VIRUS EVOLUTION

Virus Evolution, 2016, 2(2): vew022

doi: 10.1093/ve/vew022
Review article

Challenges in the analysis of viral metagenomes

Rebecca Rose[1,*,†,‡], Bede Constantinides[2,‡], Avraam Tapinos[2,‡], David L Robertson[2] and Mattia Prosperi[3,*]

[1]BioInfoExperts, Norfolk, VA, USA, [2]Computational and Evolutionary Biology Faculty of Life Sciences, University of Manchester, Manchester, UK and [3]Department of Epidemiology, University of Florida, Gainesville, FL, USA

*Corresponding author: E-mail: rebecca.rose@bioinfox.com and ahnven@gmail.com

[†]http://orcid.org/0000-0002-3480-3819

[‡]These authors contributed equally to this work.

Abstract

Genome sequencing technologies continue to develop with remarkable pace, yet analytical approaches for reconstructing and classifying viral genomes from mixed samples remain limited in their performance and usability. Existing solutions generally target expert users and often have unclear scope, making it challenging to critically evaluate their performance. There is a growing need for intuitive analytical tooling for researchers lacking specialist computing expertise and that is applicable in diverse experimental circumstances. Notable technical challenges have impeded progress; for example, fragments of viral genomes are typically orders of magnitude less abundant than those of host, bacteria, and/or other organisms in clinical and environmental metagenomes; observed viral genomes often deviate considerably from reference genomes demanding use of exhaustive alignment approaches; high intrapopulation viral diversity can lead to ambiguous sequence reconstruction; and finally, the relatively few documented viral reference genomes compared to the estimated number of distinct viral taxa renders classification problematic. Various software tools have been developed to accommodate the unique challenges and use cases associated with characterizing viral sequences; however, the quality of these tools varies, and their use often necessitates computing expertise or access to powerful computers, thus limiting their usefulness to many researchers. In this review, we consider the general and application-specific challenges posed by viral sequencing and analysis, outline the landscape of available tools and methodologies, and propose ways of overcoming the current barriers to effective analysis.

Key words: metagenomics, assembly, next-generation sequencing, classification, surveillance, epidemic

1. Introduction

In the last decade, at least seven separate viral outbreaks have caused tens of thousands of human deaths (Woolhouse, Rambaut, and Kellam, 2015), and the ever-increasing density of livestock, rate of habitat destruction, and extent of human global travel provides a fertile environment for new pandemics to emerge from host switching events (Delwart 2007; Fancello, Raoult, and Desnues 2012), as was the case for SARS, Ebola, Middle East Respiratory Syndrome (MERS), and influenza-A (H1N1) (Castillo-Chavez et al. 2015). At present we have a limited grasp of the extent of viral diversity present in the environment: the 2014 database release from the International Committee for the Taxonomy of Viruses classified just 7 orders, 104 families, 505 genera, and 3286 species (http://www.ictvonline.org/virustaxonomy.asp); yet, one study estimated that there are at least 320,000 virus species infecting mammals alone (Anthony et al. 2013).

Table 1. Discussed software for the analysis of viral (meta)genomes.

Name	Application	Distribution	Interface	Platform	License	Description	URL
Kraken (Wood and Salzberg 2014)	Taxonomic assignment	Source	Command line	Linux, MAC OS	GNU GPL	Fast in-memory k-mer search and LCA assignment of short reads using a comprehensive sequence database	https://ccb.jhu.edu/software/kraken/
Kaiju (Menzel, Ng, and Krogh 2016)	Taxonomic assignment	Source	Web, command line	Website, Linux	GNU GPL	Fast in-memory, k-mer seeded protein search and LCA taxonomic assignment	http://kaiju.binf.ku.dk/
CLARK (Ounit et al. 2015)	Taxonomic assignment	Source	Command line	Linux, MAC OS	GNU GPL	Fast in-memory k-mer search and LCA assignment of short reads using a comprehensive sequence database	http://clark.cs.ucr.edu/
Lambda (Hauswedell, Singer, and Reinert 2014)	Protein homology search	Binary, source	Command line	Linux, Mac OS	GNU GPL	Fast BLAST compatible nucleotide and reduced alphabet protein homology search	https://seqan.github.io/lambda/
Diamond (Buchfink, Xie, and Huson 2015)	Protein homology search	Binary, source	Command line	Linux, Mac OS, FreeBSD	–	Fast BLAST compatible nucleotide and reduced alphabet protein homology search	http://ab.inf.uni-tuebingen.de/software/diamond/
NCBI BLAST + (Altschul et al. 1990)	Nucleotide and protein homology search	Binary, source	Web, command line	Linux, Windows, Mac OS	Public domain	Nucleotide and protein homology search	https://blast.ncbi.nlm.nih.gov/Blast.cgi
VirusHunter (Zhao et al. 2013)	Virus discovery	Source	Command line	Linux	GNU GPL	Automated viral discovery pipeline for use with a computing cluster	http://www.ibridgenetwork.org/wustl/virushunter
MetaVir (Roux et al. 2011)	Taxonomic assignment	Web application	Web	–	–	Annotation and visualization of viral reads and assemblies	http://metavir-meb.univ-bpclermont.fr/
VirSorter (Roux et al. 2015)	Virus and prophage discovery	Source	Command line	Linux, Mac OS, Docker	GNU GPL	Reference-based and reference-independent annotation of assembled virus genomes	https://github.com/simroux/VirSorter
One Codex (Minot, Krumm, and Greenfield)	Taxonomic assignment	Web application	Web, web API	–	–	Web portal for assignment, visualization and comparison of metagenomes using a bespoke comprehensive sequence database	https://www.onecodex.com/
PHYMMBL (Brady and Salzberg 2011)	Taxonomic assignment	Source	Command line	Linux	–	Hybrid taxonomic assignment using Phymm interpolated Markov models and BLAST results	https://ccb.jhu.edu/software/phymmbl/index.shtml
IVA (Hunt et al. 2015)	Viral genome assembly	Python package, source	Command line	Linux, Mac OS	GNU GPL	Consensus genomic assembly of diverse viral populations using paired-end short reads	http://sanger-pathogens.github.io/iva/
Vicuna (Yang et al. 2012)	Viral genome assembly	Source	Command line	Linux	Broad academic license	Consensus genomic assembly of diverse viral populations using short reads	http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/vicuna

(continued)

