Chapter 6

Analysis of Next-generation Sequencing Data in Virology - Opportunities and Challenges

Sunitha M. Kasibhatla, Vaishali P. Waman, Mohan M. Kale and Urmila Kulkarni-Kale

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61610

Abstract

Viruses are the most abundant and the smallest organisms, which are relatively simple to sequence. Genome sequence data of viruses for individual species to populations outnumber that of other species. Although this offers an opportunity to study viral diversity at varying levels of taxonomic hierarchy, it also poses challenges for systematic and structured organization of data and its downstream processing. Extensive computational analyses using a number of algorithms and programs have opened exciting opportunities for virus discovery and diagnostics, apart from augmenting our understanding of the intriguing world of viruses. Unravelling evolutionary dynamics of viruses permits improved understanding of phenomena such as quasispecies diversity, role of mutations in host switching and drug resistance, which enables the tangible measurements of genotype and phenotype of viruses. Improved understanding of geno-/serotype diversity in correlation with antigenic diversity will facilitate rational design and development of efficacious vaccines against emerging and re-emerging viruses. Mathematical models developed using the genomic data could be used to predict the spread of viruses due to vector switching and the (re)emergence due to host switching and, thereby, contribute towards designing public health policies for disease management and control.

Keywords: Virus/viral evolution, population diversity, recombination, selection pressure, phylogeny and typing

1. Introduction

1.1. Viruses: Special class of organisms

Viruses form a major class of biological entities encompassing diverse environments ranging from algae in marine ecosystems to soil, plant, human and animal systems. Several metage-

nomic studies have revealed the possibility of viruses being the dominant species of our biosphere [1]. Deep sequencing efforts have shown that viruses form 10^6–10^9 particles per millilitre of seawater [2]. It is also interesting to note that ~90% of the reads obtained from such experiments did not encode proteins, which are reported in other organisms, including viruses, that have been characterised so far. This clearly demonstrates that the actual viral diversity has not been sampled in an adequate manner so far. A crucial aspect of viral studies is the disease burden associated with them, which is known to be enormous with serious economic implications. World Health Organization documents that the global burden of communicable diseases (of which viral diseases form a major chunk) is ~15 million annually [3].

Beyond abundance aspects, study of viral evolution and genetic variations enabled the proposal of the virocentric standpoint of the evolution. Viruses gained centre stage for reasons such as being smallest replicating entities, having short generation time, large population sizes and high replication and mutation rates. Attributes such as variation in genome sizes, gene pool, shape and assembly of particles are responsible for viruses to attain pivotal role in the study of evolution [4]. It has been observed that all plausible replication and expression strategies have been employed by viruses to dynamically adapt to the ever-changing environments. Processes like complementation, recombination, reassortment, high mutation rate and existence as quasispecies enable the viruses to outgrow and outcompete the host immune system. The molecular forces driving these processes can be delineated by sequencing and the subsequent analyses.

1.2. Viral sequencing methods

The distinction of complete genome ever to be sequenced belongs to bacteriophage ΦX174 with a genome size of 5,386 bases and was achieved through the Sanger's shotgun-sequencing approach [5]. The major aim of early sequencing projects was to characterize the genomic content of an organism in terms of its coding potential. Over the last few years, the unprecedented growth in the area of sequencing technologies has had a huge impact on the way viral genomes are being addressed. The scale of generating and handling data, which was unimaginable previously, has become a reality today due to the advent of Next-Generation Sequencing (NGS) technologies. Advantages of NGS over the conventional Sanger sequencing approach are the rapid generation of sequencing data on a very massive scale and at affordable cost. NGS also provides scope for wide range of studies that include transcriptomics, gene expression and regulation (DNA–protein interaction), single-nucleotide polymorphism (SNP) and RNA profiling. Sequencing of viruses, in particular, has been important to understand the spread of epidemics, the circulating viral particles and the improvement of strains for vaccine design. Different technologies such as Roche 454 [6], Illumina [7], Ion Torrent [8] and more recently the fourth-generation sequencing methodologies popularly called single-cell sequencing, *viz.* Oxford Nanopore [9] and Pacific Biosciences [10], are available for sequencing.

Sample preparation and enrichment are the prerequisites for sequencing the viromes. Filtration and centrifugation on caesium chloride density gradient have proved to enrich

http://dx.doi.org/10.5772/61610

the virus-like particles. A strategy like depletion of host rRNAs is also known to increase the virus fraction and has been attributed to the discovery of several novel RNA viruses [11]. In plant virology, use of CF11 cellulose spin column is routinely used for deep sequencing of dsRNA.

There exist several scenarios for sequencing viral genomes such as sequencing of individual strains or population [12]. Sequencing of individual genomes helps to catalogue the genes encoded in a particular strain and is a vital step for in-depth characterization studies. Sequencing of multiple isolates/strains/species enables understanding of the factors responsible for varying virulence using comparative genomic approaches [13]. For understanding the co-evolution of viral and host genomes, in particular, archaea and bacteria, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) spacer sequencing is used [14]. CRISPR are found in archaea and bacteria that serve as an antiviral mechanism in which viral genomic sequences are integrated as CRISPR spacers into the host, thereby making it immune to viral infection [15]. Understanding complex dynamics of virus–host interactions in higher organisms using sequencing provides valuable insights into transmission between animal reservoirs [16]. Sequencing of 'Auxiliary metabolic genes', which are involved in processes like motility and transcriptional repression, enables to unravel the viral genes that influence host machinery in diverse ways [17].

1.3. Data assembly and annotations

Output from NGS technologies results in gigabases of raw sequence data per experiment. Extensive computational analysis using a number of algorithms and applications is required to infer biological significance. Generic steps include mapping of reads using either *de novo* approach or re-sequencing approach, identification of SNPs and detection of insertions/deletions (indels) and further downstream processing.

The various steps involved in data preprocessing are:

i. Removal of adaptors and low-quality sequences: This is an important step in data pre-processing, and tools such as FASTX [18] and FASTQC [19] are used for this purpose. Care should be taken in case of paired-end sequences to ensure that the reads trimmed based on the quality is reflected in both the forward and the reverse FASTQ files. In case of multiplex sequencing data, an additional step of 'de-multiplexing' based on barcodes is mandatory.

ii. Screening host sequences: Despite the methods being available for viral enrichment, it has been observed that contamination of host/vector sequences is a routine scenario. Filtering of such data ensures that no error is propagated.

Following preprocessing, reference-based mapping or *de novo* assembly of the processed reads can be carried out.

1.3.1. Reference-mapping

Alignment with a reference genome is a method of choice for most NGS experiments. Preprocessed reads when mapped to a well-annotated reference genome ensure transfer of annotations to the query genome in a hassle-free manner with statistical confidence, especially in indel-free regions. Polymorphic regions can also be identified, which account for the isolate-specific variants that may be responsible for the observed phenotype. The algorithms generally rely on indexing of either the query reads or the reference genome using suffix tree or hashing strategy [20–22]. Indexing the reference genome has been proved to be computationally advantageous and is widely preferred. Indexing is followed by gapped or ungapped alignment based on either Smith–Waterman [23] or Needleman–Wunsch dynamic programming approaches [24]. Gaps indicate indels and are important to gain strain-/species-specific properties. The quality of the reference alignment can be improved by using large inserts available in paired-end reads as compared to single-end reads wherein forward and reverse orientation of reads cannot be calculated. Downstream processing of aligned and assembled reads involves delineating the variant regions followed by annotation. It is also important to remove polymerase chain reaction (PCR) artefacts before variant calling as the duplicated reads hamper its sensitivity. Discovery of *Schmallemberg* virus, a new member of genus *Orthobunyavirus* that causes foetal abnormalities in ruminants [25], is attributed to a reference-based assembly approach.

Delineation of variant regions: All deviations from reference genome can be delineated as variants, which include SNPs and indels. Variant regions contribute to the nucleotide diversity in virus populations and hence play a vital role in their evolution and dynamics. One of the main parameters indicative of nucleotide diversity is the comparison of synonymous to non-synonymous codon substitution. Synonymous mutations result in neutral substitution, which enable in maintaining the phenotype, as compared to non-synonymous substitutions, which lead to amino acid alteration and hence may affect phenotype. It is interesting to note that the existence of overlapping reading frames in viruses often constrains synonymous substitutions. Hence, computation of the magnitude of synonymous and non-synonymous polymorphism within viral populations will provide a handle to assess the role of neutral evolution and genetic drift in viral evolution. A more detailed discussion of the role of these substitution ratios in adaptive evolution of viruses is given in Section 4.5.

Tools like SNPgenie [26] and VirVarSeq [27] have been developed with a focus on calling SNPs from pooled viral samples by including codon information in an explicit manner and hence are more sensitive than traditional SNP callers [28, 29].

