



Master Thesis in Bioinformatics

# **Inference of population genomic parameters in the human gut microbiome**

**Vasilarou Maria**

Supervisor : Dr. Pavlos Pavlidis



University of Crete  
Department of Medicine

Tsaris



UNIVERSITY OF CRETE  
DEPARTMENT OF MEDICINE

---

**Master Thesis**

for the graduate program in:

**Bioinformatics**

Defended by

**Vasilarou Maria**

Thesis supervisor

**Dr. Pavlos Pavlidis**

Heraklion  
2018

# Abstract

The human body carries a dense and complex community of microorganisms that contribute to our health and wellbeing, called the human microbiota. The adult gut in particular, is the most densely populated habitat on the human body. It harbours a dynamic ecosystem that offers many benefits to the host, through a range of physiological functions. The advent of culture independent approaches such as high-throughput sequencing methods along with sophisticated computational methods have greatly improved our ability to explore the breadth of the gut microbiota, while its role in a large number of intestinal and extraintestinal diseases has become apparent. There is extensive research in exploring the vast amount of genetic diversity that characterizes the human microbiome, with the pioneering work of the Human Microbiome Project (HMP) notably influencing the field. However, combinations of molecular and ecological approaches to investigate the evolutionary forces that drive this microbial variability have been limited. In the current study, we explored the taxonomic and genetic diversity of microbial communities colonizing the gut of healthy individuals from the 1<sup>st</sup> phase of the HMP. By applying population genetic methodologies on reconstructed whole genome sequence data, we modeled the phylogenetic history and population demography of prevalent bacterial species in the human gut.



*“Essentially, all models are wrong, but some are useful.”*

George Box

# Acknowledgments

I owe my sincere gratitude to my mentor and supervisor, Dr. Pavlos Pavlidis. I joined his research team merely by chance, when I was looking for internship opportunities for my gap year. Having graduated with a bachelor's degree in Molecular Biology and Genetics, at first I was reluctant to pursue Bioinformatics. Pavlos was so kind, supportive and patient to introduce me to this new exciting field. He encouraged me to apply to the master's program in Bioinformatics, even when I was doubting my suitability and computational skills. Over time, I was inspired by the unique way he shares his love for evolution. His dedication and scientific enthusiasm kept me motivated during this thesis. Lastly, I deeply value all his efforts to keep me calm and concentrated during hard days when everything seemed to go extremely wrong.

My journey towards this degree would not have been possible without the support of my family and friends. To my family, thank you for always being by my side. I am so grateful for loving support and encouragement in all of my pursuits. Thanks for believing in me and listening to all the times I tried to explain the aim and methodology of this thesis to you. Some special words of gratitude go to my friends for their love and comfort in times of need.

# Table of contents

<b>Introduction</b>	<b>8</b>
1.1 The human gut microbiome	9
1.2 Metagenomics	11
1.3 The Human Microbiome Project	13
1.4 Exploring the Human Microbiome	14
1.4.1 Assembly	15
1.4.2 Binning	16
1.4.3 Marker gene analysis	16
1.4.5 A daunting choice	17
1.5 Population genomics in the gut microbiome	18
1.6 Demographic inference	21
1.6.1 Bayesian statistics	21
1.6.2 Approximate Bayesian Computation (ABC)	22
1.6.3 Inferring the demographic history of gut bacterial populations from genetic variation data	23
1.7 Phylogenetic inference	24
1.8 Study overview	25
<b>Methods</b>	<b>26</b>
2.1 Data preprocessing	26
2.2 Taxonomic profiling	27
2.4 Demographic analysis	29
2.4.1 Summary statistics	30
2.4.2 Simulations	31
2.4.2 ABC inference	33
<b>Results</b>	<b>35</b>
3.1 Quality assessment	35
3.2 Microbial profiles	38
3.3 Obtaining strain-level resolution of species	40
3.4 Examining the ‘Enterotypes’ Hypothesis	42
3.5 Investigating a sex-bias in gut microbial composition and phylogeny	43
3.6 Modeling the demographic history of gut species	45
<b>Conclusions</b>	<b>54</b>
<b>References</b>	<b>56</b>
<b>Supporting information</b>	<b>66</b>

# Introduction

Life on Earth is dominated by the smallest habitants, microbes. Bacteria, archaea and microeukaryotes form complicated ecological communities that thrive in almost all environments, even in conventionally inhospitable conditions. Extremophiles live in extreme environments that are detrimental to most life: from geothermal hot springs (Aanniz et al., 2015; Madigan, M. T., Martinko J, M, Parker J, 2006, p. 31) to utterly cold waters in the deep sea (Kato, Sato, & Horikoshi, 1995) and in acid mine drainage communities (Zagury, Kulnieks, & Neculita, 2006); (Tyson et al., 2004). Microorganisms are not only ubiquitous, they actually influence profoundly all aspects of life and human activities. But in order to understand their role in the biosphere, the human society and body as well as in plants and animals, it is crucial to undertake a genomic study of them as well.

The human body carries a dense and complex community of microorganisms that contribute to our health and wellbeing, called the human microbiota. The microbiota is not limited only to the bacterial domain of life but also includes archaea, viruses as well as eukaryotes such as protozoa and fungi (D'Argenio, 2018). Although initially it has been suggested that bacterial cells outnumber human cells by a factor of 10, recent studies tuned down this ratio to 1:1 (Sender, Fuchs, & Milo, 2016a, 2016b). Still, we are made at least half of bacteria and these cells colonize almost every surface of the body. They can be characterized as beneficial symbiotes, harmless commensals or pathogenic parasites (Stilling, Dinan, & Cryan, 2014) and are crucial for immunologic, hormonal and metabolic homeostasis of the human host.

For years, the research community has accepted the “sterile womb paradigm” stating that humans were born sterile and that microbial colonization began immediately at birth. According to this scientific consensus, the microbial acquisition, during and after birth, happens both vertically (from the mother) and horizontally (from other humans or the environment) (Perez-Muñoz, Arrieta, Ramer-Tait, & Walter, 2017). Recent studies challenge this dogma by suggesting that infants incorporate an initial microbiome before birth, reporting findings of microbial colonization in the placenta, the amniotic fluid and the umbilical cord blood (D'Argenio, 2018), (Collado, Rautava, Aakko, Isolauri, & Salminen, 2016); (Kundu, Blacher, Elinav, & Pettersson, 2017); (Jiménez et al., 2005)). However, these findings have been criticized as highly controversial and inconclusive (Perez-Muñoz et al., 2017), thus further studies are required to address this issue.

## 1.1 The human gut microbiome

During early stages of life, the microbial colonization of the gastrointestinal tract is rapid and is considered as vital for immunological and physiological development. The infants' microbiome composition is characterized by low diversity (Thursby & Juge, 2017) and can be shaped by various environmental factors, with one of the most intensively investigated factor being the mode of delivery. The microbiome of infants born with vaginal delivery resembles the one of their mothers' microbiota, while those delivered by cesarean section are deprived of contact with the birth canal (D'Argenio, 2018; Salminen, Gibson, McCartney, & Isolauri, 2004; Thursby & Juge, 2017). It has been described that caesarean section or early exposure to antibiotics are associated with the development of asthma, inflammatory bowel disease and obesity (Collado et al., 2016). During the following few years of childhood, the microbial diversity increases and becomes more stable, converging to a distinct adult-like microbial profile (Kundu et al., 2017; Thursby & Juge, 2017).

The adult gut is the most densely populated habitat on the human body, with an estimated 0.2 kg of microbial biomass (Sender et al., 2016b). Its composition is relatively stable featured by increased microbial richness and complexity. In fact, it is commonly stated that this stable part of the adult microbiota is almost similar across adult healthy individuals, because of the development of a core community of permanent colonizers that protect from exogenous insults like stress and exposure to antibiotics and help restore the original microbial configuration upon these changes (D'Argenio, 2018; Kundu et al., 2017). Notably, this shared microbial composition is present at phylum level, while a significant variability has been described at lower phylogenetic levels (i.e. at species and strain levels) suggesting that "the gut microbial composition is unique in each individual, even if the microbial functions as a whole result the same" (Kundu et al., 2017). Hence the concept that a 'healthy' microbiome is defined by an ideal set of specific microbes is no longer a practical definition, introducing the alternative hypothesis of a healthy "functional core", consisting of a set of metabolic and other molecular functions that are performed by the microbiome but are not necessarily provided by the same organisms in different people (Lloyd-Price, Abu-Ali, & Huttenhower, 2016)

More recently, it has been hypothesized that the adult gut microbiota arrangements can be classified into distinct 'community types', called 'enterotypes' (Arumugam et al., 2011; Costea et al., 2017). The proposed enterotypes were identified by their enrichment in the *Bacteroides* genus (enterotype 1), the *Prevotella* genus (entero-type 2) and the *Ruminococcus* genus (enterotype 3), and were unrelated to nationality or host characteristics such as body mass index, age or gender. These were proposed as a useful method to stratify human gut microbiomes, since subjects were found to be dominated by

one of three different taxa, and this classification could in turn be used to guide diagnostics and treatment options. However, evidence surrounding the existence and formation of these enterotypes has been highly controversial, raising a 'topic of heated debate' after the publication of the original work by Arumugam et al. (2011), thoroughly reviewed by Jefferey (2012) and Knights et al. (2014). Indeed, numerous studies thereafter aimed in examining the hypothesis using more careful clustering analyses. Overall, these additional studies demonstrated that support remained for only two community types: one dominated by *Prevotella* and one dominated by *Bacteroides* or members of the Firmicutes (Gorvitovskaia, Holmes, & Huse, 2016).

Even though in adulthood the gut composition is considered fairly stable, it is still amenable to changes associated with life events. Indeed, there is a wide range of factors that are able to alter the structure of the adult gut microbiota, including age, nutrition, host genetics, hormonal status, physical activity, lifestyle, smoking, depression, climate and geographical location (D'Argenio, 2018; Kundu et al., 2017; O'Hara & Shanahan, 2006; Shapira, 2016; Yatsunenkov, Rey, Manary, & Trehan, 2012). Recently, a separation of the adult gut microbiome into two components has been hypothesized to solve this discrepancy. One part being the core microbiome is considered stable and merely influenced by external stresses, while the other being more flexible is responsible for the microbial plasticity in different environmental changes (D'Argenio, 2018).

The vital role of the gut microbiome is highlighted in numerous studies, emphasizing its involvement in healthy status acquisition and maintenance. The gastrointestinal microbiota is important for the development of the intestinal mucosal and systemic immune system (Thursby & Juge, 2017), the regulation of immunological function, as well as the establishment of physiological epithelial homeostasis (Ewald & Sumner, 2018; Thursby & Juge, 2017). Moreover, it provides a natural defence mechanism against invading pathogens. The physical presence of beneficial bacteria is able to prevent excessive colonization by pathogenic organisms, by competing for mucosal attachment sites or essential nutrients and/or oxygen and by producing of peroxides, antimicrobials, or bacteriocidins to inhibit other bacteria (Ewald & Sumner, 2018; Jarchum & Pamer, 2011). The gut microbiota may also stimulate host responses via its structural components and metabolites (Thursby & Juge, 2017).

In addition, it is also intimately involved in digestion and metabolism of food. Intestinal bacteria are essential for the extraction of nutrients from indigestible components of food. Diets high in plant polyphenols and polysaccharides, which are naturally resistant to digestion and absorption in the small intestine, are fermented by bacterial carbohydrate-active enzymes into small phenolic compounds and short-chain fatty acids (SCFAs) (Ewald & Sumner, 2018). These SCFAs are rapidly absorbed by the epithelium and participate in various molecular functions including gene expression, chemotaxis,

differentiation and apoptosis (Thursby & Juge, 2017). The gut microbiota is also important for the *de novo* synthesis of essential vitamins, since humans lack the biosynthetic capacity for most vitamins (LeBlanc, Milani, de Giori, & Sesma, 2013). Lastly, studies have shown that bacteria are capable of producing highly specific metabolic products and growth factors, that induce genes in epithelial cells to digestive enzymes that are required for physiological digestive processes (Ewald & Sumner, 2018).

The human health status may be disturbed by microbial alterations that are able to impair these essential functions. Changes in the gut microbiome composition have been related to a number of diseases including obesity, diabetes, ulcerative colitis, inflammatory bowel disease, cancers, stress, and even neurodegenerative disorders (Kundu et al., 2017; Qin et al., 2010; Turnbaugh et al., 2006). It becomes clear that our comprehension of the human physiology needs to take into account also these aspects and how these microbial groups interact with one another and with the human host. Fortunately, our ability to investigate the breadth of the gut microbiota has gravely improved the last decade, with the advent of culture independent approaches such as high-throughput sequencing methods.

## 1.2 Metagenomics

Prokaryotes form diverse communities consisting of a multitude of species. Just a fraction of these species has been cultivated in the laboratory, studied experimentally and has a known genome sequence (Strous, Kraft, Bisdorf, & Tegetmeyer, 2012). Advances in high-throughput sequencing approaches have enabled genomic analyses of all microbes in a sample, not just those that are amenable to cultivation. Metagenomics refer to the direct sequencing and analysis of DNA obtained from complete microbial communities with an environmental sample (Quince, Walker, Simpson, Loman, & Segata, 2017).

Advances in sequencing technologies and biocomputing enable the genomic study of the uncultured part of human-associated microbial communities and have considerably shaped the way metagenome research is performed. Next-generation sequencing (NGS) refers to modern post-Sanger sequencing technologies that allow sequencing of millions of small fragments of DNA (i.e. reads) in parallel (Behjati & Tarpey, 2013; Strous et al., 2012). There is a number of different available NGS platforms which provide rapid, effective and low-cost genomic characterization. The Illumina platform is an increasingly popular choice in metagenomics owing to its wide availability, very high outputs and high accuracy (with a typical error rate of 0.1–1%) (Quince et al., 2017).

There are two commonly used NGS metagenomic approaches to explore the genetic diversity of microbiota: targeted metagenomic sequencing and whole-genome shotgun (WGS). Commonly used for the identification of bacterial and archaea species is targeted sequencing studies of the 16S rRNA locus. The 16S ribosomal RNA gene has been used as

a reliable molecular clock because it contains both slowly evolving regions that can be used to design broad-spectrum PCR primers and fast-evolving regions that can be used to taxonomically classify organisms (Kuczynski et al., 2011). However, this powerful tool is not without limitations. Sequencing errors and chimeric reads, due to biases associated with PCR amplification, degrade the accuracy of the analysis, which is typically limited to taxonomic and phylogenetic studies of microbial samples. Still, there are cases where the estimation of the microbial community is not conclusive. For example, for *B. pseudomallei*, *B. thailandensis*, and closely related *Staphylococcus* species, the 16S rRNA gene sequences cannot be used to distinguish their taxonomic identity (Woo, Lau, Teng, Tse, & Yuen, 2008).

WGS metagenomics provide access to the complete genomic content of a human-associated microbial sample by random sampling all genomes present in that habitat. Here, DNA from all cells in such sample is extracted and broken into small fragments that are independently sequenced (Sharpton, 2014; Wang, 2016). Compared with 16S rRNA-based investigations, shotgun sequencing allows taxonomic identification with a higher resolution and provides insights into functional characteristics of microbes, as well as their biological processes and metabolic potential (Kuczynski et al., 2011; Quince et al., 2017).

Shotgun metagenomics produces a great volume of complex data that is very challenging to analyze. A metagenomic dataset, with a volume of several orders of magnitude larger than a single genomics NGS dataset, contains DNA fragments sequenced from an unknown number of species. In many cases, based on the complexity of the microbiome under study, it is difficult or even impossible to separate this large number of produced reads and to determine the genomic origin of each read. Also, because of the tremendous community diversity contained in a metagenome, it is hard to capture the full potential of the DNA pool, and thus many genomes might not be completely represented by reads. To further complicate the situation, many metagenomic computational approaches rely on reference genomes. However the reservoir of available microbial genomes is biased toward model organisms, pathogens and easily cultivable bacteria. In conclusion, this complex structure of metagenomic data and more potential experimental biases (Quince et al., 2017) are complicating the informatics analysis of diverse microbiomes, like the human gut microbiota, as well as the biological interpretations.

These limitations challenge bioinformaticians to distill these huge and complex data into valuable information. Yet, there is active research in the development of an entirely discrete set of computational tools and pipelines, designed to address these unique characteristics of metagenomic datasets. These promising improvements in bioinformatics make the opportunities of metagenic studies seem vast. With the majority of the microbes not being able to grow and obtained in pure culture, there is still an enormous amount of microbial variability to be captured. Now, we have access to the compositional and functional profiles of these communities. When studying microbiomes obtained from diverse communities , rather than isolated populations, sampled directly from their natural habitat, we can have



insight into the community biodiversity and structure, as well as the interactions between its members. This way, we are finally able to understand our relationship with all those microbes that live inside our bodies, how they affect our well-being and health and investigate the causal interactions of our habits (e.g. stress, nutrition or drug usage) with the microbiota composition and function. Finally, metagenomics can be complemented with metabolomic, metatranscriptomic and metaproteomic approaches to investigate the functional potential of a community (Oulas et al., 2015).

## 1.3 The Human Microbiome Project

There is extensive research in exploring the vast amount of microbial variability that characterize the human microbiome, with the pioneering work of the Human Microbiome Project notably influencing the field. The US National Institutes of Health funded Human Microbiome Project (HMP) Consortium is a large-scale initiative to explore the microbial communities associated with the human body and their role in human health and disease. HMP is not a single project, but rather a summation of multiple projects that were launched all over the world, including the USA, the European Union, and Asia (Micah, Claire, Rob, & Others, 2007; National Research Council, Division on Earth and Life Studies, Board on Life Sciences, & Committee on Metagenomics: Challenges and Functional Applications, 2007, p. 117). This international effort was divided in two separate phases.

Launched in 2008, the pilot phase of the project, HMP1 focused on an initial characterization of the normal microbiota of healthy adults in a Western population. This interdisciplinary initiative characterized the microbial communities from 300 healthy individuals, across several different sites on the human body: nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract (Human Microbiome Project Consortium, 2012a). The analyses performed on this large cohort included: 16S rRNA gene sequencing, whole genome shotgun (WGS) metagenomic sequencing, de novo assembly of a subset of the available WGS samples, and alignment of the assembled sequences to reference microbial genomes found in the human body. The HMP1 Data Browser provides access publically available resources like: clinical specimens, reference genomes, sequencing and annotation protocols, methods and analyses.

The second phase, known as the Integrative Human Microbiome Project (iHMP) was established in 2014 in order to examine the role of the microbiome in human health and disease through a study of three cohort datasets of microbiome-associated human conditions. The iHMP encompasses three projects: (1) pregnancy & pre-term birth; (2) onset of inflammatory bowel disease (IBD); and (3) onset of type 2 diabetes. Study methods included 16S rRNA gene sequencing, WGS metagenomic sequencing, metatranscriptomics,

metabolomics and proteomics (Human Microbiome Project Consortium, 2012b; Integrative HMP (iHMP) Research Network Consortium, 2014).

## 1.4 Exploring the Human Microbiome

Results from the HMP, as well as previous studies (Integrative HMP (iHMP) Research Network Consortium, 2014), shed light on the enormous genetic variability of the human microbiome revealing that the taxonomic composition of the microbiomes can vary significantly between subjects. The typical steps involved in such sequenced-based metagenomics project are: (1) the experimental pipeline, (2) the quality control and the pre-processing of the sequencing reads, (3) the sequence analysis to profile taxonomic, functional and genomic features of the microbiome and (4) the statistical and biological post-processing analysis (Wooley, Godzik, & Friedberg, 2010). In this introduction, we will not be covering the first step, that involves the sample processing, the DNA extraction and sequencing, because this field falls outside the scope of this thesis. Nevertheless, we need to emphasize that this is probably the most crucial step in any metagenomic project, since experimental artifacts challenge the analysis and the interpretation of the data. The 2<sup>nd</sup> and 4<sup>th</sup> steps are discussed in further detail in the 'Methods' section and here, we shall introduce a brief overview of the bioinformatic approaches applied in the 3<sup>rd</sup> step of the analysis. In our study, we are primarily interested in investigations exploring the taxonomic diversity of a microbiome and we won't be reviewing the functional analysis.

Taxonomic profiling of a microbial community involves the identification of the organisms present in the metagenomic sample and the estimation of their abundances. The taxonomic diversity of a microbiome can be quantified by (1) assembling reads into genomes, (2) 'binning' reads into distinct taxonomic groups or (2) using taxon-informative marker genes. These approaches are not mutually exclusive and may work in synergy (Sharpton, 2014; Thomas, Gilbert, & Meyer, 2012), while they also share a common feature: they all rely, on some degree, on reference sets of genomes to assign taxonomic labels to the reads. This available knowledge provided by already sequenced genomes is necessary since there is no prior taxonomic information provided by the sequence itself (Izard & Rivera, 2014, Chapter 5). Therefore, we shall review the 3 main approaches in an order based on how directly they make use of reference information.

### 1.4.1 Assembly

Genome assembly merges overlapping sequence reads into contiguous sequences, called contigs, in order to reconstruct the original sequence (Kunin, Copeland, Lapidus, Mavromatis, & Hugenholtz, 2008) and can be useful in recovering the genome of uncultured organisms. The two computational strategies are: reference-based assembly and *de novo* assembly. The first approach relies on the use of available reference genomes as a

“backbone” in order to create contigs representing a complete or nearly-complete genome, that is similar but not necessarily identical to the mapping sequence. These algorithms are fast and computationally efficient and may perform well in metagenomic studies, when the dataset is originated from extensively studied areas with a rich diversity of reference genomes (e.g. the human gut) (Human Microbiome Jumpstart Reference Strains Consortium et al., 2010; Thomas et al., 2012). *De novo* assembly includes sophisticated graph theory algorithms that require much larger computational resources. For single draft genome assemblies, a de Bruijn graph is constructed by breaking each read into overlapping subsequences of a fixed length  $k$  (i.e.  $k$ -mers) and then by modeling the contiguous sequence overlap in order to reconstruct the genome (Compeau, Pevzner, & Tesler, 2011). However, even when assembling a single genome, sequencing errors and repetitive elements confuse the assembler and can lead to miss-assemblies and fragmentations (Quince et al., 2017).

Reconstructing a full genome is a complex task even for isolate genomes. In such studies, assemblers assume an approximately uniform sequencing coverage all across the genome that is used to distinguish true sequence from sequencing errors. This difficulty is significantly compounded in metagenomic studies when the coverage of each constituent genome depends on the abundance of each genome in the community. Low-abundance organisms cannot be recovered because the assemblies might end up fragmented if overall sequencing depth is insufficient to form connections in the graph (Quince et al., 2017). Additionally, wide differences in coverage, caused by variably abundant organisms, makes it hard to identify genomic repeats in the reconstructed sequences. Finally, true differences between closely related organisms cannot be easily distinguished from sequencing errors and this can even lead to the generation of chimeras (i.e. contigs assemblies from distinct genomes into one, because of shared sequence similarity) (Izard & Rivera, 2014, Chapter 4; Sharpton, 2014).

Despite these challenges, early metagenomic studies relied on assembly tools developed for isolate genomes (Izard & Rivera, 2014, Chapter 5), like SOAPdeNovo in the HMP (Human Microbiome Project Consortium, 2012a). Recently, a number of metagenome-specific assemblers have been developed (Li, Liu, Luo, Sadakane, & Lam, 2015; Namiki, Hachiya, Tanaka, & Sakakibara, 2012; Peng, Leung, Yiu, & Chin, 2011) that generally utilize graph-based reconstruction algorithms which are adapted in addressing these problematic features of metagenomic data. Nevertheless, it's hard to say which is the most accurate tool (Bradnam et al., 2013) since performance depends on various factors, like the underlying microbial community structure and complexity of even the sequencing platform characteristics and coverage (Quince et al., 2017). Usually, it's extremely hard to recover the complete genomes of the more dominant members of the community, and the output assemblies are highly fragmented. Further analyses are necessary to identify the sets of contigs that belong to a same genome.

### 1.4.2 Binning

Binning is the process of clustering DNA sequences into groups that might represent an individual genome, or genomes from closely related organisms (Thomas et al., 2012). This process can be applied on sequencing reads or assembled contigs, even though most tools claim that the algorithm accuracy improves as sequence lengths increase (Sharpton, 2014). There are two distinct binning strategies: Composition-based binning and similarity-based binning. The first category of methods, also known as unsupervised binning, compares intrinsic sequence properties to group sequences into taxonomic classes. These characteristics are used as signatures and include variations in GC-content, codon usage bias, and the distribution of k-mers of variable length (Alneberg et al., 2014; Izard & Rivera, 2014, Chapter 5). The sequences can be separated based on the theoretical assumption that sequences from the same genome will have similar coverage values within each sample. However in principal, intra-genome GC content variation and increased read depth around bacterial origins of replication can challenge this (Quince et al., 2017). The second group of methods, supervised binning, utilize information from external sequence data resources to classify sequences into taxonomic groups. The reference genomes provide a more accurate substrate for identifying the taxonomic affiliation of each sequence (Izard & Rivera, 2014, Chapter 4), while in general, this approach is computationally more expensive but characterized by greater robustness (Strous et al., 2012).

### 1.4.3 Marker gene analysis

The whole length of reference genomes can arguably be characterized at best uninformative for taxonomic assignment, owing to the presence of evolutionary conserved sequences or occasionally even misleading, because of horizontally transferred genes. By preprocessing the available references to remove redundant and non-discriminating sequences, it is feasible to focus on the most taxonomically informative markers (Izard & Rivera, 2014, Chapter 5). Metagenomic sequences are compared to a database of such signature gene families (i.e., marker genes) to estimate their relative representation (Sharpton, 2014).

The most frequently used marker genes can be classified in two categories: universal markers and clade-specific markers. The characteristics of members of the first class is that they are common in all microbial genomes while they possess variable regions that can be exploited as taxonomic or phylogenetic tags (Izard & Rivera, 2014, Chapter 5). Probably the most traditional example of a universal marker is the 16S rRNA gene. Yet its reliability as a marker has been criticized because there is evidence of horizontal gene transfer in such regions (Wooley et al., 2010), while the presence of variable copy numbers in bacterial genomes may trouble abundance estimation (Louca, Doebeli, & Parfrey, 2018; Větrovský & Baldrian, 2013). Alternative markers, like single-copy housekeeping genes may lead to more

accurate estimations of taxonomic abundance. These markers also assure greater robustness, since the housekeeping functionality makes them less susceptible to horizontal gene transfer. Clade-specific markers on the other hand are unique fingerprints of each microbial clade, defined as core genes that share no sequence similarity with genomic regions in any other clade (Huang et al., 2014; Wooley et al., 2010). For example MetaPhlAn2 (Truong et al., 2015) (see methods) utilizes a set of about a million markers spanning all the 3 domains of life (i.e. archaea, bacteria and eukarya) for profiling the composition of complex microbial communities, providing high accuracy, quantitative estimation, and deeper resolution than 16S rRNA investigations (Quince et al., 2017).

Researchers make great effort for the cautious identification of marker genes that represent the overall diversity of the tree of life. Under the current genomics revolution, the substantial cost reduction in NGS has dramatically increased the number of available reference genomes (Steinberg et al., 2017), suitable for the extraction of marker genes. Extensively researched areas, like study efforts on human health and biotechnology, have larger contributions in existing databases. Indeed, microbiomes with such rich diversity of available reference genomes (e.g. the human gut microbiome) can be adequately and efficiently profiled. More diverse environments, on the contrary, are under-represented in public databases (Choi et al., 2017). Therefore, an assembly-based approach is advisable in environments like soil and oceans (Quince et al., 2017).

#### 1.4.5 A daunting choice

It becomes apparent that assessing the taxonomic diversity of a microbial community remains a challenging task. Each methodology possesses unique strengths and weaknesses. The success of either one is, in general, highly dependent on the underlying microbial composition and complexity, the sequencing depth, the volume and richness of the generated data as well as the available computational resources (Quince et al., 2017). There is no single path to a successful strategy. During the crucial step of the study design the researcher needs to take into careful consideration the features of the targeted microbiome, the characteristics of the experimental steps, and by all means the project objective, in order to choose the most appropriate analytical approach. Even though assembly-based methods usually have an advantage when exploring novel environments that include a large proportion of previously unknown microorganisms, they may miss low-abundance organisms in the microbiome due to lack of sufficient sequencing coverage and depth (Quince et al., 2017). Metagenomic assembly is also impractical for large and extremely diverse datasets. In contrast, marker-based approaches by preprocessing the reference sequences to reduce their size and increase their discriminating power, are characterized by fast, computationally efficient and accurate results while they can better capture low-abundance members of the community that are particularly challenging to assemble (Izard & Rivera, 2014, Chapter 5). An assembly-free approach is arguably more suitable for well-characterized environments with a substantial fraction of microbial diversity covered by reference genomes. In fact, the

representation of human-associated microbiomes, like those colonizing the human gut, in reference genome databases is rather extensive for a successful profiling using a marker-gene strategy.