Table 1. Continued

Name	Application	Distribution	Interface	Platform	License	Description	URL
PRICE (Ruby, Bellare, and Derisi 2013)	Viral genome assembly	Source	Command line	Linux	GNU GPL	Assembly of low abundance viral sequences in metagenomes with paired-end short reads	http://derisilab.ucsf.edu/software/price/
SPAdes (Bankevich et al. 2012)	Genome assembly	Binary, source	Command line	Linux, Mac OS	GNU GPL	Microbial genome assembly with long and short paired-end reads	http://bioinf.spbau.ru/spades
MetaSPAdes (Nurk et al. 2016)	Metagenome assembly	Binary, source	Command line	Linux, Mac OS	GNU GPL	Metagenome assembly with long and short paired-end reads	http://bioinf.spbau.ru/spades
MEGAHIT (Li et al. 2015)	Metagenome assembly	Source	Command line	Linux, Mac OS	GNU GPL	Fast metagenome assembly for complex metagenomes with short reads	https://github.com/voutcn/megahit
IDBA-UD (Peng et al. 2012)	Metagenome assembly	Source	Command line	Linux, Mac OS	GNU GPL	Metagenome assembly for short paired-end reads	https://github.com/loneknightpy/idba
MetaVelvet (Afiahayati, Sato, and Sakakibara 2015)	Metagenome assembly	Source	Command line	Linux, Mac OS	GNU GPL	Metagenome assembly for short paired-end reads	http://metavelvet.dna.bio.keio.ac.jp/
ShoRAH (Zagordi et al. 2011)	Viral haplotype reconstruction	Source	Command line	Linux	GNU GPL	Probabilistic viral haplotype reconstruction from short reads	https://github.com/ozagordi/shorah
QuRE (Prosperi and Salemi 2012)	Viral haplotype reconstruction	Java package, source	Command line, graphical interface	Most OSs	GNU GPL	Probabilistic viral haplotype reconstruction from short reads	https://sourceforge.net/projects/qure/
PredictHaplo (Prabhakaran et al. 2014)	Viral haplotype reconstruction	Source	Command line	Linux	GNU GPL	Probabilistic viral haplotype reconstruction from short paired-end reads	http://bmda.cs.unibas.ch/HivHaploTyper/

High throughput (or so-called 'next generation') sequencing of viruses during the most recent outbreaks of MERS in South Arabia (Gire et al. 2014; Carroll et al. 2015; Park et al. 2015) and Ebola in West Africa (Quick, J et al. 2016) has facilitated rapid identification of transmission chains, rates of viral evolution, and evidence of the zoonotic origin of these outbreaks. Access to such information during initial stages of an outbreak would offer invaluable insight into when, where, and how an epidemic might emerge, informing intervention and mitigation measures or even stopping it altogether. A major step towards this goal is therefore to identify existing zoonotic and environmental pathogens with pandemic potential. This is a significant undertaking, demanding considerable investment and close collaboration between government, NGOs and academia, for example, the USAID program PREDICT http://www.vetmed.ucdavis.edu/ohi/predict/index.cfm, as well as on the ground surveillance by local authorities and scientists in areas of the world most at risk.

The characterization of unknown viral entities in the environment is now possible with modern sequencing; however, current tooling for exploiting these data represents a practical and methodological bottleneck for effective data analysis. Practically, most available software tools are inaccessible to the majority of potential users, demanding expertise and computing resources often lacked by the researchers from diverse backgrounds involved in sample collection, sequencing, and analysis. There is a need for robust and intuitive analytical tools without requirements for fast internet connectivity, which may be unavailable in remote or developing regions. More fundamentally, the intended scope of published analytical tools and workflows is often less than clear, and given the diverse applications of viral sequencing, it can be difficult to gauge the relevance of newly published tools without first testing them. For example, a fast sequence classifier might fail entirely to detect a novel strain of a well-characterized virus, and equally might perform well with Illumina sequences yet deliver poor results for data generated with the Ion Torrent platform. Furthermore, results arising from these analyses should be replicable, intelligible, and useful to the end user, with provision for quality control and error management. Software tools that target expert users should be tested, documented and robustly distributed as packages or containers so as to streamline the processes of installation and generating results.

Methodologically, most genomic sequence analysis software is not well suited for viral genomes. Generic tools that are able to address the challenges posed by viral sequences are often applicable only in limited circumstances. Choosing between approaches is made difficult due to an abundance of disparate yet functionally equivalent methodologies and in general a lack of rigorous benchmarks for viral datasets. While there is much ongoing research in this area, both the sensitive detection of previously characterized viruses and viral discovery remain key challenges open for innovation. Here we survey the landscape of available approaches for analyzing both known and unknown viruses within genomic and metagenomic samples, with focus on their practical and methodological suitability for use by a broad spectrum of researchers seeking to characterize viral metagenomes.

2. Viral sequence enrichment: physical and *in silico* approaches

Within metagenomes the proportion of viral nucleic acids is typically far lower than that of host or other microbes, limiting the amount of signal available for analysis after sequencing. To

mitigate this issue, enrichment and amplification approaches are widely used prior to sequencing viral samples. Size filtration or density-based enrichment by centrifugation are two effective methods for increasing virus yield, although such methods may bias the observed composition of viral populations (Ruby, Bellare, and Derisi 2013). Alternatively, PCR amplification may be used to generate an abundance of specific viral sequences present in a sample, a widely used strategy, which was employed in the identification and analysis of MERS coronavirus (Zaki et al. 2012; Cotten et al. 2013, 2014), although effective primer design can be challenging in the presence of high genomic diversity in the target viral species. Conversely, an excess of sequencing coverage can lead to the construction of overly complex and unwieldy *de novo* assembly graphs in the presence of high genomic diversity, reducing assembly quality. Using *in silico* normalisation (Crusoe et al. 2015), excess coverage may be reduced by discarding sequences containing redundant information. This approach increases analytical efficiency when dealing with high coverage sequence data, and we have shown that it can benefit *de novo* assembly of viral consensus sequences. Another *in silico* strategy for increasing analytical efficiency by discarding unneeded data is to filter sequences from known abundant organisms through alignment with one or more reference genomes using an aligner or specialist tool (approaches reviewed in Daly et al. 2015).

3. Choosing a sequencing platform

There are several sequencing technologies in widespread use that are capable of reading hundreds of thousands to billions of DNA sequences per run (Reuter, Spacek, and Snyder 2015). The current market leader, Illumina, manufactures instruments capable of generating billions of 150 base pair (bp) paired end reads (see 'Glossary') per run, with read lengths of up to 300 bp. The Illumina short read platform is widely used for analyses of viral genomes and metagenomes, and, given sufficient sequencing coverage, enables sensitive characterization of low-frequency variation within viral populations (e.g. HIV resistance mutations as low as 0.1% (Li et al. 2014)). Ion Torrent (ThermoFisher) is capable of generating longer reads than Illumina at the expense of reduced throughput and a higher rate of insertion and deletion (indel) error (Eid et al. 2009). Single molecule real-time sequencing commercialized by Pacific Biosciences (PacBio) produces much longer (>10 kbp) reads from a single molecule without clonal amplification, which eliminates the errors introduced in this step. However, this platform has a high (~10%) intrinsic error rate, and remains much more expensive than Illumina sequencing for equivalent throughput. The Nanopore platform from Oxford Nanopore Technologies, which includes the pocket sized MinION sequencer, also implements long read single molecule sequencing, and permits truly real-time analysis of individual sequences as they are generated. Although more affordable than PacBio single molecule sequencing, the Nanopore platform also suffers from high error rates in comparison with Illumina (Reuter, Spacek, and Snyder 2015). However, the technology is maturing rapidly and has already demonstrated potential to revolutionize pathogen surveillance and discovery in the field, as well as enabling contiguous assembly of entire bacterial genomes at relatively low cost (Feng et al. 2015; Quick et al. 2015; Hoenen et al. 2016). Hybrid sequencing strategies using both long and short reads leverage the ability of long reads to resolve repetitive DNA regions while benefitting from the high accuracy of short reads, at the expense of additional sequencing, library preparation and data analysis (Madoui et al. 2015).