1.3.2. De novo assembly

Preprocessed reads are assembled using *de novo* approaches, when a closely related homologue is unavailable to serve as a reference. It should be mentioned that genome assembly is computationally challenging and also requires trained manpower. Sequencing depth plays a major role in determining the quality of the assembly as does the length of the reads. Popularly used assemblers are based on de Bruijin graph approach in which reads are divided into

http://dx.doi.org/10.5772/61610

subsequences called k-mers of length k [30]. The *k*-mers form the nodes of a graph, which are linked when a *k-1*mer is shared among them. The overall process requires large amounts of computer memory (RAM) and specialized compute clusters.

The steps involved in assembly process are:

i. Based on Overlap–Layout–Consensus principle, information stored in scattered reads are used to make contiguous regions termed 'contigs', which are generally devoid of polymorphisms.

ii. Using insert information, 'contigs' are combined to form 'scaffolds'. Gaps between contigs are usually filled with nucleotides (*N*s).

iii. Scaffolds in conjunction with synteny and geneorder information are used to build larger scaffolds.

Building a draft genome is an iterative process and involves parameter optimization, and it is advised that more than one type of assembler be used as each of them has been built for a definite purpose and has unique features. The final assembled genome is evaluated on the basis of N50 parameter. N50 is the median of assembled sequence lengths, in which longer sequences are given more weightage. Mis-assemblies due to wrong orientation of reads and low-complexity regions are, however, not accounted for in N50 parameter and tools like *amosvalidate*, which combines multiple validation procedures, are recommended [31].

One of the major limitations of *de novo* assembly using NGS data is its reporting of large proportion of incorrect recombinants. This arises mainly due to overlapping of short reads of varying quality and coverage, which in turn pave way for the introduction of spurious SNPs, ultimately resulting in artefacts in assembly. The *in silico* chimeras thus produced amplify diversity estimation and complicate true recombination detection. Efforts are being made to overcome this issue using probabilistic method, which assumes that true SNPs are under selection pressure and hence co-occur within a haplotype as compared to random SNPs [32]. Methods such as Iterative Virus Assembler (IVA) [33] and Paired-Read Iterative Contig Extension (PRICE) [34] have also been developed to overcome caveats associated with varying read depths and enable detection of regions with extensive genomic diversity. Assembly pipelines like VirAmp [35], VICUNA [36], SPAdes [37] offer many choices of tools and parameters for carrying out hassle-free assembly of viral genomes.

Novel approaches are also being introduced with special emphasis on viral metagenomic projects, *viz.* Progressive Filtering of Overlapping small RNAs (PFOR) [38]. PFOR is capable of identifying replicating circular RNAs by separating terminal small RNAs from internal small RNAs based on *k*-mer overlap. PFOR2, a multi-threaded version of PFOR, has recently been developed, which reduced the running time of filtering step by 90%. Novel viroids like *Hop stunt viroid* (HpSVd), *Grapevine yellow speckle viroid* (GYSVd) and *Grapevine hammerhead viroid-like RNA* (GHVd RNA) have been identified using this tool. Hence, *de novo* assembly has tremendous scope in unravelling the vast virome that has been unaddressed previously and there exists need for development of more efficient assembly algorithms, which will make it more tractable for use by larger scientific community.

2. Genome databases

Initial effort towards sequencing of viral genomes resulted in accumulation of genomic data in primary repositories such as GenBank [39], European Molecular Biology Laboratory (EMBL) [40] and DNA Data Bank of Japan (DDBJ) [41] and now continues to rise in International Nucleotide Sequence Database Collaboration (INSDC) [42]. Genome databases and resources dedicated to viruses were developed subsequently [43–47]. Lists of useful databases, resources and analysis tools have also been compiled previously [13, 48]. Most of these resources archive complete genome sequences, their annotations and derived data such as viral variations, multiple sequence alignments (MSAs) and phylogenetic trees, to name a few. Some of the viral genome resources are briefly described below.

2.1. National Center for Biotechnology Information (NCBI) viral genome resource

This reference resource is designed to catalogue publicly available genomic sequences of viruses deposited in INSDC [49]. It attempts to curate reference genome sequences and leverages on the knowledge of experts to annotate as well as to identify important viral sequences.

2.2. *ViralZone*

This resource is developed and maintained at the Swiss Institute of Bioinformatics. The objective of the resource is to link textbook knowledge, fact sheets and images to the genomic and proteomic data with an objective to facilitate the study of viral diversity [50].

2.3. Virus Pathogen Database and Analysis Resource (ViPR)

The ViPR [51] is supported under the Bioinformatics Resource Centers (BRC) programme of National Institute of Allergy and Infectious Diseases (NIAID). The database currently provides access to molecular data of viruses including complete genomes of 14 viral families. Analytical and visualization tools for metadata-driven statistical sequence analysis, data filtering, analytical workflows and utility of personal workbench are provided to the users.

In addition to these, several organism-specific resources have been developed such as HCV Database [52] for *Hepatitis C virus* and IVDB [53] for *Influenza virus* and HIV [54].

Annotation of the sequence (gene/genome/protein) records is an integral step in downstream processing of database entries. A well-curated reference record serves as template for transfer of annotation in terms of features such as gene boundaries, associated functions (molecular/cellular/pathway) and non-coding regions [49]. Such annotations will be highly useful in subsequent analysis and model building. The challenges of managing dedicated resources for viral genomes are relatively different as compared to the genomic databases of model and other organisms. The pace of sequencing and the quantum of genomic data being generated are affecting identification of reference genomes and annotations of genomes of strains and isolates. Additionally, to study the spatio-temporal evolution and to model the viral popula-

http://dx.doi.org/10.5772/61610

tions, it is desirable to tag metadata such as the place and date of isolation of viruses with the corresponding genomic entries.

3. Impact of NGS technologies on virology

Molecular analysis of viruses using data generated by NGS has revolutionized virology. While understanding the sequence–structure–function relationships, it has also resulted in the development of new areas of research such as phyloinformatics and immunoinformatics, which translates raw data into information. The information generated from these independent yet interlinked areas, when put together fits as pieces of jigsaw puzzle (Figure 1), leading to an improved understanding of the viral diseases and, thereby, the development of antiviral therapies.

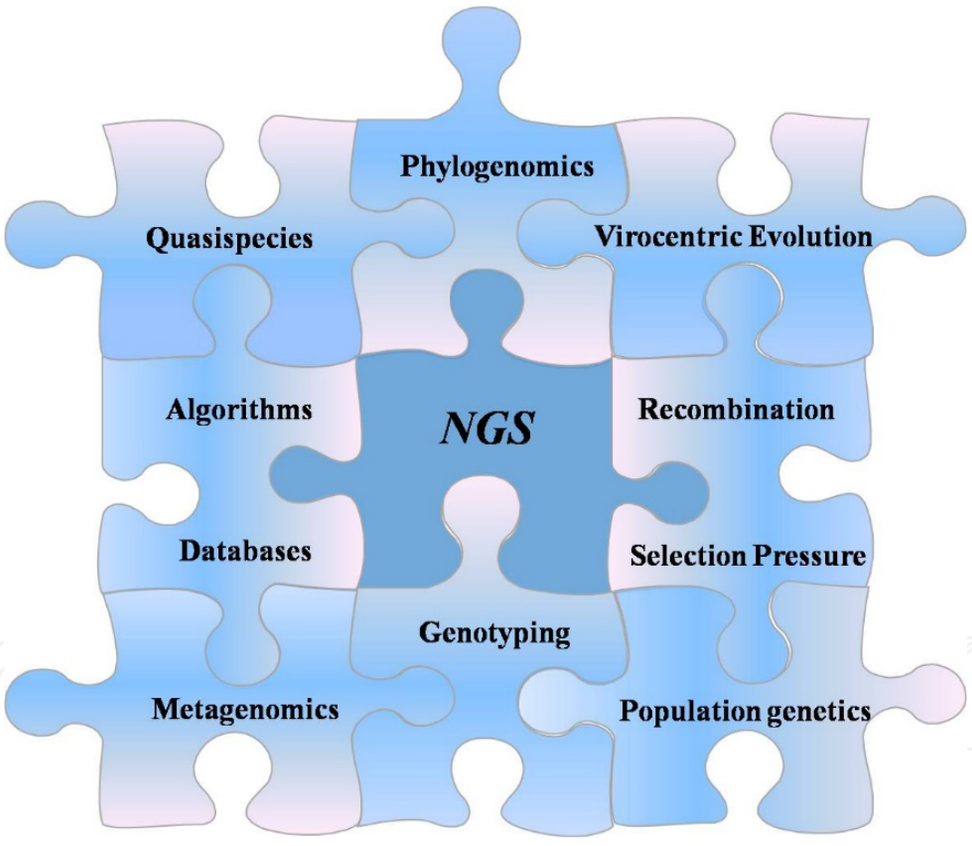

Figure 1. Scope of research in virology enabled and augmented due to availability of NGS data.

