# CHAPTER 11

# Pan-genomics of virus and its applications

Marta Giovanetti[a,b], Alvaro Salgado[b], Vagner de Souza Fonseca[a,b,c], Fraga de Oliveira Tosta[b], Joilson Xavier[a,d], Jaqueline Goes de Jesus[a,d], Felipe Campos Melo Iani[b,e], Talita Emile Ribeiro Adelino[b,e], Fernanda Khouri Barreto[d,f], Nuno Rodrigues Faria[g], Tulio de Oliveira[c], Luiz Carlos Junior Alcantara[a,b]

[a]Laboratório de Flavivírus, IOC, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil
[b]Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[c]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa
[d]Laboratório de Patologia Experimental, Instituto Gonçalo Moniz, Fiocruz Bahia, Salvador, Brazil
[e]Fundação Ezequeil Dias (Funed), Belo Horizonte, Brazil
[f]Instituto Multidisciplinar em Saúde—IMS, Universidade Federal da Bahia (UFBA), Salvador, Brazil
[g]Department of Zoology, University of Oxford, Oxford, United Kingdom

## 1 Next-generation sequencing strategies

Exploring the genetic information of viruses has been made possible due to the technology of DNA sequencing. Currently, there are 4958 described virus species, according to the International Committee on Taxonomy of Viruses (ICTV, 2018). Although many species still do not have their genomes sequenced, viral species with medical, biotechnological, and environmental relevance usually have more than one complete or partial publicly deposited genomes [1].

This large and diverse viral genetic information available in public databases allows us to begin addressing the genetic complexity of viruses, which originates from several molecular mechanisms, including insertion/deletion events, different rates of nucleotide substitution, as well as intra- and inter-genotype recombination and reassortment events [2]. These mechanisms directly affect the genetic repertoire of viral populations of different hosts and habitats, leading to important implications in molecular diagnosis, pathogenesis, and viral epidemiology [3].

Partial viral genomes sequencing has been used to: (i) detect drug resistance in both DNA and RNA viruses [4, 5] and (ii) perform phylogenetic analyses for the assignment of genotypes [6]. Despite the broad array of discoveries and advancements brought by that approach, whole genome sequencing (WGS) consistently provides more information than the sequencing of a reduced number of genes. Therefore, WGS allows for: (i) the detection of all known drug-resistant variants and the identification of new

*Pan-genomics: Applications, Challenges, and Future Prospects*
https://doi.org/10.1016/B978-0-12-817076-2.00011-1

ones; (ii) the identification of mutations associated with disease transmission or severity; (iii) better phylogenetic resolution; and (iv) genomic surveillance [7–10].

Regarding the different methods of DNA sequencing, the first strategy applied was that of chain termination with dideoxynucleotides (ddNTPs)—(Sanger sequencing) [11]. Commonly used for confirmation in some cases, given its high accuracy, it has an extremely low throughput, as well as being laborious and time consuming. The scenario started to change in 2004, with the emergence of the first DNA sequencing technologies known as NGS (next-generation sequencing), allowing for a new approach of large-scale sequencing (HTS—high-throughput sequencing) [12].

In the following years, several *Second-Generation Sequencing* platforms were developed, based mainly on the following technologies: (i) sequencing by ligation (SOLiD), (ii) ion sensing synthesis technology (IonTorrent), and (iii) sequencing by synthesis (Illumina) [13]. Second-generation platforms allowed for a more in-depth characterization of the genomic variability of viruses, while providing large amounts (millions of reads) of data for each individual sequence for the same clone or amplicon. Despite this, its major limitation is the size of each individual sequenced, not being possible to obtain complete viral genomes in a single sequencing reaction.

Recently, the development of single-molecule third-generation sequencing approaches is now providing the first promising results in the sequencing of complete viral genomes [14,15]. Two platforms are currently available: Pacific Biosciences (PacBio) RS and RS II systems, and the Oxford Nanopore Technologies (ONT) systems (MinION, GridION, and PromethION). PacBio uses Single-Molecule, Real Time Technology Sequencing (SMRT) (PacBio, http://www.pacb.com/).

Each SMRT cell of the PacBio RS II system has a typical throughput of 0.5–1 GB, with an average read length of 10 kb. Despite this, PacBio reads still present a significantly higher error rate when compared to second- and first-generation sequencing technologies (>10%–15%) [16].

In 2014, the MinION from ONT was released to early access users [17], heralding the potential for highly portable "lab-in-a-suitcase" sequencing, which is capable of sequencing DNA or RNA in a real-time scale, with ultra-long-reads. The MinION is pocket sized and is controlled and powered through a laptop USB connection.

In this technology, the DNA or RNA strands passes through various nanopores, which connect the two sides of a semiconductive layer, anchored by specialized proteins. A voltage is applied between the surfaces of the layer and, as the DNA or RNA strands move through the nanopores, each of its nucleotides creates a characteristic disruption in the electrical current flowing through the pore. This nanopore signal, which is different for each type of nucleotide, is used to determine the sequence of bases on the DNA or RNA strand [18].