Glossary

Contigs: Contiguous nucleotide sequences assembled from multiple overlapping reads.
Coverage: The number of times a genome (or part thereof) has been sequenced.
de Bruijn graph: A network of nodes and edges, where each edge represents a k-mer found in the collection of reads, and each node represents either the prefix or suffix of the k-mer.
De novo assembly: Reconstruction of short sequences into longer sequences (or contigs), without use of a reference sequence
Digital signal processing data transformation: Analytical techniques for transforming sequential data into a domain representative of data features.
Discrete Fourier transform: A spectral analysis technique for identifying sine and cosine frequency components in numerical signal data.
Discrete wavelet transform: A spectral analysis technique for decomposing data to its frequency and spatial components.
k-mer: A subsequence of length k. Many genomic analyses involve decomposition of sequences into all possible subsequences of a specified length k.
Numerical sequence representation: Numerical mapping of nucleotide sequences, permitting the application of signal processing transformation approaches.
Paired-end reads: Reads generated from both 5 and 3 ends of the same DNA molecule. Depending on the length of the molecule and that of the reads, these pairs may or may not overlap in the middle.
Read overlap graphs: A network of nodes and edges, where each edge represents a read and each vertex represents an overlap between two nodes.
Reference-based alignment: Orientation/alignment of reads with respect to a specified reference sequence.
Scaffolds: DNA sequences comprising contigs with gaps between them, often generated using read pairing information.
Suffix array: A sorted array of all suffixes of a string, such as a DNA sequence, enabling efficient sequence comparison.

4. Assembling genomes: *de novo* and reference-based assembly

The reconstruction of sequencing reads into full length genes and genomes can be performed by means of either reference-based alignment or *de novo* assembly, a decision dependent on experimental objectives, read length, quality and data complexity. In reference-based approaches, reads are mapped to similar regions of a supplied template genome, a well-studied and computationally efficient process implemented with a suffix array index of the reference genome. In contrast, *de novo* assembly is computationally exhaustive but important in cases where either a target genome is poorly characterized or reconstruction of genomes of *a priori* unknown entities in metagenomes is sought, such as in surveillance studies. For short read data, the increased sequence length afforded by assembly can be necessary to distinguish members of highly conserved gene families from one another. Assembly is also widely used for generating whole genome consensus sequences to facilitate analyses of viral variation, and is a typical starting point for analyses of diverse populations of well-characterized viruses. Even where

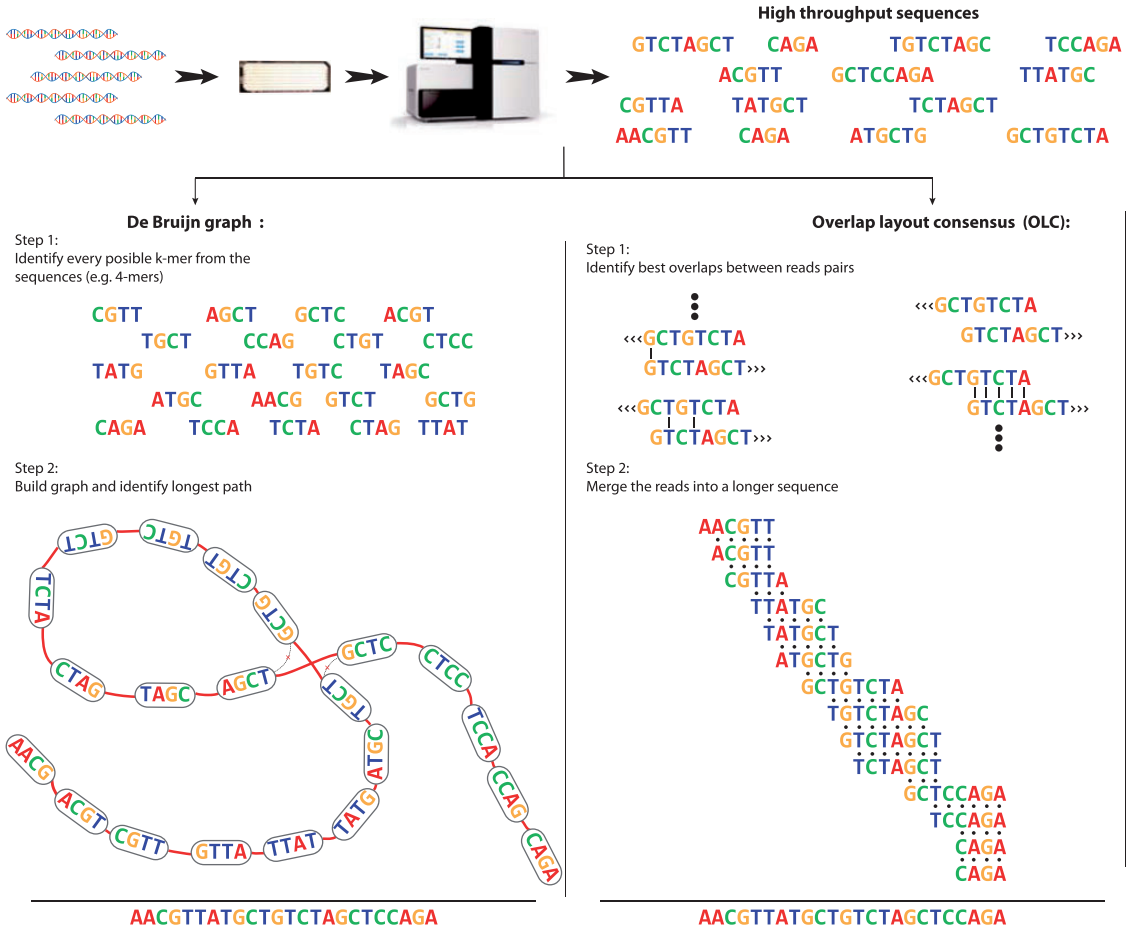

Figure 1. Two widely used methodologies in *de novo* assembly of short reads. Reads are not represented explicitly within a de Bruijn graph; they are instead decomposed into distinct subsequence 'words' of length *k*, or *k*-mers, which can be linked together via overlapping *k*-mers to create an assembly graph. In OLC, a pairwise comparison of all reads is performed, identifying reads with overlapping regions. These overlaps are used to construct a read graph. Next, overlapping reads are bundled into aligned contigs in what is referred to as the layout step, before finally the most likely nucleotide at position is determined through consensus. This figure is simplified to demonstrate the theory for the assembly of single genomes; note that the process has additional complexities for the reconstruction of metagenomes.

long reads are available, assembly plays an important role in mitigating the high error rates associated with single molecule sequencing technologies, yielding accurate consensus sequences from inaccurate individual reads.

