Bacterial Pan-Genomics

2

Arash Iranzadeh and Nicola Jane Mulder

Abstract

Due to their tendency to have a high recombination rate, bacterial genomes are highly diverse across different strains. This diversity may even be in the form of the presence or absence of entire genes; therefore, each strain might have its own combination of genes. The pan-genome represents the complete gene pool of a species. It is made up of the core genome (genes shared by all strains) and the accessory genome (genes shared by some strains and not all). The pan-genome can be considered to be a comprehensive reference genome for computational biology, and several tools have been developed for pan-genomics applications. The tools enable scientists to explore bacterial genomes with more flexibility considering all types of genetic variations. Pan-genomics has many applications in medicine such as the development of vaccines and drugs against pathogenic bacteria. In this chapter, we discuss the fundamental principles and algorithms for pan-genome analysis and introduce and compare the most recent computational tools.

2.1 Introduction

Despite the fact that not all microbes are harmful, more than 17 million people are killed from infectious diseases caused by microbes each year ('WHO | Press release' 2013). There have been several deadly bacterial pandemics through history such as the *plague*, *cholera*, and *typhus* in which millions of people perished, reshaping the ancient and medieval world populations. Fortunately, the existence of microbes was

A. Iranzadeh (✉) · N. J. Mulder
Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
e-mail: arash.iranzadeh1980@gmail.com

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_2

discovered for the first time by *Robert Hooke* and *Antoni van Leeuwenhoek* in the seventeenth century (Gest 2004). Later, in the nineteenth century, the science of bacteriology was established. *Louis Pasteur* demonstrated the germ theory of disease and the relationship between microbes and diseases (Lanska 2014), and *Robert Koch* discovered the bacterium *Bacillus anthracis* as the cause of *anthrax* (Blevins and Bronze 2010).

Bacteria are one of the most important types of organisms that cause diseases. These single-celled prokaryotes can be found almost everywhere, and their existence on earth dates back to about 3.5 billion years ago (Kara and Robert 2018). Since they can live in extreme environmental conditions on earth, they might exist on other planets or even other galaxies in the universe (Grebennikova et al. 2018).

To understand the outbreak and pathogenesis of bacterial infections, their genomes must be studied and analyzed precisely. The first step is to sequence the genome, which contains information about the origin of the species and its evolution. Two major sequencing techniques have been developed: early sequencing techniques or *first-generation sequencing* developed in the 1970s, which include the *Maxam-Gilbert* and *Sanger* methods (Sanger et al. 1977), and modern sequencing or *next-generation sequencing* (*NGS*) technologies that have been developed in the twenty-first century. *NGS* is also called *deep sequencing*, *high-throughput sequencing*, or *massive sequencing*. Examples include *Illumina (Solexa)*, *Roche 454*, *and SOLiD* (Goodwin et al. 2016). *Sanger* sequencing can cover a long stretch of DNA with higher quality and is sometimes used to sequence small pieces of DNA such as bacterial plasmids or for validation. However, it is more expensive and time-consuming than *NGS* and is inefficient for sequencing entire genomes. For instance, *Sanger sequencing* took over a decade to deliver a draft human genome, while *NGS* takes only a single day to sequence an entire human genome (Behjati and Tarpey 2013). The whole genome of a bacterium can be sequenced by *NGS* technology for a few hundreds of dollars (Mengoni et al. 2015). As a result, today, an enormous amount of genetic sequences are publicly available in databases such as the *National Center for Biotechnology Information's (NCBI) GenBank* (Ostell and McEntyre 2007), the *DNA Databank of Japan (DDBJ)* (Miyazaki et al. 2004), and the *European Nucleotide Archive (ENA)* (Leinonen et al. 2011), among others.

Owing to the availability of large amounts of sequencing data, a bacterial species can be described through an inclusive reference genome called the *pan-genome.* Computational algorithms and tools have been developed for *pan-genome* building and analysis. The pan-genome is a kind of reference genome that represents all genes in a collection of bacterial isolates. The pan-genome is usually defined for bacteria and viruses because they are highly recombinogenic and have a small genome and many isolates can be cloned and sequenced quickly. Nonetheless, the pan-genome approach can also be applied for eukaryotes like plants to investigate similarities and differences in individuals of the same species. In this chapter, the bacterial pan-genome definition, algorithms, and the computational tools available for pan-genomics will be discussed.

2.2 Pan-Genome

For most studies in comparative genomics starting with NGS data, a reference genome is required, and it must be defined for data analysis. This reference genome can be:

1. The genome of one strain.
2. The consensus sequence drawn from all strains.
3. A comprehensive genome that contains all genetic variants.

The remarkable capability of bacteria to adapt to their environment is enabled by their ability to exchange their genetic material by homologous recombination and horizontal gene transfer. It allows bacteria to have a dynamic, adaptable, and diverse genome (Maloy 2013). This genomic plasticity is even considerable across different strains of the same bacterial species. Therefore, a single genome sequence cannot necessarily represent the entire range of genetic variation in bacteria. A pan-genome is actually a type of reference genome that displays all variants, including all possible genes. In 2005 and for the first time, the term, *comparative pan-genomics*, became official when eight strains of *Streptococcus agalactiae* were compared (Tettelin et al. 2005). Since then, the pan-genome has been defined as the following: "For a collection of closely related strains, pan-genome is the entire gene set that exists in those strains."

The pan-genome contains three types of genes according to their availability among strains (Fig. 2.1):

1. *Core genes* that exist in all strains.
2. *Accessory genes* (*dispensable genes*, *variable genes*, or *adaptive genes*) that are present in some strains but not all.
3. *Unique genes* (*specific genes*) that are a particular form of accessory genes that are present only in one strain.