3.1. Unravelling mutational landscapes in viral quasispecies

Viral quasispecies are mutant swarms generated mainly by RNA viruses during replication, which is known to be error-prone due to the lack of proofreading activity of RNA-dependent RNA polymerase. The resulting mosaic is a dynamic distribution of non-identical but related

replicons that cannot be detected using conventional sequencing approaches. Hence, quasispecies remained unexplored for a considerable time, even though the theoretical concept for quasispecies was put forth by Eigen in 1970 [55]. With the advent of NGS technologies, the generation of large genomic datasets became a reality. Due to the sequencing error issues, it was still tough to demarcate true genetic variations. Circular Sequencing (CirSeq), a novel experimental approach that creates template of tandem repeats of circularized genomic RNA fragments has been developed by Andino's group [56]. CirSeq reduces the sequencing error drastically as the repeats get sequenced in a redundant manner for every genomic fragment. A consensus reduces the theoretical error close to 10^{-11}, which enables capture of the entire mutational spectrum of RNA virus populations. CirSeq was employed to study seven serial passages of *Poliovirus* replicated in HeLa cells. Mutation frequency was computed for every passage and their fitness was determined by mapping onto the 3D structure of proteins. As expected, majority of the mutations detected were neutral substitutions, thus highlighting robustness as driving force for adaptation and evolution [56]. This study clearly delineates the viral mutations responsible for quasispecies structure and highlights the extent of genetic variation that can be maintained in a population.

Microevolution in an evolving quasispecies population is responsible for the sequence diversity in *Porcine reproductive and respiratory syndrome virus* (PRRSV). PRRSV is the causative agent of late-term reproductive failure in sows and respiratory distress in pigs and hence has large economic impact. Genomic complexity of PRRSV due to multiple circulating genotypes results in antigenic diversity, which, in turn, is responsible for lack of effective vaccine development [57]. Sanger sequencing has identified open reading frames ORF5 and ORF7 as the polymorphic regions of the virus genome, encoding major immunogenic epitopes. In order to study the genome-wide polymorphisms, deep sequencing of PRRSV was carried out and amino acid substitutions in ORFs 2–7 in PRRSV strains obtained from pigs that lack B and T cells were studied [58]. By analysing nucleotide substitutions over time followed by comparative genomics with non-pathogenic variants, the role of mutation and selection in preserving the pathogenesis or fitness of PRRSV was well documented in this study.

3.2. Detection of low-frequency variants

Low-frequency variants or minority quasispecies are the variants that occur with a frequency of <20–25% in a viral population [59]. Minority quasispecies refers to the memory genomes that were dominant at an earlier phase of quasispecies evolution and can play an important role in conferring drug resistance in viruses such as *Human Immunodeficiency Virus type-1* (HIV-1) and *Influenza virus*. Minority quasispecies of drug-resistant viruses can rapidly re-emerge as major populations after the reintroduction of drug pressure. In case of HIV-1, presence of such low-frequency variants has been linked with early failure to the antiretroviral therapy [59, 60]. Emergence of highly pathogenic subtype of *Avian Influenza virus*es (HPAI) has also been explained on the basis of low-frequency variants. Ultra-deep sequencing was used to study the emergence of HPAI from that of less pathogenic (Low Pathogenic Avian Influenza (LPAI)) progenitor viruses [61].

http://dx.doi.org/10.5772/61610

3.3. Inter- and intra-host genetic diversity

The rate of viral evolution and the effectiveness of its transmission are determined by inter- and intra-host genetic diversity. Mutation rate and selection pressure ascertain viral diversity. Factors like mixed infections and random processes such as genetic drift and population bottlenecks also contribute to the genetic diversity of viruses both within and among hosts. Transmission fitness influences the effective spread of viruses and is responsible for its stable maintenance in the environment [62].

Intra-host genetic diversity in *Zucchini yellow mosaic virus* (ZYMV), a plant RNA virus known to infect *Cucurbitaceae* plants, has been studied using NGS [63]. Population bottlenecks were investigated for this aphid-borne virus and are thought to occur during both inter-host vector transmission and systemic movement within an individual plant. ZYMV populations infecting cucumbers with and without vector were sequenced followed by *de novo* assembly and variant calling. Analysis revealed that the low-frequency mutants present in the initial population got fixed rapidly in vector-transmitted viruses, whereas the same continued to remain as minor variants in mechanically inoculated viruses. In addition, regions known to be responsible for vector transmission were conserved in all samples. It is interesting to know that previous studies using Sanger sequencing of the coat protein of ZYMV, which is involved in interaction with aphids, could not detect mutations when transmitted between or within plants. However, this study reported six mutations in coat protein with frequency of occurrence as low as ~3%. Such studies provide an insight into the complex dynamics of genetic diversity of an emerging viral infection with implications in disease management.

3.4. Viral metagenomics

NGS has revolutionized metagenomics in a major way by ensuring high data throughput and by removing the hassles of cultivation/isolation by providing cost-effective options. Metagenomics involves sequencing of samples from diverse environments spanning across the biosphere [64]. The initial attempts at characterizing the viral metagenomes were more of an enumeration nature [65] and provided a glimpse of the enormous diversity underlying the previously unculturable communities. NGS has paved way for extensive characterization of the functional role of virome in hosts harbouring them [66, 67]. Analysis of metagenomics data is challenging as it includes simultaneous assembly of multiple genomes/transcriptomes and the complex interplay between them. Two major methods based on 'sequence-similarity' and 'sequence composition' are usually used for categorization of samples in metagenomics. It has been observed that the alignment-free 'sequence composition'-based methods provide better means of classifying viral samples as 'sequence similarity'-based methods could only classify up to 30% of the reads [68].

In a major study involving analysis of dsDNA viruses from 43 ocean samples obtained from across the globe revealed several intriguing observations [69]. Genes shared across different samples were used as 'core genes' for comparison. 'Niche-differentiation' of different viral populations based on the layer of the ocean they occupy was observed. As viruses rely on the host machinery to replicate, a direct relationship was observed between the community

structures of both viruses and hosts. Environmental factors like salinity also influenced the viral persistence and hence their diversity. Technological advances in viral metagenomics would help to unravel the underlying rules of viral evolution and ecology, the so-called 'Genomic rulebook of viruses' [70].

3.5. Genotype–phenotype correlation studies

3.5.1. Receptor switching

A key event during any viral infection is the interaction of viruses with the host receptors on the plasma membrane. This serves as an entry point for viruses to access resources of the host cell and is very crucial for tropism. This interaction is known to be very specific and is responsible for activation of the signalling processes that recruit cellular machinery of the host for viral replication. The specificity of receptor binding defines host range that a virus can infect and the extent of tissue tropism that a virus can display. Switching of receptors thus enables the virus to increase its host range and/or gain access to the previously unaffected cell types.

HIV-1 enters the target host cell by binding to CD4 receptor along with a co-receptor (in majority of cases, chemokine C-C motif receptor 5 (CCR5)) using its spike protein. Monitoring of the co-receptor usage using phenotype-based assays provided clues for the likely shift from CCR5 to chemokine C-X-C motif receptor 4 (CXCR4). However, due to the low resolution of these procedures, this transition could not be captured effectively. NGS of the variable loop region (V3) of the envelope gene containing determinants of co-receptor usage revealed the stepwise mutational pathway involved in the transition from CCR5 to CXCR4 [71]. The observation of the low-frequency intermediate variants provided an insight into the fitness landscape of *HIV-1* and provided clues to tackle the disease progression in a rational manner.

3.5.2. Immune escape

The *de novo* sequencing approach has helped to analyse the heterogeneity of *Influenza A virus* (strain A/Nagano/RC1-L/200 or H1N1) isolated from 2009 pandemic. The amino acid changes in haemagglutinin protein (G172E and G239N) were observed to be associated with the immune escape [72].

4. Bioinformatics methods for viral genomics

Bioinformatics approaches help to estimate and analyse population diversity by studying genetic recombination, mutation, selection and, thereby, assist in correlation of genotype to phenotype. The methods relevant to these aspects are discussed below with emphasis on the analysis of viral populations.

4.1. Methods for quasispecies reconstruction

Quasispecies reconstruction refers to the estimation of number of viral variants and their frequency. Each viral variant in a quasispecies is considered as a haplotype. Tools available for this purpose include Short Read Assembly into Haplotypes (ShoRAH) [73], Quasispecies Reconstruction algorithm (QuRe) [74] and QuasiRecomb [75].

4.1.1. Short Read Assembly into Haplotypes (ShoRAH)

Principle: This method uses Bayesian principle to estimate the genetic diversity of mixed samples obtained through NGS by incorporating subroutines for correction of sequencing errors [73]. It can detect viral haplotypes with frequencies as low as 0.1%.