4.1 *De novo* assembly methodologies

Modern *de novo* assemblers generally leverage either de Bruijn graphs or read overlap graphs as part of the approach known as overlap layout consensus (OLC). Figure 1 illustrates the differences between the two methods. OLC assemblers use the similarity of whole reads in order to construct a graph wherein each read is represented by a node, and subsequently merge overlapping reads into consensus contigs (Deng et al. 2015). OLC is relatively time and memory intensive, scaling poorly to millions of reads and beyond. However, the fewer, longer reads generated by emerging single molecule sequencing technologies tend to be well suited to OLC assembly, which can be easily implemented to tolerate long and noisy sequences (Compeau, Pevzner, and Tesler 2011). Older, notable, *de novo* assemblers implementing OLC include CAP3 (Huang and Madan 1999) and Celera (http://www.jcvi.org/cms/research/projects/cabog/overview/), while MHAP (Berlin et al. 2015), Canu (Berlin et al. 2015), and Miniasm (Li 2016) represent the current state of the art. There also exist a number of OLC assemblers intended for use with viral sequences: VICUNA was designed for short, non-repetitive and highly variable reads from a single population (Yang et al. 2012), and PRICE (Ruby, Bellare, and Derisi, 2013) iteratively assembles low to moderate complexity metagenomes (e.g. Runckel et al. 2011; Grard et al. 2012;) using a similar algorithm to the actively developed consensus assembler IVA (Hunt et al. 2015), which like VICUNA is designed for single virus populations rather than metagenomes (see Table 1 for additional details on programs).

A de Bruijn or *k*-mer graph represents a set of reads in terms of its *k*-mer composition, where *k*-mers are subsequences of a length *k*, specified by the user. Each *k*-mer is assigned to an edge in a graph, where the nodes are *k*-1 prefixes and suffixes of the *k*-mer. The assembler identifies the path through the graph in which each edge is visited only once (reviewed in Compeau, Pevzner, and Tesler 2011). De Bruijn graphs are much more efficient to construct than overlap graphs and are suited to large numbers of short reads, and where coverage is high, since redundant *k*-mers occupy negligible random access memory (RAM). However, with this efficiency comes a lack of error tolerance in identifying overlaps, less tolerance of repeated sequences in comparison to overlap graphs, and a loss of read coherence, meaning that *k*-mers originating from different reads may be co-assembled. Examples of assemblers using de Bruijn graphs include SOAPdenovo (Luo et al. 2012), ALLPATHS

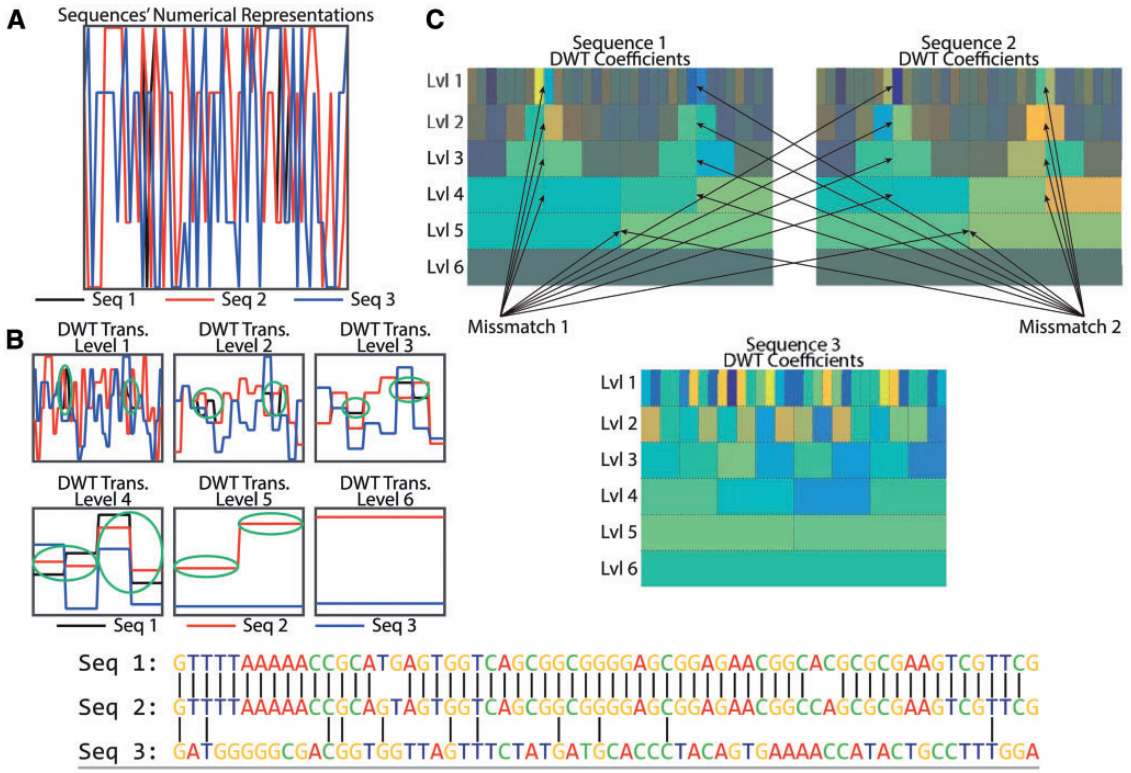

Figure 2. Proposed DWT signal processing approach for nucleotide sequence analysis. Sequences 1 and 2 are subsequences of the HIV-1 HXB2 genome (the reference genome for HIV), and sequence 3 is a subsequence of the *Mycoplasma genitalium* genome (all three sequences appear at the bottom of the figure). **(A)** illustrates the integer number representations of the three sequences—sequence 1 is depicted as a black line, sequence 2 is depicted as a red line and sequence 3 is depicted as a blue line. The sequences are mapped into numerical space with the integer representation method enabling the application of transformation approaches. **(B)** illustrates the DWT transformations of the three sequences' numerical representations at varying resolutions. The three sequences are each shown consecutively transformed into six reduced resolution representations. The minor sequence mismatches between sequences 1 and 2 (indicated with green circles) can be easily detected at different transformation resolutions despite reduction in information content from the transformation process. Similar nucleotide sequences give rise to similar DWT transformations and thus can be intuitively identified even at low resolution (level 6), where sequences are represented by a single numerical value. Depicted in **(C)** are the coefficient matrices obtained from each sequence's DWT transformation. Coefficient matrices can be used to approximately identify the sites of the mismatch positions between the two sequences. Sequences 1 and 2 differ only at sites 16–17 and 48–49. The exact location of minor differences can be detected at transformation level 4 where each sequence is compressed to four wavelets. Darker colored positions in between the matrices of sequence 1 and 2 indicate matching coefficients, and lighter colored positions indicate dissimilar coefficients.

(Butler et al. 2008), SPAdes (Bankevich et al. 2012), and ABySS (Simpson et al. 2009).

