] G F Steward, A I Culley, J Mueller, E M Wood-Charlson, M Belcaid, and G Poisson. Are we missing half of the viruses in the ocean? *The ISME journal*, 7(3):672–9, March 2013.

[27] S Roux, F Enault, A Robin, V Ravet, S Personnic, S Theil, J Colombet, T Sime-Ngando, and D Debroas. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE*, 7(3):e33641, March 2012.

[28] B Bolduc, D P Shaughnessy, Y I Wolf, E V Koonin, F F Roberto, and M Young. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated yellowstone hot springs. *Journal of virology*, 86(10):5562–73, May 2012.

[29] R Garrett, D Prangishvili, S Shah, M Reuter, K O Stetter, and X Peng. Metagenomic analyses of novel viruses and plasmids from a cultured environmental sample of hyperthermophilic neutrophiles. *Environmental microbiology*, 12(11):2918–30, November 2010.

[30] G Rice, K Stedman, J Snyder, B Wiedenheft, D Willits, S Brumfield, T McDermott, and M J Young. Viruses from extreme thermal environments. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13341–13345, November 2001.

[31] T Schoenfeld, M Patterson, P M Richardson, K E Wommack, M Young, and D Mead. Assembly of viral metagenomes from yellowstone hot springs. *Applied and environmental microbiology*, 74(13):4164–74, July 2008.

[32] F Santos, P Yarza, V Parro, C Briones, and J Antón. The metavirome of a hypersaline environment. *Environmental microbiology*, 12(11):2965–76, November 2010.

[33] A Lopez-Bueno, J Tamames, D Velazquez, A Moya, A Quesada, and A Alcami. High diversity of the viral community from an antarctic lake. *Science (New York, N.Y.)*, 326(5954):858–61, November 2009.

[34] T Sime-Ngando, S Lucas, A Robin, K P Tucker, J Colombet, Y Bettarel, E Desmond, S Gribaldo, P Forterre, M Breitbart, and D Prangishvili. Diversity of virus-host systems in hypersaline lake retba, senegal. *Environmental microbiology*, 13(8):1956–72, August 2011.

[35] N Fierer, M Breitbart, J Nulton, P Salamon, C Lozupone, R Jones, M Robeson, R Edwards, B Felts, S Rayhawk, R Knight, F Rohwer, and R B Jackson. Metagenomic and

small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and environmental microbiology*, 73(21):7059–66, November 2007.

[36] K-H Kim, H-W Chang, Y-D Nam, S W Roh, M-S Kim, Y Sung, C O Jeon, H-M Oh, and J-W Bae. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and environmental microbiology*, 74(19):5975–5985, October 2008. PMID: 18708511.

[37] B Rodriguez-Brito, Li Li, Li Wegley, M Furlan, F Angly, M Breitbart, J Buchanan, C Desnues, E Dinsdale, R Edwards, B Felts, M Haynes, H Liu, D Lipson, J Mahaffy, A B Martin-Cuadrado, A Mira, J Nulton, L Pasić, S Rayhawk, J Rodriguez-Mueller, F Rodriguez-Valera, P Salamon, S Srinagesh, T F Thingstad, T Tran, R V Thurber, D Willner, M Youle, and F Rohwer. Viral and microbial community dynamics in four aquatic environments. *The ISME journal*, 4(6):739–751, June 2010.

[38] T W Whon, M-S Kim, S W Roh, N-R Shin, H-W Lee, and J-W Bae. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *Journal of virology*, 86(15):8221–31, August 2012.

[39] C Desnues, B Rodriguez-Brito, S Rayhawk, S Kelley, T Tran, M Haynes, H Liu, M Furlan, L Wegley, B Chau, Y Ruan, D Hall, F Angly, R A Edwards, L Li, R V Thurber, R P Reid, J Siefert, V Souza, D L Valentine, B K Swan, M Breitbart, and F Rohwer. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, 452(7185):340–343, March 2008. PMID: 18311127.

[40] D M Morens, G K Folkers, and A S Fauci. The challenge

of emerging and re-emerging infectious diseases. *Nature*, 430(6996):242–9, July 2004.

[41] Fatimah S Dawood, Seema Jain, Lyn Finelli, Michael W Shaw, Stephen Lindstrom, Rebecca J Garten, Larisa V Gubareva, Xiyan Xu, Carolyn B Bridges, and Timothy M Uyeki. Emergence of a novel swine-origin influenza a (H1N1) virus in humans. *The New England journal of medicine*, 360(25):2605–15, June 2009.

[42] V Shinde, C B Bridges, T M Uyeki, B Shu, A Balish, X Xu, S Lindstrom, L V Gubareva, V Deyde, R J Garten, M Harris, S Gerber, S Vagasky, F Smith, N Pascoe, K Martin, D Dufficy, K Ritger, C Conover, P Quinlisk, A Klimov, J S Bresee, and L Finelli. Triple-reassortant swine influenza a (h1) in humans in the united states, 2005-2009. *The New England journal of medicine*, 360(25):2616–25, June 2009.

[43] G J D Smith, D Vijaykrishna, J Bahl, S J Lycett, M Worobey, O G Pybus, S K Ma, C L Cheung, J Raghwani, S Bhatt, J S M Peiris, Y Guan, and A Rambaut. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. *Nature*, 459(7250):1122–5, June 2009.

[44] D Hamre and J J Procknow. A new virus isolated from the human respiratory tract. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)*, 121(1):190–3, January 1966.

[45] K McIntosh, J H Dees, W B Becker, A Z Kapikian, and R M Chanock. Recovery in tracheal organ cultures of novel viruses from patients with respiratory disease. *Proceedings of the National Academy of Sciences of the United States of America*, 57(4):933–40, April 1967.

[46] C Drosten, S Günther, W Preiser, S van der Werf, H-R Brodt, S Becker, H Rabenau, M Panning, L Kolesnikova, R A Fouchier, A Berger, A-M Burguière, J Cinatl, M Eickmann, N Escriou, K Grywna, S Kramme, J-C Manuguerra, S Müller, V Rickerts, M Stürmer, S Vieth, H-D Klenk, A D Osterhaus, H Schmitz, and H W Doerr. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *The New England journal of medicine*, 348(20):1967–76, May 2003.

[47] R A Fouchier, N G Hartwig, T M Bestebroer, B Niemeyer, J C de Jong, J H Simon, and A D Osterhaus. A previously undescribed coronavirus associated with respiratory disease in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6212–6, April 2004.

[48] L van der Hoek, K Pyrc, M F Jebbink, W Vermeulen-Oost, R Berkhout, K Wolthers, P Wertheim-van Dillen, J Kaandorp, J Spaargaren, and B Berkhout. Identification of a new human coronavirus. *Nature medicine*, 10(4):368–73, April 2004.

[49] P Woo, S Lau, C-M Chu, K-H Chan, H-W Tsoi, Y Huang, B Wong, R Poon, J Cai, W-K Luk, L Poon, S Wong, Y Guan, J Peiris, and K-Y Yuen. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *Journal of virology*, 79(2):884–95, January 2005.

[50] S van Boheemen, M de Graaf, C Lauber, T Bestebroer, V Raj, A M Zaki, A Osterhaus, B Haagmans, A Gorbalenya, E J Snijder, and R Fouchier. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio*, 3(6), January 2012.

[51] N Anderson, J Gerin, and N Anderson. Global screening

for human viral pathogens. *Emerging Infectious Diseases*, 9(7):768–774, July 2003. PMID: 12890315.

[52] W Lipkin. The changing face of pathogen discovery and surveillance. *Nature reviews. Microbiology*, 11(2):133–41, February 2013.

[53] M Breitbart, I Hewson, B Felts, J M Mahaffy, J Nulton, P Salamon, and F Rohwer. Metagenomic analyses of an uncultured viral community from human feces. *Journal of Bacteriology*, 185:6220–6223, October 2003.

[54] M Breitbart, M Haynes, S Kelley, F Angly, R Edwards, B Felts, J Mahaffy, J Mueller, J Nulton, and S Rayhawk. Viral diversity and dynamics in an infant gut. *Research in Microbiology*, 159:367–373, June 2008.

[55] S Minot, R Sinha, J Chen, H Li, S Keilbaugh, G Wu, J Lewis, and F Bushman. The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research*, August 2011.

[56] A Reyes, M Haynes, N Hanson, F Angly, A Heath, F Rohwer, and J Gordon. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466:334–338, July 2010.

[57] T Zhang, M Breitbart, W Lee, J-Q Run, C L Wei, S W L Soh, M L Hibberd, E Liu, F Rohwer, and Y Ruan. RNA viral community in human feces: Prevalence of plant pathogenic viruses. *PLoS Biology*, 4:e3, 2006.

[58] S Nakamura, C-S Yang, N Sakon, M Ueda, T Tougan, A Yamashita, N Goto, K Takahashi, T Yasunaga, K Ikuta, T Mizutani, Y Okamoto, M Tagami, R Morita, N Maeda, J Kawai, Y Hayashizaki, Y Nagai, T Horii, T Iida, and T Nakaya. Direct metagenomic detection of viral pathogens in nasal and

fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE*, 4:e4219, January 2009.

[59] D Willner, M Furlan, R Schmieder, J A Grasis, D T Pride, D A Relman, F Angly, T McDole, R P Mariella, F Rohwer, and M Haynes. Colloquium paper: Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proceedings of the National Academy of Sciences*, 108:4547–4553, June 2010.

[60] D Willner, M Furlan, M Haynes, R Schmieder, F Angly, J Silva, S Tammadoni, B Nosrat, D Conrad, and F Rohwer. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*, 4:e7370, October 2009.

[61] D T Pride, J Salzman, M Haynes, F Rohwer, C Davis-Long, R A White, P Loomer, G C Armitage, and D A Relman. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *The ISME Journal*, 6(5):915–926, December 2011.

[62] R Robles-Sikisaka, M Ly, T Boehm, M Naidu, J Salzman, and D T Pride. Association between living environment and human oral viral ecology. *The ISME journal*, page 1–15, April 2013.

[63] V Foulongne, V Sauvage, C Hebert, O Dereure, J Cheval, M Gouilh, K Pariente, M Segondy, A Burguière, J-Cl Manuguerra, V Caro, and M Eloit. Human skin microbiota: High diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS ONE*, 7(6):e38499, June 2012.

[64] G Palacios, J Druce, L Du, T Tran, C Birch, T Briese, S Conlan, P-L Quan, J Hui, J Marshall, J F Simons, M Egholm,

C Paddock, W-J Shieh, C Goldsmith, S R Zaki, M Catton, and W I Lipkin. A new arenavirus in a cluster of fatal transplant-associated diseases. *New England Journal of Medicine*, 358:991–998, March 2008.

[65] T Briese, J T Paweska, L K McMullan, S Hutchison, C Street, G Palacios, M Khristova, J Weyer, R Swanepoel, M Egholm, S Nichol, and W I Lipkin. Genetic detection and characterization of lujo virus, a new hemorrhagic fever-associated arenavirus from southern africa. *PLoS Pathogens*, 5(5):e1000455, May 2009.

[66] Laura K McMullan, Mike Frace, Scott A Sammons, Trevor Shoemaker, Stephen Balinandi, Joseph F Wamala, Julius J Lutwama, Robert G Downing, Ute Stroeher, Adam MacNeil, and Stuart T Nichol. Using next generation sequencing to identify yellow fever virus in uganda. *Virology*, 422(1):1–5, January 2012.

[67] S van Boheemen, M de Graaf, C Lauber, T M Bestebroer, V Raj, A M Zaki, A Osterhaus, B L Haagmans, A E Gorbalenya, E J Snijder, and R Fouchier. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio*, 3(6), January 2012.

[68] G Grard, J N Fair, D Lee, E Slikas, I Steffen, J-J Muyembe, T Sittler, N Veeraraghavan, J G Ruby, C Wang, M Makuwa, P Mulembakani, R B Tesh, J Mazet, A W Rimoin, T Taylor, B S Schneider, G Simmons, E Delwart, N D Wolfe, C Y Chiu, and E M Leroy. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS pathogens*, 8(9):e1002924, September 2012.

[69] J G Victoria, A Kapoor, L Li, O Blinkova, B Slikas, C Wang,

A Naeem, S Zaidi, and E Delwart. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *Journal of Virology*, 83:4642–4651, February 2009.

[70] P F Sullivan, T Allander, F Lysholm, S Goh, B Persson, A Jacks, B Evengård, N L Pedersen, and B Andersson. An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. *BMC Microbiology*, 11:2, 2011.

[71] T Allander. From the cover: Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proceedings of the National Academy of Sciences*, 102:12891–12896, September 2005.

[72] A L Greninger, E C Chen, T Sittler, A Scheinerman, N Roubinian, G Yu, E Kim, D R Pillai, C Guyard, T Mazzulli, P Isa, C F Arias, J Hackett, G Schochetman, S Miller, P Tang, and C Y Chiu. A metagenomic analysis of pandemic influenza a (2009 H1N1) infection in patients from north america. *PLoS ONE*, 5(10):e13381, October 2010.

[73] F Lysholm, A Wetterbom, C Lindau, H Darban, A Bjerkner, K Fahlander, A M Lindberg, B Persson, T Allander, and B Andersson. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS ONE*, 7(2):e30875, February 2012.

[74] J L Mokili, B E Dutilh, Y W Lim, B S Schneider, T Taylor, M Haynes, D Metzgar, C Myers, P J Blair, B Nosrat, N D Wolfe, and F Rohwer. Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PloS one*, 8(3):e58404, January 2013.

[75] G Yu, A L Greninger, P Isa, T G Phan, M A Martínez, M de la Luz Sanchez, J F Contreras, J Ig Santos-Preciado, J Parsonnet, S Miller, J L DeRisi, E Delwart, C F Arias, and C Y Chiu. Discovery of a novel polyomavirus in acute diarrheal samples from children. *PloS one*, 7(11):e49449, January 2012.

[76] S R Finkbeiner, A F Allred, P I Tarr, E J Klein, C D Kirkwood, and D Wang. Metagenomic analysis of human diarrhea: Viral detection and discovery. *PLoS Pathogens*, 4:e1000011, February 2008.

[77] P-L Quan, T Wagner, T Briese, T Torgerson, M Hornig, A Tashmukhamedova, C Firth, G Palacios, A Baisre-De-Leon, C Paddock, S K Hutchison, M Egholm, S R Zaki, J E Goldman, H D Ochs, and W I Lipkin. Astrovirus encephalitis in boy with x-linked agammaglobulinemia. *Emerging infectious diseases*, 16(6):918–25, June 2010.

[78] N L Yozwiak, P Skewes-Cox, M D Stenglein, A Balmaseda, E Harris, and J L DeRisi. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Neglected Tropical Diseases*, 6(2):e1485, February 2012.

[79] F Meyer, D Paarmann, M D'Souza, R Olson, Em Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, J Wilkening, and Ra Edwards. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[80] S Sun, J Chen, W Li, I Altintas, A Lin, S Peltier, K Stocks, E Allen, M Ellisman, J Grethe, and J Wooley. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Research*, 39:D546–D551, November 2010.

[81] S Roux, M Faubladier, A Mahul, N Paulhe, A Bernard, D Debroas, and F Enault. Metavir: a web server dedicated to virome analysis. *Bioinformatics*, 27(21):3074–3075, 2011.

[82] K E Wommack, J Bhavsar, S W Polson, J Chen, M Dumas, S Srinivasiah, M Furman, S Jamindar, and D J Nasko. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in genomic sciences*, 6(3):427–39, July 2012.

[83] M Haynes and F Rohwer. The Human Virome. page 63–77. Springer New York, New York, NY, 2011.

[84] D Willner and M Furlan. Deciphering the role of phage in the cystic fibrosis airway. *Virulence*, 1(4):309–313, August 2010.

[85] R Kuper and S Kröpelin. Climate-controlled Holocene occupation in the Sahara: motor of Africa's evolution. *Science (New York, N.Y.)*, 313(5788):803–7, August 2006.

[86] E Prestel, S Salamitou, and M S DuBow. An examination of the bacteriophages and bacteria of the Namib desert. *Journal of microbiology (Seoul, Korea)*, 46(4):364–72, August 2008.

[87] M Prigent, M Leroy, F Confalonieri, M Dutertre, and M S DuBow. A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles : life under extreme conditions*, 9(4):289–96, August 2005.

[88] S Pääbo, H Poinar, D Serre, V Jaenicke-Despres, J Hebler, N Rohland, M Kuch, J Krause, L Vigilant, and M Hofreiter. Genetic analyses from ancient DNA. *Annual review of genetics*, 38:645–79, January 2004.

[89] E Willerslev and A Cooper. Ancient DNA. *Proceedings. Biological sciences / The Royal Society*, 272(1558):3–16, January 2005.

[90] M Drancourt and D Raoult. Palaeomicrobiology: current issues and perspectives. *Nature reviews. Microbiology*, 3(1):23–35, January 2005.

[91] S Marennikova, E Shelukhina, O Zhukova, N Yanova, and V Loparev. Smallpox diagnosed 400 years later: results of skin lesions examination of 16th century Italian mummy. *Journal of hygiene, epidemiology, microbiology, and immunology*, 34(2):227–31, January 1990.

[92] S Bédarida, O Dutour, A P Buzhilova, P de Micco, and P Biagini. Identification of viral DNA (Anelloviridae) in a 200-year-old dental pulp sample (Napoleon's Great Army, Kaliningrad, 1812). *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 11(2):358–62, March 2011.

