Received: 18 January 2024 | Revised: 24 May 2024 | Accepted: 5 June 2024 | Published online: 12 June 2024

REVIEW

Medinformatics
yyyy, Vol. XX(XX) 1–5
DOI: 10.47852/bonviewMEDIN42022496

Bacterial Pangenome: A Review on the Current Strategies, Tools and Applications

Nikita Chordia Golchha[1], Anand Nighojkar[2]and Sadhana Nighojkar[3,*]

1 School of Biotechnology, Devi Ahilya University, India. nikita.chordia25@gmail.com, ORCID: https://orcid.org/0000-0002-1149-1090
2 Maharaja Ranjit Singh College of Professional Sciences, India. nanandphd@gmail.com, ORCID: https://orcid.org/0000-0002-9072-7259
3 Mata Gujri College of Professional Studies, India. sadhana0301@gmail.com, ORCID: https://orcid.org/0000-0001-8389-5919

***Corresponding author:** Sadhana Nighojkar, Mata Gujri College of Professional Studies, India. sadhana0301@gmail.com

Abstract: The genomic research has grown by the advancement of sequencing technologies, which aim to capture the "totality" of life diversity by progressing from single strain to species or even higher taxa. Out of this scientific progress and inquisitiveness, the term "pangenome" was created. The goal of a pangenome is to represent all of a species genetic variation. It divides the genome into core, accessory and unique genes. Since the introduction of pangenomics, many software programs have been actively used for pangenome analysis. This review focuses on providing a snapshot of useful tools with their use, execution and computational details. In addition to tools for pangenome, the components, approaches, mechanism of variation, applications and challenges associated with pangenome analysis are also discussed in this article. Here, we have also discussed the increasing trend of bacterial pangenome studies with few studies and their significant findings. It has been proved that pangenomics is useful for understanding the clade of bacteria that helps in developing strategies for various application based on biological similarities and differences.

Keywords: pangenome, bacteria, genome, variation, software, gene duplication

1. Introduction

From the discovery of first sequencing technology till now, many organisms have their complete gene repertoire, creating the era of genomics. The breakthrough in sequencing technologies reduces the cost and time for sequencing which leads to the flood of genome sequences in the public databases. According to the statistics of the Genome OnLine Database (GOLD), more than 470,000 genome sequencing projects are registered from which more than 210,000 are bacterial genomes (https://gold.jgi.doe.gov/statistics) as visited on December 2023 [1]. Bacteria are easily sequenced because they are widely distributed, easy to isolate and their genomes are small, compact and does not contain repetitive sequences. In addition, bacterial genome sequences have vast application in development of vaccines, new therapeutics, identification of antibiotic resistance genes, virulence factors and industrially important products [2, 3].

Studies on comparative genomics are enhanced by the availability of sequenced genomes. The complete genome of one organism is compared against other organisms for better understanding of evolution, gene function, conserved regions, non-coding genes, genome design, gene regulation, biochemical, metabolic and physicochemical pathways [4]. But comparing genome with a single reference genome is found to be inadequate in cataloguing the entire genes of the species. These inadequacy can be overcome by sequencing multiple genomes of species that results in the introduction of pangenome concept [5].

Sigaux coined the term "pangenome" in 2000 to create a public database that contains pangenome computation of major types of tumor genomes and its transcriptome alterations [6]. Later in 2005, Tettelin and co-workers proposed the concept of a pangenome and depicted the pangenome of *Streptococcus agalactiae* [group B *Streptococcus* (GBS) [5]. It is the very first pangenome, and since then, the idea of a pangenome has gained popularity. Nowadays, there are many examples for bacteria [7-9] and other higher organisms including yeast [10], maize [11] and goat [12].

A single reference genome is inappropriate for representing the genetic diversity of the species because of different structural variants. It has been observed that differences in gene copy number and the presence or absence of specific genes

generate changes in genomic content in individuals [13]. Duplications, insertions, and deletions caused by variations in gene copy number result in altered gene expression [14]. Numerous other genes' transcription can be impacted by a gene's presence or absence. [15]. Any variation in the copy number of gene and presence/absence of gene has been studied for various organisms which include both eukaryotes and prokaryotes. It is reported that many genes are evolved due to gene duplication; after which they rapidly adapt and exhibit novel functions [16]. When enough mutations accumulate in a gene, the gene may occasionally cease to function. Gene conservation is another potential evolutionary result of duplicate genes. Two approaches are used for accomplishing this: filtering selection, which prevents duplicates from diverging, or cooperative evolution, in which recurrent gene conversion preserves sequence identity within the genome. [17].

To capture the full information on variation, the use of single reference genome is not suitable. As the number of sequenced genomes is steadily increasing, pangenome has been introduced. It takes the genome in many alternative forms that can make sense in different context. The alternative forms of pangenome can be:

- a consensus genome derived from all individuals of a species
- a minimum ‘functional’ genome that is required for the survival
- a genome of selected individual
- a maximum genome that comprises every sequence ever reported

These alternative forms have rapidly widened because of the comprehensive identification of several individual genome sequences from specific populations, taxonomic groups, or habitats using modern sequencing technologies [18]. This demands a radical shift from the use of a single reference genome towards the use of a "pangenome", which reflects the entirety of the genomic material within a given species or phylogenetic group.

2. Computational Pangenomics

Pangenome is an established concept that refers to any combination of genetic sequences that will be combined for analysis or used as a reference. In other words, a microbial pangenome is the amalgamation of genes present in all strains (core genome) and genes absent from one or more of the strains (accessory or flexible genome). Pangenome analysis provides a scaffold for estimating the genomic diversity and estimating the number of additional whole genome sequences that will be required to adequately define that diversity [19]. Efficient data structures and algorithms give rise to ‘computational pangenomics’. It has offered a simple framework for classifying, storing, and comprehending genetic data. Additionally, it has made it possible for sophisticated two-dimensional genome browsers to be developed instead of linear text [20, 21].

3. Components of Pangenome

In the era of fast sequencing technologies, a plenty of genome sequences are accessible for different isolates of the same species which has provided an easy way of investigating several genomic characteristics. It is done by pangenome analysis which compares the genome of many species within a single genus, or even distinct strain within a single species. Pangenome represents the total number of genes that are present in a specific dataset and are not redundant [22]. It generally consists of three components: the core genome, the accessory or dispensable genome, and the species- or strain-specific gene. The core genome is made up of genes that are shared by all genomes and are usually involved in essential biological processes. Genes found in a small number of isolates make up the accessory or dispensable genome, and species- or strain-specific genes are found only in a specific genome [23]. Genes of core genome are present in all species and are important for the survival and growth. They are associated with the preservation of life's fundamental elements and are primarily connected to maintenance of cellular homeostasis, replication, and translation. A larger core genome and more shared genes are expected to be found in organisms with phylogenetically linked genomes. In contrast, diverse organisms show smaller core genome [24]. The genes shared by some organisms, or those that are present in some of the investigated organisms, are referred to as the accessory or dispensable genome. It has been understood to be alterations or variations in gene sequences that may result in the emergence of novel functions. This is accomplished by paraphyletic evolution and horizontal gene transfer, which result from gene duplication and mutation [25]. As the name suggests, species specific genes work at the inter species level and are present in a single species, whereas strain-specific genes are at intra-species level and are only present in one strain. They are mainly involved in niche adaptation and are emerged through horizontal gene transfer. Figure 1 represents all three components of pangenome.

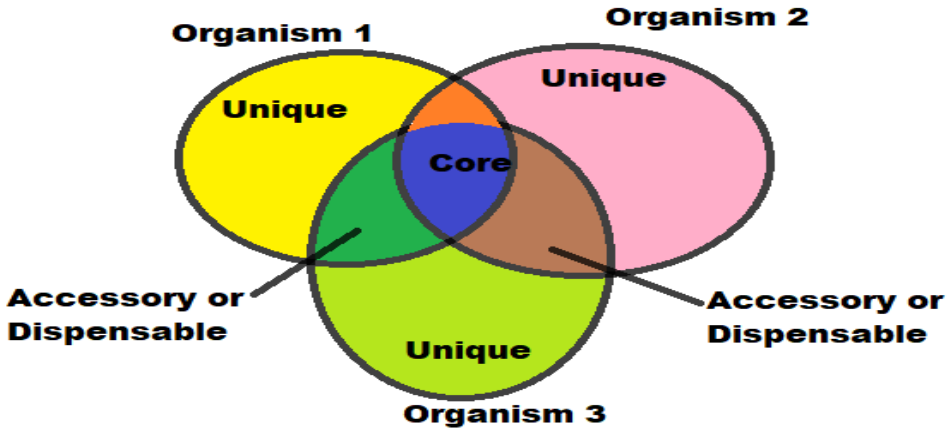

Figure 1. Venn diagram representing the components of pangenome

4. Open and Closed Pangenome

Depending on the size of pangenome, a pangenome can be of two types: open or closed. This classification is based on the proportionality among gene gain and loss events. When a pangenome is open, the number of genes increases along with the number of strains that have been further sequenced. Sympatric species typically have large genomes and an open pangenome since they live in communities. They have many ribosomal operons and a high rate of horizontal gene transfer. [26]. In simple words, species that can survive in many environments having mixed microbial communities have various methods of genetic material exchange and therefore, continually expand their repertoire of genes and have open pangenome. Since *Legionella pneumophila* is intracellular and coexists with other microorganisms within amoeba, it has an open pangenome [27].

