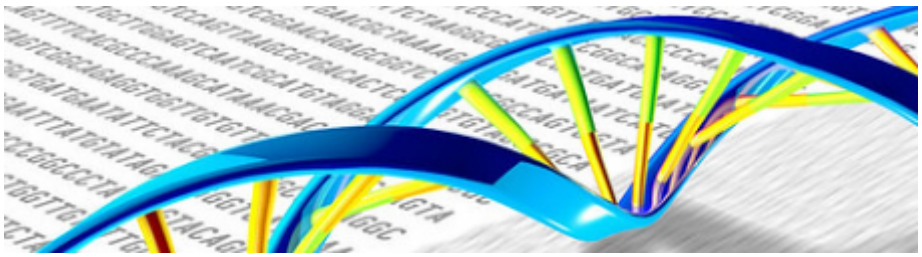
TÉCNICO LISBOA



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

Dinis Duarte Robalo Martins

Thesis to obtain the Master of Science Degree in

Biological Engineering

Supervisor: Dr. Paulo Jorge Moura Pinto da Costa Dias

Examination Committee

Chairperson: Prof. Gabriel António Amaro Monteiro
Supervisor: Dr. Paulo Jorge Moura Pinto da Costa Dias
Member of the Committee: Prof. Sílvia Andreia Bento da Silva Sousa Barbosa

December 2023

Declaration

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.

Acknowledgments

I would like to thank all my friends and family who've been with me every step of the way, you were very valuable, and I couldn't have done it without you.

I would also want to express my gratitude to Dr. Paulo Jorge Dias, my supervisor, who had the patience, flexibility and enthusiasm to work with me.

invictus maneo

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 different software suites - Roary, Panaroo and Pagoo - to create *Burkholderia* pangenomes. We compared the capability of those tools in creating good representations of genomic diversity, particularly in the identification of core genes. Using Pagoo, we generated a multi-species pangenome for the eight species. We calculated the genomic fluidity of *Burkholderia*, estimated the pangenome sizes, and identified neutral genes within the core genomes. We performed comparison assays between pangenomes, identifying unique core gene annotations and their corresponding metabolic pathways. 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.

Resumo

Pangenomas são objetos computacionais que capturam a diversidade genética num clado, competindo com genomas de referência única por lidarem melhor com o aumento exponencial de dados de sequenciamento de nova geração. *Burkholderia* é um género de bactérias gram-negativas altamente diverso, compreendendo mais de cem espécies incluindo microorganismos patógenicos e ecologicamente relevantes. Construímos pangenomas para oito espécies de *Burkholderia* para caracterizar os seus genomas e identificar elementos genómicos únicos para cada espécie. Pangenomas baseados em genes representam o conjunto dos genes essenciais e acessórios dum clado. Usámos vários *softwares* - Roary, Panaroo e Pagoo - para criar pangenomas de *Burkholderia*. Comparámos a sua capacidade de criar boas representações da diversidade genómica, particularmente a identificar genes *core*. Usando Pagoo, gerámos um pangenoma contendo as oito espécies. Calculámos a fluidez genómica de *Burkholderia*, estimámos o tamanho dos pangenomas e identificámos genes neutros contidos nos genes *core*. Fizemos análises comparativas entre pangenomas, identificando anotações de genes *core* únicos e as vias metabólicas correspondentes. Grafos de pangenoma representam sequências genómicas num formato facilmente navegável, em que os nós do grafo representam sequências alinhadas. Criámos pangenomas de espécie única: AnchorWave foi utilizado para alinhar o genoma par-a-par, e Seqwish foi utilizado para induzir os grafos. Identificámos variantes genómicas (bolhas) através de Bubblegun. Por último, criámos árvores filogenéticas para cada espécie utilizando a informação contida nos pangenomas. Este trabalho demonstra a utilidade de técnicas *in silico* como a pangenómica no estudo da diversidade genómica em espécies relacionadas, proporcionando uma visão sobre a origem das suas características distintas.

Palavras Chave

Bioinformática; espécies de *Burkholderia*; Modelo de grafo de variação; Pangenoma baseado em genes; Pangenoma baseado em grafo; Variação genética.

Contents

List of Equations

List of Figures

List of Tables

Acronyms

2D	Two Dimensional
ALN	Alignment Format
BAM	Binary Alignment Map
Bcc	Burkholderia cepacia complex
BED	Browser Extensible Data
BGC	Biosynthethic Gene Cluster
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
BLASTP	Protein Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
bp	base pairs
BSRG	Biological Sciences Research Group
BWA	Burrows–Wheeler Aligner
CDS	Coding Sequences
CNV	Copy Number Variation
COG	Cluster of Orthologous Genes
CPU	Central Processing Unit
DB	Database
DNA	Deoxyribonucleic Acid
GB	Gigabyte
GBWT	Graph Burrows–Wheeler Transform
GC	Guanine Cytosine

GFA	Graphical Fragment Assembly
GFF	General Feature Format
GML	Graph Modeling Language
HPRC	Human Pangenome Reference Consortium
indel	Insertion and Deletion
json	JavaScript Object Notation
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAF	Multiple Alignment Format
Mbp	Million base pairs
MC	Minigraph-Cactus
MCL	Markov Clustering algorithm
mRNA	Messenger Ribonucleic acid
MSA	Multiple Sequence Alignment
NGS	Next-Generation Sequencing
ODGI	Optimized Dynamic Genome Graph Implementation
OG	ODGI Graph
OS	Operating System
PAF	Pairwise mApping Format
PCA	Principal Component Analysis
PG	Path Graph
PGGB	Pangenome Graph Builder
PHB	Polyhydroxybutyrate
PNG	Portable Network Graphic
RAM	Random Access Memory
RNA	Ribonucleic Acid
SAM	Sequence Alignment Map
SNP	Single-Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SQL	Structured Query Language

STR	Short Tandem Repeat
SV	Structural Variant
T2T	Telomere-To-Telomere
TB	Terabyte
tRNA	transfer Ribonucleic Acid
TSV	Tab-Separated Values
TXT	Text
VCF	Variant Call Format
VG	Variation Graph
WFA	WaveFront Alignment

1

Introduction

Contents

1.1 Introduction

Next-Generation Sequencing (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.

An ideal pangenome possesses four fundamental attributes: completeness, encompassing all functional elements and adequate sequence space to serve as a reference for additional individuals; stability, featuring uniquely identifiable features enabling study by various researchers at different times; comprehensibility, facilitating understanding of genome complexities across individuals or species; and efficiency, organizing data to accelerate downstream analyses. Achieving these attributes presents challenges, including mechanisms for data sharing, organizational curation [7], and defining coordinate systems on pangenomes [8].

Pangenome studies for closely related taxa could potentially move beyond gene-level analysis, employing whole-genome multiple alignments (locally collinear blocks) or raw read datasets. This approach could unveil not only protein-coding sequences but also non-protein coding elements like promoters, small Ribonucleic Acids (RNAs), and repeat structures [9]. The decreasing average evolutionary distance between sequenced species predicts the release of thousands of genomes in upcoming years

through projects such as the Vertebrate Genomes Project/Genome 10K [10], Bat 1K, 200 Mammals, Insect 5K [11], and the Earth Biogenome Project [12]. Consequently, many comparative genomics projects will focus on aligning hundreds to thousands of closely related genomes rather than a few distantly related ones.

A major advantage of pangenome analysis, as demonstrated in humans, lies in the discovery of variants absent from a single reference genome. These missing variants can significantly impact both beneficial and harmful traits. For instance, in a human-sequencing project conducted in Iceland, a 766-base pairs (bp) insertion associated with reduced myocardial infarction risk was discovered [13]. Initiatives such as the Human Pangenome Reference Consortium (HPRC) aim to sequence and assemble a diverse set of individual genomes, culminating in a draft human pangenome that captures global genomic diversity across multiple individuals and haplotypes [14].

1.2 The *Burkholderia* genus

The *Burkholderia* genus is characterized by Gram-negative, obligately aerobic, rod-shaped bacteria with motility through polar flagella, except for *Burkholderia mallei* [15]. *Burkholderia* species do not produce sheaths or prosthecae and can utilize Polyhydroxybutyrate (PHB) for growth. These bacteria encompass both animal and plant pathogens, and they share a conserved RNA structure, the anti-hemB RNA motif. [16]. *Burkholderia* species exhibit a high average Guanine Cytosine (GC) content (66.73%), and an average genome size of 7.57 Million base pairs (Mbp). [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. Additionally, *Burkholderia* species exhibit a high degree of shared Biosynthethic Gene Clusters (BGCs) and distinct BGC counts, reflecting their metabolic potential [17, 19, 20].

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 [21, 22].

The Burkholderia cepacia complex (Bcc) is one of the groups part of the *Burkholderia* genus, including at least 20 different species [23]. The Bcc is often associated with pneumonia in immunocompromised individuals, particularly those with underlying lung diseases, such as cystic fibrosis [24]. 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 [25]. Person-to-person transmission has been documented, leading to strict isolation precautions in healthcare settings [26]. 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 [27]. The Bcc, along with other *Burkholderia* sub-groups, are shown in figure 1.1.

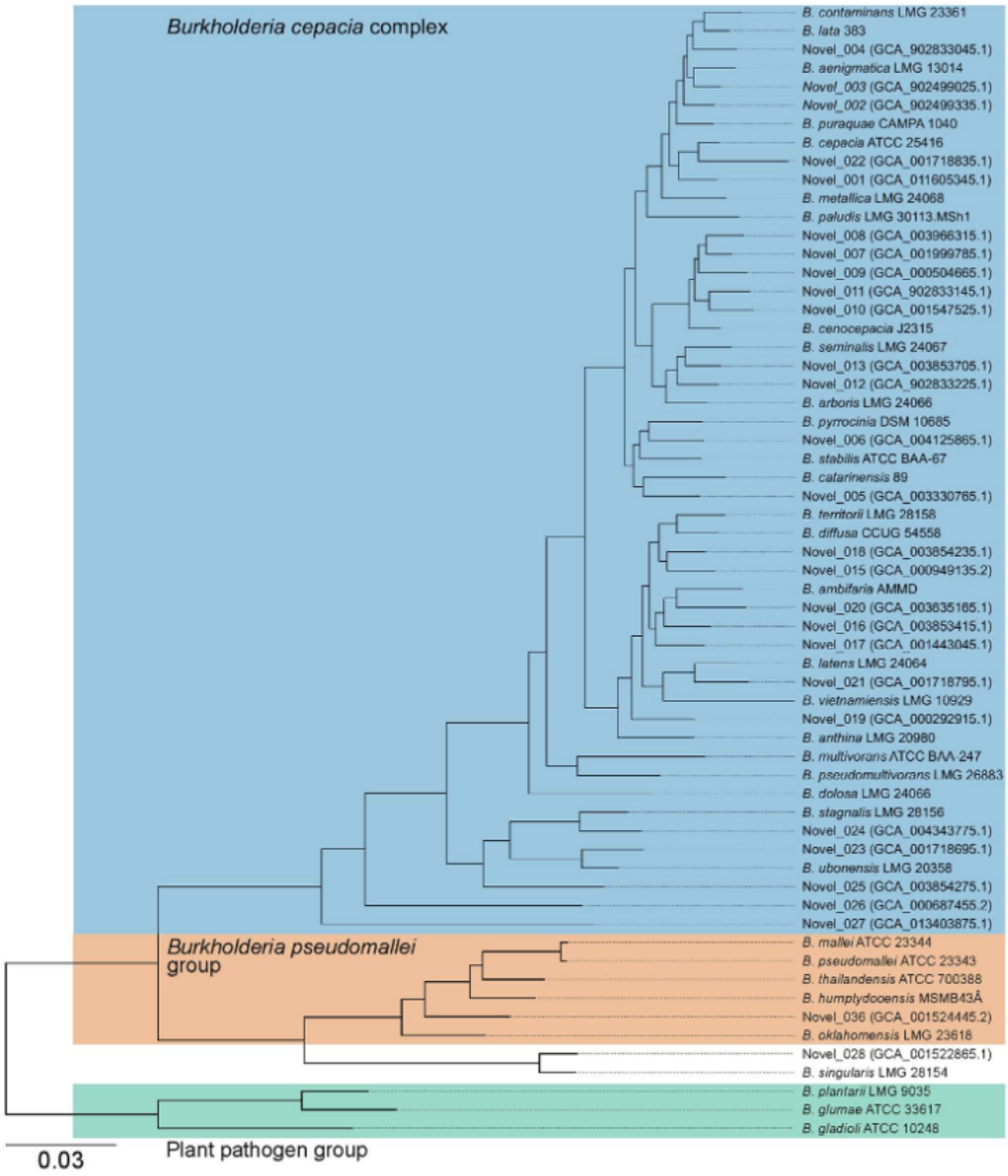


Figure 1.1: *Burkholderia* genus phylogenetic tree highlighting the 3 main groups: Bcc, mallei group and plant pathogen group. Figure courtesy of Mullins and Mahenthiralingam, 2021 [17].

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 [28]. It lacks genes necessary for survival in the soil and exhibits characteristics suitable for an intracellular lifestyle [29]. The genome of *B. mallei* 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 [28]. The organism is resistant to various antibiotics, with no available vaccine for humans or animals [30]. *B. pseudomallei*, on the other hand, has the ability to invade cells [31], polymerize actin, and spread from cell to cell, causing cell fusion and multinucleated giant cell formation [32]. It possesses a unique type VI secretion system and various toxins [33]. *B. pseudomallei* is intrinsically resistant to several antimicrobial agents through its efflux pump mechanism [34]. This mediates resistance to aminoglycosides (AmrAB-OprA), tetracyclines, fluoroquinolones, and macrolides (BpeAB-OprB).becA-R Is a exopolysaccharide biosynthetic gene cluster in *pseudoamallei* biofilm formation. Deletion of this gene cluster results in a significant decrease in biofilm formation [35].

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. Through a complex interplay of genetic traits and environmental cues, these pathogenic strains of *Burkholderia* can undermine plant health, potentially resulting in reduced crop yield, wilting, and other detrimental effects on agricultural productivity.*B. gladioli*, for example, causes decay in onion bulbs and rice [36]. 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 [17].

1.3 Concepts relevant to the understanding of pangenomics

1.3.1 Genome annotation

Genome annotation is the process of identifying functional elements within a genome assembly, encompassing various elements like protein-coding genes, noncoding transcripts [37], chromatin configuration [38], DNase hypersensitivity [39], CpG islands [40], and population variation [41, 42]. This essential task has been evolving since the mid-1990s when full-length genomes were first available [43]. Genome annotation methods typically fall into two categories: *ab initio* prediction, which employs statistical models to predict exon-intron structures computationally, and sequence alignment-based approaches that map expressed sequence tags, complementary Deoxyribonucleic Acid (DNA), or protein sequences onto assembled sequences to discover transcripts [44]. Some annotation pipelines integrate both methods to create comprehensive annotation sets [45, 46]. Advancements in sequencing tech-

nologies, including long-read [47] and linked-read technologies [48], have made high-quality genome assembly more affordable and accessible. This has led to the formation of consortia like the Vertebrate Genome Project, which aims to produce genome assemblies on a larger scale, fostering collaborative efforts in genomics research [12, 49]. Whole-genome annotation is the process of identifying features of interest in a set of genomic DNA sequences, and labelling them with useful information.

1.3.2 Orthologous versus paralogous homologous genes

Homologous genes are genes that share a common ancestry; they are derived from the same ancestral gene in a common ancestor. These genes may have similar sequences, functions, or both due to their evolutionary relationship. They provide valuable information about the evolutionary relationships between species and can help in understanding the conservation of genetic information and functions across different organisms. In bacterial pangenome analysis, a common challenge is categorizing the homologous genes into either orthologous or paralogous clusters based on shared sequence identity. Orthologs have a common ancestor linked to a speciation event, while paralogs originate from gene duplication events. It's essential to distinguish between these categories to identify core genes essential for basic biological processes (orthologs), explore genetic variations and adaptations within a species or strains (paralogs) and infer evolutionary relationships, speciation events, and divergence patterns among bacterial populations. To address this issue, many pangenome analysis programs incorporate location information to differentiate between paralogs. Additionally, they identify xenologs, which result from gene duplications acquired through horizontal gene transfer. This differentiation allows for a more comprehensive understanding of the functional and regulatory diversity of genes within bacterial pangenomes [50].

1.3.3 Open vs closed pangenome

The concept of an open versus closed pangenome is crucial in characterizing the gene repertoire of bacterial species. In an open pangenome scenario, there exists an extensive and indeterminate number of genes that can be discovered by sequencing additional genomes of the species. Conversely, a species with a closed pangenome implies that further genome sequencing efforts will not uncover new genes, as the complete gene repertoire of the species has been extensively cataloged, under the assumption that the sampling of sequenced strains is unbiased. These observations are drawn from extrapolations based on the analysis of the current dataset of bacterial genomes. To illustrate this, pangenome accumulation curves are often fitted to the power law equation (equation 1.1) [9].

$$y = kx^a \tag{1.1}$$

Where y represents the total number of genes discovered or observed in a bacterial species, x represents the number of sequenced genomes of that species, k is a constant multiplier that signifies the initial gene count or the intercept of the curve (in this context, it might represent the estimated number of genes present in the species initially sequenced), and a is the exponent or power in the power law equation, indicating the rate at which new genes are discovered as more genomes are sequenced. A smaller value of a implies a slower increase in the gene count with each additional sequenced genome, while a larger value of a indicates a more rapid increase.

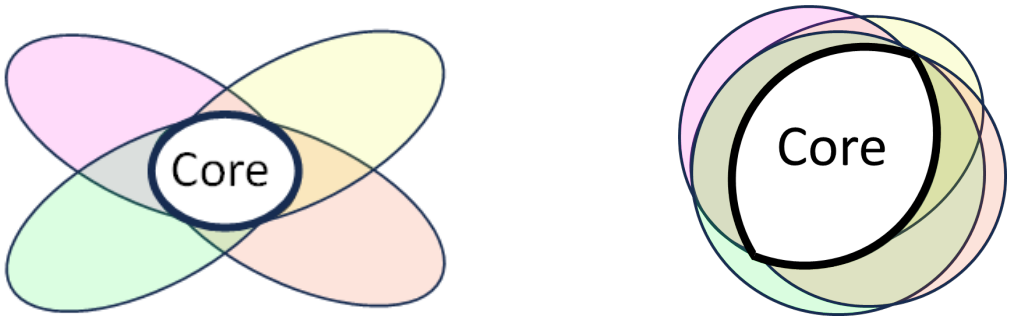


Figure 1.2: An example of an open pangenome (a), and a closed pangenome (b).

The open versus closed pangenome concept relates to the behavior described by the power law equation. In an open pangenome scenario (Fig. fig. 1.2a), where there is an extensive and indeterminate number of genes yet to be discovered, the Power Law equation accommodates this by allowing for a continual increase in the total number of genes y as more genomes x are sequenced. Conversely, in a closed pangenome scenario (Fig. 1.2b), where further sequencing efforts are not expected to uncover new genes, the Power Law equation might suggest a situation where the gene count y approaches a limit or stabilizes as the number of sequenced genomes x increases. This scenario is represented by an exponent a that might approach zero or result in a much slower growth rate of new genes as more genomes are sequenced.

1.3.4 Neutral evolution and gene neutrality

Gene neutrality and neutral evolution are fundamental concepts in evolutionary biology, essential for comprehending how natural selection and other evolutionary forces contribute to genetic diversity within populations. Gene neutrality refers to instances where a specific gene or genetic variant undergoes neither selective advantage nor disadvantage within a population. Essentially, the genetic variation at that locus does not significantly impact an organism's survival, reproduction, or competitive abilities in its environment. Such neutral genes are predominantly influenced by genetic drift (random changes in

allele frequencies due to chance events) and mutation, rather than selective pressures.

Neutral evolution posits that a substantial portion of observed genetic variation within populations arises from random mutations that do not affect an organism's fitness. These mutations typically occur within non-coding regions of the genome or within regions where changes in the DNA sequence do not alter the function of the encoded protein. Consequently, these mutations are considered selectively neutral and accumulate within populations primarily due to genetic drift, rather than being shaped by natural selection.

Two widely used tests to explore neutral evolution include the dN/dS ratio and Tajima's D test. The dN/dS ratio compares rates of nonsynonymous (dN) to synonymous (dS) substitutions in protein-coding genes. Nonsynonymous substitutions alter the amino acid sequence of a protein, while synonymous substitutions do not. A dN/dS ratio greater than 1 indicates positive or diversifying selection, favoring nonsynonymous mutations due to potential adaptive changes. Conversely, a ratio close to 1 suggests neutral evolution, where both types of mutations occur at similar rates, signifying minimal selective pressure. A ratio less than 1 indicates purifying or negative selection, where synonymous mutations are favored, suggesting functional constraints preserving the amino acid sequence [51, 52].

Tajima's D test aims to differentiate between DNA sequences evolving randomly ("neutrally") and those under non-random processes like selection or demographic events. A randomly evolving DNA sequence comprises mutations with no impact on an organism's fitness. Mutations under selection are considered "non-neutral." Tajima's D test helps identify scenarios such as directional or balancing selection, demographic changes, genetic hitchhiking (the process by which an allele (a particular variant of a gene) increases in frequency in a population due to its close association with a nearby gene undergoing positive natural selection), or introgression (the transfer of genetic material between species or distinct populations through hybridization and subsequent backcrossing.). A D value near 0 indicates neutrality, suggesting observed and expected variations align, and the population is likely evolving based on mutation-drift equilibrium, lacking evidence of selection. Values below 0 may signify rare alleles, recent selective sweeps, or population expansions following bottlenecks, while values above 0 might suggest scarcity of rare alleles, balancing selection, or sudden population contractions [53].

1.3.5 Phylogeny

Unambiguous phylogenomic trees of organismal or cellular lineages form invaluable input data for applications in various biomedical fields, for example to map the evolutionary dynamics of mutation patterns in genomes [54] or to understand the transfer of antibiotic resistance plasmids [55]. At the same time, the size of the pangenome often hampers the inference of such a 'tree of life' computationally as well as conceptually. One clear bonus offered by the pangenome, is that for traditional phylogenomics only the best aligned, and most well-behaved residues of a Multiple Sequence Alignment (MSA) can be

retained. In contrast, the pangenomic representation of multiple genomes allows for a clear encoding of the various genomic mutations in a model of the evolutionary events. This leads to the possibility for radical new evolutionary discoveries in fields including the origin of complex life [56], the origin of animals [57] and plants [58] or the spread of pathogens [59, 60], but also inferring the relationships between cancer lineages within a single patient [8, 61, 62].

1.3.6 Synteny

Genomic sequencing and mapping have enabled comparison of the general structures of genomes of many different species. The general finding is that organisms of relatively recent divergence show similar blocks of genes in the same relative positions in the genome. This situation is called synteny, translated roughly as possessing common chromosome sequences. For example, many of the genes of humans are syntenic with those of other mammals—not only apes but also cows, mice, and so on. Study of synteny can show how the genome is cut and pasted in the course of evolution. Shared synteny (also known as conserved synteny) describes preserved co-localization of genes on chromosomes of different species. During evolution, rearrangements to the genome such as chromosome translocations may separate two loci, resulting in the loss of synteny between them. Conversely, translocations can also join two previously separate pieces of chromosomes together, resulting in a gain of synteny between loci. Stronger-than-expected shared synteny can reflect selection for functional relationships between syntenic genes, such as combinations of alleles that are advantageous when inherited together, or shared regulatory mechanisms [63].

1.3.7 Genomic fluidity

Genomic fluidity refers to the dynamic and adaptable nature of genomes, describing the capacity of genetic material within an organism to change, evolve, and respond to various evolutionary pressures and environmental influences. It encompasses the ability of genomes to undergo alterations, including mutations, rearrangements, horizontal gene transfers, and genomic structural variations, often leading to genetic diversity and adaptation.

This concept suggests that genomes are not static entities but rather flexible and capable of undergoing modifications over time. Genomic fluidity accounts for the inherent plasticity of genomes, enabling organisms to adapt to changing environmental conditions, pressures from pathogens, and other evolutionary challenges. It encompasses mechanisms such as gene duplications, gene losses, mobile genetic elements, and genomic rearrangements that contribute to genome evolution and diversity.

Equation-based theoretical models, such as those in evolutionary biology and population genetics, can offer insights into explaining genomic fluidity. One such model is the mathematical framework for

measuring rates of unique gene families to the sum of gene families in pairs of genomes averaged over randomly chosen genome pairs from within a group of N genomes [64].

1.3.8 Online resources and databases - BLAST and KEGG

Basic Local Alignment Search Tool (BLAST) [65] is a powerful and widely used bioinformatics tool for comparing biological sequences. It helps in identifying regions of similarity between sequences, which might indicate functional, structural, or evolutionary relationships between them. BLAST is extensively used in genomics, proteomics, and other fields of molecular biology. BLAST compares a query sequence (could be DNA, RNA, or protein sequence) against a database of sequences to find similar regions. There are different variations of BLAST designed for specific types of sequences and purposes. Protein Basic Local Alignment Search Tool (BLASTP) specifically focuses on comparing protein sequences. BLASTP works by taking a protein sequence as the query and searches for similar sequences within a protein sequence database. It then uses its algorithm to find regions of local similarity between the query protein sequence and the sequences in the database. BLASTP uses a scoring system to assess the degree of similarity between sequences based on matches, mismatches, gaps, and other criteria. After the comparison, BLASTP generates a list of alignments or matches ranked by their similarity scores, helping researchers identify related or homologous proteins. BLASTP is widely utilized in various biological research areas, including functional annotation of proteins, studying evolutionary relationships between proteins, identifying conserved domains, and predicting protein structure and function based on similarities with known proteins in databases.

Kyoto Encyclopedia of Genes and Genomes (KEGG) [66] is a comprehensive resource that integrates information on biological systems, encompassing molecular-level details to higher-order functions in organisms. KEGG provides extensive databases and tools that aid in the understanding of biological pathways, genomes, diseases, drugs, and chemical substances. It contains various interconnected databases such as the ”KEGG PATHWAY”, cataloging molecular interaction networks in cells, illustrating various biological pathways including metabolic pathways, signaling pathways, and other complex biological processes. Scientists, researchers, and bioinformaticians often utilize KEGG to study and analyze biological systems, pathways, and the relationships between genes, proteins, diseases, and drugs. It serves as a valuable resource for understanding the molecular mechanisms underlying biological processes and diseases, aiding in various research endeavors and drug discovery efforts.

1.4 Computational pangenomics

1.4.1 Gene-based pangenome

The concept of pangenome reconstruction is essential in understanding the molecular evolution of bacterial populations, driven by factors such as horizontal gene transfer, changes in population size, and colonization of new environments [4, 9]. This concept is particularly relevant due to the substantial intraspecific diversity observed in bacterial genomes. Pangenome reconstruction typically involves the annotation of genes in a collection of whole-genome sequences, which are then grouped into orthologous clusters based on different similarity criteria [67]. The pangenome data provide insights into the membership of each gene in each genome, categorizing them into the core genome, the shell, and the cloud of rare and unique genes [50]. In this context, the pangenome represents the entire genomic repertoire accessible to a phylogenetic clade, which can range from species to higher taxonomic levels. It encompasses core genes found in all strains of the clade (or a high percentage usually above 95%), moderately conserved genes in the shell (usually between 15% and the core threshold), and rare or unique genes in the cloud ($<$15%). These gene classes play roles in the fundamental biology of the clade, species diversity, and various functions like adaptation, antibiotic resistance, or host colonization [68, 69].

Pangenome construction is influenced by factors including alignment algorithms, phylogenetic resolution, the sample of input genomes, modeling estimates of new genes versus the number of genomes, sequence annotation quality, and the level of comparison among all genomes (all-against all sequence similarity vs. phyletic profile of gene presence/absence regardless of sequence similarity). The pangenome concept challenges traditional species definitions [70] and classical phylogenetic tree-like structures [71]. As microbial genome fluidity becomes more evident, strictly bifurcating trees struggle to represent the complex relationships among species characterized by high rates of DNA exchange. [9, 72–74]. In response to these challenges, a phylogenetic network is suggested as a more suitable representation to capture the dynamic relationships shaping microbial evolution. The transition from single-gene phylogenies to multilocus sequence typing and, more recently, pangenomes provides a more comprehensive and reliable understanding of bacterial population phylogenetics. The pangenome concept is a fundamental framework in comparative genomics that advances our understanding of microbial evolution and challenges traditional species definitions, favoring a more realistic phylogenetic network approach.

1.4.2 Graph-based pangenome

Pangenome graphs serve as a fundamental framework for comprehending and dissecting genetic diversity within species. They provide a flexible structure to represent genomic sequences, variations, and inter-genomic relationships, crucial in overcoming reference bias and facilitating comprehensive ge-

netic variation analysis. These graphs, a form of sequence graph, depict multiple genomic sequences, enabling visualization and analysis of intricate structural variations (figure 1.3). Initially used for representing multiple sequence alignments, these graphs have expanded into genome assembly, capturing information from sequencing reads or fixed-length k-mers.

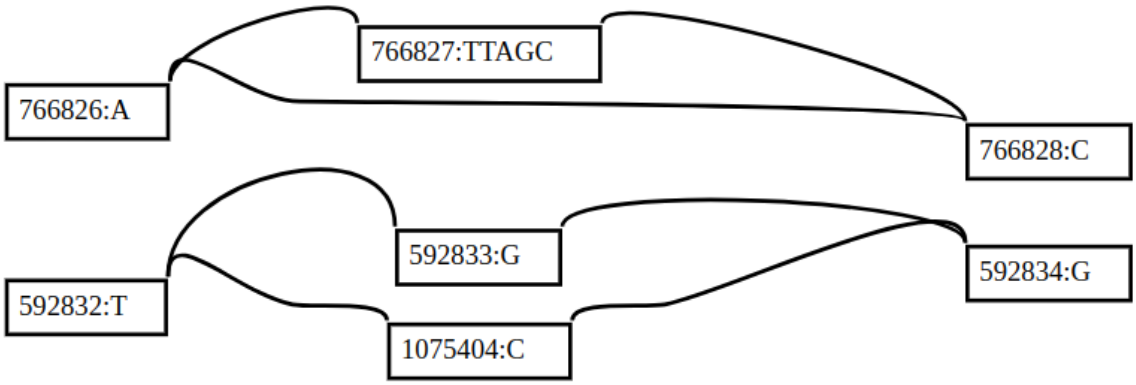


Figure 1.3: Small section of the *B. multivorans* bubble graph constructed in this work.

An extension of sequence graphs, genome graphs focus on illustrating whole-genome relationships, representing recombination events within included genomes. Further evolving this concept, variation graphs embed linear sequences as paths, offering a stable coordinate system supporting positional annotations and alignments between variation graphs and linear reference genomes. This embedding of reference sequences within graphical pangenomes establishes a coordinate system, facilitating coordinated analysis.

Eukaryotic pangenomes encompass variations in intron-exon structures and repetitive DNA sequences, differing from prokaryotic pangenomes that primarily focus on genes [75–78]. Pangenome graphs mitigate reference bias inherent in linear reference genomes, addressing inaccuracies or incompleteness in variant identification, especially in complex genomic regions [79]. They prove valuable in accurately genotyping Structural Variants (SVs) and representing nested variations within larger insertions, scenarios challenging to capture using linear references [6].

Infectious disease research benefits significantly from pangenome graphs by identifying novel genetic variants and structural variations. This aids in understanding antibiotic resistance, pathogenicity evolution, and viral genomics, providing insights into virus origin, spread, and guiding vaccination strategies [4, 27, 80, 81]. Additionally, they enable whole-genome multiple alignments and enhance variant calling across diverse genomes [82]. Bubbles within the graph effectively depict structural variants.

Microbial genomes, given their relatively small size and the availability of multiple fully closed genome sequences for many species, present an excellent opportunity for constructing pangenomes [83, 84]. Earlier pangenome studies predominantly focused on the gene level [85], but current data availability allows for sequence-level pangenomes. These encode complete sequence information of numerous

individual strains [86]. Considering microbial pangenomes aids in comparative genomics due to potential horizontal gene exchange, often leading to different phylogenetic relationships within genomes [86, 87].

However, constructing pangenome graphs poses computational challenges, raising questions about complexity and usability trade-offs. Incorporating variation into the reference system may increase ambiguity, demanding careful inclusion of relevant variation to improve utility. Moreover, working with pangenome graphs necessitates familiarity with graph-theoretic concepts uncommon among biologists. The limitations emphasize the need to deliberate on trade-offs and challenges associated with pangenome graph implementation in genomics research.