[93] P Biagini, C Theves, P Balaresque, A Geraut, C Cannet, C Keyser, D Nikolaeva, P Gerard, S Duchesne, L Orlando, E Willerslev, A N Alekseev, P de Micco, B Ludes, and E Crubezy. Variola virus in a 300-year-old Siberian mummy. *The New England journal of medicine*, 367(21):2056–7, November 2012.

[94] H C Li, T Fujiyoshi, H Lou, S Yashiki, S Sonoda, L Cartier, L Nunez, I Munoz, S Horai, and K Tajima. The presence of ancient human T-cell lymphotropic virus type I provirus DNA in an Andean mummy. *Nature medicine*, 5(12):1428–32, December 1999.

[95] S Sonoda, H C Li, L Cartier, L Nunez, and K Tajima. Ancient HTLV type 1 provirus DNA of Andean mummy. *AIDS*

research and human retroviruses, 16(16):1753–6, November 2000.

[96] R Aminov and R Mackie. Evolution and ecology of antibiotic resistance genes. *FEMS microbiology letters*, 271(2):147–61, June 2007.

[97] V Costa, C King, L Kalan, M Morar, W Sung, C Schwarz, D Froese, G Zazula, F Calmels, R Debruyne, G B Golding, H N Poinar, and G D Wright. Antibiotic resistance is ancient. *Nature*, 477(7365):457–461, 2011.

[98] P-Y Levy, G Habib, F Collart, H Lepidi, and D Raoult. Etiological diagnosis of pericardial effusion. *Future microbiology*, 1(2):229–239, August 2006.

[99] Richard W Troughton, Craig R Asher, and Allan L Klein. Pericarditis. *Lancet*, 363(9410):717–27, February 2004.

[100] P-Y Levy, R Corey, P Berger, G Habib, J-L Bonnet, S Levy, T Messana, P Djiane, Y Frances, C Botta, P DeMicco, H Dumon, O Mundler, J-J Chomel, and D Raoult. Etiologic diagnosis of 204 pericardial effusions. *Medicine*, 82(6):385–391, November 2003.

[101] K E Wommack and R Colwell. Virioplankton: viruses in aquatic ecosystems. *Microbiology and molecular biology reviews : MMBR*, 64(1):69–114, March 2000.

[102] F Ratjen and G Döring. Cystic fibrosis. *Lancet*, 361(9358):681–9, February 2003.

[103] V Waters and F Ratjen. Multidrug-resistant organisms in cystic fibrosis: management and infection-control issues. *Expert review of anti-infective therapy*, 4(5):807–19, October 2006.

[104] C Canchaya, G Fournous, S Chibani-Chennoufi, M L Dillmann, and H Brüssow. Phage as agents of lateral gene transfer. *Current opinion in microbiology*, 6(4):417–24, August 2003.

[105] F Rohwer and R V Thurber. Viruses manipulate the marine environment. *Nature*, 459(7244):207–12, May 2009.

[106] H K Allen, L A Moe, J Rodbumrer, A Gaarder, and J Handelsman. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *The ISME journal*, 3(2):243–51, February 2009.

[107] M Colomer-Lluch, J Jofre, and M Muniesa. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PloS one*, 6(3):e17549, January 2011.

[108] C S Riesenfeld, R M Goodman, and J Handelsman. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental microbiology*, 6(9):981–9, September 2004.

[109] G Torres-Cortés, V Millán, H C Ramírez-Saad, R Nisa-Martínez, N Toro, and F Martínez-Abarca. Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. *Environmental microbiology*, 13(4):1101–14, April 2011.

[110] J L Fothergill, E Mowat, M J Ledson, M J Walshaw, and C Winstanley. Fluctuations in phenotypes and genotypes within populations of Pseudomonas aeruginosa in the cystic fibrosis lung during pulmonary exacerbations. *Journal of medical microbiology*, 59(Pt 4):472–81, April 2010.

[111] C Goerke, J Köller, and C Wolz. Ciprofloxacin and trimethoprim cause phage induction and virulence modulation in

Staphylococcus aureus. *Antimicrobial agents and chemotherapy*, 50(1):171–7, January 2006.

[112] J-M Rolain, P François, D Hernandez, F Bittar, H Richet, G Fournous, Y Mattenberger, E Bosdure, N Stremler, J-C Dubus, J Sarles, M Reynaud-gaubert, S Boniface, J Schrenzel, and D Raoult. Genomic analysis of an emerging multiresistant Staphylococcus aureus strain rapidly spreading in cystic fibrosis patients revealed the presence of an antibiotic inducible bacteriophage. *Biology Direct*, 15:1–15, 2009.

[113] E Ghedin and J-M Claverie. Mimivirus relatives in the sargasso sea. *Virology journal*, 2(1):62, January 2005.

[114] A Monier, J-M Claverie, and H Ogata. Taxonomic distribution of large DNA viruses in the sea. *Genome biology*, 9(7):R106, January 2008.

[115] A Monier, J B Larsen, R-A Sandaa, G Bratbak, J-M Claverie, and H Ogata. Marine mimivirus relatives are probably large algal viruses. *Virology journal*, 5(1):12, January 2008.

[116] B La Scola, A Campocasso, R N'Dong, G Fournous, L Barrassi, C Flaudrops, and D Raoult. Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology*, 53(5):344–53, January 2010.

[117] B La Scola, T J Marrie, J-P Auffray, and D Raoult. Mimivirus in pneumonia patients. *Emerging infectious diseases*, 11(3):449–52, March 2005.

[118] D Raoult, P Renesto, and P Brouqui. Laboratory infection of a technician by mimivirus. *Annals of internal medicine*, 144(9):702–3, May 2006.

[119] J L Goodman. Marseillevirus, Blood Safety and the Human Virome. *Journal of infectious diseases*, July 2013.

[120] N Popgeorgiev, G Michel, D Lepidi, Hand Raoult, and C Desnues. Marseillevirus adenitis in an eleven month old child. submitted, 2013.

[121] J Van Etten, L Jones-Brando, E Severance, M Webster, S Kim, J Gurnon, D Dunigan, F Dicjerson, and R Yolken. Chlorella viruses: potential infectious agents of humans and experimental animals. In *Viruses of microbes. From exploration to applications in the -omics era*, July 2012. EMBO Viruses of Microbes Conference, Brussels, Belgium.

[122] P-E Fournier, F Thuny, H Richet, H Lepidi, J-P Casalta, J-P Arzouni, M Maurin, M Célard, J-L Mainardi, T Caus, F Collart, G Habib, and D Raoult. Comprehensive diagnostic strategy for blood culture-negative endocarditis: a prospective study of 819 new cases. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 51(2):131–40, july 2010.

[123] P-Y Levy, P-E Fournier, F Fenollar, and D Raoult. Systematic PCR detection in culture negative osteoarticular infections. *American Journal of Medicine*, In press, 2013.

[124] Thi Anh Dao Phung. Détection des infections à staphylococcus aureus par leurs bactériophages. Master thesis, Aix-Marseille université, Faculté de Médecine, Juin 2013.

[125] R Pantůcek, J Doskar, V Růzicková, P Kaspárek, E Orácová, V Kvardová, and S Rosypal. Identification of bacteriophage types and their carriage in staphylococcus aureus. *Archives of virology*, 149(9):1689–1703, September 2004. PMID: 15593413.

[126] F Maggi, M Pifferi, C Fornai, E Andreoli, E Tempestini, M Vatteroni, S Presciuttini, S Marchi, A Pietrobelli, A Boner, M Pistello, and M Bendinelli. TT virus in the nasal secretions of children with acute respiratory diseases: relations to

viremia and disease severity. *Journal of virology*, 77(4):2418–25, February 2003.

[127] J-Y Chung, T H Han, J W Koo, S W Kim, J K Seo, and E S Hwang. Small anellovirus infections in Korean children. *Emerging infectious diseases*, 13(5):791–3, May 2007.

[128] F Maggi, M Pifferi, E Tempestini, L Lanini, E De Marco, C Fornai, E Andreoli, S Presciuttini, M L Vatteroni, M Pistello, V Ragazzo, P Macchia, A Pietrobelli, A Boner, and M Bendinelli. Correlation between Torque tenovirus infection and serum levels of eosinophil cationic protein in children hospitalized for acute respiratory diseases. *The Journal of infectious diseases*, 190(5):971–4, September 2004.

[129] F Maggi and M Bendinelli. Human anelloviruses and the central nervous system. *Reviews in medical virology*, 20(6):392–407, November 2010.

[130] P Biagini, R Uch, M Belhouchet, H Attoui, J-F Cantaloube, N Brisbarre, and P de Micco. Circular genomes related to anelloviruses identified in human and animal samples by using a combined rolling-circle amplification/sequence-independent single primer amplification approach. *The Journal of general virology*, 88(Pt 10):2696–701, October 2007.

[131] Nikolay Popgeorgiev and Sarah Temmam. Describing the silent human virome with an emphasis on giant viruses. accepted for publication in Intervirology, 2013.

[132] I Stefanska, M Romanowska, S Donevski, D Gawryluk, and L B Brydak. Co-infections with influenza and other respiratory viruses. *Advances in experimental medicine and biology*, 756:291–301, January 2013.

[133] J K Ball, R Curran, S Berridge, a M Grabowska, C L Jameson, B J Thomson, W L Irving, and P M Sharp. TT virus sequence heterogeneity in vivo: evidence for co-infection with multiple genetic types. *The Journal of general virology*, 80 (Pt 7), July 1999.

[134] P Biagini, P Gallian, H Attoui, J F Cantaloube, P de Micco, and X de Lamballerie. Determination and phylogenetic analysis of partial sequences from TT virus isolates. *The Journal of general virology*, 80 (Pt 2):419–24, February 1999.

[135] W L Irving, J K Ball, S Berridge, R Curran, a M Grabowska, C L Jameson, K R Neal, S D Ryder, and B J Thomson. TT virus infection in patients with hepatitis c: frequency, persistence, and sequence heterogeneity. *The Journal of infectious diseases*, 180(1):27–34, July 1999.

[136] Y E Khudyakov, M E Cong, B Nichols, D Reed, X G Dou, S O Viazov, J Chang, M W Fried, I Williams, W Bower, S Lambert, M Purdy, M Roggendorf, and H A Fields. Sequence heterogeneity of {tt} virus and closely related viruses. *J. Virol.*, 74(7):2990–3000, April 2000.

[137] C Luo, D Tsementzi, N Kyrpides, T Read, and K Konstantinidis. Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE*, 7(2):e30087, February 2012.

[138] H Heydari, S Mamishi, G-T Khotaei, and S Moradi. Fatal type 7 adenovirus associated with human bocavirus infection in a healthy child. *Journal of medical virology*, 83(10):1762–3, October 2011.

[139] H Brüssow, C Canchaya, and W-D Hardt. Phages and the evolution of bacterial pathogens: from genomic rearrangements to

lysogenic conversion. *Microbiology and molecular biology reviews : MMBR*, 68(3):560–602, table of contents, September 2004.

[140] H W Virgin, E J Wherry, and R Ahmed. Redefining chronic viral infection. *Cell*, 138:30–50, July 2009.

[141] F Tobar-Tosse, A C Rodríguez, P E Vélez, M M Zambrano, and P A Moreno. Exploration of noncoding sequences in metagenomes. *PloS one*, 8(3):e59488, January 2013.

[142] T M Conrad, N E Lewis, and B Ø Palsson. Microbial laboratory evolution in the era of genome-scale science. *Molecular systems biology*, 7:509, January 2011.

[143] V A Portnoy, D Bezdan, and K Zengler. Adaptive laboratory evolution–harnessing the power of biology for metabolic engineering. *Current opinion in biotechnology*, 22(4):590–4, August 2011.

[144] A Heinken, S Sahoo, R M T Fleming, and I Thiele. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut microbes*, 4(1):28–40, 2013.

[145] S Stolyar, S Van Dien, K L Hillesland, N Pinel, T J Lie, J A Leigh, and D A Stahl. Metabolic modeling of a mutualistic microbial community. *Molecular systems biology*, 3:92, January 2007.

[146] K Zhuang, M Izallalen, P Mouser, H Richter, C Risso, R Mahadevan, and D R Lovley. Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *The ISME journal*, 5(2):305–16, February 2011.

[147] K Zengler and B O Palsson. A road map for the development of community systems (CoSy) biology. *Nature reviews. Microbiology*, 10(5):366–72, May 2012.

[148] N Popgeorgiev, P Colson, I Thuret, J Chiarioni, P Gallian, X de Lamballerie, D Raoult, and C Desnues. Marseillevirus prevalence in multitransfused patients suggests blood transmission. submitted, 2013.

Appendices

During my thesis I had the opportunity to collaborate on other viral metagenomic studies.
In the work presented in appendix A, I participated in the discovery of a new giant virus in the viral metagenome generated from the serum of a blood donor.

In the work reported in appendix B, I used my competences in viral metagenomics to review the available human metagenomes to search for sequences related to large nucleocytoplasmic DNA viruses and giant viruses. This study originated from the recovery of a giant virus in the feces of a healthy adult in Senegal and aimed at investigating the presence of giant viruses in humans.

In another work, presented in appendix C, I had the opportunity to collaborate on a review of the role of bacteriophages as reservoirs and vehicles of resistance genes in cystic fibrosis.

Appendix A. Article 5. Giant Blood Marseillevirus recovered from asymptomatic blood donors

Giant Blood Marseillevirus recovered from asymptomatic blood donors.

Popgeorgiev Nikolay[1], Boyer Mickael[1,2], Fancello Laura[1], Monteil Sonia[1], Robert Catherine[1], Rivet Romain[1], Nappez Claude[1], Azza Said[1], Chiaroni Jacques[3], Raoult Didier[1] and Desnues Christelle[1,*]

Accepted in Journal of Infectious Diseases 2013. In press.

[1] Aix-Marseille Univ., Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE) UM 63 CNRS 7278 IRD 198 INSERM U1095, Facultés de Médecine et de Pharmacie, Marseille, France;
[2] Current address: Danone Research, 92190 Meudon, France.
[3] Aix-Marseille Université, Etablissement Français du Sang, Anthropologie bio-culturelle, CNRS UMR6578, 13005 Marseille, France

[*] Corresponding author. Email: christelle.desnues@univ-amu.fr

Journal of Infectious Diseases Advance Access published July 2, 2013

Giant Blood Marseillevirus recovered from asymptomatic blood donors

Nikolay Popgeorgiev[1], Mickaël Boyer[1,2], Laura Fancello[1], Sonia Monteil[1], Catherine Robert[1], Romain Rivet[1], Claude Nappez[1], Said Azza[1], Jacques Chiaroni[3], Didier Raoult[1] and Christelle Desnues[1,*]

[1]Aix Marseille Université, URMITE, UM63, CNRS 7278, IRD 198, Inserm 1095, 13005 Marseille, France

[2]Current address: Danone Research, 92190 Meudon, France

[3]Aix-Marseille Université, Etablissement Français du Sang, Anthropologie bio-culturelle, CNRS UMR6578, 13005 Marseille, France

[*]Corresponding author: Christelle Desnues, Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE CNRS-IRD UMR 7278, Aix-Marseille Université, Faculté de médecine, 27 Bd Jean Moulin, Marseille 13385, France. E-mail: christelle.desnues@univ-amu.fr, Phone: 00 33 4 91 32 46 30

Abstract

The study of the human virome is still in its infancy, especially with regards to the viral content of the blood of people that are apparently disease free. In this study, the genome of a new giant virus that is related to the amoeba-infecting pathogen Marseillevirus was recovered from blood donors using high-throughput sequencing. Viral antigens were identified by an immunoconversion assay. The virus was visualized with transmission electron microscopy and fluorescence *in situ* hybridization and was grown in human T-cell lymphocytes. Specific antibody reactions were used to identify viral proteins in the PCR-positive blood donor. Finally, we tested 20 additional blood donors. Three had antibodies directed against this virus and two had circulating viral DNA. This study shows that giant viruses, which are missed by the use of ultrafilters, are part of the human blood viromes. The putative pathogenic role of giant viruses in humans remains undefined.

Introduction

Viruses are the most abundant and diverse entities in the human body[1]. Viral communities have been described in various anatomical sites, including the skin, oropharynx, gut and blood[2-4]. Human blood harbors heterogeneous viral flora, and the majority of these communities are commensal and rarely cause disease in healthy people[5, 6]. Determining when a given virus is a component of the typical floral community rather than an invasive pathogen has direct clinical applications, but remains a challenging process. The detection of new or highly divergent viruses is difficult due to limitations in isolation and cultivation methods and the lack of conserved genetic elements among genomes[7, 8]. During the last decade, the use of sequence- and culture-independent techniques to characterize genetic material from viral populations (this process is known as viral metagenomics) has allowed for the discovery of previously uncharacterized viruses[9-12]. However, despite the increase in knowledge gained during the metagenomics era, viral diversity has been mainly assessed on the ultrafilterable fraction (*i.e.*, on particles recovered after 0.22 µm filtration), which potentially filters out larger viruses and creates a gap in the complete description of the human virome.

The order of Megavirales, also known as nucleocytoplasmic large DNA viruses, contains seven viral families: *Poxviridae, Iridoviridae*, *Ascoviridae, Mimiviridae, Phycodnaviridae, Asfaviridae* and *Marseilleviridae*[13, 14]. Marseillevirus, the first representative of the *Marseilleviridae* family, was isolated due to its ability to prey on amoebae[15]. Marseillevirus possesses a 368 kb double-stranded DNA genome, the sixth largest known viral genome, enclosed in a 250 nm icosahedral capsid.

