*Rev. Sci. Tech. Off. Int. Epiz.*, 2016, **35** (1), 271-285

# Bioinformatics tools for analysing viral genomic data

R.J. Orton, Q. Gu, J. Hughes, M. Maabar, S. Modha, S.B. Vattipally, G.S. Wilkie & A.J. Davison*

MRC – University of Glasgow Centre for Virus Research, Sir Michael Stoker Building, 464 Bearsden Road, University of Glasgow, Glasgow G61 1QH, United Kingdom
*Corresponding author: andrew.davison@glasgow.ac.uk

**Summary**
The field of viral genomics and bioinformatics is experiencing a strong resurgence due to high-throughput sequencing (HTS) technology, which enables the rapid and cost-effective sequencing and subsequent assembly of large numbers of viral genomes. In addition, the unprecedented power of HTS technologies has enabled the analysis of intra-host viral diversity and quasispecies dynamics in relation to important biological questions on viral transmission, vaccine resistance and host jumping. HTS also enables the rapid identification of both known and potentially new viruses from field and clinical samples, thus adding new tools to the fields of viral discovery and metagenomics. Bioinformatics has been central to the rise of HTS applications because new algorithms and software tools are continually needed to process and analyse the large, complex datasets generated in this rapidly evolving area. In this paper, the authors give a brief overview of the main bioinformatics tools available for viral genomic research, with a particular emphasis on HTS technologies and their main applications. They summarise the major steps in various HTS analyses, starting with quality control of raw reads and encompassing activities ranging from consensus and *de novo* genome assembly to variant calling and metagenomics, as well as RNA sequencing.

**Keywords**
Bioinformatics – *De novo* assembly – High-throughput sequencing – Metagenomics – Next-generation sequencing – Quasispecies – Software – Variant calling – Virus.

## Introduction

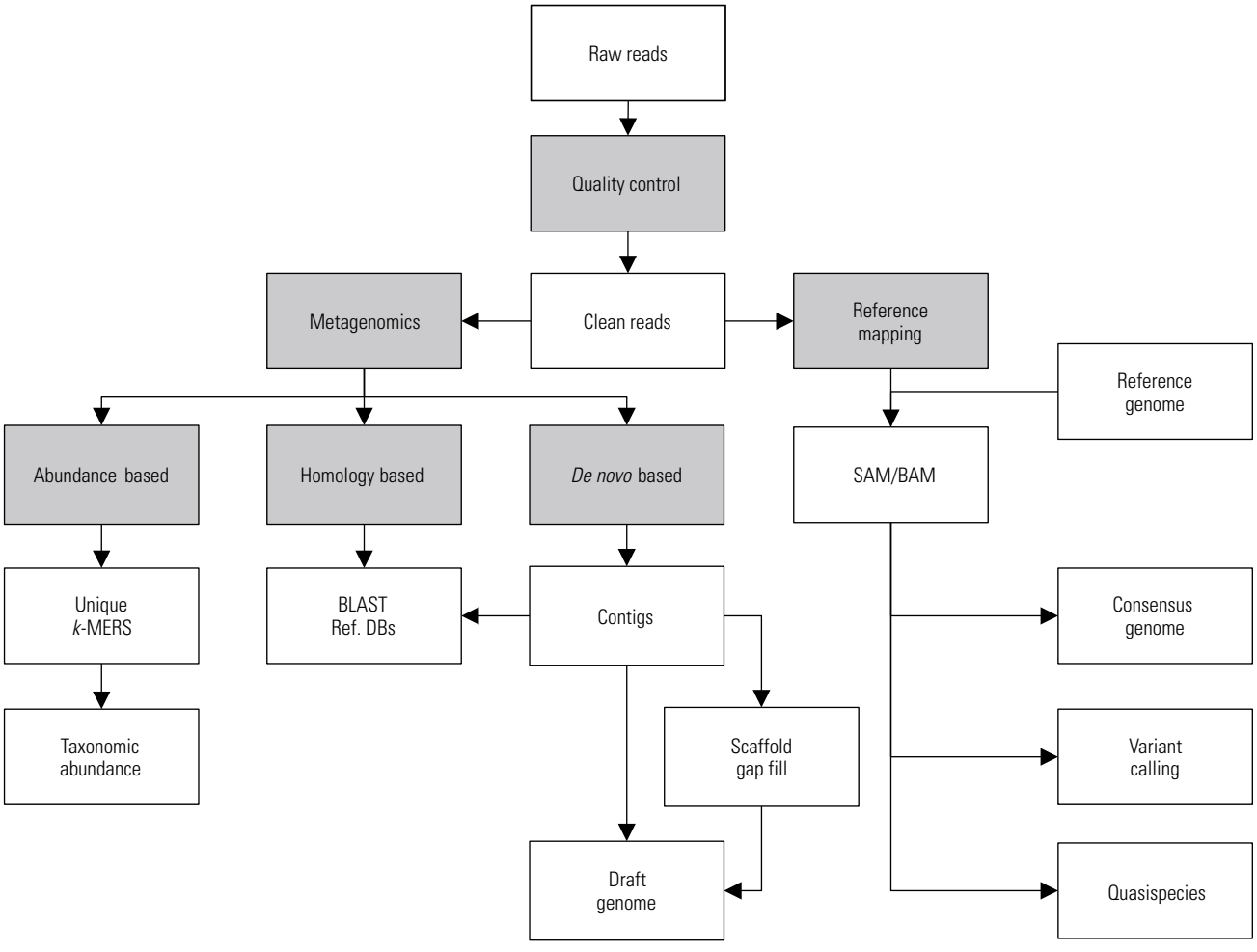
Since the discovery by Ivanovski in 1892 that tobacco mosaic disease is caused and transmitted by fine pore filtrates, viruses have been isolated and characterised from animals, plants, protists, bacteria and even other viruses (1). Viruses have been invaluable model systems in the development of molecular biology and genomics. They can also be highly contagious pathogens with devastating effects on human and animal health, and have consequently been studied in detail for decades. Viruses evolve rapidly because of their large population sizes and high replication rates. RNA viruses have particularly high mutation rates due to the poor fidelity of their RNA polymerases, which enables them to adapt rapidly to new host environments and to selective pressures such as drug treatments (2).

The field of viral genomics and bioinformatics is now experiencing a strong resurgence owing to high-throughput sequencing (HTS) technology, which provides a means for the rapid and cost-effective sequencing and subsequent assembly of large numbers of viral genomes (3, 4). In addition, the unprecedented power of HTS technologies has enabled the analysis of intra-host viral genetic diversity and quasispecies dynamics relevant to important biological questions on viral transmission, vaccine resistance and host jumping. HTS also enables the rapid identification of both known and potentially new viruses (for example,

doi: http://dx.doi.org/10.20506/rst.35.1.2432

Middle East respiratory syndrome coronavirus [5]) from field and clinical samples, thus adding new tools to the fields of viral discovery and metagenomics. Bioinformatics has been central to the rise of HTS because new algorithms and software tools are continually needed to process and analyse the large and complex datasets generated in this rapidly evolving area. However, bioinformatics is not a new field. It has been an integral part of biological research for many years and is routinely used for genome alignment and annotation, and to identify functional motifs, recombination events and phylogenetic relationships. Nonetheless, it is safe to say that HTS has provided a great impetus for the ongoing development of bioinformatics. Most bioinformatics tools are publicly available (although many are limited to UNIX-based operating systems) and utilise common formats that facilitate data exchange and further software development. There are also a number of key database resources that store a vast array of viral genomic and associated meta-data, such as GenBank (6) and the Virus Pathogen Database (7).

In this paper, the authors give a brief overview of the main bioinformatics tools available for viral genomic research, with a particular emphasis on HTS technologies and their major applications. They summarise the major steps involved in various HTS analyses, starting with quality control (QC) of raw reads and encompassing activities ranging from consensus and *de novo* genome assembly to variant calling and metagenomics (as illustrated in Fig. 1). It is not feasible to describe all available bioinformatics tools in this paper; however, readers are directed to Figure 2 for examples of the major tools available for each analytical step.