Ongoing improvements to the launched barcoding kits in the nanopore sequencing technology had the potential to increase the number of generated genomes per sequencing run from 12 to 96, which could also increase the number of genome sequences available from affected regions and allow more detailed investigations of the association between pathogens mutations and environmental context with less costs.

However, nanopore technology also has a lower accuracy when compared to older technologies, with an error profile <10% insertion-deletion mutations (indels) rate [19].

For data analysis, most bioinformatics tools take FASTA or FASTQ files as input, where base calling has already been done during the sequencing process or off line with the sequencer. For new platforms in their early stages, however, original raw data files may be useful for some applications. Currently, the MinION outputs one FAST5 file per read. Much like the h5 file format adopted by PacBio, the FAST5 file format is based on the hierarchical data format 5 (HDF5) standard (https://www.hdfgroup.org). FAST5 files have a hierarchical structure, meaning that they can store both the metadata associated with a read, along with the events (such as aggregated bulk current measurements) preprocessed by the sequencing device [19]. Despite this, nanopore long reads simplify assembly and sequencing of repetitive regions and speed up the identification of new species and metagenomic experiments. For those reasons, the MinION sequencer is getting much attention from the genomic community, mainly for genomic viral surveillance and genomic epidemiology areas, as they can benefit from the real-time nature of this sequencing platform. Importantly, the MinION has been used in field situations, including in diagnostic tent laboratories during the Ebola epidemic [20,21] and in a roving bus-based mobile laboratory in Brazil as part of the ZIBRA project (http://www.zibraproject.org) [22]. Others have taken the MinION to more extreme environments where even the smallest traditional benchtop sequencer could not go, including the Arctic [23] and Antarctic [24], a deep mine [25], and zero gravity aboard the reduced-gravity aircraft [26], and the International Space Station [27].

The shortage of complete genomic sequences represents a limiting point for the study of viral genetic divergence as well as of population dynamics (genotypes and subgenotypes), pathogenesis and vectors associated with virus transmission among human populations [9, 28]. In this context, sequencing of viral genomes plays an important role in the fight against emerging and reemerging epidemics, as well as in the early detection and/or identification of new potential emerging pathogens through metagenomics approaches. Metagenomics, in this sense, can be used as a tool to monitor, at an early stage, the introduction of new pathogens in specific regions. It may have important applications for the epidemiological surveillance, outbreak investigation, and diagnosis of infectious diseases [29] of both known and unknown pathogens. NGS-based metagenomics, therefore, can be used as a complementary tool to monitor the emergence and spread of new human pathogens, a central concern in public health in tropical regions.

Therefore, as sequencing chemistry and technologies progress, such techniques are likely to become key tools for the construction of viral pan-genomes. We expect that computational pan-genomics will allow increased power and accuracy, for example, by allowing the pan-genome structure of a viral population to be directly compared with that of a susceptible host population. Portable genome sequencing technology and digital epidemiology platforms form the foundation for both real-time pathogen and disease surveillance systems and outbreak response efforts, all of which exist within the One Health context, in which surveillance, outbreak detection, and response span the human, animal and environmental health domains.

## 2 Genomic surveillance

Infectious diseases continue to be one of the leading causes of death worldwide [30] and pathogens such as viruses can be considered notorious mutation machines. They can evolve and spread rapidly, leading to the emergence of newly mutated human pathogens, more virulent strains, as well as antibiotic- and drug-resistant organisms [31,32]. In this context, genomic surveillance aims are: (i) to perform global surveillance of pathogens using WGS; (ii) to understand drug resistance, emergence, and spread of viral pathogens; and (iii) to provide actionable data.

Several approaches have been developed and are widely used for the quick detection and identification of viral pathogens (i.e., diagnostics). Some of them are based on different serological and molecular strategies including, for example, assays based on real-time polymerase chain reaction [33]. Even though these kinds of approaches present high sensitivity and specificity for their purpose, they are more suitable for diagnostics only and cannot provide detailed genomic information [34].

Bearing these limitations in mind, the main point of developing new genomic surveillance tools is to answer the following inquiry: what sort of questions are important for genomic surveillance that cannot be addressed by conventional RT-qPCR or serology? (i) RT-qPCR assays do not allow genotype classification, neither does it help identify particular and/or characteristic transmission routes; (ii) RT-qPCR assays also do not allow to determine how fast a viral pathogen is being transmitted and in what direction it is spreading; (iii) serological and molecular assays also cannot help identify epidemiologically linked individuals, neither predict future outbreaks; and (iv) finally, serological and some molecular approaches cannot help to identify novel pathogenic agents and are, therefore, unsuitable for pathogen discovery [34].

NGS technologies produce significantly more raw data than other molecular diagnostic assays, including Sanger sequencing, and are also capable of informing not just pathogen diagnostics but also epidemiology [35]. This is why WGS of viral genomes by using new technologies plays an important role in the fight against emerging and reemerging epidemics [36,37]. The availability of high-throughput sequencing has also provided immense insights into the ecology of health-care-associated pathogens [38]. Therefore, real-time sequencing of entire pathogen genomes has become a standard and indispensable research tool for the critical role of genomic surveillance in the prevention and control of emerging infectious diseases [39], which justifies why NGS can be considered a powerful strategy that also allows the discovery of novel potential viral pathogens [34,40].