Algorithm steps:

i. Alignment: The program requires a FASTA input file of NGS reads along with a reference sequence. It performs pairwise alignment of all reads to the reference sequence and generates a multiple sequence alignment (MSA).

ii. Error correction (local haplotype reconstruction): Using MSA as a starting point, a set of overlapping windows is analysed by employing a model-based probabilistic clustering algorithm to obtain (i) haplotype sequences, (ii) their frequencies, (iii) corrected reads and (iv) posterior probability of the reconstruction.

iii. Global haplotype reconstruction: The set of corrected reads is analysed under parsimony principle, which results in identification of set of unique reads of maximum length.

iv. Frequency estimation: Using maximum likelihood (ML) and expectation maximization algorithm, the frequencies of the reconstructed haplotypes are estimated.

4.1.2. Quasispecies Reconstruction algorithm (QuRe)

Principle: QuRe [74] is based on a heuristic algorithm and automatically reconstructs a set of error-free, full-gene/genome variants from a collection of long NGS reads (>100 bp).

Algorithm steps:

i. Overlaps between the reference genome and reads are generated in terms of *k*-mers.

ii. Mapping of *k*-mers is then carried out to obtain genomic co-ordinates.

iii. Generates a multinomial distribution based on the alignment scores of true matches along with the matches with randomly shuffled reads.

iv. Coverage, nucleotide content and entropy of each mapped genomic position are then calculated.

v. Errors are corrected based on Poisson distribution model, parameterized differently for homopolymeric and non-homopolymeric regions.

vi. Reconstruction of quasispecies is carried out using the sliding window approach by calculating maximal coverage and read diversity, which reduces the false positives, *i.e.*, *in-silico* recombinants.

4.1.3. QuasiRecomb

Principle: It employs the jumping Hidden Markov Model (HMM)-based probabilistic statistics for inference of viral quasispecies, especially for estimating the intra-patient viral haplotype distribution [75]. This method assumes that the true genetic diversity is generated by a few sequences (called generators) through mutation and recombination, and that the observed diversity results from additional sequencing errors.

Algorithm steps:

i. Distribution of haplotypes in a given population is modelled to account for either point mutation or recombination in the form of probability tables and jumping HMM states respectively.

ii. Expectation maximization algorithm is used to estimate posterior probabilities associated with rare events of mutation and recombination.

4.2. Methods to study viral population genetics

Genetic structure of a population refers to the number of distinct subpopulations, identified using a characteristic set of allele frequencies [76]. A model-based population analysis can be performed using the STRUCTURE program [77] based on genomic data. The program can infer the genetic structure in haploid, diploid and polyploid species [78].

4.2.1. STRUCTURE program

Principle: This method is based on Bayesian clustering approach and employs Markov Chain Monte Carlo (MCMC) algorithm to identify genetically distinct subpopulations based on allele frequencies. It assigns individuals to subpopulations based on likelihood estimates. In case of haploids, the program assumes that the loci are in linkage equilibrium or only weakly linked [78]. The program accounts for recombination by incorporating ancestry models such as admixture and linkage models. An admixed strain is assigned with a membership score to belong to two or more subpopulations, to indicate its mixed ancestry. Linkage model is an extension of admixture model to account for weak linkage that arises as a result of admixture linkage disequilibrium (LD). Therefore, the extent of linkage equilibrium within the markers needs to be tested prior to usage of the STRUCTURE program. The relevant linkage analysis (LIAN) programs and measures are discussed in Section 4.3.

Input genotype data: A wide range of markers such as multi-locus genotype data, microsatellites, SNPs can be used as an input. In case of viruses, the polymorphic sites or more specifically the parsimony-informative (PIs) sites obtained from genome-based alignment are suitable markers for population genetic analyses. A PI site contains at least two types of nucleotide bases and at least two of which occur with a minimum frequency of two. The position of each

PI corresponds to a locus. At every locus, any of the four bases (A, T, G and C) and the gap is considered as an allele.

Algorithm steps:

i. Carry out MSA of complete genomes and extract PI sites.

ii. Estimate the degree of linkage equilibrium and test the null hypothesis about the same.

iii. Simulate data using burn-in and burn-length with values in the range of 10,000–1,00,000. Check the convergence of parameters and consistency of clustering results.

iv. Estimate the appropriate number of clusters (*K*) using independent runs with varying values of *K*.

v. Determine the best *K* either by comparing mean of log likelihoods [77] or based on an *ad hoc* statistic, ΔK [79].

vi. Validate the genetic structure hypothesis using Analysis of MOlecularVAriance (AMOVA) based on Fixation index (F_{ST}) as implemented in ARLEQUIN software [80]. F_{ST} represents the extent of genetic differentiation among subpopulations and ranges between 0 (no differentiation) and 1 (complete differentiation).

Salient features of the STRUCTURE program:

i. This method is advantageous over traditional molecular phylogenetic methods in terms of classification of recombinant strains.

ii. User can incorporate prior information such as geographic location of samples.

Limitations:

i. Variation in sample size may affect the clustering.

ii. This method is not suitable for datasets having high linkage disequilibrium.

Case studies:

The ability of the admixture model to account for recombination has been used to analyse the extent of recombination and its role in determining the population structure of viruses such as *Hepatitis B virus* [81] and *Rhinoviruses* [82].

Population genomic study of *Hepatitis B virus* (HBV) was carried out using both admixture and linkage models (with burn-in of 20,000 and burn-length of 40,000). HBV is an enveloped DNA virus and belongs to the genus *Orthohepadnavirus* and family *Hepadnaviridae*. It is known to consist of eight genotypes designated as A–H, each of which has characteristic geographic distribution. This method helped to resolve the hierarchical nature of population subdivision with the presence of four major clusters (F_{ST} = 0.497, $p < 0.0001$) and eight sub-clusters. The extent of recombination was observed to be low [81].

Rhinoviruses represent the highly diverse members of genus *Enterovirus* and family *Picornaviridae*. They are ss (+) RNA viruses with genome of ~7,200 bases. There are three species, *viz.*

Rhinovirus A, -B and *-C*, each of which is further subdivided into distinct serotypes. The STRUCTURE-based analysis revealed a strong evidence for existence of seven genetically distinct subpopulations (with $F_{ST} = 0.45$, $p = 0$). *Rhinovirus A* and *Rhinovirus C* were subdivided into four and two subpopulations respectively, whereas *Rhinovirus B* species remain undivided. Furthermore, usage of both the admixture and the linkage models (with burn-in of 20,000 and burn-length of 40,000) helped to resolve the role of recombination in diversification of subpopulations. In case of *Rhinovirus A*, intra-species recombination was common, whereas in case of *Rhinovirus* C, intra- and inter-species recombination were observed to cause diversity [82].

4.3. Methods to compute linkage disequilibrium

Linkage equilibrium refers to the statistical independence of alleles at all loci and indicates evidence of free recombination [83]. Thus, linkage disequilibrium is a measure of the correlation between the occurrences of nucleotides at different loci of the genome. The extent to which recombination occurs can be estimated in terms of the degree of linkage disequilibrium [84] using measures made available by specialized programs such as Linkage Analysis (LIAN) [83] and DNA Sequence Polymorphism (DnaSP) [85]. The extent of linkage can be inferred based on the following parameters.

i. **Standardized index of association, $I^S A$:** It is a measure of the degree of haplotype-wide linkage derived from a given dataset. $I^S A$ is computed using a formula, $I^S A = [1/(e-1)]\,[(V_D/V_E)-1]$, where '$V_D$' represents the observed variance of pairwise distances between haplotypes and 'V_E' represents the expected variance when all loci are in linkage equilibrium. The term $[(V_D/V_E)-1]$ is the function of rate of recombination, which is zero in case of linkage equilibrium. The number of loci analysed is denoted by '*e*'. The value of $I^S A$ can be computed by using the program called LIAN (for Linkage Analysis), which requires haplotype data as an input. This program implements both a Monte Carlo and an algebraic method to test the null hypothesis: $V_D = V_E$.

ii. **$|D'|$ and r^2**: The $|D'|$ measure is the absolute value of the difference between the observed and the expected haplotype frequency in the absence of linkage disequilibrium, which is normalized by the maximum (or minimum) possible value of this difference. The squared value of the difference between the observed and the expected haplotype frequency normalized by the variance of the allele frequency is denoted by r^2. These measures can be computed using DnaSP program [85]. The values for these measures can range between 0 (no linkage disequilibrium) and 1 (complete linkage disequilibrium) [84, 86].

Case studies:

LD provides a good measure for analysing the extent of recombination in viruses [82, 87]. For example, in case of *Rhinoviruses,* low values for LD measures ($I^S A = 0.0666$, $p < 10^{-4}$; $|D'| = 0.5409$ and that of $r^2 = 0.0613$) were observed and correlated well with the evidence of recombination obtained using independent methods [82]. Similarly, LD analyses in serotypes of *Foot and mouth disease virus* [87] helped to reveal low values of $|D'|$ and r^2, supporting high recombination.

http://dx.doi.org/10.5772/61610

4.4. Methods for detection of recombination

In addition to undergoing mutations, viruses are known to generate new variants through genetic recombination. Genetic recombination refers to the exchange of genetic material between strains of the same or different species of viruses [88]. Within a host, co-infected with viruses, the recombination occurs either by homologous recombination or by reassortment [89]. Homologous recombination can occur between highly similar RNA genomes usually through the process called 'copy-choice' or 'template-switching' mechanism, whereas reassortment involves exchange of genomic regions between viruses that have segmented genomes. Presence of recombinants can hamper analyses pertaining to molecular clock [90], selection pressure, phylogenetic classification [91, 92] and thus need to be detected prior to such analyses.