BAM: binary alignment/map
Ref. DBs: reference databases
SAM: sequence alignment/map

**Fig. 1**
**Flow chart of the major analytical steps in various types of high-throughput sequencing analysis**
The major steps involved in consensus sequencing, *de novo* assembly and metagenomics are illustrated (explained in detail in the text)

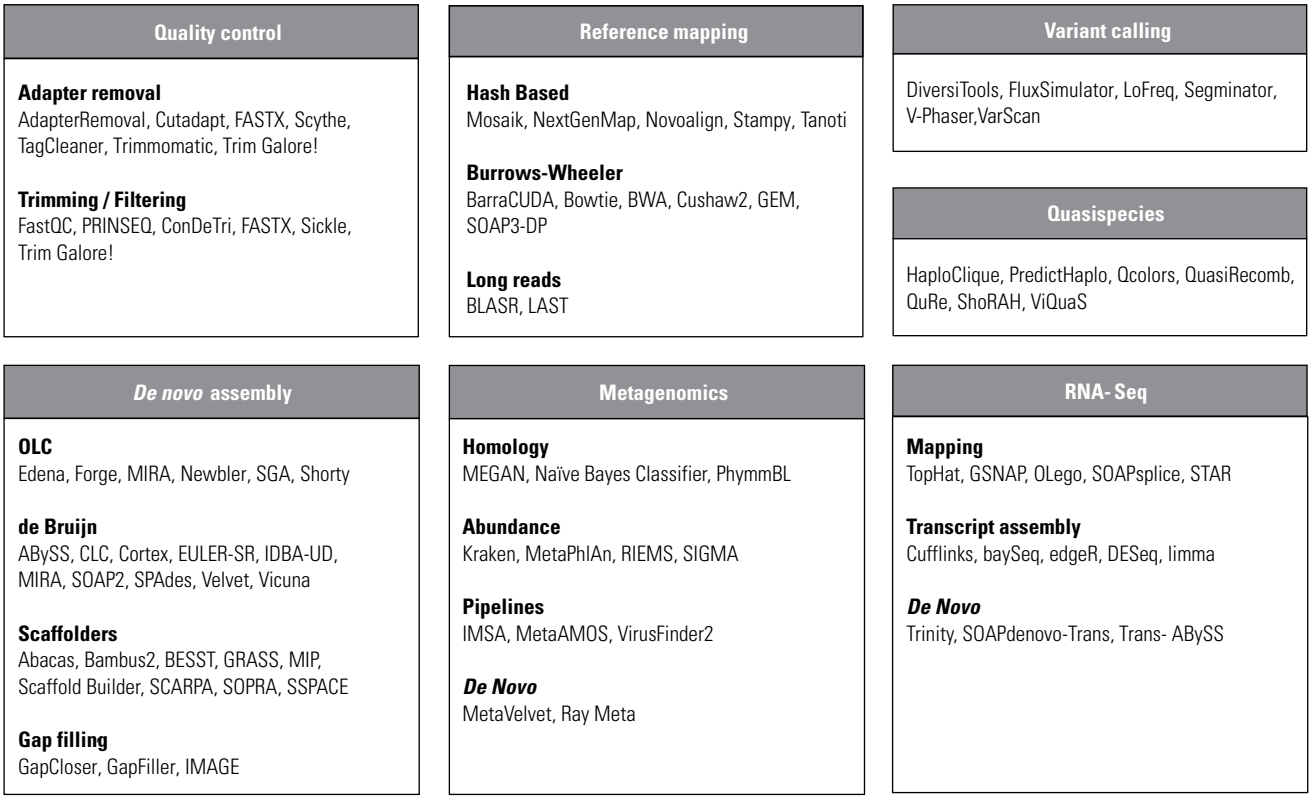



**Fig. 2**
**Bioinformatics tools for viral high-throughput sequencing data**
Examples of the available bioinformatics tools, categorised by the relevant high-throughput sequencing analysis step

# Read quality control

The first step in all HTS analyses is QC. Typically, the output of a sequencing run is a file containing millions of reads that represent DNA sequences originating from the analysed sample. These reads are either output in, or can be readily converted to, the standard FASTQ format (8), which is used for storing biological sequences with their associated quality scores. Sequencing artefacts (e.g. primer/adapter contamination) and sequencing errors (e.g. base miscalls) are common in HTS reads, making QC extremely important for accurate downstream analysis. Adapter sequences vary depending on the library preparation protocol, and these need to be removed because they can hinder the correct mapping of reads and influence single-nucleotide polymorphism calling and other analyses. Two of the most widely used tools for removing adapter sequences are Cutadapt (cutadapt.readthedocs.org/en/stable) and Trimmomatic (9). HTS reads are also usually trimmed to remove poor-quality bases from the ends of reads (typically the 3′ end because quality tends to decrease along the length of the read) and then filtered. Filtering involves the complete removal of some reads from the dataset, such as those of poor average quality or short length, or those containing ambiguous bases. In some analyses (e.g. *de novo* assembly), it can also be beneficial to remove exact read duplicates from the dataset. Two of the most widely used tools for read filtering and trimming are Trim Galore! (www.bioinformatics.babraham.ac.uk/projects/trim_galore) and PRINSEQ (10). It can often be useful to run a host sequence depletion step in which reads are first aligned to the host genome of the sample. Only the unmapped (unaligned) reads are then used for mapping against a viral genome or for *de novo* assembly. An additional QC step can be performed for reads generated by the Ion Torrent and Roche 454 platforms, using tools such as RC454 (11) and Coral (12) to correct for carry forward and incomplete extension errors (CAFIE), particularly at homopolymeric regions. In addition, tools such as PyroCleaner (13), pbh5tools (github.com/PacificBiosciences/pbh5tools) and PoreTools (14) can process 454 (sff format), PacBio (hd5 format) and Oxford Nanopore Technologies (FAST5 format) reads, respectively, in their native formats.

# Mapping and consensus sequence generation

One of the most common HTS applications for viral samples is consensus sequencing of full-length viral genomes. After QC, HTS reads can be mapped to a known reference

genome sequence, which is typically closely related to the genome of interest. Mapping is a critical step in HTS analysis because it determines where each read aligns on the reference genome, and thus affects all downstream analyses, such as variant calling. Most mapping programs are hash-based tools (e.g. Mosaik [15] and Stampy [16]) or Burrows–Wheeler transform (BWT)-based tools (e.g. BWA [17] and Bowtie2 [18]). However, specialist mapping tools are typically needed for longer reads: BLASR (19) and LAST (20) are commonly used for reads generated by PacBio and Oxford Nanopore Technologies, respectively. BWT-based mapping programs can rapidly align reads to a reference genome using low computational resources; in contrast, hash-based programs are more sensitive tools for aligning diverse reads to distantly related reference genomes. This makes reference genome selection an important step: if the reference genome is too distantly related to the sample, mapping programs may struggle to map the majority of reads, resulting in poor or incomplete coverage.

The vast majority of mapping tools utilise the Sequence Alignment/Map (SAM) format (21) to store all read mappings. A SAM file is usually converted into the Binary Alignment/Map (BAM) format, which holds the same data but in a binary format. This makes the file smaller and it is therefore faster to sort and index the reads. The entire BAM alignment of every read to the reference genome can be visualised using tools such as Tablet (22) and IGV (23). This enables users to inspect coverage and variation visually across the genome. SAMtools is a key bioinformatics tools that provides various utilities for manipulating alignments in the SAM/BAM format, including sorting, merging, indexing and generating alignments in a per-position format (21). To call a consensus sequence, one must first identify the nucleotide differences (mutations and indels) in the sample relative to the reference genome. SAMtools can be used in conjunction with VCFtools (24) to identify variants from the reference genome and generate a consensus sequence for the viral sample; an alternative tool for consensus sequence generation is VarScan (25). The consensus sequence is a critical output and can be used for a vast range of subsequent analyses.

