Gene-based and Graph-based Pangenomes of 8 Burkholderia Species: Construction, Visualization and Analysis

Dinis Duarte Robalo Martins
dinis.martins@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2023

Abstract

Pangenomes are computational objects that capture genetic diversity within a clade, competing with single-reference genomes due to their better handling of exponentially growing next-generation sequencing (NGS) data. *Burkholderia* is a highly diverse genus of gram-negative bacteria comprising over a hundred species, including human-infecting pathogens and ecologically relevant microorganisms. We built pangenomes for eight *Burkholderia* species to characterize their genomes and to identify genomic elements unique to each species. Gene-based pangenomes represent the set of essential and accessory genes of a clade. We used the software suite Pagoo to create *Burkholderia* pangenomes using in-house databases comprising genomic information about each gene encoded in the 90 *Burkholderia* genomes under analyses in this work. We analyzed the capability of this tool in creating good representations of genomic diversity, particularly in the identification of core genes. We also calculated the genomic fluidity of *Burkholderia*, estimated the pangenome sizes, and identified neutral genes within the core genomes. Pangenome graphs depict genomic sequences in an easily navigable way, using graph nodes to represent aligned sequences. We created single-species pangenome graphs for *Burkholderia*: we performed whole-genome pairwise alignment with Anchorwave, then induced the graphs with Seqwish. We identified genomic variants (bubbles) using Bubblegun. Finally, we created phylogenetic trees for each species using information contained in the pangenomes. This work showcases how *in silico* techniques such as pangenomics can be used to study genomic diversity in closely related species, providing insights on the origin of their distinct characteristics.

Keywords: Bioinformatics; *Burkholderia* species; Gene-Based Pangenomes; Genetic Variation; Graph-based pangenomes; Variation Graph Model.

1. Introduction

NGS technologies have led to an exponential growth of genome sequence data within public databases, shifting genome analyses from individual or limited genomes to the exploration of hundreds or even thousands of genomes [1]. This shift has highlighted that the bottleneck in contemporary genomic research no longer lies in data scarcity but rather in managing the accumulation of genomic data.

The historical trajectory of genomics primarily revolved around the creation of reference genomes, an endeavor demanding substantial resources and time [2]. Nonetheless, it has become increasingly evident that reliance on a single reference genome poses constraints on genetic studies. Pangenome projects, initially centered on bacteria due to their smaller genomes, offered insights into pathogenicity, virulence, and drug resistance [3, 4]. With the evolution of sequencing technologies, pangenomic investigations expanded to encompass plants and animals, unveiling substantial genetic diversity within these populations [5]. To counter reference bias, a paradigm shift towards pangenomic reference systems has been proposed [6].

Conceptually, a pangenome signifies the complete genomic repertoire of a species or a related group of organisms, encapsulating all genetic elements, sequences, and variations within that specific set of individuals. It transcends specific methodologies, focusing on capturing the entirety of available genetic information within a species. This encompasses core genomic elements shared universally among individuals and the accessory or variable elements specific to certain individuals or subgroups.

A pangenome represents not only the genes but

also encompasses non-coding regions, regulatory elements, structural variations, and other genomic features. By considering the entire spectrum of genetic content within a species or population, the concept of the pangenome offers a more comprehensive understanding of genomic diversity, evolution, and adaptation.

Pagoo [7] is a versatile tool tailored for analyzing pangenome data from various reconstruction software. Operating within R, it simplifies data management by integrating orthologous clusters, sequences, annotations, and metadata into a single, easily shareable object. This object supports querying, handling, and data subsetting, along with standard statistical analyses and dynamic visualizations, making it ideal for bacterial population genomics studies. One standout aspect of Pagoo is its object-oriented design based on R6 classes (PgR6, PgR6M, PgR6MS), offering diverse functionalities from basic data handling to advanced statistical methods and visualization tools. Importantly, third-party applications can extend these classes to suit specific needs. Pagoo also allows interaction with the pangenome object without altering the original data, enabling dynamic modifications like hiding organisms, adjusting gene definitions, or extracting specific gene information. Its interactive application facilitates statistical analysis and visualization, providing customizable plots and a built-in R-Shiny app for exploring evolutionary trends. Pagoo's flexibility allows users to set core levels, influencing gene categorizations within the pangenome object. It fills a gap in existing tools by offering both visualizations and the capability to revisit data for further analyses, all within a single framework. R's flexibility further enhances the transition between data exploration and in-depth analysis.

AnchorWave (Anchored Wavefront Alignment) [8] identifies collinear regions via conserved anchors (full-length CDS and full-length exon have been implemented currently) and breaks collinear regions into shorter fragments, i.e., anchor and inter-anchor intervals. By performing sensitive sequence alignment for each shorter interval via a 2-piece affine gap cost strategy and merging them together, AnchorWave generates a whole-genome alignment for each collinear block. AnchorWave implements commands to guide collinear block identification with or without chromosomal rearrangements and provides options to use known polyploidy levels or whole-genome duplications to inform alignment. AnchorWave takes the reference genome sequence and gene annotation in GFF3 as input and extracts reference full-length CDS to use as anchors.

Seqwish plays a pivotal role in generating pangenome graphs by constructing a complete variation graph based on a collection of sequences and their corresponding alignments [9]. The resultant graph paths offer precise and complete reconstruction of the input sequences, while the graph topology faithfully represents all variants inferred from the input alignments. Seqwish allows for the representation of complex genomic structures, enabling researchers to explore genetic variations, including SNPs, insertions, deletions, and structural rearrangements, thereby facilitating a comprehensive understanding of genome evolution and diversity. Seqwish has the capability to handle large-scale genomic datasets with speed and minimal memory requirements. Its ability to process vast amounts of genomic information without sacrificing accuracy has made it a valuable tool for researchers working with extensive datasets .