Regarding the methodology selection, in some situations, the combination of the discussed strategies is another option since they are not mutually exclusive. For example, it is common practice to perform composition-based binning on assembled datasets (Oulas et al., 2015), whereas in complex communities researchers often bin reads and then assemble independently each refined genome bin, so they may mitigate the risk of generating assembly chimeras (Luo, Tsementzi, Kyrpides, & Konstantinidis, 2012; Sharpton, 2014). In other cases, a pipeline of assembling, binning and re-assembling is also implemented. (Sangwan, Xia, & Gilbert, 2016; Sharon et al., 2013). Even many widely-used assemblers, like MetaVelvet and Meta-IDBA (Namiki et al., 2012), employ a combined binning and assembly approach.

## 1.5 Population genomics in the gut microbiome

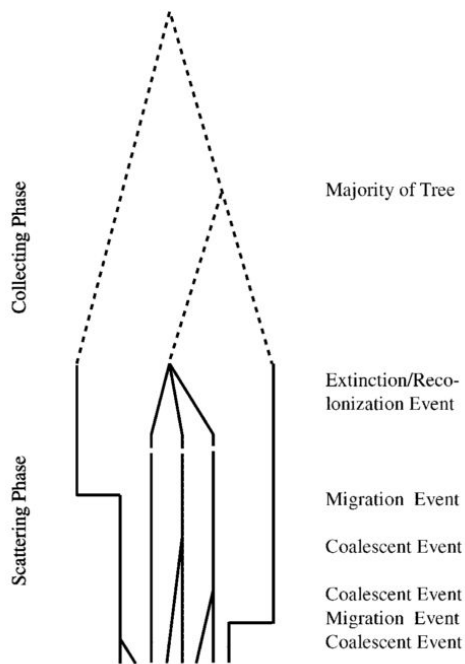
The human microbiome is a diverse ecosystem characterized by increased genetic variation and rapid evolution (Shapira, 2016). Metagenomic studies can explore the genomic composition of these natural ecosystems by sampling their diversity across time points and/or replicates. Bringing these investigation to another level, theories of population genetics can be employed to explain the observed genetic variation and understand how it is influenced by evolutionary processes. Indeed, combinations of molecular and ecological approaches have been applied to study the evolution of host microbiomes in order to interpret features like colonization resistance, host nutrition or immune development (Foster, Schluter, Coyte, & Rakoff-Nahoum, 2017).

The advent of NGS technologies has revolutionized the field of microbial ecology and can be utilized as an innovative and rather promising way of addressing fundamental questions of evolutionary biology. Population genomics applies established population genetic methodologies to whole genome sequence data to improve our understanding of the evolutionary forces that affect variation across genomes and populations (Luikart, England, Tallmon, Jordan, & Taberlet, 2003). Increasing our investigation resolution, it enables the separation of evolutionary forces that alter individual loci (e.g. mutation, recombination, selection) from those that influence the genome as a whole unit (e.g., population bottlenecks, genetic drift). Genome-wide changes reliably inform us about population demography and phylogenetic history, whereas locus-specific effects more likely reflect selective pressures underlying bacterial adaptation (Ashley Robinson, Feil, & Falush, 2010, Chapter 7).

The development of the coalescent theory (Kingman, 1982) has strongly influenced many areas of population genetics, becoming a central concept for the study of variation at a sequence level. This is a mathematical model for the description of gene genealogies used as a standard framework for the retrospective statistical analysis of genetic data (Ashley Robinson et al., 2010, Chapter 1). Genealogies are family trees which depict the ancestors and descendents of individuals in a population and can be quite informative about historical demography and the processes that have acted to shape the diversity of populations (Salemi, Vandamme, & Lemey, 2009, Chapter 17). During a retrospective inference we aim to understand the past of a population through analysis of present-day sequences. Coalescent theory can be used to make inferences about population genetic parameters, such as mutation, recombination or demographic parameters.

The coalescent is a continuous-time approximation to the Wright–Fisher model for large populations. This basic model describes an idealized panmictic haploid population of constant size  $N$  and represents the transmission of genes from one generation to the next. Generations are non-overlapping and there are no selective forces acting on the population, while all individuals have an equal chance of producing offspring (Salemi et al., 2009, Chapter 17). Of course, most natural populations fail to satisfy one or more of the assumptions. Yet with the Wright-Fisher model we can study how introducing more complex evolutionary forces can affect a simple model (Hein, Schierup, & Wiuf, 2004, Chapter 1). This model can be used to generate many theoretical genealogies, and then compare observed data to these simulations to test assumptions about the demographic history of a population.

Microbial species colonizing human hosts are subdivided in geographically distant habitats (i.e. the gastrointestinal tract of different individuals) forming a complex metapopulation as a whole. A metapopulation describes a population that is subdivided into a large number of discrete demes, each of which is subject to random extinction and recolonization (Levins, 1969). A simple genealogical process exists for samples from a metapopulation and Wakeley (2001) has shown that, when the number of demes is large, the genealogy includes two phases (Fig. 1.1). During the ‘scattering phase’ sample genealogies exhibit a recent burst of coalescent events among samples taken from the same locality. The ‘collecting phase’ is a Kingman-type coalescence process that describes a more ancient historical process for the remaining ancestral lineages.



**Figure 1.1 |** Example of scattering and collecting phase of a metapopulation coalescence. Reprinted from “Gene genealogies in a metapopulation”, by Wakeley, J., & Aliacar, N. (2001) *Genetics*, 159(1997), 893–905. An example of a genealogy of sample size eight from a single deme. In this case, during the scattering phase, there are two migration events (to some unoccupied demes that are not pictured) and then an extinction/recolonization event with  $k = 2$  in which all of the lineages remaining in the deme are descended from a single common ancestor. The coalescent collecting phase of the three resulting lineages is shown above. The relative duration of the scattering phase is greatly exaggerated for purposes of illustration.

## 1.6 Demographic inference

Coalescent theory forms the basis for likelihood calculations in genealogical models, and enables the application of Bayesian approaches to infer demographic parameters (Beaumont & Rannala, 2004).

### 1.6.1 Bayesian statistics

Bayesian statistics have gained a rising popularity in population genetics and in demographic studies in particular. This method for statistical inference provides a framework for summarizing the uncertainty of each parameter in probability distributions which represent one's ‘degrees of belief’ (Yang, 2014, Chapter 6). The prior distribution of the parameter reflects our uncertainty about the parameter before the analysis of the data while the posterior distribution represents our updated beliefs after examining the data. In a Bayesian analysis our interest is calculating the posterior distribution (i.e. the conditional probability distribution of a parameter  $\theta$  given the observed data  $X$ ,  $P(\theta|X)$ ) using the Bayesian rule:

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)},$$



where  $P(\theta)$  is the prior distribution,  $P(X|\theta)$  is the likelihood of the data and  $P(X)$  is the marginal distribution which represents the probability distribution of the data irrespective of the parameters and acts as a normalizing factor.

Thus, the posterior distribution can be obtained as the product of the prior and the likelihood and is used to infer the value of the parameter  $\theta$ . One common method in summarizing this distribution is providing point estimates of  $\theta$ , such as the mean, the median or the mode of the distribution (Shoemaker, Painter, & Weir, 1999). For interval estimation, one can use a region in which the parameter value is expected to be located, such as the 2.5% and the 95.5% quantiles of the posterior density.

Summing up, the following steps can describe the necessary elements in a Bayesian data analysis (Glickman & van Dyk, 2007): (1) formulating a model, (2) quantifying the uncertainty regarding the parameter in the form of a prior distribution. This expresses our knowledge and experience gained from past relevant evidence, (3) constructing the likelihood function based on the chosen model on the 1st step and given the observed data, and finally (4) estimating the posterior distribution, by combining information from the prior distribution and the likelihood of the data, and summarize the result using point or interval estimates

### 1.6.2 Approximate Bayesian Computation (ABC)

Even though Bayesian approaches have been proven especially valuable in advanced genetics problems, many of these methods are limited by the difficulty of analytically defining likelihood functions for complex population genetics models or realistically large datasets. Approximate Bayesian Computation approaches bypass the exact likelihood computation by using stochastic simulations and summary statistics. The approximation of the likelihood distribution is performed by fitting a local-weighted linear regression of simulated parameter values on simulated summary statistics, and then substituting the observed summary statistics into the regression equation (Beaumont, Zhang, & Balding, 2002).

Specifically, the basic ABC-rejection algorithm is typically implemented in the following steps. Firstly, a large number of simulations is generated. The parameters of the model in each simulation are not chosen deterministically, but rather are drawn from prior distributions that are predefined by the researcher. The generated data are then reduced to summary statistics and compared to the observed data. The sampled parameters are accepted if the vector of statistics is sufficiently close to the observed summary statistics, with respect to some metric in the space of used summary statistics. Finally, the posterior distribution of each model parameter is approximated by a local regression adjustment on the statistics and the fitted parameters of the accepted simulations (Csilléry, Blum, Gaggiotti, & François, 2010; Elleouet & Aitken, 2018).

An important approximation in such analysis, is the replacement of the full data by a set of summary statistics. These are numerical values calculated from the data, so that they represent the maximum amount of information in the simplest possible form (Csilléry et al., 2010). They have a wide-spread use in population genetics correlating various demographic parameters like population sizes, migration rates and time of population changes while, typically, can be classified in 4 groups describing (Laurent, 2011): (1) the amount of genetic variation, (2) the shape of the polymorphisms frequency distribution, (3) the linkage disequilibrium, measuring the non-random association of alleles in a chromosome (Doris, 2002) and (4) the amount of genetic differentiation between individuals sampled from different populations. The selection of summary statistics is a critical part in an ABC analysis, significantly influencing the acceptance rate of the algorithm. One approach to capture most of the information present in the data would be to increase the number of statistics, but this aggravates the statistical noise included in the posterior estimation (Joyce & Marjoram, 2008), decreasing the accuracy and stability of ABC (Sunnåker et al., 2013). Instead, a better strategy is focussing on more representative sets of summary statistics. In fact, there is active research in improving methods for model testing and the associated choice of summary statistics (Elleouet & Aitken, 2018).

Another essential element of an ABC analysis, as in Bayesian inference in general, is the specification of the priors. A prior should be chosen wisely to reflect our degree of belief about the parameter before the collection and analysis of the data. It can incorporate external information gained in past experiments about the biological system, such as mutation rates, recombination rates, dates of demographic events based on fossil records or other ecological informations (Laurent, 2011). Unfortunately, conclusion may be sensitive to the choice of priors (Sunnåker et al., 2013), thus the selection must be cautious.

### 1.6.3 Inferring the demographic history of gut bacterial populations from genetic variation data

Molecular data contains a significant amount of information about the demographic processes that have affected the evolution of natural populations. Researchers attempt to uncover events in the history of populations that generated the present data. However, demographic models have many parameters (e.g. the effective population size, population growth/decline rates, population structure) and the complicated likelihood functions challenge the inference. ABC analysis demonstrate valuable strengths dealing with these issues, establishing itself as a prevalent choice in demographic inference.

The population structures of bacteria species are extraordinarily diverse. Indeed, bacterial populations demonstrate extensive demographic variations across space and time, such as frequent expansions and bottlenecks (Lapierre, Blin, Lambert, Achaz, & Rocha, 2016).

Exploring the demographic dynamics of the human gut microbiota will improve our understanding of its biology and health impact. For example, bacterial populations can go through dramatic bottlenecks due to drug usage or during shifts from one host to the next for pathogenic or commensal species (Ashley Robinson et al., 2010, Chapter 1). Arguably, the characterization of such demographic changes in gut microbial populations would be of great interest in itself. For example, investigating changes among populations of infectious agents could inform epidemiological studies and guide public health interventions (Lapierre et al., 2016; Rocha, 2018).

Demographic processes affect the possible patterns of genetic variation in a population in a similar way with selective events. Unfortunately, most methods aiming at identifying demographic changes are not robust to the presence of selection nor, sometimes, recombination. Disentangling the effects of the two processes poses a serious challenge and the application of the existing methods of dissociating the two has been criticized as inappropriate when it come to bacterial populations (Rocha, 2018) . Likewise, variations in population size of a species have an important impact on the shape of the phylogenetic tree, by altering the patterns of genetic diversity.

## 1.7 Phylogenetic inference

In order to understand and predict the dynamics of the human gut microbiota, it is necessary to unravel how bacterial species are genetically structured on a population scale. By far the most widely used method to describe population structure is phylogeny. If one can reconstruct the correct phylogeny for a given sample, then the evolutionary relationships between its members become apparent. Molecular phylogenetic analysis, in particular, establishes the relationships between genes or gene fragments, by inferring their common history. This evolutionary history is depicted in a mathematical structure, called a phylogenetic tree (Rokas, 2011) and can be inferred using various computational approaches.

There are numerous available methods for reconstructing phylogenetic trees from molecular data (Salemi et al., 2009, Chapter 1). These can be grouped according to the type of data they use: character-base methods handle discrete character states while distance-based methods rely on a measure of "genetic distance" between the sequences being classified. The Neighbor Joining algorithm as well as the UPGMA method (abbreviation for unweighted pair group method with arithmetic mean) are the most notable representatives of the second class. The first class consists of methods that apply the maximum parsimony criterion or maximum likelihood in order to infer phylogenies, along with bayesian inference methods.

In recent decades, the field of phylogenetics has found applications in the analysis of genomic scale data. With the rapid improvement of high-throughput sequencing platforms and the increase of sophisticated computational tools, the analysis of hundreds to thousands of loci has become routine. Genome-wide information has been extensively utilized to resolve complex phylogenetic problems, creating a new area of research, named phylogenomics.

## 1.8 Study overview

In the present study, we explored the gut microbial composition of 137 healthy individuals from the 1st phase of the HMP. We characterized the genetic structure of thousands of bacterial strains, by reconstructing consensus sequence variants within species-specific marker genes. The estimated taxonomic composition as well as the microbial phylogenetic structure was tested for significant associations with available metadata annotation about the subjects' gender. Next, we focused on modeling and inferring the demography of specific prevalent bacterial species, using approximate Bayesian computation (ABC).

# Methods

## 2.1 Data preprocessing

We were interested in analysing stool samples of healthy individuals from the 1st phase of the Human Microbiome Project, in order to construct the genomic variation landscape of the bacterial species colonizing the human gut. We obtained raw sequence data in fastq format originated from 137 stool samples (available at: <ftp://public-ftp.ihmpdcc.org/HMASM/WGS/stool/>). To assure download integrity, md5sums were also retrieved and confirmed upon download completion. The HMP Consortium performed the sequencing using the Illumina GAIIx platform with 101 bp paired-end reads. The samples were subjected to quality control assessment that included the identification and removal of human reads, the removal of duplicated reads, and the trimming of low quality bases (HMP ref1). Reads trimmed to less than 60 bp were removed, and their partners, if longer than 60 bp were placed in a separate 'singletons' file. In such manner, each sample's reads were organized in three files: (1) <sampleID>.1.fastq, (2) <sampleID>.2.fastq and (3) <sampleID>.singleton.fastq.

The quality control step is an essential part of every NGS analysis that should be carried out cautiously. Quality problems, like low-confidence bases, PCR artifacts, 3'/5' positional bias, untrimmed adapters and sequence contamination, typically originate either in the sequencing procedure itself or during the library preparation (Korpelainen, Tuimala, Somervuo, Huss, & Wong, 2014, Chapter 3). These problems can seriously affect downstream analyses and lead to erroneous conclusions. Thus we decided to proceed with a more strict quality filtering of the preprocessed reads downloaded from the HMP.

We used PRINSEQ (Schmieder & Edwards, 2011) to prepare our dataset prior to the downstream analysis. The standalone graphs version of PRINSEQ was used to review the suggested summary statistics that are commonly used for quality control of NGS data and we only adopt the filtering steps which we reckon that are fitted to our diverse metagenomic dataset (See results). The summary statistics that were selected include base quality scores and the occurrence of Ns and poly-A/T tails. More precisely, using the standalone lite version of PRINSEQ, we removed all reads with mean quality scores of less than 25 with the argument '-min\_qual\_mean 20', and removed all sequences with more than 20 ambiguous bases (N) with '-ns\_max\_n 20'. Sequences were trimmed from both ends using a quality score threshold of 20, using the options '-trim\_qual\_right 20' and '-trim\_qual\_left 20'.

Additionally, all repeats of As or Ts with at least a length of 5 were trimmed from the sequence ends with '-trim\_tail\_left 5' and '-trim\_tail\_right 5'. Finally, after the trimming operations, any sequence shorter than 60 bp was discarded using the argument '-min\_len 60'. The identification of sequence duplicates and contamination with reads of human origin, performed by the HMP Consortium, are also crucial but we trusted that do not need further examination.

## 2.2 Taxonomic profiling

The composition of the microbial communities present in the stool samples was estimated using MetaPhlAn2 (Truong et al., 2015) and the StrainPhlAn module. We are briefly introducing the overall workflow designed by Tryong research team (Truong, Tett, Pasolli, Huttenhower, & Segata, 2017) for the identification of the dominant strain of each detected species in each metagenomic sample. The first step is performed using MetaPhlAn2 which uses a library of species-specific markers spanning the bacterial, archaeal, viral, and eukaryotic phylogenies. These marker sequences are selected so that they are strongly conserved within each species and do not possess substantial sequence similarity with genomic regions in other species (Truong et al., 2015).

Following, the strain-level microbial profiling is performed using the StrainPhlAn module. The reads mapped against the MetaPhlAn2 database are used to reconstruct the consensus sequence of each detected species-specific marker. The consensus is independent from the sequence of the marker used as a backbone for the mapping (strain paper). Some filtering operations are then applied: (i) removal of reconstructed markers with a percentage of ambiguous bases higher than 20% and (ii) trimming markers by removing the first and last 50 bases. A species is considered as present in a sample if the number of corresponding reconstructed markers are more than the 80% of the total number of markers available for that species in the MetaPhlAn2 database.

Finally, the reconstructed marker sequences for each present species in the metagenomic samples are aligned using MUSCLE(ref). The multiple sequence alignment (MSA) is then processed in order to filter poorly covered sequence regions with the default StrainPhlAn settings: (i) trimming of the MSA by removing the first and last bases, until the fraction of gaps in each position is less than 20%, (ii) removal of MSA parts that are present in only 30% fraction of the samples and (iii) removal of alignment columns with ambiguous nucleotides (i.e. "Ns"), if the number of those containing at least one "N" is less than the 80% of the total number of columns. The MSA are concatenated, for each of the the present species, and the remaining "Ns" in the alignment are replaced with gaps. Strains that have gaps in more 20% of the concatenated alignment are excluded from the downstream analysis.

We thus used MetaPhlAn2 to estimate the relative abundance of microbial cells in each sample, by mapping the reads (performed with Bowtie2, version 2.2.8 (Langmead & Salzberg, 2012) against the set of clade-specific marker sequences in the mpa\_v20\_m200 markers database (available at: <http://huttenhower.sph.harvard.edu/metaphlan2>). The clades represented in the taxonomy tables were filtered to include only those with a relative abundance above 0.5% present at least once in the samples. All the complementary data handling and visualization was performed in the R programming environment. For the enterotypes investigation in particular, we used the R packages "cluster", "factoextra" and "magrittr". In the principal components analysis (PCA) the visualization was inspired by Gorvitovskaia's (2016) work, labeling as Prevotella-dominated or Bacteroides-dominated any sample in which either Prevotella or Bacteroides was the most abundant taxon, respectively. All remaining samples were classified together as "Other". The variables were scaled to have unit variance before the PCA analysis took place. The cladogram of the detected taxa was created with the GraPhlAn tool and the "export2graphlan" conversion software.

To increase the taxonomic resolution, we applied StrainPhlAn to the microbial clades identified by the taxonomic profiling tool MetaPhlAn2, in order to reconstruct the specific strain of a given species within a metagenome. StrainPhlAn allows the analysis of any strain with sufficient sequencing depth per sample (i.e. at least 2X coverage for each species' detected marker). More precisely, the output MSAs of the reconstructed markers were used to infer the phylogenies of each species as well as for the demographic analysis. In addition, the reconstructed alignment of each species was used to generate a phylogenetic distance matrix that contains the pairwise nucleotide substitution rate between strains. This was employed with the distmat package from the EMBOSS software, applying the Kimura 2-parameter correction method (Kimura, 1980) to correct the observed substitution rates to more accurately reflect the true evolutionary distance. We applied Multidimensional scaling (MDS) on the distance matrix of each species to create two-dimensional graphical representations. Lastly, The reads-to-markers alignments of the 12 species were used internally from the software, in order to calculate descriptive statistics of the percentage of the polymorphic sites per strain per species.

The phylogenetic inference was performed with RaxM-NG (Kozlov, 2018) utilizing the MSA obtained for each species' concatenated markers. RaxML-NG (Randomized Accelerated Maximum Likelihood- Next Generation) is a phylogenetic tree inference tool which uses maximum-likelihood (ML) optimality criterion. In brief, for a given tree, at each site, the likelihood is determined by evaluating the probability that a certain evolutionary model has generated the observed data. The likelihoods for each site are then multiplied to provide likelihood for each tree. The search for the ML tree for each species was therefore performed individually, using the GTR+G model of DNA evolution and five randomized parsimony starting trees. The GTR model considers that the six possible kinds of

substitutions among the different bases occur at different rates, while the '+G' indicated that there is rate heterogeneity among sites (Salemi et al., 2009). In order to assess how well supported the reconstruction is, one hundred bootstrap replicates were drawn from the input alignment during each RAxML run.

Due to limited public availability of clinical metadata of the HMP samples, outside of dbGaP's authorized access system, the only statistical association we could test was the one between the gut microbial composition and the subjects' gender. So, in order to study the issue of sex-biased differences, a statistical hypothesis test was performed, exploiting the estimated relative abundance of the clades across female and male subjects. In detail, we postulated and tested the null hypothesis proposing that the observed distributional differences in relative abundance of microbial species among the two genders is just due to chance. We had to take into account that in the case of metagenomic data, the normal distribution that is implied in many statistical tests does not properly describe this type of data. In the case of the widely diverse gut microbiota in particular, a typical bacterial species is estimated at zero or low abundance in the majority of samples while only a small fraction of the population carries high abundance of this species (Odintsova, Tyakht, & Alexeev, 2017). To overcome this problem, we choose to apply a non parametric method that makes no assumptions about the underlying distribution. For simplicity, interactions between the individual species was ignored and the abundance is compared individually for each one of the species, resulting in many statistical tests. It was therefore necessary to control for the number of 'false discoveries' by adjusting the P-values via the Benjamini-Hochberg (1995) method that controls the false discovery rate (FDR) (i.e. the expected proportion of false discoveries amongst the rejected hypotheses). Each individual comparison was carried out with the non-parametric Mann–Whitney–Wilcoxon test, while the significance threshold was set at 0.05.

## 2.4 Demographic analysis

We employed an Approximate Bayesian Computation (ABC) approach to investigate the demographic histories of specific bacterial metapopulations colonizing the 137 stool samples of the HMP. A metapopulation describes a population that is subdivided into a large number of discrete demes, each of which is subject to random extinction and recolonization. In our study, the detected bacterial species were subdivided in geographically distant habitats, i.e. the gastrointestinal tract of different healthy individuals. For a given bacterial metapopulation, the scattering phase consists of coalescent events occurring inside each deme, that go back in time until they coalesce into the ancestral lineage of the demes. Each deme is under different host-associated pressures that shape it's evolution. During the collecting phase we are modeling the longer coalescent of each deme's ancestral lineage, representing the common history of the bacterial metapopulation between different hosts.



For our demographic analysis, we focused exclusively on the collecting phase of the coalescent. For each species, we assumed that each subject's reconstructed dominant strain represents the ancestral state of the scattering phase of that deme. This way, we were able to model the history of that bacterial metapopulation between hosts.

### 2.4.1 Summary statistics

For each species metapopulation understudy, the MSA of the reconstructed marker genes produced during the taxonomic profiling step constitutes our 'observed data', while each marker gene represents one independent loci. An important approximation in the ABC framework is the replacement of the full dataset by a set of summary statistics. Therefore, for each loci across the genome of a species, we computed a vector of summary statistics, in the program msABC (Pavlidis, Laurent, & Stephan, 2010), to make use of the coalescent information contained within each sequence. Then for each metapopulation, we calculated the mean and variance values of the summary statistics across all the species' loci.

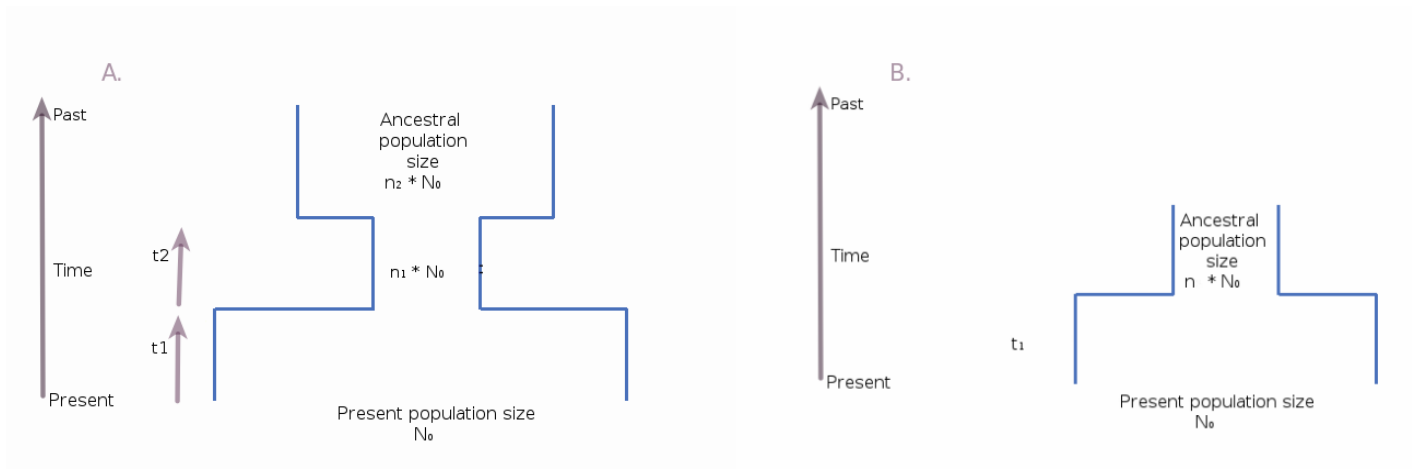
The calculated summary statistics include (1) estimates of genetic diversity: the Watterson's estimator  $\theta_w$  (Watterson, 1975) and the mean pairwise differences of sequences  $\theta\pi$  (Tajima, 1983), (2) summary of the site frequency spectrum in the form of Tajima's D (1989), (3) the average pairwise correlation coefficient  $ZnS$  (Kelly, 1997) as a measure of linkage disequilibrium, (4) two haplotype-based statistics (Depaulis & Veuille, 1998): the number of haplotypes calculated by the Depaulis and Veuille  $K$  ( $DVK$ ) and the haplotype diversity measured by the Depaulis and Veuille  $H$  ( $DVH$ ) and (5) the Thomson estimator of TMRCA and its variance. The mean and variance values of this group of summary statistics constitutes our 'observed vector'. Even though msABC provides the calculation of two additional statistics, we agreed to exclude them from the analysis. Specifically, the number of segregating sites  $S$  was ignored as a summary statistic owing to its correlation with  $\theta_w$ . It won't only have no impact on the inference, but it may add statistical noise in the posterior estimation. Neither did we include the Fay & Wu's  $H$  statistic (Fay & Wu, 2000), since we are unaware of the derived and the ancestral states in the alignments.

A preliminary exploratory analysis of the observed summary statistics was performed using hierarchical clustering. Firstly, we combined the 'observed vectors' of all the species in one table, and the features were normalized with the min/max normalization. A distance matrix was calculated, using the euclidean distance metric. An agglomerative clustering algorithm was implemented using the Ward's minimum variance method (Murtagh & Legendre, 2014), which minimizes the total within-cluster variance. The number of clusters was determined using the gap statistic (Tibshirani, Walther, & Hastie, 2001) with 500 Monte Carlo samples.

## 2.4.2 Simulations

In order to investigate the history of a gut bacterial metapopulation, we created sets of simulations of neutral polymorphism data under specific demographic scenarios. The coalescent simulations were performed using msABC. This program implements a simulation process based on Hudson's ms software (Hudson, 2002). In brief, the generation of the simulation samples is performed assuming the standard coalescent approximation to the Wright–Fisher model. For each sample, the program creates a random genealogical history of each locus and conditional on the genealogy, it adds random mutations on the genealogy according to the assumption that the number of mutations on a branch is Poisson distributed with mean given by the product of the mutation rate and the branch length. msABC enables the simulation of multiple independent loci and the simulation process is described by the ancestral recombination graph and not by a tree, due to the presence of recombination within the loci. Furthermore, it computes a multitude of summary statistics from the simulated data.