# Intra-host viral diversity and quasispecies reconstruction

High-throughput sequencing enables the level of diversity within the whole viral population to be examined and monitored either within (intra) or among (inter) individual hosts in order to investigate evolutionary events such as selection and bottlenecks (Fig. 3). Furthermore, the high sequencing depth of viral samples enables the identification of important variants present at low frequencies within the viral population, such as those that increase pathogenicity or convey drug resistance. For example, HTS has been used to investigate foot and mouth disease virus evolution at the intra- and inter-host levels in a cow transmission chain (26), and to detect high-pathogenicity avian influenza mutations in low-pathogenicity samples from an early epidemic stage (27). However, it is hard to distinguish low-frequency viral variants from errors introduced during sample preparation, such as those originating from reverse transcription or polymerase chain reaction (PCR) (28). More errors are introduced by the sequencer itself in the form of base miscalls, CAFIE errors (29) or systematic errors that occur more frequently around certain motifs, such as GGC and GGX on the Illumina platform (30).

A number of computational tools are available for calling variants at all frequencies from viral samples, such as Lo-Freq (31) and V-Phaser (32), which consider the sources that may have introduced errors. Lo-Freq uses read quality scores to model base miscalls and identify strand-biased variants. In strand bias, a variant is predominantly observed on reads oriented in a single direction, which suggests that the variant is an artefact. It is a characteristic of many systematic errors because the causative sequence motif is not present in both orientations. V-Phaser can potentially detect variants at lower frequencies by utilising information on the co-occurrence of variants on individual reads. However, these tools operate on the basis of certain assumptions and, moreover, do not consider reverse transcription or PCR errors (28). In alternative approaches, variant calling uses modifications to standard protocols, such as circular re-sequencing (33) and incorporating unique barcodes into sample DNA (34, 35). These approaches have been applied successfully to viral population and fitness analyses. Similar circular re-sequencing approaches are also used to correct for sequencing errors in long reads in the SMRTbell technology (developed by PacBio) and in 2D consensus sequences generated by Oxford Nanopore Technologies applications.

Variant callers identify individual variants across the genome but do not identify which variants are located together in individual genomes (unless they are located within a read length of each other). RNA viruses, in particular, have high mutation rates and exist within their hosts as large, complex and heterogeneous populations comprising a spectrum of related but non-identical genome sequences termed the ‘quasispecies’. Viral quasispecies represent a group of interactive genomes rather than a collection of diverse variants, and it has been shown that the quasispecies population, rather than the individual variants, is the target of evolutionary selection (36). Therefore, characterisation of the viral quasispecies and identification of individual viral haplotypes can be a valuable analytical step. Given the short length and error-prone nature of HTS reads, quasispecies reconstruction is computationally challenging. However,

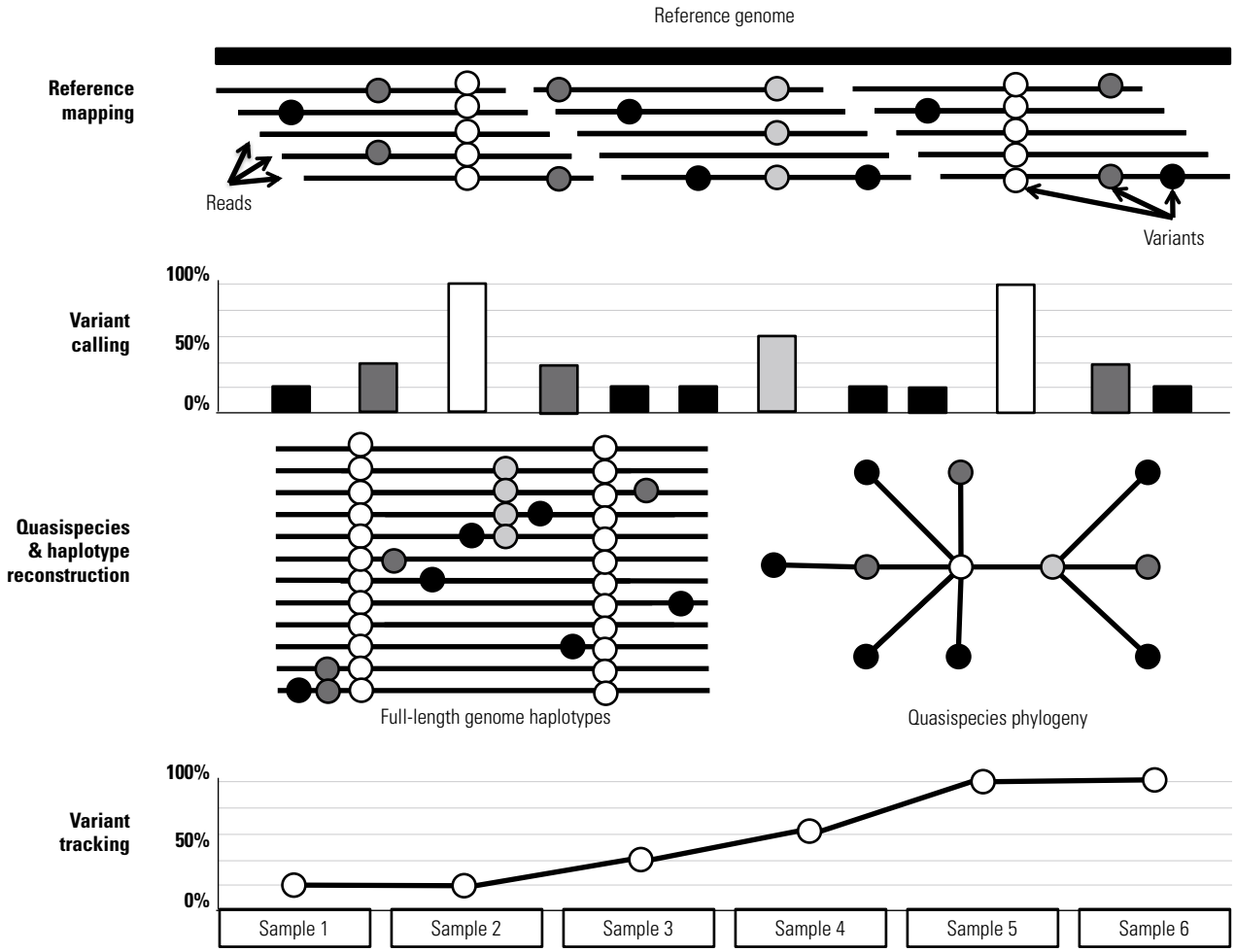



**Fig. 3**
**Schematic diagram of viral population analysis**
The various stages involved in applying high-throughput sequencing technologies to viral populations. Reads are first mapped to a reference sequence (circles, reads containing variants/mutations compared with the reference). Variant frequencies are then calculated (shading represents different frequencies). Subsequent analyses include viral haplotype reconstruction and tracking particular variants of interest through related samples

computational tools such as QuRe (37) and ShoRAH (38) can construct overlapping windows on a genome using read alignments for local haplotype reconstruction, and then collect the results from all individual windows to reconstruct global haplotypes and estimate their frequencies.

# Recombination, phylogeny and selection

High-throughput sequencing applications can go beyond defining the consensus sequence to explore the composition and dynamics of the underlying viral population. However, the consensus sequence remains critical for many analyses, such as those focusing on recombination, phylogenetics and selection. Recombination is the exchange of genetic information between non-segmented viruses. It is therefore a powerful evolutionary process that enables viruses to acquire new genetic combinations, which can assist the process of immune system evasion or cross-species transmission. For example, Western encephalitis virus arose through recombination between a Sindbis-like virus and an Eastern equine encephalitis virus, which could explain its successful establishment and widespread distribution (39). As recombination can lead to the misquantification of selection pressures and phylogenetic estimations, screening for recombination is essential in phylogenetic analyses. Methods for detecting recombination can be broadly

split into four categories (40), all of which have been implemented in a plethora of bioinformatics tools (Fig. 4):

– **distance methods** use genetic differences among sequences at different positions across the genome to identify the presence of recombination

– **phylogenetic methods** explore inconsistencies between the tree topologies of different parts of the genome

– **compatibility methods** are phylogenetic approaches that test on a site-by-site basis whether each site is compatible with the same tree

– **substitution distribution methods** test for the fit to an expected statistical distribution or for significant clustering of substitutions.