The VG toolkit capabilities include read mapping, variant calling, and visualization tools, making it a versatile choice for NGS data analysis [10]. Notably, the toolkit facilitates the integration of various genomic variants within pangenome graphs, enabling the genotyping of variants that are challenging to achieve using a single linear reference. This is a substantial advantage that reduces reference bias and improves data analysis efficiency [11]. In the context of structural variant (SV) genotyping, the VG toolkit demonstrates strong performance across datasets. It is particularly robust when faced with minor inaccuracies in SV breakpoint locations (up to 10 bp). VG's ability to genotype arbitrary combinations of SVs simultaneously, using the snarl decomposition, is a significant advantage [12]. VG can fine-tune SV breakpoints by augmenting the graph with observed differences from read alignments, effectively correcting small errors in SV breakpoints [13]. VG may struggle with variants having higher uncertainty in breakpoint location, primarily those discovered through read coverage analysis. Nevertheless, the tool offers flexible and efficient solutions for SV genotyping. The Optimized Dynamic Genome Graph Implementation (ODGI) toolkit complements VG by facilitating graph manipulation tasks like visualization and the extraction of distances among paths in the graph, supporting phylogenetic analysis [14]. ODGI also offers a multitude of other graph processing commands, allowing graph compression, variation analysis, graph traversal, aligning sequences onto the graph representation, read mapping, efficient data access and interoperability. ODGI similarity is a command introduced in April 2023 that provides a similarity or distance matrix for paths of a given variation graph, allowing phylogenetic comparison between paths. The VG and ODGI toolkits offer a promising approach to

NGS data analysis, improving the efficiency and accuracy of SV genotyping, and reducing reference bias. They empower researchers to work with more complex genomic variations and enhance our understanding of genetic diversity within species [15].

The *Burkholderia* genus is characterized by Gram-negative, obligately aerobic, rod-shaped bacteria with motility through polar flagella, except for *Burkholderia mallei* [16]. These bacteria encompass both animal and plant pathogens, and they share a conserved RNA structure, the anti-hemB RNA motif. [17]. Burkholderia species are known to produce a wide range of specialized metabolites with properties like cytotoxicity, antimicrobial activity, and virulence functions [18]. The presence of these metabolites varies among different clades, with some showing higher specialized metabolite capacities than others.

Recent research areas related to *Burkholderia* species includes metabolomic responses to antibiotics, contact-dependent interactions between bacterial communities, and genomic potential for beneficial product synthesis. In particular, certain antibiotics, such as trimethoprim, have been shown to induce metabolic responses and upregulate silent secondary metabolite gene clusters in *Burkholderia thailandensis*. Moreover, research has revealed that closely related cystic fibrosis-associated *Burkholderia* species exhibit personalized metabolomic responses to trimethoprim [19, 20].

The *Burkholderia cepacia complex* (Bcc) is one of the groups part of the *Burkholderia* genus, including at least 20 different species [21]. The Bcc is often associated with pneumonia in immunocompromised individuals, particularly those with underlying lung diseases, such as cystic fibrosis [22]. It can also affect plants and exhibits the ability to digest oil. Bcc organisms are commonly found in water and soil, showing relatively poor virulence. They possess virulence factors such as adherence to plastic surfaces, enzyme production, and resistance to neutrophil attacks [23]. Person-to-person transmission has been documented, leading to strict isolation precautions in healthcare settings [24]. More importantly, homologous recombination contributed more genetic variation to a large number of genes and largely maintained the genetic cohesion in Bcc. This high level of recombination between Bcc species blurs their taxonomic boundaries, which leads Bcc species to be difficult to distinguish phenotypically and genotypically [25].

The mallei group is a *burkholderia* group comprised of closely related species, mainly *B. mallei* and *B. pseudomallei*, sharing 99% identity in conserved genes. *B. mallei* has undergone genomic reduction, likely evolving from *B. pseudomallei* after infecting an animal host[26]. It lacks genes necessary for survival in the soil and exhibits characteristics suitable for an intracellular lifestyle [27]. *B. mallei*'s genome is composed of two circular chromosomes, with chromosome 1 housing metabolism-related genes and capsule formation information, while chromosome 2 contains virulence-associated genes and secretion systems[26]. The organism is resistant to various antibiotics, with no available vaccine for humans or animals [28]. *B. pseudomallei*, on the other hand, has the ability to invade cells[29], polymerize actin, and spread from cell to cell, causing cell fusion and multinucleated giant cell formation [30]. *B. pseudomallei* is intrinsically resistant to several antimicrobial agents through its efflux pump mechanism [31].

The Plant Pathogen Group within *Burkholderia* encompasses various strains known for their ability to cause diseases in plants. These strains possess specific mechanisms and virulence factors that enable them to colonize host plants, leading to infections and subsequent damage. *B. gladioli*, for example, causes decay in onion bulbs and rice [32]. The Bcc and mallei groups, along with the plant pathogens, make up the 3 big categories of the *Burkholderia* genus, with the Bcc being the more variable group with more complex interactions [33].

In this extended abstract, we showcase a summary of the thesis work:

- Gene-based pangenomic assays performed on data contained in an in-house genome database using the Pagoo framework;
- The construction and analysis of graph-based pangenomes and resulting structures;
- Some of the phylogeny work, mainly the construction of phylogenetic trees, both on gene-based and graph-based approaches.

This work aimed to advance the understanding of genomic diversity within the *Burkholderia* genus, employing a dual approach of gene-based and graph-based pangenomic analyses. By achieving these objectives, we seeked to provide valuable insights into the genetic makeup, phylogenetic relationships, and potential functional implications of the examined *Burkholderia* species. We aimed to contribute to the broader field of genomic research by offering methodologies and visualizations that enhance the exploration and comprehension of pangenomic data.

