UNIVERSITÀ DI TRENTO

Dipartimento di
Biologia Cellulare, Computazionale e Integrata - CIBIO

International PhD Program in Biomolecular Sciences

Department of Cellular, Computational
and Integrative Biology - CIBIO

XXXIII Cycle

Integrative computational microbial genomics for large-scale metagenomic analyses

Tutor

Prof. Nicola Segata
Department of Cellular, Computational and Integrative Biology - CIBIO
University of Trento, Italy

Ph.D. Thesis of

Francesco Beghini

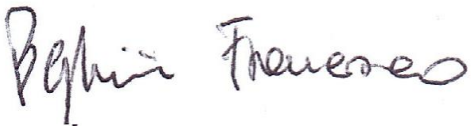
Department of Cellular, Computational and Integrative Biology – CIBIO
University of Trento, Italy

Academic Year 2019-2020

Declaration

I, **Francesco Beghini**, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Trento, 01st October 2020



Index

Abstract

Advancements of DNA sequencing technologies and improvement of analytic methods changed the way we analyze complex microbial communities (metagenomics). In only a few years, these methods have evolved so far as to ease a more precise community profiling and to allow high-level strain resolution. A typical computational metagenomic analysis relies on mapping raw DNA sequencing reads against sets of “reference” microbial genomes usually obtained through single-isolate sequencing. With an almost exponential increase in the number of reference genomes deposited daily in public data sets, current computational methods are incapable of managing and exploiting such a rich reference set, limiting the potential of metagenomic investigations.

In my doctoral thesis, I will present my contribution towards fully exploiting the available reference data for metagenomic analysis. I developed ChocoPhlAn, an integrated pipeline for automatic retrieval, organization, and annotation of reference genomes and gene families as the foundation for bioBakery 3, an improved set of computational methods for the analysis of shotgun metagenomics data. Using the latest set of microbial genomic reference data available and processed through ChocoPhlAn, the six bioBakery 3 tools that I updated resulted in more comprehensive and higher resolution taxonomic and functional profiling of microbiomes and allowed strain-level characterization of their constituent strains. After extensive benchmarks with previous versions and competitors, we applied those methods on more than 10,000 real metagenomes and showed how metagenomics can be a more powerful tool for identifying novel links between the gut microbiome and disease conditions such as colorectal cancer and Inflammatory Bowel Disease. Accurate strain-level phylogeny reconstruction and pangenomic analysis of 7,783 metagenomes revealed novel functional, phylogenetic, and geographic diversity of *Ruminococcus bromii,* a common and high-prevalent gut inhabitant.

We then focused on the influence of the Eukaryotic fraction of the human microbiome and its potential impact on human gut health, which is a frequently overlooked aspect of microbial communities. To this end, we assessed the presence of the Eukaryotic parasite *Blastocystis spp.*, in more than 2,000 metagenomes from 5 continents for understanding associations with disease statuses and geographic conditions. We showed that *Blastocystis* is the most common Eukaryotic colonizer of the human gut, and it is particularly prevalent in healthy subjects and non-westernized populations. We further explored intra-subtype diversity by reconstructing and functionally profiling new metagenomic-assembled *Blastocystis* genomes, showing how metagenomics can be valuable to unravel protists' genomics and providing a genomic resource for additional integration of non-bacterial taxa in metagenomic pipelines.

By developing and implementing ChocoPhlAn and the new bioBakery tools, we provided the community with improved and efficient microbiome profiling tools and started identifying novel patterns of association between host and niche-associated microbiomes and discovering previously uncharacterized species from human and non-human hosts.

Introduction to the thesis

The human microbiome

The microbiome is commonly defined as the whole ensemble of microorganisms that live and occupy a specific ecological niche. These are not only limited to bacteria, but also include archaea, viruses, and micro-eukaryotes such as protozoa, algae, and fungi (Berg *et al*, 2020). Since the last half of the XVII century when Antoni van Leeuwenhoek was able to first observe bacteria, or *animalculum* as referred by him, the knowledge about the microbial world has increased enormously. Microbes have been studied in several ecological environments, ranging from freshwater (Lee *et al*, 2016), oceans (Venter *et al*, 2004; Sunagawa *et al*, 2015), rivers (Van Rossum *et al*, 2015), soil (Jansson & Hofmockel, 2018; Brown *et al*, 2015), permafrost (Mackelprang *et al*, 2011) and also in outer space (Voorhies *et al*, 2019). Several studies were carried out to understand the capabilities of microbes to adapt and live in urban environments such as subways (Afshinnekoo *et al*, 2015; MetaSUB International Consortium, 2016) or even public restrooms (Flores *et al*, 2011). Microbiomes are associated with living hosts, such as human beings (Human Microbiome Project Consortium, 2012), mice (Lesker *et al*, 2020), cattle (Wallace *et al*, 2015; Stewart *et al*, 2019), and even mosquitos (Segata *et al*, 2016).

Several efforts have been made in the last fifteen years to specifically understand the importance of the human microbiome in healthy and diseased conditions, and to unravel its impact on host physiology (Sommer & Bäckhed, 2013; Levy *et al*, 2017) or the influence of drugs e.g. antibiotics (Langdon *et al*, 2016), probiotics (Petschow *et al*, 2013), and different lifestyles (Keohane *et al*, 2020; Barone *et al*, 2019; Devkota, 2020). Large scale investigations were conducted in Europe by the MetaHIT Consortium (Qin *et al*, 2010) and in the US by the Human Microbiome Project (Human Microbiome Project Consortium, 2012). Both projects, considered milestones in the microbiome field, generated and released an enormous amount of data, starting from a curated catalog of genes, reference genomes, and metagenomes that first defined the composition of microbiomes associated with the skin (Oh *et al*, 2014; Grice & Segre, 2011), the oral cavity (Donati *et al*, 2016; Segata *et al*, 2012a; Human Microbiome Project Consortium, 2012), the nasal cavities (Human Microbiome Project Consortium, 2012), the gastrointestinal tract (Human Microbiome Project Consortium, 2012; Qin *et al*, 2010), and the urogenital tract (Aagaard *et al*, 2012). Reference genomes released with the Human Microbiome Project were a key factor and enabled new lines of research as well as greatly improving tools for microbiome analysis.

Although it is not simple to exactly define what a healthy microbiome is and which is its composition, the Human Microbiome Project (Human Microbiome Project Consortium, 2012) provided some insight into the differences between distinct body sites. The richness of bacterial species is comparatively different when looking, for example, at the gut microbiome or the vaginal one: while the latter displays a low diversity and is mainly dominated by *Lactobacillus* species, the gut displays an intermediate diversity and is mainly composed of the phyla Bacteroidetes and Firmicutes (Human Microbiome Project Consortium, 2012). The Firmicutes-to-Bacteroidetes ratio has been proposed as a potential biomarker for obesity (Ley *et al*, 2006) after observing that the gut microbiome of obese subjects displayed an increased abundance of Firmicutes and reduced abundance of Bacteroidetes. A similar pattern of increased abundance of Bacteroidetes has been observed in people living traditional lifestyles that were not subjected to westernization (Obregon-Tito *et al*, 2015; Rampelli *et al*, 2015; Brito *et al*, 2016) in which *Prevetella* and *Treponema* were reported as the most abundant genera, genera that are generally less abundant in westernized gut microbiomes. This is because the gut microbiome composition can be shaped by external factors like diet or antibiotic intake, able to promote or deplete certain phyla. Comparison of non-westernized microbiomes to those of westernized populations highlights a striking difference in terms of species richness: this is potentially due to a possible loss of specific microbiome members caused by a multiplicity of reasons that can range from the adoption of medical and public health practices to the shift to a different type of diet, mostly high-fat and low-fiber based (Segata, 2015). The increased richness in non-westernized gut microbiomes can be partially explained by the "microbial dark matter", which is the unexplored fraction of the microbiome that can account for up to 60% of a metagenome and encompasses novel and not yet described species (Pasolli *et al*, 2019). The reduction of the amount of known and identifiable species is due to the fact that most microbiome studies so far sampled westernized regions such as Europe or the US, leaving other regions of the world largely undersampled.

Although advances in sequencing technologies allowed for a deeper investigation of the microbiome composition, most of the conducted studies are focused on bacteria, while a relevant portion of the microbiome composed of eukaryotes has been neglected (Parfrey *et al*, 2014; Laforest-Lapointe & Arrieta, 2018). In the context of the human gut microbiome, microbial eukaryotes, microeukaryotes, has been traditionally associated with a negative outcome, such as the onset of disease after colonization by parasitic organisms, but recent works untangled this issue showing that may have a beneficial role as probiotics (McFarland & Bernasconi, 1993) or even be common inhabitants of a healthy gut (Nash *et al*, 2017; Beghini *et al*, 2017). This opens up to a lot of questions, such as their role in the gut ecology or the interplay between host, microeukaryotes, and other members of the microbiome.

With its advent, Next Generation Sequencing was found to be a good fit for analyzing microbiomes. The development of high throughput platforms paved the way for studying microbiomes through shotgun metagenomics or high throughput amplicon sequencing of the 16S or 18S ribosomal RNA genes (16S and 18S rRNA sequencing). These technological advances allowed for a long trail of successful studies which tried to characterize and associate the microbiome composition with several diseases, such as colorectal cancer (Thomas *et al*, 2019; Yu *et al*, 2017; Zeller *et al*, 2014; Wirbel *et al*, 2019), type 1 (Heintz-Buschart *et al*, 2016) and type 2 (Karlsson *et al*, 2013; Qin *et al*, 2012) diabetes, obesity (Ley *et al*, 2006; Le Chatelier *et al*, 2013), inflammatory bowel disease (Nielsen *et al*, 2014; Lloyd-Price *et al*, 2019; Morgan *et al*, 2012), arthritis (Chen *et al*, 2018; Zhang *et al*, 2015), liver diseases (Qin *et al*, 2014; Caussy & Loomba, 2018) , or atherosclerosis (Koren *et al*, 2011), to mention a few. Identifying disease-specific bacterial signatures could allow in the future to perform non-invasive diagnosis (Thomas *et al*, 2019; Yu *et al*, 2017; Chen *et al*, 2018; Cameron *et al*, 2017; Ghensi *et al*, 2020).

Although enormous efforts have been done to understand the dynamics of microbial communities and their compositions in different environments and settings, there is still a lot to explore. Recent efforts in identifying novel species without the availability of reference genomes allowed the recovery of more than 5,000 species from metagenomes, leading to an substantial increase of reference data (Pasolli *et al*, 2019; Nayfach *et al*, 2019; Almeida *et al*, 2019, 2020). This enormous amount of data is posing new challenges: in particular, there is the need to leverage such resources and try to unravel the many facets of the interplay of the human microbiome and the host, further exploring missing links with the host's lifestyle, health, and geographic association.

Shotgun metagenomics for the human microbiome

Shotgun metagenomics is the untargeted sequencing of all the genetic material present in a sample associated with a specific environment (Quince *et al*, 2017b). Since it is untargeted, it also allows for studying not-so-well characterized and cultivation-recalcitrant microorganisms. Ideally, by using a high-throughput sequencing approach, we are able to sequence all the genetic material, even from relatively low-abundant microorganisms. Different from amplicon sequencing, which targets particularly genes, shotgun metagenomics can capture the entire microbial genetic repertoire and allows for taxonomic and functional potential analysis (Scholz *et al*, 2016; Quince *et al*, 2017b)

The first steps in a metagenomic workflow are sample collection, DNA extraction, and sequencing (Quince *et al*, 2017b; Knight *et al*, 2018). Standard best practices were established and have been used for years in order to rely on the same protocol and make results of

different studies comparable (Costea *et al*, 2017b; Vandeputte *et al*, 2017; Thompson *et al*, 2017). Sequencing of the genetic material through a high-throughput sequencer generates a high volume of data that needs to be analyzed using a computational approach. A high number of computational methods were developed in recent years to analyze such kind of data, but still, challenges are present when it comes to handling the increasing amount of generated data and increasing profiling resolution.

From a practical perspective, we can divide sequence analysis methods into two categories: assembly-based and read-based (or assembly-free) (Quince *et al*, 2017b). There is no “best method” since most of the time analyses use combinations of the two approaches. Both types of analyses want to answer three simple questions, “who is there?”, hence which taxa are present in the microbiome, and “what are they capable of doing?”, hence which functions are encoded in their genomes, and “who is doing what”, hence which functions are carried out by a taxa.

The first question, “who is there?”, addresses the problem of taxonomic classification, namely the determination of the species present in the metagenome. Several computational methods have been described and developed using different approaches in the literature in order to answer this question. The second question, “what are they capable of doing?”, answers to the functional potential of microbial communities by looking at the whole gene content. In this case, we call it “functional potential analysis” since we are looking at the genes encoded by genomes instead of looking at transcriptomes, which, by adding information about gene expression, enables for “functional activity analysis” (Franzosa *et al*, 2018). The last question, “who is doing what?” combines taxonomic classification, functional profiling, and genome reconstruction in order to obtain evidence of the functional activity of a genome, enabling comparative analyses.

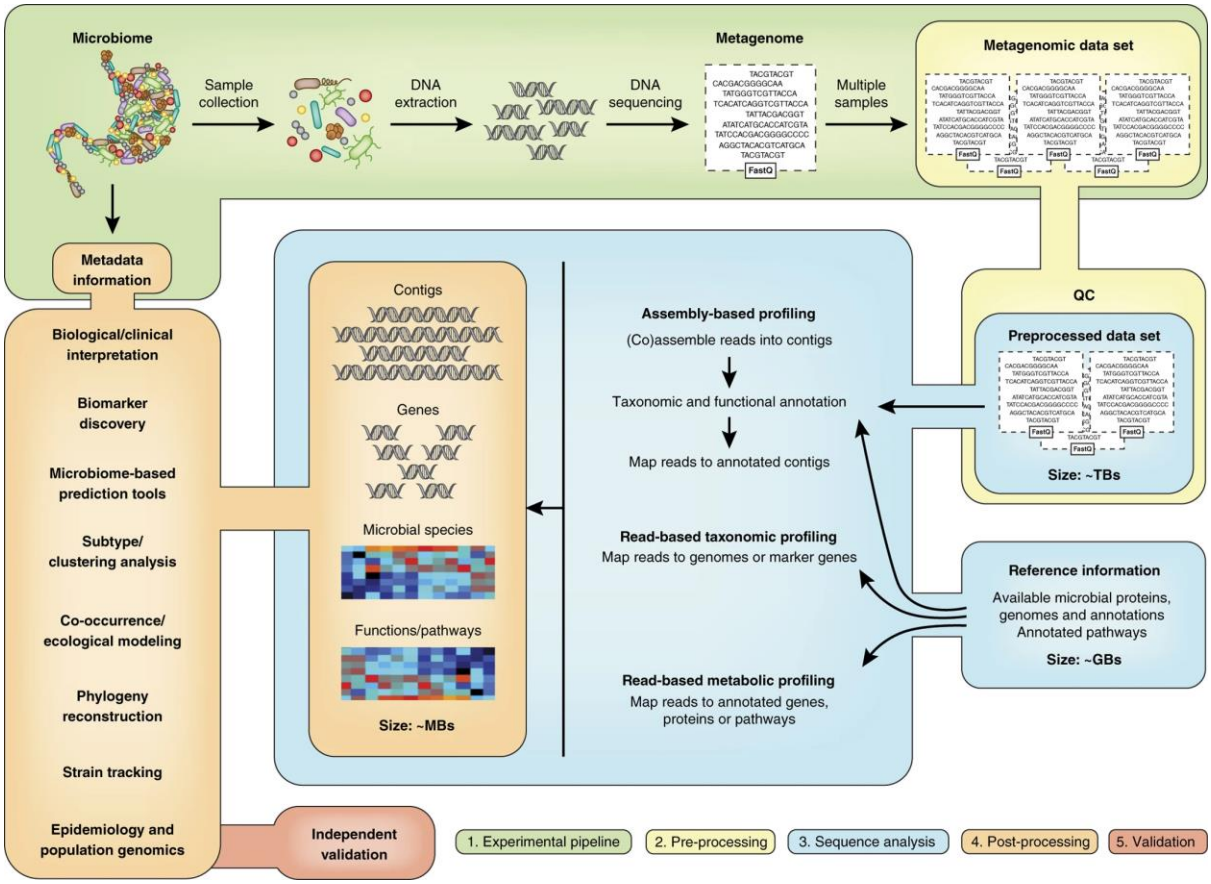


Figure 1: An example of a metagenomic workflow, courtesy of (Quince et al. 2017).

Challenges and opportunities for human metagenomics

Computational metagenomics methods rely largely on the availability of “reference” microbial genomes generally obtained through single-isolate sequencing. However, the number of these references deposited in public datasets is increasing exponentially due to the increased availability of sequencers and lower costs related to sample processing and sequencing. Sequencing technologies are improving at a faster pace than computational infrastructures, a faster pace that, so far, a DNA megabase can be sequenced with no more than 1 dollar cent (Wetterstrand, 2020).