This way, inferring the demographic history of one metapopulation at a time, we generated sets of simulations corresponding to a specific demographic scenario. We considered two models, one of instantaneous population growth and one population bottleneck model. Every evolutionary scenario was defined by a set of parameters while every parameter was characterized by a prior distribution. Given the current size of a population  $N_0$ , the expansion model is characterized by three parameters: the scaled population mutation rate  $\theta$ ,  $n$ ; the stepwise population change from  $N_0$  to  $n \cdot N_0$  that occurred at time  $t_1$  (Fig. 2.1.A.). A population bottleneck is defined by:  $\theta$ , the first population change  $n_1$  (as in  $n_1 \cdot N_0$ ) that happen at time  $t_1$  and a second change,  $n_2$ , that occurred later in the population's past, at time  $t_2$  from the first event (Fig. 2.1.B.). Since this demographic scenario includes multiple population size changes, in order to assure the correct order of the events, we used the 'duration mode' on the msABC simulations. This way, we were able to specify the relative time of the second event compared to the previous one, while the duration and not the absolute time is drawn from the prior. Time is measured backwards in time in units of  $4N_0$  generations and  $\theta$  is defined as  $2N_0\mu$  for haploid organisms, where  $\mu$  is the mutation rate per generation (Hein, Schierup, & Wiuf, 2005).



**Figure 2.1 |** Schematic representation of the two evolutionary models. A. Population expansion model with parameters:  $n$ ,  $t_1$  and  $\theta$ . B. Population bottleneck defined by the parameters:  $n_1$ ,  $t_1$ ,  $n_2$ ,  $t_2$  and  $\theta$ .

For each one of the demographic scenarios, we created sets of 500,000 coalescent simulations. The number of individuals sampled per simulation was equal to the number of sequences present in the MSA of that population's marker genes. Each DNA sequence in the MSA FASTA file belongs to the dominant strain present in one sample and therefore it has been assigned with the same name of the respective sample ID. Moreover, msABC enables the simulation from a multitude of loci of variable lengths and sample sizes. The attributes of the simulated loci were specified in a supporting file and were in agreement with the features of the observed data, i.e. the marker genes. The program also allows the generation of sites with missing information, i.e. non-identified nucleotides symbolized as 'N'. Since incomplete information on sites affect the values of summary statistics and may bias the inference, we included the simulation of missing data by specifying the coordinates (position and sequence) of each 'N' in the alignment according to their positional distribution in the observed data.

The specification of the parameter priors is an essential element of an ABC analysis. We were cautious with their selection, since it has a large impact on the performance of the algorithm as well as on the inferred conclusions. When priors distributions covered different orders of magnitude, they were chosen on a logarithmic scale, so the sampling won't concentrate on larger values. Table 1. reports the set of parameter priors that were used for the simulation of each demographic model. Since we are simulating multiple loci, the program msABC internally rescales  $\theta$  for each fragment according to its length measurement. The scaled recombination rate  $\rho$  was set to zero.

**Table 1 |** Prior distributions for the demographic models.

	Population expansion		
Parameter	Prior distribution		
	Min	Max	Distribution
$\theta$	0.1	5	Log Uniform
$n_1$	0.01	1	Log Uniform
$t_1$	0.01	1	Log Uniform
	Population Bottleneck		
Parameter	Prior distribution		
	Min	Max	Distribution
$\theta$	0.1	5	Log Uniform
$n_1$	0.01	1	Log Uniform
$t_1$	0.01	1	Log Uniform
$n_2$	1	3	Uniform
$t_2$	0.02	2	Log Uniform

### 2.4.2 ABC inference

To estimate the posterior probabilities of different demographic models and posterior distributions of the parameters of these models, we employed an ABC approach. During the model selection procedure, the posterior probabilities of different predefined demographic models were estimated on the basis of the Euclidean distance between the observed summarized dataset and the simulated summarized datasets of all models. The inference procedure consists in retaining simulations for which the Euclidean distance between the set of simulated summary statistics and the observed set is sufficiently small. The percentage of accepted simulations is determined by the tolerance value, which was set to 0.05. Then a local linear adjustment to correct for the discrepancy between the simulated and the observed statistics. The method assigns weights to the parameters according to how well the simulated summary statistics adhere to the observed ones and performs linear regression between the summary statistics and the weighted parameters in the vicinity of observed summaries. The obtained regression coefficients are used to correct sampled parameters in the direction of observed summaries.

The estimation of the best model's parameters was also performed within an ABC framework. The tolerance value however for this analysis was set to 0.5. A correction for heteroskedasticity was performed for the regression step. Moreover, the parameters were 'logit' transformed prior to estimation, using as bounds the distributions' minimum and maximum values that were used during simulations for each parameter. After the regression estimation, the parameters were back-transformed to their original scale.

# Results

The 137 stool samples obtained from the Human Microbiome project were downloaded and analysed in a 64-bit Linux cluster with 24 nodes. The size of the compressed downloaded files was about 1 terabyte, so the manipulation of the uncompressed data required cautious manipulations for space management. In order to ensure that the data used for downstream analysis is not compromised of low - quality sequences, sequence artifacts, or sequence contamination that might lead to erroneous conclusions, we proceeded with a more strict quality filtering of the preprocessed reads downloaded from the HMP.

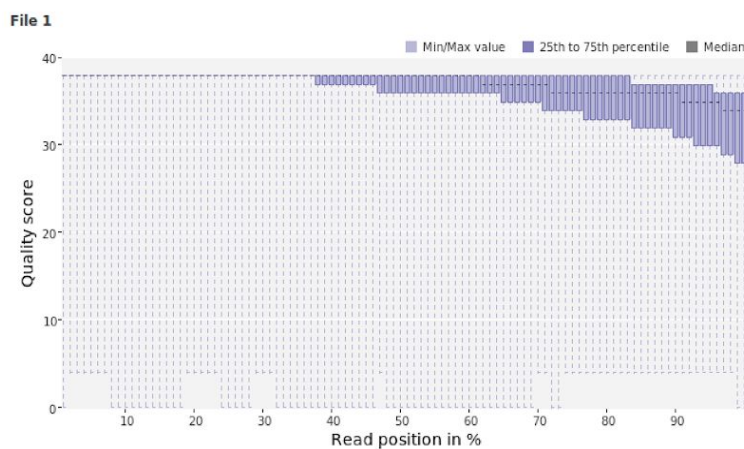
## 3.1 Quality assessment

The data preprocessing operations were determined by examining the unique characteristics of our diverse dataset, the type of library being sequenced and the sequencing technology used to produce the data. This was achieved by reviewing the summary statistics produced by Prinseq. The estimated summary statistics include read length and GC content distributions, tables with statistical metrics for read length and GC content, base quality distributions, occurrence of Ns and poly-A/T tails, base frequencies at the sequence ends and the probability of tag sequences, estimations of sequence duplication, sequence complexity distributions, the dinucleotide odds ratios as an indication of contamination and the dinucleotide relative abundance. The calculation of all those statistics was performed for each metagenomic sample, and was visualized in graphical format (Fig. 3.1).

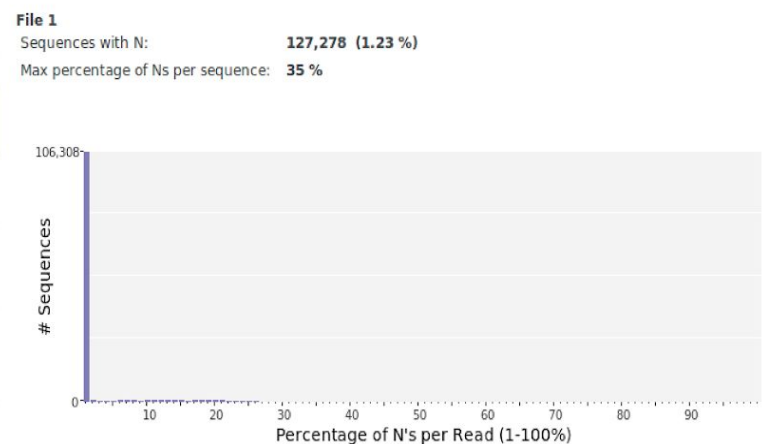
Reviewing the graphical representation of the summary statistics supported the selection of the preprocessing parameters. To begin with, base calling accuracy is the most widely used metric assessing the accuracy of a sequencing platform. It indicates the probability that a given base is called incorrectly (Illumina, 2011) and it is measured by the Phred quality score. A custom script was used to identify that the FASTQ files were encoded in ASCII\_BASE=33 Phred format (i.e. Phred+33). Next - generation sequencers produce data with linearly degrading quality across the read (Fig. 3.1.A.) that should be trimmed to discard low-quality positions. We used a quality score threshold of 20. The average quality scores of the reads are also informative, as suggested by Huse et al. (Huse, Huber, Morrison, Sogin, & Welch, 2007), and reads with values smaller than 25 were removed. A high number of ambiguous bases (i.e. Ns) has been considered as another sign of a low quality sequence, so a filtering of reads containing a significant number of Ns was applied (i.e. around 2% of the read length) (Fig. 3.1.B.). Moreover, we observed a small presence of poly-A/T tails in

the dataset's reads (Fig. 3.1.D.). Indeed, despite an early common misconception, polyadenylation not only occurs in eukaryotes (Kushner, 2015), so it was not unexpected to detect it's signatures inside our diverse dataset of microorganisms. We, therefore, chose to perform trimming on poly - A/T tails, because this can reduce the number of false positives during database searches, as long tails tend to align well to sequences with low complexity or sequences with tails (e.g. viral sequences) in the database (Schmieder & Edwards, 2011). Finally, the reads length distribution was not as informative, in fact it is intended mainly for other platforms, like PacBio and 454/Roche, which may produce variable read lengths. Illumina sequencers produce more discrete peaks (Fig. 3.1.F.). A length threshold, however,

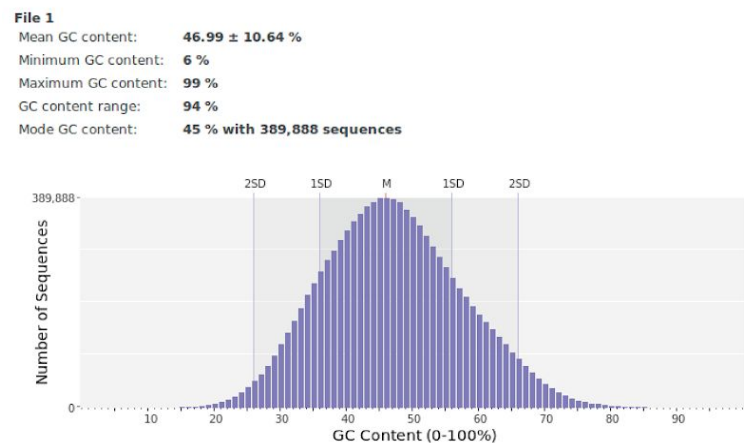
### A. Base Quality Distribution



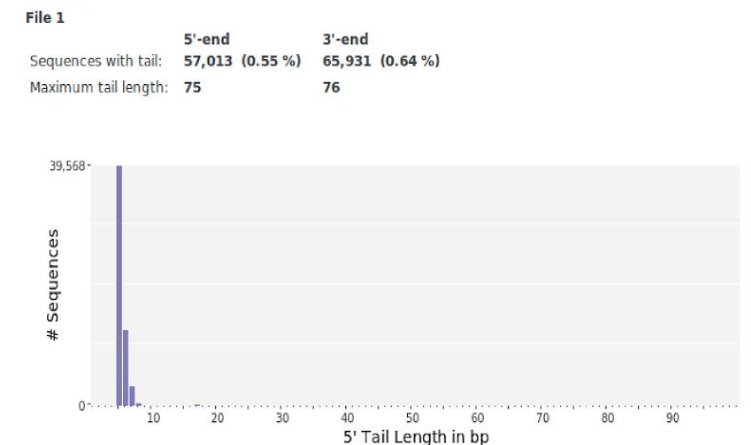
### B. Occurrence of N



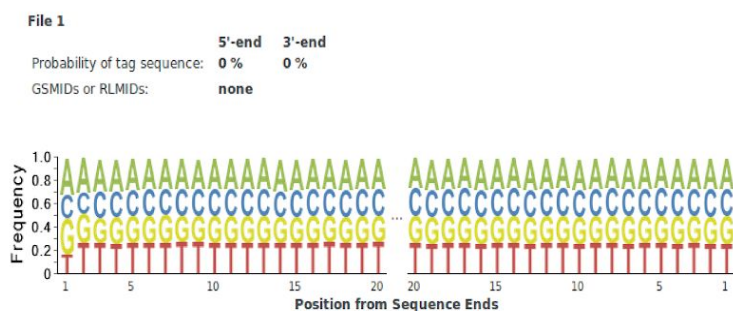
### C. GC Content Distribution



### D. Poly-A/T Tails



### E. Tag Sequence Check



### F. Length Distribution

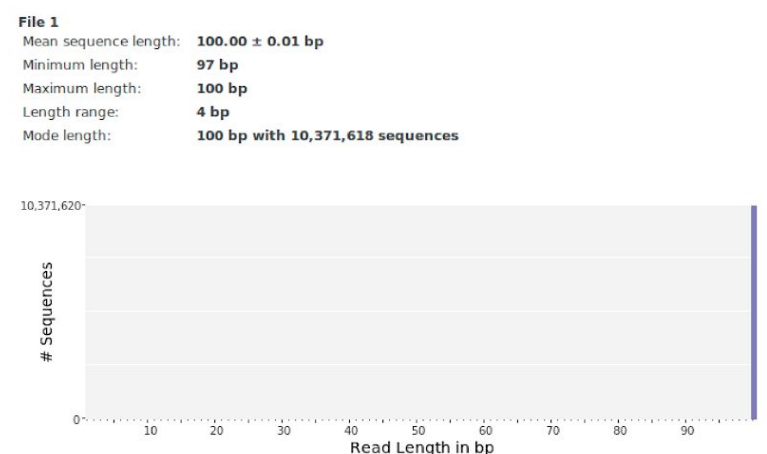
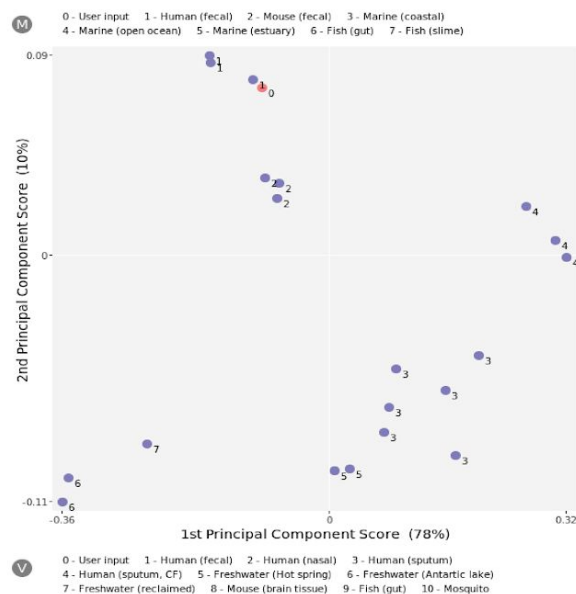
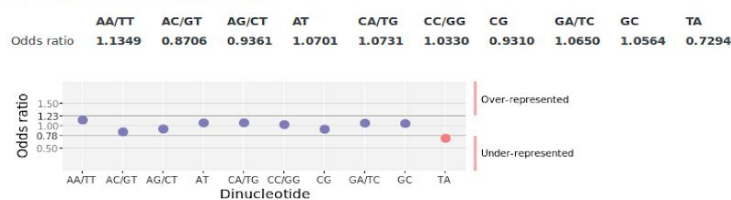


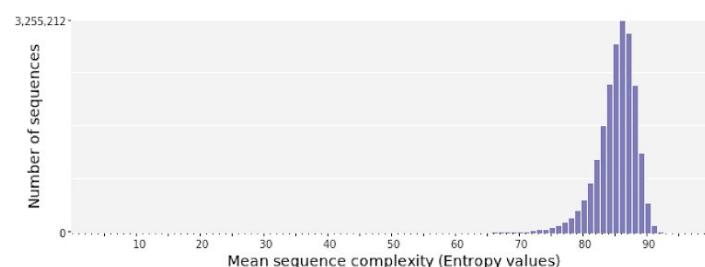
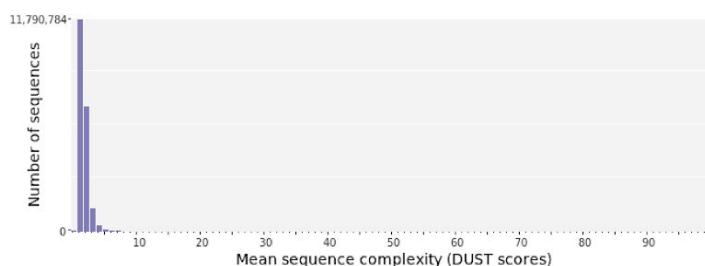
Figure continues on next page.

## G. Dinucleotide Odds Ratios



## H. Sequence Complexity

	Value	Sequence
Minimum DUST score:	0	GACCTCGGTACTTCACCATACAAGTTCATCAGTTCGTGTTGAACGTAAAGGACTCC
Maximum DUST score:	80	GATGCCGACGTTATATTAAATCTTCACCTCCATCGAGA
Minimum Entropy value:	13	TACCGGACGGGAAATAGATCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
Maximum Entropy value:	94	CTGGTCCGTGGGTGACACCCGGAAAAAAGAAAAAAGAAAAAAGAAAAAAGAAAAA
		AAAAAAGAAAAAAGAAAAAAGAAAAAAGAAAAAAGAAAAAAGAAAAAAGAAAAA
		AGTTCTTGGATCCCAAGCCACTGACACCTTGTGCTTCGAGCCGTACCGATAAGGCGG
		TCTATCTTGTGGAAATACAAAGTAGGCAGCTCATCCACGA



**Figure 3.1** | Graphical representation of the summary statistics of a random metagenomic sample (i.e. SRS022524), as produced by the standalone graphs version of PRINSEQ. A. Box plots of the quality scores across the reads. B. Histogram representing the occurrence of ambiguous based in the sample's reads. C. GC content plot marking the mean GC content (M) and the GC content for one and two standard deviations (1SD and 2SD). D. Histogram demonstrating the presence of poly-A/T tails in the sample's reads. E. Plot of base frequencies across the end of the reads, as an examination of residual tag sequences. F. Reads length information, summarized in a table of statistical metrics and a histogram of the read length distribution. G. The set of dinucleotide odds ratio values, expressed in relative abundance, constituting a signature of each DNA genome; Principal component analysis (PCA) of metagenomes from similar environments based on dinucleotide abundances as an investigation of the sequence variation compare to other microbial metagenomes. H. Sequence complexity estimated with the DUST and Entropy approaches (Morgulis, Gertz, Schäffer, & Agarwala, 2006).

applied, filtering sequences shorter than 60 bp. Short sequences may result in false positive functional or taxonomical assignments, since they are more likely to match at a random position by chance than longer sequences.

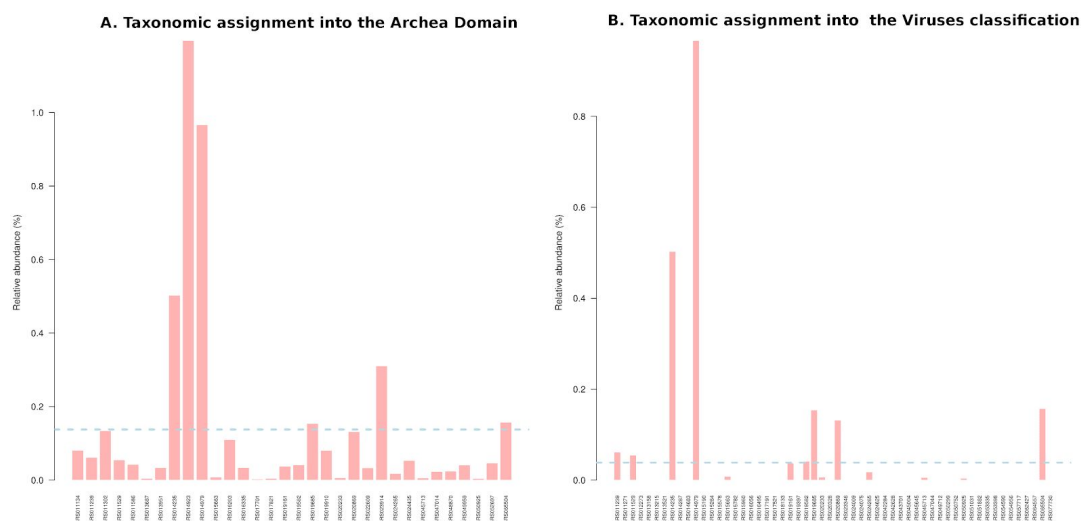
Some of the remaining summary statistics that are commonly used for quality control of NGS data, were evaluated as uninformative in the case our metagenomic dataset. The microbial composition of metagenomic samples is extraordinarily diverse, rendering the application of those statistics obsolete (Fig. 3.1.C, 3.1.E-H.). Indeed, due to the absence of compositional information, at this preliminary stage of the analysis, measurements of sequence complexity



were ignored. Additionally, no filtering operations were applied regarding the GC content metric, because of the presence of multiple microbial organisms with different GC content signatures. Finally, the use of dinucleotide odds ratios as an indication of possible contamination was omitted, along with the examination of residual ‘tag sequences’, since these preprocessing procedures were performed by the HMP Consortium and we trusted that need no further examination.

## 3.2 Microbial profiles

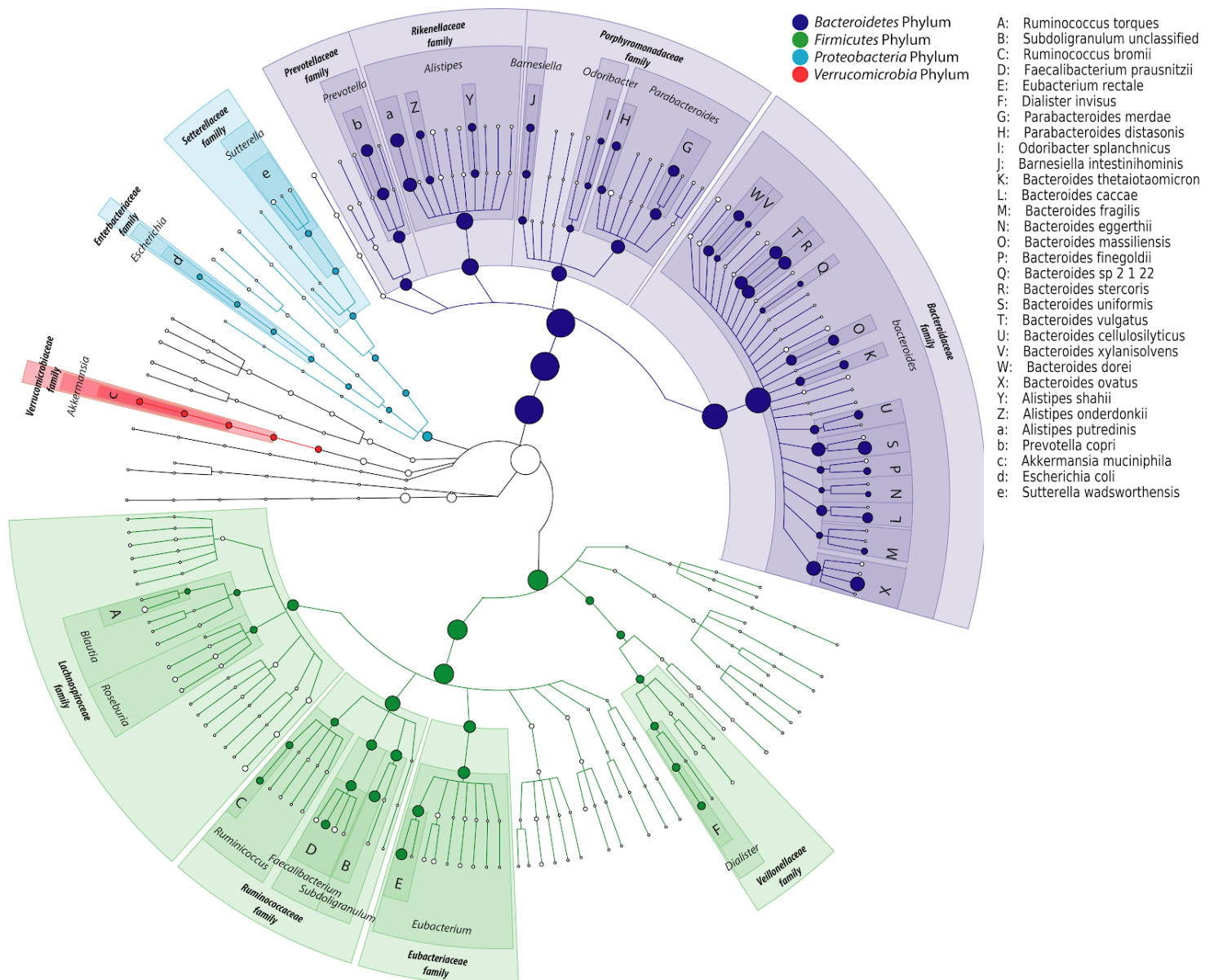
The composition of the microbial communities present in the stool samples was profiled using MetaPhlAn2. This tool estimates the relative abundance of microbial cells by mapping reads against a set of clade-specific marker sequences that are computationally preselected from coding sequences. Almost all of the taxonomically assigned reads belonged to the Bacteria Domain of life, with a 99.91% mean value of relative abundance across all samples. Although bacteria predominate the human gut, archaea and viruses were also identified at a small but detectable fraction of the microbial abundance (Fig. 3.2). More precisely, the Archaea Domain was represented by two genera belonging to the class *Methanobacteria*, which was detected at 32 samples with a mean relative abundance of 0.13%. The detected dominant archaeon *Methanobrevibacter smithii* in particular, has a well-documented role in the human gut, affecting the digestion of dietary polysaccharides by other microbes (Samuel et al., 2007). In 56 samples, an even smaller amount of the *Myoviridae* family of bacteriophage viruses was detected at a mean relative abundance of about 0.11%.



**Figure 3.2 |** Relative abundance of the archaea and viruses that were present in a subset of the samples (x-axis). The dashed lines indicate the mean relative abundance across those samples.