Once recombination has been accounted for, one may want to reconstruct the evolutionary and epidemiological dynamics of the non-recombining part of the viral genome. Phylogenetic reconstruction can be either distance based (e.g. neighbour-joining) or character based (e.g. maximum parsimony, maximum likelihood or Bayesian inference). A vast number of tools are available for phylogenetic analysis; it is therefore impractical to list all available resources in this paper. However, readers are directed to a recent review of these methods (41) and a detailed catalogue of the phylogenetic packages maintained by the Felsenstein laboratory (evolution.genetics.washington.edu/phylip/software.html). One tool that has become increasingly popular in recent years is BEAST (42), which uses time-measured phylogenetic trees for Bayesian evolutionary analyses such as coalescent-based population genetics, phylodynamics and phylogeography.

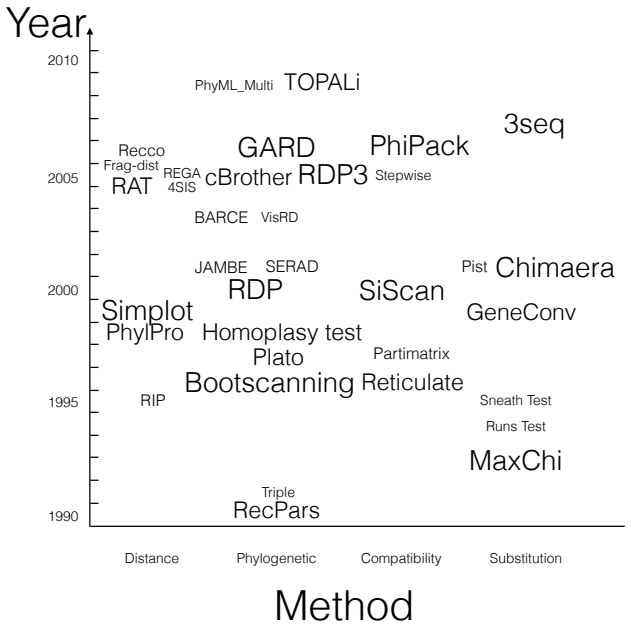



**Fig. 4**

**Overview of tools for analysing recombination**

The plethora of recombination tools is classified according to the method and year of publication. The font size is proportional to the number of citations in Google Scholar

Another important analytical step is to identify the selection pressures that have shaped the molecular evolution of a virus. The methods available for this can be split into three classes (43):

– **counting methods** that enumerate the numbers of non-synonymous and synonymous substitutions along the phylogeny

– **random effects models** that assume a distribution of rates across sites and infer the rates of individual sites according to the distribution

– **fixed effects models** that estimate the rate of non-synonymous to synonymous substitutions on a site-by-site basis.

When a sufficiently large dataset of related sequences is available (>40), the three approaches provide similar results. However, it is advisable to apply all three methods and use the consensus to avoid false results (43). The two main packages that implement these models are PAML (44) and HYPHY (43; the latter is also available on the Datamonkey web server).

# *De novo* genome assembly

If the reference sequence is significantly divergent from the sample or if no reference is available, then it is necessary to generate a consensus sequence by *de novo* assembly of the reads. As an example, this approach was used to construct the genome sequence of elephant endotheliotropic herpesvirus from samples in which as little as 0.04% of the DNA was viral (45). Sequencing errors disrupt the assembly process, so it is essential to trim poor-quality bases and remove adapter sequences from the read datasets before running assemblies. Most algorithms used for *de novo* assembly fall into two groups: overlap layout consensus (OLC) assemblers and de Bruijn graph assemblers. OLC assemblers, such as MIRA (46) and Edena (47), work by first identifying pairs of reads that overlap and then constructing a graph in which reads are represented by nodes in the graph, with overlapping reads connected by edges (lines). The graph is then analysed to find paths through the graph that traverse multiple edges, thus enabling reads to be tiled in the correct

order to generate a genome sequence. However, the OLC approach does not typically scale well because the overlap graph can become very large.

Other assemblers, such as AbySS (48) and Velvet (49), utilise a de Bruijn graph algorithm, which reduces the computational effort by breaking reads into shorter strings of a fixed length *k* (called *k*-mers). The de Bruijn graph captures overlaps of length *k*-1 between these *k*-mers, which avoids the need to calculate overlaps between long sequences. The reads themselves are not modelled but are instead represented by paths through a graph; thus, this approach has proved highly effective. However, there is an upper limit to the length of *k* (generally about 128) that can be handled, even by a powerful computer.

*De novo* assemblies typically consist of a number of long, contiguous sequences (contigs) rather than complete genomes because sequencing errors, repeat regions and areas with low coverage tend to confound the assembly process. Contigs may be joined together to produce a draft genome by alignment to a related viral reference sequence using software such as Abacas (50) or Scaffold Builder (51). If a reference is not available, paired-end reads or mate-pair reads (i.e. long-range read pairs, typically spanning 2–10 kilobases) can be utilised to scaffold the contigs into the correct linear order and produce a draft genome containing gaps. Many *de novo* assembly packages carry out this scaffolding step automatically on assembled contigs when given paired-end read data, but stand-alone scaffolders are also available, including Bambus2 (52) and BESST (53). Alternatively, assemblies made using short-read data, such as those from the Illumina or Ion Torrent platforms, can be improved by using a second dataset with longer reads to join contigs. Gap-filling software (e.g. IMAGE [54] and GapFiller [55]) may be used to close some of the gaps remaining in the draft genome. Their algorithms require paired-end data and identify specific read pairs in which one member matches the end of a contig and the other falls within the gap. Such read subsets are used to extend the contigs iteratively and close gaps by *k*-mer overlap or local *de novo* assembly. However, repeat regions with a period longer than the read length cannot be resolved; these require either PCR followed by Sanger sequencing or data from an HTS platform that yields longer reads.

Since *de novo* assemblers make errors, it is important to check the draft genome, for example by mapping the reads back to the assembled consensus genome and inspecting the alignment for issues such as miscalled bases, short indels and regions with no coverage. Although this process is time-consuming, packages such as ICORN2 (56) have been developed to automate the checking and error correction of viral genomes, while some assemblers (e.g. SPAdes [57]) are capable of carrying out most of the processes described in this section.