The collection of core genes is called the *core-genome*, and the collection of accessory genes is called the *accessory-genome*. Therefore the pan-genome = core-genome + accessory-genome. The total gene number in the pan-genome is:

Total genes = core genes + accessory genes
Pan-genome = core-genome + accessory-genome

The genes in the core-genome are often the signals of identity and make a species what it is. Core genes, also called *the minimal gene set*, are essential for normal cell functions such as DNA replication, transcription, and translation and are universally conserved. The genes in the accessory-genome are not necessary for basic life, at least for all conditions that bacteria encounter. The existence of these genes causes some strains to gain specific traits such as virulence and antibiotic resistance or the

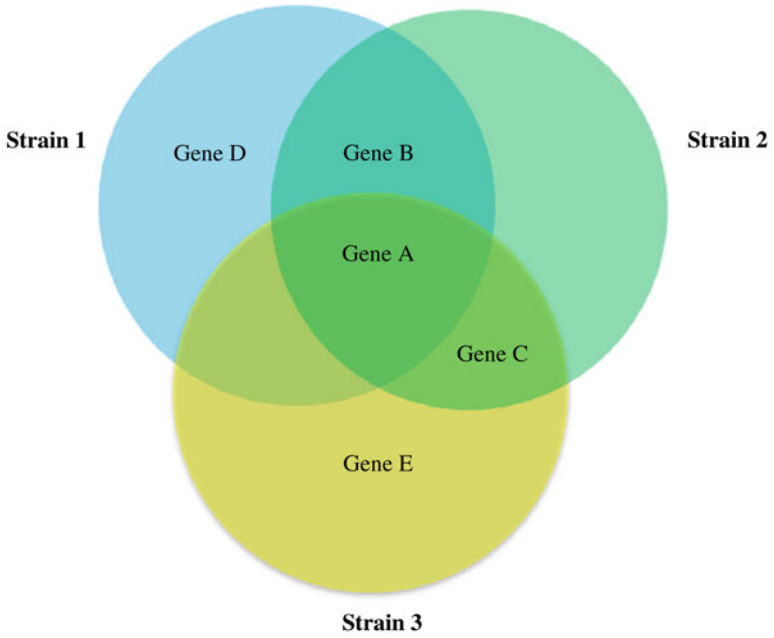

Fig. 2.1 Pan-genome of three strains. Gene A is a core gene as it exists in all strains. Gene B and C are accessory genes because they exist in two strains and not all. Gene D and E are unique genes specific to strain 1 and 3, respectively. This pan-genome has five genes A, B, C, D, and E in total

ability to occupy niche environments. The total number of genes in the pan-genome is usually larger than the number of genes in one single strain.

In some species, after adding a certain number of genomes from different strains, the total number of genes in the pan-genome does not increase further. This pan-genome that reaches a plateau is called a *closed* pan-genome. Meanwhile, for other species, every new strain adds new genes to the pan-genome. Such species have an *open* pan-genome that does not reach a plateau (Fig. 2.2). Species that are dormant and live in an isolated environment often have a closed pan-genome, whereas metabolically active species that have a diverse genome and horizontally transfer genes have an open pan-genome (Rouli et al. 2015).

It is worth noting that, although the pan-genome usually has a gene-based definition which means it refers to the entire gene set existing in different strains of one species, it can also have a sequence-based characterization which means it refers to all sequences found in different strains of one species. The gene-based definition considers variations at gene levels such as gene presence/absence and gene copy number variations (CNVs), while the sequence-based description is complete and considers all small-scale variants such as single nucleotide polymorphisms (SNPs), insertion/deletions (Indels), and structural variants (SVs) in coding and noncoding sequences. Nonetheless, the fundamental reason for the pan-genome definition is that a single individual genome is unable to show all genetic variants

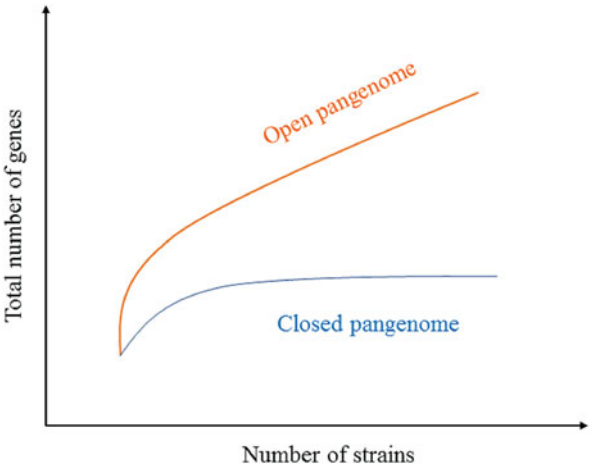

Fig. 2.2 In a closed pan-genome, the number of total genes will not increase after adding a certain number of strains. In open pan-genome, any new strain adds new genes to the pan-genome

found in the species. Therefore, a pan-genome is a hypothetical combination of variants that do not exist in reality. Thus, a more comprehensive pan-genome definition is:

Pan-genome is the entire set of all DNA sequences including genes and noncoding regions found in individuals of the species.

2.3 Computational Pan-Genomics

The discipline of computational pan-genomics refers to all computational principles that are applied for pan-genome visualization, statistical analysis, and software development.

2.3.1 Pan-Genome Graphical Representation

The main idea in pan-genomics is to replace traditional consensus and linear reference genomes by a pan-genome structure that captures all variants in one species. The two main methods to represent the pan-genome structure are:

1. A multiple sequence alignment (MSA)
2. A graph data structure

The structure of the pan-genome in Fig. 2.1 is visualized in Fig. 2.3 by an MSA and a graph.