Considering pathogen surveillance in mind, bioinformatics tools and the combination of genomic and epidemiological data from viral infections can give essential information for understanding the past and the future of an epidemic, because genomic data generated by real-time sequencing can provide important information on how and when viruses were introduced in a particular site, their pattern, and determinants of dissemination in neighboring locations and the extent of genetic diversity, that is, its dynamics, making it possible to establish an effective surveillance framework on tracking the spread of infections to other geographic regions [28,40]. In this context, recently established international networks for real-time, portable genomic sequencing, genomic surveillance, and data analysis made it possible to monitor the evolution of viral genomes, to understand the origins of outbreaks and epidemics, to predict future outbreaks and to assist in the maintenance of updated diagnostic methods [40–43]. In addition, genomic surveillance framework allows to determine, through genome sequencing, the real-time molecular epidemiology of viruses circulating and cocirculating in different regions in a specific area, and also to detect and characterize the early emergence of new pathogens in large urban centers, generating data that can inform outbreak control responses [28,43]. Generated data regarding the molecular, epidemiological, phylogenetic, and geographical aspects of circulating viral pathogens in a specific setting contribute to a better understanding of those viral infections in a national and international context, assuming an important role in solving issues relevant to Public Health [44]. As a result, studies involving more in-depth molecular and dispersion analysis of circulating pathogens may help the World Health Organization appropriately adopt measures to control epidemics and to monitor the dynamics and spreading of new viral strains. However, even though NGS has advantages over diagnostics routine, all of the different strategies and technologies, developed by Illumina, Thermo Scientific, Oxford Nanopore, and others, are not yet considered a panacea. Remaining challenges include dealing with high data throughput, which requires sophisticated computational processing as well as the annotation of large amounts of sequencing data, high DNA or RNA input sample requirements (in some cases hundreds of nanograms), which often raises the need for previous PCR-based amplification approaches. On top of all this, there are relatively few researchers in the area with sufficient bioinformatics expertise and who are able to engage in near-patient or disease surveillance activities [44].

## 3 Genomic epidemiology

Genomic epidemiology has been applied to many outbreaks in the past few years and is becoming a widely accepted method to investigate outbreaks [45]. The use of WGS to understand infectious disease transmission and epidemiology is crucial to understanding the direction of an outbreak both in national and international contexts. The characterization of the evolutionary history and the geographic and temporal dissemination of viral pathogens could allow the identification of strains associated with a greater epidemic potential, suggesting targets for the development of more effective therapeutic interventions, and then allowing the establishment of an effective surveillance framework in the tracking of the spread of these strains to other geographic regions [46]. The goal of this kind of approach is to use the population structure of the pathogen to understand the overall dynamics of the epidemic [47]. Moreover, with improvements in sequencing technology and continuing optimization and standardization of bioinformatics algorithms, genomic epidemiology investigations can now be conducted during the course of an ongoing outbreak to provide real-time guidance for infection control interventions [47].

In addition, the rapid development of sequencing technologies has led to an explosion of pathogen sequencing data, which are increasingly collected as part of routine surveillance or clinical diagnostics. While sequencing has become cheaper, the analysis of sequence data has become a critical bottleneck.

Molecular epidemiological techniques can reconstruct the temporal and spatial spread of an outbreak. Similarly, by linking samples that originate from different geographic locations, phylogeographic methods can reconstruct the geographic spread and can differentiate distinct introductions. In this context, the use of powerful bioinformatics tools in the field of phylodynamics, defined as the study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies, can support genomic surveillance and epidemiology. Phylodynamic models may aid in dating epidemic and pandemic origins and viral spread by mapping the geographic movement of a particular pathogen population in a specific area. Phylodynamic approaches have also been used to better understand viral transmission dynamics and spread within infected hosts. Such approaches can also be useful in ascertaining the effectiveness of viral control efforts, particularly for diseases with low reporting rates [48].

The potential exists to move from pathogen genomics, providing static "snapshots" of epidemics, often months after the cases occurred, to a situation where data are produced in real time, providing a detailed picture of the epidemic that is only a few days old. Such rapid results are crucial if the intention is to intervene in an outbreak rather than simply document it in retrospect.

## 4 Bioinformatic tools

NGS techniques have transformed genomic studies from the analysis of single or few genomes to an ever-increasing amount of genomic data, bringing with it the need to develop novel techniques to efficiently treat, novel tools to assemble, analyze, and derive useful information from overwhelmingly large datasets.

One of the ways to derive meaningful and useful information from a large genomic dataset is through pan-genomics. According to Vernikos et al. [49], a pan-genome defines the whole genetic repertoire of a phylogenetic clade and describes the set of all sequence entities (ORFs, genes, etc.) belonging to the genomes of interest. The union, intersection, and subsetting of units in the pan-genome can be classified as core genes, dispensable genes, and strain-specific genes.

