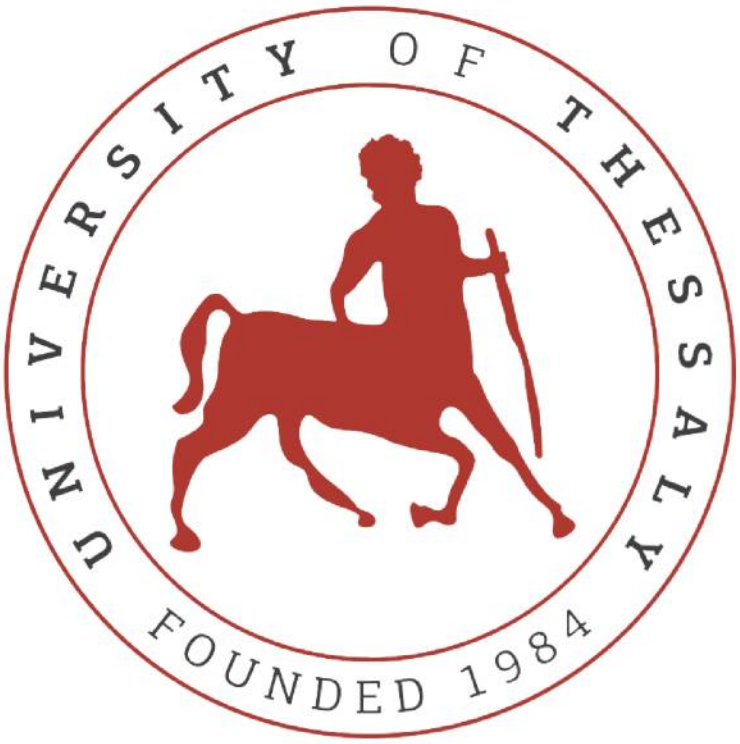
Department of Biochemistry and Biotechnology

University of Thessaly



Comparative and evolutionary genomic analysis of the *Prevotella* genus with Bioinformatics methods

Συγκριτική και εξελικτική γονιδιωματική ανάλυση του γένους *Prevotella* με μεθόδους Βιοπληροφορικής

Πετσούλας Χρήστος

του Αθανασίου

This study was performed in the Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, Faculty of Health Sciences, University of Thessaly.

Three-member Advisory Committee:

Supervisor: Amoutzias Grigorios, Professor of Bioinformatics with emphasis in Microbiology, Department of Biochemistry and Biotechnology, School of Health Sciences, University of Thessaly, Larisa, Greece

Dahle Håkon, Professor, Group Leader at Computatonal Biological UnitX, Department of Biological Sciences & Computational Biology Unit, University of Bergen, Bergen, Norway

Joshi Anagha, Professor of Bioinformatics, Department of Clinical Sciences, Computational Biology Unit, University of Bergen, Bergen, Norway

Acknowledgements

I would like to thank my supervisor, Professor Grigorios Amoutzias for the guidance and supervision of this study, as well as for the trust he showed in me. Next, I would like to warmly thank the other members of the three-member advisory committee, Professor Håkon Dahle and Professor Anagha Joshi who honored me with their presence on the committee and with their help during the year. At this point, I would like to especially thank the former PhD student of the bioinformatics laboratory Marios Nikolaidis, who provided me with constant help and support that I needed, both as a colleague and more so as a friend.

Finally, I would like to sincerely thank all my relatives and especially my family, who were constantly by my side from the first steps of my academic life until the end of my study.

Abstract 5
Περίληψη 5
1 Introduction 6
1.1 Pan-Genome, Core genomes and Fingerprint Genomes 6
1.1.1 Pan-genome 6
1.1.2 Core genomes 8
1.1.3 Fingerprint genes 10
1.2 Bacterial species demarcation 10
1.3 The Prevotellaceae family 14
2 Materials and methods 17
2.1 Software 17
2.1.1 Linux Ubuntu 24.04 LTS 17
2.1.2 PERL programing language 17
2.1.3 Visual Studio Code (VS Code) v.1.89.1 17
2.2 NCBI 18
2.3 pyPGCF, a comparative bacterial genomics python software 18
2.4 Multiple Sequence alignment with MUSCLE 19
2.5 Gblocks 19
2.6 IQ-Tree2 20
2.7 Interactive Tree of Life (iTOL). 20
2.8 FastANI 21
2.9 dbCAN2 identification of Carbohydrate-Active enZYmes (CAZymes) 21
3 Results – Discussion 23
4 Bibliography 40

Abstract

Prevotellaceae are spherical Gram-negative, anaerobic bacteria that colonize various organs in both humans and animals. Despite being key members of the gut, oral and vaginal microbiome, some species within this family have been characterized as pathogenic. By utilizing 1,180 publicly available genomes, we performed large-scale evolutionary and phylogenomic analyses of the *Prevotellaceae* family and the *Prevotella* genus. We identified i) miss-classified species, ii) we analyzed the host distribution, iii) the genomic characteristics and core proteome of the entire family, iv) the core and fingerprint proteins of the various species and v) the amount of carbohydrate degrading enzymes. The fingerprints are important for understanding the lineage-specific adaptations.

Περίληψη

Η οικογένεια *Prevotellaceae* είναι σφαιρικά Gram-αρνητικά, αναερόβια βακτήρια που αποικίζουν διάφορα όργανα τόσο σε ανθρώπους όσο και σε ζώα. Παρά το γεγονός ότι αποτελούν βασικό μέλος του μικροβιώματος του εντέρου, της στοματικής κοιλότητας και του κόλπου, ορισμένα είδη αυτής της οικογένειας έχουν χαρακτηριστεί ως παθογόνα. Χρησιμοποιώντας 1.180 διαθέσιμα γονιδιώματα, πραγματοποιήσαμε ευρείας κλίμακας εξελικτικές και φυλογενωμικές αναλύσεις της οικογένειας *Prevotellaceae* και του γένους *Prevotella*. Εντοπίσαμε i) εσφαλμένα χαρακτηρισμένα είδη, ii) αναλύσαμε τη κατανομή στους ξενιστές, iii) τα γονιδιωματικά χαρακτηριστικά και το πρωτέωμα πυρήνα ολόκληρης της οικογένειας, iv) τις βασικές και τις χαρακτηριστικές πρωτεΐνες των διάφορων ειδών και v) τον αριθμό των ενζύμων που αποδομούν τους υδατάνθρακες. Οι χαρακτηριστικές πρωτεΐνες είναι σημαντικές για την κατανόηση των προσαρμογών που αφορούν συγκεκριμένες γενεαλογίες.

1 Introduction

1.1 Pan-Genome, Core genomes and Fingerprint Genomes

1.1.1 Pan-genome

The bacterial pan-genome concept revolutionizes our understanding of microbial diversity and evolution by transcending the traditional view of a fixed genome for a species. It encompasses the core genome and the accessory genomes (Tettelin et al., 2008). The core genome consists of the genes that are shared among all strains of a species or an evolutionary group, while the accessory genes are unique to certain strains or sub-groups. This dynamic genomic landscape reflects the adaptability and plasticity of bacteria in response to diverse environmental pressures and ecological niches. By characterizing the pan-genome, researchers gain insights into the genetic reservoir available to bacteria, shaping their ability to colonize new habitats, interact with hosts, and evolve resistance to antibiotics and other stressors (Medini et al., 2005).

Understanding the bacterial pan-genome has noteworthy implications in various fields, including microbiology, pharmaceuticals, and biotechnology. In clinical settings, understanding the pan-genome helps in following the development and spread of antibiotic resistance among bacterial populations, and has implications for infection control and antibiotic stewardship. Moreover, knowledge of the pan-genome contributes to the development of novel helpful approaches, such as phage treatment and probiotics, which use bacterial differing qualities to target particular pathogens or modulate host-microbe interactions (Ventola, 2015).

Bacterial pan-genome research continues to evolve with advances in sequencing technologies, bioinformatics and computational tools. High-throughput sequencing platforms empower comprehensive genomic studies of bacterial populations, for the identification of core and accessory genes across assorted strains and species (Page et al., 2015). As researchers analyse the vast genomic data from microorganisms (the genomic pool), they identify processes by which genetic

material is shared or transferred between different species and strains. This exchange of genes, termed horizontal gene transfer, influences the diversity of microbes, their ecosystems, and their evolutionary paths. Understanding these relationships allows for the development of new strategies and innovations in fields like healthcare (e.g., targeted antibiotic treatments), agriculture (e.g., crop resistance), and biotechnology (e.g., bioengineering tools) (McInerney et al., 2017). Also, by exploring the pan-genome, researchers can identify enzymes, biosynthetic pathways, and other biocatalysts that are useful for producing biofuels, aiding in bioremediation (cleaning up pollutants), and synthesizing pharmaceuticals. Harnessing this genetic diversity allows for the creation of robust microbial systems that can support sustainable production and environmental solutions (Zhang et al., 2018).

In conclusion, the bacterial pan-genome expands our knowledge of microbial diversity, evolution, and adaptability. Scientists can learn more about the genetics underlying bacterial behavior, their function in ecosystems, their capacity to cause illness, and their potential use in biotechnology by examining the pan-genome. In a hypothetical diagram Venn in **Figure 1**, we can observe an example of the pan-genome, where the pan-genome is outlined within the green box.

The pan-genome is further classified as either open or closed. An open pan-genome keeps increasing with the addition of newly sequenced strains, while the addition of new genomes has a negligible effect in closed pan-genomes

Bacterial species that live in varied or dynamic habitats, where horizontal gene transfer is common, are likely to have an open pan-genome. With accessory genes that provide adaptation in domains like antibiotic resistance, virulence, or environmental niche specialization, this enables the gene pool to continue to grow. This concept is best shown by bacteria with the ability to quickly adapt and develop, such as *Streptococcus pneumoniae* and *Escherichia coli* (Medini et al., 2005). These species benefit greatly from an open pan-genome in dynamic settings, which increases genetic variety and promotes ecological success.

In contrast, species with a closed pan-genome exhibit a relatively fixed and stable gene pool, with limited gene acquisition and a smaller accessory genome.

These species typically reside in more stable, niche environments, such as obligate intracellular habitats, where new genetic material is less accessible. As a result, their pan-genomes are largely composed of core genes that are highly conserved across strains. Examples include *Mycobacterium tuberculosis* and *Bacillus anthracis*, where adaptation occurs primarily through small mutations rather than gene acquisition (Snipen & Ussery, 2010). This closed model reflects evolutionary stability but limits rapid adaptation to changing conditions.