Allopatric species with restricted access to the global microbial gene pool, such as those that occupy isolated habitats, have closed pangenomes. For those species, pangenome cannot increase with the increase in sequenced strains. Their pangenome is covered even by small number of sequenced strains. *Bacillus anthracis* and *Yersinia pestis* are soil-dwelling organisms with closed pangenomes that exhibit an allopatric way of lifestyle. *B. anthracis* divides in its host and exists in the soil as a dormant sporulated form. It has very few possiblity to exchange genes, hence it is having closed pangenome [28]. Figure 2 shows the subgroups of open and closed pangenome through a venn diagram.

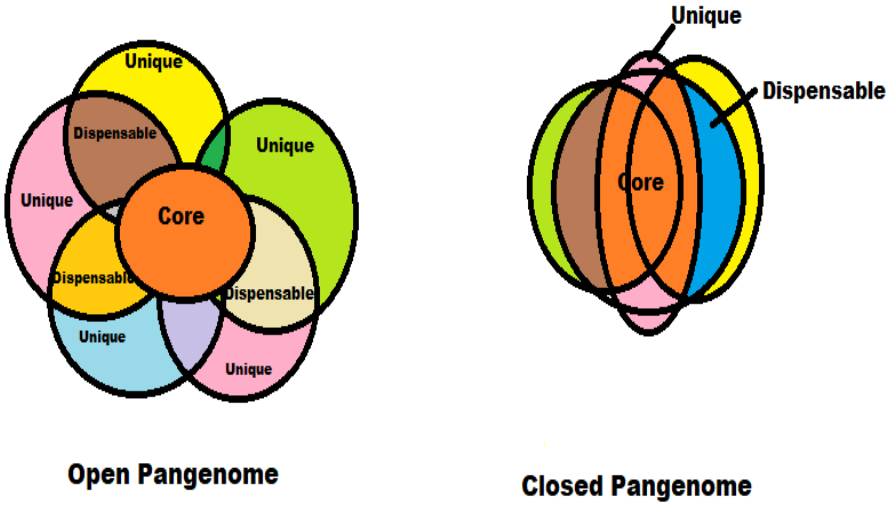

Figure 2. Venn Diagram representing an open and closed pangenome.

5. Mechanism of Pangenome Variation

One of the important part of pangenome studies is to identify the different sources of new genes. A bacteria's accessory genome's diversity and variability relies on the nature and frequency of horizontal gene transfer (HGT). In addition to HGT, a

substantial portion of novel genes also emerge from DNA duplication, followed by sequence divergence [29]. Modes of HGT include transformation, transduction and conjugation. In transformation, the exogenous DNA is being uptaken and integrated from the environment. Plasmid transmission has long been known to result in the transfer of antibiotic resistance among clinical isolates, and many bacterial species are naturally capable of incorporating DNA from the environment. The second mode of HGT is conjugation, in which DNA is directly transferred through a pilus between two bacterial cells, which mainly involves transposons and plasmids. The third route of HGT is transduction and is considered to be a major cause of HGT and is fundamental for gene innovation [30]. In transduction, bacterial DNA is transferred by the virus, by enclosement in bacteriophage capsids and enters a new host cell. It is especially involved in transferring virulence factors and invasion-related functions [31]. Bacteria differ in using each of these processes for HGT among their populations and in their mechanisms for DNA uptake.

Finally, it has been demonstrated that bacterial phenotypes can get altered as a result of both gene expansions and deletions. Studies have demonstrated that when specific genes are knocked off in *E. coli*, additional mechanisms which aid the bacteria's virulence are activated. And when the gene is restored by experimental insertion, the bacteria show non virulent phenotype. Thus, it can be concluded that the different traits of strains within a population should be considered the outcome of novel functionalities coded by additional genes, as well as the result of a regulatory network that can be changed by a smaller gene set and by different gene interactions. [32].

6. Approaches for Pangenome Studies

6.1. Network reconstruction and flux balance analysis

Genome-scale network reconstructions provide a possibility for computing species phenotypic characteristics. A high-quality reconstruction can be used to map information from a reference strain to other strains of the target organism. In network reconstruction, information from annotated genes are encoded to their biological products and their interactions. Numerous biological functions, including replication, repair, control of transcription and translation, expression, and metabolism, are subject to these reconstructions. It is made up of all known metabolic genes, the proteins they encode, and the reactions they catalyse. The data required for the network reconstructions is synthesized by using various accumulations of primary literature, organism-specific databases and high-throughput data. They are auto - updated based on the latest research via technology advancement [33].

Biological processes organized in network are assembled into pathways, pathways into individual system and finally into genome-scale networks. This represents biological processes at multiple levels. Metabolic reconstructions are more informative when they are transformed into a mathematical representation using computational interrogation. This mathematical model can then be used for flux balance analysis to predict organism phenotype [34]. It anticipates a wide range of phenotypes, including metabolite secretions, the impact of gene knockouts, and the ability to grow on various substrates. Using this approach, phenotype of variety of sequenced strains within a species can now be calculated with much more ease.

6.2 Multiple strain sequencing

One of the method for prokaryotic pangenome studies is to sequence many strains of a particular species. After sequencing, the variability of phenotypes at the genetic level is identified using comparative genomics integrated with genome-scale modeling. This technique is useful to understand strain-specific metabolic diversity and to know the evolutionary trajectories of a bacterial species. In this method, life style diversity can be predicted by auxotrophic study [35]. Multiple strain approach was used to determine functional differences and common metabolic capabilities between different strains of *E. coli* species. They have expanded their research to include 1122 *E. coli* strains and found that variations in gene products result from substantial genetic differences between species. [36].

Numerous studies on various organisms were done, the first study being on *E. coli*. In 2016, Fouts et al. applied this multi-strain sequencing approach on *Leptospira* to demonstrate that the pathogenic species only possess the capacity to produce vitamin B12. This may be because the body produces vitamin B12 in a human host, giving them a survival advantage [37]. In addition, this approach is applied to many species of bacteria as *Salmonella serovars* [38], *Staphylococcus aureus* [39] and many more.

6.3 Hybridization based techniques

Numerous studies on the genomic level of bacteria reveal how these organisms evolved under various ecological conditions and how they adapted to specialised niches. One of the powerful approach to study these processes is microarray. In microaaray, a chip is prepared by spotting ORFs of the representative species sequence that serve as a probe. This chip can be used to identify a strain with an unknown gene composition or for polymorphisms, where variations in the ratio of fluorescence emitted can be used to identify gene deletions and insertions. [40]. By using this hybridization technique, it is possible to compare strains within the same or related species or identify genetic differences even between different bacterial species. [41].

However, this technology has one adverse limitation that sequences are always determined in relation to the reference genome, whether they are present or not. As a result, unique genes on the strains that were studied cannot be found. Suppression subtractive hybridization (SSH) provides an effective solution to this issue. In SSH, two closely related DNA are distinguished

based on PCR technology. DNA from test strain is hybridized with control (reference) strain; the sequences which are hybridized are removed. The remaining DNA contains unique genes of test strain, which are then cloned and/or sequenced. This method is applied to plant pathogen *Serratia marcescens* [42].

7. Impact of SI and CG Thresholds in Pangenome Analysis

In pangenomics, the two parameters SI (sequence identity) and CG (cytosine guanine) thresholds are very important parameters that affects pangenome analysis results. SI is used to define homology; for example, two genes are considered homologous if they share at least 90% of their sequences. If we change this threshold, then the size of the core genes would also change. Several alignment threshold values are employed for analysis in pangenome investigations. For instance, Tettelin used the 50/50 rule to identify conserved genes and/or proteins in various species' genomes if the sequences shared at least 50% of the total length of the gene or protein, whereas Hiller et al.[43] used a more stringent threshold of 70% sequence identity/sequence length.

Similarly, the CG threshold of 100% indicates that CGs must be present in 100% of genomes, and a CG threshold of 90% indicates that CGs must be present in more than 90% of genomes. Different CG thresholds were compared utilizing Roary, BPGA, and Anvi'tools at 100%, 99%, 98%, 95%, 92%, and 90% in a study conducted by Li and Yin in 2022 [44]. Likewise, Roary alone was used to evaluate various SI thresholds at 95%, 90%, 85%, and 80%. They came to the conclusion that in order to reduce core genome loss and enhance gene functional enrichment analysis, a lower CG criterion (99% or 95%) should be applied.

8. Tools and Software Used in Pangenome Analysis

- **Prokaryotic Genome Analysis Tool (PGAT):** PGAT is a web-based database software for comparative analysis of closely related bacterial genomes. It analyzes the gene content and sequences of several microbial genomes. Its novel algorithm is being used for analysis of pangenome and synteny, comparison of genes in metabolic pathways, identification of genes in a dataset, comparison of SNPs (single-nucleotide polymorphism) in orthologous genes and improvement of functional annotation. Based on these ortholog mappings, an online interface makes it easier to search for and analyse the genetic similarities and differences across the various bacterial strains [45].