The analysis of pan-genomes can uncover significant information regarding the genomes of interest. According to Carlos Guimaraes et al. [50], pan-genomic studies can help understand pathogen evolution, niche adaptation, population structure, and host interaction. Furthermore, it can help in vaccine and drug design, as well as in the identification of virulence genes.

In the context of virus investigations, pan-genomics, and bioinformatics in general face great challenges. Rapid extraction of genomic features with an evolutionary signal will facilitate evolutionary analyses ranging from the reconstruction of species phylogenies to tracing epidemic outbreaks.

### 4.1 Bioinformatic tools used in pan-genomic studies

According to Xiao et al. [51], Panseq [52], and PGAP (pan-genomes analysis pipeline) [53] were ranked as the two top most popular packages based on cumulative citations of peer-reviewed scientific publications at the end of 2014. Other tools applicable to virus pan-genomics include: EDGAR, ITEP, GET_HOMOLOGUES, CASTOR, and Genome Detective. Most pan-genome bioinformatic tools are based on orthologous and paralogous gene identification [50], and the functions of these software packages and tools usually include categorizing orthologous genes, calculating pan-genomic profiles, integrating gene annotations, and constructing phylogenies [51].

#### *4.1.1 Panseq—Pan-genome sequence analysis program*

As mentioned by Carlos Guimaraes et al. [50], Panseq is a freely available web-tool written in BioPerl, which is available at http://76.70.11.198/panseq. Panseq defines the core and accessory genome based on the sequence identity and segmentation length. The NRF (novel region finder) module first splits the genome sequence into fragments with predefined sizes, then the MUMmer alignment program [54] identifies the sequences and contiguous regions that are present or absent in the database. Next, the CAGF module (Core and Accessory Genome Finder) compares each individual fragment sequence to all sequences, adding single sequences that fit in with predefined parameters to the pan-genome. Each newly added to fragment sequence is used for subsequent comparisons, continuing this loop until all of the fragment sequences have been tested [53]. Panseq, according to Laing et al. [52], is able to determine core and accessory regions of genome assemblies and identify SNPs among the core genomic regions. In addition, it can select the most discriminatory loci among the accessory loci or core gene SNPs. Panseq, however, is not able to provide pan-genomic profile and functional enrichment analysis that is important for discriminating the functional relevance of the pan-genomic elements.

#### 4.1.2 PGAP—Pan-genome analysis pipeline

PGAP is a stand-alone tool available at http://pgap.sf.net developed by Laing et al. [52] to perform pan-genome analysis, genetic variation, evolution, and function analysis of gene clusters [50]. The software uses two methods to calculate all of the analyses: (i) the GF method to detect homologous genes, and (ii) the MP method to detect orthologous genes.

The GF method is based on the protein BLAST and MCL (Markov clustering) algorithms. All of the protein sequences are brought together, and protein BLAST is performed; the results are filtered and clustered using the MCL algorithm [55,56]. The MP method is based on two algorithms: (i) Inparanoid to search orthologous and paralogous genes using BLAST. Then, the pairwise ortholog clusters are moved to (ii) MultiParanoid, which was specifically developed to search for gene clusters among multiple strains [50,55,57–59].

#### 4.1.3 EDGAR (efficient database framework for comparative genome analyses using BLAST score ratios)

EDGAR is a web-tool available at https://edgar.computational.bio.uni-giessen.de/ [50]. It is designed to automatically perform genome comparisons in a high-throughput approach. It provides novel analysis features and significantly simplifies the comparative analysis of related genomes. The software supports a quick survey of evolutionary relationships and simplifies the process of obtaining new biological insights into the differential gene content of kindred genomes. Visualization features, like synteny plots or Venn diagrams, are offered to the scientific community through a web-based and therefore platform-independent user interface, where the precomputed data sets can be browsed [60]. According to Carlos Guimaraes et al. [50], this software performs homology analyses based on a specific cutoff that is automatically adjusted to the query data. The orthology analysis to calculate pan-genome, core-genome, and singletons is performed using BLAST score ratio values.

#### 4.1.4 ITEP—Integrated toolkit for the exploration of microbial pan-genomes

ITEP is a stand-alone toolkit that is available for download at https://price.systemsbiology.net/itep [50]. It was developed to predict protein families, orthologous genes, functional domains, pan-genome, and metabolic networks for related microbial species [61]. Its workflow consists of a three-step process: data input, database building (startup scripts), and database analysis [50]. ITEP receives three different types of data: GenBank file format, organism file format, and groups file format, and all of the inputs require preprocessing before running the ITPEP toolkit (for more details, see the ITEP documentation). In database building, scripts are run to predict the gene locations, BLAST results, and clustering results. Finally, the package can perform core and variable genes analyses, phylogenies, metabolic reconstructions, and gene gain and loss patterns [50].

### *4.1.5 GET_HOMOLOGUES*

GET_HOMOLOGUES is a stand-alone and open-source toolkit that was written in Perl and R that can be installed on personal machines. It was developed to perform pan-genome and comparative-genomic analysis [50,62].