Results

Pyrosequencing DNA extracted from the viral-enriched fraction of the blood of asymptomatic blood donors generated 20,238 reads (NCBI accession number: PRJNA183996), which were classified by their taxonomic distribution (fig. 1). A best BLAST hits (BBH) comparison against the GenBank database (BLASTn; E-value < 10^{-5}) indicated that 17,257 metagenomic reads (85.3%) had significant hits, and 67.7% of these reads were related to viruses. The majority of these sequences aligned with viruses from the *Anelloviridae* family including Torque Teno virus (TTV), TTV-like, SENV, TTV midi and TTV-like mini (Table 1), which has been previously reported[5, 12, 16]. However, we unexpectedly found that 2.5% of the viral metagenome matched the giant Marseillevirus. We were initially unconvinced that these data were correct, as this virus has only been found in our laboratory.

A genomic assembly of the Marseillevirus-related reads identified two large contigs that mapped on two separate regions that were localized at positions between 111,593-125,241 (13,649 bp) and 210,367-220,539 (10,173 bp) on the Marseillevirus genome (fig. S1). Coding DNA sequence (CDS) predictions on these contigs showed nucleotide differences between the Marseillevirus and metagenome genetic maps. Some of these differences corresponded to stop codons that modify CDS lengths, which strongly suggests the identification of a new virus related to *Marseilleviridae* (Supplementary table S1). We thus performed two sets of polymerase chain reactions (PCR) on the sera pool using orf 268 primers specific to the metagenomic sequences and orf 152 primers, which target both the Marseillevirus genome and the metagenome contigs. PCR orf 152 amplified a 198 bp amplicon in both metagenome and Marseillevirus genomic DNA whereas PCR orf 268 resulted in the amplification of a

608 pb fragment specifically in the metagenomic DNA (fig. S1C) thus confirming the identification of new Marseillevirus-like DNA sequences. We next performed PCR orf 152 on each serum sample (n=10) separately. We detected Marseillevirus-like DNA in the serum from one blood donor, #27725 (fig. S1D). The viral fraction from this blood donor serum was concentrated using three Marseillevirus-specific mouse monoclonal antibodies to allow for additional SOLiD sequencing (NCBI accession number: PRJNA185405). Genomic mapping was performed with CLC software using the Marseillevirus genome (NC_013756.1) as a matrix, which allowed for the isolation of 114,278 mapped reads. These reads were organized into a 357,433 bp consensus sequence (%GC=44.73; fraction coverage=0.95; average coverage=7.68) that showed a close resemblance between Marseillevirus and this virus. This virus was named Giant Blood Marseille-like virus (GBM virus). When CDS predictions of these regions were performed, a total of 617 CDSs were identified, 436 of which were homologous to Marseillevirus, 33 were homologous to Lausannevirus and 148 demonstrated no significant BLAST hit (E-value $<10^{-5}$) (fig. 2A, Supplementary table S1). A neighbor-joining phylogeny analysis based on concatenated alignments of the D5-helicase primase and the A32 ATPase confirmed the clustering of GBM in the *Marseilleviridae* family (fig. 2B).

A two-dimensional Western blot (2DWB) was performed, coupled with mass spectrometry on a concentrated viral fraction of the serum using an α-Marseillevirus polyclonal antibody to identify potential GBM antigens (fig. 3A). Fifteen spots were selected for identification, three of which were further identified containing viral peptides associated with this virus. These peptides corresponded to two hypothetical GBM proteins (ORF 137 and 543) and one flavin-containing amine oxidoreductase

(GBM ORF 123). To observe the viral particle we used transmission electron microscopy (TEM), and observations of the same sample allowed for detection of virus-like particles (VLPs) with an average size of 216.7 nm (±4.7 nm; n=3), which is compatible with Marseillevirus family members (fig. 3B). Additionally, fluorescence in situ hybridization (FISH) using DNA-probe hybridization in the *orf 152 - orf 153* region identified particles of the same size on serum that had been purified and concentrated (fig. 3C).

We next performed a systematic screen of human cell lines and primary cells to evaluate the ability of GBM to infect human cells. These screens included blood cell lines, such as monocytes (THP1), lymphocyte T-cells, (Jurkat) and anti-CD14 purified primary human macrophages. A serum inoculation of 14 days did not result in GBM detection in the supernatant of HeLa cells, THP1, or macrophages. The PCR results were positive for the Jurkat supernatant, which suggested that GBM was able to infect lymphocyte T-cells (fig. 4A). In addition, Jurkat supernatant ultracentrifugation combined with DNAse treatment still showed the presence of GBM DNA, which strongly suggested the production and release of encapsidated viral particles from these cells (fig. 4A). Additional FISH and TEM observations of Jurkat cell inclusions confirmed these results. GBM DNA and viral particles were detected inside the Jurkat cytosol 21 days post-infection (fig. 4B).

ELISA was used to locate specific antibodies with Marseillevirus as an antigen in the GBM-infected blood compared to a PCR-negative blood control. A high level of IgG antibodies (1:1,000) was detected in the infected sample compared to none in the control (fig. 5A). This serum was subsequently submitted to a 2DWB on the Marseillevirus proteome. We were able to identify 26 protein spots that were found in the genome of GBM, with the most reactive being the major viral capsid protein (MAR_ORF342) (fig. 5B).

Finally, we evaluated the prevalence of GBM infection in twenty additional blood donors by performing combined IgG detection and PCR amplification/sequencing on 20 (Male= 65%; median age=46.5 years) sera sampled from asymptomatic blood donors (Montpellier, France). We found three blood donors, with IgG levels above the calculated ELISA threshold (OD_{490nm}>0.260). Moreover, two were PCR positive for GBM and one presented elevated levels of IgG (Table 2; Supplementary table S2).

Discussion

In this study, we definitively established the presence of a novel giant virus in the blood of an asymptomatic blood donor by complete genome sequencing, antigen detection, morphological visualization (TEM and FISH), and cell culture. Although these findings should be further confirmed, contamination is unlikely because of the multiple procedures used to identify the virus, the fact that this is the first Marseillevirus growing in human cells, and because of its original genomic sequence.

Recently, another virus of the *Marseilleviridae* family, Senegalvirus, was isolated from a human stool sample in an asymptomatic patient[17]. The present study confirms that some giant viruses of the *Marseilleviridae* family such as GBM are associated with humans. Because giant viruses escape current screening techniques, human infections by giant viruses related to the *Mimi*- and *Marseilleviridae* families may be underestimated. The fact that GBM infection was found in apparently healthy blood donors suggests that this infection may present asymptomatically or with only mild symptoms. Additionally, the presence in some cases of both IgG antibodies and viruses in the blood suggests possible chronic carriage. Altogether, these results place *Marseilleviridae* in the human virome and raise questions about the long-term consequences of this viral carriage.

Methods

Blood sampling

Viral metagenomics studies were performed on blood samples from healthy donors (n=10) collected at the "Etablissement Français du Sang" Marseille (France). Epidemiological study was performed on additional blood samples (n=20) collected at the "Etablissement Français du Sang" Montpellier (France). Blood pockets were stored at 4°C before further processing.

Viral purification and immunoprecipitation

Blood samples (40 mL) were centrifuged at 3,000 × g at 4°C for 10 minutes to pellet the blood cells and cellular debris. 10 milliliters of each cell-free plasma sample were aliquoted and filtered through 0.45 µm Whatman filters. The filtrate was loaded onto a CsCl step gradient consisting of 1 mL each of 1.7, 1.5, and 1.2 g/mL CsCl in phosphate-buffered saline (PBS; 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.4 mM KH_2PO_4, pH 7.3) and then centrifuged at 55,000 × g at 4°C for 2 h. The 1.2–1.5 g/mL fraction was collected and centrifuged at 55,000 × g at 4°C for 2 h to pellet viral particles, which were then resuspended in 200 µL of 0.02 µm Whatman-filtered PBS. The purified viral particles were then incubated with 0.2 volumes of chloroform for 10 min at room temperature. The chloroform solution was pelleted, and the supernatant was removed and treated with 2.5 U DNase I (Sigma-Aldrich) per microliter of sample for 2 h at 37°C to remove residual host DNA. For GBM immunoprecipitation, serum from blood donor #27725 was filtered through 0.45 µm Whatman filters and then concentrated on a 0.1 µm filter. Attached viral particles were resuspended in Page's Amoeba Saline Solution. Primary anti-Marseillevirus antibody and secondary antibody incubations were performed overnight with constant agitation. All steps were carried out at 4 °C unless otherwise stated. Briefly, pre-cleared serum solution (80 µL diluted in 0.5 mL of 0.02 µm-filtered PBS) and sheep anti-mouse IgG-coated beads (10 µL) were incubated with a mixture of three mouse monoclonal anti-Marseillevirus antibodies (2 µg). On the next day, the solution was incubated with magnetic sheep anti-mouse IgG coated beads (3.5 x 10^7 beads, Dynabeads®, Invitrogen). Beads containing immunoprecipitated viral particles were magnetically pulled down and washed three times with PBS and then resuspended in 0.2 mL of PBS. Viral DNA was then extracted using the High Pure Viral Nucleic Acid Kit (Roche) according to the manufacturer's protocol.

High-throughput sequencing

Purified viral DNA was amplified with Genomiphi (GE Healthcare) to generate sufficient material for shotgun 454 pyrosequencing library preparations. The resulting amplified DNA was purified with silica columns (Qiagen) to remove the enzyme, dNTPs and primers. DNA was pyrosequenced on a Roche 454 Life Sciences GS FLX Titanium platform (1/16 plate used), generating raw data (11.9 Mb) with an average length of 389 bp. The same protocol was used for the serum from blood donor #27725, and the immunoprecipitated fraction was then further sequenced on an AB SOLiD 4 System platform using 1/96 of the plate.

Read processing, assembly and contig analysis

The sequences obtained from the 454 pyrosequencing were screened to remove exact and nearly identical duplicates, which are a common artifact of the pyrosequencing technology. Duplicate removal was performed using the CD-HIT-454 program available under the CAMERA 2.0 web portal. This process resulted in 8.2 Mb of non-redundant sequences with an average length of 406 bp, which were then subsequently taxonomically classified by a BLASTn search against the GeneBank nucleotide database with an E-value<10^{-5}.

Read assembly was performed by Newbler software (Roche). We chose a minimum overlap length of 35 bp and a minimum overlap identity of 98%. Only contigs longer than 400 bp were kept for further analyses. Large contigs were classified as those spanning more than 1,500 bp. Contigs were annotated by performing a BLASTx search against the BLAST non-redundant database nr with an E-value<10^{-5}. SOLiD sequencing generated 15,568,200 paired sequences. Marseillevirus was used as a reference for genome mapping (CLC software). The generated consensus genomic sequence was further analyzed for open reading frames (ORFs) using Prodigal. Predicted ORFs were then compared to GenBank nr using BLASTp with an E-value<10^{-5}.

PCR and primer design for targeting metagenome and Marseillevirus homologue genes

To evaluate the presence of human and mitochondrial DNA contamination, the sample was checked by PCR using specific primers: H18S F5'-TCAAGAACGAAAGTCGGAGG-3', H18S R5'-CAGCTTTGCAACCATACTCC-3', Mit3130 F5′-AGGACAAGAGAAATAAGGCC-3′, and Mit3301 R5′-AGGACAAGAGAAATAAGGCC-3′. The presence of GBM in human sera was evaluated using primers amplifying *orf 152* (orf 152F 5' AGACCCAAACTCGCAGCTTA-3' and orf 152R 5'-CCGGAAGATTCCAAGTTTCA-3') and orf 268F (5'-ACAACCTCCTACCTTCACC-3' and orf 268R 5'-AATTCTCCTCCGCCTTCA-3') for *orf 268* amplification. Amplification using *Phusion* DNA Polymerase (New England Biolabs) started with an initial denaturation step at 98 °C for 30 s followed by 35 cycles of 98 °C for 10 s, 53 °C for 30 s, and 72 °C for 30 s. Sequencing reactions were carried out with the reagents of the ABI Prism dye terminator cycle sequencing ready reaction kit (Perkin Elmer Applied Biosystems, Foster City, Calif.) according to the manufacturer's instructions.

Enzyme-linked immunosorbent assay (ELISA)

The presence IgG antibodies specific to Marseillevirus like particles were evaluated using the ELISA assay with Marseillevirus as the antigen. ELISA negative controls were obtained from specific pathogen free three-month-old balb/c mice sera (n=10) (Charles River, Lyon), which were supposed exempted from GBM infection. ELISA assay was performed as follows: purified Marseillevirus (20 ng), was coated with carbonate buffer (15 mM Na_2CO_3, 35 mM $NaHCO_3$, 0.2 g/L NaN_3, pH 9.6) overnight at 4°C in a flat-bottom 96-well ELISA microplate. The plate was then washed 2 times with PBS, saturated with 5% Bovin serum album (Sigma-Aldrich) for 2 h and incubated with sera at 1/100 dilution. IgG titters of #27725 and #20363 blood donors were estimated using serial dilutions ranging from 1/10 to 1/5000. For a positive control, mouse anti-Marseillevirus antibody was used at 1/1000 dilution. Following sera incubation, the plate was washed 3 times with PBS 0.1% Tween 20 and incubated for 1 h with secondary HRP-conjugated anti-human IgG at 1/5000 dilution (Jackson ImmunoResearch). Detection was performed at 490 nm using o-Phenylenediamine dihydrochloride substrate (Sigma-Aldrich). Average results for IgG levels were obtained from two independent experiments. Negative ($OD_{490\ nm}$=0.057) and positive controls ($OD_{490\ nm}$=0.474) were included for each plate. ELISA threshold ($OD_{490\ nm}$=0.260) was calculated using relative percentage of positivity (RPP%) formula $RPP\% = \frac{ODtreshold - ODneg.control}{ODpos.control - ODneg.control}$. RPP=50% was used for which 99.9% relative specificity and 90.4% relative sensibility of the ELISA results were estimated [18].

Cell culture and viral infection

Human cells were maintained under standard culture conditions (37°C and 5% CO_2). THP1 cells were cultured in MEM media supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, and 1% penicillin/streptomycin (PS). HeLa cells were cultured in DMEM media supplemented with 10% FBS and 1% PS. Jurkat cells were cultivated in RPMI supplemented with 10% FBS and 1% PS. Human primary macrophages were obtained from fresh peripheral blood mononuclear cells (PCR negative for possible GBM infection) incubated with anti-CD14 magnetic beads by following the manufacturer's recommendations (Dynabeads® CD14, Life Technologies). Purified cells were incubated for two days to allow attachment and differentiation. Infections with GBM containing serum were performed by inoculating cells at 100% confluence with a purified viral suspension diluted 1/10 in cell media for 48 hours. At seven and fourteen days post inoculation, 200 µL of cell suspension was harvested and centrifuged at 1350 rpm for 5 min. DNA from the cell supernatant and cell pellet was extracted using a High Pure Viral Nucleic Acid Kit and then amplified by PCR using the orf 152 and H18S primers. For DNAse treatment, 35 mL of Jurkat T-cell supernatant fourteen days post inoculation was centrifuged at 22,000 rpm for 2 hours to pellet viral particles. The pellet was resuspened in 0.5 mL of sterile 0.02-filtered PBS. Viral suspension (0.2 mL) was treated with 4 U of TURBO DNAse (Life Technologies) for 1H at 37°C following by DNAse inactivation by adding 0.1 volume of DNAse inactivation buffer. Nucleic acids were extracted using the High Pure Viral Nucleic Acid Kit (Roche) according to the manufacturer's protocol.

Fluorescence in situ hybridization and microscopy observations

For serum, the fluorescence in situ hybridization concentrated viral suspension (50 µL) was fixed in 4% formol for 1 h at room temperature, pelleted at 56,000 × g for 2 h and then embedded in sterile agarose. The agarose plug was dehydrated in 100% ethanol for 2 h, embedded in paraffin and cut into thin sections. Agarose sections were deposited onto glass slides and deparaffinized. The sections were incubated with 50 ng of DIG-labeled DNA probe (*orf152-orf153*) for 5 min at 95°C and then for 20 h at 37°C. The sections were washed 3 times in 1:1 (vol:vol) formamide SSC 2x followed by 3 washes with SSC 2x for 3 minutes, and they were then washed for 30 min in TNB buffer (100 mM Tris-HCl, pH 7.5, 150 mM NaCl, Blocking buffer 1X (Roche)). Anti-DIG FITC mouse antibodies (1/200) were incubated for 30 min followed by 3 washes for 3 minutes in TNT (100 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% Tween 20) buffer, counterstained with DAPI and stored at 4°C for further observation using a Leica SP5 confocal microscope. For epifluorescence observations, the purified viral suspension

(20 µL) was diluted in PBS containing 2% PFA and filtered through a 0.02 µm Whatman filter. The filter containing viral particles was stained with 2.5x SYBR® Gold (Molecular Probes) for 10 min, washed two times with PBS 1X and then observed under an epifluorescence inverted microscope (Leica DMI6000). The average VLP number was estimated using ImageJ software and was calculated from three representative fields. For electron microscopy detection, Formvar-coated grids were deposited in a 40 µL viral suspension drop and incubated for 30 min at 37°C. The grids were then incubated for 10 seconds on 1% ammonium molybdate, dried on absorbing paper and observed with a transmission electron microscope Morgagni 268D (Philips).