- **Integrated Toolkit for Exploration of Microbial Pangenomes (ITEP):** It is a collection of Python and BASH programmes that interacts with a SQLite database. It contains a large number of existing programs for sequence alignment, metabolic analysis, clustering analysis and protein prediction to assemble and interpret genomic content across related genomes. Numerous features are available through ITEP, including cross-genome analysis integration, determining core and variable genes, de novo protein family prediction, evaluation of functional domains, gene region identification, sequence alignments and tree generation, annotation curation, and metabolic networks for the study of their evolution. [46].

- **EDGAR (Efficient Database Framework for Comparative Genome Analyses Using BLAST Score Ratios):** It uses high throughput approach to automatically perform genome comparisons. It also provides precomputed database of various pangenome and visualization tools like Venn diagrams or synteny plots. Data from the NCBI genomes database were used in the initial release of EDGAR in 2009, which includes a comparative examination of 582 genomes from 75 genus groups [47]. The EDGAR platform has undergone continual development and advancement. In EDGAR 2.0, the total number of projects that are accessible to the public has grown to 167 genera and 2160 genomes with phylogenetic and taxonomic analysis [48]. Now soon EDGAR 3.0 will be available.

- **ClustAGE:** The BLAST technique is used to analyse the accessory genomes of a group of bacterial isolates. A set of nucleotide sequences (AGEs) of accessory genomic elements is grouped in order to estimate the minimum set of AGEs in the population as well as the distribution of each AGE among the genomes. It efficiently aids in understanding the role that accessory genomes play in a species' evolution, adaptability, and pathogenesis. [49].

- **GET_HOMOLOGUES:** It was created in Perl and R to analyse bacterial strains' comparative genomes and pangenomes. It can use OrthoMCL, bidirectional best-hit clustering, COGtriangles, or other clustering algorithms to organise homologous gene families. It is a fully automatic analysis pipeline that is highly customizable. This program uses BLAST+ and HMMER tool to build groups of orthologous genes. After this, sequences, intergenes and its features are extracted, indexed and sorted [50].

- **Micropan:** It is written in R and its package is mainly for the study of prokaryotic pangenomes with a number of functions. It is openly accessible from The Comprehensive R Archive Network (CRAN). R is used as the working environment for the implementation of several options of pangenome models, statistical analysis and graphics. However, it transfers complex calculations to external software e.g., HMMER3. A data set for the gram-positive bacterium *Enterococcus faecalis* has been benchmarked. [51].

- **PanCake:** It is exclusively based on sequence data and pairwise alignments. It has a command line interface that has sub-commands to add sequences, group sequences (i.e. genomes), annotate the results of pairwise alignments computed by BLAST or nucmer, identify the core regions, and singleton regions and retrieve chromosome sequences. Benchmarking of the tool has been performed on three genera, namely *Yersinia*, *Burkholderia* and *Pseudomonas*[52].

- **Pangenome Analysis Pipeline (PGAP and PGAP-X):** In 2012, PGAP was developed in Perl language that works on the Linux platform. One of the exciting features of PGAP is that by using only one command, it executes five analytical tasks. These include analysis of clusters, their functions enrichment, pangenome profiles, genetic variation of functional genes, and species evolution [53].
  A PGAP-X extension has been created in order to interpret and illustrate the results in an effective manner. Its benchmarking has been done on datasets of *Chlamydia trachomatis* and *S. pneumoniae*[54].

- **PanFunPro:** PanFunPro is PAN-genome Analysis Based on FUNctional PROfiles. It uses the functional domain information to classify homologous proteins into families by combining domains from three different Hidden Markov Models (HMM) collections. Additionally, it calculates the pan/core genome and analyses the GO data. The four steps of PanFunPro [55] are :
  - Step 1 –Selection of genome
  - Step 2 –Prediction of functional domains
  - Step 3 – Protein groupings and construction of functional profiles
  - Step 4: Analysis of pan, core and accessory genomes

- **SplitMEM:** It directly plots the compressed de Bruijn graph against the total number of genomes of interest in terms of time and space. It generates compressed de Bruijn pangenome graphs and refers suffix trees (for the genomes). Nodes in a graph represent conserved or strain-specific pangenome sequences. Its benchmarking reveals the core-genome properties that has been demonstrated on 62 strains of *E. coli* and 9 strains of *Bacillus anthracis* datasets [56].

- **Panakeia:** It is a novel tool based on Python for Pangenome analysis and visualization. It uses clustering approach to cluster the proteins and using their paralogs and synteny information, it generates the pangenome graph. It works closely with Pantagruel pipeline that provides the phylogenetic and evolutionary information. This combination advances our knowledge of the structure, adaptation, and evolution of microbial genomes. This pipeline reduces the manual work as it extracts and highlights the important features of the genome without finding them in long lists, tables and huge graphs [57].

- **BPGA:** BPGA (Bacterial Pan Genome Analysis tool) is written in Perl and can be executed in both windows and linux. It performs Pangenome analysis in stepwise fashion from clustering of input sequences followed by matrix generation and running seven function modules (analysis of pangenome profile, phylogeny creation, analysis of function and pathway, pangenome statistics, atypical GC content analysis, subset analysis and exclusive absent gene analysis). BPGA is integrated with various clustering method (USEARCH, CD-HIT and OrthoMCL), minimum running prerequisites, user friendly command line and high quality output [58].

- **PATO:** PATO (Pangenome Analysis Toolkit) has been developed as R package for large data analysis. It handles many types of file formats and provides 20-30X faster results with accuracy. In addition to Pangenome, it also identifies and visualizes core genome and accessory genome. PATO uses MMSeqs2, MASH and Minimap2 for core functions, VFDB for pathogenecity, and Resfinder for antibiotic resistance. It also uses Refseq database for representative and reference bacterial genomes. Its functions are divided into quality control, main functions, analysis and visualization [59].

- **PANPROVA:** PANgenomic PROkaryotic eVolution of full Assemblies is a benchmark tool that simulates the prokaryotes Pangenome evolution. Its primary component was created in C++, and its secondary procedures were created using Python and Bash scripts. For analysis, full genomic sequences are taken as input and sequences are evolved considering sequence alteration, gene variations and horizontal acquisition of genes. Its output generates the set of isolates which can be used for an artificial fragmentation. It also provides parental relationships and association between their genes [60].

- **Roary:** Roary is implemented in Perl Language. Using Roary, pangenome can be constructed for thousands of prokaryotic samples using desktop in short time without compromising results. To reduce the protein sequences, coding regions are removed from input sequences and iteratively pre-clustered with CD-HIT. It is followed by the 'All against All' comparison using BLAST and then sequences are clustered using MCL. Results from pre-clustering CD-HIT and MCL are merged and final graph is plotted as output providing context for each gene [61].

- **BGDMdocker:** BGDMdocker (Bacterial Genome Data Mining docker) uses docker technology for pangenome analysis, genome annotation, gene clusters mining, and visualization of results. It has three integrated toolkits: Prokka, panX, and antiSMASH. Docker has three core structural components: Image, Container, and Repository. Its benchmarking was done on 44 *Bacillus amyloliquefaciens* strains [62].

- **DeNoGAP**: De-Novo Genome Analysis Pipeline (DeNoGAP) integrates many tools and databases for pangenome analysis, gene prediction, ortholog prediction, homolog protein family assignment, chimera prediction and functional annotation. It is developed using Perl, BioPerl and SQLite. Using this pipeline, both complete and draft genomes with minimal computational requirement can be analysed using iteratively refined hidden Markov models clustering strategy [63].

- **Harvest**: Harvest is suitable for alignment of large number of genome. It integrates three modules Parsnp (for core genome alignment), Gingr tools (for output visualization) and HarvestTools (meta-analysis) of genome. Using these tools together allows for the generation of phylogenetic trees, variant calls, and recombination detection. Three simulated datasets of *Escherichia coli* and three actual datasets of *Peptoclostridium difficile*, *Streptococcus pneumoniae*, and *Mycobacterium tuberculosis* were used to assess its performance. [64].

- **NGSPanPipe:** It can directly identify pangenome from short sequence reads. It takes input in FastQ format, merge them and generates contigs to cover the maximum genome. Afterwards, a matrix is created depicting the presence and absence of gene with respect to reference genome. Other pangenomics tools can use this resulting matrix as input. Benchmarking has been done on simulated reads of *Mycobacterium tuberculosis* [65].

- **PanACEA**: PanACEA (Pan-genome Atlas with Chromosome Explorer and Analyzer) is an open source Perl program that can be used to visualize the pangenome data generated from other clustering tools. This interactive visualisation tool can also show annotation elements like gene ontology and antibiotic resistance. It is a user friendly, platform independent, interoperatable across pangenome software and gives both high-level and detailed views of the data [66].

- **PanACoTA:** PanACoTA (PANgenome with Annotations, Core identification, Tree and corresponding Alignments) focuses on modular method to download the genomes of a species, building a database with redundancy controls, then annotate them and construct their pangenome with alignments and phylogenetic tree. All these steps can be carried out either simultaneously or individually and several times to focus on the parameters of interest. It is developed in Python. It is illustrated with two data sets of *K. pneumoniae* genomes. It is rapid, accurate and portable software that can be evolved in future [67].