4.4.1. Virus Recombination Mapper (ViReMa)

ViReMa is developed to analyse the recombinants within the viral genome data derived through NGS [93]. It can detect inter-virus or virus–host recombination. This method can also detect insertion and substitution events and multiple recombination junctions within a single read.

Algorithm steps:

i. Alignment of 5' end of each read to the reference genome(s) using seed-based approach.

ii. Dynamic generation of a new read segment: 3' end of the read that fail to align is extracted or the first nucleotide from the read is trimmed. This step is iterated until all the reads are either mapped or trimmed or a combination of both.

iii. For each read, all possible recombinations are reported.

4.4.2. Recombination Detection Program version 4 (RDP4) package

In order to detect recombination, various methods have been developed and are provided in RDP4 package [94]. It identifies the significant evidence of recombination events based on the *p-value* and identifies the potential recombinant sequences and its both parents (major and minor). The main strength of the package is that it does not need any prior knowledge pertaining to non-recombinant set of reference sequences. The starting point of analysis is MSA of genomic sequences.

Algorithm steps:

i. RDP4 package sequentially tests every combination of three sequences in MSA (a triplet) for potential evidence that one of the three is a recombinant and the other two are its parents. Various recombination detection methods, such as the Ramer–Douglas–Peucker algorithm (RDP) method [95], BOOTSCAN [96, 97], maximum Chi-square (MAXCHI) method [98, 99], CHIMAERA [99], 3'-end sequencing for expression quantification (3SEQ) [100], gene conversion method (GENECONV) [101], Sister

Scanning method (SISCAN) [102], LARD [103], Topal/Difference of Sums of Squares (DSS) [104] and DNA distance plot, are used.

ii. Following the detection of a 'recombination signal', RDP4 determines approximate breakpoint positions using HMM and then identifies the recombinant sequence using various methods such as phylogenetic profiling (PHYLPRO) [105] and Visual Recombination Detection (VisRD) [106].

iii. The minimum number of recombination events that are needed to account for these signals are then inferred. It involves sequential disassembly of the identified recombinant sequences into respective components and iteratively rescanning the resulting expanded dataset until no further recombination signals are evident.

Salient feature:

RDP4 package provides a unified interface for multiple methods and facilitates visualization of recombination events using genomic data (up to 2,500 sequences).

Limitations:

i. The genomic dataset up to 200 million nucleotides can be analysed and is reported to have operational limits for large genomic datasets.

ii. Recombination analysis is likely to fail in case of poor alignments, if recombinant sequences are used as reference and sequences having ambiguous characters are included.

4.5. Methods for selection pressure analysis

Natural selection is one of the fundamental evolutionary processes that shape the genetic structure of viral populations. The ratio of non-synonymous substitution rate (*dN*) to synonymous substitution rate (*dS*) is a useful means to infer selection pressure based on a codon alignment for a particular gene. Positive selection ($dN/dS > 1$) increases the frequency of advantageous alleles, whereas the negative selection ($dN/dS < 1$) is responsible for purging (removal) of deleterious alleles.

Broadly, the selection pressure can be classified as pervasive and episodic. Pervasive selection acts across all the lineages in a phylogenetic tree, whereas the episodic selection operates on a few lineages of a tree. Various statistical methods for analysis of pervasive and episodic selection are available at the Datamonkey web-server of Hypothesis testing using Phylogenies (HyPhy) software package [107–109].

4.5.1. Single Likelihood Ancestor Counting (SLAC)

Principle: This method belongs to a class called counting methods [110]. It is suitable for pervasive selection analysis and involves estimating the number of non-synonymous and synonymous changes that have occurred at each codon throughout the evolutionary history of the sample. It involves reconstructing the ancestral sequences using likelihood-based method [111].

Algorithm steps:

i. Nucleotide model fit: Using maximum likelihood (ML), a nucleotide model of time-reversible class is fitted to the data and tree, to obtain branch lengths and substitution rates. If multiple segments are present in the input codon alignment, base frequencies and substitution rates are inferred jointly from the whole alignment, while branch lengths are estimated for each segment separately.

ii. Codon model fit: To obtain a global $\omega = dN/dS$ ratio, the branch lengths and substitution rate parameters are considered constant at the values estimated in 'step i'. A codon model is obtained using a combination of MG94 model and the nucleotide model of 'step i' and then fitted to the data.

iii. Ancestral sequence reconstruction: Based on the parameter estimates obtained using steps i and ii, codons of ancestral sequences are reconstructed site by site using maximization of the likelihood of the data at the site over all possible ancestral character states. Inferred ancestral sequences are treated as known for the next step.

iv. Inference of selection at each site: For every variable site, four quantities, *viz.* the normalized expected (*ES* and *EN*) and the observed numbers (*NS* and *NN*) are calculated for synonymous and non-synonymous substitutions respectively. SLAC estimates $dN = NN/EN$ and $dS = NS/ES$, and if $dN < dS$, a codon is called negatively selected or if $dN > dS$, it is said to be positively selected. A *p*-value is derived to assess the significance. The test assumes that under neutrality, a random substitution will be synonymous with probability $p = ES/(ES + EN)$.

4.5.2. Fixed-Effect Likelihood (FEL) and Internal Fixed-Effect Likelihood (IFEL)

Principle: These belong to a class of methods called 'fixed effects'. It analyses pervasive selection and involves fitting substitution rates on a site-by-site basis by assuming that the synonymous substitution rate is the same for all sites. Thus, FEL and IFEL assume the same dN/dS (ω) ratio, which is applicable to all branches and to interior branches, respectively [111].

Algorithm steps:

i. Nucleotide and codon model fitting procedure in these methods is similar to those of SLAC method as detailed in Section 4.5.1.

ii. Site-by-site likelihood ratio test (LRT):

FEL method: For every site, based on the parameter estimates obtained using nucleotide- and codon-fit procedure, two rate parameters namely α and β are first fitted independently and then under the constraint of $\alpha = \beta$. Here, the parameter α represents the instantaneous synonymous site rate, while β represents the instantaneous non-synonymous site rate. Furthermore, LRT is performed to infer whether α is different from β and a *p*-value is computed. If the *p*-value is significant, the site is classified based on whether $\alpha > \beta$ (indicates negative selection) or $\alpha < \beta$ (indicates positive selection).

IFEL method: It differs from FEL in following aspects:

- The selection is only tested for internal branches of the phylogenetic tree.
- Each site has three rate parameters, α, β_I (instantaneous non-synonymous site rate for internal branches) and β_L (instantaneous non-synonymous site rate for terminal branches). Here, the null model assumes that $\alpha = \beta$_I.

4.5.3. Mixed Effects Model of Evolution (MEME)

Principle: MEME is categorized under the 'branch-site random effects' phylogenetic methods [112]. Though this method is a generalization of FEL method, it differs from FEL and IFEL, by accounting for episodic positive selection that particularly affects a subset of lineages. MEME uniquely allows the distribution of *dN/dS (ω)* to vary from site to site (the fixed effect) and also from branch to branch at a site (the random effect).

Algorithm steps:

i. The steps 'i' and 'ii' are same as that of the SLAC method (Section 4.5.1), whereas there is variation in step 'iii' as follows:

ii. The ω ratio is modelled across lineages at an individual site, i.e., each site is treated as a fixed-effect component of the model using a two-bin random distribution with $\omega- \leq 1$ (proportion p) and $\omega+$ (unrestricted, proportion $1-p$). Thus, a proportion (p) of branches at a site evolve neutrally (or under negative selection), while the remaining ($1-p$) may evolve under diversifying selection. To test for evidence of episodic selection, a likelihood ratio test is applied.

4.6. Methods for reconstruction of molecular phylogeny

Molecular phylogenetic analyses are the most commonly performed studies in virology with major applications in viral taxonomy, systematics and genotyping. Methods for reconstruction of phylogenetic tree are broadly classified into three main categories, *viz.* distance-based, character-based and Bayesian-based and are reviewed earlier [113, 114]. Distance-based methods use pairwise distance matrix as an input for tree building. Neighbour-joining [115], minimum evolution [116] and least square [117, 118] methods are widely used methods under this category. These methods are computationally efficient and suitable for the analysis of large datasets with low levels of sequence divergence. However, these methods do not perform equally well in case of highly divergent sequences with low levels of sequence similarity. Moreover, uncertainties can be introduced due to positioning of gaps in the MSA. Character-based methods assume each site in MSA to evolve independently. The two classical methods under this category are maximum parsimony and maximum likelihood [119], which estimate the tree score based on the minimum number of changes and the log-likelihood value respectively. However, it needs to be mentioned that alignment-based phylogenetic methods are observed to misclassify taxa with mixed ancestry and/or recombination [91, 92].

