First Steps in the Analysis of Prokaryotic Pan-Genomes

Sávio Souza Costa[1,2], Luís Carlos Guimarães[1] (ID), Artur Silva[1,2], Siomar Castro Soares[3] and Rafael Azevedo Baraúna[1,2] (ID)

[1]Centro de Genômica e Biologia de Sistemas, Universidade Federal do Pará, Belém, Brazil.
[2]Laboratório de Engenharia Biológica, Espaço Inovação, Parque de Ciência e Tecnologia Guamá, Belém, Brazil. [3]Instituto de Ciências Biológicas e Naturais, Universidade Federal do Triângulo Mineiro, Uberaba, Brazil.

ABSTRACT: Pan-genome is defined as the set of orthologous and unique genes of a specific group of organisms. The pan-genome is composed by the core genome, accessory genome, and species- or strain-specific genes. The pan-genome is considered open or closed based on the alpha value of the Heap law. In an open pan-genome, the number of gene families will continuously increase with the addition of new genomes to the analysis, while in a closed pan-genome, the number of gene families will not increase considerably. The first step of a pan-genome analysis is the homogenization of genome annotation. The same software should be used to annotate genomes, such as GeneMark or RAST. Subsequently, several software are used to calculate the pan-genome such as BPGA, GET_HOMOLOGUES, PGAP, among others. This review presents all these initial steps for those who want to perform a pan-genome analysis, explaining key concepts of the area. Furthermore, we present the pan-genomic analysis of 9 bacterial species. These are the species with the highest number of genomes deposited in GenBank. We also show the influence of the identity and coverage parameters on the prediction of orthologous and paralogous genes. Finally, we cite the perspectives of several research areas where pan-genome analysis can be used to answer important issues.

KEYWORDS: Pan-genome, core genome, accessory genome

Introduction

Pan-genome was a term coined by Tettelin et al[1] to describe the gene content of several strains of *Streptococcus agalactiae*. The pan-genome is divided into core genome, dispensable or accessory genome, and singleton genes (ie, species-specific genes). Fifteen years after the publication of Tettelin's article, the number of genomes sequenced and available in databases have grown exponentially surpassing 30 000 complete and draft genomes in 2020 (https://gold.jgi.doe.gov/statistics). The evolution of sequencing technologies from classic chain termination method to fourth-generation sequencing, based on a massively parallel analysis, has been facilitating cost reduction over the years.[2] However, in many countries, the sequencing value still exceeds the prediction of USD 1000 per genome.[3] Despite criticisms about the use of draft genomes in pan-genome analysis, several new software have been developed to improve the assembly of these draft genomes.[4] For example, *Escherichia coli* has 15 275 genomes in scaffold or contigs available in GenBank (https://www.ncbi.nlm.nih.gov/genome/genomes/167) and their use for pan-genome studies is considered limited.

A search in PubMed database using the words "pan genome" or "pan-genome" returns a total of 494 works published in the last 5 years (2015 to date). This number tends to increase because the results of pan-genome analyses are becoming more accurate. Zeng et al[5] used a new pan-genome reverse vaccinology approach and found 121 cell surface-exposed proteins belonging to the core genome of *Leptospira interrogans*. These proteins proved to

be highly antigenic and widely distributed in the species. Thus, these proteins are potential candidates for vaccine development. Pan-genome analysis was also applied to the discovery of antiphage defense systems,[6] in RNAseq analysis,[7] and evolutionary studies of adaptation to different hosts.[8]

In this review, we present the main concepts and software used for the analysis of prokaryotic pan-genomes. First, introduction to basic concepts is presented followed by an up-to-date description of the most recent software for pan-genome analysis. We also present a pan-genome analysis of the 9 bacterial species with the highest number of genomes deposited in GenBank.

Basic Concepts

Pan-genome structure

The sequence of a single genome does not reflect the entire genetic variability of a bacterial species. Complex analysis such as evolutionary genomics and molecular pathogenesis require a large number of sequenced genomes.[1,9] Fortunately, the constant evolution of sequencing technologies has been allowing the reduction of sequencing time and cost. Consequently, an exponential increase in the number of genomes available in the databases has been observed. New research fields have emerged such as comparative genomics, whose principle is to compare the genetic content of several taxonomically related microorganisms.[10] For example, in the pre-genomic era, 2 strains were classified in the same species
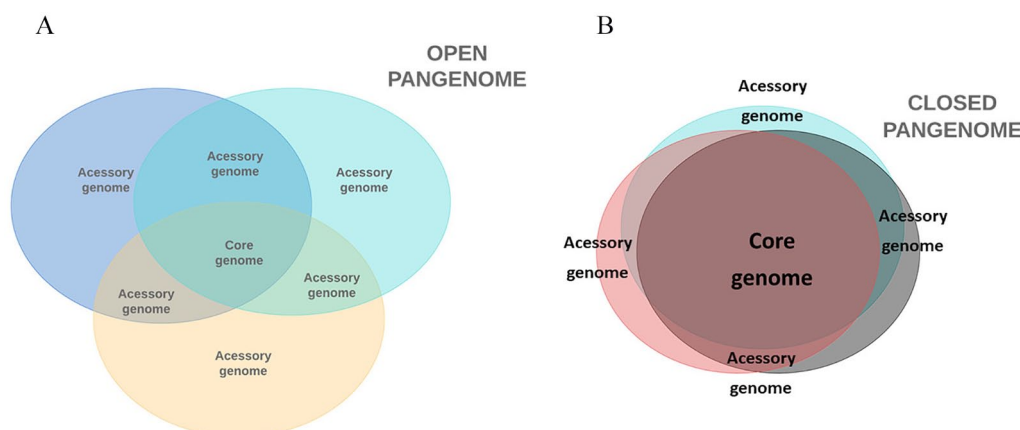
Figure 1. Venn diagram representing the subgroups of a pan-genome. Each set represents the gene families detected in a genome. The intersection of these sets represents the core genome. The number of gene families in the core genome corresponds to the size of the intersection. The fraction of genes corresponding to the core genome in an open pan-genome (A) is smaller than in a closed pan-genome (B). In contrast, the fraction of genes corresponding to the accessory genome in an open pan-genome (A) is higher than in a closed pan-genome (B).