- **PANINI:** Panini (Pangenome Neighbour Identification for Bacterial Populations) is a web browser for online visualization of core and accessory genome components. To determine the neighbor of each isolate in the data set, it employs machine learning with stochastic neighbour embedding, which is based on the t-SNE (t-distributed stochastic neighbour embedding) technique. Its working is demonstrated using many single- and multi-clone pneumococcal populations [68].

- **PanOCT**: PanOCT (Pangenome Ortholog Clustering Tool) is used when homology based clustering method cannot be used, i.e., PanOCT analyzes pangenome of closely related prokaryotic species or strains. It is written in Perl. In contrast to conventional graph-based ortholog tools, it precisely places proteins into orthologous clusters by combining homology and micro synteny or conserved gene neighbourhood (CGN) [69].

- **Panseq**: It is freely available online program written in Perl. It is used for quick identification of pangenome and accessory genome with its component. For local sequence comparisons, it employs the BLASTn method; for full genome comparisons, it uses the MUMmer alignment algorithm. Based on user defined parameters, it quickly identifies SNP's, unique region in genome, constructs file for phylogeny program and produce graphical view of the output [70].

- **PanTetris :** Its standard coordinate-based arrangement allows for the visual inspection of gene occurrences in a pangenome table, representing several genomes. Its aggregation technique is inspired by the Tetris game with the option of editing and changing the pangenome groups. Its benchmarking has been done on 32 *Staphylococcus aureus* genomes. For computation, it uses progressive MAUVE algorithm [71].

- **Spine and AGEnt:** Spine is used to identify the core genome from group of genomic sequences. Afterwards, accessory genome is obtained using AGEnt, it computationally subtracts previously obtained core genome sequences from complete genome sequence. The program requires two inputs: a core genome sequence and the sequence of a query genome. To align sequences, both the programs use nucmer. However the selection of reference sequence for comparison alters the results. Spine also allows to include the incomplete or draft genome. The tool was benchmarked with the twelve genome sequences of *Pseudomonas aeruginosa* [72].

- **Pyseer :** It is the python reimplementation of SEER (Sequence Element Enrichment Analysis). It uses linear model with K-mers of variable length as input and a particular phenotype is evaluated. It produces output in VCF or Rtab files that can be used in other tools also. As compared to SEER, new algorithms and tools are also introduced in the pyseer to help in processing of output, example; use of machine learning to predict traits in sequences. [73].

- **Seq-pan-seq:**It aligns the whole genome and provides the pangenome data structure. It allows user to update the genome by addition or deletion of sequences from set of aligned genomes. It uses progressive Mauve for alignment of whole genome and its results are optimized throughout the whole sequential process and results are arranged in linear representation. The resultant sequence can be used as reference genome for further analysis such as variant detection and read mapping. [74].

- **Ptolemy:** It compares full genome assemblies using a top-down method. It is capable of managing both dynamic and conserved genomic elements. The genome is represented as quivers, or labeled multi-directed graphs, which combine to form a single quiver. Afterwards using synteny analysis, gene anchors are identified that make canonical quivers. This canonical quivers are alignment of all genomes encoding structural variations. It is tested on different datasets of *Mycobacterium, Escherichia, Saccharomyces* and *Shigella* species [75].

- **PPanGGOLiN**: It uses a statistical approach and graphical model to construct the pangenome. This uses a Markov Random Field in conjunction with a multivariate Bernoulli Mixture Model-based Expectation-Maximization algorithm to partition the graph's nodes. The graph's edges depict the genomic neighborhood, while the node symbolizes the gene family. Based on the presence or lack of genes, the gene families are categorized into persistent, cloud, and one or more shell partitions. Genomes from 439 species and 78 species' metagenome-assembled genomes are used to illustrate the method. [76].

There are two types of computational pangenome analysis tools: online which are accessible though their websites and computer software that needs to be downloaded and installed. Although user-friendly, online platforms might only support a limited number of genomes, whereas large dataset analysis may demand powerful computers and informatics expertise to install the software. All the tools and databases available for pangenome analysis are provided in Table 1.

Table 1. Various tools and databases available for pangenomic analysis

Tools/Databases	Offline or Online	Website	Reference
PGAT	Online	http://tools.uwgenomics.org/pgat/	[45]
ITEP	Offline	https://price.systemsbiology.net/itep	[46]
EDGAR	Offline	https://edgar.computational.bio.uni-giessen.de/	[47, 48]
ClustAGE	Offline	https://clustage.sourceforge.io	[49]
GET_HOMOLOGUES	Offline	http://www.eead.csic.es/compbio/soft/gethoms.php http://maya.ccg.unam.mx/soft/gethoms.php	[50]
Micropan	Offline	The Comprehensive R Archive Network (CRAN)	[51]
PanCake	Offline	https://bitbucket.org/CorinnaErnst/pancake	[52]
PGAP and PGAP-X	Offline	http://pgap.sf.net http://pgapx.ybzhao.com	[53, 54]
PanFunPro	Offline	https://zenodo.org/record	[55]
SplitMEM	Offline	http://splitmem.sourceforge.net	[56]
Panakeia	Offline	https://github.com/BioSina/Panakeia	[57]
BPGA	Offline	https://iicb.res.in/bpga/	[58]
PATO	Offline	https://github.com/irycisBioinfo/PATO	[59]
PANPROVA	Offline	https://github.com/InfOmics/PANPROVA	[60]
Roary	Online and offline both	http://sanger-pathogens.github.io/Roary	[61]
BGDMdocker	Offline	https://github.com/cgwyx/debian_prokka_panx_antismash_biodocker	[62]
DeNoGAP	Offline	https://sourceforge.net/projects/denogap/	[63]
Harvest	Offline	http://github.com/marbl/harvest	[64]
NGSPanPipe	Offline	https://github.com/Biomedinformatics/NGSPanPipe	[65]
PanACEA	Offline	https://github.com/JCVenterInstitute/PanACEA/	[66]
PanACoTA	Offline	https://github.com/gem-pasteur/PanACoTA	[67]
PANINI	Online and offline both	http://panini.pathogen.watch	[68]
PanOCT	Offline	http://panoct.sourceforge.net/	[69]

Panseq	Online	http://76.70.11.198/panseq	[70]
PanTetris	Online	http://bit.ly/1vVxYZT	[71]
Spine and AGEnt			[72]
Pyseer	Offline	https://github.com/mgalardini/pyseer	[73]
Seq-pan-seq	Offline	https://gitlab.com/rki_bioinformatics	[74]
Ptolemy	Offline	https://github.com/AbeelLab/ptolemy	[75]
PPanGGOLiN	Offline	https://github.com/labgem/PPanGGOLiN	[76]

9. Applications of Pangenome

9.1. Evolutionary studies

Pangenomics is a promising method for evolutionary studies because of widespread coverage of genome for a specific species. Genes of a species undergo many variations by virtue of evolutionary forces. Pangenome provides the aggregated picture of all the genetic variations of all the strains of species. It provides the evolutionary signature of species in the form of SNP (single nucleotide polymorphism) and gene content. The obtained evolutionary signature is useful for evolutionary studies and determining the phylogenetic relationships among different strains of species. Pangenome study for pathogenic bacteria reveals the genetic variation associated to their distinct pathophysiology and eventual emergence of antibiotic resistance [77, 78]. This will further help in detection and treatment of the pathogen associated infection. Similarly, pangenome of industrially important bacterium helps to understand the evolution of gene responsible for the production. By understanding the evolution of gene in species, industrial productivity can be increased by further variations in gene [79]. On expanding the pangenome study across many species and genera with the help of better data structure, tools and novel method would help in strengthening our evolutionary studies.

9.2. Developing novel therapeutics

To combat resistant pathogens, the development of novel treatments is urgently needed. The drug target for this therapeutics can be an essential gene or any virulence factor. However, the genes that are taken as essential or virulent for one strain of species can be essential and required for infection in other strain in genetic context. Therefore, the genetic variability provided by pangenomics should be considered while developing the new therapeutics at species level. In 2020, Dar et al., identified the drug target for *Mycobacterium tuberculosis* after pangenome analysis. They took core genes from pangenomic analysis and found essential genes from them to identify drug target [80].

9.3. Probiotics characterization

Probiotics are live bacteria that, when ingested by humans, provide positive health effects. It manifests itself in numerous illnesses, including gastrointestinal sickness. [81], dermatitis [82], cancer [83], diabetes [84], obesity [85] and many others. Many microbial strains of the *Lactobacillus, Leuconostoc, Streptococcus, Bifidobacterium, Bacillus, Enterococcus, Pediococcus, Escherichia coli* and *Saccharomyces* genera have been reported to be used as probiotics [86]. Pangenomics has allowed to study and characterize the properties of these organisms, and provides the new ways for better understanding of their genome. It provides the presence or absence of some features related with the Probiotics. For example, high acid and bile salts tolerance, antibiotic activity, cell adhesion property and modulating the immune system [87]. This information can be further used for screening and improvement of strains, safety assessments and process improvements.