Two-dimensional protein analysis

A two-dimensional protein analysis was performed as previously described[15]. Following protein migration, gels were silver-stained. Selected excised spots were further identified using MALDI-TOF mass spectrometry. For Western blot analyses, the samples were transferred to a 0.45 µm nitrocellulose membrane and immunoblotted overnight at 4°C using a mouse anti-Marseillevirus polyclonal antibody (1/2500) or #27725 serum (1/1000) in PBS + 0.3% Tween 20 with 5% non-fat dried milk. Secondary HRP-conjugated anti-mouse IgG goat antibody was used at a 1/5000 dilution.

Acknowledgments: We are grateful to Audrey Averna and Audrey Borg for their help with electron microscopy, Lina Barassi for amoeba cultures and Marseillevirus production, and Nicholas Armstrong for mass spectrometry identification. **Conflict of**

interest: The authors declare no conflict of interest. **Funding:** This work was supported by the Starting Grant #242729 from the European Research Council to C. Desnues.

Footnote:

Conflict of interest: The authors declare no conflict of interest.

Funding: This work was supported by the Starting Grant #242729 from the European Research Council to C. Desnues.

Author contributions: NP, MB, DR and CD designed the experiments. NP, MB, LF, SM, CR, RR, CN and SA performed the experiments. NP, MB, DR and CD analyzed the data. NP, MB, DR and CD wrote the paper. JC provided the blood samples.

This work was presented at the poster session of the Viruses of Microbes 2012 meeting Brussels, Belgium (16 - 20 July, 2012). Abstract # 222.

Corresponding author : Christelle Desnues, Unité de recherche sur les maladies infectieuses et tropicales émergentes, URMITE CNRS-IRD UMR 7278, Aix-Marseille Université, Faculté de médecine, 27 Bd Jean Moulin, Marseille 13385, France. E-mail: christelle.desnues@univ-amu.fr

Phone: 00 33 4 91 32 46 30

Fax: 00 33 4 91 38 77 72

Mickaël Boyer : Current address: Danone Research, 92190 Meudon, France.

References:

1. Matthew Haynes FR. The Human Virome. Metagenomics of the human body **2011**:63-77.

2. Reyes A, Haynes M, Hanson N, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature **2010**; 466:334-8.

3. Foulongne V, Sauvage V, Hebert C, et al. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. PloS one **2012**; 7:e38499.

4. Kim C, Ahmed JA, Eidex RB, et al. Comparison of nasopharyngeal and oropharyngeal swabs for the diagnosis of eight respiratory viruses by real-time reverse transcription-PCR assays. PloS one **2011**; 6:e21610.

5. Bernardin F, Operskalski E, Busch M, Delwart E. Transfusion transmission of highly prevalent commensal human viruses. Transfusion **2010**; 50:2474-83.

6. Fryer JF, Delwart E, Hecht FM, et al. Frequent detection of the parvoviruses, PARV4 and PARV5, in plasma from blood donors and symptomatic individuals. Transfusion **2007**; 47:1054-61.

7. Rose TM, Schultz ER, Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic acids research **1998**; 26:1628-35.

8. Specter S. Clinical virology manual, 2nd ed., **1992** Elsevier, New York).

9. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol **2012**; 2:63-77.

10. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B. Cloning of a human parvovirus by molecular screening of respiratory tract samples.

Proceedings of the National Academy of Sciences of the United States of America **2005**; 102:12891-6.

11. Lysholm F, Wetterbom A, Lindau C, et al. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. PloS one **2012**; 7:e30875.

12. Breitbart M, Rohwer F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques **2005**; 39:729-36.

13. Colson P, de Lamballerie X, Fournous G, Raoult D. Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. Intervirology **2012**; 55:321-32.

14. Koonin EV, Yutin N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology **2010**; 53:284-92.

15. Boyer M, Yutin N, Pagnier I, et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proceedings of the National Academy of Sciences of the United States of America **2009**; 106:21848-53.

16. Nishizawa T, Okamoto H, Konishi K, Yoshizawa H, Miyakawa Y, Mayumi M. A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. Biochemical and biophysical research communications **1997**; 241:92-7.

17. Lagier JC, Armougom F, Million M, et al. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect **2012**.

18. Desquesnes M. [International and regional standardization of immunoenzyme tests: methods, concerns and limitations]. Rev Sci Tech **1997**; 16:809-23.

Figures Legends

Figure 1. Classification of the human blood virome metadata.

Taxonomic distribution of metagenomic reads based on GenBank taxonomic classification of the best BLASTn hit (E-value < 10^{-5}).

Figure 2. Genome mapping and phylogenetic analysis of the GBM genome.

(A) Genomic map of the predicted CDS of the GBM chromosome using 454 pyrosequencing and SOLiD data. Forward/reverse CDS are represented with blue/red lines, respectively. The two contigs obtained using 454 pyrosequencing are represented by light green boxes. The BLASTp results of the predicted CDS matching Marseillevirus ORFs are presented in orange, and those matching Lausannevirus are presented in purple boxes. GC deviation and GC skew from the average are represented as blue/red graphs, respectively. **(B)** A neighbor-joining tree (number of bootstrap replications=1000) based on concatenated alignments of two NCLDV core proteins: D5-helicase primase and A32 ATPase. The GBM branch is indicated with a red arrow. The NCLDV family is represented as follows: (○) *Iridoviridae*, (●) *Ascoviridae,* (□) *Marseilleviridae,* (■) *Mimiviridae,* (Δ) *Phycodnaviridae*. (◊) *Asfaviridae* and (♦) *Poxviridae*.

Abbreviations: **AtV** (Ambystoma tigrinum virus), **SGI** (Singapore grouper iridovirus), **LDV-1** (Lymphocystis disease virus 1), **LDV-isolate China** (Lymphocystis disease virus isolate China), **ISKNV** (Infectious spleen and kidney necrosis virus), **IIV3** (Invertebrate iridescent virus 3), **IIV6** (Invertebrate iridescent virus 6), **TnA-2c** (Trichoplusia ni ascovirus 2c), **HvA-3e** (Heliothis virescens ascovirus 3e), **SfAV-1a**

(Spodoptera frugiperda ascovirus 1a), **OsV5** (Ostreococcus virus 5), **EhV86** (Emiliania huxleyi virus 86), **PBCV-FR483** (Paramecium bursaria Chlorella virus FR483), **FsV** (Feldmannia species virus), **EsV1** (Ectocarpus siliculosus virus 1), **ASFV** (African swine fever virus), **MCV subtype 1** (Molluscum contagiosum virus subtype 1).

Figure 3. Detection of GBM infection and immune response in blood donor #27725.

(A) Bidimensional protein analysis by silver staining and Western blot analysis of blood serum #27725. The presence of viral proteins in serum #27725 was evaluated using a mouse polyclonal α-Marseillevirus antibody. Selected spots (black arrows) were cut and analyzed by MALDI-TOF mass spectrometry. Positive spots were reported on the 2D gel. **(B)** Negative staining of virus-like particles present in the virus-purified fraction from serum of blood donor #27725; scale bar = 200 nm. **(C)** Epifluorescence microscopy images obtained from fluorescence in situ hybridization on GBM experiments on serum #27725 and #20363 DNA-probes synthesized using the Marseillevirus genomic region *orf 152 – orf 153* (upper panels). Sections were counterstained with DAPI (middle panels). The merged green and blue channels are presented on the bottom; scale bar = 2 µm.

Figure 4. GBM infection of human T lymphocytes.

(A) PCR results from serum #27725 inoculation (day 14 post-inoculation) of various types of cells lines, including HeLa cells, THP1 monocytes, primary anti-CD14 purified human macrophages, and Jurkat T-cell lymphocytes. Pluses and minuses represent the presence or the absence of serum inoculation and/or 22,000 rpm + DNAse treatment.

GBM DNA was detected in the lymphocyte T cells supernatant by PCR amplification of the viral D5 helicase (*orf 152)*. PCR amplification after viral pelleting at 22,000 rpm coupled with DNAse treatment showed the presence of encapsidated viral DNA. 18S rDNA was used as internal PCR control to verify the absence of DNA contamination from host cell and thus the success of 22,000 rpm centrifugation and DNAse treatment of the culture. **(B)** Confocal microscopy images of Jurkat T-cells fourteen days post-inoculation with negative control serum (#20363) or with #27725 serum. Viral detection was performed using DIG-labeled DNA probe hybridization to the *orf* 152 - *orf* 153 genomic region. Cell nuclei were stained with H33342 dye (middle panel). Merged channels are presented on the right. Scale bar = 10 µm (**C**) Electron microscopy images of Jurkat cell supernatant at J14 post-inoculation spun at 22,000 rpm (upper panel). Viral particle detection in Jurkat embedded cells; **C** (cytosol); **M** (cellular membrane). Scale bar = 100 nm.

Figure 5. Detection of GBM immune response in blood donor #27725.

(A) Graphic representation of the ELISA testing results for serial dilutions of total IgG extracted from GBM-positive (#27725) and GBM-negative (#20363) sera. **(B)** Bidimensional protein analysis by silver staining and Western blot analysis using Marseillevirus as an antigen, which showed the existence of anti-GBM IgG antibodies in serum #27725. Positive spots were reported on the 2D gel and WB (black arrowheads). Control serum (PCR and ELISA, negative for infection) was used as a negative control.

Table 1. Viral reads classification.

Viral class	Viral species	Reads (n)	Percentage (%)
ssDNA viruses	TTV	11080	81.40
	TTV-like	1575	11.57
	SENV	576	4.23
	TTV midi	12	0.09
	TTV-like mini	11	0.08
	RSM1	6	0.04
	RSM3	11	0.08
dsDNA viruses	Marseillevirus	333	2.45
	Enterobacteria phage λ	7	0.05

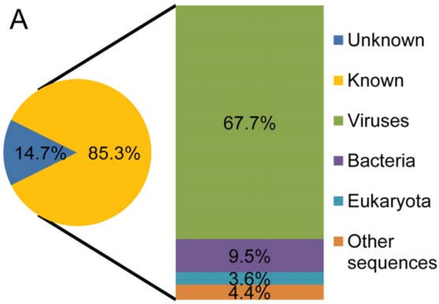
Viral reads classification was based on BLASTn significant hits (E-value < 10-5) from searches against the GenBank non-redundant database. Abbreviations: TTV: Torque Teno Virus. SENV: SEN virus. RSM: *Ralstonia solanacearum* phage. ssDNA: single-stranded DNA viruses; dsDNA: double-stranded DNA viruses.

Table 2. Serological and molecular survey for GBM infection performed on 20 blood donors

	Blood Donors	IgG+	PCR +	IgG+/PCR +
Sample number	20	3/20 (15%)	2/20 (10%)	1/30 (5%)
M/F (M %)	13/7 (65%)	2/1 (66,7%)	2/0 (100%)	1/0 (100%)
Median age (range)	46.5 (22-65)	53 (44-59)	49 (45-53)	53

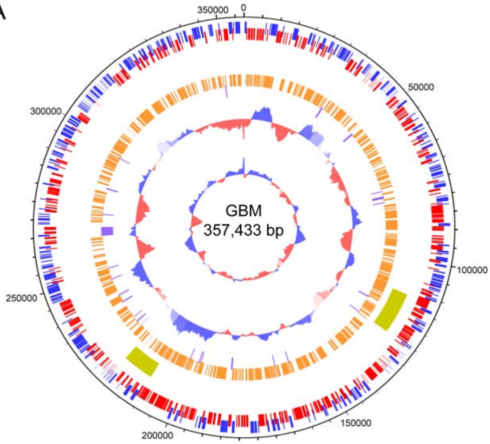
Summary results from serological and molecular testing for the presence of GBM IgGs and DNA in serum performed on 20 blood donors. IgG positive threshold was fixed at 0.26 OD. PCR positive sera were systematically verified by DNA sequencing. See Materials and methods section for further information.

23


A
14.7%
85.3%
67.7%
9.5%
3.6%
4.4%
Unknown
Known
Viruses
Bacteria
Eukaryota
Other sequences

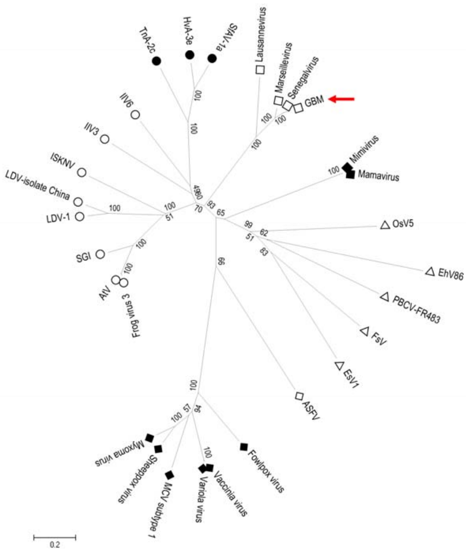
24

A



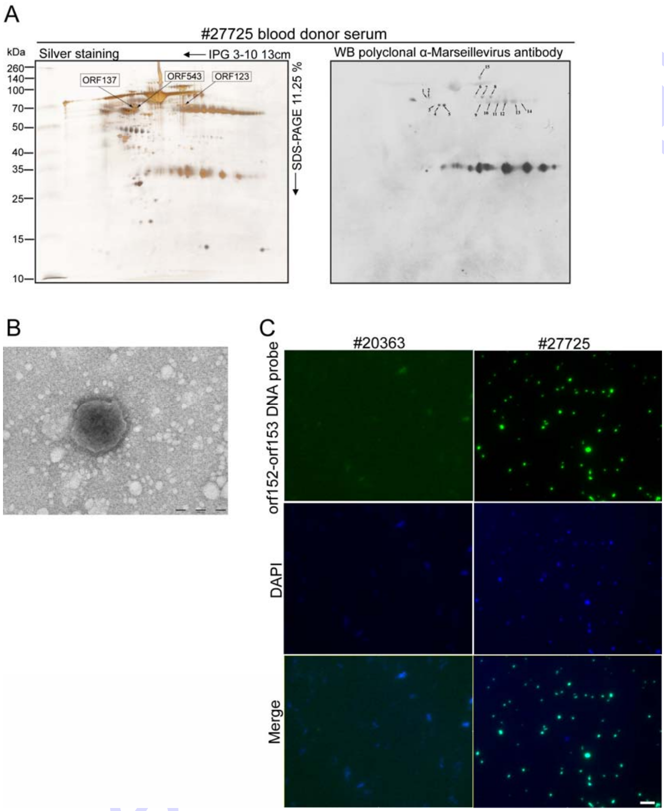
GBM
357,433 bp
350000
0
50000
100000
150000
200000
250000
300000

B

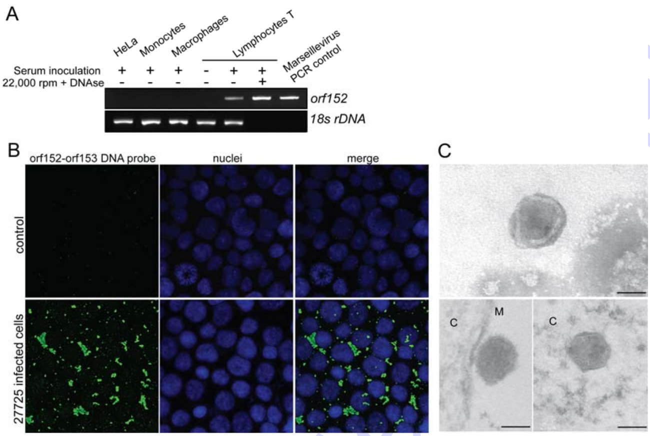


TnA-2c
HvA-3e
SfAV-1a
Lausannevirus
Marseillevirus
Senegalvirus
GBM
IIV6
IIV3
ISKNV
LDV-isolate China
LDV-1
SGI
AIV
Frog virus 3
Mimivirus
Mamavirus
OsV5
EhV86
PBCV-FR483
FsV
EsV1
ASFV
Myxoma virus
Sheeppox virus
MCV subtype 1
Variola virus
Vaccinia virus
Fowlpox virus
0.2

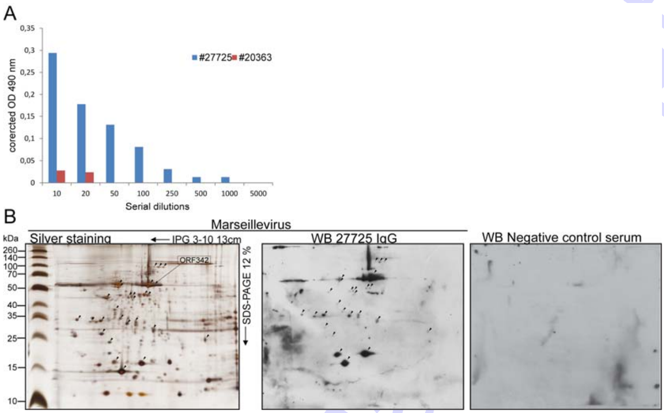
25


A
#27725 blood donor serum
kDa
Silver staining
IPG 3-10 13cm
WB polyclonal α-Marseillevirus antibody
260
140
100
70
50
40
35
25
15
10
ORF137
ORF543
ORF123
SDS-PAGE 11.25 %
B
C
#20363
#27725
orf152-orf153 DNA probe
DAPI
Merge

26


A
HeLa
Monocytes
Macrophages
Lymphocytes T
Marseillevirus PCR control
Serum inoculation
22,000 rpm + DNAse
orf152
18s rDNA
B
orf152-orf153 DNA probe
nuclei
merge
control
27725 infected cells
C
M
C
C

27



A
corercted OD 490 nm
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
#27725
#20363
10
20
50
100
250
500
1000
5000
Serial dilutions
B
Marseillevirus
Silver staining
IPG 3-10 13cm
ORF342
SDS-PAGE 12 %
WB 27725 IgG
WB Negative control serum
kDa
260
140
100
70
50
40
35
25
15
10

Appendix B. Article 6. Evidence of the megavirome in humans

Evidence of the megavirome in humans.