### PanFunPro: PAN-genome analysis based on FUNctionalPROfiles

PanFunPro is a stand-alone tool for pan-genome analysis using functional domains from HMM (hidden Markov models) to group homologous proteins into families based on their functional domain content [50,63,64]. In addition to pan-genome analyses, the software performs homology detection and genome annotation using HMM, genome and proteome estimation as well as gene ontology (GO) information [65].

### *4.1.6 CASTOR*

The classification and annotation of virus genomes constitute important assets in the discovery of genomic variability, taxonomic characteristics, and disease mechanisms. Existing classification methods are often designed for specific well-studied families of viruses. Thus, the viral comparative genomic studies could benefit from more generic, fast, and accurate tools for classifying and typing newly sequenced strains of diverse virus families [66].

According to Rose et al. [67], CASTOR is a virus classification platform based on machine learning methods, inspired by a well-known technique in molecular biology: restriction fragment length polymorphism. It simulates, in silico, the restriction digestion of genomic material by different enzymes into fragments. It uses two metrics to construct feature vectors for machine learning algorithms in the classification step. The performance of CASTOR, its genericity, and robustness could permit performing novel and accurate large-scale virus studies. The CASTOR web platform provides an open access, collaborative, and reproducible machine learning classifiers. CASTOR can be accessed at http://castor.bioinfo.uqam.ca.

### 4.1.7 Genome Detective

According to Vilsker et al. [68], the analysis of viral genomes is especially challenging because of their high variability and deviation from reference genomes. This is aggravated by the increasing speed of identification, the continuous emergence of new viruses, and the relative rareness of viral fragments in metagenomic analyses.

Genome Detective (http://www.genomedetective.com/app/typing tool/virus/) was developed to address this problem [69]. It is an easy to use web-based software application that assembles the genomes of viruses quickly and accurately, designed to generate and analyze whole or partial viral genomes directly from NGS reads within minutes. The application gains accuracy by using a novel alignment method that uses a combination of amino acids and nucleotide scores to construct genomes by the reference-based linking of de novo contigs. Speed and accuracy were also gained by using DIAMOND with a UniProt90 reference dataset to sort viral taxonomy units. The use of DIAMOND and UniRef90 allowed Genome Detective to identify viral short reads at least 1000 times faster than if we used Blastn and the viral nt database of NCBI [70]. The software was optimized using synthetic datasets to represent the great diversity of virus genomes. The application was then validated with NGS data of hundreds of viruses. User time is minimal, and it is limited to the time required to upload the data [69]. According to the authors [69], Genome Detective accepts unprocessed paired-end or single reads generated by NGS platforms in FASTQ format and/or processed FASTA sequences. Candidate viral reads are identified using the protein-based alignment method, DIAMOND [70]. It uses the viral subset of the Swiss-Prot UniRef90 protein database, which contains representative clusters of proteins linked to taxonomy IDs, to improve sensitivity and speed, which was also improved by first sorting short reads into groups, or buckets. The objective is to run a separate metagenomic de novo assembly in each bucket; so, all reads of one virus species have to be assigned to the same bucket. Each bucket is then identified using the taxonomy ID of the lowest common ancestor of the hits identified by DIAMOND.

Once all of the reads have been sorted in buckets; each bucket is then de novo assembled separately using SPAdes [71] for single-ended reads or metaSPAdes [71] for paired-end reads. Blastx and Blastn are used to search for candidate reference sequences against the NCBI RefSeq virus database. Genome Detective combines the results for every detected contig at the amino acid and nucleotide (nt) level by calculating a total score that is a sum of the total nt score plus total amino acid score. It then chooses the five best scoring references for each contig to be used during the alignment. The contigs for each individual species are joined using Advanced Genome Aligner (AGA) [72]. AGA is designed to compute the optimal global alignment considering simultaneously the alignment of all annotated coding sequences of a reference genome. This makes alignments using Genome Detective more sensitive and accurate as both nt and protein scores are taken into account in order to produce a consensus sequence from the de novo contigs. A report is generated, referring to the final contigs and consensus sequences, available in

FASTA format. The report also contains detailed information on filtering, assembly, and consensus sequence. Web-based graphics are also available. In addition, the user can produce a bam file with BWA [73] using the reference or de novo consensus sequence by selecting the detailed report and access viral phylogenetic identification tools [74] directly from the interface. The authors found that, for large NGS and metagenomic datasets, Genome Detective substantially reduces computational cost without compromising the quality of the result. However, the construction of de novo whole genomes from metagenomic samples depends on the number of reads, the virus genome size, and read length. Genome Detective is linked to popular virus-specific typing tools [74], which allow phylogenetic classification below species level.

## 4.2 Future improvements

According to Xiao et al. [51], additional annotation information, such as that of epigenetics, noncoding RNAs, insertion elements, conserved structural elements, and pseudogenes remains to be implemented into the relevant software packages. The authors highlight the transition from the representation of reference genomes as strings to representations as graphs as a prominent example for a computational paradigm shift. In addition, improvements on genome assembly using machine learning techniques are proposed by Padovani de Souza et al. [75]. Finally, in order to help better use all the information acquired by high-throughput real-time sequencing and its analysis, text mining and knowledge discovery techniques, integrated with medical and scientific literature and gene family and metabolic pathway databases, could help generate new insights and speed up discoveries.