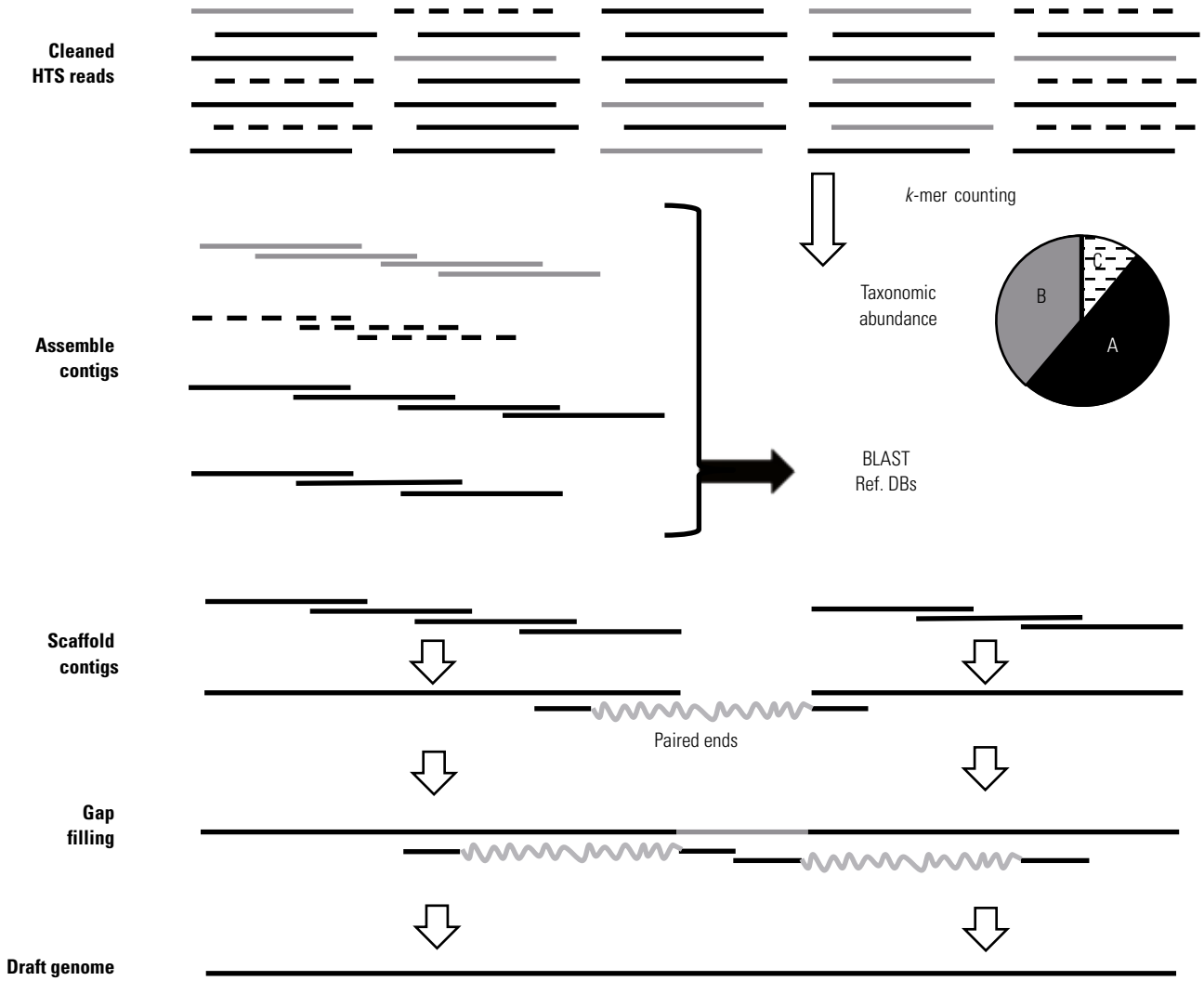
# Metagenomic analyses

Traditional methods of viral detection are based on the isolation and culture of viral pathogens, but often the virus cannot be cultivated under laboratory conditions. This limitation constitutes a substantial barrier to viral discovery (58). Metagenomics can be defined as the sequence-based analysis of the whole collection of viral genomes isolated directly from a sample (Fig. 5) (59). This overcomes the limitation because viral cultivation or prior knowledge of which viruses are present in the sample is not required. However, the method of sample isolation and library preparation can affect the types of virus retrieved (58). One of the main challenges in analysing a metagenomic sample is phylogenetic classification of the raw sequence reads into groups representing the same or similar species. Metagenomics data analysis can be broadly divided into three major approaches based on homology, abundance and *de novo* assembly.

Homology-based approaches classify reads using sequence alignment tools such as BLAST (Basic Local Alignment Search Tool) (60) and BLAT (BLAST-like alignment tool) (61) to align reads directly to reference databases. Although BLAST was not designed for metagenomic sequence classification, tools such as MEGAN (62) and PhymmBL (63) integrate BLAST with Markov models for speed and lowest common ancestor algorithms to assign taxonomic identifications to individual reads and produce summaries at various taxonomic levels.

Abundance-based approaches are much faster, and can be used to obtain summary-level characterisation of the organisms in a given sample. They work by creating a database of known sequences that are representative of particular viruses, for example in the form of clade-specific, short sequences (i.e. *k*-mers). The aim of these methods is to produce an overview of the viral content of the sample by taxonomically identifying and labelling each read using tools such as MetaPhlAn (64) and Kraken (65). Kraken can build a *k*-mer-based database of all known viral genomes by associating unique *k*-mers with individual viral species and higher taxonomic units. This enables it to process a metagenomics dataset rapidly and identify the presence of viral species based on the occurrence of their unique *k*-mers in the reads. Such tools typically utilise dedicated *k*-mer counting tools such as Jellyfish (66) and KAnalyze (67). Tools such as Krona (68) can then be used to visualise the presence and abundance of organisms in the sample.

*De novo* assembly-based approaches can provide a better idea of the breadth and depth of the genome sequences present in a metagenomic dataset. They are typically run as part of an analysis pipeline, such as MetAMOS (69) and VirusFinder2 (70), that also integrates scaffolding and




Ref. DBs: reference databases

**Fig. 5**

**Schematic diagram of high-throughput sequencing metagenomic analysis**

The main stages of a typical HTS metagenomic analysis are illustrated. Shading is used to represent reads originating from different species. Reads can be analysed via abundance-based approaches to determine the relative frequency of each taxon or can be assembled *de novo* to form contigs. Contigs can be scaffolded together using a related reference genome or paired-end reads, which can also be used for gap filling to yield a draft genome. Both contigs and original reads can be 'BLASTed' against reference databases for taxonomic classification

subsequent searching against sequence databases. The resulting scaffolds can then be searched against known sequence databases using similarity-based methods such as BLAST and BLAT to identify the organisms present in the sample or find the most closely related species. For example, the use of HTS metagenomics in the first identification of the Schmallenberg virus (71) demonstrates the power of these approaches.

# RNA-Seq

The transcriptome is defined as the complete, quantified set of transcripts in a single cell or cell population at a specific developmental stage or under specific physiological conditions. Understanding the transcriptome is essential for interpreting the functional elements of a genome, for revealing the molecular constituents of cells and tissues,

and for understanding the processes of development and disease. RNA-Seq, also called whole-transcriptome shotgun sequencing, is a technology that uses the capabilities of HTS to reveal a snapshot of the type and quantities of RNA molecules expressed from a genome at a given moment in time (72). Compared to microarrays, the technology previously used for transcriptomics, RNA-Seq offers increased specificity and sensitivity for the enhanced detection of gene transcripts, and can detect novel transcripts, gene fusions and gene variants. It is important to note that for RNA-Seq experiments replicates are needed to ensure the statistical significance of differences in gene expression (73). Examples of RNA-Seq applied to virology include characterising the response of bovine cells to Schmallenberg virus infection (74) and studying the transcriptomes of viral genomes themselves (75).

RNA-Seq reads must be mapped to a reference genome. However, many reads will fail to map to a standard genome because they span exon junctions. Therefore, novel mapping and downstream analysis tools are needed to handle RNA-Seq data effectively. One of the most commonly used pipelines involves TopHat (76) combined with Cufflinks (77, 78). TopHat aligns reads to the reference genome and detects splice junctions *ab initio* by analysing all fully mapped reads to identify nearby exon splice junctions, and then mapping the reads against these junctions. Cufflinks then uses the mapping data to assemble transcripts, estimate their abundances, and test for differential expression and regulation in RNA-Seq samples. However, numerous alternatives to TopHat and Cufflinks are available, such as STAR (79) and DESeq (80), respectively. If a reference genome is not available, then it is possible to reconstruct a transcriptome *de novo* from RNA-Seq data using tools such as Trinity (81) and SOAPdenovo-Trans (82). Downstream analysis tools such as DAVID (83) can then be used for functional annotation, pathway analysis and gene regulation analysis.

# Discussion

In this paper, the authors have provided a brief overview of the bioinformatics tools available for viral genomic analyses, particularly those that utilise HTS data. They have given exemplar tools and studies at each step, ranging from consensus sequencing and analyses of intra-host diversity to *de novo* assembly and metagenomics, as well as RNA-Seq. However, viral genomics is a broad field and there are areas that could not be covered, such as searching for endogenous viral elements in mammalian genomes (84), detecting novel functional elements (85) and reconstructing transmission chains (86). Additional bioinformatics tools that should be mentioned include Galaxy (87), a user-friendly web-based platform for configuring and running many of the steps and tools described in this paper, and the Genome Analysis Toolkit (88), a software package developed for many different types of HTS data analysis, with a primary focus on human data.