Philippe Colson[1,2], Laura Fancello[1], Gregory Gimenez[1], Fabrice Armougom[1], Christelle Desnues[1], Ghislain Fournous[1], Niyaz Yoosuf[1], Matthieu Million[1], Bernard La Scola[1,2], Didier Raoult[1,2,*]

Published in Journal of Clinical Virology 57 (2013) 191–200.

[1] Aix-Marseille Université, Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE) UM 63 CNRS 7278 IRD 198 INSERM U1095, Facultés de Médecine et de Pharmacie, Marseille, France.

[2] IHU Méditerranée Infection, Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone, Assistance Publique – Hôpitaux de Marseille, Marseille, France.

[*] Corresponding author. E-mail: didier.raoult@gmail.com

Journal of Clinical Virology 57 (2013) 191–200

Contents lists available at SciVerse ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Evidence of the megavirome in humans

Philippe Colson[a,b], Laura Fancello[a], Gregory Gimenez[a], Fabrice Armougom[a], Christelle Desnues[a], Ghislain Fournous[a], Niyaz Yoosuf[a], Matthieu Million[a], Bernard La Scola[a,b], Didier Raoult[a,b,*]

[a] Aix-Marseille Univ., Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE) UM 63 CNRS 7278 IRD 198 INSERM U1095, Facultés de Médecine et de Pharmacie, Marseille, France

[b] IHU Méditerranée Infection, Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone, Assistance Publique – Hôpitaux de Marseille, Marseille, France

ARTICLE INFO

Article history:
Received 7 January 2013
Received in revised form 14 March 2013
Accepted 29 March 2013

Keywords:
Mimivirus
Giant virus
Marseillevirus
Megavirales
Metagenomics
Humans
Next-generation sequencing
Amoeba
Infectious diseases

ABSTRACT

Background: Megavirales is a proposed new virus order composed of Mimivirus, Marseillevirus and closely related viruses, as well as members of the families *Poxviridae*, *Iridoviridae*, *Ascoviridae*, *Phycodnaviridae* and *Asfarviridae*. The *Megavirales* virome, which we refer to as the megavirome, has been largely neglected until now because of the use of technical procedures that have jeopardized the discovery of giant viruses, particularly the use of filters with pore sizes in the 0.2–0.45-μm range. Concurrently, there has been accumulating evidence supporting the role of Mimivirus, discovered while investigating a pneumonia outbreak using amoebal coculture, as a causative agent in pneumonia.

Objectives: In this paper, we describe the detection of sequences related to Mimivirus and Marseillevirus in the gut microbiota from a young Senegalese man. We also searched for sequences related to *Megavirales* in human metagenomes publicly available in sequence databases.

Results: We serendipitously detected Mimivirus- and Marseillevirus-like sequences while using a new metagenomic approach targeting bacterial DNA that subsequently led to the isolation of a new member of the family *Marseilleviridae*, named Senegalvirus, from human stools. This discovery demonstrates the possibility of the presence of giant viruses of amoebae in humans. In addition, we detected sequences related to *Megavirales* members in several human metagenomes, which adds to previous findings by several groups.

Conclusions: Overall, we present convergent evidence of the presence of mimiviruses and marseilleviruses in humans. Our findings suggest that we should re-evaluate the human megavirome and investigate the prevalence, diversity and potential pathogenicity of giant viruses in humans.

1. Background

The story of giant viruses that infect phagocytic protists began in 1992 in England during the investigation of a pneumonia outbreak that led to isolate obligate intra-amoebal microorganisms in water of a cooling tower.[1–3] Then, in 2003, Mimivirus was discovered as part of this collection of intra-amoebal parasites.[1–4] Subsequently, Mamavirus, Marseillevirus and other giant viruses infecting phagocytic protists have been discovered; all of these viruses have been linked to nucleocytoplasmic large DNA viruses (NCLDVs), a monophyletic class of viruses composed of *Poxviridae*, *Iridoviridae*, *Ascoviridae*, *Phycodnaviridae* and *Asfarviridae*[5–11] for which we recently proposed the reclassification into a new viral order, the *Megavirales*.[12]

The question of pathogenicity of mimiviruses initially focused on the capability of Mimivirus to cause pneumonia due to the setting of its initial discovery as well as the involvement of some water-associated amoebae-resistant bacteria, including *Legionella pneumophila*, in such infections.[1,13,14] Experimentally, Mimivirus was found to be capable of inducing pneumonia in mice[15] and infecting macrophages through phagocytosis.[16] In humans, serological studies showed seroconversion to Mimivirus in several patients presenting with pneumonia.[13,18] Antibodies to Mimivirus were associated with pneumonia, re-hospitalization after discharge[13] and a poorer outcome in mechanically ventilated pneumonia patients (Table 1).[19] In contrast, several studies have failed to isolate Mimivirus from patients with pneumonia and Mimivirus DNA testing was positive in only a single patient.[13,20–23,25] However, DNA detection may have been hampered in these studies

* Corresponding author at: Aix-Marseille Université, Unité des Rickettsies, URMITE UM3 CNRS 7278 IRD 198 INSERM U1095, Faculté de Médecine, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05, France. Tel.: +33 491 324 375; fax: +33 491 387 772.

E-mail address: didier.raoult@gmail.com (D. Raoult).

1386-6532/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jcv.2013.03.018

Table 1
Summary of the clinical, microbiological and metagenomic data on mimivirus and pneumonia.

Reference	Year(s) of sample collection	Country	Population size	Subgroup size	Main characteristics of patients	Respiratory sample type	PCR	Serology
13	1985–1997	Canada	887	121	Adults; ambulatory/com-acquired pneumonia patients, Nova Scotia	N.t.	N.t.	36 positive (9.7%)
				255	Adults; patients hospitalized with com-acquired pneumonia, multiple centers across Canada			
				511	Adults; healthy control subjects, Nova Scotia	N.t.	N.t.	12 positive (2.3%)
	2003–2004	France	129	32	Adults; patients with ICU-acquired pneumonia, one-year survey	BAL	1 positive sample of 32 (3.1%)	5 positive samples of 26 (19.2%)
				50	Adults; controls (patients tested for anti-*Rickettsia* spp. antibodies)	N.t.	N.t.	All negative
				21	Adults; intubated control patients in the ICU who did not present with pneumonia	BAL	All negative	N.t.
17	2003	France	157		Adults; ICU pneumonia patients (pneumonia was com-acquired or ventilator-associated)	N.t.	N.t.	7 cases with a high level of evidence and 7 additional cases with a low level of evidence
18	2004	France	1		38-year-old laboratory technician	N.t.	N.t.	Positive
19	2006–2008	France	300		Adults; ventilated patients in the ICU with a suspicion of a ventilator-associated pneumonia and positive serology for Mimivirus (cases)	N.t.	N.t.	59 positive (19.7%)
20	2005–2006	Austria	214		Children hospitalized for respiratory tract infections; 209 were non-immunocompromised; six-month survey during the fall and winter seasons	NP aspirate samples	All negative	N.t.
21	2000–2001	Urban USA	496	124	Children < 5 y; com-acquired pneumonia cases	Nasal swabs	All negative	N.t.
	2003–2004	Rural Thailand		120	Adults, children; com-acquired pneumonia cases	NP swabs	All negative	N.t.
	2002–2004	USA		71	Geriatric; nosocomially acquired pneumonia outbreak, retirement centers	Nasal or NP swabs	All negative	N.t.
		USA		5	Adults, children; com-acquired pneumonia outbreak (familial cluster)	Lower respiratory samples	All negative	N.t.
	2001–2003	USA		87	Adults; bone marrow transplant recipients	NP aspirates, nasal wash, or NP swabs	All negative	N.t.
	2002–2003	Canada		89	Adults; lung transplant recipients	NP swabs	All negative	N.t.
22	–	Urban Italy	63 (69 BAL) specimens	30 BAL	Patients receiving mechanical ventilation for at least 48 h	BAL	All negative	N.t.
				39 BAL	Non-ventilated patients from different clinical settings, including lung transplant recipients	BAL	All negative	N.t.
23	2003–2004	Queensland, Australia	315 (477 specimen)		People with suspected acute respiratory tract infections; the subjects ranged in age from 1 day to 80.3 years (mean = 7.7 years); 79% were ≤5 years of age; an additional 81 consecutively collected summer specimens formed a secondary population	Predominantly (92.4%) nasopharyngeal aspirates	All negative	N.t.
24	2004–2005	Stockholm, Sweden	210		Patients with respiratory tract infections	NP aspirates	Metagenomics identified 2 reads matching mimivirus	–
25	2009–2010	The Netherlands	109 (220 sputum samples)		COPD patients during stable conditions and during exacerbations of COPD, referred for pulmonary rehabilitation 115 sputum samples were collected from 84 patients during the stable phase, and 105 samples were collected from 74 patients during an acute exacerbation Mimivirus serology was performed for 118 serum samples, 88 collected during the stable phase and 30 during an exacerbation	Sputum	All negative	3 positive (2 during an acute exacerbation; 1 during the stable phase)

BAL, Bronchoalveolar lavage; Com., Community; COPD, chronic obstructive pulmonary disease; NP, naso-pharyngeal.

P. Colson et al. / Journal of Clinical Virology 57 (2013) 191–200

Table 2
Studies that identified sequences closely related to those of *Megavirales* in human and animal samples.

Reference	Mean	Enrichment in viruses	Sample	Name of virus(es) (Number of hits)	City, continent, country
41	Shotgun library (with strand displacement polymerase amplification), standard sequencing	Cesium chloride gradient	Human blood from 20 healthy donors	Cowpox virus (2)	San Diego, California, USA
42	Random primer amplification, standard sequencing	Filtration through 0.45-µm filters	Fecal samples from 12 distinct pediatric patients suffering from acute diarrhea	Mimivirus (5)	Melbourne, Australia; Seattle, USA
43	454 pyrosequencing	Cesium chloride gradient for sewage (not described for serum)	Sewage and human sera from 199 healthy volunteers and patients with acute febrile illness	*Asfarviridae* (6)	Serum: Middle East; Sewage: Barcelona, Spain
44	454 pyrosequencing	Filtration through 0.22- or 0.45-µm filters	Serum samples from 45 pairs of monozygotic twins affected and unaffected with chronic fatigue illness	*Asfarviridae* (1), *Iridoviridae* (3), *Mimiviridae* (2), *Phycodnaviridae* (2), *Poxviridae* (5)	Sweden
45	Illumina	None described	Human sera from 123 Nicaraguan patients presenting with dengue-like symptoms (but testing negative for dengue virus)	*Asfarviridae*	Nicaragua
24	454 pyrosequencing	Filtration through 0.22- and 0.45-µm filters	210 nasopharyngeal aspirate samples from patients with respiratory tract infection	Mimivirus (2), *Paramecium bursaria* Chlorella virus A1 (2), African swine fever virus (2)	Stockholm, Sweden
46	454 pyrosequencing or Illumina sequencing	None described	50 nasopharyngeal swabs and 23 plasma samples from children under 3 years of age with unexplained fever	*Asfarviridae*	Washington DC, USA
47	Retrospective study of large metagenomic datasets	–	Human gut	Virophages (65)	–
Present study	454 pyrosequencing	None	Human stools (nondiarrheic)	Marseillevirus (9), Mimivirus (44)	Senegal
48	454 pyrosequencing	Tangential flow filtration, ultrafiltration and cesium chloride gradient	Stools from 5 pigs	*Mimiviridae* (0.11% of reads), *Poxviridae* (0.52%), *Iridoviridae* (0.06%), *Phycodnaviridae* (0.58%)	Berlin, Germany
49	454 pyrosequencing	Filtration through 0.45-µm filters and cesium chloride gradient	Three mosquitoes	*Poxviridae*	San Diego, California, USA
50	454 pyrosequencing (transcriptomics)	None described	Gypsy moth (*Lymantria dispar*)-derived IPLB-Ld652Y cell line	Mimivirus, *Cafeteria roenbergensis* virus BV-PW1 *Poxviridae*,	Baltimore, USA

because the PCR primers used targeted only the Mimivirus genome, whereas currently at least 18 close relatives of Mimivirus that exhibit considerable genetic diversity have been described.[12,26,27] Besides, positive serology for the virophage of mimiviruses, initially described in 2008,[6] was recently observed in two people who experienced fever while returning from Laos.[28] In addition, a mimivirus named Lentillevirus has been isolated from contact lens storage case liquid.[29,30] During the past decade, mimiviruses and marseilleviruses have been isolated from freshwater, seawater, and soil samples in six countries located on three continents (http://maps.google.fr/maps/ms?vps=2&hl=fr&i.e.=UTF8&oe=UTF8&msa=0&msid=200914559094835369589.0004beba4af112f60dcf2), and isolation rates reached ≈20% from water samples,[31] suggesting common exposure to these viruses of humans. In addition, the currently identified hosts of these viruses are widespread in water and soil,[14] and Mimivirus-like particles have been observed using light microscopy within Acanthamoeba species in treated sewage sludge from a wastewater treatment plant in the UK.[32] Concurrently, searches for *Megavirales* sequences in multiple environmental metagenomes enabled the identification of sequences similar to those of Mimivirus[33–39] and other members of the families *Asfarviridae*, *Poxviridae*, *Phycodnaviridae* and *Iridoviridae*.[36–39] Furthermore, sequences related to *Megavirales* members as well as virophages described to infect mimiviruses[40] have also been retrieved from human and animal metagenomes (Table 2).[24,41–50]

2. Objectives

In this paper, we describe the detection of sequences related to Mimivirus and Marseillevirus in the gut microbiota from a young Senegalese man. We also searched for sequences related to *Megavirales* in human metagenomes publicly available in sequence databases.

3. Study design

3.1. Investigation of the gut microbiota from a young Senegalese man

A previous study conducted in our laboratory consisted of ultra-deep sequencing of bacterial 16S ribosomal DNA (rDNA) in the stools of a healthy 20-year-old man living in rural Senegal.[51] A new method to avoid PCR amplification bias prior to sequencing was used in addition to the classical method based on PCR amplification of the V6 region of 16S rDNA with universal primers 917F and 1391R.[51] The new sequencing method consisted of complete enzymatic digestion of the fecal sample DNA with *Eco*O190I and *Brs*GI enzymes that are able to cleave sites inside primers 917F and 1391R and are therefore able to generate fragments corresponding to the 16S rDNA V6 region (see supplementary methods). Sequencing of the products from

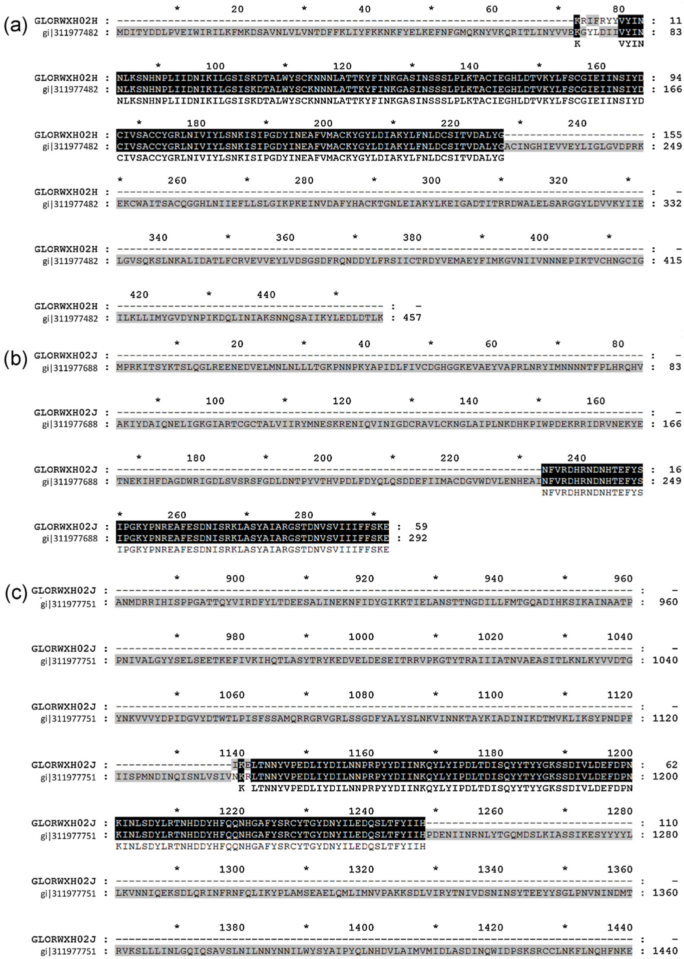


Fig. 1. Alignments of amino acid sequences of Mimivirus (a putative ankyrin repeat protein (gi|311977482) (a), a putative protein phosphatase 2C (gi|311977688) (b), and a putative ATP-dependent RNA helicase gi|311977751) (c)) and three different metagenomic reads obtained from the feces of a young Senegalese. The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

both the new and classical procedures was performed using the 454 FLX Titanium instrument (Roche, USA).[51] Sequencing products were trimmed and analyzed by BLAST searches[52] and with the QIIME pipeline.[53] Among the reads generated by enzymatic digestion, 99.9% were unrelated to bacterial 16S rDNA. These results led us to search for Mimivirus and Marseillevirus-related sequences among the initially rejected reads. The reads were mapped onto the Mimivirus and Marseillevirus genomes with CLC bio software (http://www.clcbio.com/index.php?id=479) using default parameters (50% minimum coverage, 80% minimum similarity) and tBLASTn searches were performed with the giant viruses against the reads (e-value threshold $1e{-}6$). Senegalvirus isolation, sequencing and assembly have been previously described.[51] Senegalvirus genome annotation is described in the supplementary methods.