9.4. Host pathogen interaction

Pangenomics can be used to elucidate the genes associated with the pathogenesis. It compares the genes of pathogenic strain with the nearest neighbor non pathogenic strain. This genomic comparison provides the genes responsible for the pathogenesis, strain to strain variation and helps to understand its interaction with the host. The antagonistic relationship between the host and the pathogen is always evolving as both of them adapt to secure their survival. In open pangenome, a species acquire new genes that may alter the host pathogen interaction in novel ways. It increases the gene repertoire involved in pathogenesis [88]. On performing pangenome analysis on 17 *E. coli* strains, the proportion of accessory gene was found to be very high (almost 83% of pangenome) which shows the adaptive evolution. In another study, on comparing the pathogenic strain of *E. coli* with the non pathogenic strains, it is revealed that the horizontal gene transfer is responsible for pathogenicity [89].

9.5. Finding novel genes

The homology between recently generated DNA sequences and those that are already available in the archives serves as the basis for pangenomic comparison. Orthologs are genes that develop as a result of speciation, whereas paralogs develop as a result of DNA duplication processes. The pangenome is categorized as core and accessory based on the orthologus relationship. The

orthology is identified using Bidirectional Best Hits (BBH) which is in accordance to the hypothesis that compared to any other sequences in the genome, orthologous genes are more similar to each other [90]. On pangenomic analysis using orthology concept, biosynthetic gene clusters were identified from *Virgibacillus* strains. These produce variety of metabolites having antimicrobial activities against multidrug-resistant bacteria, which is very useful for pharmaceutical industry [79]. Similarly, on pangenomic analysis of 524 *E. coli* strains, it was identified that almost 767 gene families (approximately 2.5% of *E. coli* pangenome) are remodeled genes that can be due to alternative reading frames of ORF's that provides novel genes in *E. coli* [91].

9.6.Taxonomic studies

The conventional bacterial classification is based on the principle of discontinuous variation. With the advancement in technology, pangenomics can be helpful for taxonomic studies. As it can help in better visualization of gene content differences, and their adaptation in different habits and habitats, it can be used to classify species according to discontinuous genomic content. A difference in the ration of core/pangenome indicates significant variation between genomes. The same core/pangenome ratio is calculated for different *Klebsiella* species that helps in the re-classification of *Klebsiella* [92]. In another case, as in *E. coli* and *Shigella* species, there are many differences amongst them, but their mechanism of pathogenecity is same. Therefore, authors argued to keep both of them in same group as the difference between them is due to medical diagnosis. On pangenome analysis, both have been clustered in two groups, each containing a mix of *E. coli* and *Shigella* species [93].

10. Bacterial Pangenome Status

Bacteria are the most widely studied organisms in the field of biology and medicine. Scientists are much interested in understanding the evolutional repertoire. This will not only enhance their knowledge, but also help in the development of therapies for wide variety of bacterial infections. The improvement in sequencing technology leads to the advancement in the pangenome studies. The pangenome data is useful in various aspects of pathogen research, phylogenomics, functional genomics and GWAS (Genome-wide association studies).

Pangenome data is available for variety of microorganisms [94-96]. A bacterium is having very small genome and the availability of hundreds of sequence strains helps to create draft genome assemblies. However, there is one fact that, besides sequencing data, pangenome is possible due to availability of algorithms and data structure. They help in consistently working with pangenomes and are constantly improving by transition to advanced and efficient system. One of the recent transitions is from using linear genome representation to graphical genome representation [97].

From the first study of pangenome in 2005 on *Streptococcus agalactiae* isolates [5], several pangenome studies has been performed on bacteria that generates novel biological insights. In Table 2, a few studies of pangenomics on different bacterial species have been provided. It has a large spectrum of applications in various fields of biology. As per analysis of Anani et al., in 2020 [98], there is an increasing trend in publication in pangenomics. A large number of annual publications in Pubmed with 'Bacteria', 'pangenome', 'pan genome' and 'pan-genome' as keywords have been evaluated and a sharp increase in publications is found as shown in Figure 3.

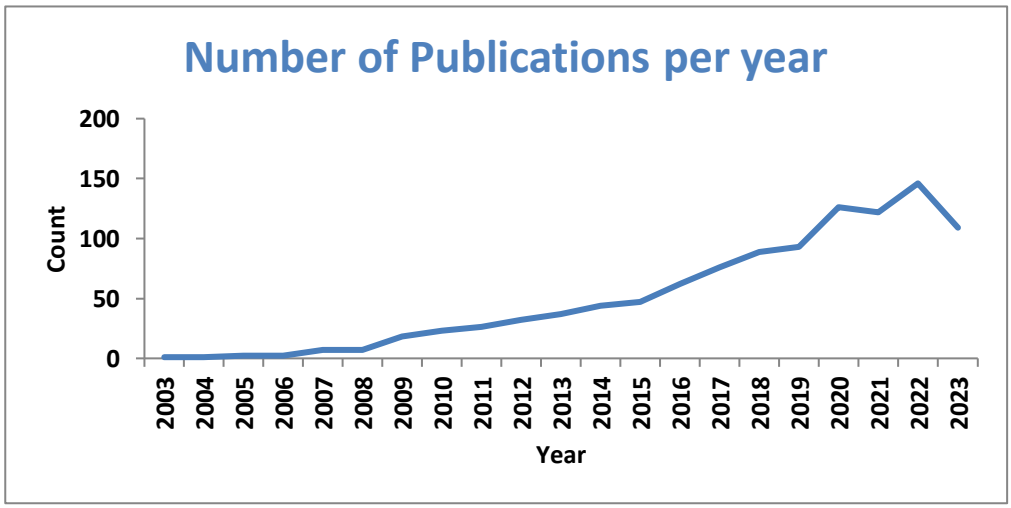

Figure 3. Annual number of publications available with 'Bacteria', 'pangenome', 'pan genome' and 'pan-genome' as keywords in Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/).

Table 2. Few examples of bacterial pangenomic studies.

Year	Bacteria	Findings	Reference
2005	*Streptococcus agalactiae*	All isolates share a core genome that makes up about 80% of bacteria.	[5]
2012	*Mycobacterium abscessus*	481,453 bp of accessory regions with a GC content of 61% and 4,857,651 bp of core regions with a GC content of 61%	[99]
2013	*Bifidobacterium animalis*	Closed pangenome structure of *B.animalis*	[100]
2014	*Streptococcus mutans*	Genetically heterogeneous group of strains of *S. mutans* and development of Microbial Genome Comparison (MGC) tool.	[101]
2015	*Lactobacillus reuteri*	A total pangenome of 5225 gene clusters including 851 genes in the core genome	[102
2016	*Staphylococcus epidermidis*	Contains multiple examples of chimeric mobile genetic elements.	[103]
2017	*Helicobacter pylori*	lineage-specific genes of *H. pylori* has been reported	[104]
2018	*Vibrio diabolicus*	1,109 conserved coding sequences between *V. diabolicus* and other synonym species	[105]
2019	*Acinetobacter seifertii*	Association of Antimicrobial Resistance and Metal Tolerance genes	[106]
2020	*Salmonella infantis*	Highest diversity suggesting a faster evolution	[107]
2021	*Enterococcus faecalis*	High degree of genomic fluidity and strain EF-2001lacked multiple drug-resistance, virulence and pathogenesis genes	[108]
2022	*Clostridioides difficile*	High genetic diversity and open pangenome in *C. difficile* and 1,481 core genes	[109]
2023	*Escherichia coli*	4.7 hotspots occupied by defense system-containing mobile elements	[110]
2024	*Acinetobacter baumannii*	High genomic flexibility and a propensity for horizontal genetic transfer can transfer mobile genetic components among themselves	[111]

11. Challenges

One of the computational challenges in pangenome analysis is the data storage and presentation. It possesses all of the characteristics of "Big Data," including volume, diversity, velocity, and veracity. The enormous amount of generated data can be handled with the introduction of more databases for pangenome and use of cloud storage. Another problem is with the heterogeneous and complex data of pangenome, which leads to big challenge for software and algorithm development. In addition to heterogeneous data, there are widely accepted standards for file formats, because of which developing an algorithm and software for vast variety of file formats is also a challenge for computational scientist. There is also a requirement of distributed and parallel processing of the data so that data can be analyzed and processed in cheap, flexible, reliable and secure manner [112].

The lack of comprehensive and well-annotated genome sequences presents another difficulty for researchers working on pangenome investigations. Scientists sometimes face difficulties aligning fragmented genomes because the entire genome of a species may not always be available. In order to ensure that genomes are produced with good quality, new technologies must be devised that can facilitate the assembly of long reads.

12. Conclusion

The challenge of analyzing rapidly increasing number of genome can be solved by pangenome studies. In pangenome studies, genome is classified into core, accessory and unique genes. Bacterial pangenome studies reveal the concept of evolution that can be further implemented on higher organisms. Many databases and tools are available for pangenome analysis, but they are not sufficient for such vast data, due to algorithm limitation and file format differences. There are many challenges in the pangenome studies that need to be sorted out in the near future. For the rapid extension of computational pangenomics, different

computation methods and paradigms are required. In near future, pangenomics will be a new branch for research in computational biology.