However, the number of assembled isolates available in public databases like NCBI is restricted to a few species. To date, the ten most sequenced species are pathogens (NCBI, 2020) and account for nearly 25% of the total sequenced genomes. This is a good example of the usefulness of big data in microbiology in the context of bacterial outbreaks: high-throughput sequencing and big data allow researchers to conduct comparative analyses to understand the differences between the bacterial strains responsible for the outbreak (Relman, 2013). Large-scale genomic epidemiologic studies of recent viral and bacterial outbreaks (Lu *et al*, 2020; Gire *et al*, 2014; Weill *et al*, 2017; Guthrie & Gardy, 2017) have

made been possible thanks to metagenomics and high-throughput sequencing, allowing to scale up, a not-so-feasible approach with single isolate sequencing. Clinical metagenomics still remains a niche technique limited to very few research laboratories worldwide but it is an outstanding technique for the identification of unusual pathogens (Wilson *et al*, 2018).

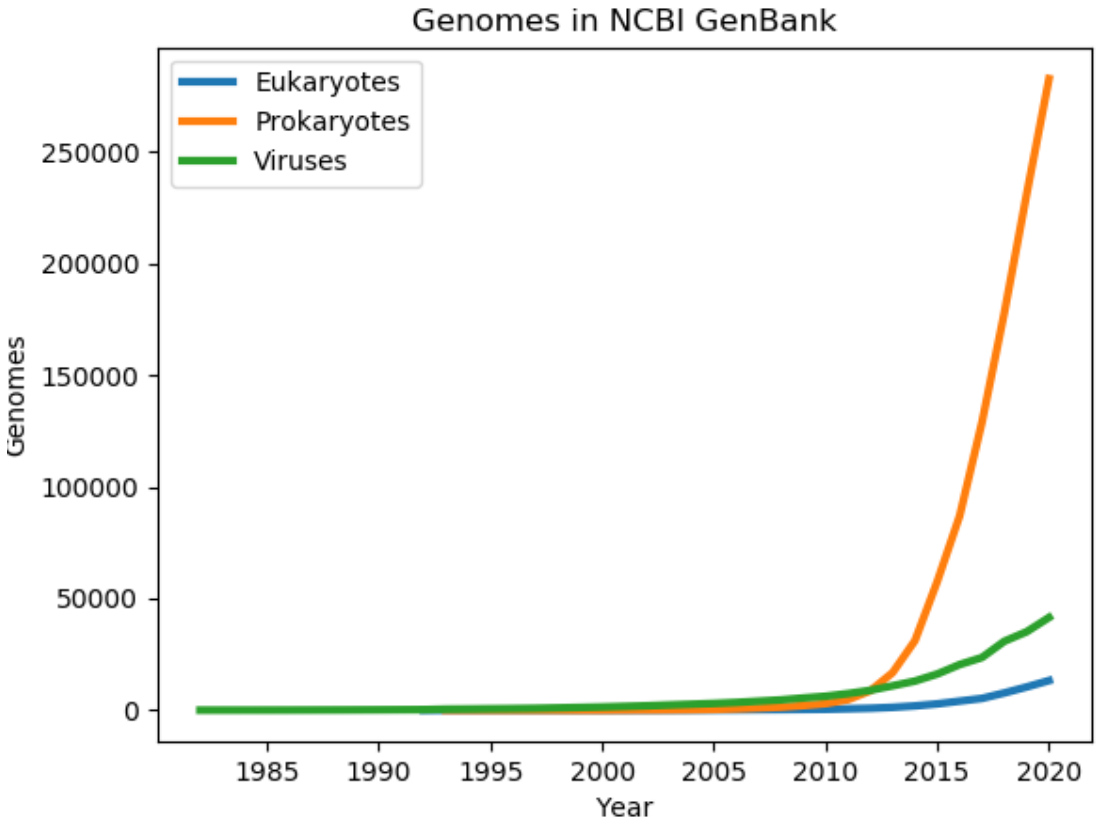


Figure 2: Exponential growth of genomes deposited into NCBI GenBank. Adapted from (Zynda, 2020)

Pathogenic species isolated and sequenced in the context of surveillance projects are certainly important for disease-specific surveillance, but in the context of the human microbiome, many prevalent and abundant species are poorly characterized. In this context, in **Section 1**, I will describe the characterization and the genetics of *Ruminococcus bromii*, a highly prevalent bacterium (50%) in human intestinal communities, but for which only 15 isolates are available. This opens the door to exploring species that currently are described by a paucity of reference genomes and discovering new species that are not characterized. By leveraging metagenomic assembly and binning, recent works have been able to reconstruct thousands of genomes of species (Karcher *et al*, 2020; Tett *et al*, 2019; Almeida *et al*, 2020; Manara *et al*, 2019; Delmont *et al*, 2020). This further expands our knowledge of the composition of microbiomes, host-associated and not. In order to successfully manage this large amount of data and access the intrinsic information, computational methods need to effectively handle big data.

Although having more and more reference genomes available is compelling, methods and pipelines that were developed in the last 10 years are not capable of dealing with such a large amount of data or were designed in a way that expansion with new reference data is really hard to do or utilize a static database. But there are some drawbacks to having more reference genomes available, for example, (Nasko *et al*, 2018) describe a critical issue afflicting *k*-mer based taxonomic classifiers, resulting in a major decrease of precision due to the high number of species' reference genomes. Here is the importance of having consistent methods able to keep up with the growth of available data. Advances in computational methods with faster algorithms capable of performing raw data processing in a more efficient paradigm and new methods able to scale-up regardless the size of the input data are essential for the development, implementation, and maintenance of computational pipelines.

In addition to data coming from single isolate sequencing and metagenomes, sample-related metadata is integral to any microbiome big data analysis. Particularly with larger datasets available, metadata provides answers to complex questions through comparative analyses. Following efforts to ensure the integration of useful metadata for reference genomes (Field *et al*, 2008) and metagenome-assembled genomes (Bowers *et al*, 2017), new additional efforts are required to ensure the integration of minimal sample-related metadata integrated with metagenomic data (Pasolli *et al*, 2017). Comparative metagenomic analyses and meta-analyses require comprehensive annotations of the origin of the sample and the state of the environment at the time of sampling. Unfortunately, many available datasets have incomplete or unavailable metadata or, worse, are difficult to access (e.g., "availability upon request"). Second, a common standard must be adopted to accurately and consistently report metadata (Huttenhower *et al*, 2014), avoiding the end-user picking on different repositories associated with the study (e.g. supplementary materials). Despite the efforts made in the past and the continuous progress that is still being made in this area of study, several open challenges are still posed, challenges that have been the driving force of the work discussed in this thesis.

Aims

The work presented in this thesis is intended to address the challenges described earlier. In particular, the main goal of this work is the development of innovative computational metagenomic strategies to improve the profile of microbiomes at different levels. This primary goal will be discussed in **Section 1.2**. In more detail, this manuscript aimed to:

- Develop an integrated system to organize all the microbial genomic information.
- Use the newly organized genomic information to produce more accurate taxonomic and functional profiles.
- Apply the improved methods on real data for novel findings.

Structure of the thesis

The work presented in this thesis will cover all the aspects outlined in this introductory chapter, and the thesis is divided into two sections. **Section 1** will address the problem of maintaining methods to overcome the growing number of available reference sequences and aims to answer the previously discussed goals by developing an integrated ecosystem for metagenomic analyses. I will present the main project carried out during the three years of the Doctoral Programme that resulted in the development of the ChocoPhlAn pipeline for the generation of a uniform genomic repository for the bioBakery 3 tools, tools that underwent a consistent process of update and refinishing.

Section 2 will present additional side projects I have been involved in, projects aimed to better understand the directions to be taken for the work presented in **Section 1.2,** and to apply different methodologies to metagenomic data. These projects have resulted in various publications, two of which with major contributions that will be described in **Section 2.1** and **Section 2.2**. My contributions to other published works in which I am included as co-author will be described in **Section 2.3**.

Section 2.1 will present a work focused on the investigation of one member of the eukaryome, the eukaryotic portion of the human gut microbiome, the micro-eukaryote *Blastocystis* spp. Using a metagenomic approach, we investigated its presence in more than 2,000 metagenomes from 10 different countries and we provide evidence to consider *Blastocystis* as a common member of the human gut microbiome and associations with health conditions. In this work, I contributed by performing all the analyses using the available reference genomes and metagenomic datasets available to date. I also developed a custom pipeline for the coverage calculation, the assessment of the limit of detection of *Blastocystis* in the

metagenomes, and the reconstruction and annotation of the metagenome-assembled genomes.

In the context of bodysite-associated microbiomes, **Section 2.2** focuses on the impact of tobacco and nicotine products on the oral microbiome. Unlike the other works discussed here, this study was conducted using 16S rRNA amplicon sequencing instead of shotgun metagenomics. I was personally involved in all the computational, statistical, and epidemiological analyses while sample and data collection were performed by the New York City Health Department, and the laboratory work for sample extraction, processing and sequencing was performed by the Burk laboratory at the Albert Einstein College of Medicine.

Applications of some of the tools described in **Section 1.2** are presented in **Section 2.3**, together with other works to which I have contributed. I also report here below other contributions that I made to other software developed by our laboratory and were instrumental to the wide adoption of the corresponding methods:

- I maintained the previous version of MetaPhlAn, MetaPhlAn2, and led the development of MetaPhlAn 3, as outlined in **Section 1** [1].
- I developed the ChocoPhlAn pipeline, presented in **Section 1**[2], and started the development of ChocoPhlAnSGB.
- I maintained, updated, and ported to Python 3 LEfSe[3].
- I maintained and ported to Python 3 hclust2[4], a visualization tool for generating heatmaps from taxonomic and functional profiles.

In the context of reproducibility and software packaging, for MetaPhlAn, PanPhlAn, hclust2, and CMSeq I have created Anaconda recipes, set up a pip-based installer, and made available the software via the PyPi and Bioconda platforms

Figures included in **Sections 1.2** and **2.2** were extracted from the open-access preprint made available before submission to the journal. The materials included in the open-access preprint are licensed under the CC BY 4.0 license which allows the full sharing and adaptation of the work after attribution to the original material. In addition, Annals of Epidemiology, the journal in which was published the manuscript included in **Section 2.2**, permits the reuse of the work by the authors. The paper included in **Section 2.1** is licensed under the CC BY-NC-SA 4.0 license which allows the non-commercial sharing and adaptation of the work after attribution to the original material.

[1] MetaPhlAn GitHub repository: https://github.com/bioBakery/metaphlan
[2] ChocoPhlAn GitHub repository: https://github.com/SegataLab/chocophlan
[3] LEfSe GitHub repository: https://github.com/SegataLab/lefse
[4] hclust2 GitHub repository: https://github.com/SegataLab/hclust2

Section 1

1.1. Background

Unlike amplicon-based sequencing, shotgun metagenomic sequencing allows high-resolution and culture-independent exploration of several aspects of microbial communities, including microbial composition and functional analysis. Computational analyses are essential for understanding the characteristics and dynamics of the microbial community and several analytical strategies have been developed during the years. This section provides an overview of the most common computational metagenomic tasks.

Metagenomic assembly

Reconstruction of microbial genomes from metagenomes is possible via a technique called *de-novo* metagenomic assembly. Starting from metagenomic reads, metagenomic assembly pipelines generates longer stretches of contiguous DNA sequences known as contigs. A common method employed to reconstructing contigs is based on the *de Brujin* graphs (Pevzner *et al*, 2001). In a *de Brujin* graph, all the overlapping sub-sequences of length *k* (*k*-mers) extracted from the reads are the node set, and two nodes are connected by an edge if the two sequences are contiguous (the suffix of one *k*-mer is the prefix of another *k*-mer). Different metagenomic assemblers have been developed trying to take these problems into account and overcome them using different approaches.

To date, two popular assemblers are metaSPAdes (Nurk *et al*, 2017) and MEGAHIT (Li *et al*, 2015), both implementing assembly through *de Brujin* graphs. metaSPAdes is an evolution of the SPAdes assembler (Bankevich *et al*, 2012), which was originally developed for single-cell assembly, and employs a multi *k*-mer iterative search, an approach previously implemented by metaIDBA (Peng *et al*, 2011), in which at every iteration, after increasing the value of *k,* the *de Brujin* graph is rebuilt extracting *k*-mers from the contigs and the reads, obtaining a more refined assembly.

Metagenomic assembly presents a variety of challenges. Since metagenomes samples contains multiple species present at different abundances, in order to be able to perform metagenomic assembly with high accuracy, the sequencing depth should be uniform along the genome. Shallow sequencing or uneven coverage could result in an incomplete or incorrect assembled genome, especially for those species present at low abundance. DNA repeats also impacts the assembly process since they could create ambiguous sequences: this problem can be partially tackled by using longer *k*-mers, but longer *k*-mers are more prone to include sequences with sequencing errors (Olson *et al*, 2019). The presence of multiple closely related organisms or strains requires to perform sequencing as deeply as possible to allow identification of genomic variants.

Different metagenomic assemblers can produce a whole different set of contigs and some of them are more suitable to a specific task, for example the Critical Assessment of Metagenome Interpretation (CAMI) Challenge (Sczyrba *et al*, 2017) showed that MEGAHIT is able to generate larger assemblies while metaSPAdes generates more accurate contigs (van der Walt *et al*, 2017).

Evaluation of the generated assemblies can be performed using MetaQUAST (Mikheenko *et al*, 2016) that reports on genome quality statistics (N50, genome length, number of contigs, etc.) and allows for comparisons with closely related reference genomes to detect misassembles or structural variants.

Taxonomic analyses

Taxonomic analysis approaches can be summarized under two main categories:

1. Taxonomic classifiers
2. Taxonomic profilers

Among the various analysis tools available, it is necessary to distinguish between taxonomic classifiers and taxonomic profilers. Taxonomic classifiers assign a taxonomic label to all the reads contigs within a metagenome, whereas taxonomic profilers do not assign taxonomies to each read, but rather report presence and relative abundances of the detected taxa. The two terms are used interchangeably in error, although estimates of relative abundance from sequence profilers may be done after correction of the total read count for the genome size of each clade (Lu *et al*, 2017).

Alignment-based taxonomic classifiers internally align sequences with tools like Bowtie2 (Langmead & Salzberg, 2012), BLAST (Altschul *et al*, 1990), bwa (Li & Durbin, 2009), or HMMER (Eddy, 2011) to identify matches against a reference database. An incredible number of metagenomic classifiers have been developed over the years to overcome this challenge. Early popular tools like MEGAN (Huson *et al*, 2007) or MG-RAST (Meyer *et al*, 2008) used under the hood the BLAST suite to profile metagenomes using either the NCBI-NR/NT databases or a custom database of reference genomes. Both methods were inflating the number of detected species by improperly detecting low abundant species, increasing the false positive rate in the lower tail of the relative abundance distribution (Peabody *et al*, 2015). Alongside the aforementioned methods, which implemented an approach based on sequence homology, PhyloPhytia (McHardy *et al*, 2007) and its successive versions PhyloPhytiaS (Patil *et al*, 2011), and PhyloPhytiaS+ (Gregor *et al*, 2016), are *k*-mer based approaches performing supervised classification by leveraging a support vector machine trained on reference

sequences. Since supervised methods rely on genomic references, they are more precise than unsupervised methods, which do not require any training (Dröge & McHardy, 2012). Another early leader, also the first implementing a *k*-mer based approach for classification, is TETRA (Teeling *et al*, 2004), which used an intrinsic characteristics of the sequences, DNA tetranucleotide frequencies, to classify sequences. PhymmBL (Brady & Salzberg, 2009), attempted to address the problem of high false positive rate by refining the BLAST mapping results using interpolated Markov models and this can be considered the first method displaying higher accuracy, compared to the other software available at the time working with short reads (about 100bp). Taxonomic classifiers performed using large databases, such as non-redundant databases (NCBI-NT) or whole genomes, can be memory-consuming since a taxonomic label has to be assigned, hypothetically, to all the reads present in a metagenome.

Classifiers implementing a *k*-mer matching approach have the potential to use a large collection of genomes, such as the complete RefSeq catalog, and to profile all the reads in a metagenomic sample in a relatively small amount of time. The method is based on the calculation of the frequencies of *k*-length oligonucleotides extracted from all the DNA sequences present in the genomic database. State-of-the-art methods like Kraken (Wood & Salzberg, 2014), Kraken2 (Wood & Salzberg, 2014; Wood *et al*, 2019), KrakenUniq (Breitwieser *et al*, 2018), CLARK (Ounit *et al*, 2015), or Centrifuge (Kim *et al*, 2016) have different approaches to classify *k-mers*: Kraken and Kraken2 assign *k*-mers in the lowest common ancestor between two taxa. CLARK, instead, uses a discriminative approach to identify unique *k*-mers able to characterize each species. To lessen the amount of space required by the database and the redundancy, Centrifuge uses the Burrows–Wheeler Transform (BWT) and an FM-index to store and index the genome database, a similar approach also used in Kraken2, which implements minimizers. These methods are capable of detecting species represented by a small number of reads but are subject to a very high rate of false positives, in fact, a tradeoff between sensitivity and specificity must be determined. Whereas the use of longer *k*-mers can lead to a potential failure to match specific species, the use of shorter *k*-mers will yield non-specific matches to many sequences. This issue was addressed by KrakenUniq by the introduction of the HyperLogLog algorithm to count unique *k*-mer in order to reduce the false-positive rate. Another downside of these methods is the size of the database: by increasing the *k*-mer size, it automatically increases the computational cost since more matching needs to be done.