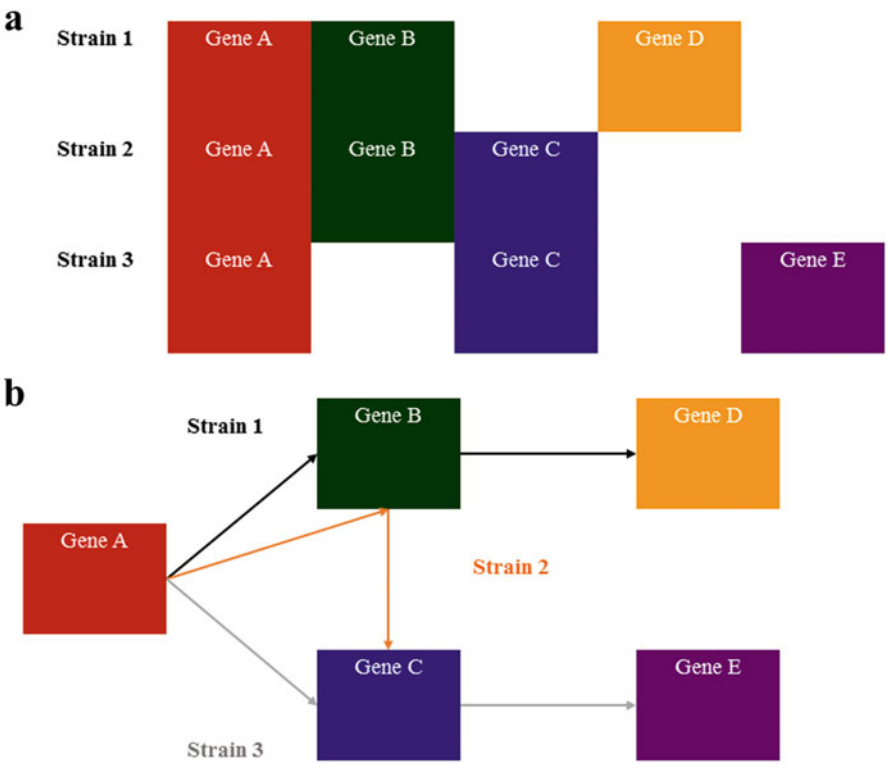

Fig. 2.3 (**a**) The pan-genome in Fig. 2.1 is shown by an MSA and (**b**) a directed graph. Each rectangle is a node that represents a gene, and each arrow is a directed edge that joins two nodes. Each path, which is a combination of edges on the graph, indicates the set of genes that are present in one strain; the black path shows genes found in strain 1, the orange path for strain 2, and the gray path for strain 3

2.3.1.1 Pan-Genome as an MSA

Representing a pan-genome as an MSA generates a large and sophisticated structure, the analysis of which is complicated. The reason is that each gene must be represented as many times as the number of isolates it exists in. For example, in Fig. 2.3a, Gene A is present in three strains, and Gene B is present in two strains. Therefore, Gene A appears three times and Gene B twice. Moreover, although the MSA can spot SNPs and Indels, it is not able to identify gene duplications and chromosomal structural variants. For almost all pan-genome analyses, researchers prefer graph data structures over an MSA.

2.3.1.2 Pan-Genome as a Graph

In molecular biology, the terms *graph* and *network* are used interchangeably (Huber et al. 2007). Before explaining how to represent a pan-genome as a graph data structure, it is helpful to introduce the basic definitions in the *graph theory* briefly.

The graph *G(V, E)* is a set of nodes or vertices (*V*) that are joined by a set of edges (*E*) (Fig. 2.3b). An edge that joins two nodes *u* and *v* is an *incident* on them and is denoted by *(u,v)*. Two nodes that are joined by an edge are called *adjacent nodes*, and two edges that are joined by a node are called *adjacent edges*. The edge *(u,u)* that joins the node *u* to itself is called a *loop*. The edge (*u*,*v*) that joins two different nodes u and v is named a *proper edge*. Multiple edges that join the same two nodes are called *multi-edges*. A complete graph is a graph where every pair of nodes is joined by an edge. A *directed edge* is an edge that has a specified direction and joins a start node called *tail node* to an end node called *head node*. The head node is the *successor* of the tail node, and the tail node is the *predecessor* of the head node. If all edges in a graph have direction, the graph will be called a *directed graph* or *digraph*. A *walk* is a way of getting from one node to another through a sequence of edges. A *path* is a walk in which every vertex appears only once. The length of the shortest path between two vertices is called the *distance* between them. A walk visiting every edge exactly once is called the *Eulerian walk* (Wilson 2006).

To design a graph that is able to represent all genetic variations in the pan-genome, the genome sequences of different strains are split into their substrings called *k-mers* which are subsequently arranged in a graph. A *k-mer* is a substring of certain length *k* (*k* is a natural number) that can be obtained from the sequence *S* of length *n* while $1 <= k < n$. The substring from position *i* to *j* is shown as *S*[*i*:*j*]. Therefore, any *k-mer* of sequence *S* is defined as below:

$$k\text{-}mers = S[i : i + k - 1] \quad \text{(Inclusive)} \quad (1 <= i <= n - k + 1) \& (1 <= k <= n).$$

The total number of *k-mers* will be $(n - k + 1)$.

Example

Sequence: S = "CGCTGAGCT"
Example of substrings: S[1:4] = "CGCT", S[2:4] = "GCT", S[5:5] = "G"
Sequence length: n = 9
K-mer length: k = 3
Kmers = S[i:i + k-1] = S[i:i + 2] (1 < = i < = 7)
A total number of k-mers of length 3 (3-mers) obtained from sequence S of length 9:
N – K + 1 = 9–3 + 1 = 7
All possible 3-mers obtained from S, note that there is a repeat of "GCT":
3-mers = ["CGC", "GCT", "CTG", "TGA", "GAG", "AGC", "GCT"]

The *k* value is critical here and must be selected carefully. It depends on the genome length, the available computing resources, and the type of analysis. To count all *k-mers* in a sequence, many space-efficient algorithms have been developed. Examples are *disk streaming of k-mers* (Rizk et al. 2013), *k-mer counter* (Kokot et al. 2017), and *Squeakr* (Pandey et al. 2018).

All *k-mers* derived from sequence *S* are arranged in a directed graph called a *de Bruijn graph* (DBG) denoted by *G(S,k)*. This graph contains a node for each distinct k-mer of *S*, and a directed edge *(u,v)* connects two nodes *u* and *v* if

$$u = S[i : i + k - 1] \text{ and } v = S[i + 1 : i + k]$$

The gene content of each strain is illustrated by an Eulerian walk on the graph.

To save memory and space, the DBG is compressed by merging its nodes to produce a *compressed DBG*. Two nodes *u* and *v* are allowed to be merged into a single node if

Node u is the only predecessor of node v, and node v is the only successor of node u. There may be multiple edges between them.