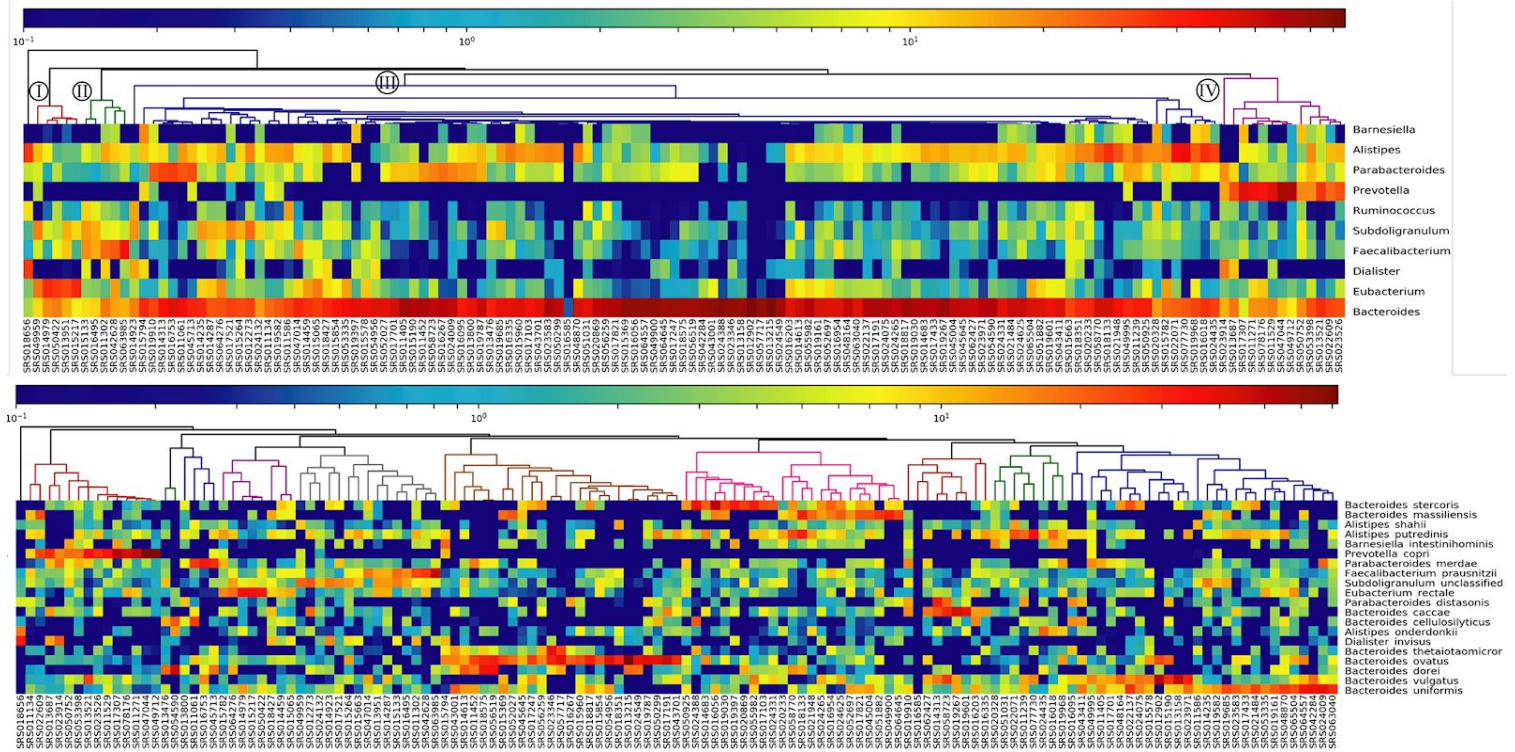
To that end, we focused on the bacterial fraction of the gut microbiota. At a phylum level, we observe notably poor variability, in fair agreement with relative studies (Arumugam et al., 2011; Integrative HMP (iHMP) Research Network Consortium, 2014). On average, 96% of

the gut bacteria were members of only two phyla, *Bacteroidetes* and *Firmicutes*. In higher taxonomic resolution, we identified 52 genera belonging to the Bacteria Kingdom, with a relative abundance above 0.5% present at least once in the samples. The taxonomic relatedness of the detected clades was captured in a cladogram, which also visualizes their average relative abundance. (Fig 3.3). *Bacteroides* was consistently the most abundant genus, demonstrating the highest mean relative abundance across all samples, confirming its well-established role in healthy adult microbiomes. (Wexler & Goodman, 2017). It was also estimated that it ranged from being the dominant bacterial genus in 110 samples, to a minority in others who carried a greater amount of *Prevotella*. Other species achieving dominance in multiple samples belong to the genera: *Faecalibacterium*, *Eubacterium* and *Alistipes*.



**Figure 3 |** Taxonomic cladogram reporting all clades present the samples ( $\geq 0.5\%$  relative abundance in  $\geq 1$  sample). The nodes' size reflect their relative abundance. The labeled annotation lists the 30 most abundant species.

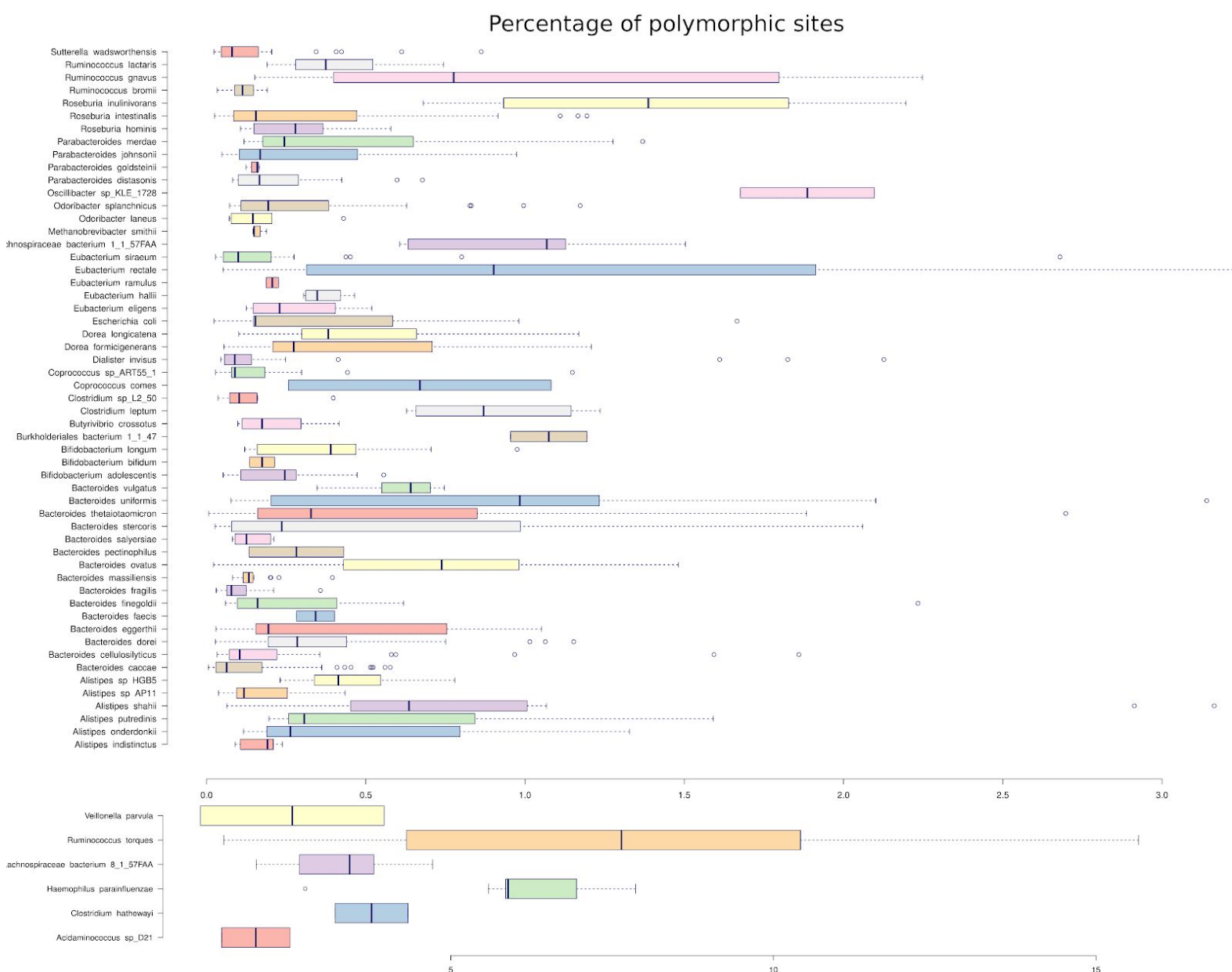
Genus and species level taxonomic profiles of the most abundant clades were summarized and hierarchically clustered (Fig. 3.4). In some cases, discrete clustering patterns were apparent. On a genus level, high abundance values in *Prevotella*, *Eubacterium* and *Faecalibacterium* genera could be considered as distinct characteristics of clusters IV, I and II respectively. Similar patterns can be observed on a species taxonomic resolution. Clusters tend to feature the dominance of one specific species over all the other detected organisms across each specific cluster.



**Figure 3.4 | A.** Genus-level taxonomic profile of the 10 most abundant genera, hierarchically clustered (average linkage) with the Pearson correlation coefficient. **B.** Species-level taxonomic profile of the 20 most abundant species, hierarchically clustered (average linkage) with the Pearson correlation coefficient.

### 3.3 Obtaining strain-level resolution of species

Increasing the taxonomic resolution, we applied StrainPhlAn to the microbial clades identified by the taxonomic profiling tool MetaPhlAn2. In total, MetaPhlAn2 identified 134 individual species with a relative abundance above 0.5% present at least once in the samples. Out of these, StrainPhlAn targeted samples of 61 species (Supplemental Table S1) that exhibit sufficient enough coverage to be analysed (Supplemental Fig. S1) and reconstructed the dominant strain of a given species in each metagenomic sample.



**Figure 3.5 | Intra-species heterogeneity.** StrainPhlAn estimates the percentage of polymorphic sites for each sample's detected strain of a given species, while this box-plot represents the distributional characteristics of the estimated values, grouped by species.

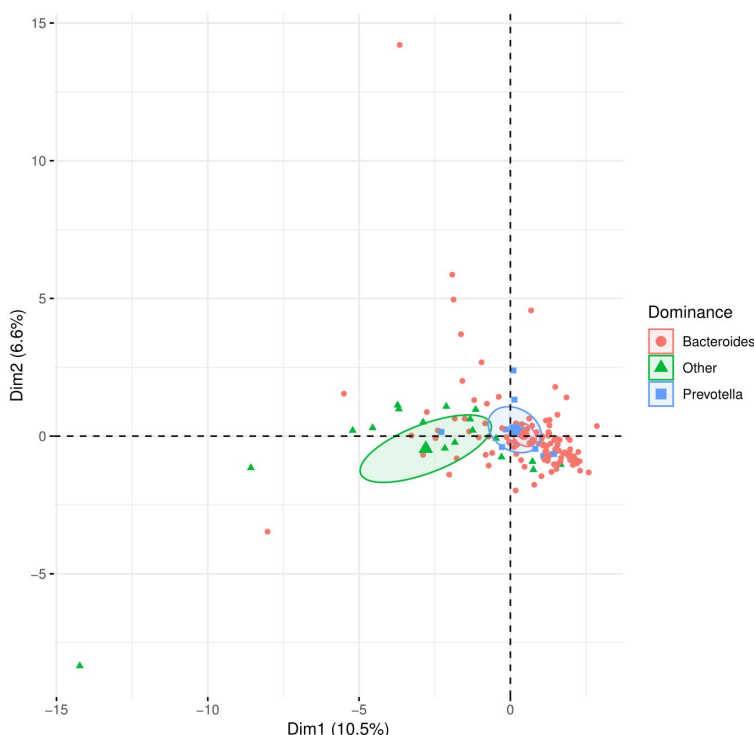
StrainPhlAn reconstructs each species' most abundant strain per sample, and it can quantify the strain heterogeneity within a sample by identifying single nucleotide polymorphisms (SNPs). In total, we observed intra-species variability across all the 61 detected species in the metagenomic samples (Fig. 3.5). A few species in particular (e.g. *C. hathewayi*, *H. parainfluenzae*, *R. torques*, *V. parvula* and *Acidaminococcus* sp. D21), demonstrate substantially high numbers of polymorphic sites. A number of plausible hypothesis can explain this observation. For instance, a high mutation rate or a big effective population size can be the cause of this intra-species variability. Another possible scenario is that more than



one strain of the species is present within the sample. At the same time, low polymorphism estimations characterize a number of species, with *B. massiliensis*, *B. fragilis*, *D. Invisius*, *M. Smithii* and *P. goldsteinii* being the most notable examples. That is in fact, most often the case, where many of the detected species carry less than 0.5% polymorphic sites. Accordingly, a lower mutation rate, a smaller effective population size or the representation of just a single dominant strain might explain these data. Lastly, we postulate that the widely-spread interquartile range of some species (e.g. *B. uniformis*, *B. thetaiotaomicron*, *E. rectale*, *R. gnavus* and *R. inulinivorans*) implies that in some samples they might be represented by only one dominant strain while in other samples, they coexist in multiple strains.

### 3.4 Examining the ‘Enterotypes’ Hypothesis

These results can lead our own investigation on the presence of enterotypes. It has been hypothesized that the adult gut microbiota arrangements can be classified into distinct ‘community types’, called ‘enterotypes’ (Arumugam et al., 2011; Costea et al., 2017). The proposed enterotypes were identified by their enrichment in the *Bacteroides* genus (enterotype 1), the *Prevotella* genus (entero-type 2) and the *Ruminococcus* genus (enterotype 3). Arumugam’s research team (2011) initially proposed this hypothesis, based on clustering results of genus composition seen in a principal components analysis (PCA). Similarly, we performed a PCA on the genus-level abundance table, the samples were colored by the most dominant taxon (Fig. 3.6). We reckon that the separation of the three groups was apparently unclear. In any case, the groups were not that obviously discrete and distant as the ones observed in the supporting work. However, based on this preliminary analysis we can not claim, at least prima facie, that our data reject the enterotypes hypothesis.



**Figure 3.5 |** PCA plot of the genus-level taxonomic abundance table. Samples are colored by their most prominent taxon. If the sample is dominated neither by *Prevotella* nor *Bacteroides*, it is classified as other. Confidence ellipses are produced around the three categories.

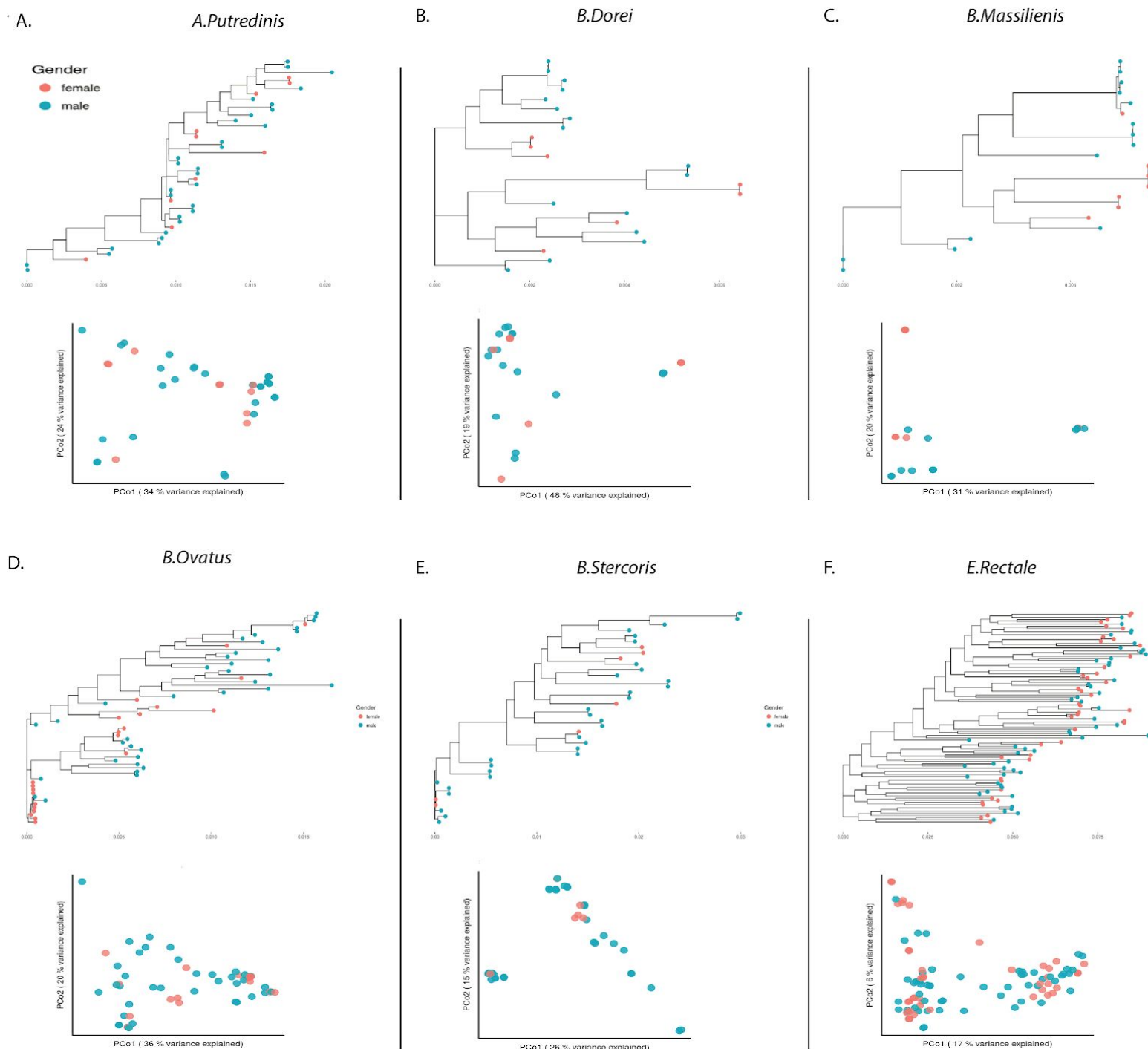
An exploratory PCA can certainly be informative, but the low-dimensional representation of the samples can often miss a lot of valuable information. Indeed, the first two principal components that were used for visualization only explain the 17,1% of the variability present in our dataset. Exploiting the results of the hierarchical clustering (Fig. 3.4), we were able to examine in greater detail the issue. The clustering of the samples becomes clear, demonstrating noticeable enterotype-like patterns. Indeed, the subjects that were separated in cluster IV, characterized by the highest *Prevotella* genus abundance, may resemble enterotype 2. In fact, all samples are characterized either by moderately high *Prevotella* abundance or it's nearly-complete absence (Supplemental Fig. S4). Many more samples were characterized by highly abundant *Bacteroides* genus that might bear resemblance to enterotype 1. But *Bacteroides*' relative abundance formed a continuum across samples (Supplemental Fig. S4) spanning over different inferred clusters (Fig. 3.4.A). Likewise, on a species level, *Bacteroides* seemed to be visibly diversified (Fig. 3.4.B). For instance, *B. ovatus*, *B. stercoris*, *B. vulgatus* along with *B. uniformis*, are spread into three separate clusters. Ultimately, we might claim that our results supports the existence of more complex community patterns than those captured by Arumugam et al.

### 3.5 Investigating a sex-bias in gut microbial composition and phylogeny

Several studies have demonstrated that males and females have gender-specific differences in their gut microbiota composition (Fransen et al., 2017; Haro et al., 2016; Taneja, 2017). Even though the question of whether these differences in gut microbiota composition are a cause or consequence of observed gender-specific characteristics of the immune system remains unanswered, there is compelling evidence suggesting the two-way dynamic influence between sex hormones and the gut microbiota.

In order to study the issue of sex-biased differences in the gut microbial composition, we investigated which bacterial species demonstrate a significant difference in their relative abundance between male and female subjects. The abundance is compared individually for each one of the species, but since the values were severely skewed (Supplemental Fig. S2), we chose to perform a non-parametric test, instead of a two-sample t-Test. The Mann–Whitney–Wilcoxon (MWW) test, as it compares the sums of ranks, is less likely than the t-Test to spuriously indicate statistical significance because of the presence of outliers. MWW is a test of both location and shape (Hart, 2001) used to determine whether the two independent samples were selected from populations having the same distribution. Adjusting the test's pvalues (Supplemental Table S2), to correct for the multiple comparisons that were performed, we estimated a statistical significant support for two bacterial species. The abundance of *Bacteroides xylanisolvens* follows statistically different distributions (Mann–Whitney U=3113,  $n_1=56$ ,  $n_2=78$ , adjusted pvalue=0.028) between Male and Female

subjects (Supplemental Fig. S3.A.). *Sutterella wadsworthensis* was also differentially distributed (Mann–Whitney  $U=3024$ ,  $n_1=56$ ,  $n_2=78$ , adjusted  $p$ value=0.028) among samples of different gender (Supplemental Fig. S3.B).



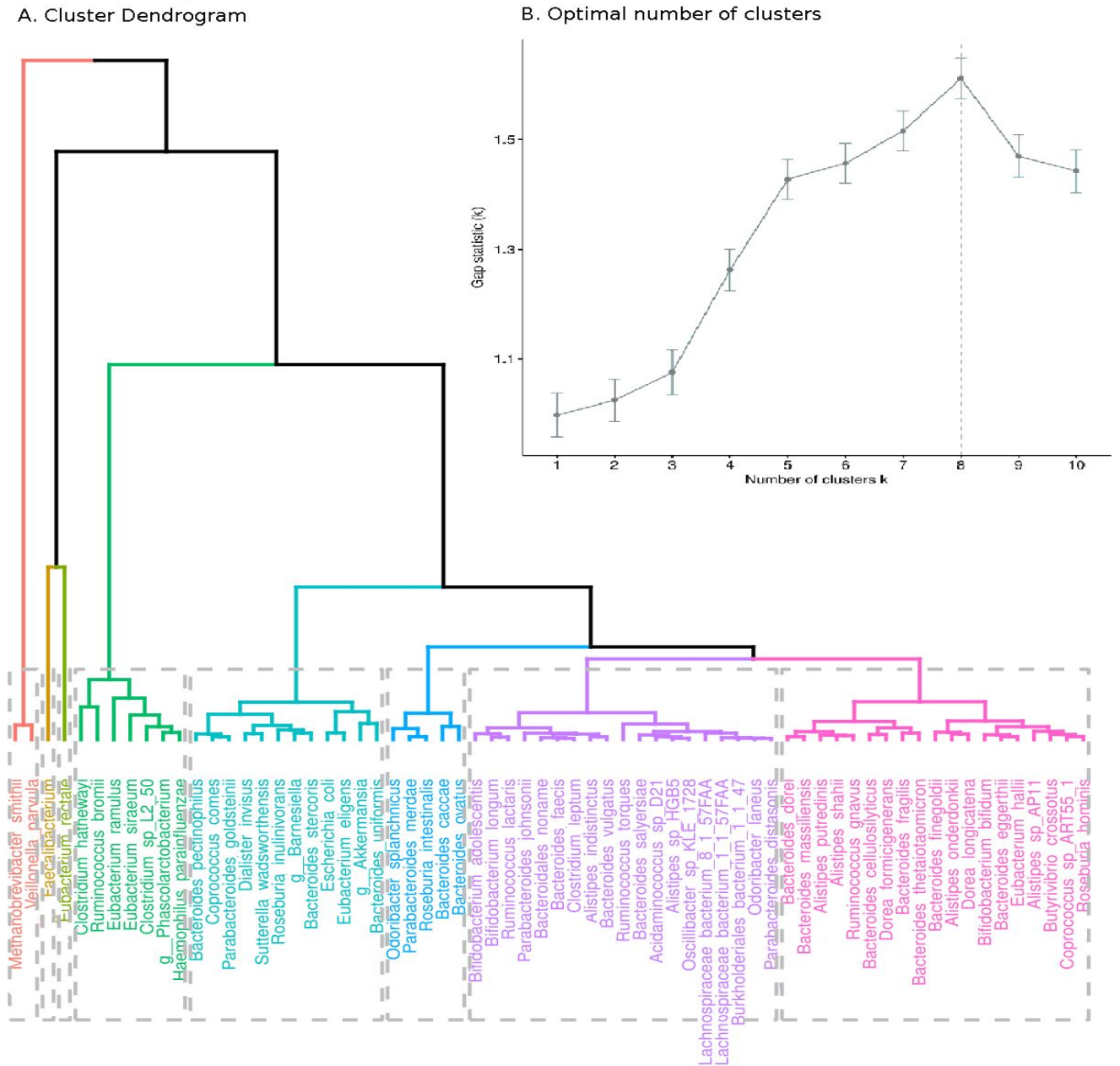
**Figure 3.6 |** Population structure of 6 prevalent species detected with StrainPhlAn. We report the results of A. *A. putredinis*; B. *B. Dorei*; C. *B. massiliensis*; D. *B. Ovatus*; E. *B. stercoris* and F. *E. rectale*. while the rest of the analysed species are reported in Supplemental Figs. S5-15. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.

This issue can be further examined on a different level with a phylogenetic approach. The concatenated alignment of the reconstructed markers was used to infer the strain-level phylogenetic trees using RAxML-ng of each species inferred by StrainPhlAn (Fig. 3.6, Supplemental Figs. S5-15). The trees are annotated with the subjects' gender to investigate associations with the genetic structure of the species. In addition, the evolutionary distance between every pair of strains in a specific species was visualized by ordination with the classical Multidimensional Scaling (MDS) method, annotated with the gender information (Fig. 3.6, Supplemental Figs. S5-15). The MDS plot did not reveal any apparent separation of the gender information. However, we were able to identify a few features based on the species phylogenies. In many species, we observed well-defined subtrees, usually on external branches, that were uniquely composed of strains belonging to same-gender subjects. Moreover, in the majority of the species, even those lacking notable gender discrete subclades, strains on the cherry branches often evidenced same-gender assortment. We postulate that this observation can be explained by the HMP's sampling procedure. According to the database's protocols "all enrolled subjects were sampled at one visit, with a subset of subject sampled at up to three visits". So at least a number of the cherry nodes might correspond to different time-point collection of the same individual's microbiome. Unfortunately, the number of visit of each individual was not publicly available outside of dbGaP's authorized access system.

### 3.6 Modeling the demographic history of gut species

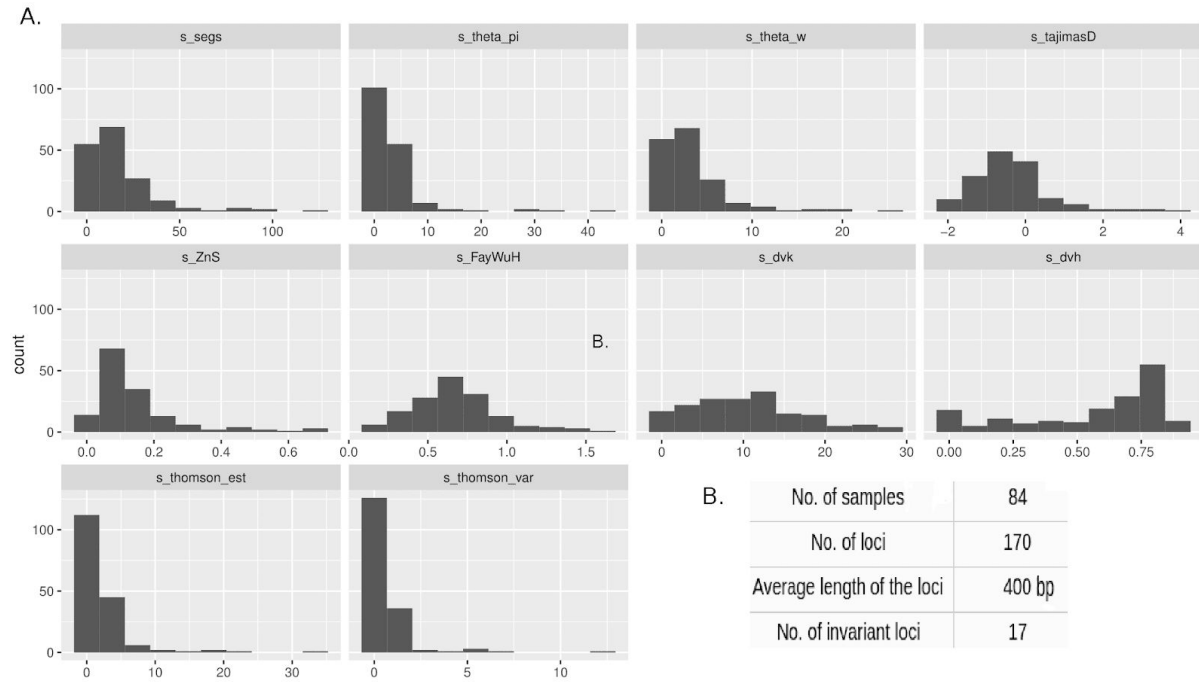
Exploring the demographic dynamics of the human gut microbiota is important in understanding its biology and health impact. Here, we focused on modeling and inferring the demographic history of specific bacterial species metapopulations colonizing the 137 stool samples of the HMP. At the beginning of this analysis, we first examined the set of the observed summary statistics that describe each species metapopulation. We hierarchical clustered the bacteria based on the summary statistics vectors (Fig. 3.8.A.), attempting to find structure patterns between them. The number of clusters was determined using the gap statistic (Tibshirani et al., 2001) (Fig. 3.8.B.).





**Figure 3.8 | Hierarchical clustering of each species' observed summary statistics. A.** Clustering dendrogram demonstrating the results of hierarchical clustering. **B.** Display of the gap-statistic values (500 simulations for each K-value) for each choice of number of clusters from K=1 up to K=10.

Arguably it has been impractical to revisit the demography of each one of the 61 species. Hence we focused on specific metapopulations of interest. Firstly, we employed an ABC approach to investigate the demographic history of *B. caccae*, one of the most prevalent bacteria in the human gut. The MSA of the 170 *B. caccae*'s reconstructed marker genes, produced during the taxonomic profiling step, constitutes our 'observed data' that were replaced by a set of summary statistics (Fig. 3.9).



**Figure 3.9 |** Summary information of *B. caccae*. **A.** Histograms of the observed summary statistics of *B. caccae*'s reconstructed loci. **B.** Table of genetic information about the loci included in the demographic analysis.

The results of the model choice procedure showed that a population bottleneck in the history of *B. Caccae* fits the observed data better than an expansion. The Bayes factor (Kass & Raftery, 1995) was greater than 100 (viz. 28726) supporting with strong evidence the bottleneck model, but it any case does it report any information about the absolute fit of this model to the data. The posterior probabilities of the model's parameters were inferred in a ABC framework, as well.