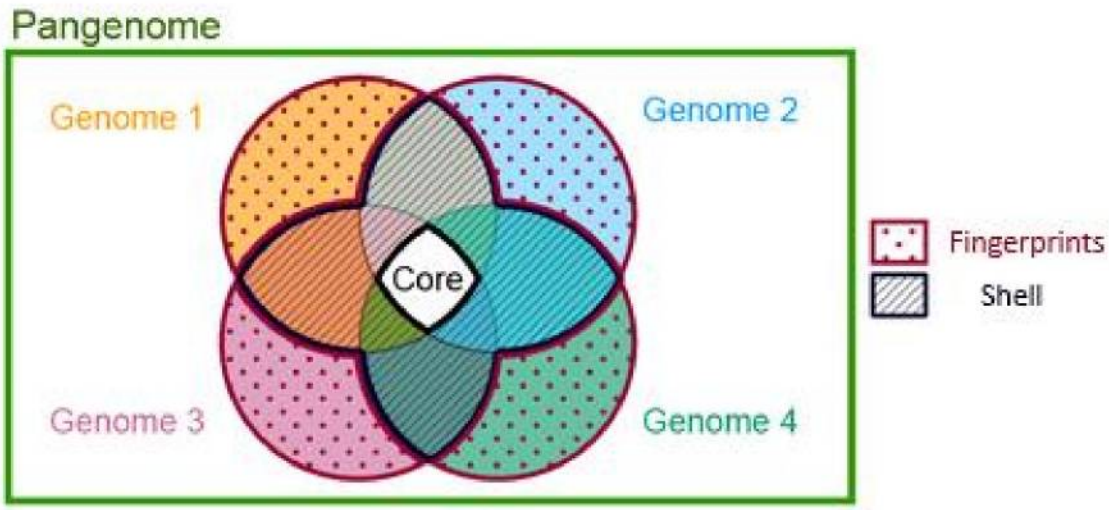


Figure 1: Hypothetical example of pan-genome visualization as a Venn diagram using four organisms (circles). In the middle, in white, is the core genome of these organisms. The unique genes (Fingerprints) of each representative are marked with dots. With the lines, genes shared by most genomes are highlighted (Shell) Picture downloaded from https://en.m.wikipedia.org/wiki/File:Parts_of_the_pangenome.png.

1.1.2 Core genomes

The bacterial core genome embodies the conserved genomic components that are shared by all members of a species, which form the foundation of the species' genetic identity. These shared genes are crucial for basic survival and defining

characteristics, and they remain conserved across different strains (Gordienko et al., 2013).

The core genome represents the genome passed down from ancestral species to subsequent ones, reflecting both the evolutionary history and ecological niche of a bacterial species. This collection of essential genes is shared among all strains of the species, ensuring survival and continuity. As such, the core genome serves as a genetic record of how the species has adapted to its environment over time, offering insights into its evolutionary trajectory (Vernikos et al., 2015). Core genes encode key metabolic pathways, regulatory networks, and auxiliary components vital for bacterial growth, survival, and adaptation to assorted environments (Medini et al., 2005). Through comparative genomics, analysts can distinguish the core genome from accessory genomic components, and also identify species-specific characteristics, evolutionary relationships, and ecological interactions inside microbial communities (Zhang et al., 2018).

Deciphering the variety, virulence, and evolution of microorganisms requires an understanding of the bacterial pangenome. Antibiotic resistance, pathogenicity-causing proteins, and other characteristics necessary for bacterial growth and survival are frequently found in core genes (Méric et al., 2015). Researchers can learn more about the genetic foundation of bacterial diseases using core genome analysis, which informs treatment, prevention, and diagnosis approaches (Croucher et al., 2017). Furthermore, evolutionary processes such as genetic drift, natural selection, and horizontal gene transfer that drive bacterial diversity may be shown by comparing core genomes (Page et al., 2015). The population structure and evolutionary paths of bacterial species are revealed by insights into the core genome, which aids in microbial surveillance, epidemiology, and public health initiatives (McInerney et al., 2017). Advances in high-throughput sequencing technologies, genomics and bioinformatics have enabled the study of core-genomes. Through high-throughput sequencing platforms it is possible to quickly gather large amounts of bacterial genome data (Page et al., 2015). The concept of bacterial core genomes holds promise for various applications in healthcare, agribusiness, and biotechnology. The

set of conserved genes may serve as an important target for diagnostic measures, antibody development, and antimicrobial compound discovery, thereby aiding in the control and management of infections (Méric et al., 2015

1.1.3 Fingerprint genes

The concept of species fingerprint genes encapsulates the unique genetic signatures that distinguish one bacterial species from all the others, akin to a molecular fingerprint. These genes represent a subset of genes and genetic elements conserved across all or almost all strains of a certain species, while being absent from all other species that are included in the analysis, thus providing a genetic identity unique to that species (Gordienko et al., 2013). These fingerprints consist of genes that are essential for species-specific traits, including metabolic pathways, virulence factors, and regulatory networks (Vernikos et al., 2015). By characterizing these genetic fingerprints, researchers gain insights into the genetic basis of species-specific phenotypes, ecological adaptations, and pathogenicity mechanisms.

1.2 Bacterial species demarcation

The demarcation of bacterial species, historically was performed with the 16S ribosomal RNA (rRNA) sequence. This genetic component, about 1,500 base pairs long, contains both conserved and hypervariable regions, making it a great condidate for elucidating phylogenetic relationships among microbes (**Figure 2**) (Johnson et al., 2019).

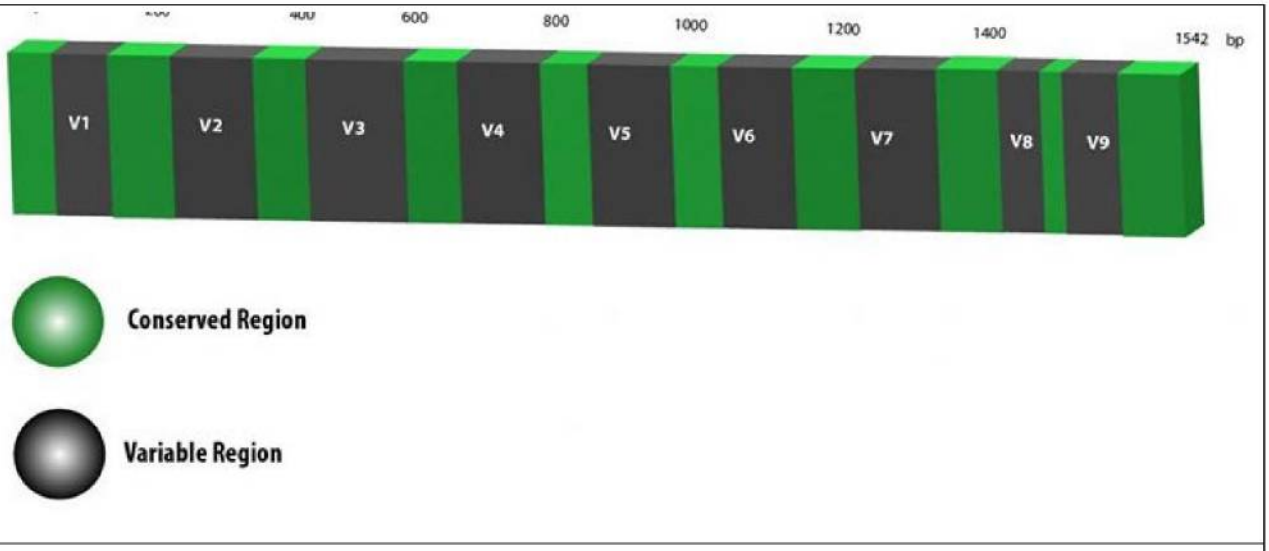


Figure 2: The conserved and variable regions of 16s rRNA which make it an excellent tool for sequencing. Downloaded from https://www.mrdnalab.com/what-is-16s-rrna.html.

The broad application of 16S rRNA sequencing has enormously improved our understanding of microbial differences. By comparing 16S rRNA groupings, analysts can construct phylogenetic trees that outline the evolutionary relationships among microorganisms (Martiny et al., 2017). This molecular approach outperforms traditional phenotypic methods, which can be susceptible to homoplasy (Parks et al., 2018). As a result, 16S rRNA sequencing provides a more reliable method for bacterial identification and taxonomy (Schloss, 2016).

However, there are limitations to using 16S rRNA for species demarcation. Horizontal gene transfer involving the 16S rRNA, although less common than in other genes, can create phylogenetic inconsistencies (Vos et al., 2015). Another significant challenge is its limited resolution for closely related species, which often have very similar 16S rRNA sequences, making it difficult to distinguish between them (Johnson et al., 2019). This can lead to complications in species-level identification and numerous misclassifications. To improve resolution, scientists may turn to additional genes or whole-genome sequencing, which offers much higher resolution (Jain et al., 2018). In particular, average nucleotide identity (ANI) analysis, as demonstrated in Jain et al. (2018), has shown clear species boundaries across

large-scale genomic datasets, providing a more accurate method for delineating species.

Multilocus Sequence Typing (MLST) is one technique that employs many genes to identify different species of bacteria. In order for MLST to function, certain sections of many housekeeping genes are sequenced, and the sequences from various strains are compared. A strain's sequence type (ST) is determined by combining the allele counts of all the distinct sequences at each gene locus to get an allele number (Maiden et al., 1998). High precision and reliability make this approach a useful tool for researching the evolution and epidemiology of bacteria (Jolley & Maiden, 2010).

MLST has proven to be highly effective in distinguishing bacterial species and sub-species, especially within clinically important groups. By analyzing multiple loci, MLST provides a more comprehensive view of the genetic relatedness among strains, allowing for the detection of recombination events and the tracking of clonal lineages (Maiden et al., 2013). This multilocus approach surpasses single-locus methods, such as 16S rRNA sequencing, in resolving closely related strains and understanding their population structures (Maiden et al., 1998).

The implementation of MLST has significantly advanced the knowledge of microbial taxonomy and pathogenesis. Important cases have been those of *Staphylococcus aureus*, *Neisseria meningitidis*, and *Escherichia coli*, where MLST was used to identify and characterize pathogenic clones (Enright et al., 2000). The data generated from MLST studies are often stored in publicly accessible databases, facilitating global surveillance and comparison of bacterial strains (Jolley et al., 2012).