Hash-based *k*-mer based taxonomic profilers exploit containment estimation using hash functions implemented in techniques like MinHash (Broder, 1997), an algorithm based on the Jaccard similarity and originally designed for web search engines, or Bloom Filters (Bloom,

1970). Hash-based methods like Metalign (LaPierre *et al*, 2020) and ganon (Piro *et al*, 2020) perform hashing on the identified *k*-mers and containment to perform taxonomic assignation.

Taxonomic profilers are compositionality-driven and reports presence and relative abundances of detectable taxa. In particular, marker-based approaches make it possible to identify taxa by detecting the presence of specific marker sequences. These markers can be extracted from universally conserved genes, an approach used by mOTUs (Sunagawa *et al*, 2013; Milanese *et al*, 2019), which profiles metagenomes using a set of 40 prokaryotic conserved marker genes, or clade-specific marker genes as used by MetaPhlAn (Segata *et al*, 2012b; Truong *et al*, 2015). A marker-based approach minimizes the number of mapped sequences, thus reducing the computational burden. This is one of the advantages of this type of approach, as well as the very low rate of false positives (Sczyrba *et al*, 2017; Breitwieser *et al*, 2019a), and the estimation of less-biased relative abundance profiles. One of the disadvantages of these methods lies in the number of species identifiable: in order for a species to be identified, the marker genes of the species of interest should be identified with a certain degree of confidence, and failure of this procedure will lead to the missing identification of the species. Regardless of the approach used, marker-based taxonomic profilers accurately determine the relative abundance of all the species that can be profiled, a challenge when it comes to identifying low-abundance species. *k*-mers are also employed in taxonomic profilers like FOCUS (Silva *et al*, 2014) which uses non-negative least squares to report taxa abundances using 7-mers extracted from a set of reference genomes.

The proliferation of new methods requires continuous benchmarking with the already established and state-of-the-art methods. Over the past few years, several works have dealt with the issue of benchmarking in the context of metagenomic classification (Peabody *et al*, 2015; Sczyrba *et al*, 2017; McIntyre *et al*, 2017; Ye *et al*, 2019), in particular the Critical Assessment of Metagenome Interpretation (CAMI) initiative released OPAL (Meyer *et al*, 2019) and AMBER (Meyer *et al*, 2018), a framework for systematic comparisons of taxonomic profilers using predetermined criteria This is important because not all tools behave in the same way and it is necessary to understand which one is more suitable when performing different types of analyses.

Genome binning

Regardless of the method chosen, metagenomic assemblers will not produce a single, but a multitude of fragmented contigs, usually of short length. It is then necessary to use genome binners to obtain a metagenome-assembled genome (MAG) from contigs. Genome binners are capable of grouping together contigs coming from the same genome using a variety of approaches such as GC content, tetranucleotide frequencies, or genome coverage.

A simple supervised binning approach can be implemented by classifying contigs by searching for nucleotide similarity using BLAST or HMMER. However, as a supervised approach, it requires previous knowledge of the taxonomy of the genomes used as reference. Hybrid taxonomic-driven approaches like MetaWatt (Strous *et al*, 2012) combines taxonomic signatures, tetranucleotide frequencies, and total coverage all interpolated using Markov modeling via Glimmer (Delcher *et al*, 2007) but the process is not completely automatic since requires input from the user.

Compositional-based binners use intrinsic contig properties such as GC content or nucleotide frequencies. While it could not be classified as a contig binner *per-se*, Canopy (Nielsen *et al*, 2014) performs gene clustering extracted from the contigs using their abundances and then groups contigs by sample co-abundance in a so called co-abundance gene group (CAG). MyCC (Lin & Liao, 2016), a fully automatic binner, primarily uses *k*-mer frequencies to assign contigs to bins, but coverage information and conserved marker genes can be used to improve the binning process. CONCOCT (Alneberg *et al*, 2014) combines contig's *k*-mer frequencies and coverage into a matrix and apply principal component analysis to perform dimensionality reduction. Thanks to this last step, CONCOCT performs well when it is required to bin contigs generated from complex microbial communities and in case of co-assembly of reads.

MetaBAT2 (Kang *et al*, 2019) and MaxBin2 (Wu *et al*, 2016b) both uses probabilistic models built on distances between *k*-mer frequencies and coverage. Those distances are then compared to a probabilistic model built using inter and intra-species distances extracted from known genomes. Due to the ease of computing probability distances, both methods are suitable when many samples are taken into consideration, however a main limitation of these methods can be found in the pre-built model used to compare those distances since it was generated using a limited number of available reference genomes.

A useful process that can be done after contig binning is to map the metagenomic reads to the bins in order to produce bin abundances or perform genome curation in order to close gaps or correcting assembly errors. Tools such as Anvi'o (Eren *et al*, 2021) make it possible to manually inspect the produced bins and also enables investigations of strain-level variants

across multiple metagenomic datasets. Recovered genomes can be checked for completeness and contamination using CheckM (Parks *et al*, 2015), which uses lineage-specific marker genes, or single-copy orthologous genes as implemented in BUSCO (Simão *et al*, 2015). In order to determine the completeness of the bins, CheckM first determines the most probable lineage by searching for the presence of the lineage-specific single-copy marker genes. To be considered as complete, a bin is required to contain as much as possible lineage-specific marker genes. Failure in identifying them leads to a decrease in the completeness score. A similar method is used for assessing the contamination. Since the marker genes are lineage specific, presence of multiple lineage marker genes indicates that the bin contains contigs reconstructed from different species. One limitation in the evaluation of bins with CheckM is that eukaryotic genomes will be reported with low values of completeness since the single-copy marker are extracted from bacterial and archaeal genomes. A good alternative to CheckM is BUSCO, which alongside prokaryotic genomes, it is able to evaluate eukaryotic ones. Other binning evaluators able to work with eukaryotic genomes are EukCC (Saary *et al*, 2020) and EukRep (West *et al*, 2018).

It may be worth noting that evaluation of reconstructed genomes using single copy genes may lead to errors in the evaluation: a genome reported with a low completeness value using single copy genes could be due to the fact that the taxa is lacking the genes considered for the evaluation, while they can be present in other taxa (Brown *et al*, 2015; Chen *et al*, 2020).

Functional potential analysis

The functional potential analysis addresses the problem of identifying the whole gene content of a metagenome. Usually, this is done by mapping the metagenome to a database of functionally characterized protein sequences, like KEGG (Kanehisa & Goto, 2000), UniRef (Suzek *et al*, 2007), COG (Tatusov *et al*, 2000), or eggNOG (Jensen *et al*, 2008; Huerta-Cepas *et al*, 2016b). Functional databases classify proteins into gene families, sets of proteins sharing a common sequence and, more often, a similar function (Dayhoff, 1976). Typically, gene families are then used to perform pathway reconstruction, using definitions from MetaCyc (Caspi *et al*, 2014) or KEGG, for successive analyses aimed to link the functional profile of the microbial community with phenotypes.

The MG-RAST web server, previously introduced as a taxonomic classifier, also implements a pipeline to perform gene prediction for subsequent functional annotation against the SEED classification of subsystems (Overbeek *et al*, 2005, 2014). As of MEGAN4 (Huson *et al*, 2011), the MEGAN suite also introduced functional analysis using SEED, KEGG, InterPro, and eggNOG. Initially, both tools relied on the BLAST suite for performing sequence alignment, and with the release of MEGAN6, NCBI-BLAST was replaced in favor of DIAMOND (Buchfink

et al, 2015), a high-throughput DNA-to-protein aligner. The approach used is based on BLAST: reads from the metagenomic sample or open reading frames extracted from assembled genomes are mapped to a database of interest using translated BLAST (BLASTX), and the best hit is used to assign the function to the gene.

In order to determine the functional potential of a community, it is necessary to look at the comprehensive set of all the reads, a task that can be computationally intensive and not avoidable by using, for example, a marker-based approach. Furthermore, it is possible to characterize the functions encoded by the uncharacterized species that make up the microbial "dark matter" by simply searching for orthologous genes. More recent and better performing methods have been developed to overcome the increasing number of available sequences, such as the introduction of faster protein search algorithms like DIAMOND or RAPSearch2 (Zhao *et al*, 2012). The latter method has been used by SUPER-FOCUS (Silva *et al*, 2016), which internally uses the taxonomic profiler FOCUS in order to identify genera present in the metagenomic sample, and also by ShotMAP (Nayfach *et al*, 2015).

In addition to the other tools previously described, as part of the Human Microbiome Project, HUMAnN (Abubucker *et al*, 2012) was developed to provide a scalable method able to handle the large amount of data generated from the seven body sites considered in the study. HUMAnN allows users to directly profile metagenomes from the raw reads rather than assemble reads into contigs. To reduce the bottleneck introduced by the translated search, the second version of the method, HUMAnN2 (Franzosa *et al*, 2018) introduced a tiered approach that limits the number of reads mapped with the translated search by first mapping the raw reads against functional annotated species-level pangenomes, allowing for faster and accurate profiling.

Strain-level analysis

With the increasing availability of reference genomes and studies aimed to characterize them, phenotypes have been observed to be usually associated with a specific bacterial strain (Segata, 2018). Numerous events like antibiotic resistance gene acquisition (Manara *et al*, 2018), pathogenicity (Leimbach *et al*, 2013), or host-to-host transmission (Asnicar *et al*, 2017; Ferretti *et al*, 2018; Yassour *et al*, 2018) occur at the strain level. Strain characterization is a task that has traditionally been done for only a limited number of cultivable species relying on single isolate sequencing. Recent advances in sequencing and the continuous decreasing cost of sequencing allowed researchers for a high-throughput investigation of microbial strains using different computational approaches. There is no single definition of what a "strain" is. Taking into account each single nucleotide variant (SNV) introduced into the genome, all isolated microorganisms could be potentially classified as different strains. A thorough

analysis of 90K bacterial genomes conducted by (Jain *et al*, 2018) defines two genomes assigned to the same species if they share more than 95% average nucleotide identity (ANI) calculated over the whole genome sequence, but a clear-cut threshold may not exists for an operational definition of strains.

Several tools have been developed to perform strain-level taxonomic profiling using different approaches, ranging from classical microbiology tools applied to metagenomes, an approach chosen by MetaMLST (Zolfo *et al*, 2017) and MG-MLST (Bangayan *et al*, 2020), while other tools like ConStrains (Luo *et al*, 2015), DESMAN (Quince *et al*, 2017a), StrainPhlAn (Truong *et al*, 2017), and MetaSNV (Costea *et al*, 2017a) identify species' strains by analyzing the SNV present in the whole genome or within marker genes. Another family of methods relies on gene content instead of SNV: methods that use this approach are PanPhlAn (Scholz *et al*, 2016) and MIDAS (Nayfach *et al*, 2016). Methods that identify strains using SNV can provide a more detailed phylogenetic reconstruction but lack functional characterization, a perk of gene-content-based methods. The latter allows for tracking the gene content variation over time and are able to track low abundant species.

Despite several methods that have been developed to tackle this problem, identifying and quantifying strains from metagenomes is still a challenge in the field. One big limitation is given by the fact that it is not feasible to resolve strains based on genome sequence variants for most of the species having very few sequenced genomes available in public databases. This is crucial for aligning-based methods: while they can outperform on human-associated microbiomes, which have a lot of sequenced reference genomes, they could be penalized on not-so-well-characterized microbiomes like free-living microbial communities. Reproducible and robust methods to identify strains and genetic variants can enable future studies to identify specific variants linked to the onset of a specific disease, a similar approach used in genome-wide association studies (GWAS) (Power *et al*, 2017), or even identify specific organisms that are beneficial for therapeutic treatments (e.g. fecal microbiota transplant (Smillie *et al*, 2018))

Integrated pipelines for metagenomic analysis

Although several methods have been developed as individual and standalone tools focusing on a single task that can be taxonomic profiling, functional profiling, or strain-level profiling, suites that integrate individual computational platforms are available.

Two classic tools as MEGAN (Huson *et al*, 2007) and MG-RAST (Meyer *et al*, 2008) are all-in-one solutions that alongside taxonomic and functional profiling tools also include visualization and statistical tools in order to analyze microbiomes using a single application. Other integrated pipelines like MIDAS (Nayfach *et al*, 2016) allow for species-level and strain-

level taxonomic profiling leveraging a custom database clustered at species level mainly including human-associated species.

A different approach adopted for integrating pipelines is to implement a workflow-based pipeline, as implemented both by the bioBakery suite (McIver *et al*, 2018) and QIIME2 (Bolyen *et al*, 2019). The bioBakery suite includes AnADAMA, a scientific workflow manager that wraps the several individual tools included in the suite. QIIME2 architecture is based on plugins allowing for the integration of different tools inside the ecosystem (e.g., the q2-metaphlan2 plugin allows to run MetaPhlAn2 through QIIME 2 and then process the output using other QIIME2 plugins). Plugins can be coordinated using workflow systems such as Jupyter Notebooks (Kluyver *et al*, 2016), Common Workflow Language (CWL) (Amstutz *et al*, 2016), or Snakemake (Köster & Rahmann, 2012) that can be run on a single machine or scaled-up to a grid or cloud computing environment. Both the bioBakery suite and QIIME2 allow the end-user to perform a full metagenomic analysis as described in **Figure 1** since both include tools able to perform each step of the theoretical metagenomic workflow. Online resources like the Galaxy framework (Afgan *et al*, 2018) allow researchers to process, analyze, and visualize metagenomic data using a web server. Within the Galaxy framework, all the tasks can be performed singularly but the key feature is the availability of a workflow manager to orchestrate the execution of tools.

Integrated workflow-based pipelines are key for reproducible research: they allow to track all the tasks that are needed to be run for a single analysis and more importantly, analyses are performed transparently, also tracking which exact version of the software was used, an important factor when comes to identifying possible changes in results due to software bugs. Computational workflows improve efficiency by parallelizing the same set of tasks on multiple samples, potentially the ease of the parallelization is offered by providing only information about the location of the input and output files.

Databases for microbial genomics

Microbial databases are crucial when it comes to performing metagenomic analyses. To date, several general purpose and specific databases are available storing crucial information on genes, proteins, genomes, expression data, gene interactions, and metabolic pathways. NCBI RefSeq is the *de facto* microbial database chosen by the majority of researchers for depositing reference genomes, but several other databases are available (e.g., PATRIC, Ensembl, JGI 1K, FungiDB, IMG). However, researchers choose to submit reference genomes exclusively to a single database, resulting in a non-concordance of the microbial genomes in those databases (Loeffler *et al*, 2020). Many databases have unique content not shared with other databases and this makes it difficult for tools to rely on the same set of reference genomes.

A good way to handle proteins is to classify them into different protein families. To this end, different databases organize proteins according to different criteria, just to mention a few, PFAM (Finn *et al*, 2014) classifies protein sequences according to conserved protein domains or protein structure, EggNOG (Jensen *et al*, 2008) and COG (Tatusov *et al*, 2000) provides a phylogenetic classification of proteins. Reconstructing the metabolism of an organism starting from the genome is a challenging task that combines high-throughput data and classical biochemistry. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) provides an extensive catalog of organism-specific metabolic pathways and genomic information. Along the lines of KEGG, the MetaCyc database (Caspi *et al*, 2014), part of BioCyc, makes available metabolic pathways, enzymes, reactions, and metabolites for a multitude of species. Although at a first sight these resources can look independent one from the other, efforts in creating a "one-stop-shop" database have positively resulted in portals hosted by the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute at the European Molecular Biology Laboratory (EMBL-EBI), which store genomes submitted to the International Nucleotide Sequence Database Consortium (INSDC) (Cochrane *et al*, 2016), a joint initiative between the DNA Data Bank of Japan (DDBJ), EMBL-EBI and NCBI. EMBL-EBI offers a protein-centered portal that cross-links databases and tools, UniProt (UniProt Consortium, 2019), a comprehensive resource for protein sequence and annotation data. The UniProt Knowledgebase (UniProtKB) is a collection of manually curated (SwissProt) and unreviewed (TrEMBL) proteins extracted from the genomes deposited in the INSDC and linked via cross-references to several different functional, phylogenomic, and protein domain databases (KEGG, KO, EggNOG, GO, EC, Pfam) (Kanehisa & Goto, 2000; Huerta-Cepas *et al*, 2016b; The Gene Ontology Consortium, 2019; El-Gebali *et al*, 2019).

Such levels of integration and cross-reference with multiple databases facilitates the development of automated, high-throughput analyses, even if a level of manual curation is needed. One aim of this thesis is to provide an integrated system to organize the microbial genomic information, here we want to present ChocoPhlAn, a genomic repository of high-quality fully-annotated microbial reference genomes and the corresponding functionally-annotated gene families.

