CHAPTER 2

Bioinformatics approaches applied in pan-genomics and their challenges

Yan Pantoja, Kenny da Costa Pinheiro, Fabricio Araujo, Artur Luiz da Costa Silva, Rommel Ramos
Institute of Biological Sciences, Federal University of Pará (UFPA), Belem, Brazil

1 Introduction

Since the advent of next-generation sequencing (NGS), it became possible to evaluate an increasing number of genomes and, consequently, genetically related organisms [1]. Currently, it is known that there are a great number of genomic variations within a particular bacterial population or species. Thus, the functional annotation of such variants is now possible as well as the analysis of different strains that constitute a particular bacterial species. And the trend is that this scenario will be even bigger and more complex in the future [2].

As the number of genomes available in biological databases increased due to NGS technologies, it became necessary to rethink the idea of a "reference" genome that represents a particular species and aids in research [3]. This reference genome can be shaped in many forms, including:

- the genome of a single individual selected;
- a consensus from an entire population;
- a "functional" genome (without disabling mutations of any gene); and
- a maximum genome that captures every sequence of a given species already detected.

Depending on the context, each one of these options might be best suited for a particular research approach. However, many initial reference sequences did not contain any of the previously mentioned characteristics [3].

In this context, in order to take the most advantage of the data produced by NGS platforms, using a reference, it was necessary to do a paradigm shift: instead of focusing only on a single reference genome, use a "pan-genome," that is, a representation of the entire gene repertoire of a particular species or phylogenetic clade [3].

A decade after the beginning of the genomic era, identifying the number of genomes that could describe a bacterial species became the target of the major questions. Understanding the genomic versatility has become particularly relevant for the study of disease-causing bacteria, which frequently have a large number of variable genes [4].

Pan-genomics: Applications, Challenges, and Future Prospects
https://doi.org/10.1016/B978-0-12-817076-2.00002-0

However, species classification was never simple. Since the first use of the term in a biological context by the English naturalist John Ray in the 17th century [5], the definition of species has been repeated several times, based on different criteria; from shared physical characteristics or ability to produce viable descendants until a shared pattern, niche, or evolutionary history. But regardless of the used definition, the frontier between one taxonomic group and the next is not always clear. While a reproductive definition effectively organizes most multicellular animals into distinct taxonomic groups, the bacteriologists community has not yet been able to establish a uniformly accepted definition for bacterial species due to the fact that these microorganisms possess high levels of genomic diversity and because of their complexity in terms of cultivability, in addition to the high level of horizontal transfer observed [6].

Facing such complexity, some researchers are developing a more subtle view. In prokaryotes, where the lines between taxonomic units are more diffuse, pan-genome analysis (which divides the genome into core and variable genes depending on their presence or absence among species) could offer a more effective way to distinguish closely related organisms when compared to the traditional alternative approaches. While most current methods compare the sequences of only one or a few genes (such as the 16S rRNA gene, or housekeeping genes in the case of multilocus sequence typing) to determine relationships between organisms, pan-genome analysis compare and contrast whole genomes of several individuals, providing an expanded view of similarities and differences between organisms [7, 8].

2 Pan-genome analysis

The pan-genome analysis in the last decade has allowed researchers to develop universal vaccines that could be effective against all strains of one species, or even against several related species. In 2005, the work of Tettelin and colleagues on *Streptococcus agalactiae* (or group B Streptococcus [GBS]) led to the creation of a potentially universal vaccine based on the combination of four bacterial surface proteins [4]. And in June of 2016, researchers at the University of California, San Diego, published a study on methicilin-resistant hospital superbug *Staphylococcus aureus* (MRSA). This study started with 64 strains as a starting point for the development of a vaccine that is widely effective against MRSA [9].

Now that pan-genome approach is widely accepted as a useful way of organizing bacterial diversity, efforts are concentrated on incorporating such studies into phylogenetics, taxonomy, and even into metagenomics, in a more recent metapangenome area [10].

As pan-genome research in microbiology continues to increase, observed intraspecific variation also influences the genomic descriptions of other taxons. As an example, it can be mentioned the eukaryotic species, where the horizontal transfer is even more complex when compared to the prokaryotes. It is also noted that the sequencing of multiple individuals of the same species begins to reveal an extensive genomic diversity that

goes far beyond the small differences observed between genes. Besides, the horizontal transfer events can also occur between prokaryotes and eukaryotes, increasing the diversity of these taxons [11, 12].

Researchers from San Marcos, California State University, realized the importance of such genomic variation a few years ago, shortly after the assembly of a reference genome for the eukaryotic phytoplankton *Emiliania huxleyi*. This species can be found in several ocean sites all over the world. Suspecting that the organism's ability to adapt to varied conditions may depend on single-nucleotide polymorphisms (SNPs) within genes, scientists started to work on the sequencing of more isolate organisms [13].

After sequencing 13 distinct strains, researchers were surprised to find that the size of the genome, originally estimated at about 30,000 genes, varied widely among the analyzed strains, with some strains losing more than 2000 genes. When they performed a pan-genome analysis, the researchers found that only two-thirds of the genes they had identified initially were shared by all sequenced isolates. In particular, there was a high degree of variability in genes encoding metal-binding proteins—key components in the adaptation of *E. huxleyi* to the environment [13].

Given the lack of evidence for horizontal gene transfer in *E. huxleyi*, it is unlikely that the availability of the total genetic pool for each individual is similar to that of prokaryotes. But it is believed that the bigger pan-genome in relation to the central genome of an individual supports the adaptability of this unicellular eukaryote [13].

Emiliania huxleyi hardly is the only one to have this diversity in its DNA. Large-scale sequencing projects were applied to thousands of whole genomes of model eukaryotic organisms, such as *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. These also revealed significant numbers of duplicate new genes. And in cultivated plants, whose genomes often contain large duplicate regions, some studies already support the correlation between the presence or the absence of "variable" genes, disease resistance, metabolite production, and stress responses, showing that the genetic difference has a great impact [13].

2.1 Pan-genome approaches

Computational methods to find more efficient data structures, algorithms, and statistical methods to perform bioinformatic analyses of pan-genomes give rise to a new area known as "computational pan-genomics." This field has desirable characteristics [3]:

- *Completeness*: The presence of all functional elements.
- *Stability*: To present unique identifiable characteristics that can be studied.
- *Comprehensibility*: Understanding the complexity of the genome structure from many species.
- *Efficiency*: Organization of data in a way that accelerates downstream analysis.

The main objective of pan-genome analysis is to determine the genomic diversity of the available dataset, and to predict, via extrapolation, how many genomic sequences would

be necessary to characterize the whole pan-genome or repertoire of genes [14]. Most of the pan-genome projects that emerged after 2005 had as their main differences: the number of genomes/strains analyzed, the phylogenetic resolution, the mathematical prediction model used, the threshold of orthology definition, the algorithm used for alignment and search beside the parameters of percentage of alignment, and completeness of the product [8].