3.2. Searches for Megavirales-related sequences in metagenomes recovered from human samples

Reads annotation was automatically performed from the metagenomics RAST (MG-RAST) server (http://metagenomics.anl.gov/) for viral metagenomes from eleven published studies and obtained from human samples, including stools,[54–59] nasopharyngeal aspirates,[24,60] saliva,[61] oropharyngeal swabs,[62] and sputum[63] (supplementary Table S1) against the NCBI GenBank protein sequence database (*e*-value threshold, 1*e*−5). Only hits with alignment lengths ≥40 amino acids were considered. In addition, metagenomes from 9 of the previous studies, and from two other studies that analyzed lung samples,[64–65] were downloaded and preprocessed using several tools including Prinseq[66] for removal of duplicate reads as well as low quality and low complexity reads (supplementary Fig. S1). The remaining sequences were annotated with an in-house strategy using BLASTn searches[52] against all genomes of the *Megavirales* members and virophages available in the NCBI GenBank sequence database as well as those not yet released, available in our laboratory (supplementary Table S2). We considered only matches with ≥30% coverage and ≥90% identity and identical coverage and identity for the corresponding reads through BLAST searches against the NCBI GenBank nucleotide sequence database. BLASTn searches were also performed for all genomes of members of the order *Megavirales* and virophages (supplementary Table S2) against four human gut bacterial metagenomes (supplementary Table S3) through the CAMERA portal (http://camera.calit2.net/), using default parameters. Metagenome reads identified as matching with *Megavirales* sequences were extracted and manually tested against the NCBI nonredundant protein sequence database (nr) using BLASTx (*e*-value threshold, 1*e*−05) to determine whether a *Megavirales* sequence was among the best matches.

Finally, amino acid BLAST (BLASTp) searches were performed for the published proteomes of mimiviruses (Mimivirus, Mamavirus, *Cafeteria roenbergensis* virus (Crov; isolated from *Cafeteria roenbergensis*, a widespread marine unicellular flagellate), *Megavirus chilensis*), marseilleviruses (Marseillevirus, Lausannevirus) and the Sputnik virophage against annotated bacterial metagenomes recovered from eleven different body sites ((Supplementary Table S3; Table 5) as part of the human microbiome project (http://www.hmpdacc.org/HMGI/). Hits were considered significant based on the *e*-value threshold of 1*e*−04 and amino acid identity and coverage above 30% and 70%, respectively.

4. Results

4.1. Serendipitous identification of Mimivirus- and Marseillevirus-related sequences in the stool metagenome of a young Senegalese male

Among the metagenomic reads recovered from the feces of a young Senegalese male,[51] 44 and 9 (supplementary Table S4) could be mapped to the Mimivirus and Marseillevirus genomes, respectively, and 12 reads > 300 bp could be mapped to Mimivirus DNA with >90% identity and coverage (supplementary Table S4; Table 3; Fig. 1a–c). In addition, tBLASTn searches with the Mimivirus and Marseillevirus genomes against the stool metagenome yielded 2306 and 259 hits, respectively. These findings prompted us to inoculate one gram of the young Senegalese stools on *Acanthamoeba polyphaga*,

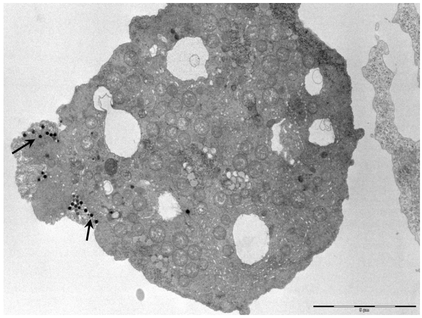


Fig. 2. Electron microscopy image of Senegalvirus in *Acanthamoeba polyphaga*. The scale bar represents 5 μm.

as previously reported,[51] with the purpose of confirming the presence of the giant virus from amoeba in the human feces. Indeed, amoebal culture enabled the isolation of a new marseillevirus, named Senegalvirus (Fig. 2), and the sequencing of its genome.[51] The Senegalvirus double-stranded DNA genome is ≈372,690 base pairs (bp) in length, currently making this genome the largest among marseilleviruses; it is ≈4200 bp larger than that of Marseillevirus and ≈26,000 bp larger than that of Lausannevirus. Comparison of the 479 protein sequences of Senegalvirus predicted using GeneMarkS[67] to those of Marseillevirus or Lausannevirus by all-against-all BLASTp searches yielded bidirectional best hits for 351 and 253 Marseillevirus and Lausannevirus proteins, respectively. Thus, overall, 384 Senegalvirus proteins could be considered *bona fide* orthologs to Marseillevirus or Lausannevirus proteins, because these Senegalvirus proteins are involved in pairs of bidirectional best hits with predicted proteins of Marseillevirus and/or Lausannevirus. The mean (±standard deviation (SD)) amino acid identity between Senegalvirus and Marseillevirus protein in pairs was 97 ± 7%, whereas the mean identity for Senegalvirus and Lausannevirus protein pairs was 59 ± 16%. We detected Senegalvirus orthologs to three histone-like proteins first described in Marseillevirus, as well as proteins containing bacterial-like membrane occupation and recognition nexus (MORN) repeat domains (proteins described as mediating membrane-membrane or membrane-cytoskeleton interactions[5]) and serine/protein kinases, including a unique kinase shared by marseilleviruses, iridoviruses and ascoviruses.[5,8] Homology for the Senegalvirus proteins was greater with their Marseillevirus counterparts than their Lausannevirus counterparts. Congruently with comparative genomics, phylogeny reconstruction based on the family-B DNA polymerase showed that Senegalvirus was clustered with Marseillevirus within the family *Marseilleviridae* (Fig. 3).

4.2. Blast searches for Megavirales-like sequences in metagenomes

A few reads from human stools and oropharyngeal viromes[58,61,62] available on the MG-RAST server were found to match Mimivirus sequences. They were predicted to encode a collagen-like protein 6 (MIMI_L668), an uncharacterized HNH endonuclease (MIMI_L245) and a hypothetical protein (MIMI_R892) (Supplementary Table S6). Although a BLASTx search using these metagenomic reads against the NCBI GenBank non-redundant protein sequence database did not find *Acanthamoeba polyphaga*

Table 3
Description of reads longer than 300 nucleotides that map to Mimivirus sequences.

Read length (nt)	Best matches	BLASTn (nucleotide)		BLASTp (amino acid)	
		Eval	Identities	*e*-Val	Identities
504	YP_003986602.1 (L112): putative ankyrin repeat protein	0.0	504/505 (99%)	8*e*−154	148/148 (100%)
403	YP_003986629.1 (L137): Hypothetical proteins	0.0	399/405 (99%)	2*e*−99	82/83 (99%)
334	YP_003986695 (L199): Hypothetical proteins	9*e*−157	329/339 (97%)	6*e*−82	94/119 (79%)
361	YP_003986808 (R306/R307): putative protein phosphatase 2C	0.0	361/363 (99%	2*e*−50	59/59 (100%)
350	YP_003986871.1 (R366): putative ATP-dependent RNA helicase	4*e*−180	349/350 (99%)	3*e*−109	107/107 (100%)
475*	AEJ34636.1 (R398): hypothetical protein	0.0	467/476 (98%)	6*e*−49	44/44 (100%)
397*	YP_003986902.1 (R398): putative phosphoesterase	0.0	393/398 (99%)	9*e*−37	32/32 (100%)
447	YP_003987107.1 (R592): putative helicase	0.0	442/447 (99%)	3*e*−90	88/148 (59%)
403	YP_003987195.1 (L673): putative serine/threonine-protein kinase	0.0	400/403 (99%)	1*e*−100	112/144 (78%)
331	YP_003987287.1 (R757): putative F-box protein	2*e*−165	329/332 (99%)	2*e*−52	73/73 (100%)
376	YP_003987388.1 (R857): hypothetical protein	0.0	374/376 (99%)	9*e*−41	49/49 (100%)
406	YP_003987407.1 (L872): hypothetical protein	0.0	406/407 (99%)	2*e*−131	128/128 (100%)

* Two overlapping reads; nt, nucleotide.

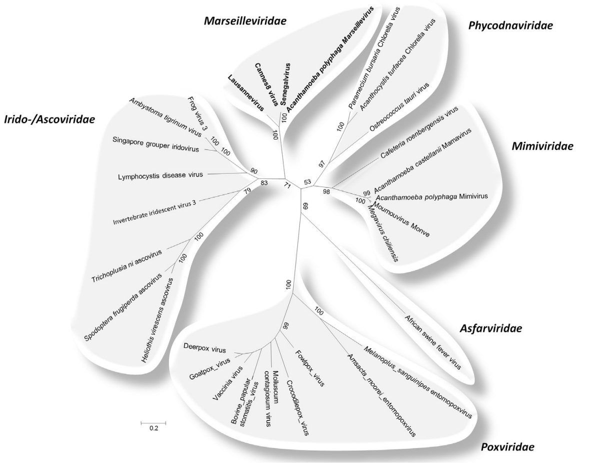


Fig. 3. Phylogeny reconstruction based on an alignment (generated by Muscle (http://www.ebi.ac.uk/Tools/msa/muscle/)) of DNA polymerase of marseilleviruses and other *Megavirales* members, using the Maximum Likelihood method with the Mega 5 software (http://www.megasoftware.net/). Probabilities are shown near branches as a percentage and are used as confidence values of tree branches. Scale bar represents the number of estimated changes per position for a unit of branch length.

mimivirus as the top hit, mimivirus proteins were among the best hits, with *e*-values ranging from 1*e*−6 to 1*e*−14 and amino acid identities ranging from 28 to 58%. In addition, when searching using tBLASTn with the three mimivirus proteins against 55 human metagenomes with the NCBI genomic BLAST tool (http://www.ncbi.nlm.nih.gov/sutils/blast_table.cgi?taxid=Environmental&taxidinf=environ_info&selectall), significant matches, with *e*-values ranging from 8*e*−24 to 3*e*−25 and amino acid identities

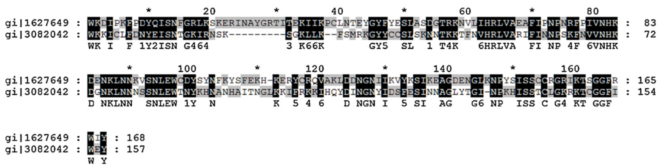


Fig. 4. Amino acid alignments for translated sequences of metagenomic reads (GenBank Accession number BAAZ01000542.1 (gi|162764931)) recovered from a human gut metagenome and the uncharacterized Mimivirus HNH endonuclease (accession no. ADO18080.1 (gi|308204279)). The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

Table 4
Number of BLASTn hits obtained for 4 gut metagenomes analyzed through the CAMERA portal.

		CAM_PROJ_Human Distal Gut Human distal gut biome project[69]	CAM_PROJ_Human Gut 13 healthy human gut metagenomes[68]	CAM_PROJ_Human Gut Diagnosis Metagenomic diagnosis of bacterial infections [70]	CAM_PROJ_Twin Study A core gut microbiome in obese and lean twins [71]
Mimiviridae	Group I – lineage A	10	40	0	18
Mimiviridae	Group I – lineage B	32	15	0	105
Mimiviridae	Group I – lineage C	4	20	0	50
Mimiviridae	Group II	0	2	0	3
Mimiviridae	Any	46	77	0	176
Phycodnaviridae		9	29	2	36
Poxviridae		8	1	0	0
Total		63	107	2	212

ranging from 36 to 41%, were found for the uncharacterized HNH endonuclease (Fig. 4) and the putative oxidoreductase against sequences recovered in fecal samples from healthy individuals in Japan.[68]

Additional searches of viral metagenomes from 11 studies (supplementary Table S1) using cleaning and trimming of reads and comparative BLASTn searches against our database of *Megavirales* members and virophages and the NCBI sequence database enabled detection of two reads that displayed significant hits with genomes of mimiviruses (Supplementary Figs. S2–S12). In particular, a 130-nucleotide-long read (no. SRR101483.9794578) recovered from a metagenomic dataset corresponding to pulmonary microbiota from patients with acute exacerbation of idiopathic pulmonary fibrosis[65] (Supplementary Fig. S12) found a fragment of a putative bifunctional dihydrofolate reductase/thymidylate synthase (GenBank Accession no. YP_003970135.1) of Crov as the best match, through BLASTn, BLASTx as well as tBLASTx searches against the NCBI sequence databases. Nucleotides 55–130 of this metagenomic read matched amino acids 214–237 of the 452 amino acid-long putative Crov protein (Fig. 5). In addition, 384 hits, including 299 for mimiviruses, 76 for phycodnaviruses and 9 for poxviruses, were found by BLASTn searches against four human gut bacterial metagenomes through the CAMERA portal (Table 4; Supplementary Table S3). The top hits obtained for the metagenomes were from mimiviruses (Pointe-Rouge2 virus, Moumouvirus and Ochan virus)[31] in three cases and *Paramecium bursaria chlorella* virus 1 in the remaining case (Fig. 6a–d); the corresponding metagenomic reads did not find a member of the *Megavirales* among the best matches through BLAST searches against the NCBI sequence databases. Finally, BLASTp searches with the proteomes of mimiviruses and marseilleviruses against microbial metagenomes from 11 body sites from the human microbiome project showed from 2 to 54 significant hits per virus (Table 5), including for instance a putative metalloendopeptidase protein (Crov ORF_67) and an asparaginyl-tRNA synthetase (*Megavirus chilensis* ORF_743) in a saliva metagenome, and a 70-kDa heat-shock protein and a DNA Topoisomerase IA of Mimivirus (ORF_393, ORF_221) in a vagina metagenome. Nonetheless, BLASTp searches using the corresponding annotated proteins from the metagenomes against the NCBI GenBank non-redundant protein sequence database did not find mimiviruses or marseilleviruses proteins as best hits.

Overall, *Megavirales*-related sequences were recovered through various strategies from a variety of samples such as saliva (2), oropharynx (1), lung (1) and stools (&).

Fig. 5. Amino acid alignments for translated sequences of a metagenomic read recovered from human lung[65] and the putative bifunctional dihydrofolate reductase/thymidylate synthase (Cafeteria roenbergensis virus BV-PW (Crov; accession no. YP_003970135.1)). The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

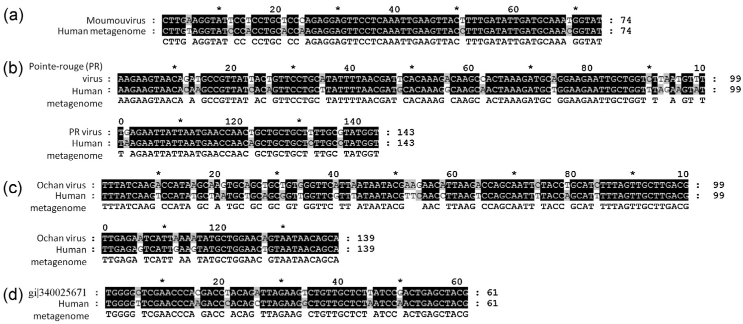


Fig. 6. Nucleotide alignments for metagenomic reads recovered from human gut and the DNA genome of Moumouvirus (5a), Pointe-Rouge2 virus (5b), Ochan virus (5c), and *Paramecium bursaria* Chlorella virus 1 (gi|340025671) (5d). The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

Table 5
Significant hits (number) for proteomes of *Megavirales* members against protein-encoding sequences from various human metagenomes of the human microbiome project (http://www.hmpdacc.org/HMGI/).

Megavirales		Human body sites (number of metagenomes)										
		Anterior nares (87)	Throat (7)	Palatine tonsils (6)	Subgingival plague (7)	Attached keratinized gingiva (6)	Saliva (3)	Right retroauricular crease (17)	Left retroauricular crease (9)	Posterior fornix (51)	Vaginal Introitus (3)	Mid vagina (2)
Mimiviridae												
Group I – lineage A	*Acanthamoeba polyphaga mimivirus*	26	32	31	33	25	11	42	34	16	11	11
	Acanthamoeba castellanii mamavirus	27	37	36	36	28	14	49	42	17	12	11
Group I – lineage B	Moumouvirus	32	36	38	40	41	21	54	45	25	14	13
Group I – lineage C	*Megavirus chilensis*	30	30	32	33	29	13	45	40	19	10	10
Group II	*Cafeteria roenbergensis* virus	20	22	24	22	21	11	35	22	13	9	10
Marseilleviridae	Marseillevirus	11	6	5	8	6	5	12	12	5	2	2
	Lausannevirus	8	7	6	8	5	4	12	14	4	4	4

5. Conclusions

In the present study, we showed association of *Megavirales* members with humans using different strategies and samples. Our attention was drawn to the presence of giant viruses in a patient stool sample through the serendipitous detection of Mimivirus- and Marseillevirus-like sequences while using a new metagenomic approach targeting bacterial DNA. Subsequently, Senegalvirus, a new member of the family *Marseilleviridae*, was isolated from this stool sample demonstrating the possibility of the presence of giant viruses in humans. In addition, we detected sequences matching DNA of *Megavirales* members in several human metagenomes, which adds to previous findings in human nasopharyngeal, fecal and blood samples by other laboratories (Table 2).[24,41–47]

Members of the order *Megavirales* represent a technical problem in the current investigation of the virome due to their size. Indeed, viruses are still usually considered small agents,[1,12] which leads the vast majority of research groups to perform viral purification by filtering samples through small-pore (0.2–0.45 μm) filters prior to metagenomic analysis, thus preventing the detection of viruses larger than the filter pores.[72–74] This biased technical approach has most likely led to considerable underestimation of the prevalence of *Megavirales* members in environmental and human samples. In addition, the present work underscores that the detection of giant viruses in humans may benefit from the concurrent use of culture and metagenomics. Accordingly, dramatic differences between the set of bacteria isolated by means of a large panel of culture conditions, the so-called culturomics approach, and the set of bacteria identified through metagenomics were recently unveiled.[51] The Senegalvirus discovery highlights that a virus may be cultured but not molecularly detected. The isolation of Senegalvirus represented the first isolation of a marseillevirus from a human sample.[51] In another report, we also described the isolation of a, Lentillevirus, from contact lens liquid.[29,30] More importantly, we have recently isolated a mimivirus, LBA111, by amoebal culture from a bronchoalveolar sample collected in a Tunisian woman presenting with pneumonia.[75] In addition, sequences related to Marseillevirus DNA and the genome of a new giant virus, named Giant Blood Marseille-like virus, were recovered from the blood of asymptomatic blood donors using high-throughput sequencing.[76]. Previous studies have also shown the presence of poxvirus- and asfarvirus-related sequences in human blood from apparently healthy persons (Table 2),[41,43] raising questions about the asymptomatic carriage of giant viruses and their role over the short and long term.