2. Methodology

The raw data used consisted in 90 *Burkholderia* strain Complete genomes belonging to 8

species (4 *B. gladioli* genomes, 5 *B. contaminans* genomes, 14 *B. multivorans* genomes, 15 *B. cepacia* genomes, 18
B. cenocepacia genomes, 16 *B. pseudomallei* genomes, 9 *B. mallei* genomes, 9 *B. thailandensis* genomes) in fasta format and their respective annotations in General feature format (GFF) downloaded from the the "Burkholderia Genome Database" [34, 35]. The specific information for each strain can be found in a publicly available table that is provided as part of this thesis [36] that contains each customized acronym given to each strain and their respective original name and download address.

Two SQLite Databases built in-house by the Biological Sciences Research Group (BSRG) group were supplied for the realization of this work. The first database, henceforth referred to as GenomeDB, comprises genomic information about each gene encoded in the 90 *Burkholderia* genomes under analyses in this work. The second database, henceforth referred to as BlastDB, comprises a BLASTp network containing amino acid sequence similarity information about the all-against-all pairwise comparisons between the proteins contained in the GenomeDB. The GenomeDB contains, most importantly, the unique internal identifier for each gene and the corresponding strain custom acronym, the gene sequence, coding sequences (CDS) annotation and family code for each BLAST query E-value, which ranged from E-30 to E-110. This data was already pre-calculated and it was provided as input for the work performed in this thesis.

Pagoo [7] is an encapsulated, object-oriented class system for analyzing bacterial pangenomes. We used the sqldf [37] R library with the help of scripting to pull the gene specific acronym, organism name, gene family and annotation for each *Burkholderia* strain from the GenomeDB SQL Database to produce a text file, containing all the strains' data for a single species, that will be Pagoo's input. A separate text file was created for each different E-value for the gene families present in this database. We also pulled the Gene sequences from the GenomeDB and created FASTA files that are also a required input to create a Pagoo pangenome object. To create a pangenome object, we issued the command `pg < pagoo(data, sequences)`, where "pg" is the pangenome object, "data" is the input text and "sequences" are the FASTA files. We utilized the `pg$summary_stats` command to generate info on the number of core and accessory genes in each pangenome. We chose the optimal E-value for each species, which corresponded to the highest core gene number in the summary statistics. We then used the commands `pg\$gg\_curves`, `pg\$gg\_barplot`, `pg\$gg\_pca` and `pg\$gg\_pie` on that dataset to generate figures for further analysis. This process was repeated for every species and for the following combinations of species: All Bcc species (*cenocepacia, cepacia, multivorans, contaminans*), all mallei group species (*mallei, pseudomallei, thailandensis*), mallei group + gladioli (*mallei, pseudomallei, thailandensis, gladioli*) and all eight species together.

We used various libraries within Pagoo's framework to perform further analyses of the pangenome objects, including core and pangenome size estimation, genomic fluidity calculation, gene neutrality testing and maximum likelihood phylogenetic tree building. All assays shown in the subsections below were performed as described in Pagoo's recipes page [38].

We calculated the Genomic fluidity of all "pg" objects using the R library "micropan" [39] applied on Pagoo's `pg\$pan\_matrix` (Matrix of the pangenome object, in which rows are organisms, and columns are groups of orthologous) using the `fluidity` command. The genomic fluidity is obtained using equation 1:

$$\phi = \frac{2}{N(N-1)} \sum_{\substack{k,l=1\ldots N \\ k<l}} \frac{U_k + U_l}{M_k + M_l} \qquad (1)$$

This equation calculates the probability of recombination between genetic elements (e.g., genes, alleles, or genomic regions) in a population of size N (i.e the number of genomes considered in the analysis). ϕ represents the probability of recombination between genetic elements (i.e fluidity). U_k and U_l represent the number of unique nucleotides (or in our case, gene families) in the kth and lth elements, respectively. M_k and M_l represent the total number of nucleotides (gene families) of the kth and lth elements, respectively. The summation term denotes a sum over all unique pairs of elements within the population (where k and l range from 1 to N and k¡l). This equation quantifies the genomic fluidity by considering the ratio of the sum of unique gene families to the sum of the total gene families in these elements. Higher values of ϕ (fluidity coefficient) suggest a higher likelihood of recombination between genetic elements, indicating increased potential for genetic exchange and mixing of genetic material within the population. If it is 1, the two genomes are non-overlapping. If it is 0, the two genomes contain identical gene clusters. The micropan library first calculates the genomic fluidity between 2 random genomes that is then averaged over N random pairs of genomes to obtain a population estimate. The default value for N was used (N=100). The difference between ge-

nomic fluidity and a Jaccard distance is small, they both measure overlap between genomes, but fluidity is computed for the population by averaging over many pairs, while Jaccard distances are computed for every pair.

We estimated the pangenome and core sizes of our pangenome objects with the R library "micropan" applied on `pg\$pan\_matrix` using the `binomixEstimate` command.

A binomial mixture model can be used to describe the distribution of gene clusters across genomes in a pangenome. The central idea is that every gene had a probability of being present in a genome. Genes who are always present are the core genes and have a probability of 1. The rest of the genes are present in a probability less than 1 because they are only present in a subset of the genomes. A binomial mixture model with "K" components estimates "K" detection probabilities. This model separates the pangenome in "K" categories that can be more than the commonly used 3 (Core,Shell,Cloud). To choose an optimal "K" value, `binomixEstimate` computes the Bayesian information criterion (BIC) criterion [40]. As the number of genomes become higher, the tendency is to observe an increasing number of gene clusters. When the 'K'-component binomial mixture has been fitted, the number of clusters not yet observed is estimated, and thereby the pangenome size. Also, as the number of genomes grows fewer core genes are observed. The fitted binomial mixture model gives an estimate of the final number of core gene clusters, i.e. those still left after having observed 'infinite' many genomes. The micropan command will output 2 tables. The first table presents the Core and pangenomes sizes estimated for each K value and respective BIC value. The lowest BIC value will correspond to the optimal K value. The second table shows the detection probabilities for each estimated category (= K), and the proportion of genes having that probability.