if they presented 70% DNA-DNA reassociation.[11] In the post-genomic era, several other methods can be applied to evaluate taxonomic relationships such as average nucleotide identity (ANI).[12] A broader discussion about the concept of bacterial species will be accomplished later. Recently, the tree of life was updated based on the comparative analysis of a large number of bacterial genomes.[13]

A pan-genome is determined through comparative genomic analyses. A pan-genome consists in the set of non-redundant gene families belonging to a taxonomically related group of organisms. The pan-genome is divided into 3 sub-groups as demonstrated in Figure 1. Core genome is the set of genes shared by all analyzed microorganisms. Most of these genes are involved in vital roles for bacterial survival.[1,14,15] However, genes of the core genome may also be involved in pathogenicity and virulence in some bacterial species.[16,17] Accessory or dispensable genome is composed by the set of genes that are present in 2 or more genomes but not all.[18] Singleton genes such as species- or strain-specific genes are those present in only one genome. Usually, accessory and singleton genes are acquired by horizontal gene transfer (HGT) or evolved due to mutations in pre-existing genes. They are commonly related to a specific metabolism, virulence, antibiotic resistance mechanism, or other environmental adaptation.[19] A pan-genome is classified as open or closed depending on the probability of detecting new gene families as new genomes are added into the analysis. In an open pan-genome, the number of gene families will continuously increase with the addition of new genomes to the analysis. In contrast, in a closed pan-genome, the number of gene families will not increase considerably.

Genomic plasticity and the accessory genome

The flexible portion of the pan-genome (accessory and strain-specific genes) is the main genetic component responsible for the adaptation of a bacterial population to different environmental stresses. In this context, it is important to distinguish the terms genomic plasticity and accessory genome. Genomic plasticity is used to describe the mobile genetic elements (MGEs) and hypervariable regions that transform the genome into a dynamic molecule. Therefore, it is a concept used to discuss the genetic variability of a single or multiple genome without necessarily making use of a pan-genomic approach. Accessory genome is the set of genes that, after a pan-genomic analysis, are present in 2 or more genomes but not all. Thus, it comprises the variable portion of a pan-genome. In some cases, MGEs and hypervariable regions comprise most of the accessory genome.[20] In strains of *Bacillus amyloliquefaciens*, most gene clusters for the production of secondary metabolites are present in the accessory genome of the species.[21]

Orthologous and paralogous genes

Most comparative genomic analyses begin by identifying the homologous characteristics between 2 or more prokaryotic genomes. These homologies range from large chromosomal segments to genes or even point mutations. In a pan-genome analysis, the genes are the main characteristics evaluated. From an evolutionary perspective, a gene is classified as homologous or analogous. Homologous genes are those originated from a common ancestor, whereas the analogous genes evolved independently through convergent evolution. In both cases, they will present the same function but in different organisms. About 15% of the genes of a bacterium are acquired through HGT.[22] Thus, it is difficult to apply the concept of analogous genes in the Bacteria domain.

A pan-genomic analysis searches for homologous genes within the set of analyzed genomes. These homologous genes are divided into orthologous and paralogous genes.[23] Orthologous genes diverged via evolutionary speciation.
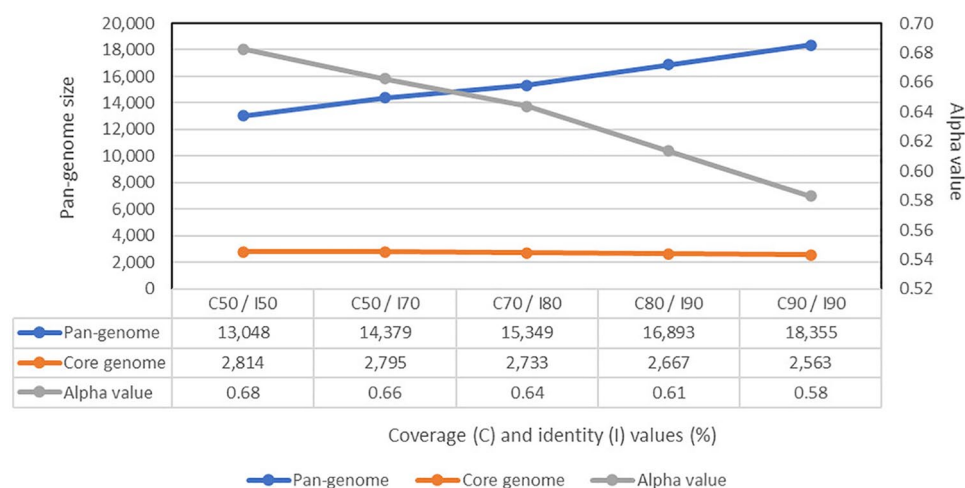
Figure 2. Pan-genome, core genome, and alpha value for each one of the five analyses performed using 30 genomes of *E coli* randomly downloaded from GenBank. In a more stringent analysis (high values of coverage and identity), the trend is the increase of the pan-genome size and the decrease of the alpha value. C indicates coverage; I, identity.

Paralogous genes diverged via gene duplication. Thus, orthologous genes are those shared by 2 or more bacteria and have equivalent biological function. It is worth noting that orthologous genes tend to be more conserved than paralogous genes.[24] In contrast, paralogous genes commonly undergo several mutations after their duplication leading to a change in their biological function.[25]

As pan-genomic analysis is based on sequence homology, some parameters such as coverage and identity must be carefully chosen. These parameters will strongly influence in the detection of orthologous genes. To demonstrate this issue, we calculated the pan-genome of *E coli* using different coverage and identity values. Thirty genomes were randomly downloaded from GenBank. Five analyses were performed starting from 50% coverage and 50% identity up to 90% coverage and 90% identity (Figure 2). The size of the pan-genome and core genome as well as the alpha value of the Heap's law changed significantly (Figure 2). The pan-genome increased from 13 000 to about 18 000 gene families. The alpha value decreased from 0.68 to 0.58.