1.2. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3

This chapter contains a work made available online via a preprint on bioRxiv and currently submitted for publication in a scientific journal.

In this project I designed and implemented the methods for ChocoPhlAn and led the update of the source code for MetaPhlAn. I also was in charge for the creation of standardized packaging for most of the bioBakery tools. I performed all the validation of the methods and software for ChocoPhlAn and updated them based on validation issues and external feedback, I also led the analysis related to the evaluation of MetaPhlAn, processed all the samples analyzed in the meta-analysis of the CRC datasets, performed all the statistical analysis except for the random forest classification and the standardized mean differences based meta-analysis, and wrote the first draft of the manuscript.

Francesco Beghini, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Andrew Maltez Thomas, Paolo Manghi, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A. Franzosa, Nicola Segata

Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3

bioRxiv preprint (2020) - https://doi.org/10.1101/2020.11.19.388223

Abstract

Culture-independent analyses of microbial communities have advanced dramatically in the last decade, particularly due to advances in methods for biological profiling via shotgun metagenomics. Opportunities for improvement continue to accelerate, with greater access to multi-omics, microbial reference genomes, and strain-level diversity. To leverage these, we present bioBakery 3, a set of integrated, improved methods for taxonomic, strain-level, functional, and phylogenetic profiling of metagenomes newly developed to build on the largest set of reference sequences now available. Compared to current alternatives, MetaPhlAn 3 increases the accuracy of taxonomic profiling, and HUMAnN 3 improves that of functional potential and activity. These methods detected novel disease-microbiome links in applications to CRC (1,262 metagenomes) and IBD (1,635 metagenomes and 817 metatranscriptomes). Strain-level profiling of an additional 4,077 metagenomes with StrainPhlAn 3 and PanPhlAn 3 unraveled the phylogenetic and functional structure of the common gut microbe *Ruminococcus bromii*, previously described by only 15 isolate genomes. With open-source

implementations and cloud-deployable reproducible workflows, the bioBakery 3 platform can help researchers deepen the resolution, scale, and accuracy of multi-omic profiling for microbial community studies.

Introduction

Studies of microbial community biology continue to be enriched by the growth of culture-independent sequencing and high-throughput isolate genomics (Pasolli *et al*, 2019; Almeida *et al*, 2019; Zou *et al*, 2019; Almeida *et al*, 2020; Parks *et al*, 2017; Forster *et al*, 2019; Poyet *et al*, 2019). Shotgun metagenomic and metatranscriptomic (i.e. “meta-omic”) measurements can be used to address an increasing range of questions as diverse as the transmission and evolution of strains in situ (Ferretti *et al*, 2018; Yassour *et al*, 2018; Asnicar *et al*, 2017; Truong *et al*, 2017), the mechanisms of multi-organism biochemical responses in the environment (Blaser *et al*, 2016; Alivisatos *et al*, 2015), or the epidemiology of the human microbiome for biomarkers and therapy (Zeller *et al*, 2014; Le Chatelier *et al*, 2013; Gopalakrishnan *et al*, 2018; Thomas *et al*, 2019). Using such analyses for accurate discovery, however, requires efficient ways to integrate hundreds of thousands of (potentially fragmentary) isolate genomes with community profiles to detect novel species and strains, non-bacterial community members, microbial phylogeny, and evolution, and biochemical and molecular signaling mechanisms. Correspondingly, this computational challenge has necessitated the continued development of platforms for the detailed functional interpretation of microbial communities.

The past decade of metagenomics has seen remarkable growth both in the biology accessible via high-throughput sequencing and in the methods for doing so. Beginning with the now-classic questions of “who’s there?” and “what are they doing?” in microbial ecology (Human Microbiome Project Consortium, 2012), shotgun metagenomics provide a complementary means of taxonomic profiling to amplicon-based (e.g. 16S rRNA gene) sequencing, as well as functional profiling of genes or biochemical pathways (Quince *et al*, 2017b; Segata *et al*, 2013a; Morgan *et al*, 2013). More recently, metagenomic functional profiles have been joined by metatranscriptomics to also capture community regulation of gene expression (Lloyd-Price *et al*, 2019). Methods have been developed to focus on all variants of particular taxa of interest within a set of communities (Pasolli *et al*, 2019), to discover new variants of gene families or biochemical activities (Franzosa *et al*, 2018; Kaminski *et al*, 2015), or to link the presence and evolution of closely related strains within or between communities over time, space, and around the globe (Tett *et al*, 2019; Karcher *et al*, 2020; Beghini *et al*, 2017). Critically, all of these analyses (and the use of the word “microbiome” throughout this manuscript) are equally applicable to both bacterial and non-bacterial community members (e.g. viruses and eukaryotes) (Beghini *et al*, 2017; Olm *et al*, 2019; Yutin *et al*, 2018). Finally, although not

addressed in depth by this study, shotgun meta-omics have increasingly also been combined with other community profiling techniques such as metabolomics (Sun *et al*, 2018; Heinken *et al*, 2019; Lloyd-Price *et al*, 2017) and proteomics (Xiong *et al*, 2015) to provide richer pictures of microbial community membership, function, and ecology.

Methods enabling such analyses of meta-omic sequencing have developed in roughly two complementary types, either relying on metagenomic assembly or using largely assembly-independent, reference-based approaches (Quince *et al*, 2017b). The latter is especially supported by the corresponding growth of fragmentary, draft, and finished microbial isolate genomes, and their consistent annotation and clustering into genome groups and pan-genomes (Pasolli *et al*, 2019; Almeida *et al*, 2019, 2020). Most such methods focus on addressing a single profiling task within (most often) metagenomes, such as taxonomic profiling (Wood *et al*, 2019; Lu *et al*, 2017; Truong *et al*, 2015; Milanese *et al*, 2019), strain identification (Truong *et al*, 2017; Scholz *et al*, 2016; Luo *et al*, 2015; Nayfach *et al*, 2016), or functional profiling (Franzosa *et al*, 2018; Kaminski *et al*, 2015; Nazeen *et al*, 2020; Nayfach *et al*, 2015). In a few cases, platforms such as the bioBakery (McIver *et al*, 2018), QIIME 2 (Bolyen *et al*, 2019), or MEGAN (Mitra *et al*, 2011) integrate several such methods within an overarching environment. While not a primary focus of this study, metagenomic assembly methods enabling the former types of analyses (e.g. novel organism discovery or gene cataloging (Lesker *et al*, 2020; Stewart *et al*, 2019)) have also advanced tremendously (Li *et al*, 2015; Nurk *et al*, 2017) and are now reaching a point of integrating microbial community and isolate genomics, particularly for phylogeny (Asnicar *et al*, 2020; Zhu *et al*, 2019). These efforts have also led to increased consistency in microbial systematics and phylogeny, facilitating the types of automated, high-throughput analyses necessary when manual curation cannot keep up with such rapid growth (Chaumeil *et al*, 2019; Asnicar *et al*, 2020).

Here, to further increase the scope of feasible microbial community studies, we introduce a suite of updated and expanded computational methods in a new version of the bioBakery platform. The bioBakery 3 includes updated sequence-level quality control and contaminant depletion guidelines (KneadData), MetaPhlAn 3 for taxonomic profiling, HUMAnN 3 for functional profiling, StrainPhlAn 3 and PanPhlAn 3 for nucleotide- and gene-variant-based strain profiling, and PhyloPhlAn 3 for phylogenetic placement and putative taxonomic assignment of new assemblies (metagenomic or isolate). Most of these tools leverage an updated ChocoPhlAn 3 database of systematically organized and annotated microbial genomes and gene family clusters, newly derived from UniProt/UniRef (Suzek *et al*, 2007) and NCBI (NCBI Resource Coordinators, 2014). Our quantitative evaluations show each individual tool to be more accurate and, typically, more efficient than its previous version and other comparable methods, increasing sensitivity and specificity by sometimes more than 2-fold

(e.g., in non-human-associated microbial communities). Biomarker identifications in 1,262 colorectal cancer (CRC) metagenomes, 1,635 inflammatory bowel disease (IBD) metagenomes, and 817 metatranscriptomes show both the platforms efficiency and its ability to detect hundreds of species and thousands of gene families not previously profiled. Finally, in 4,077 human gut metagenomes containing *Ruminococcus bromii*, the bioBakery 3 platform permits an initial integration of assembly- and reference-based metagenomics, discovering a novel biogeographical and functional structure within the clade's evolution and global distribution. All components are available as open-source implementations with documentation, source code, and workflows enabling provenance, reproducibility, and local or cloud deployment at http://segatalab.cibio.unitn.it/tools/biobakery and http://huttenhower.sph.harvard.edu/biobakery.

Results

The bioBakery provides a complete meta-omic tool suite and analysis environment, including methods for individual meta-omic (and other microbial community) processing steps, downstream statistics, integrated reproducible workflows, standardized packaging and documentation via open-source repositories (GitHub, Conda, PyPI, and R/Bioconductor), grid- and cloud-deployable images (AWS, GCP, and Docker), online training material and demonstration data, and a public community support forum. For any sample set, quality control, taxonomic profiling, functional profiling, strain profiling, and resulting data products and reports can all be generated with a single workflow, while maintaining version control and provenance logging. All of the methods themselves, the associated training material, quality control using KneadData, and packaging for distribution and use have been updated in this version. For example, Docker images have been scaled down in size to optimize use in cloud environments, and workflows have been ported to AWS (Amazon Web Services) Batch and Terra/Cromwell (Google Compute Engine) to reduce costs through the use of spot and pre-emptive instances, respectively. All base images and dependencies have been updated as well, including the most recent Python (v3.7+) and R (v4.0+, see **Methods**). New and updated documentation of all tools, including detailed instructions on installation in different environments and package managers, is available at http://huttenhower.sph.harvard.edu/biobakery.

High-quality reference sequences for improved meta-omic profiling

The majority of methods within the bioBakery 3 suite leverage a newly-updated reference genome and gene cataloging procedure, the results of which are packaged as ChocoPhlAn 3 (**Fig. 1.2.1A**) (McIver *et al*, 2018). ChocoPhlAn uses publicly available genomes and standardized gene calls and gene families to generate markers for taxonomic and strain-level

profiling of metagenomes with MetaPhlAn 3, StrainPhlAn 3, and PanPhlAn 3, phylogenetic profiling of genomes and MAGs with PhyloPhlAn 3, and functional profiling of metagenomes with HUMAnN 3.

ChocoPhlAn 3 is based on a genomic repository of 99.2k high-quality, fully annotated reference microbial genomes from 16.8k species available in the UniProt Proteomes portal as of January 2019 (UniProt Consortium, 2019) and the corresponding functionally-annotated 87.3M UniRef90 gene families (Suzek *et al*, 2015). From this resource, ChocoPhlAn initially generates annotated species-level pangenomes associating each microbial species with its sequenced genomes and repertoire of UniRef-based gene (nucleotide) and protein (amino acid sequence) families. These pangenomes provide a uniform shared resource for subsequent profiling across bioBakery 3. HUMAnN 3 and PanPhlAn 3 are directly based on complete pangenomes for overall functional and strain profiling, whereas other tools use additional information annotated onto the catalog. PhyloPhlAn 3 focuses on the subset of conserved core gene families (i.e., present in almost all strains of a species) for inferring accurate phylogenies, and both MetaPhlAn 3 and StrainPhlAn 3 further refine core gene families into species-specific unique gene families to generate unambiguous markers for metagenomic species identification and strain-level genetic characterization.

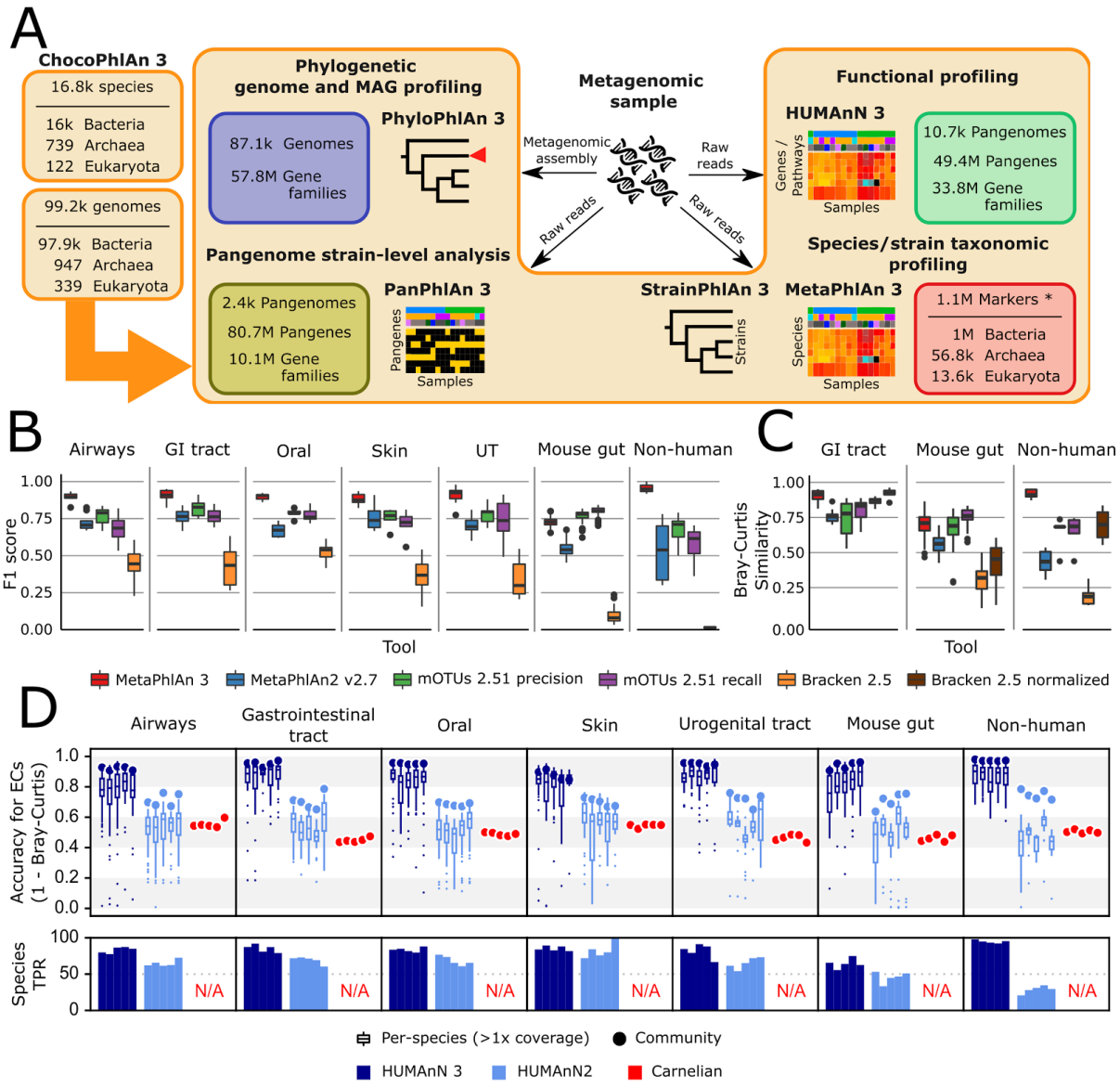


Figure 1.2.1: bioBakery 3 includes new microbial community profiling approaches that outperform previous versions and current methods. (A) The newly developed ChocoPhlAn 3 consolidates, quality controls, and annotates isolate-derived reference sequences to enable metagenomic profiling in subsequent bioBakery methods. (*The 1.1M MetaPhlAn 3 markers also comprise for 61.8k viral markers from MetaPhlAn 2 (Truong *et al*, 2015); other version descriptions in (Truong *et al*, 2017; Scholz *et al*, 2016; Asnicar *et al*, 2020)) **(B)** MetaPhlAn 3 was applied to a set of 113 total evaluation datasets provided by CAMI (Fritz *et al*, 2019) representing diverse human-associated microbiomes and 5 datasets of non-human-associated microbiomes (**Table S1.2.1**). MetaPhlAn 3 showed increased performance compared with the previous version MetaPhlAn 2 (Truong *et al*, 2015), mOTUs2 (Milanese *et al*, 2019), and Bracken 2.5 (Lu *et al*, 2017). We report here the F1 scores (harmonic mean of the species-level precision and recall, see **Fig. S1.2.1** for other evaluation scores). **(C)** MetaPhlAn 3 better recapitulates relative abundance profiles both from human and murine gastrointestinal metagenomes as well from non-human-associated communities compared to the other currently available tools (full results in **Fig. S1.2.1**). Bracken is reported both using its original estimates based on the fraction of reads assigned to each taxa and after re-normalizing them using the genome lengths of the taxa in the gold standard to match the taxa abundance estimate of the other tools. **(D)** Compared with HUMAnN 2 (Franzosa *et al*, 2018) and Carnelian (Nazeen *et al*, 2020), HUMAnN 3 produces more accurate estimates of EC abundances and displays a higher species true positive rate compared to HUMAnN 2.