4.2 *De novo* assembly for metagenomes

Typical *de novo* assemblers are designed to reconstruct genomes with uniform sequencing coverage across their length. This is problematic for metagenomes (including viromes) where coverage typically varies considerably both among different genomes and within individual genomes. To address this problem, dedicated metagenome assemblers have been developed. Omega (Haider et al. 2014) is an OLC-based method that uses a minimum cost flow analysis of the OLC graph to generate initial contigs, merging these to create longer contigs and scaffolds using mate-pair information. Genovo (Laserson, Jojic, and Koller 2011) is another OLC-based method that generates a probabilistic model for the dataset and subsequently uses an iterative approach to reconstruct the most likely genome contigs. MEGAHIT (Li et al. 2015) prioritizes speed, leveraging a succinct de Bruijn graph to rapidly reconstruct high complexity metagenomes, such as those of soil or seawater, on a single computer. Noteworthy is the iterative de Bruijn graph assembler SPAdes, which although not initially intended for metagenome assembly, has been widely adopted for its effectiveness in assembling variable coverage metagenomes of limited complexity. MetaSPAdes (Nurk et al. 2016) is a metagenome-specific release of the SPAdes pipeline with refinements to its graph simplification and repeat resolution algorithms, counter-intuitively capable of leveraging rare strain information so as to improve its consensus reconstruction capabilities. Other de Bruijn graph metagenome assemblers based on their genomic counterparts include Ray-Meta (Boisvert et al. 2012), MetAMOS (Treangen et al. 2013), MetaVelvet (Namiki et al. 2012; Afiahayati, Sato, and Sakakibara 2015), and IDBA-UD (Peng et al. 2012).

For example, unlike the genome assembler Velvet, MetaVelvet's de Bruijn graph is decomposed into many subgraphs (using coverage difference and graph connectivity), and scaffolds are built independently for each subgraph. MetaVelvet-SL addresses limitations with MetaVelvet, using supervised learning to detect and classify chimeric nodes within the de Bruijn graph. IDBA-UD partitions a de Bruijn graph into isolated components, constructs a multiple alignment, and subsequently identifies variation within these partitions using multiple depth relative thresholds to remove erroneous *k*-mers. Ray Meta (Boisvert et al. 2012) extends the massively distributed assembly model of Ray to variable coverage metagenomes, while MetAMOS (Treangen et al. 2013) is both a metagenomic extension and successor to the AMOS genome assembler.

We recently proposed a method based on numerical sequence representations and digital signal processing data transformation (SPDT) approaches to reduce the size of working

datasets, permitting fast and sensitive read alignment and *de novo* assembly of diverse viral populations (Tapinos et al. 2015). SPDT methods, such as the discrete Fourier transform (DFT) (Agrawal, Faloutsos, and Swami 1993), and discrete wavelet transform (DWT) (Percival and Walden 2006) (Fig. 2), are used to reduce sequences into lower dimensional space, preserving only prominent data characteristics. Analysis is subsequently performed with these lower dimensionality transformations, enabling faster data comparison. Since SPDT methodologies such as the Fourier and wavelet transforms are applicable only to numerical sequences, nucleotide sequences must first be numerically transformed with one of several techniques including real number representations (Chakravarthy et al. 2004), complex number representations (Anastassiou 2001), the DNA walk (Lobry 1996), and the Voss method (Voss 1992).

Although metagenome assemblers generally outperform single genome assemblers in reconstructing different genomes simultaneously, the complexity of this task stipulates their tendency to collapse variation at or beneath strain level into consensus sequences. Even to this end, their effectiveness may be limited as a consequence of extreme variation within specific RNA virus populations due to mutation and recombination, and low and/or uneven sequencing coverage across a particular genome. Furthermore, it should be noted that *de novo* assembly is particularly sensitive to the quality of input sequences, meaning that problems during sample extraction, enrichment and library preparation can be highly detrimental to downstream analyses. Of key importance therefore are quality control methods for detecting, and where appropriate correcting, problems associated with contamination (Darling et al. 2014; Orton et al. 2015), primer read-through and low quality reads (reviewed in Leggett et al. 2013).

5. Haplotype reconstruction in specific viral populations

Viral genomes and metagenomes comprising high intraspecific variation can be challenging targets for assembly, giving rise to complex assembly graphs and fragmented assemblies. This is often the case for clinical samples from HIV and Hepatitis C patients, in which high rates of mutation and long durations of infection can contribute to extreme population divergence, but can also be observed in environmental samples. Where such diversity exists, alignment based probabilistic population reconstruction approaches can be effective, permitting the reconstruction of individual viral variants into 'haplotypes' exceeding read length. This problem has been well studied, and tools such as ShoRAH, QuRE, and PredictHaplo (Giallonardo et al. 2014) are designed for haplotyping viral populations. ShoRAH (Zagordi et al. 2011) extracts local alignments of a specified window length, reconstructs haplotypes for each 'cluster' in that window, and removes mutations from sequences in the cluster not matching the reconstructed haplotype using a model-based probabilistic clustering algorithm. QuRe (Prosperi and Salemi 2012; Prosperi et al. 2013) removes nucleotide substitutions and indels with a Poisson model and reconstructs haplotypes using a heuristic algorithm based on a multinomial distribution. Both approaches have the advantage of reporting probabilities for the reconstructed haplotypes. PredictHaplo is notable for taking into account the read pairing information in Illumina data. A limitation of all of these approaches; however, is their reliance upon a single reference sequence with which to perform the initial alignment, a process which assumes a degree of sequence similarity which may not always be observed in diverse regions, such as regions encoding envelope proteins, of RNA virus genomes. This can be mitigated through construction of a data-specific template through iterative reference mapping and consensus refinement strategies (Archer et al. 2010; Břinda, Boeva, and Kucherov 2016). Other possibilities for broader utility of these approaches include the use of multiple viral reference sequences, either through consideration of multiple linear sequences or by direct alignment of sequences to a variation graph [https://github.com/vgteam/vg], an emerging approach for modeling genomic variation.

6. Sequence classification

Sequence classification is one of the most studied problems in computational biology, and taxonomic assignment is a key objective of metagenome analysis. All classification methods, to some extent, depend upon detecting similarity between a query sequence and a collection of annotated sequences. Classification may be undertaken using either unassembled reads or the reconstructed contigs arising from the assembly process. The computational requirements of available approaches vary dramatically according to their ability to detect homology in divergent sequences; for example, exact *k*-mer matching approaches permit rapid sequence classification, yet typically struggle to identify divergent sequences of viral origin, while high-sensitivity protein alignment searches may be prohibitively slow, especially in application to entire sequencing datasets. Some of the more contemporary and speed-optimized taxonomic assignment approaches also have high RAM requirements, limiting scope for their use with readily available computer hardware. The output of sequence homology search tools is not itself easily interpreted, requiring post-processing in order to yield meaningful classifications. Retroactive taxonomic assignment using these results is non-trivial, requiring additional database lookups, for example, for determination of a conservative 'lowest common ancestor' (LCA) taxon shared by all matches for each query sequence. This kind of complexity necessitates the need for the integration of different tools within application-specific 'pipelines'.