The approach to estimate the pan-genome size, the core genome, and the novel gene discovery rate was started by Tettelin and colleagues; intuitively, starting from a small pan-genome model (i.e., two genomes) and adding more genomes to it, a large number of new genes will be found, since the repertoire of the starting genes were small; conversely, the size of the central genome will decrease, since genes will be less likely to be shared by all genomes. The higher the number of genomes added, the greater the pan-genome and the lower the number of new genes that will be revealed. In parallel, the size of the core genome will decrease. It is possible that a point of "saturation" will be reached, in the sense that the addition of new genomes will not increase the size of the core genome, while the ratio of new genes will be asymptotically stabilized at a given value. For a closed pan-genome, this value is higher than 1 and the pan-genome size can be estimated; for an open pan-genome, this value is lower than 1, and the size of the pan-genome cannot be estimated (i.e., it will probably grow "indefinitely"). Since the number of shared genes and the number of specific genes for a pan-genome depends on how many strains are taken into account, the approach used by Tettelin and colleagues was to use eight genomes of pathogenic strains of *S. agalactiae* and to compute all possible comparisons among n genomes (i.e., eight possible combinations for pan-genome of $n = 2$ genomes) [15].

Plotting the number of shared genes and the number of new genes for each comparison as a function of the n strains considered, Tettelin and colleagues were able to fit exponential decaying function curves over the data which asymptotically reached the values of 1806 shared genes and 33 novel genes, corresponding to the estimate of core-genome size and novel gene discovery rate. The latter value was used for extrapolating the *S. agalactiae* pan-genome size [15].

Users interested in pan-genome analysis have the option of implementing methods such as alignment of multiple nucleotide sequences (complete genomes) to improve sensitivity, for comparisons of high resolution in the species/subspecies or at strain level. They may also use amino acid similarity, protein grouping, structural alignment, and metabolic pathway information at higher levels to reduce noise and eliminate artifacts resulting from nucleotide sequence alignment [8].

The original implementation of the algorithm or workflow pipeline for pan-genome analysis, while conceptually intuitive, has several potential technical pitfalls, some of which are essential enough to impact the conclusions drawn. Issues include the prediction of an open versus closed pan-genome, a rapid or slow pan-genome growing (the rate

at which new genes identified from additional genomes expands the pan-genome), genes that are assigned to the core genome versus accessory genome (the choice of parameters affects whether genes are considered shared/core or noncore), and determining the size of the core genome (the asymptote for the extrapolation of the core genome tends to decrease as more genomes are added to the analysis) [8].

In addition, there is the combinatorial aspect of this approach, where all possible permutations when adding a genome to a set of previously analyzed genomes is considered. The number of comparisons (n) used to calculate the number of new genes, genes belonging to the core, and genes shared in the nth genome can be modeled with the following function, where C is the total number of combinations and N is the total number of genomes in the analysis [8]:

$$C = \frac{N!}{(n-1)!*(N-n)!} \tag{1}$$

These combinations can be represented in the form of a boxplot that can be drawn for both pan- and core-genomes. The combinations from 1 to the total number of samples are placed in the x-axis of the graph, being that in combination 1, the number of genes found in each individual genome is determined. In the combination 2, all possible combinations of 2×2 genomes are observed. In the combination 3, all possible combinations of 3×3 genomes are observed and so on, until reaching the maximum combination that corresponds to the set of all samples [8] (Figs. 1 and 2).

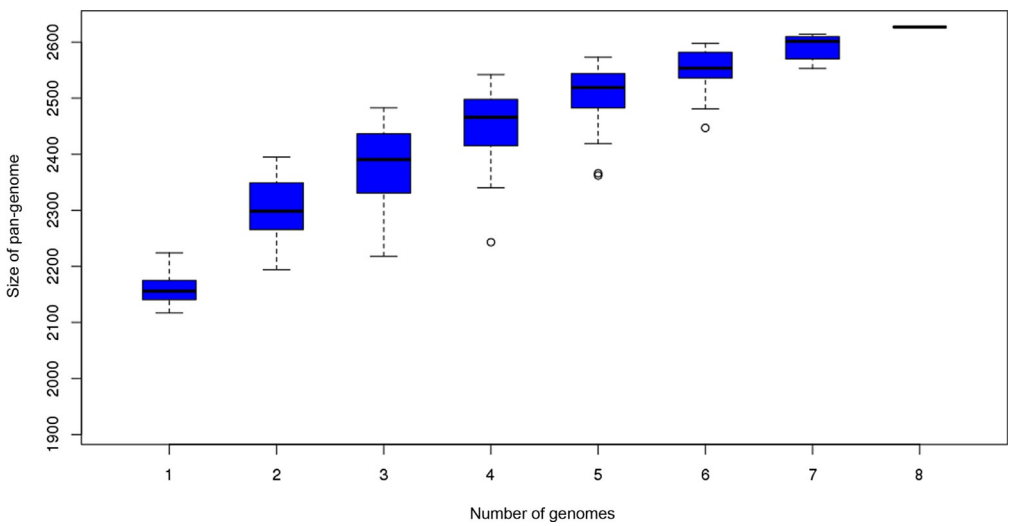

Fig. 1 Pan-genome being displayed graphically. Combinations 1–8 are presented as boxplot (*blue*). It is possible to note that as the number of samples inserted in the combinations increases, the pan-genome also increases.

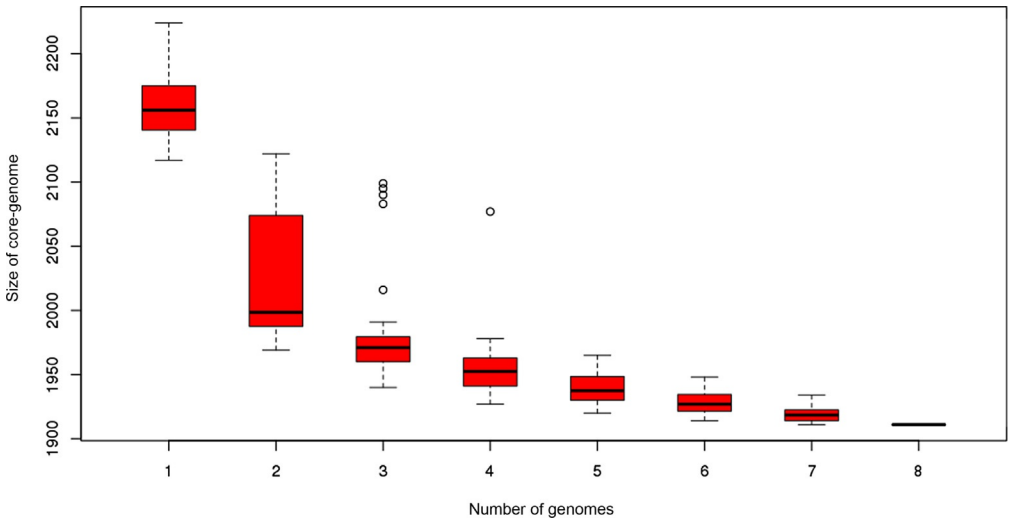

Fig. 2 Core genome being displayed graphically. Combinations 1–8 are shown as box distributions (*red*). It is possible to note that as the number of samples inserted in the combinations increases, the core genome decreases.