MetaPhlAn 3 increases the accuracy of quantitative taxonomic profiling

MetaPhlAn estimates the relative abundance of microbial taxa in a metagenome using the coverage of clade-specific marker genes (Segata *et al*, 2012b; Truong *et al*, 2015). Such marker genes are chosen so that essentially all of the strains in a clade (species or otherwise) possess such genes, and at the same time no other clade contains orthologs close enough to incorrectly map metagenomic reads. MetaPhlAn 3 incorporates 13.5k species (more than twice than MetaPhlAn 2) with a completely new set of 1.1M marker genes (84±47 mean±SD markers per species) selected by ChocoPhlAn 3 from the set of 16.8k species pangenomes. The adoption of UniRef90 gene families permitted the efficient expansion of the core-gene identification procedure, which is followed by a mapping of potential core genes against all available whole microbial genomes to ensure unique marker identification (see **Methods**). This restructuring of the marker selection process has been combined with several improvements and extensions of the algorithm, including optimized quality control during marker alignments and an estimation of the metagenome fraction composed of unknown microbes (**Table S1.2.2**).

We evaluated the taxonomic profiling performance of MetaPhlAn 3 using 118 synthetic metagenomes spanning 113 synthetic samples from the 2nd CAMI Challenge (Fritz *et al*, 2019; Sczyrba *et al*, 2017) through the OPAL benchmarking framework (Meyer *et al*, 2019) These represent typical microbiomes from five human-associated body sites and the murine gut, and we complemented them with 5 additional newly-generated synthetic non-human-associated metagenomes (see **Methods**). In addition to MetaPhlAn 3, the comparative evaluation considered MetaPhlAn 2.7 (Truong *et al*, 2015), mOTUs 2.51 (Milanese *et al*, 2019) (latest database available as of July 2020), and Bracken 2.5 (using a database built after the April 2019 RefSeq release) (Lu *et al*, 2017; Wood *et al*, 2019). These three profiling tools have consistently been shown to outperform other methods across multiple evaluations (Truong *et al*, 2015; Milanese *et al*, 2019; Meyer *et al*, 2019; Sczyrba *et al*, 2017; McIntyre *et al*, 2017; Ye *et al*, 2019).

MetaPhlAn 3 outperformed all the other profilers across all considered types of communities when assessing the F1 score (**Fig. 1.2.1B**), which is a measure combining the fraction of species actually present in the metagenomes that are correctly detected (recall, **Fig. S1.2.1**) and the fraction of species predicted to be present that were actually included in the synthetic metagenome (precision, **Fig. S1.2.1**). With a very low number of false positive species detected, MetaPhlAn 3 (avg 8.51 s.d. 5.12) also maximized precision (**Fig. S1.2.1**) with respect to the other tools (avg 9 s.d. 4.78 for mOTUs in high precision mode, the closest competitor on precision). On recall, Bracken and mOTUs in high-recall mode were in several

cases superior to MetaPhlAn 3, but at the cost of a very high number of false positives (on average 729 species for Bracken and 39 for mOTUs high-recall, for a total of 86,077 and 4,655 false positive species across the synthetic metagenomes). MetaPhlAn can further minimize false positives by requiring a higher fraction of positive markers for positive species ("--stat_q" parameter, **Fig. S1.2.2**), but overall the F1 measure with default settings remains higher than the other evaluated tools across the panel of synthetic metagenomes in our evaluation.

In addition to more accurate species detection, MetaPhlAn 3 also quantified taxonomic abundance profiles more accurately compared to MetaPhlAn 2, mOTUs2, and Bracken based on Bray-Curtis dissimilarities in most datasets (**Table S1.2.3**, **Fig. 1.2.1C**). While it was slightly outperformed by mOTUs (only in high-recall mode) on the synthetic mouse gut dataset, even in this case, correlation-based measures (Pearson Correlation Coefficient between estimated and expected relative abundances) found MetaPhlAn 3 to be more accurate (r=0.73) than the other considered profilers (MetaPhlAn 2 r=0.63, mOTUs2 precision r=0.60, mOTUs2 recall r=0.71, Bracken r=0.43). Additionally, because Bracken estimates the fraction of reads belonging to each taxa rather than the relative abundance of each taxa, we also re-normalized its abundances based on genome length of the target species. This post-hoc re-normalization improved Bracken's performance on taxonomic abundances (but not false positives and false negatives) that however remained comparable with MetaPhlAn 3 only in some of the simulated environments (**Fig. S1.2.1**). Overall, this confirms that MetaPhlAn 3 is superior to its previous version and is more accurate than other currently available tools in the large majority of simulated environment-specific datasets.

In addition to improvements in accuracy, MetaPhlAn 3's computational efficiency also compares favorably with alternatives and with its previous version. It is >3x faster than MetaPhlAn 2 (10.0k vs. 2.9k reads/second on a Xeon Gold 6140) and almost matches the speed of Bracken (11k reads per second). MetaPhlAn 3 memory usage is slightly higher (2.6Gb for a complete taxonomic profiling run) than MetaPhlAn 2 (2.1Gb), but outperforms the other methods (4.3 Gb for mOTUs and 32.5 Gb for Bracken, **Fig. S1.2.2**, **Table S1.2.4**).

HUMAnN 3 accurately quantifies species' contributions to community function

HUMAnN 3 functionally profiles genes, pathways, and modules from metagenomes, now using native UniRef90 annotations from ChocoPhlAn species pangenomes. We compared its performance against HUMAnN 2 (Franzosa *et al*, 2018), and the recently published Carnelian (Nazeen *et al*, 2020) when profiling the 30 CAMI and 5 additional synthetic metagenomes introduced above (see **Methods** and **Fig. 1.2.1**). Carnelian was selected because it was published subsequent to HUMAnN 2 and, more importantly, follows the HUMAnN strategy of

estimating the relative abundance of molecular functions directly from shotgun meta-omic sequencing reads rather than assembled contigs (albeit by a different approach). While HUMAnN 2 and 3 can both natively estimate the relative abundances of a wide variety of functional features from a metagenome (by first quantifying and then manipulating UniRef90 or UniRef50 abundances), we selected level-4 enzyme commission (EC) categories as a basis for comparison with Carnelian, as the method's authors provided a precomputed index for EC quantification (Nazeen *et al*, 2020).

HUMAnN 3 produced highly accurate estimates of community-level EC abundances across the 30 CAMI metagenomes (mean ± SD of Bray-Curtis similarity=0.93 ± 0.03, **Fig. 1.2.1**). HUMAnN 2 followed with an accuracy of 0.70 ± 0.04 and Carnelian at 0.49 ± 0.04. While HUMAnN 3 benefits in part from access to a more up-to-date sequence database, we note that HUMAnN 2's database (c. 2014) predates the Carnelian method by several years, and so recency cannot be the only explanation for this trend. For example, Carnelian uses a mean sequence length per EC during abundance estimation, a choice which may contribute additional error relative to HUMAnN's sum over per-sequence estimates. We observed similar trends in accuracy among the three methods using F1 score to prioritize presence/absence calls over abundance (**Fig. S1.2.3**). HUMAnN 2 and Carnelian were notably similar with respect to sensitivity (0.72 ± 0.05 vs. 0.74 ± 0.04, respectively) but not precision (0.95 ± 0.02 vs. 0.60 ± 0.08). This difference is attributable in part to HUMAnN's use of database sequence coverage filters (see **Methods**) to reduce false positives, an approach introduced for translated search in HUMAnN 2 and expanded to nucleotide search in HUMAnN 3 (**Fig. S1.2.4**).

One of the main advantages of HUMAnN 3 (and 2) compared with other functional profiling systems (including Carnelian) is their ability to stratify community functional profiles according to contributing species. This feature is additionally more accurate and useful in HUMAnN 3 as a function of its broader pangenome catalog. Across the CAMI metagenomes, EC accuracy for species with at least 1x mean coverage depth was 0.81 ± 0.16 for HUMAnN 3 and 0.51 ± 0.15 for HUMAnN 2 (mean ± SD within-species Bray-Curtis similarity; **Fig. 1.2.1**). HUMAnN 3 (via MetaPhlAn 3) additionally tended to detect more expected species in this coverage range compared with HUMAnN 2, a major driver of its improved community-level accuracy. As previously noted (Franzosa *et al*, 2018), HUMAnN's within-species function sensitivity is naturally lower for species below 1x coverage in a sample, as many of their genes will not have been sampled at all during the sequencing process. Per-species precision, however, remained high with HUMAnN independent of coverage and, following refinements in alignment post-processing, was slightly improved in v3 compared with v2 (0.95 ± 0.08 vs. 0.91 ± 0.07).

Carnelian was the most computationally efficient of the three methods, analyzing the CAMI metagenomes in 26.4 ± 2.7 CPU-hours (mean ± SD) compared with 38.1 ± 12.8 CPU-hours for HUMAnN 2 and 52.5 ± 19.2 CPU-hours for HUMAnN 3 (**Fig. S1.2.3**). Trends in peak memory use (MaxRSS) were similar, with Carnelian requiring 11.9 ± 0.0 GB versus HUMAnN 2's 17.0 ± 0.3 GB and HUMAnN 3's 21.5 ± 1.9 GB. We attribute these differences in large part to the sizes of the sequence spaces over which the methods search: while Carnelian focuses only on a subset of sequences annotatable to EC terms, HUMAnN aims to first quantify 10s of millions of unique UniRef90s, of which only 12.5% are ultimately annotated by ECs. The increased runtime of HUMAnN 3 compared to HUMAnN 2 is likewise attributable to the former's larger translated search database (87.3M vs. 23.9M UniRef90 sequences), as the translated search tier is the rate-limiting step of the HUMAnN algorithm even when most sample reads are explained in the preceding nucleotide-level search tiers (**Fig. S1.2.5**). This phenomenon also explains the greater runtime variability of HUMAnN, as runtimes vary inversely with the (a priori unknown) fraction of sample reads explained before the translated search tier (Franzosa *et al*, 2018). Notably, by bypassing the translated search step, HUMAnN 3 could explain the majority of CAMI metagenomic reads (70.9 ± 9.6% per sample) in only 5.8 ± 0.8 CPU-hours (a 9x speed-up; **Fig. S1.2.5**), although this is generally only appropriate for communities known to be well-covered by related reference sequences.

Evaluations on a set of synthetic metagenomes enriched for non-human-associated species resulted in similar relative accuracy and efficiency trends among the three methods (**Fig. 1.2.1** and **Fig. S1.2.3**). Hence, HUMAnN 3's strong performance is not restricted to microbial communities assembled from host-associated species. Moreover, MetaPhlAn 3's improved sensitivity for non-host-associated species increased both the accuracy and performance of HUMAnN 3 relative to HUMAnN 2 (by enabling a larger fraction of reads to be explained during the faster and more accurate pangenome search step). Finally, we evaluated HUMAnN 3's accuracy at the level of individual UniRef90 protein families (**Fig. S1.2.5**). As previously noted (Franzosa *et al*, 2018), the challenge of differentiating globally homologous UniRef90 protein sequences using short sequencing reads results in a reduction of community and per-species accuracy relative to broader gene families. However, because these homologs tend to share similar functional annotations, this error is smoothed out when individual UniRef90 abundances are combined in HUMAnN's downstream steps (as seen in the EC-level evaluation; **Fig. 1.2.1**).

MetaPhlAn 3 and HUMAnN 3 expand the link between the microbiome and colorectal cancer with a meta-analysis of 1,262 metagenomes

To illustrate the potential of bioBakery 3's updated profiling tools and to extend our understanding of the microbial signatures in colorectal cancer (CRC), we expanded our previous work to meta-analyze both existing and newly available CRC metagenomic cohorts for a total of 1,262 samples (600 control and 662 CRC samples) from 9 different datasets spanning 8 different countries (Gupta *et al*, 2019; Feng *et al*, 2015; Thomas *et al*, 2019; Vogtmann *et al*, 2016; Wirbel *et al*, 2019; Yachida *et al*, 2019; Yu *et al*, 2017; Zeller *et al*, 2014). The resulting integrated profiles are available for download (**Table S1.2.5**) and included in the new release of curatedMetagenomicData (Pasolli *et al*, 2017).

MetaPhlAn 3 identified a total of 1,083 species detected at least once (172 considered "prevalent" when defined as present in >5% of samples at >0.1% relative abundance), of which 505 species (52 prevalent) were previously not reported by MetaPhlAn 2 due to the expansion of the genome database (or in some cases because of changes in the NCBI taxonomy). In addition, 82 species present in the MetaPhlAn 2 database were not detected by MetaPhlAn 2 but are now identified in the samples by MetaPhlAn 3, due to the expanded sequence catalog, improved marker discovery procedure, and increased sensitivity to low-abundance species (Thomas *et al*, 2019).

We found 121 species significantly associated with CRC (FDR < 0.05 and Q-test for heterogeneity > 0.05; **Table S1.2.6**) by a meta-analysis of standardized mean differences using a random-effects model on arcsine-square-root-transformed relative abundances (see **Methods**). Association coefficients were also concordant with previous MetaPhlAn 2-based results using a fraction of the samples (Thomas *et al*, 2019), including the three species with the highest effect sizes: *Fusobacterium nucleatum*, *Parvimonas micra*, and *Gemella morbillorum*. We also identified three additional species not present in the previous MetaPhlAn 2 database that were among those most strongly associated with CRC (effect size >0.35): *Dialister pneumosintes*, *Ruthenibacterium lactatiformans*, and *Eisenbergiella tayi* (**Fig. 1.2.2B**, **Fig. S1.2.6**, **Fig. S1.2.7A**). Among these species, *Dialister pneumosintes* is typically oral, further reinforcing the role of oral taxa in CRC, and *Ruthenibacterium lactatiformans* was reported as part of a consortium of bacteria able to increase colonic IFNγ+ T-cells (Tanoue *et al*, 2019). The expanded number of species detectable by MetaPhlAn 3 also strengthened the previously-observed pattern of increased richness in CRC-associated microbiomes - in contrast to the stereotype of decreased diversity during dysbiosis - in large part due to low-level addition of typically oral microbes to the baseline gut microbiome (**Fig. 1.2.2C**).

Functional profiling of this expanded CRC meta-analysis with HUMAnN 3 identified 4.3M UniRef90 gene families, corresponding to 549 MetaCyc pathways and 2,895 ECs. 120 MetaCyc pathways were significantly associated with CRC (Wilcoxon rank-sum test FDR < 0.05 and Q-test for heterogeneity > 0.05) (**Fig. S1.2.7B**), of which 59 (49.1%) were in concordance with previous results and included the increased abundance of starch degradation pathway V (**Table S1.2.6**) in the healthy individuals. This pathway encodes functions for extracellular breakdown of starch by an *amylopullulanase* enzyme, which has both pullulanase and α-amylase activity (Flint *et al*, 2012). *Bifidobacterium breve* and other *Bifidobacterium* spp have been shown to encode amylopullulanases and attach to starch particles and were also reported for their potential protective role against carcinogenesis (Sivan *et al*, 2015). Among the 20 disease-associated pathways with the highest significance, only 3 were present in the previous meta-analysis with the majority exhibiting significant heterogeneity in the random effects model, possibly due to the inclusion of additional geographically distinct cohorts. Large and geographical heterogeneous CRC cohorts combined with improved taxonomic and functional profiling available via MetaPhlAn 3 and HUMAnN 3 thus have the possibility to extend and refine microbiome biomarkers of CRC .

Improvements in HUMAnN 3 also allowed us to directly test functional hypotheses in the context of the CRC microbiome. Specifically, we previously showed that the abundance of the microbial gene encoding for the choline trimethylamine-lyase (*cutC*) is significantly higher in CRC patients (Thomas *et al*, 2019), using a customized ShortBRED database (Kaminski *et al*, 2015) due to incompleteness of reference sequences previously available to HUMAnN 2. HUMAnN 3 was instead able to directly profile relative abundances of 113 UniRef90 gene families annotated as *cutC* orthologs and identified 909 metagenomes in this data collection carrying at least one UniRef90 gene family annotated as *cutC*. These confirmed an increase of *cutC* relative abundance in CRC samples compared to controls (Wilcoxon rank-sum test P<0.05 in 6 of the 9 datasets, meta-analysis P<0.0001) and thus a potential role of TMA-producing dietary choline metabolism in the gut for this malignancy. Interestingly, a meta-analysis performed on the relative abundances of the L-carnitine dioxygenase gene (*yeaW*), a gene also involved in the trimethylamine synthesis, revealed only weak associations with disease status (Wilcoxon rank-sum test P<0.05 in 3 of the 9 datasets, meta-analysis P=0.095, **Fig. S1.2.8**, **Fig. S1.2.9**), possibly reflecting a stronger effect of dietary choline on CRC risk compared to carnitine.