Variation graphs, resembling train tracks, implicitly incorporate their reverse complements, allowing representation of inversions without duplicating the inverted sequence. This minimizes duplication of annotations and information about variation. These graphs, while complex, offer a standard model matching the Graphical Fragment Assembly (GFA)-encoded model, ensuring versatility in usage.

1.4.2.A Alignment formats

The representation of genome alignments is a fundamental aspect of genomics, and it poses specific challenges due to the potential for rearrangements and duplications in genomes. Collinear alignment formats, such as aligned-FASTA, are inadequate for accommodating these complex genomic events as they solely depict alignments through insertions, deletions, and substitutions. At present, the prevailing format for genome alignments is Multiple Alignment Format (MAF). MAF can effectively capture referenced multiple alignments, even when they involve rearrangements in the genomes. However, MAF is structured in a column/block-oriented fashion, which presents limitations in representing intricate orthology relationships in a reference-free manner [88]. Other prominent formats used in genomics include the Sequence Alignment Map (SAM) and its binary counterpart Binary Alignment Map (BAM). SAM is a tab-delimited text format for storing sequence alignment data, whereas BAMis the binary representation of the same information, optimized for efficient storage and retrieval. These formats are widely utilized for storing read alignments against reference genomes in various sequencing applications due to their versatility and compactness [89]. Additionally, the Pairwise mApping Format (PAF) has gained attention for representing pairwise mappings between sequences. Unlike MAF, PAF is more focused on representing pairwise alignments and is commonly used in applications where aligning long sequencing reads to a reference or between different long reads is crucial, such as in long-read sequencing technologies like PacBio or Oxford Nanopore. Each of these formats — MAF, SAM, BAM, and PAF — serve distinct purposes in genomics, catering to different needs in representing and storing genome alignment data, depending on the complexities and specific requirements of the analyses involved.

1.4.2.B Graph and graph index formats

To facilitate the exchange of pangenome graphs, a subset of the GFA format is commonly used within the community. This format, initially designed for assembly graphs, serves as a practical means to represent pangenome graphs while allowing compatibility with many genome assembly tools. GFA is a textual format for labeled graphs. The Graph Burrows–Wheeler Transform (GBWT) plays a significant role in the efficient storage and indexing of variation graphs, focusing on maintaining data locality and minimizing global information in favor of local, vertex-based data. The GBWT accomplishes this by storing sets of paths, with the variation graph being inferred from these paths. Each path is represented as a sequence of vertex identifiers, rather than storing the labels. The essence of the variation graph is thus a collection of such strings. This data structure has broad applications in haplotype analysis and variant genotyping at the population level. In addition to GBWT, several other graph-based index formats have emerged, each with its specific focus and utility. The succinctly named succinct representation for the variation graph (XG) format provides a compact way to store variation graphs and associated metadata. It's optimized for fast random access and enables efficient querying of graph structures and paths within the graph. The Variation Graph (VG) format, on the other hand, allows for the representation of complex genome variation in the form of bidirected graphs, enabling alignment and analysis of diverse genomes and genomic structures. Furthermore, the Path Graph (PG) format emphasizes the encoding of genomic variation through paths, enabling the representation of multiple genomes and their alterations. PG is particularly useful for exploring and analyzing the structural variations among multiple individuals or populations. Alongside these formats, the Optimized Dynamic Genome Graph Implementation (ODGI) toolkit introduces the ODGI Graph (OG) file format, offering efficient handling and manipulation of overlay and De Bruijn graphs. ODGI files contain graph-related information utilized within the toolkit for scalable representation and processing of complex genomic graph data. While GBWT excels in haplotype-aware indexing and its applications in personalized medicine and RNA-seq data analysis [90], XG, VG, PG, and OG formats each offer distinct advantages in representing and analyzing variation graphs and genomic structures. They find utility in diverse applications, ranging from understanding population-level variations to elucidating complex relationships between genomic elements, thus contributing significantly to the field of genomics and bioinformatics.

1.4.2.C Structural variants, bubbles and variant call format files

SVs, which are genomic mutations involving 50 or more bps, encompass diverse forms like deletions, insertions, inversions, translocations, or complex events, often exerting a greater influence on phenotype than smaller mutations. These SVs have long been associated with developmental disorders, cancer, and other complex diseases [82]. However, SVs have been less studied compared to Single Nucleotide Variants (SNVs) and small Insertions and Deletions (indels) due to the limitations of short-read sequenc-

ing. Short-read sequencing struggles with SV detection because of challenges in mapping reads to the reference genome, particularly when substantial differences exist between the sample and the reference. SVs, which are often found in repeat-rich regions, further complicate read mapping. Short reads are more suitable for genotyping known SVs, but their use is limited by the cost and scalability issues associated with large-scale studies [83, 84]. The representation of known SVs presents challenges. The standard Variant Call Format (VCF) is suboptimal for expressing nested or complex SVs. Integrating SVs into a linear pangenome reference via alternative contigs also increases mapping ambiguity and scalability concerns. Pangenomic graph reference representations offer an attractive solution for storing genetic variations, as they seamlessly represent both SVs and point mutations using the same framework. By incorporating known variants in the reference, read mapping, variant calling, and genotyping become variant-aware, enhancing accuracy and sensitivity. This coherency allows different types of variants to be called and scored simultaneously in a unified framework [91, 92].

A graph-centric approach in characterizing variant sites within pangenome graphs introduces the concept of defining sites based on motifs inherent in the genome graph structure. To classify variants in pangenome graphs, graph decomposition is employed to identify "bubble" subgraphs that correspond to non-overlapping variant sites. These variant sites are then categorized into small variants ($<$50bp) and SVs ($>$50bp) of different types. Two motifs, "superbubbles" (in directed graphs) [93] and "ultrabubbles" (a generalization for bidirected graphs) [82], have been proposed for this purpose. Superbubbles and ultrabubbles represent directed acyclic sub-graphs that connect to the rest of the graph through a single source node and a sink node. These motifs often emerge when new variants are added to the graph and are particularly useful in identifying nesting and overlapping relationships between structural variants, offering a more comprehensive site definition compared to linear reference coordinates. Paten et al.(2017) [82] demonstrate how the nesting of these sites can be naturally organized using a Cactus graph, a structure that does so globally without the reliance on an existing reference genome. However, it's important to note that this approach does not perfectly partition all variation into these bubbles, and it lacks the simplicity of linear reference coordinates.

More recently, Bubblegun [94], a tool for detecting Bubbles and Superbubbles in De Bruijn graphs, shows promise when compared to VG's snarl detection, especially in bigger graphs, in time efficiency, while maintaining the same level of efficacy.

1.5 Software tools for pangenomics

1.5.1 Prokka annotation

Prokka [95] is a versatile software tool that has emerged as a prominent solution for the rapid and accurate annotation of bacterial genomes. Prokka represents a valuable annotation tool designed

specifically for prokaryotic genomes, offering an automated and high-throughput annotation pipeline. Prokka integrates various bioinformatic algorithms and databases to annotate genomic features, including protein-coding genes, non-coding RNAs, and other elements such as CRISPR arrays and plasmids. Its core functionality involves the prediction of Coding Sequences (CDS) using software like Prodigal, followed by the assignment of putative functions through homology searches against multiple databases (e.g., UniProtKB, Pfam, and TIGRFAM). Furthermore, Prokka employs HMMER for domain identification and integrates tools for tRNA, rRNA, and CRISPR detection [96, 97]. Prokka's modular design and compatibility with commonly used file formats enhance its flexibility and usability. It has gained widespread popularity due to its speed, accuracy, and user-friendly interface, making it accessible to both bioinformaticians and bench scientists. Comparative evaluations against other annotation tools have demonstrated Prokka's robust performance in terms of accuracy and computational efficiency. Moreover, its ability to handle large-scale genome datasets with minimal manual intervention makes it particularly advantageous for high-throughput genomic studies. While Prokka excels in prokaryotic genome annotation, certain limitations exist, such as potential challenges in the annotation of novel or divergent genes lacking homologs in existing databases. Continuous updates and improvements to its underlying databases and algorithms could enhance Prokka's annotation accuracy and expand its applicability to a broader range of organisms.

1.5.2 Software dedicated to the construction of gene-based pangenomes

Various approaches are available to infer the Gene-based pangenome of a collection of bacterial isolates, including tools like Roary, OrthoMCL, PanOCT, PIRATE, PanX, PGAP, COGsoft, MultiParanoid, PPanGGoLiN, and MetaPGN [98, 99]. These methods predominantly employ two similar strategies to determine the pangenome. The process commonly begins by assessing similarity between predefined gene sequences through homology search tools like CD-HIT [100], BLAST [65], or DIAMOND [101, 102]. Subsequently, a pairwise distance matrix is constructed, and genes are grouped into orthologous clusters. This clustering is achieved either using the Markov Clustering algorithm (MCL) or by examining triangles of pairwise best hits [103, 104]. Some of these methods also incorporate gene adjacency information to construct a graphical representation of the pangenome. This graph is utilized to further differentiate orthologous clusters into paralogs. Roary, PIRATE, PPanGGoLiN, and MetaPGN also provide this graphical representation as an output file. In some pipelines, a final step involves classifying the resulting clusters into core and accessory categories based on their prevalence within the dataset. This categorization typically relies on predefined thresholds; however, recent approaches have suggested model-based extensions to this process [50, 105].

Panaroo [50] is an innovative gene-based pangenome clustering tool designed to tackle the challenges associated with prokaryotic genome assembly annotation, particularly errors, contamination,

and fragmented segments. It adopts an algorithm that harnesses information from multiple genomes to enhance annotation precision and the clustering of orthologs and paralogs within the pangenome. Its primary strength lies in its capacity to correct errors introduced during annotation, filter out contamination, merge fragmented gene segments, and rediscover missing genes. The key feature of Panaroo is its ability to construct a full graphical representation of the pangenome, where nodes represent Clusters of Orthologous Genes (COGs) , and edges connect nodes if they appear adjacent on a contig in any sample from the population. Panaroo offers users the flexibility to choose from predefined threshold modes, including 'strict' and 'sensitive.' In 'strict' mode, the tool aggressively eliminates contamination and erroneous annotations, making it ideal for scenarios where the focus is not on rare plasmids or when stringent control over erroneous clusters is essential. Conversely, 'sensitive' mode retains all gene clusters and is suitable for studying rare plasmids that might be challenging to distinguish from contamination. Panaroo produces diverse output formats, such as a gene presence-absence matrix and a fully annotated graph in Graph Modeling Language (GML) format for visualization in software like Cytoscape [106]. Furthermore, it seamlessly interfaces with various pangenome analysis packages, facilitating the exploration of associations between phenotypes, gene presence/absence, and structural variations within the graph. When assessed alongside other pangenome tools, Panaroo consistently excels by accurately identifying core genes while yielding more conservative estimates for the accessory genome. This high performance holds even in the presence of contamination and fragmented assemblies, making Panaroo a standout choice for researchers seeking lower error rates and precise core and accessory genome reconstructions. The Panaroo algorithm builds a comprehensive graphical representation of the pangenome. It uses gene adjacency to rectify errors introduced during genome annotation, accepting annotated assemblies in General Feature Format (GFF)3, as output by Prokka [95]. Notably, Panaroo aims to preserve the complete global context of each gene within the graph, differing from some other pangenome software that relies solely on local gene context, like Roary [50].

Pagoo is a versatile pangenome post-processing tool designed for the standardized analysis of pangenome data generated by various pangenome reconstruction software tools. It is built on an object-oriented design within the R programming environment, offering several key features and advantages. First and foremost, Pagoo provides a well-structured data storage format for pangenome information, encompassing orthologous clusters, sequences, annotations, and metadata within a single object that can be easily shared and saved as a single file. This object is highly adaptable, with user-friendly methods for querying, handling, and subsetting the data. In addition, Pagoo incorporates a range of standard statistical analyses and dynamic visualizations, making it a powerful platform for downstream comparative analyses in bacterial population genomics. One distinctive aspect of Pagoo is its use of an encapsulated, object-oriented design based on the R6 package. This approach integrates data and methods, with the Pagoo object built on three R6 classes: PgR6, PgR6M, and PgR6MS. These classes

offer different levels of functionality, from basic data handling and subsetting to advanced statistical methods, visualization tools, and biological sequence manipulation. Moreover, third-party applications can easily inherit and extend these classes to accommodate specific needs. Another noteworthy feature of Pagoo is that it allows users to interact with the pangenome object without altering the underlying raw data. Active bindings enable dynamic modifications to the object, such as temporarily hiding specific organisms, adjusting core gene definitions, or extracting specific information from genes, clusters, or sequences. Class-specific methods for generic subset operators facilitate the extraction of relevant data fields from the object using standard R notation. Pagoo's interactive application provides a rich environment for statistical analysis and visualization. Users can generate customized plots and perform statistical analyses directly from the pangenome object using active bindings or through a built-in R-Shiny application. This application includes a general dashboard with interactive summary statistics and a specific dashboard for exploring evolutionary trends, including accessory gene distances, Principal Component Analysis (PCA), gene presence/absence matrices, and more. Using an interpreted language like R for post-processing offers significant advantages, allowing users to seamlessly transition between data exploration and in-depth analysis within a single environment. One of the key features of Pagoo is the ability to define the core level, which determines the minimum percentage of genomes in which a gene must be present to be considered a core gene. By default, Pagoo sets the core level at 95%, but users can modify it to create more or less stringent core genome definitions, thereby affecting the state of the pangenome object and resulting in different core, shell, and cloud gene sets. Pagoo addresses a current limitation in pangenome data analysis tools by providing an end-to-end solution within a single framework. Unlike other tools that produce visualizations but do not allow users to return to the data for further analyses, Pagoo offers a comprehensive and customizable platform for pangenome analysis.

1.5.3 Software dedicated to the alignment of genome sequences

1.5.3.A Local genome alignment tools

Local genome alignment tools are essential in genome alignment, as genome alignment tools frequently depend on local alignments. Due to the impractical time and memory requirements of finding all-against-all optimal local alignments, genome alignment typically employs approximate local aligners akin to BLAST [65]. These aligners operate by identifying exact matches referred to as "seeds," which may encompass positions that are allowed to vary for increased sensitivity. The alignment is then extended from these seeds. Notably, local aligners for genome alignment differ from read aligners likeBurrows–Wheeler Aligner (BWA) [107]. Genome alignment local aligners must handle greater evolutionary distances, as opposed to read aligners optimized for aligning reads to nearly identical reference

genomes. Several local alignment tools are in use. BLAST-Like Alignment Tool (BLAT) [108] is a popular, fast local alignment tool suitable for short evolutionary distances and is equipped to manage longer evolutionary distances with its "translated BLAT" mode. BLASTZ [109] and its successor, LASTZ [110], are designed for increased sensitivity compared to normal BLAST, employing Pattern Hunter-like spaced seeds and allowing transitions. Similarly, LAST [111] is another sensitive aligner, capable of using smaller seeds while efficiently handling partial matches. Smith-Waterman and Needleman-Wunsch algorithms are capable of producing alignments with a fixed order and orientation, limiting edit operations to insertions, deletions, and substitutions. While these algorithms are suited for short or well-conserved sequences like genes, they prove insufficient for large evolutionary distances and extensive genomic regions. Genomes often exhibit complex rearrangements like inversions, transpositions, and duplications, which disrupt order and orientation. Therefore, constant order and orientation constraints in alignments fail to capture these rearrangements. The impractical runtime of global alignment algorithms like Needleman-Wunsch and Smith-Waterman led to the development of tools that produce approximately optimal global alignments. These tools employ high-confidence anchors within a single order and orientation, partitioning the alignment into smaller, more efficiently solvable problems. However, this approach lacks the flexibility to detect rearrangements due to its reliance on constant order and orientation, which is unsuitable for capturing complex genomic variations.

1.5.3.B Whole-genome alignment tools

Most genome aligners, at a high level, work in two stages: filtering, in which a large number of local alignments are generated and filtered down to remove spurious false-positive alignments and identify homologous, rearrangement-free regions [112], and refinement, in which the homologous regions undergo alignment with a collinear aligner. (Some aligners keep a subset of the original local alignments as anchors to be included in the final alignment, whereas others throw away all the original local alignments and align the rearrangement-free regions from scratch.) The filtering step can take many different forms, but many involve constructing a graph representation of the alignment and using various heuristics to simplify the graph (for a review, see [113]).

Here we list some software of the 2 types of whole-genome alignment tools:

1. **Pairwise genome alignment tools:** MUMmer, Chains and nets, Shuffle_LAGAN, AnchorWave
2. **Multiple genome alignment tools:** TBA, MUGSY, MULTIZ, ABA, EPO, VISTA-LAGAN, Mauve, Cactus [12]

AnchorWave (Anchored Wavefront Alignment) [114] 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. Using a splice aware alignment program (minimap2 [115] and GMAP [116] have been tested) to lift over the start and end position of reference full-length CDS to the query genome (step 1). AnchorWave then identifies collinear anchors using one of three user-specified algorithm options (step 2) and uses the WaveFront Alignment (WFA) algorithm to perform alignment for each anchor and inter-anchor interval (step 4). Some anchor/inter-anchor regions cannot be aligned using our standard approach due to high memory and computational time costs. For these, AnchorWave either identifies novel anchors within long inter-anchor regions (step 3), or for those that cannot be split by novel anchors, aligns using the `ksw_extd2` function implemented in minimap2 or a reimplemented sliding window approach (step 4). AnchorWave concatenates base pair sequence alignment for each anchor and inter-anchor region and outputs the alignment in MAF format (step 5) [117].

1.5.4 Software dedicated to the construction of graph-based pangenomes: Minigraph, Minigraph-Cactus, PGGB and Seqwish

Minigraph [118] extends the minimap2 [115] alignment chaining model to work on graphs. It applies this alignment model to progressively build out a pangenome graph from a series of genomes that contains large sequences ($>$ 250 bp) that were not previously seen in other genomes. The resulting pangenome does not contain all input sequences and variation between them but rather a representative subset and large structural variants.

Progressive Cactus [119], another tool, constructs ancestral sequences guided by a phylogenetic tree, eliminating the need for a global reference assembly and successfully handling transitive collapsing of SV hotspots within repetitive sequences. However, a precise phylogenetic tree is vital. At each step, the LASTZ alignments are used as anchors to construct a Cactus graph [120], which in turn is used to filter and then refine the alignment [79].

The Minigraph-Cactus pipeline combines Minigraph's assembly-to-graph mapping speed with Cactus's base aligner, yielding base-level pangenome graphs at a scale applicable to hundreds of vertebrate haplotypes. Such graphs can enhance short-read mapping, variant calling, and structural variant genotyping. The pipeline constructs graphs for a variety of species, with construction times varying based on cluster resources.

The Pangenome Graph Builder (PGGB) [85] offers a comprehensive approach to generating pangenome graphs, consisting of three distinct phases. In the initial alignment phase, the wfmash aligner is used

to establish all-vs-all alignments for input sequences. The subsequent graph induction stage transforms the input FASTA sequences and PAF-format alignments produced by wfmash [121] into a graph in GFA format using Seqwish [122], preserving the input alignments and sequences as a lossless transformation. The final step involves graph normalization, where the smoothxg [123] normalization algorithm is applied to simplify complex motifs, such as those found in Short Tandem Repeats (STRs) and other repetitive sequences, mitigating potential under-alignment issues.

Seqwish, as a tool, plays a pivotal role in generating pangenome graphs by constructing a complete variation graph based on a collection of sequences and their corresponding alignments [122]. The resulting 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 Single-Nucleotide Polymorphisms (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 differentiation among pangenome graphs primarily stems from their handling of Copy Number Variation (CNV) sequences (genomic alterations that result in an abnormal number of copies of one or more genes). PGGB often merges CNVs, taking a distinct approach from Minigraph-Cactus (MC) graphs, which represent CNV copies as independent sub-graphs. The choice between these methods depends on specific applications and necessitates further experimentation and community input. A unique strength of the PGGB method is its retention of centromeric and satellite sequences (very large arrays of tandemly repeating, non-coding DNA), distinguishing it from MC graphs that tend to prune these sequences. This characteristic makes the use of MC graphs for read alignment applications practical in current methods. However, the pruning of such sequences is seen as an unsatisfactory solution, demanding further research, especially as Telomere-To-Telomere (T2T) assembly becomes more prevalent. Although PGGB retains centromeric and satellite sequences, its initial population-genetic analysis of these regions raises concerns about assembly accuracy and alignment, particularly in regions with significantly higher mutation rates. PGGB's significant feature is its creation of a lossless model that treats all input assemblies equivalently. This property allows every pangenome assembly to serve as a reference system, facilitating a comprehensive exploration of pangenome variation. The transitively collapsed SV hotspots in all-to-all pairwise alignments, including centromeres, lead to a reduction in variant sizes found in repetitive sequences. Unlike MC graphs, PGGB refrains from filtering rapidly evolving satellite sequences or unaligned regions, resulting in increased complexity and size compared toMC graphs. This complexity, however, enables annotations and coordinates from all pangenome as-

semblies to be related to the graph structure, facilitating downstream analyses. In the landscape of pangenome construction,PGGB stands out as a method with comprehensive and unbiased capabilities. It incorporates an "all-versus-all" alignment approach, treating each input genome equally, and offers a base-level representation of the pangenome. It encompasses variants of all scales, from SNPs to large SVs, allowing every included genome to function as a reference for subsequent analyses [124]. PGGB has been employed in constructing the draft human pangenome [85], emphasizing its versatility and credibility [125]. PGGB's approach to generating pangenome graphs makes it a valuable tool for researchers seeking a comprehensive and unbiased representation of genomic variation.

1.5.5 Graph-based pangenome toolkits: vg and odgi

The VG toolkit capabilities include read mapping, variant calling, and visualization tools, making it a versatile choice for NGS data analysis [83]. 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 [76]. In the context of 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 [126]. VG can fine-tune SV breakpoints by augmenting the graph with observed differences from read alignments, effectively correcting small errors in SV breakpoints [92]. 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 ODGI toolkit complements VG by facilitating graph manipulation tasks like visualization and the extraction of distances among paths in the graph, supporting phylogenetic analysis [127]. 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 [85].

1.5.6 Graph visualization

Visualization tools for pangenome graphs can be classified based on the types of graphs they are

meant to depict and their ability to linearize the graph. Some tools, such as Bandage [128], gfaestus [129] and GfaViz [130], are primarily designed for assembly graph interpretation. These tools emphasize the overall structure of the graph without specific visual features for pangenomic representation. In contrast, tools initially created for variation graph visualization pay more attention to base-level structure and pangenomic relationships. Additionally, ODGI [127] `odgi viz` (Optimized Dynamic Genome Graph Implementation visualization) employs binning and direct rendering to generate rasterized visualizations of gigabase-scale pangenomes, building on the linear layout technique of VG viz [83]. Older tools like Cytoscape [106] were originally designed for visualizing, analyzing, and interpreting complex biological networks, but were also adapted to interpret these networks as pangenome graphs, which can be seen in Panaroo [50], a tool that will be explored in this work. There remains an open challenge in interactive visualization, however, especially when dealing with human-genome-scale pangenome graphs and retaining coherency across various zoom levels [6].

1.6 Aims

We set to construct, visualize, and analyze gene-based and graph-based pangenomes of eight *Burkholderia* species. The specific objectives guiding this thesis are enumerated as follows:

1. To compile comprehensive genomic data from eight distinct *Burkholderia* species, and employing it to assemble gene-based pangenomes through the integration and comparison of the gene content;

2. To employ a graph-based approach for the creation of pangenome representations, enlightening the structural relationships among genes within and across the studied species;

3. To visualize and interpret the constructed pangenomes, employing state-of-the-art graphical representations and visualization tools to elucidate the genetic diversity, shared genes, and unique features among the *Burkholderia* species;

4. To conduct a detailed analysis of the constructed pangenomes, elucidating functional annotations and genomic variations that contribute to species-specific traits or evolutionary adaptations;

5. To derive insights from the analysis results, offering a comprehensive understanding of the genomic architecture and phylogenetic dynamics across the studied *Burkholderia* species.

This work aims 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, this thesis seeks to provide valuable insights into the genetic makeup, phylogenetic relationships, and potential functional implications of the examined *Burkholderia* species. Moreover, it aims to

contribute to the broader field of genomic research by offering methodologies and visualizations that enhance the exploration and comprehension of pangenomic data.

1.7 Thesis organization

In this first chapter of the thesis, we phrased the main limitations of sequencing data storage and single-reference systems and offered pangenomics as the solution. We introduced the *Burkholderia* genus, the object of this study; covered the basic concepts relevant to understand pangenomics, defined the various meanings of the pangenome and showcased several computational tools to manipulate them. The subsequent chapters are organized to address the research questions posed in the introduction. Chapter 2 outlines the methodology employed, detailing the data collection method, pipelines and software suites utilized and analysis techniques. Following this, Chapter 3 presents the findings, while Chapter 4 engages in the discussion of these results in relation to the research questions and concludes the thesis by summarizing the key findings, and presenting the questions to be answered at a later time, suggesting avenues for future research. This arrangement is designed to hopefully present a coherent and logical progression of ideas, facilitating a comprehensive understanding of the outcome of this work.

2

Methodology

Contents

2.1 Pipeline overview

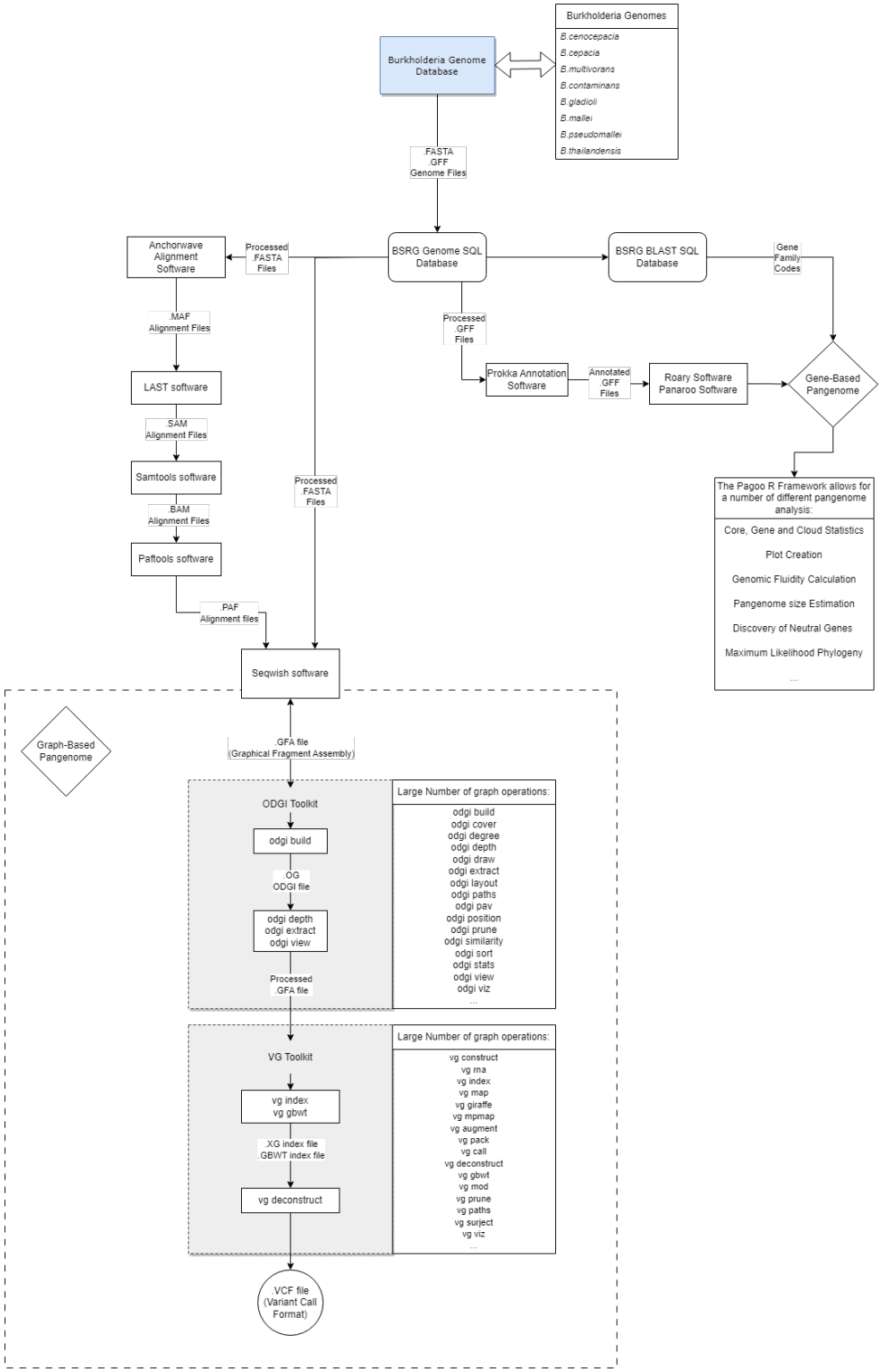


Figure 2.1: Flowchart representing an overview of the pipelines and methodologies used in this work.

2.2 Data sources and annotation

2.2.1 Raw data availability

The raw data used consisted in 90 *Burkholderia* strain Complete genomes belonging to 8 species (4 *Burkholderia gladioli* genomes, 5 *Burkholderia contaminans* genomes, 14 *Burkholderia multivorans* genomes, 15 *Burkholderia cepacia* genomes, 18
Burkholderia cenocepacia genomes, 16 *Burkholderia pseudomallei* genomes, 9 *Burkholderia mallei* genomes, 9 *Burkholderia thailandensis* genomes) in fasta format and their respective annotations in GFF format downloaded from the the "Burkholderia Genome Database" [131, 132]. The specific information for each strain can be found in a publicly available table that is provided as part of this thesis [133] that contains each customized acronym given to each strain and their respective original name and download address.

2.2.2 SQLite databases

Two Structured Query Language (SQL)ite Database (DB)s 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, 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.

2.2.3 Data annotation with Prokka

We used prokka [95] to annotate all the genomes using the downloaded fasta genome files as input. Prokka produced several output files, most notably a GFF file required to build certain gene-based pangenomes. These GFF files contain the genomic feature annotations of each genome (gene predictions, Messenger Ribonucleic acid (mRNA) transcriptions or transfer Ribonucleic Acid (tRNA)), their attributes such as their unique identifiers, locus tags or gene names; the genomic coordinates (start and end positions) of each annotated feature relative to the input genome sequence and information about the orientation (forward or reverse strand) of these features. Prokka's output also contains a Text (TXT)

file with summary statistics of the genomic features and the input genomes, such as the genome sizes, amount of CDS, tRNA and contigs.

2.2.4 Computing power

This work was performed mainly in three Computer setups described by Table 2.1.

Table 2.1: Specifications of the three workstations used to conduct this work. Central Processing Unit (CPU) and Random Access Memory (RAM) power are extremely relevant to conduct some analyses that are particularly workload heavy. It is also necessary to have a high storage capacity and an updated Operating System (OS)

Workstation	CPU threading (threads)	RAM memory (Gigabyte (GB))	Storage (Terabyte (TB))	OS
BSRG1	64	64	3	Ubuntu 22.04
BSRG2	20	64	4	Ubuntu 22.04
Home Computer	24	32	2	Ubuntu 20.04

2.3 Building gene-based pangenomes

2.3.1 Using Roary

For each species we built a gene-based pangenome using Roary [98]. We gathered all Prokka annotated GFF files and sequence fasta files intended for each pangenome in a single folder as input for Roary. Roary requires the user to specify the identity (-i) and core definition (-cd) values, and whether or not to perform a core alignment (-e). The Identity score is how similar the gene sequence has to be to the compared BLASTP database to be considered in the analysis. Core definition is the presence threshold for a gene to be considered a core gene. A first run was done for each of the 8 species using identity and core definition of 100%. A second run was done changing the identity value to 98%. A last run was done using Roary's default values of 95% identity and 99% core definition.

Roary's outputs consisted in a summary statistics TXT file, with the number of core,shell, cloud and total genes in the pangenome, a gene Presence-absence Tab-Separated Values (TSV) table which lists each gene and which strains it is present in, a pangenome reference FASTA file containing a single representative nucleotide sequence from each of the clusters in the pangenome, several "RTab" files used as input for plot creating scripts, a newick format tree roughly grouping isolates together based on their accessory genome, and a core gene alignment Alignment Format (ALN) file for tree building, if the core alignment is performed. All Roary runs performed a core alignment using PRANK [134] and were ran using 10 CPU threads. Pangenome building times took a few seconds and the core alignments took from a few minutes to an hour, depending on pangenome size, on the Home Computer. From Roary's

Github page [135] we downloaded two scripts, Roary_plots.py and create_pan_genome_plots.R, which use Roary's "RTab" output files as input, and created plots for analysis.