High values of coverage and identity lead to overestimation of the pan-genome size and may separate groups of orthologous genes. The opposite is also true. Low values of coverage and identity lead to the clustering of non-orthologous genes. One way to determine the best values for these parameters is checking the clustering of known orthologous genes in such a way that they will serve as an internal control during the analysis. Other characteristics that should be taken into consideration include the genomic plasticity and the taxonomic level of the microorganisms being compared.

Some software allows the user to modify other parameters that also strongly influence the results of the analysis. For example, GET_HOMOLOGUES[26] offers the possibility to use DIAMOND algorithm instead of BLAST to perform the alignment. This software also allows user to choose the

algorithm used for bidirectional search of best hits, between COGtriangle and orthoMCL. A broader discussion about the types of orthology analysis used by each software will be further presented.

Pan-genomic concept of bacterial species

In 1942, Ernst Mayr proposed a species concept that is widely used for eukaryotes: "species are groups of interbreeding natural populations that are reproductively isolated from the other such groups." The microbial world, basically composed by microorganisms that reproduce asexually, is therefore one of the great bottlenecks of the species concept proposed by Mayr. Bacteria are able to exchange genetic material through HGT.

Nevertheless, taxonomy and systematics are extremely important for basic analysis in microbiology. Species is the fundamental taxonomic unit and in the absence of a concept that encompasses all living beings, new ideas were presented or rediscussed.[27] The Bergey's Manual of Systematics of Archaea and Bacteria has a broader concept of species: "a distinct group of strains that have certain distinguishing features and that generally bear a close resemblance to one another in the more essential features of organization." Due to the emergence of modern (or molecular) microbiology that is based on the genetic analysis of cultivable or uncultivable strains, the concept of bacterial species is becoming increasingly divergent from the concept of Mayr and clearer than the concept provided by the Bergey's manual. Currently, bacteria with >70% DNA-DNA hybridization and >97% 16S rRNA sequence identity are classified in the same species.[28] Even so, taxonomic classification of environmental bacteria is a bottleneck. The overwhelming majority of free-living microorganisms are uncultivable, which makes the hybridization analysis of complete DNA molecules unfeasible. To circumvent this bias, environmental microbiologists began to use the Operational
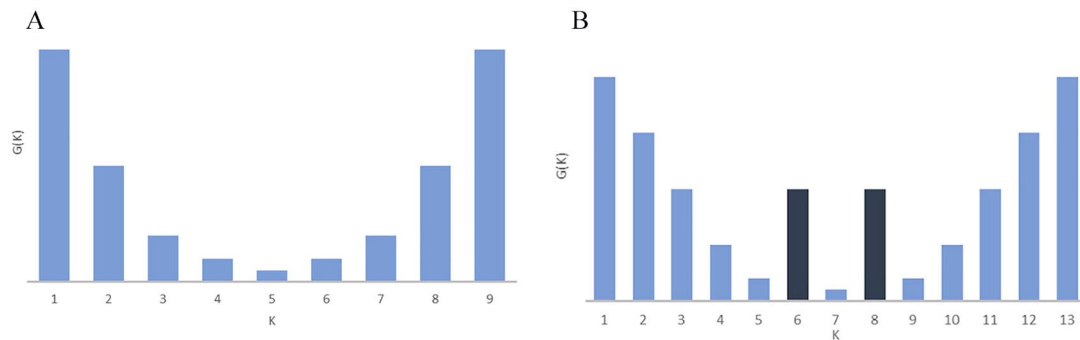
A

B

Figure 3. *G*(*k*) function for 2 sets of genomes. (A) A U-like shape chart indicates that the genomes analyzed belong to strains of the same species. (B) Internal peaks (highlighted in dark blue) suggest that the genomes belong to different species.

Taxonomic Unit (OTU) definition to replace the biological species concept in microbial ecology analyses.[29]

One of the most recent concepts of bacterial species was proposed by Bobay and Ochman.[30] They proposed that bacterial strains should be classified in the same species if they present an intra-group rate of gene flow higher than the rate between that group and any other strains. HGT is the main mechanism responsible for the spread of these genes within the bacterial population. HGT also allows exchange of genetic material between taxonomically distinct bacterial populations but on a smaller scale.[30] A hypothesis presented by Baumdicker et al[31] supports the concept proposed by Bobay and Ochman.[30] Baumdicker et al[31] argue that in the microbial world, individual bacterial cells maintain compact genomes whereas a higher number of genes exists at the population level. This idea was called the distributed genome hypothesis. The distributed genome of a group of bacteria can be accessed by calculating its pan-genome. The most recent methods used to characterize a pan-genome will be described in a later section.

Moldovan and Gelfand[32] proposed a new method for defining bacterial species using pan-genome data. The pan-genome can be represented by a gene frequency spectrum *G*(*k*) that correlates the number of orthologous genes groups (OGGs) containing genes from exactly *k* genomes. When the chart of the *G*(*k*) function presents a U-like shape, it is said that this set of genomes is homogeneous (Figure 3).[32] Moldovan and Gelfand then proposed that a population of isolates should be classified in the same species if they obey 3 criteria:

(a) Must be monophyletic in a sequence-based tree;
(b) Should be composed of a homogeneous set of genomes;
(c) Should be the maximal set of strains satisfying conditions a and b.

Ecological perspective and the pan-selectome hypothesis

Theoretically, a bacterial species whose population is highly clonal (closed pan-genome) is more successful in colonizing stable environments such as the human or animal tissues. In contrast, free-living microorganisms have a greater gene variability to adapt to different environmental conditions. The coagulase-negative staphylococci *Staphylococcus lugdunensis* is an example of commensal bacteria with closed pan-genome.[33] However, several other analyses demonstrated that this theory is not a rule.[34,35] The genome of the oncogenic *Helicobacter pylori* appears to be quite different depending on the geographical location of the isolate.[34] It is worth noting that Lapierre and Gogarten[18] demonstrated that the whole bacteria domain appears to have an open pan-genome. Therefore, it is difficult to define whether closed pan-genomes are true evidence of species with limited gene frequency or if they are only artifacts from analysis with a limited number of genomes.