High-throughput sequencing technologies have revolutionised the field of viral genomics by enabling the rapid and cost-effective sequencing of viral genomes in large numbers, the assembly of novel viral genomes, the analysis of viral populations in unprecedented depth and detail, and the detection of viral agents in clinical samples. Bioinformatics is central to the exploitation of HTS, as software tools are needed to manage and analyse the large datasets that this technology produces. Furthermore, new HTS technologies are continually being developed to increase read length and improve quality, which again requires the development of appropriate bioinformatics tools. For example, the MinION (Oxford Nanopore Technologies) offers the potential to sequence entire viral genomes in a single read. Although still in its infancy, this improved technology offers enormous potential for all the areas of viral bioinformatics discussed in this paper. Furthermore, technology is likely to be ultra-portable in the future; therefore, one can envisage its direct use in the field during disease outbreaks to provide clinicians and veterinarians with rapid diagnoses.

## Acknowledgements

The MRC – University of Glasgow Centre for Virus Research is an OIE Collaborating Centre for Viral Genomics and Bioinformatics. This work was supported by Medical Research Council grant numbers MC_UU_12014/3 and G0801822.

# Des outils bio-informatiques pour l'analyse des données de génomique virale

R.J. Orton, Q. Gu, J. Hughes, M. Maabar, S. Modha, S.B. Vattipally, G.S. Wilkie & A.J. Davison

**Résumé**

Le champ de la génomique virale et de la bio-informatique connaît actuellement un nouvel essor grâce à la technologie du séquençage à haut débit (SHD), qui permet de séquencer puis d'assembler rapidement un très grand nombre de génomes viraux, à un coût abordable. De surcroît, grâce à la puissance sans précédent des technologies du SHD, il est désormais possible d'analyser la diversité des virus au sein d'un hôte ainsi que la dynamique des quasi-espèces afin d'élucider d'importantes questions biologiques ayant trait à la transmission virale, à la résistance vis-à-vis des vaccins et au passage d'un hôte à l'autre. Le SHD permet également d'identifier rapidement des virus connus ou potentiellement nouveaux dans des échantillons de terrain ou cliniques, ce qui apporte de nouveaux outils pour la découverte des virus et la métagénomique. La bio-informatique joue un rôle central dans le développement des applications du SHD car ce domaine en constante évolution génère des séries de données aussi nombreuses que complexes dont le traitement et l'analyse requièrent en permanence de nouveaux algorithmes et logiciels. Les auteurs font rapidement le point sur les principaux outils de la bio-informatique utilisés dans la recherche sur les génomes viraux, en mettant particulièrement l'accent sur les technologies du SHD et sur leurs applications les plus importantes. Ils décrivent schématiquement les grandes étapes de différents types d'analyse recourant au SHD, depuis le contrôle qualité des lectures brutes jusqu'aux activités telles que l'assemblage de séquences consensus et *de novo* du génome, l'appel de variants et la métagénomique, et enfin le séquençage d'ARN.

**Mots-clés**

Appel de variants – Assemblage *de novo* – Bio-informatique – Logiciel – Métagénomique – Quasi-espèces – Séquençage à haut débit – Séquençage de nouvelle génération – Virus.

# Herramientas de bioinformática para analizar datos de genómica vírica

R.J. Orton, Q. Gu, J. Hughes, M. Maabar, S. Modha, S.B. Vattipally, G.S. Wilkie & A.J. Davison

**Resumen**

El campo de la genómica vírica y la bioinformática conoce hoy un renovado dinamismo gracias a las técnicas de secuenciación de alto rendimiento, que permiten secuenciar con rapidez y rentabilidad, y a continuación ensamblar, un gran número de genomas víricos. Además, la potencia sin precedentes que

ofrecen estas técnicas ha hecho posible analizar la diversidad vírica dentro de los anfitriones y la dinámica de cuasiespecies en relación con importantes interrogantes biológicos tocantes a la transmisión de virus, la resistencia a las vacunas o el salto de un anfitrión a otro. Con la secuenciación de alto rendimiento también es posible identificar con celeridad los virus tanto conocidos como eventualmente nuevos que estén presentes en muestras clínicas u obtenidas sobre el terreno, lo que aporta nuevas herramientas al arsenal disponible en los campos del descubrimiento de virus y la metagenómica. La bioinformática ha sido un factor capital en el auge de las aplicaciones de técnicas de secuenciación de alto rendimiento, pues continuamente se necesitan nuevos algoritmos y programas informáticos para procesar y analizar los vastos y complejos conjuntos de datos que se generan en un ámbito sujeto a tan rápida evolución. Los autores repasan brevemente las principales herramientas bioinformáticas que existen para la investigación en genómica vírica, prestando especial atención a las técnicas de secuenciación de alto rendimiento y sus principales aplicaciones. Asimismo, resumen las etapas básicas de diversos procedimientos de análisis por secuenciación de alto rendimiento, empezando por el control de calidad de las lecturas brutas y pasando por labores que van desde el ensamblaje del genoma con creación de secuencia consenso o ensamblaje *de novo* hasta la asignación de variantes (*variant calling*) o la metagenómica, sin olvidar la secuenciación de ARN.

**Palabras clave**
Asignación de variantes – Bioinformática – Cuasiespecies – Ensamblaje *de novo* – Metagenómica – Programas informáticos – Secuenciación de alto rendimiento – Secuenciación de próxima generación – Virus.

# References

1. Seto D. (2010). – Viral genomics and bioinformatics. *Viruses*, **2** (12), 2587–2593. doi:10.3390/v2122587.

2. Holland J., Spindler K., Horodyski F., Grabau E., Nichol S. & VandePol S. (1982). – Rapid evolution of RNA genomes. *Science*, **215** (4540), 1577–1585.

3. Orton R.J., Wright C.F., Morelli M.J., Juleff N., Thébaud G., Knowles N.J., Valdazo-González B., Paton D.J., King D.P. & Haydon D.T. (2013). – Observing micro-evolutionary processes of viral populations at multiple scales. *Philos. Trans. Roy. Soc. Lond., B, Biol. Sci.*, **368** (1614), 20120203. doi:10.1098/rstb.2012.0203.

4. Van Borm S., Belák S., Freimanis G., Fusaro A., Granberg F., Höper D., King D.P., Monne I., Orton R. & Rosseel T. (2015). – Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases? *Meth. Molec. Biol.*, **1247**, 415–436. doi:10.1007/978-1-4939-2004-4_30.

5. Zaki A.M., van Boheemen S., Bestebroer T.M., Osterhaus A.D. & Fouchier R.A. (2012). – Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.*, **367** (19), 1814–1820. doi:10.1056/NEJMoa1211721.

6. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J. & Wheeler D.L. (2005). – GenBank. *Nucleic Acids Res.*, **33** (Suppl. 1), D34–D38. doi:10.1093/nar/gki063.

7. Pickett B.E., Greer D.S., Zhang Y., Stewart L., Zhou L., Sun G., Gu Z., Kumar S., Zaremba S., Larsen C.N., Jen W., Klem E.B. & Scheuermann R.H. (2012). – Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses*, **4** (11), 3209–3226. doi:10.3390/v4113209.

8. Cock P.J., Fields C.J., Goto N., Heuer M.L. & Rice P.M. (2010). – The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38** (6), 1767–1771. doi:10.1093/nar/gkp1137.

9. Bolger A.M., Lohse M. & Usadel B. (2014). – Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30** (15), 2114–2120. doi:10.1093/bioinformatics/btu170.

10. Schmieder R. & Edwards R. (2011). – Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27** (6), 863–864. doi:10.1093/bioinformatics/btr026.