6.1. Sequence similarity searches

Viral identification approaches typically depend on similarity searches against a database using an aligner such as BLAST (Altschul et al. 1990). Comprehensive databases (e.g. GenBank) or smaller custom databases containing for example, only viral sequences of interest may be used, although the latter can generate misleading results. ProViDE (Ghosh et al. 2011) uses virus-specific alignment parameters and thresholds to assign viruses at different taxonomic levels from BLAST matches to a protein database. VIROME (Wommack et al. 2012) is a multifaceted tool integrating results from searches of several sequence and function databases. MEGAN (Huson et al. 2011) is a generally applicable metagenomic classifier, which uses BLAST results to infer the LCA for a given sequence and provides functional analyses through a graphical interface. Automatic pipelines which combine various homology search strategies to identify a final set of viral reads include VirusHunter (Zhao et al. 2013), a Perl script that automates viral identification using BLAST prior to assembly; MetaVir (Roux et al. 2011), a web application that compares users' datasets to published viral sequences; and VirSorter (Roux et al. 2015), which identifies prophages and viruses by comparison with custom datasets. With the exception of web

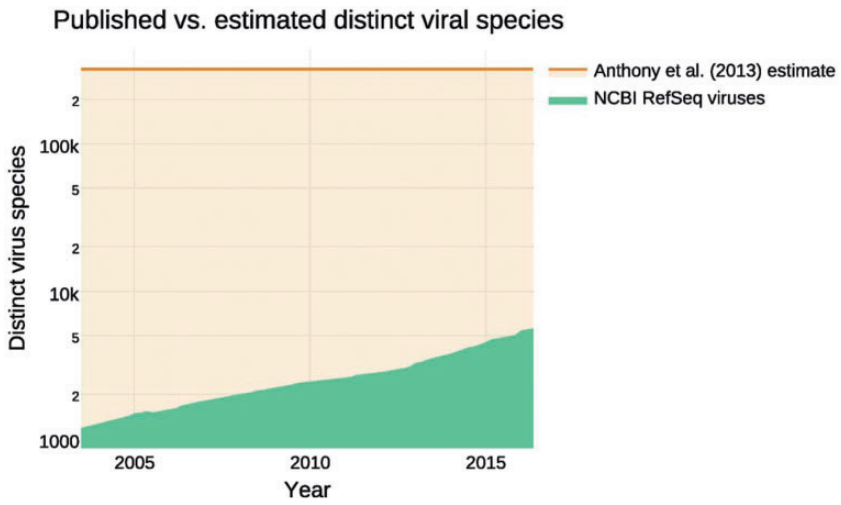

Figure 3. Distinct viral species in the NCBI RefSeq releases from June 2003 – May 2015 (data from ftp://ncbi.nlm.nih.gov/refseq/release/release-statistics/viral.acc_taxid_growth.txt).

applications, however, these are not intuitive tools for the majority of users, requiring manual configuration and installation of software dependencies. Furthermore, similarity search approaches are in general extremely resource-intensive, and performing sensitive BLAST-like database searches with millions of reads is intractable without use of specialist computational resources. To address this problem, tools have emerged leveraging optimized search algorithms and prebuilt databases so as to increase the tractability of classifying millions of reads. For example, Kraken (Wood and Salzberg 2014) and Clark (Ounit et al. 2015) are fast exact *k*-mer matching approaches that use prebuilt databases of viruses, bacteria, human, and fungi, although custom databases may also be built. One Codex is a proprietary web-based metagenome analysis platform with an integrated fast *k*-mer matching engine (similar to that of Kraken) which is both fast, very easy to use, and free for academic use (Minot, Krumm, and Greenfield). Lambda (Hauswedell, Singer, and Reinert 2014) and Diamond (Buchfink, Xie, and Huson 2015) are sensitive and heavily optimized BLAST-like aligners which leverage alphabet reduction to permit protein searches three to five orders of magnitude faster than BLAST, offering prebuilt database indexes for common applications.

6.2 Alternatives to similarity searches

Although exhaustive BLAST-like methods can detect homology in divergent sequences, these methods are in general limited by the relatively few validated viral sequences deposited in public databases, the high diversity within viral families which can obscure relatedness, and the lack of a defined set of core genes common to all viruses that can be used to distinguish species (e.g. the 16S gene for bacteria) (Fancello, Raoult, and Desnues 2012). These features make it difficult to assign similarity thresholds for classification that are applicable to all potential viruses in a sample (Simmonds 2015). Comparison methods that do not rely on sequence similarity include PhyloPythia (McHardy et al. 2007), which uses nucleotide frequencies to classify reads, and PHYMM (Brady and Salzberg 2009), which uses interpolated Markov models to find variable length oligonucleotides that characterize species in the NCBI RefSeq database. Although these approaches are less accurate than BLAST searches, PHYMMBL (Brady and Salzberg 2011) combines PHYMM and BLAST and outperforms either one on its own. Alignment-free comparison approaches, for example, based on dinucleotide frequencies, codon usage patterns, or small but conserved regions of family wide ubiquitous genes, may be more robust to the limitations of the database than sequence similarity searches. These features may also reduce the computation required and highlight evolutionary relationships otherwise obscured by high sequence variability.

A fundamental challenge in the classification of viral sequences with any of these methods remains their limited representation within curated sequence databases. While the rate at which new viruses are being added to NCBI's RefSeq collection has increased considerably, from a year average 0.34 species/day in 2010 to 2.5 species/day in 2015 (Fig. 3), our documented understanding of the extent of viral diversity remains superficial (Anthony et al. 2013). Reads of true viral origin are therefore liable to be missed in many cases. The rate of database growth also highlights the need to maintain frequently updated search indexes for sequence classification, construction of which often demands specialist servers equipped with hundreds of gigabytes of RAM. Even if up-to-date indexes are maintained inside a public repository, their file sizes are substantial, demanding users have access to a fast internet connection. Consequently, complete outsourcing of sequence classification to remote web services is a compelling prospect for those with adequate internet connections but without powerful computing hardware, increasing scope for conducting analyses with portable computers.

7. Conclusion

We see several barriers to realizing the goal of active, on-the-ground surveillance and early detection of viruses with epidemic potential.

1. The emergence of virus-specific assembly and metagenomic tools is a relatively recent phenomenon, with many of the methodologies in use today repurposing one or more existing algorithms. These tools mostly target a small audience

of expert users and, as with most research software, decay after initial release due to a lack of ongoing funding, poor software development practices and/or authors' change of circumstances (Duck et al. 2016). There is a need for a better balance between research software presenting novel methodologies and for sustainably developed, documented and tested software distributed through robust and user friendly channels such as package managers so as to increase the useful life of viral informatics software. Researchers and granting agencies should consider the importance of this step and allocate resources accordingly.

2. Democratisation of routine analyses through development of user friendly, locally installable software and remote web services is critical. Preconfigured cloud virtual machines offer a convenient, low cost way to run analyses, yet must permit straightforward sequence database and software version updates so as to remain relevant after their initial release.
3. Maintaining up to date indexes of large sequence databases is a problem all classification tools must address, stipulating access either to powerful computers for index construction or the ability to download the prebuilt indexes over a fast connection. Furthermore, classification of viral sequences is critically dependent upon the quality of curated viral databases such as RefSeq, to which submitting newly discovered sequences can be prohibitively time consuming. A solution might involve the creation of a central database containing for any given sequencing project both raw reads as well as filtered, assembled and/or annotated reads, and analysed using a single central pipeline. On a regular basis, the database could report sequences and corresponding metadata for unclassified 'dark matter', which is often discarded and yet is likely to contain sequences belonging to novel pathogens. By combining the dark matter from multiple studies, trends within these unclassified reads may be identified that could lead to greater power to identify new biological entities.
4. Benchmarking of software also remains an open problem within the field, which lacks standardized test datasets that are used across multiple studies. Often benchmarking datasets are chosen to highlight the advantages of the method under study, and therefore may be quite specific for a given application. Thus the field needs to agree upon a set of standard, well-characterized reference datasets for virus-focused studies.