The maintenance of gene frequency in a pan-genome has been subject of several studies. Rodriguez-Valera et al[36] raised the hypothesis that the pan-genome of a bacterial population is maintained and equalized by phage predation. Phages are more abundant than bacteria in several environments.[37] Many works analyze the relationship between microbial communities and abiotic factors. However, bacteria also need to adapt to biotic factors such as phage predation. A bacterial population under constant phage predation is also under constant modulation of its gene content. This process is called pan-selectome, and the pan-genome is a snapshot of the gene frequency in a given population under constant phage predation.[36] Subsequently, Rodriguez-Valera et al[36] postulated that this pan-selectome is the evolutionary unit of selection in the microbial world (therefore, at the genomic level the unit of selection is the pan-genome).

Bioinformatic Tools

The first step in a pan-genomic analysis is the homogenization of the genome annotation followed by gene clustering based in gene orthology and, finally, the curve fitting (Figure 4). In addition, some software performs phylogeny analyses based on core genome and single-nucleotide polymorphism (SNP) calling. The homogenization of the annotation avoids the wrong classification of core genes into the shared subset and shared genes assigned to singletons. It should be performed using genome annotation software like RAST[38] and Prokka.[39] Alternatively,
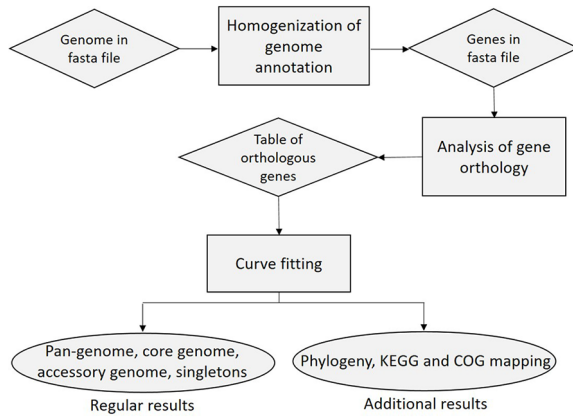
Figure 4. Flow diagram representing the main steps in a pan-genomic analysis. Each process (represented by blocks) can be performed by different methods. Table 1 details the methods used by different software. COG indicates Clusters of Orthologous Groups of proteins; KEGG, Kyoto Encyclopedia of Genes and Genomes.

the researcher may use refseq-annotated genomes from NCBI[40] or gene prediction using software such as FGENES[41] and GeneMark,[42] as long as the methodology is the same for all genomes being analyzed.

The clustering analysis is normally achieved by first performing an all-vs-all bidirectional BLAST analysis followed by the use of an orthology identification software, such as OrthoMCL[43] and Orthofinder.[44] Orthofinder is capable of eliminating gene length bias in orthogroup detection. Emms and Kelly[44] showed that Orthofinder performed 25% better than OrthoMCL. Another strategy is to use the results from BLAST to define the orthology directly from the size of the alignment and the identity of the sequence alignment, by setting a threshold for both. Also, a strategy described by Lerat performs the orthology identification by means of the score rate value (SRV), a normalization of the bit score from blast analyses. Afterward, as previously described, genes that are present in all strains are assigned to core genome, whereas genes that are shared by more than 2 but not all strains are assigned to shared genome and unique genes, that are only present in one strain, are assigned to singletons.[45]

The complete table with all the orthologous genes may then be used for pan-genome, core genome, and singletons development analyses, which will fit the specific curve generated from permutations of all genomes in all positions. Normally, the software performs curve fitting of the pan-genome using Heaps law or Power Law, whereas the curve fitting of shared genome and singletons are performed by means of exponential regression decay. The formula for the Heaps Law is

$$n = kN^{-\alpha}$$

where n is the number of genes, N is the number of genomes, and k and α are constants defined to fit the formula, whereas the formula for least square fit of the exponential regression decay is represented by

$$n = ke^{-x/\tau} + tg\theta$$

where n is the number of genes, x is the number of genomes, e is Euler number, and k, τ, and $tg\theta$ are constants defined to fit the formula.[46]

Interestingly, the α value is representative of the openness or closeness of the pan-genome, where an α lower than 1 is representative of an open pan-genome and an α higher than 1 is representative of a closed pan-genome. Also, $tg\theta$ represents the number of genes that will be found in the core genome after genome stabilization, whereas in singleton analysis, it represents the number of genes that will be added to the analysis for each newly added genome.[47]

BPGA[48] software uses USEARCH,[49] CD-HIT,[50] and OrthoMCL[43] software for the orthology analyses and power-law regression and exponential curve fit for the pan-genome and core genome developments (Table 1), respectively. It also implements other relevant analyses such as core/pan/MLST (Multi Locus Sequence Typing) phylogeny, subset analysis, and KEGG[51] & COG[52] mapping. EDGAR,[53] on the other hand, uses SRV analyses for the orthology identification followed by Heaps law and exponential curve fitting of pan-genome and core genome development analyses, respectively. The website also plots venn diagrams, allows the analyses of subgroups of genomes, and exports multiple sequence alignments for phylogeny analyses.

GET_HOMOLOGUES[26] uses bidirectional best-hit, COGtriangles[26] or OrthoMCL[43] for orthology analyses and performs pan-genome and core development analyses using the script plot_pancore_matrix.pl. The software also generates high-quality graphics, computes pan-genome trees, and performs syntenic cluster analyses. Pandelos[54] uses a dictionary-based method for orthology analyses and introduces a measure based on K-mer multiplicity and computation of Jaccard similarity. Panseq[55] uses MUMmer[56] and BLASTn[57] for the orthology analyses and it exports the accessory genome and binary presence/absence data, and core genome and SNPs for phylogeny analyses. PanX[58] uses Diamond[59] and MCL[60] for the Orthology analysis and it displays an alignment, a phylogenetic tree for each gene cluster, besides mapping mutation and inferring gain and loss of genes on the core genome phylogeny. It also has introduced a divide-and-conquer strategy for large datasets, by clustering small batches of genomes and combining different batches.

PGAP[61] uses Inparanoid, MultiParanoid, and Gene Family for the orthology assignment and Heap law and power law for the pan-genome and core genome developments. PGAP[61] is based in 5 modules: cluster, pan-genome, genetic variation, species evolution, and function enrichment analyses. PANWeb[62] and PGAWeb[63] are graphical interfaces for the use of PGAP.