In a compressed DBG, every node (except the start node) has at least two different predecessors, or its single predecessor has at least two different successors, and every node (except the end node) has at least two different successors, or its only successor has at least two different predecessors (Beller and Ohlebusch 2016). A compressed DBG can be constructed by identifying maximal exact matches using a suffix tree (Marcus et al. 2014) or more efficiently by a combination of FM index, compressed suffix tree, and Burrows-Wheeler transform (Baier et al. 2015).

The k-mer-based representation of the pan-genome in a graph has many advantages such as simplicity, speed, and robustness. It is not always necessary to use fixed-length *k-mers* as the pan-genome can be arranged in acyclic and cyclic graphs (Marschall et al. 2016). The pan-genome graph is able to highlight all genetic diversities found in a species from SNPs and Indels to gene presence and absence. It renders a compact graphical portrait of the pan-genome that characterizes the variants among individuals. Moreover, graph-based pan-genomics provides access for retrieving data and defining a suitable coordinate system. It highlights the variable and conserved regions across the genomes. All genes are represented only once on the graph (no matter in how many strains they are present), and each strain is characterized by an exclusive walk on the graph (Fig. 2.4).

2.3.2 Pan-Genome Computational Analysis

High-performance and parallel computing are necessary for pan-genome computational analysis, particularly when a high number of strains are involved in the investigation. The computational pipeline often needs significant RAM and storage space.

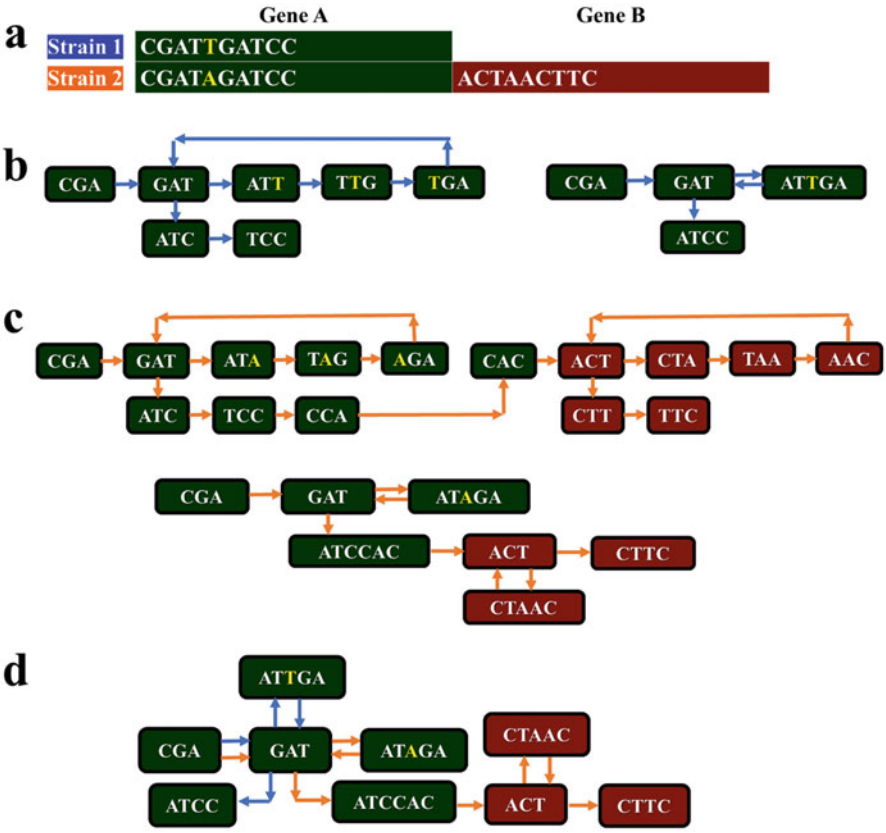

Fig. 2.4 (**a**) Genomes of strain 1 and strain 2; Gene A shown in green is a core gene, and gene B shown in red is an accessory gene; there is an SNP on position 5 of Gene A highlighted in yellow. (**b**) The genome of strain 1 is demonstrated by its 3-mers in the form of a DBG on the left and a compressed DBG on the right. (**c**) The genome of strain 2 is shown by its 3-mers in the form of a DBG on the top and a compressed DBG on the bottom. **(d)** The pan-genome of both strains is shown as a compressed DBG, the blue Eulerian walk indicates the genome of strain 1, and the orange Eulerian walk indicates the genome of strain 2; both variants in the form of SNP and gene presence/absence are identifiable on the graph; the yellow letter indicates the SNP, the green nodes show gene A, and the red nodes show gene B

A bacterial pan-genome analysis starts with a set of *whole genome sequencing (WGS)* short reads obtained from several closely related strains, preferably from the same species. The pipeline for a pan-genome analysis has four main steps (Fig. 2.5):

1. Reads quality control, preprocessing, and cleaning
2. Genome assembly and annotation
3. Pan-genome construction
4. Pan-genome downstream analysis