The alignment-free methods have been developed as an alternative and can be classified into four categories based on the underlying principles employed. They are *k*-mer/word composition, substring theory, information theory and graphical representation [120].

http://dx.doi.org/10.5772/61610

Whole genome-based phylogenetic trees are widely used for various viruses owing to their small genome sizes and conservation of genomic structure. Phylogenomics field has gained importance as whole genome data became available enabling the study of evolution in general and epidemiology and disease surveillance, in particular. This field when analysed in the context of spatio-temporal data helps to understand the disease spread and progression during outbreaks. The program such as Bayesian Evolutionary Analysis by Sampling Trees (BEAST) has been exclusively designed for phylogeography studies [121] and is used widely to study spatio-temporal dynamics of viruses at population scale.

BEAST software provides a Bayesian Markov chain Monte Carlo (MCMC) framework for parameter estimation and hypothesis testing of evolutionary models from molecular sequence data. It brings together a large number of evolutionary models into a single coherent framework for evolutionary inference. Available evolutionary models include substitution, insertion–deletion, demographic, tree shape priors, node calibration and relaxed clock models. This combinatorial principle is advantageous as it provides a flexible system to specify models to understand various aspects of virus evolution. BEAST uniquely incorporates the time-scale data to explicitly model the rate of molecular evolution on each branch in the tree. Under the uniform rate assumption over the entire tree, the molecular clock model becomes applicable. It is the first software to incorporate the relaxed molecular clock model that does not assume constant rate across lineages.

4.7. Methods for typing of viruses

Phylogenetic analysis, whether alignment-based or alignment-free, is routinely used for genotyping/serotyping of viruses. Such analysis is carried out using the regions that are identified as markers for the purpose of classification by the expert evolutionary virologists and the International Committee of viruses (ICTV) [122]. It has been observed that genotype information for less than 10% of the viral genomes is available as part of their sequence records. As NGS technologies are producing a large number of genomic sequences for various strains, isolates and viral species, the genotype assignment gap is ever-increasing. Several tools for genotyping have been developed using both alignment-based and alignment-free methods and are most often organism-specific. NCBI Genotyping Tool is based on the sequence similarity for identifying the genotype of recombinant and non-recombinant viral sequences [123]. Similar tools exist for *Influenza virus, viz.* FluGenom [124]. Alignment-free method for phylogeny and genotyping of viruses based on the concept of Return Time Distribution has been developed *in-house* and its applicability for genotyping of viruses such as *Mumps virus*, *Dengue virus* and *West Nile virus* has been demonstrated [125–127].

5. The way forward

NGS has proved to be extremely useful and has become an integral part of virus research and opened up new vistas in studying viral evolution. Ample proof of the same is the characteri-

zation of the *Ebola virus* infection in West Africa (2014 outbreak), wherein the patient samples were sequenced using NGS to trace the origin and transmission of the infection as part of the global epidemic surveillance strategy [128]. The discovery followed by the development of vaccine [129] has been made in a short time span owing to the genomics-enabled translational research. In order to harness the use of NGS in virology, care needs to be exerted to avoid misinterpretation and over-interpretation of the data. It must be noted that starting from sample collection, DNA/RNA extraction, PCR amplification, library preparation up to sequencing are prone to errors, which have been explained [130] very comprehensively. Circumventing these issues, application of NGS in virology has enabled basic and applied research to take a quantum leap. The thorough understanding of the intricacies of a quasispecies structure aids in tracing the mutational network operational due to selection pressures. Furthermore, characterization of intra- and inter-host viral evolution helps in understanding the role of host immune system on the genetic variability of viruses. Such data when analysed in the context of population genetics provide constructs to understand emergence of new strains/lineages. Reverse vaccinology [131] enabled via genomics is expected to accelerate the rate of vaccine discovery, thereby, reducing the virus-associated disease burden.

Acknowledgements

The authors would like to acknowledge the Department of Biotechnology (DBT), Government of India for the infrastructural facilities. Dr. Urmila Kulkarni-Kale acknowledges Centre of Excellence (CoE) Grant to the Bioinformatics Centre from the DBT. Sunitha M Kasibhatla acknowledges the support of Bioinformatics Resources and Applications Facility (BRAF), C-DAC, Pune. Vaishali P. Waman acknowledges DBT fellowship (2010–2015).

Author details

Sunitha M. Kasibhatla[1,2], Vaishali P. Waman[1], Mohan M. Kale[3] and Urmila Kulkarni-Kale[1*]

*Address all correspondence to: urmila@bioinfo.net.in; urmila.kulkarni.kale@gmail.com

1 Bioinformatics Centre, Savitribai Phule Pune University (formerly University of Pune), Ganeshkhind, Pune, Maharashtra, India

2 Bioinformatics Group, Centre for Development of Advanced Computing, Pune University Campus, Ganeshkhind, Pune, Maharashtra, India

3 Department of Statistics, Savitribai Phule Pune University (formerly University of Pune), Ganeshkhind, Pune, Maharashtra, India

References

[1] Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. 2010;18(1):11–9. DOI: 10.1016/j.tim.2009.11.003.

[2] Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the marine virosphere using metagenomics. PLoS Genet. 2013;9(12):e1003987. DOI:10.1371/journal.pgen.1003987.

[3] Dye C, Mertens T, Hirnschall G, Mpanju-Shumbusho W, Newman RD, Raviglione MC, Savioli L, Nakatani H. WHO and the future of disease control programmes. Lancet. 2013;381(9864):413–8. DOI:10.1016/S0140-6736(12)61812-1.

[4] Koonin EV, Dolja VV. A virocentric perspective on the evolution of life. Curr Opin Virol. 2013;3(5):546–57. DOI:10.1016/j.coviro.2013.06.008.

[5] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265(5596):687–95. DOI:10.1038/265687a0.

[6] Roche 454. Available from: http://www.454.com/ [Accessed: 2015-08-10]

[7] Illumina. Available from: http://www.illumina.com/ [Accessed: 2015-08-10]

[8] Ion Torrent. Available from: https://www.lifetechnologies.com [Accessed: 2015-08-10]

[9] Oxford Nanopore. Available from: https://www.nanoporetech.com/ [Accessed: 2015-08-10]

[10] Pacific Biosciences. Available from: http://www.pacificbiosciences.com/ [Accessed: 2015-08-10]

[11] Marston DA, McElhinney LM, Ellis RJ, Horton DL, Wise EL, Leech SL, David D, de Lamballerie X, Fooks AR. Next generation sequencing of viral RNA genomes. BMC Genomics. 2013;14(1):444. DOI:10.1186/1471-2164-14-444.

[12] Quiñones-Mateu ME, Avila S, Reyes-Teran G, Martinez MA. Deep sequencing: becoming a critical tool in clinical virology. J Clin Virol. 2014;61(1):9–19. DOI:10.1016/j.jcv.2014.06.013.

[13] Kulkarni-Kale U, Waman V, Raskar S, Mehta S, Saxena S. Genome to vaccinome: role of bioinformatics, immunoinformatics & comparative genomics. Curr Bioinform. 2012;7(4):454–66. DOI:10.2174/15748936113089990014.

[14] Anderson RE, Brazelton WJ, Baross JA. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol. 2011;77(1):120–33. DOI:10.1111/j.1574-6941.2011.01090.x.

[15] Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327(5962):167–70. DOI:10.1126/science.1179555.

[16] Wylie KM, Weinstock GM, Storch GA. Virome genomics: a tool for defining the human virome. Curr Opin Microbiol. 2013;16(4):479–84. DOI:10.1016/j.mib.2013.04.006.

[17] Sharon I, Battchikova N, Aro EM, Giglione C, Meinnel T, Glaser F, Pinter RY, Breitbart M, Rohwer F, Béjà O. Comparative metagenomics of microbial traits within oceanic viral communities. ISME J. 2011;5(7):1178–90. DOI:10.1038/ismej.2011.2.

[18] FastX toolkit. Available from: http://hannonlab.cshl.edu/fastx_toolkit/index.html [Accessed: 2015-08-10]

[19] FastQC. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed: 2015-08-10]

[20] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95. DOI:10.1093/bioinformatics/btp698.

[21] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. DOI:10.1093/bioinformatics/btp324.

[22] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. DOI:10.1186/gb-2009-10-3-r25.

[23] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.

[24] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53.

[25] Rosseel T, Scheuch M, Höper D, De Regge N, Caij AB, Vandenbussche F, Van Borm S. DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe. PLoS One. 2012;7(7):e41967. DOI:10.1371/journal.pone.0041967.

[26] Nelson CW, Moncla LH, Hughes AL. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. Bioinformatics. 2015. pii: btv449.