PGAT[64] uses BLASTp for Orthology analysis with a sequence alignment of >80% and sequence identity >91% to

Table 1. Main software used for pan-genomic analysis and their respective algorithms.

SOFTWARE	ORTHOLOGY ANALYSIS	PAN-GENOME DEVELOPMENT	CORE GENOME DEVELOPMENT	REFERENCES
BPGA	USEARCH, CD-HIT and OrthoMCL	Power-law regression	Exponential curve fit	Chaudhari et al[48]
EDGAR and EDGAR 2.0	Score ratio values	Heaps' law	Exponential curve fit	Blom et al[53]
GET_HOMOLOGUES	Bidirectional best-hit, COGtriangles, or OrthoMCL	*plot_pancore_*matrix.pl	*plot_pancore_*matrix.pl	Contreras-Moreira and Vinuesa[26]
PanDelos	Dictionary-based method	—a	—a	Bonnici et al[54]
Panseq	MUMmer and BLASTn	—a	—a	Laing et al[55]
PanWeb	PGAP	PGAP	PGAP	Pantoja et al[62]
PanX	Diamond and MCL	—a	—a	Ding et al[58]
PGAP	Inparanoid, MultiParanoid and Gene Family	Heaps' law	Exponential curve fit	Zhao et al[61]
PGAT	BLASTp (sequence alignment of >80% and sequence identity >91%-92%)	—a	—a	Brittnacher et al[64]
PGAweb	PGAP and PGAP-x modules	PGAP and PGAP-x modules	PGAP and PGAP-x modules	Chen et al[63]
Piggyb	Roary	—a	—a	Thorpe et al[66]
Roary	CD-HIT, BLAST and MCL	—a	—a	Page et al[65]

aNot mentioned in manuscript.
bPan-genome analysis of intergenic regions.

92%. PGAT[47] also allows comparison of sequence polymorphisms, multigenome display of regions surrounding a query gene, comparisons of metabolic pathways, and manual community annotation. Roary[65] uses CD-HIT,[49] BLAST,[57] and MCL[60] for the orthology analyses. Roary[65] runs in thousands of genomes in pan-genome analyses in standard desktops. Also, it uses the context provided by conserved gene neighborhood information for orthology analyses. Piggy[66] emulates Roary,[65] but for intergenic regions (Table 1).

Pan-Genome Assessment of 9 Bacterial Species

Pan-genomic analyses become more accurate the greater the number of genomes used. Thus, the 9 bacterial species with the higher number of complete sequenced genomes were selected in GenBank and their pan-genomes were calculated using PGAP v.1.2.0.[61] Fifty genomes of each species were used, totaling 450 genomes analyzed (Table 2). The gene annotation was normalized submitting all fasta files of each genome to the RAST[38] server. The PGAP[61] parameters were 50% identity, 60% coverage, and an *e*-value of 0.00001. *E coli* is a bacterium found in the microbiome of warm-blooded animals and in environmental habitat. In our analysis, this species presented only 19% of its genes in the core genome (2290/11714 gene families) (Table 2). Rasko et al (2008)[67] have previously performed an analysis with 17 genomes and

found a pan-genome composed by 13 000 gene families. A large accessory genome is expected in *E coli* because this species is adapted to different habitat and this adaptation is directly related to the genome content. MGEs are the portions of a genome that significantly contributes to the diversity of gene families. Mobilome is the set of MGEs present in a genome. Plasmids and prophages are the main components of the mobilome in prokaryotes. The Shiga Toxin-producing *E coli* (STEC) has mobilome that comprises 19.8% of its genome.[68] This mobilome carries genes involved in virulence and resistance to antibiotics.[69]

Staphylococcus aureus, Listeria monocytogenes, and *Streptococcus pneumoniae* presented a pan-genome size of 5197, 5075, and 4404, respectively (Table 2). A total of 32% (1672), 41% (2114), and 26% (1152) of the gene families were present in the core genome of *S aureus, L monocytogenes*, and *S pneumoniae*, respectively. All 3 species are considered human pathogens. They presented an open pan-genome (Table 2). Bosi et al[70] analyzed 64 strains of *S aureus* and found a pan- and core-genome composed of 7457 and 1441 gene families, respectively.

Mycobacterium tuberculosis is a human pathogen and one of the biggest global threats to public health. A previous study analyzed 36 genomes of *M tuberculosis* and found a pan-genome composed by 5765 gene families being 3679

belonging to the core genome.[71] These values were very different from our results. However, in both cases, the analysis indicated an open pan-genome which reinforces the idea that there is some exchange of genetic material among *M tuberculosis* strains.[71] *Pseudomonas aeruginosa* is considered a metabolically versatile species with the ability to adapt to different habitat. Several other studies have assessed the pan-genome of this species.[17,72,73] Freschi et al[73] performed an analysis using 1311 genomes and observed that only 1% of the gene families were identified in the core genome. To the best of our knowledge, this is the smallest bacterial core genome described so far. We found a pan-genome containing 7850 gene families and a core genome corresponding to 36% of the entire pan-genome (Table 2). Comparing these results, the importance of evaluating a large number of genomes to achieve more accurate results in pan-genomic analysis becomes clear. *Campylobacter jejuni* presented the lowest number of gene families on its core genome (1211). This number represents 37% of the entire pan-genome that was characterized as open because the alpha value was lower than 1 (0.79) (Table 2). *Campylobacter jejuni* as well as *Bordetella pertussis* have a low number of published pan-genome analyses. However, they are pathogenic bacteria with the highest number of genomes available in databases. Tettelin et al[15] demonstrated that *B pertussis* have a closed pan-genome. In our analysis, *B pertussis* presented the highest alpha value; however, it was lower than 1. *B pertussis* also presented the highest core genome comprising 59% of the pan-genome (Table 2). Due to this low genomic plasticity, antibiotics and vaccines are quite effective against this species.[74]

Today, a basic step in genomics is the prediction and annotation of coding sequences (CDS) using bioinformatics. The need for homogenization of genome annotation was discussed in a previous section. Several software that use different methods for prediction of CDS were developed such as Glimmer3,[75] Prodigal,[76] or GeneMarkS-2.[42] Prediction of false-positive genes is common.[42] Theoretically, in a pan-genome analysis, even a small rate of prediction error can change the results significantly due to the high number of genomes analyzed. False positives are commonly predicted as CDSs encoding uncharacterized proteins. We determined the average percentage of uncharacterized proteins in the genomes of the 9 species (Table 2). *S aureus* and *L monocytogenes* presented the highest number of uncharacterized proteins: $26.4\% \pm 12.9\%$ and $25.4\% \pm 14.3\%$, respectively. The high value of standard deviation indicates that the genomes of some strains are better characterized than others (Table 2). We also calculated the pan-genome of those species removing the uncharacterized proteins from the genbank files. This process slightly altered the alpha value; however, all pan-genomes remained open ($\alpha > 1$) (Table 2).