Acknowledgement

The authors acknowledge the facilities of the Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi (DBT), under the Bioinformatics Sub Centre as well as M.Sc. Biotechnology programme used in the present work.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

[1] Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J. C., Lee, J., ... & Reddy, T. B. K. (2021). Genomes OnLine Database (GOLD) v. 8: overview and updates. Nucleic acids research, 49(D1), D723-D733.https://doi.org/10.1093/nar/gkaa983

[2] Qin, S., Xiao, W., Zhou, C., Pu, Q., Deng, X., Lan, L., ... & Wu, M. (2022). Pseudomonas aeruginosa: pathogenesis, virulence factors, antibiotic resistance, interaction with host, technology advances and emerging therapeutics. Signal transduction and targeted therapy, 7(1), 199.https://doi.org/10.1038/s41392-022-01056-1

[3] Uddin, T. M., Chakraborty, A. J., Khusro, A., Zidan, B. R. M., Mitra, S., Emran, T. B., ... & Koirala, N. (2021). Antibiotic resistance in microbes: History, mechanisms, therapeutic strategies and future prospects. Journal of infection and public health, 14(12), 1750-1766.https://doi.org/10.1016/j.jiph.2021.10.020

[4] Bornstein, K., Gryan, G., Chang, E. S., Marchler-Bauer, A., & Schneider, V. A. (2023). The NIH Comparative Genomics Resource: addressing the promises and challenges of comparative genomics on human health. BMC genomics, 24(1), 575.https://doi.org/10.1186/s12864-023-09643-4

[5] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... & Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences, 102(39), 13950-13955.https://doi.org/10.1073/pnas.0506758102

[6] Sigaux, F. (2000). Cancer genome or the development of molecular portraits of tumors. Bulletin de L'academie Nationale de Medecine, 184(7), 1441-7.

[7] Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M. C., Jolley, K. A., & Sheppard, S. K. (2014). A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter. PloS one, 9(3), e92798.https://doi.org/10.1371/journal.pone.0092798

[8] Hisham, Y., & Ashhab, Y. (2018). Identification of cross-protective potential antigens against pathogenic Brucella spp. c

[15] Millán-Zambrano, G., Burton, A., Bannister, A. J., & Schneider, R. (2022). Histone post-translational modifications—cause and consequence of genome function. Nature Reviews Genetics, 23(9), 563-580.https://doi.org/10.1038/s41576-022-00468-7

[16] Copley, S. D. (2020). Evolution of new enzymes by gene duplication and divergence. The FEBS journal, 287(7), 1262-1283.https://doi.org/10.1111/febs.15299

[17] Schirrmeister, B. E., Dalquen, D. A., Anisimova, M., & Bagheri, H. C. (2012). Gene copy number variation and its significance in cyanobacterial phylogeny. BMC microbiology, 12, 1-15.https://doi.org/10.1186/1471-2180-12-177

[18] Rogers, J., & Gibbs, R. A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. Nature Reviews Genetics, 15(5), 347-359.https://doi.org/10.1038/nrg3707

[19] Gouda, S. K., Kumari, K., Panda, A. N., & Raina, V. (2024). Computational Tools for Whole Genome and Metagenome Analysis of NGS Data for Microbial Diversity Studies. In S. Das, & H. R. Dash (Eds.), Microbial Diversity in the Genomic Era (pp. 43-71). Academic Press.https://doi.org/10.1016/B978-0-443-13320-6.00032-9

[20] Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., ... & Garrison, E. (2020). Pangenome graphs. Annual review of genomics and human genetics, 21, 139-162.https://doi.org/10.1146/annurev-genom-120219-080406

[21] Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics, 14(2), 178-192.https://doi.org/10.1093/bib/bbs017

[22] Snipen, L., Almøy, T., & Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. BMC genomics, 10, 1-8.https://doi.org/10.1186/1471-2164-10-385
[23] Mira, A., Martín-Cuadrado, A. B., D'Auria, G., & Rodríguez-Valera, F. (2010). The bacterial pan-genome: a new paradigm in microbiology. International Microbiology, 13(2), 45-57.https://doi.org/10.2436/20.1501.01.110
[24] Fang, X., Lloyd, C. J., & Palsson, B. O. (2020). Reconstructing organisms in silico: genome-scale models and their emerging applications. Nature Reviews Microbiology, 18(12), 731-743.https://doi.org/10.1038/s41579-020-00440-4
[25] Croll, D., & McDonald, B. A. (2012). The accessory genome as a cradle for adaptive evolution in pathogens. PLoS pathogens, 8(4), e1002608.https://doi.org/10.1371/journal.ppat.1002608
[26] Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomics comes of age: from bacteria to plant and animal applications. Trends in Genetics, 36(2), 132-145.https://doi.org/10.1016/j.tig.2019.11.006
[27] Gimenez, G., Bertelli, C., Moliner, C., Robert, C., Raoult, D., Fournier, P. E., & Greub, G. (2011). Insight into cross-talk between intra-amoebal pathogens. BMC genomics, 12, 1-14.https://doi.org/10.1186/1471-2164-12-542
[28] Rouli, L., Merhej, V., Fournier, P. E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. New microbes and new infections, 7, 72-85.https://doi.org/10.1016/j.nmni.2015.06.005
[29] Arnold, B. J., Huang, I. T., & Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. Nature Reviews Microbiology, 20(4), 206-218.https://doi.org/10.1038/s41579-021-00650-4
[30] Emamalipour, M., Seidi, K., Zununi Vahed, S., Jahanban-Esfahlan, A., Jaymand, M., Majdi, H., ... & Zare, P. (2020). Horizontal gene transfer: from evolutionary flexibility to disease progression. Frontiers in cell and developmental biology, 8, 229.https://doi.org/10.3389/fcell.2020.00229
[31] Boyd, E. F., & Brüssow, H. (2002). Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends in Microbiology, 10(11), 521-529.https://doi.org/10.1016/S0966-842X(02)02459-9
[32] Branger, C., Zamfir, O., Geoffroy, S., Laurans, G., Arlet, G., Thien, H. V., ... & Denamur, E. (2005). Genetic background of Escherichia coli and extended-spectrum β-lactamase type. Emerging infectious diseases, 11(1), 54.https://doi.org/10.3201/eid1101.040257
[33] Oberhardt, M. A., Puchałka, J., Martins dos Santos, V. A., & Papin, J. A. (2011). Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. PLoS computational biology, 7(3), e1001116.https://doi.org/10.1371/journal.pcbi.1001116
[34] Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis?. Nature biotechnology, 28(3), 245-248.https://doi.org/10.1038/nbt.1614
[35] Norsigian, C. J., Fang, X., Palsson, B. O., & Monk, J. M. (2020). Pangenome flux balance analysis toward panphenomes. In H. Tettelin, & D. Medini (Eds.)., The pangenome: diversity, dynamics and evolution of genomes (pp. 219-232). Springer Nature. https://doi.org/10.1007/978-3-030-38281-0_10
[36] Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., ... & Palsson, B. O. (2017). i ML1515, a knowledgebase that computes Escherichia coli traits. Nature biotechnology, 35(10), 904-908.https://doi.org/10.1038/nbt.3956
[37] Fouts, D. E., Matthias, M. A., Adhikarla, H., Adler, B., Amorim-Santos, L., Berg, D. E., ... & Vinetz, J. M. (2016). What makes a bacterial species pathogenic?: comparative genomic analysis of the genus Leptospira. PLoS neglected tropical diseases, 10(2), e0004403.https://doi.org/10.1371/journal.pntd.0004403
[38] Seif, Y., Kavvas, E., Lachance, J. C., Yurkovich, J. T., Nuccio, S. P., Fang, X., ... & Monk, J. M. (2018). Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. Nature communications, 9(1), 3771. https://doi.org/10.1038/s41467-018-06112-5
[39] Choudhary, K. S., Mih, N., Monk, J., Kavvas, E., Yurkovich, J. T., Sakoulas, G., & Palsson, B. O. (2018). The Staphylococcus aureus two-component system AgrAC displays four distinct genomic arrangements that delineate genomic virulence factor signatures. Frontiers in Microbiology, 9, 1082.https://doi.org/10.3389/fmicb.2018.01082
[40] Gibson, G. (2002). Microarrays in ecology and evolution: a preview. Molecular ecology, 11(1), 17-24.https://doi.org/10.1046/j.0962-1083.2001.01425.x
[41] Lindroos, H. L., Mira, A., Repsilber, D., Vinnere, O., Näslund, K., Dehio, M., ... & Andersson, S. G. (2005). Characterization of the genome composition of Bartonella koehlerae by microarray comparative genomic hybridization profiling. Journal of bacteriology, 187(17), 6155-6165.https://doi.org/10.1128/jb.187.17.6155-6165.2005
[42] Zhang, Q., Melcher, U., Zhou, L., Najar, F. Z., Roe, B. A., & Fletcher, J. (2005). Genomic comparison of plant pathogenic and nonpathogenic Serratia marcescens strains by suppressive subtractive hybridization. Applied and environmental microbiology, 71(12), 7716-7723.https://doi.org/10.1128/AEM.71.12.7716-7723.2005
[43] Hiller, N. L., Janto, B., Hogg, J. S., Boissy, R., Yu, S., Powell, E., ... & Hu, F. Z. (2007). Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. Journal of Bacteriology, 189(22), 8186 - 8195. https://doi.org/10.1128/jb.00690-07
[44] Li, T., & Yin, Y. (2022). Critical assessment of pan-genomic analysis of metagenome-assembled genomes. Briefings in Bioinformatics, 23(6), bbac413.https://doi.org/10.1093/bib/bbac413
[45] Brittnacher, M. J., Fong, C., Hayden, H. S., Jacobs, M. A., Radey, M., & Rohmer, L. (2011). PGAT: a multistrain analysis resource for microbial genomes. Bioinformatics, 27(17), 2429-2430.https://doi.org/10.1093/bioinformatics/btr418
[46] Benedict, M. N., Henriksen, J. R., Metcalf, W. W., Whitaker, R. J., & Price, N. D. (2014). ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC genomics, 15, 1-11.https://doi.org/10.1186/1471-2164-15-8