Let x_j be the number of genomes in which we observe domain gene family j in the pangenome matrix. Let y_g be the number of families found in g genomes (number of x_j's with value g). Then y_g is also a random variable. The probability density of this variable can be described by a K component binomial mixture model (2).

$$\theta_y = \sum_{k=1}^{K} \pi_k f\left(y; \rho_k\right), \quad y = 0, \ldots, g \tag{2}$$

where π_k is the mixing proportion (which sum to 1) and

$$f\left(y; \rho_k\right) = \binom{g}{y} \rho_k^y (1-\rho)^{g-y}, \quad k = 0, 1, 2, \ldots, K \tag{3}$$

is a binomial probability mass function with detection probability ρ_k. Summing y_1, y_2, ..., y_g we get the number of domain sequence families seen so far, i.e. the sample pangenome size. From the binomial mixture model we can also predict y_0, the number of families not yet seen, and in this way we can estimate the population pangenome size [41].

The final part of the estimation procedure is to find the proper number of components K in the binomial mixture, i.e how many binomial probability mass functions do we need to approximate the distribution of the observed data. The BIC selects the proper model complexity. Hence, we look for a K where

$$\mathrm{BIC}(K) = -2l(\pi, \rho|K) + (2K-2)log(n) \tag{4}$$

is minimized, where (2K - 2) is the number of free parameters in the model since the sum of mixing proportions is always 1 and the core component has a fixed detection probability ρ_1 [42]. n is the sample pangenome size.

DECIPHER was used to align the core genome at a level of 100%. Then, we applied the Tajima's neutrality test by using the "pegas" library command `Tajima.test` on the `pg$core_seqs_4_phylo` Pagoo object [43]. Tajima's D test is done by applying the formula below (Equation 5).

$$D = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{\pi - \theta}{\sqrt{\frac{a_1(n-1)}{2} + \frac{a_2(n^2+n+3)}{6(n+1)}}} \tag{5}$$

π represents the average number of pairwise differences between sequences in a sample. It measures the nucleotide diversity within a population. It is calculated as the average number of differences at a given site between pairs of sequences. θ: This symbolizes the population mutation rate, which is an estimate of the effective population size multiplied by the mutation rate per generation per base pair. It is an estimator of genetic diversity based on the number of segregating sites (sites in DNA sequences where at least two different nucleotides are present in the sample). n: This variable represents the sample size, i.e., the number of sequences or individuals in the sample. a_1 and a_2 : These are coefficients derived from population genetics theory and represent the sum of the inverse squares $(1/i^2)$ and the sum of the inverse squares of the differences $(1/i^2 - 1/i)$, respectively, where i ranges from 1 to $n-1$. a_1 and a_2 are coefficients derived from theoretical expectations under neutrality.

From the result of the test we produced a table with the number of genes considered as evolving neutrally (the default suggested values were -0.2 $<$ D $<$ 0.2). We also retrieved the top 10 lowest and highest Tajima scores for *B. cenocepacia*, and their corresponding gene family most common annotation.

It is worth noting that calculating a conventional "p-value" associated with any Tajima's D value that is obtained from a sample is impossible. Briefly, this is because there is no way to describe the distribution of the statistic that is independent of the true, and unknown, theta parameter (no pivot quantity exists).

A Maximum Likelihood phylogenetic tree was built for all pangenomes using the method implemented in the "Phangorn" Package [44]. The core genomes were first aligned using "DECIPHER" [45].

Anchorwave[8] was used to create a pairwise alignment for every pair of strains within every one of the eight studied *Burkholderia* species, using as input the processed version of the reference genome gene annotation in GFF3 format, and the query genome in FASTA format. AnchorWave extracted the full-length CDS from the reference genome using the reference genome and annotation. The start and end positions of the reference full-length CDS to the query genome were lifted over using GMAP [46], a splice-aware sequence alignment program. Anchorwave used its algorithm to identify collinear anchors, then aligned the base pair sequences within each anchor and inter-anchor, finishing with concatenating all alignments to generate the final alignment for each collinear block. Anchorwave then outputted the alignment in Multiple alignment format (MAF).

The alignment MAF files produced by Anchorwave were converted to Sequence alignment format (SAM) using the command `maf.convert` made avaialble by the the "Last" software [47]. SAM files were converted to BInary alignment map (BAM) format with samtools[48] `samtools view -bt` command and reconverted into SAM with `samtools view -h` command. Paftools' `sam2paf` command [49] converted the SAM files into pairwise mApping format (PAF) files. We moved All PAF and FASTA files into a folder of their own, we compressed and concatenated the FASTAs into a single file, then indexed them with samtools's `faidx` command. The PAF files were concatenated, and used as input along with the indexed FASTA files, to induce a graph in Graphical fragment assembly (GFA) format utilizing seqwish [9] and 10 threads.

Bubblegun [50] was used to detect bubble and superbubble chains in graphs in GFA format. Using the GFA as input, the commands `Bubblegun bchains --bubble_json` and `Bubblegun bchains --chains_gfa` were used to output a JavaScript Object Notation (json) file with information about the bubbles, and a GFA graph containing only the bubble chains, respectively.

We built the ODGI graph file using the `odgi build` command, using the GFA graph as input. With the odgi graph file we ran statistics using the `odgi stats` command, with -S and -W as options, which summarize the graph properties and shows the weakly connected components, respectively. We converted the GFA files to packed graphs (PG) with the `vg convert` command, then calculated the number of Sub-graphs with the `vg stats -s` command.