Table 2. Genome, pan-genome, and core genome size of each species.

SPECIES	AVERAGE GENOME SIZE (BP)	PAN-GENOME SIZE (# GENE FAMILIES)	CORE GENOME SIZE (# GENE FAMILIES)	PERCENTAGE OF UNCHARACTERIZED PROTEINS	α VALUE (WITH UNCHARACTERIZED PROTEINS)	α VALUE (WITHOUT UNCHARACTERIZED PROTEINS)
Escherichia coli	5 139 550	11 614	2290	13.7% ± 0.3%	0.7393	0.7984
Pseudomonas aeruginosa	6 591 640	12 483	3200	22.7% ± 7.9%	0.7551	0.8253
Bordetella pertussis	4 104 300	4462	2635	18% ± 4.9%	0.8942	0.8885
Mycobacterium tuberculosis	4 383 450	7850	2868	20.9% ± 0.3%	0.745	0.7398
Klebsiella pneumoniae	5 585 920	10 476	3096	18.7% ± 5.2%	0.7687	0.8564
Campylobacter jejuni	1 698 660	3320	1211	16.5% ± 3.9%	0.7947	0.8323
Listeria monocytogenes	3 007 780	5075	2114	25.4% ± 14.3%	0.8312	0.845
Streptococcus pneumoniae	2 085 860	4404	1152	15.9% ± 8.1%	0.7777	0.8137
Staphylococcus aureus	2 837 870	5197	1672	26.4% ± 12.9%	0.7931	0.8296

Abbreviation: α, alpha value of Heap law.
The average percentage of uncharacterized proteins is also presented. To remove the uncharacterized proteins, we opened the genbank files in the genome browser Artemis[77] and using the feature selector tool we excluded all coding sequences that contained the names "uncharacterized protein" or "hypothetical protein" in their qualifier/product. The alpha value of each pan-genome analyzed with and without uncharacterized proteins is also shown.

Applications and Future Perspectives

A pan-genomic analysis presents all the gene variability of a group of organisms. The set of genes shared among all organisms as well as species- or strain-specific genes are also extremely useful information. All of these data allow an improvement of time and technology in different areas of biology and bioinformatics.

Comparison of several bacterial genomes is a valuable aid for reverse vaccinology analysis. Reverse vaccinology was a pioneering method first applied to the serogroup B meningococci.[78] The method has greater advantages over classic approaches of vaccine development because it is less laborious, less costly, and more accurate in choosing a gene target. Several reverse vaccinology studies have now used pan-genome to determine the main targets for vaccine development.[79,80]

The host-pathogen interaction can be evaluated at the genomic level through genes that are responsible for processes such as adhesion, invasion, and toxin production. Therefore, a pan-genome analysis helps to define which virulence genes are shared among all pathogenic species, as well as which genes are specific to one isolate. This has direct implications for understanding the evolution of pathogenic species.

In addition to the examples cited above, pan-genomic analysis has been increasingly used to assist in the taxonomic classification of microorganisms,[81] to determine a set of molecular markers for phylogenomic analysis,[82] among other applications. Therefore, determining the pan-genome of a group of organisms is sometimes the initial step of a research. Several downstream analyses depend on a good prediction of the pan-genome. Thus, knowledge about the basic concepts and the correct choice of software, algorithms, and parameters are extremely important to the success of the research.

Author Contributions

All authors contributed equally to the conception and writing of this manuscript. SSC performed the pan-genomic analysis.

ORCID iDs

Luís Carlos Guimarães (iD) https://orcid.org/0000-0002-7147-9448

Rafael Azevedo Baraúna (iD) https://orcid.org/0000-0001-6654-2118

REFERENCES

1. Tettelin H, Masignani V, Cieslewicz MJ, et al. Erratum. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc Natl Acad Sci USA*. 2005;102:16530. doi:10.1073/pnas.0508532102.
2. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High throughput sequencing: an overview of sequencing chemistry. *Indian J Microbiol*. 2016;56:394-404. doi:10.1007/s12088-016-0606-4.
3. Plöthner M, Frank M, von der Schulenburg JMG. Cost analysis of whole genome sequencing in German clinical practice. *Eur J Heal Econ*. 2017;18:623-633. doi:10.1007/s10198-016-0815-0.
4. Veras A, Araujo F, Pinheiro K, et al. Pan4Draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes. *Sci Rep*. 2018;8:9670. doi:10.1038/s41598-018-27800-8.