Taken together, present and previous data provide convergent evidence for the presence of mimiviruses and marseilleviruses in humans, which raises further questions about their potential pathogenicity. We recommend discarding technical procedures that are too stringent and may lead to the neglect of the study of the 'megavirome' while investigating the human virome.

Funding

Researches of LF and CD are financed through a starting grant number 242729 from the European Research Council.

Competing interests

None for all authors.

Ethical approval

None required.

Acknowledgements

None.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jcv.2013.03.018.

References

1. Raoult D, La Scola B, Birtles R. The discovery and characterization of Mimivirus, the largest known virus and putative pneumonia agent. *Clin Infect Dis* 2007;**45**(July (1)):95–102.
2. Raoult D. Giant viruses from amoeba in a post-Darwinist viral world. *Intervirology* 2010;**53**(5):251–3.
3. La Scola B, Audic S, Robert C, Jungang L, de L, Drancourt XM, et al. A giant virus in amoebae. *Science* 2003;**299**(March (5615)):2033.
4. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1. 2-megabase genome sequence of Mimivirus. *Science* 2004;**306**(November (5700)):1344–50.
5. Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA* 2009;**106**(December (51)):21848–53.
6. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, et al. The virophage as a unique parasite of the giant mimivirus. *Nature* 2008;**455**(September (7209)):100–4.
7. Fischer MG, Allen MJ, Wilson WH, Suttle CA. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA* 2010;**107**(November (45)):19508–13.
8. Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, et al. Lausannevirus, a giant amoebal virus encoding histone doublets. *Environ Microbiol* 2011;**13**(June (6)):1454–66.
9. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci USA* 2011;**108**(October (42)):17486–91.
10. Iyer LM, Balaji S, Koonin EV, Aravind L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* 2006;**117**(April (1)):156–84.
11. Iyer LM, Aravind L, Koonin EV. Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* 2001;**75**(December (23)):11720–34.
12. Colson P, de L, Fournous X, Raoult GD. Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology* 2012;**55**(April (5)):321–32.
13. La Scola B, Marrie TJ, Auffray JP, Raoult D. Mimivirus in pneumonia patients. *Emerg Infect Dis* 2005;**11**(March (3)):449–52.
14. Greub G, Raoult D. Microorganisms resistant to free-living amoebae. *Clin Microbiol Rev* 2004;**17**(April (2)):413–33.
15. Khan M, La Scola B, Lepidi H, Raoult D. Pneumonia in mice inoculated experimentally with Acanthamoeba polyphaga mimivirus. *Microb Pathog* 2007;**42**(February (2/3)):56–61.
16. Ghigo E, Kartenbeck J, Lien P, Pelkmans L, Capo C, Mege JL, et al. Ameobal pathogen mimivirus infects macrophages through phagocytosis. *PLoS Pathog* 2008;**4**(June (6)):e1000087.
17. Berger P, Papazian L, Drancourt M, La Scola B, Auffray JP, Raoult D. Ameba-associated microorganisms and diagnosis of nosocomial pneumonia. *Emerg Infect Dis* 2006;**12**(February (2)):248–55.
18. Raoult D, Renesto P, Brouqui P. Laboratory infection of a technician by mimivirus. *Ann Intern Med* 2006;**144**(May (9)):702–3.
19. Vincent A, La Scola B, Forel JM, Pauly V, Raoult D, Papazian L. Clinical significance of a positive serology for mimivirus in patients presenting a suspicion of ventilator-associated pneumonia. *Crit Care Med* 2009;**37**(January (1)):111–8.
20. Larcher C, Jeller V, Fischer H, Huemer HP. Prevalence of respiratory viruses, including newly identified viruses, in hospitalised children in Austria. *Eur J Clin Microbiol Infect Dis* 2006;**25**(November (11)):681–6.
21. Dare RK, Chittaganpitch M, Erdman DD. Screening pneumonia patients for mimivirus. *Emerg Infect Dis* 2008;**14**(March (3)):465–7.
22. Costa C, Bergallo M, Astegiano S, Terlizzi ME, Sidoti F, Solidoro P, et al. Detection of Mimivirus in bronchoalveolar lavage of ventilated and nonventilated patients. *Intervirology* 2011;**55**(November (4)):303–5.
23. Arden KE, McErlean P, Nissen MD, Sloots TP, Mackay IM. Frequent detection of human rhinoviruses, paramyxoviruses, coronaviruses, and bocavirus during acute respiratory tract infections. *J Med Virol* 2006;**78**(September (9)):1232–40.
24. Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, et al. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS ONE* 2012;**7**(2):e30875.
25. Vanspauwen MJ, Franssen FM, Raoult D, Wouters EF, Bruggeman CA, Linssen CF. Infections with mimivirus in patients with chronic obstructive pulmonary disease. *Respir Med* 2012;**12**(September):10.
26. Colson P, Raoult D. In: Scheld WM, Grayson ML, Hughes JM, editors. *Is Acanthamoeba polyphaga Mimivirus an emerging causative agent of pneumonia?*. Washington, DC: ASM Press; 2010.
27. Vincent A, La Scola B, Papazian L. Advances in Mimivirus pathogenicity. *Intervirology* 2010;**53**(5):304–9.
28. Parola P, Renvoisé A, Botelho-Nevers E, La Scola B, Desnues C, Raoult D. Acanthamoeba polyphaga Mimivirus virophage seroconversion in patients returning from Laos. *Emerg Infect Dis* 2012;**18**(9):1500–2.
29. Cohen G, Hoffart L, La Scola B, Raoult D, Drancourt M. Ameba-associated Keratitis, France. *Emerg Infect Dis* 2011;**17**(July (7)):1306–8.
30. Desnues C, La Scola B, Yutin N, Fournous G, Robert C, Azza S, et al. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci USA* 2012;**109**(44):18078–83.
31. La Scola B, Campocasso A, N'Dong R, Fournous G, Barrassi L, Flaudrops C, et al. Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology* 2010;**53**(5):344–53.
32. Gaze WH, Morgan G, Zhang L, Wellington EM. Mimivirus-like particles in Acanthamoebae from sewage sludge. *Emerg Infect Dis* 2011;**17**(June (6)):1127–9.
33. Ghedin E, Claverie JM. Mimivirus relatives in the Sargasso sea. *Virol J* 2005;**2**:62.
34. Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H. Marine mimivirus relatives are probably large algal viruses. *Virol J* 2008;**5**:12.
35. Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, et al. Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. *J Invertebr Pathol* 2009;**101**(July (3)):172–80.
36. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* 2010;**18**(January (1)):11–9.
37. Correa AM, Welsh RM, Vega Thurber RL. Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *ISME J* 2012;**7**(July (1)):13–27.
38. Monier A, Claverie JM, Ogata H. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* 2008;**9**(7):R106.
39. Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW, Brami D, Thiagarajan M, et al. Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* 2012;**7**(10):e42047.
40. Desnues C, Raoult D. Virophages question the existence of satellites. *Nat Rev Microbiol* 2012;**10**(February (3)):234–43.
41. Breitbart M, Rohwer F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* 2005;**39**(November (5)):729–36.
42. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D. Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog* 2008;**4**(February (2)):e1000011.
43. Loh J, Zhao G, Presti RM, Holtz LR, Finkbeiner SR, Droit L, et al. Detection of novel sequences related to african Swine Fever virus in human serum and sewage. *J Virol* 2009;**83**(December (24)):13019–25.
44. Sullivan PF, Allander T, Lysholm F, Goh S, Persson B, Jacks A, et al. An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. *BMC Microbiol* 2011;**11**(January (2)):2.
45. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis* 2012;**6**(2):e1485.
46. Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA. Sequence analysis of the human virome in febrile and afebrile children. *PLoS ONE* 2012;**7**(6):e27735.
47. Zhou J, Zhang W, Yan S, Xiao J, Zhang Y, Li B, et al. Diversity of virophages in metagenomic datasets. *J Virol* 2013;**87**(April (8)):4225–36.
48. Sachsenroder J, Twardziok S, Hammerl JA, Janczyk P, Wrede P, Hertwig S, et al. Simultaneous identification of DNA and RNA viruses present in pig faeces using process-controlled deep sequencing. *PLoS ONE* 2012;**7**(4):e34631.
49. Ng TF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C, et al. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* 2011;**6**(6):e20579.
50. Sparks ME, Gundersen-Rindal DE. The Lymantria dispar IPLB-Ld652Y cell line transcriptome comprises diverse virus-associated transcripts. *Viruses* 2011;**3**(November (11)):2339–50.
51. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, et al. Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin Microbiol Infect* 2012;**18**(December (12)):1185–93.
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;**215**(October (3)):403–10.
53. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;**7**(May (5)):335–6.
54. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, Felts B, et al. Viral diversity and dynamics in an infant gut. *Res Microbiol* 2008;**159**(June (5)):367–73.
55. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, et al. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 2003;**185**(October (20)):6220–3.
56. Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, et al. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol* 2009;**83**(May (9)):4642–51.
57. Kim MS, Park EJ, Roh SW, Bae JW. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl Environ Microbiol* 2011;**77**(November (22)):8062–70.
58. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* 2011;**21**(October (10)):1616–25.

59. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, et al. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* 2006;**4**(January (1)):e3.
60. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* 2009;**4**(1):e4219.
61. Pride DT, Salzman J, Haynes M, Rohwer F, Davis-Long C, White III RA, et al. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J* 2012;**6**(May (5)):915–26.
62. Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, Relman DA, et al. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci USA* 2011;**108**(March (Suppl. 1)):4547–53 [Epub, 2010 June].
63. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, et al. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 2009;**4**(10):e7370.
64. Willner D, Haynes MR, Furlan M, Schmieder R, Lim YW, Rainey PB, et al. Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J* 2012;**6**(February (2)):471–4.
65. Wootton SC, Kim DS, Kondoh Y, Chen E, Lee JS, Song JW, et al. Viral infection in acute exacerbation of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011;**183**(June(12)):1698–702.
66. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;**27**(March (6)):863–4.
67. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 2005;**33**(July (Web Server issue)):W451–4.
68. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 2007;**14**(August (4)):169–81.
69. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal gut microbiome. *Science* 2006;**312**(June (5778)):1355–9.
70. Nakamura S, Maeda N, Miron IM, Yoh M, Izutsu K, Kataoka C, et al. Metagenomic diagnosis of bacterial infections. *Emerg Infect Dis* 2008;**14**(November (11)):1784–6.
71. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. *Nature* 2009;**457**(January (7228)):480–4.
72. Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, et al. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* 2009;**5**(December (12)):e1000593.
73. Edwards RA, Rohwer F. Viral metagenomics. *Nat Rev Microbiol* 2005;**3**(June (6)):504–10.
74. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 2009;**4**(4): 470–83.
75. Saadi H, Pagnier I, Colson P, Kanoun Cherif J, Beji M, Boughalmi M, et al. First isolation of Mimivirus in a patient with pneumonia. *Clin Infect Dis* 2013; in press.
76. Popgeorgiev N, Boyer M, Fancello L, Monteil S, Robert C, Rivet R, et al. Giant Blood Marseillevirus recovered from asymptomatic blood donors. *J Infect Dis* 2013; in press.

Appendix C. Article 7. Bacteriophages as vehicles of the resistome in cystic fibrosis

Bacteriophages as vehicles of the resistome in cystic fibrosis.

Rolain Jean Marc [1*], Fancello Laura[1], Desnues Christelle[1] and Raoult Didier[1]

Published in: Journal of Antimicrobial Chemotherapy. 2011 Nov; 66(11):2444-7.

[1] URMITE UMR CNRS-IRD 6236, IFR48, Faculté de Médecine et de Pharmacie, Université de la Méditerranée, Marseille, France.

[*] Corresponding author. Email: jean-marc.rolain@univmed.fr

J Antimicrob Chemother 2011; **66**: 2444–2447
doi:10.1093/jac/dkr318 Advance Access publication 4 August 2011

Journal of Antimicrobial Chemotherapy

Bacteriophages as vehicles of the resistome in cystic fibrosis

Jean Marc Rolain*, Laura Fancello, Christelle Desnues and Didier Raoult

URMITE UMR CNRS-IRD 6236, IFR48, Faculté de Médecine et de Pharmacie, Université de la Méditerranée, Marseille, France

*Corresponding author. URMITE, CNRS-IRD UMR 6236, Faculté de Médecine et de Pharmacie, 27 Boulevard Jean Moulin, 13385 Marseille cedex 5, France. Tel: +33-4 91-32-43-75; Fax: +33-4-91-38-77-72; E-mail: jean-marc.rolain@univmed.fr

Environmental microbial communities and human microbiota represent a huge reservoir of mobilizable genes, the 'mobilome', including a pool of genes encoding antimicrobial resistance, the 'resistome'. Whole-genome sequencing of bacterial genomes from cystic fibrosis (CF) patients has demonstrated that bacteriophages contribute significantly to bacterial genome alterations, and metagenomic analysis of respiratory tract DNA viral communities has revealed the presence of genes encoding antimicrobial resistance in bacteriophages of CF patients. CF airways should now be considered as the site of complex microbiota, where bacteriophages are vehicles for the adaptation of bacteria to this specific environment and for the emergence and selection of multidrug-resistant bacteria with chimeric repertoires. As phages are already known to be mobilized during chronic infection of the lungs of patients with CF, it seems particularly important to improve the understanding of the mechanisms of phage induction to prevent the spread of virulence and/or antimicrobial resistance determinants within the CF population as well as in the community. Such a modern point of view may be a seminal reflection for clinical practice in the future since current antimicrobial therapy guidelines in the context of CF may lead to the emergence of genes encoding antimicrobial resistance.

Keywords: mobilome, multidrug resistance, microbiota, transduction, phage induction

The environment as a source of mobilizable genes encoding antimicrobial resistance: the pan-microbial resistance genome (resistome)

Metagenomic analysis and fully sequenced genomes of bacteria have provided new insights into the relationship between the ecology and genome evolution of bacteria. Genome analysis has revealed that genome size and gene content in bacteria are correlated with their lifestyles.[1] Bacteria can be classified according to their lifestyle as mutualistic (sympatric speciation) and parasitic (allopatric speciation) organisms.[1] The evolutionary transitions underlying pathogenic and symbiotic lifestyles are mainly due to gene transfer and gene loss occurring within bacterial lineages. Compared with free-living bacteria, host-dependent bacteria are characterized by specialization by massive gene loss and genome size reduction.[1] Conversely, genome sequence analysis has shown that evolutionary pressure exerted by an ecological niche selects for a similar genetic repertoire in those prokaryotes that occupy the same niche, and that this is due to both vertical and horizontal transmission.[2] Genetic diversity in independent genetic worlds could be considered as a network structure with many possibilities for genetic exchange between these genetic worlds, thanks to plasmids and phages.[3] Moreover, environmental microbial communities represent a huge reservoir of mobilizable genes—the so called mobilome—that includes a pool of genes encoding antimicrobial resistance—the resistome—that may penetrate from the pan-microbial genome into taxonomically and ecologically distant bacterial populations, including pathogens.