Despite its strengths, MLST is not without limitations. The method can be labor-intensive and costly, as it requires the sequencing of multiple gene loci for each strain (Jolley & Maiden, 2010). Additionally, the selection of loci can influence the resolution and discriminatory power of the analysis, and different schemes may be needed for different bacterial species (Maiden et al., 2013). These factors necessitate careful consideration in the design and interpretation of MLST studies.

In conclusion, MLST is a powerful and versatile method for bacterial species demarcation, offering high resolution and reliability for epidemiological and phylogenetic studies. While it provides significant advantages over single-locus methods, challenges such as cost and labor intensity must be addressed (Jolley & Maiden, 2010). Advances in sequencing technologies and the integration of MLST data with other genomic information continue to enhance the utility and impact of this method in microbial research (Maiden et al., 2013).

With the development of sequencing technologies, affordable whole genomes are nowadays available for analysis. This availability has led to the development of genome comparison techniques, such as the Average Nucleotide Identity (ANI). This is a genomic similarity metric used for bacterial species demarcation, which compares the nucleotide sequences of whole genomes between pairs of microorganisms. ANI is calculated by comparing genomic fragments from one genome to another and determining the average nucleotide identity of these regions (Jain et al., 2018). This method provides a quantitative measure of genomic similarity, making it a powerful tool for distinguishing bacterial species (Richter & Rossello-Mora, 2009).

The ANI method has proven to be more accurate and reliable than traditional methods, such as DNA-DNA hybridization (DDH), for bacterial species demarcation, an ANI value of 95% generally corresponds to the traditional species boundary defined by 70% DDH similarity, offering a more standardized and reproducible criterion for species identification (Richter & Rossello-Mora, 2009). This higher resolution enables researchers to distinguish closely related species that may not be differentiated by 16S rRNA gene sequencing alone (Jain et al., 2018).

ANI's application has altogether upgraded our understanding of microbial taxonomy and evolution. By providing an exact degree of hereditary relatedness, ANI encourages the identification of novel species and the reclassification of existing ones based on whole-genome comparisons (Konstantinidis & Tiedje, 2005). This genomic approach makes a difference in constructing more precise phylogenetic trees and understanding the evolutionary relationships among bacteria (Rodriguez-R &

Konstantinidis, 2014). In conclusion, ANI has risen as a significant and high resolution method for bacterial species demarcation.

1.3 The Prevotellaceae family

The *Prevotellaceae* family is a group of Gram-negative, anaerobic microbes found within the human gut. These microbes are also found in other mucosal surfaces of the body, such as the oral cavity and the vagina. Members of the *Prevotellaceae* family are known for their capacity to break down complex carbohydrates (Larsen, 2017; Ley, 2016).

Prevotellaceae are very interesting because of their association with different diseases. A change in *Prevotellaceae* populations, known as dysbiosis, has been linked to conditions such as inflammatory bowel disease (IBD), weight, and metabolic disorders (Alvarez, 2016; Kovatcheva-Datchary et al., 2015). Inside the oral cavity, some species within this family are implicated in periodontal disease. Despite these associations, the precise means by which Prevotellaceae influence well-being and disease remain to be elucidated. Advances in metagenomic and microbiome research are shedding light on the complex interplay between *Prevotellaceae* and humans, highlighting their potential role in both the maintenance of well-being and the pathogenesis of infections (Larsen, 2017).

The *Prevotellaceae* family includes a few genera, with *Prevotella* being the most well-known. These organisms show metabolic flexibility, permitting them to adjust to different natural conditions inside the human body. Their genome encodes numerous enzymes capable of degrading a wide range of polysaccharides, which not only benefits the host by aiding in the digestion of dietary fiber but also helps maintain a balanced gut microbiome (Rock et al., 2012). Past the intestine, *Prevotella* species have been considered for their part within the oral microbiome, where they take part in biofilm formation. Members of the *Prevotella* genus have been isolated from polymicrobial infections like dental root canal infections as well as skin wound infections. *Prevotella* species have a wide range of metabolic capabilities which help

them to grow in an anaerobic state through fermenting diverse carbohydrates. These bacteria are crucial to the different states of equilibrium in which they help keep various parts of the human microbiome and, as a result, assist with functions including digestion and immune regulation. The production of SCFAs by these strains in the oral and gut microbiota highlights their role in human health and disease (Ley, 2010; Ley et al., 2006).

One of the best studied species inside this genus is *Prevotella intermedia*, regularly related with periodontal diseases. *Prevotella intermedia,* together with other species like *Prevotella nigrescens*, is associated with the advancement of conditions such as gingivitis and periodontitis. These microorganisms are part of the complex microbial communities that form biofilms on teeth and gums, driving to aggravation and tissue destruction when the balance of these communities is disturbed (Darveau, 2010). It has also been associated with other diseases like ghout (see figure 3).

Along their part in oral health, *Prevotella* species are critical components of the intestine's microbiome. They are involved in the production of short-chain fatty acids (SCFAs) like butyrate, propionate, and acetic acid, which are useful for colon wellbeing (Rock et al., 2012). The relative abundance of *Prevotella* within the intestine has been connected to dietary preferences (Wu et al., 2011). Studies suggest that a *Prevotella*-dominant intestine microbiota may protect against inflammatory diseases, although this association may vary in the context of other microbial interactions and host factors (Kovatcheva-Datchary et al., 2015).

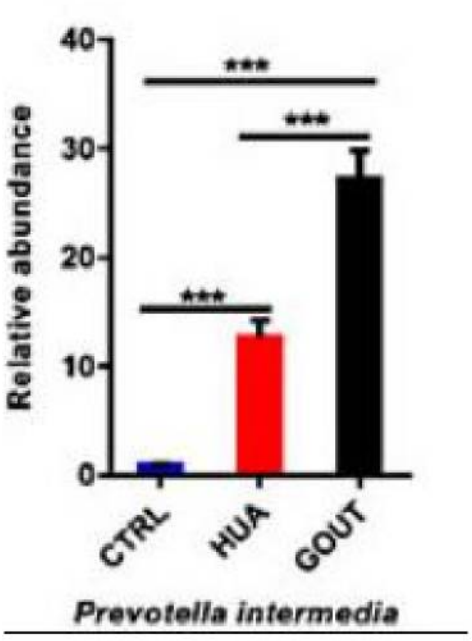


Figure 3: The level of *Prevotella intermedia* was at the highest in the patients with gout, while at the lowest in the healthy controls (***P< 0.001) (Juan Liu et al., 2018).

Aim of the study

The aim of this study was two-fold. First, it was delineate the various species of the *Prevotella* genus and the other closely-related bacterial genera of the *Prevotellaceae* family. Secondly it was to perform comparative analyses of these species and identify their core-, fingerprint- and carbohydrate-degrading genes.

2 Materials and methods

2.1 Software

2.1.1 Linux Ubuntu 24.04 LTS

Ubuntu is an open-source operating system based on the Linux kernel, known for its user-friendly interface. Created by Canonical Ltd., Ubuntu is broadly utilized by people, businesses, and teachers for its soundness, security, and adaptability. It is especially favored for its ease of use. Ubuntu is accessible in different versions, as desktop, server, and cloud forms, each custom-made to particular needs. With standard upgrades, Ubuntu proceeds to be a driving choice for those looking for a solid and customizable computing environment. Also, the management of large volumes of data and process automation make it an external tool for bioinformatics analyses (Ubuntu: Enterprise Open Source and Linux). The most important tool included in Ubuntu is the terminal, which features BASH, allowing for the efficient and rapid analysis of large datasets. This powerful functionality makes it an essential resource for handling complex data tasks with speed and precision.

2.1.2 PERL programing language

Perl is a high-level, general-purpose programming language, known for its adaptability and control. Over the years, Perl has advanced, with Perl 5 being broadly utilized. In spite of confronting competition from more up to date languages, Perl remains a trusted tool (Christiansen, 2012) (The Perl Programming Language – www.perl.org).

2.1.3 Visual Studio Code (VS Code) v.1.89.1

Visual Studio Code (VS Code) is a prevalent, open source code editor created by Microsoft. Firstly published in 2015, it became widely known due to its flexibility, speed and ease of use. VS Code can be extended with addons that enable users to

customize their editor to their needs while supporting a wide variety of programming languages This code editor is available forWindows, macOS, and Linux.

2.2 NCBI

The first step in conducting this study was to obtain and download a large genomic dataset of the bacterial family *Prevotellaceae,* which is categorized with the taxonomic identification (taxid) 171552. The NCBI Taxonomy database, which offers comprehensive taxonomic information, and the NCBI Assembly database, a popular repository for genomic assemblies, were the sources from which the data was taken. The collection contained every genomic sequence that was currently available for *Prevotellaceae* family members. This includes chromosomes, contigs, scaffolds, and complete genomes. 1,203 genomes were obtained from NCBI, for our comparative analyses.

2.3 pyPGCF, a comparative bacterial genomics python software

Then, in order to proceed with the necessary analyses, the pyPGCF (python PhyloGenomics, Core genome, Fingerprint) pipeline (Nikolaidis et al., 2024), which was developed by the Bioinformatics Laboratory team of the Department of Biochemistry and Biotechnology of the University of Thessaly, Larissa, Greece, was used.

The specific pipeline uses a series of scripts written in Python and is executed in the Linux terminal, providing both phylogenomic analyses, as well as core and fingerprint gene identification analyses, using the appropriate modules. This software can also be used for species demarcation, secondary metabolites identification with antiSMASH and functional annotation through the EggNOG Database (**Figure 4**).