Novel command `odgi similarity` from the ODGI toolkit was used to create a distance matrix based on path or path-group similarity. The odgi graph file for *B. cenocepacia* was used as input and the "-d" option was used to provide distances(dissimilarities) rather than similarities. Since the odgi graph file contains a path for each contig (chromosome or plasmid), the "sed" command was used to integrate the delimiter # so that the name of the strain could be interpreted by the command `odgi similarity`. Then, this new odgi graph file, which combines paths from the same strain into a single path-group, was generated. The `odgi similarity` command was repeated for this new odgi graph file to provide a distance matrix between strains rather than contigs. The distance matrix files obtained contain various distances, mainly Jaccard distance, which was used to quickly convert into a phylogenetic tree in the R environment, by means of scripting.

3. Results & discussion

As explained in the methodologies, the GenomeDB contains clusters of genes based on gene family for many different E-values for the 8 species of the *Burkholderia* genus. Using Pagoo, we successfully generated a pangenome object for each species and for each E-value from 30 to 100. The methodology used for the creation of these pangenomes proved to be simple and fast. To perform further analyses and comparisons, we selected only the pangenomes for each species with the E-value that contained the highest number of core genes, i.e the largest core genome (Table 1. It was decided in this study to adopt an optimal criteria for the selection of the E-value threshold to obtain the clusters, the one E-value that maximized the number of clusters in the core. We are assuming the strains are highly related since they are from the same species, so high cores are expected. Indeed, the pangenomes obtained with Pagoo contained high and consistent

Table 1: Summary table containing all assays performed on the Pagoo pangenome objects, originated from the GenomeDB and BlastDB. K represents the optimal number of components obtained with the binomial mixture model of the pangenomes.

Species name	Optimal E-value	Core genes	Shell genes	Cloud genes	Mean fluidity	K	Estimated core size	Estimated size	No. neutral genes	% Neutral genes in core
B. thailandensis	80	3749	512	8754	0.233	5	1739	276978	16	0.4267%
B. pseudomallei	80	3906	928	19251	0.248	5	3299	1102389	1034	26.47%
B. multivorans	85	3495	1855	14126	0.258	4	3404	246089	1968	56.31%
B. mallei	80	3108	546	8640	0.259	4	2128	433570	306	9.845%
B. gladioli	90	5084	171	4541	0.218	3	2963	194694	4	0.07868%
B. contaminans	80	4532	1291	5004	0.233	3	4251	9366582	500	11.03%
B. cepacia	80	2890	3862	15473	0.318	6	1211	215631	442	15.29%
B. cenocepacia	80	3284	3653	18046	0.281	5	2554	238898	100	3.045%

core genomes.

We identified species with large cloud genomes, which is indicative of a high amount of Horizontal transfers, although with the amount of datasets studied, the species with highest cloud genomes corresponded to the ones with the largest datasets. To properly compare the sizes of the core, shell and cloud genomes directly, we would require: larger amount of datasets, a consistent number of datasets for all species.

Genomic fluidity is between 0.218 and 0.259 for all the pangenomes except *B. cenocepacia* and *B. cepacia*, which have the highest fluidity, the latter having 0.317, corroborating the higher variability of these two species. The average fluidity for the 8 species was 0.256; 0.273 for the Bcc; and 0.247 for the mallei group, indicating that the Bcc is more fluid than mallei group species. Overall, the gene clusters are much more identical than fluid, since they are much closer to 0 than 1, which is expected from single-species pangenomes of the same genus.

The pangenome size estimation test determines if our pangenomes could be explained by a different model than just the 3 categories named Core, Shell and Cloud. After fitting the estimation model to each pangenome, it is also able to estimate their size (Table 1). The Optimal number of categories is represented by the K value corresponding to the lowest BIC value calculated. For *B. gladioli* and *B. contaminans*, the pangenome is better represented by 3 categories, while *B. multivorans* and *B. mallei* result in a 4 category pangenome. The binomial mixture model suggests a number of 6 categories for *B. cepacia*. The rest of the Species are fitted to a 5 category model. All species then, with the exception of *B. gladioli* and *B. contaminans*, who have the smallest datasets, fitted a binomial mixture model with 4 or more components. This could imply that in this genus, an extra category for gene presence, besides core, shell and cloud, should be added to better fit the data.

Table 1 also contains the number of neutral genes and the percentage of the core genome those genes represent. *B. thailandensis*, *B. gladioli* and *B. cenocepacia*'s core genome contained less than 3% neutral genes, while *B. cepacia*, *B. contaminans* and *B. mallei* contained around 10-15% neutral genes. These species are very highly subject to environmental and selective pressure, which is understandable on a genus that is highly regarded as containing highly pathogenic specimens and capable of surviving many antibiotics. Over 50% of *B. multivorans* core genes are evolving neutrally, suggesting that *B. multivorans* is not evolving mostly through selective pressures but through random mutation events. This result was unexpected, as *B. multivorans* is part of the Bcc. Additional study on this case should be done, for example, testing the reliability of Tajima's D test.