2.3.2 Using Panaroo

Similarly to Roary, we used Panaroo [50] to build pangenomes using a folder containing all the prokka annotated GFF files and all the sequence fasta files. We were also required to specify a core threshold, identity and core alignment. Additionally, it is required to choose to run Panaroo in one of three stringency modes (strict, moderate and sensitive). 4 runs for each species using Panaroo were done. The first and second run used the strict mode and 98% and 100% identity, respectively; the third and fourth run used sensitive mode for 98% and 100% identity. Core threshold was 100% on all runs. Core alignment was performed for *B. cenocepacia* using the PRANK [134] aligner. Panaroo's outputs were mostly similar to Roary: A gene presence-absence TSV table, a pangenome reference FASTA file, "RTab" files for plot creation, and a core gene alignment ALN file if the core alignment is performed. Additionally, Panaroo produces a graphical visualization of a gene-based pangenome, a GML graph file which will be discussed below. Panaroo's output is compatible with Roary's supplementary script Roary_plots.py, but these plots were not created in this work. All runs performed on Panaroo used 8 CPU threads. The runs took less than a minute, with the exception of the core alignments, taking approximately 30 minutes on the Home Computer.

As mentioned, one of Panaroo's outputs is a GML graph file, which contains all the meta-information of the pangenome, with the genes being represented as nodes and edges connecting the nodes if two genes appear adjacent to one another on at least one contig. We loaded the graph in Cytoscape [106] for visualization, selecting the yFiles Organic Layout. When loaded, the graph displayed was convoluted and difficult to analyze. The graph was simplified for analysis utilizing the script "reference_based_layout.py" available on Panaroo's Github page [136]. This script created a TXT file of a table with all edges that break up long-range connections. We imported the table into Cytoscape and cut the corresponding edges, which created a more appealing graph.

2.3.3 Phylogenetic tree-building using FastTree

FastTree [137] was used to build a phylogenetic Tree based on approximate maximum likelihood for all Roary and Panaroo assays in which core alignment was performed using the core alignment output file (.ALN) as input. The output was a newick tree file which was imported into Dendroscope [138], an interactive tree-viewing software, for further analysis.

2.3.4 Creating and exploring a pangenome object using Pagoo

Pagoo [67] is an encapsulated, object-oriented class system for analyzing bacterial pangenomes. We used the sqldf [139] 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 DB to produce a TXT file, containing all the strains' data for a single species, that will be Pagoo's input. A separate TXT 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 TXT 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 (*B. cenocepacia, B. cepacia, B. multivorans, B. contaminans*), all mallei group species (*B. mallei, B. pseudomallei, B. thailandensis*), mallei group + gladioli (*B. mallei, B. pseudomallei, B. thailandensis, B. gladioli*) and all eight species together. To note that unlike Roary and Panaroo, that use the default that genes in ¡15% of genomes belong to the cloud, Pagoo's cloud includes the genes that are only present in 1 genome.

2.3.4.A Analyzing the pangenome objects in the Pagoo framework

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 [140].

2.3.4.B Genomic fluidity calculation

We calculated the Genomic fluidity of all "pg" objects using the R library "micropan" [141] 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 2.1:

$$\phi = \frac{2}{N(N-1)} \sum_{\substack{k,l=1\ldots N \\ k<l}} \frac{U_k + U_l}{M_k + M_l} \tag{2.1}$$

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 genomic 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.

2.3.4.C Binomial estimation of the pangenome and core sizes

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 [142]. 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 (2.2).

$$\theta_y = \sum_{k=1}^{K} \pi_k f\left(y; \rho_k\right), \quad y = 0, \ldots, g \tag{2.2}$$

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

$$f\left(y; \rho_k\right) = \binom{g}{y} \rho_k^y (1-\rho)^{g-y}, \quad k = 0, 1, 2, \ldots, K \tag{2.3}$$

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 [143].

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

$$\mathrm{BIC}(K) = -2l(\pi, \rho | K) + (2K - 2) log(n) \tag{2.4}$$

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 [144]. n is the sample pangenome size.

2.3.4.D Gene neutrality and evolution testing using Tajima's D scores

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 [145]. Tajima's D test is done by applying the formula below (Equation 2.5).

$$D = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{\pi - \theta}{\sqrt{\frac{a_1(n-1)}{2} + \frac{a_2(n^2+n+3)}{6(n+1)}}} \tag{2.5}$$

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 results 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).

2.3.4.E Maximum likelihood phylogenetic tree building

A Maximum Likelihood phylogenetic tree was built for all pangenomes using the method implemented in the "Phangorn" Package [146]. The core genomes were first aligned using "DECIPHER" [147].

For a small population genomics assay, we searched for the source of our *B. cenocepacia* strains in the "Burkholderia Genome Database". We created a table which defines three populations for the strains: Geographic location (in continents), Medical Vs Environmental strains, and Body location (where the strain was sampled from) (Table A.8). We then created a new Pagoo object using the metadata from this table keeping all other specifications equally, using the command `pg < Pagoo(data, sequences, org_meta)`, in which org_meta is the population metadata. Using the same methodology as previously for the Maximum Likelihood trees, we created three new *B. cenocepacia* Trees and PCA plots, which now identify each of the populations defined, respectively.

2.4 Discovering unique genes in a dataset using group operations

We used Pagoo's commands `pg$core_genes`, `pg$shell_genes`, `pg$cloud_genes` and `pg$genes` to retrieve core, shell, cloud and pangenome annotation metadata for each pangenome object produced from the E-value of 70 (the optimal for the All species pangenome) and created a SQLite table for each. With the sqldf library, we performed a combination of minus and intersect commands on pairs of tables

belonging to the same pangenome category (core,shell,cloud or pangenome) but of separate species, with the intent of obtaining the gene family codes that were unique among that pair (A minus B, and B minus A), as well as the gene family codes in common with those two species (A intersect B). We performed this assay for all combinations of categories and species. We obtained these gene family codes in TXT format, and we used them in combination with the BLAST DB to obtain the annotation that belonged to those genes codes. We used the R library "VennDiagram"'s [148] command `venn.diagram` to create Venn diagrams for the pairwise comparisons of the species for every pangenome category, using the unique and the intersection of gene codes in the pairs of species. We also created venn diagrams for the Bcc and mallei groups.

We chose some Pairwise comparisons (*B. cenocepacia* Versus *B. cepacia*, *B. pseudomallei* versus *B. mallei*) and Bcc versus mallei group) in order to further analyze the unique core gene codes in these pairs of pangenomes. We enriched this data via hypergeometric testing with the bonferroni correction. The hypergeometric test can be explained by equation 2.6 [149].

$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{2.6}$$

Where N is the population size, or the total core genes in the tested pangenome. K is the number of success states in the population, or the number of genes belonging to the target cluster from the total core genes in the tested pangenome. n is the number of draws (i.e. quantity drawn in each trial), or the number of core genes that are unique to the tested pangenome (in relation to the pair of pangenomes). k is the number of observed successes, or the number of unique core genes that belong to the target cluster. $\binom{a}{b}$ is a binomial coefficient.

The classical application of the hypergeometric distribution is sampling without replacement. We want the probability of drawing k unique core genes that belong to the target cluster in the number of unique core genes in the pangenome out of the total core genes in the pangenome. The resulting Probability is a p-value that we used to determine if a gene family is enriched or not. This test was repeated for every cluster/gene family in the unique core genes of the tested pangenome (unique when compared to its pair). The resulting p-values were corrected with the bonferroni correction. Statistical hypothesis testing involves the rejection of the null hypothesis when the observed data probability under the null hypothesis is low. When numerous hypotheses are examined, the chance of encountering an unusual event rises, elevating the risk of mistakenly rejecting a null hypothesis (referred to as a "False Positive" error). The Bonferroni correction method is notably conservative, aiming to counteract this escalation by assessing each individual hypothesis at a significance level of a/m, where a represents the preferred significance level (p-value), and 'm' stands for the total number of hypotheses being tested, in our case, the number of clusters tested [150].

The Enriched dataset contained all the annotations from the unique core genes with a p-value under 0.05 and with a frequency superior to 20% (i.e only gene annotations from a gene family that represented by over 20% of the genes in the gene family).

2.4.1 Uncovering their metabolic pathways through KEGG

We also employed KEGGREST [151] as a client interface to query KEGG on the unique enriched gene annotations in the pairs of species/Bcc and mallei group pair, obtaining a table with the most common KEGG pathways unique to each pangenome in the pair.

2.5 Construction of a preliminary tree for the *B. cenocepacia* strains using Mashtree

We built a *B. cenocepacia* tree utilizing mashtree [152], using the tool's fast mode and all *B. cenocepacia* FASTA genome files as input. Mashtree works by first generating MinHash sketches for each input genome. These sketches are compressed representations of the genomic content derived from k-mers (short subsequences of fixed length) extracted from the sequences in the FASTA files. Next, Mashtree computes pairwise similarities between the MinHash sketches of the genomes. It calculates Jaccard distances (a measure of set similarity) between the sketches, which estimate the genomic similarity between pairs of genomes based on shared k-mers. Using the computed Jaccard distances, Mashtree employs a phylogenetic algorithm, a variant of the Neighbor-Joining algorithm, to build a tree structure that represents the evolutionary relationships among the genomes. Mashtree constructs a phylogenetic tree, where the branches represent the inferred evolutionary distances or relationships between the genomes. The lengths of the branches in the tree correspond to the computed similarities or distances between the genomes.

2.6 Pairwise genome alignment using the AnchorWave software

The raw data was processed through a series of steps intended to create the adequate input files for alignment using in-house developed scripting. The FASTA genome file names were changed and their headers were annotated into a serialized code to improve the smoothness of the process. The GFF file was converted into GFF3 with similar annotations. AnchorWave [114] 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 [116], a splice-aware sequence alignment program. AnchorWave used its algorithm to identify collinear anchors, then aligned the base pair sequences within each anchor and interanchor, finishing with concatenating all alignments to generate the final alignment for each collinear block. AnchorWave then outputted the alignment in MAF format.

2.7 DotPlot construction using the pafR library

For every alignment PAF file produced by paftools, (i.e for every pairwise alignment between strains), the R library "pafr" [153] was used to generate two dotPlots to represent the genome alignment.

For the "Single" dotPlot, the `dotplot(ali, label\_seqs=TRUE)` was used. For the "All" dotPlot, `dotplot(ali, label\_seqs=TRUE, order\_by='qstart')`.

2.8 Pangenome induction using Seqwish

The alignment MAF files produced by AnchorWave were converted to SAM format using the command `maf.convert` made avaialble by the the "Last" software [111]. SAM files were converted to BAM format with samtools [154] `samtools view -bt` command and reconverted into SAM with `samtools view -h` command. Paftools' `sam2paf` command [155] converted the SAM files into 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 GFA format utilizing seqwish [122] and 10 threads.

2.9 Detecting bubbles and compacting graphs with Bubblegun

Bubblegun [94] was used to detect bubble and superbubble chains in the pangenome 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. The `Bubblegun compact` command was used to generate a compacted version of the GFA graph.

2.10 ODGI and VG toolkits

2.10.1 Graph statistics and complex region detection

We built the ODGI graph file using the `odgi build` command, using the GFA graph as input. With the OG 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 PG graphs with the `vg convert` command, then calculated the number of Sub-graphs with the `vg stats -s` command. Complex regions of the genome were detected using the `odgi depth` command and the bedtools [156] toolkit to produce a Browser Extensible Data (BED) file containing the mean depth of each region of the pangenome.

2.10.2 Jaccard distance matrix and tree building using odgi similarity

Novel command `odgi similarity` from the ODGI toolkit was used to create a distance matrix based on path or path-group similarity. The OG 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 OG 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.

2.10.3 Construction of a VCF file for *B. cenocepacia* and *B. pseudomallei*

From the OG file, utilizing the ODGI toolkit, a BED file was produced using `odgi depth`, using depth intervals "-w 1000:0:25:0" as specification. This means that a BED file of path intervals will be printed where the node depth is between 0 and 25, and merging regions are not separated by more than 1000bp. A new, pruned, OG file was produced from the original OG using the BED file as filter, with the `odgi extract` command. Thanks to the BED file, all regions that were too complex (node depth higher than 25) were pruned. This immensely reduced computing time for the extract. Using the `odgi view` command, the pruned OG was converted into GFA format. A XG index file was created from this GFA file using the VG toolkit's command `vg index` [83]. A GBWT index file was produced with the `vg gbwt` command using the XG file as input, using the option "-E" to build the index from embedded paths in the graph. Finally, the VCF file was produced using the command `vg deconstruct`, using the GBWT

file and providing all the pangenome paths. For *B. cenocepacia*, this entire process took an immense amount of time, taking approximately 30 days on BSRG1, using 60 threads.

2.10.4 Viewing the graph-based pangenomes

Bandage [128] was used to visualize the pangenome graphs. The program accepts GFA format as input and draws the graph in an interactive platform for viewing and analyzing. Gfaestus [157] was also used to view GFA graph pangenomes. Additionally, Gfaestus requires a path-guided Two Dimensional (2D) layout input. This input was created using ODGI's `odgi layout` command with -T as prefix, which created the layout in TSV format. The ODGI toolkit also provides a command named `odgi viz`, that produces a linear and static visualization of an OG graph in Portable Network Graphic (PNG) format. We used this command to produce 1-dimensional visualizations of the pangenomes. We used `odgi sort` to sort the graph nodes in the OG file from left to right, then we created an additional `odgi viz` visualization of the now sorted pangenome.

3

Results

Contents

3.1 Pre-pangenome analyses

3.1.1 Preliminary statistics of the *Burkholderia* genomes

We performed a whole-genome annotation of the eight *Burkholderia* species using Prokka. Prokka generated outputs that provide statistical data on each genome annotated, which we collected and plotted (Figure 3.1).

The eight studied *Burkholderia* species have between one to five ”contigs” (this is just the prokka assigned name, but in fact, every reference of a contig is referring to a chromosome or a plasmid as the genomes were sequenced ”end-to-end”), with the majority tending to have two or three. *B. gladioli* has the overall highest amount of contigs. Size-wise, the species have a genome size of around 5Mbp to 9Mbp, *B. mallei* having the shortest genome overall, and *B. gladioli*, *B. contaminans* and *B. cepacia* being the longest. On the amount of tRNA found, for the most part, the genomes encode between 70 and 83. *B. pseudomallei* is the outlier in this statistic, with most strains having the highest amount of tRNA between the genus. In terms of the amount of CDS, the same tendencies as in genome size are mostly maintained, with the largest genomes having higher amounts of CDS and vice-versa. At first glance, however, it seems that *B. cenocepacia* has a lower CDS proportion to genome size, but this may not be true, as the data showcased here isn't conclusive in any form. GC content in the genus varies minimally, with the species having between 66-69% , which is consistent with recent assays [17]. A benefit of this preliminary statistical experiment is to identify possible contaminated or bad sources of data. If an overly distant outlier presents itself, it may be wise to remove it from our dataset, before proceeding to other analyses.

3.1.2 Preliminary phylogenetic tree for *B. cenocepacia* using mashtree

We also used the fasta genome files from the raw data belonging to *B. cenocepacia* and employed mashtree to create a fast, preliminary similarity tree of the species (Figure 3.2).

Mashtree's tree indicates two large *B. cenocepacia* strain clusters, cluster A (Strains 3, 4, 5, 8, 11, 13, 16, 17) and cluster B (Strains 1, G1, 2, 9, 10, 12, 14, 15) , and two outlier clusters, cluster C (Strain 18) and cluster D (Strains 6 and 7).

3.1.3 Visualization of the pairwise anchorwave alignments

Pafr is a R environment tool that plots the pairwise alignment files (PAF) from AnchorWave so that we can see what is happening in the alignments between every strain. We show the example of the *B. cenocepacia*'s alignments of strain 4 Versus all others. Strain 4 contains 3 chromosomes and 1 plasmid which makes it ideal to compare to the alignment of the other strains, as it is the maximum amount of

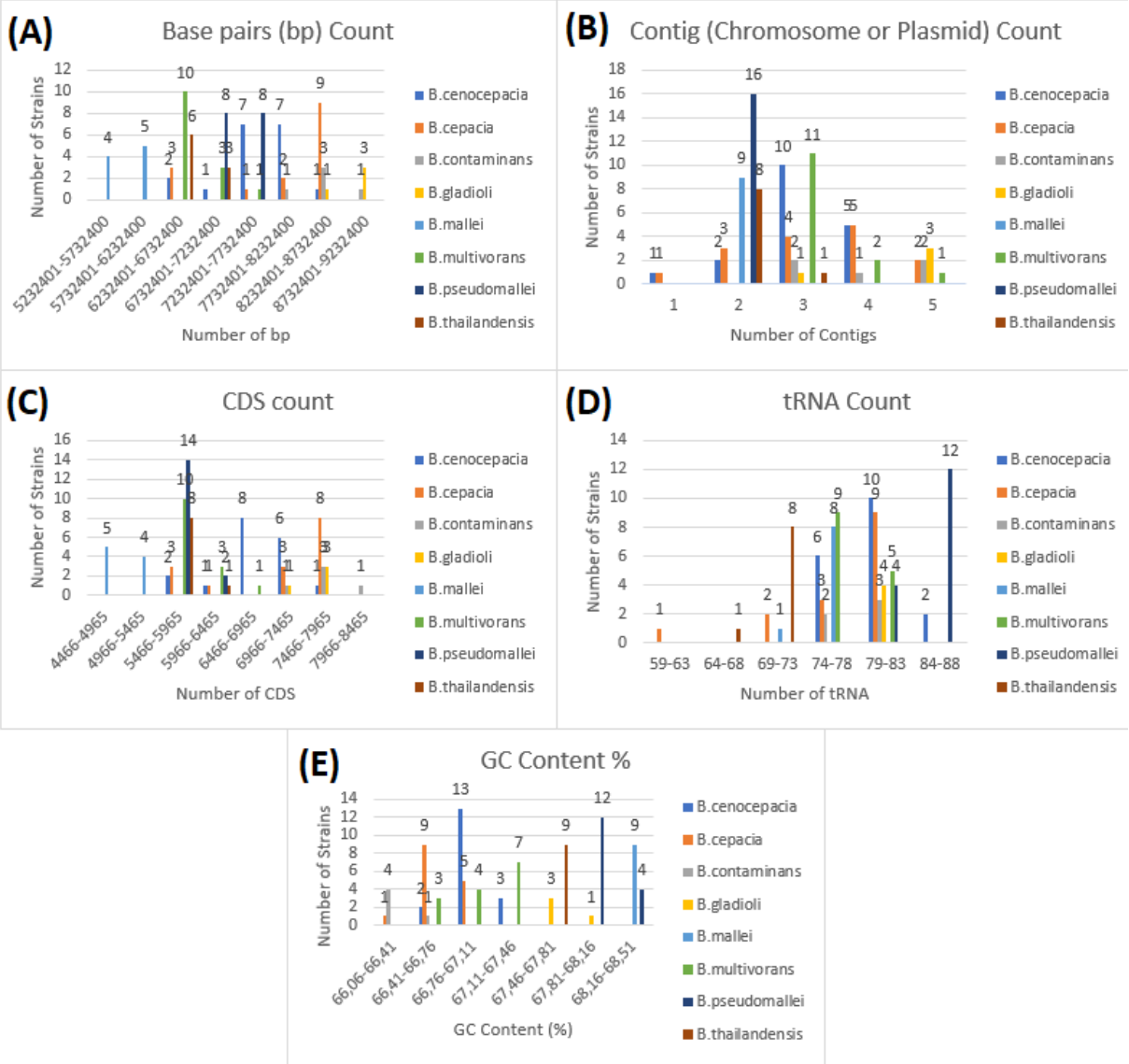


Figure 3.1: Histograms representing and comparing basic statistics of the Prokka-based genome annotation performed of the 8 *Burkholderia* species selected for analysis in this work. (A):Number of strains showing genomes sequences between 5232401bp and 9232400bp in size.(B): Number of strains showing a number of "contigs" (chromosomes and/or plasmids) between 1 to 5. (C): Number of strains showing the number of CDS in the genome between 4466 and 8465. (D): Number of strains showing the number of tRNAs in the genome between 59 and 88 . (E): Number of strains showing the percentage of GC content in the genome between 66.06% and 68.51% . This last statistic does not come directly from Prokka, but by calculating the percentage of G's and C's in the genomes after Prokka's annotation.

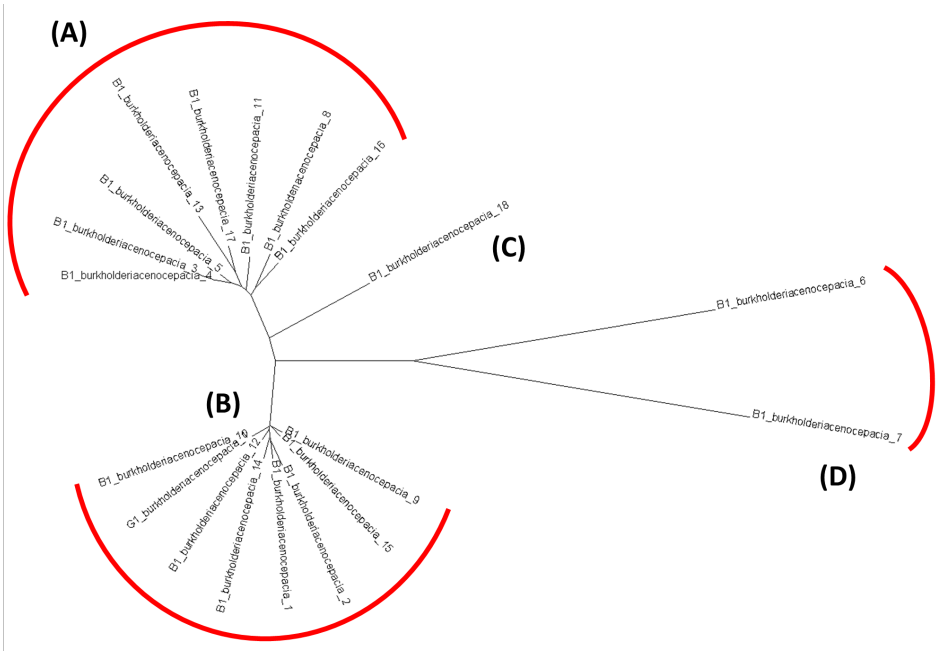
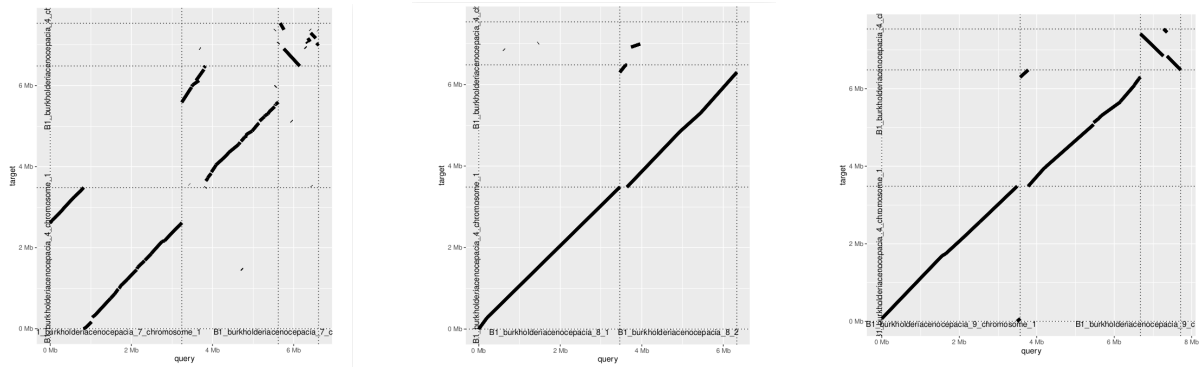


Figure 3.2: *B. cenocepacia* confidence tree generated by mashtree. Two large *B. cenocepacia* strain clusters (A,B) and two outlier clusters (C,D) can be observed. Visualization was performed on Dendroscope.

chromosomes+plasmids. We can see examples of the alignment of strain 4, with strains 7, 8 and 9 in figure 3.3. We can observe rearrangings in the first and second chromosomes in the alignment with strain 7, a portion of chromosome 3 in strain 4 that is present in chromosome 2 of strain 8, and an inversion of chromosome 3 in strain 9 in relation to strain 4. The rest of the alignments are shown in A.1 where we can see more of the diversity that occurs in the pairwise alignments.



(a) (b) (c)

Figure 3.3: PafR dot plots for three pairwise *B. cenocepacia* strain alignments. (A) *B. cenocepacia* strain 4 vs Strain 7. (B) *B. cenocepacia* strain 4 vs Strain 8 (C) *B. cenocepacia* strain 4 vs Strain 9

3.2 Roary + Panaroo gene-based pangenome

3.2.1 Roary-based pangenome

To study gene-based pangenomes and compare the methodologies used to create them, we started with Roary, which compares the annotated genes using its algorithm of similarity-based clustering inspired by CD-HIT, and has been a staple in pangenomic studies for many years, to build the first pangenomes.

Table 3.1: Core, shell, cloud, and total gene number in Roary-based pangenomes.

Species Name	No. of Genomes	No. of Genes	Gene Type	100% Identity	98% Identity	95% Identity
B. thailandensis	9	51921	Core	698	741	2524
			Shell	5523	5442	3519
			Cloud	625	7269	5006
			Total	6846	13452	11049
B. pseudomallei	16	95686	Core	4425	4739	5084
			Shell	2594	2119	1478
			Cloud	985	2987	2150
			Total	8004	9845	8712
B. multivorans	14	83675	Core	3148	3364	4489
			Shell	4767	4546	2477
			Cloud	3184	6398	3890
			Total	11099	14308	10856
B. mallei	9	44631	Core	3592	3817	3854
			Shell	1861	1576	1512
			Cloud	402	589	471
			Total	5855	5982	5837
B. gladioli	4	29896	Core	4758	5063	5909
			Shell	2765	4550	2849
			Cloud	0	0	0
			Total	7523	9613	8758
B. contaminans	5	38560	Core	4256	4543	5734
			Shell	5526	7235	4611
			Cloud	0	0	0
			Total	9782	11778	10345
B. cepacia	15	108323	Core	187	217	1422
			Shell	7762	7698	6593
			Cloud	6546	30822	19925
			Total	14495	38737	27940
B. cenocepacia	18	124210	Core	273	305	1518
			Shell	11062	10912	7133
			Cloud	3440	27758	17818
			Total	14775	38975	26469

Roary outputs the core and accessory genes of each pangenome, represented in Table 3.1, as well as a presence/absence Matrix. When looking for optimal results, we aim to achieve a high core gene count among strains. This was the operational criteria chosen for this work because we are assuming, as the strains belong to the same species, that the homologous genes will have high similarity. This means that, by choosing high identity percentage (or E-values) in our assays, we are ”exposing” the genes that are common in all strains, i.e the set of essential genes/proteins that ensure the viability nad growth of these bacteria. *B. thailandensis*, *B. cepacia* and *B. cenocepacia* showed extremely low number of core genes when Roary was ran with 100% identity. Running Roary at 100% identity means that Roary's

algorithm will only consider genes with perfect sequence similarity. Generally, core gene number being low in percentage should mean that the samples are highly diverse in their essential genes. Since we know the samples belong to the same species, this indicates poor performance on Roary's part in defining the clusters. We also tried running Roary for lower identity values. For 98%, the cloud genes increased dramatically for *B. thailandensis* but the core genes only slightly increased. The core genes for *B. thailandensis*, *B. cenocepacia* and *B. cepacia* remained very low. For 95%, the core genome for those species largely increased. Note that total size of the pangenomes decreased. When Roary's algorithm recalculated the clusters with 95% identity, some clusters must have merged. *B. gladioli* and *B. contaminans* pangenomes present no cloud genes. No cloud genes could be partially explained by the low amount of datasets for *B. gladioli* and *B. contaminans*, with only 4 and 5 genomes, respectively, which is considerably low to view relevant variability in a pangenome, but we point to most likely being a sign of Roary's poor performance in defining the clusters in these datasets. The Identity value of 100% was chosen as the default for this work to compare pangenome tools. The reasoning behind this was to compare only perfect matches to reduce any kind of bias. Lowering Roary's identity threshold did not seem to improve the pangenomes significantly, however. Roary also aligned core genomes when building the pangenome. With the core alignment performed, the ALN file was input to FastTree to create Maximum-Likelihood phylogenetic trees for some species. In the case of cenocepacia, we can compare Roary's 100% identity tree to the preliminary mashtree's tree. Fasttree produced a similar tree to mashtree's (Figure 3.4). In Roary's tree, strain number 18 was included in strain 6 and 7's branch (C).

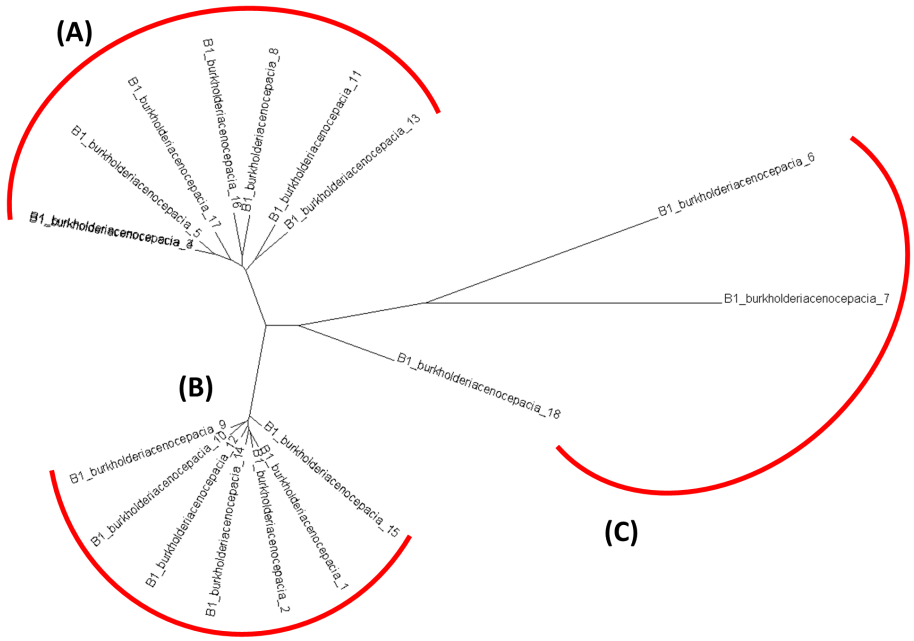


Figure 3.4: *B. cenocepacia* FastTree maximum-likelihood tree, using Roary's core alignment output, using 100% identity. We can observe 3 clusters: cluster A (Strains 3, 4, 5,8,11,13,16,17),cluster B (Strains 1,G1,2,9,10,12,14,15) , and cluster C (Strain 6,7,18). Visualization from Dendroscope.

Roary also offers complementary python scripts that allow plot creation using Roary's output files. These plots contain detailed information about the genes in the pangenome. We created the plots for *B. cenocepacia*, which show that Roary suggests the existence of a closed pangenome (Figure A.2).

3.2.2 Panaroo-based pangenome

Panaroo is a newer alternative to Roary, which merges the same gene-based concepts as Roary but implements a graph-based strategy to enhance annotation precision and clustering of orthologs and paralogs in the pangenome.

Table 3.2: Core, shell, cloud, and total gene number in Panaroo-constructed pangenomes for all 8 species studied. The first two columns of values represent the output when running Panaroo with the "Strict" stringency mode, with the default Identity specification of 98% and 100% identity on the first and second columns respectively . The last two columns represent the same identity values organized as previously but with the pangenome constructed using the "Sensitive" stringency mode.