11. Henn M.R., Boutwell C.L., Charlebois P., Lennon N.J., Power K.A., Macalalad A.R., Berlin A.M., Malboeuf C.M., Ryan E.M., Gnerre S., Zody M.C., Erlich R.L., Green L.M., Berical A., Wang Y., Casali M., Streeck H., Bloom A.K., Dudek T., Tully D., Newman R., Axten K.L., Gladden A.D., Battis L., Kemper M., Zeng Q., Shea T.P., Gujja S., Zedlack C., Gasser O., Brander C., Hess C., Gunthard H.F., Brumme Z.L., Brumme C.J., Bazner S., Rychert J., Tinsley J.P., Mayer K.H., Rosenberg E., Pereyra F., Levin J.Z., Young S.K., Jessen H., Altfeld M., Birren B.W., Walker B.D. & Allen T.M. (2012). – Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.*, **8** (3), e1002529. doi:10.1371/journal.ppat.1002529.

12. Salmela L. & Schroder J. (2011). – Correcting errors in short reads by multiple alignments. *Bioinformatics*, **27** (11), 1455–1461. doi:10.1093/bioinformatics/btr170.

13. Jerome M., Noirot C. & Klopp C. (2011). – Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Res. Notes*, **4**, 149. doi:10.1186/1756-0500-4-149.

14. Loman N.J. & Quinlan A.R. (2014). – Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, **30** (23), 3399–3401. doi:10.1093/bioinformatics/btu555.

15. Lee W.P., Stromberg M.P., Ward A., Stewart C., Garrison E.P. & Marth G.T. (2014). – MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*, **9** (3), e90581. doi:10.1371/journal.pone.0090581.

16. Lunter G. & Goodson M. (2011). – Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.*, **21** (6), 936–939. doi:10.1101/gr.111120.110.

17. Li H. & Durbin R. (2009). – Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25** (14), 1754–1760. doi:10.1093/bioinformatics/btp324.

18. Langmead B., Trapnell C., Pop M. & Salzberg S.L. (2009). – Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10** (3), R25. doi:10.1186/gb-2009-10-3-r25.

19. Chaisson M.J. & Tesler G. (2012). – Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238. doi:10.1186/1471-2105-13-238.

20. Frith M.C., Hamada M. & Horton P. (2010). – Parameters for accurate genome alignment. *BMC Bioinformatics*, **11**, 80. doi:10.1186/1471-2105-11-80.

21. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. & the 1000 Genomes Project Data Processing Subgroup (2009). – The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25** (16), 2078–2079. doi:10.1093/bioinformatics/btp352.

22. Milne I., Stephen G., Bayer M., Cock P.J., Pritchard L., Cardle L., Shaw P.D. & Marshall D. (2013). – Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform.*, **14** (2), 193–202. doi:10.1093/bib/bbs012.

23. Thorvaldsdottir H., Robinson J.T. & Mesirov J.P. (2013). – Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.*, **14** (2), 178–192. doi:10.1093/bib/bbs017.

24. Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. & the 1000 Genomes Project Analysis Group (2011). – The variant call format and VCFtools. *Bioinformatics*, **27** (15), 2156–2158. doi:10.1093/bioinformatics/btr330.

25. Koboldt D.C., Chen K., Wylie T., Larson D.E., McLellan M.D., Mardis E.R., Weinstock G.M., Wilson R.K. & Ding L. (2009). – VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25** (17), 2283–2285. doi:10.1093/bioinformatics/btp373.

26. Morelli M.J., Wright C.F., Knowles N.J., Juleff N., Paton D.J., King D.P. & Haydon D.T. (2013). – Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Vet. Res.*, **44**, 12. doi:10.1186/1297-9716-44-12.

27. Monne I., Fusaro A., Nelson M.I., Bonfanti L., Mulatti P., Hughes J., Murcia P.R., Schivo A., Valastro V., Moreno A., Holmes E.C. & Cattoli G. (2014). – Emergence of a highly pathogenic avian influenza virus from a low-pathogenic progenitor. *J. Virol.*, **88** (8), 4375–4388. doi:10.1128/JVI.03181-13.

28. Orton R.J., Wright C.F., Morelli M.J., King D.J., Paton D.J., King D.P. & Haydon D.T. (2015). – Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics*, **16**, 229. doi:10.1186/s12864-015-1456-x.

29. Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L., Jarvie T.P., Jirage K.B., Kim J.B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F. & Rothberg J.M. (2005). – Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437** (7057), 376–380. doi:10.1038/nature03959.

30. Meacham F., Boffelli D., Dhahbi J., Martin D.I., Singer M. & Pachter L. (2011). – Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451. doi:10.1186/1471-2105-12-451.

31. Wilm A., Aw P.P., Bertrand D., Yeo G.H., Ong S.H., Wong C.H., Khor C.C., Petric R., Hibberd M.L. & Nagarajan N. (2012). – LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.*, **40** (22), 11189–11201. doi:10.1093/nar/gks918.

32. Yang X., Charlebois P., Macalalad A., Henn M.R. & Zody M.C. (2013). – V-Phaser 2: variant inference for viral populations. *BMC Genomics*, **14**, 674. doi:10.1186/1471-2164-14-674.

33. Acevedo A., Brodsky L. & Andino R. (2014). – Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, **505** (7485), 686–690. doi:10.1038/nature12861.

34. Wu N.C., Young A.P., Al-Mawsawi L.Q., Olson C.A., Feng J., Qi H., Chen S.H., Lu I.H., Lin C.Y., Chin R.G., Luan H.H., Nguyen N., Nelson S.F., Li X., Wu T.T. & Sun R. (2014). – High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Sci. Rep.*, **4**, 4942. doi:10.1038/srep04942.

35. Mangul S., Wu N.C., Mancuso N., Zelikovsky A., Sun R. & Eskin E. (2014). – Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics*, **30** (12), i329–i337. doi:10.1093/bioinformatics/btu295.

36. Vignuzzi M., Stone J.K., Arnold J.J., Cameron C.E. & Andino R. (2006). – Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, **439** (7074), 344–348. doi:10.1038/nature04388.

37. Prosperi M.C. & Salemi M. (2012). – QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, **28** (1), 132–133. doi:10.1093/bioinformatics/btr627.

38. Zagordi O., Bhattacharya A., Eriksson N. & Beerenwinkel N. (2011). – ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, **12**, 119. doi:10.1186/1471-2105-12-119.

39. Hahn C.S., Lustig S., Strauss E.G. & Strauss J.H. (1988). – Western equine encephalitis virus is a recombinant virus. *Proc. Natl Acad. Sci. USA*, **85** (16), 5997–6001.

40. Posada D. (2002). – Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.*, **19** (5), 708–717.

41. Yang Z. & Rannala B. (2012). – Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.*, **13** (5), 303–314. doi:10.1038/nrg3186.

42. Drummond A.J., Suchard M.A., Xie D. & Rambaut A. (2012). – Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29** (8), 1969–1973. doi:10.1093/molbev/mss075.

43. Kosakovsky Pond S.L. & Frost S.D. (2005). – Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.*, **22** (5), 1208–1222. doi:10.1093/molbev/msi105.

44. Yang Z. (2007). – PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24** (8), 1586–1591. doi:10.1093/molbev/msm088.

45. Wilkie G.S., Davison A.J., Watson M., Kerr K., Sanderson S., Bouts T., Steinbach F. & Dastjerdi A. (2013). – Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases. *J. Virol.*, **87** (12), 6700–6712. doi:10.1128/JVI.00655-13.

46. Chevreux B., Wetter T. & Suhai S. (1999). – Genome sequence assembly using trace signals and additional sequence information. *In* Computer science and biology. Proceedings of the German Conference on Bioinformatics, 4–6 October, Hanover. German Conference on Bioinformatics, Vol. 99, 45–56. Available at: www.bioinfo.de/isb/gcb99/talks/chevreux/ (accessed on 18 April 2016).

47. Hernandez D., Francois P., Farinelli L., Østerås M. & Schrenzel J. (2008). – *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.*, **18** (5), 802–809. doi:10.1101/gr.072033.107.

48. 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** (6), 1117–1123. doi:10.1101/gr.089532.108.

49. Zerbino D.R. & Birney E. (2008). – Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, **18** (5), 821–829. doi:10.1101/gr.074492.107.