[27] Verbist BM, Thys K, Reumers J, Wetzels Y, Van der Borght K, Talloen W, Aerssens J, Clement L, Thas O. VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. Bioinformatics. 2015;31(1):94–101. DOI:10.1093/bioinformatics/btu587.

[28] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93. DOI:10.1093/bioinformatics/btr509.

[29] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8. DOI:10.1038/ng.806.

[30] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 May;18(5):821–9. DOI:10.1101/gr.074492.107.

[31] Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 2008;9(3):R55. DOI:10.1186/gb-2008-9-3-r55.

[32] Yang X, Charlebois P, Macalalad A, Henn MR, Zody MC. V-Phaser 2: variant inference for viral populations. BMC Genomics. 2013;14(1):674. DOI: 10.1186/1471-2164-14-674.

[33] Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, Nastouli E, Keane JA, Kellam P, Otto TD. IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics. 2015;31(14):2374–6. DOI:10.1093/bioinformatics/btv120.

[34] Ruby JG, Bellare P, Derisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda). 2013;3(5):865–80. DOI: 10.1534/g3.113.005967.

[35] Wan Y, Renner DW, Albert I, Szpara ML. VirAmp: a galaxy-based viral genome assembly pipeline. Gigascience. 2015;4(1):19. DOI:10.1186/s13742-015-0060-y.eCollection 2015.

[36] Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ, Qu J, Ryan EM, Zody MC, Henn MR. De novo assembly of highly diverse viral populations. BMC Genomics. 2012;13(1):475. DOI:10.1186/1471-2164-13-475.

[37] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77. DOI:10.1089/cmb.2012.0021.

[38] Zhang Z, Qi S, Tang N, Zhang X, Chen S, Zhu P, Ma L, Cheng J, Xu Y, Lu M, Wang H, Ding SW, Li S, Wu Q. Discovery of replicating circular RNAs by RNA-seq and computational algorithms. PLoS Pathog. 2014;10(12):e1004553. DOI:10.1371/journal.ppat.1004553.

[39] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2013;41(Database issue):D36–42. DOI:10.1093/nar/gks1195.

[40] Cochrane G, Alako B, Amid C, Bower L, Cerdeño-Tárraga A, Cleland I, Gibson R, Goodgame N, Jang M, Kay S, Leinonen R, Lin X, Lopez R, McWilliam H, Oisel A,

Pakseresht N, Pallreddy S, Park Y, Plaister S, Radhakrishnan R, Rivière S, Rossello M, Senf A, Silvester N, Smirnov D, Ten Hoopen P, Toribio A, Vaughan D, Zalunin V. Facing growth in the European Nucleotide Archive. Nucleic Acids Res. 2013;41(Database issue):D30–5. DOI:10.1093/nar/gks1175.

[41] Kodama Y, Mashima J, Kosuge T, Katayama T, Fujisawa T, Kaminuma E, Ogasawara O, Okubo K, Takagi T, Nakamura Y. The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data. Nucleic Acids Res. 2015;43(Database issue):D18–22. DOI:10.1093/nar/gku1120.

[42] INSDC. Available from: http://www.insdc.org/ [Accessed: 2015-08-10]

[43] Hiscock D, Upton C. Viral Genome DataBase: storing and analyzing genes and proteins from complete viral genomes. Bioinformatics. 2000;16(5):484–5.

[44] Albà MM, Lee D, Pearl FM, Shepherd AJ, Martin N, Orengo CA, Kellam P. VIDA: a virus database system for the organization of animal virus genome open reading frames. Nucleic Acids Res. 2001;29(1):133–6. DOI:10.1093/bioinformatics/16.5.484.

[45] Kulkarni-Kale U, Bhosle S, Manjari GS, Kolaskar AS. VirGen: a comprehensive viral genome resource. Nucleic Acids Res. 2004;32(Database issue):D289–92.

[46] Hirahata M, Abe T, Tanaka N, Kuwana Y, Shigemoto Y, Miyazaki S, Suzuki Y,Sugawara H. Genome Information Broker for Viruses (GIB-V): database for comparative analysis of virus genomes. Nucleic Acids Res. 2007;35(Database issue):D339–42. DOI: 10.1093/nar/gkl1004.

[47] Chang S, Zhang J, Liao X, Zhu X, Wang D, Zhu J, Feng T, Zhu B, Gao GF, Wang J, Yang H, Yu J, Wang J. Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. Nucleic Acids Res. 2007;35(Database issue):D376–80.

[48] Sharma D, Priyadarshini P, Vrati S. Unraveling the web of viroinformatics: computational tools and databases in virus research. J Virol. 2015;89(3):1489–501. DOI:10.1128/JVI.02027-14.

[49] Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2015;43(Database issue):D571–7. DOI:10.1093/nar/gku1207.

[50] Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, Le Mercier P. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. 2011;39(Database issue):D576–82. DOI:10.1093/nar/gkq901.

[51] Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen CN, Jen W, Klem EB, Scheuermann RH. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses. 2012;4(11):3209–26. DOI:10.3390/v4113209.

[52] Kuiken C, Hraber P, Thurmond J, Yusim K. The hepatitis C sequence database in Los Alamos. Nucleic Acids Res. 2008;36(Database issue):D512–6. DOI:10.1093/nar/gkm962.

[53] Chang S, Zhang J, Liao X, Zhu X, Wang D, Zhu J, Feng T, Zhu B, Gao GF, Wang J, Yang H, Yu J, Wang J. Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. Nucleic Acids Res. 2007;35(Database issue):D376–80. DOI:10.1093/nar/gkl779.

[54] HIV Sequence databases. Available from: http://www.hiv.lanl.gov [Accessed: 2015-08-10]

[55] Eigen M. Self organization of matter and the evolution of biological macromolecules. Naturwissenschaften. 1971;58(10):465–523.

[56] Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature. 2014;505(7485):686–90. DOI: 10.1038/nature12861.

[57] Lu ZH, Archibald AL, Ait-Ali T. Beyond the whole genome consensus: unravelling of PRRSV phylogenomics using next generation sequencing technologies. Virus Res. 2014;194(Pt 2):167–74. DOI:10.1016/j.virusres.2014.10.004.

[58] Chen N, Dekkers JC, Ewen CL, Rowland RR. Porcine reproductive and respiratory syndrome virus replication and quasispecies evolution in pigs that lack adaptive immunity. Virus Res. 2015;195(2):246–9. DOI:10.1016/j.virusres.2014.10.006.

[59] Metzner K. The significance of minority drug-resistant quasispecies. In: Geretti AM, editor. Antiretroviral resistance in clinical practice. London: Mediscript; 2006. Chapter 11.

[60] Li JZ, Paredes R, Ribaudo HJ, Svarovskaia ES, Metzner KJ, Kozal MJ, Hullsiek KH, Balduin M, Jakobsen MR, Geretti AM. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. JAMA. 2011;305(13):1327–35. DOI:10.1001/jama.2011.375.

[61] Monne I, Fusaro A, Nelson MI, Bonfanti L, Mulatti P, Hughes J, Murcia PR, Schivo A, Valastro V, Moreno A. Emergence of a highly pathogenic avian influenza virus from a low-pathogenic progenitor. J Virol. 2014;88(8):4375–88. DOI:10.1128/JVI.03181-13.

[62] Rodpothong P, Auewarakul P. Viral evolution and transmission effectiveness. World J Virol. 2012;1(5):131–4. DOI:10.5501/wjv.v1.i5.131.

[63] Simmons HE, Dunham JP, Stack JC, Dickins BJ, Pagán I, Holmes EC, Stephenson AG. Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. J Gen Virol. 2012;93(Pt 8):1831–40. DOI:10.1099/vir.0.042622-0.

[64] Rosario K, Breitbart M. Exploring the viral world through metagenomics. Curr Opin Virol. 2011;1(4):289–97. DOI:10.1016/j.coviro.2011.06.004.

[65] Yang J, Yang F, Ren L, Xiong Z, Wu Z, Dong J, Sun L, Zhang T, Hu Y, Du J, Wang J, Jin Q. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol. 2011;49(10):3463–9. DOI:10.1128/JCM.00273-11.

[66] Lecuit M, Eloit M. The human virome: new tools and concepts. Trends Microbiol. 2013;21(10):510–5. DOI:10.1016/j.tim.2013.07.001.

[67] Stobbe AH, Roossinck MJ. Plant virus metagenomics: what we know and why we need to know more. Front Plant Sci. 2014;5:150. DOI:10.3389/fpls.2014.00150.

[68] Soueidan H, Schmitt LA, Candresse T, Nikolski M. Finding and identifying the viral needle in the metagenomic haystack: trends and challenges. Front Microbiol. 2015;5(29):739. DOI:10.3389/fmicb.2014.00739.

[69] Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, AlbertiA,Chaffron S, Cruaud C, de Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S; Tara Oceans Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 2015;348(6237):1261498. DOI: 10.1126/science.1261498.