# 5 Conclusions

High-throughput real-time NGS projects have transformed the field of bioinformatics from single-genome studies to pan-genome analyses. The limiting factor now is no longer data rarity, but immense data availability and dimensionality. In this new context, bottom-up analysis stemming from big data provides great challenges but also great rewards.

# References

[1] ICTV Master Species List 2018a v1, International Committee on Taxonomy of Viruses (ICTV), Available from https://talk.ictvonline.org/files/master-species-lists/m/msl/7992%3e, 2018.

[2] S. Duffy, L.A. Shackelton, E.C. Holmes, Rates of evolutionary change in viruses: patterns and determinants, Nat. Rev. Genet. 9 (2008) 267–276.

[3] E. Domingo, Mechanisms of viral emergence, Vet. Res. 41 (2010) 38–312.

[4] C.J. Houldcroft, J.M. Bryant, D.P. Depledge, B.K. Margetts, J. Simmonds, S. Nicolaou, H.J. Tutill, R. Williams, A.J.J. Worth, S.D. Marks, P. Veys, E. Whittaker, J. Breuer, Detection of low frequency

multi-drug resistance and novel putative maribavir resistance in immunocompromised pediatric patients with cytomegalovirus, Front. Microbiol. 7 (2016) 13–17.

[5] H. Zaraket, R. Saito, Y. Suzuki, T. Baranovich, C. Dapat, I. Caperig-Dapat, H. Suzuki, Genetic makeup of amantadine-resistant and oseltamivir-resistant human influenza A/H1N1 viruses, J. Clin. Microbiol. 48 (2010) 1085–1092.

[6] T. Gräf, H. Machado Fritsch, R.M. de Medeiros, D. Maletich Junqueira, S. Esteves de Matos Almeida, A.R. Pinto, Comprehensive characterization of HIV-1 molecular epidemiology and demographic history in the Brazilian region most heavily affected by AIDS, J. Virol. 90 (2016) 8160–8168.

[7] S. Ramirez, L.S. Mikkelsen, J.M. Gottwein, J. Bukh, Robust HCV genotype 3a infectious cell culture system permits identification of escape variants with resistance to sofosbuvir, Gastroenterology 151 (2) (2016) 973–985.

[8] L. Yuan, X.-Y. Huang, Z.-Y. Liu, F. Zhang, X.-L. Zhu, J.-Y. Yu, X. Ji, Y.-P. Xu, G. Li, C. Li, H.-J. Wang, Y.-Q. Deng, M. Wu, M.-L. Cheng, Q. Ye, D.-Y. Xie, X.-F. Li, X. Wang, W. Shi, B. Hu, P.-Y. Shi, Z. Xu, C.-F. Qin, A single mutation in the prM protein of Zika virus contributes to fetal microcephaly, Science 358 (2017) 933–936.

[9] N.R. Faria, J. Quick, I.M. Claro, J. Theze, J.G. de Jesus, M. Giovanetti, et al., Establishment and cryptic transmission of Zika virus in Brazil and the Americas, Nature 546 (2017) 406–410.

[10] J.L. Gardy, N.J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system, Nat. Rev. Genet. 19 (1) (2018) 9–20.

[11] F. Sanger, S. Nicklen, A.R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. U. S. A. 74 (12) (1977) 5463–5467.

[12] T. Jarvie, Next generation sequencing technologies, Drug Discov. Today Technol. 2 (3) (2005) 255–260.

[13] S. Goodwin, J.D. McPherson, W.R. McCombie, Coming of age: ten years of next-generation sequencing technologies, Nat. Rev. Genet. 17 (2016) 333–351.

[14] J. Wang, N.E. Moore, Y.M. Deng, D.A. Eccles, R.J. Hall, MinION nanopore sequencing of an influenza genome, Front. Microbiol. 6 (2015) 766.

[15] N. Beerenwinkel, H.F. Günthard, V. Roth, K.J. Metzner, Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data, Front. Microbiol. 3 (2012) 329.

[16] N. Nagarajan, M. Pop, Sequence assembly demystified, Nat. Rev. Genet. 14 (2013) 157–167.

[17] M. Jain, H.E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community, Genome Biol. 17 (2016) 239.

[18] C.L. Ip, M. Loose, J.R. Tyson, M. de Cesare, B.L. Brown, M. Jain, et al., MinION analysis and reference consortium: phase 1 data release and analysis, F1000Res 4 (2015) 1075.

[19] T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, et al., Assessing the performance of the Oxford Nanopore technologies MinION, Biomol. Detect. Quantif. 3 (2015) 1–8.

[20] J. Quick, N.J. Loman, S. Duraffour, J.T. Simpson, E. Severi, L. Cowley, Real-time, portable genome sequencing for Ebola surveillance, Nature 530 (2016) 228.

[21] T. Hoenen, A. Groseth, K. Rosenke, R.J. Fischer, A. Hoenen, S.D. Judson, Nanopore sequencing as a rapidly deployable Ebola outbreak tool, Emerg. Infect. Dis. 22 (2016) 331.