Species Name	No. of Genomes	No. of Genes	Gene Type	Strict 98% Identity	Strict 100% Identity	Sensitive 98% Identity	Sensitive 100% Identity
B. thailandensis	9	51921	Core	4009	2907	4011	2907
			Shell	1913	3014	1953	3014
			Cloud	2629	3584	3004	4207
			Total	8551	9505	8968	10128
B. pseudomallei	16	95686	Core	5254	5202	5284	5202
			Shell	1147	1230	1231	1230
			Cloud	1463	1620	1468	1626
			Total	7864	8052	7983	8058
B. multivorans	14	83675	Core	4721	4679	4724	4679
			Shell	2122	2113	2148	2113
			Cloud	3037	3221	3045	3276
			Total	9880	10013	9917	10068
B. mallei	9	44631	Core	4097	4021	4107	4021
			Shell	1013	1117	1017	1119
			Cloud	88	85	97	94
			Total	5198	5223	5221	5234
B. gladioli	4	29896	Core	6137	6059	6146	6059
			Shell	2217	2275	2285	2441
			Cloud	0	0	0	0
			Total	8354	8334	8431	8500
B. contaminans	5	38560	Core	6074	5736	6078	5736
			Shell	3709	4351	3729	4393
			Cloud	0	0	0	0
			Total	9783	10087	9807	10129
B. cepacia	15	108323	Core	2901	1891	2901	1891
			Shell	5230	6091	5248	6091
			Cloud	10281	14132	10636	16556
			Total	18412	22114	18785	24538
B. cenocepacia	18	124210	Core	3051	2235	3051	2235
			Shell	4859	5726	4897	5726
			Cloud	10843	12363	11629	15653
			Total	18753	20324	19577	23614

Like Roary, Panaroo outputs the core and accessory genes of the pangenomes (Table 3.2). Panaroo produced results for both Sensitive and Strict stringency mode. Strict mode aggressively removes contamination or erroneous annotations while sensitive mode retains all gene clusters. Overall, the mode difference didn't cause significant changes in the pangenome. The number of core and shell genes from

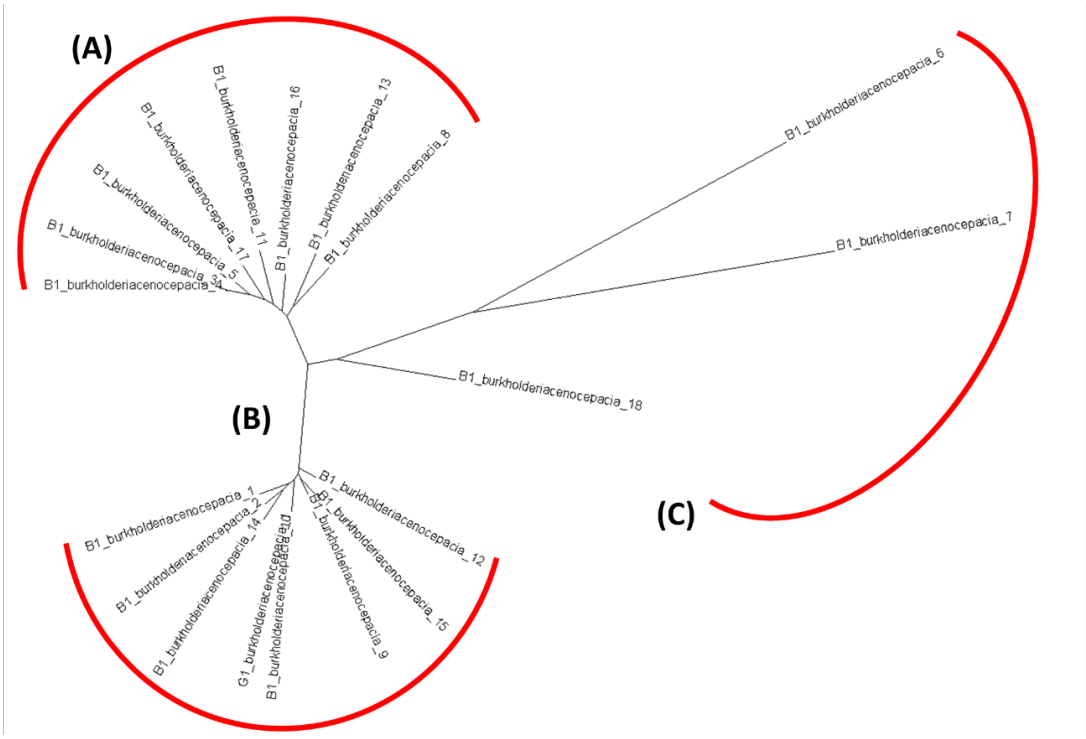


Figure 3.5: *B. cenocepacia* FastTree maximum likelihood tree, using Panaroo's core alignment output, using 100% identity and Sensitive mode. Visualization from Dendroscope.

Strict mode are very slightly lower than Sensitive mode, with cloud genes having the highest differences. This indicates that our dataset contains very low amounts of contamination or bad annotations. like in Roary, *B. mallei*, *B. gladioli* and *B. contaminans* had an exceedingly low cloud gene count, with the latter two having no cloud genes at all. This indicates that Panaroo, like Roary, is overly merging the genes in very large clusters, which, in addition to the low amount of datasets, causes the clusters to always have presence above 15% (which defines the cloud genes). Unlike Roary, however, Panaroo's pangenomes contained much higher counts of core genes, even at 100% identity, despite lower values compared to 98% identity. The change is seen more notably in *B. thailandensis B. cepacia* and *B. cenocepacia*, with the core genome being 5 to 7 times larger than in Roary, which seems to indicate that Panaroo is able to perform better than Roary when it comes to correcting the very low core gene numbers. Like Roary, Panaroo can produce a core alignment file which can be input into FastTree to compute a maximum-likelihood Tree of the pangenome of a specific species. To compare the softwares, we produced the alignment of *B. cenocepacia*'s core genes and built the phylogenetic tree (Figure 3.5).

There were no major differences in the tree produced using Panaroo's core genome and Roary's core genome, suggesting that regardless of the core size, the contents of the core genome aligned very similarly.

As previously mentioned, Panaroo's output also includes a graph in GML format, representing a

(a) (b)

Figure 3.6: *B. cenocepacia* pangenome graph of the gene-based pangenome produced by Panaroo. (A): Original graph. (B): Graph with the long distance edges cut.

graphical visualization of a Gene-based pangenome, in which genes are nodes and edges connect nodes if two genes appear adjacent to one another on at least one contig. We opened the graph GML files using Cytoscape [106] to view the Strict mode, 98% pangenomes, but the graphs were too convoluted to analyze, as exemplified in Figure 3.6a. We simplified the graphs removing long distance connections and got much more appeasing graphs, as shown in Figure 3.6b. The Panaroo graphs for the other 7 species can be consulted in the thesis' github repository [158].

More linear and simple sections of the graphs tend to represent the core parts of the pangenome, that is, the genes common in most genomes. The more complex parts show the lack of consensus between which genes are connected with which, i.e representing the differences between genes of each strain, the Shell and the Cloud. Additionally, We can use these graphs and take advantage of cytoscape's user interface to pinpoint a particular gene, and see its connections to other genes in the pangenome.

3.3 Gene-based pangenome exploration using Pagoo

As explained in Chapter 2, the GenomeDB, containing all genes for the 8 species of the *Burkholderia* genus contains clusters of genes based on gene family for many different E-values. Using Pagoo, We generated a pangenome object for each species and for each E-value from 30 to 100. Tables A.3 and A.2 contain the core and accessory genes of all the pangenomes. 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 3.3. As mentioned previously, 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.

3.3.1 Single-species pangenomes

Compared to Roary and Panaroo, the pangenomes produced differ greatly. Cloud Genes massively increased and shell genes lowered. In the case of *B. gladioli* and *B. contaminans*, where cloud genes were absent in both Roary and Panaroo and most accessory genes belonged in the shell gene category, were almost interchanged in Pagoo's case. Shell gene decrease was less noticeable in the species belonging to the Bcc. Core genome size results, compared to the other methods, were varied. The change is less noticeable in the smallest datasets, *B. contaminans* and *B. gladioli*, with core decreases in *B. pseudomallei*, *B. mallei* and *B. multivorans*, and increases in *B. cepacia*, *B. cenocepacia* and *B. thailandensis*. This is consistent with Roary and Panaroo using low thresholds for the separation of genes, merging many gene families, causing the cloud genes to move to the shell. This also causes an artificial increase in the core genome, since the shell gene families can reach 100% presence due to the merging. The proportions of Core and accessory genomes in the pangenomes are represented in Figure 3.7.

Table 3.3: Core, shell, cloud, and total genes calculated by Pagoo's command`summary_stats` for all 8 studied species and the respective optimal E-value.

Species name	No. of genomes	Total No. of genes	Optimal E-value	Core genes	Shell genes	Cloud genes	Pangenome genes
B. thailandensis	9	51862	80	3749	512	8754	13015
B. pseudomallei	16	100728	80	3906	928	19251	24085
B. multivorans	14	89714	85	3495	1855	14126	19476
B. mallei	9	45517	80	3108	546	8640	12294
B. gladioli	4	28947	90	5084	171	4541	9796
B. contaminans	5	37400	80	4532	1291	5004	10827
B. cepacia	15	104273	80	2890	3862	15473	22225
B. cenocepacia	18	128637	80	3284	3653	18046	24983

Tendentiously, the higher the amount of strains in the pangenome, the smaller in size ratio the core-genome is. We expected to see higher shell gene ratios with smaller cores, and while that is true for Bcc pangenomes, it was not the case for the mallei group. This can also be observed in the bar plots in Figure 3.8. By observation of the Pie and Bar plots, we can suggest that, With the exception of *B. contaminans* and *B. gladioli*, which only have 5 and 4 genomes, respectively, the behaviour of the other 6 strains with a more robust amount of genomes is consistent. Which indicates that Pagoo with the gene families defined by the GenomeDB and BlastDB have a consistent behaviour.

We can also analyze the core genome and pangenome to check whether these 8 pangenomes are open or closed. An open pangenome is indicative of a species where the continual exchange of genetic material results in the constant discovery of new genes within its bacterial populations. If the tendency of adding genomes to a pangenome is the stagnation of the number of its clusters, then we can say the pangenome is closed. All the *Burkholderia* pangenomes present appear to be open 3.9, as there is no visible stagnation of the number of pangenome clusters with the increase of genomes. The core

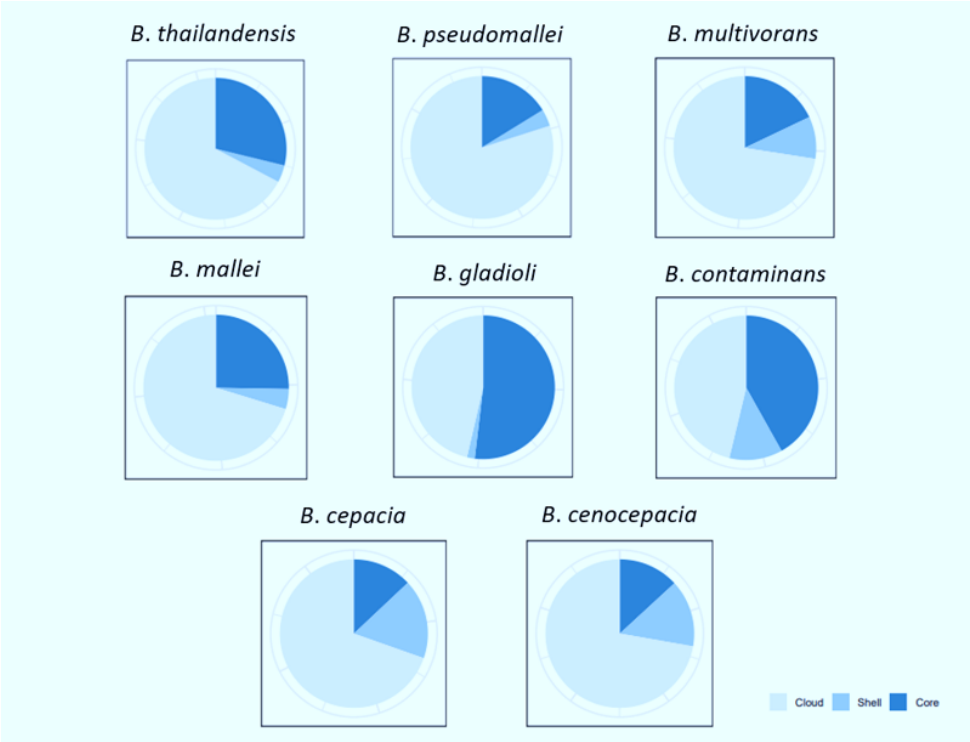


Figure 3.7: Pie charts representing the percentages of the pangenome that are cloud, shell or core genes, for each of the 8 species. From lighter to darker: Cloud, shell, core genes.

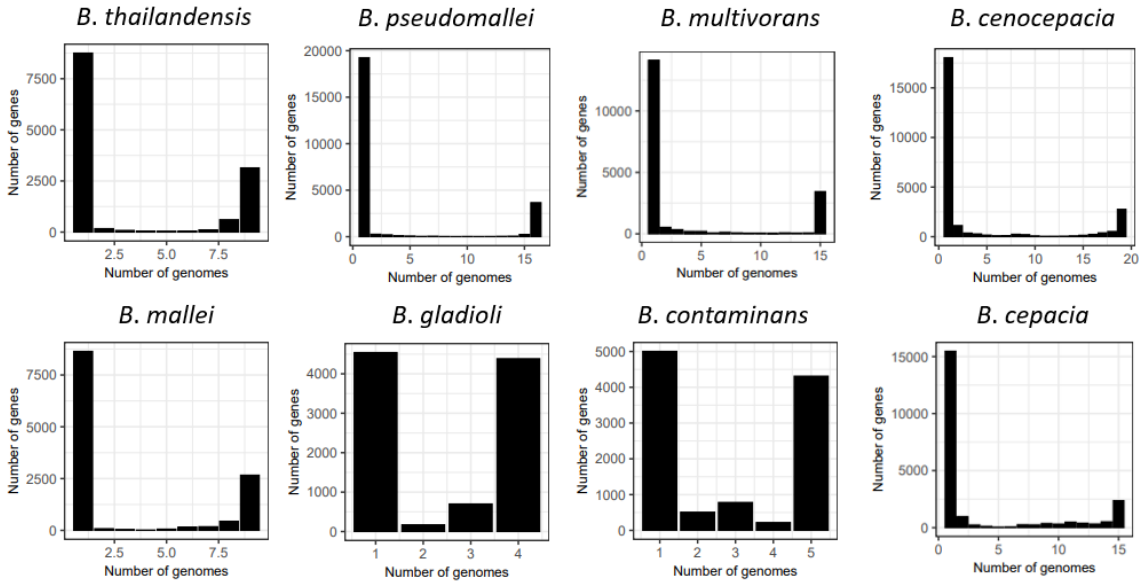


Figure 3.8: Bar plot created by Pagoo's `pg$gg_barplot` command that represents in how many genomes/strains a certain number of genes are present, for every species. In this case, the core is represented by the last bar in each plot (i.e the number of genes present in all genomes), while the cloud is mostly represented by the first bar (i.e the number of genes that are present in only one genome).

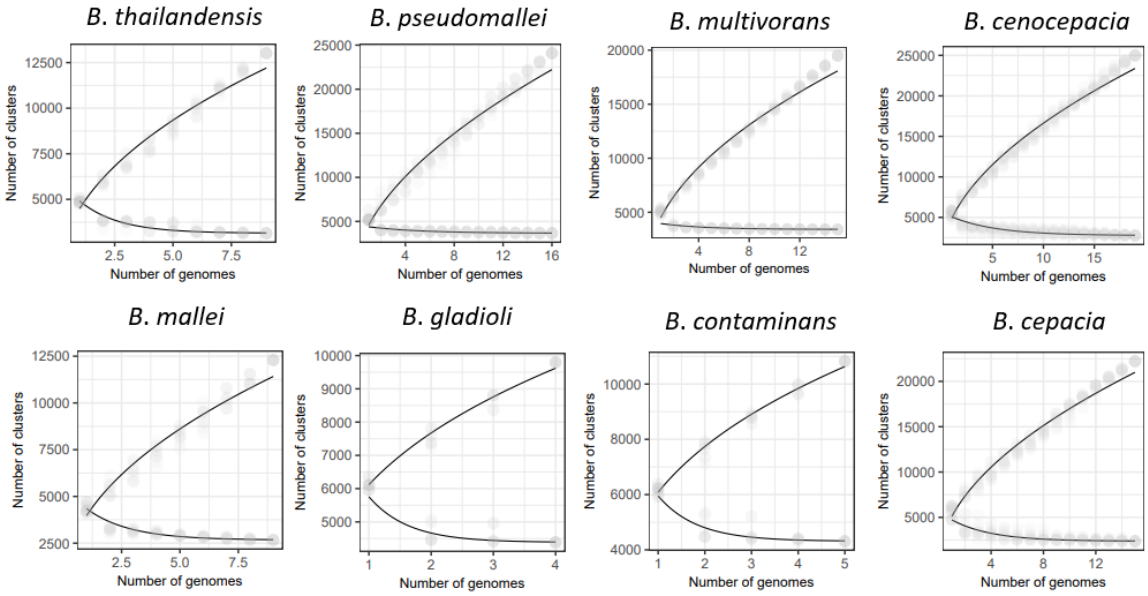


Figure 3.9: Curve plots for every pangenome representing two curves: The upper curve represents the pangenome's increase in clusters the more genomes are added into itself. Lower curve represents the core's decrease in clusters the more genomes are added into the pangenome. Created with Pagoo's `pg$ggcurves` command.

genome, however, appears to be stabilized over most pangenomes (it is inconclusive in *gladioli* and *contaminans*, due to the low amount of genomes in the dataset, but the tendency seems to be the same). In the case of *B. cenocepacia*, this data contradicts the results obtained by Roary and its plots, which indicated a tendency towards a closed pangenome.

Pagoo also generates a PCA plot (Figure 3.10), which can be used to identify the existence of genome clusters and point out variation between the genomes. In those plots, approximately 50% of the variation in the *B. gladioli* and *B. contaminans* pangenomes can be explained by the PC1 (the first principal component). We can observe that one strain and two strains, respectively, in the corresponding plots, that are completely apart from the others. In the case of PC2 (the second principal component), we can see that 27% and 22%, respectively, of the *B. thailandensis* and *B. gladioli* pangenomes' variation can be explained by that component, as we can see one strain in *B. thailandensis* and two strains in *B. gladioli* that are responsible for the high gene variance explained by PC2.

3.3.2 Analyses in the Pagoo framework

3.3.2.A Genomic fluidity

The mean genomic fluidity of the *Burkholderia* species pangenomes is represented in Table 3.4. Genomic fluidity is between 0.218 and 0.259 for the pangenomes with the exception of *B. cenocepacia*

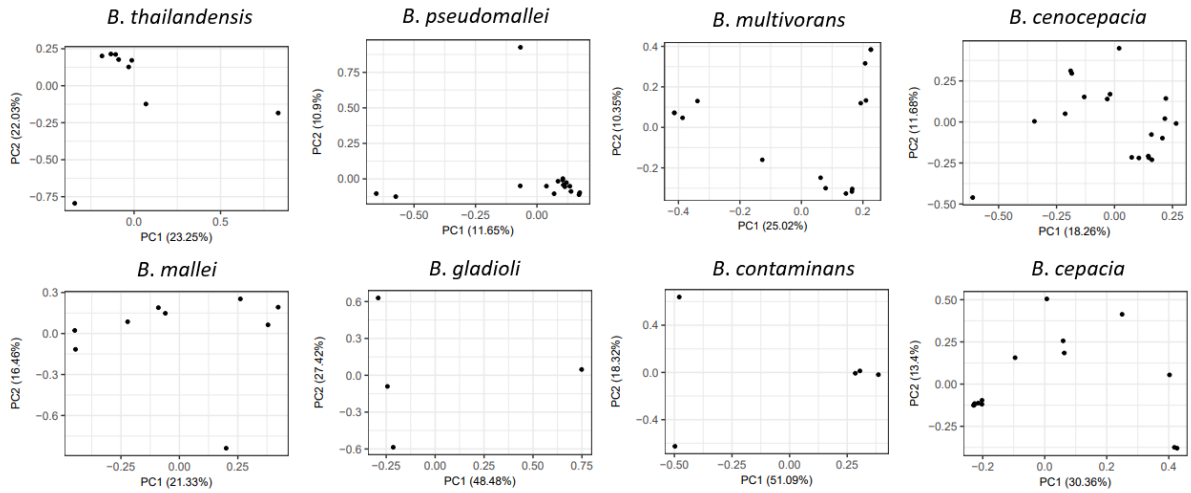


Figure 3.10: PCA plots for *Burkholderia* pangenomes obtained by Pagoo's `pg$gg_pca` command. Each dot represents an individual strain/genome.

and *B. cepacia*, which have the highest fluidity, the latter having 0.317, but with a high standard deviation of $\pm$0.106. The average fluidity for all eight species is 0.256, with the average fluidity for the Bcc being 0.273 and 0.247 for the mallei group (Table A.1).

Table 3.4: Genomic Fluidity of the *Burkholderia* species studied generated by Pagoo's framework. The "mean" values represent the genomic fluidity between two genomes/strains averaged over n random pairs of genomes/strains to obtain a population estimate. The default value for n was used (n = 100). The "standard deviation" is the sample standard deviation over those n computed values.

Species name	Mean Fluidity	Standard Deviation
B. thailandensis	0.233	0.0403
B. pseudomallei	0.248	0.0430
B. multivorans	0.258	0.0220
B. mallei	0.259	0.0410
B. gladioli	0.218	0.0401
B. contaminans	0.233	0.0537
B. cepacia	0.317	0.106
B. cenocepacia	0.281	0.0464

3.3.2.B Pangenome size estimation

The pangenome size estimation test determines if our pangenomes could be explained by a different model than just 3 categories named Core, Shell and Cloud and after fitting the model to the pangenome, it is also able to estimate its size. The Optimal number of categories is represented by the K value corresponding to the lowest BIC value (Table 3.5). 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. *B. cepacia* is the single Species with an ideal number of categories equal to 6. The rest of the Species are fitted to a 5 category model. These models, represented in table A.4, contain the estimated detection

probabilities for each component of the mixture models, which relate to the thresholds where we would consider the core and accessory genes. Indeed, the detection probability of 1 refers to the core genes, as there is a 100% probability of finding a core gene in any genome. The mixing proportion indicates the proportion of gene clusters having the corresponding detection probability.

Table 3.5: BIC values, core and pangenome size estimation for a array of K ranges for all species. Optimal BIC value for each pangenome is in Bold.

K range	Core size	Pangenome size	BIC	K range	Core size	Pangenome size	BIC
B. thailandensis				*B. pseudomallei*			
3	2812	117987	25389.	3	3663	214417	38677.
4	2810	122049	25406.	4	3564	917945	35098.
5	1739	276978	**24715.**	5	3299	1102389	**34848.**
6	12	276171	24726.	6	22	745613	35079.
B. multivorans				*B. mallei*			
3	3424	127952	41205.	3	2513	347612	22607.
4	3404	246089	**38862.**	4	2128	433570	**22439.**
5	3404	257307	38881.	5	655	466079	22464.
B. gladioli				*B. contaminans*			
3	2963	194694	**19157.**	3	4251	9366582	**24736.**
4	1064	76981	19175.	4	4248	76407	24801.
5	24	84033	19193.	5	0	70849	24830.
B. cepacia				*B. cenocepacia*			
3	2346	91687	59014.	3	2755	75066	69603.
4	2235	107134	57440.	4	2587	143820	60157.
5	2209	158495	57387.	5	2554	238898	**59437.**
6	1211	215631	**56795.**	6	650	240039	60008.

3.3.2.C Tajima's D test and identification of neutral genes

Table 3.6 contains the number of neutral genes and the percentage of the core genome those genes represent. The highest ratio of neutral genes in the core is in *B. multivorans*. Over half of its core genome is neutral, indicating that this species is under the least selective pressure out of all the samples, while *B. thailandensis* and *B. gladioli* possess almost no neutral genes, suggesting these species are under strong selective pressures. With the exception of *B. multivorans*, the *Burkholderia* genus appears to be constantly evolving through natural selection pressure.

Table 3.6: Number of neutral genes in each pangenome obtained by Pagoo's framework using Tajima's D test, and percentage of core genes that are neutral.

Species name	No. of neutral genes	Neutral genes in core
B. thailandensis	16	0.427%
B. pseudomallei	1034	26.47%
B. multivorans	1968	56.31%
B. mallei	306	9.845%
B. gladioli	4	0.0787%
B. contaminans	500	11.03%
B. cepacia	442	15.29%
B. cenocepacia	100	3.045%

We chose the pangenome for *B. cenocepacia* to delve deeper into the results obtained by the Tajima's D test. We retrieved most common gene annotations belonging to the gene families identified by the test deemed to be evolving neutrally (Table A.7), as well as the top 10 gene families with the highest and lowest tajima values (Tables A.5 and A.6, respectively). The genes evolving neutrally were identified most commonly as "hypothetical protein", "LysR family transcriptional regulator", "type VI secretion system tip protien VgrG" and "pyridine nucleotide-disulfide oxidoreductase". Many ABC transporters were also identified. As for the highest tajima scores (values between 2.700 and 2.286), these are the genes under the highest selective pressure, and we can identify the "HAD-IB family hydrolase" as the annotation corresponding to the most under pressure cluster. The lowest tajima scores (values between -4.186 and -3.506) relate to the genes with an exacerbated proportion of rare alleles, and the lowest value belongs to a "chromosomal replication initiator protein DnaA".

3.3.2.D Maximum likelihood phylogeny

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 3.11. The remaining 7 trees are available in the appendix (Figures A.4, A.5, A.7, A.6, A.8, A.3, A.9).

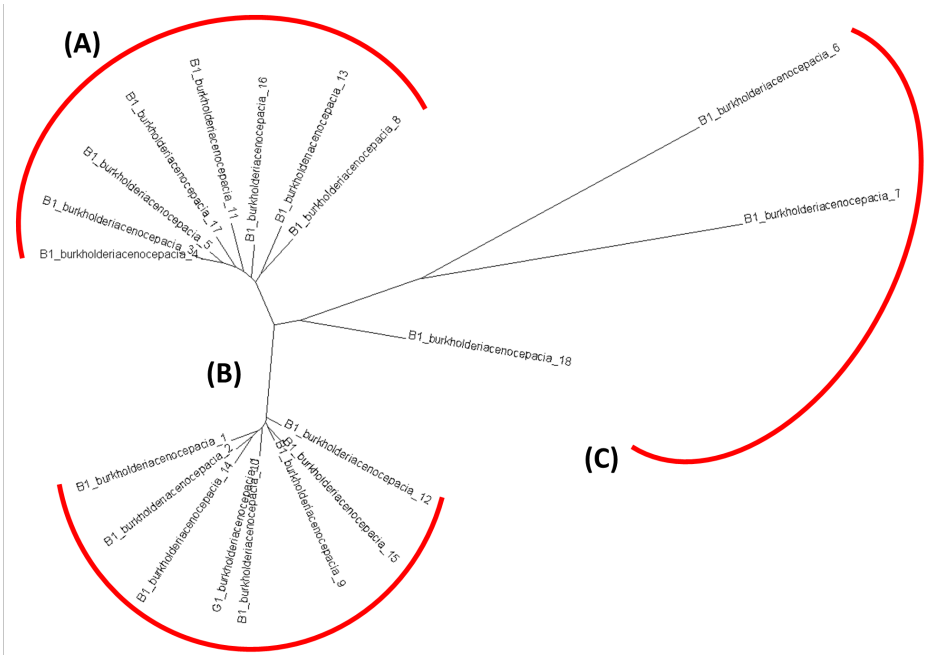


Figure 3.11: Maximum-likelihood *B. cenocepacia* phylogenetic tree obtained with Pagoo. Cluster A: Strains 3, 4, 5, 8, 11, 13, 16, 17; Cluster B: Strains 1, 2, 9, 10, 12, 14, 15, G1; Cluster C: Strains 6, 7, 18. Visualization on Dendroscope.

The tree for *B. cenocepacia* identifies 3 clusters similar to Roary and Panaroo, unlike mashtree's initial tree, which didn't include strain 18 in cluster C. In *B. pseudomallei* tree, we can observe the highest amount of lone strains (strains 1, 4, 10, 12 and 13), but we can identify 4 clusters: Cluster A with Strains 11 and 14; Cluster B With both strains 2 and strain 8; Cluster C with both strains 3 and strain 9; and Cluster D with Strains 5, 6 and 7. In the *B. mallei* tree, we can observe 3 Clusters. Cluster A with Strains 1, 3, 4 and 6; Cluster B with strains 7 and 8; and Cluster C with Strains 2, 5 and 9, where 2 and 5 are almost identical. In the *B. cepacia* tree, we see a high amount of lone strains (Strain 6, 10, 7 and 2) but we do see 3 Clusters: Cluster A with Both strains 9; Cluster B with Strains 3 and 8; Cluster C with Both Strain 4, Both strain 1, Strain 5 and strain 11. In the *B. contaminans* tree, we can observe 2 clusters: Cluster A with Strain 1 and 2; and Cluster B with strains 3, 4 and 5. In the *B. gladioli* tree, we can observe a single cluster with Strains 1 and 2. In the *B. multivorans* tree, we can observe 5 Clusters: Cluster A with strains 2, 3, 5, 8 and 9; Cluster B with Strains 10 and 12; Cluster C with strains 11 and 7; Cluster D with Both strains 1; and Cluster E with Strains 4 and 13. In the *B. thailandensis* tree, we can observe 3 Clusters: Cluster A with strains 2 and both 4s; Cluster B with strains 3, 6 and 9; And Cluster C with strains 1 and 8.

3.3.2.E Population genetics assay for *B. cenocepacia*

We Identified three populations for *B. cenocepacia* (Table A.8), Geographic Location, Medical Vs Environmental strain, and Body location of the strain. We recreated the same Pagoo trees with these populations in mind and obtained in Figure 3.12, 3.13 and 3.14.

In terms of geographic location, we can identify A cluster with only species from North America, which corresponds to cluster A containing strains 3, 4, 5, 8, 11, 13, 16 and 17. We can also see Strain 6 and 7, which are close and both from Oceania. In terms of Isolation source, it becomes harder to define a meaningful cluster but if we disregard the strain with unknown source (green), we see a cluster with 6 strains of a medical source, corresponding to cluster B, where the water source strain is strain 12, and the rest of the Medical strains are strain 1, 2, 9, 10, 14 and 15. The tree defining populations by body part found was inconclusive.

3.3.3 Multi-species pangenomes

With Pagoo being sufficiently optimized to run more complex datasets thanks to its integration to the R environment, we were able to build not only intraspecies pangenomes but interspecies ones as well. We grouped the 8 *Burkholderia* species into 4 datasets: the Bcc (*B. cenocepacia*, *B. cepacia*, *B. contaminans*, *B. multivorans*); the mallei group(*B. mallei*, *B. pseudomallei*, *B. thailandensis*); mallei group with *B. gladioli*; and all species. The reasoning to group *gladioli* to the mallei group was to be able to include it into the group vs group analyses if we wanted to (phylogeny wise, *B. gladioli* is closer to *B.*

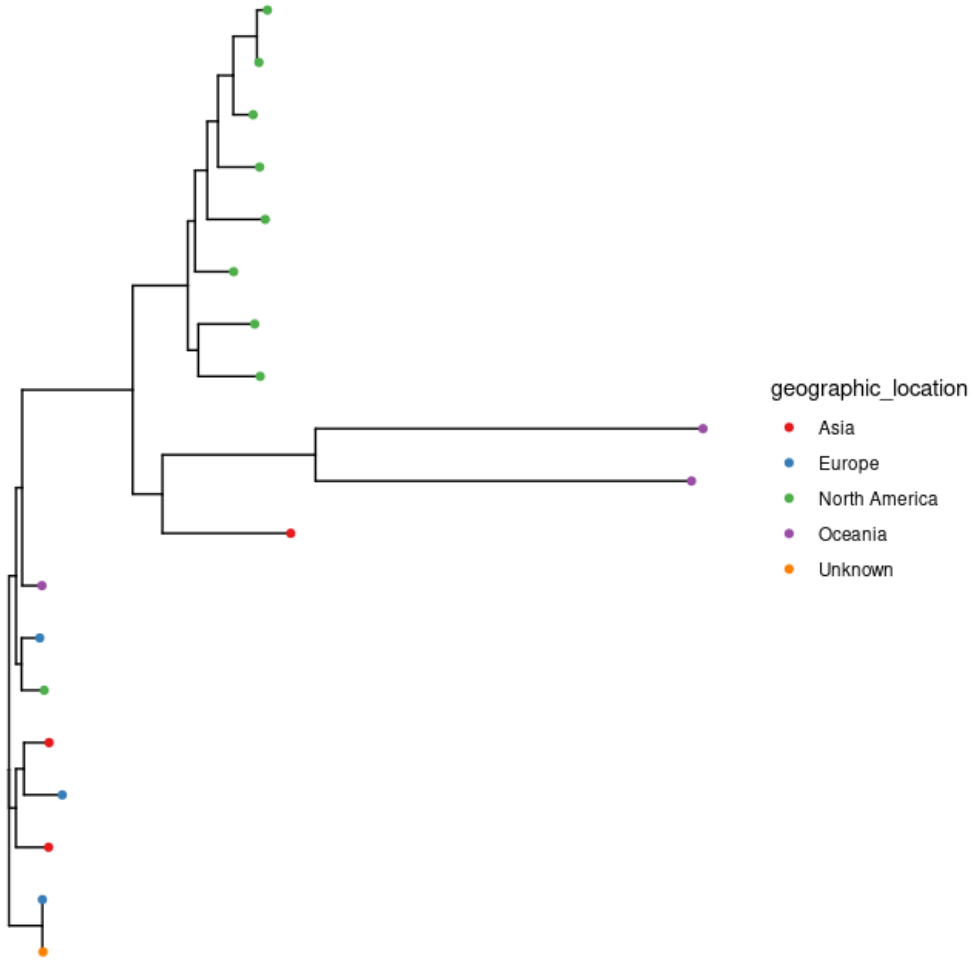


Figure 3.12: *B. cenocepacia* - Phylogenetic tree with populations defined by geographic location.

mallei than the Bcc), and see how integrating a species from a different group would cause the results to alter. As we did in the Intraspecies cases, we tested various E-value in search for the one with the largest core-genome (Table A.13).