We were able to build graph-based pangenomes for our eight species, through the use of a Anchorwave+seqwish pipeline. Anchorwave produced good pairwise alignments for *B. cenocepacia*. Although the process was orders of magnitude longer, using hardware specifications higher than the common computer setup at home, we believe our pipeline was successful in creating a variation graph and has potential for use in future work. Through VG and ODGI's statistical output we can compare the pangenome graphs (table 2). *B. pseudomallei* and *mallei*'s small graph sizes may indicate less complex graphs, and explain smaller distances between the genomes, which is corroborated by the gene-based pangenomes in Pagoo. *B. gladioli* and *B. cepacia* present the largest graph length. A graph sums up to many subgraphs, which are sections that do not align with the rest of graph. They were obtained with the PG file and it identified over 10 sub graphs for all species, with *B. cenocepacia* and *B. cepacia* having 38 and 36 sub graphs, respectively. The amount of sub-graphs in these species imply many portions of the genomes that did not align, such as plasmids. We calculated bubbles with Bubblegun, identifying variations in the graphs such as insertions, simple bubbles and super bubbles (Table 2). Bubblegun's run time for each pangenome was about one week. For most species, the bubble graphs contained a high number of bubbles, with over 2000 simple bubbles and 300 insertions, representing high amount of genetic variation, as bubbles represent alternative genotypes. *B. mallei* and *B. pseudomallei* were outliers by far, as these two species contained 38

Table 2: Output generated by `odgi stats`, `vg stats -s` and Bubblegun with a pangenome graph as input, for all 8 burkholderia species in the study.

Species name	Avg. genome size (bp)	Length	Nodes	Edges	Sub graphs	Paths	Simple bubbles	Super bubbles	Insertions
B. thailandensis	6.745.675	3.308.693	1.122.022	2.270.461	11	19	3263	3	357
B. pseudomallei	7.223.274	645.677	368.175	737.633	10	32	3	0	1
B. multivorans	6.649.370	2.629.212	1.437.430	2.880.174	22	46	2364	12	572
B. mallei	5.696.930	402.140	246.482	503.140	14	18	28	0	7
B. gladioli	8.741.840	10.692.791	5.965.340	12.698.168	20	18	9782	322	491
B. contaminans	8.587.033	4.816.475	3.335.895	7.411.740	12	20	5448	213	358
B. cepacia	7.981.127	14.657.785	3.504.760	6.857.892	36	49	-	-	-
B. cenocepacia	7.653.437	3.170.570	1.784.153	3.504.743	38	55	12191	2046	686

and 5 variations, respectively. This suggests that the strains of these 2 species are incredibly homozygous.

When it comes to phylogeny, we were able to perform the analysis on all fronts. All gene-based and the graph-based methods were able to create phylogenetic trees. Pagoo's framework allows for quick phylogeny assessments but also more time consuming maximum likelihood phylogenic assays. We performed the maximum likelihood assays for every pangenome and created a phylogenetic tree for each of them, still using the dataset with the optimal E-value. The Tree for *B. cenocepacia* can be observed in Figure 1(a). Pagoo's *B. cenocepacia* tree contains 3 clusters: Cluster A: Strains 3, 4, 5, 8, 11, 13, 16, 17; Cluster B: Strains 1, 2, 9, 10, 12, 14, 15, G1; Cluster C: Strains 6, 7, 18.

We calculated the distances between each path-group (i.e the set of paths that constitute a strain) in the variant graph for *B. cenocepacia* using ODGI. We created a phylogenetic tree from the jaccard distances obtained, represented in Figure 1(b). The graph-based pangenome similarity tree shows differences between clusters compared to the gene-based tree. Unlike Pagoo's tree that presented 3 clear clusters, the Graph-based tree presents much more scattered strains. Small Clusters can be observed such as between strains 7 and 8, strains 3 and 16, strains 5 and 15, strains 9 and 12 and strains 1 and 2. The significant differences in the trees can be explained by the graph-based tree being built from the entire pangenome dissimilarities, unlike the gene-based tree, which was made completely with the core alignment distances.

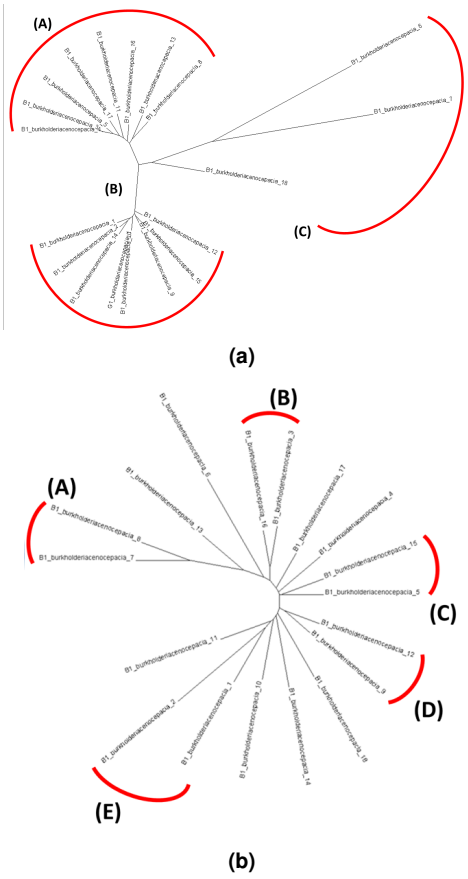

Figure 1: Comparison between the gene-based maximum likelihood and the graph-based Jaccard distance phylogenetic trees. (a): Maximum-likelihood *B. cenocepacia* phylogenetic tree. Cluster A: Strains 3, 4, 5, 8, 11, 13, 16, 17; Cluster B: Strains 1, 2, 9, 10, 12, 14, 15, G1; Cluster C: Strains 6, 7, 18. (b): *B. cenocepacia* phylogenetic tree from the graph-based pangenome using Jaccard distance. Cluster A: Strains 7 and 8; Cluster B: Strains 3 and 16; Cluster C: Strains 5 and 15; Cluster D: Strains 9 and 12; Cluster E: Strains 1 and 2.