It is now accepted that environmental bacteria are more intrinsically resistant to antimicrobials compared with the commensal bacteria that are the main causes of infectious diseases, which contain only a tiny fraction of the resistome.[4] Moreover, the majority of the mechanisms of antimicrobial resistance in clinical isolates have originated from the environmental resistome and have been transferred to other bacteria through mobile genetic elements, including plasmids and/or phages.[4,5] Recent functional metagenomic studies on soil samples revealed that soil bacteria display a high level of genetic diversity and are a reservoir of antimicrobial resistance genes.[6,7] Similarly, it has been demonstrated recently that phages from environmental samples carry antimicrobial resistance genes that are able to confer antimicrobial resistance on bacterial strains.[8] Finally, human gut microbiota also carry antimicrobial resistance genes, as recently demonstrated by culture of human faecal microbiota.[9] Selective pressure provided by antimicrobials contributes considerably to the mobilization and distribution of antimicrobial resistance genes from the resistome throughout microbial populations. A genome sequence can be considered to be the result of adaptation in response to evolutionary pressure. Genetic mechanisms involved in the acquisition of antimicrobial resistance genes from the environmental 'resistome' by commensal and pathogenic bacteria are mainly driven by horizontal gene transfer (HGT), including transformation, conjugation and transduction.[3] Interestingly, it has recently been demonstrated that clinical isolates of carbapenem-resistant *Acinetobacter baumannii* may release outer membrane vesicles as a mechanism of HGT to transfer carbapenem resistance to surrounding carbapenem-susceptible *A. baumannii* bacterial isolates.[10] Viruses, especially bacteriophages, have the ability to manipulate the life histories and evolution of their hosts in remarkable ways.[11–13] Transduction is one of the most powerful means of dissemination of genes that allow bacteria to become more pathogenic and resistant to antimicrobials.[14] Recent data have demonstrated that HGT by phages is much more prevalent than previously thought and that the environment plays a critical role in the transfer of antimicrobial resistance determinants by phages.[8,15] It has been demonstrated that bacteriophages in sewage and in environmental samples contribute to the spread of several β-lactamase genes.[8,15] Similarly, bacteriophages have been shown to play a role in the transfer of antimicrobial resistance determinants in enterococci, especially in interspecies transduction.[16]

Thus, current knowledge on extensive HGT and recombination processes in the prokaryotic world has led to the emergence of the concept of open compartments in microbial communities, with antimicrobial resistance being a consequence of evolution, ecology and gene exchanges in a specific niche.[3] These recent studies give new insights into mechanisms that could explain the emergence and diffusion of antimicrobial resistance in human bacterial pathogens. Besides the environmental resistome, there are other antimicrobial resistance gene pools, especially in humans, including gut microbiota[13,17] and the respiratory tract microbiota, which may be exposed to antimicrobials. In this way, some infectious diseases may be defined as a prokaryotic community resulting from a complex genomic assembly within the environment rather than single pathogenic species. Thus, it is assumed that the size of a microbial community is correlated with the size of the mobilome, including the resistome, and the probability of having genetic exchanges that may influence the pan-microbial genome, leading to the adaptation of this microbial community to a specific niche under selective pressure.

Cystic fibrosis (CF) is a polymicrobial disease and the lung is the site of a specific microbiota for emergence of antimicrobial resistance

CF, the most common genetic disease in Caucasian populations, is characterized by impaired mucociliary clearance, leading to colonization of the airways by pathogenic bacteria.[18] Repeated bronchopulmonary infections are responsible for a progressive decrease in the lung function and, eventually, death. Although the life expectancy of patients has increased to more than 36 years over the last decade,[19] mainly due to the improvement of care in specialized centres and intensive antimicrobial therapy, emerging multidrug-resistant pathogens are becoming more difficult to eradicate.[18] Recent studies of CF mucus by molecular methods have changed our thinking about CF lung infections. Besides the most commonly found pathogens (*Pseudomonas aeruginosa, Staphylococcus aureus, Burkholderia cepacia* complex, *Haemophilus influenzae* and atypical mycobacteria), analysis of mucus in CF airways has revealed the presence of a wide variety of bacteria at the same time, including anaerobes.[18,20,21] In our previous study using 16S PCR amplification, cloning and sequencing of 760 clones from 25 sputum samples from CF patients, we were able to detect 53 different bacterial species with a mean of about 7 different bacterial species per sputum sample.[20] Interestingly, the mean number of bacterial species detected increased when more clones were sequenced, indicating a huge complexity of microbial communities in these patients.[20] In a similar study, Harris *et al*.[21] detected 65 different bacterial species from 28 bronchoalveolar lavage samples from CF patients. Moreover, metagenomic analysis of DNA viruses in CF sputa highlighted the importance of the phage community in this niche.[22] It has been suggested that the human respiratory tract DNA virome in CF patients is defined by metabolism rather than by taxonomy, suggesting that CF therapeutics could be changed to focus on changing the environment of the airways rather than targeting specific common bacterial pathogens.[22] All these studies led us to consider CF mucus and airways as the site of a complex microbiota, where bacteria and viruses live together in a mutualistic lifestyle and adapt to their environment. This represents a new approach to the disease in which CF is considered a chronic disease with a complex microbial community resulting from genomic assemblage within the CF lung microbiota. Inside this microbiota, cross-talk between microbes and HGT are believed to contribute to the adaptation of bacteria to this specific environment over time, as in the human gut[23] or marine ecosystems.[24] In these communities with high levels of selective pressure, including antimicrobials, microbes must co-evolve and copy antimicrobial resistance strategies to compete for resources with microbial neighbours, mainly by HGT. Once the infection is installed in CF patients, eradication of these bacteria is impossible and iterative antimicrobial therapy is the rule, leading to the emergence of multidrug-resistant bacteria. Thus, lung infection and chronic colonization in CF patients is radically different from that of acute pneumonia, where only a few bacteria and viruses are pathogens. Moreover, the microbial community in the CF lung is highly variable over time and in relation to antimicrobial treatment and is probably linked to a balance between antimicrobial use and the acquisition of genes encoding antimicrobial resistance.

Bacterial genomes in CF patients are manipulated by phages

Whole-genome sequencing of many genomes of bacteria has recently demonstrated that lysogeny (integration of phage DNA into the bacterial chromosome as prophages) is a very frequent event, with many bacteria carrying multiple prophages.[13,25] More specifically, phages are important vectors for the transfer of DNA between bacterial strains and contribute to the genetic individuality of a bacterial strain.[25] The presence of *P. aeruginosa* tailed phages in sputum samples from CF patients was reported previously, suggesting that they may play a role in the phenotypic changes of *P. aeruginosa* during chronic lung infection.[26] Moreover, fluctuations in phenotypes and genotypes within populations of *P. aeruginosa* in the CF lung during short periods of pulmonary exacerbation and intravenous antimicrobial therapy were found to be mainly due to widespread phage activity and genomic rearrangements.[27] Similarly, genome analysis of strains of *B. cepacia*[28] and *P. aeruginosa*[29] isolated from CF patients revealed a particular genomic organization due to the presence of numerous prophages, which were demonstrated to be essential for the adaptation of these strains to the CF lung. Interestingly, the accessory genome of these CF strains contains many antimicrobial resistance genes as well as specific functions for adaptation specifically to the respiratory tract that were found to be encoded in prophages.[28–30] Recently, we have also reported the detection of an original multidrug-resistant *S. aureus* strain (CF-Marseille) spreading in the CF community in Marseilles, France; this strain displayed a novel phenotype and presented a new phage closely related to phiETA3, suggested to be instrumental in acquired resistance.[30] Thus, phage mobilization is believed to contribute significantly to genome alteration in the context of CF.

The paradigm of antimicrobial treatment, phage induction and diffusion of antimicrobial resistance genes by transduction

There is now evidence that antimicrobials can cause phage induction in *S. aureus* and *P. aeruginosa* isolates from CF patients.[30,31] We have reported that several antimicrobials commonly used in CF patients, including tobramycin, ciprofloxacin, co-trimoxazole and imipenem, are able to induce phages in *S. aureus* strain CF-Marseille.[30] Similarly, it was reported recently that ciprofloxacin could induce high levels of phage production in *P. aeruginosa* strains from CF patients.[31] Moreover, free phages from *P. aeruginosa* were detected directly in CF patient sputum samples by both plaque (40% positive) and PCR (76% positive) assays.[31] A recent study by Willner *et al.*[22] characterizing viral communities in CF patients versus non-CF patients demonstrated that the phage community was common to all CF patients and different from that of healthy subjects, and these phages presented with a specific metabolic profile reflecting adaptation to the particular nature of CF mucus.[22] Specifically, it was found that 16059 annotated sequences were specific to the CF viromes,[32] including genes related to virulence. Annotation of sequences in the 'virulence' dataset identified many proteins associated with antimicrobial resistance genes.[32] Along with this review we report an original analysis of these specific sequences to reinforce our discussion (see article by Fancello *et al.*[33] in this issue). Looking at these specific sequences, we found that CF viromes contain a huge and significantly higher number of sequences putatively encoding resistance to antimicrobials from various origins (efflux pumps, fluoroquinolone resistance and β-lactamases) compared with non-CF viromes.[33] As CF patients are regularly treated with antimicrobials, we speculate that this phenomenon could explain in part the emergence of antimicrobial resistance via HGT and generalized transduction.

CF lung as a niche for creation of new species and multidrug-resistant bacteria with chimeric repertoires

Intensive antimicrobial therapy has undeniably increased the life expectancy of CF patients.[19] However, collateral effects of antimicrobial treatment, i.e. transduction of virulence genes or toxins or antimicrobial resistance determinants, often appear at concentrations subinhibitory for bacterial growth.[34] As antimicrobial treatment is adapted to the bacterium responsible for the disease, this situation may be considered unlikely from a clinical perspective. It is important to be aware that it can easily occur in CF, where bronchial exacerbation is now considered polymicrobial.[18] Indeed, antimicrobials employed in cystic fibrosis target the predominant bacterium found in sputum culture; in the case of a co-infection or co-colonization, antimicrobials which might be inefficient against, for example, *S. aureus* could be used to treat *P. aeruginosa*, and favour induction of *S. aureus* prophages.

These findings, along with genomic data, strongly suggest that the CF microbiota continuously generates new species, especially multidrug-resistant bacteria with chimeric repertoires, as exemplified by genome analysis of *S. aureus*,[30] *P. aeruginosa*[29] and *B. cepacia*[28] isolated from CF patients, which may be selected for their adaptation to this niche during the course of the disease and recurrent antimicrobial therapy. These new 'superbugs' may be able to spread in this community as well as on a worldwide scale. We postulate that CF patients have complex lung microbiota that include eukaryotes, bacteria, fungi and viruses within which there are many genetic exchanges occurring by HGT, especially generalized transduction due to antimicrobial treatment and selective pressure, which promote both the emergence and selection of multidrug-resistant bacteria that may spread in this population and the evolution of the microbial community that is specific to the disease. This paradigm is very similar to that of amoebas as a melting pot for the creation of new species with chimeric repertoires that may succeed and be selected if adapted in a specific niche.[35] Further work is needed to isolate and characterize phage particles containing genes encoding antimicrobial resistance directly from sputum samples and test their ability to transduce the antimicrobial resistance determinants to diverse host strains, as recently demonstrated in bacteriophages from environmental samples.[8] If phages are induced by antimicrobials among this complex microbiota, including bacteria, fungi and viruses, we can imagine the innumerable possibilities of HGT, particularly in the presence of phages with broad host range, which are able to transduce genes between different species. As phages are already known to be mobilized during chronic infection of the lungs of patients with CF, it seems particularly important to improve our understanding of the mechanisms of phage induction in order to prevent the spread of virulence and/or antimicrobial resistance determinants within CF population as well as in the community. These findings are of great concern for clinical practice in the future since current antimicrobial therapy guidelines in the context of CF may lead to the emergence and spread of genes encoding antimicrobial resistance.

Funding

This work was partly supported by CNRS, France and a Starting Grant (number 242729) to C. Desnues from the European Research Council.

Transparency declarations

None to declare.

References

1 Merhej V, Royer-Carenzi M, Pontarotti P *et al*. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* 2009; **4**: 13.

2 Suen G, Goldman BS, Welch RD. Predicting prokaryotic ecological niches using genome sequence analysis. *PLoS ONE* 2007; **2**: e743.

3 Halary S, Leigh JW, Cheaib B *et al*. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci USA* 2010; **107**: 127–32.

4 Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol* 2007; **5**: 175–86.

5 Canton R. Antibiotic resistance genes from the environment: a perspective through newly identified antibiotic resistance mechanisms in the clinical setting. *Clin Microbiol Infect* 2009; **15** Suppl 1: 20–5.

6 Torres-Cortes G, Millan V, Ramirez-Saad HC *et al*. Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. *Environ Microbiol* 2011; **13**: 1101–14.

7 Riesenfeld CS, Goodman RM, Handelsman J. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol* 2004; **6**: 981–9.

8 Colomer-Lluch M, Jofre J, Muniesa M. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS ONE* 2011; **6**: e17549.

9 Goodman AL, Kallstrom G, Faith JJ *et al*. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc Natl Acad Sci USA* 2011; **108**: 6252–7.

10 Rumbo C, Fernandez-Moreira E, Merino M *et al*. Horizontal transfer of the OXA-24 carbapenemase gene via outer membrane vesicles: a new mechanism of dissemination of the carbapenem resistance genes in *Acinetobacter baumannii*. *Antimicrob Agents Chemother* 2011; **55**: 3084–90.

11 Rohwer F, Thurber RV. Viruses manipulate the marine environment. *Nature* 2009; **459**: 207–12.

12 Canchaya C, Fournous G, Chibani-Chennoufi S *et al*. Phage as agents of lateral gene transfer. *Curr Opin Microbiol* 2003; **6**: 417–24.

13 Luo C, Walk ST, Gordon DM *et al*. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci USA* 2011; **108**: 7200–5.

14 Brussow H, Canchaya C, Hardt WD. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* 2004; **68**: 560–602.

15 Muniesa M, Garcia A, Miro E *et al*. Bacteriophages and diffusion of β-lactamase genes. *Emerg Infect Dis* 2004; **10**: 1134–7.

16 Mazaheri Nezhad FR, Barton MD, Heuzenroeder MW. Bacteriophage-mediated transduction of antibiotic resistance in enterococci. *Lett Appl Microbiol* 2011; **52**: 559–64.

17 Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 2009; **325**: 1128–31.

18 Bittar F, Rolain JM. Detection and accurate identification of new or emerging bacteria in cystic fibrosis patients. *Clin Microbiol Infect* 2010; **16**: 809–20.

19 O'Sullivan BP, Freedman SD. Cystic fibrosis. *Lancet* 2009; **373**: 1891–904.

20 Bittar F, Richet H, Dubus JC *et al*. Molecular detection of multiple emerging pathogens in sputa from cystic fibrosis patients. *PLOS One* 2008; **3**: e2908.

21 Harris JK, De Groote MA, Sagel SD *et al*. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. *Proc Natl Acad Sci USA* 2007; **104**: 20529–33.

22 Willner D, Furlan M, Haynes M *et al*. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 2009; **4**: e7370.

23 Kurokawa K, Itoh T, Kuwahara T *et al*. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 2007; **14**: 169–81.

24 Sobecky PA, Hazen TH. Horizontal gene transfer and mobile genetic elements in marine systems. *Methods Mol Biol* 2009; **532**: 435–53.

25 Canchaya C, Fournous G, Brussow H. The impact of prophages on bacterial chromosomes. *Mol Microbiol* 2004; **53**: 9–18.

26 Ojeniyi B, Birch-Andersen A, Mansa B *et al*. Morphology of *Pseudomonas aeruginosa* phages from the sputum of cystic fibrosis patients and from the phage typing set. An electron microscopy study. *APMIS* 1991; **99**: 925–30.

27 Fothergill JL, Mowat E, Ledson MJ *et al*. Fluctuations in phenotypes and genotypes within populations of *Pseudomonas aeruginosa* in the cystic fibrosis lung during pulmonary exacerbations. *J Med Microbiol* 2010; **59**: 472–81.

28 Holden MT, Seth-Smith HM, Crossman LC *et al*. The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. *J Bacteriol* 2009; **191**: 261–77.

29 Winstanley C, Langille MG, Fothergill JL *et al*. Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res* 2009; **19**: 12–23.

30 Rolain JM, Francois P, Hernandez D *et al*. Genomic analysis of an emerging multiresistant *Staphylococcus aureus* strain rapidly spreading in cystic fibrosis patients revealed the presence of an antibiotic inducible bacteriophage. *Biol Direct* 2009; **4**: 1.

31 Fothergill JL, Mowat E, Walshaw MJ *et al*. Effect of antibiotic treatment on bacteriophage production by a cystic fibrosis epidemic strain of *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother* 2011; **55**: 426–8.

32 Willner D, Furlan M. Deciphering the role of phage in the cystic fibrosis airway. *Virulence* 2010; **1**: 309–13.

33 Fancello L, Desnues C, Raoult D *et al*. Bacteriophages and diffusion of genes encoding antimicrobial resistance in cystic fibrosis sputum microbiota. *J Antimicrob Chemother* 2011; doi:10.1093/jac/dkr315.

34 Stanton TB, Humphrey SB, Sharma VK *et al*. Collateral effects of antibiotics: carbadox and metronidazole induce VSH-1 and facilitate gene transfer among *Brachyspira hyodysenteriae* strains. *Appl Environ Microbiol* 2008; **74**: 2950–6.

35 Boyer M, Yutin N, Pagnier I *et al*. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA* 2009; **106**: 21848–53.

Acknowledgements

In the first place I would like to record my gratitude to Doctor Christelle Desnues for her supervision, advice and guidance. Her scientific and human qualities as well as her enthusiasm and courage are an example for me.

I would also like to thank Professor Didier Raoult for accepting me in his unit and Professor Bernard La Scola for his acceptance to be the president of the committee.
I gratefully thank Professor Patrick Forterre, Professor Renaud Mahieux and Doctor François Enault for their acceptance to be part of my thesis committee.

I am very grateful to all present and past people who have been part of my team: Dao, Ikram, Mickael, Nikolay, Priscilla, Sarah, Sonia and Emmanuel. I would also like to thank every member of the URMITE. I apologize that I cannot mention one by one every person who welcomed, supported and adviced me in these years.
I thank all other fellow PhD students for their support, company and friendship.

I would like to thank all my friends in Marseille, Milano and elsewhere in the world to be part of my life.

I won't write the most important aknowledgement because I don't know how to... but this is for YOU.

At last, thanks to all my family who supported my adventure here with enthusiasm and pride. I'm proud of you too :)