The parameter inference procedure consists in retaining only the simulations for which the Euclidean distance between the set of simulated summary statistics and the observed set is sufficiently small. Visualizing in a contour plot the percentage of accepted simulations that was determined with a tolerance of 0.5 (Fig. 3.10), demonstrates how these simulations sample the space around the observed data. The summary statistics occupy a space of 16 dimensions. So we notice that for some statistics (i.e. dimensions), the contour lines diverge slightly from the observed value, because the algorithm keeps the simulations that minimize the 16-dimensional Euclidean distance with the observed vector. An improvement to the simple rejection ABC algorithm was employed, in order to correct for the discrepancy between the simulated and the observed statistics. The posterior probability of the parameters was approximated applying local linear regression to the retained simulations (Fig. 3.11). This correction influences significantly the results. In the case of  $\theta$  and  $n_1$  it supports with stronger evidence a narrower area or parameter values (Fig. 3.11: red lines), than the distribution estimated by the ABC accepted simulations prior to the regression step

(Fig. 3.11: blue lines). While for  $t_2$  and  $n_2$  (Fig. 3.11) the correction shifts the distributions' shape and skewness. On a final note, our data do not contain enough information about the  $t_1$  parameter, so the posterior is determined mainly by the prior and the regression correction.

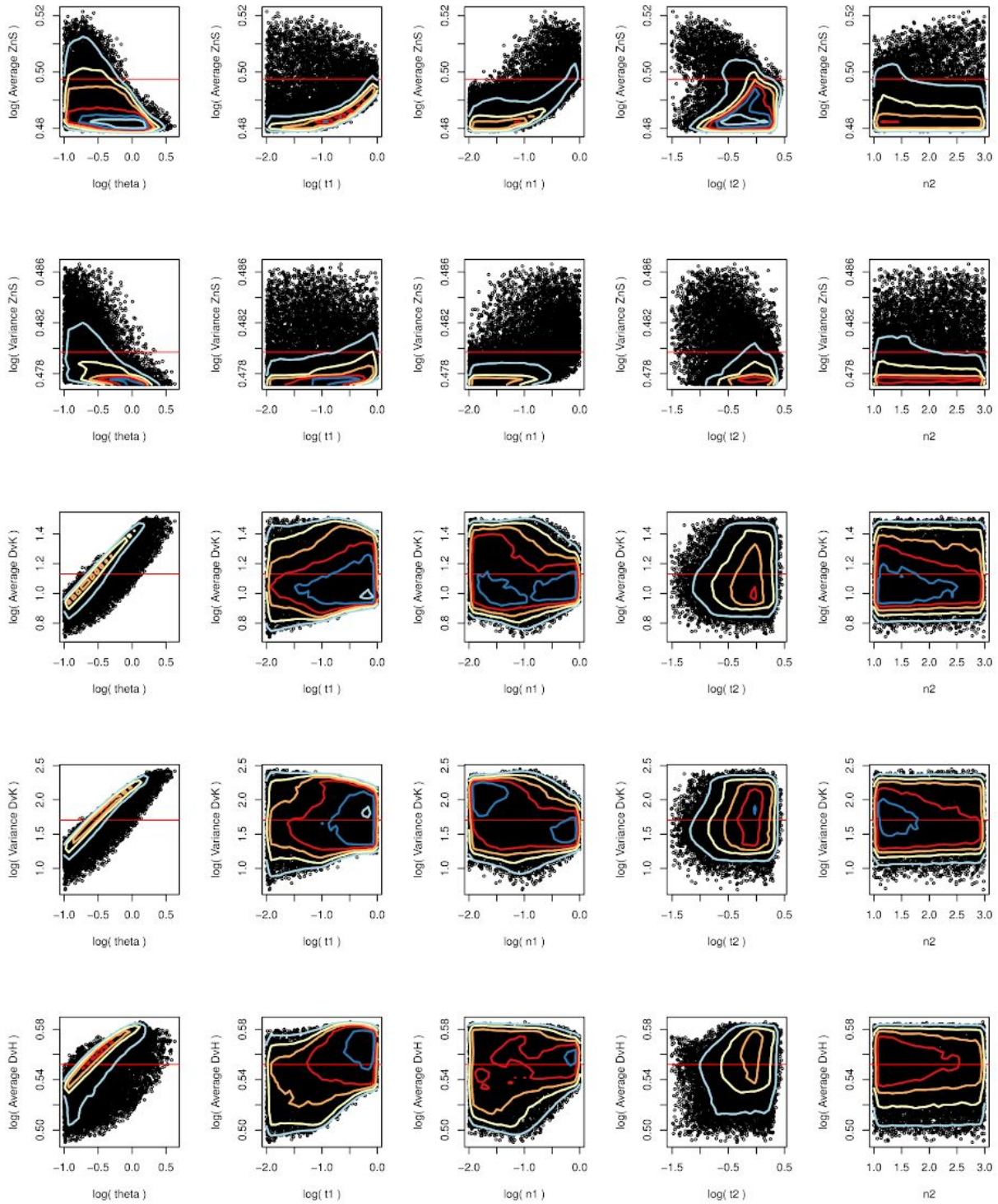


Figure continues on next page



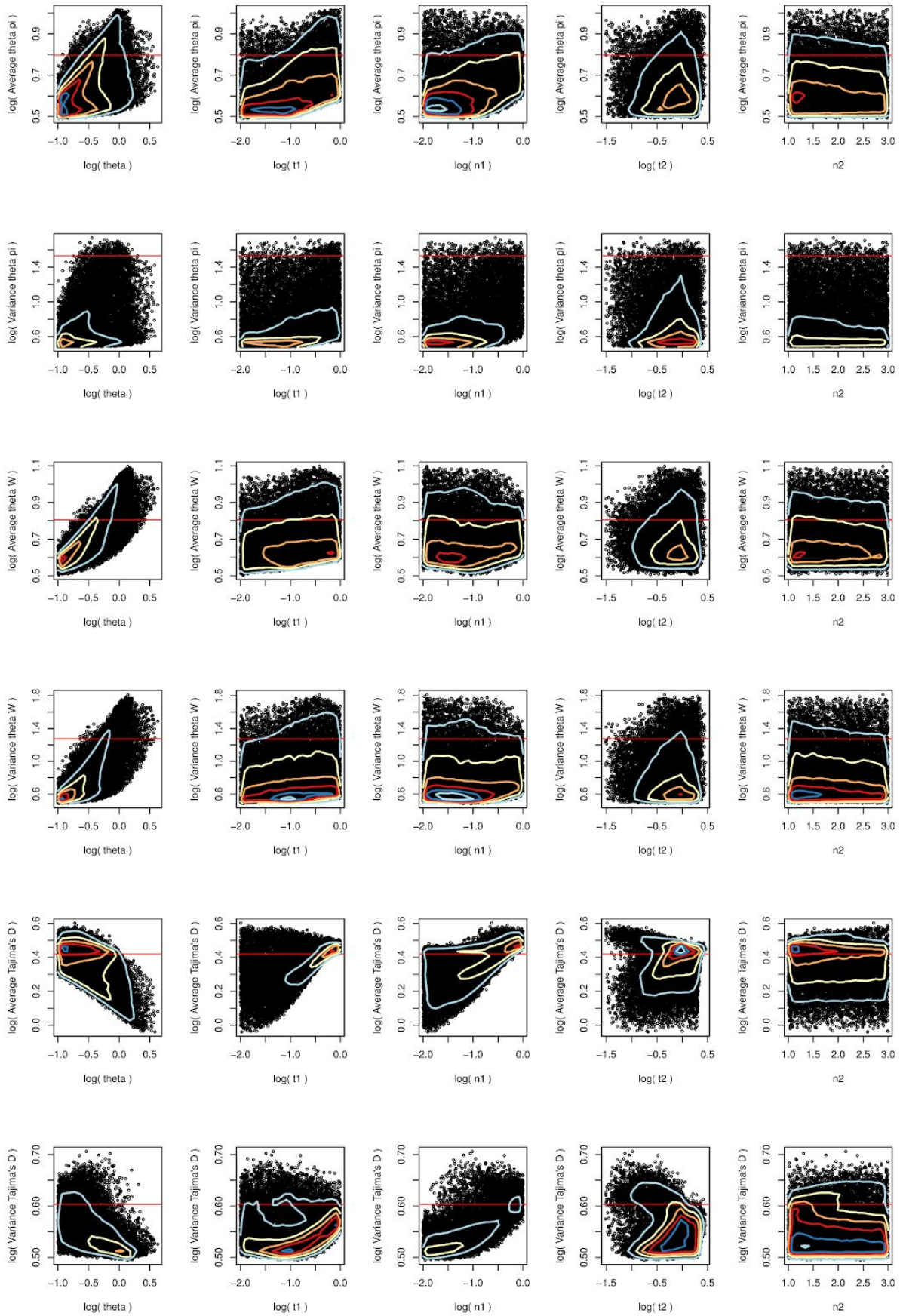
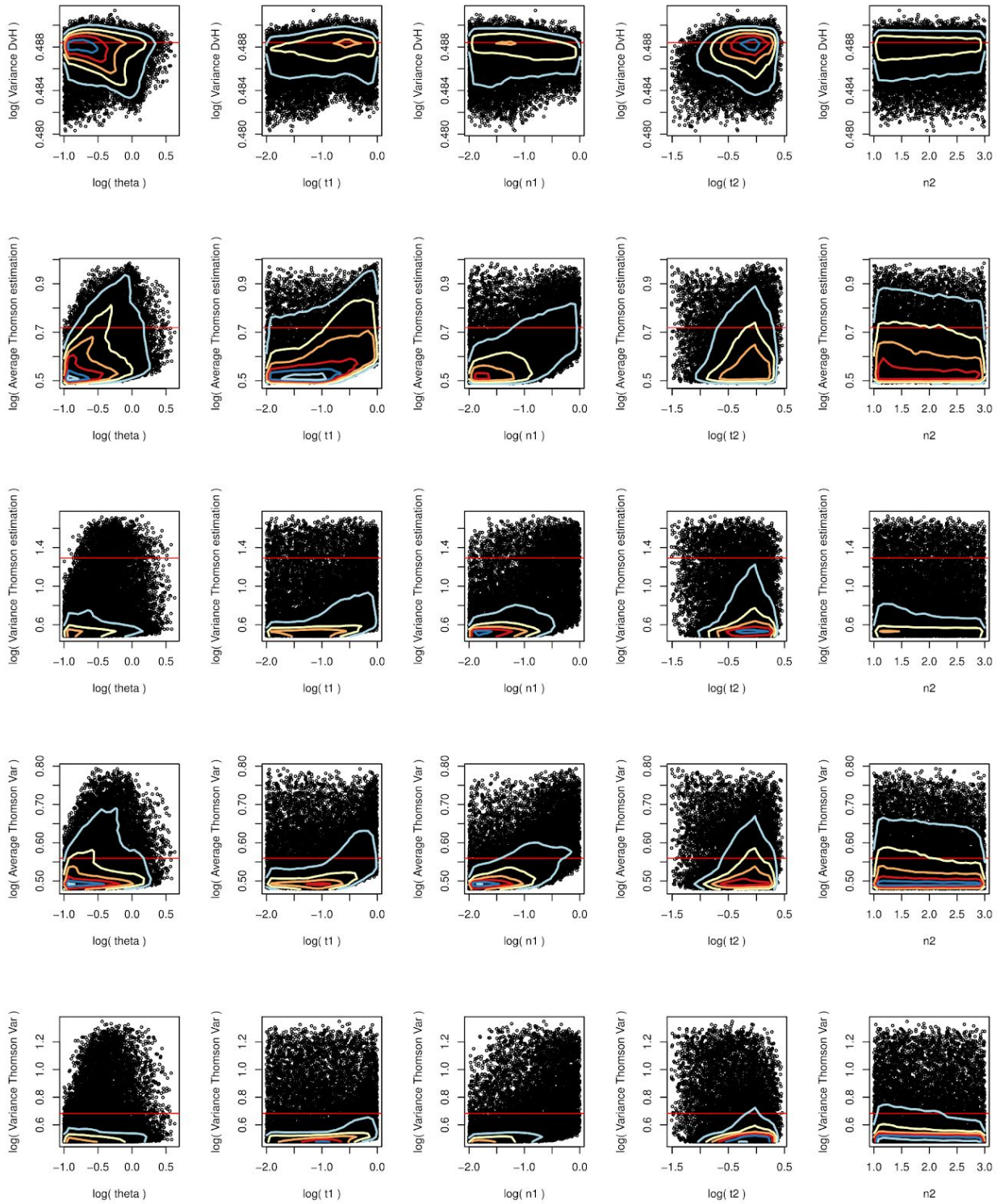
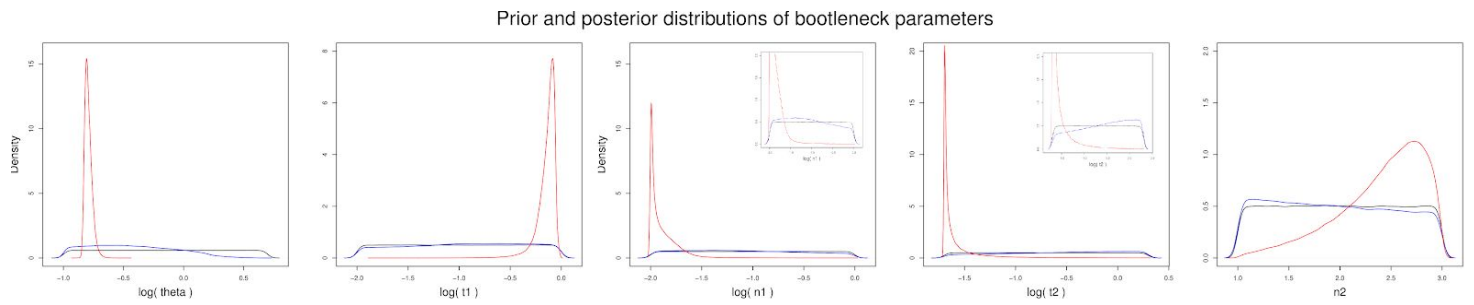


Figure continues on next page



**Figure 3.10** | The space of the ABC accepted simulations that were determined with a tolerance of 0.5. The contour lines on the scatterplot indicate the density estimation.

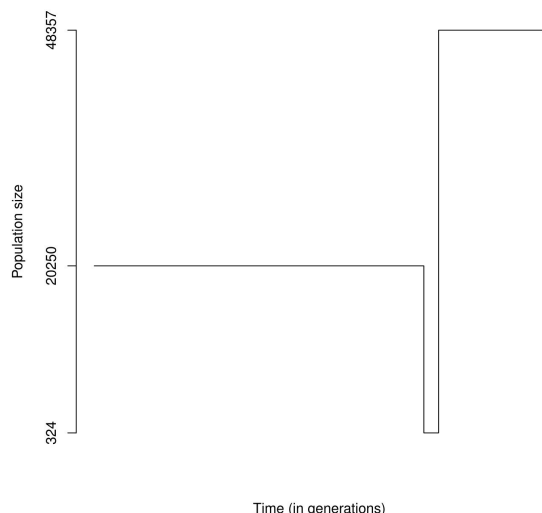


**Figure 3.11** | Probability distributions of *B. caccae*' bottleneck parameters. Black line: Prior distribution; Blue line: Density distribution or the ABC accepted simulations that were determined with a tolerance of 0.5, before applying the regression correction; Red line; Posterior probability approximated applying nonlinear regression using a feed-forward neural networks.

We used the mean value of each inferred parameter (Table 2) to summarize our estimation about *B. caccae*' history. The inferred value of  $\theta$  was scaled with the average length of the loci (i.e. 400bp), to obtain a per site estimation. Given this estimation, which is equal to  $2N_0\mu$ , we needed the mutation rate to calculate  $N_0$ . Mutation rates vary widely in bacteria: from  $10^{-8}$  to lower than  $10^{-10}$  per site per generation (Rocha, 2018), so we set the mutation rate at an average of  $10^{-9}$  and compute that the current population size equals to 2,050. We, hence, estimated that 58,563 ( $=0.992 \cdot N_0$ ) generations ago, the population size was reduced to 324 for a duration of 2,616 generations. Before that, we inferred that the ancestral population size was 2.279 time it's present size, viz. 48,357 .

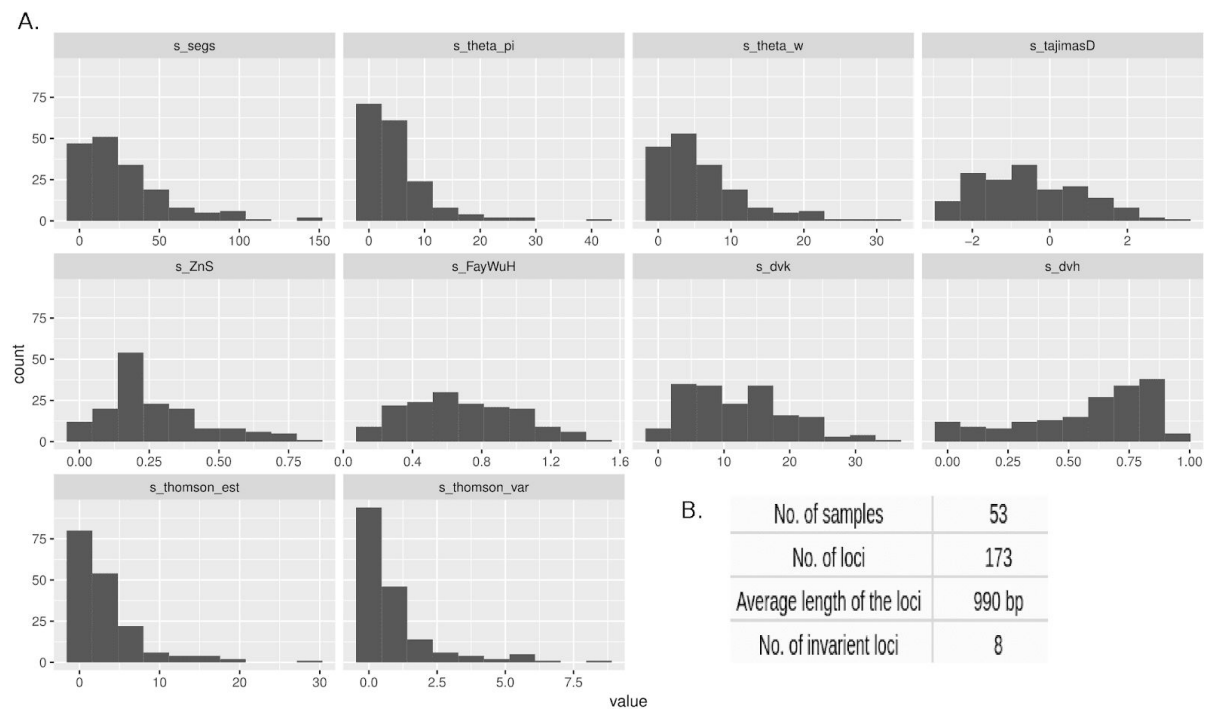
**Table 2** | Parameter estimation of *B. caccae*'s population bottleneck model.

	$\theta$	$t_1$	$n_1$	$t_2$	$n_2$
Weighted 2.5% percentile	0.145	0.392	0.010	0.020	1.366
Weighted median	0.159	0.748	0.012	0.021	2.477
Weighted mean	0.162	0.723	0.016	0.036	2.388
Weighted mode	0.155	0.794	0.0111	0.022	2.706
Weighted 97.5% percentile	0.191	0.910	0.035	0.129	2.950



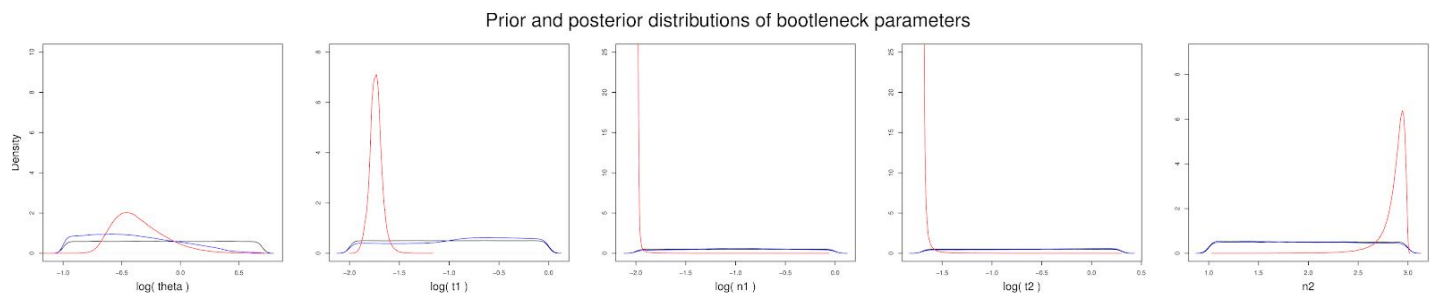
**Figure 3.12** | Changes in *B. caccae*'s metapopulation size during its evolutionary history.

Next, we were interested in the demographic dynamics of *Bacteroides ovatus*. The hierarchical clustering (Fig. 3.8) yield that *B. ovatus* is the closest species to *B. caccae*, based on their vectors of observed summary statistics. We hypothesized that since they share such similarities in their summarized information (Figs. 3.9, 3.13), they might have experience analogous demographic changes in their past. On that premise, we performed a similar ABC analysis on *B. ovatus*, using the same model priors (Table 1).



**Figure 3.13 |** Summary information of *B. ovatus*. **A.** Histograms of the observed summary statistics of *B. ovatus*’ reconstructed loci. **B.** Table of genetic information about the loci included in the demographic analysis.

In this case, the model choice procedure supported the bottleneck model over the population expansion, as well. The parameter inference however, was not conclusive (Fig. 3.14, Supplemental Fig. S16). The prior-posterior plots (Fig. 3.14) demonstrate that the observed data contradict with the information provided by the prior. Therefore the inference did not find almost any information by the data in the parameter space defined by the priors. At the same time, the posterior densities of  $n_1$  and  $t_2$  were asymptotically shifted towards the minimum values of the prior distribution. This behavior suggest that the data seem to drive the posterior in an area outside from the predefined prior. Ultimately, it becomes apparent that the preselected priors do not contain the sufficient information needed to make an accurate inference. We therefore rejected the current estimated model for *B. ovatus*, challenging our original hypothesis.



**Figure 3.14** | Probability distributions of *B. ovatus*' estimated bottleneck parameters. Black line: Prior distribution; Blue line: Density distribution of the ABC accepted simulations that were determined with a tolerance of 0.5, before applying the regression correction; Red line; Posterior probability approximated applying nonlinear regression using a feed-forward neural networks.



## Conclusions

We are witnessing the beginning of an era of population-scale whole-microbiome epidemiology. While many current studies of the human microbiome focus on disease, a better understanding of the healthy state is necessary, before we can determine the impact of the microbiota on disease predisposition and pathogenesis. To that end, we choose to study the taxonomic and genetic diversity of microbial communities colonizing the gut of healthy individuals from the 1<sup>st</sup> phase of the HMP.

Advances in sequencing technologies and biocomputing enable the genomic study of the uncultured part of human-associated microbial communities and have considerably shaped the way metagenome research is performed. A great variety of computational algorithms and pipelines have been developed to address the unique characteristics of metagenomic datasets and the previously discussed challenges that they present. What is missing from the literature is a systematic evaluation of those tools. Computational pipelines yet attaining a comparable degree of standardization for results derived from different workflows (Oulas et al., 2015; Segata et al., 2013).

There is extensive research in exploring the vast amount of genetic and taxonomic diversity that characterize the human microbiome, with the pioneering work of the Human Microbiome Project (HMP) notably influencing the field. Strain-level metagenomics is a very active area of research and has the potential of increasing the profiling resolution to the level of sequencing cultured single isolates (Quince et al., 2017). In this study, we employed a marker-based approach, to explore the taxonomic diversity of the metagenomic samples. This approach can mitigate assembly problems, using external sequence data resources. A notable limitation though exists due to the dependance on publicly available reference genomes. Extensively researched areas, like study efforts on human health and biotechnology, have larger contributions in existing databases. Indeed, microbiomes with such rich diversity of available reference genomes (e.g. the human gut microbiome) can be adequately and efficiently profiled. Unfortunately, metagenomes from more diverse environments, such as soil or water, are particularly affected by this problem since they are under-represented in reference databases (Choi et al., 2017). The reference provided by already sequenced genomes arguably provides the most reliable substrate for identifying the taxonomic origin of individual metagenomic reads, but it is still apparent that the enrichment in reference genome sequence data spanning the whole phylogenetic diversity of the microbial tree of life is more than necessary.

Assessing the taxonomic diversity of microbial communities colonizing the stool samples, we were able to investigate the controversial ‘enterotypes’ hypothesis. In a series of studies, enterotypes have been used to classify gut microbial samples into distinct community types. However, evidence surrounding the existence and formation of these enterotypes has raised questions and numerous studies thereafter aimed in examining the hypothesis using more careful clustering analyses (Huse, Ye, Zhou, & Fodor, 2012; Jeffery et al., 2012; Knights et al., 2014; Koren et al., 2013; Wu et al., 2011). In agreement with Gorvitovskaia et al. (2016), our results of genus information support the existence of more complex community patterns than discrete clusters. Prevalent genera, like *Bacteroides*, form continuous gradients across different samples demonstrating no particular enterotype-like separation. Categorizing the gut microbiota into discrete groups might be appealing, but Knight et al. (2014) cautioned that putative clusters can hide potentially important variation.

The human microbiome is a diverse ecosystem characterized by increased genetic variation and rapid evolution (Shapira, 2016). The dominating bacterial part of gut communities, is a perfect subject of study for researchers looking for complex problems in population genetics (Rocha, 2018). Unfortunately, there is no such thing as a ‘unified field of bacterial population genetics’ (Ashley Robinson et al., 2010, Chapter 6). The need to adapt the methodology to the specific research question and the species under study, the unique evolutionary nature of those species and the lack of explicit analytical frameworks is what makes bacterial population genetics one of the most challenging, yet exciting, fields.

Arguably, this field is moving rapidly and data are accumulating on an unprecedented scale. Combinations of novel models in population genetics with molecular approaches is likely to be a very rewarding one. We are still far from capturing the full extent of genetic diversity hidden in the microbial branches of the tree of life. Will computational capacity keep pace?

# References

- Aanniz, T., Ouadghiri, M., Melloul, M., Swings, J., Elfahime, E., Ibijbijen, J., ... Amar, M. (2015). Thermophilic bacteria in Moroccan hot springs, salt marshes and desert soils. *Brazilian Journal of Microbiology: [publication of the Brazilian Society for Microbiology]*, 46(2), 443–453.
- Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11), 1144–1146.
- Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., ... Bork, P. (2011). Enterotypes of the humangut microbiome. *Nature*, 473(7346), 174–180.
- Ashley Robinson, D., Feil, E. J., & Falush, D. (2010). *Bacterial Population Genetics in Infectious Disease*. John Wiley & Sons.
- Beaumont, M. A., & Rannala, B. (2004). The bayesian revolution in genetics. *Nature Reviews. Genetics*, 5(4), 251–261.
- Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4), 2025–2035.
- Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, 98(6), 236–238.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 57(1), 289–300.
- Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., ... Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1), 10.
- Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., ... Howe, A. (2017). Strategies to improve reference databases for soil microbiomes. *The ISME Journal*, 11(4),

829–834.