MetaPhlAn 3 and HUMAnN 3 also proved accurate when combining CRC microbiomes using more purely discriminative models such as random forests (RFs), reaching 0.85 average AUC for CRC (vs. control) sample classification in leave-one-dataset-out evaluations using taxonomic features (LODO, minimum 0.76 for the YachidaS_2019 and ThomasAM_2019_a

datasets, maximum 0.97 for the GuptaA_2019 dataset; **Fig. 1.2.4F**, **Fig. S1.2.10**). As in previous studies (Thomas *et al*, 2019; Pasolli *et al*, 2016), RFs using functional features performed similarly (0.69 Cross Validation and 0.71 LODO ROC AUC on pathways relative abundance), indicating a tight link between strain-specific taxonomy and gene carriage in this setting. When the classification model was used for assessing features' importance, several new taxa were identified compared to MetaPhlAn 2 and metabolic pathways or EC-numbers relative to HUMAnN 2 (**Fig. S1.2.10**), further confirming the relevance of the new reference sequences and annotations available to be profiled in bioBakery 3.

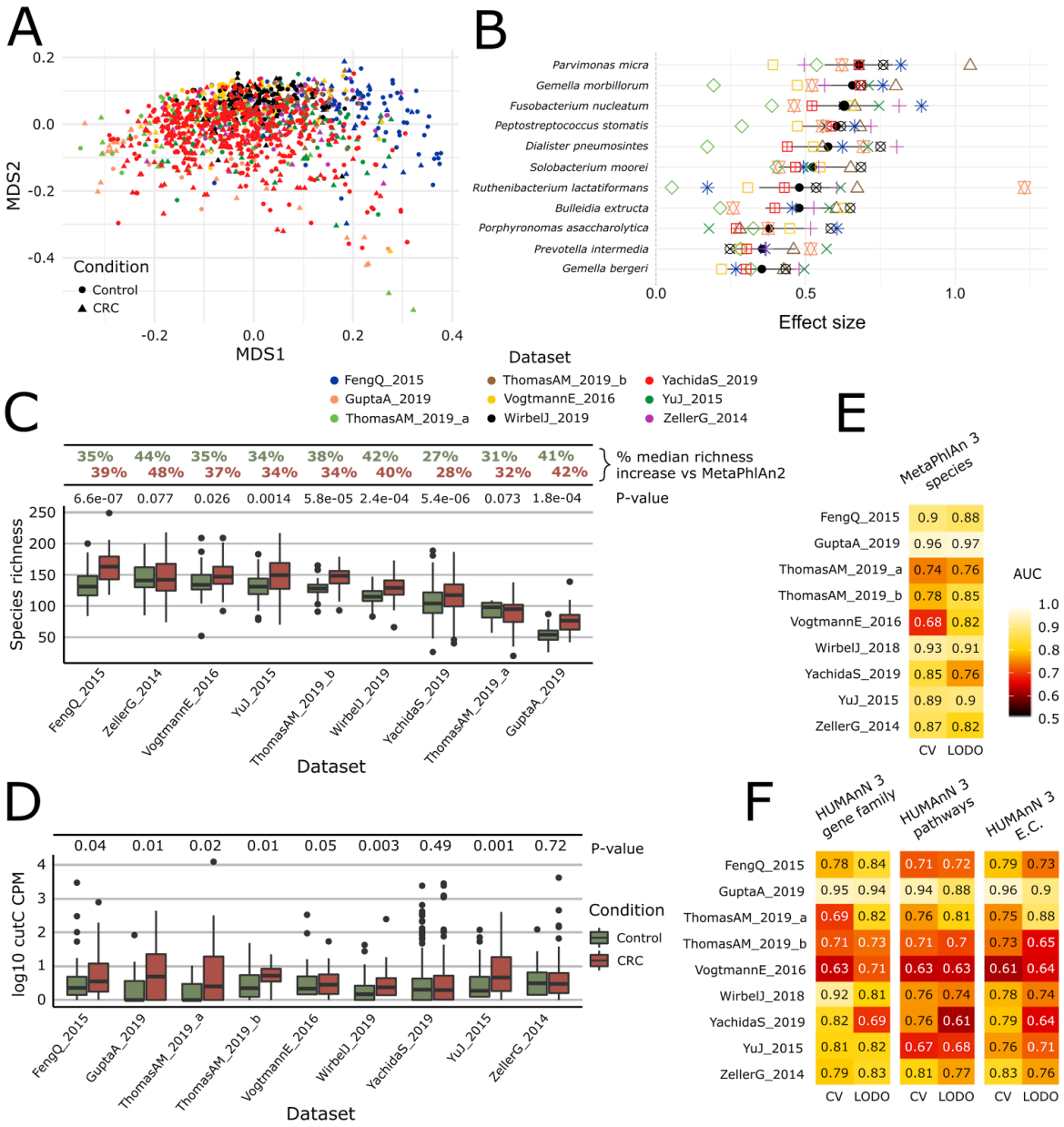


Figure 1.2.2: Meta-analysis with MetaPhlAn 3 and HUMAnN 3 expands taxonomic and functional associations with the CRC microbiome. (A) We considered a total of nine independent datasets (1,262 total samples) that highly but not completely overlap in composition based on ordination (multidimensional scaling) of weighted UniFrac distances (Lozupone & Knight, 2005) computed from the MetaPhlAn 3 species relative abundances. **(B)** Meta-analysis based on standardized mean differences and a random effects model yielded 11 MetaPhlAn 3 species significantly (Wilcoxon rank-sum test FDR P<0.05) associated with colorectal cancer at effect size>0.35 (see **Methods**). **(C)** Species richness is significantly higher in CRC samples compared to control (Wilcoxon rank-sum test P<0.05 in 7/9 datasets), and the expanded MetaPhlAn 3 species catalog detects more species compared to MetaPhlAn 2 (CRC mean median increase 37.1%, controls mean median increase 36.3%). **(D)** Distribution of *cutC* gene relative abundance (log10 count-per-million normalized) from HUMAnN 3 gene family profiles supporting the potential link between choline metabolism and CRC (Thomas *et al*, 2019). **(E)** Random forest (RF) classification using MetaPhlAn 3 features and HUMAnN 3 features **(F)** confirms that CRC patients can be predicted at (treatment-naive) baseline from the composition of their gut microbiome with performances reaching ~0.85 cross-validated or leave-one-dataset-out (LODO) ROC AUC (see **Methods**).

Longitudinal taxonomic and functional meta-omics of IBD

To further demonstrate the utility of MetaPhlAn 3 and HUMAnN 3 on combined meta-omic sequencing datasets, including identification of expression-level biomarkers, we applied the updated methods to 1,635 shotgun metagenomes (MGX) and 817 shotgun metatranscriptomes (MTX) derived from the stool samples of the HMP2 Inflammatory Bowel Disease Multi-omics Database (IBDMDB) cohort (http://ibdmdb.org; see **Methods**). Compared with previously published profiles of the samples generated with MetaPhlAn 2 and HUMAnN 2 (Lloyd-Price *et al*, 2019) (**Fig. 1.2.3A**), the v3 methods' profiles 1) identified more species pangenomes (MGX medians 40 vs. 48, MTX medians 40 vs. 47); 2) explained larger fractions of sample reads by mapping to pangenomes (MGX medians 54 vs. 63%, MTX medians 12 vs. 22%); and 3) explained larger total fractions of sample reads after falling back to translated search (MGX medians 69 vs. 75%, MTX medians 20 vs. 31%). Note that reduced MTX mapping rates (relative to MGX rates) result from enrichment for high-quality but non-coding RNA reads, which are unmapped by design in both HUMAnN 2 and 3. The v3 profiles thus promise increased understanding even of an already well-characterized dataset.

To that end, we applied a mixed-effects model to identify microbial biomarkers of disease activity within the Crohn's disease (CD) and ulcerative colitis (UC) subpopulations of the HMP2 cohort (see **Methods**). More specifically, we examined abundance profiles of EC families from 817 paired HMP2 metagenomes and metatranscriptomes in search of differences in functional activity between active (dysbiotic) and inactive (non-dysbiotic) time points from longitudinally sampled CD and UC patients. We identified 558 ECs whose residual expression was significantly different (FDR q<0.05) in active CD compared with inactive CD and a single EC that was differentially expressed in active UC (protein O-GlcNAcase, EC 3.2.1.169; **Fig. 1.2.3B**). The relative absence of biomarkers for active UC may result both from its generally more benign phenotype (Lloyd-Price *et al*, 2019) and from the smaller number of active UC samples (n=23) compared with active CD samples (n=76); as a result, we focused our subsequent analyses on expression differences within the CD subcohort.

Of the >500 significantly differentially expressed ECs in active CD, all but one were "over-expressed" (i.e. their residual expression after controlling for DNA copy number was higher than expected in active CD; see **Fig. 1.2.3B**). Hence, while many species (and their encoded functions) are known to be lost entirely during active IBD (Lloyd-Price *et al*, 2019), it seems to be rare for functions to be maintained by the community but not utilized. The one notable example of an "under-expressed" function was galactonate dehydratase (EC 4.2.1.6; **Fig. 1.2.3C**). This enzyme was encoded and highly expressed by *Faecalibacterium prausnitzii* in both control and inactive CD samples. While galactonate dehydratase was still

metagenomically abundant in active CD (where it was contributed primarily by *Escherichia coli*), it was not highly expressed under those conditions. Related observations were made previously using a mouse model of colitis monocolonized with commensal *E. coli* (Patwa *et al*, 2011). There, microarray-based measurements found a number of enzymes in the galactonate utilization pathway, including galactonate dehydratase, to be among the most strongly down-regulated in comparison with wild-type mice. These results suggest that galactonate metabolism is either infeasible (e.g. due to low bioavailability) or otherwise suboptimal (e.g. due to the presence of preferred energy sources) in the inflamed gut, thus leading to its down-regulation by "generalist" pathobionts like *E. coli*.

From the many over-expressed functions in active CD, we focused for illustrative purposes on examples that were encoded non-trivially by species either new or newly classified in MetaPhlAn 3 ("3.0-new species"; **Fig. 1.2.3C**). To aid in this process, we defined an *h*-index-inspired "novelty" score (*s*) for each EC equal to the largest percentile *p* of samples with the EC in which *p* percent of its copies were contributed by 3.0-new species. For example, an EC with *s*=0.25 indicates that at least 25% of the EC's copies were from 3.0-new species in at least 25% of samples with the EC. The previously mentioned galactonate dehydratase thus had a low novelty score (s=0.11) resulting from dominant contributions of *F. prausnitzii* and *E. coli* (which are not new to MetaPhlAn 3).

Conversely, the highest novelty score was observed for glutamyl-tRNA reductase (EC 1.2.1.70, s=0.46), a highly-transcribed housekeeping gene that received large contributions from the 3.0-new species *Roseburia faecis*, *Phascolarctobacterium faecium*, and *Ruminococcus bicirculans*. Betaine reductase (EC 1.21.4.4, s=0.43), conversely, is much more specific and was contributed in part by 3.0-new species *Hungatella hathewayi*; this is notable as a rare example of a function that was often detectable from community RNA but not DNA (indicating high expression from a small pool of gene copies). Pyruvate, water dikinase (EC: 2.7.9.2) and Ribonuclease E (EC 3.1.26.12) were among the strongest signals of over-expression in active CD by both effect size and statistical significance; these functions were also characterized by large contributions of 3.0-new species (s=0.36 and 0.38, respectively). Ribonuclease E and a final example, hydrogen dehydrogenase NADP(+) (EC 1.12.1.3), are also representative of the degree to which metagenomic copy number (DNA abundance) tends to be a strong driver of transcription (RNA abundance) in the gut microbiome, and thus the need to account for the former when estimating functional activity. The 3.0-new *H. hathewayi* expresses this enzyme highly in a subset of active CD samples, thus contributing to the enzyme's overall association with active CD.

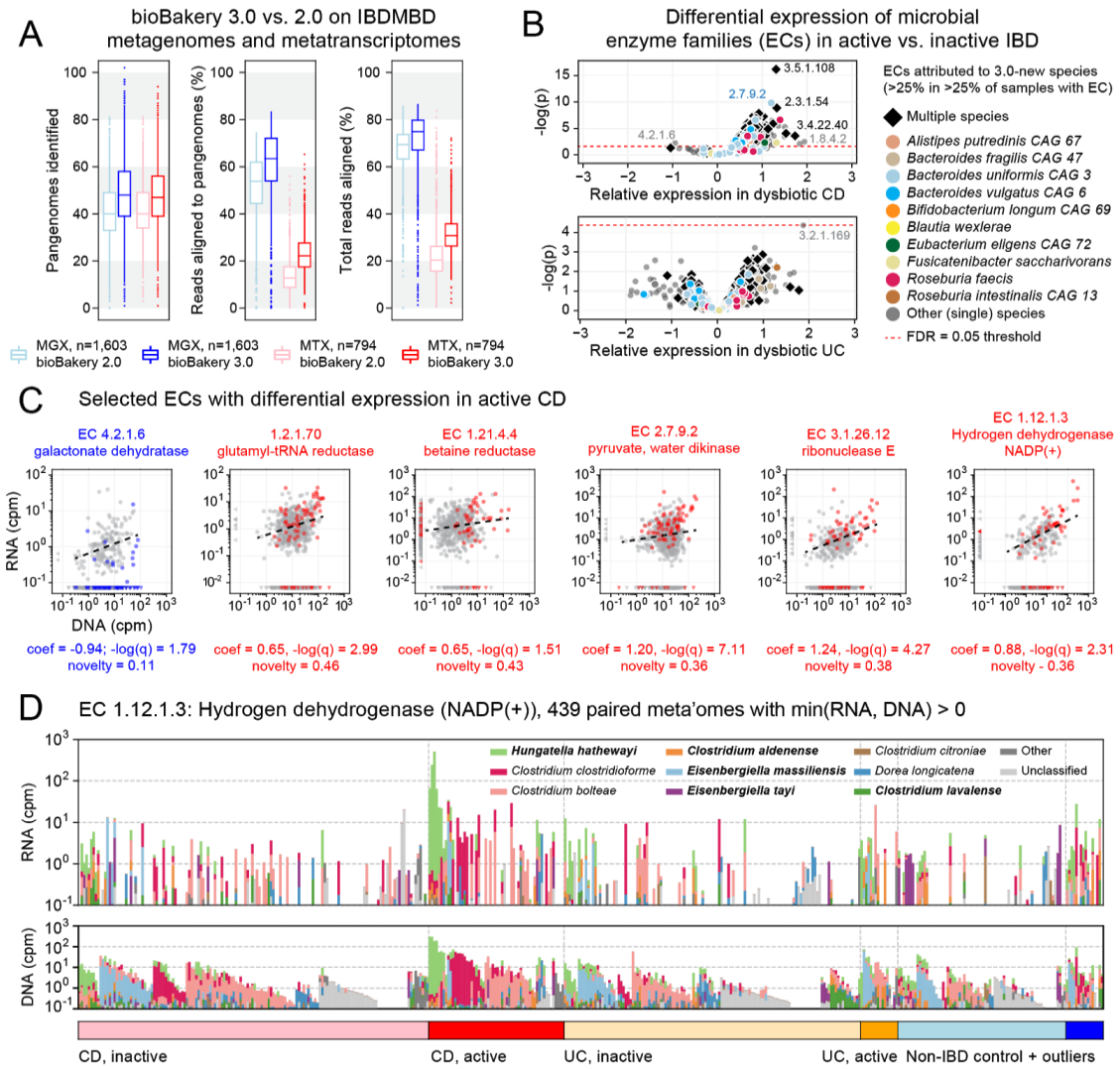


Figure 1.2.3: Longitudinal taxonomic and functional meta-omics of IBD. (A) Comparison of MetaPhlAn and HUMAnN profiles of IBDMDB metagenomes and metatranscriptomes using v2 and v3 software (sequencing data and v2 profiles downloaded from http://ibdmdb.org). **(B)** >500 Enzyme Commission (EC) families were significantly [linear mixed-effects (LME) models, FDR q<0.05] differentially expressed in active CD relative to inactive CD; only a single EC met this threshold for active UC. ECs (points) are colored to highlight large contributions from one or more species that were new or newly classified in MetaPhlAn 3 (independent of the strength of their association with active IBD). **(C)** Selected examples of EC families that were differentially expressed in active CD. Colored points correspond to active CD samples; all other samples are gray. The first example (blue) is the only EC to be down-regulated in active CD (as indicated by CD active samples falling below the best-fit RNA vs. DNA line). To match the associated LME models (see **Methods**), best-fit lines exclude samples where an EC's RNA or DNA abundance was zero (such samples are shown as triangles in the x:y margins). **(D)** Species contributions to RNA (top) and DNA (bottom) abundance of EC 1.12.1.3. The 7 strongest contributing species are colored individually; bold names indicate new species in MetaPhlAn 3. Samples are sorted according to the most abundant contributor and then grouped by diagnosis. The tops of the stacked bars (representing community total abundance) follow the logarithmic scale of the y-axis; species' contributions are linearly scaled within that height.