50. Assefa S., Keane T.M., Otto T.D., Newbold C. & Berriman M. (2009). – ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, **25** (15), 1968–1969. doi:10.1093/bioinformatics/btp347.

51. Silva G.G., Dutilh B.E., Matthews T.D., Elkins K., Schmieder R., Dinsdale E.A. & Edwards R.A. (2013). – Combining *de novo* and reference-guided assembly with scaffold_builder. *Source Code Biol. Med.*, **8** (1), 23. doi:10.1186/1751-0473-8-23.

52. Koren S., Treangen T.J. & Pop M. (2011). – Bambus 2: scaffolding metagenomes. *Bioinformatics*, **27** (21), 2964–2971. doi:10.1093/bioinformatics/btr520.

53. Sahlin K., Vezzi F., Nystedt B., Lundeberg J. & Arvestad L. (2014). – BESST: efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, **15,** 281. doi:10.1186/1471-2105-15-281.

54. Tsai I.J., Otto T.D. & Berriman M. (2010). – Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.*, **11** (4), R41. doi:10.1186/gb-2010-11-4-r41.

55. Nadalin F., Vezzi F. & Policriti A. (2012). – GapFiller: a *de novo* assembly approach to fill the gap within paired reads. *BMC Bioinformatics*, **13 (**Suppl. 14), S8. doi:10.1186/1471-2105-13-S14-S8.

56. Otto T.D., Sanders M., Berriman M. & Newbold C. (2010). – Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*, **26** (14), 1704–1707. doi:10.1093/bioinformatics/btq269.

57. Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A. & Pevzner P.A. (2012). – SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19** (5), 455–477. doi:10.1089/cmb.2012.0021.

58. Fancello L., Raoult D. & Desnues C. (2012). – Computational tools for viral metagenomics and their application in clinical research. *Virology*, **434** (2), 162–174. doi:10.1016/j.virol.2012.09.025.

59. Höper D., Mettenleiter T.C. & Beer M. (2016). – Metagenomic approaches to identify infectious agents. *In* Potential applications of pathogen genomics (P.R. Murcia, M. Palmarini & S. Belák, eds). *Rev. Sci. Tech. Off. Int. Epiz.*, **35** (1), 83–93.

60. Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990). – Basic local alignment search tool. *J. Molec. Biol.*, **215** (3), 403–410. doi:10.1016/S0022-2836(05)80360-2.

61. Kent W.J. (2002). – BLAT: the BLAST-like alignment tool. *Genome Res.*, **12** (4), 656–664. doi:10.1101/gr.229202.

62. Huson D.H., Auch A.F., Qi J. & Schuster S.C. (2007). – MEGAN analysis of metagenomic data. *Genome Res.*, **17** (3), 377–386. doi:10.1101/gr.5969107.

63. Brady A. & Salzberg S.L. (2009). – Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods*, **6** (9), 673–676. doi:10.1038/nmeth.1358.

64. Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O. & Huttenhower C. (2012). – Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9** (8), 811–814. doi:10.1038/nmeth.2066.

65. Wood D.E. & Salzberg S.L. (2014). – Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15** (3), R46. doi:10.1186/gb-2014-15-3-r46.

66. Marçais G. & Kingsford C. (2011). – A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, **27** (6), 764–770. doi:10.1093/bioinformatics/btr011.

67. Audano P. & Vannberg F. (2014). – KAnalyze: a fast versatile pipelined K-mer toolkit. *Bioinformatics*, **30** (14), 2070–2072. doi:10.1093/bioinformatics/btu152.

68. Ondov B.D., Bergman N.H. & Phillippy A.M. (2011). – Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385. doi:10.1186/1471-2105-12-385.

69. Treangen T.J., Koren S., Sommer D.D., Liu B., Astrovskaya I., Ondov B., Darling A.E., Phillippy A.M. & Pop M. (2013). – MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.*, **14** (1), R2. doi:10.1186/gb-2013-14-1-r2.

70. Wang Q., Jia P. & Zhao Z. (2015). – VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med.*, **7** (1), 2. doi:10.1186/s13073-015-0126-6.

71. Hoffmann B., Scheuch M., Höper D., Jungblut R., Holsteg M., Schirrmeier H., Eschbaumer M., Goller K.V., Wernike K., Fischer M., Breithaupt A., Mettenleiter T.C. & Beer M. (2012). – Novel orthobunyavirus in cattle, Europe, 2011. *Emerg. Infect. Dis.*, **18** (3), 469–472. doi:10.3201/eid1803.111905.

72. Wang Z., Gerstein M. & Snyder M. (2009). – RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10** (1), 57–63. doi:10.1038/nrg2484.

73. Liu Y., Zhou J. & White K.P. (2014). – RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30** (3), 301–304. doi:10.1093/bioinformatics/btt688.

74. Blomström A.L., Gu Q., Barry G., Wilkie G., Skelton J.K., Baird M., McFarlane M., Schnettler E., Elliott R.M., Palmarini M. & Kohl A. (2015). – Transcriptome analysis reveals the host response to Schmallenberg virus in bovine cells and antagonistic effects of the NSs protein. *BMC Genomics*, **16**, 324. doi:10.1186/s12864-015-1538-9.

75. Gatherer D., Seirafian S., Cunningham C., Holton M., Dargan D.J., Baluchova K., Hector R.D., Galbraith J., Herzyk P., Wilkinson G.W. & Davison A.J. (2011). – High-resolution human cytomegalovirus transcriptome. *Proc. Natl Acad. Sci. USA*, **108** (49), 19755–19760. doi:10.1073/pnas.1115861108.

76. Trapnell C., Pachter L. & Salzberg S.L. (2009). – TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25** (9), 1105–1111. doi:10.1093/bioinformatics/btp120.

77. Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., van Baren M.J., Salzberg S.L., Wold B.J. & Pachter L. (2010). – Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28** (5), 511–515. doi:10.1038/nbt.1621.

78. Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D.R., Pimentel H., Salzberg S.L., Rinn J.L. & Pachter L. (2012). – Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7** (3), 562–578. doi:10.1038/nprot.2012.016.

79. Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M. & Gingeras T.R. (2013). – STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29** (1), 15–21. doi:10.1093/bioinformatics/bts635.

80. Anders S. & Huber W. (2010). – Differential expression analysis for sequence count data. *Genome Biol.*, **11** (10), R106. doi:10.1186/gb-2010-11-10-r106.

81. Haas B.J., Papanicolaou A., Yassour M., Grabherr M., Blood P.D., Bowden J., Couger M.B., Eccles D., Li B., Lieber M., Macmanes M.D., Ott M., Orvis J., Pochet N., Strozzi F., Weeks N., Westerman R., William T., Dewey C.N., Henschel R., Leduc R.D., Friedman N. & Regev A. (2013). – *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8** (8), 1494–1512. doi:10.1038/nprot.2013.084.

82. Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S., Huang W., He G., Gu S., Li S., Zhou X., Lam T.W., Li Y., Xu X., Wong G.K. & Wang J. (2014). – SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30** (12), 1660–1666. doi:10.1093/bioinformatics/btu077.

83. Huang D.W., Sherman B.T. & Lempicki R.A. (2009). – Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4** (1), 44–57. doi:10.1038/nprot.2008.211.

84. Arriagada G. & Gifford R.J. (2014). – Parvovirus-derived endogenous viral elements in two South American rodent genomes. *J. Virol.*, **88** (20), 12158–12162. doi:10.1128/JVI.01173-14.

85. Firth A.E. (2014). – Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res.*, **42** (20), 12425–12439. doi:10.1093/nar/gku981.

86. Morelli M.J., Thébaud G., Chadœuf J., King D.P., Haydon D.T. & Soubeyrand S. (2012). – A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.*, **8** (11), e1002768. doi:10.1371/journal.pcbi.1002768.

87. Goecks J., Nekrutenko A., Taylor J. & the Galaxy Team (2010). – Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11** (8), R86. doi:10.1186/gb-2010-11-8-r86.

88. McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. & DePristo M.A. (2010). – The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20** (9), 1297–1303. doi:10.1101/gr.107524.110.