4. Conclusions

Pangenomes are the new technological development to handle the continuously increasing number of genomic datasets and the consequent increasing unreliability of linear reference methods. The exploration of the different computer-based tools for building and analyzing pangenomes are essential to simplify the methodologies that often plague biologists who don't want to deal with learning the intricacies of informatics and programming languages to review their genomic data. By studying gene-based pangenomes and the more novel graph-based pangenomes, and applying these *in silico* tools on a common, but relevant bacterial genus such as *Burkholderia*, who are notable for their diversity and versatility, clinical importance and biotechnical potential, this work has contributed to the advance in both the methodology and in the unraveling of key genomic knowledge.

Acknowledgements

This thesis was developed and written during the second semester of 2022/2023, in the Department of Bioengineering, Instituto Superior Técnico, Lisboa, under the supervision of Dr. Paulo Jorge Moura Pinto da Costa Dias. I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

References

[1] G. S. Vernikos, "The pyramid of knowledge," *Nature Reviews Microbiology*, vol. 8, no. 2, p. 91–91, 2010.

[2] T. H. G. Project, "The human genome project faq."

[3] L. Rouli, V. Merhej, P.-E. Fournier, and D. Raoult, "The bacterial pangenome as a new tool for analysing pathogenic bacteria," *New Microbes and New Infections*, vol. 7, p. 72–85, 2015.

[4] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, "The microbial pan-genome," *Current Opinion in Genetics & Development*, vol. 15, no. 6, p. 589–594, 2005.

[5] T. G. P. Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, p. 56–65, 2012.

[6] J. M. Eizenga, A. M. Novak, J. A. Sibbesen, S. Heumos, A. Ghaffaari, G. Hickey, X. Chang, J. D. Seaman, R. Rounthwaite, J. Ebler, and et al., "Pangenome graphs," *Annual Review of Genomics and Human Genetics*, vol. 21, no. 1, p. 139–162, 2020.

[7] I. Ferrés and G. Iraola, "An object-oriented framework for evolutionary pangenome analysis," *Cell Reports Methods*, vol. 1, no. 5, p. 100085, 2021.

[8] B. Song, S. Marco-Sola, M. Moreto, L. Johnson, E. S. Buckler, and M. C. Stitzer, "Anchorwave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, 2021.

[9] E. Garrison and A. Guarracino, "Unbiased pangenome graphs," *Bioinformatics*, vol. 39, no. 1, 2022.

[10] E. Garrison, J. Sirén, A. M. Novak, G. Hickey, J. M. Eizenga, E. T. Dawson, W. Jones, S. Garg, C. Markello, M. F. Lin, and et al., "Variation graph toolkit improves read mapping by representing genetic variation in the reference," *Nature Biotechnology*, vol. 36, no. 9, p. 875–879, 2018.

[11] Z. Yang, A. Guarracino, P. J. Biggs, M. A. Black, N. Ismail, J. R. Wold, T. R. Merriman, P. Prins, E. Garrison, and J. de Ligt, "Pangenome graphs in infectious disease: A comprehensive genetic variation analysis of neisseria meningitidis leveraging oxford nanopore long reads," *Frontiers in Genetics*, vol. 14, 2023.

[12] B. Paten, J. M. Eizenga, Y. M. Rosen, A. M. Novak, E. Garrison, and G. Hickey, "Superbubbles, ultrabubbles, and cacti," *Journal of Computational Biology*, vol. 25, no. 7, p. 649–663, 2018.

[13] G. Hickey, D. Heller, J. Monlong, J. A. Sibbesen, J. Sirén, J. Eizenga, E. T. Dawson, E. Garrison, A. M. Novak, and B. Paten, "Genotyping structural variants in pangenome graphs using the vg toolkit," *Genome Biology*, vol. 21, no. 1, 2020.

[14] A. Guarracino, S. Heumos, S. Nahnsen, P. Prins, and E. Garrison, "Odgi: Understanding pangenome graphs," *Odgi: Understanding Pangenome Graphs*, 2021.

[15] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, and et al., "A draft human pangenome reference," *Nature*, vol. 617, no. 7960, p. 312–324, 2023.

[16] S. I. Paul, M. M. Rahman, M. A. Salam, M. A. Khan, and M. T. Islam, "Identification of marine sponge-associated bacteria of the saint martin's island of the bay of bengal emphasizing on the prevention of motile aeromonas septicemia in labeo rohita," *Aquaculture*, vol. 545, p. 737156, 2021.

[17] Z. Weinberg, J. E. Barrick, Z. Yao, A. Roth, J. N. Kim, J. Gore, J. X. Wang, E. R. Lee, K. F. Block, N. Sudarsan, and et al., "Identification of 22 candidate structured rnas in bacteria using the cmfinder comparative genomics pipeline," *Nucleic Acids Research*, vol. 35, no. 14, p. 4809–4819, 2007.

[18] S. Kunakom and A. S. Eustáquio, "burkholderia as a source of natural products," *Journal of Natural Products*, vol. 82, no. 7, p. 2018–2037, 2019.

[19] B. K. Okada, Y. Wu, D. Mao, L. B. Bushin, and M. R. Seyedsayamdost, "Mapping the trimethoprim-induced secondary metabolome of burkholderia thailandensis," *ACS Chemical Biology*, vol. 11, no. 8, p. 2124–2130, 2016.

[20] A. C. McAvoy, O. Jaiyesimi, P. H. Threatt, T. Seladi, J. B. Goldberg, R. R. da Silva, and N. Garg, "Differences in cystic fibrosis-associated burkholderia spp.. bacteria metabolomes after exposure to the antibiotic trimethoprim," *ACS Infectious Diseases*, vol. 6, no. 5, p. 1154–1168, 2020.

[21] Y. Jin, J. Zhou, J. Zhou, M. Hu, Q. Zhang, N. Kong, H. Ren, L. Liang, and J. Yue, "Genome-based classification of burkholderia cepacia complex provides new insight into its taxonomic status," *Biology Direct*, vol. 15, no. 1, 2020.