[22] N.R. Faria, E.C. Sabino, M.R. Nunes, L.C.J. Alcantara, N.J. Loman, O.G. Pybus, Mobile real-time surveillance of Zika virus in Brazil, Genome Med. 8 (2016) 97.

[23] A. Edwards, A.R. Debbonaire, B. Sattler, L.A. Mur, A.J. Hodson, Extreme Metagenomics Using Nanopore DNA Sequencing: A Field Report From Svalbard, 78N, 2016.

[24] S.S. Johnson, E. Zaikova, D.S. Goerlitz, Y. Bai, S.W. Tighe, Real-time DNA sequencing in the Antarctic dry valleys using the Oxford Nanopore sequencer, J. Biomol. Tech. 28 (2017) 2–7.

[25] A. Edwards, A. Soares, S. Rassner, P. Green, J. Felix, A. Mitchell, Deep sequencing: intra-terrestrial metagenomics illustrates the potential of off-grid Nanopore DNA sequencing, bioRxiv (2017).

[26] A.B. McIntyre, L. Rizzardi, M.Y. Angela, N. Alexander, G.L. Rosen, D.J. Botkin, Nanopore sequencing in microgravity, NPJ Microgravity 2 (2016) 16035.

[27] S.L. Castro-Wallace, C.Y. Chiu, K.K. John, S.E. Stahl, K.H. Rubins, A.B. McIntyre, Nanopore DNA sequencing and genome assembly on the International Space Station, Sci. Rep. 7 (2017) 18022.

[28] N.R. Faria, M.U. Kraemer, S. Hill, J.G. de Jesus, R.S. de Aguiar, F.C. Iani, et al., Genomic and epidemiological monitoring of yellow fever virus transmission potential, Science (2018) https://doi.org/10.1126/science.aat7115.

[29] S. Sardi, S. Somasekar, S.N. Naccache, A.C. Bandeira, L.B. Tauro, G.S. Campos, et al., Co-infections from Zika and chikungunya virus in Bahia, Brazil identified by metagenomic next-generation sequencing, J. Clin. Microbiol. 54 (9) (2016) 2348–2353.

[30] D.M. Morens, G.K. Folkers, A.S. Fauci, The challenge of emerging and re-emerging infectious diseases, Nature 430 (2004) 242–249.

[31] P. Daszak, A.A. Cunningham, A.D. Hyatt, Emerging infectious diseases of wildlife—threats to biodiversity and human health, Science 287 (2000) 443–449.

[32] S.S. Morse, Factors in the emergence of infectious diseases, Emerg. Infect. Dis. 1 (1995) 7–15.

[33] J. Versalovic, J.R. Lupski, Molecular detection and genotyping of pathogens: more accurate and rapid answers, Trends Microbiol. 10 (2002) 15–21.

[34] A.J. Sabat, A. Budimir, D. Nashev, R. Sá-Leão, J.M. van Dijl, F. Laurent, et al., Overview of molecular typing methods for outbreak detection and epidemiological surveillance, Euro Surveill. 18 (2013) 20380.

[35] J. Shendure, H. Ji, Next-generation DNA sequencing, Nat. Biotechnol. 26 (2008) 1135–1145.

[36] B.L. Haagmans, A.C. Andeweg, A.D.M.E. Osterhaus, The application of genomics to emerging zoonotic viral diseases, PLoS Pathog. 5 (2009).

[37] A.C. McHardy, B. Adams, The role of genomics in tracking the evolution of influenza A virus, PLoS Pathog. 5 (2009).

[38] P. Tang, J.L. Gardy, Stopping outbreaks with real-time genomic epidemiology, Genome Med. 6 (2014) 104.

[39] E.C. Holmes, Viral evolution in the genomic age, PLoS Biol. 5 (2007).

[40] J. Gardy, N.J. Loman, A. Rambaut, Real-time digital pathogen surveillance—the time is now, Genome Biol. 16 (2015) 155.

[41] J. Quick, N.D. Grubaugh, S.T. Pullan, I.M. Claro, A.D. Smith, K. Gangavarapu, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples, Nat. Protoc. 12 (2017) 1261.

[42] N.D. Grubaugh, J.T. Ladner, M.U. Kraemer, G. Dudas, A.L. Tan, K. Gangavarapu, Genomic epidemiology reveals multiple introductions of Zika virus into the United States, Nature 546 (2017) 401.

[43] J. Thézé, T. Li, L. du Plessis, J. Bouquet, M.U. Kraemer, S. Somasekar, Genomic epidemiology reconstructs the introduction and spread of Zika virus in Central America and Mexico, Cell Host Microbe 23 (2018) 855–864.

[44] N.J. Loman, C. Constantinidou, J.Z.M. Chan, M. Halachev, M. Sergeant, C.W. Penn, et al., High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity, Nat. Rev. Microbiol. 10 (2012) 599–606.

[45] K.J. Popovich, E.S. Snitkin, Whole genome sequencing—implications for infection prevention and outbreak investigations, Curr. Infect. Dis. Rep. 19 (2017) 15.