2.2 Mathematical model: Heaps' law

It is common to adjust the regression curves of box charts using a power law model (Heaps' law) rather than an exponential decay. Heaps' law is an empirical law that describes the number of distinct words in a document (or set of documents) as a function of document length, and is represented by the formula [15]:

$$n = k^* N^{-\alpha} \tag{2}$$

where n is the expected number of genes for a given set of genomes and N is the number of genomes in a given analysis. K and α are the free coefficients of the regression.

Heaps' law is used in pan-genome analysis to determine whether a given pan-genome is open or closed. This is done after adjusting the regression curve where it is possible to get the values of the alpha coefficient (α). This way it can be inferred that a certain pan-genome is open when the value of α is less than 1. On the other hand, we have that a pan-genome is considered closed when the observed value of α is greater than 1 [15] (Fig. 3).

To obtain the complete gene repertoire of a given microbial species, it is necessary to identify how many extra genes can be added to each new genome sequenced. If each new genome sequenced increases the amount of new genes inserted considerably, we say this pan-genome is open. Generally, open pan-genomes can be observed in species that undergo frequent horizontal gene transfer and colonize multiple environments. In contrast, microorganisms that are more conserved and that live in more isolated niches and consequently have a low capacity to acquire new genes have greater tendency to have a

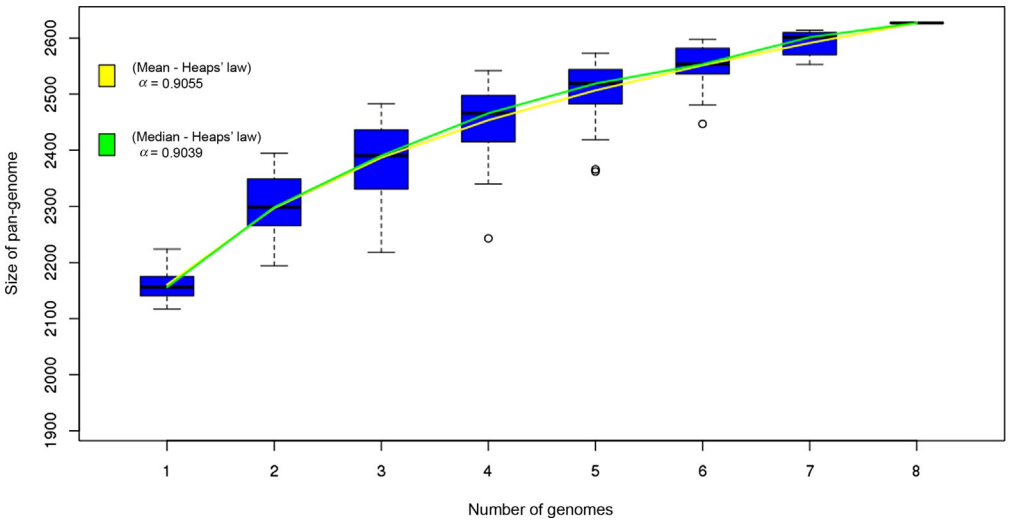

Fig. 3 Pan-genome being plotted along with the regression curves. The curves are adjusted for both the median (*green*) and the mean (*yellow*) values of each distribution. It can be observed in the figure that the values of α (alpha) are close to 0.9 considering a pan-genome near to being closed.

closed pan-genome [7]. It is important to note that a closed pan-genome is not always synonymous with the same phenotype for all the bacterial strains analyzed, because different SNPs can confer different characteristics to different strains [4].

2.3 Software packages and tools

Existing software packages and tools responsible for performing pan-genome analysis have some common functions, such as the search and identification of orthologous and paralogous genes, calculation of the pan-genome profile, and definition of the core genome, accessory genome, and strain-specific genes [7].

2.3.1 Composition and annotation

In order to evaluate the composition and later annotation, the search for orthologs is performed in order to estimate the composition of the pan-genome (core genes, accessory genes, and unique genes). This search is made with tools and algorithms most often used in bioinformatics such as BLAST [16] or OrthoMCL [17]. OrthoMCL uses the Markov clustering algorithm, a method based on a graph flow theory that determines the transition probabilities among the nodes in the graphs, eventually producing clusters of nodes representing groups of orthologous proteins between two or more species [17]. In the later steps, to characterize the sequences found (annotation), tools such as COG (Cluster of Orthologous Groups), InterPro, and KEGG (Kyoto Encyclopedia of Genes and Genomes) are used to obtain data on how the function of the genes is distributed

within the core and accessory genome as well as assessing the metabolic pathways found [7].

Another important factor is the study of the regulation of protein expression and related transcription factors, since the identification of these elements in one or more isolates may help to explain some characteristics that distinguish the different strains. A very useful online tool for this purpose is P2RP (Predicted Prokaryotic Regulatory Proteins), which was developed to make this type of search feasible for all researchers and not only for bioinformaticians, since it has a user friendly interface and is simple, fast, and effective [18].

In addition to the regulatory elements, another important factor is the definition of homology relations between genes belonging to different genomes. Basically, there are two types of situations: when genes descend from an event of speciation (orthologs) and when the genes come from a duplication event (paralogs) from a common ancestor. To find these two groups, it is often used alignment and sequence comparison tools. Homologous genes are conceptualized as corresponding genes in different species. The approach used to find such sequences (genes or proteins) is based on similarity and on the assumption that they are more similar to each other than in any other genome sequence, or they are bidirectional best hits (BBHs). Thus, it is common to assume that BBHs are composed of orthologs that serve to identify families of genes. However, this approach does not take into account the duplication events that may have occurred after a speciation event, since it captures only one-to-one orthological relationships. To overcome this problem, other approaches can be used as COGs proteins and InParanoid/MultiParanoid, which are, respectively, used to call orthologs in pairwise comparison and multiple genome comparison [18a].

InParanoid [19] was initially designed to find orthologous sequences in pairwise genome analysis. Subsequently, the algorithm called MultiParanoid [20] was created to complement and extend the InParanoid approach by taking as input the pairwise orthologous clusters and thus producing clusters of orthologous genes. The comparison of the results obtained using these two different methods showed that there are only small differences in performance between them (Fondi, 2015).

There are several bioinformatics tools capable of predicting microbial genes from genomic sequences. Among them, we can cite GeneMarkHMM [21], Glimmer [22], or Prodigal [23], which depend on statistical methods of learning such as the hidden Markov model to accomplish this task. Tools that use unsupervised learning (Prodigal) are simpler to use since they do not require a trained data set and are able to infer algorithm parameters from the provided genomic sequence.

In global alignment, MAUVE can be used [24], or it can be possible to try a multiple alignment [25] to perform the phylogeny. The MEGA [26] or MAFFT [27] tools are recommended for the reconstruction of trees in the study of phylogeny, and the algorithms most used for this purpose are: neighbor joining and maximum parsimony.

The search for SNPs in the core genome can be used to estimate the age of the species of interest. However, it is necessary that the genomes of the analyzed species are very close in order to study in detail the mutational events that led to the separation in two distinct species. As an example we can mention the work that was carried out in *Yersinia pestis*, in which a comparative analysis was performed with *Yersinia pseudotuberculosis* and *Yersinia enterocolitica* [7].