[70] Wommack KE, Nasko DJ, Chopyk J, Sakowski EG. Counts and sequences, observations that continue to change our understanding of viruses in nature. J Microbiol. 2015;53(3):181–92. DOI:10.1007/s12275-015-5068-6.

[71] Bunnik EM, Swenson LC, Edo-Matas D, Huang W, Dong W, Frantzell A, Petropoulos CJ, Coakley E, Schuitemaker H, Harrigan PR, van 't Wout AB. Detection of inferred CCR5- and CXCR4-using HIV-1 variants and evolutionary intermediates using ultra deep pyrosequencing. PLoS Pathog. 2011;7(6):e1002106.

[72] Kuroda M, Katano H, Nakajima N, Tobiume M, Ainai A, Sekizuka T, Hasegawa H, Tashiro M, Sasaki Y, Arakawa Y. Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. PLoS One. 2010;5(4):e10256. DOI:10.1371/journal.pone.0010256.

[73] Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinform. 2011;12:119. DOI:10.1186/1471-2105-12-119.

[74] Prosperi MC, Salemi M. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics. 2012;28(1):132–3.

[75] Töpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, Beerenwinkel N. Probabilistic inference of viral quasispecies subject to recombination. J Comput Biol. 2013;20(2):113–23. DOI:10.1093/bioinformatics/btr627.

[76] Chakraborty R. Analysis of genetic structure of populations: meaning, methods, and implications. In: Majumder P, editor. Human population genetics. New York: Springer 1993. p. 189–206. DOI:10.1007/978-1-4615-2970-5_14.

[77] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.

[78] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164(4): 1567–87.

[79] Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14(8):2611–20. DOI:10.1111/j.1365-294X.2005.02553.x.

[80] Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2005;1:47.

[81] Szmaragd C, Balloux F. The population genomics of hepatitis B virus. Mol Ecol. 2007;16(22):4747–58. DOI:10.1111/j.1365-294X.2007.03564.x.

[82] Waman VP, Kolekar PS, Kale MM, Kulkarni-Kale U. Population structure and evolution of rhinoviruses. PloS One. 2014;9(2):e88981. DOI:10.1371/journal.pone.0088981.

[83] Haubold B, Hudson RR. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Bioinformatics. 2000;16(9):847–9.

[84] Slatkin M. Linkage disequilibrium understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477–85. DOI:10.1038/nrg2361.

[85] Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2. DOI:10.1093/bioinformatics/btp187.

[86] Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics. 1995;29(2):311–22.

[87] Haydon DT, Bastos AD, Awadalla P. Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments. J Gen Virol. 2004;85(Pt 5):1095–100. DOI:10.1099/vir.0.19588-0.

[88] Alejska M, Kurzyñska-Kokorniak A, Broda M, Kierzek R, Figlerowicz M. How RNA viruses exchange their genetic material. Acta Biochem Pol. 2001;48(2):391–408.

[89] Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? Nat Rev Microbiol. 2011;9(8):617–26. DOI:10.1038/nrmicro2614.

[90] Schierup MH, Hein J. Recombination and the molecular clock. Mol Biol Evol. 2000;17(10):1578–9.

[91] Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol. 2002;54(3):396–402.

[92] Martin DP, Lemey P, Posada D. Analysing recombination in nucleotide sequences. Mol Ecol Resour. 2011;11(6):943–55. DOI:10.1111/j.1755-0998.2011.03026.x.

[93] Routh A, Johnson JE. Discovery of functional genomic motifs in viruses with ViReMa–a Virus Recombination Mapper–for analysis of next-generation sequencing data. Nucleic Acids Res. 2014;42(2):e11–e11. DOI:10.1093/nar/gkt916.

[94] Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010;26(1):2462–3. DOI:10.1093/bioinformatics/btq467.

[95] Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 2000;16(6):562–3. DOI:10.1093/bioinformatics/16.6.562.

[96] Salminen MO, Carr JK, Burke DS, McCutchan FE. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses. 1995;11(11):1423. DOI:10.1089/aid.1995.11.1423.

[97] Martin D, Posada D, Crandall K, Williamson C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses. 2005;21(1):98–102. DOI:10.1089/aid.2005.21.98.

[98] Smith JM. Analyzing the mosaic structure of genes. J Mol Evol. 1992;34(2):126–9.

[99] Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci. 2001;98(24):13757–62. DOI:10.1073/pnas.241370698.

[100] Boni MF, Posada D, Feldman MW. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics. 2007;176(2):1035–47. DOI:10.1534/genetics.106.068874.

[101] Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. Virology. 1999;265(2):218–25. DOI:10.1006/viro.1999.0056.

[102] Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics. 2000;16(7):573–82. DOI:10.1093/bioinformatics/16.7.573.

[103] Holmes EC, Worobey M, Rambaut A. Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol. 1999;16(3):405–9.

[104] McGuire G, Wright F. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics. 2000;16(2):130–4. DOI:10.1093/bioinformatics/16.2.130.

[105] Weiller GF. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol. 1998;15(3):326–35.

[106] Lemey P, Lott M, Martin DP, Moulton V. Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. BMC Bioinform. 2009;10(1):126.

[107] Delport W, Poon AF, Frost SD, Pond SLK. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010;26(19):2455–7. DOI: 10.1093/bioinformatics/btq429.

[108] Pond SLK, Frost SD. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 2005;21(10):2531–3. DOI:10.1093/bioinformatics/bti320.

[109] Pond SLK, Muse SV. HyPhy: hypothesis testing using phylogenies. In: Nielsen R, editor. Statistical methods in molecular evolution. New York: Springer; 2005. p. 125–181. DOI:10.1093/bioinformatics/bti079.

[110] Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Mol Biol Evol. 1999;16(10):1315–28.

[111] Pond SLK, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22(5):1208–22.

[112] Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SK. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8(7):e1002764. DOI:10.1093/molbev/msi105.

[113] Kolekar P, Kale M, Kulkarni-Kale U. In: Lopes H, editor. Molecular evolution & phylogeny: what, when, why & how? Croatia: InTech Open Access Publisher; 2011. DOI: 10.5772/20225.

[114] Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012;13(5):303–14. DOI:10.1038/nrg3186.

[115] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.

[116] Cavalli-Sforza LL, Edwards AW. Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet. 1967;19(3):233.

[117] Fitch WM, Margoliash E. Construction of phylogenetic trees. Science. 1967;155(3760): 279–84.

[118] Desper R, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol. 2002;9(5):687–705. DOI: 10.1089/106652702761034136.

[119] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.

[120] Cheng J, Cao F, Liu Z. AGP: a multimethods web server for alignment-free genome phylogeny. Mol Biol Evol. 2013;30(5):1032–7. DOI:10.1093/molbev/mst021.

[121] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–73. DOI:10.1093/molbev/mss075.

[122] Kolekar PS, Kale MM, Kulkarni-Kale U. Genotyping of Mumps viruses based on SH gene: Development of a server using alignment-free and alignment-based methods. Immunome Res. 2011;7(3):1–7.

[123] Kolekar P, Kale M, Kulkarni-Kale U. Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping. Mol Phylogenet Evol. 2012;65(2):510–22. DOI:10.1016/j.ympev.2012.07.003.

[124] Kolekar P, Hake N, Kale M, Kulkarni-Kale U. WNV Typer: a server for genotyping of West Nile viruses using an alignment-free method based on a return time distribution. J Virol Methods. 2014;198(1):41–55. DOI:10.1016/j.jviromet.2013.12.012.

[125] King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ, editors. Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses. San Diego: Elsevier Academic Press; 2012.

[126] Rozanov M, Plikat U, Chappey C, Kochergin A, Tatusova T. A web-based genotyping resource for viral sequences. Nucleic Acids Res. 2004;32(Web Server issue):W654–9.

[127] Lu G, Rowley T, Garten R, Donis RO. FluGenome: a web tool for genotyping influenza A virus. Nucleic Acids Res. 2007;35(Web Server issue):W275–9.

[128] Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JL, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345(6202):1369–72. DOI:10.1126/science.1259657.

[129] Huttner A, Dayer JA, Yerly S, Combescure C, Auderset F, Desmeules J, Eickmann M, Finckh A, Goncalves AR, Hooper JW, Kaya G, Krähling V, Kwilas S, Lemaître B, Matthey A, Silvera P, Becker S, Fast PE, Moorthy V, Kieny MP, Kaiser L, Siegrist CA; VSV-Ebola Consortium. The effect of dose on the safety and immunogenicity of the VSV Ebola candidate vaccine: a randomised double-blind, placebo-controlled phase 1/2 trial. Lancet Infect Dis. 2015; 15(10): 1156 - 1166. DOI:10.1016/S1473-3099(15)00154-1.

[130] McElroy K, Thomas T, Luciani F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microb Inform Exp. 2014;4(1): 1. DOI:10.1186/2042-5783-4-1.

[131] Rappuoli R. Vaccines, emerging viruses, and how to avoid disaster. BMC Biol. 2014;12(1):100. DOI:10.1186/s12915-014-0100-6.

IntechOpen