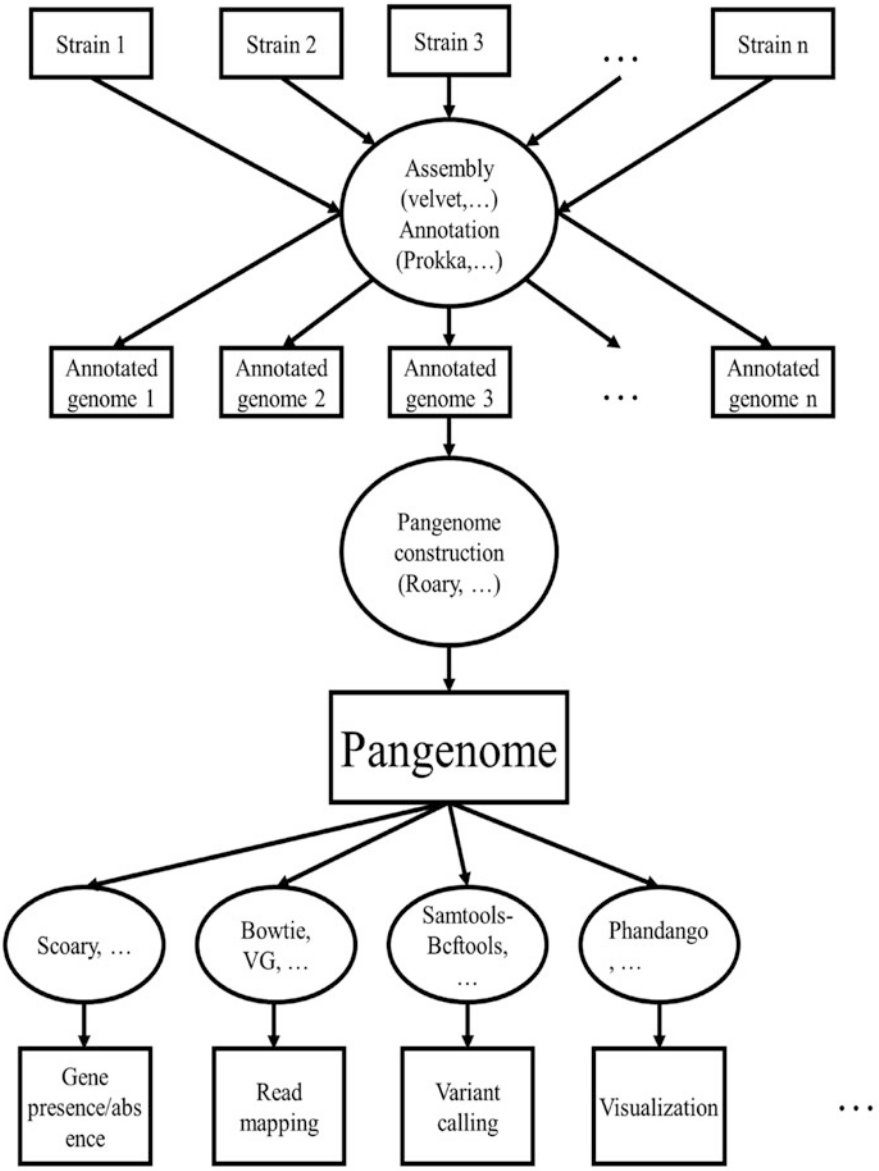

Fig. 2.5 Overview of pan-genomic analysis

2.3.2.1 Reads Quality Control, Pre-processing and Cleaning

The sequencing short reads are stored in standard file formats like *fastq* (Cock et al. 2009). The quality of the reads is evaluated by tools such as *fastqc* (Andrews 2010), the adapter sequences are trimmed, and sequences with low

quality are removed by tools like *FASTX toolkit* (Gordon and Hannon 2010). The clean sequences that are of high quality are supplied to the genome assemblers.

2.3.2.2 Genome Assembly and Annotation

The next step is to assemble the genomes of all strains. Typically, a pan-genome analysis is useful when working with a bacterial species whose genome is highly divergent across different strains. Defining a linear reference genome for such a diverse species is difficult. Thus, de novo *assembly* (Paszkiewicz and Studholme 2010), which is reference-free, is desired in computational pan-genomics. The genome assembly can be achieved by some publicly available tools such as *VelvetOptimiser* (Gladman and Seemann 2008) and *SOAPdenovo* (Luo et al. 2015). For a successful pan-genome analysis, the assembled genome should be of high quality, and contigs should have a minimum length of 500 base pairs. Tools like *Quast* (Gurevich et al. 2013) can be used to evaluate the quality of the assembled genomes.

To define the pan-genome and determine its core and accessory genes, all assembled genomes must be annotated coherently with a tool that is compatible with the pan-genome builder. Accurate assembly and annotation produce a pan-genome with high quality enabling a productive analysis. Annotated genomes are saved in standard file formats, such as *BED*, *GFT*, *GFF*, *and GFF3*. Many tools have been developed for bacterial gene prediction and annotation. Examples include *Glimmer* (Delcher et al. 2007) and *Prokka* (Seemann 2014). *Prokka* is specifically designed for prokaryotic genome annotation and works based on the integration of several tools and databases such as *SignalP* (Petersen et al. 2011), *Aragorn* (Laslett and Canback 2004), *HMMER3* (Finn et al. Finn et al. 2011), *Rfam* (Nawrocki et al. 2015), and *Infernal* (Nawrocki and Eddy 2013). *Prokka* is a well-run annotator that produces its outputs in various file formats, at least one of which will be compatible with one of the tools used for pan-genome construction. For annotation from scratch, *Pannotator* (Santos et al. 2013) is suitable, and to improve the available annotations, *eCAMBer* (Wozniak et al. 2014) and *Mugsy-Annotator* (Angiuoli et al. 2011) can be applied.

2.3.2.3 Pan-Genome Construction

As discussed earlier, the pan-genome can be defined either as a collection of genes or genome sequences from multiple strains of one species. For this reason, two types of tools have been developed for pan-genome construction and analysis (Zekic et al. 2018):

1. Gene-based tools
2. Sequence-based tools

To use the gene-based tools, all genomes must be annotated, and the gene content of each strain must be determined. These tools first use graph-based methods to assign orthologous genes found in strains and then construct the pan-genome. Some of the most popular gene-based tools developed so far are *EDGAR* (Blom et al. 2016),

PGAT (Brittnacher et al. 2011), *PGAP* (Zhao et al. 2012), *PanOCT* (Inman et al. 2018), *GET_HOMOLOGOUS* (Contreras-Moreira and Vinuesa 2013), *PanFunPro* (Lukjancenko et al. 2013), *ITEP* (Benedict et al. 2014), *PanGP* (Zhao et al. 2014), *LS-BSR* (Sahl et al. 2014), *Roary* (Page et al. 2015), *Micropan* (Snipen and Liland 2015), *Piggy* (Thorpe et al. 2018), *BPGA*, and *Pyseer* (Lees et al. 2018).