2.3.2 Pan-genome tools

In an effort to compute standardized pan-genome analysis, several online tools and software suites have been developed. Among the early-developed packages, Panseq [28] and PanCGHweb [29] were published in 2010, followed by Prokaryotic-genome Analysis Tool (PGAT) [30] in 2011. Panseq is a software suite that supports core/dispensable gene mapping and classification of a collection of genome sequences. This tool defines the core and accessory genome based on the sequence identity and segmentation length and not on the predicted proteins. For this purpose, the Novel Region Finder (NRF) module was developed. The module first splits the genome sequence into fragments with predefined sizes, and then the MUMmer alignment program [31] identifies the sequences and contiguous regions that are present or absent in the database [28].

Subsequently, a second module called Core and Accessory Genome Finder (CAGF) is executed and through it a comparison of a single sequence file is performed against all other sequences. The sequence will be added to the pan-genome if it fits in with the predefined parameters, and then, the newly added to fragment sequence is used for subsequent comparisons, and the looping continues until all of the fragment sequences have been tested [28]. PanCGHweb is a web tool for pan-genome microarray analysis based on PanCGH algorithm [32]. It enables users to group genes into orthologs and to construct gene-based phylogenies of related strains and isolates. However, this tool is rather specific to analyze microarray data but not RNA-seq data. The package PGAT integrates several functions, such as identifying SNPs among orthologs and syntenic regions, plotting the presence and the absence of genes among members of a pan-genome, comparing gene orders among different strains and isolates, providing KEGG pathway analysis tools, and searching for genes through different annotations such as the COGs of proteins, PSORT, SignalP, the transmembrane hidden Markov model, and Pfam. However, PGAT is just a database with a limited number of species curated and it cannot perform analysis for new sequencing data from users [33].

GET_HOMOLOGUES [34] is a customizable and detailed pan-genome analysis platform for microorganisms addressed to nonbioinformaticians that was written in Perl and R and can be installed on personal machines. The program starts using BLAST [16] and HMMER [35] to build clusters of orthologous groups. Then, the sequences, features, and intergenes are extracted, sorted, and indexed. Next, the genomes are classified by size being the smallest used as a reference, and then the paralogous genes that arose by

duplication after the speciation process are identified, this whole process is performed through the bidirectional best hit (BBH) algorithm. Subsequently, new genomes are added and compared with the reference genome, and their BBHs are annotated; in the last step, clusters that comprise at least one sequence per genome are conserved [34]. Concomitantly, the results are submitted to OrthoMCL [36] and COGtriangles [37]. Another software that performs pan-genome analysis is called PanGP [38] that implements two sampling algorithms totally random and distance guide on combinations of N strains and generates pan-genome, core genome, and new gene graphs similar to Tettelin and colleagues [4]. The basic difference between the totally random and distance guide algorithms consists of estimating the sample size, where the totally random algorithm repeats randomly the samples in nonredundant combinations for all possible combinations, and the distance guide algorithm has a variable amplification coefficient, which controls the sample size for evaluating the genome diversity of all of the combinations. Tests performed by the authors showed that the distance guide algorithm has better efficiency [38].

PanOCT [39] and PGAP [40] perform scalable pan-genome analyses and require an all-against-all comparison using BLAST, with the running time growing approximately quadratically with the size of input data and are computationally infeasible with large datasets. They also have quadratic memory requirements, quickly exceeding the RAM available in high-performance servers for large datasets. PanOCT is a graph-based ortholog clustering tool for pan-genome analysis of closely related prokaryotic genomes exploiting conserved gene neighborhood information to separate recently diverged paralogs into distinct clusters of orthologs [39]. PGAP executes five analysis modules: cluster analysis of functional genes (the core module), pan-genome profile analysis, genetic variation analysis of functional genes, species evolution analysis, and function enrichment analysis of gene clusters. 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 algorithms. All of the protein sequences are brought together, and protein BLAST is performed; the results are filtered and clustered using the MCL algorithm [16, 41]. The MP method is based on two algorithms: (i) Inparanoid to search orthologous and parologous 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 [20, 42]. Large-scale BLAST score ratio (LS-BSR) introduces a preclustering step that makes it an order of magnitude faster than PGAP; however, it is less sensitive [43].

The software Roary [44] and BPGA [45] were created to address the computational issues related to performance and execution time. Roary performs a rapid clustering of highly similar sequences, which can reduce the running time of BLAST [16] substantially, and carefully manage RAM usage so that it increases linearly, both of which make

it possible to analyze datasets with thousands of samples using commonly available computing hardware without compromising on the accuracy of results [44]. The Bacterial Pan Genome Analysis tool (BPGA) is written in perl programming language but complied in executable files for both Windows and Linux so that no module installation is required. The tool is an ultrafast computational pipeline with seven functional modules for comprehensive pan-genome studies and downstream analyses, these include (i) pan-genome profile analysis, (ii) pan-genome sequence extraction, (iii) exclusive gene family analysis, (iv) atypical GC content analysis, (v) pan-genome functional analysis, (vi) species phylogenetic analysis, and (vii) subset analysis. Other notable features include user friendly command-line interface and high-quality graphics outputs [45].

In the work of Page et al., an accuracy study was performed between four similar stand-alone pan-genome applications. They accurately analyzed the clustering quality of the programs by performing simulated data analysis based on *Salmonella enterica* serovar Typhi (*S. typhi*) CT18 (accession no. AL513382) and they used a single processor (AMD Opteron 6272) and provided 60 GB of RAM. For the study, 12 genomes with 994 identical central genes and 23 accessory genes in various combinations were created and they concluded that all the applications created clusters that are within 1% of the expected results and that the overlap of clusters is almost identical among all applications, except LS-BSR, as shown in Table 1 [44].

The tools and software packages shown so far are the main and best-known available in the scientific community. Although these tools perform different approaches in their pan-genome analysis process, most have common features and functions. Table 2 shows, briefly, each step performed by the cited tools [33, 45].

It is known that in a pan-genome analysis the greater the amount of genomes taken to the analysis the greater will be the computational costs, that is, the discovery of a pan-genome content is an NP-hard problem because comparisons between all sets of genes are necessary to solve the task [46]. The task of recognizing homologous genes becomes even more difficult in the presence of phylogenetically distant genomes, due to the variability introduced in duplication and gene transmission. This research field has the challenge of designing similarity measures that are fast and adaptive, in order to find an adequate homology pan-genome structure [46]. Therefore, in the study of Bonnici

Table 1 Accuracy of each pan-genome application on a dataset of simulated data [44]

	Core genes	Total genes	Incorrect merge
Expected	994	1017	0
PGAP	991	1012	4
PanOCT	993	1015	1
LS-BSR	974	994	23
Roary	994	1017	0

Table 2 Features of each pan-genome application

Name tools	Link	Platform	Main features
BPGA	http://iicb.res.in/bpga/index.html	Windows Linux	a, b, c, d, e, f, g, h
PGAP	https://sourceforge.net/projects/pgap/	Linux	b, c, d, e, f
PGAT	http://nwrce.org/pgat/	Online	b, h
LS-BSR	https://github.com/jasonsahl/LS-BSR	Linux	b
Roary	https://sanger-pathogens.github.io/Roary/	Linux	b, c, d, e
Panseq	https://lfz.corefacility.ca/panseq/	Online Windows Linux	b, e
GET_HOMOLOGUES	http://github.com/eead-csic-compbio/get/_homologues/	MacOS Linux	b, d, e
PanCGHweb	http://bamics2.cmbi.ru.nl/websoftware/pancgh/	Online	b
PanOCT	http://bamics2.cmbi.ru.nl/websoftware/pancgh/	Online	b
PanGP	https://pangp.zhaopage.com/	Windows Linux	c, d