- Collado, M. C., Rautava, S., Aakko, J., Isolauri, E., & Salminen, S. (2016). Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Scientific Reports*, 6, 23129.
- Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11), 987–991.
- Costea, P. I., Hildebrand, F., Manimozhiyan, A., Bäckhed, F., Blaser, M. J., Bushman, F. D., ... Bork, P. (2017). Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3(1), 8–16.
- Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7), 410–418.
- D’Argenio, V. (2018). Human Microbiome Acquisition and Bioinformatic Challenges in Metagenomic Studies. *International Journal of Molecular Sciences*, 19(2). <https://doi.org/10.3390/ijms19020383>
- Depaulis, F., & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution*, 15(12), 1788–1790.
- Doris, P. A. (2002). Hypertension genetics, single nucleotide polymorphisms, and the common disease:common variant hypothesis. *Hypertension*, 39(2 Pt 2), 323–331.
- Elleouet, J. S., & Aitken, S. N. (2018). Exploring Approximate Bayesian Computation for inferring recent demographic history with genomic markers in nonmodel species. *Molecular Ecology Resources*, 18(3), 525–540.
- Ewald, D. R., & Sumner, S. C. J. (2018). Human microbiota, blood group antigens, and disease. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 10(3), 1–44.
- Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3), 1405–1413.
- Foster, K. R., Schluter, J., Coyte, K. Z., & Rakoff-Nahoum, S. (2017). The evolution of the host microbiome as an ecosystem on a leash. *Nature*, 548(7665), 43–51.
- Fransen, F., van Beek, A. A., Borghuis, T., Meijer, B., Hugenholtz, F., van der Gaast-de Jongh, C., ... de Vos, P. (2017). The impact of gut microbiota on gender-specific differences in immunity. *Frontiers in Immunology*, 8(JUN). <https://doi.org/10.3389/fimmu.2017.00754>

- Glickman, M. E., & van Dyk, D. A. (2007). Basic Bayesian methods. *Methods in Molecular Biology*, 404, 319–338.
- Gorvitovskaia, A., Holmes, S. P., & Huse, S. M. (2016). Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome*, 4, 1–12.
- Haro, C., Rangel-Zúñiga, O. A., Alcalá-Díaz, J. F., Gómez-Delgado, F., Pérez-Martínez, P., Delgado-Lista, J., ... Camargo, A. (2016). Intestinal Microbiota Is Influenced by Gender and Body Mass Index. *PloS One*, 11(5), e0154090.
- Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ*, 323(7309), 391–393.
- Hein, J., Schierup, M., & Wiuf, C. (2004). *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. Oxford University Press, USA.
- Hein, J., Schierup, M., & Wiuf, C. (2005). *Gene Genalogies Variation and Evolution*.
- Huang, K., Brady, A., Mahurkar, A., White, O., Gevers, D., Huttenhower, C., & Segata, N. (2014). MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Research*, 42(Database issue), D617–D624.
- Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2), 337–338.
- Human Microbiome Jumpstart Reference Strains Consortium, Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., ... Zhu, D. (2010). A catalog of reference genomes from the human microbiome. *Science*, 328(5981), 994–999.
- Human Microbiome Project Consortium. (2012a). A framework for human microbiome research. *Nature*, 486(7402), 215–221.
- Human Microbiome Project Consortium. (2012b). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207–214.
- Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7), R143.
- Huse, S. M., Ye, Y., Zhou, Y., & Fodor, A. A. (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. *PloS One*, 7(6), e34242.
- Illumina. (2011). Quality Scores for Next-Generation Sequencing.

- [Http://Res.Illumina.Com/Documents/Products/Technotes/Technote\\_Q-Scores.Pdf](http://res.illumina.com/Documents/Products/Technotes/Technote_Q-Scores.Pdf), 1–2.
- Integrative HMP (iHMP) Research Network Consortium. (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host & Microbe*, 16(3), 276–289.
- Izard, J., & Rivera, M. C. (2014). *Metagenomics for Microbiology* (pp. 1–175).
- Jarchum, I., & Pamer, E. G. (2011). Regulation of innate and adaptive immunity by the commensal microbiota. *Current Opinion in Immunology*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0952791511000239>
- Jeffery, I. B., Claesson, M. J., & O'toole, P. W. (2012). Categorization of the gut microbiota: enterotypes or gradients? *Nature Reviews*. Retrieved from <https://www.nature.com/articles/nrmicro2859>
- Jiménez, E., Fernández, L., Marín, M. L., Martín, R., Odriozola, J. M., Nueno-Palop, C., ... Rodríguez, J. M. (2005). Isolation of commensal bacteria from umbilical cord blood of healthy neonates born by cesarean section. *Current Microbiology*, 51(4), 270–274.
- Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1). <https://doi.org/10.2202/1544-6115.1389>
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795.
- Kato, C., Sato, T., & Horikoshi, K. (1995). Isolation and properties of barophilic and barotolerant bacteria from deep-sea mud samples. *Biodiversity & Conservation*, 4(1), 1–9.
- Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, 146(3), 1197–1206.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2), 111–120.
- Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, 13(3), 235–248.
- Knights, D., Ward, T. L., McKinlay, C. E., Miller, H., Gonzalez, A., McDonald, D., & Knight, R. (2014). Rethinking enterotypes. *Cell Host & Microbe*, 16(4), 433–437.
- Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., ... Ley, R. E. (2013). A

- guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Computational Biology*, 9(1), e1002863.
- Korpelainen, E., Tuimala, J., Somervuo, P., Huss, M., & Wong, G. (2014). *RNA-seq Data Analysis A Practical Approach* (p. 322).
- Kozlov, A. (2018). *Amkozlov/Raxml-Ng: Raxml-Ng V0.6.0 Beta*. Zenodo.  
<https://doi.org/10.5281/ZENODO.593079>
- Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., & Knight, R. (2011). Experimental and analytical tools for studying the human microbiome. *Nature Reviews. Genetics*, 13(1), 47–58.
- Kundu, P., Blacher, E., Elinav, E., & Pettersson, S. (2017). Our Gut Microbiome: The Evolving Inner Self. *Cell*, 171(7), 1481–1493.
- Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews: MMBR*, 72(4), 557–578, Table of Contents.
- Kushner, S. R. (2015). Polyadenylation in *E. coli*: a 20 year odyssey. *RNA*, 21(4), 673–674.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359.
- Lapierre, M., Blin, C., Lambert, A., Achaz, G., & Rocha, E. P. C. (2016). The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. *Molecular Biology and Evolution*, 33(7), 1711–1725.
- Laurent, S. (2011). *Statistical inference of complex demographic models in Drosophila melanogaster and two wild tomato species* (Text.PhDThesis). Ludwig-Maximilians-Universität München.  
Retrieved from <https://edoc.ub.uni-muenchen.de/12641/>
- LeBlanc, J. G., Milani, C., de Giori, G. S., & Sesma, F. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Current Opinion in*. Retrieved from  
<https://www.sciencedirect.com/science/article/pii/S095816691200119X>
- Levins, R. (1969). Some Demographic and Genetic Consequences of Environmental Heterogeneity for Biological Control. *Bulletin of the Entomological Society of America*, 15(3), 237–240.
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node

- solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10), 1674–1676.
- Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, 8(1), 51.
- Louca, S., Doebeli, M., & Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, 6(1), 41.
- Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews. Genetics*, 4(12), 981–994.
- Luo, C., Tsementzi, D., Kyrpides, N. C., & Konstantinidis, K. T. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *The ISME Journal*, 6(4), 898–901.
- Madigan, M. T., Martinko J, M, Parker J. (2006). Brock biology of microorganisms. Prentice Hall.  
Retrieved from <http://walikota-travel.com/brocks-biology-of-microorganisms-8th.pdf>
- Micah, H., Claire, F.-L., Rob, K., & Others. (2007). The Human Microbiome Project: Exploring the Microbial Part of Ourselves in a Changing World. *Nature*, 449(7164), 804–810.
- Morgulis, A., Gertz, E. M., Schäffer, A. A., & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 13(5), 1028–1040.
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31(3), 274–295.
- Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40(20), e155.
- National Research Council, Division on Earth and Life Studies, Board on Life Sciences, & Committee on Metagenomics: Challenges and Functional Applications. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. National Academies Press.
- Odintsova, V., Tyakht, A., & Alexeev, D. (2017). Guidelines to Statistical Analysis of Microbial Composition Data Inferred from Metagenomic Sequencing. In *Metagenomics: Current Advances and Emerging Concepts*. Caister Academic Press.



- O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Reports*. Retrieved from [http://embor.embopress.org/content/7/7/688.abstract?casa\\_token=Dul0I7AR3hwAAAAA:7viV1ik\\_UU9bpNMVLGUbRCPzpJK4mjF\\_SfQEBnloBuhBp8psgXD5bu1AM6Ifjl9fgfSJopGzaWjNoepVaw](http://embor.embopress.org/content/7/7/688.abstract?casa_token=Dul0I7AR3hwAAAAA:7viV1ik_UU9bpNMVLGUbRCPzpJK4mjF_SfQEBnloBuhBp8psgXD5bu1AM6Ifjl9fgfSJopGzaWjNoepVaw)
- Oulas, A., Pavlodi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., ... Iliopoulos, I. (2015). Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies, 75–88.
- Pavlidis, P., Laurent, S., & Stephan, W. (2010). msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, 10(4), 723–727.
- Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2011). Meta-IDBA: A de Novo assembler for metagenomic data. *Bioinformatics*, 27(13), 94–101.
- Perez-Muñoz, M. E., Arrieta, M.-C., Ramer-Tait, A. E., & Walter, J. (2017). A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome*, 5(1), 48.
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., ... Wang, J. (2010). Human gut microbial gene catalogue established by metagenomic sequencing. In *Nature* (Vol. 464, pp. 59–65).
- Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9), 833–844.
- Rocha, E. P. C. (2018). Neutral theory , microbial practice : challenges in bacterial population genetics. *Molecular Biology and Evolution*, 35(April), 1–26.
- Rokas, A. (2011). Phylogenetic Analysis of Protein Sequence Data Using the Randomized Accelerated Maximum Likelihood ( RAXML ) Program. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, (October), 1–14.
- Salemi, M., Vandamme, A.-M., & Lemey, P. (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press.
- Salminen, S., Gibson, G. R., McCartney, A. L., & Isolauri, E. (2004). Influence of mode of delivery on gut microbiota composition in seven year old children. *Gut*, 53(9), 1388–1389.
- Samuel, B. S., Hansen, E. E., Manchester, J. K., Coutinho, P. M., Henrissat, B., Fulton, R., ...

- Gordon, J. I. (2007). Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. *Proceedings of the National Academy of Sciences of the United States of America*, 104(25), 10643–10648.
- Sangwan, N., Xia, F., & Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, 4(1), 8.
- Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), 863–864.
- Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., & Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Molecular Systems Biology*, 9(1), 1–15.
- Sender, R., Fuchs, S., & Milo, R. (2016a). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, 164(3), 337–340.
- Sender, R., Fuchs, S., & Milo, R. (2016b). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8), e1002533.
- Shapira, M. (2016). Gut Microbiotas and Host Evolution: Scaling Up Symbiosis. *Trends in Ecology & Evolution*, 31(7), 539–549.
- Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., & Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1), 111–120.
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5(June), 209.
- Shoemaker, J. S., Painter, I. S., & Weir, B. S. (1999). Bayesian statistics in genetics: A guide for the uninitiated. *Trends in Genetics: TIG*, 15(9), 354–358.
- Steinberg, K. M., Schneider, V. A., Alkan, C., Montague, M. J., Warren, W. C., Church, D. M., & Wilson, R. K. (2017). Building and Improving Reference Genome Assemblies. *Proceedings of the IEEE*, 105(3), 422–435.
- Stilling, R. M., Dinan, T. G., & Cryan, J. F. (2014). Microbial genes, brain & behaviour - epigenetic regulation of the gut-brain axis. *Genes, Brain, and Behavior*, 13(1), 69–86.
- Strous, M., Kraft, B., Bisdorf, R., & Tegetmeyer, H. E. (2012). The binning of metagenomic contigs for

- microbial physiology of mixed cultures. *Frontiers in Microbiology*, 3(DEC), 1–11.
- Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1).  
<https://doi.org/10.1371/journal.pcbi.1002803>
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2), 437–460.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585–595.
- Taneja, V. (2017). Microbiome: Impact of Gender on Function & Characteristics of Gut Microbiome. *Principles of Gender-Specific Medicine: Gender in the Genomic Era: Third Edition*, 569–583.
- Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1), 3.
- Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal*, 474(11), 1823–1836.
- Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 63(2), 411–423.
- Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., ... Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10), 902–903.
- Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes, 626–638.
- Turnbaugh, P. J., Ley, R. E., Mahowald, M. a., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122), 1027–1031.
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., ... Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978), 37–43.
- Větrovský, T., & Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and

- its consequences for bacterial community analyses. *PloS One*, 8(2), e57923.
- Wakeley, J., & Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics*, 159(2), 893–905.
- Wang, X. (2016). *Next-Generation Sequencing Data Analysis* (p. 213).
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2), 256–276.
- Wexler, A. G., & Goodman, A. L. (2017). An insider's perspective: Bacteroides as a window into the microbiome. *Nature Microbiology*, 2, 17026.
- Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, 6(2), e1000667.
- Woo, P. C. Y., Lau, S. K. P., Teng, J. L. L., Tse, H., & Yuen, K.-Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, 14(10), 908–934.
- Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., ... Lewis, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052), 105–108.
- Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford University Press.
- Yatsunenکو, T., Rey, F. E., Manary, M. J., & Trehan, I. (2012). Human gut microbiome viewed across age and geography. *Nature*. Retrieved from <https://www.nature.com/articles/nature11053>
- Zagury, G. J., Kulnieks, V. I., & Neculita, C. M. (2006). Characterization and reactivity assessment of organic substrates for sulphate-reducing bacteria in acid mine drainage treatment. *Chemosphere*, 64(6), 944–954.



# Supporting information

## Supporting Information Contents

### 1. Supplemental Tables

**Supplemental Table S1** | Information about the 61 species that were reconstructed by StrainPhlAn.

**Supplemental Table S2** | Statistical significant results of the Mann–Whitney-Wilcoxon test on bacterial abundance between male and female subjects,

### 2. Supplemental Figures

**Supplemental Figure S1** | Coverage estimation of each species' strain.

**Supplemental Figure S2** | Species' abundance distributions violate *t*-Tests's assumption of normality.

**Supplemental Figure S3** | Kernel density distributions of *Bacteroides Xylanisolvans*' and *Sutterella Wadsworthensis*' relative abundance between male and female subjects.

**Supplemental Figure S4** | *Bacteroides* and *Prevotella* sorted values of relative abundance across all samples.

**Supplemental Figure S5** | Population structure of *Alistipes shahii*, *Alistipes* sp AP11, and *Alistipes* HGB5.

**Supplemental Figure S6** | Population structure of *Bacteroides caccae*, *Bacteroides cellulosilyticus* and *Bacteroides eggerthii*.

**Supplemental Figure S7** | Population structure of *Bacteroides finegoldii*, *Bacteroides fragilis* and *Bacteroides massiliensis*.

**Supplemental Figure S8** | Population structure of *Bacteroides salyersiae*, *Bacteroides uniformis* and *Bacteroides thetaiotaomicron*.

**Supplemental Figure S9** | Population structure of *Bacteroides vulgatus*, *Bifidobacterium adolescentis* and *Bifidobacterium longum*.

**Supplemental Figure S10** | Population structure of *Clostridium* sp L2\_50, *Coprococcus* sp ART55\_1 and *Dialister invisus*.

**Supplemental Figure S11** | Population structure of *Dorea formicigenerans*, *Escherichia coli* and *Eubacterium siraeum*.

**Supplemental Figure S12** | Population structure of *Lachnospiraceae* *Bacterium* 8\_1\_57FAA, *Odoribacter splanchnicus* and *Parabacteroides distasonis*.

**Supplemental Figure S13** | Population structure of *Parabacteroides johnsonii*, *Parabacteroides merdae* and *Roseburia hominis*.

**Supplemental Figure S14** | Population structure of *Roseburia intestinalis*, *Roseburia inulinivorans* and *Ruminococcus bromii*.

**Supplemental Figure S15** | Population structure of *Ruminococcus gnavus*, *Ruminococcus lactaris* and *Sutterella wadsworthensis*.

**Supplemental Figure S16** | *The space of the ABC accepted simulations for B. ovatus' bottleneck model.*

**Supplemental Table S1 |** Information about the 61 species that were reconstructed by StrainPhlAn.

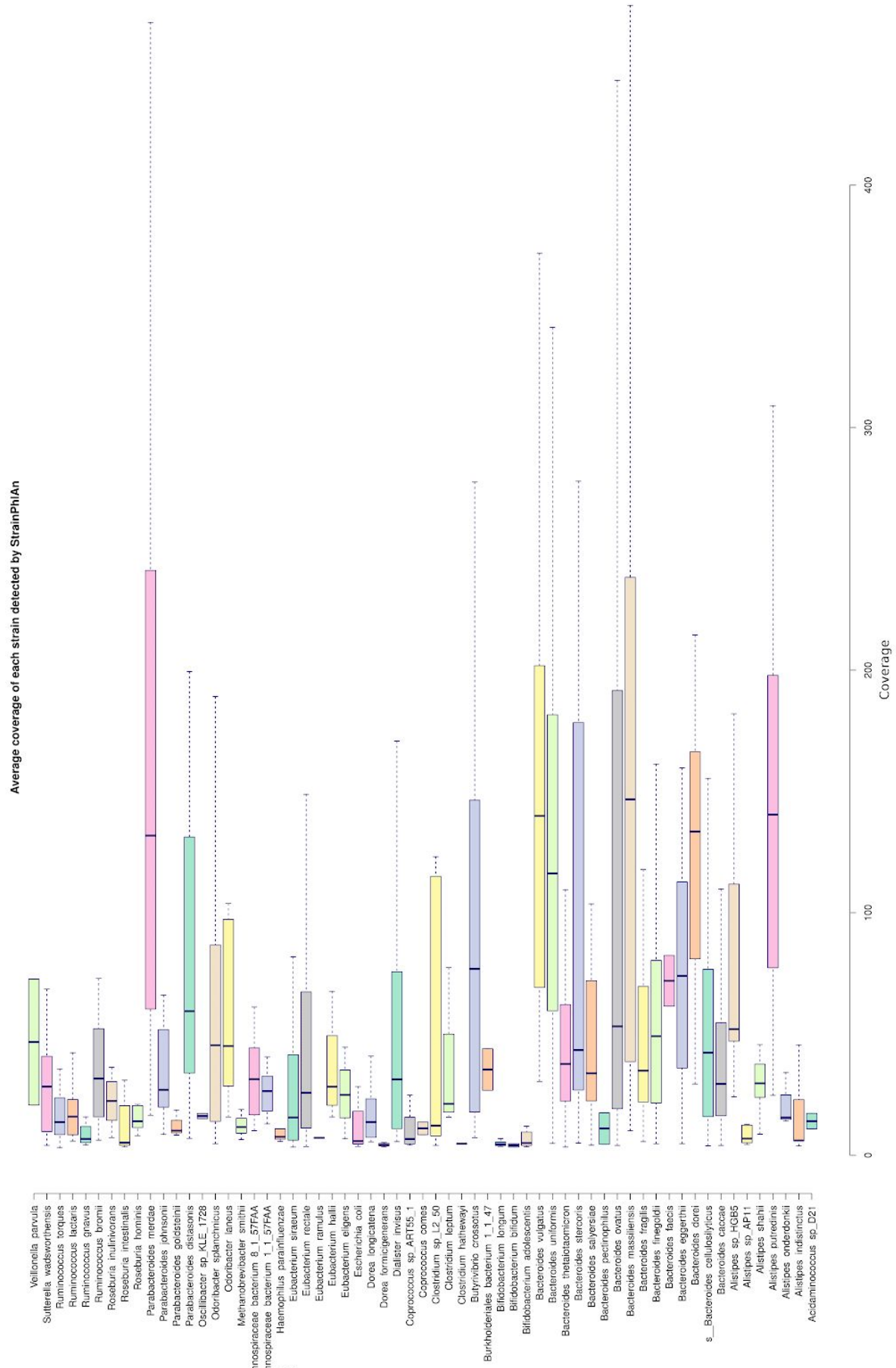
Species	Number of samples	Number of markers in the DB	Number of reconstructed markers by StrainPhlAn
Acidaminococcus_sp_D21	2	13	10
Alistipes_indistinctus	5	200	167
Alistipes_nderdonkii	3	200	154
Alistipes_putredinis	40	200	157
Alistipes_shahii	13	200	148
Alistipes_sp_AP11	14	200	148
Alistipes_sp_HGB5	7	200	164
Bacteroides_caccae	84	200	170
Bacteroides_cellulosilyticus	30	200	165
Bacteroides_dorei	23	200	152
Bacteroides_eggerthii	10	200	158
Bacteroides_faecis	2	200	159
Bacteroides_finegoldii	19	200	162
Bacteroides_fragilis	31	200	156
Bacteroides_massiliensis	21	200	152
Bacteroides_ovatus	59	200	173
Bacteroides_pectinophilus	2	200	166
Bacteroides_salysiae	12	200	179
Bacteroides_stercoris	38	200	148
Bacteroides_thetaiotaomicron	53	200	170
Bacteroides_uniformis	38	200	151
Bacteroides_vulgatus	10	200	143
Bifidobacterium_adolescentis	19	200	173
Bifidobacterium_bifidum	2	200	161
Bifidobacterium_longum	13	200	174
Burkholderiales_bacterium_1_1_47	2	200	162
Butyrivibrio_crossotus	8	200	156
Clostridium_hathewayi	2	200	158
Clostridium_leptum	4	200	149
Clostridium_sp_L2_50	7	200	163
Coprococcus_comes	2	200	149
Coprococcus_sp_ART55_1	17	200	154
Dialister_invisus	29	200	148
Dorea_formicigenerans	18	200	166
Dorea_longicatena	5	200	155
Escherichia_coli	7	24	21
Eubacterium_eligens	4	200	169
Eubacterium_hallii	4	200	137
Eubacterium_ramulus	2	200	153
Eubacterium_rectale	98	200	166
Eubacterium_siraeum	32	200	184
Haemophilus_parainfluenzae	9	200	162
Lachnospiraceae_bacterium_1_1_57FAA	7	38	31
Lachnospiraceae_bacterium_8_1_57FAA	16	21	16
Methanobrevibacter_smithii	3	200	178
Odoribacter_laneus	6	200	177
Odoribacter_splanchnicus	74	200	158
Oscillibacter_sp_KLE_1728	2	187	140
Parabacteroides_distasonis	13	200	172
Parabacteroides_goldsteinii	3	200	150
Parabacteroides_johnsonii	10	200	173
Parabacteroides_merdae	81	200	177
Roseburia_hominis	11	200	155
Roseburia_intestinalis	52	200	175
Roseburia_inulinivorans	11	200	151
Ruminococcus_bromii	22	200	152
Ruminococcus_gnavus	14	200	171
Ruminococcus_lactaris	18	200	174
Ruminococcus_torques	56	8	8
Sutterella_wadsworthensis	36	200	187
Veillonella_parvula	2	200	164



**Supplemental Table S2** | Statistical significant results of the Mann–Whitney–Wilcoxon test on bacterial abundance between male and female subjects, with a significance level of 0.05. The sample size of the male and female groups was 78 and 56, respectively.

Species	Mann-Whitney U	Pvalue	Adjusted Pvalue
<i>Bacteroides clarus</i>	2816	0.0042	0.1214
<i>Bacteroides eggerthii</i>	2648	0.023	0.2822
<i>Bacteroides massiliensis</i>	2729	0.0387	0.3043
<i>Bacteroides stercoris</i>	2960	0.0033	0.1214
<i>Bacteroides uniformis</i>	1804	0.0308	0.2822
<i>Bacteroides xylanisolvens</i>	3113	0.0004	0.02826
<i>Butyricimonas synergistica</i>	2714.5	0.0266	0.2822
<i>Parabacteroides johnsonii</i>	2611.5	0.0454	0.3043
<i>Eubacterium siraeum</i>	1853	0.0435	0.3043
<i>Ruminococcus obeum</i>	1840	0.0453	0.3043
<i>Butyrivibrio crossotus</i>	2532	0.0263	0.2822
<i>Coprococcus eutactus</i>	2596	0.0045	0.1214
<i>Lachnospiraceae bacterium 3 1 57FAA CT1</i>	1861	0.024	0.2822
<i>Roseburia intestinalis</i>	2813.5	0.0256	0.2822
<i>Roseburia inulinivorans</i>	2789.5	0.0337	0.2822
<i>Bacteroides clarus</i>	2740	0.0332	0.2822
<i>Catenibacterium mitsuokai</i>	2532	0.0263	0.2822
<i>Clostridium spiroforme</i>	2030	0.0253	0.2822
<i>Sutterella wadsworthensis</i>	3024	0.0004	0.02826
<i>Escherichia coli</i>	1824	0.0328	0.2822

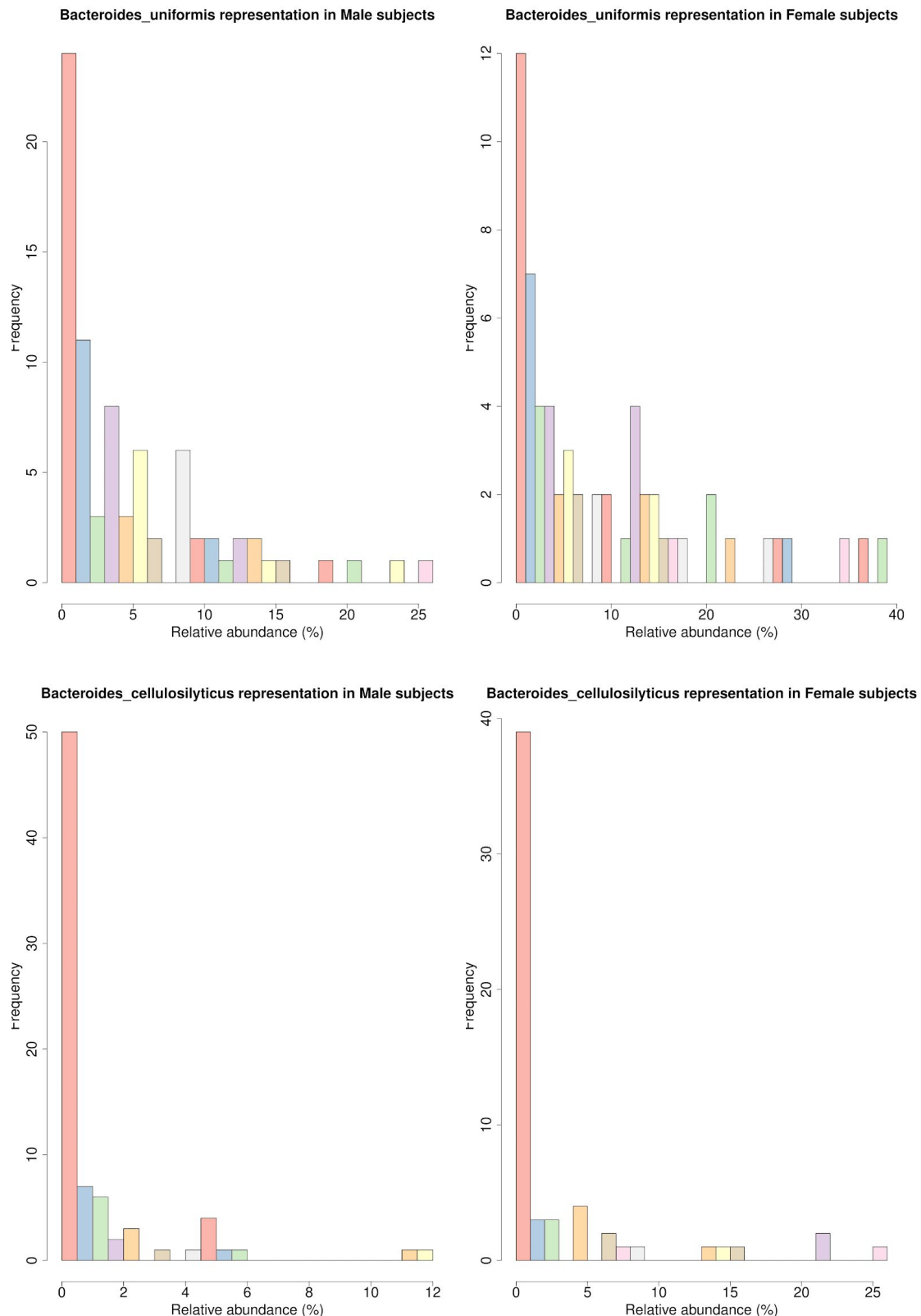
**Supplemental Figure S1** | Coverage estimation of each species' strain. StrainPhlAn calculates the average coverage value for each sample's detected strain of a given species, while this box-plot represents the distributional characteristics of these values, grouped by species.



---

**Supplemental Figure S2** | Species' abundance distributions violate  $t$ -Tests's assumption of normality. Most of the detected species demonstrate severely skewed distributions. For demonstration we present the distributions of *Bacteroides Uniformis* and *Bacteroides Cellulolyticus*. If we had applied a  $t$ -Tests, those species would have produced statistically significant difference in mean values between the male and female groups, driven by the presence of outliers. The characteristic of the distributions lead us to perform a non parametric test, instead.