[46] S. Reuter, M.J. Ellington, E.J.P. Cartwright, C.U. Köser, M.E. Török, T. Gouliouris, et al., Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology, JAMA Intern. Med. 173 (2013) 1397–1404.

[47] M.R. Halachev, J.Z.-M. Chan, C.I. Constantinidou, N. Cumley, C. Bradley, M. Smith-Banks, et al., Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant Acinetobacter baumannii in Birmingham, England, Genome Med. 6 (2014) 70.

[48] A.J. Drummond, O.G. Pybus, A. Rambaut, R. Forsberg, A.G. Rodrigo, Measurably evolving populations, Trends Ecol. Evol. 18 (2003) 481–488.

[49] G. Vernikos, D. Medini, D.R. Riley, H. Tettelin, Ten years of pan-genome analyses, Curr. Opin. Microbiol. 23 (2015) 148–154.

[50] L. Carlos Guimaraes, et al., Inside the pan-genome—methods and software overview, Curr. Genomics 16 (4) (2015) 245–252.

[51] J. Xiao, Z. Zhang, J. Wu, J. Yu, A brief review of software tools for pangenomics, Genomics Proteomics Bioinformatics 13 (1) (2015) 73–76.

[52] C. Laing, et al., Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions, BMC Bioinformatics 11 (1) (2010) 461.
[53] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, J. Yu, PGAP: pan-genomes analysis pipeline, Bioinformatics 28 (3) (2012) 416–418.
[54] S. Kurtz, et al., Versatile and open software for comparing large genomes, Genome Biol. 5 (2) (2004) 34–87.
[55] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[56] A.J. Enright, S. Van Dongen, C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Res. 30 (7) (2002) 1575–1584.
[57] A. Alexeyenko, I. Tamas, G. Liu, E.L.L. Sonnhammer, Automatic clustering of orthologs and inparalogs shared by multiple proteomes, Bioinformatics 22 (14) (2006) 9–15.
[58] G. Ostlund, et al., InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, Nucleic Acids Res. 38 (2010) 196–203.
[59] M. Remm, C.E. Storm, E.L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, J. Mol. Biol. 314 (5) (2001) 1041–1052.
[60] J. Blom, et al., EDGAR: a software framework for the comparative analysis of prokaryotic genomes, BMC Bioinformatics 10 (2009) 154.
[61] M.N. Benedict, J.R. Henriksen, W.W. Metcalf, R.J. Whitaker, N.D. Price, ITEP: an integrated toolkit for exploration of microbial pan-genomes, BMC Genomics 15 (2014) 8.
[62] B. Contreras-Moreira, P. Vinuesa, GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis, Appl. Environ. Microbiol. 79 (24) (2013) 7696–7701.
[63] S.R. Eddy, Hidden Markov models, Curr. Opin. Struct. Biol. 6 (3) (1996) 361–365.
[64] O. Lukjancenko, M.C. Thomsen, M. Voldby Larsen, D.W. Ussery, PanFunPro: PAN-genome analysis based on FUNctionalPROfiles, F1000Res. 2 (2013) 15.
[65] The Gene Ontology Consortium, The gene ontology project in 2008, Nucleic Acids Res. 36 (2008) 440–444.
[66] M.A. Remita, A. Halioui, A.A. Malick Diouara, B. Daigle, G. Kiani, A.B. Diallo, A machine learning approach for viral genome classification, BMC Bioinformatics 18 (1) (2017) 208–239.
[67] R. Rose, B. Constantinides, A. Tapinos, D.L. Robertson, M. Prosperi, Challenges in the analysis of viral metagenomes, Virus Evol. 2 (2) (2016) 207–323.
[68] M. Vilsker, Y. Moosa, S. Nooij, V. Fonseca, Y. Ghysens, K. Dumon, R. Pauwels, L.C. Alcantara, E. VandenEynden, A.M. Vandamme, K. Deforche, T. de Oliveira, Genome Detective: an automated system for virus identification from high-throughput sequencing data, Bioinformatics 4 (2018) 32–103.
[69] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (2014) 59.
[70] A. Bankevich, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (5) (2012) 455–477.
[71] K. Deforche, An alignment method for nucleic acid sequences against annotated genomes, bioRxiv (2017).
[72] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.
[73] T. de Oliveira, K. Deforche, S. Cassol, M. Salminen, D. Paraskevis, C. Seebregts, J. Snoeck, E.J. van Rensburg, A.M. Wensing, D.A. van de Vijver, C.A. Boucher, R. Camacho, A.M. Vandamme, An automated genotyping system for analysis of HIV-1 and other microbial sequences, Bioinformatics 21 (19) (2005) 3797–3800.
[74] Computational pan-genomics: status, promises and challenges, Brief. Bioinform. 19 (1) (2018) 118–135.
[75] K. Padovani de Souza, J.C. Setubal, A.C. Ponce de Leon F. de Carvalho, G. Oliveira, A. Chateau, R. Alves, Machine learning meets genome assembly, Brief. Bioinform. 18 (1) (2018) 533.