Population-scale subspecies genetics (StrainPhlAn) and pangenomics (PanPhlAn) of *Ruminococcus bromii*

Strain-level characterization of taxa directly from metagenomes is an effective cultivation-free means to profile the population structure of a microbial species across geography or other conditions (Truong *et al*, 2017; Scholz *et al*, 2016) and to track strain transmission (Ferretti *et al*, 2018). These functionalities are incorporated into (i) StrainPhlAn 3, which infers strain-level genotypes by reconstructing sample-specific consensus sequences from MetaPhlAn 3 markers (Zolfo *et al*, 2019) (ii) PanPhlAn 3, which identifies strain-specific combination of genes from species' pangenomes; and (iii) PhyloPhlAn 3, which performs precise phylogenetic placement of isolate and metagenome-assembled genomes (MAGs) using global and species-specific core genes (Asnicar *et al*, 2020) (see **Methods**). ChocoPhlAn 3 automatically quantifies and annotates the distinct types of conservation metrics necessary to identify these markers, all updated in bioBakery 3 (**Table S1.2.2**).

Ruminococcus bromii is a common gut microbe that is surprisingly understudied due to its fastidious anaerobicity and general non-pathogenicity (Ze *et al*, 2012) but it is prevalent in over half of typical gut microbiomes. This made its population genetics, geographic association, and genomic variability of particular interest to assess via StrainPhlAn and PanPhlAn. From the meta-analysis of 7,783 gut metagenomes integrated for a previous study (Pasolli *et al*, 2019), we considered the 4,077 metagenomes in which *R. bromii* was found present with a relative abundance above 0.05% according to MetaPhlAn 3. StrainPhlAn SNV-based analysis of the 124 *R. bromii*-specific marker genes across the 3,375 samples with sufficient markers' coverage (see **Methods**) revealed a complex population structure not previously recapitulated by the only fifteen genomes available from isolate sequencing (**Fig. 1.2.4A**). Sub-clade prediction (see **Methods**) highlighted two sub-species clades that are particularly divergent within the phylogeny (**Fig. S1.2.11C-D**); interestingly, the first one (Cluster 1) is mainly composed of strains retrieved from Chinese subjects and from cohorts with a non-Westernized lifestyle (**Fig. 1.2.4A**; Cluster 1). StrainPhlAn 3 can thus rapidly reconstruct complex strain-level phylogenies from metagenomes (5,700 seconds using 20 CPUs), and with the integration of PhyloPhlAn 3's improvements specifically for strain-level manipulation of alignments and phylogenies (Asnicar *et al*, 2020), surpasses the previous version of the software in accuracy and in sensitivity (67.4% more strain profiled, **Fig. S1.2.11A-B**).

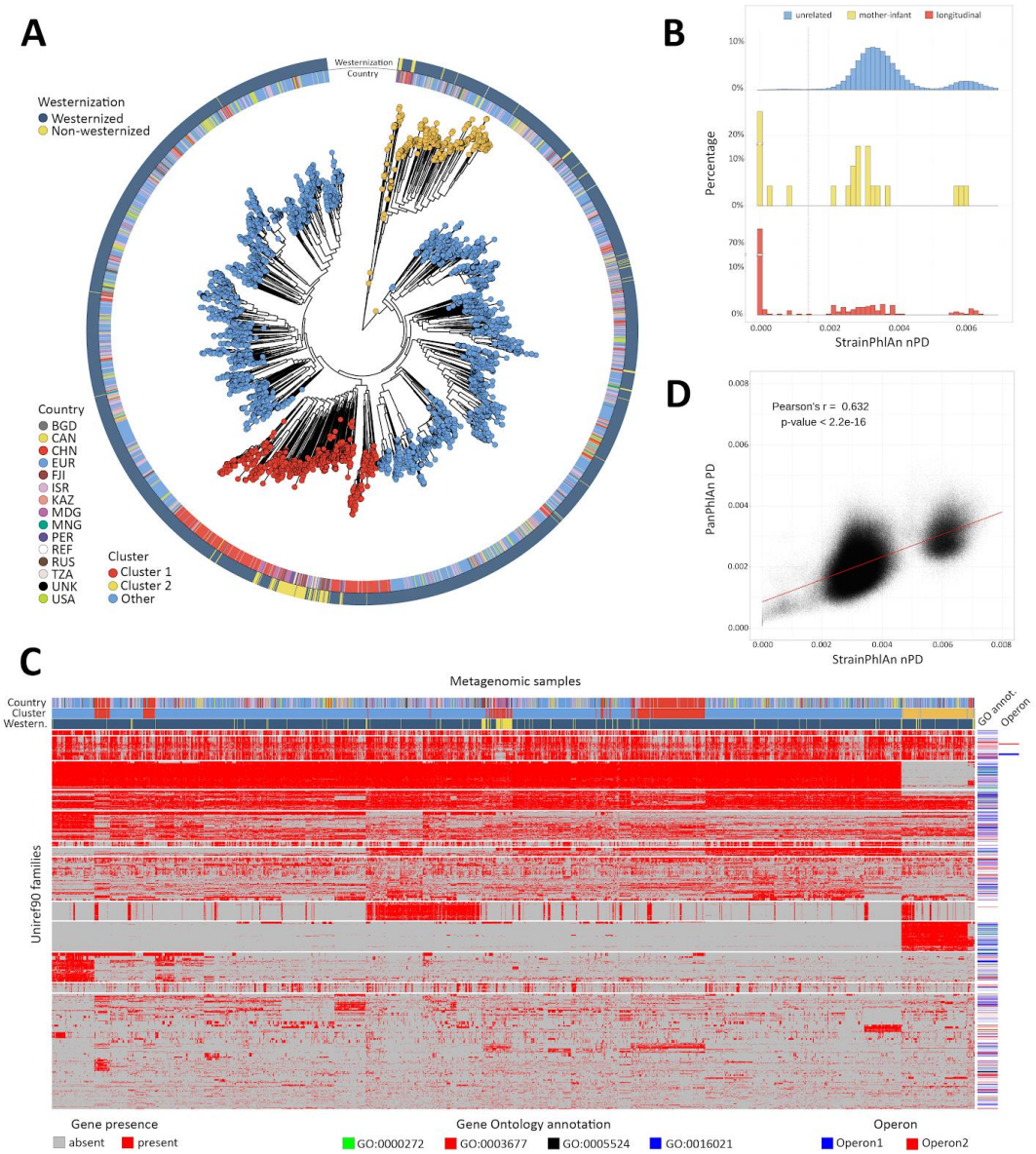


Figure 1.2.4: Population-scale strain-level phylogenetic and pangenomic analyses of *Ruminococcus bromii* from over 4,000 human gut metagenomes. (A) StrainPhlAn 3 profiling revealed stratification of *Ruminococcus bromii* clades with genetic content and variants frequently structured with respect to geographic origin and lifestyle. Genetically divergent subclades were identified, labeled as “Cluster 1” (mainly composed of strains retrieved from Chinese subjects) and a subspecies-like Cluster 2. **(B)** Strain tracking of *R. bromii*. While unrelated individuals from diverse populations very rarely share highly genetically similar strains, pairs of related strains are readily detected by StrainPhlAn from longitudinal samples from the same individuals (quantifying short- and medium-term strain retation at about 75%) and in mother-infant pairs (confirming this species is at least partially vertically transmitted). Normalized phylogenetic distances (nPD) were computed on the StrainPhlAn tree. **(C)** PanPhlAn 3 gene profiles of *R. bromii* strains from metagenomes highlights the variability and the structure of the accessory genes across datasets (core genes were removed for clarity). A total of 6,151 UniRef90 gene families from the *R. bromii* pangenome were detected across the 2,679 of the 4,077 samples in which a strain of this species was present at a sufficient abundance to be profiled by PanPhlAn. The 13 highest-rooted gene clusters are shown, highlighting co-occurrence of blocks likely to be functionally related. The most common GO annotations are also reported together with two operons containing genes verified to be on the same locus by analysis of the reference genomes in the PanPhlAn 3 database. **(D)** Genetic (SNV on marker genes from StrainPhlAn 3) and genomic (gene presence/absence from PanPhlAn 3) distances between *R. bromii* strains are correlated (Pearson's r=0.632, p-value<2.2e-16) pointing at generally consistent functional divergence in this species.

StrainPhlAn 3 also extends the ability of reference-based approaches to infer the genetic identity of strains across samples as previously explored (Truong *et al*, 2017; Ferretti *et al*, 2018). Specifically for *R. bromii*, different individuals tend to carry different strains diverged with a roughly normal distribution of genetic identities (mean 3.54e-3 normalized phylogenetic distance, **Fig. 1.2.4B**). However, the genetic differences between Cluster 1 and Cluster 2 were generally greater, with a lower peak and higher distances (mean 6.1e-3, **Fig. 1.2.4B**). For carriers of either clade, within-subject strain retention tended to be high as expected (i.e. low divergence); at distinct time points (average 261.35 s.d. 239.86 days, first quartile 72 days, third quartile 386 days, 3,537 comparisons in total), most of the strain distances (76.4%) approached zero (compared to 1% of comparisons for inter-individual differences, **Fig. 1.2.4B**). In addition to detecting these two genetically distinct clades and quantifying within-individual strain retention, a final distribution of higher intra-individual distances clearly captured (rare) strain replacement by *R. bromii* strains (i) in the same or (ii) in a different main cluster in the species' phylogeny. Mother-infant pairs showed a similar dynamic (**Fig. 1.2.4B**), with sporadic vertical transmission (~33.3%) (Ferretti *et al*, 2018; Korpela *et al*, 2018; Yassour *et al*, 2018) mixed with strain loss, replacement, and acquisition from other environmental or human sources (Korpela *et al*, 2018). This analysis highlighted the high precision of StrainPhlAn 3 in detecting strain identity across samples and thus the potential of using it for tracking the transmission network of specific individual strains within and between subjects.

PanPhlAn 3 provides a complementary form of strain analysis by constructing pangenome presence-absence (rather than individual nucleotide variant) genotypes (see **Methods**). Using 8 *R. bromii* reference genomes, PanPhlAn 3 revealed the presence of 6,151 UniRef90 pangenes across 2,679 samples with sufficient depth to permit confident strain-specific gene repertoire reconstruction (**Fig. 1.2.4C**). This mirrored the genetic divergence of *R. bromii* Clusters 1 and 2, while also highlighting a range of functional differences annotatable to genes unique to the two clusters: Cluster 1 and Cluster 2 showed a total of 797 and 601 UniRef90 families specific to them (Fisher's exact test, FDR < 0.05). Although most of these gene families do not have precise functional annotations, these sets of genes should be prioritized in experimental characterization efforts to unravel the sub-species diversity of *R. bromii*, and UnireRef90-to-GO ID mapping also highlighted an enrichment of membrane proteins in Cluster 2. Interestingly, other clusters of co-occurring genes were independent of phylogenetic structure and also verified to be on the same locus on at least two reference genomes in the PanPhlAn 3 database (**Fig. 1.2.4C**) providing a new approach at identifying and annotating potential laterally-mobile elements.

StrainPhlan 3 and PanPhlAn 3 can thus be combined with PhyloPhlAn 3 (Asnicar *et al*, 2020) and HUMAnN 3 to provide multiple, complementary, culture-independent means to investigate

the strain-level diversity of taxa in the microbiome, from new data or by re-using thousands of publicly available metagenomes. It is notable that these approaches tend to be consistent with each other (e.g. for *R. bromii,* Pearson's r=0.632, P<2.2e-16, **Fig. 1.2.4D**), while providing different benefits and drawbacks: PanPhlAn used with HUMAnN input is computationally efficient, used from whole pangenomes has higher sensitivity, and StrainPhlAn tends to have higher specificity. Together, the bioBakery 3 components provide an integrated platform for applying strain-level comparative genomics, taxonomic, and functional profiling to meta-omic microbial community studies.

Discussion

Here, we introduce and validate the set of expanded microbial community profiling methods making up the bioBakery 3 platform, including quality control (KneadData), taxonomic profiling (MetaPhlAn), strain profiling (StrainPhlAn and PanPhlAn), functional profiling (HUMAnN), and phylogenetics (PhyloPhlAn), largely relying on the underlying data resource of ChocoPhlAn 3 genomes and pangenomes. These modules are each more accurate and, often, more efficient than their previous versions and current alternatives, particularly for challenging (e.g. non-human-associated) metagenomes and for multi-omics (e.g. metatranscriptomes). In the process of these evaluations, we detected three species newly associated with CRC (*Dialister pneumosintes*, *Ruthenibacterium lactatiformans*, and *Eisenbergiella tayi*), over 500 enzyme families metatranscriptomically upregulated by diverse microbes in IBD, and two new phylogenetically, genomically, and biogeographically distinct subclades of *Ruminococcus bromii*.

These results highlight the degree to which meta-omic approaches can now realize the potential of culture-independent sequencing for characterizing microbial community dynamics, interactions, and evolution that are only active *in situ* and not *in vitro*. Since early studies of environmental and host-associated microbial communities (Venter *et al*, 2004; Tyson *et al*, 2004; Gill *et al*, 2006), it has been clear that many aspects of intercellular and inter-species signaling, short- and long-term evolution, and regulatory programs are exercised by microbes in their natural settings and extremely difficult to recapitulate in a controlled setting. This is supported by the extent to which “dark matter” not previously characterized in the laboratory pervades host-associated and (especially) environmental metagenomes (Parks *et al*, 2017), with most communities containing a plurality, majority, or sometimes supermajority of novel and/or uncharacterized sequences (Pasolli *et al*, 2019; Almeida *et al*, 2019). The bioBakery 3 begins to overcome this challenge by combining a greatly expanded set of reference sequences with ways of “falling back” gracefully when encountering new sequences, while also paving the way for further integration of assembly-based discovery in the future (discussed below). Critically, this now permits large collections of meta-omes to be used in ways only previously possible with large isolate genome or transcriptome collections, e.g. strain-level integrative comparative genomics, near-real-time epidemiology and evolution, and detailed gene content prediction and metabolic modeling. Results such as the heterogeneity of maternal-infant strain transmission and retention, or the globally stratified distribution of subspecies clades, would be extremely challenging to discover by other means.

Methodologically, it is notable that these new meta-omic analysis types have been enabled by several years of improved experimental fidelity, denoising, and quality control approaches.

These effectively retain only the “best” subset of reads from large, noisy meta-omes for each analysis of interest, e.g. only the most unique sequences for taxonomic identification, or only the most evolutionarily informative loci for phylogeny. Meta-omes are uniquely positioned for broad reuse and discovery since different “best” subsets of each dataset can be used to answer different questions. The development of meta-omic analysis methods thus parallels that of genome-wide association studies or transcriptomics, inasmuch as early methods were later refined to provide much greater accuracy and scalability through removal of low-quality measurements, within- and between-study normalization approaches, statistical methods to reliably separate signal from noise, and biological annotation of previously uncharacterized loci. Similarly, methods for amplicon-based community profiling have progressed from noise- and chimera-prone stitching and clustering to near-exact sequence variant tracking (Callahan *et al*, 2016). Fortunately, continued decreases in sequencing prices and increases in protocol efficiency have now made shotgun meta-omics nearly as affordable as amplicon sequencing in many settings. The challenge, of course, is that each metagenome combines many different noise sources: there is no single, whole genome to finish; host, microbial, and contaminant sequences are not always easily differentiated; there is no one set of “true” underlying variants (since each organism might be represented by multiple strains); and millions of microbial gene products remain functionally uncharacterized (Thomas & Segata, 2019).

Notably, the bioBakery provides one of very few environments currently capable of integrating both metagenomes and metatranscriptomes to begin overcoming these uncertainties (Franzosa *et al*, 2018). As introduced above, microbial community transcriptomes can be highly unintuitive to interpret, as transcript abundance is always influenced both by expression level and by underlying DNA copy number, i.e. abundance of the expressing taxon. Since both sequence-based DNA and RNA profiles are typically compositional (relative, not absolute, abundances), there is not always a simple way to account for these effects. HUMAnN 3 provides initial within- and between-species normalization options that can be combined with the statistical models of differential expression described here, making e.g. the >500 transcripts overexpressed in Crohn’s disease particularly noteworthy. *Hungatella hathewayi* was uniquely responsible for many of these, an organism not previously associated with IBD in humans (Schaubeck *et al*, 2016). While many of its overexpressed transcripts are core or housekeeping processes, indicative of general bioactivity in the inflamed gut (comparable to that of e.g. *Escherichia coli* (Lloyd-Price *et al*, 2019)), others such as betaine reductase are much more specific. This enzyme contributes directly to trimethylamine (TMA) formation (Rath *et al*, 2019), one of the more noteworthy microbial metabolites implicated in human disease via its transformation to proatherogenic trimethylamine-oxide (TMAO) (Tang *et al*, 2013). Conversely, the only transcript differentially regulated in ulcerative colitis, underexpressed *F.*

prausnitzii galactonate dehydratase, contrasts its utility in polysaccharide degradation under non-inflamed conditions with the upregulation of alternative, more host-antagonistic energy sources in E. coli during inflammation (Lloyd-Price *et al*, 2019). Both of these examples are only analyzable due to the highly specific assignment of meta-omic reads to individual community members' gene families, in combination with appropriate downstream statistical methods for multi-omics.