Notes: The main features are represented by letters: (a) Preparation step; (b) clustering; (c) matrix generation (pan-matrix); (d) pan-genome profile analysis; (e) phylogeny construction; (f) function and pathway analysis; (g) pan-genome statistics; and (h) atypical GC content analysis.
Source: (a) From N. Chaudhari, V. Gupta, C. Dutta, BPGA—an ultra-fast pan-genome analysis pipeline, Sci. Rep. 6 (2016) 24373.

et al. [46], a computational tool called PanDelos was developed with the purpose of minimizing these challenges. It is an autonomous dictionary-based tool for the discovery of pan-genome contents among distant genomes phylogenetically.

Pan-genome analysis can be applied in many different application domains. Table 3 summarizes the main fields.

The approaches to pan-genome content discovery need to take into account that duplication and gene transmission may introduce sequence changes [30, 45]. These variations hamper the task of recognizing homologous genes, especially when ancestral genomes are no longer available. The sequences present in the core genome are transferred almost without any change, since the genes present in the core genome are often under strong evolutionary selection. The process is different for the genes present in the accessory genome because these dispensable genes have a number of inconstant and varied variations, and depending on the phylogenetic distance, the similarity between the homologous sequences tends to decrease. Organisms very close phylogenetically, when

Table 3 Description of pan-genome applications [3]

Application	Description
Microbes	Important to understand the functional and evolutionary repertoire of microbial genomes, which opens possibilities for the development of therapies and engineering applications
Metagenomics	In the metagenome, there is the possibility of revealing common adaptations to the environment, as well as the coevolution of the interactions through the pan-genome
Viruses	One of the goals of pan-genomics, both in virology and in medical microbiology, will be to fight infectious disease
Plants	A pan-genome available for a certain crop that includes its wild relatives provides a unique coordinate system to anchor all known phenotype and variation information, and will allow the identification of new genes from the available germplasm that are not present in the genome of reference(s)
Human genetic diseases	Pan-genome data structures are able to handle combinations of genomic variants with comprehensive functional annotations—for example, epigenomic datasets or gene expression
Cancer	A pan-genome of somatic cancer, representing variability in the inferred rate of change throughout the genome, would increase the identification of disease-related genomic changes based on their recurrence among individuals
Phylogenomics	The pan-genome extracts genomic features with an evolutionary signal, such as gene content tables, alignments of shared marker gene sequences, genomic SNPs, or transcribed internal spacer sequences, depending on the level of kinship of the included organisms

they are analyzed, reasonable thresholds are applied in the similarity of the sequences so that recognition of gene families occurs [46].

The Roary and EDGAR tools [47] are based on sequence alignment; however, some alternative strategies can be used to retrieve domain architecture between homologous genes [48] or for the detection of horizontal gene transfer [49], through the exploration of free alignment techniques.

PanDelos uses a different strategy, the tool seeks to discover pan-genome content in phylogenetically distant organisms based on the information theory and network analysis. The use of parameters is not a requirement of the software and the limits are automatically deduced from the context. PanDelos avoids sequence alignment by introducing a measure of similarity based on k-mers multiplicity, rather than the simple presence/absence of mers. Strategy confidence is supported by a nonempirical choice of the most appropriate k-mer length. In addition, when two sequences are identified as homologous, the

selection of the least similarity between them is based on the knowledge from the mapping of the readings that were used in the sequence sequencing and reconstruction processes [46].

To infer thresholds for the discovery of paralogs, the best results from the 1vs1 comparison of the genome that was made previously, aiming at the discovery of orthologous genes, are used. The homology relationships between organisms are incorporated and form part of a global network and the groups of homologous genes used in the analysis are extracted from that network using applications with detection algorithms. According to Bonnici et al. [46], the PanDelos tool overcomes existing tools such as Roary and EDGAR in terms of execution time and accuracy of analysis, both in real applications and in synthetic analysis with simulated data.

2.3.3 Machine learning applied to pan-genome

Machine learning techniques have been widely used in the field of bioinformatics [50]. Techniques such as supervised classification, grouping, and probabilistic graphical models for discovery of knowledge, as well as deterministic and stochastic heuristics for optimization [50]. The rapidly growing data diversity, produced by modern molecular biology and made available in public databases, has stimulated the need for accurate classification and prediction algorithms [51]. With this exponential growth in the amount of biological data, computational problems arise such as the proper storage and management of this astronomical amount of information being generated, as well as problems for extracting useful information from such data. The second problem is one of the main challenges of computational biology [52]. Therefore, there is a need in the development of methods and tools capable of transforming all this heterogeneous data into biological knowledge about the fundamental mechanisms. These tools and methods should allow us to provide knowledge in the form of testable models and not just describe the content present in those data. By means of this simplifying abstraction that constitutes a model, we can obtain predictions from the system [52].

Machine learning techniques basically consist of developing algorithms for computers to optimize one performance criteria using example data or past experience. The optimized criteria can be the precision provided by a predictive model—in a modeling problem—and the value of a function of adequacy or evaluation—in an optimization problem [52].

The techniques and computational methods of machine learning are applied in several biological fields such as genomics, proteomics, microarrays, systems biology, evolution, text mining [52], and even pan-genome analysis because researchers face challenges such as processing and maintaining large datasets, while providing accurate and efficient analysis approaches. Genomics is one of the most important fields of bioinformatics, mainly because of the exponential increase in the number of available sequences that need to be processed. The initial step is to obtain and extract the location and structure of the genes,

either by prediction or genomic annotation, from genome sequences [50]. In addition, it is possible to further identify regulatory elements and RNA noncoding genes present in intergenic regions.

In the field of proteomics, the main application of computational methods is the prediction of protein structure. Proteins are very complex macromolecules and therefore, the number of possible structures is enormous. This makes the prediction of protein structure a very complicated combinatorial problem, where optimization techniques are required [52].

The management of the large amount of complex experimental data is another application in which computational methods of machine learning can be used [52]. The microarray assays are one of the best known, but not the only, fields where this type of data is collected. Complex experimental data raise two different problems: first, the data need to go through a preprocessing step, that is, they need to be formatted to be used properly by machine learning algorithms. The second problem would be the analysis of the data itself, which will depend on what it is searched. In the case of microarray data, the most typical applications are identification of patterns of expression, classification, and induction of genetic networks [52]. Systems biology is another field in which biology and machine learning work very well together as it is very complex to model the life processes that occur within the cell. Thus, computer learning techniques are extremely useful in the modeling of biological networks, especially genetic networks, signal transduction networks, and metabolic pathways.