For a sequence-based pan-genome analysis, the sequences of different genomes are indexed. To increase efficiency regarding required time and memory, graph-based methods are applied. DBG is usually employed here as the analysis does not require a reference sequence or alignment. Examples of sequence-based tools are *Panseq* (Laing et al. 2010), *Harvest* (Treangen et al. 2014), *SplitMEM* (Marcus et al. 2014), *TwoPaCo* (Minkin et al. 2017), and *Bloom Filter Trie* (Holley et al. 2016).

2.3.2.4 Pan-Genome Downstream Analysis

Most of the tools mentioned above can perform some downstream analysis. The downstream analysis includes tasks such as multiple sequence alignment of the core-genomes, phylogenetic tree construction, alignment of the short reads to the pan-genome, variant calling, studying of genes in different metabolic pathways, pan-genome visualization, and various statistical analyses.

The multiple sequence alignment of core-genomes is sometimes produced by the pan-genome builder. This alignment is then used to extract the variant sites in the core genes which are used to draw an initial phylogenetic tree. This tree can be colored according to the sample phenotypes and provides an overview of the association between samples from different phenotypes. *Snp-sites* (Keane et al. 2016) is a tool which is useful to extract all variant sites from the multiple sequence alignment of all the core genes. The phylogenetic tree can be drawn by considering only those variant sites in the core-genome of the different strains. However, if sufficient computing resources are available, the phylogenetic tree can be drawn directly from the alignment of the core-genomes. Tools like *ClustalW* (Larkin et al. 2007) and *FastTree* (Price et al. 2010) are appropriate for this tree drawing. The tree coloring and visualization can be performed with the help of tools such as *Evolview* (He et al. 2016).

The software *Scoary* (Brynildsrud et al. 2016) was developed to score genes in the pan-genome for their association with an observed trait. This tool finds genes whose presence or absence are strongly associated with a phenotype and considers the influence of the population stratification. *Piggy* (Thorpe et al. 2018), on the other hand, is a tool that examines the variation in intergenic regions in bacteria. Apart from genes, the presence or absence of some intergenic regions affects the phenotypic behavior of the bacterium. Both *Scoary* and *Piggy* can use the output of *Roary* as their input.

To perform a pan-genome-based GWAS, the sequence of the genes in the pan-genome can be utilized as a reference to identify SNPs and Indels in each strain and determine whether they are in the core genes or accessory genes. In this case, the pan-genome is usually saved in a fasta file in which each record represents the consensus sequence of a gene drawn from the entire population. Then the short reads are aligned to this reference sequence and variants are called. The SNPs in the core

genes reflect the age of the species. To reduce the analysis workload, specific informative SNPs should be selected in a process called representative SNP selection (Hurgobin and Edwards 2017). As explained earlier, the pan-genome can be saved as a graph. Several tools have been developed to align short reads directly to the pan-genome graph. Examples are *BGREAT* (Limasset et al. 2016) and VG.[1] The pan-genome graph can also be used for reference-free variant calling (Iqbal et al. 2012). For further details about the tools, their algorithms, and performance, refer to (Xiao et al. 2015), (Vernikos et al. 2015), and (Zekic et al. 2018). Many scripts written in *R* and *Python* are available for pan-genome visualization; some of the more versatile tools for this are *Phandango* (Hadfield et al. 2018), *Panx* (Ding et al. 2017), and *Panviz* (Pedersen et al. 2017).

2.4 Pan-Genomics Research Examples

As mentioned earlier, the importance of pan-genomics in medicine and microbiology was first considered in 2005, when (Tettelin et al. 2005) utilized the term pan-genome for the first time and investigated six strains of *Streptococcus agalactiae*, which is the primary cause of neonatal infection in human. The research concluded that multiple strains must be sequenced to study a bacterial species because many of them have an open pan-genome. In some cases, hundreds or even thousands of strains must be sequenced. It is essential to consider their large accessory-genomes to identify potential candidates for the design of effective drugs or universal vaccines. To date, many studies have been conducted based on pan-genomics algorithms and principles.

As an example, (Rasko et al. 2008) carried out a pan-genome analysis on 17 samples of *E. coli*. Their dataset was composed of commensal, extraintestinal pathogenic, and diarrheagenic samples. They identified an open pan-genome which was made up of more than 13,000 genes including 2200 core genes. They found isolate-specific genes that led them to assume that each isolate can develop its virulence independently. They suggested that extraintestinal pathogenic samples share a significant level of similarity; however, in general, this study showed that the *E. coli* pathovars are not distinct on the molecular level (Rasko et al. 2008).

Pan-genomics can also be applied to investigate the genome of one bacterial species in association with other species. For example, (Donati et al. 2010) analyzed the genome of 44 strains of pathogenic *Streptococcus pneumoniae* and compared them with strains of nonpathogenic *Streptococcus mitis*. According to their results, *Streptococcus pneumoniae* has an open pan-genome that enables the bacterium to respond to the different environments. They determined that homologous recombination is the primary evolutionary process used by *Streptococcus pneumoniae*. The genetic materials can be exchanged within

[1] https://github.com/vgteam/vg

the species or with other species like *Streptococcus mitis*. They correlated the age of clones with the number of acquired genes (Donati et al. 2010).

In another study carried out by (D'Auria et al. 2010), they built the pan-genome of five strains of *Legionella pneumophila*. They compared the gene content of a persistent strain from Spain to the genome of four other strains from other countries such as England, France, and the United States. Out of their constructed pan-genome, they identified 53 genes specific to the pathogenic strain from Spain. The research demonstrated that the accessory-genome contains new traits that can be exchanged through horizontal gene transfer and the virulence of the bacterium is promoted by part of its core-genome. In this study, pan-genomics was applied to compare samples from different geographical locations (D'Auria et al. 2010).

To improve the accuracy of a pan-genome study, especially for species that have a very diverse genome, hundreds or even thousands of samples must be included in the study. This type of data analysis requires excessive computing power and time. This issue has been resolved with the availability of *high-performance computing (HPC)* algorithms and clusters. A pan-genome study conducted by (Azarian et al. 2018) used 937 *Streptococcus pneumoniae* samples to investigate the evolutionary impact of vaccination. They concluded that the introduction of the *pneumococcal conjugate vaccine (PCV)* reduced the pan-genome size and the genetic diversity and changed frequencies of genes. However, the genetic diversity expanded again through in-migration of non-vaccine lineages, and frequencies of genes returned to the original value by selection (Azarian et al. 2018). In this study, pan-genomics was applied to compare samples from different time points.