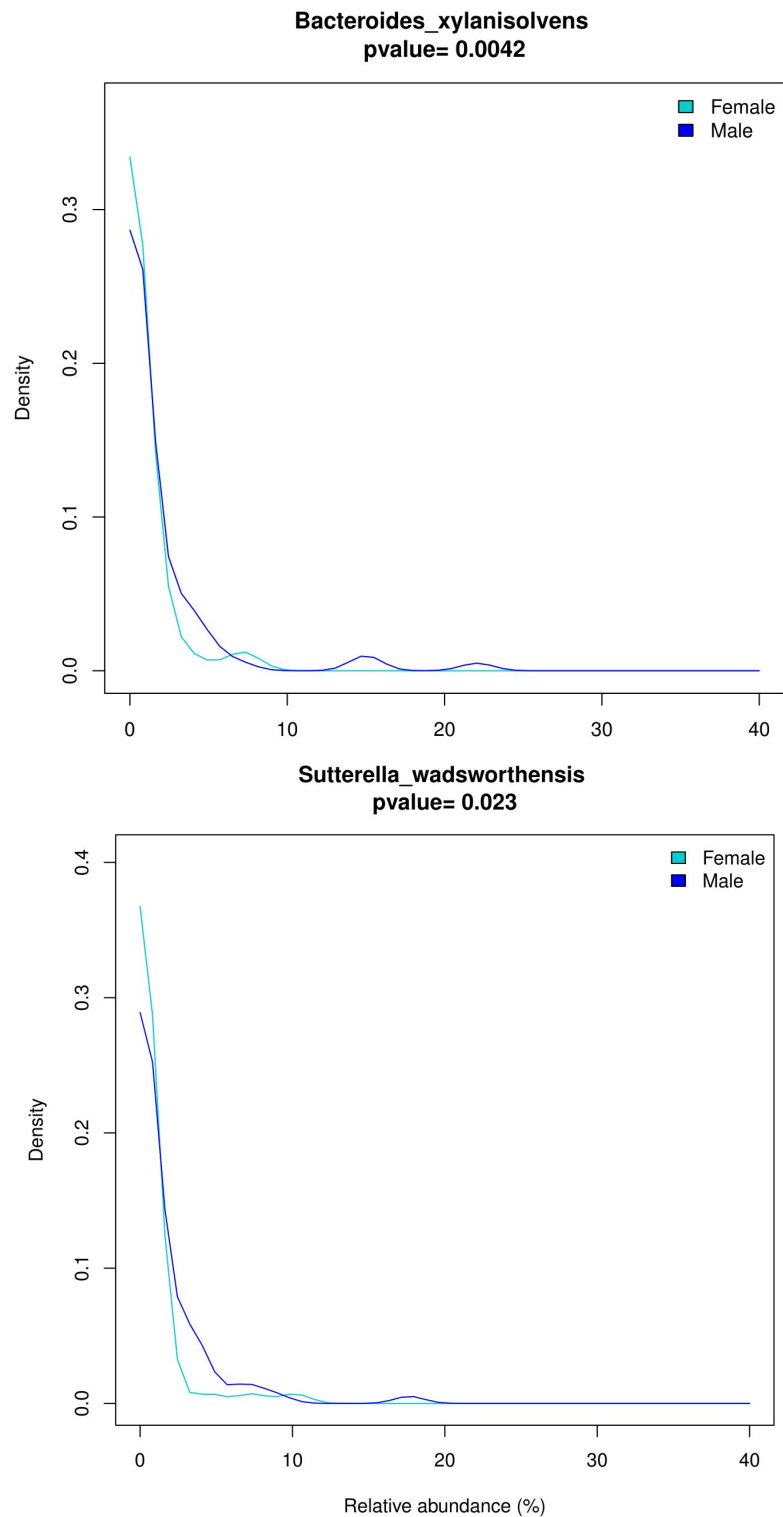
---



---

**Supplemental Figure S3** | Kernel density distributions of *Bacteroides Xylanisolvens*' and *Sutterella Wadsworthensis*' relative abundance between male and female subjects. The Mann-Whitney-Wilcoxon test estimated a statistical significant difference between the two gender groups.

---

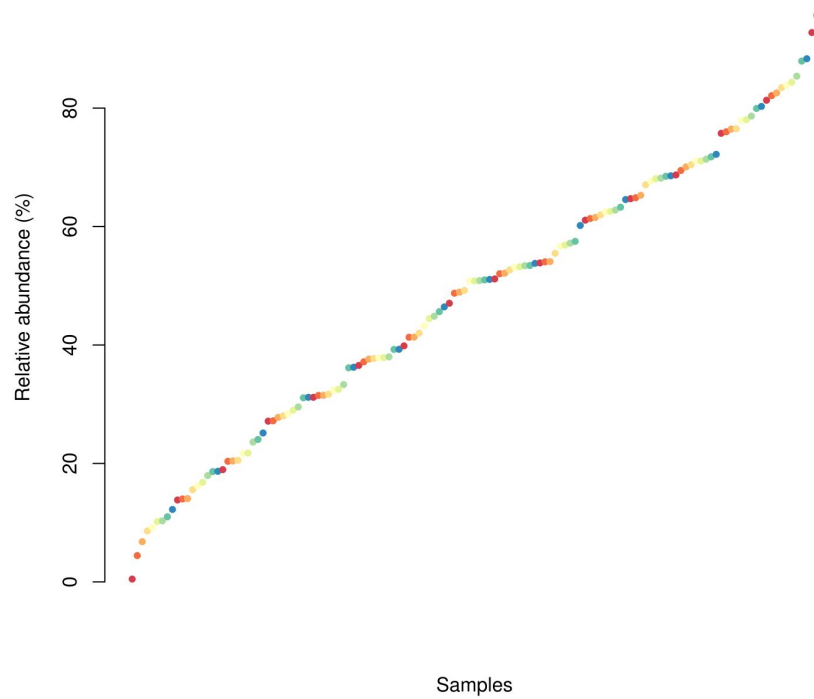


---

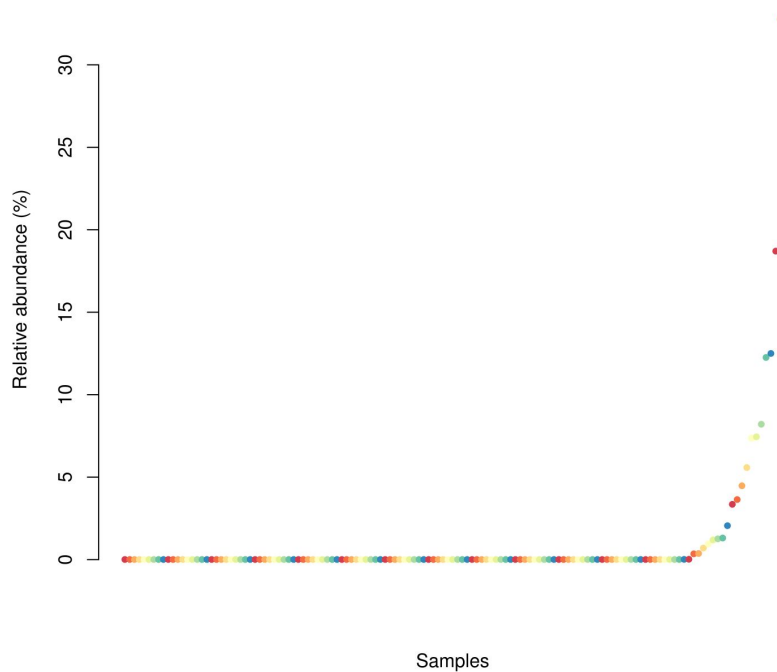
**Supplemental Figure S4** | *Bacteroides* and *Prevotella* sorted values of relative abundance across all samples. In *Bacteroides* the values form a continuous gradient, while in *Prevotella* the distribution is negatively skewed, with a long left tail and a discrete-like increase of values following the tail.

---

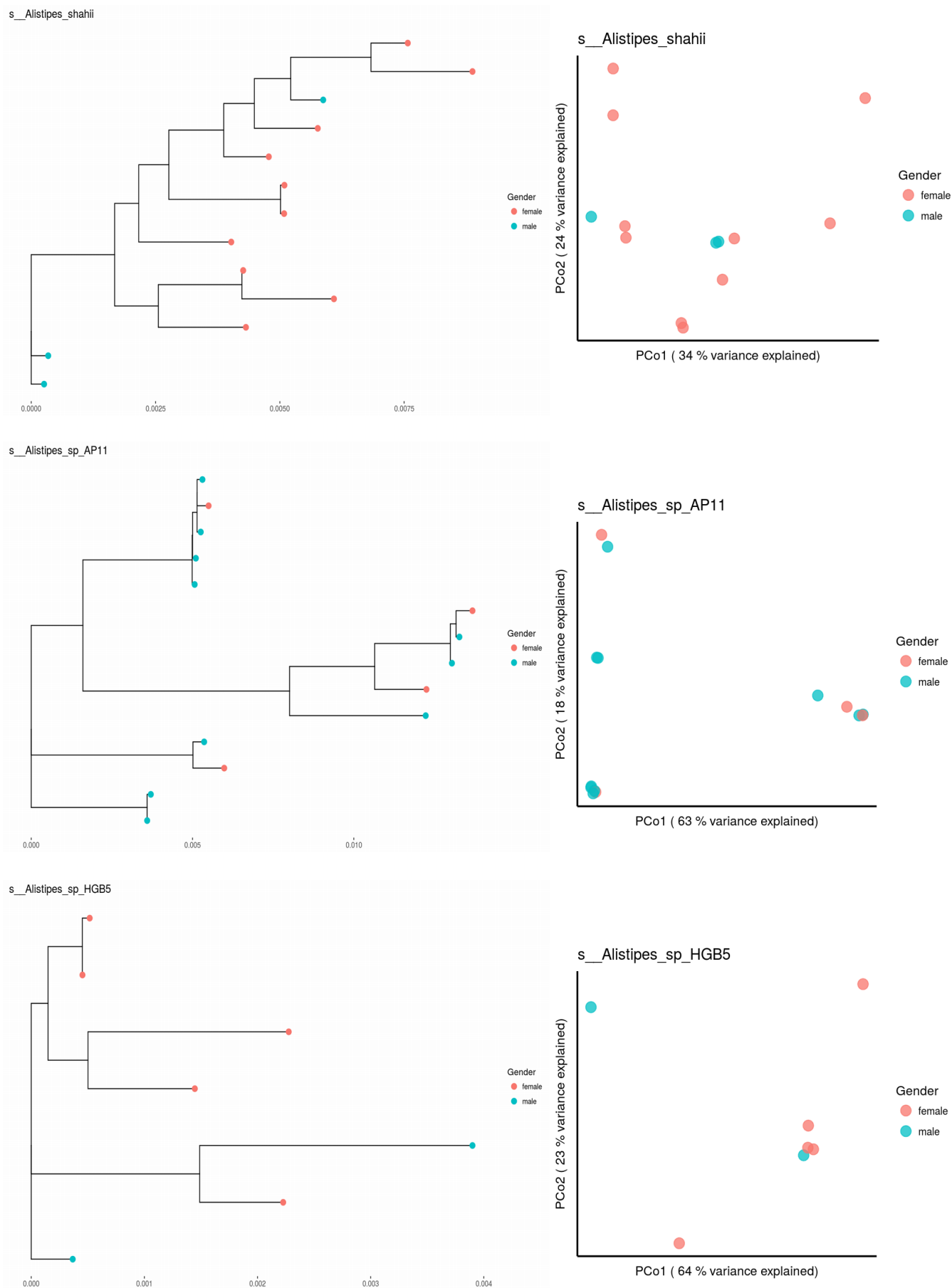
### Bacteroides sorted abundance values



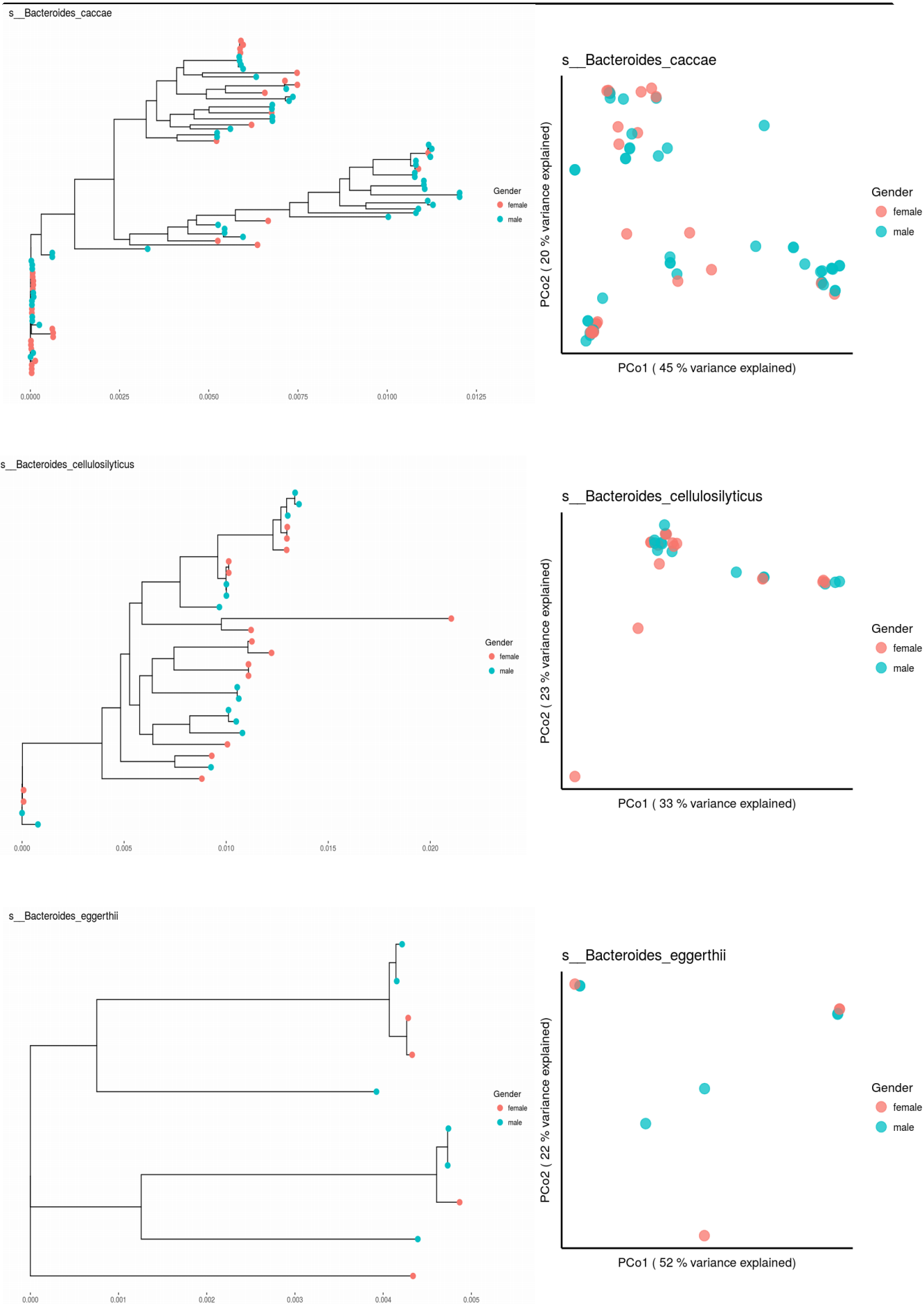
### Prevotella sorted abundance values



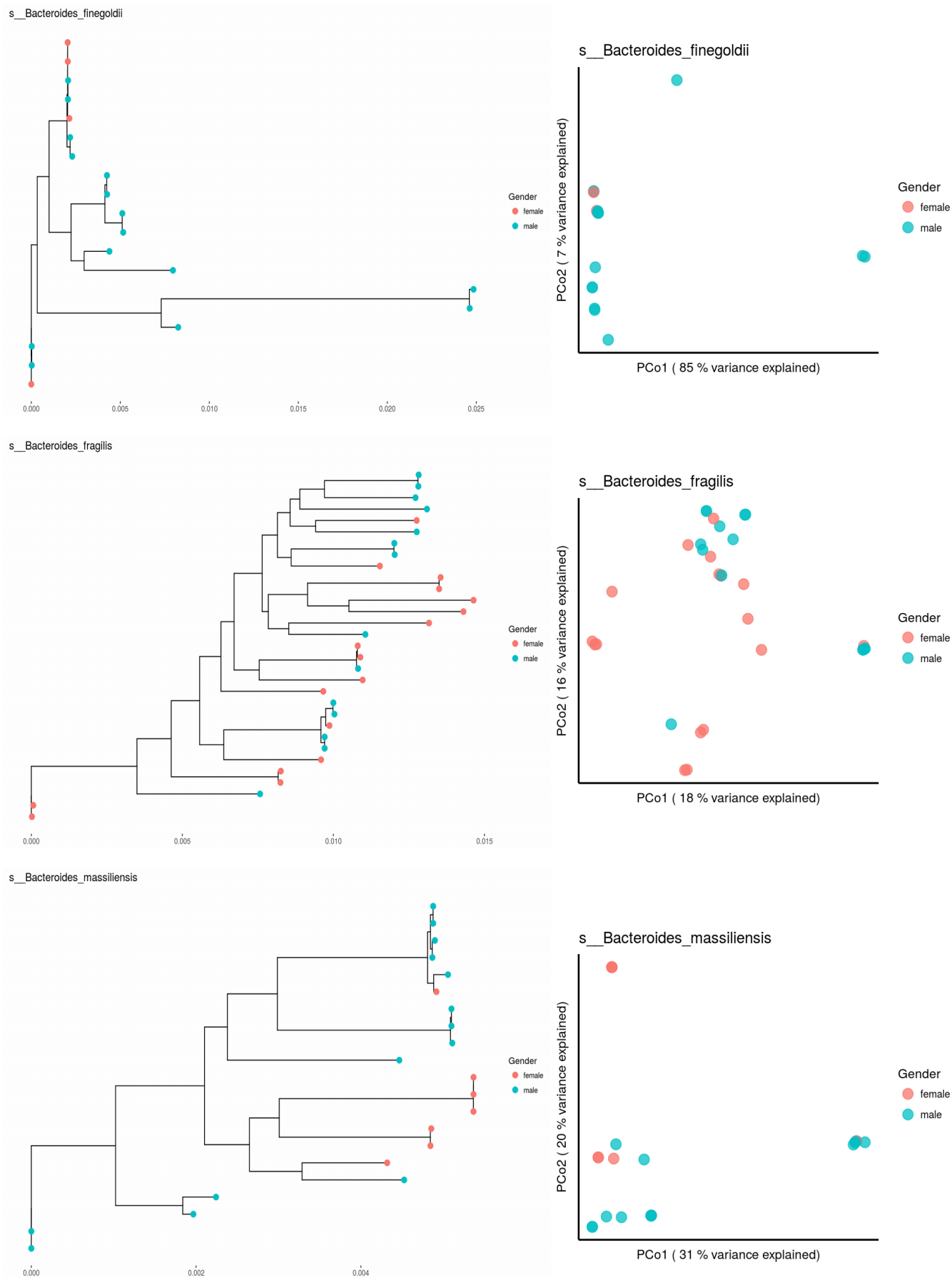
**Supplemental Figure S5 |** Population structure of *Alistipes shahii*, *Alistipes* sp AP11, and *Alistipes* HGB5. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.



**Supplemental Figure S6 |** Population structure of *Bacteroides caccae*, *Bacteroides cellulosilyticus* and *Bacteroides eggerthii*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.

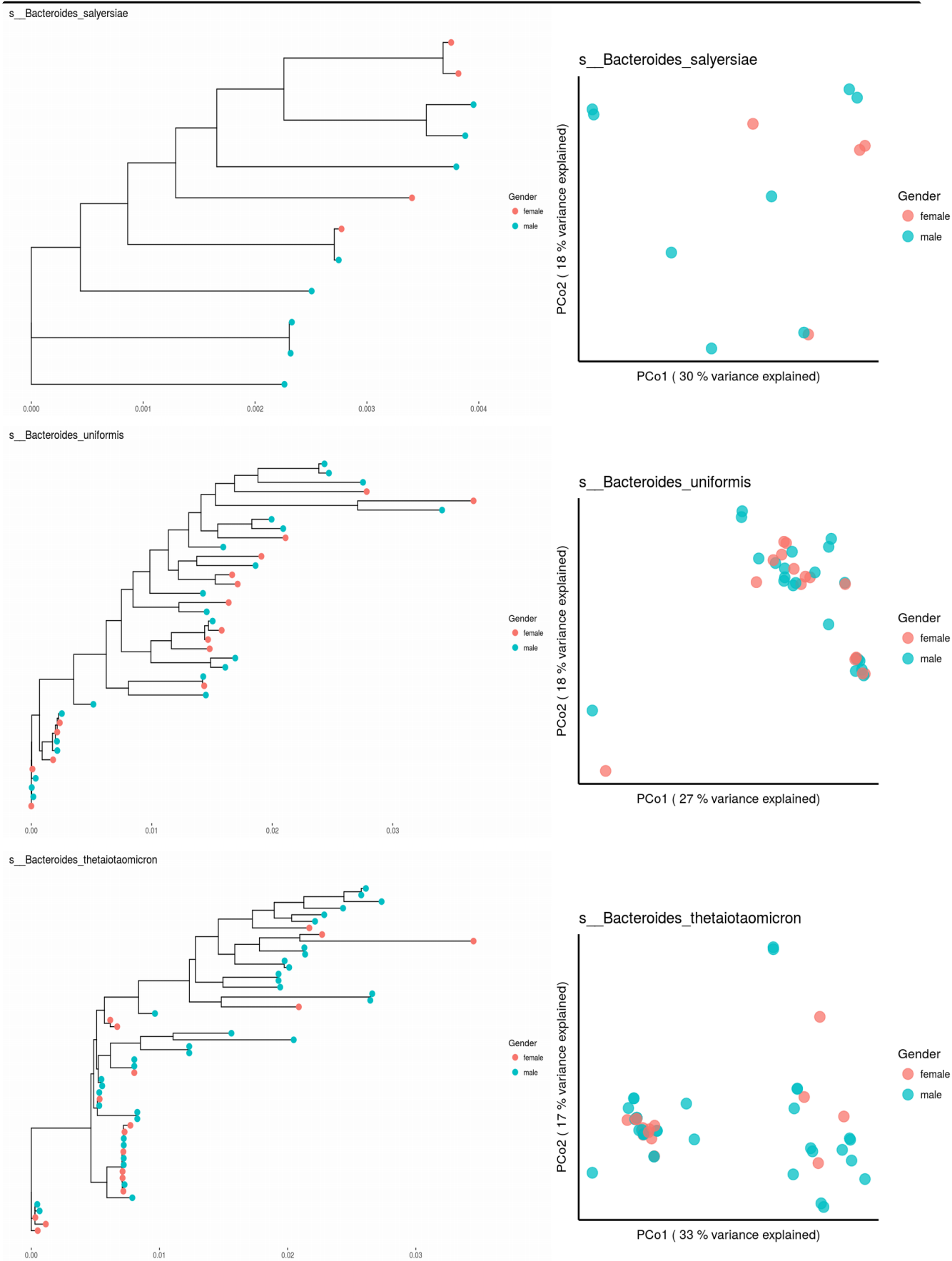


**Supplemental Figure S7 |** Population structure of *Bacteroides finegoldii*, *Bacteroides fragilis* and *Bacteroides massiliensis*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.

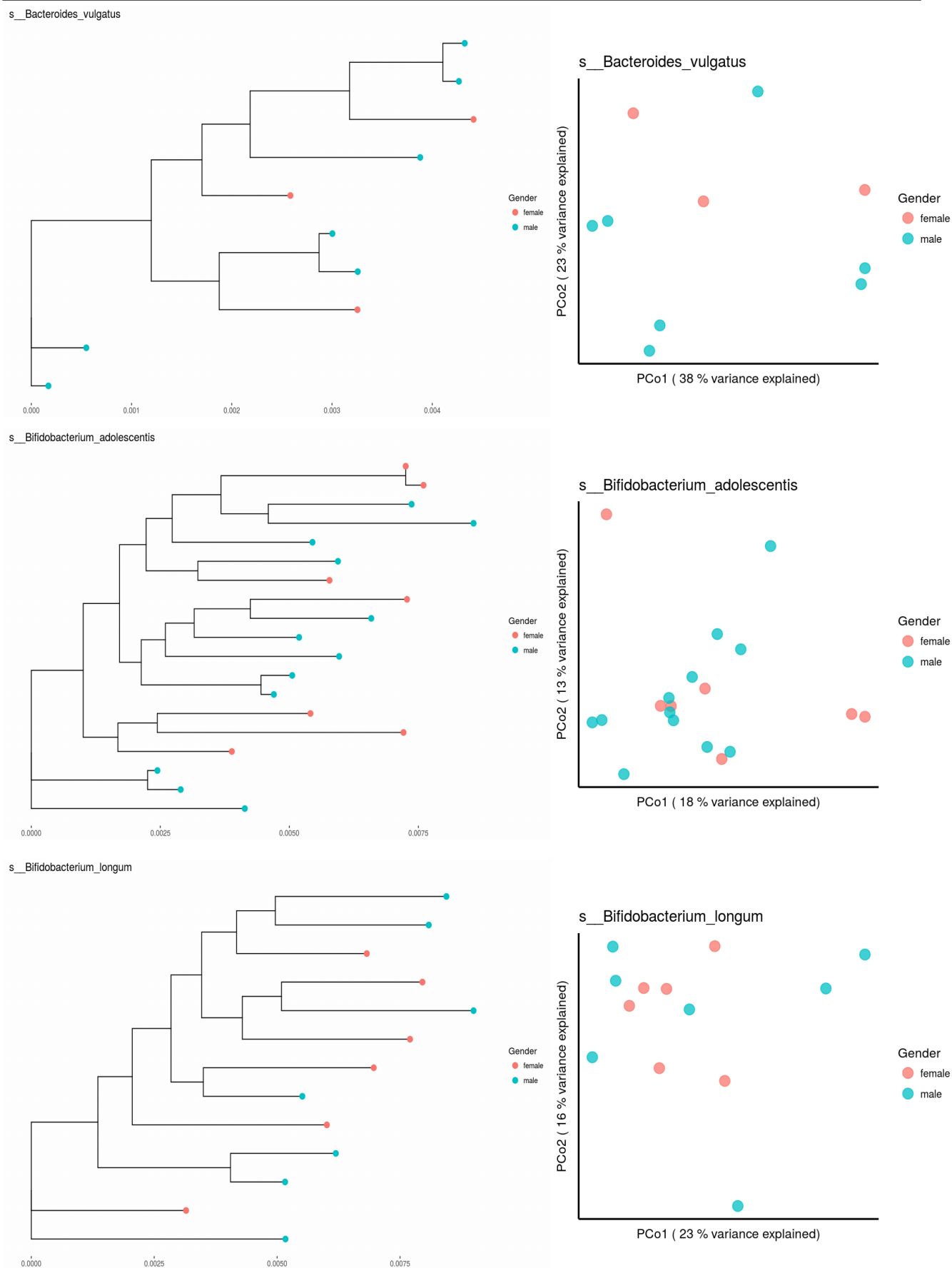




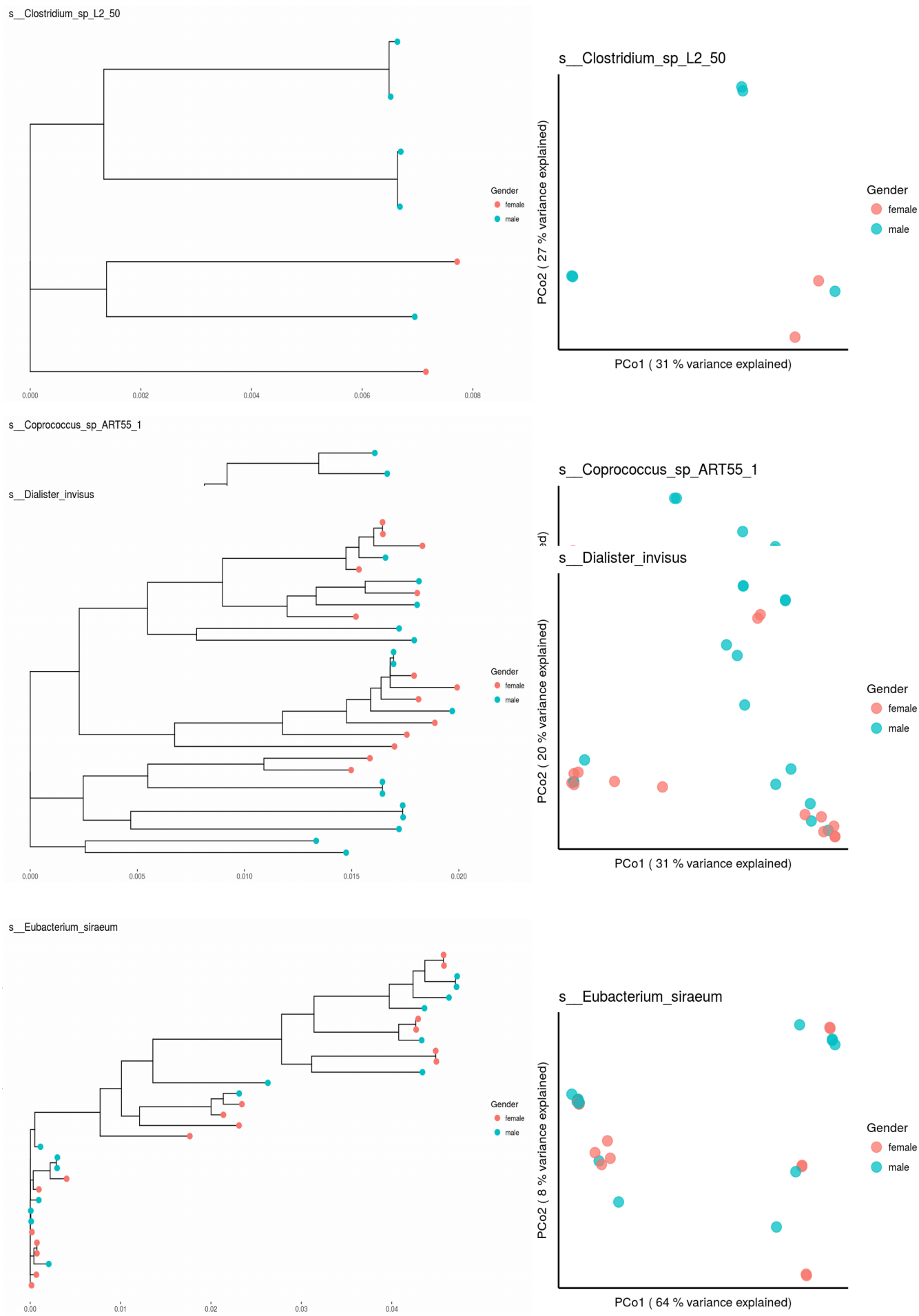
**Supplemental Figure S8 |** Population structure of *Bacteroides salyersiae*, *Bacteroides uniformis* and *Bacteroides thetaiotaomicron*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.



