Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

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

ARTICLE INFO

Available online 11 October 2012

Keywords:
Virus
Metagenomics
Computational tools
Clinical research
Virome
Emerging disease

ABSTRACT

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.

© 2012 Elsevier Inc. All rights reserved.

Contents

Viral infections and the need for better viral discovery tools . . . 163
Viral metagenomics and its first applications . . . 163
Identifying human-associated viral communities (the human virome) . . . 163
Bacteriophages in the human virome . . . 164
Clinical applications: discovery of human pathogens . . . 164
General considerations on technical issues and potential biases in metagenome preparation . . . 164
Computational tools and algorithms in clinical viral metagenomics . . . 166
Pre-processing and quality control . . . 166
Annotation, assembly and estimation of the community diversity and structure . . . 167
Taxonomic classification . . . 167
Assembly . . . 168
Genotype abundances, community diversity and structure . . . 169
Statistical tools for the analysis of clinical metagenomic samples . . . 169
Characterization of the "unknown" . . . 170
Next-generation sequencing technologies and the need for a common standardized pipeline analysis . . . 170
Conclusions . . . 171
Acknowledgments . . . 171
References . . . 171

* 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:

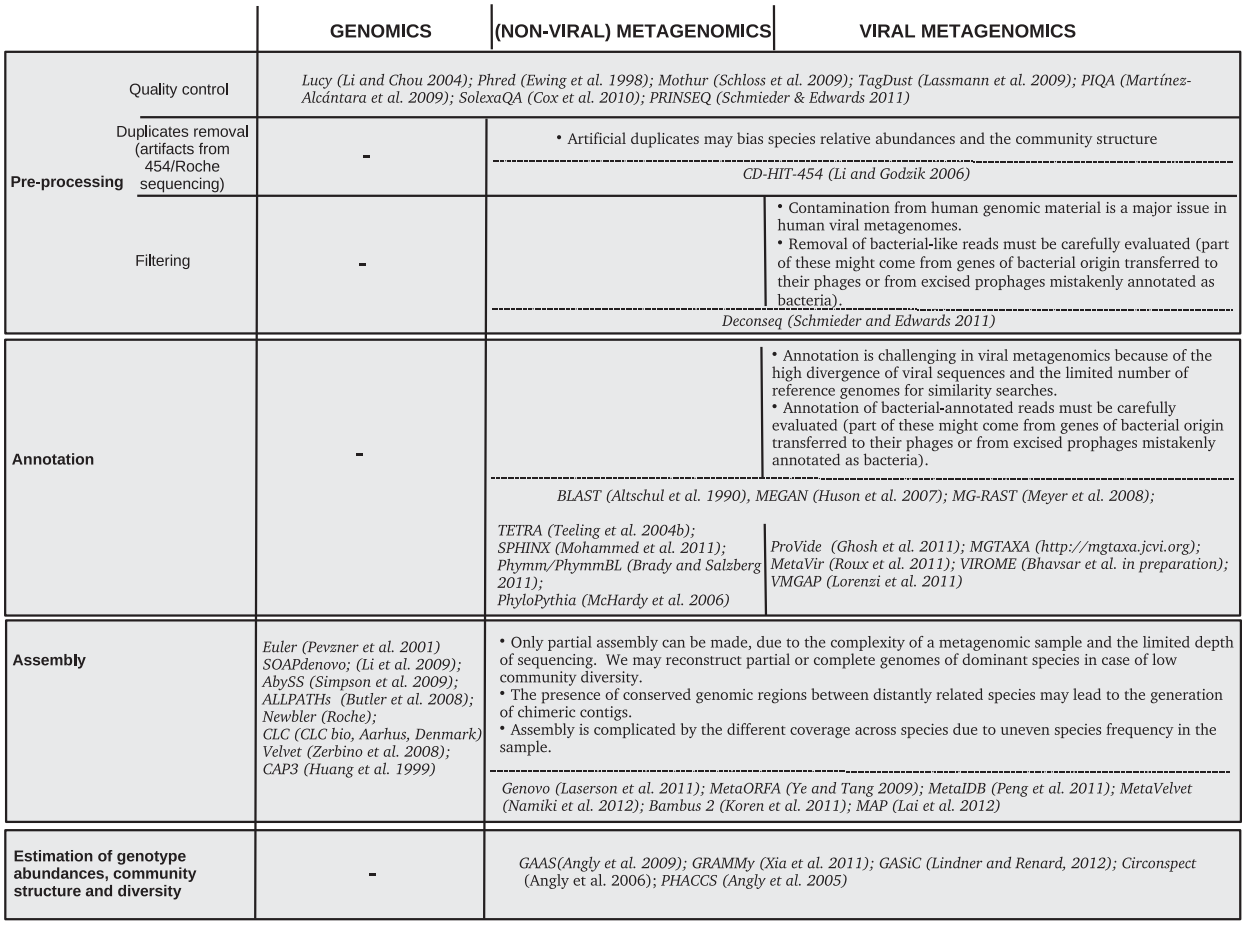

Fig. 1. Overview of the main issues and tools for computational analysis in genomics, metagenomics and viral metagenomics. For each step of the computational analysis, we reported specific issues, if any, relative to (non viral) metagenomic and viral metagenomics. Corresponding computational tools are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