The future of the field is promising, with emerging technologies showing potential to eliminate certain challenges. Single molecule sequencing, for example, permits the sequencing of whole viral genomes as single reads, with forthcoming portable and smartphone operated sequencers promising potentially revolutionary analyses in the field. Innovative analytical approaches are constantly being published, and it is evident that the motivation, creativity and expertise needed to meet these challenges exists within the community. Broader communication among developers and end users is essential, and in conjunction with well-funded international initiatives directed at this goal, intelligent viral surveillance could soon be realized.

Acknowledgements

The Virogenesis project receives funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 634650. Bede Constantinides receives funding through a Biotechnology and Biological Sciences Research Council (BBSRC) Doctoral Training Program and Avraam Tapinos receives funding from a BBSRC project grant, BB/M001121/1. We thank Katrina Lithgoe and two anonymous reviewers for their helpful edits and suggestions.

Conflict of interest: None declared.

References

Afiahayati, K, Sato, Y., and Sakakibara, (2015) 'MetaVelvet-SL: An Extension of the Velvet Assembler to a De Novo Metagenomic Assembler Utilizing Supervised Learning', *DNA Research*, 22/1: 69–77.

Agrawal, R C., Faloutsos, A., and Swami, (1993) *Efficient Similarity Search in Sequence Databases*. Heidelberg: Springer.

Altschul, S. F. et al. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215/3: 403–10.

Anastassiou, D. (2001) 'Genomic Signal Processing', *Signal Processing Magazine, IEEE*, 18/4: 8–20.

Anthony, S. J. et al. (2013) 'A Strategy to Estimate Unknown Viral Diversity in Mammals', *MBio*, 4/5: e00598–13

Archer, J. et al. (2010) 'The Evolutionary Analysis of Emerging Low Frequency HIV-1 CXCR4 Using Variants Through Time–An Ultra-Deep Approach', *PLoS Computational Biology*, 6/12: e1001022.

Bankevich, A. et al. (2012) '*SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing*', *Journal of Computational Biology*, 19/5: 455–77.

Berlin, K. et al. (2015) 'Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing', *Nat Biotechnol*, 33/6: 623–30.

Boisvert, S. et al. (2012) 'Ray Meta: Scalable De Novo Metagenome Assembly and Profiling', *Genome Biology*, 13/12: R122.

Brady, A. and Salzberg, S. (2011) 'PhymmBL Expanded: Confidence Scores, Custom Databases, Parallelization and More', *Nature Methods*, 8/5: 367.

—— and —— (2009) 'Phymm and PhymmBL: Metagenomic Phylogenetic Classification with Interpolated Markov Models', *Nature Methods*, 6/9: 673–6.

Břinda, K V., Boeva, G., and Kucherov, (2016) 'Dynamic read mapping and online consensus calling for better variant detection', arXiv:1605.09070v1.

Buchfink, B., Xie, C., and Huson, D. H. (2015) 'Fast and Sensitive Protein Alignment Using DIAMOND', *Nature Methods*, 12/1: 59–60.

Butler, J. et al. (2008) 'ALLPATHS: De Novo Assembly of Whole-Genome Shotgun Microreads', *Genome Research*, 18/5: 810–20.

Carroll, M. W. et al. (2015) 'Temporal and Spatial Analysis of the 2014-2015 Ebola Virus Outbreak in West Africa', *Nature*, 524/7563: 97–101.

Castillo-Chavez, C. et al. (2015) 'Beyond Ebola: Lessons to Mitigate Future Pandemics', *Lancet Global Health*, 3/7: e354–5

Chakravarthy, N. et al. (2004) 'Autoregressive Modeling and Feature Analysis of DNA Sequences', *EURASIP Journal on Applied Signal Processing*, 2004, pp. 13–28.

Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011) 'How to Apply de Bruijn Graphs to Genome Assembly', *Nature Biotechnology*, 29/11: 987–91.

Cotten, M. et al. (2013) 'Full-Genome Deep Sequencing and Phylogenetic Analysis of Novel Human Betacoronavirus', *Emerging Infectious Diseases*, 19/5: 736–42B.

—— et al. (2014) 'Spread, Circulation, and Evolution of the Middle East Respiratory Syndrome Coronavirus', *MBio*, 5/1: e01062–13.

Crusoe, M. R. et al. (2015) 'The khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis', *F1000Res*, 4: 900.

Daly, G. M. et al. (2015) 'Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data', *PLoS One*, 10/6: e0129059.

Darling, A. E. et al. (2014) 'PhyloSift: Phylogenetic Analysis of Genomes and Metagenomes', *PeerJ*, 2: e243.

Delwart, E. L. (2007) 'Viral Metagenomics', *Reviews in Medical Virology*, 17/2: 115–31.

Deng, X. et al. (2015) 'An Ensemble Strategy that Significantly Improves de novo Assembly of Microbial Genomes from Metagenomic Next-Generation Sequencing Data', *Nucleic Acids Research*, 43/7: e46.

Duck, G. et al. (2016) 'A Survey of Bioinformatics Database and Software Usage through Mining the Literature', *PLoS One*, 11/6: e0157989.

Eid, J. et al. (2009) 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Science*, 323/5910: 133–8.

Fancello, L., Raoult, D., and Desnues, C. (2012) 'Computational Tools for Viral Metagenomics and Their Application in Clinical Research', *Virology*, 434/2: 162–74.

Feng, Y. et al. (2015) 'Nanopore-based Fourth-generation DNA Sequencing Technology', *Genomics Proteomics Bioinformatics*, 13/1: 4–16.

Ghosh, T. S. et al. (2011) 'ProViDE: A Software Tool for Accurate Estimation of Viral Diversity In Metagenomic Samples', *Bioinformation*, 6/2: 91–4.

Giallonardo, F. D. et al. (2014) 'Full-Length Haplotype Reconstruction to Infer the Structure of Heterogeneous Virus Populations', *Nucleic Acids Research*, 42/14: e115.

Gire, S. K. et al. (2014) 'Genomic Surveillance Elucidates Ebola Virus Origin and Transmission During the 2014 Outbreak', *Science*, 345/6202: 1369–72.

Grard, G. et al. (2012) 'A Novel Rhabdovirus Associated with Acute Hemorrhagic Fever in Central Africa', *PLoS Pathogens*, 8/9: e1002924.

Haider, B. et al. (2014) '*Omega: An Overlap-Graph De Novo Assembler for Metagenomics*', *Bioinformatics*, 30/19: 2717–22.

Hauswedell, H., Singer, J., and Reinert, K. (2014) 'Lambda: The Local Aligner for Massive Biological Data', *Bioinformatics*, 30/17: i349–55.