5. Zeng LB, Wang D, Hu NY, et al. A novel pan-genome reverse vaccinology approach employing a negative-selection strategy for screening surface-exposed antigens against leptospirosis. *Front Microbiol*. 2017;8:396. doi:10.3389/fmicb.2017.00396.
6. Doron S, Melamed S, Ofir G, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*. 2018;359:eaar4120. doi:10.1126/science.aar4120.
7. Chaves-Moreno D, Wos-Oxley ML, Jáuregui R, Medina E, Oxley APA, Pieper DH. Application of a novel "pan-genome"-based strategy for assigning RNAseq transcript reads to *Staphylococcus aureus* strains. *PLoS ONE*. 2015;10:e0145861. doi:10.1371/journal.pone.0145861.
8. Lebreton F, Manson AL, Saavedra JT, Straub TJ, Earl AM, Gilmore MS. Tracing the enterococci from paleozoic origins to the hospital. *Cell*. 2017;169:849.e13-861.e13. doi:10.1016/j.cell.2017.04.027.
9. Ogier JC, Calteau A, Forst S, et al. Units of plasticity in bacterial genomes: new insight from the comparative genomics of two bacteria interacting with invertebrates, Photorhabdus and Xenorhabdus. *BMC Genomics*. 2010;11:568. doi:10.1186/1471-2164-11-568.
10. Sivashankari S, Shanmughavel P. Comparative genomics—a perspective. *Bioinformation*. 2007;1:376-378. doi:10.6026/97320630001376.
11. Wayne LG. International Committee on Systematic Bacteriology: announcement of the report of the Ad Hoc committee on reconciliation of approaches to bacterial systematics. *Syst Appl Microbiol*. 1988;10:99-100. doi:10.1016/S0723-2020(88)80020-1.
12. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol*. 2014;64:346-351. doi:10.1099/ijs.0.059774-0.
13. Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*. 2018;36:996-1004. doi:10.1038/nbt.4229.
14. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pangenome. *Curr Opin Genet Dev*. 2005;15:589-594. doi:10.1016/j.gde.2005.09.006.
15. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 2008;11:472-477. doi:10.1016/j.mib.2008.09.006.
16. Donati C, Hiller NL, Tettelin H, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol*. 2010;11:R107. doi:10.1186/gb-2010-11-10-r107.
17. Mosquera-Rendón J, Rada-Bravo AM, Cárdenas-Brito S, Corredor M, Restrepo-Pineda E, Benítez-Páez A. Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. *BMC Genomics*. 2016;17:45. doi:10.1186/s12864-016-2364-4.
18. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. *Trends Genet*. 2009;25:107-110. doi:10.1016/j.tig.2008.12.004.
19. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res*. 2001;11:555-565. doi:10.1101/gr.GR-1660R.
20. Kuenne C, Billion A, Mraheil MA, et al. Reassessment of the *Listeria monocytogenes* pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. *BMC Genomics*. 2013;14:47. doi:10.1186/1471-2164-14-47.
21. Belbahri L, Bouket AC, Rekik I, et al. Comparative genomics of *Bacillus amyloliquefaciens* strains reveals a core genome with traits for habitat adaptation and a secondary metabolites rich accessory genome. *Front Microbiol*. 2017;8:1438. doi:10.3389/fmicb.2017.01438.
22. Paquola ACM, Asif H, Pereira CAB, et al. Horizontal gene transfer building prokaryote genomes: genes related to exchange between cell and environment are frequently transferred. *J Mol Evol*. 2018;86:190-203. doi:10.1007/s00239-018-9836-x.
23. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol*. 2012;8:e1002514. doi:10.1371/journal.pcbi.1002514.
24. Chen X, Zhang J. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput Biol*. 2012;8:e1002784. doi:10.1371/journal.pcbi.1002784.
25. Gabaldón T, Martin T, Marcet-Houben M, et al. Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics*. 2013;14:623. doi:10.1186/1471-2164-14-623.
26. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol*. 2013;79:7696-7701. doi:10.1128/AEM.02411-13.
27. De Queiroz K. Ernst Mayr and the modern concept of species. *Proc Natl Acad Sci USA*. 2005;102:6600-6607.
28. Stackebrandt E, Frederiksen W, Garrity GM, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol*. 2002;52:1043-1047. doi:10.1099/00207713-52-3-1043.

29. Blaxter M, Mann J, Chapman T, et al. Defining operational taxonomic units using DNA barcode data. *Philos Trans R Soc B Biol Sci*. 2005;360:1935-1943. doi:10.1098/rstb.2005.1725.

30. Bobay L-M, Ochman H. Biological species are universal across life's domains. *Genome Biol Evol*. 2017;9:491-501. doi:10.1093/gbe/evx026.

31. Baumdicker F, Hess WR, Pfaffelhuber P. The infinitely many genes model for the distributed genome of bacteria. *Genome Biol Evol*. 2012;4:443-456. doi:10.1093/gbe/evs016.

32. Moldovan MA, Gelfand MS. Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus* spp. *Front Microbiol*. 2018;9:428. doi:10.3389/fmicb.2018.00428.

33. Argemi X, Matelska D, Ginalski K, et al. Comparative genomic analysis of *Staphylococcus lugdunensis* shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC Genomics*. 2018;19:621. doi:10.1186/s12864-018-4978-1.

34. Kawai M, Furuta Y, Yahara K, et al. Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC Microbiol*. 2011;11:104. doi:10.1186/1471-2180-11-104.

35. Lu QF, Cao DM, Su LL, et al. Genus-wide comparative genomics analysis of *Neisseria* to identify new genes associated with pathogenicity and niche adaptation of *Neisseria* pathogens. *Int J Genomics*. 2019;2019:6015730. doi:10.1155/2019/6015730.

36. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, et al. Explaining microbial population genomics through phage predation. *Nat Rev Microbiol*. 2009;7:828-836. doi:10.1038/nrmicro2235.

37. Breitbart M. Marine viruses: truth or dare. *Ann Rev Mar Sci*. 2012;4:425-448. doi:10.1146/annurev-marine-120709-142805.

38. Aziz RK, Bartels D, Best A, et al. The RAST server: rapid annotations using subsystems technology. *BMC Genomics*. 2008;9:75. doi:10.1186/1471-2164-9-75.

39. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068-2069. doi:10.1093/bioinformatics/btu153.

40. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44:D733-D745. doi:10.1093/nar/gkv1189.

41. Rogic S, Mackworth AK, Ouellette FBF. Evaluation of gene-finding programs on mammalian sequences. *Genome Res*. 2001;11:817-832. doi:10.1101/gr.147901.

42. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res*. 2018;28:1079-1089. doi:10.1101/gr.230615.117.

43. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13:2178-2189. doi:10.1101/gr.1224503. candidates.

44. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16:157. doi:10.1186/s13059-015-0721-2.

45. Lerat E, Daubin V, Moran NA. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria. *PLoS Biol*. 2003;1:101-109. doi:10.1371/journal.pbio.0000019.