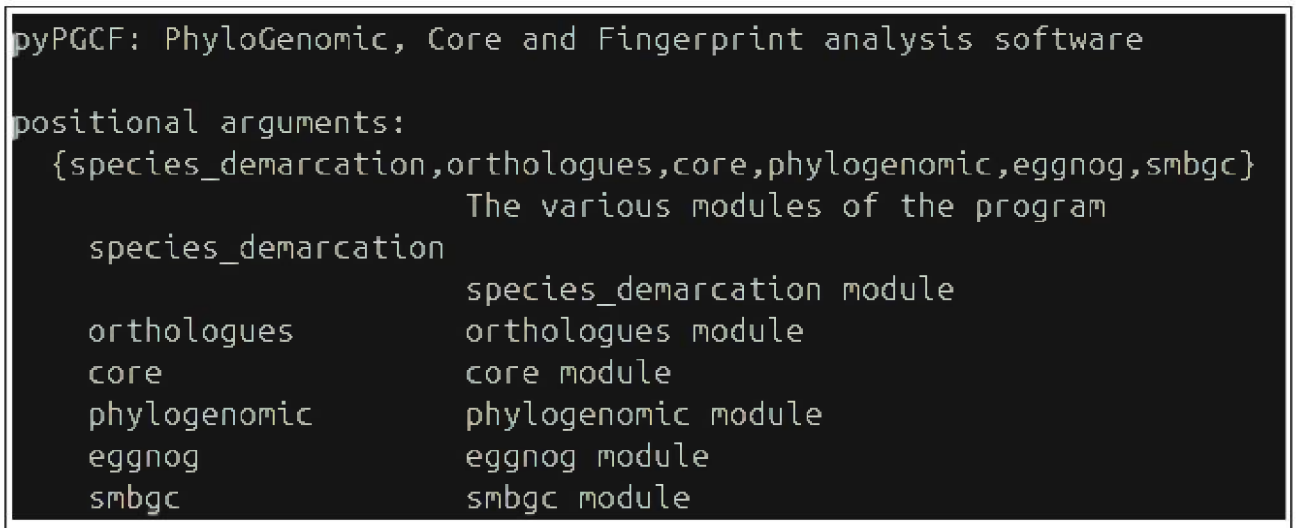


Figure 4: The various pyPGGF modules (species_demarcation, orthologues, core etc.) (Nikolaidis et al, 2024). For the analyses described in this thesis the species_demarcation, orthologues, core and phylogenomic modules were used.

2.4 Multiple Sequence alignment with MUSCLE

MUSCLE (https://www.ebi.ac.uk/jdispatcher/msa/muscle) is a commonly used bioinformatics application for aligning multiple nucleotide and protein sequences. This tool is available either online or the command line and can be used to precisely and quickly identify conserved regions. It is an important step for estimating phylogenies.

2.5 Gblocks

Gblocks is a bioinformatics tool designed to improve the accuracy of phylogenetic analysis by selecting conserved regions from multiple sequence alignments. When working with DNA, RNA, or protein sequences, certain parts may contain gaps, ambiguities, or non-homologous regions that can interfere with accurate evolutionary interpretations. Gblocks addresses this by automatically identifying and removing poorly aligned or highly variable regions, leaving only the most conserved,

reliable segments for further analysis. This process helps ensure that the remaining sequences are robust for inferring phylogenetic relationships.

The algorithm behind Gblocks uses strict selection criteria to retain blocks of aligned sequences that are evolutionarily conserved. These blocks must meet thresholds for length, number of sequences, and level of conservation across taxa, which helps reduce noise and improve the signal for evolutionary studies. By focusing on these conserved regions, Gblocks enhances the reliability of downstream analyses like tree-building, minimizing errors caused by misaligned or irrelevant sequence data.

2.6 IQ-Tree2

IQ-TREE2 (http://www.iqtree.org/) is a fast application that can reconstruct the phylogenetic tree from sequence data. It is an improvement of its predecessor, the first IQ-TREE program, using Maximum Likelihood (ML) method to estimate phylogenies. Large datasets can be handled by incorporating advanced models in phylogenetic inference with precision and accuracy. These include best-fit substitution models which are automatically chosen as well as branch length optimization.

It takes input in many formats such as nucleotide, amino acid sequences and runs on both personal computers and high performance computing clusters. This software also contains tools for model testing, bootstrapping and statistical analysis of tree reliability hence providing comprehensive support for phylogeny inference.

2.7 Interactive Tree of Life (iTOL).

Interactive Tree if Life (TOL) (https://itol.embl.de/) is a versatile web-based tool for phylogenetic tree manipulation, visualization and analysis. It is commonly applied in biology, genetics and ecology due to its ability of displaying phylogenetic

trees and annotating them with other data. Users can annotate their phylogenetic trees with labels and color the tree and add other features to improve the interpretation of the trees. Also iTOL has a wide range of annotation files templates with which information can be integrated into the tree and modified. Users can export trees in a variety of formats including images and PDFs.

2.8 FastANI

FastANI is a bioinformatics application designed to calculate Average Nucleotide Identity (ANI) from whole genome sequences in an efficient manner. It compares two microbial genomes using ANI as a proxy for species-level relatedness. Traditional methods for computing ANI on large datasets can be inefficient and time-consuming. FastANI addresses this issue by offering a highly optimized algorithm that significantly speeds up the calculation time without sacrificing much precision. The algorithm works by fragmenting genomes into small sequences, rapidly aligning them, and then computing ANI based on these aligned fragments, ensuring both speed and accuracy even on large-scale genomic datasets. The new tool is especially important for microbial genomics because it can be much faster than other methods at comparing thousands of genomes to determine genetic similarity and resolve taxonomic relationships.

2.9 dbCAN2 identification of Carbohydrate-Active enZYmes (CAZymes)

The CAZymes (Carbohydrate-Active enZYmes) database (http://www.cazy.org/) is an extensive resource for enzymes involved in biomass breakdown. The CAZymes database categorizes enzymes in the eukaryotic, bacterial and archaeal genomes into different categories based on their functional as well as structural features namely Glycoside hydrolases (GHs), glycolsyltransferases (GTs),

carbohydrate esterases (CE), polysaccharides lyases and accessory proteins. The database groups these enzymes based on sequence similarities and conserves catalytic mechanisms, providing a framework with which to understand key enzyme activities within the context of carbohydrate metabolism in diverse organisms.

The CAZymes database is an important tool for investigators exploring the universe of carbohydrate-active enzymes across the life forms and their biotechnological applications. For example, the information is important for biofuel production to learn how enzymes like cellulases degrade plant biomass and in health research where metabolic diseases are studied using amylases and glucosidase (needed not only for primary digestion of sugar complexes but also as potential therapeutic agents). Such detailed annotation of enzyme families and the ability to compare results across protein groups makes CAZymes a key tool for both basic research in carbohydrate biology, as well as applied science.

The dbCAN web server (https://bcb.unl.edu/dbCAN2/) was developed in 2012 (Zhang et al, 2018) to provide a public service for automated CAZyme annotation for newly sequenced genomes. dbCAN2 is an upgraded version of the original dbCAN, providing more accurate and comprehensive annotations by integrating several computational tools. These tools are designed to predict CAZymes based on protein or nucleotide sequences, assisting researchers in studying the structure, function, and diversity of carbohydrate-related enzymes in various organisms.

3 Results – Discussion

Initially, 1,203 genomes of the *Prevotellaceae* family were downloaded from NCBI and filtered through a series of steps in order to ensure that only high quality genomes would be retained for further analyses.

More specifically, genomes that exhibited an N% greater than 1% were excluded from further consideration. The rationale behind this threshold is that a significant proportion of unknown nucleotides (N%) can compromise the accuracy of downstream analyses, potentially leading to misleading results. In addition to the automated filtering process, a manual review of the genomes was conducted, and those with a disproportionately small number of proteins or a genome size compared to the other member of the family were also discarded. As a result of this filtering process, 23 genomes that did not meet the established quality criteria were rejected. Consequently, a final set of 1,180 genomes was retained, representing high-quality sequences.

Then, the species_demarcation module from the pyPGCF pipeline (Nikolaidis et al, 2024) was used to demarcate the 1,180 remaining genomes into species-clusters using the ANI value cut-off of 95%, resulting in 145 different species-clusters. From each cluster, a genome was defined as a reference genome based on the number of proteins, except for *Prevotella melaninogenica*, for which GCF_000144405.1 was defined as a reference based on NCBI.

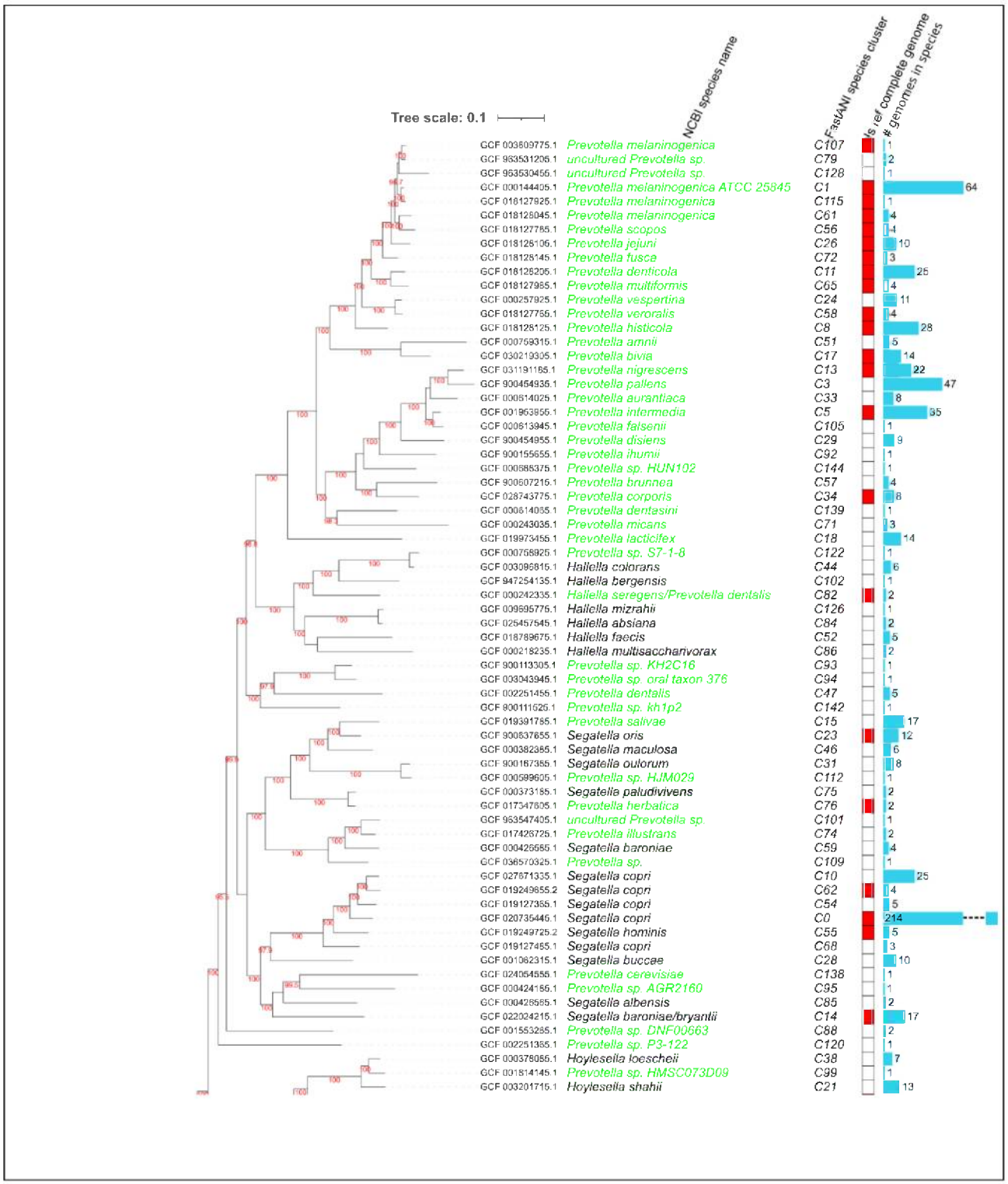
After the proteins from each reference were collected in FASTA format, they were used to find orthologous proteins through the orthologues module of pyPGCF (Nikolaidis et al, 2024). The orthologues module identifies orthologous groups across the genomes being analyzed using a reference-based best reciprocal BLAST (Basic Local Alignment Search Tool) search. For protein sequences, DIAMOND is employed, while BLASTN is used for nucleotide sequences. Both searches are performed with an e-value threshold of 1e-5 by default, though this parameter can be adjusted by the user.