In addition to bacteria, pan-genomics can be applied for eukaryotes and is particularly useful for plants. A study by (Gordon et al. 2017) explored 54 lineages of the grass *Brachypodium distachyon*. The number of genes in the pan-genome was twice the number of genes existing in an individual genome. As expected, they demonstrated that the core genes are essential for basic biological functions, while accessory genes are required for beneficial functions such as defense and development. They concluded that genes in the accessory-genome have critical roles in the phenotype of different individuals and transposable elements are significantly involved in the pan-genome evolution.

2.5 Conclusion

Pan-genomics has revolutionized our understanding of microbiology. Thanks to advances in sequencing technologies, researchers can access hundreds or even thousands of sequenced genomes from the same species. Different strains of a bacterial species often have different gene contents, and each of them has its combination of genes. The pan-genome has been developed to address this genomic diversification in bacteria. For some bacterial species, the diversity is high, and the ratio of *core-genome/pan-genome* is low. For example, only 3.7% of the pan-genome obtained from 2000 *Escherichia coli (E.coli)* strains (Land et al.

2015) and 3.5% of the pan-genome obtained from 53 strains of *Campylobacter concisus* (Gemmell et al. 2018) are core and shared by all strains.

On the other hand, in higher eukaryotic organisms, the gene contents of different individuals are almost the same, and their genomic diversity is mainly in the form of SNPs, Indels, SVs, and CNVs, which affect gene regulation or function instead of the gene content. For instance, the genetic diversity across human genomes is about 0.6% (Auton et al. 2015), and they share more than 99% of their gene repertoire. If we define a eukaryotic species as a group of individuals that share more than 99% of their genomes, we should redefine it for bacteria as many of them do not meet this threshold.

The considerable variation in gene content between bacterial strains is attributed to their capability to exchange their genetic material through transformation, transduction, and conjugation. They can use this ability to adapt themselves to their environment, survive in extreme conditions, become pathogenic, and acquire resistance to antibiotics. From the medical perspective, the pan-genome is an ideal reference genome to highlight all variants among bacterial strains. It helps researchers to differentiate pathogenic isolates in a large population and understand their pathology. The pan-genome can be applied as a reference to find genes or sequences that are significantly liked to the particular phenotypes such as invasiveness or antibiotic resistance and is thus crucial for facilitating bacterial disease control, vaccine development, and drug design.

References

Andrews S (2010) FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angiuoli SV et al (2011) Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinform. https://doi.org/10.1186/1471-2105-12-272

Auton A et al (2015) A global reference for human genetic variation. Nature. https://doi.org/10.1038/nature15393

Azarian T et al (2018) The impact of serotype-specific vaccination on phylodynamic parameters of Streptococcus pneumoniae and the pneumococcal pan-genome. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1006966

Baier U, Beller T, Ohlebusch E (2015) Graphical pan-genome analysis with compressed suffix trees and the burrows-wheeler transform. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv603

Behjati S, Tarpey PS (2013) What is next generation sequencing? Arch Dis Child Educ Pract Ed 98 (6):236–238. https://doi.org/10.1136/archdischild-2013-304340

Beller T, Ohlebusch E (2016) A representation of a compressed de Bruijn graph for pan-genome analysis that enables search. Algorithms Mol Biol. https://doi.org/10.1186/s13015-016-0083-7

Benedict MN et al (2014) ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC Genomics. https://doi.org/10.1186/1471-2164-15-8

Blevins SM, Bronze MS (2010) Robert Koch and the "golden age" of bacteriology. Int J Infect Dis. https://doi.org/10.1016/j.ijid.2009.12.003

Blom J et al (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw255

Brittnacher MJ et al (2011) PGAT: a multistrain analysis resource for microbial genomes. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr418

Brynildsrud O et al (2016) Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 17(1):238. https://doi.org/10.1186/s13059-016-1108-8
Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. https://doi.org/10.1128/AEM.02411-13
D'Auria G et al (2010) *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics. https://doi.org/10.1186/1471-2164-11-181
Delcher AL et al (2007) Identifying bacterial genes and endosymbiont DNA with glimmer. Bioinformatics. https://doi.org/10.1093/bioinformatics/btm009
Ding W, Baumdicker F, Neher RA (2017) panX: pan-genome analysis and exploration. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx977
Donati C et al (2010) Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biol. https://doi.org/10.1186/gb-2010-11-10-r107
Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr367
Gemmell MR et al (2018) Comparative genomics of campylobacter concisus: analysis of clinical strains reveals genome diversity and pathogenic potential. Emerg Microb Infect. https://doi.org/10.1038/s41426-018-0118-x
Gest H (2004) The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, fellows of the Royal Society. Notes Records R Soc. https://doi.org/10.1098/rsnr.2004.0055
Gladman S, Seemann T (2008) Velvet optimiser. Free Softw Found. https://doi.org/10.1016/S0925-8574(99)00040-3
Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. https://doi.org/10.1038/nrg.2016.49
Gordon A, Hannon GJ (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools, http://hannonlab.cshl.edu/fastx_toolkit/
Gordon SP et al (2017) Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat Commun. https://doi.org/10.1038/s41467-017-02292-8
Grebennikova TV et al (2018) The DNA of bacteria of the world ocean and the earth in cosmic dust at the international Space Station. Sci World J. https://doi.org/10.1155/2018/7360147
Gurevich A et al (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt086
Hadfield J et al (2018) Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx610
He Z et al (2016) Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw370
Holley G, Wittler R, Stoye J (2016) Bloom filter Trie: an alignment-free and reference-free data structure for pan-genome storage. Algorithms Mol Biol. https://doi.org/10.1186/s13015-016-0066-8
Huber W et al (2007) Graphs in molecular biology. BMC Bioinform. https://doi.org/10.1186/1471-2105-8-S6-S8
Hurgobin B, Edwards D (2017) SNP discovery using a Pangenome: has the single reference approach become obsolete? Biology 6(1):21. https://doi.org/10.3390/biology6010021
Inman JM et al (2018) Large-scale comparative analysis of microbial Pan-genomes using PanOCT. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty744
Iqbal Z et al (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. https://doi.org/10.1038/ng.1028
Kara R, Robert JK (2018) Bacteria | cell, evolution, & classification | Britannica.com. Encyclopaedia Britannica, Inc
Keane JA et al (2016) SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genom. https://doi.org/10.1099/mgen.0.000056