[47] Blom, J., Albaum, S. P., Doppmeier, D., Pühler, A., Vorhölter, F. J., Zakrzewski, M., & Goesmann, A. (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC bioinformatics, 10, 1-14.https://doi.org/10.1186/1471-2105-10-154
[48] Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C., & Goesmann, A. (2016). EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic acids research, 44(W1), W22-W28.https://doi.org/10.1093/nar/gkw255
[49] Ozer, E. A. (2018). ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. BMC bioinformatics, 19, 1-12.https://doi.org/10.1186/s12859-018-2154-x
[50] Contreras-Moreira, B., & Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Applied and environmental microbiology, 79(24), 7696-7701.https://doi.org/10.1128/AEM.02411-13
[51] Snipen, L., & Liland, K. H. (2015). micropan: an R-package for microbial pan-genomics. BMC bioinformatics, 16, 1-8.https://doi.org/10.1186/s12859-015-0517-0
[52] Ernst, C., & Rahmann, S. (2013). PanCake: a data structure for pangenomes. In German conference on bioinformatics 2013, 34, 35-45. https://doi.org/10.4230/OASIcs.GCB.2013.35
[53] Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., & Yu, J. (2012). PGAP: pan-genomes analysis pipeline. Bioinformatics, 28(3), 416-418.https://doi.org/10.1093/bioinformatics/btr655
[54] Zhao, Y., Sun, C., Zhao, D., Zhang, Y., You, Y., Jia, X., ... & Xiao, J. (2018). PGAP-X: extension on pan-genome analysis pipeline. BMC genomics, 19, 115-124.https://doi.org/10.1186/s12864-017-4337-7
[55] Lukjancenko, O., Thomsen, M. C., Voldby Larsen, M., & Ussery, D. W. (2013). PanFunPro: PAN-genome analysis based on FUNctional PROfiles. F1000Research, 2, 265.https://doi.org/10.12688/f1000research.2-265.v1
[56] Marcus, S., Lee, H., & Schatz, M. C. (2014). SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics, 30(24), 3476-3483.https://doi.org/10.1093/bioinformatics/btu756
[57] Beier, S., & Thomson, N. R. (2022). Panakeia-a universal tool for bacterial pangenome analysis. BMC genomics, 23(1), 265.https://doi.org/10.1186/s12864-022-08303-3
[58] Chaudhari, N. M., Gupta, V. K., & Dutta, C. (2016). BPGA-an ultra-fast pan-genome analysis pipeline. Scientific reports, 6(1), 24373.https://doi.org/10.1038/srep24373
[59] Fernández-de-Bobadilla, M. D., Talavera-Rodríguez, A., Chacón, L., Baquero, F., Coque, T. M., & Lanza, V. F. (2021). PATO: pangenome analysis toolkit. Bioinformatics, 37(23), 4564-4566.https://doi.org/10.1093/bioinformatics/btab697
[60] Bonnici, V., & Giugno, R. (2022). PANPROVA: pangenomic prokaryotic evolution of full assemblies. Bioinformatics, 38(9), 2631-2632.https://doi.org/10.1093/bioinformatics/btac158
[61] Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., ... & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31(22), 3691-3693.https://doi.org/10.1093/bioinformatics/btv421
[62] Cheng, G., Lu, Q., Zhou, Z., Ma, L., Zhang, G., Yilei, W. U., & Chen, C. (2017). BGDMdocker: a Docker workflow for analysis and visualization pan-genome and biosynthetic gene clusters of bacterial. BioRxiv, 098392.
[63] Thakur, S., & Guttman, D. S. (2016). A De-Novo Genome Analysis Pipeline (DeNoGAP) for large-scale comparative prokaryotic genomics studies. BMC bioinformatics, 17, 1-18.https://doi.org/10.1186/s12859-016-1142-2
[64] Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome biology, 15, 1-15.https://doi.org/10.1186/s13059-014-0524-x
[65] Kulsum, U., Kapil, A., Singh, H., & Kaur, P. (2018). NGSPanPipe: a pipeline for pan-genome identification in microbial strains from experimental reads. In Infectious Diseases and Nanomedicine III: Second International Conference (ICIDN-2015), 39-49. https://doi.org/10.1007/978-981-10-7572-8_4
[66] Clarke, T. H., Brinkac, L. M., Inman, J. M., Sutton, G., & Fouts, D. E. (2018). PanACEA: a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes. BMC bioinformatics, 19, 1-6.https://doi.org/10.1186/s12859-018-2250-y
[67] Perrin, A., & Rocha, E. P. (2021). PanACoTA: a modular tool for massive microbial comparative genomics. NAR genomics and bioinformatics, 3(1), lqaa106.https://doi.org/10.1093/nargab/lqaa106
[68] Abudahab, K., Prada, J. M., Yang, Z., Bentley, S. D., Croucher, N. J., Corander, J., & Aanensen, D. M. (2019). PANINI: pangenome neighbour identification for bacterial populations. Microbial Genomics, 5(4), e000220.https://doi.org/10.1099/mgen.0.000220
[69] Fouts, D. E., Brinkac, L., Beck, E., Inman, J., & Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic acids research, 40(22), e172-e172.https://doi.org/10.1093/nar/gks757
[70] Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., ... & Gannon, V. P. (2010). Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC bioinformatics, 11, 1-14.https://doi.org/10.1186/1471-2105-11-461
[71] Hennig, A., Bernhardt, J., & Nieselt, K. (2015). Pan-Tetris: an interactive visualisation for Pan-genomes. BMC bioinformatics, 16, 1-11.https://doi.org/10.1186/1471-2105-16-S11-S3
[72] Ozer, E. A., Allen, J. P., & Hauser, A. R. (2014). Characterization of the core and accessory genomes of Pseudomonas aeruginosa using bioinformatic tools Spine and AGEnt. BMC genomics, 15, 1-17.https://doi.org/10.1186/1471-2164-15-737