A table of best matching hits for all query proteomes was then created, with the different rows corresponding to the orthologous proteins and the columns to the different proteomes. This table was then used to calculate the core genome of the entire family through the core module (Nikolaidis et al, 2024). The phylogenomic species-tree of the entire family was calculated using the phylogenomic module (Nikolaidis et al, 2024) and visualized through the interactive tree of life (iTOL). More specifically, through the phylogenomic module each group of orthologs was subjected to multiple alignment using MUSCLE and then the different multiple sequence alignments were concatenated, creating a superalignment that was then filtered with Gblocks (default parameters). This final superalignment was used to generate a phylogenomic tree using IQ-TREE2 software. Assessment of branch support was performed with the approximate Likelihood Ratio Test (aLRT).

As input file to iTOL, the .nwk file was used (as it uses Newick, Nexus or PhyloXML format files). This file was generated in the previous step to construct the phylogenetic tree (Figure 5), incorporating each reference genome from the 145 clusters resulting from the species_demarcation_module (Nikolaidis et al., 2024). In the accompanying tree, we can observe the various species within the *Prevotellaceae* family, each assigned to the corresponding cluster to which they belong. Additionally, the image highlights which of these genomes are complete genomes—indicated by those colored red in the first column. There is also a column that displays the number of genomes present in each cluster, providing a clear overview of genomic representation. Notably, the largest number of genomes is found in cluster C0, which corresponds to the species *Segatalla copri*. Furthermore, the *Prevotellaceae* family contained a total of 137 core genes, and the alignment used for the construction of the tree comprised 27,127 sequence positions, underscoring the extensive genetic information utilized in this analysis.

Also, in the tree, the species that are characterized as members of the genus *Prevotella* are shown in green. According to studies, species with ANI $\geq$ 78% seem to belong to the same genus. Thus, after FastANI analysis, it appeared that the family contains 7 known genera to which the species of the tree belong. As for the *Prevotella*

genus, it consists of the first 29 species with the last one being *P. laxticifex*, so the rest marked with green colour below that were misclassified as species of the genus and should be renamed. Also, *Prevotella* is the genus which includes the most species-clusters.



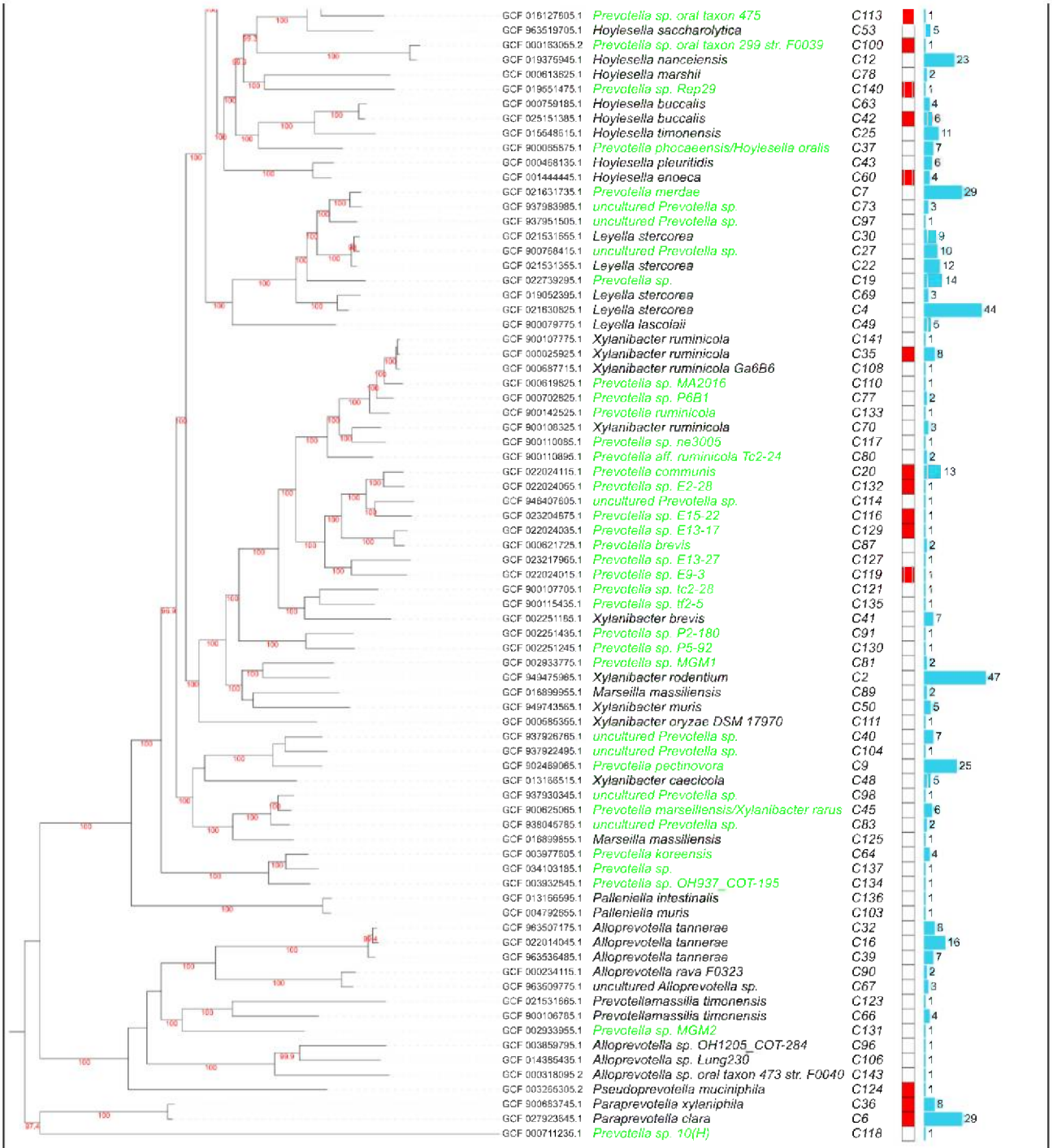


Figure 5: In the tree above, we observe the species of the family. Next to them is the number of the cluster they belong to as well as a column indicating which of them is a complete genome and another column showing the total number of genomes of each cluster. The family had 137 core genes and the alignment for the tree had 27,127 sites and iTOL was used to build the tree.

Next, the phylogenomic tree of the 1,180 genomes of the entire *Prevotellaceae* family, was also computed through the phylogenomics module of pyPGCF (Nikolaidis et al., 2020). The selected maximum likelihood model in this case was the LG+I+F+G4 model. By manually inspecting the phylogenomic tree, species-clusters that contained more than 5 genomes were identified, for a total of 56. In order to facilitate more meaningful comparisons and account for the sampling bias of certain important species-clusters each species was reduced to a maximum of 5 genomes, preferring complete genomes over draft ones.

This new, normalized set of genomes, consisted of 433 genomes and was utilized for the calculation of core and fingerprints genes of the 56 aforementioned species. For each of these 56 species, the orthologues of the reference strain against the other 432 strains were calculated using the orthologues module within the pyPGCF software. Then, the tables with the orthologous proteins obtained from the orthologues module were used as input files for the core module along with the relevant species-demarcation information. This approach enabled the calculation of both the core genomes and the fingerprint genes corresponding to each species. The results of these analyses are shown in Table 1.

Reference Assembly	NCBI Organism name	# core genes	# fingerprints	Species-Cluster
GCF_020735445.1	*Segatella copri DSM 18205*	1705	5	C0
GCF_018128005.1	*Prevotella melaninogenica*	1726	0	C1
GCF_949475965.1	*Xylanibacter rodentium*	2247	62	C2
GCF_900454935.1	*Prevotella pallens*	1604	9	C3
GCF_021630825.1	*Leyella stercorea*	1566	3	C4
GCF_002763955.1	*Prevotella intermedia*	1662	3	C5
GCF_027923645.1	*Paraprevotella clara*	2063	18	C6
GCF_021631735.1	*Prevotella sp.*	1705	4	C7
GCF_018128125.1	*Prevotella histicola*	1740	18	C8

GCF_902469065.1	*uncultured Prevotella sp.*	1590	12	C9
GCF_027671335.1	*Segatella copri*	1833	3	C10
GCF_018128205.1	*Prevotella denticola*	1822	4	C11
GCF_019375945.1	*Hoylesella nanceiensis*	1596	4	C12
GCF_018127865.1	*Prevotella nigrescens F0103*	1737	14	C13
GCF_022024235.1	*Segatella baroniae B14*	2129	48	C14
GCF_019391785.1	*Prevotella salivae*	1887	14	C15
GCF_022014045.1	*Alloprevotella tannerae*	1522	5	C16
GCF_030219305.1	*Prevotella bivia*	1556	17	C17
GCF_019973455.1	*Prevotella lacticifex*	2961	219	C18
GCF_022739295.1	*Prevotella sp.*	1725	2	C19
GCF_022024115.1	*Prevotella communis*	2115	18	C20
GCF_003201715.1	*Hoylesella shahii DSM 15611*	1913	25	C21
GCF_021531355.1	*Leyella stercorea*	1764	10	C22
GCF_900637655.1	*Segatella oris*	1770	3	C23
GCF_000257925.1	*Prevotella sp. oral taxon 306 str. F0472*	1542	4	C24
GCF_015548615.1	*Hoylesella timonensis*	1547	4	C25
GCF_018128105.1	*Prevotella jejuni*	1978	4	C26
GCF_900768415.1	*uncultured Prevotella sp.*	1453	0	C27
GCF_001062315.1	*Segatella buccae*	1903	17	C28
GCF_000759225.1	*Prevotella disiens DNF00882*	1519	17	C29
GCF_021531655.1	*Leyella stercorea*	1644	8	C30
GCF_900167385.1	*Segatella oulorum*	1586	1	C31
GCF_963507175.1	*Alloprevotella tannerae*	1500	1	C32
GCF_000614025.1	*Prevotella aurantiaca JCM 15754*	1448	19	C33
GCF_028743775.1	*Prevotella corporis*	1655	24	C34
GCF_000025925.1	*Prevotella ruminicola 23*	2048	5	C35
GCF_900683745.1	*Paraprevotella xylaniphila YIT 11841*	2222	23	C36
GCF_900065875.1	*Prevotella phocaeensis*	1989	34	C37
GCF_000378085.1	*Hoylesella loescheii DSM 19665*	1941	14	C38
GCF_963536485.1	*Alloprevotella tannerae*	1522	1	C39
GCF_937926765.1	*uncultured Prevotella sp.*	1686	9	C40
GCF_002251185.1	*Prevotella sp. P5-126*	1914	24	C41

GCF_025151385.1	*Hoylesella buccalis ATCC 35310*	1719	2	C42
GCF_000468135.1	*Hoylesella pleuritidis F0068*	1501	12	C43
GCF_003096815.1	*Hallella colorans*	1631	6	C44
GCF_900625065.1	*Prevotella marseillensis*	1974	12	C45
GCF_000382385.1	*Segatella maculosa DSM 19339*	1444	8	C46
GCF_002251455.1	*Prevotella sp. P4-51*	1931	86	C47
GCF_013166515.1	*Xylanibacter caecicola*	1528	21	C48
GCF_900079775.1	*Leyella lascolaii*	1767	46	C49
GCF_949743565.1	*Xylanibacter muris*	2332	94	C50
GCF_000759315.1	*Prevotella amnii DNF00058*	1383	16	C51
GCF_018789675.1	*Hallella faecis*	1814	20	C52
GCF_963519705.1	*Hoylesella saccharolytica*	1454	10	C53
GCF_019127365.1	*Segatella copri*	2069	4	C54
GCF_019249725.2	*Segatella hominis*	1632	5	C55

Table 1: Table with the number of core genes and fingerprint genes for each species-cluster with more than 5 genomes.

Then, the 7 genera from the phylogenetic tree were organized into distinct groups to calculate the core and fingerprint genes for each group separately. To accomplish this, all the clusters associated with each genus were merged together. Once these merged clusters were established, a new set of clusters was formed, representing the combined genomic profiles of each genus. This reorganization allowed for a more accurate calculation of the core and fingerprint genes, which were now based on the merged clusters rather than the individual ones. The results of this reorganization and clustering process are presented in Table 2, illustrating how the species were grouped. Additionally, the genomes listed in the table were selected as reference genomes for each genus.

Genus	Included clusters	Reference Genome
Segatella	C0, C55, C68, C54, C10, C62	GCF_009495075.1
Hallella	C122, C44, C102, C82, C126, C84, C86, C52	GCF_009695775.1
Hoylesella	C99, C38, C21, C113, C53, C100, C12, C78, C140, C42, C63, C25, C37, C43, C60	GCF_015548615.1
Leyella	C30, C27, C22, C73, C7, C97, C19, C69, C4, C49	GCF_021630825.1
Xylanibacter	C35, C141, C108, C110, C77, C133, C117, C70, C80, C114, C116, C132, C20, C87, C129, C127, C119, C41, C135, C121, C130, C91, C50, C89, C81, C2	GCF_949475965.1
Alloprevotella	C16, C32, C39, C90, C67, C123, C66, C131, C106, C96, C143, C124	GCF_900106785.1
Prevotella	C107, C79, C128, C1, C115, C61, C56, C26, C72, C65, C11, C58, C24, C8, C17, C51, C3, C13, C33, C105, C5, C29, C92, C144, C34, C57, C71, C139, C18	GCF_000144405.1

Table 2: The table above shows the genera for which core and fingerprint genes were calculated as well as the clusters of each genus.

The reference genome for each genus was selected based on the number of proteins it had and its assembly status, similar to previous analyses. An exception to this selection process was made for the genus *Prevotella,* where the genome GCF_000144405.1 of *P. melaninogenica* was chosen as a reference genome because it was listed as the reference of the genus in the NCBI database.

Thus, following a similar approach as in previous steps, the resulting tables generated after running the orthologues module were subsequently utilized as input files for the core module. As a result of this analysis, the core genes, which represent the essential genetic components conserved across all members of a genus, and the fingerprint genes, which highlight the unique genetic markers of each genus, were identified. These findings are listed in Table 3.

Genus	Reference Genome	# core genes	# fingerprints
Segatella	GCF_009495075.1	1419	1
Hallella	GCF_009695775.1	948	0
Hoylesella	GCF_015548615.1	787	0
Leyella	GCF_021630825.1	907	0
Xylanibacter	GCF_949475965.1	909	0
Alloprevotella	GCF_900106785.1	858	3
Prevotella	GCF_000144405.1	524	0

Table 3: Number of core and fingerprint genes of each genus. For these calculations the corresponding genome that appears next to each genus was used as reference.

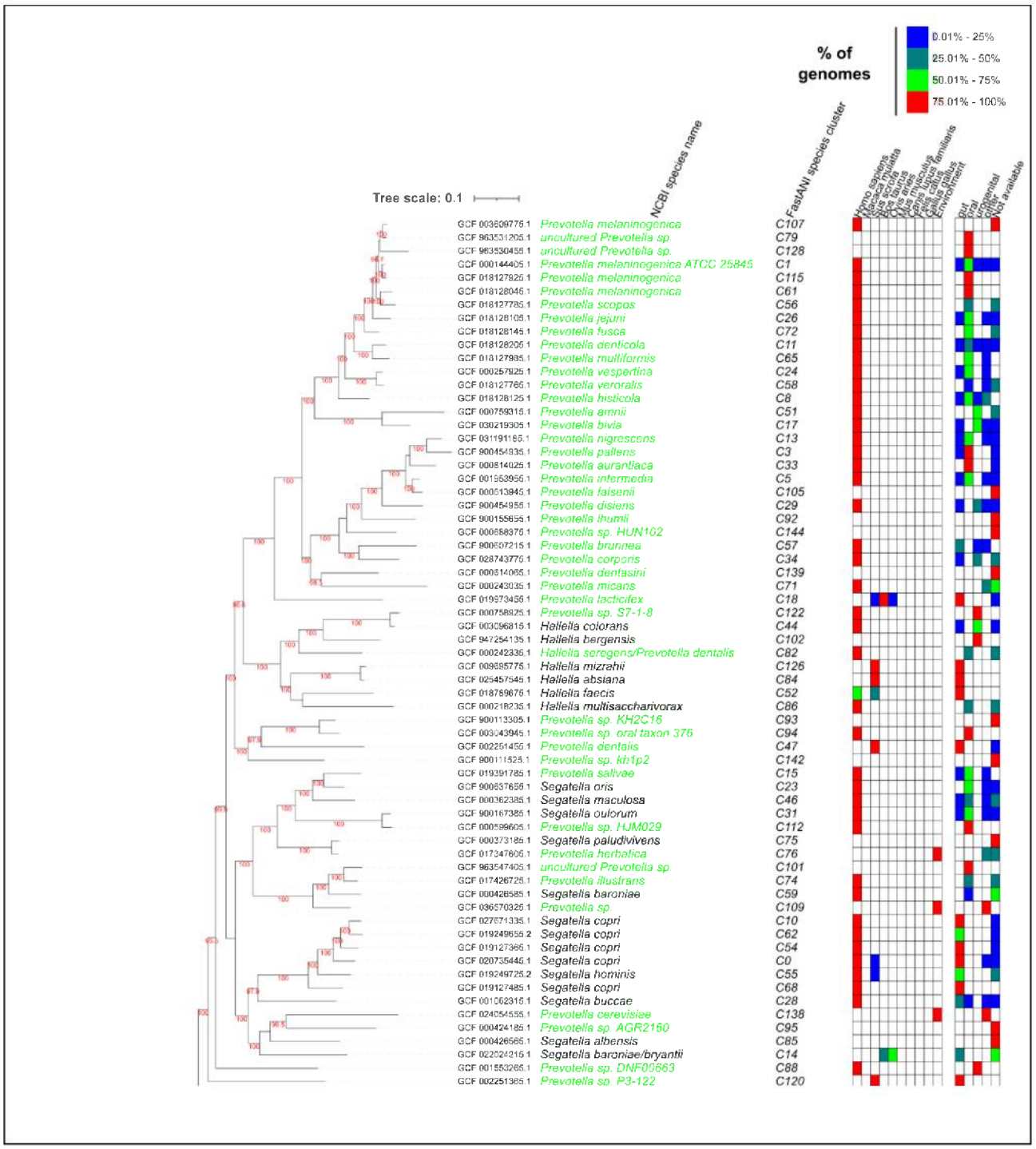
In the table above, it is observed that the genera *Segatella* and *Alloprevotella* possess 1 and 3 fingerprint genes, respectively. Further details regarding these fingerprint genes can be found in Table 4, where the corresponding proteins associated with these genes are listed. Additionally, the NCBI annotations of these proteins were retrieved. We note that for this part there is not much information available about the specific proteins other than the protein found in the genus *Segatalla* which is proposed to bind to ATP.

Genus	Reference Genome	Protein	NCBI annotation
Alloprevotella	GCF_900106785.1	WP_071122601.1	hypothetical
Alloprevotella	GCF_900106785.1	WP_071122726.1	hypothetical
Alloprevotella	GCF_900106785.1	WP_071123210.1	hypothetical
Segatella	GCF_009495075.1	WP_194256733.1	ATP_binding protein

Table 4: The table shows the proteins that correspond to the fingerprint genes for the genera *Alloprevotella* and *Segatalla*, as well as their role.

Then, using Perl scripts, essential information was extracted from the files acquired from NCBI. The data extracted included key details, such as information regarding the organism from which each genome had been isolated, the country in which the organism was observed, and other relevant metadata. In the tree presented in **Figure 6**, two heatmaps are displayed, representing the host and the isolation source for each species, respectively. Concerning the host, it is evident that the majority of species were isolated from humans (*Homo sapiens*), comprising a significant portion of the dataset. This correlation helps to visualize the distribution of species across different hosts.

Observing the second heatmap in **Figure 6**, which illustrates the isolation source, it becomes evident that the species of the *Prevotellaceae* family were most frequently identified at the gut, oral cavity, and urogenital tract. Additionally, data was available for other microbiomes, while some species lacked any information regarding their isolation source. This visualization highlights that certain species from specific genera are found in both the gut and the oral cavity, indicating their presence across multiple organs. This result may suggest that these organisms may have the ability to adapt to diverse environments within the host, indicating their ecological flexibility and potentially broader roles in host-microbiome interactions.



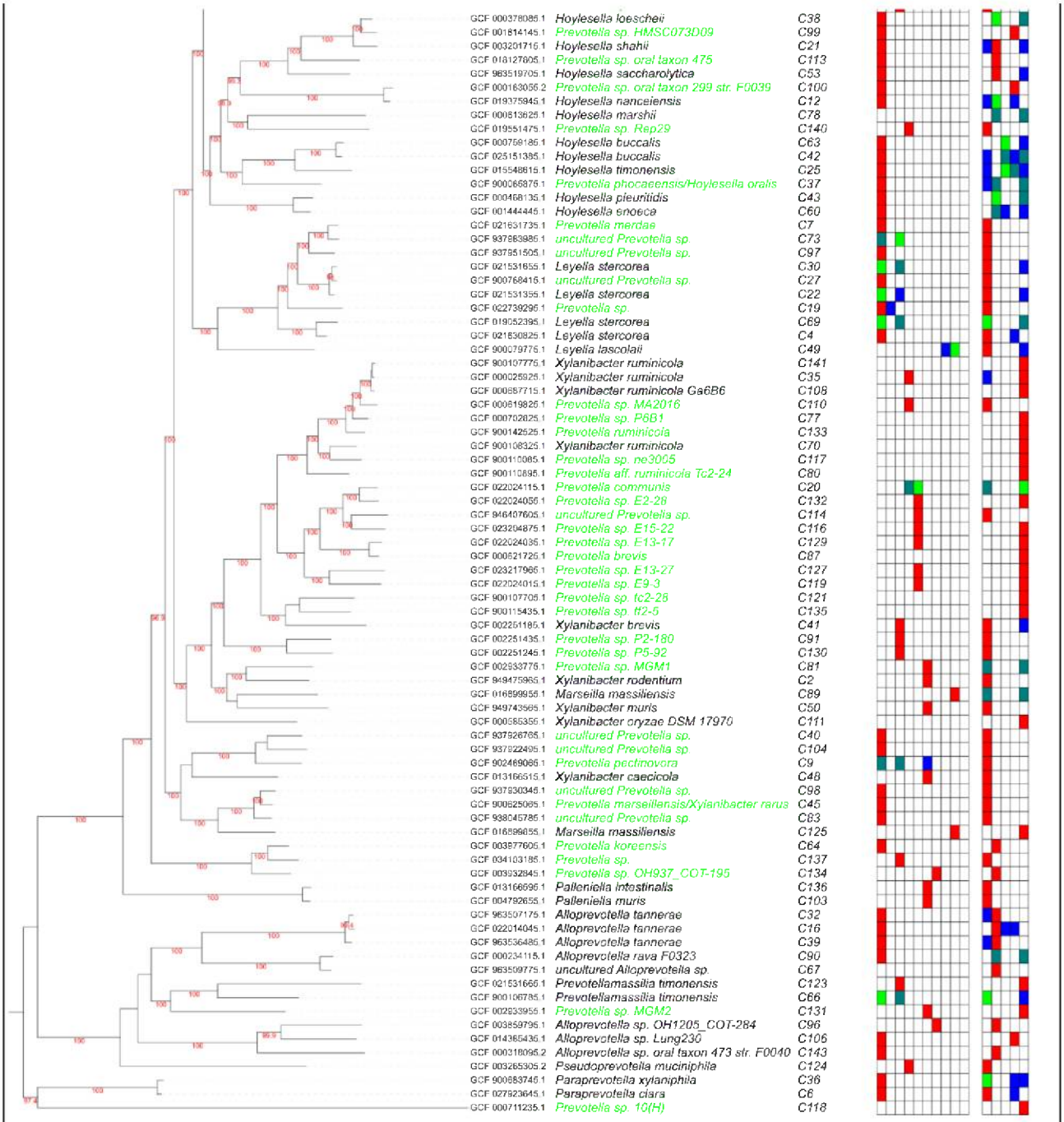


Figure 6: Heatmaps showing the host (left heatmap) and isolation source (right heatmap) where the genomes were isolated from. Blue color corresponds to 0.01% - 25% of genomes, gray color to 25.01% - 50% genomes, green color to 50.01% - 75% of genomes and red color to 75.01% - 100% of genomes.

Subsequently, the CAZymes of the 145 species belonging to the family, as depicted in Figure 5, were identified. The dbCAN2 tool was employed to accurately assign the enzymes to their respective CAZyme categories.

The results of the CAZyme analysis is summarized in Table 5, where the various types of carbohydrate-active enzymes are listed according to their functional classification. Also, in the chart of Figure 7, we can observe the percentage of each category of CAZymes to which the species of the tree were assigned.

In the chart we can observe that the most frequent class of CAZymes identified is GH (Glycoside Hydrolases). The fact that the species have many enzymes that catalyze the hydrolysis and rearrangement of glycosidic linkages may suggest that their diet is flexible, enabling them to use a range of polysaccharides. These enzymes assist bacteria in breaking down complex carbohydrates and adjusting to environments that are high in organic materials, such as plant detritus. They could also suggest a function in the recycling of organic matter and, in some situations, be linked to antibiotic resistance.

Cazymes Categories	Functional Classification
Glycoside Hydrolases (GHs)	Hydrolysis and/or rearrangement of glycosidic bonds
GlycosylTransferases (GTs)	Formation of glycosidic bonds
Polysaccharide Lyases (PLs)	Non-hydrolytic cleavage of glycosidic bonds
Carbohydrate Esterases (CEs)	Hydrolysis of carbohydrate esters
Auxiliary Activities (AAs)	Redox enzymes that act in conjunction with CAZymes
Carbohydrate-Binding Modules (CBMs)	Adhesion to carbohydrates

Table 5: The above table shows the categories of CAZymes with a brief description of their function. Data was collected from Cazy - Home (http://www.cazy.org/).

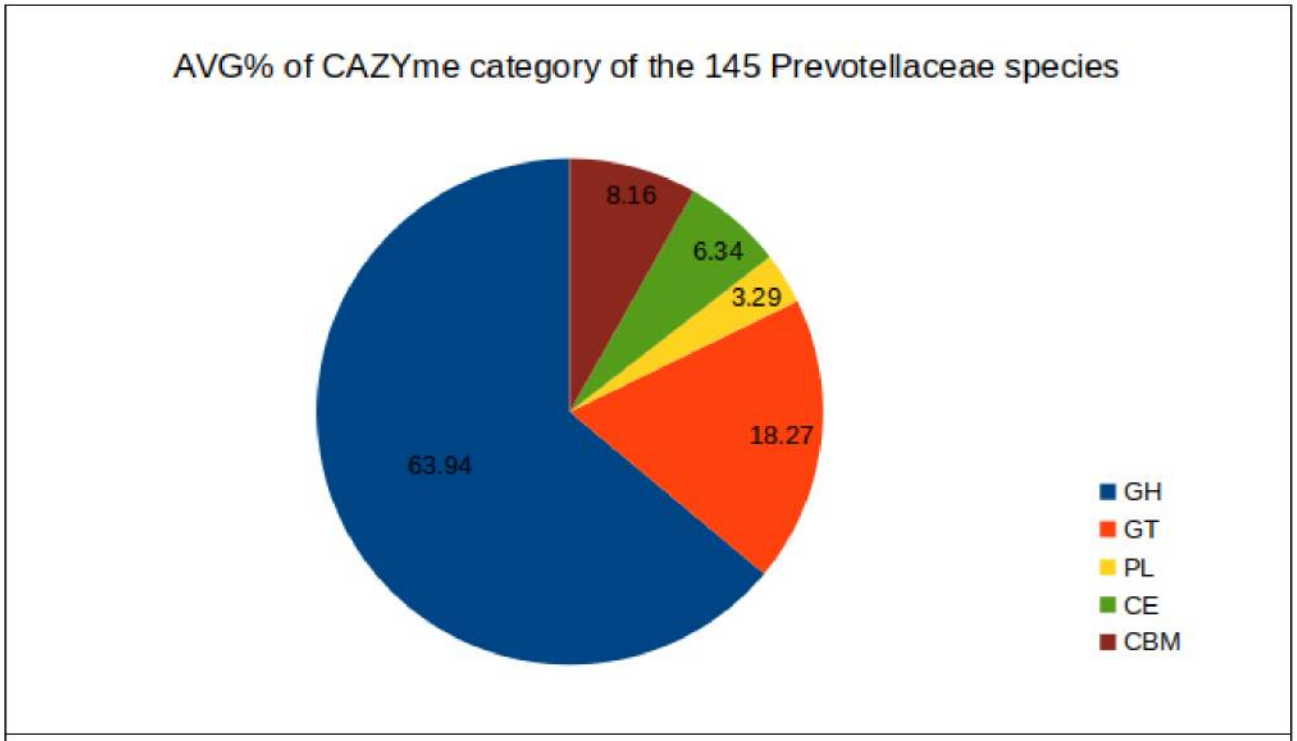


Figure 7: Piechart of each category of CAZymes found in the 145 species of the tree is observed.