[22] E. Mahenthiralingam, T. A. Urban, and J. B. Goldberg, "The multifarious, multireplicon burkholderia cepacia complex," *Nature Reviews Microbiology*, vol. 3, no. 2, p. 144–156, 2005.

[23] E. Torok, E. Moran, and F. J. Cooke, *Oxford Handbook of Infectious Diseases and Microbiology*. Oxford University Press, 2017.

[24] J. E. Bennett, R. Dolin, M. J. Blaser, G. L. Mandell, and R. G. Douglas, *Mandell, Douglas, and Bennett's principles and practice of infectious diseases*. Elsevier / Saunders, 2015.

[25] J. Zhou, H. Ren, M. Hu, J. Zhou, B. Li, N. Kong, Q. Zhang, Y. Jin, L. Liang, and J. Yue, "Characterization of burkholderia cepacia complex core genome and the underlying recombination and positive selection," *Frontiers in Genetics*, vol. 11, 2020.

[26] G. C. Whitlock, D. Mark Estes, and A. G. Torres, "Glanders: Off to the races with burkholderia mallei," *FEMS Microbiology Letters*, vol. 277, no. 2, p. 115–122, 2007.

[27] L. Losada, C. M. Ronning, D. DeShazer, D. Woods, N. Fedorova, H. Stanley Kim, S. A. Shabalina, T. R. Pearson, L. Brinkac, P. Tan, and et al., "Continuing evolution of burkholderia mallei through genome reduction and large-scale rearrangements," *Genome Biology and Evolution*, vol. 2, p. 102–116, 2010.

[28] I. Fong and K. Alibek, *Bioterrorism and infectious agents a new dilemma for the 21st Century*. Springer US, 2005.

[29] W. J. Wiersinga, T. van der Poll, N. J. White, N. P. Day, and S. J. Peacock, "Melioidosis: Insights into the pathogenicity of burkholderia pseudomallei," *Nature Reviews Microbiology*, vol. 4, no. 4, p. 272–282, 2006.

[30] W. Kespichayawattana, S. Rattanachetkul, T. Wanun, P. Utaisincharoen, and S. Sirisinha, "burkholderia pseudomallei induces cell fusion and actin-associated membrane protrusion: A possible mechanism for cell-to-cell spreading," *Infection and Immunity*, vol. 68, no. 9, p. 5377–5384, 2000.

[31] T. Mima and H. P. Schweizer, "The bpeab-oprb efflux pump of burkholderia pseudomallei does not play a role in quorum sensing, virulence factor production, or extrusion of aminoglycosides but is a broad-spectrum drug efflux system," *Antimicrobial Agents and Chemotherapy*, vol. 54, no. 8, p. 3113–3120, 2010.

[32] M. Stoyanova, I. Pavlina, P. Moncheva, and N. Bogatzevska, "Biodiversity and incidence of burkholderia species," *Biotechnology & Biotechnological Equipment*, vol. 21, no. 3, p. 306–310, 2007.

[33] A. J. Mullins and E. Mahenthiralingam, "The hidden genomic diversity, specialized metabolite capacity, and revised taxonomy of burkholderia sensu lato," *Frontiers in Microbiology*, vol. 12, 2021.

[34] G. L. Winsor, B. Khaira, T. Van Rossum, R. Lo, M. D. Whiteside, and F. S. Brinkman, "The burkholderia genome database: Facilitating flexible queries and comparative analyses," *Bioinformatics*, vol. 24, no. 23, p. 2803–2804, 2008.

[35] "Burkholderia genome database." Accessed: 2023-12-1.

[36] "Thesis extra files - dbfinaltable." Accessed: 2023-12-1.

[37] G. Grothendieck, *sqldf: Manipulate R Data Frames Using SQL*, 2017. R package version 0.4-11.

[38] "Pagoo recipes." Accessed: 2023-12-1.

[39] L. Snipen and K. H. Liland, "Micropan: An r-package for microbial pan-genomics," *BMC Bioinformatics*, vol. 16, no. 1, 2015.

[40] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, 1978.

[41] L.-G. Snipen and D. W. Ussery, "A domain sequence approach to pangenomics: Applications to escherichia coli," *F1000Research*, vol. 1, p. 19, 2012.

[42] L. Snipen, T. Almøy, and D. W. Ussery, "Microbial comparative pan-genomics using binomial mixture models," *BMC Genomics*, vol. 10, no. 1, p. 385, 2009.

[43] E. Paradis, "pegas: an R package for population genetics with an integrated–modular approach," *Bioinformatics*, vol. 26, pp. 419–420, 2010.

[44] K. P. Schliep, "Phangorn: Phylogenetic analysis in r," *Bioinformatics*, vol. 27, no. 4, p. 592–593, 2010.

[45] S. Wright, Erik, "Using decipher v2.0 to analyze big biological sequence data in r," *The R Journal*, vol. 8, no. 1, p. 352, 2016.

[46] T. D. Wu and C. K. Watanabe, "Gmap: A genomic mapping and alignment program for mrna and est sequences," *Bioinformatics*, vol. 21, no. 9, p. 1859–1875, 2005.

[47] M. C. Frith, R. Wan, and P. Horton, "Incorporating sequence quality data into alignment improves dna read mapping," *Nucleic Acids Research*, vol. 38, no. 7, 2010.

[48] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li, "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, 02 2021. giab008.

[49] K. B. I. t. Royal Botanic Gardens, "pypaftol: Python module for paftol."

[50] F. Dabbaghie, J. Ebler, and T. Marschall, "Bubblegun: Enumerating bubbles and superbubbles in genome graphs," *Bubblegun: Enumerating bubbles and superbubbles in genome graphs*, 2021.