Not very different, the analysis of evolution and, especially, the reconstruction of phylogenetic trees is also used of the techniques of machine learning. Phylogenetic trees are schematic representations of organisms' evolution [52]. Generally, they were constructed according to different characteristics of the organisms (morphological characteristics, metabolic characteristics, etc.) but, nowadays, with the great amount of biological sequences available in public databases, phylogenetic tree-building algorithms are based on comparison between different genomes [50]. This comparison is made through the alignment of multiple sequences, where optimization techniques, used with machine learning algorithms, are very useful.

In the paper by Her et al. [53], a machine learning approach based on pan-genome was developed to predict antimicrobial resistance (AMR) activities in *Escherichia coli* strains. Machine learning approaches were applied in the pan-genome to better define and predict AMR. According to the authors, AMR is becoming a major problem in the developed and developing countries, and the identification of resistant or susceptible strains of certain antibiotics is essential in the fight against antibiotic-resistant pathogens [53].

Antimicrobial-resistant pathogens (AMR) have an ultrarapid mutation rate which renders most of the existing drugs against superbugs unavoidable, and existing classes of antibiotics are probably the best there will ever be [54]. Another study published in 2013 also identified that additional economic costs due to AMR could reach $55 billion

and that trivial bacterial infections, such as hip replacements, for example, could increase the mortality rate from approximately 0% to 30% [55].

Pan-genome was also used in the analysis of diversity, virulence, and AMR phenotypes in the organism *Klebsiella pneumoniae* [56]. In this study, they found that *K. pneumoniae* can be divided into three distinct groups, and that certain branches in all three groups may be hypervirulent or resistant to multiple drugs [56]. In addition, in another study a computational approach, called Scoary, was developed to make an association between the genetic components found in the pan-genome with the observed phenotypic traits and to identify the gene pools that were associated with activities of high level of AMR, such as resistance to linezolid in *Staphylococcus epidermidis* [57]. These examples have suggested that the pan-genome idea can be very useful in defining genetic components that can contribute to the phenotypes of living organisms.

The PATRIC database is known as one of the most comprehensive antibiotic resistance databases that collects genes, proteins, and genomic information related to the resistance or susceptibility of pathogens to various antibiotics [58]. PATRIC has a collection of more than 80,000 bacterial genomes available in its database allowing scientists to understand the mechanisms of AMR in terms of genes, proteins, and genomes.

Thus, it was developed a pan-genome-based approach to characterize strains that are resistant to antibiotics and strains of *E. coli* were used as a model in which 59 strains of *E. coli* from the PATRIC database were selected [58]. By using machine learning techniques through genetic algorithms (GA), it was obtained better predictive performance than the sets of genes established in the literature, suggesting that gene sets selected by GA may justify a more in-depth analysis in investigating more details on how *E. coli* fights against antibiotics.

3 Challenges

The data analyzed in a pan-genome study have characteristics of Big Data such as volume, variety, speed, and veracity. These studies presented great challenges to algorithm and software developers, especially due to the size of the data generated by the new-generation sequencers, the data heterogeneity, and their complex interaction [3].

The International Cancer Genome Consortium has accumulated a dataset of more than two petabytes in just 5 years, resulting in the need to store data in clouds, providing a scalable, dynamic and parallel way of processing data in an inexpensive, flexible, reliable, and safe manner. Currently, there are large providers with a complex computing infrastructure and large public repositories (e.g., National Center for Biotechnology Information, European Bioinformatics Institute, and DNA Data Bank of Japan) that assist both researchers who choose to download/upload data for analysis, but also provide a secure and reliable storage environment for this large set of information. Distributed and parallel

computing has also been used as a resource to deal with the considerable volume of data stored in public databases [3].

Pan-genome has also introduced new challenges for data visualization. As the relationships between several genomes can be highly complex and the homology relations can vary widely with each dataset studied, it became necessary to obtain new ways of visualizing these relations in their total complexity without loss of information. In general, mathematical approaches to comparing sets are used to evaluate homology relationships such as Venn and Flower Plots diagrams [3].

New data visualization packages for pan-genome are developed to facilitate the research and generate a better visualization of the relations of homology existing in the genomes. As an example we can mention the UpSetR package that has provided users with an improved alternative to the Venn chart; while a normal Venn graph accepts up to five data sets at most (five genomes), the visualization offered by UpSetR does not have limit to the data set analyzed [59].

3.1 Pan-genome analysis with draft genomes

Pan-genome analysis are usually performed using complete genomes to analyze the complete gene repertoire. However, depositing a complete genome of an organism in a public database is not an easy task, the finalization of this process is directly linked to a number of variables, and therefore, the number of drafts deposited genomes increases exponentially, thus increasing the number of projects that use this type of genome in pan-genome analysis. According to the Genomes OnLine Database [60], the number of complete and draft genomes deposited in public databases in 2017 reached 4311 and 31,332, respectively. Bacteria have a greater number of reports of genomes being deposited, due to their compact nature, being relatively less complex in the sequencing process, and due to the importance of their application in various fields, such as biotechnology, agriculture, medicine, etc. [61].

Working with draft genomes in any type of analysis, even in pan-genome analysis, brings a series of challenges and requires greater attention precisely because it is not yet finalized, that is, the genomic repertoire of this genome is not yet totally represented. In addition, draft genomes may contain a number of errors, such as broken products or frameshifts. Several factors may explain the reason why a given genome was not yet fully finalized, such as errors in sequencing, assembly, or even genomic annotation errors. In this case, there may be a lot that has not yet been represented, such as important products and functions for the bacteria, which may imply errors in the final result of a given analysis, such as pan-genome. Therefore, an important step before using a draft genome in any type of analysis is to seek to represent its gene repertoire as much as possible. In the study by Veras et al. [62], for example, a computational tool was developed in JAVA programming language, called Pan4draft, especially to work with drafts genomes in

pan-genome analysis. Pan4draft uses the PGAP software pipeline to perform the pan-genome analysis, but performs a series of previous steps, automatically integrating several tools, responsible for seeking a better representation of the gene repertoire of these genomes drafts, thus increasing the accuracy of the pan-genome analysis [62].

3.2 Perspectives for pan-genome applied to the human genome

The human genome project was founded in 1990, and after numerous surveys carried out in several centers, it is now known that *Homo sapiens* cannot be described only by a single reference sequence. Although the variation occurring in the human genome is inferior in comparison to microbes and plants, the first attempt to construct a human pan-genome in 2009 (based on the human reference genome and other two genomes) estimated that up to 40 megabases of sequence including the coding regions of proteins, were absent from the reference genome [63]. Still in 2009, researchers estimated that gene counts ranged from 73 to 87 genes found in two randomly selected individuals [64].

Such observed differences are increasingly associated with genetic disorders such as autism, Parkinson's disease, and Alzheimer's, causing research to turn even further to the study of these variations observed in our species [65, 66]. Researchers at the Case Western Reserve University have identified that more than 300 small sequences absent from the reference genome were present in at least 1% of the human population, leading to a reconsideration of the whole concept of the reference genome used not only for prokaryotes but also for eukaryotes [67].

In this way, it is possible to evaluate that we can still improve in many aspects the approaches and methodologies used in pan-genomic studies. The main objective in overcoming such challenges is to find a more complete scenario that presents all the desired characteristics when analyzing certain species of both prokaryotes and eukaryotes.