Multi-species pangenome core and accessory gene number for each optimal E-value is contained in Table 3.7. As expected, adding more genomes decreased core gene count, and *B. gladioli* caused significant reduction in core gene count when coupled with the mallei group, lower even than the pangenome of all species as well.

Table 3.7: Core, shell, cloud, and total genes calculated by Pagoo's command `pg$summary_stats` for each group.

Group name	Optimal E-value	Core genes	Shell genes	Cloud genes	Total genes
Bcc group	80	2913	7792	49080	59785
mallei group	70	2852	2874	31238	36964
mallei group + *B. gladioli*	63	2192	5962	30233	38387
All species	70	2243	12530	72181	86954

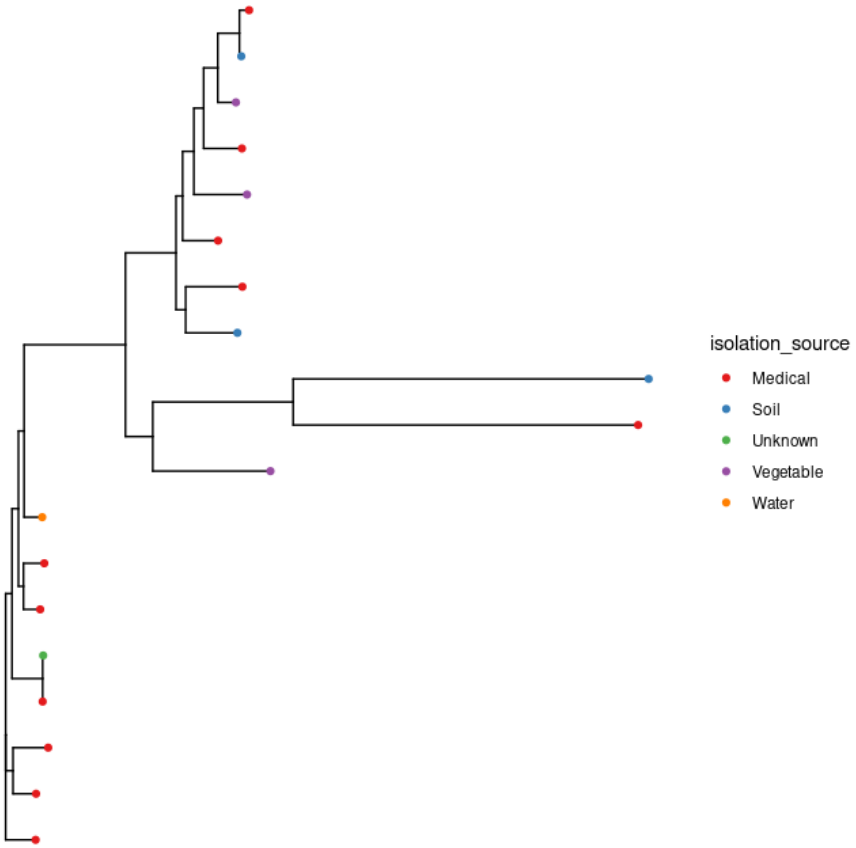


Figure 3.13: *B. cenocepacia* - Phylogenetic tree with populations defined by isolation source.

The Bcc, like in individual species of the group, seem to have higher shell ratio compared to the mallei group (Figure 3.21a). The group pangenomes also appear to be open like the individuals 3.21c. PCA plot shows clusters that indicates the distinguishing between species instead of strains. One example is the clear distancing between *B. gladioli* and the rest of the mallei group species (3.21d). We built a Phylogenetic tree with all the genomes in this work. The creation of the phylogenic tree using Pagoo was lengthy, taking approximately one week in BSRG2 to perform the core alignments of the whole dataset.

Phylogeny was able to separate the 3 groups (Bcc, mallei, plant pathogens) into clusters successfully (Figure 3.22). *B. pseudomallei* and *mallei* are very close and with low variability within the species. *B. thailandensis*, as expected, belongs to the mallei group but with some clear distance. In regards to the PCA plot of the mallei group, we can it seems that *mallei* and *pseudommalei* are related in the first Component, and *thailandensis* is the further away cluster. On the Bcc front, it proved difficult to distinguish between the different species, as we could see many examples of outliers and confusing branches. In a very large pangenome with all these groups included, there could have been a lack of resolution, or in another words, too much noise, which would make differenciation between the Bcc, which are already a very diverse and high variability group, very challenging in this tree. Regardless

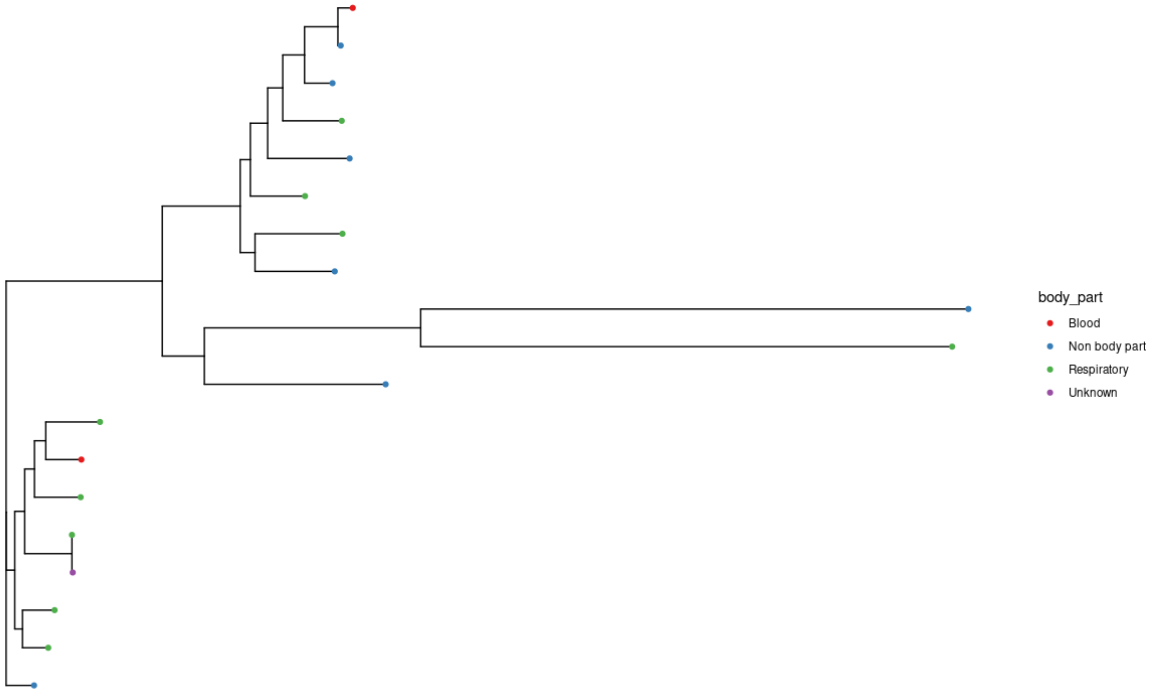


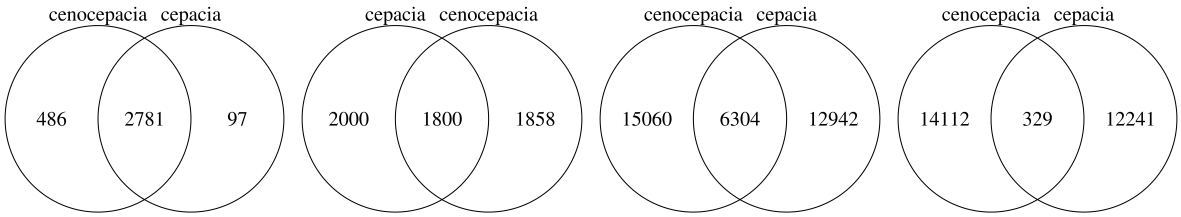
Figure 3.14: *B. cenocepacia* - Phylogenetic tree with populations defined by body part found.

of this, the ”All Species” phylogenetic tree fulfilled its function of distinguishing the different groups in the Genus. To further analyze the Bcc, it would be required to do a core alignment of the Bcc group pangenome.

3.3.4 Highlighting the differences between the pangenomes

With access to the pangenome genes and gene family codes, we used simple mathematical operations and manipulated the datasets to compare them in different ways. For example, we can Intersect the pangenome of 2 species to obtain all genes common to those 2 species. From that we can also subtract the intersection from one of the genomes to obtain all the unique genes of one species in regard to the other. We performed this and we created Venn diagrams showing the contribution of each species to the Core, shell and cloud genes. We decided to showcase 4 examples of these pairwise Venn diagrams. Figure 3.15 represents the pairwise venn diagrams between *B. cenocepacia* and *B. cepacia*. These 2 species share 2781 core genes but 486 of the core genes are unique do *B. cenocepacia* while 97 are unique to *B. cepacia*. We see that most of the common genes between these two species are in the core, but there are still 1800 common genes in the shell, which may represent a significant portion of the reason these 2 species are very alike. Even in the cloud genes there are still 329 common genes.

In figure A.12, we see the same diagrams for the comparison between *B. pseudomallei* and *B. mallei*, that so far are presented as very similar species. 3045 genes are common in the core, an overwhelming



(a) Core **(b)** Shell **(c)** Pangenome **(d)** Cloud

Figure 3.15: *B. cenocepacia* pangenome versus *B. cepacia* pangenome Venn diagrams.

proportion. *B. mallei* only contains 50 unique genes in comparison with *B. pseudomallei*'s 831. These 2 species present 112 common shell genes. Figure A.13 represents the comparison between all Bcc species. We can see here that by far, the pair *B. cenocepacia* and *B. cepacia* share the most shell and cloud genes unique to those 2 species, when compared to the other comparisons (1138 shell genes vs 133 between *B. cepacia* and *B. contaminans*, 37 between *B. contaminans* and *B. multivorans* and 254 between *B. multivorans* and *B. cenocepacia*, for example). In terms of the core genes, *B. contaminans* contains 1044 unique core genes, much higher than the others. Finally, Figure A.14 shows the same group comparison for the mallei group. As expected, *B. thailandensis* contains the highest amount of unique core genes.

3.3.5 Unique annotations and KEGG pathways between pangenome pairs

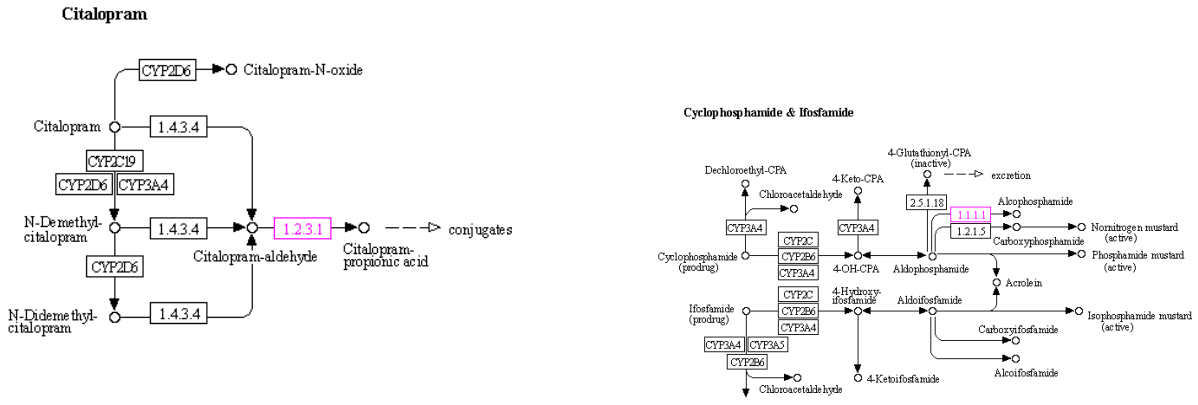
We selected the comparisons between *B. cenocepacia* and *B. cepacia*, *B. pseudomallei* and *B. mallei* and created a new comparison, a pairwise core genome comparison between the whole Bcc group and the mallei group (this comparison is represented by the venn diagram in figure A.11), to further explore the unique core genes found in these comparisons. We enriched the data through an hypergeometric test and obtained the most common unique annotations between the cores of these pairs. Table 3.8 shows the count of the most common core gene annotations from the core genes that are unique among the Bcc and the mallei group. We see annotations that show in both groups, like "membrane protein", "LysR family transcriptional regulator", "MFS transporter" and "AraC family trancriptional regulator". This means that there are different core genes coding the same type of proteins. We can see type III and type VI secretion proteins that are only present in high quantity in the mallei group, as we can see MarR and GntR families transcriptional regulators only in the Bcc. Tables A.9 and A.10 show the same results but for the two other pair comparisons. We can see many families of trancriptional regulators unique to *B. cenocepacia* and we can see the "biopolymer transporter ExbD" unique to *B. cepacia*. It becomes hard to observe the comparison in *B.pseudommalei* and *B. mallei* since there are very few unique core genes in the latter compared to the first.

We employed KEGG to identify metabolic pathways in the unique core genes for each of these

Table 3.8: Core genes unique to Bcc versus core genes unique to mallei group.

Core Genes Unique to mallei Group	Count	Core Genes Unique to Bcc	Count
hypothetical protein	6511	hypothetical protein	4694
membrane protein	941	LysR family transcriptional regulator	1622
type VI secretion protein	377	AraC family transcriptional regulator	653
LysR family transcriptional regulator	291	membrane protein	590
type III secretion system protein	273	ABC transporter substrate-binding protein	451
AraC family transcriptional regulator	242	alpha/beta hydrolase	446
transcriptional regulator	202	MFS transporter	413
MFS transporter	161	ABC transporter permease	366
ABC transporter permease	148	MarR family transcriptional regulator	262
polyketide cyclase	134	GntR family transcriptional regulator	244
sensor histidine kinase	123	Lrp/AsnC family transcriptional regulator	238
lipoprotein	111	N-acetyltransferase	238
type III secretion protein	110	cytochrome c	232
capsular polysaccharide biosynthesis protein	110	TetR family transcriptional regulator	224
glycosyl transferase	109	porin	208

comparisons. Table 3.9 shows the metabolic pathway category for each core gene in the Bcc versus mallei group comparison. The mallei group contains pathways not mentioned in the Bcc like nitrogen and pyrimidine metabolisms, while the Bcc contains Arginine biosynthesis and Fructose and mannose metabolism. Common pathways between the two can be seen in Bold, like the Pentose phosphate pathway, Purine metabolism and Cytochrome P450 drug metabolism. When we search the unique identifiers for Cytochrome P450 drug metabolism in KEGG, we see that the Bcc has unique genes in the Citalopram metabolism, while the mallei group has unique genes in the Cyclophosphamide and Ifosfamide metabolism (Figure 3.16). Unique gene features in the Pentose Phosphate pathway can be seen in Figure A.10.



(a) Bcc **(b)** mallei group

Figure 3.16: KEGG Cytochrome P450 Drug Metabolism - Bcc Citalopram metabolism (a), mallei group Cyclophosphamide and Ifosfamide metabolism (b)

Table 3.9: KEGG pathways - core genes unique to Bcc vs core genes unique to mallei group. Pathways present in both datasets are marked in bold.

Core Genes Unique to mallei Group - Pathways	Count	Core Genes Unique to Bcc - Pathways	Count
Non metabolic	596	Non metabolic	708
Metabolic pathways	128	Metabolic pathways	148
Biosynthesis of secondary metabolites	77	Biosynthesis of secondary metabolites	74
Pentose phosphate pathway	49	Non allowed character, can't be tested	63
Microbial metabolism in diverse environments	35	**Pentose phosphate pathway**	31
Non allowed character, can't be tested	14	Microbial metabolism in diverse environments	30
Nitrogen metabolism	11	Pentose and glucuronate interconversions	13
Pyrimidine metabolism	8	Arginine biosynthesis	12
Starch and sucrose metabolism	8	Fructose and mannose metabolism	12
Drug metabolism - other enzymes	7	**Purine metabolism**	10
Methane metabolism	6	**Glutathione metabolism**	9
Purine metabolism	6	Tryptophan metabolism	8
Drug metabolism - cytochrome P450	5	Valine, leucine and isoleucine degradation	8
Glycolysis / Gluconeogenesis	5	Alanine, aspartate and glutamate metabolism	6
Metabolism of xenobiotics by cytochrome P450	5	Butanoate metabolism	6
Pyruvate metabolism	5	**Drug metabolism - cytochrome P450**	6
Sulfur metabolism	5	Glyoxylate and dicarboxylate metabolism	6
Glutathione metabolism	4	Amino sugar and nucleotide sugar metabolism	5
Glycerolipid metabolism	4	Arginine and proline metabolism	5
Glycine, serine and threonine metabolism	4	Cysteine and methionine metabolism	5

3.4 Graph-based pangenome

3.4.1 Graph statistics

Through VG and ODGI's statistical output we can compare the pangenome graphs.

Table 3.10: Output generated by `odgi stats` and `vg stats -s` 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	Steps
B. thailandensis	6.745.675	3.308.693	1.122.022	2.270.461	11	19	58.455.554
B. pseudomallei	7.223.274	645.677	368.175	737.633	10	32	115.289.278
B. multivorans	6.649.370	2.629.212	1.437.430	2.880.174	22	46	91.844.470
B. mallei	5.696.930	402.140	246.482	503.140	14	18	51.096.412
B. gladioli	8.741.840	10.692.791	5.965.340	12.698.168	20	18	28.985.008
B. contaminans	8.587.033	4.816.475	3.335.895	7.411.740	12	20	40.657.651
B. cepacia	7.981.127	14.657.785	3.504.760	6.857.892	36	49	106.624.449
B. cenocepacia	7.653.437	3.170.570	1.784.153	3.504.743	38	55	136.339.539

B. pseudomallei and *mallei*'s small graph sizes may preliminarily explain less distance 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.

3.4.2 Viewing the pangenome

Pangenome graphs are complex in nature, and searching for a way to visualize the pangenome without consuming many resources is very important. Here, we showcase two different ways to observe the pangenome. Gfaestus produces an interactable interface to view and manipulate the pangenome graph, and the `odgi viz` command produces a static 1D image of the variation graph. Although Bandage also produces a full interactable drawing of the graph pangenome, it failed to load the largest graphs, so we only used Bandage for smaller graphs, which will be shown further. Figure 3.17 shows these visualization techniques applied to *B. cenocepacia*'s pangenome graph. In the sorted graph (Figure 3.17b) we can observe clearly the lack of alignment in the last sections of the graph, corresponding to the plasmids and chromosome 3. The remaining visualizations for the other 7 species can be consulted in the thesis' github repository [159].

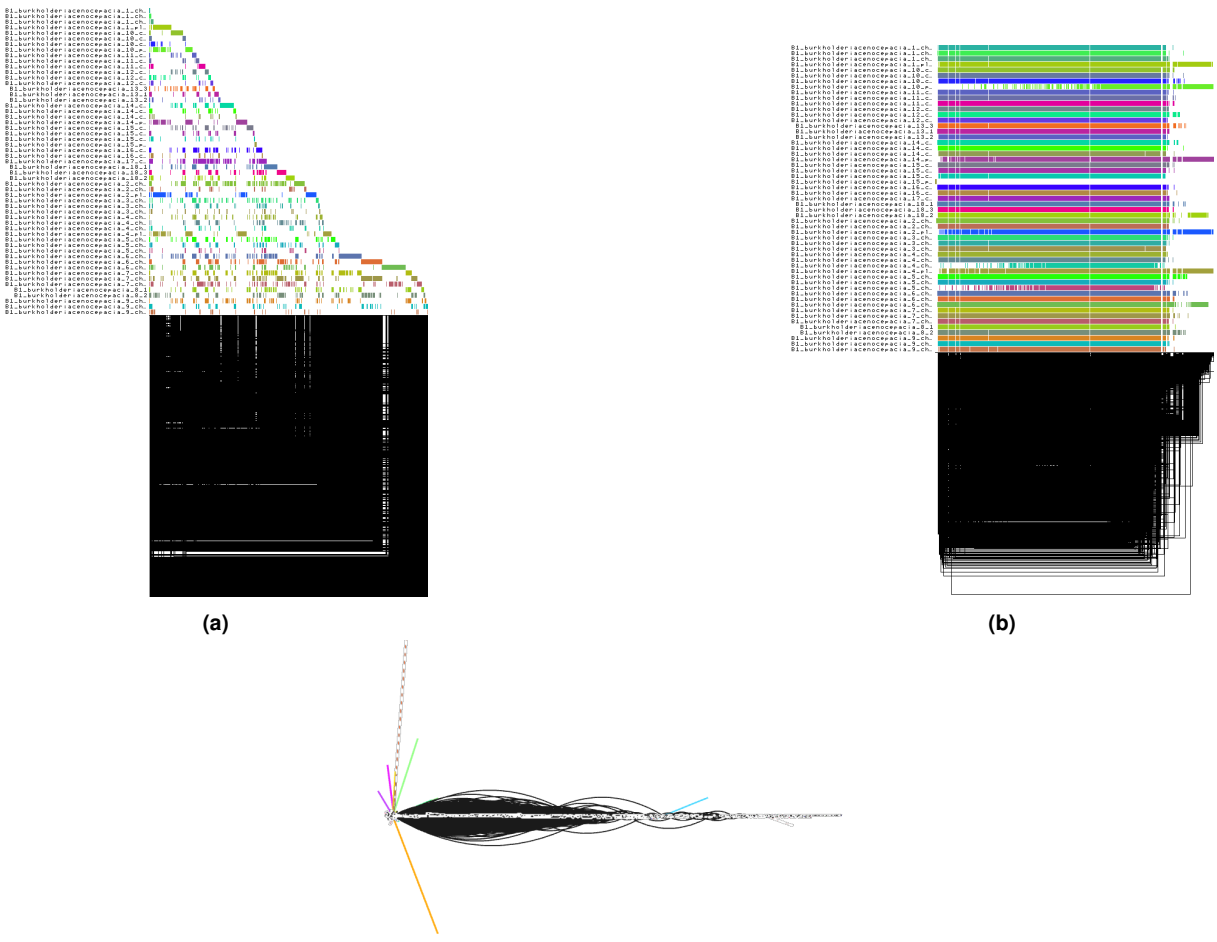


Figure 3.17: *B. cenocepacia* - variation graph visualization:(A) odgi viz unsorted, (B) odgi viz sorted, (C) gfaestus.

3.4.3 Detection of complex regions in the graph pangenomes

The ODGI toolkit was used to attempt to detect complex regions of the pangenome through the use of mainly the `odgi depth` command. `odgi depth` allows the identification of repetitive sequences, the number of times in which a node is crossed by all paths in the graph. We then calculated the mean depth per region (a region is just a section of a graph with start and ending positions, in this case, a region was 1000 graph positions), and plotted the result. The scatter plots, containing millions of dots, were complex and convoluted. Overall, the depth for every chromosome was pretty constant in very high values but often with different mean depth values in each chromosome. In some positions, the depth decreased, but only in very rare occasions there were a significantly large section of a chromosome with different depth that would indicate us towards a particularly complex region, as lower node depth corresponds to regions with less paths going through those nodes, meaning that those nodes are only used by certain strains. The idea of detecting complex regions this way originated from human pangenomes, which most likely work very differently to our bacterial pangenomes.

3.4.4 Jaccard distance-based phylogeny using odgi similarity (*B. cenocepacia*)

Using `odgi similarity`, we calculated the distances between each path in the variant graph for *B. cenocepacia*. We created a phylogenetic tree from the jaccard distances obtained, represented in Figure 3.18. We can take a look at the similarities between chromosomes on the tree on the left. We can observe Two large clusters with some sub clusters. Cluster A: Contains most chromosomes 3 and plasmids (Cluster A1: 10 Chromosomes 3; Cluster A2, 5 chromosomes 3; Cluster A3: 5 plasmids. Cluster B: Contains most chromosomes 1 and 2 (Cluster B1: 13 Chromosomes 1; Cluster B2: 12 Chromosomes 2). The graph-based pangenome similarity tree, on the right, shows differences between clusters compared to the gene-based trees. 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.

3.4.5 Bubble data and bubble graphs

Bubblegun computes the bubbles in a variation graph. This process took approximately a week per graph on the Home computer. Graphs that presented very low amount of bubbles like *mallei* and *pseudomallei* took a day each to compute. The amount of bubbles is synonymous with genetic variation as bubbles represent alternative genotypes. For *B. cenocepacia*, for example, 4% of nodes are located in bubbles. On the other pangenomes, 1% or lower of nodes are located in them.

Long chains represent genetic variations that involve mutations with high bp amount, i.e very large insertions or deletions. *B. cenocepacia* has chains as big as 19000bp. We can view the simple and

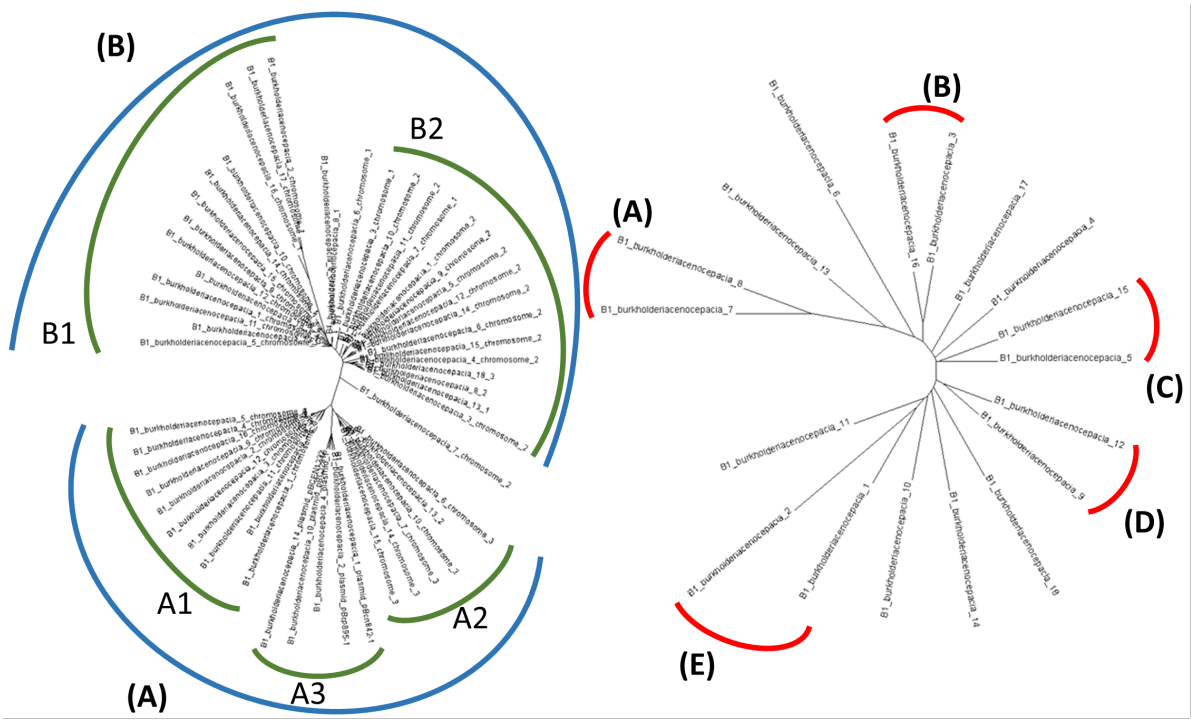


Figure 3.18: (Left) *B. cenocepacia* tree showing chromosome and plasmid similarity. Each branch is a path in the graph. We can observe Two large clusters with some sub clusters. Cluster A: Contains most chromosomes 3 and plasmids (Cluster A1: 10 Chromosomes 3; Cluster A2, 5 chromosomes 3; Cluster A3: 5 plasmids. Cluster B: Contains most chromosomes 1 and 2 (Cluster B1: 13 Chromosomes 1; Cluster B2: 12 Chromosomes 2). (Right) The same tree with the paths grouped into the 18 *cenocepacia* strains, showing the similarities between 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.

Table 3.11: Statistics about the bubbles present in the pangenome graphs generated by Bubblegun. Coverage is the percentage of the sequence/nodes that can be represented by bubble chains. Data for *B. cepacia* is missing due to time constrains.

Species name	Simple Bubbles	Super Bubbles	Insertions	Sequence Coverage (%)	Node Coverage (%)	Longest Chain (bp)
B. thailandensis	3263	3	357	1.916	1.039	25378
B. pseudomallei	3	0	1	0.0040	0.0041	12
B. multivorans	2364	12	572	0.837	0.727	5329
B. mallei	28	0	7	0.647	0.0499	1250
B. gladioli	9782	322	491	0.935	0.696	11790
B. contaminans	5448	213	358	2.297	0.700	58369
B. cepacia	-	-	-	-	-	-
B. cenocepacia	12191	2046	686	6.525	3.772	19367

superbubbles, as bubblegun can also output a graph that only contains these variations. The bubble graph drawn by Bandage for *B. cenocepacia* can be seen in Figure 3.19, where we can see the 19367bp chain clearly. The remaining bubble graphs are showcased in Figures A.16, A.18, A.15, A.19 and A.17. We can see the largest bubbles mentioned in table 3.11 for *B. contaminans* and *B. multivorans*. We can see more clearly examples of simple bubbles in *B. mallei* and *B. pseudomallei*, and in the latter we can

see an insertion (bottom left).

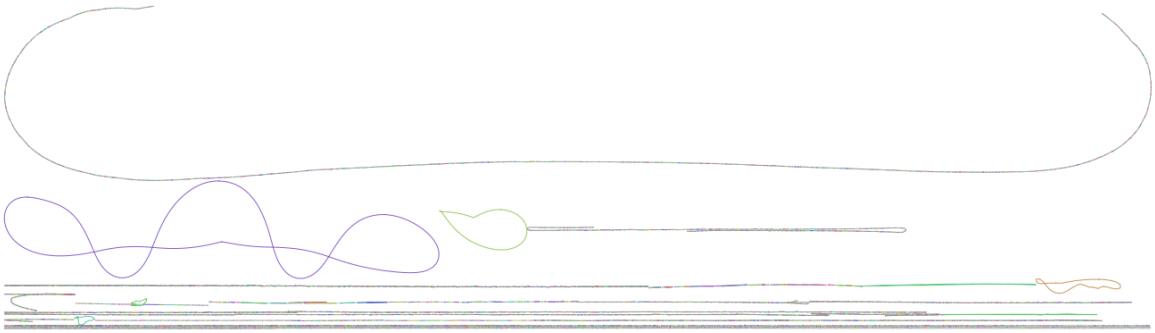


Figure 3.19: *B. cenocepacia* bubble graph. Visualization on Bandage.