Kokot M, Dlugosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. Bioinformatics (Oxford, UK). https://doi.org/10.1093/bioinformatics/btx304
Laing C et al (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinform. https://doi.org/10.1186/1471-2105-11-461
Land M et al (2015) Insights from 20 years of bacterial genome sequencing. Funct Integrat Genom. https://doi.org/10.1007/s10142-015-0433-4
Lanska DJ (2014) Pasteur, Louis. In: Encyclopedia of the neurological sciences. https://doi.org/10.1016/B978-0-12-385157-4.00973-8
Larkin M et al (2007) ClustalW and ClustalX version 2. Bioinformatics. https://doi.org/10.1093/bioinformatics/btm404
Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. https://doi.org/10.1093/nar/gkh152
Lees JA et al (2018) pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty539
Leinonen R et al (2011) The European nucleotide archive. Nucleic Acids Res 39(Suppl 1). https://doi.org/10.1093/nar/gkq967
Limasset A et al (2016) Read mapping on de Bruijn graphs. BMC Bioinform. https://doi.org/10.1186/s12859-016-1103-9
Lukjancenko O et al (2013) PanFunPro: PAN-genome analysis based on FUNctional PROfiles. F1000 Res. https://doi.org/10.12688/f1000research.2-265.v1
Luo R et al (2015) Erratum to "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler" [GigaScience, (2012), 1, 18]. GigaScience. https://doi.org/10.1186/s13742-015-0069-2
Maloy S (2013) Bacterial genetics. In: Encyclopedia of biodiversity: second edition. https://doi.org/10.1016/B978-0-12-384719-5.00431-7
Marcus S, Lee H, Schatz MC (2014) SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu756
Marschall T et al (2016) Computational Pan-genomics: status, promises and challenges. bioRxiv. https://doi.org/10.1101/043430
Mengoni A, Galardini M, Fondi M (2015) Bacterial Pangenomics: methods and protocols. Methods Mol Biol. https://doi.org/10.1007/978-1-4939-1720-4
Minkin I, Pham S, Medvedev P (2017) TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. Bioinformatics (Oxford, UK). https://doi.org/10.1093/bioinformatics/btw609
Miyazaki S et al (2004) DDBJ in the stream of various biological data. Nucleic Acids Res 32 (Database issue):D31–D34. https://doi.org/10.1093/nar/gkh127
Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt509
Nawrocki EP et al (2015) Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. https://doi.org/10.1093/nar/gku1063
Ostell J, McEntyre J (2007) The NCBI handbook. NCBI Bookshelf:1–8. https://doi.org/10.4016/12837.01
Page AJ et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31 (22):3691–3693. https://doi.org/10.1093/bioinformatics/btv421
Pandey P et al (2018) Squeakr: an exact and approximate k-mer counting system. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx636
Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. Brief Bioinform. https://doi.org/10.1093/bib/bbq020
Pedersen TL et al (2017) PanViz: interactive visualization of the structure of functionally annotated pangenomes. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw761
Cock PJA et al (2009) The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. https://doi.org/10.1093/nar/gkp1137

Petersen TN et al (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. https://doi.org/10.1038/nmeth.1701

Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. https://doi.org/10.1371/journal.pone.0009490

Rasko DA et al (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol. https://doi.org/10.1128/JB.00619-08

Rizk G, Lavenier D, Chikhi R (2013) DSK: K-mer counting with very low memory usage. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt020

Rouli L et al (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microb New Infect 7:72–85. https://doi.org/10.1016/j.nmni.2015.06.005

Sahl JW et al (2014) The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. Peer J. https://doi.org/10.7717/peerj.332

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Santos AR et al (2013) PANNOTATOR: an automated tool for annotation of pan-genomes. Genet Mol Res. https://doi.org/10.4238/2013.August.16.2

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14):2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Snipen L, Liland KH (2015) micropan: an R-package for microbial pan-genomics. BMC Bioinform. https://doi.org/10.1186/s12859-015-0517-0

Tettelin H et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci 102 (39):13950–13955. https://doi.org/10.1073/pnas.0506758102

Thorpe HA et al (2018) Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. GigaScience. https://doi.org/10.1093/gigascience/giy015

Treangen TJ et al (2014) The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. https://doi.org/10.1186/s13059-014-0524-x

Vernikos G et al (2015) Ten years of pan-genome analyses. Curr Opin Microbiol. https://doi.org/10.1016/j.mib.2014.11.016

'WHO | Press release' (2013) WHO. World Health Organization. Available at: http://www.who.int/whr/1996/media_centre/press_release/en/. Accessed 12 Sept 2018

Wilson RJ (2006) Graph theory. In: History of topology. https://doi.org/10.1016/B978-044482375-5/50018-3

Wozniak M, Wong L, Tiuryn J (2014) ECAMBer: efficient support for large-scale comparative analysis of multiple bacterial strains. BMC Bioinform. https://doi.org/10.1186/1471-2105-15-65

Xiao J et al (2015) A brief review of software tools for pangenomics. Genomics Proteom Bioinform. https://doi.org/10.1016/j.gpb.2015.01.007

Zekic T, Holley G, Stoye J (2018) Pan-genome storage and analysis techniques. Methods Mol Biol. https://doi.org/10.1007/978-1-4939-7463-4_2

Zhao Y et al (2012) PGAP: Pan-genomes analysis pipeline. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr655

Zhao Y et al (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu017