KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING
FACULTY OF ENGINEERING SCIENCE
FACULTY OF MEDICINE
FACULTY OF SCIENCES

Large-scale genomic characterization of *Bacteroides fragilis* reveals multiple lineages associated with functional distinct traits

Promoters:
Prof. Jeroen Raes
Dr. Rodrigo Bacigalupe
Department of Microbiology, Immunology and Transplantation
Laboratory of Molecular Bacteriology (Rega Institute)

Dissertation presented in
fulfillment of the requirements
for the degree of Master of Science:
Bioinformatics

Joon KLAPS

June 2021

FOREWORD

The writing of a thesis can be difficult and exhausting but it is the people who surround you that make it endurable and memorable. This is why a word of gratitude is in order.

First, I would like to thank Prof. Jeroen Raes for giving me the opportunity to be part of one of his projects and providing me with the necessary facilities.

Furthermore, I would like to express my gratitude towards my parents, brother and girlfriend Elise for proofreading my thesis multiple times and for their endless support.

But most importantly, I would like to thank my supervisor Dr. Rodrigo Bacigalupe for his feedback and guidance in the project and my future career. I am very grateful for the hours and hours of online discussions we had and I feel privileged to have had you as my supervisor.

Thank you!

Joon Klaps

ABSTRACT

Bacteroides fragilis is one of the most abundant Gram-negative bacteria in the human gut microbiota and has been associated with inflammatory bowel disease, colorectal cancer and various other diseases. Interestingly, the nontoxigenic counterpart of *B. fragilis* is increasingly suggested as a probiotic due to its beneficial interactions with the host. We hypothesized that this contradiction could be explained by the existence of diversifying subspecies or the exchange of mobile genetic elements encoding for pathogenic and virulence factors. In this master thesis research, we analysed the genomic diversity of *B. fragilis* aiming to comprehend the exceptional but varying properties of this species, which remain poorly understood. We first compiled all the available genomes from multiple geographic locations and sources into a large-scale dataset of over a thousand sequences. Phylogenetic and population structure analyses revealed numerous lineages, which were supported by pangenomic analyses, suggesting a common diversification of the core and accessory genomes, and reinforcing the concept of subspecies in *B. fragilis*. In total, we identified 18 distinct lineages, 15 of them were significantly associated with multidrug resistance encoding genes, and several others with distinct carbohydrate metabolism genes. Additionally, multiple mobile elements were detected containing virulence factors and stress resistance genes. In summary, our study confirms the widely underestimated genetic diversity of *Bacteroides fragilis*.

LIST OF ABBREVIATIONS

Abbreviation	Meaning
ANI	Average Nucleotide Identity
B.	*Bacteroides*
B1-2	*Bacteroides* 1-2 (enterotype)
bft	*Bacteroides fragilis* toxin
BLAST	Basic local alignment search tool
bps	Basepairs
COG(s)	Cluster of Orthologous Group(s)
ETBF	Enterotoxigenic *Bacteroides fragilis*
GWAS	Genome-Wide Association Study
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HGT	Horizontal Gene Transfer
IBD	Inflammatory Bowel Disease
MAG(s)	Metagenome-Assembled Genome(s)
N_e	Effective population sizes
NTBF	Nontoxigenic *Bacteroides fragilis*
P.	*Prevotella*
PCoA	Principal Coordinate Analysis
PS (A-C)	Polysaccharide (A-C)
R.	*Ruminococcaceae*
RBCs	Red Blood Cells
RCC	Raeslab Cultrure Collection
ROC	Receiver Operating Characteristic
s	Gene selection coefficients
spp.	Species Pluralis (multiple species)
SRA	Sequence Read Archive
SRR	SRA run accession
UHGG	Unified Human Gastrointestinal Genome
UMAP	Uniform Manifold Approximation and Projection
vWF	Von Willebrand factor

LIST OF TABLES AND FIGURES

TABLE OF CONTENTS

CONTEXT AND AIMS

Human bodies are colonized by trillions of microorganisms that play a critical role in our physiology, with the gut having the largest abundance of bacteria as well as being the most diverse in many different species[1]. The massive implementation of high-throughput sequencing technologies on the gut ecosystem, and the obtention of metagenome-assembled genomes (MAGs) has transformed the current understanding of the gut microbiome composition. However, there is still a large knowledge gap in our understanding of the causal factors of a healthy microbiome, including the genes, functions and species involved. An important species of the gut flora is *Bacteroides fragilis*, a common pathobiont that is regularly found in the gut of humans but also in the gut of other animals. Interestingly, *B. fragilis* has been associated with inflammatory bowel disease (IBD), colitis, intra-abdominal abscesses, colon cancer, and several other diseases[2]. However, some *B. fragilis* strains have also been proposed as a probiotic for intra-abdominal abscesses, multiple sclerosis, asthma, IBD, and autism, due to their beneficial interactions with the host[3]. A handful of metabolic subproducts that induce the aforementioned traits have been identified in *B. fragilis*. Notably, *B. fragilis* strains display a considerable variation in the production and potency of the metabolics. Studies showed that this variation is induced by mobile genetic elements as well as the considerable genetic diversity of *B. fragilis* strains[4,5]. However, the global diversity and population structure of *B. fragilis* has yet to be addressed.

This thesis aims to characterize the genomic diversity of *B. fragilis* to better comprehend its opposing traits through a large-scale study analysing the population structure of over 1,000 publicly available *B. fragilis* genomes. We identify which lineages enclosed by the population of *B. fragilis* are associated with specific functional traits or mobile elements. The detection of coinciding genes, as well as recombinant regions, are necessary to identify mobile elements but also the influence of mobile elements and recombination on the evolution of *B. fragilis*.

1 BACKGROUND

1.1 *Bacteroides fragilis*, an emerging multidrug-resistant pathogen or a next-generation probiotic?

The phylum *Bacteroidetes* make ~30% of a normal gut microbiota, with the genus *Bacteroides* being the most predominant anaerobe[6,7]. Members of the *Bacteroides* genus are anaerobic, bile-resistant, non-spore-forming, and gram-negative rod bacteria[7]. *Bacteroides* species are commonly isolated from the gastrointestinal tract of animals and humans, but can also occur outside of the gut[2]. However, when they do leave the gut through a rupture in the gastrointestinal tract or by surgery, they frequently become highly pathogenic and cause abscess formation as well as bacteremia[7,8]. Of all the *Bacteroides spp., Bacteroides fragilis* is recognized as the most pathogenic due to its virulence factors and has received increasing attention for its growing resistance to antimicrobials, especially against carbapenems and metronidazole, two commonly used antibiotics against severe infections[9,10]. Notably, *B. fragilis* has also been proposed as a probiotic (for multiple sclerose[11,12], IBD[13–15], asthma[16], and autism[17]) due to its beneficial immunomodulatory effects on the host[3].

The controversy around *B. fragilis'* influence on the host is partially explained by the two known phenotypic subtypes of the species. Some strains produce a toxin, called the *Bacteroides fragilis* toxin (*bft*), and are known as enterotoxigenic strains (ETBF)[18]. Conversely, the nontoxigenic *Bacteroides fragilis* (NTBF) strains lack this toxin encoding gene[3,18]. The toxin cleaves E-cadherin, type IV collagen, fibrinogen and a variety of other proteins[19,20]. It is encoded in a pathogenicity island of a conjugative transposon that also encodes several other proteins -including the pathogenic metalloprotease II- and can be transferred from an ETBF to an NTBF through horizontal gene transfer (HGT)[21,22]. Three homologous isoforms of *bft* are known, *bft*1 the most common one (two-thirds of the isolates)[23], *bft*2 which has been found in 25% of the ETBF strains[24], and *bft*3 which is mostly present in Southeast Asia (10% of ETBF)[25]. Studies analysing the pathogenicity of ETBF strains demonstrated that ETBF strains can induce diarrhea[26–28], colitis[29–31], colon cancer[32–34] and Alzheimer's disease[35]. Notably, most of these studies are based on germ-free mice inoculated with a handful of well-characterised ETBF strains (high-*bft*-expressing strain I-1345, ETBF strain 86-

54432-2) or even on the *bft* protein itself. Yet, multiple diarrhoea studies[26–28] and a study on 13,096 patients from Kwong *et al.* (2018) demonstrated an increased risk of colon cancer in patients with bacteremia from *B. fragilis*[34]. Hence, previous pathogenicity studies of ETBF based on a small number of strains should be interpreted carefully since they are missing the overall genetic and phenotypic divergence of *B. fragilis* strains.

Similarly, multiple studies based on a small number of NTBF strains (*B. fragilis* NCTC 9343, *B. fragilis* ZY-312) have shown that NTBF strains inhibit inflammation in the intestinal tract, lungs, peritoneum and brain[11–16,36]. Additionally, evidence has been found that these strains inhibit the infection of other invasive pathogenic bacteria as well as the suppression of colorectal cancer pathogenesis[37–39]. An important component responsible for *B. fragilis* immunomodulatory effect is the zwitterionic capsular polysaccharide A (PSA), one of the 9 capsular polysaccharides (A-H) in *B. fragilis* [2,36,40]. PSA stimulates the maturation of dendritic cells in the gut which in turn facilitates the proliferation of CD4+ T cells, triggering a cascade of immune responses resulting in the aforementioned benefits as a probiotic[15,36,40,41]. Of note, a recent study suggested that mice colonized with PSA-competent and PSA-deficient NTBF strains, both showed prophylaxis against colitis-inducing ETBF strains[42]. These examples show that the molecular mechanisms used by *Bacteroides fragilis* to influence its host remain largely unknown.

However, the capsular polysaccharide also has a downside; when *B. fragilis* leaves the gastrointestinal tract, the polysaccharide can cause inflammation and ultimately abscess formation[43]. Additionally, PSA, PSB, and PSC hemagglutinate red blood cells (RBCs) and by interfering with either PSB or PSC, the overall capacity of *B. fragilis* to hemagglutinate decreases considerably[44]. It is a well-known strategy of pathogenic bacteria to target blood antigens to adhere to the host and to utilize host glycans as a nutrient source[45,46]. Notably, the host's blood group influences the susceptibility towards pathogens, including *B. fragilis* where the agglutination of type A and type B RBCs is significantly reduced in comparison to the type O RBCs[44,47]. In addition, a recent genome-wide association study (GWAS) analyzing the microbiome of individuals revealed significant associations between the histo-blood group types with single bacteria, like *B. fragilis,* and the overall microbiome composition, implying an impact of the blood group on the microbiome composition[48].

1.2 *Bacteroides fragilis* is a key member of the human microbiome

Multiple studies of the human gastrointestinal tract have shown considerable inter-and intra-individual species variation[49–53]. However, a substantial fraction of this diversity tends to overlap between humans and is generally stratified into subgroups and are referred to as enterotypes[54,55]. These enterotypes are labelled according to their principal taxonomic identifiers *Prevotella* (P), *Ruminococcaceae* (R), *Bacteroides1* (B1), and *Bacteroides2* (B2)[55]. Moreover, the B2 enterotype distinguishes itself from the B1 enterotype by its decreased relative abundance of the *Faecalibacterium*[54]. While *Bacteroides fragilis* is present in all enterotypes, it is the most abundant *Bacteroides* species in the B2 enterotype (Figure 1.1)[56]. The B2 enterotype is a community composition that contains a reduced microbial load in comparison to the other enterotypes and has increased prevalence among patients with Crohn's disease, depression, obesity, and loose stools[54,56–58]. For these reasons, it was hypothesized that the *Bacteroides2* enterotype could be of a dysbiotic nature[54,58].

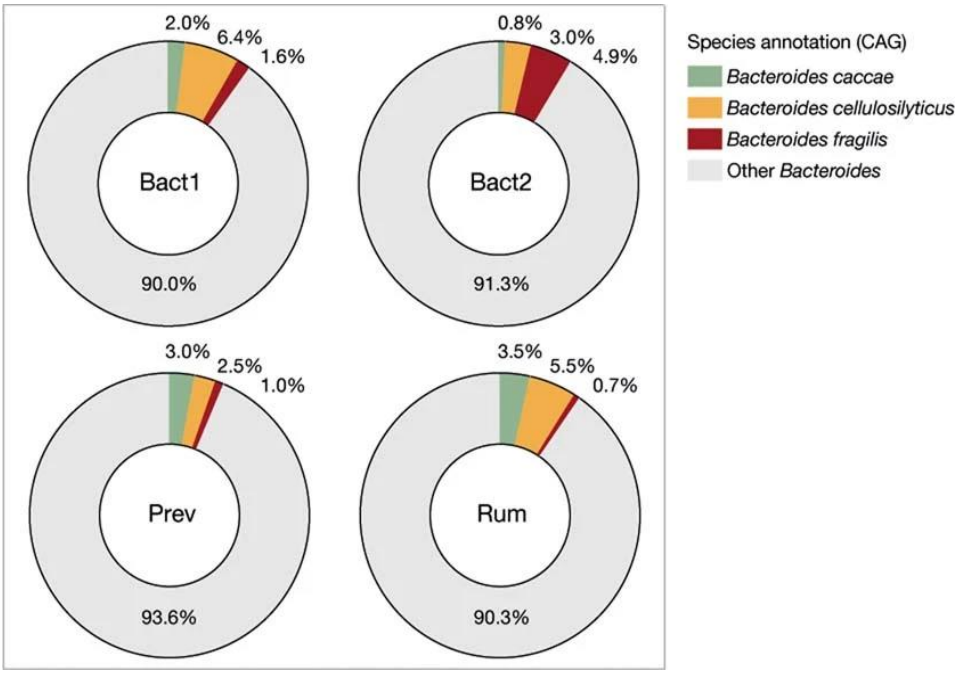


Figure 1.1 *Bacteroides* species fraction in the different enterotypes for a non-statin-medicated study group (Source: Vieira-Silva, et al.[57])

1.3 The evolution and genetic variation of *B. fragilis*

The influence of the phenotypic strain diversity of *B. fragilis* (ETBF vs NTBF) in the microbiome on host health is well recognized[5]. Notably, besides phenotypic clusters, two distinct genotypic clusters (labelled *cfiA*-positive and *cfiA*-negative) have also been identified at a genetic distance of approximately 0-70 by multilocus enzyme electrophoresis and by matrix-assisted laser desorption ionization-time of flight mass spectrometry[59,60]. Both clusters were not associated with the *bft* gene, diseases or geographical origin. A more recent, high-resolution whole genome sequence analysis of *B. fragilis* strain diversity revealed substantial de novo nucleotide and mobile element diversity in *B. fragilis* populations[4]. Additionally, multiple genes displayed parallel evolution within individuals suggesting that natural selection plays a considerable part in shaping the intra-individual *B. fragilis* populations[4].

The within-species genomic variation introduced through mutation and gene flow is shaped by genetic drift and natural selection. Here, natural selection preserves or eliminates specific mutations depending on the induced fitness advantage or disadvantage, while genetic drift eliminates mutations randomly, regardless of the influence on the organism's fitness[61]. Biotic and abiotic factors drive the natural selection of a species and therefore determines the fate of the composition of microbial communities at the species and within species level[62]. By comparing genomic similarities, a structured population can be observed containing distinct clusters or subpopulations through a combination of soft selective sweeps (Figure 1.2), genetic drift, as well as dispersion into new or akin ecological niches[62]. Additionally, if the within-species recombination rate drops and the rate of mutation remains high, these subpopulations can become more internally cohesive relative to one another and may result in the establishment of subspecies[63]. Subspeciation can be accelerated by blocking the gene flow between subspeciating groups when physical or geographical barriers are present, making the subspeciation strictly dependent on natural selection or drift[64]. Without spatial separation (sympatric populations), subspeciation can still occur but there is likely a selective advantage necessary for specialization[65]. However, the extreme distribution of bacteria and archaea make the complete prevention of HGT rare and an in-between scenario more plausible[62]. The presence of HGT through transformation, conjugation and transduction in addition to mutation and rapid multiplication can result in bacteria with very flexible genomes, *i.e.* a lot of strain-specific

genomic variation[66]. In order to describe this variation, the terms “core” and “accessory” genome were introduced[67].

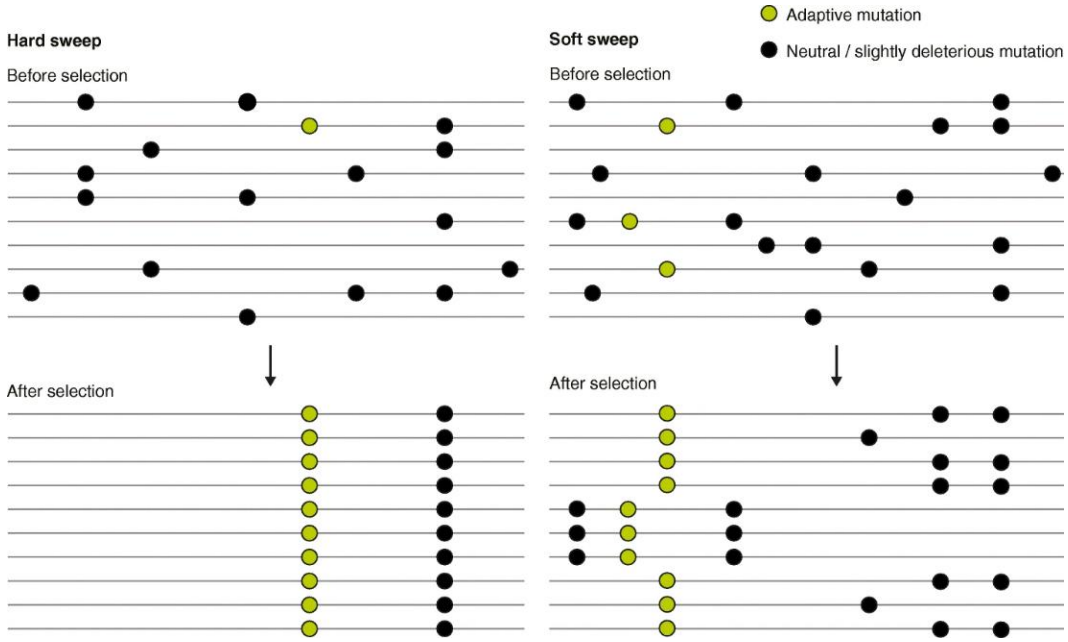


Figure 1.2: Schematic overview of hard selective sweeps (left panel) and soft selective sweeps (right panel). Each row depicts the haploid genome of a specific individual where green dots represent adaptive mutations and neutral or slightly deleterious mutations are black dots. A hard selective sweep results in a single beneficial allele at a specific locus and replaces other alleles in the population. A soft selective sweep replaces genetic variation of the population where multiple beneficial alleles of a locus gain prevalence. (Source: McCoy and Akey[68]).

The core genome represents the genes present in all members of one species and is suggested to contain essential gene families. On the contrary, the accessory genome refers to the genes that are not shared across all individuals also considered dispensable genes[69]. The pangenome refers to the combination of the core and accessory genomes and consists of all genes that can be found in a species[70]. Therefore, in the pangenome most of the genes are either very common or exceptional, resulting in the characteristic asymmetric U-shaped frequency distribution of genes in bacterial populations[71–73]. Two distinct types of pangenomes have been observed: (I) open pangenomes, in prokaryotic species that have an extensive accessory genome, and (II) closed pangenomes, in those that have limited gene diversity (Figure 1.3)[70,74]. The pangenome size is strongly correlated with the environment and niche of the bacteria. For instance, bacteria with sympatric lifestyles that are in contact with other organisms are much more likely to have an open pangenome. Whereas, allopatric

species, which live in isolated habitats, tend to have a closed pangenome with a small accessory genome. This is because the accessory genes are transmitted through HGT and if the environment includes multiple other organisms, the gene transfer will be facilitated resulting in a larger and more open pangenome[75]. It has been suggested that the evolution of *B. fragilis* strains within the microbiome does not depend on a linear diversification process but rather through HGT from a mixture of related organisms[5], resulting in a more open pangenome.

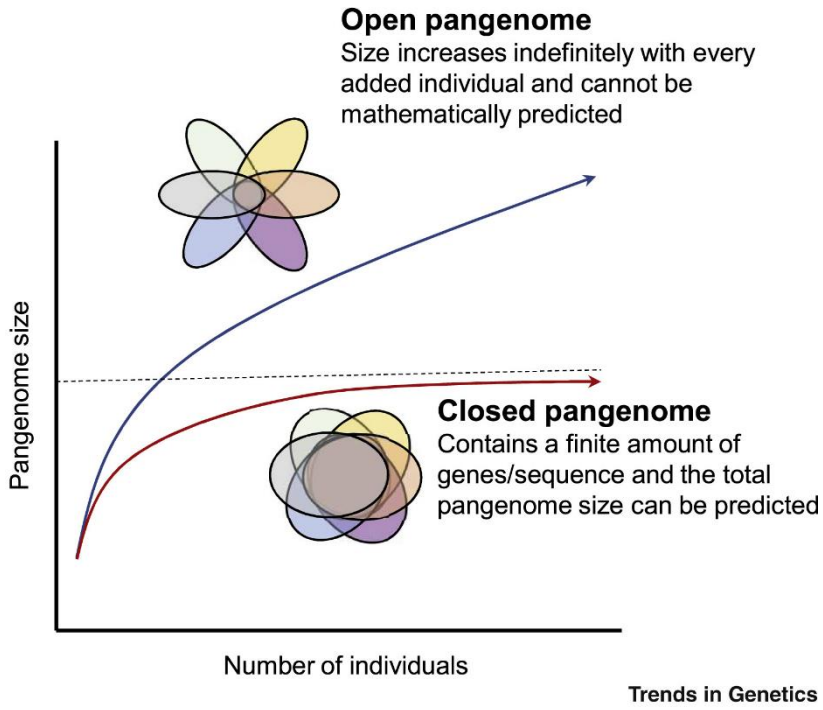


Figure 1.3: Pangenome size as a function of the number of individuals observed

(Source: Golicz et al.[70]).

Identifying core genes essential to the survival of the bacteria provide exceptional targets for antibiotic drug and vaccine development[76,77], while the analysis of accessory genes may reveal those responsible for an adaption to dynamic niches, pathogenicity, antibiotic resistance, or colonization of a new host[78,79]. Notably, there is still a lot of debate whether the nature of the accessory genome is adaptive, neutral, or deleterious[80–82]. A study from Sela *et al.,* suggested the accessory genome was of an adaptive nature by showing that only good fits to the empirical distribution could be found when the predicted genome size model was based on theoretical models including a positive mean fitness distribution[81]. Oppositely, a model driven by effective population size (N_e; Drift-Barrier model, Figure 1.4) has been suggested which indicate

that accessory genes with low selection coefficients (*s*) are close to neutral when drift dominates over selection (*i.e.* $s << 1/N_e$), resulting in the loss of genes with little adaptive values[83,84]. Hence, the latter model suggests a pangenome driven by neutral evolution and is therefore strongly dependent on population size but also on the population size of co-occurring prokaryotic populations due to HGT[82].

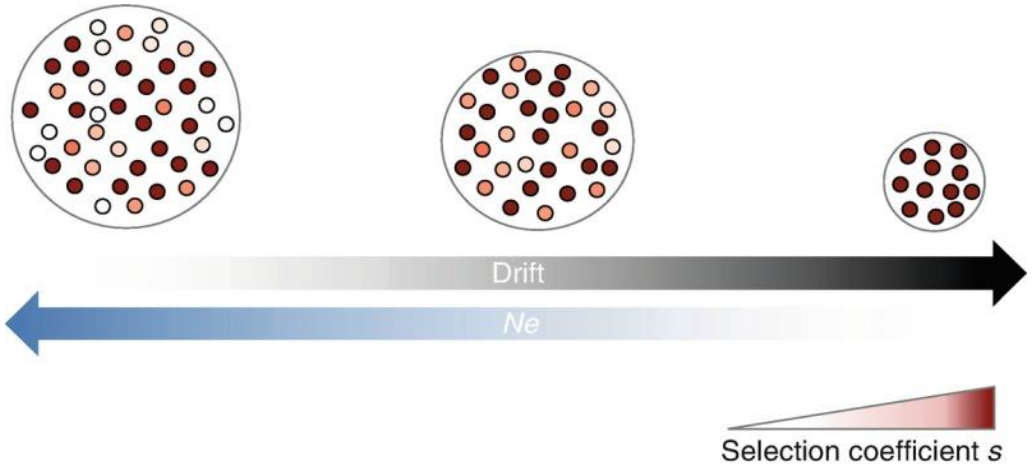


Figure 1.4: The Drift-Barrier model as one of the two models used to explain the nature of the accessory genome. Each encapsulating circle represents a pangenome and the smaller circles represent individual genes. The colour gradient reflects the selective coefficient of genes (s), *i.e.* the beneficial value of a gene. Species with larger effective population size (N_e) are less driven by drift and can therefore retain genes with smaller selective coefficients (left). In a pangenome with a small population size, the genetic drift is strong and only genes with high beneficial value will be retained (right; Source: Bobay[85]).

In this study, we obtained insights into the evolution, population structure and pangenome characteristics of *B. fragilis* by analysing the genetic diversity of over 1,000 genomes of the species. We identified numerous distinct lineages which were tested for association with genes and geographical origin. Coinciding gene groups and recombination events were determined to better comprehend the evolution of *B. fragilis*. Finally, we discuss in more detail several mobile elements including the gene group containing the *bft* encoding gene.

2 MATERIAL AND METHODS

2.1 Genome collection

All publicly available *Bacteroides fragilis* genomes were downloaded (September 2020) to create a large-scale genomic dataset of *B. fragilis* with its encompassing diversity. The retrieved sequenced genomes include isolates and MAGs that were recovered from all over the world as well as multiple genomes originating from single hosts. In addition, our data set contains isolates from 1955 to 2019, but most (78%) were isolated after 2008. The organized data collections that were used to download sequences from are the Unified Human Gastrointestinal Genome (UHGG) collection[86], the NCBI database[87], a recent evolutionary study of *B. fragilis* in the gut also referred to as the '*Zhao study set'* in this thesis[4], animal genomes[88], and the Raeslab culture collection (RCC). For the Zhao study set and the RCC samples, we performed genome assemblies. Raw reads were trimmed using Trimmomatic v0.36[89] (default settings) and assembled using SPAdes v3.13.0[90] (default settings).

The corresponding metadata of the genomes was acquired with E-utilities by either a direct search of the biosample or by querying for the Sequence Read Archive (SRA) or BioProject identifier and linking it to BioSample.

2.2 Quality control and filtering of genomes

Duplicated genomes were detected through the comparison of Sequence Read Archive run accession (SRR) identifiers. A total of 39 duplicates were observed and removed from the data set. Assembled genome statistics and quality summary statistics were calculated with QUAST v5.0.2[91] (default settings) and CheckM v1.1.3[92] (lineage_wf option, default settings) respectively. We removed genomes with contamination higher than 5% (43 genomes), completeness below 90% (5 genomes), smaller than 4,250 kbps or larger than 5,750 kbps (12 genomes), having more than 600 contigs (17 genomes), or more than 100 N's per 100 kbp (17 genomes).

2.3 Selection of the genotypic subtype and non-redundant assemblies

Distinct species within the remaining dataset were detected by estimating the pairwise Average Nucleotide Identity (ANI) of all genomes with FastANI[93] v1.1 and were analysed with the bactaxR[94] v0.1.0 package in R. Here 60 genomes formed a distinct cluster with less than 88% ANI from all other genomes and were therefore also discarded (Supplementary figure 1). An overview of the remaining genomes is given in Figure 2.1.

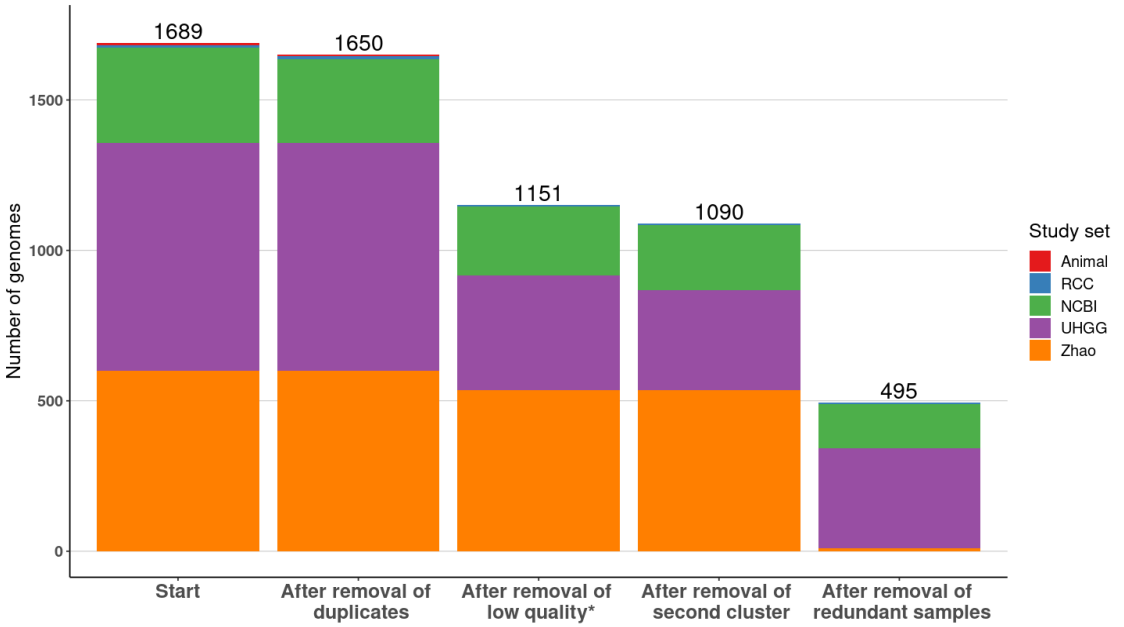


Figure 2.1: Overview of the number of genomes used in this study at each data filter step. *Low-quality genomes were defined as genomes with contamination higher than five, completeness below 90, smaller than 4,250 kbps or larger than 5,750 kbps, more than 600 contigs, more than 100 N's per 100 kbp (0.1%)

A preliminary exploration of the pangenome and population structure was done with Pangenome Analysis Toolkit (PATO)[95] v1.0.2, a community-driven R package for analysing thousands of genomes using conventional computers based on MASH[96], MMSeqs2[97], and Minimap2[98]. To obtain a preliminary overview of the population structure that guided our further analysis, PATO was used because it provides numerous ordination methods for the gene absence-presence matrix as well as tools for inferring the phylogenetic trees from nucleotide and protein alignments. We used Panaroo[99] v1.2.3 for pangenome and population structure analysis with parameters --

clean-mode moderate (genes present in at least 1% of the genomes), *--remove-invalid-genes* (removal of invalid genes), *--core_threshold 0.95* (core gene threshold of 95%), *-c 0.95* (sequence identity threshold of 95%), and *--len_dif_percent 0.95* (length difference cutoff of 95%). For epidemic and overrepresented clusters (*i.e.* clusters with multiple assemblies originating from the same location and host, or very high ANI values), which add redundancy to the analysis, only representative genomes were selected (Supplementary figure 2). This resulted in the final dataset of 495 non-redundant, high-quality genomes.

2.4 Phylogenetic and population structure analyses

The remaining 495 non-redundant genomes were analysed again with Panaroo[99] v1.2.3 using the same settings as specified above, resulting in a gene absence-presence matrix and a core genome alignment. Pangenome sequences retrieved by Panaroo were annotated with eggnog-mapper[40] v2.1.2 using eggnogDB[101] v5.0.2 as a reference database. Acquired antimicrobial or virulence genes were searched using Abricate v1.0.1 (https://github.com/tseemann/Abricate), against the NCBI AMRFinderPlus[102] v3.10 database.

The core genome alignment from Panaroo was used to construct an approximately-maximum-likelihood phylogenetic tree with FastTree v.2.1.0. Next, we used the R package fastbaps[103] v1.0.5, a robust clustering technique based on hierarchical Bayesian clustering with a prior Dirichlet distribution to optimise symmetry was used on the recombinant accommodated genome alignment from Gubbins[104] v2.3.4 (see also recombinant analysis) to identify distinct lineages. The absence-presence matrix of the accessory genome was plotted by an Uniform Manifold Approximation and Projection (UMAP) as well as through calculating the Jaccard distances of the matrix and performing a Principal Coordinate Analysis (PCoA) on these distances. Based on the UMAP coordinates, a Hierarchical Density-Based Spatial Clustering of Applications with Noise[105] (HDBSCAN) was applied to determine accessory based clusters.

2.5 Recombination analysis

To identify recombinant regions, a whole genome alignment was created using Snippy[106] v4.6.0 with the highest quality genome as a reference (GCA_000210835.1). Next, the whole genome alignment was used as an input for Gubbins[104] v2.3.4 (fasttree as tree builder and other default options) which iteratively identified regions with high densities of nucleotide substitutions. The output of Gubbins was visualised in Phandango[107] v1.3.0. Here, high-density areas were marked as recombinant sites (top 10% loci) using bedtools[108] and R.

2.6 Genome-wide association analysis

We identified genes associated with fastbaps and HDBSCAN clusters, sample continents, sample study sets, and genome types using Scoary[109] v1.6.16. Scoary performs numerous Fisher exact tests and accommodates for population structure through the provided phylogenetic tree, settings were kept to their default value. By using the output of Scoary, we can quantify the overlap of the core and accessory genome clusters by counting the number of shared genes. This gives an indication which lineages correspond to which accessory genome clusters.

An enrichment analysis using a hypergeometric distribution was performed to compare clusters of orthologous groups (COGs) functional categories of core and accessory genes within each lineage and recombinant genes versus all genes using the phyper function of the R package stats v3.6.3. COG functional categories of both the accessory and core genome of each lineage were compared to the combined accessory and core of all other lineages through a Fisher exact test.

Genes that tend to co-occur with other genes were detected with Coinfinder[110] v1.0.2. Coinfinder provides reliable results for observing coinciding genes utilizing a Bonferroni-corrected binomial exact test. In order to be able to link gene clusters with specific lineages, we first had to determine thresholds for (I) when an assembly contains a gene cluster and (II) when a lineage contains a gene cluster. We considered that assemblies having at least 80% of the genes from the corresponding Coinfinder groups contained the gene cluster (presence, I). A lineage contained a gene cluster when it had at least a single assembly with that gene cluster (II). We identified significantly associated Coinfinder groups with continents based on a chi-squared test.

We defined gene clusters as mobile genetic elements based on three properties of the cluster. (I) Gene clusters that are present in all lineages are most likely not mobile elements, hence, the cluster needs to be present in at least one lineage but not in all lineages. (II) The gene cluster needed to be of the right size where it had to contain at least 3 genes but no more than 1,000 genes. (III) Lastly, if the gene cluster was dispersed along the prokaryotic genome, it is likely not to be a mobile element as well. We were able to map the dispersion of a gene cluster by analyzing the locus tag of genes in the assemblies and taking the standard deviation of the locus tags. If the previous two properties hold and the standard deviation was smaller than 50 the gene cluster is marked as a mobile genetic element. Finally, three gene clusters were visualised using SnapGene v5.2.5 (https://www.snapgene.com/) and ggtree[111].

3 RESULTS AND DISCUSSION

3.1 Compiling a diverse genome collection of *B. fragilis*

Sequences of 1,689 *Bacteroides fragilis* genomes, primarily corresponding with bacteria from human faecal samples, were retrieved from various sources (methods). After quality filtering, the removal of duplicates, the distinct genotypic subtype based on ANI, and redundant samples, 495 high-quality genomes remained. Of these, 40.40% were obtained from pure isolates, and 59.60% were assembled from metagenomes (MAGs). Genomes originated from all continents, but most of them corresponded with samples from Europe (163 genomes, 68.26%) and were primarily associated with humans (N=439, 88.69%). A large fraction of the genomes were sampled from faecal material (N=304, 61.14%) with a few samples being from appendix tissue, blood or pus (Figure 3.1).

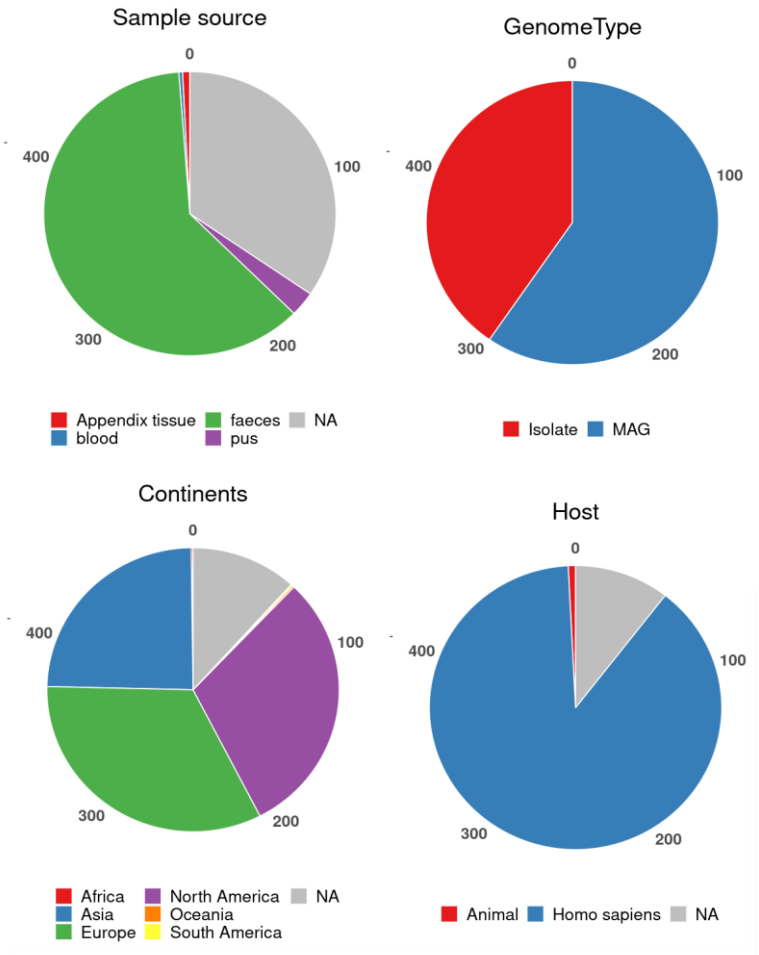


Figure 3.1: Pie charts of genome metadata including the continent of a sample, sample host, sample source, and genome type.

The median genome size was 5.2Mbp (4.3Mbp to 5.8Mbp) and consisted of 48 contigs (range 1-202 contigs ≥1 kbp). The GC content ranged from 42.44% to 44.36% with a median of 43.29%. Genomes from isolates were significantly larger than MAGs (Wilcoxon-test, p-value<2.2e-16; Figure 3.2), suggesting that MAGs, despite being considered high-quality assemblies (completeness>90, contamination<5), may lack numerous genes.

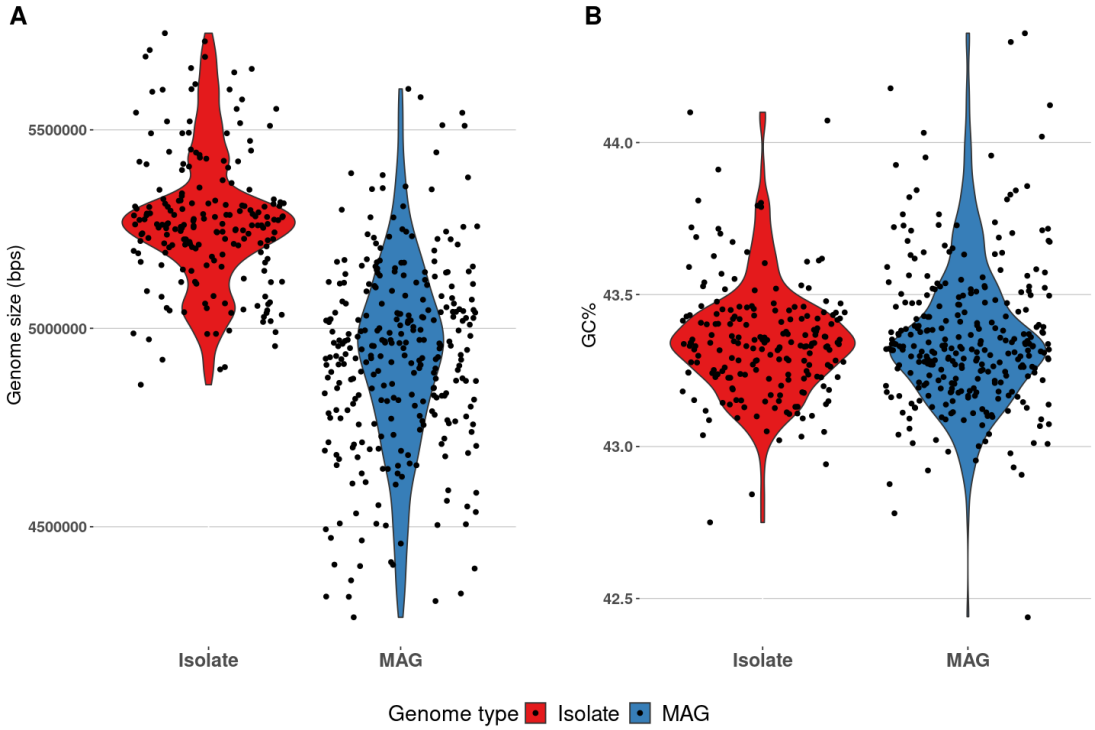


Figure 3.2: Jitter-violin plot of isolate and MAGs assembly sizes (A) and GC% (B). MAGs were significantly smaller than isolates (Wilcoxon-test, p-value<2.2e-16).

3.2 The unrevealed lineages of *B. fragilis*

To identify if any distinct species were included in the dataset, a pairwise ANI dendrogram was created (Supplementary figure 1). In here, two distinct clusters showing approximately 88% whole genome ANI were identified, representing the previously identified genotypic subtypes of *B. fragilis* (*cfiA* + and *cfiA* -)[59,60]. Notably, all assemblies had a relatively high 16S similarity to the *B. fragilis* type strain NCTC:9343 (ANI > 97.84%; RefSeq NR_074784.2). Given that the average nucleotide identity is below the empirically defined species ANI threshold of 95%[112], it is most likely that these two clusters represent two distinct species. For this reason, only one cluster was

considered for further analysis which was the cluster containing the type strain of *B. fragilis*. Fortunately, this was also the genotypic subtype with the largest dataset of assemblies. A BLASTp search with *cfiA* (UniProt: P25910) nucleotide sequence as a query, revealed that the genotypic subtype with type strain is the *cfiA*-negative subtype since no significant BLAST hits were returned (all e-values > 2.5). Whereas running the search on the removed cluster, 59 assemblies (of the 61, 96.72%) contained a significant hit (e-value< 1e-50), indicating that the analysis was continued with the *cfiA*-negative subtype.

Furthermore, an approximate maximum-likelihood-tree of the selected genotypic subtype showed several high clonal clusters. These genomes correspond with overrepresented clusters and would therefore introduce bias when estimating the pangenome (Supplementary figure 2). This is because genes or proteins belonging to these overrepresented clusters are more likely to be found significant since the diversity of the dataset is no longer homogeneously distributed. Hence, redundant genomes based on host, ANI and location were removed and a single assembly as a cluster representative was used instead. Notably, almost all redundant genomes were from a single study set (Zhoa et al.[4]). Moreover, multiple assemblies from a specific host (S01) from the Zhoa[4] dataset showed considerable high core genome similarity with multiple other assemblies not from their study but from Genbank instead. A more detailed analysis of the meta- and phylogenetic- data suggested that all these genomes originate from the same host and therefore a single representative was selected. This indicates that humans can have a specific gut bacterial fingerprint.

Once the redundant assemblies were removed, we visualized Panaroo results by projecting the gene absence-presence matrix of the accessory genome using UMAP and PCoA (Supplementary figure 3A and B). Displaying the genomes using an UMAP approach directly on the absence-presence table of accessory genes showed a better separation of clusters and therefore outperformed the PCoA projection. Additionally, the 15 clusters identified by fastbaps based on the core genome alignment coincided with the UMAP projection of the accessory genome and its 23 HDBSCAN clusters. Moreover, the projection also shows that a single core based cluster (labelled origin, baby pink in Supplementary figure 3) overlaps with multiple accessory genome clusters, which suggests that additional clusters based on the core genome could be identified. Therefore, we accounted for any potential effect of recombination, considering that recombinant events are known to obscure population structures[62].

Using the recombination accommodated genome alignment as input for fastbaps revealed now 19 clusters (Figure 3.3). HDBSCAN clusters were kept the same since the absence-presence matrix of the accessory genome remained the same. Figure 3.3 shows that some lineages, like lineage E and K, have a more unique accessory genome due to their isolated position in the UMAP projection. A single accessory based cluster (HDBSCAN) did not distinguish itself nicely because its assemblies are not clearly arranged in distinct groups and was therefore not drawn in Figure 3.3. Notably, most of these assemblies were also assigned to the single core genome cluster (origin) which is scattered in both the UMAP projection as well as in the phylogenetic tree (Figure 3.4) and can therefore not be considered as a distinct lineage. The fastbaps algorithm creates such a large cluster when it is unable to further segregate the core genomes without over-partitioning the data.

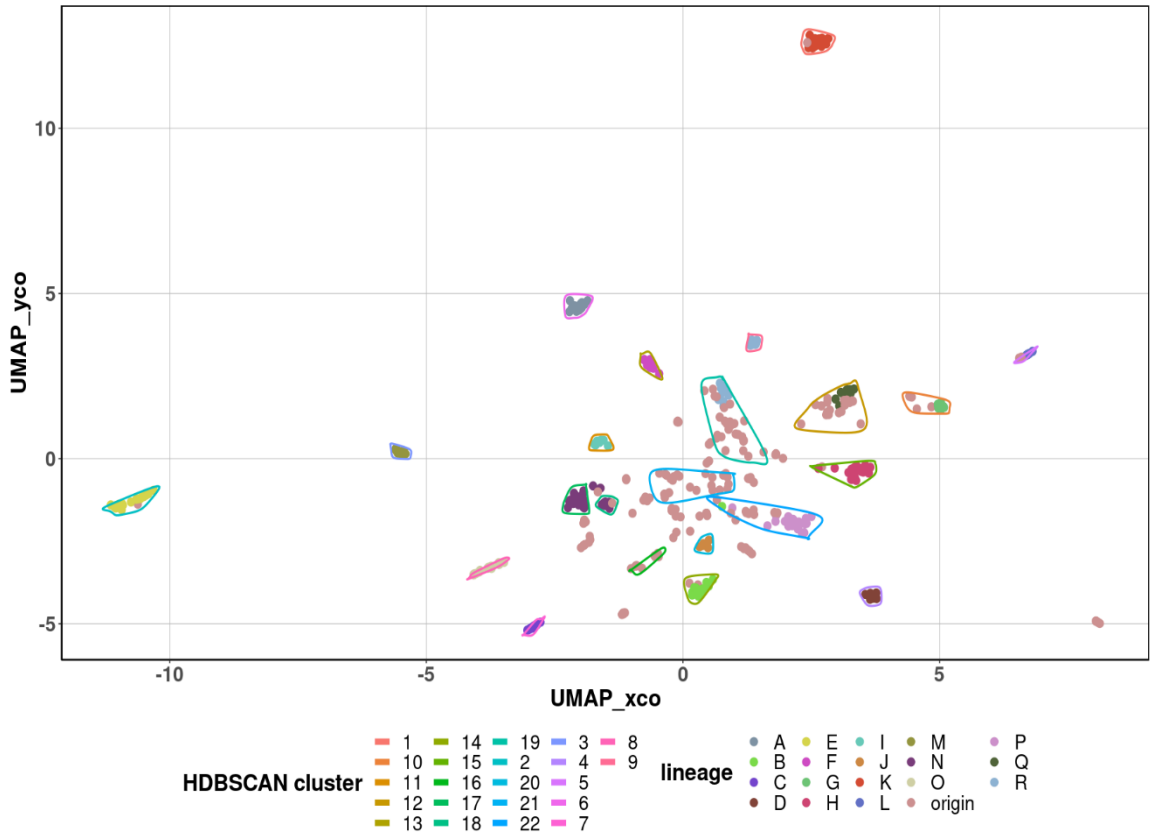


Figure 3.3: UMAP projection of absence-presence matrix of the accessory genome. Assemblies are coloured by lineages based on fastbaps clustering of the recombination accommodated whole genome alignment. Assemblies are encircled by HDBSCAN clusters based on the UMAP projection. HDBSCAN cluster 0 is not drawn.

The iteratively created phylogenetic tree by GUBBINS[104] is based on a genome alignment where only substitutions outside of recombinant regions are considered and are shown in Figure 3.4 with the 18 lineages highlighted. All clusters tend to be strictly monophyletic, except for the large cluster (origin), which was paraphyletic. However, lineage R has two inner subclusters making it the most phylogenetically divergent lineage. These inner two subclusters can also be seen in Figure 3.3. Two lineages are from one single continent and seem to be highly clonal, whereas all others originate from multiple continents. Additionally, no lineages were made up of only MAGs nor were there any clusters with only samples from appendix tissue, blood or pus. Notably, by taking recombinant events into account the genetic diversity of each cluster decreased considerably. In other words, the total heterozygosity is very large in comparison to the genetic diversity within subpopulations which is common in epidemic populations. Therefore, the *B. fragilis* population tends to be highly clonal but due to bottlenecks or selective sweeps, the population emerges as independent lineages[113]. Epidemic population structures are common for pathogenic bacteria, for instance, *Mycobacterium tuberculosis, Bordetella pertussis or Yersinia pestis* all have epidemic population structures[114]*.* However, epidemic population structures are rather unique for species that occur in the gut microbiome as well as the broad geographical distribution of lineages since strain-level genetics in the microbiome were found to be strongly associated with geographically separated host populations[115].

The rate of recombination was approximately 16-fold less than the mutation rate (r/theta 0.06). However, the recombination rate was estimated to contribute 2.4-fold more (r/m 2.43) to the population diversity than mutation did since each recombination event can induce multiple nucleotide changes. The influence of the recombination rate on the genetic diversity and the clonal expansion of *B. fragilis*, suggests that recombination has had a large impact on the evolutionary history of this bacteria and therefore its population structure. Note that the r/m estimation is a lower bound of the actual level of sequence exchange. This is because GUBBINS requires recombination events that import a sufficient level of sequence diversity[104]. Furthermore, the r/m rates deviate over time but also from strain to strain and therefore through the detected lineages. The overall standard deviation of r/m across all strains was 5.51 showing that *B. fragilis* lineages are highly divergent in their recombination rates (Supplementary figure 4). Lineage C and N showed the highest mean r/m of 6.53 and 6.44 (respectively), lineage E and H the lowest with both a r/m of 0.65.

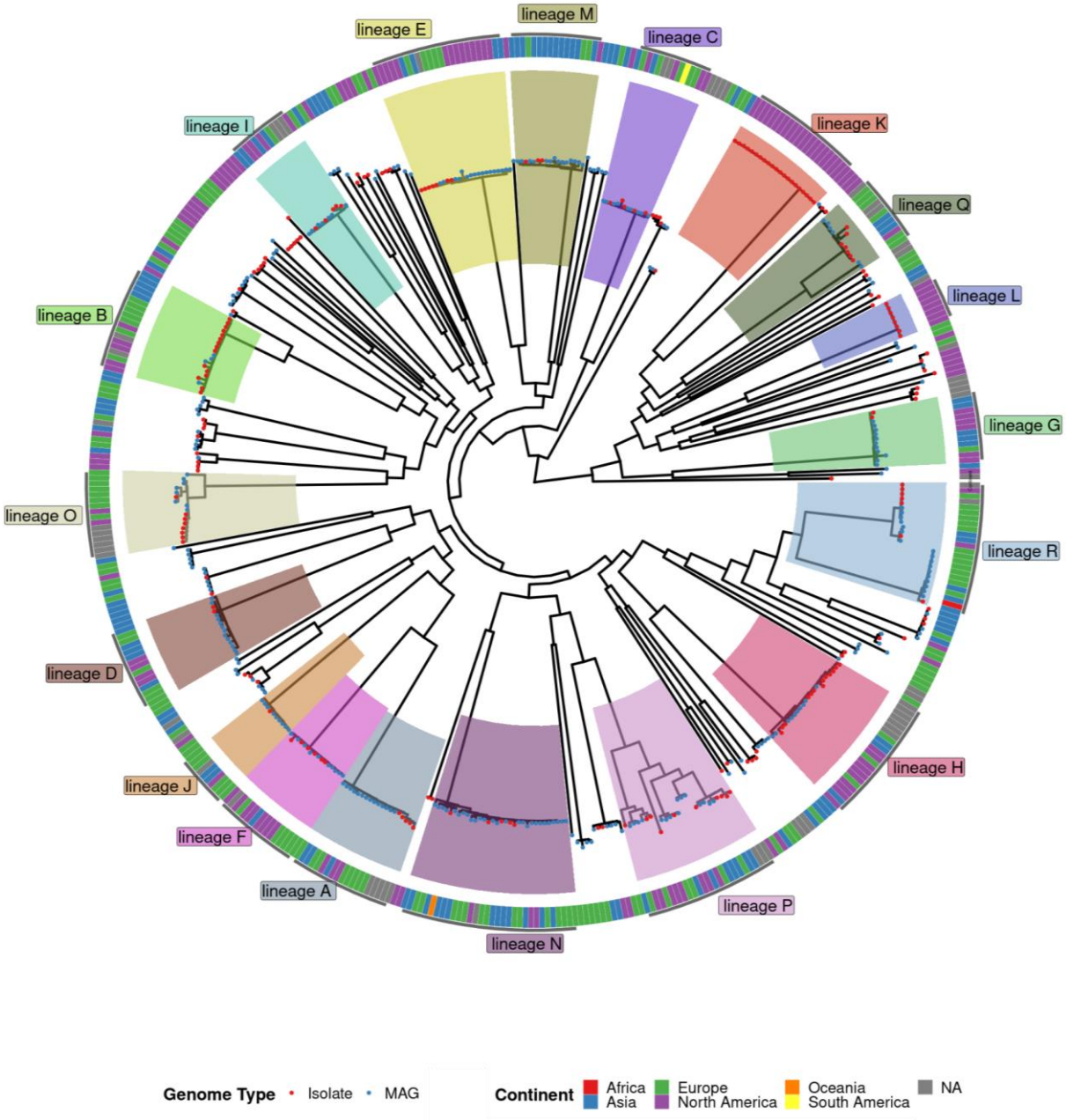


Figure 3.4: An approximate maximum likelihood using a recombination corrected whole genome alignment created by GUBBINS[104] with a general time reversible model. Assemblies belonging to the origin cluster are not highlighted. The dendrogram was annotated using ggtree[111].

3.3 The open pangenome of *B. fragilis* suggests the presence of genes with a small beneficial value

The *Bacteroides fragilis* genome has on average 4,419 genes (median 4,385) where an assembly has a mean of 2,737 core genes and 1,620 accessory genes. However, a total of 20,922 accessory genes were detected (Table 3.1). Meaning that the pangenome has a lot of strain-specific genes and suggesting that it's a rather open pangenome (Figure 3.5A). Previous studies have demonstrated that pangenome and effective population sizes are significantly correlated in a positive fashion[83,116]. Therefore, the open pangenome of *B. fragilis* suggests that *B. fragilis* has a large effective population size and is less subjected to genetic drift according to the Drift-Barrier model, *i.e.* it retains genes of small beneficial value as well. However, given that the exact nature of the pangenome is still heavily debated, one cannot exclude the possibility that only genes with high selective coefficients are retained. Notably, despite the large detected genomic diversity, it is probable that we are still underestimating the pangenome. Approximately 60% of our assemblies used to describe the pangenome were MAGs, and MAGs are not able to incorporate the accumulation of within-species diversity and is, therefore, a big limitation of our study.

Table 3.1: Summarizing overview of all genes with moderate evidence (5% abundance)

Gene type	Strain coverage	Number of genes
Core genes	(99% <= strains <= 100%)	1,577
Soft core genes	(95% <= strains < 99%)	1,255
Shell genes	(15% <= strains < 95%)	2,414
Cloud genes	(5% <= strains < 15%)	18,508

By plotting the occurrence frequency of genes, the characteristic U-shape for bacteria can be observed, where most genes are very rare or either very common (Figure 3.5B). Eggnog was able to find homologs for nearly two-thirds of the pangenome which demonstrates that there is still a huge knowledge gap when it comes to gene annotation for *B. fragilis* or even the gut microbiome. More specifically, Figure 3.5C and D shows that, as expected, a considerably larger amount of accessory genes could not be annotated in comparison to core genes.

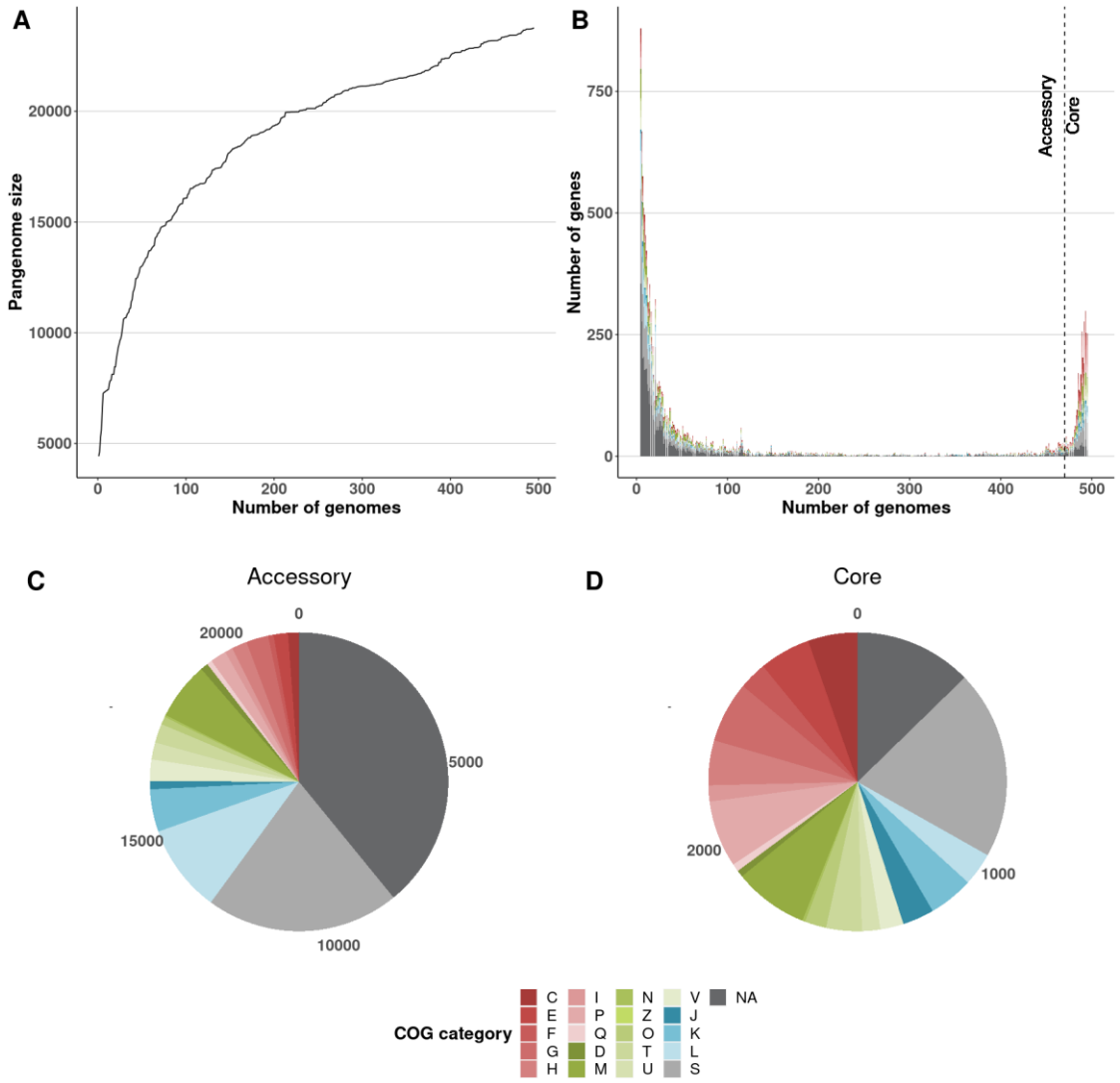


Figure 3.5: Overview of the pangenome characteristics. The pangenome size as a function of the number of genomes (A). The gene occurrence distribution with its typical U-shape, coloured by eggnog's annotated functional COG categories (B). C and D display the number of annotated accessory and core genes, respectively.

3.4 Identification of lineage marker genes

Through a series of fisher exact tests, we detected that all lineages had significantly associated genes (Benjamini & Hochberg corrected p-value <0.05), where values ranged from 268 to 2,500 (median=1,050). To better comprehend how rare some of these genes are, we made a volcano plot as well as a ROC like plot where a gene is represented as the fraction of true positives over the fraction of false positives for each

lineage (Supplementary figures 5 and 6, respectively). Here, lineages that contain a small genetic diversity, like lineage K and L, had much more unique genes with infinite odds ratios in comparison to other lineages with more genetic diversity.

A proportional heatmap (Figure 3.6) representing the overlap of significantly associated genes revealed that most core-based clusters (lineages) correspond with a specific accessory cluster (HDBSCAN), except for lineage R, which is presented by two subclusters (12 and 18). This could also be seen in the UMAP projection (Figure 3.3). The co-evolving core and accessory genome are interesting because it is often assumed that the core genome represents the selectively important differences in gene content[117]. Whereas most *B. fragilis* strains belonging to the identified lineages are likely to share accessory genes, the lineages are not defined by it. Instead, they are characterized by multiple stable genomic islands of core and accessory genes. But note that sequencing data is but a snapshot in time and the identified lineages are part of dynamic communities in the human microbiome. Hence, these lineages diverge and can become subspecies or eventually a new species[62].

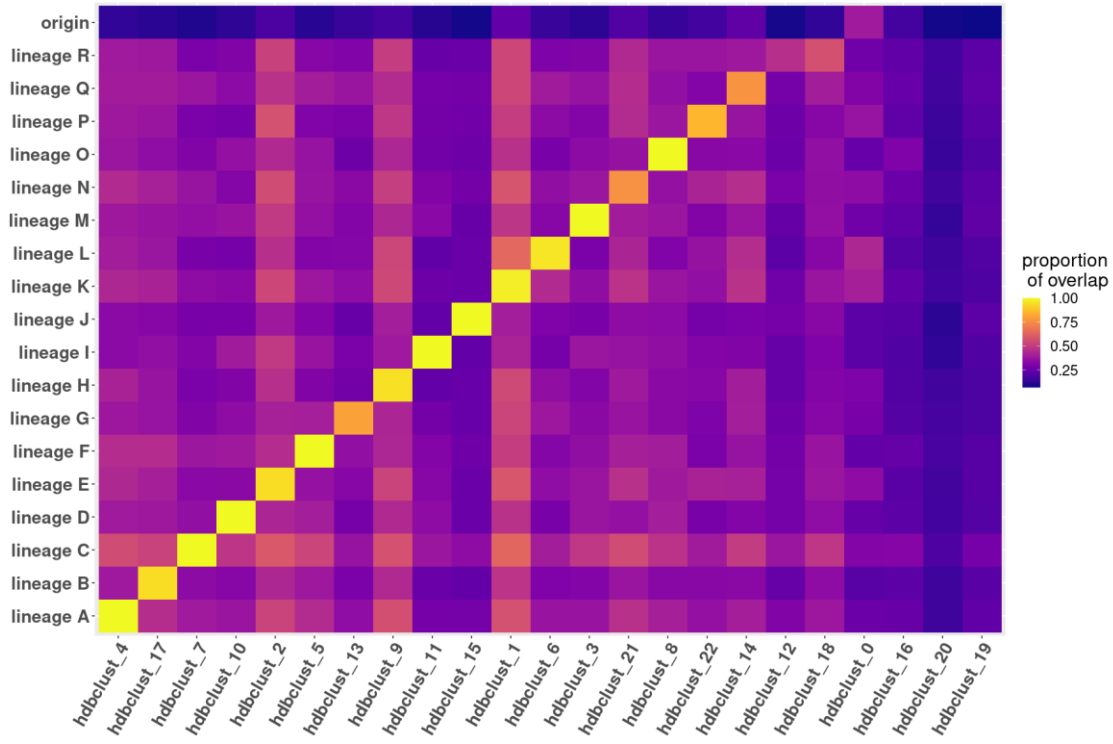


Figure 3.6: Proportional heatmap of overlapping associated genes from core genome clusters (lineages) with accessory genome clusters (HDBSCAN).

Combining the results of Scoary and Abricate allowed us to search for lineage-specific virulence factors. A total of 39 virulent genes in all genomes were detected with Abricate, which is not unexpected, given the numerous antibiotic resistance genes previously detected and the diseases associated with *B. fragilis*[3,7]. Notably, the *bft* genes were not present in the reference database and were therefore not detected. Therefore, it is most likely that there are a lot more virulent genes present in *B. fragilis* aside from the 39 detected genes. From these 39 virulent genes, two of them were found significantly associated with lineages. The aminoglycoside 6-adenylyltransferase (aadK) which mediates bacterial resistance to streptomycin was found to be significant with lineage K (Fisher exact test, p-value=1.77e-25 Benjamini & Hochberg corrected). The other virulent gene is the tetracycline resistance protein (*tetM)*, which abolishes the inhibitory effect of tetracycline, had multiple variants and was significantly associated with lineages E, L, and K (Benjamini & Hochberg corrected p-values 9.09e-5, 7.85e-3, and 5.20e-7, respectively). However, a total of 15 lineages (A, B, C, D, E, F, G, H, K, M, N, O, P, Q and R) were associated with one or more antibiotic resistance genes based on eggnog annotation descriptions (Fisher exact test, Benjamini & Hochberg corrected p-value <0.05). Furthermore, multiple lineages contained specific carbohydrate transferases and/or synthases. For instance, lineage C had various copies of *fabG* and *fabH* genes significantly associated with it. These genes are involved in the pathway fatty acid biosynthesis, which is part of lipid metabolism[118,119].

An enrichment analysis to compare the functional categories of the core and accessory genomes within each lineage showed that all lineages had the COG functional categories K (transcription), L (replication, recombination and repair), and M (cell wall/membrane/envelope biogenesis) significantly enriched in the accessory genome (Figure 3.7A and B). We hypothesized that the enrichment in these three categories is due to the abundant number of recombination genes in the accessory genome and by HGT[120]. Functional category J (Translation, ribosomal structure and biogenesis) was enriched in 11 lineages which will also be most likely due to recombinant regions. Lineage Q and N had secondary metabolites biosynthesis, transport and catabolism (Q) significantly enriched. Additionally, lineage R has significantly more genes in the accessory genome involved in defence mechanisms, which might suggest that members of this lineage have been exposed to environmental stress such as antibiotics, bacteriocins or even the host's immune system. Other enriched functional categories were U (intracellular trafficking, secretion, and vesicular transport) in lineages B, L, and

N as well as functional category D (cell cycle control, cell division, chromosome partitioning) in lineages K, L and D.

No significant differences were found in both the core and the accessory genome when comparing each lineage to all other lineages (Fisher exact test with Benjamini & Hochberg corrected p-value >0.05).

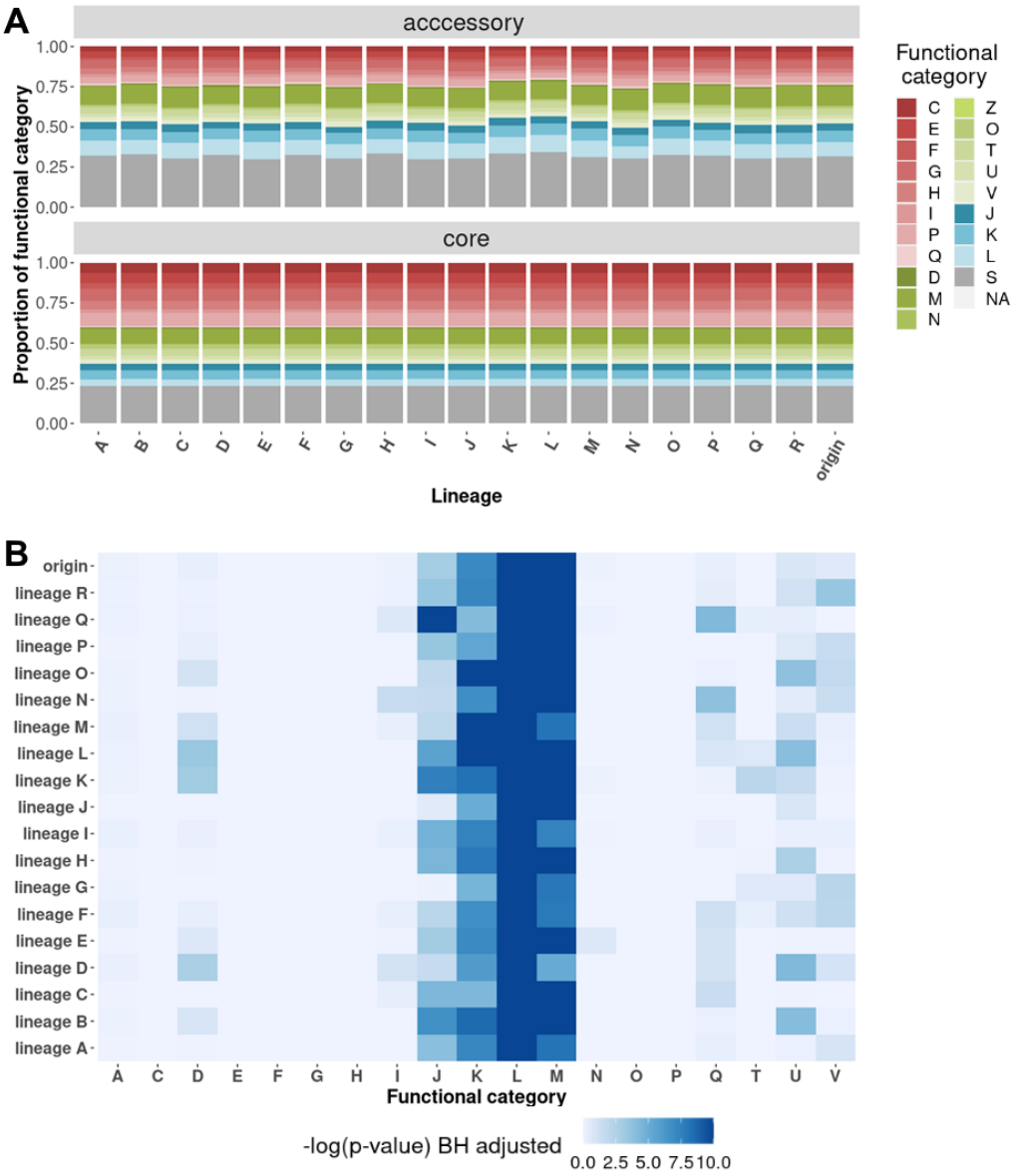


Figure 3.7: Analysis and comparison of the annotated genomes of each lineage. A proportional bar chart for each lineage representing the fraction of COG's functional categories (A). Heatmap result of enrichment analysis for comparing the core genome's COG functional categories with the accessory genome's COG functional categories coloured by the negative logarithm of the Benjamini-Hochberg corrected p-value (B).

3.5 Recombinant regions of *B. fragilis* are largely made up of genes with unknown functions

The frequency of the high SNP density areas calculated by SNIPPY and Gubbins was used to identify highly recombinant regions. Unfortunately, 19 of the 463 recombinant genes could not be annotated. Functional category L (Replication, recombination and repair) and category S (function unknown) were significantly enriched in recombinant regions (hypergeometric test, p-value=1.03e-05,9.65e-8, respectively; Figure 3.8). Furthermore, six *SusC* family proteins (membrane protein) were identified as recombinant regions. Interestingly, multiple *SusC* family proteins were submitted to parallel evolution according to the study from *Zhoa et al.*[4]. Therefore, our results further endorse the previously made hypothesis which suggested that these genes are under pressure to change their interaction with the host's immune system or to circumvent phage infection[4,121]. No other genes submitted to parallel evolution identified by *Zhao* et al.[4] were detected as recombinant regions.

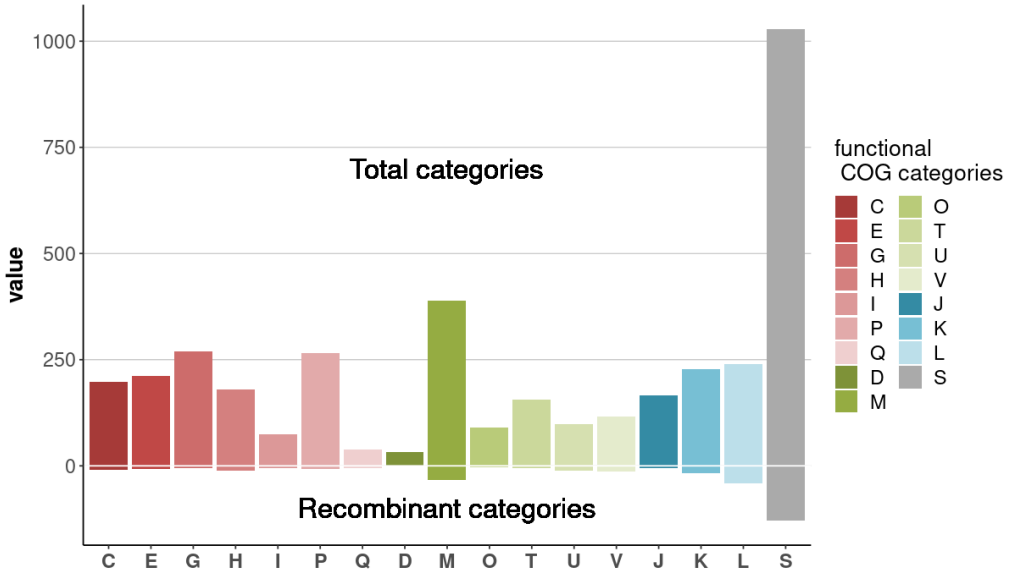


Figure 3.8: Barplot of functional COG categories detected in recombinant regions (bottom) and in all regions (top).

3.6 Identifying lineage-specific genes and mobile genetic elements

Coinfinder identified a total of 114 co-occurring gene clusters. The size of these gene clusters deviated considerably from 2 to 3,763 genes (median 3 genes, second largest 135 genes). An association network with the Fruchterman-Reingold layout of the determined gene clusters is given in Figure 3.9A. This figure illustrates that the biggest gene cluster (orange) contains multiple other subclusters and tends to be overestimated in size. One might guess that this large cluster is mostly made up of core genes, however, according to Coinfinder's developers, this cannot be the case since genes present in all genomes are removed before the analysis[110].

A total of 41 coinciding gene groups were distributed significantly different than their genome's sample origin (chisq-test, p-value <0.05 Benjamini-Hochberg corrected). Notably, four of these gene clusters were also later identified as mobile elements (groups 18, 47, 71 and 82). Groups 47 and 71 were significantly more present in Asia and Europe in comparison to North America. Groups 18 and 82 showed bias towards North America.

By combining Coinfinder gene clusters with the locus tag of the genes for all genomes, we were able to estimate the size of these clusters and filter for gene clusters containing genes that are not dispersed across the genomes. The remaining 45 coinciding gene clusters were considered as potential mobile elements and a proportional heatmap was created by calculating the proportion of genomes in a lineage that have at least 80% of the genes (Figure 3.9C). From here, we highlighted mobile genetic elements that were absent in more than 50% of all assemblies (Figure 3.9B and C). The annotation of these mobile elements showed that most groups might be involved in pathogenic activities due to genes encoding for virulence factors, including sex pili, stress resistance, exotoxins, and amino acid and carbohydrates metabolism genes.

By summarising the description of eggnog annotations of all the highlighted mobile gene clusters, we selected two relatively large mobile groups with promising genes as well as a part of the group containing the *Bacteroides fragilis* toxin to discuss in more detail (Figure 3.10-12).

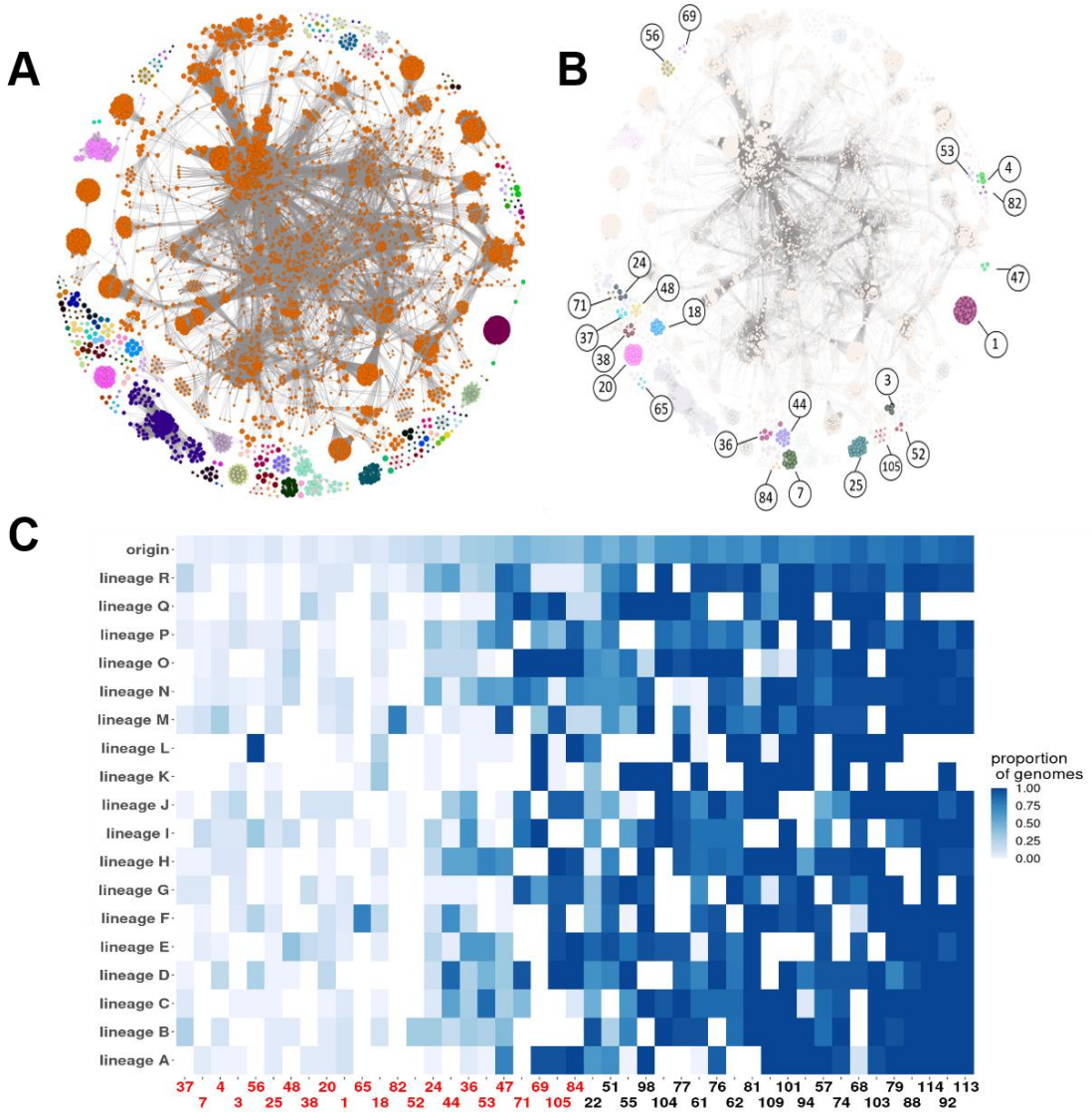


Figure 3.9: The co-occurring gene clusters and mobile elements of *B. fragilis*. Association network of coincident genes detected by Coinfinder coloured by the set of genes showing associative relationships (A), highlighted mobile coincident gene groups in the association network (B). Heatmap displaying the proportion of genomes in a specific lineage containing at least 80% of the genes in potential mobile Coinfinder gene groups, gene groups in red are absent in more than 50% of the assemblies (C).

The *Bacteroides fragilis* toxin (*bft*) was detected in a large Coinfinder group (group 23) of 135 genes. By analyzing the locus tag of the genes in multiple isolates, we noticed that only a fraction tends to be located close to each other and other coinciding genes were scattered along the genome. For this reason, only 62 (46%) genes of Coinfinder group 23 are shown in Figure 3.10. The region contained genes involved in bacterial conjugation, a peptidase, the *bft* encoding genes, a sporulation initiation inhibitor, a CRISPR-associated helicase, a lot of unknown and hypothetical proteins and multiple

membrane proteins like *susC* and *susD* which are part of a complex that binds and degrades starch into oligosaccharides while transporting it to the cytoplasm[122]. Additionally, fimbriae encoding genes, proteins known to mediate cell adhesion and biofilm formation, were also present. These results concur with previous studies demonstrating that the toxin is encoded in a conjugative transposon which occurs globally[5]. Notably, the absence-presence matrix shows that a large chunk of the conjugative transposon without the *bft* gene can still be transferred or inherited.

Coinfinder group 56 contains a total of 9 genes including two von Willebrand factor (*vWF*) type-A domain encoding genes, an important protein in hemostasis (Figure 3.11). A total of 18 *vWF* type-A domain protein-encoding genes were found in *B. fragilis'* pangenome. It has been previously shown that pathogenic bacteria use *vWF* to promote bacterial attachment and thereby contribute to the pathogenesis of the bacteria[123]. Other coincident genes of this group included a stress-responsive protein, phosphatases, and a kinase. Recombinant regions encapsulate this virulence enhancing, mobile element. Notably, multiple genes despite showing similar functionalities and being located in between the group's genes were not included in the associating gene group and were, therefore, left blank in Figure 3.11.

Another mobile group we discuss is the associating gene group 25 identified by Coinfinder containing a total of 22 genes. In here multiple proteases/peptidases can be found as well as a ligase, an epimerase, and recombinases but also protective genes against stress. The *tpx* gene is a thiol-specific peroxidase that acts as a lipid peroxidase to inhibit bacterial membrane oxidation and as an antioxidant during anaerobic growth[124]. Aside from the *tpx* gene, the mobile element also includes the pstSCAB-phoU operon, an essential part of the phosphate (Pho) regulon that regulates the phosphate homeostasis to cope with inorganic phosphate stravation[125]. The Pho regulon can also be part of complex networks for bacterial virulence, tolerance to antibiotics and tress response[125]. Finally, the group contains a final recombinase in addition to multiple restriction sites.

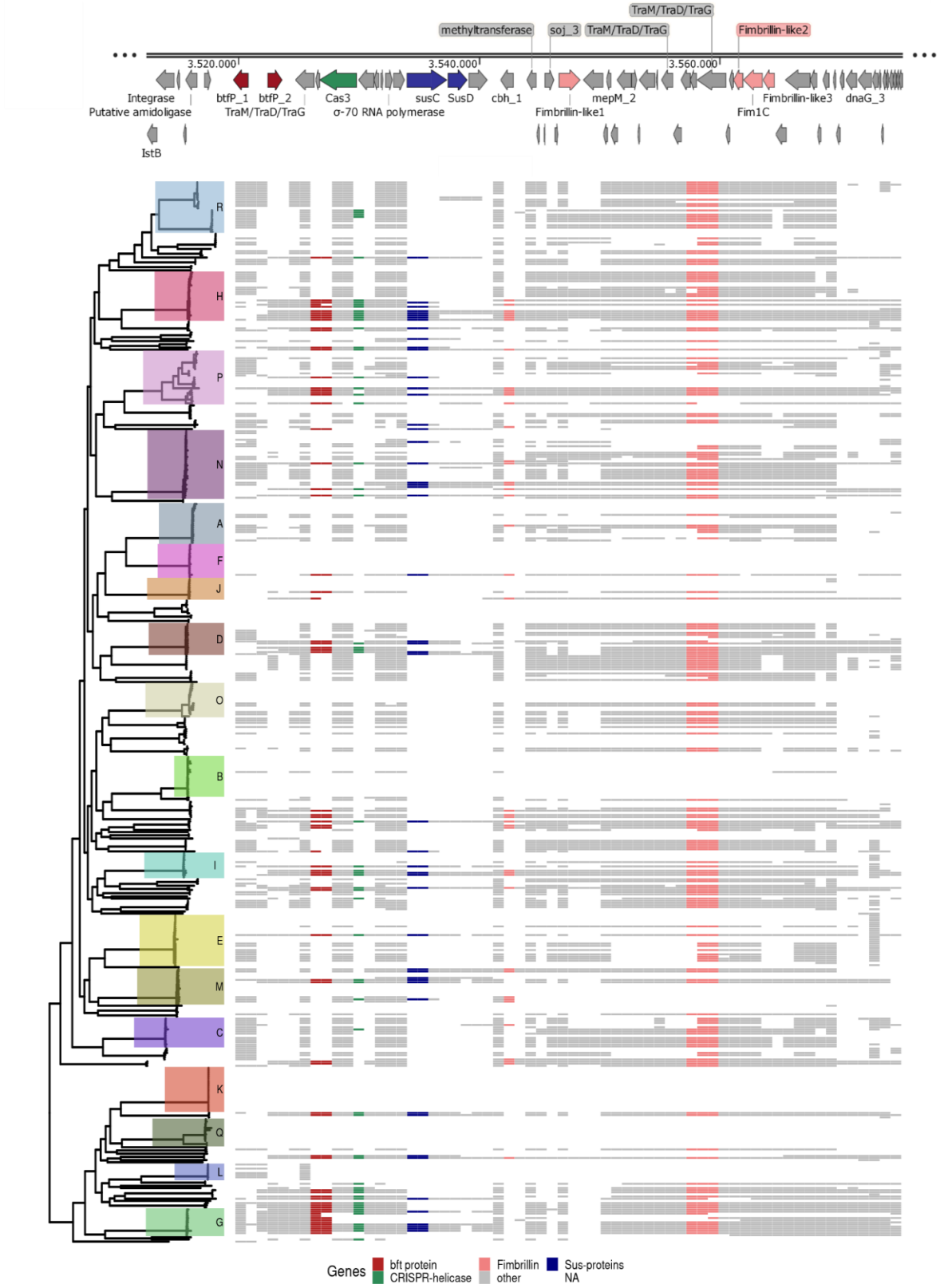


Figure 3.10: Absence-presence matrix of a fraction of Coinfinder group 23 containing the *Bacteroides fragilis* toxin encoding genes aligned with maximum likelihood tree. Interesting genes are highlighted in a specific colour and known annotations of genes are shown on top of the matrix based on the ETBF_BOB25 assembly.

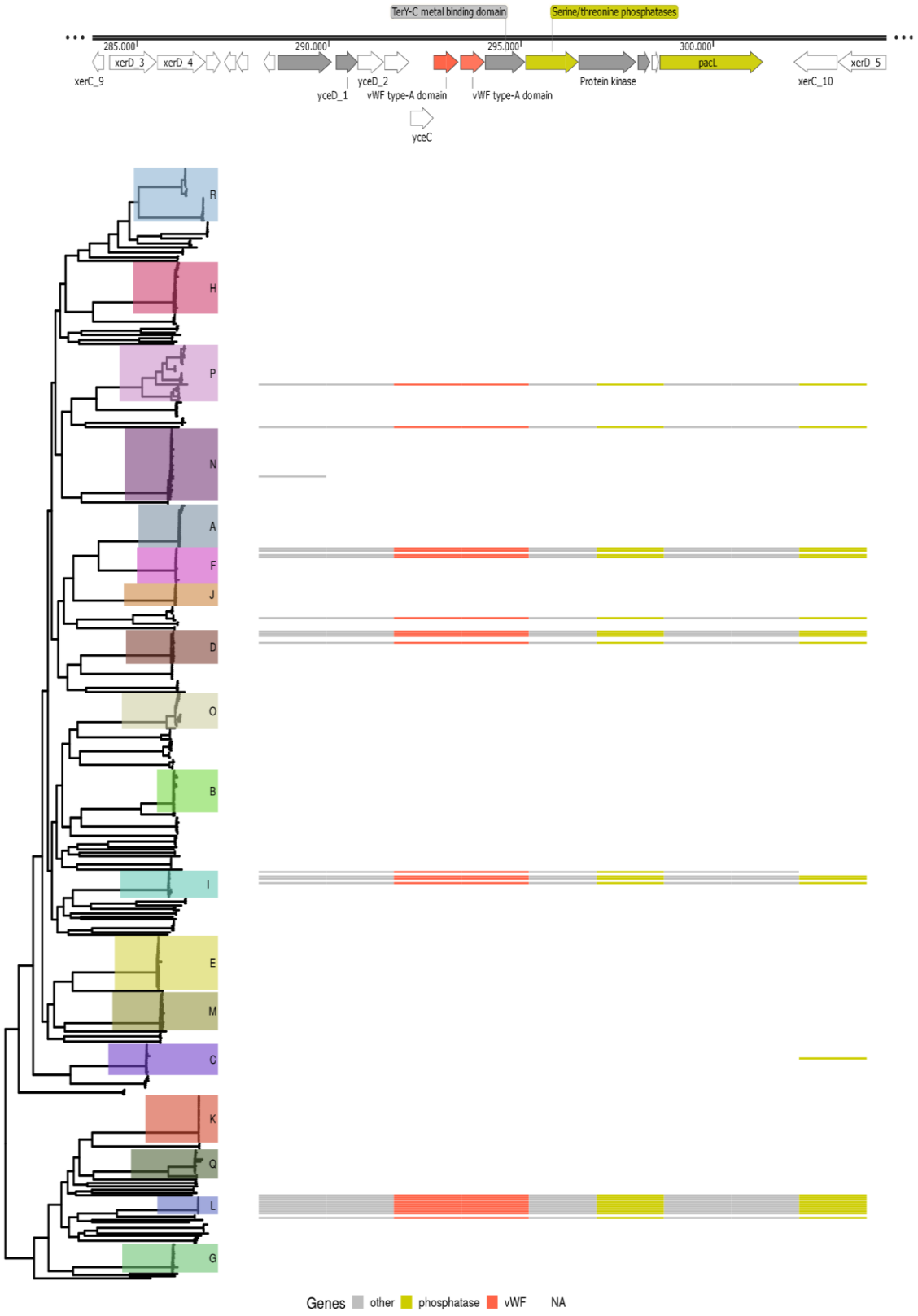


Figure 3.11: Absence-presence matrix of Coinfinder group 56 containing the von Willebrand factor aligned with maximum likelihood tree. Interesting genes are highlighted in a specific colour (genes not belonging to the Coinfinder group 56 are white) and known annotations of genes are shown on top of the matrix based on the BFR_KZ06 assembly.

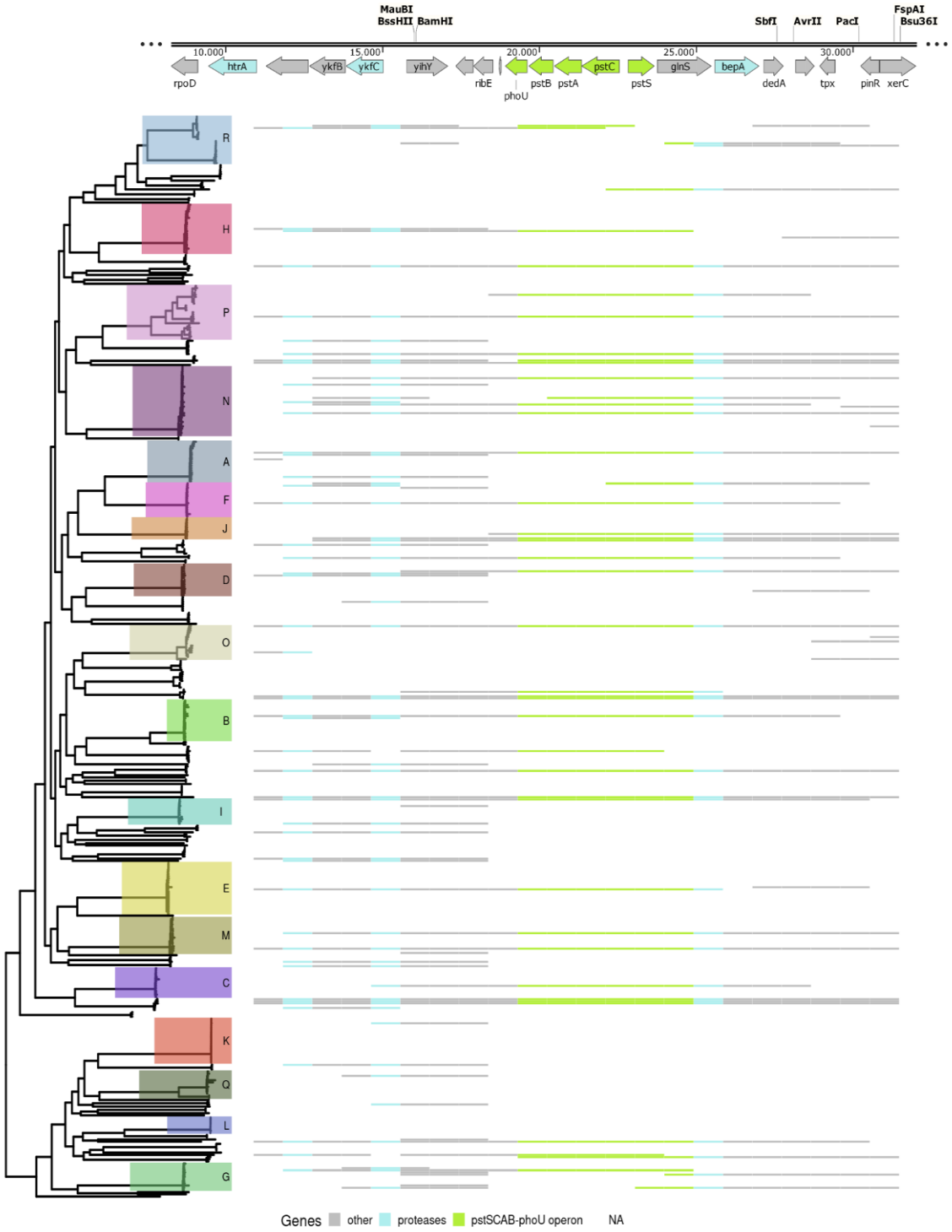


Figure 3.12: Absence-presence matrix of Coinfinder group 25 containing pstSCAB-phoU region aligned with maximum likelihood tree. Interesting genes are highlighted and known annotations of genes are shown on top of the matrix based on the S23_R14 assembly.

These three highlighted mobile groups are a few examples of the accessory gene pool of *B. fragilis* that contribute to the species genetic diversity. They illustrate that not a single strain or even two subtypes (ETBF vs NTBF) can fully represent the diversity of the rest of the species and partly explains why strains exhibit different levels of virulence or benefit on the host as previously found[5]. In addition to the first limitation of having a large fraction of MAGs used in this study, is that we can not elaborate on pathogenic properties or probiotic capacity of the identified lineages or mobile elements due to lack of appropriate metadata. The health status or age of sampled hosts was rarely reported. Regardless of these limitations, our study has provided insights into the contribution of mobile elements and recombination events on the rapid evolution of *B. fragilis* but also its large genotypic diversity. Where *B. fragilis* strains have used HGT to exchange multidrug-resistant encoding genes and other virulence factors. In addition, we hypothesized that the established *B. fragilis* lineages are a result of distinct gut microbiome ecosystems. Here, each lineage occupies a specific ecological niche within that gut microbiome where two lineages are likely to fulfil two different functions in different gut ecosystem. Finally, despite a previous study showing that *B. fragilis* strains adapt within individual microbiomes through years-long of coexistence, high genotypic similarities across geographic locations are still present, which clearly supports the clonal expansion model of the evolution of this species.

4 CONCLUSION

In conclusion, our large-scale genomic characterization of *B. fragilis* confirms its underestimated genomic diversity. The phylogenetic and population structure analyses endorse the two previously identified genotypic subtypes of *B. fragilis* and revealed a total of 18 lineages in the *cfiA-negative* subtype. All lineages were supported by pangenomic analyses, suggesting a co-evolving core and accessory genome, and reinforcing the concept of subspecies in *B. fragilis*. Multidrug resistance encoding genes were detected in a total of 12 lineages as well as distinct carbohydrate metabolism genes, some of which are embedded in mobile genetic elements. This suggests that recombination and mobile genetic elements have had a large impact on the genetic diversity of *B. fragilis*.

Future research will be able to integrate the identified lineages in their studies for more specific genotypic-phenotypic associations of *B. fragilis.* Here the phenotypic traits can include the wide array of diseases induced by *B. fragilis* or its beneficial interactions, which will provide potential targets for a multi-locus sequence identification of lineages allowing faster and more direct treatment in clinical interventions. Moreover, subsequent studies that provide a better understanding of the pangenome evolution and give insights into the dynamics of *B. fragilis* pathogenic but also beneficial properties in the microbiome are required for the design of *Bacteroides fragilis* based therapeutics.

REFERENCES

1. Quigley, E. M. M. Gut Bacteria in Health and Disease. *Gastroenterol. Hepatol.* 9, 560–569 (2013).

2. Purcell, R. V. Chapter 4 - *Bacteroides fragilis*. in *Colorectal Neoplasia and the Colorectal Microbiome* (ed. Floch, M. H.) 57–77 (Academic Press, 2020). doi:10.1016/B978-0-12-819672-4.00004-0.

3. Sun, F. *et al.* A potential species of next-generation probiotics? The dark and light sides of *Bacteroides fragilis* in health. *Food Res. Int.* 126, 108590 (2019).

4. Zhao, S. *et al.* Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* 25, 656-667.e8 (2019).

5. Pierce, J. V. & Bernstein, H. D. Genomic Diversity of Enterotoxigenic Strains of *Bacteroides fragilis*. *PLOS ONE* 11, e0158171 (2016).

6. Magne, F. *et al.* The Firmicutes/Bacteroidetes Ratio: A Relevant Marker of Gut Dysbiosis in Obese Patients? *Nutrients* 12, (2020).

7. Wexler, H. M. *Bacteroides*: the Good, the Bad, and the Nitty-Gritty. *Clin. Microbiol. Rev.* 20, 593–621 (2007).

8. Elliott, D., Kufera, J. A. & Myers, R. A. The microbiology of necrotizing soft tissue infections. *Am. J. Surg.* 179, 361–366 (2000).

9. Ghotaslou, R. *et al.* Mechanisms of *Bacteroides fragilis* resistance to metronidazole. *Infect. Genet. Evol.* 64, 156–163 (2018).

10. Ferløv-Schwensen, S. A., Sydenham, T. V., Hansen, K. C. M., Hoegh, S. V. & Justesen, U. S. Prevalence of antimicrobial resistance and the *cfiA* resistance gene in Danish *Bacteroides fragilis* group isolates since 1973. *Int. J. Antimicrob. Agents* 50, 552–556 (2017).

11. Wang, Y. *et al.* An intestinal commensal symbiosis factor controls neuroinflammation via TLR2-mediated CD39 signalling. *Nat. Commun.* 5, 4432 (2014).

12. Wang, Y. *et al.* A commensal bacterial product elicits and modulates migratory capacity of CD39(+) CD4 T regulatory subsets in the suppression of

neuroinflammation. *Gut Microbes* 5, 552–561 (2014).

13. Shen, Y. *et al.* Outer membrane vesicles of a human commensal mediate immune regulation and disease protection. *Cell Host Microbe* 12, 509–520 (2012).

14. Chang, Y.-C. *et al.* TLR2 and interleukin-10 are involved in *Bacteroides fragilis*-mediated prevention of DSS-induced colitis in gnotobiotic mice. *PloS One* 12, e0180025 (2017).

15. Round, J. L. & Mazmanian, S. K. Inducible *Foxp3*+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proc. Natl. Acad. Sci. U. S. A.* 107, 12204–12209 (2010).

16. Johnson, J. L., Jones, M. B. & Cobb, B. A. Bacterial capsular polysaccharide prevents the onset of asthma through T-cell activation. *Glycobiology* 25, 368–375 (2015).

17. Gilbert, J. A., Krajmalnik-Brown, R., Porazinska, D. L., Weiss, S. J. & Knight, R. Towards effective probiotics for autism and other mental disorders? *Cell* 155, 1446–1448 (2013).

18. Franco, A. A. *et al.* Molecular evolution of the pathogenicity island of enterotoxigenic *Bacteroides fragilis* strains. *J. Bacteriol.* 181, 6623–6633 (1999).

19. Moncrief, J. S. *et al.* The enterotoxin of *Bacteroides fragilis* is a metalloprotease. *Infect. Immun.* 63, 175–181 (1995).

20. Wu, S., Lim, K. C., Huang, J., Saidi, R. F. & Sears, C. L. *Bacteroides fragilis* enterotoxin cleaves the zonula adherens protein, E-cadherin. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14979–14984 (1998).

21. Moncrief, J. S., Duncan, A. J., Wright, R. L., Barroso, L. A. & Wilkins, T. D. Molecular characterization of the fragilysin pathogenicity islet of enterotoxigenic *Bacteroides fragilis*. *Infect. Immun.* 66, 1735–1739 (1998).

22. Franco, A. A. The *Bacteroides fragilis* Pathogenicity Island Is Contained in a Putative Novel Conjugative Transposon. *J. Bacteriol.* 186, 6077–6092 (2004).

23. Van Tassell, R. L., Lyerly, D. M. & Wilkins, T. D. Purification and characterization of an enterotoxin from *Bacteroides fragilis*. *Infect. Immun.* 60, 1343–1350 (1992).

24. Franco, A. A. *et al.* Cloning and characterization of the *Bacteroides fragilis* metalloprotease toxin gene. *Infect. Immun.* 65, 1007–1013 (1997).

25. Chung, G.-T. *et al.* Identification of a Third Metalloprotease Toxin Gene in Extraintestinal Isolates of *Bacteroides fragilis*. *Infect. Immun.* 67, 4945–4949 (1999).

26. Zhang, G., Svenungsson, B., Kärnell, A. & Weintraub, A. Prevalence of enterotoxigenic *Bacteroides fragilis* in adult patients with diarrhea and healthy controls. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* 29, 590–594 (1999).

27. Sack, R. B. *et al.* Enterotoxigenic *Bacteroides fragilis*: epidemiologic studies of its role as a human diarrhoeal pathogen. *J. Diarrhoeal Dis. Res.* 10, 4–9 (1992).

28. Ji, D.-D. *et al.* Prevalence and characterization of enterotoxigenic *Bacteroides fragilis* and toxigenic *Clostridium difficile* in a Taipei emergency department. *J. Microbiol. Immunol. Infect.* 50, 83–89 (2017).

29. Wu, S. *et al.* The *Bacteroides fragilis* Toxin Binds to a Specific Intestinal Epithelial Cell Receptor. *Infect. Immun.* 74, 5382–5390 (2006).

30. Wu, S., Rhee, K.-J., Zhang, M., Franco, A. & Sears, C. L. *Bacteroides fragilis* toxin stimulates intestinal epithelial cell shedding and γ-secretase-dependent E-cadherin cleavage. *J. Cell Sci.* 120, 1944–1952 (2007).

31. Rhee, K.-J. *et al.* Induction of Persistent Colitis by a Human Commensal, Enterotoxigenic *Bacteroides fragilis*, in Wild-Type C57BL/6 Mice. *Infect. Immun.* 77, 1708–1718 (2009).

32. Wu, S. *et al.* A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat. Med.* 15, 1016–1022 (2009).

33. Thiele Orberg, E. *et al.* The myeloid immune signature of enterotoxigenic *Bacteroides fragilis* -induced murine colon tumorigenesis. *Mucosal Immunol.* 10, 421–433 (2017).

34. Kwong, T. N. Y. *et al.* Association Between Bacteremia From Specific Microbes and Subsequent Diagnosis of Colorectal Cancer. *Gastroenterology* 155, 383-390.e8 (2018).

35. Zhao, Y. & Lukiw, W. J. Bacteroidetes Neurotoxins and Inflammatory Neurodegeneration. *Mol. Neurobiol.* 55, 9100–9107 (2018).

36. Mazmanian, S. K., Round, J. L. & Kasper, D. L. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* 453, 620–625 (2008).

37. Sommese, L. *et al.* Evidence of *Bacteroides fragilis* Protection from *Bartonella*

henselae-Induced Damage. *PLOS ONE* 7, e49653 (2012).

38. Sittipo, P. *et al.* Toll-Like Receptor 2-Mediated Suppression of Colorectal Cancer Pathogenesis by Polysaccharide A From *Bacteroides fragilis*. *Front. Microbiol.* 9, (2018).

39. Li, Z. *et al.* Bioluminescence Imaging to Track *Bacteroides fragilis* Inhibition of *Vibrio parahaemolyticus* Infection in Mice. *Front. Cell. Infect. Microbiol.* 7, (2017).

40. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An Immunomodulatory Molecule of Symbiotic Bacteria Directs Maturation of the Host Immune System. *Cell* 122, 107–118 (2005).

41. Round, J. L. *et al.* The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science* 332, 974–977 (2011).

42. Chan, J. L. *et al.* Non-toxigenic *Bacteroides fragilis* (NTBF) administration reduces bacteria-driven chronic colitis and tumor development independent of polysaccharide A. *Mucosal Immunol.* 12, 164–177 (2019).

43. Surana, N. K. & Kasper, D. L. The yin yang of bacterial polysaccharides: lessons learned from *B. fragilis* PSA. *Immunol. Rev.* 245, 13–26 (2012).

44. Arnolds, K. L., Moreno-Huizar, N., Stanislawski, M. A., Palmer, B. & Lozupone, C. Hemagglutination by *B. fragilis* is mediated by capsular polysaccharides and is influenced by host ABO blood type. *bioRxiv* 2020.08.19.258442 (2020) doi:10.1101/2020.08.19.258442.

45. Weng, M. & Walker, W. A. The role of gut microbiota in programming the immune phenotype. *J. Dev. Orig. Health Dis.* 4, 203–214 (2013).

46. Ewald, D. R. & Sumner, S. C. J. Blood type biochemistry and human disease. *WIREs Syst. Biol. Med.* 8, 517–535 (2016).

47. Cooling, L. Blood Groups in Infection and Host *Susc*eptibility. *Clin. Microbiol. Rev.* 28, 801–870 (2015).

48. Rühlemann, M. C. *et al.* Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *Nat. Genet.* 53, 147–155 (2021).

49. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214 (2012).

50. Eckburg, P. B. *et al.* Diversity of the Human Intestinal Microbial Flora. *Science* 308, 1635–1638 (2005).

51. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* 486, 222–227 (2012).

52. Lay, C. *et al.* Colonic microbiota signatures across five northern European countries. *Appl. Environ. Microbiol.* 71, 4153–4155 (2005).

53. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* 352, 560–564 (2016).

54. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551, 507–511 (2017).

55. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* 473, 174–180 (2011).

56. Vieira-Silva, S. *et al.* Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat. Microbiol.* 4, 1826–1831 (2019).

57. Vieira-Silva, S. *et al.* Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* 581, 310–315 (2020).

58. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* 4, 623–632 (2019).

59. Wybo, I. *et al.* Differentiation of *cfiA*-negative and *cfiA*-positive *Bacteroides fragilis* isolates by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.* 49, 1961–1964 (2011).

60. Gutacker, M., Valsangiacomo, C. & Piffaretti, J.-C. Identification of two genetic groups in *Bacteroides fragilis* by multilocus enzyme electrophoresis: distribution of antibiotic resistance (*cfiA*, *cepA*) and enterotoxin (*bft*) encoding genes. *Microbiol. Read. Engl.* 146 (Pt 5), 1241–1254 (2000).

61. Hallatschek, O., Hersen, P., Ramanathan, S. & Nelson, D. R. Genetic drift at expanding frontiers promotes gene segregation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19926–19930 (2007).

62. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* 18, 491–506

(2020).

63. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* 13, 960 (2017).

64. Retchless, A. C. & Lawrence, J. G. Temporal fragmentation of speciation in bacteria. *Science* 317, 1093–1096 (2007).

65. Shapiro, B. J. What Microbial Population Genomics Has Taught Us About Speciation. in *Population Genomics: Microorganisms* (eds. Polz, M. F. & Rajora, O. P.) 31–47 (Springer International Publishing, 2018). doi:10.1007/13836_2018_10.

66. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3, 711–721 (2005).

67. Young, J. P. W. *et al.* The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* 7, R34 (2006).

68. McCoy, R. C. & Akey, J. M. Selection plays the hand it was dealt: evidence that human adaptation commonly targets standing genetic variation. *Genome Biol.* 18, 139 (2017).

69. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955 (2005).

70. Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J. & Edwards, D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends Genet.* 36, 132–145 (2020).

71. Domingo-Sananes, M. R. & McInerney, J. O. Selection-based model of prokaryote pangenomes. *bioRxiv* 782573 (2019) doi:10.1101/782573.

72. Colquhoun, R. M. *et al.* Nucleotide-resolution bacterial pan-genomics with reference graphs. *bioRxiv* 2020.11.12.380378 (2020) doi:10.1101/2020.11.12.380378.

73. Bolotin, E. & Hershberg, R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. *Genome Biol. Evol.* 7, 2173–2187 (2015).

74. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* 2, 1–5 (2017).

75. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* 7, 72–85 (2015).

76. Davies, M. R. *et al.* Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nat. Genet.* 51, 1035–1043 (2019).

77. Poulsen, B. E. *et al.* Defining the core essential genome of Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci.* 116, 10072–10080 (2019).

78. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–594 (2005).

79. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154 (2015).

80. Vos, M. & Eyre-Walker, A. Are pangenomes adaptive or not? *Nat. Microbiol.* 2, 1576–1576 (2017).

81. Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci.* 113, 11399–11407 (2016).

82. Shapiro, B. J. The population genetics of pangenomes. *Nat. Microbiol.* 2, 1574–1574 (2017).

83. Bobay, L.-M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* 18, 153 (2018).

84. Ohta, T. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* 246, 96–98 (1973).

85. Bobay, L.-M. The Prokaryotic Species Concept and Challenges. in *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (eds. Tettelin, H. & Medini, D.) 21–49 (Springer International Publishing, 2020). doi:10.1007/978-3-030-38281-0_2.

87. Kitts, P. A. *et al.* Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44, D73-80 (2016).

88. Youngblut, N. D. *et al.* Large-Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity. *mSystems* 5, (2020).

89. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120 (2014).

90. Nurk, S. *et al.* Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. in *Research in Computational Molecular Biology* (eds. Deng, M., Jiang, R., Sun, F. & Zhang, X.) 158–170 (Springer, 2013). doi:10.1007/978-3-642-37195-0_13.

91. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075 (2013).

92. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043 (2015).

93. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 5114 (2018).

94. Carroll, L. M., Wiedmann, M. & Kovac, J. Proposal of a Taxonomic Nomenclature for the *Bacillus cereus* Group Which Reconciles Genomic Definitions of Bacterial Species with Clinical and Industrial Phenotypes. *mBio* 11, (2020).

95. Fernández-de-Bobadilla, M. D. *et al. PATO: Pangenome Analysis Toolkit*. http://biorxiv.org/lookup/doi/10.1101/2021.01.30.428878 (2021) doi:10.1101/2021.01.30.428878.

96. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 132 (2016).

97. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028 (2017).

98. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100 (2018).

99. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* 21, 180 (2020).

100. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122 (2017).

101. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314 (2019).

102. Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob. Agents Chemother.* 63, (2019).

103. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* 47, 5539–5549 (2019).

104. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 43, e15–e15 (2015).

105. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. in *Advances in Knowledge Discovery and Data Mining* (eds. Pei, J., Tseng, V. S., Cao, L., Motoda, H. & Xu, G.) 160–172 (Springer, 2013). doi:10.1007/978-3-642-37456-2_14.

107. Hadfield, J. *et al.* Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 34, 292–293 (2018).

108. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).

109. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 17, 238 (2016).

110. Whelan, F. J., Rusilowicz, M. & McInerney, J. O. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb. Genomics* 6, e000338 (2020).

111. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* 69, e96 (2020).

112. Bobay, L.-M. & Ochman, H. Biological Species Are Universal across Life's Domains. *Genome Biol. Evol.* 9, 491–501 (2017).

113. Lan, R. & Reeves, P. R. When does a clone deserve a name? A perspective on

bacterial species based on population genetics. *Trends Microbiol.* 9, 419–424 (2001).

114. Spratt, B. G. Exploring the Concept of Clonality in Bacteria. in *Genomics, Proteomics, and Clinical Bacteriology: Methods and Reviews* (eds. Woodford, N. & Johnson, A. P.) 323–352 (Humana Press, 2004). doi:10.1385/1-59259-763-7:323.

115. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27, 626–638 (2017).

116. Kuo, C.-H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19, 1450–1454 (2009).

117. Croucher, N. J. *et al.* Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun.* 5, (2014).

118. Choi, K.-H., Heath, R. J. & Rock, C. O. β-Ketoacyl-Acyl Carrier Protein Synthase III (FabH) Is a Determining Factor in Branched-Chain Fatty Acid Biosynthesis. *J. Bacteriol.* 182, 365–370 (2000).

119. Lai, C.-Y. & Cronan, J. E. Isolation and Characterization of β-Ketoacyl-Acyl Carrier Protein Reductase (*fabG*) Mutants of Escherichia coli and Salmonella enterica Serovar Typhimurium. *J. Bacteriol.* 186, 1869–1878 (2004).

120. Kurokawa, K. *et al.* Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. *DNA Res.* 14, 169–181 (2007).

121. Merino, S. & Tomás, J. M. Bacterial Capsules and Evasion of Immune Responses. in *eLS* 1–10 (American Cancer Society, 2015). doi:10.1002/9780470015902.a0000957.pub4.

122. Foley, M. H., Cockburn, D. W. & Koropatkin, N. M. The *Sus* operon: a model system for starch uptake by the human gut Bacteroidetes. *Cell. Mol. Life Sci. CMLS* 73, 2603–2617 (2016).

123. Steinert, M., Ramming, I. & Bergmann, S. Impact of Von Willebrand Factor on Bacterial Pathogenesis. *Front. Med.* 7, (2020).

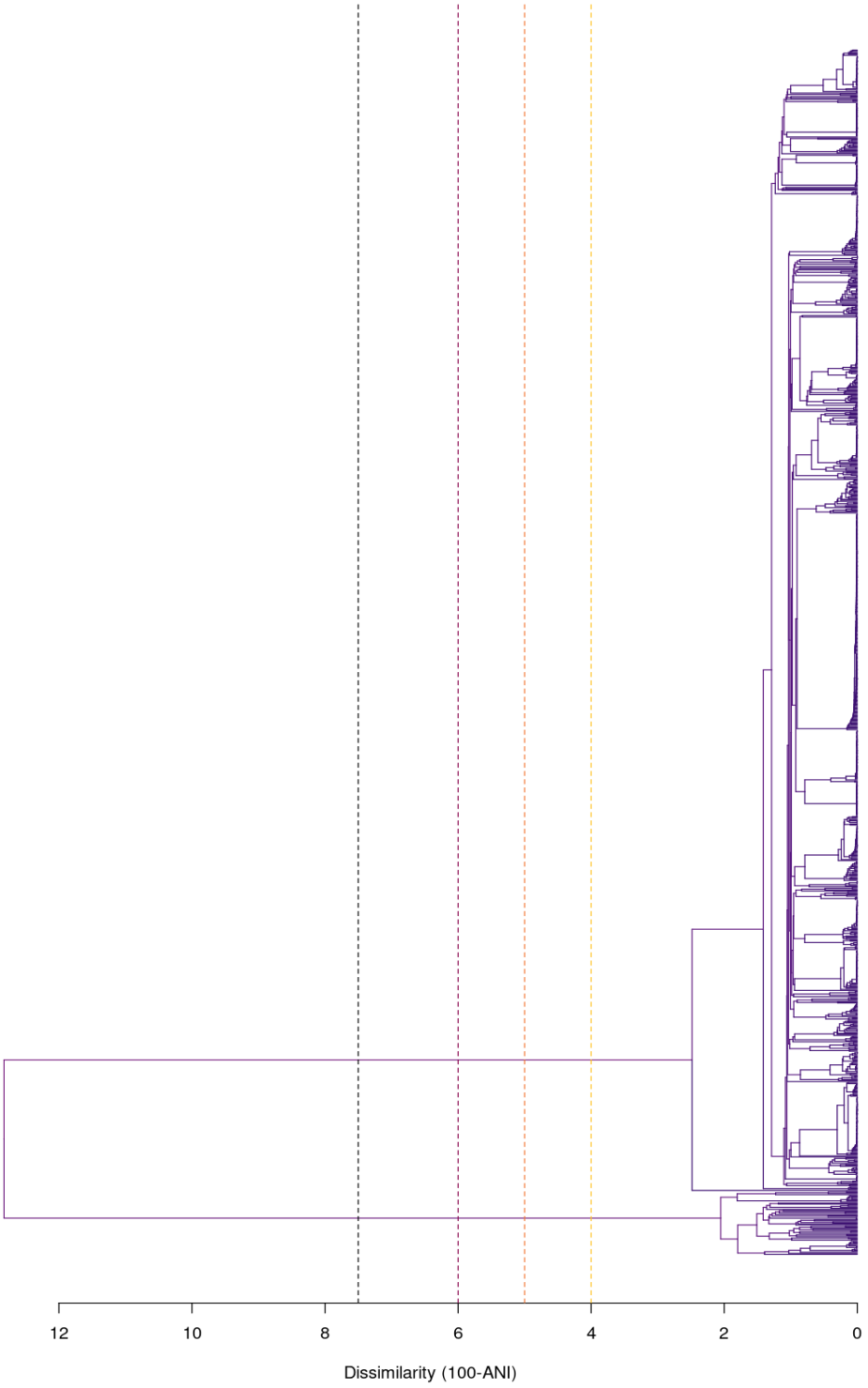
124. Cha, M.-K., Kim, W.-C., Lim, C.-J., Kim, K. & Kim, I.-H. *Escherichia coli* Periplasmic Thiol Peroxidase Acts as Lipid Hydroperoxide Peroxidase and the Principal Antioxidative Function during Anaerobic Growth *. *J. Biol. Chem.* 279,

8769–8778 (2004).

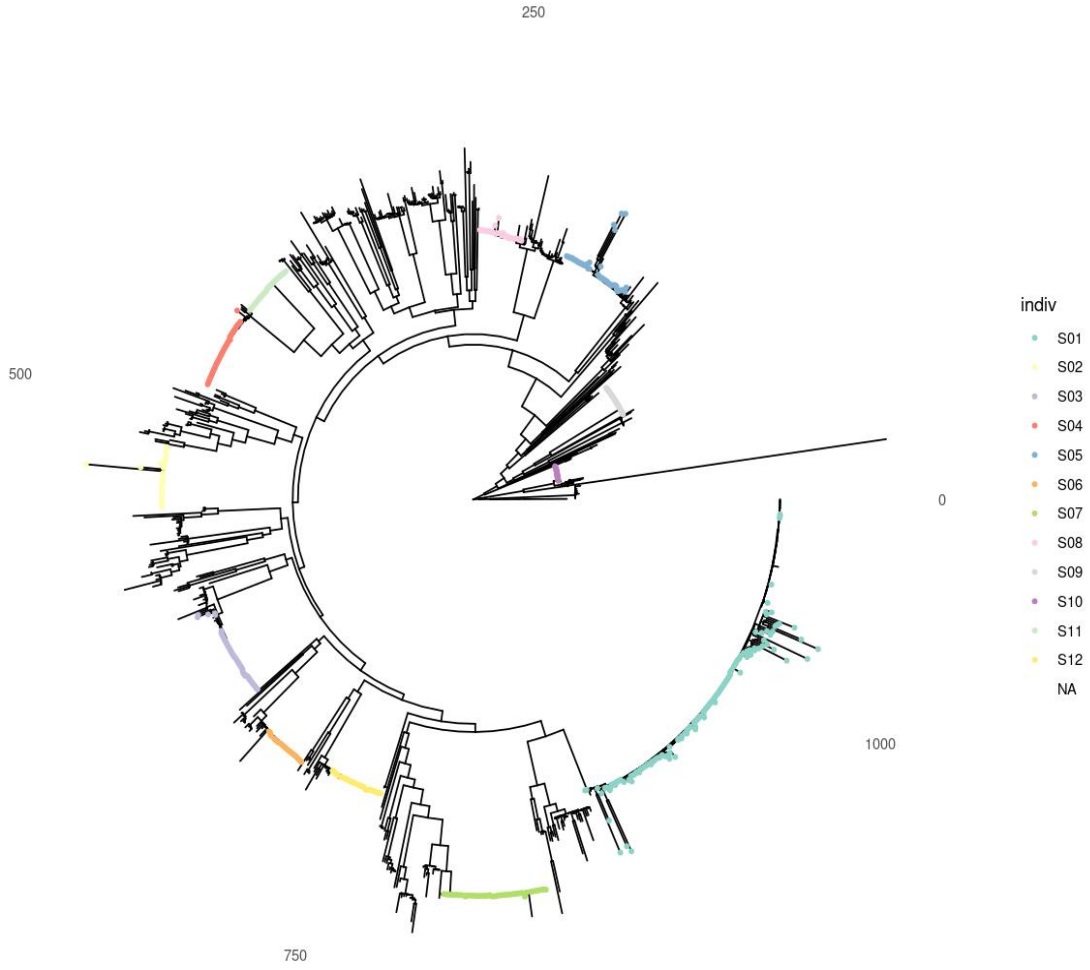
125. Santos-Beneit, F. The Pho regulon: a huge regulatory network in bacteria. *Front. Microbiol.* 6, (2015).

APPENDIXES

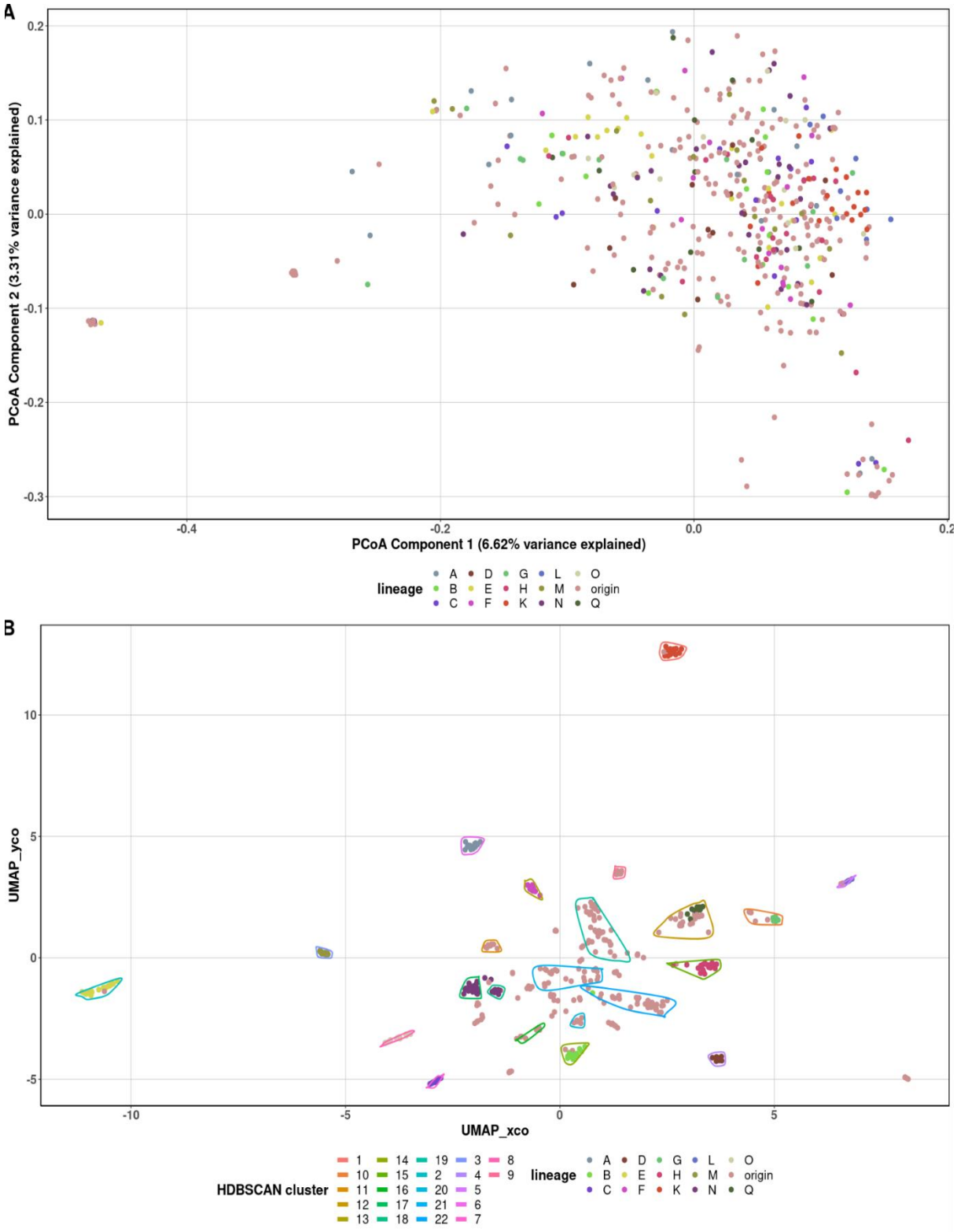
Supplementary figure 1: Dendrogram based on Average Nucleotide Identity using the BactaxR[94] pipeline. The two genotypic subtypes are separated at approximately 88% ANI.



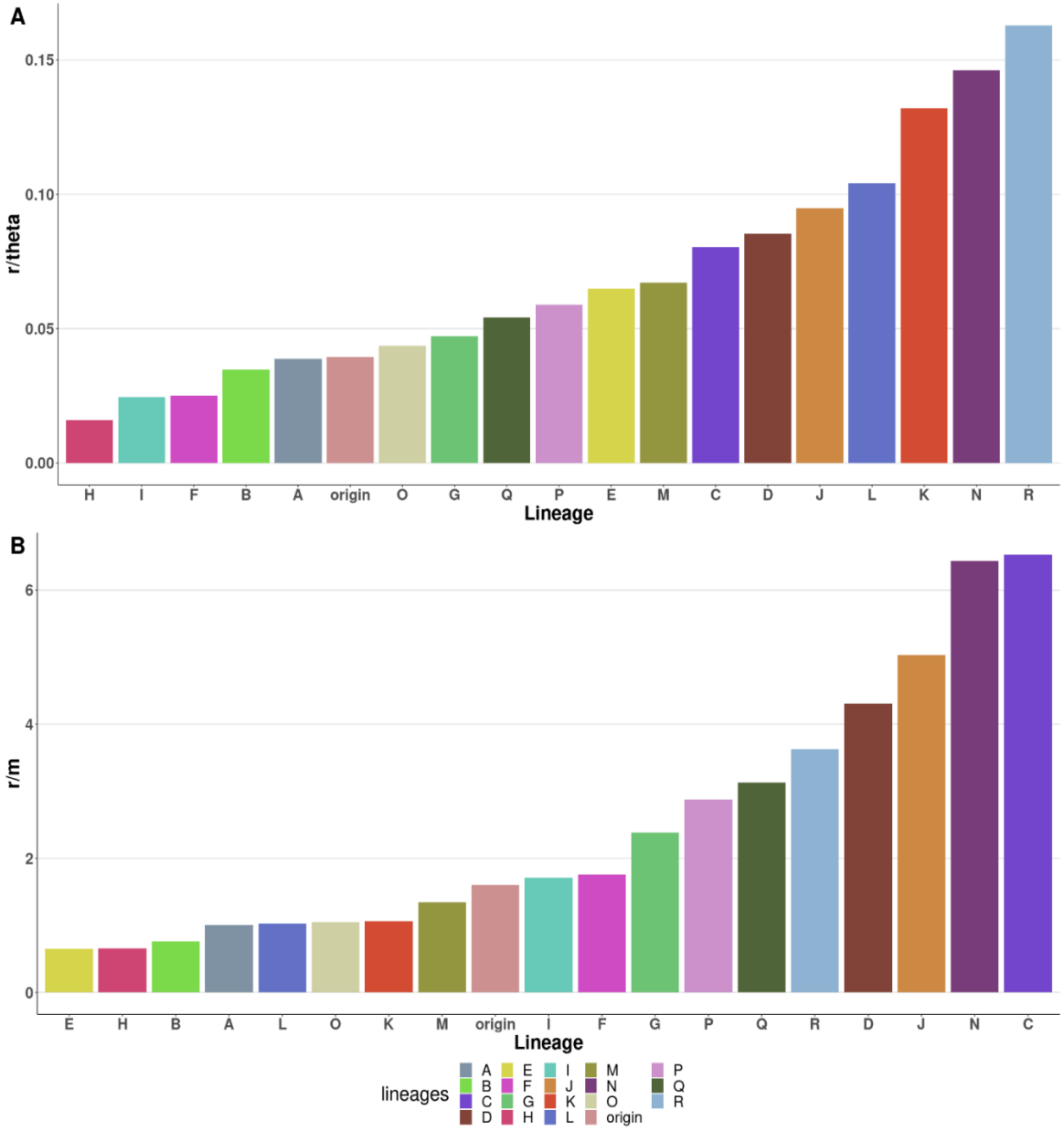
Supplementary figure 2: Dendrogram based on core genome alignment of all assemblies estimated by Panaroo[99]. Assemblies of over represented clusters based on ANI, sample location, and sample host are visualized.



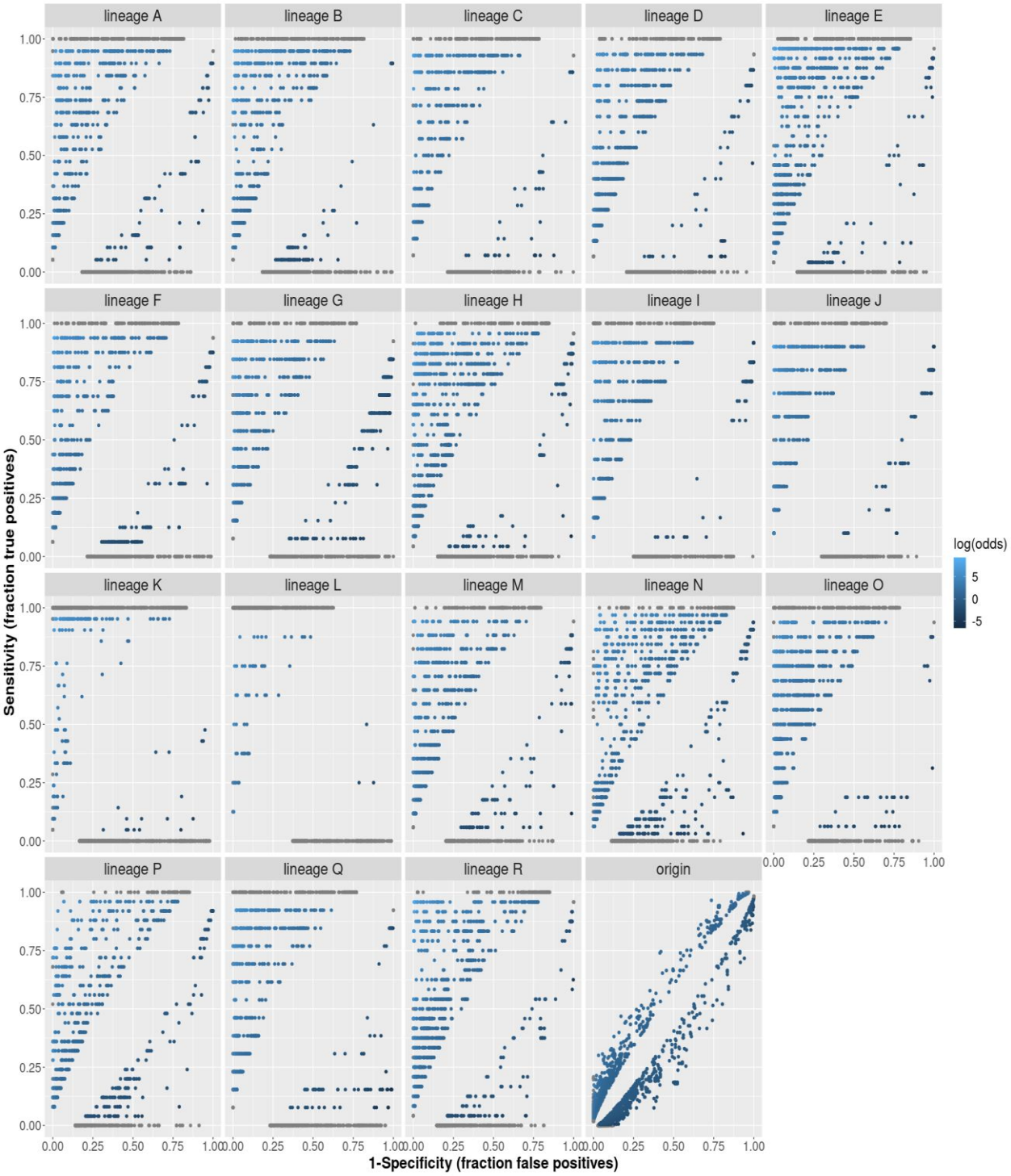
Supplementary figure 3: PCoA (A) and UMAP (B) projection of absence-presence matrix of accessory genome. Assemblies are coloured by lineages based on fastbaps clustering of the recombination accommodated whole genome alignment. Assemblies are encircled by HDBSCAN clusters based on the UMAP projection. HDBSCAN cluster 0 is not drawn.



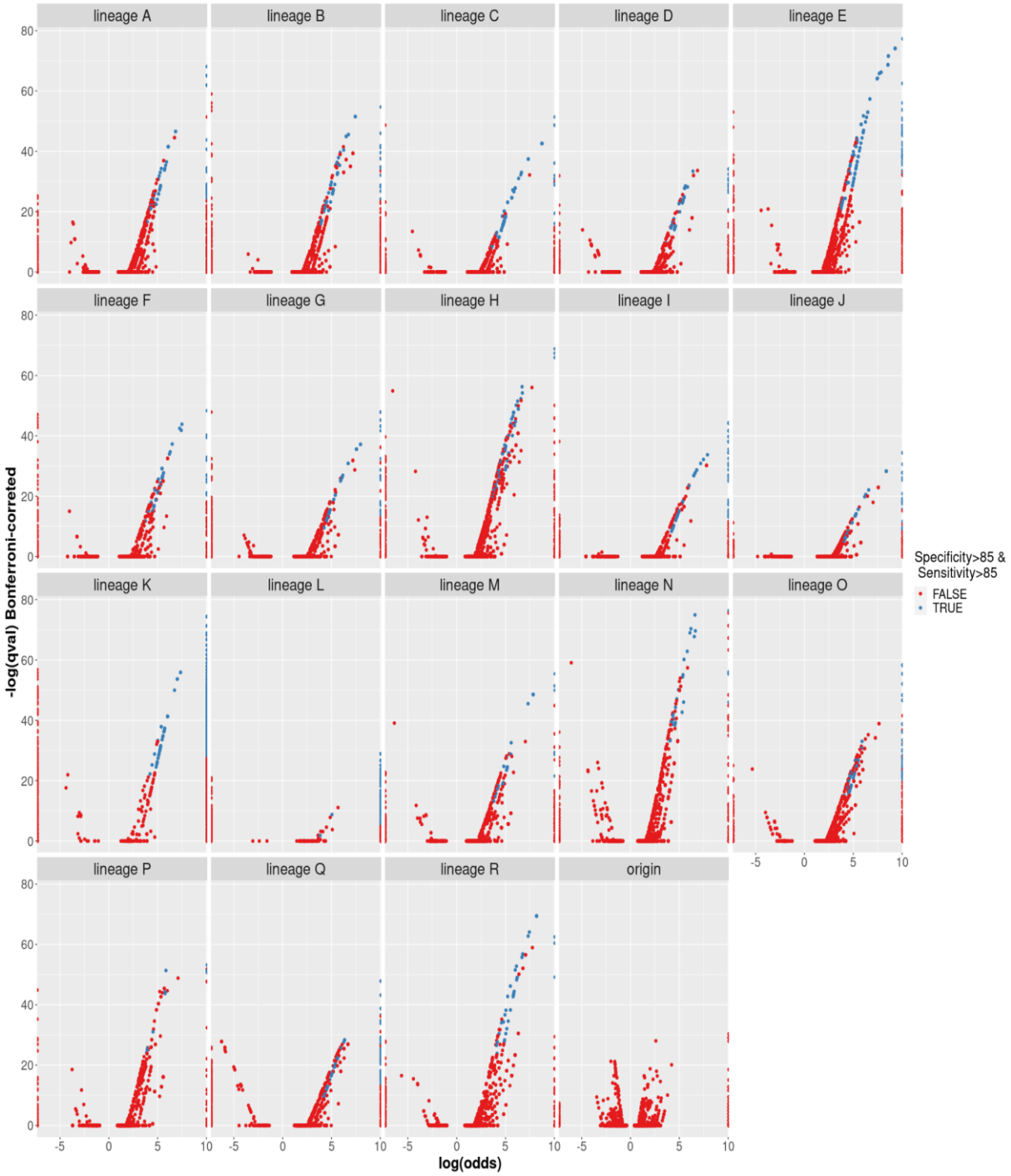
Supplementary figure 4: The rate of recombination was over the mutation rate (r/theta; A) and the estimated to contribution to population diversity of recombination rate over mutation (r/m; B) .



Supplementary figure 5: Results of GWAS study with sensitivity plotted over 1-specificity, coloured by the logarithm of the odds ratio. Note, that there are multiple points in grey on the very top and the very bottom of each graph. This is due to an infinite odds value.



Supplementary figure 6: Volcano plots of GWAS for each lineage, coloured blue if genes had a sensitivity and specificity higher than 85%, else the coloured is red. Note, that there are multiple points on the outer right and left edge of each graph. This is due to an infinite odds value.



POPULARISED SUMMARY

Microorganisms are essential to human health. It has been estimated that the total number of microorganisms living in and on a human is 100 trillion (10^{14}), where the vast majority lives in the large intestine. Some bacterial species in the colon are known to induce pathogenesis like inflammation, colorectal cancer and various other diseases whereas others are known to reduce signs of inflammation, multiple sclerosis and even autism. Some species are known to do both and are called 'pathobionts'. One of these species is *Bacteroides fragilis.* To investigate this controversial property of *B. fragilis*, we downloaded over a thousand genomes from multiple geographic locations and sources and analysed them. The analysis revealed that over time, *B. fragilis* developed numerous distinct lineages and that some of these lineages were partly defined by multidrug-resistant genes as well as distinct carbohydrate metabolism genes. Furthermore, we noticed that some groups of genes liked to be transferred from strain to strain and that some groups contained virulence factors and stress resistance genes. In summary, we showed that the *B. fragilis* is genetically a very diverse species which will help future research to link diseases or beneficial traits, not to the *B. fragilis* species but one or multiple lineages.


1425