3.4.6 Index files and VCF variant data

We obtained a vcf file for *B. cenocepacia* and *B. pseudomallei* with a bubble type annotation. It is simultanealy a FASTA storage file and a file that represents a multiple genome alignment that is contained in the graph. It contains the position in every chromosome with a variation, as well as indicating the reference and alternative alleles and the traversal of the allele as a path in the graph. A snapshot of the VCF file for *B. cenocepacia* obtained can be seen in Figure 3.20, containing 14933 variations. The VCF for *B. pseudomallei* contains 4 variations.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	631	>118665>118667	T	C	60	.	AC=0;AF=0;AN=0;AT=>118665>118666>118667,>118665>548710>118667;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	679	>118679>118681	T	C	60	.	AC=0;AF=0;AN=0;AT=>118679>118680>118681,>118679>548713>118681;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	769	>118703>118705	G	C	60	.	AC=0;AF=0;AN=0;AT=>118703>118704>118705,>118703>548717>118705;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	894	>118732>118734	A	G	60	.	AC=0;AF=0;AN=0;AT=>118732>118733>118734,>118732>548721>118734;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1351	>118854>118856	C	T	60	.	AC=0;AF=0;AN=0;AT=>118854>118855>118856,>118854>548752>118856;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1390	>118866>118868	C	A	60	.	AC=0;AF=0;AN=0;AT=>118866>118867>118868,>118866>548757>118868;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1407	>118872>118873	T	TCGGGG	60	.	AC=0;AF=0;AN=0;AT=>118872>118873,>118872>548762>118873;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1434	>118882>118884	GG	GAC	60	.	AC=0;AF=0;AN=0;AT=>118882>118883>118884,>118882>548770>118884;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1603	>118932>118934	G	A	60	.	AC=0;AF=0;AN=0;AT=>118932>118933>118934,>118932>548790>118934;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1792	>118982>118984	T	C	60	.	AC=0;AF=0;AN=0;AT=>118982>118983>118984,>118982>548812>118984;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1853	>119000>119002	G	A	60	.	AC=0;AF=0;AN=0;AT=>119000>119001>119002,>119000>548817>119002;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1936	>119026>119028	T	G	60	.	AC=0;AF=0;AN=0;AT=>119026>119027>119028,>119026>548831>119028;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	1978	>119038>119040	G	T	60	.	AC=0;AF=0;AN=0;AT=>119038>119039>119040,>119038>548836>119040;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	2002	>119047>119049	T	C	60	.	AC=0;AF=0;AN=0;AT=>119047>119048>119049,>119047>548839>119049;NS=0;LV=0	GT	
B1_burkholderiacenocepacia_10_chromosome_1:1222905-1315219	2118	>119083>119085	A	G	60	.	AC=0;AF=0;AN=0;AT=>119083>119084>119085,>119083>548864>119085;NS=0;LV=0	GT	

Figure 3.20: Snapshot of the *B. cenocepacia* VCF file, showing the variations found in the exact positions in the graph.

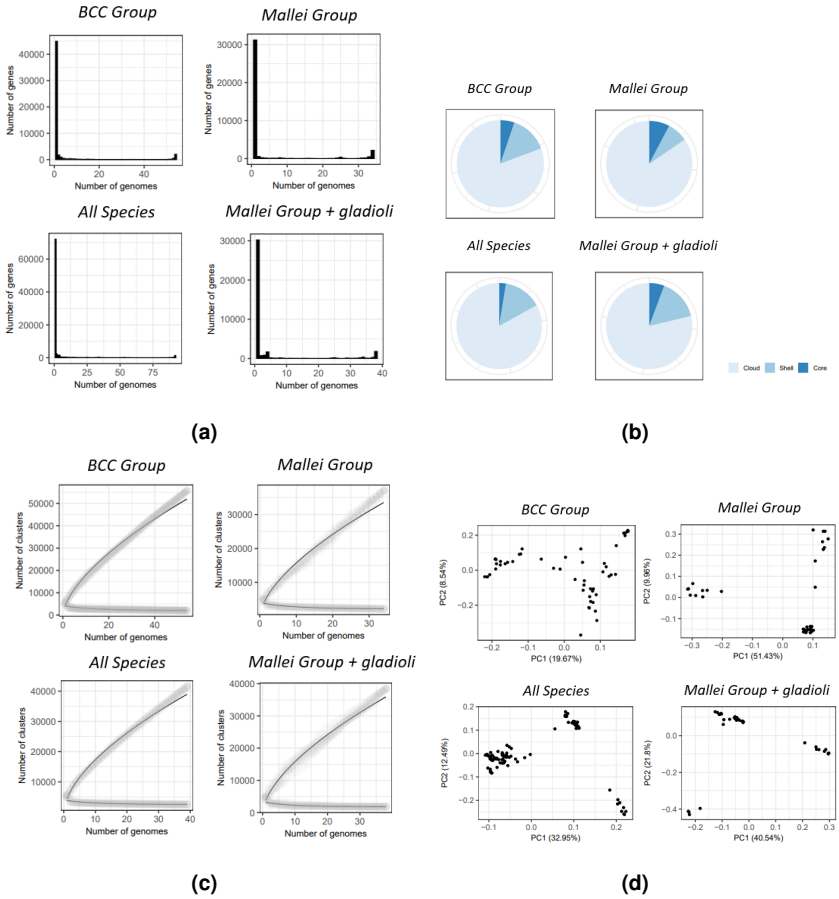


Figure 3.21: Pagoo generated plots for the 4 species groups: (A) `pg$gg_pie` plots representing pie charts separating the pangenome in the proportions of core, shell and cloud genes.(B) `pg$gg_barplot`. Relaying the previous bar plot's description, it represents in how many genomes/strains a certain number of genes are present, for every species. The core is represented by the last bar in each plot (i.e the number of genes present in all genomes), while the cloud is mostly represented by the first bar (i.e the number of genes that are present in only one genome).(C) `pg$gg_curves`. It shows an upper curve which represent the increase in pangenome clusters with the increase in pangenome genomes and a lower curve representing the decrease in core clusters with the increase in pangenome genomes. (D) Principal component analysis (PCA) plots obtained by Pagoo's `pg$gg_pca` command. Each dot represents an individual strain/genome.

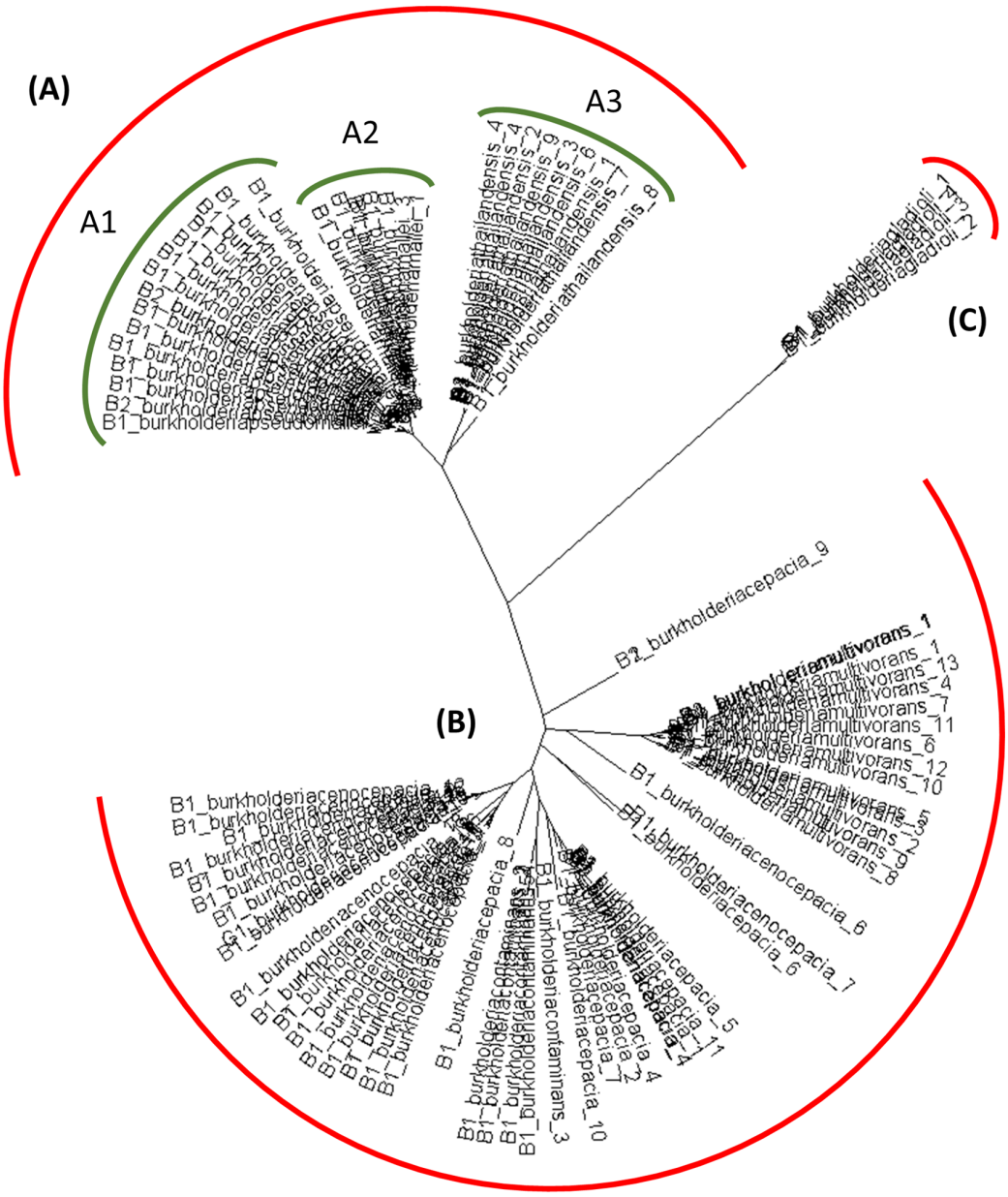


Figure 3.22: *Burkholderia* genus maximum-likelihood phylogenic tree obtained with the Pagoo framework. Contains the phylogeny relationships between *B. thailandensis, B. mallei, B. pseudomallei, B. gladioli, B. multivorans, B. cenocepacia, B. cepacia and B. contaminans*. Visualization using Dendroscope: (A) mallei group (We can see the clear separation between *B. pseudomallei* (A1), *B. mallei* (A2) and *B. thailandensis* (A3). (B): Bcc group. (*B. multivorans, B. cenocepacia, B. cepacia and B. contaminans*). (C): *B. gladioli*, representing the plant pathogens group.

4

Discussion and Final Remarks

Contents

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.

4.1 Interpretation of findings

We successfully built gene-based pangenomes with Roary, Panaroo and Pagoo for eight *Burkholderia* species. In Pagoo's case, we've also built multi-species pangenomes, including one with all the datasets. The methodology used for these pangenomes proved to be simple and fast. In terms of the pangenomic analysis, we discovered that Roary produces the smallest core-genome sizes by far in comparison to the other two methods. Between Panaroo and Pagoo, there are some discrepancies between species, some with higher number of core-genes with Panaroo, others with Pagoo. However, Panaroo and Roary both failed to produce cloud genomes in some cases, reducing the viability of these softwares. If we had the same pie/bar figures for Roary and Panaroo as we did in Pagoo, they would have most likely demonstrated that the behaviour of the 2 former softwares is more erratic, and that they are dependant of a set of reference proteins that is short for the number of bacterial phylogenetic clades that these softwares were applied on, while in Pagoo, we had all the family codes in the GenomeDB database. With Pagoo, 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: a larger amount of datasets, a consistent number of datasets for all species. 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 maximum-likelihood trees. Smaller or larger core sizes seemed to be less relevant in this approach as the clusters identified differed minimally. Our tree containing all eight studied species was able to separate the 3 groups, Bcc, mallei group and plant pathogens, very consistently. It was even able, in the mallei group case, to distinguish between all three species (*B. mallei*, *B. pseudomallei* and *B. thailandensis*. However, it was not able to separate the species inside the Bcc cluster. This is because knowing the species of this group

are highly variable and difficult to distinguish, trying to separate them into clear clusters inside of a tree with a lot of noise, i.e comparing with the genes of the other species in the other groups as well as the ones inside the Bcc proved too difficult to provide clear clusters. The small populations genetics assay on *B. cenocepacia* was able to show a cluster of the strains entirely from north america, suggesting an environmental genomic difference between the strains of the species. In terms of genomic fluidity, the average fluidity for the 8 species was 0.256, indicating that the clusters are much more identical than fluid, since they are much closer to 0 than 1. Out of all the species, the fluidity of *B. cepacia* and *B. cenocepacia* were the highest, corroborating the variability of these two species. In the size estimation test, all species, 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. Neutral genes introduced some interesting results. *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. Over 50% of *B. multivorans*'s 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. When it comes to *B. cenocepacia*'s gene annotations from Tajima's D test, we identified that most genes evolving neutrally are involved in processes like transportation of substrates across cellular membranes, enzymes involved in redox reactions and delivery of toxins into target cells. The gene family under most selective pressure was the "HAD-IB family hydrolase" which is an enzyme that plays a crucial role in breaking down haloalkanoic acids, phosphates and sugar phosphates. The gene family with the lowest Tajima score was the "chromosomal replication initiator protein DnaA", which plays a fundamental role in the regulation and start of DNA replication.

The Venn diagrams produced gave us insight of what differentiates pairs of *Burkholderia* species and their groups. We identified that *B. cenocepacia* and *B. cepacia*'s cores are very similar with 2781 common core gene families in a total of 3364, and had many similarities even at the shell level, with 1800 common gene families. *B. pseudomallei* and *B. mallei*, on the other hand, are highly differentiated in the shell genome. *B. pseudomallei* contains 831 unique core genes versus *B. mallei*'s 50. This could be due to the amount of datasets available (16 vs 9), but if *B. mallei* originated when it differentiated from *B. pseudomallei* my losing a fair amount of genes, then this difference is justified and we could suggest that the genes *B. pseudomallei* uniquely has in the core are, in some part, the ones lost by *B. mallei* when it evolved. The Bcc Venn diagrams show a couple of interesting remarks. In the core diagram, we see that *B. contaminans* is the most unique species, with 1044 unique core genes, followed by *B.*

multivorans with 301. Most of *B. cenocepacia* and *B. cepacia* core genes in the group belong to the whole group. In the shell, we continue to see what happened in the *B. cenocepacia* - *B. cepacia* pair, Where most of the connected shell genes in the group are also part of that pair's shell. In the cloud, it even becomes clearer that *B. cenocepacia* and *B. cepacia* are the most alike with 308 shared cloud genes, about 5 times as much as any other pair connection. In the mallei group diagrams, We learn that more of the group similarities are mostly core genomes, as the common shell and cloud genes are scarce. When finding the unique core gene annotations for some of the pairs we discovered type III and type VI secretion proteins only present in high quantities in the mallei group, indicating this group's species use these secretion systems to inject virulence factors into host cells and is a high remark in pathogenecy involvement. In the Bcc group, we found MarR and GntR family trancriptional regulators, which are essential in bacterial adaptation and survival in environmentally difficult conditions, such as antibiotics or oxidative stress. The biopolymer transporter ExbD was present only in *B. cepacia* in the pairwise comparison with *B. cenocepacia*, suggesting us that *B. cepacia*, unlike *B. cenocepacia* is capable of transporting biopolymers such as iron-chelating compounds or vitamin B12 into the bacterial cell. When we looked at the corresponding KEGG pathways involving these core genes, we found, in the Bcc vs mallei group comparison, that the mallei group contains metabolic pathways not found in abundance in the Bcc such as nitrogen and pyrimidine metabolisms, while the Bcc are highly capable of Arginine biosynthesis and fructose and mannose metabolism. Arginine biosynthesis is a vital part of nitrogen metabolism as well, but also plays a role in bacteria virulence factors, which suggests a link between Bcc's virulence and arginine biosynthesis. The literature conveys that both mallei group and Bcc strains can metabolize pyrimidine and fructose, but we suggest that the higher presence of each of these metabolisms on the respective groups represent the environments where these bacteria groups are more prepared to deal with. We identified the metabolism of citalopram as a unique metabolism in the Bcc. They demonstrate the capacity to convert citalopram-aldehyde to citalopram propionic acid, and it could indicate a unique metabolic potential within these bacteria, as citalopram is a commonly used antidepressant. The mallei group was capable of converting aldophosphamide into alcophophamide, which are intermediates in the metabolism of cyclophosphamide and ifosfamide, which are both drugs used in chemotherapy. This could also indicate a unique metabolic potential in the mallei group.

We were also 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 GFA graph and has potential for use in future work. Gfaestus and `odgi viz` were able to create visualizations of our graphs. Unfortunately, Bandage could not open the full graphs as they were too big. We created bubble graphs with Bubblegun and were able to view those in Bandage, and identify variations in the

graphs such as insertions, simple bubble and super bubbles. For most species, the bubble graphs contained a high number of bubbles, with over 2000 simple bubbles and 300 insertions. *B. mallei* and *B. pseudomallei* were outliers by far, as these two species contained 38 and 4 variations, respectively. This suggests that the strains of these 2 species are incredibly homozygous, which corroborates what was seen in the all-species tree. We observed the tree for *B. cenocepacia* obtained with the graph pangenome. This tree was built from the entire pangenome dissimilarities, unlike the gene-based trees which were made completely with the core alignment distances. This shows why there is a significant difference in the tree.

We were able to take a peak at the variation data of *B. cenocepacia* and *B. pseudomallei* VCF files, which contained 14933 and 4 variations, respectively. This is comparable to the number of variations computed by Bubblegun, which was 14923 and 4. This implies at least, some consistency in the two methods to obtain the number of variations in a pangenome graph. However, while Bubblegun's run time for each pangenome was about one week, the run time for vg's production of a VCF file spanned from one to several months.

4.2 Difficulties and challenges

The *B. cenocepacia* species was the most explored species in this work, compared to the other seven *Burkholderia* species. This was due to time constraints caused by the graph-based methodologies being much more computer intensive. *B. cenocepacia* was the first species out of the eight to be processed, allowing more in-depth work on it compared to the others. Some of the graph-based processes such as the construction of the VCF files, are still ongoing on BSRG1 for over 4 months. We chose to work on 8 species in this thesis, and that may have been too over-ambitious, as this amount of species caused delays and exponentially increased the time of comparisons between them. Perhaps a study of only the four Bcc species would have been more adequate.

Some of the challenges in our work include a learning curve for users unfamiliar with graph-based genomic approaches and the need for parameter optimization to maximize performance, especially when dealing with highly complex or heterogeneous datasets.

As bacterial genomic population studies have grown larger, there has not been a corresponding increase in genome annotation accuracy or genome assembly contiguity. Thus, as these databases have grown, so has the number of erroneous gene annotations. This can have profound implications for the resulting estimates of the number of gene families present, whereby a higher number genomes leads to a higher number of errors [50]. Errors can be introduced into pangenome analyses by fragmented assemblies, mis-annotation, contamination and mis-assembly. Whilst errors often lead to inflations in the estimates of the size of the accessory genome, they can also lead to missing genes when the

annotation software fails to identify a gene or where the gene is fragmented by a break in the assembly, which reduces the estimated size of the core genome.

In the context of pangenomics, the field has traditionally focused on gene-based representations for prokaryotes. Genes constitute a significant proportion of their sequences, often over 90%, and variations in gene content can greatly influence traits such as pathogenicity and drug resistance. Consequently, the division of genes into core and dispensable categoriescategories can be insightful for understanding these phenotypes in prokaryotes.

However, the gene-centric approach is less suitable for eukaryotes, particularly those with large genomes exceeding 500Mbp [160]. In eukaryotes, a substantial portion, often more than 50%, of the genome is intergenic, and genes themselves contain lengthy introns. Eukaryotes also exhibit less frequent DNA exchange compared to bacteria, resulting in more stable gene content. For species like humans, where exons account for only around 2% of the genome, limiting the pan-genome to exonic sequences provides limited insight into within-species variations [161]. Hence, in eukaryotic pangenome studies, the definition of a pangenome typically includes all DNA sequences in a collection of genomes, encompassing intergenic regions. While the terminology of 'core' and 'dispensable' genomes is occasionally borrowed in eukaryotic pangenomics, it extends to intergenic sequences and considers unique sequences as 'singletons'. This broader approach is vital for capturing the comprehensive genomic diversity within eukaryotic species [162].

4.3 Future work

For future analyses, as a whole, we would want to analyze the range of bacteria chosen more deeply, instead of a *B. cenocepacia* focus. It would be interesting to create a Bcc phylogenetic tree so that we could study and distinguish the different species inside the Bcc, as we could not with simply the tree with all species. We produced the Panaroo graphs for all species but we did not venture deeper into the interpretation of the genes inside them. A future work may include identifying the genes both in linear and complex regions of the graph and their connections to the other genes, as more biologically relevant information could be retrieved from them. If not for the delay in the production of the VCF files, we would've eventually entered into a population genetics study of the species in *Burkholderia*. The "population genetics and genomics in R" [163] primer contains several recipes to process the VCF files to create phylogenetic trees and calculate genetic distances between populations. We could have used Syri to predict genomic differences between related genomes using the whole-genomes assemblies obtained [164]. Syri could identify the syntenic path (longest set of co-linear regions), structural rearrangements (inversions, translocations, and duplications), local variations (SNPs, indels, CNVs, etc..) within syntenic and structural rearrangements, and un-aligned regions. Lastly, we could have explored

a pangenome constructed from RNA-seq data, using, for example, rpvg [165]. With this pangenome we could emphasize transcript-level distinctions between individuals, typically excluding intergenic sequences and introns but encompassing alternative splice variants and other isoforms originating from a single genomic locus, offering insights into RNA-level diversity of our species.

4.4 Conclusion

In summary, the emergence of computational pangenomics represents a significant advancement in contemporary genomics research, offering the potential to close gaps in global genomic maps and gain comprehensive insights into the extent and nature of evolution. The transition from linear reference genomes to graph-based representations is a pivotal paradigm shift, and pangenomes have the potential to improve variant calling on linear references. Although some argue that linear genomic models will remain important, we suggest that graphical reference systems will proliferate as more whole-genome sequences become available. We can expand the idea of our own bacteria graphs and envision that in the near future, pangenome graphs will be able to play a vital role even in human genetics, personalized medicine, metagenomics, transcriptomics, and epigenomics, enhancing our understanding of genomic diversity and its applications in various fields.

Bibliography

[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." [Online]. Available: https://www.genome.gov/human-genome-project/Completion-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] R. J. Hall, J. L. Draper, F. G. Nielsen, and B. E. Dutilh, "Beyond research: A primer for considerations on using viral metagenomics in the field and clinic," *Frontiers in Microbiology*, vol. 6, 2015.

[8] C. P.-G. Consortium, "Computational pan-genomics: status, promises and challenges," *Briefings in Bioinformatics*, 2018.

[9] G. Vernikos, D. Medini, D. R. Riley, and H. Tettelin, "Ten years of pan-genome analyses," *Current Opinion in Microbiology*, vol. 23, p. 148–154, 2015.

[10] K.-P. Koepfli, B. Paten, and S. J. O'Brien, "The genome 10k project: A way forward," *Annual Review of Animal Biosciences*, vol. 3, no. 1, p. 57–111, 2015.

[11] G. E. Robinson, K. J. Hackett, M. Purcell-Miramontes, S. J. Brown, J. D. Evans, M. R. Goldsmith, D. Lawson, J. Okamuro, H. M. Robertson, and D. J. Schneider, "Creating a buzz about insect genomes," *Science*, vol. 331, no. 6023, p. 1386–1386, 2011.

[12] J. Armstrong, I. T. Fiddes, M. Diekhans, and B. Paten, “Whole-genome alignment and comparative annotation,” *Annual Review of Animal Biosciences*, vol. 7, no. 1, p. 41–64, 2019.

[13] B. Kehr, A. Helgadottir, P. Melsted, H. Jonsson, H. Helgason, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Gylfason, G. H. Halldorsson, and et al., “Diversity in non-repetitive human sequences not found in the reference genome,” *Nature Genetics*, vol. 49, no. 4, p. 588–593, 2017.

[14] R. M. Sherman and S. L. Salzberg, “Pan-genomics in the human genome era,” *Nature Reviews Genetics*, vol. 21, no. 4, p. 243–254, 2020.

[15] 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.

[16] 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.

[17] 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.

[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] X. Wang, H. Zhou, H. Chen, X. Jing, W. Zheng, R. Li, T. Sun, J. Liu, J. Fu, L. Huo, and et al., “Discovery of recombinases enables genome mining of cryptic biosynthetic gene clusters in burkholderiales species,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, 2018.

[20] W. Zheng, X. Wang, H. Zhou, Y. Zhang, A. Li, and X. Bian, “Establishment of recombineering genome editing system in paraburkholderia megapolitana empowers activation of silent biosynthetic gene clusters,” *Microbial Biotechnology*, vol. 13, no. 2, p. 397–405, 2020.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] I. J. Toesca, C. T. French, and J. F. Miller, "The type vi secretion system spike protein vgrg5 mediates membrane fusion during intercellular spread by pseudomallei group burkholderia species," *Infection and Immunity*, vol. 82, no. 4, p. 1436–1444, 2014.

[34] 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.

[35] G. I. Borlee, B. A. Plumley, K. H. Martin, N. Somprasong, M. R. Mangalea, M. N. Islam, M. N. Burtnick, P. J. Brett, I. Steinmetz, D. P. AuCoin, and et al., “Genome-scale analysis of the genes that contribute to burkholderia pseudomallei biofilm formation identifies a crucial exopolysaccharide biosynthesis gene cluster,” *PLOS Neglected Tropical Diseases*, vol. 11, no. 6, 2017.

[36] 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.

[37] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, and et al., “Gencode: The reference human genome annotation for the encode project,” *Genome Research*, vol. 22, no. 9, p. 1760–1774, 2012.

[38] J. Zhu, M. Adli, Y. Zou, James, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P. L. De Jager, and et al., “Genome-wide chromatin state transitions associated with developmental and environmental cues,” *Cell*, vol. 152, no. 3, p. 642–654, 2013.

[39] E. P. Consortium, “The encode (encyclopedia of dna elements) project,” *Science*, vol. 306, no. 5696, p. 636–640, 2004.

[40] A. M. Deaton and A. Bird, “Cpg islands and the regulation of transcription,” *Genes Development*, vol. 25, no. 10, p. 1010–1022, 2011.

[41] S. T. Sherry, “Dbsnp: The ncbi database of genetic variation,” *Nucleic Acids Research*, vol. 29, no. 1, p. 308–311, 2001.

[42] T. G. P. Consortium, “A map of human genome variation from population-scale sequencing,” *Nature*, vol. 467, no. 7319, p. 1061–1073, 2010.

[43] S. Letovsky, “Gdb: The human genome database,” *Nucleic Acids Research*, vol. 26, no. 1, p. 94–99, 1998.

[44] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. García Girón, T. Hourlier, and et al., “The ensembl gene annotation system,” *Database*, vol. 2016, 2016.

[45] K. D. Pruitt, T. Tatusova, and D. R. Maglott, “Ncbi reference sequences (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins,” *Nucleic Acids Research*, vol. 35, no. Database, 2007.

[46] B. L. Cantarel, I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado, and M. Yandell, “Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes,” *Genome Research*, vol. 18, no. 1, p. 188–196, 2007.

[47] D. Gordon, J. Huddleston, M. J. Chaisson, C. M. Hill, Z. N. Kronenberg, K. M. Munson, M. Malig, A. Raja, I. Fiddes, L. W. Hillier, and et al., “Long-read sequence assembly of the gorilla genome,” *Science*, vol. 352, no. 6281, 2016.

[48] N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church, and D. B. Jaffe, “Direct determination of diploid genome sequences,” *Genome Research*, vol. 27, no. 5, p. 757–767, 2017.

[49] G. K. C. of Scientists, “Genome 10k: a proposal to obtain whole-genome sequence for 10,000 vertebrate species,” *Journal of Heredity*, vol. 100, no. 6, p. 659–674, 2009.

[50] G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, and et al., “Producing polished prokaryotic pangenomes with the panaroo pipeline,” *Genome Biology*, vol. 21, no. 1, 2020.

[51] D. C. Jeffares, B. Tomiczek, V. Sojo, and M. dos Reis, “A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome,” *Methods in Molecular Biology*, p. 65–90, 2014.

[52] S. Kryazhimskiy and J. B. Plotkin, “The population genetics of dn/ds,” *PLoS Genetics*, vol. 4, no. 12, 2008.

[53] F. Tajima, “Statistical method for testing the neutral mutation hypothesis by dna polymorphism.” *Genetics*, vol. 123, no. 3, p. 585–595, 1989.

[54] B. Boeckmann, M. Marcet-Houben, J. A. Rees, K. Forslund, J. Huerta-Cepas, M. Muffato, P. Yilmaz, I. Xenarios, P. Bork, S. E. Lewis, and et al., “Quest for orthologs entails quest for tree of life: In search of the gene stream,” *Genome Biology and Evolution*, vol. 7, no. 7, p. 1988–1999, 2015.

[55] M. de Been, V. F. Lanza, M. de Toro, J. Scharringa, W. Dohmen, Y. Du, J. Hu, Y. Lei, N. Li, A. Tooming-Klunderud, and et al., “Dissemination of cephalosporin resistance genes between escherichia coli strains from farm animals and humans by specific plasmid lineages,” *PLoS Genetics*, vol. 10, no. 12, 2014.

[56] T. A. Williams, P. G. Foster, C. J. Cox, and T. M. Embley, “An archaeal origin of eukaryotes supports only two primary domains of life,” *Nature*, vol. 504, no. 7479, p. 231–236, 2013.

[57] L. L. Moroz, K. M. Kocot, M. R. Citarella, S. Dosung, T. P. Norekian, I. S. Povolotskaya, A. P. Grigorenko, C. Dailey, E. Berezikov, K. M. Buckley, and et al., “The ctenophore genome and the evolutionary origins of neural systems,” *Nature*, vol. 510, no. 7503, p. 109–114, 2014.

[58] B. Zhong, L. Sun, and D. Penny, “The origin of land plants: A phylogenomic perspective,” *Evolutionary Bioinformatics*, vol. 11, 2015.

[59] M. Eppinger, T. Pearson, S. S. Koenig, O. Pearson, N. Hicks, S. Agrawal, F. Sanjar, K. Galens, S. Daugherty, J. Crabtree, and et al., “Genomic epidemiology of the haitian cholera outbreak: A single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic,” *mBio*, vol. 5, no. 6, 2014.

[60] M. T. Holden, L.-Y. Hsu, K. Kurt, L. A. Weinert, A. E. Mather, S. R. Harris, B. Strommenger, F. Layer, W. Witte, H. de Lencastre, and et al., “A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant staphylococcus aureus pandemic,” *Genome Research*, vol. 23, no. 4, p. 653–664, 2013.

[61] C. D. Greenman, E. D. Pleasance, S. Newman, F. Yang, B. Fu, S. Nik-Zainal, D. Jones, K. W. Lau, N. Carter, P. A. Edwards, and et al., “Estimation of rearrangement phylogeny for cancer genomes,” *Genome Research*, vol. 22, no. 2, p. 346–361, 2011.

[62] C. S. Cooper, R. Eeles, D. C. Wedge, P. Van Loo, G. Gundem, L. B. Alexandrov, B. Kremeyer, A. Butler, A. G. Lynch, N. Camacho, and et al., “Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue,” *Nature Genetics*, vol. 47, no. 4, p. 367–372, 2015.

[63] G. Moreno-Hagelsieb, “Transcription unit conservation in the three domains of life: A perspective from escherichia coli,” *Trends in Genetics*, vol. 17, no. 4, p. 175–177, 2001.

[64] A. O. Kislyuk, B. Haegeman, N. H. Bergman, and J. S. Weitz, “Genomic fluidity: An integrative view of gene diversity within microbial populations,” *BMC Genomics*, vol. 12, no. 1, 2011.

[65] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, no. 3, p. 403–410, 1990.

[66] M. Kanehisa, “Kegg: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 28, no. 1, p. 27–30, 2000.

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

[68] Y. I. Wolf, K. S. Makarova, N. Yutin, and E. V. Koonin, “Updated clusters of orthologous genes for archaea: A complex ancestor of the archaea and the byways of horizontal gene transfer,” *Biology Direct*, vol. 7, no. 1, p. 46, 2012.

[69] I. Sela, Y. I. Wolf, and E. V. Koonin, “Assessment of assumptions underlying models of prokaryotic pangenome evolution,” *BMC Biology*, vol. 19, no. 1, 2021.

[70] C. E. McClung, “Systematics: systematics and the origin of species by ernst mayr.” *Science*, vol. 97, no. 2523, p. 424–425, 1943.

[71] C. Darwin and J. Murray, “On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life,” *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*, 1859.

[72] D. H. Huson and D. Bryant, “Application of phylogenetic networks in evolutionary studies,” *Molecular Biology and Evolution*, vol. 23, no. 2, p. 254–267, 2005.

[73] W. F. Doolittle, “Phylogenetic classification and the universal tree,” *Science*, vol. 284, no. 5423, p. 2124–2128, 1999.

[74] V. Kunin, L. Goldovsky, N. Darzentas, and C. A. Ouzounis, “The net of life: Reconstructing the microbial phylogenetic network,” *Genome Research*, vol. 15, no. 7, p. 954–959, 2005.

[75] K. H. Miga and T. Wang, “The need for a human pangenome reference sequence,” *Annual Review of Genomics and Human Genetics*, vol. 22, no. 1, p. 81–102, 2021.

[76] 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.

[77] C.-S. Chin, S. Behera, A. Khalak, F. J. Sedlazeck, P. H. Sudmant, J. Wagner, and J. M. Zook, “Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes,” *Nature Methods*, vol. 20, no. 8, p. 1213–1221, 2023.