[73] Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). Pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics, 34(24), 4310-4312.https://doi.org/10.1093/bioinformatics/bty539
[74] Jandrasits, C., Dabrowski, P. W., Fuchs, S., & Renard, B. Y. (2018). seq-seq-pan: Building a computational pan-genome data structure on whole genome alignment. BMC genomics, 19, 1-12.https://doi.org/10.1186/s12864-017-4401-3
[75] Salazar, A. N., & Abeel, T. (2018). Approximate, simultaneous comparison of microbial genome architectures via syntenic anchoring of quiver representations. Bioinformatics, 34(17), i732-i742.https://doi.org/10.1093/bioinformatics/bty614
[76] Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., ... & Vallenet, D. (2020). PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. PLoS computational biology, 16(3), e1007732.https://doi.org/10.1371/journal.pcbi.1009687
[77] Livingstone, P. G., Morphew, R. M., & Whitworth, D. E. (2018). Genome sequencing and pan-genome analysis of 23 Corallococcus spp. strains reveal unexpected diversity, with particular plasticity of predatory gene sets. Frontiers in microbiology, 9, 428423.https://doi.org/10.3389/fmicb.2018.03187
[78] Rajput, A., Seif, Y., Choudhary, K. S., Dalldorf, C., Poudel, S., Monk, J. M., & Palsson, B. O. (2021). Pangenome analytics reveal two-component systems as conserved targets in ESKAPEE pathogens. Msystems, 6(1), 10-1128.https://doi.org/10.1128/msystems.00981-20
[79] Othoum, G., Bougouffa, S., Bokhari, A., Lafi, F. F., Gojobori, T., Hirt, H., ... & Essack, M. (2019). Mining biosynthetic gene clusters in Virgibacillus genomes. BMC genomics, 20, 1-10.https://doi.org/10.1186/s12864-019-6065-7
[80] Dar, H. A., Zaheer, T., Ullah, N., Bakhtiar, S. M., Zhang, T., Yasir, M., ... & Ali, A. (2020). Pangenome analysis of Mycobacterium tuberculosis reveals core-drug targets and screening of promising lead compounds for drug discovery. Antibiotics, 9(11), 819. https://doi.org/10.3390/antibiotics9110819
[81] Jonkers, D., Penders, J., Masclee, A., & Pierik, M. (2012). Probiotics in the management of inflammatory bowel disease: a systematic review of intervention studies in adult patients. Drugs, 72, 803-823. https://doi.org/10.2165/11632710-000000000-00000
[82] Rusu, E., Enache, G., Cursaru, R., Alexescu, A., Radu, R., Onila, O., ... & Radulian, G. (2019). Prebiotics and probiotics in atopic dermatitis. Experimental and therapeutic medicine, 18(2), 926-931.https://doi.org/10.3892/etm.2019.7678
[83] Lu, K., Dong, S., Wu, X., Jin, R., & Chen, H. (2021). Probiotics in cancer. Frontiers in Oncology, 11, 638148.https://doi.org/10.3389/fonc.2021.638148
[84] Bordalo Tonucci, L., Dos Santos, K. M. O., De Luces Fortes Ferreira, C. L., Ribeiro, S. M. R., De Oliveira, L. L., & Martino, H. S. D. (2017). Gut microbiota and probiotics: Focus on diabetes mellitus. Critical Reviews in Food Science and Nutrition, 57(11), 2296-2309.https://doi.org/10.1080/10408398.2014.934438
[85] Kobyliak, N., Conte, C., Cammarota, G., Haley, A. P., Styriak, I., Gaspar, L., ... & Kruzliak, P. (2016). Probiotics in prevention and treatment of obesity: a critical view. Nutrition & metabolism, 13, 1-13.https://doi.org/10.1186/s12986-016-0067-0
[86] Fijan, S. (2014). Microorganisms with claimed probiotic properties: an overview of recent literature. International journal of environmental research and public health, 11(5), 4745-4767.https://doi.org/10.3390/ijerph110504745
[87] Kim, J. A., Bayo, J., Cha, J., Choi, Y. J., Jung, M. Y., Kim, D. H., & Kim, Y. (2019). Investigating the probiotic characteristics of four microbial strains with potential application in feed industry. PloS one, 14(6), e0218922.https://doi.org/10.1371/journal.pone.0218922
[88] Reis, A. C., & Cunha, M. V. (2021). The open pan-genome architecture and virulence landscape of Mycobacterium bovis. Microbial Genomics, 7(10), 000664.https://doi.org/10.1099/mgen.0.000664
[89] Perna, N. T., Plunkett III, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., ... & Blattner, F. R. (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157: H7. Nature, 409(6819), 529-533.https://doi.org/10.1038/35054089
[90] Bosi, E., Fani, R., & Fondi, M. (2015). Defining orthologs and pangenome size metrics. In A. Mengoni, M. Galardini, & M. Fondi (Eds.), Bacterial Pangenomics: Methods and Protocols (pp. 191-202). Humana Press. https://doi.org/10.1007/978-1-4939-1720-4_13
[91] Watson, A. K., Lopez, P., & Bapteste, E. (2022). Hundreds of out-of-frame remodeled gene families in the Escherichia coli pangenome. Molecular Biology and Evolution, 39(1), msab329.https://doi.org/10.1093/molbev/msab329
[92] Caputo, A., Merhej, V., Georgiades, K., Fournier, P. E., Croce, O., Robert, C., & Raoult, D. (2015). Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: the Klebsiella paradigm. Biology Direct, 10, 1-12.https://doi.org/10.1186/s13062-015-0085-2
[93] Aggarwal, S. K., Singh, A., Choudhary, M., Kumar, A., Rakshit, S., Kumar, P., ... & Varshney, R. K. (2022). Pangenomics in microbial and crop research: progress, applications, and perspectives. Genes, 13(4), 598.https://doi.org/10.3390/genes13040598
[94] Gong, H., Huang, X., Zhu, W., Chen, J., Huang, Y., Zhao, Z., ... & Wang, X. (2023). Pan-genome analysis of the Burkholderia gladioli pv. cocovenenans reveal the extent of variation in the toxigenic gene cluster. Food Microbiology, 113, 104249.https://doi.org/10.1016/j.fm.2023.104249
[95] Gu, X., Lu, X., Lin, S., Shi, X., Shen, Y., Lu, Q., ... & Zheng, M. (2022). A comparative genomic approach to determine the virulence factors and horizontal gene transfer events of clinical Acanthamoeba isolates. Microbiology Spectrum, 10(2), e00025-22.https://doi.org/10.1128/spectrum.00025-22
[96] Yin, Z., Liu, X., Qian, C., Sun, L., Pang, S., Liu, J., ... & Xie, Z. (2022). Pan-genome analysis of Delftia tsuruhatensis reveals important traits concerning the genetic diversity, pathogenicity, and biotechnological properties of the species. Microbiology Spectrum, 10(2), e02072-21.https://doi.org/10.1128/spectrum.02072-21

[97] Liu, Y., & Tian, Z. (2020). From one linear genome to a graph-based pan-genome: a new era for genomics. Science China. Life sciences, 63(12), 1938-1941.https://doi.org/10.1007/s11427-020-1808-0
[98] Anani, H., Zgheib, R., Hasni, I., Raoult, D., & Fournier, P. E. (2020). Interest of bacterial pangenome analyses in clinical microbiology. Microbial pathogenesis, 149, 104275.https://doi.org/10.1016/j.micpath.2020.104275
[99] Ngeow, Y. F., Wee, W. Y., Wong, Y. L., Tan, J. L., Ongi, C. S., Ng, K. P., & Choo, S. W. (2012). Genomic analysis of Mycobacterium abscessus strain M139, which has an ambiguous subspecies taxonomic position.https://doi.org/10.1128/jb.01455-12
[100] Milani, C., Duranti, S., Lugli, G. A., Bottacini, F., Strati, F., Arioli, S., ... & Ventura, M. (2013). Comparative genomics of Bifidobacterium animalis subsp. lactis reveals a strict monophyletic bifidobacterial taxon. Applied and Environmental Microbiology, 79(14), 4304-4315.https://doi.org/10.1128/AEM.00984-13
[101] Argimón, S., Konganti, K., Chen, H., Alekseyenko, A. V., Brown, S., & Caufield, P. W. (2014). Comparative genomics of oral isolates of Streptococcus mutans by in silico genome subtraction does not reveal accessory DNA associated with severe early childhood caries. Infection, Genetics and Evolution, 21, 269-278.https://doi.org/10.1016/j.meegid.2013.11.003
[102] Wegmann, U., MacKenzie, D. A., Zheng, J., Goesmann, A., Roos, S., Swarbreck, D., ... & Juge, N. (2015). The pan-genome of Lactobacillus reuteri strains originating from the pig gastrointestinal tract. BMC genomics, 16, 1-18.https://doi.org/10.1186/s12864-015-2216-7
[103] Lee, J. Y., Monk, I. R., Pidot, S. J., Singh, S., Chua, K. Y., Seemann, T., ... & Howden, B. P. (2016). Functional analysis of the first complete genome sequence of a multidrug resistant sequence type 2 Staphylococcus epidermidis. Microbial genomics, 2(9), e000077.https://doi.org/10.1099/mgen.0.000077
[104] van Vliet, A. H. (2017). Use of pan-genome analysis for the identification of lineage-specific genes of Helicobacter pylori. Microbiology Letters, 364(2), fnw296.https://doi.org/10.1093/femsle/fnw296
[105] Turner, J. W., Tallman, J. J., Macias, A., Pinnell, L. J., Elledge, N. C., Nasr Azadani, D., ... & Strom, M. S. (2018). Comparative genomic analysis of Vibrio diabolicus and six taxonomic synonyms: a first look at the distribution and diversity of the expanded species. Frontiers in Microbiology, 9, 1893.https://doi.org/10.3389/fmicb.2018.01893
[106] Furlan, J. P. R., de Almeida, O. G. G., De Martinis, E. C. P., & Stehling, E. G. (2019). Characterization of an environmental multidrug-resistant Acinetobacter seifertii and comparative genomic analysis reveals co-occurrence of antimicrobial resistance and metal tolerance determinants. Frontiers in Microbiology, 10, 482791.https://doi.org/10.3389/fmicb.2019.02151
[107] Nagy, T., Szmolka, A., Wilk, T., Kiss, J., Szabó, M., Pászti, J., ... & Olasz, F. (2020). Comparative genome analysis of Hungarian and global strains of Salmonella Infantis. Frontiers in Microbiology, 11, 513428.https://doi.org/10.3389/fmicb.2020.00539
[108] Panthee, S., Paudel, A., Hamamoto, H., Ogasawara, A. A., Iwasa, T., Blom, J., & Sekimizu, K. (2021). Complete genome sequence and comparative genomic analysis of Enterococcus faecalis EF-2001, a probiotic bacterium. Genomics, 113(3), 1534-1542.https://doi.org/10.1016/j.ygeno.2021.03.021
[109] Golchha, N. C., Nighojkar, A., & Nighojkar, S. (2022). Redefining genomic view of Clostridioides difficile through pangenome analysis and identification of drug targets from its core genome. Drug Target Insights, 16, 17.https://doi.org/10.33393/dti.2022.2469
[110] Hochhauser, D., Millman, A., & Sorek, R. (2023). The defense island repertoire of the Escherichia coli pan-genome. PLoS genetics, 19(4), e1010694.https://doi.org/10.1371/journal.pgen.1010694
[111] Karampatakis, T., Tsergouli, K., & Behzadi, P. (2024). Pan-genome plasticity and virulence factors: a natural treasure trove for Acinetobacter baumannii. Antibiotics, 13(3), 257.https://doi.org/10.3390/antibiotics13030257
[112] The Computational Pan-Genomics Consortium. (2018). Computational pan-genomics: status, promises and challenges. Briefings in bioinformatics, 19(1), 118-135. https://doi.org/10.1093/bib/bbw089

How to Cite: Golchha, N. C. ., Nighojkar, A., & Nighojkar, S. (2024) Bacterial Pangenome: A Review on the Current Strategies, Tools and Applications. Medinformatics. https://doi.org/10.47852/bonviewMEDIN42022496