Finaly, in Figure 8, we observe a boxplot that visually represents the distribution of species-fingerprint genes across different genera. This distribution was derived from Table 1, where species were categorized into one of the seven genera. Upon analyzing the figure, it becomes evident that, for the most part, the distribution of fingerprint genes is relatively uniform across the majority of the genera. However, one notable exception is the genus *Xylanibacter*, which exhibits significant variability in its fingerprint genes values. This is clearly reflected in the boxplot, where we see a wide range of values indicating that the species within this genus have a large spread in their fingerprint genes. In contrast, the genus *Segatalla* shows a very different pattern. Here, there are very few species-fingerprints recorded, suggesting that the species in this genus are genetically quite similar, with minimal variation in their fingerprint genes values.

In terms of the genus *Prevotella*, the boxplot reveals relatively small variation in fingerprint gene values. However, there is a noticeable outlier in the genus, represented by the species *P. lacticifex*, which deviates significantly from the other species. Aside from this outlier, the remaining species within the genus have fingerprint values ranging from 0 (minimum) to 24 (maximum). This suggests that while the majority of the species within *Prevotella* share similar genetic characteristics, *P. lacticifex* is an outlier.

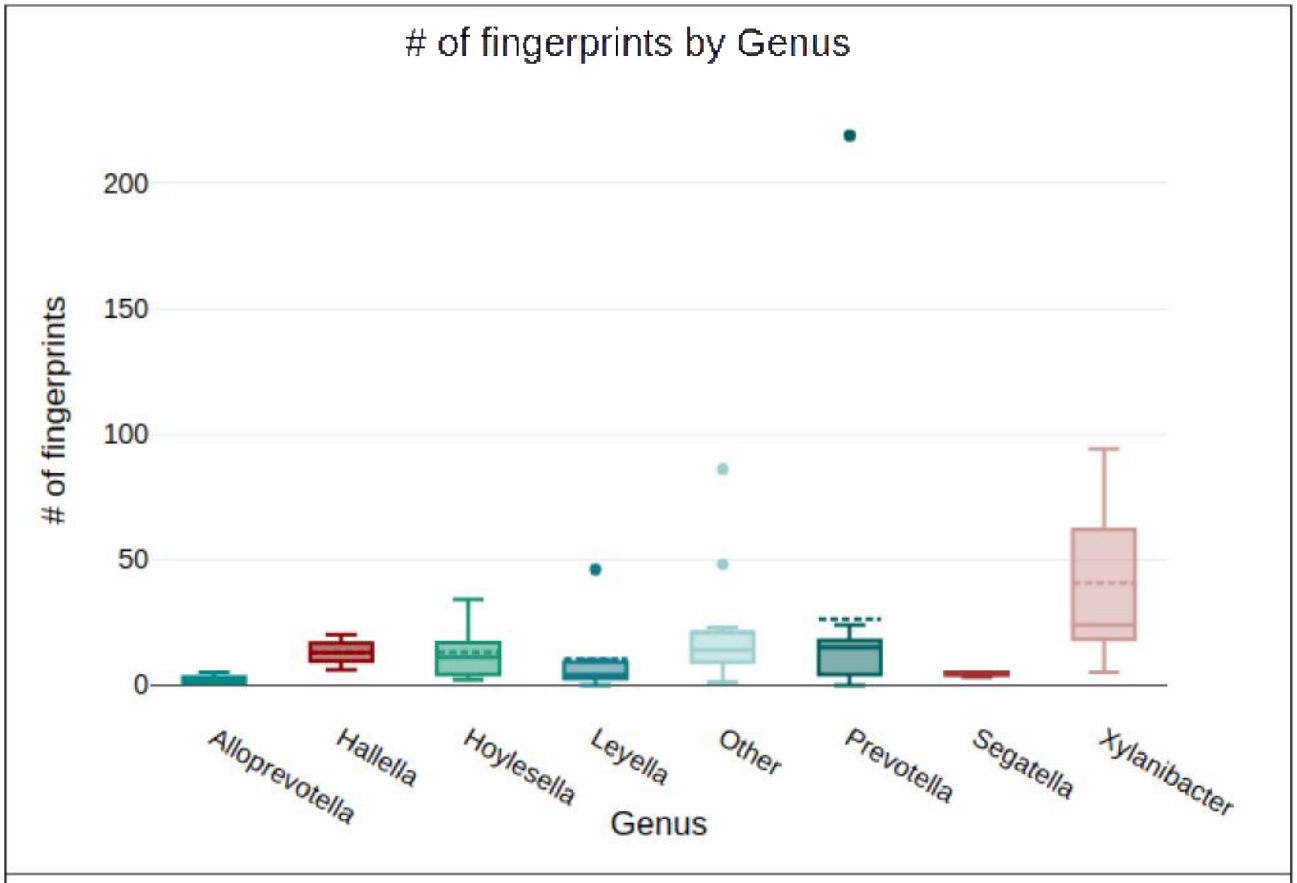


Figure 8: Boxplots in of of the number of identified fingerprints of the 56 species with 5 or more genomes each. These species were classified based on the genus they belong to.

In conclusion, important results were presented regarding the *Prevotellaceae* family. The bacterial genus *Prevotella*, a member of this family, is a center of interest, since as mentioned, according to the evolutionary and genomic analysis carried out, it is the genus of the family with the most species. However, several mis-classifications were

observed according to ANI values, so perhaps some species should be renamed. Also, it was observed that most species of *Prevotella* were isolated from humans and the greater number of them more specifically from the gut and oral cavity. But in terms of fingerprint genes, not so many have been been found.

4 Bibliography

- Alvarez, J. (2016). The microbiome and inflammatory bowel disease. Journal of Gastroenterology and Hepatology, 31(1), 14-19.
- Christiansen T. (2012). Programming Perl, 4th Edition, U.S.A.: O'Reilly Media.
- Croucher, N. J., Mostowy, R., Wymant, C., Turner, P., Bentley, S. D., Fraser, C., & Cornelissen, C. N. (2017). Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. PLoS biology, 15(11), e2000122.
- Darveau, R. P. (2010). Periodontitis: a polymicrobial disruption of host homeostasis. Nature Reviews Microbiology, 8(7), 481-490.
- Enright, M. C., et al. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. Journal of Clinical Microbiology, 38(3), 1008-1015.
- Gordienko, E. N., Kazanov, M. D., Gelfand, M. S., & osterman, A. L. (2013). Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. Journal of bacteriology, 195(12), 2786-2792.
- Jain, C., et al. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications, 9(1), 5114.
- Juan Liu., Li Cui., Xinmin Yan., Xinyuan Zhao., Jinluo Cheng., Lei Zhou., Jianbo Gao., Zhong Cao., Xinhua Ye., Shen Hu. (2018). Analysis of Oral Microbiota Revealed High Abundance of Prevotella Intermedia in Gout Patients, 49(5) 1804–1812.

➢ Johnson, J. S., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature Communications, 10(1), 5029.

➢ Jolley, K. A., & Maiden, M. C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics, 11(1), 595.

➢ Jolley, K. A., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology, 158(3), 1005-1010.

➢ Kovatcheva-Datchary, P., et al. (2015). Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of Prevotella. Cell Metabolism, 22(6), 971-982.

➢ Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences, 102(7), 2567-2572.

➢ Larsen, J. M. (2017). The immune response to Prevotella bacteria in chronic inflammatory disease. Immunology, 151(4), 363-374.

➢ Ley, R. E. (2010). Obesity and the human microbiome. Current Opinion in Gastroenterology, 26(1), 5-11.

➢ Ley, R. E. (2016). Gut microbiota in 2015: Prevotella in the gut: choose carefully. Nature Reviews Gastroenterology & Hepatology, 13(2), 69-70.

➢ Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell, 124(4), 837-848.

➢ Lozupone, C.A., et al. (2012). Diversity, stability and resilience of the human gut microbiota. Nature, 489, 220-230.

➢ Maiden, M. C. (2013). Multilocus sequence typing of bacteria. Annual Review of Microbiology, 60, 561-588.

➢ Maiden, M. C., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences, 95(6), 3140-3145.

➢ Marios Nikolaidis, Stephen G Oliver, Grigorios D Amoutzias (2024). pyPGCF: A Python Software for Phylogenomic Analysis, Species Demarcation, Identification of Core, and Fingerprint Proteins of Bacterial Genomes That Are Important for Plants. Methods Mol Biol. 2024:2788:139-155.

➢ Martiny, J. B. H., et al. (2017). Microbial legacies alter decomposition in response to simulated global change. The ISME Journal, 11(2), 490-499.

➢ McInerney, J. o., McNally, A., & o'Connell, M. J. (2017). Why prokaryotes have pangenomes. Nature microbiology, 2(4), 17040.

➢ Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. Current opinion in genetics & development, 15(6), 589-594.

➢ Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M. C., Jolley, K. A., ... & Sheppard, S. K. (2015). A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic campylobacter. PLoS one, 10(3), e0119242.

➢ Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., ... & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31(22), 3691-3693.

➢ Parks, D. H., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology, 36(10), 996-1004.

➢ Richter, M., & Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences, 106(45), 19126-19131.

➢ Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). Bypassing cultivation to identify bacterial species. Microbe Magazine, 9(3), 111-118.

➢ Schloss, P. D. (2016). Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. mSystems, 1(2), e00027-16.

➢ Snipen, L., & Ussery, D. W. (2010). Standard operating procedure for comparing bacterial pan-genomes. Journal of Bacteriology, 192(13), 3463-3465.

➢ Tettelin, H., Masignani, V., Cieslewicz, M. J., Eisen, J. A., Peterson, S., Wessels, M. R., ... & Fraser, C. M. (2008). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences, 102(39), 13950-13955.

➢ Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. P & T: a peer-reviewed journal for formulary management, 40(4), 277.

➢ Vernikos, G., Medini, D., & Riley, D. R. (2015). Bacterial pan-genomics: challenges and prospects. Computational and structural biotechnology journal, 13, 202-213.

➢ Vos, M., et al. (2015). Horizontal gene transfer in bacteria: transfer mechanisms and recipient responses. Nature Reviews Microbiology, 13(2), 89-100.

➢ Wu, G. D., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science, 334(6052), 105-108.

➢ Zhang, Hongzhan, Yohe, Taylor, Huang, Lu, Entwistle, Simon, Wu, Peng, Yang, Zhi, Kuo, Roland, & Yin, Yanbin (2018). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research, 46(W1), W95-W101.

➢ Zhang, L., & Cai, X. (2018). Pan-genome analysis links the hereditary variation of Leptospirillum ferriphilum with its evolutionary adaptation. Frontiers in microbiology, 9, 2846.