[78] H. J. Abel, D. E. Larson, A. A. Regier, C. Chiang, I. Das, K. L. Kanchi, R. M. Layer, B. M. Neale, W. J. Salerno, C. Reeves, and et al., “Mapping and characterization of structural variation in 17,795 human genomes,” *Nature*, vol. 583, no. 7814, p. 83–89, 2020.

[79] G. Hickey, J. Monlong, J. Ebler, A. M. Novak, J. M. Eizenga, Y. Gao, H. J. Abel, L. L. Antonacci-Fulton, M. Asri, G. Baid, and et al., “Pangenome graph construction from genome alignments with minigraph-cactus,” *Nature Biotechnology*, 2023.

[80] K. J. Forsberg, S. Patel, M. K. Gibson, C. L. Lauber, R. Knight, N. Fierer, and G. Dantas, “Bacterial phylogeny structures soil resistomes across habitats,” *Nature*, vol. 509, no. 7502, p. 612–616, 2014.

[81] S. M. Soucy, J. Huang, and J. P. Gogarten, “Horizontal gene transfer: Building the web of life,” *Nature Reviews Genetics*, vol. 16, no. 8, p. 472–482, 2015.

[82] B. Paten, A. M. Novak, J. M. Eizenga, and E. Garrison, “Genome graphs and the evolution of genome inference,” *Genome Research*, vol. 27, no. 5, p. 665–676, 2017.

[83] 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.

[84] G. Rakocevic, V. Semenyuk, W.-P. Lee, J. Spencer, J. Browning, I. J. Johnson, V. Arsenijevic, J. Nadj, K. Ghose, M. C. Suciu, and et al., “Fast and accurate genomic analyses using genome graphs,” *Nature Genetics*, vol. 51, no. 2, p. 354–362, 2019.

[85] 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.

[86] G. Liti, D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. A. James, R. P. Davey, I. N. Roberts, A. Burt, V. Koufopanou, and et al., “Population genomics of domestic and wild yeasts,” *Nature*, vol. 458, no. 7236, p. 337–341, 2009.

[87] B. E. Dutilh, C. C. Thompson, A. C. Vicente, M. A. Marin, C. Lee, G. G. Silva, R. Schmieder, B. G. Andrade, L. Chimetto, D. Cuevas, and et al., “Comparative genomics of 274 vibrio cholerae genomes reveals mobile functions structuring three niche dimensions,” *BMC Genomics*, vol. 15, no. 1, 2014.

[88] D. Blankenberg, J. Taylor, and A. Nekrutenko, “Making whole genome multiple alignments usable for biologists,” *Bioinformatics*, vol. 27, no. 17, p. 2426–2428, 2011.

[89] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, “The sequence alignment/map format and samtools,” *Bioinformatics*, vol. 25, no. 16, p. 2078–2079, 2009.

[90] J. A. Baaijens, P. Bonizzoni, C. Boucher, G. Della Vedova, Y. Pirola, R. Rizzi, and J. Sirén, "Computational graph pangenomics: A tutorial on data structures and their applications," *Natural Computing*, vol. 21, no. 1, p. 81–108, 2022.

[91] A. Little, Y. Hu, Q. Sun, D. Jain, J. Broome, M.-H. Chen, F. Thibord, C. McHugh, P. Surendran, T. W. Blackwell, and et al., "Whole genome sequence analysis of platelet traits in the nhlbi trans-omics for precision medicine (topmed) initiative," *Human Molecular Genetics*, vol. 31, no. 3, p. 347–361, 2021.

[92] 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.

[93] T. Onodera, K. Sadakane, and T. Shibuya, "Detecting superbubbles in assembly graphs," *Lecture Notes in Computer Science*, p. 338–348, 2013.

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

[95] T. Seemann, "Prokka: Rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, p. 2068–2069, 2014.

[96] K. Lagesen, P. Hallin, E. A. Rødland, H.-H. Stærfeldt, T. Rognes, and D. W. Ussery, "Rnammer: Consistent and rapid annotation of ribosomal rna genes," *Nucleic Acids Research*, vol. 35, no. 9, p. 3100–3108, 2007.

[97] A. Bateman, "The pfam protein families database," *Nucleic Acids Research*, vol. 32, no. 90001, 2004.

[98] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, "Roary: Rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, no. 22, p. 3691–3693, 2015.

[99] Y. Peng, S. Tang, D. Wang, H. Zhong, H. Jia, X. Cai, Z. Zhang, M. Xiao, H. Yang, J. Wang, and et al., "Metapgn: A pipeline for construction and graphical visualization of annotated pangenome networks," *GigaScience*, 2018.

[100] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "Cd-hit: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, p. 3150–3152, 2012.

[101] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, p. 1658–1659, 2006.

[102] B. Buchfink, C. Xie, and D. H. Huson, “Fast and sensitive protein alignment using diamond,” *Nature Methods*, vol. 12, no. 1, p. 59–60, 2014.

[103] A. J. Enright, “An efficient algorithm for large-scale detection of protein families,” *Nucleic Acids Research*, vol. 30, no. 7, p. 1575–1584, 2002.

[104] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, “A genomic perspective on protein families,” *Science*, vol. 278, no. 5338, p. 631–637, 1997.

[105] G. Gautreau, A. Bazin, M. Gachet, R. Planel, L. Burlot, M. Dubois, A. Perrin, C. Médigue, A. Calteau, S. Cruveiller, and et al., “Ppanggolin: Depicting microbial diversity via a partitioned pangenome graph,” *PLOS Computational Biology*, vol. 16, no. 3, 2020.

[106] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: A software environment for integrated models of biomolecular interaction networks,” *Genome Research*, vol. 13, no. 11, p. 2498–2504, 2003.

[107] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with bwa-mem,” 2013.

[108] J. M. Hancock, “Blat (blast-like alignment tool),” *Dictionary of Bioinformatics and Computational Biology*, 2004.

[109] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller, “Human–mouse alignments with blastz,” *Genome Research*, vol. 13, no. 1, p. 103–107, 2002.

[110] B. Harris, “Lastz/lastz: Program for aligning dna sequences, a pairwise aligner.” [Online]. Available: https://github.com/lastz/lastz

[111] 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.

[112] A. C. Darling, B. Mau, F. R. Blattner, and N. T. Perna, “Mauve: Multiple alignment of conserved genomic sequence with rearrangements,” *Genome Research*, vol. 14, no. 7, p. 1394–1403, 2004.

[113] B. Kehr, K. Trappe, M. Holtgrewe, and K. Reinert, “Genome alignment with graph data structures: A comparison,” *BMC Bioinformatics*, vol. 15, no. 1, 2014.

[114] 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.

[115] H. Li, “Minimap2: Pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34, no. 18, p. 3094–3100, 2018.

[116] 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.

[117] S. Wang, Q. Xu, and B. Song, “Applying anchorwave to address plant genome alignment,” *BIO-PROTOCOL*, vol. 13, no. 19, 2023.

[118] H. Li, X. Feng, and C. Chu, “The design and construction of reference pangenome graphs with minigraph,” *Genome Biology*, vol. 21, no. 1, 2020.

[119] J. Armstrong, G. Hickey, M. Diekhans, I. T. Fiddes, A. M. Novak, A. Deran, Q. Fang, D. Xie, S. Feng, J. Stiller, and et al., “Progressive cactus is a multiple-genome aligner for the thousand-genome era,” *Nature*, vol. 587, no. 7833, p. 246–251, 2020.

[120] B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, and D. Haussler, “Cactus: Algorithms for genome multiple sequence alignment,” *Genome Research*, vol. 21, no. 9, p. 1512–1528, 2011.

[121] S. Marco-Sola, J. C. Moure, M. Moreto, and A. Espinosa, “Fast gap-affine pairwise alignment using the wavefront algorithm,” *Bioinformatics*, vol. 37, no. 4, p. 456–463, 2020.

[122] E. Garrison and A. Guarracino, “Unbiased pangenome graphs,” *Bioinformatics*, vol. 39, no. 1, 2022.

[123] E. Garrison, A. Guarracino, S. Heumos, A. Novak, G. Hickey, J. Eizenga, and P. Prins, “Pangenome/smoothxg: Citation release,” Sep 2023. [Online]. Available: https://zenodo.org/records/7239231

[124] E. Garrison, A. Guarracino, S. Heumos, F. Villani, Z. Bao, L. Tattini, J. Hagmann, S. Vorbrugg, S. Marco-Sola, C. Kubica, and et al., “Building pangenome graphs,” *Building pangenome graphs*, 2023.

[125] A. Guarracino, S. Buonaiuto, L. G. de Lima, T. Potapova, A. Rhie, S. Koren, B. Rubinstein, C. Fischer, H. J. Abel, L. L. Antonacci-Fulton, and et al., “Recombination between heterologous human acrocentric chromosomes,” *Nature*, vol. 617, no. 7960, p. 335–343, 2023.

[126] 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.

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

[128] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt, “Bandage: Interactive visualization of de novo genome assemblies,” *Bioinformatics*, vol. 31, no. 20, p. 3350–3352, 2015.

[129] E. G. Christian Fischer, “Chfi/gfaestus: Gfa visualizer, gpu-accelerated using vulkan.” [Online]. Available: https://github.com/chfi/gfaestus

[130] G. Gonnella, N. Niehus, and S. Kurtz, “gfaviz: Flexible and interactive visualization of gfa sequence graphs,” *Bioinformatics*, vol. 35, no. 16, p. 2853–2855, 2018.

[131] 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.

[132] “Burkholderia genome database,” accessed: 2023-12-1. [Online]. Available: https://www.burkholderia.com

[133] “Thesis extra files - dbfinaltable,” accessed: 2023-12-1. [Online]. Available: https://github.com/Silpex/ThesisBigFiles/blob/main/DB_FINAL_1.xlsx

[134] A. Löytynoja, “Phylogeny-aware alignment with prank,” *Methods in Molecular Biology*, p. 155–170, 2013.

[135] “Roary github repository,” accessed: 2023-12-1. [Online]. Available: https://github.com/sanger-pathogens/Roary

[136] “Panaroo github repository,” accessed: 2023-12-1. [Online]. Available: https://github.com/gtonkinhill/panaroo

[137] M. N. Price, P. S. Dehal, and A. P. Arkin, “Fasttree 2 – approximately maximum-likelihood trees for large alignments,” *PLoS ONE*, vol. 5, no. 3, 2010.

[138] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp, “Dendroscope: An interactive viewer for large phylogenetic trees,” *BMC Bioinformatics*, vol. 8, no. 1, 2007.

[139] G. Grothendieck, *sqldf: Manipulate R Data Frames Using SQL*, 2017, r package version 0.4-11. [Online]. Available: https://CRAN.R-project.org/package=sqldf

[140] “Pagoo recipes,” accessed: 2023-12-1. [Online]. Available: https://cran.r-project.org/web/packages/pagoo/vignettes/Recipes.html

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

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

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

[144] 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.

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

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

[147] 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.

[148] H. Chen and P. C. Boutros, “Venndiagram: A package for the generation of highly-customizable venn and euler diagrams in r,” *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[149] J. A. Rice, *Mathematical Statistics and data analysis*. Thomson/Brooks Cole, 2007.

[150] R. Mittelhammer, D. J. Miller, and G. G. Judge, *Page 73-74*. Cambridge University Press, 2000.

[151] D. Tenenbaum and B. P. Maintainer, *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*, 2023, r package version 1.42.0. [Online]. Available: https://bioconductor.org/packages/KEGGREST

[152] L. Katz, T. Griswold, S. Morrison, J. Caravas, S. Zhang, H. Bakker, X. Deng, and H. Carleton, “Mashtree: A rapid comparison of whole genome sequence files,” *Journal of Open Source Software*, vol. 4, no. 44, p. 1762, 2019.

[153] D. Winter, “pafr: Read, manipulate and visualize pairwise mapping format data,” accessed: 2023-12-1. [Online]. Available: https://dwinter.github.io/pafr/index.html

[154] 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, no. 2, 02 2021, giab008. [Online]. Available: https://doi.org/10.1093/gigascience/giab008

[155] K. B. I. t. Royal Botanic Gardens, “pypaftol: Python module for paftol.” [Online]. Available: https://github.com/RBGKew/pypaftol

[156] A. R. Quinlan and I. M. Hall, “Bedtools: A flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, p. 841–842, 2010.

[157] “Gfaestus github repository,” accessed: 2023-12-1. [Online]. Available: https://github.com/chfi/gfaestus

[158] “Thesis extra files - panaroo graphs,” accessed: 2023-12-3. [Online]. Available: https://github.com/Silpex/ThesisBigFiles/tree/main/PanarooGraphs

[159] “Thesis extra files - graph-based pangenome visualizations,” accessed: 2023-12-3. [Online]. Available: https://github.com/Silpex/ThesisBigFiles/tree/main/Graph-based%20Pangenome%20Visualizations

[160] W. R. Francis and G. Wörheide, “Similar ratios of introns to intergenic sequence across animal genomes,” *Genome Biology and Evolution*, vol. 9, no. 6, p. 1582–1598, 2017.

[161] A. Piovesan, F. Antonaros, L. Vitale, P. Strippoli, M. C. Pelleri, and M. Caracausi, “Human protein-coding genes and gene feature statistics in 2019,” *BMC Research Notes*, vol. 12, no. 1, 2019.

[162] Y. Zou, W. Xue, G. Luo, Z. Deng, P. Qin, R. Guo, H. Sun, Y. Xia, S. Liang, Y. Dai, and et al., “1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses,” *Nature Biotechnology*, vol. 37, no. 2, p. 179–185, 2019.

[163] “Population genetics and genomics in r,” accessed: 2023-12-2. [Online]. Available: https://grunwaldlab.github.io/Population_Genetics_in_R/

[164] M. Goel, H. Sun, W.-B. Jiao, and K. Schneeberger, “Syri: Finding genomic rearrangements and local sequence differences from whole-genome assemblies,” *Genome Biology*, vol. 20, no. 1, 2019.

[165] J. A. Sibbesen, J. M. Eizenga, A. M. Novak, J. Sirén, X. Chang, E. Garrison, and B. Paten, “Haplotype-aware pantranscriptome analyses using spliced pangenome graphs,” *Nature Methods*, vol. 20, no. 2, p. 239–247, 2023.

Appendix

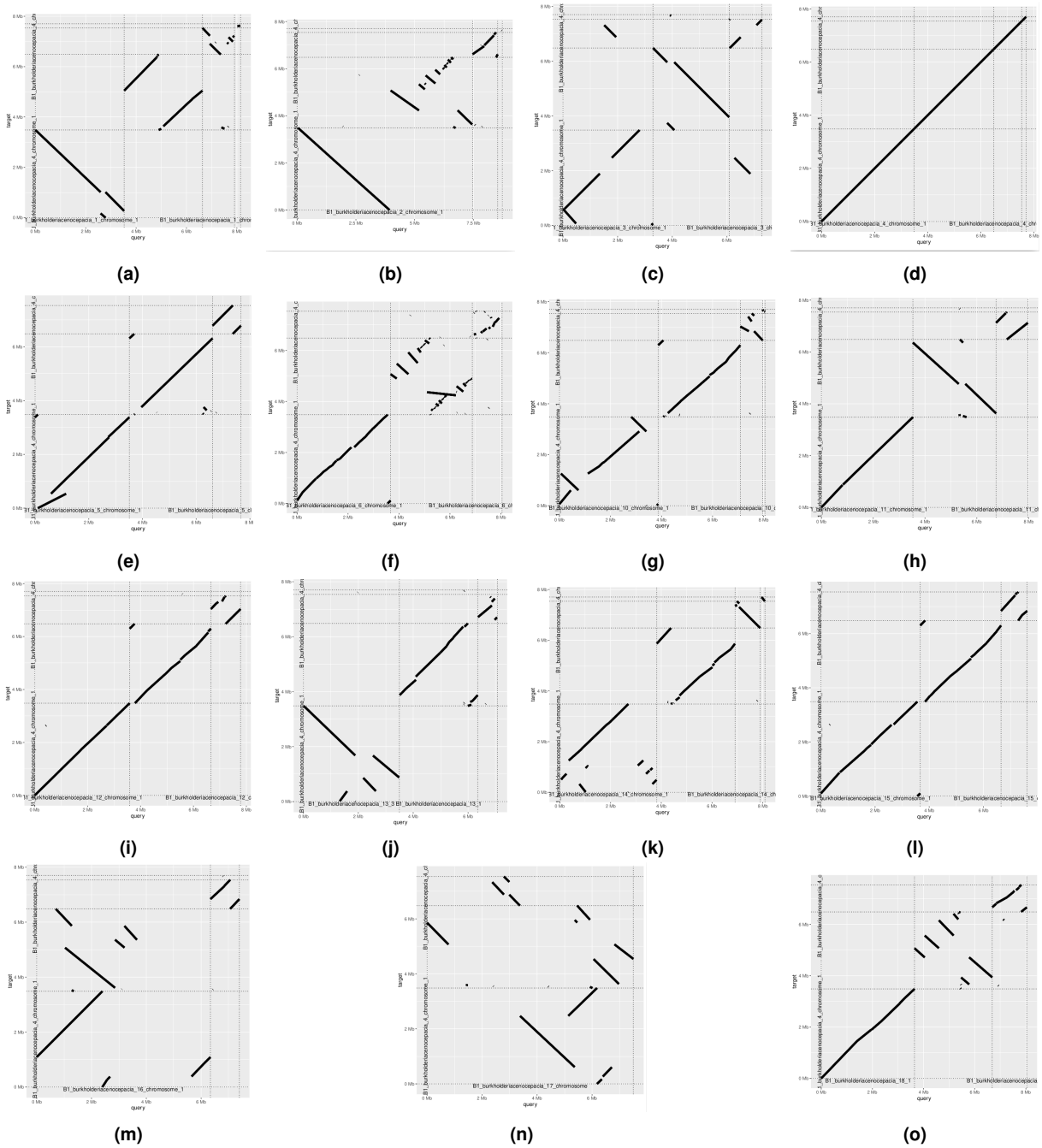


Figure A.1: PafR dot plots for the pairwise *B. cenocepacia* strain alignment of strain 4 versus all other strains, with the exception of strains 7,8 and 9, in order.

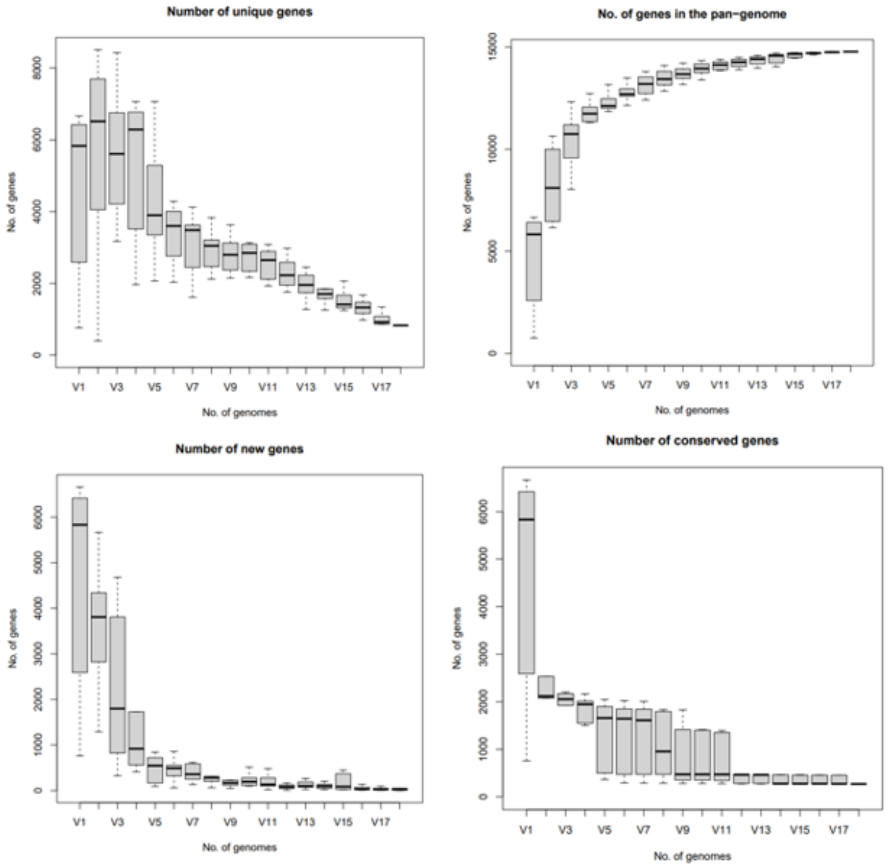


Figure A.2: Roary supplementary curve plots for the *B. cenocepacia* pangenome. The closed curve suggests a closed pangenome as per the power law equation.

Table A.1: Genomic fluidity averages between species groups.

Species Group	Average Genomic Fluidity
mallei Group	0.246
Bcc Group	0.272
All 8 Species	0.256

Table A.5: Top 10 *B. cenocepacia* clusters in the core with highest Tajima scores and their corresponding most common annotation.

Cluster	Tajima Score	Most Common Annotation
713	2,700453525	HAD-IB family hydrolase
16	2,570784361	MFS transporter
3990	2,564361837	AraC family transcriptional regulator
1692	2,456994001	mannitol dehydrogenase family protein
2460	2,427436542	nitronate monooxygenase
1775	2,393967508	sulfate adenylyltransferase subunit CysD
1921	2,375799802	quinone oxidoreductase
2011	2,359835064	histidine utilization repressor
737	2,338354815	glycolate oxidase subunit GlcE
614	2,286027423	UDP-glucose 4-epimerase GalE

Table A.6: Top 10 *B. cenocepacia* clusters in the core with lowest Tajima scores and their corresponding most common annotation.

Cluster	Tajima Score	Most Common Annotation
1236	-4,186038565	chromosomal replication initiator protein DnaA
1289	-4,087330718	adhesin
2129	-3,968750897	pseudouridine synthase
785	-3,847040612	autotransporter outer membrane beta-barrel domain-containing protein
3561	-3,749681196	flagellar hook-length control protein FliK
2	-3,694042471	long-chain fatty acid–CoA ligase
1529	-3,666143988	excinuclease ABC subunit A
2350	-3,652531875	RNA polymerase sigma factor RpoD
19	-3,610156185	PrkA family serine protein kinase
1275	-3,505673684	penicillin-binding protein 2

Table A.2: *B. contaminans*, *B. cepacia*, *B. gladioli* and *B. cenocepacia* - Number of genes classified as Core, Cloud, Shell for a wide array of E-value thresholds obtained by Pagoo.

B. contaminans (5)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	3473	3880	4152	4252	4356	4401	4426	4473	4488	**4532**	4525	4513	4487	4451
Shell genes	1134	1220	1247	1241	1266	1274	1272	1281	1303	1291	1302	1290	1288	1267
Cloud genes	908	1406	2216	2658	3029	3282	3502	3971	4564	5004	5443	5896	6447	7006
Total genes	5515	6506	7615	8151	8651	8957	9200	9725	10355	10827	11270	11699	12222	12724
B. cepacia (15)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	2354	2596	2738	2790	2828	2853	2871	2877	2883	**2890**	2880	2850	2850	2791
Shell genes	3175	3481	3610	3645	3733	3749	3758	3799	3802	3862	3884	3910	3910	3868
Cloud genes	2882	4611	7272	8575	9806	10544	11153	12569	14263	15473	16740	18101	18101	21293
Total genes	8411	10688	13620	15010	16367	17146	17782	19245	20948	22225	23504	24861	24861	27952
B. gladioli (4)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	3904	4340	4626	4742	4846	4894	4928	4985	5013	5055	5083	**5084**	5058	5022
Shell genes	206	214	190	185	177	181	181	176	168	169	171	171	168	171
Cloud genes	797	1182	1783	2116	2474	2670	2819	3170	3615	3908	4191	4541	4927	5340
Total genes	4907	5736	6599	7043	7497	7745	7928	8331	8796	9132	9445	9796	10153	10533
B. cenocepacia (18)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	2616	2892	3076	3132	3198	3220	3242	3266	3267	**3284**	3270	3243	3205	3187
Shell genes	3143	3373	3503	3559	3626	3634	3630	3657	3668	3653	3625	3618	3608	3591
Cloud genes	2998	5071	8019	9542	11099	12048	12751	14440	16541	18046	19616	21147	23001	24819
Total genes	8757	11336	14598	16233	17923	18902	19623	21363	23476	24983	26511	28008	29814	31597

Table A.3: *B. multivorans*, *B. mallei*, *B. pseudomallei* and *B. thailandensis* - Number of genes classified as Core, Cloud, Shell for a wide array of E-value thresholds obtained by Pagoo.

B. multivorans (14)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	2700	3023	3230	3293	3357	3420	3447	3454	3486	**3495**	3479	3456	3426
Shell genes	2124	2136	2051	2005	1997	1957	1936	1910	1880	1855	1863	1837	1814
Cloud genes	1612	3435	5830	7135	8214	9327	10585	12185	13162	14126	15151	16428	17669
Total genes	6436	8594	11111	12433	13568	14704	15968	17549	18528	19476	20493	21721	22909

B. mallei (9)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	2492	2756	2919	2975	3017	3074	3094	3089	**3108**	3100	3086	3058	3035
Shell genes	943	827	713	683	635	596	575	549	546	533	520	511	505
Cloud genes	2329	3473	4682	5209	6558	7149	7649	8321	8640	9122	9617	10257	10818
Total genes	5764	7056	8314	8867	10210	10819	11318	11959	12294	12755	13223	13826	14358

B. pseudomallei (16)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	3122	3466	3680	3750	3794	3865	3875	3877	**3906**	3905	3896	3877	3856
Shell genes	1868	1616	1361	1253	1169	1070	1013	950	928	909	877	835	802
Cloud genes	5483	7991	10899	12455	14044	15498	16828	18386	19251	20361	21537	22994	24251
Total genes	10473	13073	15940	17458	19007	20433	21716	23213	24085	25175	26310	27706	28909

B. thailandensis (9)	**E = 30**	**E = 40**	**E = 50**	**E = 55**	**E = 60**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E = 90**	**E = 95**	**E = 100**
Core genes	3013	3323	3522	3594	3654	3715	3740	3729	**3749**	3744	3726	3700	3675
Shell genes	782	725	644	609	589	568	559	539	512	495	501	498	486
Cloud genes	2165	3304	4676	5321	6025	6661	7259	8155	8754	9461	10110	10874	11633
Total genes	5960	7352	8842	9524	10268	10944	11558	12423	13015	13700	14337	15072	15794

Table A.4: Detection probability and mixing proportions for a array of K ranges for all species in a binomial estimation test.

K	Detection prob.	Mixing prop.	K	Detection prob.	Mixing prop.	K	Detection prob.	Mixing prop.	K	Detection prob.	Mixing prop.
B. thailandensis			*B. pseudomallei*			*B. cepacia*			*B. cenocepacia*		
3	0.00910	0.966	3	0.00617	0.980	3	0.0143	0.941	3	0.0175	0.933
3	0.868	0.0102	3	0.709	0.00290	3	0.714	0.0337	3	0.720	0.0301
3	1	0.0238	3	1	0.0171	3	1	0.0256	3	1	0.0367
4	0.00291	0.0872	4	0.00134	0.995	4	0.0119	0.948	4	0.00788	0.960
4	0.00937	0.880	4	0.261	0.000674	4	0.551	0.0142	4	0.331	0.0104
4	0.868	0.00985	4	0.909	0.000544	4	0.841	0.0167	4	0.889	0.0112
4	1	0.0230	4	1	0.00388	4	1	0.0209	4	1	0.0180
5	0.00344	0.578	5	0.00111	0.996	5	0.00355	0.623	5	0.00435	0.972
5	0.00399	0.407	5	0.201	0.000591	5	0.0159	0.342	5	0.135	0.00658
5	0.408	0.000995	5	0.614	0.000133	5	0.562	0.0100	5	0.445	0.00376
5	0.955	0.00769	5	0.961	0.000623	5	0.848	0.0109	5	0.898	0.00666
5	1	0.00628	5	1	0.00299	5	1	0.0139	5	1	0.0107
6	0.00344	0.568	6	0.00166	0.994	6	0.00515	0.968	6	0.00252	0.783
6	0.00399	0.417	6	0.305	0.000755	6	0.115	0.00693	6	0.0135	0.193
6	0.387	0.000968	6	0.879	0.000320	6	0.590	0.00617	6	0.337	0.00591
6	0.902	0.00192	6	0.996	0.00382	6	0.781	0.00605	6	0.858	0.00479
6	0.985	0.0121	6	1.00	0.00126	6	0.978	0.00726	6	0.991	0.0100
6	1	0.0000446	6	1	0.0000292	6	1	0.00562	6	1	0.00271
B. multivorans			*B. mallei*			*B. gladioli*			*B. contaminans*		
3	0.00851	0.964	3	0.00285	0.990	3	0.00609	0.973	3	0.000101	0.999
3	0.473	0.00951	3	0.806	0.00301	3	0.891	0.0116	3	0.504	0.000205
3	1	0.0268	3	1	0.00723	3	1	0.0152	3	1	0.000454
4	0.00412	0.980	4	0.00227	0.992	4	0.0147	0.466	4	0.0142	0.921
4	0.285	0.00491	4	0.631	0.000793	4	0.0185	0.467	4	0.197	0.000000884
4	0.786	0.00155	4	0.922	0.00254	4	0.950	0.0530	4	0.528	0.0232
4	1	0.0138	4	1	0.00491	4	1	0.0138	4	1	0.0556
5	0.00270	0.343	5	0.00132	0.448	5	0.0129	0.743	5	0.0140	0.513
5	0.00461	0.638	5	0.00278	0.544	5	0.0236	0.195	5	0.0178	0.403
5	0.286	0.00468	5	0.694	0.00113	5	0.847	0.00300	5	0.552	0.0237
5	0.787	0.00149	5	0.980	0.00511	5	0.966	0.0580	5	1.00	0.0600
5	1	0.0132	5	1	0.00141	5	1	0.000286	5	1	0.00000363

Table A.7: CDS annotations of the most common *B. cenocepacia* genes that are evolving neutrally, estimated by Tajima's D test. Only genes with count 15$>$ were considered.

CDS Annotation	Count
hypothetical protein	150
LysR family transcriptional regulator	118
type VI secretion system tip protein VgrG	91
pyridine nucleotide-disulfide oxidoreductase	61
ABC transporter ATP-binding protein	59
membrane protein	45
amidase	40
N-acetyl-gamma-glutamyl-phosphate reductase	36
peptide ABC transporter permease	33
NADP-dependent malic enzyme	32
citrate transporter	32
D-aminoacylase	30
sulfonate ABC transporter permease	28
esterase	26
type IV secretion protein Rhs	25
ABC transporter permease	24
TonB-dependent siderophore receptor	23
cytochrome d ubiquinol oxidase subunit II	20
transposase	18
ornithine carbamoyltransferase	18
cytochrome ubiquinol oxidase subunit I	18
D-alanyl-D-alanine dipeptidase	18
30S ribosomal protein S6	18
aspartate ammonia-lyase	17
phospholipase D family protein	17
carboxymethylenebutenolidase	17
DNA-binding response regulator	17
glutaminase	17
type VI secretion protein	17
potassium transporter	16
zinc-binding dehydrogenase	16
RNA methyltransferase	16
glyoxalase	16
TonB-dependent receptor	16
lysine transporter LysE	15
citrate synthase	15
glutathione ABC transporter permease GsiD	15

Figure A.3: *B. pseudomallei* Pagoo maximum-likelihood phylogenetic tree. we can observe the highest amount of lone strains (strains 1, 4, 10, 12 and 13), but we can identify 4 clusters: Cluster A with Strains 11 and 14; Cluster B With both strains 2 and strain 8; Cluster C with both strains 3 and strain 9; and Cluster D with Strains 5, 6 and 7. Visualization on Dendroscope.

Figure A.4: *B. mallei* Pagoo maximum-likelihood phylogenetic tree. we can observe 3 Clusters. Cluster A with Strains 1, 3, 4 and 6; Cluster B with strains 7 and 8; and Cluster C with Strains 2, 5 and 9, where 2 and 5 are almost identical. Visualization on Dendroscope.