4 Conclusion and future direction

With the development of sequencing technologies, thousands of biological data have become accessible in the past years. In this context, in order to take the most advantage of the data produced by NGS platforms, using a reference, it was necessary to do a paradigm shift: instead of focusing only on a single reference genome, use a pan-genome, that is, a representation of the entire gene repertoire of a particular species or phylogenetic clade. Thus, life sciences have entered the era of pan-genomics, which is known to represent "all" major genetic variation of a collection of genomes of interest. The search for sequence similarity is the important step in the pan-genome analysis and in comparative genomics in general.

Nowadays, the process of similarity search and pan-genome visualization are two of the wide variety of particular computational challenges that need to be considered. For

this, novel different computational methods and paradigms are needed over the years, making the computational pan-genomics a subarea of research in rapid extension.

A current pan-genome analysis can be considered a "one-dimensional" approach by mainly working with genomes only as sequences and thus concentrating on storing and analyzing sequences and relations between certain parts of subsequences, such as variant alleles and their interconnections, genes, and/or transcriptomes.

However, new technologies that are emerging in rapid development allow to infer the pan-genome with three-dimensional conformation, that is, in the medium term, one can expect to be able to raise the pan-genome in up to three dimensions. This will mean that future three-dimensional pan-genomes will not only represent all sequence variation of the species or genus, but also will encode their spatial organization, as well as their mutual relationships in this regard.

References

[1] B. Hall, G. Ehrlich, F. Hu, Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing, Microbiology 156 (2010) 1060–1068.
[2] M. Pallen, B. Wren, Bacterial pathogenomics, Nature 449 (2007) 835.
[3] Computational Pan-Genomics Consortium, Computational pan-genomics: status, promises and challenges, Brief. Bioinf. 19 (2016) 118–135.
[4] H. Tettelin, V. Masignani, M. Cieslewicz, C. Donati, D. Medini, N. Ward, S. Angiuoli, J. Crabtree, A. Jones, A. Durkin, Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome", Proc. Natl. Acad. Sci. USA 102 (2005) 13950–13955.
[5] I. Stevenson, John Ray and his contributions to plant and animal classification, J. Hist. Med. Allied Sci. 2 (1947) 250–261.
[6] L. Olendzenski, J. Gogarten, M. Gogarten, J. Gogarten, L. Olendzenski, Horizontal Gene Transfer: Genomes in Flux, Humana Press, Totowa, NJ, 2009.
[7] L. Rouli, V. Merhej, P. Fournier, D. Raoult, The bacterial pangenome as a new tool for analysing pathogenic bacteria, New Microbes New Infect. 7 (2015) 72–85.
[8] G. Vernikos, D. Medini, D. Riley, H. Tettelin, Ten years of pan-genome analyses, Curr. Opin. Microbiol. 23 (2015) 148–154.
[9] E. Bosi, J. Monk, R. Aziz, M. Fondi, V. Nizet, B. Palsson, Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity, Proc. Natl. Acad. Sci. USA 113 (2016) E3801–E3809.
[10] T. Delmont, A. Eren, Linking pangenomes and metagenomes: the *Prochlorococcus metapangenome*, PeerJ 6 (2018) e4320.
[11] K. Sieber, R. Bromley, J. Hotopp, Lateral gene transfer between prokaryotes and eukaryotes, Exp. Cell Res. 358 (2017) 421–426.
[12] J. Huang, Horizontal gene transfer in eukaryotes: the weak-link model, Bioessays 35 (2013) 868–875.
[13] B. Read, J. Kegel, M. Klute, A. Kuo, S. Lefebvre, F. Maumus, C. Mayer, J. Miller, A. Monier, A. Salamov, Pan genome of the phytoplankton *Emiliania underpins* its global distribution, Nature 499 (2013) 209.
[14] P. Lapierre, J. Gogarten, Estimating the size of the bacterial pan-genome, Trends Genet. 25 (2009) 107–110.
[15] H. Tettelin, D. Riley, C. Cattuto, D. Medini, Comparative genomics: the bacterial pan-genome, Curr. Opin. Microbiol. 11 (2008) 472–477.
[16] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[17] F. Chen, A. Mackey, C. Stoeckertjr, D. Roos, OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups, Nucleic Acids Res. 34 (2006) D363–D368.
[18] M. Barakat, P. Ortet, D. Whitworth, P2RP: a web-based framework for the identification and analysis of regulatory proteins in prokaryotic genomes, BMC Genomics 14 (2013) 269.
[18a] F. Del Chierico, M. Ancora, M. Marcacci, C. Cammà, L. Putignani, S. Conti, Bacterial pangenomics [Internet], in: A. Mengoni, M. Galardini, M. Fondi (Eds.), Methods in Molecular Biology, Springer, New York, NY, 2015, pp. 31–47. Available from: http://link.springer.com/10.1007/978-1-4939-1720-4.
[19] K. O'brien, M. Remm, E. Sonnhammer, Inparanoid: a comprehensive database of eukaryotic orthologs, Nucleic Acids Res. 33 (2005) D476–D480.
[20] A. Alexeyenko, I. Tamas, G. Liu, E. Sonnhammer, Automatic clustering of orthologs and inparalogs shared by multiple proteomes, Bioinformatics 22 (2006) e9–e15.
[21] J. Besemer, M. Borodovsky, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses, Nucleic Acids Res. 33 (2005) W451–W454.
[22] A. Delcher, K. Bratke, E. Powers, S. Salzberg, Identifying bacterial genes and endosymbiont DNA with Glimmer, Bioinformatics 23 (2007) 673–679.
[23] D. Hyatt, G.L. Chen, P.F. Locascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinf. 11 (2010) 119.
[24] A. Darling, B. Mau, N. Perna, progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement, PLoS ONE 5 (2010) e11147.
[25] A. Jacobsen, R. Hendriksen, F. Aaresturp, D. Ussery, C. Friis, The *Salmonella enterica* pan-genome, Microb. Ecol. 62 (2011) 487.
[26] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, Mol. Biol. Evol. 28 (2011) 2731–2739.
[27] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.
[28] C. Laing, C. Buchanan, E. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. Thomas, V. Gannon, Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions, BMC Bioinf. 11 (2010) 461.
[29] J. Bayjanov, R. Siezen, S. Vanhijum, PanCGHweb: a web tool for genotype calling in pangenome CGH data, Bioinformatics 26 (2010) 1256–1257.
[30] M. Brittnacher, C. Fong, H. Hayden, M. Jacobs, M. Radey, L. Rohmer, PGAT: a multistrain analysis resource for microbial genomes, Bioinformatics 27 (2011) 2429–2430.
[31] S. Kurtz, A. Phillippy, A. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. Salzberg, Versatile and open software for comparing large genomes, Genome Biol. 5 (2004) R12.
[32] J. Bayjanov, M. Wels, M. Starrenburg, J. Vanhylckamavlieg, R. Siezen, D. Molenaar, PanCGH: a genotype-calling algorithm for pangenome CGH data, Bioinformatics 25 (2009) 309–314.
[33] J. Xiao, Z. Zhang, J. Wu, J. Yu, A brief review of software tools for pangenomics, Genom. Proteom. Bioinform. 13 (2015) 73–76.
[34] B. Contreras-moreira, P. Vinuesa, GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pan-genome analysis, Appl. Environ. Microbiol. 79 (2013) 7696–7701.
[35] R. Finn, J. Clements, S. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37.
[36] L. Li, C. Stoeckert, D. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes, Genome Res. 13 (2003) 2178–2189.
[37] D. Kristensen, L. Kannan, M. Coleman, Y. Wolf, A. Sorokin, E. Koonin, A. Mushegian, A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches, Bioinformatics 26 (2010) 1481–1487.
[38] Y. Zhao, X. Jia, J. Yang, Y. Ling, Z. Zhang, J. Yu, J. Wu, J. Xiao, PanGP: a tool for quickly analyzing bacterial pan-genome profile, Bioinformatics 30 (2014) 1297–1299.
[39] D. Fouts, L. Brinkac, E. Beck, J. Inman, G. Sutton, PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species, Nucleic Acids Res. 40 (2012) e172.