46. Trost E, Al-Dilaimi A, Papavasiliou P, et al. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. *BMC Genomics*. 2011;12:383. doi:10.1186/1471-2164-12-383.

47. Soares SC, Silva A, Trost E, et al. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the Biovar ovis and equi strains. *PLoS ONE*. 2013;8:e53818. doi:10.1371/journal.pone.0053818.

48. Chaudhari NM, Gupta VK, Dutta C. BPGA—an ultra-fast pan-genome analysis pipeline. *Sci Rep*. 2016;6:24373. doi:10.1038/srep24373.

49. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26:2460-2461. doi:10.1093/bioinformatics/btq461.

50. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:3150-3152. doi:10.1093/bioinformatics/bts565.

51. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45:D353-D361. doi:10.1093/nar/gkw1092.

52. Tatusov RL. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 2001;29:22-28. doi:10.1093/nar/29.1.22.

53. Blom J, Albaum SP, Doppmeier D, et al. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*. 2009;10:154. doi:10.1186/1471-2105-10-154.

54. Bonnici V, Giugno R, Manca V. PanDelos: a dictionary-based method for pan-genome content discovery. *BMC Bioinformatics*. 2018;19:437. doi:10.1186/s12859-018-2417-6.

55. Laing C, Buchanan C, Taboada EN, et al. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics*. 2010;11:461. doi:10.1186/1471-2105-11-461.

56. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018;14:e1005944. doi:10.1371/journal.pcbi.1005944.

57. Yamada M. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Fish Sci*. 2018;84:3389-3402. doi:10.1007/s10924-010-0204-1.

58. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. *Nucleic Acids Res*. 2018;46:e5. doi:10.1093/nar/gkx977.

59. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2014;12:59-60. doi:10.1038/nmeth.3176.

60. Enright AJ. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30:1575-1584. doi:10.1093/nar/30.7.1575.

61. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: pan-genomes analysis pipeline. *Bioinformatics*. 2012;28:416-418. doi:10.1093/bioinformatics/btr655.

62. Pantoja Y, Pinheiro K, Veras A, et al. PanWeb: a web interface for pan-genomic analysis. *PLoS ONE*. 2017;12:e0178154. doi:10.1371/journal.pone.0178154.

63. Chen X, Zhang Y, Zhang Z, et al. PGAweb: a web server for bacterial pan-genome analysis. *Front Microbiol*. 2018;9:1910. doi:10.3389/fmicb.2018.01910.

64. Brittnacher MJ, Fong C, Hayden HS, Jacobs MA, Radey M, Rohmer L. PGAT: a multistrain analysis resource for microbial genomes. *Bioinformatics*. 2011;27:2429-2430. doi:10.1093/bioinformatics/btr418.

65. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31:3691-3693. doi:10.1093/bioinformatics/btv421.

66. Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *GigaScience*. 2018;7:giy015. doi:10.1093/gigascience/giy015.

67. Rasko DA, Rosovitz MJ, Myers GSA, et al. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol*. 2008;190:6881–6893. doi: 10.1128/JB.00619-08.

68. Delannoy S, Mariani-Kurkdjian P, Webb HE, Bonacorsi S, Fach P. The mobilome; a major contributor to *Escherichia coli* stx2-positive O26: H11 strains intra-serotype diversity. *Front Microbiol*. 2017;8:1625. doi:10.3389/fmicb.2017.01625.

69. Mbelle NM, Feldman C, Osei Sekyere J, Maningi NE, Modipane L, Essack SY. Publisher correction. The resistome, mobilome, virulome and phylogenomics of multidrug-resistant *Escherichia coli* clinical isolates from Pretoria, South Africa. *Sci Rep*. 2020;10:1270. doi:10.1038/s41598-020-58160-x.

70. Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson B. Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc Natl Acad Sci USA*. 2016;113:E3801-E3809. doi:10.1073/pnas.1523199113.

71. Yang T, Zhong J, Zhang J, et al. Pan-genomic study of *Mycobacterium tuberculosis* reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations. *Front Microbiol*. 2018;9:1886. doi:10.3389/fmicb.2018.01886.

72. Valot B, Guyeux C, Rolland JY, Mazouzi K, Bertrand X, Hocquet D. What it takes to be a *Pseudomonas aeruginosa*? The core genome of the opportunistic pathogen updated. *PLoS ONE*. 2015;10:e0126468. doi:10.1371/journal.pone.0126468.

73. Freschi L, Vincent AT, Jeukens J, et al. The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biol Evol*. 2019;11:109-120. doi:10.1093/gbe/evy259.

74. Carbonetti NH. Pertussis leukocytosis: mechanisms, clinical relevance and treatment. *Pathog Dis*. 2016;74:ftw087. doi:10.1093/femspd/ftw087.

75. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. 2007;23:673-679. doi:10.1093/bioinformatics/btm009.

76. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119. doi:10.1186/1471-2105-11-119.

77. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*. 2012;28:464-469. doi:10.1093/bioinformatics/btr703.

78. Pizza M, Scarlato V, Masignani V, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*. 2000;287:1816-1820. doi:10.1126/science.287.5459.1816.

79. Hisham Y, Ashhab Y. Identification of cross-protective potential antigens against pathogenic *Brucella* spp. through combining pan-genome analysis with reverse vaccinology. *J Immunol Res*. 2018;2018:1474517. doi:10.1155/2018/1474517.

80. Seib KL, Zhao X, Rappuoli R. Developing vaccines in the era of genomics: a decade of reverse vaccinology. *Clin Microbiol Infect*. 2012;18:109-116. doi:10.1111/j.1469-0691.2012.03939.x.

81. Caputo A, Fournier PE, Raoult D. Genome and pan-genome analysis to classify emerging bacteria. *Biol Direct*. 2019;14:5. doi:10.1186/s13062-019-0234-0.

82. Velsko IM, Perez MS, Richards VP. Resolving phylogenetic relationships for *Streptococcus mitis* and *Streptococcus oralis* through core- and pan-genome analyses. *Genome Biol Evol*. 2019;11:1077-1087. doi:10.1093/gbe/evz049.