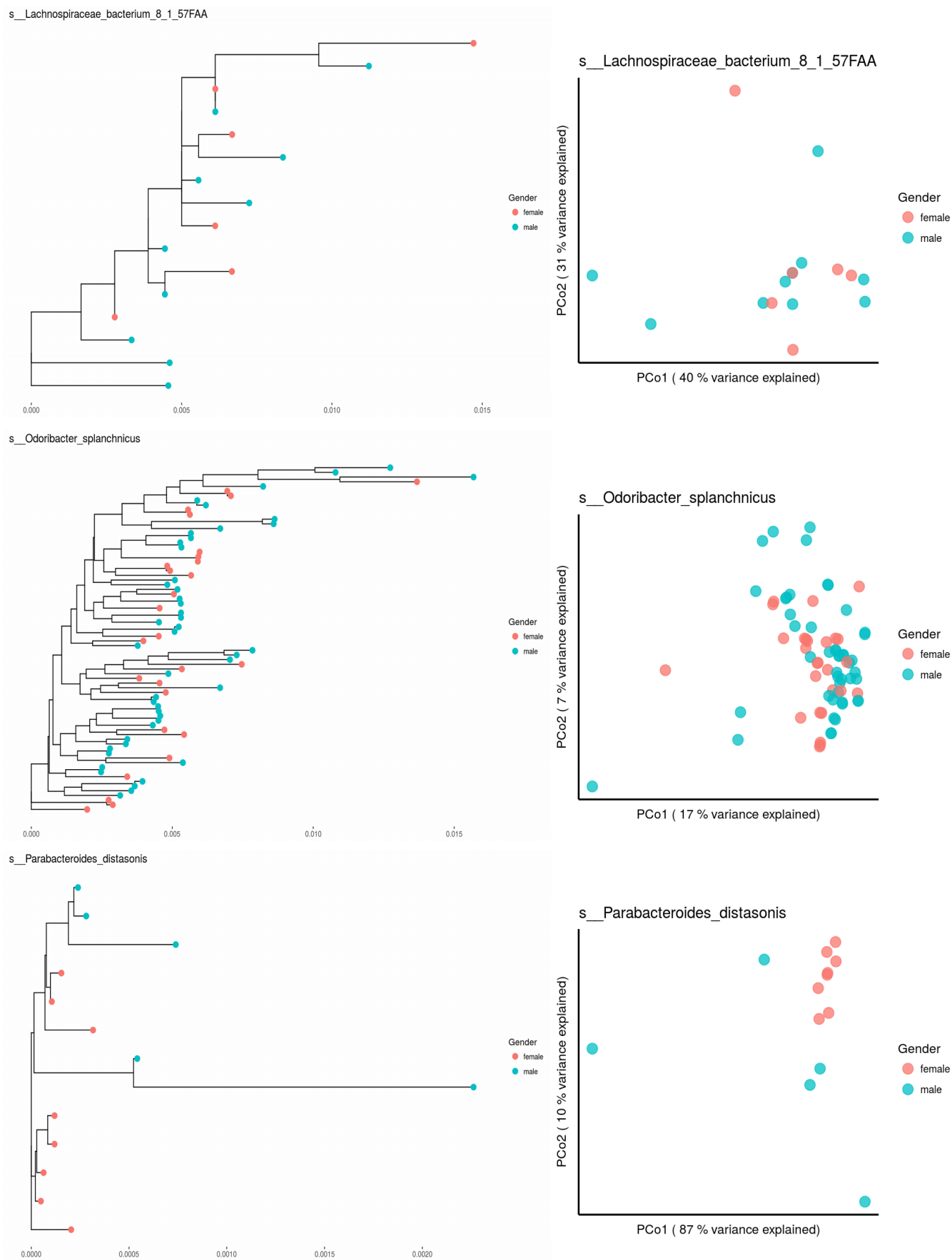
**Supplemental Figure S9 |** Population structure of *Bacteroides vulgatus*, *Bifidobacterium adolescentis* and *Bifidobacterium longum*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.



**Supplemental Figure S10 |** Population structure of *Clostridium* sp L2\_50, *Coproccoccus* sp ART55\_1 and *Dialister invisus*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.

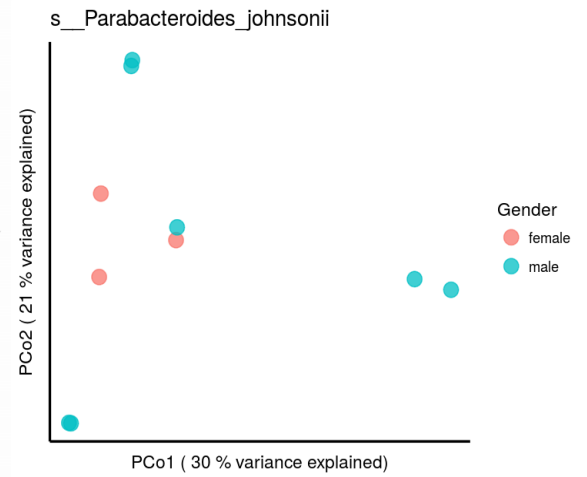
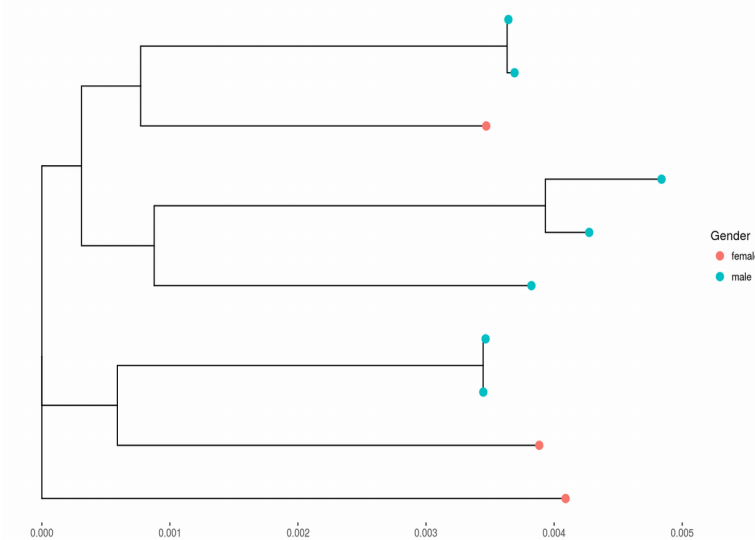


**Supplemental Figure S12 |** Population structure of *Lachnospiraceae bacterium 8\_1\_57FAA*, *odoribacter Splanchnicus* and *Parabacteroides distasonis*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.

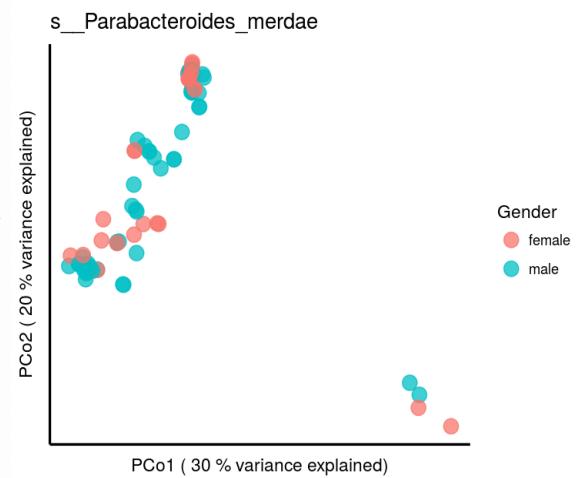
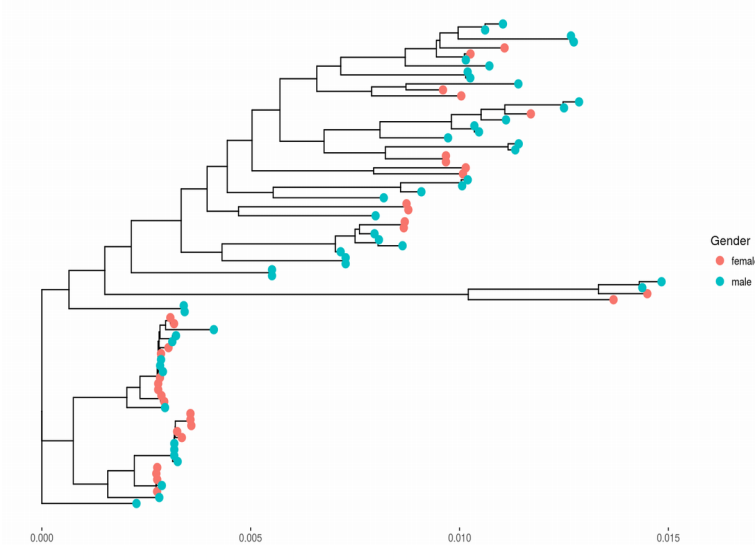


**Supplemental Figure S13** | Population structure of *Parabacteroides johnsonii*, *Parabacteroides merdae* and *Roseburia hominis*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.

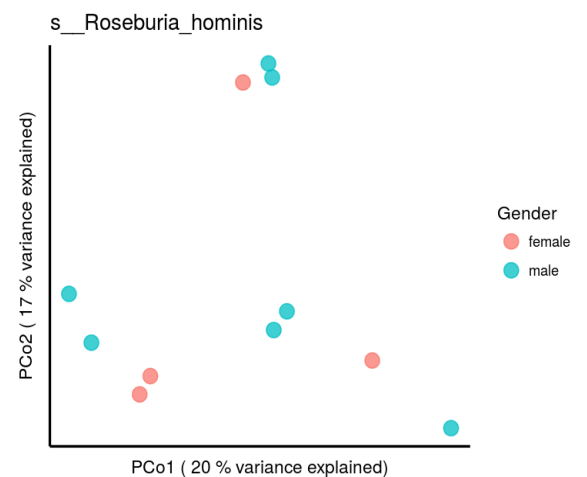
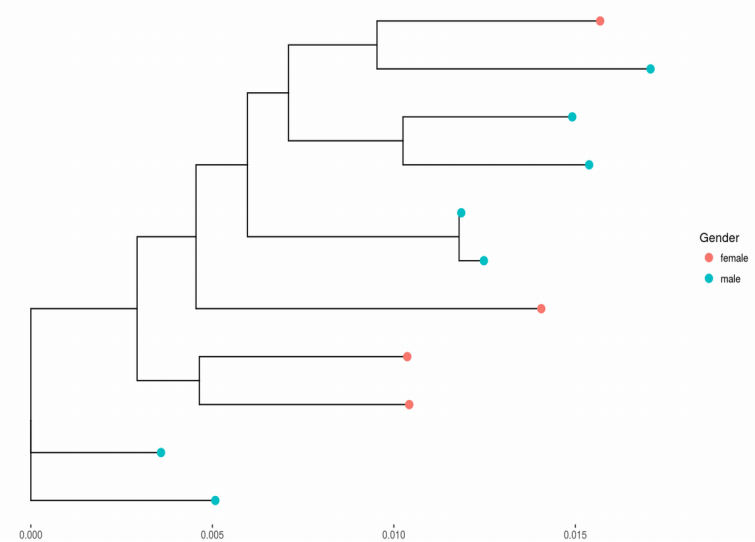
s\_\_Parabacteroides\_johnsonii



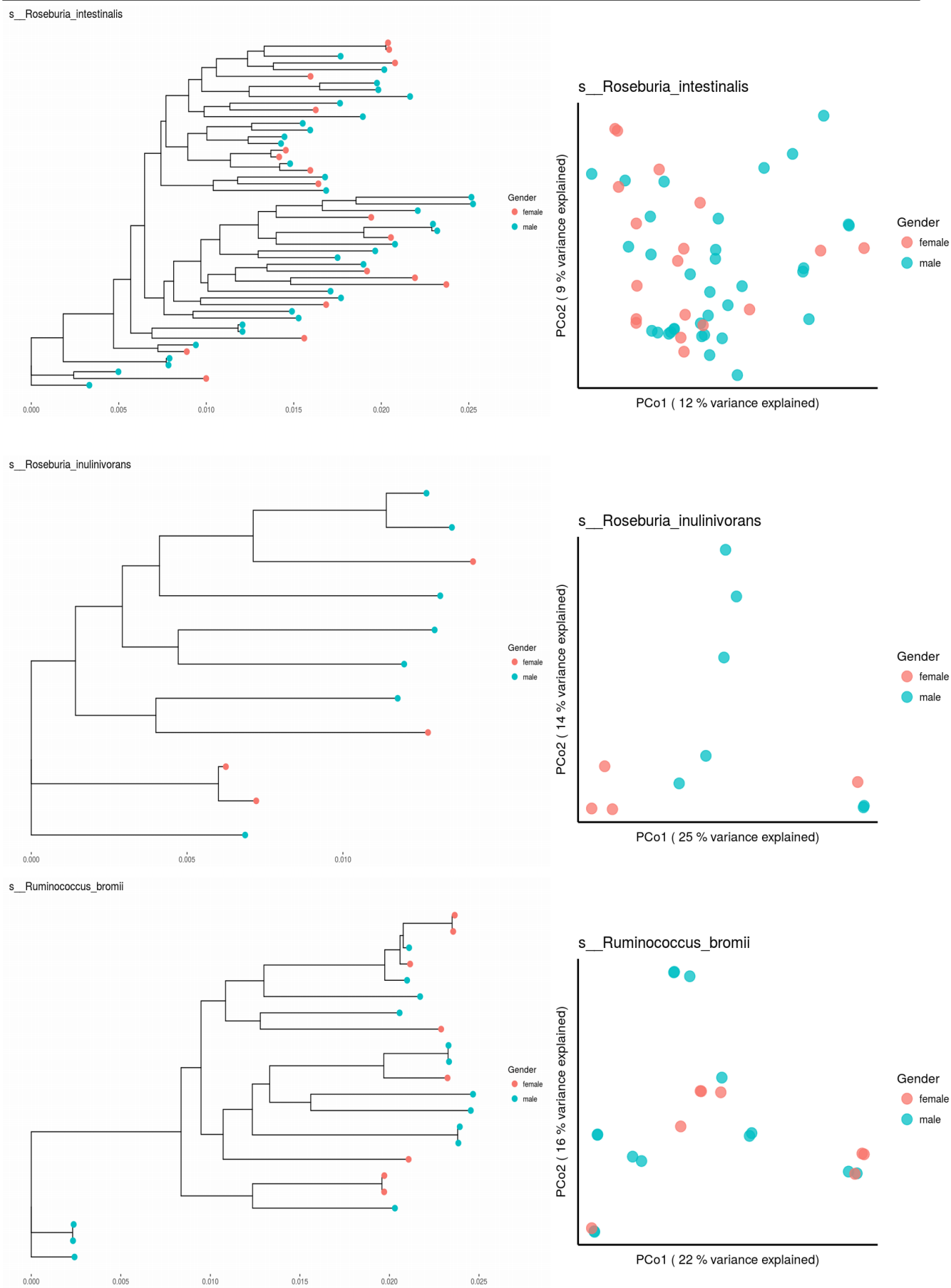
s\_\_Parabacteroides\_merdae



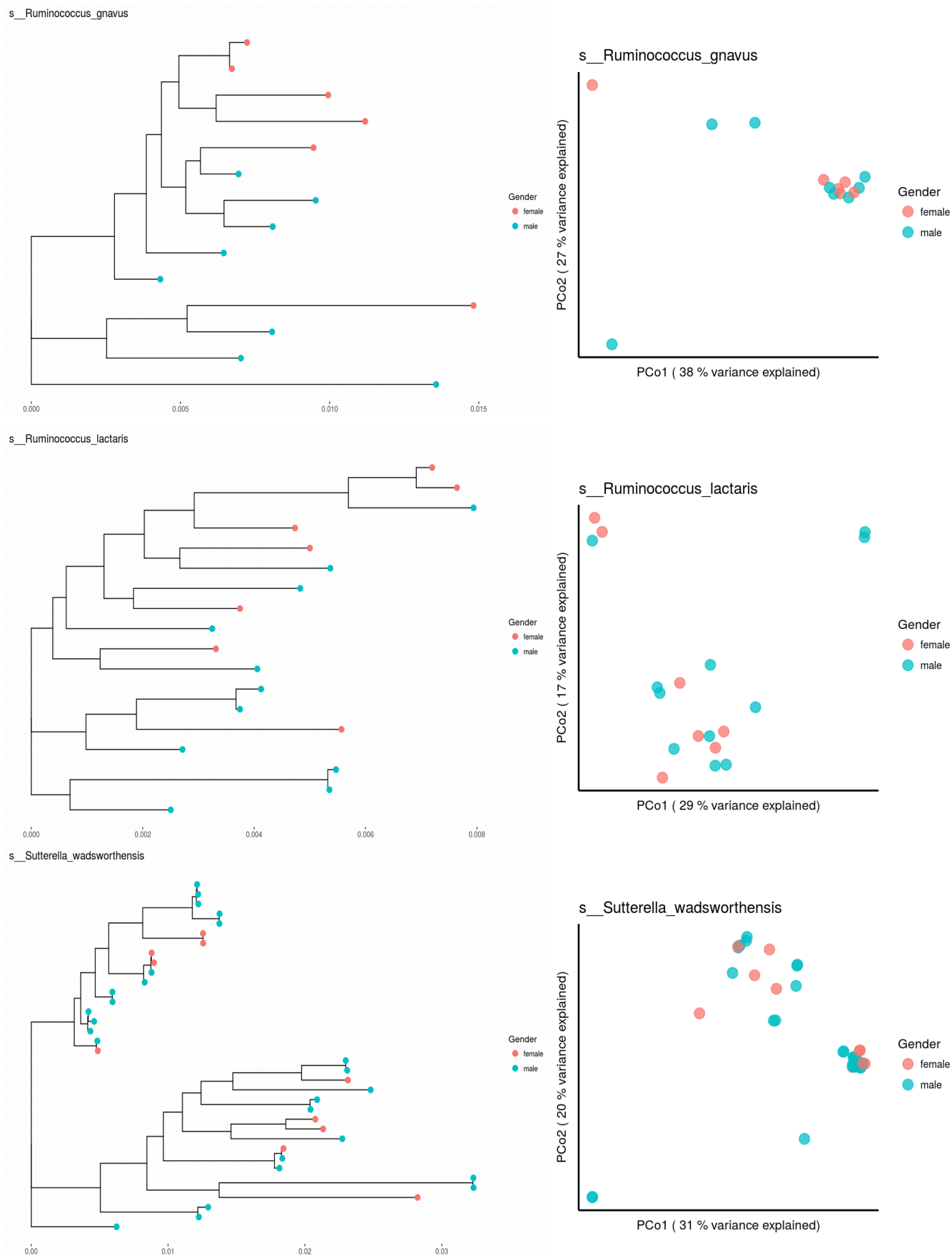
s\_\_Roseburia\_hominis



**Supplemental Figure S14 |** Population structure of *Roseburia intestinalis*, *Roseburia inulinivorans* and *Ruminococcus bromii*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.



**Supplemental Figure S15 |** Population structure of *Ruminococcus gnavus*, *Ruminococcus lactaris* and *Sutterella wadsworthensis*. Maximum-Likelihood phylogenetic trees of the reported species and ordination plots of the phylogenetic distance matrix for each species.



**Supplemental Figure S15** | The space of the ABC accepted simulations for *B. ovatus*' bottleneck model. The contour lines on the scatterplot indicate the density estimation. The ABC tolerance value was set to 0.5.

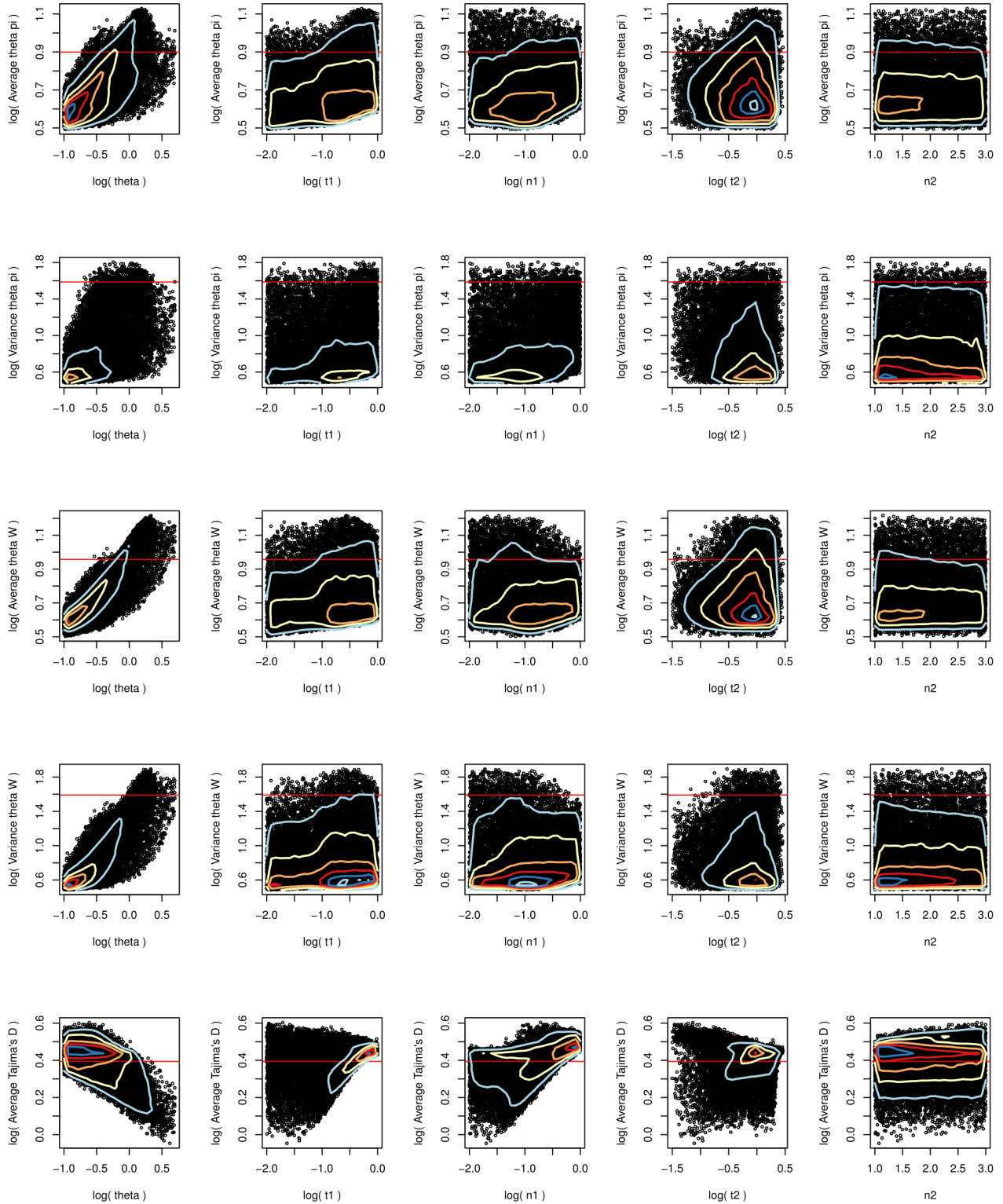


Figure continues on next page.



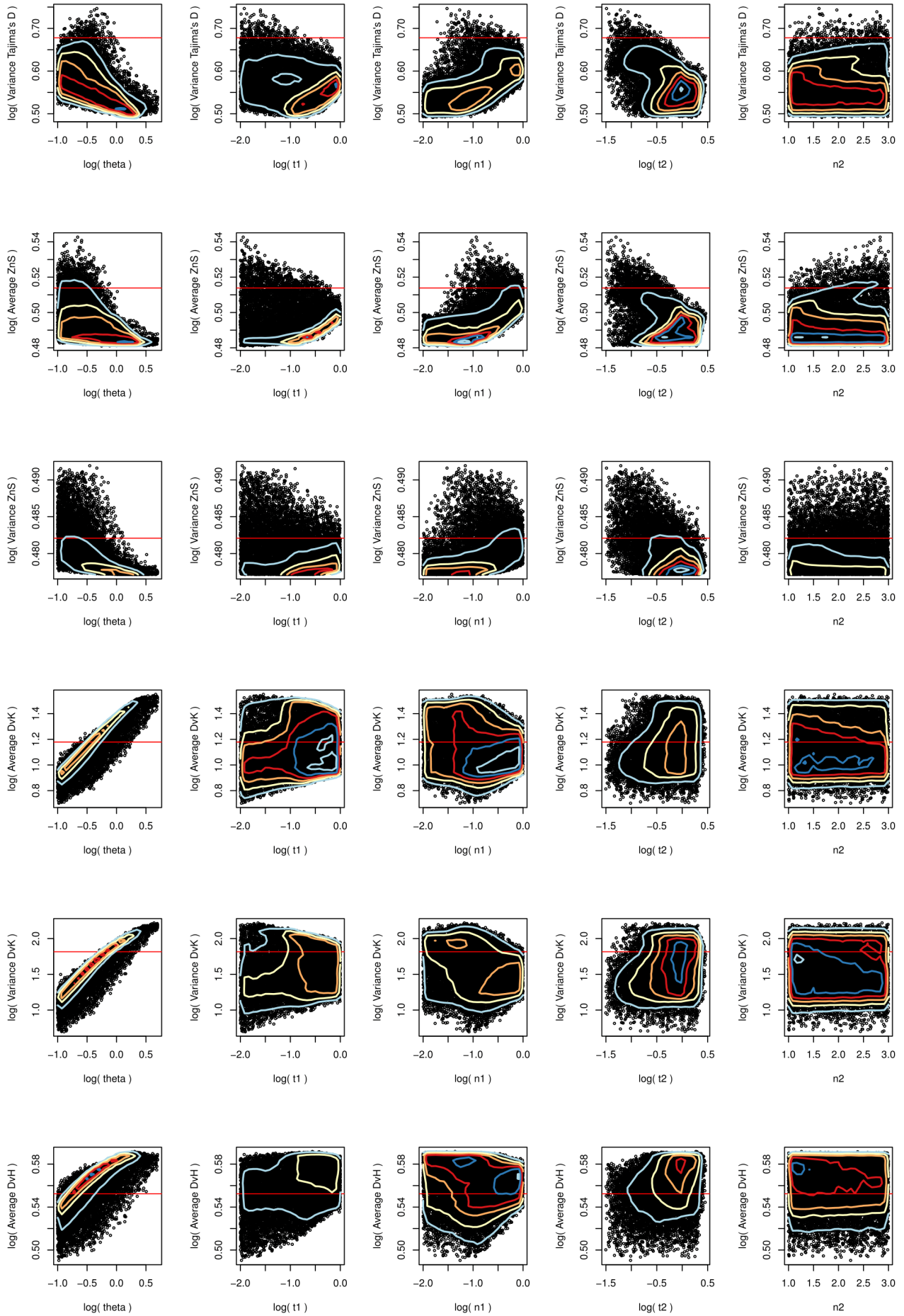


Figure continues on next page.