[40] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, J. Yu, PGAP: pan-genomes analysis pipeline, Bioinformatics 28 (2011) 416–418.
[41] A. Enright, S. Vandongen, C. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Res. 30 (2002) 1575–1584.
[42] G. Ostlund, T. Schmitt, K. Forslund, T. Köstler, D. Messina, S. Roopra, O. Frings, E. Sonnhammer, InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, Nucleic Acids Res. 38 (2009) D196–D203.
[43] J. Sahl, J. Caporaso, D. Rasko, P. Keim, The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes, PeerJ 2 (2014) e332.
[44] A. Page, C. Cummins, M. Hunt, V. Wong, S. Reuter, M. Holden, M. Fookes, D. Falush, J. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, Bioinformatics 31 (2015) 3691–3693.
[45] N. Chaudhari, V. Gupta, C. Dutta, BPGA—an ultra-fast pan-genome analysis pipeline, Sci. Rep. 6 (2016) 24373.
[46] V. Bonnici, R. Giugno, V. Manca, PanDelos: a dictionary-based method for pan-genome content discovery, BMC Bioinf. 19 (2018) 437.
[47] J. Blom, J. Kreis, S. Spänig, T. Juhre, C. Bertelli, C. Ernst, A. Goesmann, EDGAR 2.0: an enhanced software platform for comparative gene content analyses, Nucleic Acids Res. 44 (2016) W22–W28.
[48] D. Syamaladevi, A. Joshi, R. Sowdhamini, An alignment-free domain architecture similarity search (ADASS) algorithm for inferring homology between multi-domain proteins, Bioinformation 9 (2013) 491.
[49] G. Bernard, C. Chan, Y. Chan, X. Chua, Y. Cong, J. Hogan, S. Maetschke, M. Ragan, Alignment-free inference of hierarchical and reticulate phylogenomic relationships, Brief. Bioinform. 20 (2019) 426–435.
[50] P. Baldi, S. Brunak, F. Bach, Bioinformatics: The Machine Learning Approach, MIT Press, Cambridge, 2001.
[51] H. Bhaskar, D. Hoyle, S. Singh, Machine learning in bioinformatics: a brief survey and recommendations for practitioners, Comput. Biol. Med. 36 (2006) 1104–1125.
[52] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. Lozano, R. Armañanzas, G. Santafé, A. Pérez, Machine learning in bioinformatics, Brief. Bioinform. 7 (2006) 86–112.
[53] H. Her, Y. Wu, A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains, Bioinformatics 34 (2018) i89–i95.
[54] M. Cormican, A. Vellinga, Existing classes of antibiotics are probably the best we will ever have, Br. Med. J. (Online) 344 (2012).
[55] R. Smith, J. Coast, The true cost of antimicrobial resistance, BMJ 346 (2013) f1493.
[56] K. Holt, H. Wertheim, R. Zadoks, S. Baker, C. Whitehouse, D. Dance, A. Jenney, T. Connor, L. Hsu, J. Severin, Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health, Proc. Natl. Acad. Sci. USA 112 (2015) E3574–E3581.
[57] O. Brynildsrud, J. Bohlin, L. Scheffer, V. Eldholm, Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary, Genome Biol. 17 (2016) 238.
[58] A. Wattam, J. Davis, R. Assaf, S. Boisvert, T. Brettin, C. Bun, N. Conrad, E. Dietrich, T. Disz, J. Gabbard, Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center, Nucleic Acids Res. 45 (2016) D535–D542.
[59] J. Conway, A. Lex, N. Gehlenborg, UpSetR: an R package for the visualization of intersecting sets and their properties, Bioinformatics 33 (2017) 2938–2940.
[60] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, O. Verezemska, M. Isbandi, A. Thomas, R. Ali, K. Sharma, N. Kyrpides, Genomes OnLine Database (GOLD) v. 6: data updates and feature enhancements, Nucleic Acids Res. 45 (2016) D446–D456. D1.
[61] V. Wanchai, P. Patumcharoenpol, I. Nookaew, D. Ussery, dBBQs: dataBase of bacterial quality scores, BMC Bioinf. 18 (2017) 483.
[62] A. Veras, F. Araujo, K. Pinheiro, L. Guimarães, V. Azevedo, S. Soares, A. Dasilva, R. Ramos, Pan4-Draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes, Sci. Rep. 8 (2018) 9670.

[63] R. Li, Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li, Building the sequence map of the human pan-genome, Nat. Biotechnol. 28 (2010) 57.
[64] C. Alkan, J. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. Kitzman, C. Baker, M. Malig, O. Mutlu, Personalized copy number and segmental duplication maps using next-generation sequencing, Nat. Genet. 41 (2009) 1061.
[65] H. Yoo, Genetics of autism spectrum disorder: current status and possible clinical applications, Exp. Neurobiol. 24 (2015) 257–272.
[66] C. Klein, A. Westenberger, Genetics of Parkinson's disease, Cold Spring Harb. Perspect. Med. 2 (2012) a008888.
[67] Y. Liu, M. Koyutürk, S. Maxwell, M. Xiang, M. Veigl, R. Cooper, B. Tayo, L. Li, T. Laframboise, Z. Wang, Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing, BMC Genomics 15 (2014) 685.

Further reading

[68] D. Andersson, B. Levin, The biological cost of antibiotic resistance, Curr. Opin. Microbiol. 2 (1999) 489–493.
[69] J. Bower, H. Bolouri, Computational Modeling of Genetic and Biochemical Networks, MIT Press, Cambridge, 2004.
[70] J. Hogg, F. Hu, B. Janto, R. Boissy, J. Hayes, R. Keefe, J. Post, G. Ehrlich, Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains, Genome Biol. 8 (2007) R103.
[71] G. Kettler, A. Martiny, K. Huang, J. Zucker, M. Coleman, S. Rodrigue, F. Chen, A. Lapidus, S. Ferriera, J. Johnson, Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*, PLoS Genet. 3 (2007) e231.
[72] M. Krallinger, R. Erhardt, A. Valencia, Text-mining approaches in molecular biology and biomedicine, Drug Discov. Today 10 (2005) 439–445.