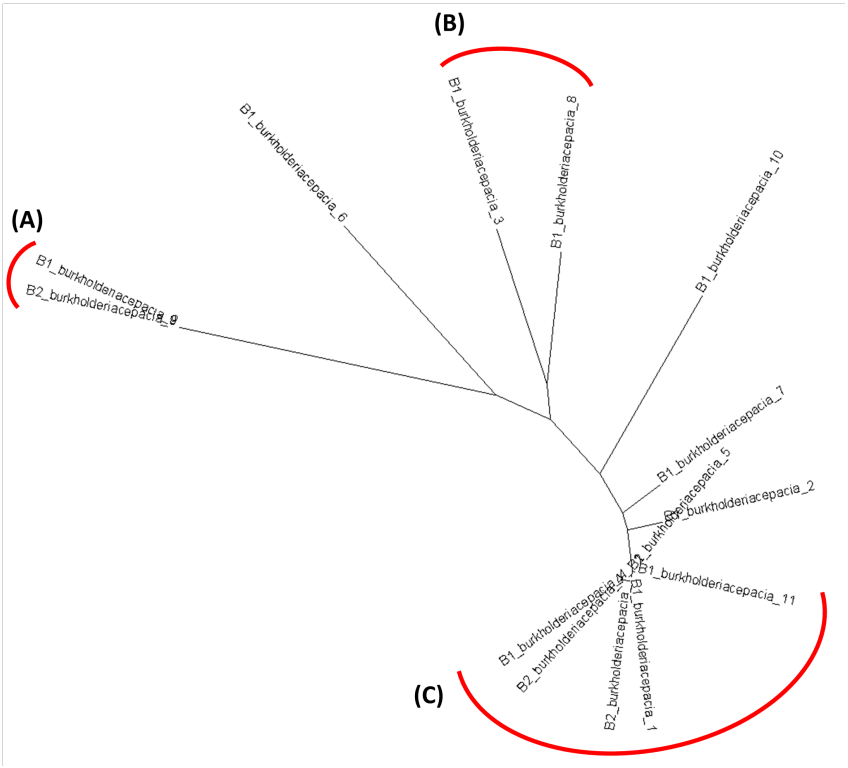


Figure A.5: *B. cepacia* Pagoo maximum-likelihood phylogenetic tree. We see a high amount of lone strains (Strain 6, 10, 7 and 2) but we do see 3 Clusters: Cluster A with Both strains 9; Cluster B with Strains 3 and 8; Cluster C with Both Strain 4, Both strain 1, Strain 5 and strain 11. Visualization on Dendroscope.

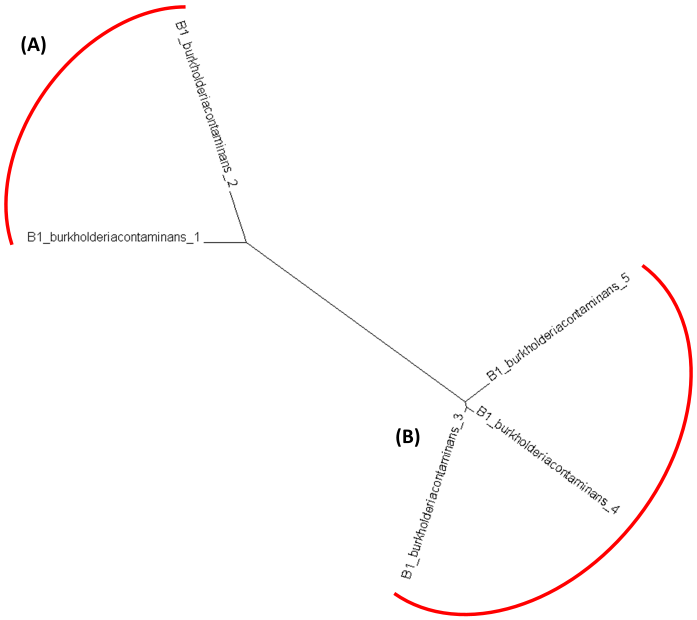


Figure A.6: *B. contaminans* Pagoo maximum-likelihood phylogenetic tree. We can observe 2 clusters: Cluster A with Strain 1 and 2; and Cluster B with strains 3, 4 and 5. Visualization on Dendroscope.

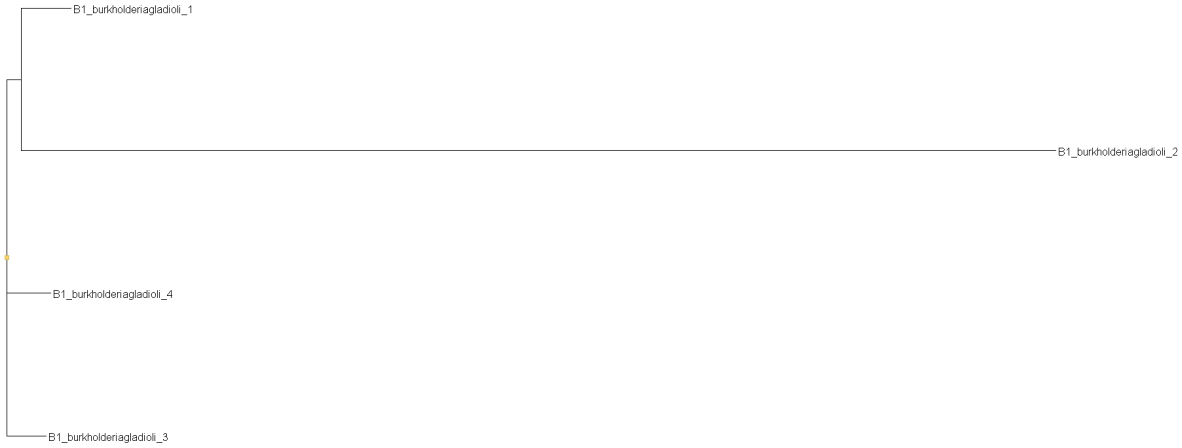


Figure A.7: *B. gladioli* Pagoo maximum-likelihood phylogenetic tree. We can observe a single cluster with strains 1 and 2. Visualization on Dendroscope.

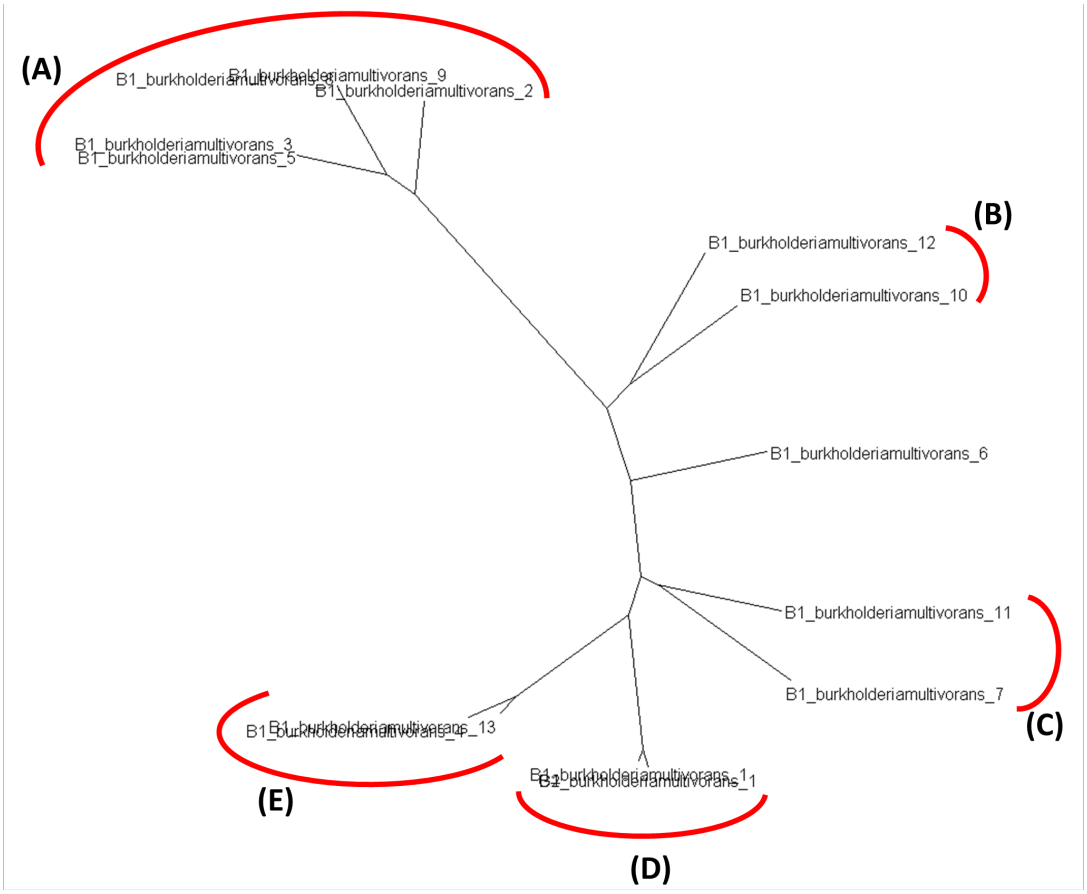


Figure A.8: *B. multivorans* Pagoo maximum-likelihood phylogenetic tree. We can observe 5 Clusters: Cluster A with strains 2, 3, 5, 8 and 9; Cluster B with Strains 10 and 12; Cluster C with strains 11 and 7; Cluster D with Both strains 1; and Cluster E with Strains 4 and 13. Visualization on Dendroscope.

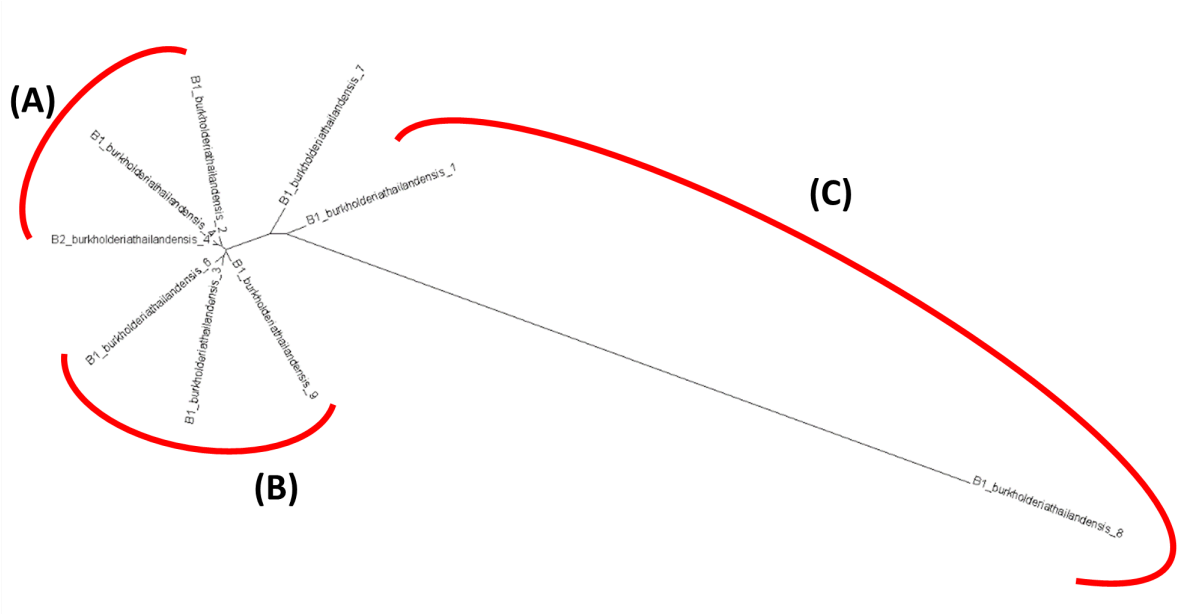


Figure A.9: *B. thailandensis* Pagoo maximum-likelihood phylogenetic tree. We can observe 3 Clusters: Cluster A with strains 2 and both 4s; Cluster B with strains 3, 6 and 9; And Cluster C with strains 1 and 8. Visualization on Dendroscope.

Table A.8: *B. cenocepacia* definition of populations by geographic location, medical vs environmental strain, and body part source.

Acronym	Continent	Medical VS Environmental	Body Part
B1_burkholderiacenocepacia_1	Asia	Medical	Respiratory
B1_burkholderiacenocepacia_2	Asia	Medical	Blood
B1_burkholderiacenocepacia_3	North America	Medical	Blood
B1_burkholderiacenocepacia_4	North America	Soil	Non body part
B1_burkholderiacenocepacia_5	North America	Vegetable	Non body part
B1_burkholderiacenocepacia_6	Oceania	Medical	Respiratory
B1_burkholderiacenocepacia_7	Oceania	Soil	Non body part
B1_burkholderiacenocepacia_8	North America	Soil	Non body part
B1_burkholderiacenocepacia_9	Europe	Medical	Respiratory
B1_burkholderiacenocepacia_10	Europe	Medical	Respiratory
B1_burkholderiacenocepacia_11	North America	Vegetable	Non body part
B1_burkholderiacenocepacia_12	Oceania	Water	Non body part
B1_burkholderiacenocepacia_13	North America	Medical	Respiratory
B1_burkholderiacenocepacia_14	Europe	Medical	Respiratory
B1_burkholderiacenocepacia_15	North America	Medical	Respiratory
B1_burkholderiacenocepacia_16	North America	Medical	Respiratory
B1_burkholderiacenocepacia_17	North America	Medical	Respiratory
B1_burkholderiacenocepacia_18	Asia	Vegetable	Non body part

Table A.9: Core Genes Unique to *B. cenocepacia* vs Core Genes Unique to *B. cepacia*.

Core Genes Unique to ***B. cenocepacia***	Count	Core Genes Unique to ***B. cepacia***	Count
hypothetical protein	2916	hypothetical protein	318
LysR family transcriptional regulator	736	biopolymer transporter ExbD	106
AraC family transcriptional regulator	348	integrase	68
membrane protein	312	energy transducer TonB	56
MFS transporter	304	ABC transporter permease	54
TetR family transcriptional regulator	158	IclR family transcriptional regulator	54
XRE family transcriptional regulator	154	amine dehydrogenase	54
LacI family transcriptional regulator	142	cytochrome c	48
sugar ABC transporter permease	120	MotA/TolQ/ExbB proton channel family protein	46
MarR family transcriptional regulator	116	FAD-dependent oxidoreductase	36
glutathione S-transferase	116	DUF1275 domain-containing protein	36
integrase	112	type VI secretion system baseplate subunit TssK	36
short-chain dehydrogenase	106	DUF159 family protein	32
oxidoreductase	102	FAD-binding oxidoreductase	32
GntR family transcriptional regulator	102	TetR family transcriptional regulator	30

Table A.10: Core genes unique to *B. pseudomallei* vs core genes unique to *B. mallei*.

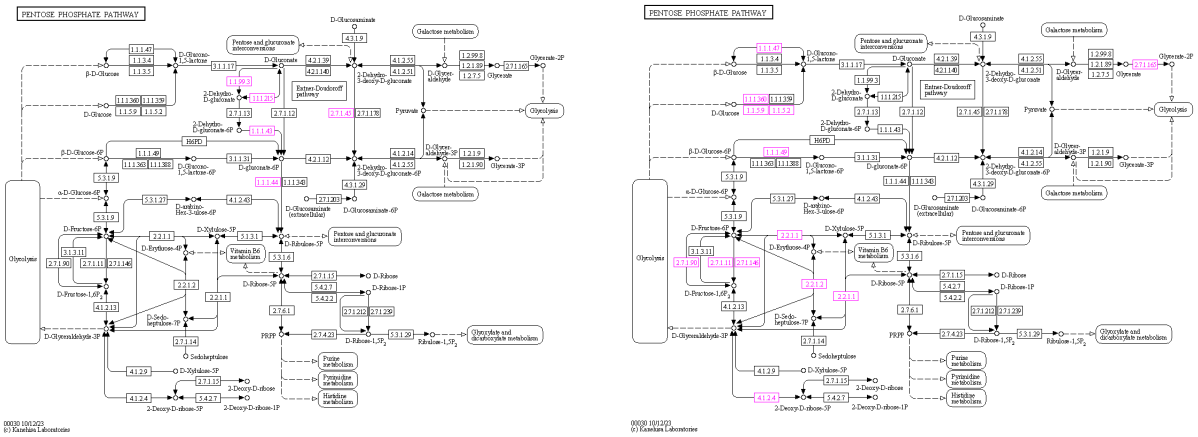
Core Genes Unique to ***B. pseudomallei***	Count	Core Genes Unique to ***B. mallei***	Count
hypothetical protein	3852	hypothetical protein	173
membrane protein	502	integrase	16
AraC family transcriptional regulator	161	porin	9
LysR family transcriptional regulator	138	MarR family transcriptional regulator	9
fimbrial protein	130	AraC family transcriptional regulator	9
lipoprotein	116	uroporphyrin-III C-methyltransferase	9
transposase	106	glycyl-tRNA synthetase subunit alpha	9
ABC transporter permease	103	50S ribosomal protein L21	9
alpha/beta hydrolase	99	glutathione S-transferase	8
LuxR family transcriptional regulator	91	ATP synthase I	7
TetR family transcriptional regulator	86	ABC transporter permease	6
glycosyl transferase	71	branched-chain amino acid ABC transporter permease	6
MFS transporter	70	mechanosensitive ion channel protein MscS	6
methyltransferase	69	chorismate lyase	5
GntR family transcriptional regulator	68	aminoglycoside 6$-acetyltransferase	5

Table A.11: KEGG pathways of the core genes unique to *B. cenocepacia* vs core genes unique to *B. cepacia*.

Pathways w/unique ***B. cenocepacia*** core genes	Count	Core Pathways w/unique ***B. cepacia*** core genes	Count
Non metabolic	663	Non metabolic	212
Metabolic pathways	127	Metabolic pathways	34
Non allowed character, can't be tested	69	Non allowed character, can't be tested	30
Biosynthesis of secondary metabolites	67	Biosynthesis of secondary metabolites	18
Microbial metabolism in diverse environments	47	Fructose and mannose metabolism	6
Butanoate metabolism	16	Glycerophospholipid metabolism	6
Drug metabolism - other enzymes	12	Microbial metabolism in diverse environments	6
Drug metabolism - cytochrome P450	10	Pentose and glucuronate interconversions	6
Glutathione metabolism	10	Amino sugar and nucleotide sugar metabolism	4
Metabolism of xenobiotics by cytochrome P450	10	**Butanoate metabolism**	4
Glyoxylate and dicarboxylate metabolism	8	Nitrogen metabolism	4
Riboflavin metabolism	8	O-Antigen nucleotide sugar biosynthesis	4
Sulfur metabolism	8	**Tryptophan metabolism**	4
Tryptophan metabolism	8	alpha-Linolenic acid metabolism	2
Aminobenzoate degradation	6	Aminobenzoate degradation	2

Table A.12: KEGG Pathways of the core genes unique to *B. pseudomallei* vs core genes unique to *B. mallei*.

Pathways w/unique ***B. pseudomallei*** core genes	Count	Pathways w/unique ***B. mallei*** core genes	Count
Non metabolic	373	Non metabolic	53
Metabolic pathways	233	Metabolic pathways	9
Biosynthesis of secondary metabolites	150	Biosynthesis of secondary metabolites	8
Pentose phosphate pathway	83	Microbial metabolism in diverse environments	4
Microbial metabolism in diverse environments	38	Non allowed character, can't be tested	3
Non allowed character, can't be tested	24	Carbon fixation pathways in prokaryotes	3
Glyoxylate and dicarboxylate metabolism	21	Citrate cycle (TCA cycle)	2
Starch and sucrose metabolism	20	alpha-Linolenic acid metabolism	2
Pyruvate metabolism	12	Caprolactam degradation	1
Penicillin and cephalosporin biosynthesis	8	Riboflavin metabolism	1
Valine, leucine and isoleucine degradation	8	Phenylalanine, tyrosine and tryptophan biosynthesis	1
Biosynthesis of unsaturated fatty acids	7	Benzoate degradation	1
Nitrogen metabolism	7	Aminobenzoate degradation	1
Purine metabolism	7	Drug metabolism - cytochrome P450	1
Pyrimidine metabolism	7	Phenazine biosynthesis	1



(a) Bcc **(b)** Mallei group

Figure A.10: KEGG - Pentose Phosphate Pathway. Bcc(a) mallei group(b)

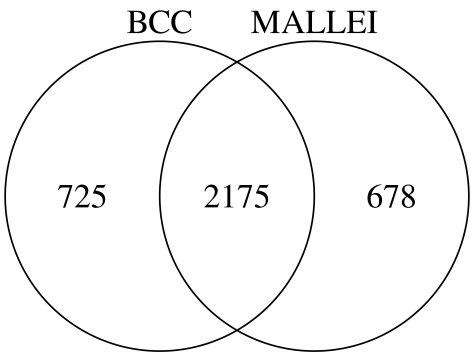
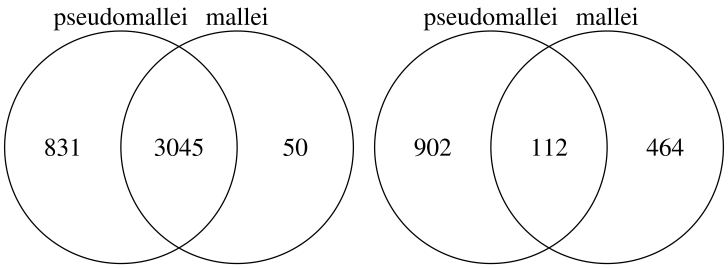
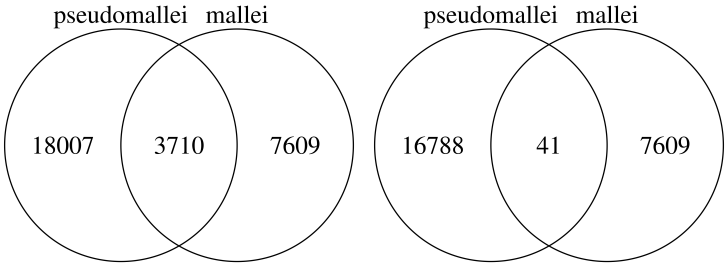


Figure A.11: Venn diagram highlighting the core genome of the Bcc vs mallei group.



(a) Core **(b)** Shell



(c) Pangenome **(d)** Cloud

Figure A.12: Venn diagrams highlighting the differences between the *B. pseudomallei* and *B. mallei* pangenomes.

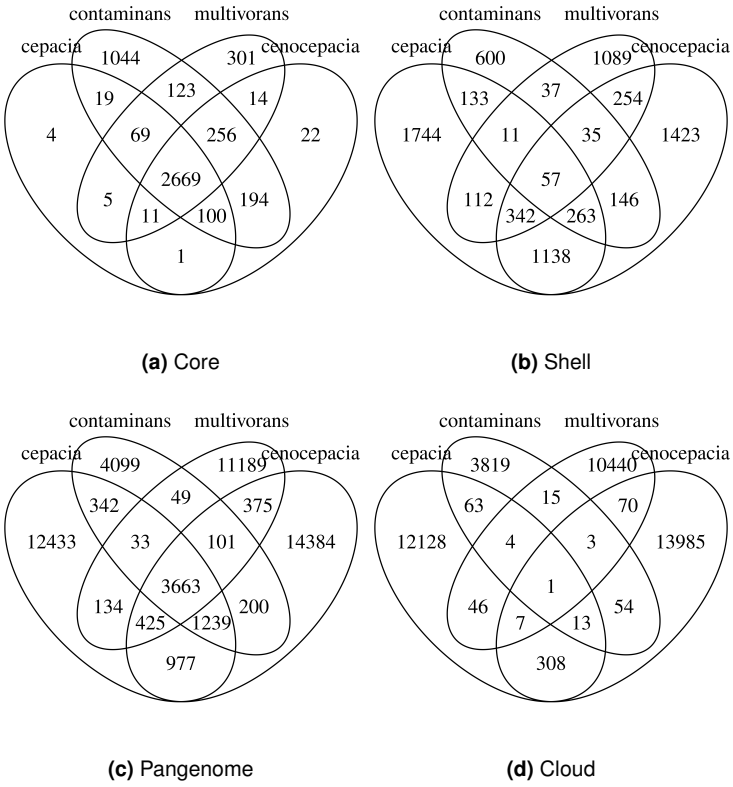


Figure A.13: Venn diagrams highlighting the differences inside the Bcc pangenomes.

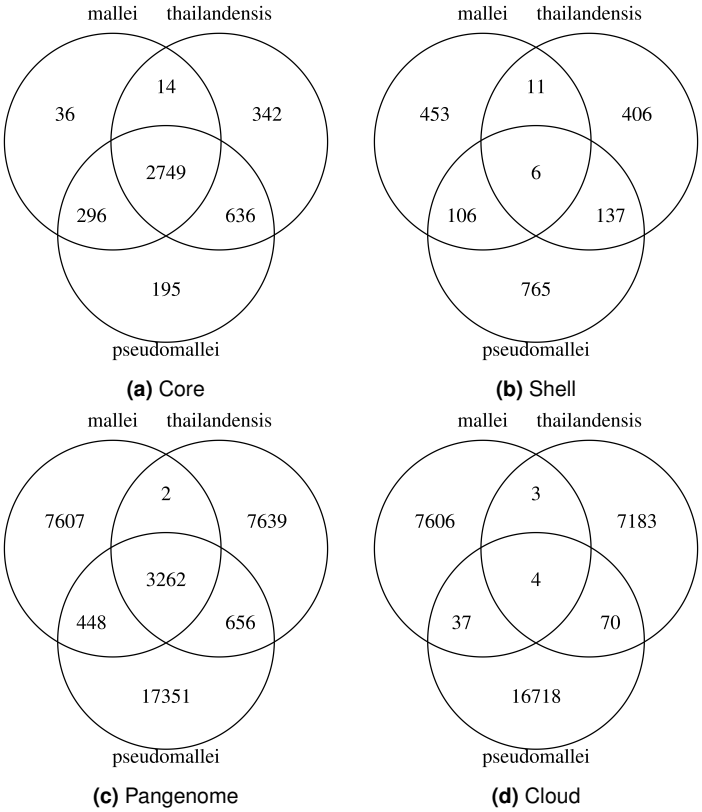


Figure A.14: Venn diagrams highlighting the differences inside the mallei group pangenomes.

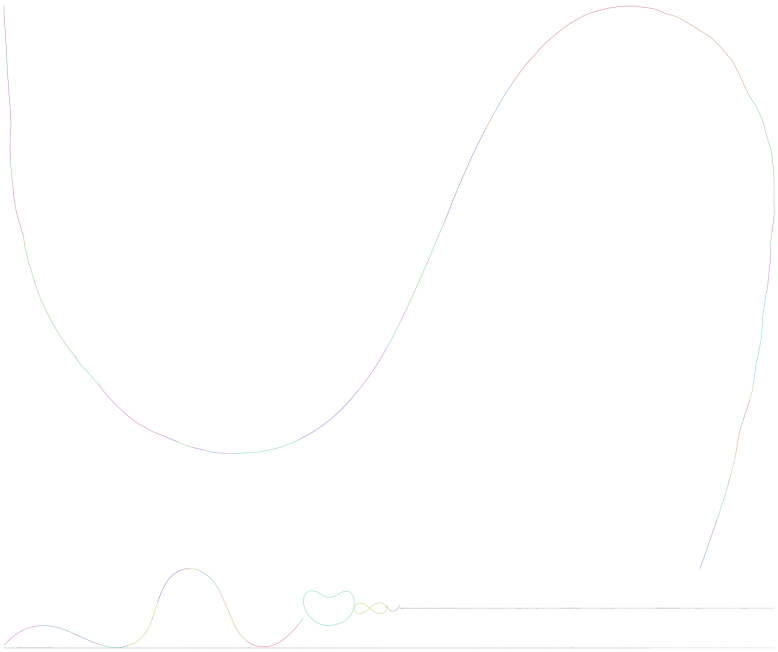


Figure A.15: *B. contaminans* bubble graph visualization on Bandage.

Table A.13: Multi-species pangenomes- number of genes classified as core, cloud, shell for a wide array of E-value thresholds.

Mallei Group	**E = 50**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E= 90**
Core	/	/	2829	2840	**2852**	2840	2844	2838	/
Shell	/	/	2948	2939	2874	2803	2784	2739	/
Cloud	/	/	27781	28780	31238	34362	36167	38481	/
Total	/	/	33558	34559	36964	40005	41795	44058	/
Bcc	**E = 50**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E= 90**
Core	/	/	/	2885	2899	2907	**2913**	2912	2873
Shell	/	/	/	7792	7838	7805	7792	7758	7860
Cloud	/	/	/	34097	38905	44924	49080	53371	57737
Total	/	/	/	44774	49642	55636	59785	64041	68470
Mallei Group + Gladioli	**E = 50**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E= 90**
Core	2135	2181	**2192**	2182	2183	2174	2165	2130	/
Shell	5982	5981	5962	6018	6003	5957	6001	6054	/
Cloud	21254	28340	30233	31379	34188	37755	39863	42460	/
Total	29371	36502	38387	39579	42374	45886	48029	50644	/
All Species	**E = 50**	**E = 60**	**E = 63**	**E = 65**	**E = 70**	**E = 75**	**E = 80**	**E = 85**	**E= 90**
Core	/	2229	2238	2242	**2243**	2234	2225	/	/
Shell	/	12467	12456	12480	12530	12446	12495	/	/
Cloud	/	56925	61465	64566	72181	81782	88050	/	/
Total	/	71621	76159	79288	86954	96462	102770	/	/

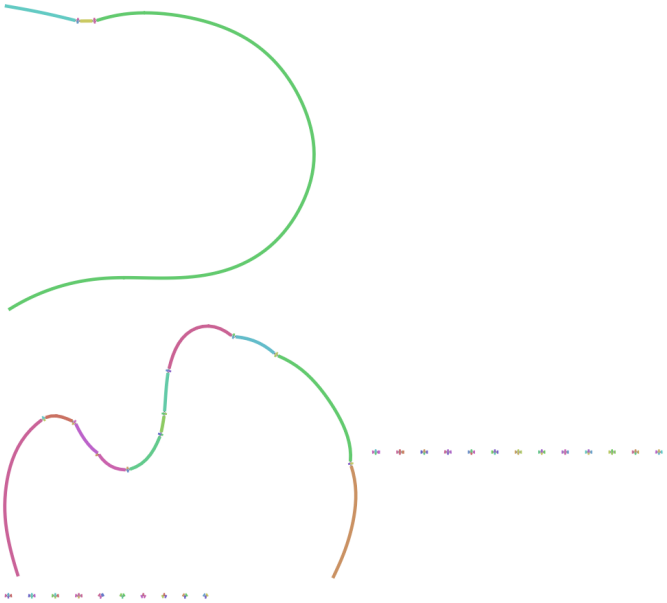


Figure A.16: *B. mallei* bubble graph visualization on Bandage.

Figure A.17: *B. multivorans* bubble graph visualization on Bandage.

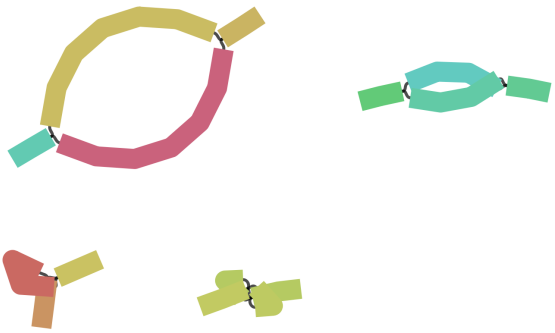


Figure A.18: *B. pseudomallei* bubble graph visualization on Bandage.

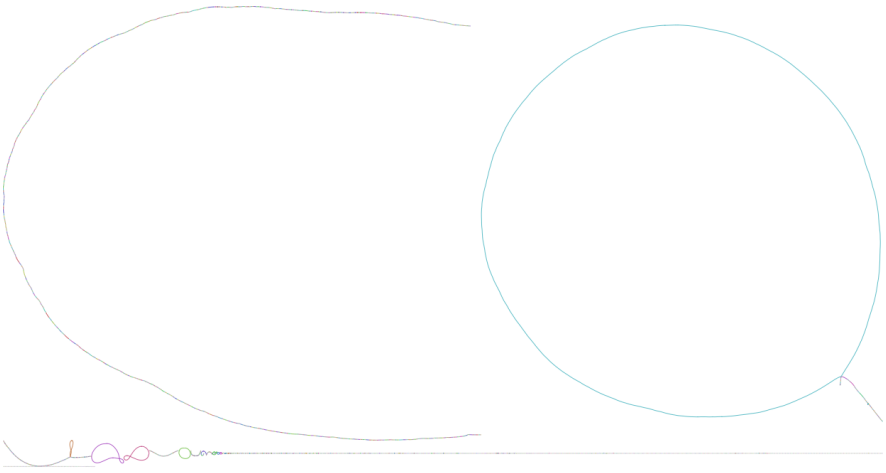


Figure A.19: *B. thailandensis* bubble graph visualization on Bandage.