Hoenen, T. et al. (2016) 'Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool', *Emerging Infectious Disease*, 22/2: 331–4.

Huang, X. and Madan, A. (1999) 'CAP3: A DNA Sequence Assembly Program', *Genome Research*, 9/9: 868–77

Hunt, M. et al. (2015) 'IVA: Accurate De Novo Assembly of RNA Virus Genomes', *Bioinformatics*, 31/14: 2374–6

Huson, D. H. et al. (2011) 'Integrative Analysis of Environmental Sequences Using MEGAN4', *Genome Research*, 21/9: 1552–60.

Laserson, J., Jojic, V., and Koller, D. (2011) 'Genovo: De Novo Assembly for Metagenomes', *Journal of Computational Biology*, 18/3: 429–43.

Leggett, R. M. et al. (2013) 'Sequencing Quality Assessment Tools to Enable Data-Driven Informatics for High Throughput Genomics', *Frontiers in Genetics*, 4: 288.

Li, D. et al. (2015) 'MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph', *Bioinformatics*, 31/10: 1674–6.

Li, H. (2016) 'Minimap and Miniasm: Fast Mapping and De Novo Assembly for Noisy Long Sequences', *Bioinformatics*, 32/14: 2103–10.

Li, J. Z. et al. (2014) 'Comparison of Illumina and 454 Deep Sequencing in Participants Failing Raltegravir-Based Antiretroviral Therapy', *PLoS One*, 9/3: e90485.

Lobry, J. (1996) 'A Simple Vectorial Representation of DNA Sequences for the Detection of Replication Origins in Bacteria', *Biochimie*, 78/5: 323–6.

Luo, R. et al. (2012) 'SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read De Novo Assembler', *Gigascience*, 1/1: 18.

Madoui, M. A. et al. (2015) 'Genome Assembly using Nanopore-Guided Long and Error-Free DNA Reads', *BMC Genomics*, 16: 327.

McHardy, A. C. et al. (2007) 'Accurate Phylogenetic Classification of Variable-Length DNA Fragments', *Nature Methods*, 4/1: 63–72

Menzel, P., Ng, K. L., and Krogh, A. (2016) 'Fast and sensitive taxonomic classification for metagenomics with Kaiju', *Nature Communications*, 7: 11257.

Minot S. S. N., Krumm N. B., and Greenfield, 'One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification', bioRxiv doi: http://dx.doi.org/10.1101/027607.

Namiki, T. et al. (2012) 'MetaVelvet: An Extension of Velvet Assembler to De Novo Metagenome Assembly from Short Sequence Reads', *Nucleic Acids Research*, 40/20: e155.

Nurk, S. et al. (2016) 'metaSPAdes: A New Versatile De Novo Metagenomics Assembler', arXiv:1604.03071v1.

Orton, R. J. et al. (2015) 'Distinguishing Low Frequency Mutations from RT-PCR and Sequence Errors in Viral Deep Sequencing Data', *BMC Genomics*, 16: 229.

Ounit, R. et al. (2015) 'CLARK: Fast and Accurate Classification of Metagenomic and Genomic Sequences Using Discriminative k-mers', *BMC Genomics*, 16: 236.

Park, D. J. et al. (2015) 'Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone', *Cell*, 161/7: 1516–26.

Peng, Y. et al. (2012) 'IDBA-UD: A De Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth', *Bioinformatics*, 28/11: 1420–8.

Percival, D. B. and Walden, A. T. (2006) *Wavelet Methods for Time Series Analysis*, vol. 4. Cambridge: Cambridge University Press.

Prabhakaran, S. et al. (2014) 'HIV Haplotype Inference Using a Propagating Dirichlet Process Mixture Model', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11/1: 182–91.

Prosperi, M. C. et al. (2013) 'Empirical Validation of Viral Quasispecies Assembly Algorithms: State-of-the-Art and Challenges', *Science Reports*, 3: 2837.

—— and Salemi, M. (2012) 'QuRe: Software for Viral Quasispecies Reconstruction from Next-Generation Sequencing Data', *Bioinformatics*, 28/1: 132–3.

Quick, J. et al. (2015) 'Rapid Draft Sequencing and Real-Time Nanopore Sequencing in a Hospital Outbreak of Salmonella', *Genome Biology*, 16: 114.

—— et al. (2016) 'Real-Time, Portable Genome Sequencing for Ebola Surveillance', *Nature*, 530/7589: 228–32.

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015) 'High-Throughput Sequencing Technologies', *Molecular Cell*, 58/4: 586–97.

Roux, S. et al. (2011) 'Metavir: A Web Server Dedicated to Virome Analysis', *Bioinformatics*, 27/21: 3074–5.

—— et al. (2015) 'VirSorter: Mining Viral Signal from Microbial Genomic Data', *PeerJ*, 3: e985.

Ruby, J. G., Bellare, P., and Derisi, J. L. (2013) 'PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data', *G3 (Bethesda)*, 3/5: 865–80.

Runckel, C. et al. (2011) 'Temporal Analysis of the Honey Bee Microbiome Reveals Four Novel Viruses and Seasonal Prevalence of Known Viruses, Nosema, and Crithidia', *PLoS One*, 6/6: e20656.

Simmonds, P. (2015) 'Methods for Virus Classification and The Challenge of Incorporating Metagenomic Sequence Data', *Journal of General Virology*, 96/Pt 6: 1193–206.

Simpson, J. T. et al. (2009) '*ABySS: A Parallel Assembler for Short Read Sequence Data*', *Genome Research*, 19/6: 1117–23.

Tapinos, A. et al. (2015) Alignment by numbers: sequence assembly using compressed numerical representation. http://biorxiv.org/content/early/2015/01/27/011940.

Treangen, T. J. et al. (2013) 'MetAMOS: A Modular and Open Source Metagenomic Assembly and Analysis Pipeline', *Genome Biology*, 14/1: R2.

Voss, R. F. (1992) 'Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences', *Physical Review Letters*, 68/25: 3805.

Wommack, K. E. et al. (2012) 'VIROME: A Standard Operating Procedure for Analysis of Viral Metagenome Sequences', *Standards in Genomic Sciences*, 6/3: 427–39.

Wood, D. E. and Salzberg, S. L. (2014) 'Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments', *Genome Biology*, 15/3: R46.

Woolhouse, M. E., Rambaut, A., and Kellam, P. (2015) 'Lessons from Ebola: Improving Infectious Disease Surveillance to Inform Outbreak Management', *Science Translational Medicine*, 7/307: 307rv5.

Yang, X. et al. (2012) 'De novo Assembly of Highly Diverse Viral Populations', *BMC Genomics*, 13: 475.

Zagordi, O. et al. (2011) 'ShoRAH: Estimating the Genetic Diversity of a Mixed Sample from Next-Generation Sequencing Data', *BMC Bioinformatics*, 12: 119.

Zaki, A. M. et al. (2012) 'Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia', *New England Journal of Medicine*, 367/19: 1814–20.

Zhao, G. et al. (2013) 'Identification of Novel Viruses Using VirusHunter–An Automated Data Analysis Pipeline', *PLoS One*, 8/10: e78470.