Finally, it is striking that metagenomically-derived comparative genomics have only recently been able to reach the scale and scope previously possible with microbial isolates. The genomic epidemiology of pathogens has driven the latter - recently in viral outbreaks such as COVID-19 (Lu *et al*, 2020) and Ebola (Gire *et al*, 2014), and in many bacterial conditions such as cholera (Weill *et al*, 2017) or pneumonia (Croucher *et al*, 2011). Since metagenomes can simultaneously access all community members with relatively little bias, such studies are now possible with organisms previously overlooked due to the absence of obvious associated phenotypes or convenient culture techniques (Pasolli *et al*, 2019; Manara *et al*, 2019). *Ruminococcus bromii* is one such example; despite being over 50% prevalent among typical human gut communities, only 15 isolates were previously sequenced, precluding any type of epidemiology or basic phylogenetics. In addition to making a novel sub-species phylogenetic and biogeographic structure apparent, the combination of MetaPhlAn, HUMAnN, PanPhlAn, StrainPhlAn, and PhyloPhlAn together confirmed that most strains are "personal" (i.e. specific of an individual) and transmissible across hosts, and that genomic differences characterize each subspecies suggesting multiple paths of functional adaptation and specialization.. Such results are in principle possible with any combination of metagenomic and isolate taxa and genes of interest, richly integrating culture-independent data with hundreds or thousands of isolate genomes.

Of course, many challenges remain both for improvement of the bioBakery platform and for the field as a whole. Both experimental and computational accessibility of non-bacterial microbial community members remains limited. While accurate, bioBakery 3's capacity for non-bacterial profiling is only slightly improved from the previous version by the expansion of available eukaryotic microbial reference sequences. These components of metagenomes - and, for RNA viruses, metatranscriptomes - are often measured with surprising heterogeneity during the initial generation of sequencing data themselves (Zolfo *et al*, 2019), suggesting necessary improvements in analytical quality control and normalization as well. The visibility of species with particularly high genetic diversity within individual communities also remains limited; in most cases, only the most dominant strain of each taxon per community is currently analyzable, again for both experimental (e.g. sequencing depth) and analytical reasons (Quince *et al*, 2017b). This is true both for reference-based and for assembly-based

approaches, the latter of which are often also stymied by highly diverse taxa (Pasolli *et al*, 2019). A final area of improvement for the bioBakery, relatedly, is the increased integration between reference-based and assembly-based approaches - begun here via PhyloPhlAn 3 - in order to better leverage MAGs (Almeida *et al*, 2020), SGBs (Pasolli *et al*, 2019), and novel gene families.

We thus anticipate improved integration of reference- and assembly-based meta-omic analyses to be one of the main areas of future development for the bioBakery, along with expanded methods for other types of multi-omics in addition to transcription. There will also be a continued focus on quality control and precision, enabling new types of functional analysis within microbial communities (e.g. bioactivity and gene function prediction) without sacrificing sensitivity to rare or novel community members. Finally, we are also committed to the platform's availability with well-documented, open-source implementations, training material, and pre-built locally-executable and cloud-deployable packaging. Feedback on any aspect of the methods or their applications in diverse host-associated or environmental microbiome settings can be submitted at https://forum.biobakery.org, and we hope the bioBakery will continue to provide a flexible, convenient, reproducible, and accurate discovery platform for microbial community biology.

Methods

The bioBakery 3 is a set of computational methods for the analysis of microbial communities from meta-omic data that produce taxonomic, functional, phylogenetic, and strain-level profiles to be interpreted directly or included in downstream statistical analyses (**Fig. 1.2.1A**). After read-level quality control by KneadData, MetaPhlAn 3 estimates the set of microbial species (and corresponding higher taxonomic clades) present in a sample and their relative abundances. StrainPhlAn 3 deepens genetic characterization by refining strain-level genotypes of species identified by MetaPhlAn 3. HUMAnN 3 focuses instead on the identification and quantification of the molecular functions encoded in the metagenome or expressed in the metatranscriptome, which can be resolved by PanPhlAn 3 into gene presence-absence strain-level genotypes. PhyloPhlAn 3, as previously reported (Asnicar *et al*, 2020), provides a comprehensive means to interpret the draft genomes produced by assembly-based metagenomic tools. These bioBakery 3 modules are generally based on an underlying dataset of functionally-annotated isolate microbial genes and genomes produced by ChocoPhlAn 3 to quality-control and annotated UniProt derivatives. This currently includes 99,227 genomes and 87.3M gene families, almost 100-fold greater than the data types included in the first bioBakery release (Segata & Huttenhower, 2011).

The AnADAMA scientific workflow manager

Most bioBakery 3 tools are integrated into reproducible workflows (the “bioBakery workflows”, http://huttenhower.sph.harvard.edu/biobakery_workflows) using the AnADAMA (Another Automated Data Analysis Management Application) task manager, currently v2 (http://huttenhower.sph.harvard.edu/anadama2). Briefly, this wraps doit (http://pydoit.org), a Python-based dependency manager, to provide a simple but scalable language for analysis task definition, version and provenance tracking, change management, documentation, grid and cloud deployment of large compute tasks, and automated reporting. AnADAMA operates in a make-like manner using targets and dependencies of each task to allow for parallelization. In cases where a workflow is modified or input files change, only those tasks impacted by the changes will be rerun. Essential information from all tasks is recorded, using the default logger and command line reporters, to ensure reproducibility. The information logged includes command line options provided to the workflow, the function or command executed for each data modification task, versions of tracked executables, and any output and data products from each task. It can optionally be used to chain together subsequent bioBakery 3 tasks and/or to parallelize them efficiently across multiple files or datasets.

KneadData read-level quality control

The bioBakery 3 includes a simple quality control module for raw sequences, KneadData (http://huttenhower.sph.harvard.edu/kneaddata), which automates a set of typical best practices for raw metagenome and metatranscriptome read cleaning and validation. These include:

- Trimming of 1) low-quality bases (default: 4-mer windows with mean Phred quality <20), 2) truncated reads (default: <50% of pre-trimmed length), and 3) adapter and barcode contaminants using Trimmomatic (Bolger *et al*, 2014).
- Removal of overrepresented sequences (default: > 0.1% frequency) using FastQC (Andrews & Others, 2010) and low-complexity sequences using TRF (Benson, 1999).
- Depletion of host-derived sequences by mapping with bowtie2 (Langmead & Salzberg, 2012) against an expanded human reference genome (including known “decoy” and contaminant sequences (Breitwieser *et al*, 2019b)) and optionally other host (e.g. mouse) reference genomes and/or transcriptomes.
- Depletion of microbial ribosomal and structural RNAs by mapping against SILVA (Yilmaz *et al*, 2014) in metatranscriptomes.

It is recommended that KneadData be applied to raw sequences prior to further analyses, and the bioBakery workflows do this for all sequence types by default.

The ChocoPhlAn 3 pipeline

We developed the ChocoPhlAn pipeline to organize microbial reference genomes according to their taxonomy and to compute the relevant sequence and annotation data for subsequent bioBakery modules. At a high level, after retrieval of UniProt genomes and gene annotations, species-specific pangenomes (i.e. the set of gene families of a species present in at least one of its genomes) are generated using all the microbial reference genomes passing initial quality control. Core genomes (i.e. gene families present in all the genomes of a species) are then identified from the whole set of pangenomes and used as markers in PhyloPhlAn 3. Core genomes are also processed for the extraction of unique marker genes (i.e. core gene families uniquely associated with one species) that constitute the marker database for MetaPhlAn 3 and StrainPhlAn 3. Finally, functionally annotated pangenomes are processed to serve as references for PanPhlAn 3 and HUMAnN 3.

Data retrieval

ChocoPhlAn relies on the UniProt core data resources (UniProt Consortium, 2019) (release January 2019) and on the NCBI taxonomy and genomes repositories (NCBI Resource Coordinators, 2014) (release January 2019). The two basic sequence data types considered in ChocoPhlAn are the raw genomes of all available microbes and all the microbial proteins/genes identified on these genomes. The main supporting structure for a genome is the underlying microbial taxonomy, whereas the microbial proteins are organized in protein families clustered at multiple stringency parameters.

We adopted the NCBI taxonomy database (NCBI Resource Coordinators, 2014) for use by ChocoPhlAn as it is the one on which our genomic repository, UniProt, is also based. The full taxonomy was downloaded from the NCBI FTP server (ftp.ncbi.nlm.nih.gov/pub/taxonomy/) on January 24 2019. We identified and tagged species with “unidentified”, “sp.”, “Candidatus”, “ bacterium “, and several other keywords as low-quality species. Specifically, the regular expressions used to filter low-quality taxonomic annotations are:

“ (C|c)andidat(e|us) | _sp(_.*|$) | (.*_|^)(b|B)acterium(_.*|) | .*(eury|)archaeo(n_|te|n$).* | .*(endo|)symbiont.* | .*genomosp_.* | .*unidentified.* | .*_bacteria_.* | .*_taxon_.* | .*_et_al_.* | .*_and_.* | .*(cyano|proteo|actino)bacterium_.*) ”

All reference genomes available through UniProt Proteomes and linked to the public DDBJ, ENA, and GenBank repositories were then considered. Genomes are included by UniProt into UniProt Proteomes only if they are fully annotated and have a number of predicted CDSs falling within a statistically defined range of published proteomes from neighbouring species (What are proteomes?). We considered all UniProt Proteomes genomes assigned to the archaeal and bacterial domain. For micro-eukaryotes, we considered all genomes assigned to the following manually selected genera: *Blastocystis, Candida, Saccharomyces, Cryptosporidium, Entamoeba, Aspergillus, Cryptococcus, Cyclospora, Cystoisospora, Giardia, Leishmania, Malassezia, Neosartorya, Pneumocystis, Toxoplasma, Trachipleistophora, Trichinella, Trichomonas,* and *Trypanosoma*.

Reference genomes (‘fasta’ format, suffix ‘.fna’) and the associated genomic annotation (‘.gff’) of each proteome were downloaded from the NCBI GenBank FTP server (ftp.ncbi.nlm.nih.gov/genomes/all/GCA) by retrieving URLs from the assembly_summary_genbank.txt file (ftp.ncbi.nlm.nih.gov/genomes/genbank/assembly_summary_genbank.txt) using the GCA accession included in the UniProt Proteomes resource (01/24/2019). Starting from a total of 111,825 UniProt Proteomes entries, we discarded 12,598 proteomes missing the GenBank accession, ending up with 99,227 genomes (997 Archaea, 97,941 Bacteria, 339 Eukaryota).

The microbial proteins (and genes) associated to at least one UniProt Proteome and considered by ChocoPhlAn are retrieved from the UniProt Knowledgebase (UniProtKB) and the UniProt Archive (UniParc) databases. Proteins included in UniProtKB have been derived from the translation of the CDSs of all available reference genomes included in UniProt Proteomes. ChocoPhlan 3 also retrieves and includes relevant data present in the UniProtKB entries (retrieved from ftp.uniprot.org/pub/databases/uniprot/ as XML files uniprot_sprot.xml.gz, uniprot_trembl.xml.gz, uniparc_all.xml.gz) such as functional, phylogenomic, and protein domain annotations (KEGG, KO, EggNOG, GO, EC, Pfam) (Kanehisa & Goto, 2000; Huerta-Cepas *et al*, 2016b; The Gene Ontology Consortium, 2019; El-Gebali *et al*, 2019), accessions for cross-referencing entries with external databases (GenBank, ENA, and BioCyc) (Clark *et al*, 2016; Leinonen *et al*, 2011; Karp *et al*, 2019), name of the gene that encodes for the protein, and proteome accession.

We processed a total of 203.9M proteins included in both UniProtKB and UniParc, and 126.9M of them were associated with a UniProt Proteome entry. The Bacteria domain tallied the highest number of proteins (194.8M), whereas Archaea and Eukaryotes accounted for 5.0M and 4.0M proteins respectively.

In order to reduce the redundancy of the database, we use the UniRef90 clustering of UniProtKB proteins provided by UniProt. In brief, UniProtKB are clustered at different thresholds of sequence identity (100, 90, 50) and made available through the UniProt Reference Clusters (UniRef) resource (Suzek *et al*, 2015). UniRef90 clusters are generated by clustering unique sequences (UniRef100, which combines identical UniProtKB proteins in a single cluster) via CD-HIT (Li & Godzik, 2006) until August 2019, and via MMseqs2 (Steinegger & Söding, 2018) afterward. Sequences in UniRef90 clusters have at least 90% sequence identity (Suzek *et al*, 2015). UniRef50 clusters are generated by clustering the UniRef90 cluster seed sequences, and each cluster contains proteins with at least 50% identity. Both UniRef90 and UniRef50 require each protein to overlap at least 80% with the cluster's longest sequence. UniRef entries considered in ChocoPhlAn 3 contain the sequence of a representative protein, the accession IDs of all the entries included in the cluster, the accessions to the UniProtKB and UniParc records, and the accessions of the other associated UniRef cluster are included in the UniProt entries.

A total of 292.1M UniRef clusters were processed (172.3M, 87.3M, and 32.5M for UniRef100, UniRef90, and UniRef50, respectively) and associated to each protein and each genome in ChocoPhlAn 3.

Pan-proteome generation

We then generate pan-proteomes for each species represented at least by one UniProt Proteome. We define a species' pan-proteome as the non-redundant representation of the species' protein-coding potential. These are obtained for each species by considering the unique UniRef90 and UniRef50 protein families present in the genomes assigned at the species level and below.

For each pan-protein, we compute several scores. We define a 'coreness' score for a UniRef90 family as the number of genomes included in the species' pan-proteome having a protein belonging to the UniRef family, and the 'uniqueness' score as the number of pan-proteomes of other species possessing the same pan-protein. We then also considered a 'uniqueness_sp' score, a variant of the 'uniqueness' score obtained excluding those species that were previously tagged as low-quality species. Alongside the 'uniqueness' score, we compute the 'external_genomes' as the number of genomes (rather than species or species' pan-proteomes) of other species' pan-proteomes possessing the same pan-protein. These scores were computed for both UniRef50 and UniRef90 protein families.

In ChocoPhlAn 3 we consider a total of 22,096 species' pan-proteomes and a total of 87.3M UniRef90 core proteins (i.e. with coreness > 0.7, avg. 3,952 s.d. 6,311 per species).

Generation of MetaPhlAn 3 markers

MetaPhlAn relies on a set of unique and species-specific nucleotide markers that were updated in MetaPhlAn 3 starting from the ChocoPhlAn 3 pan-proteomes. We initially filtered out species having taxonomies previously tagged as low quality using the species-level genome bin (SGB) system (Pasolli *et al*, 2019). "Low-quality" species that were assigned to a SGB the same SGB representative were merged and only the SGB representative was taken into account.

This merging procedure occurred for a total of 1,328 species (6%) that were merged as they were unlikely to be distinguishable in metagenomic samples and would potentially lead to false-positive taxonomic assignments (see **Table S1.2.7** for the merged species). For the cases in which multiple species included by the NCBI taxonomy into a "species-group" showed a high number of markers with a high 'uniqueness' score (>30), we proceeded to identify unique markers for the whole species groups. This occurred for the following species groups: *Streptococcus anginosus* group, *Lactobacillus casei* group, *Bacillus subtilis* group, *Enterobacter cloacae* complex, *Pseudomonas syringae* group, *Pseudomonas stutzeri* group, *Pseudomonas putida* group, *Pseudomonas fluorescens* group, *Pseudomonas aeruginosa*

group, *Streptococcus dysgalactiae* group, and *Bacillus cereus* group. In all these cases, the pangenomes were built by merging all the species-level pangenomes and treating them as a single species.

In the first step of the marker discovery procedure, we use the pan-proteome built using the UniRef90 clusters considering all proteins with a length between 150 and 1,500 amino acids. Starting from the coreness and uniqueness scores, we applied an iterative approach in order to find up to 150 unique markers whenever possible and retaining only those species with a minimum of 10 unique markers. We classify candidate markers into unique and quasi-markers according to the 'uniqueness' value: markers having zero 'uniqueness' are reported as 'unique markers'. When no unique markers can be identified, the less-stringent thresholds used in the marker discovery procedure allows the identification of the so-called 'quasi-markers', markers having non-null values of 'uniqueness'.

The iterative approach started with the definition of four tiers of unique markers according to a combination of the values of 'coreness', 'uniqueness', and 'external_genomes'. Tier 'A' includes pan-proteins with a coreness score higher than 80%, not shared with more than 2 other pan-proteomes considering both UniRef90 and UniRef50 clustering score ('Uniqueness_NR90' and 'Uniqueness_NR50'), and not present in more than 10 single genomes when considering the UniRef90 and 5 single genomes when considering UniRef50 ('External_genomes_NR90' and 'External_genomes_NR50'), respectively. Tier 'B' includes markers with 'coreness' values between 70% and 80%, 'Uniqueness_NR90', and 'Uniqueness_NR50' values of 5, and values of 'External_genomes_NR90' and 'External_genomes_NR50' lower than 15 and 10 genomes, respectively. Markers that did not meet the previous criteria were included in the 'C' tier, which includes markers with 'coreness' values between 50% and 70%, 'Uniqueness_NR90' less than 10, 'Uniqueness_NR50' less than 15, 'External_genomes_NR90' less than 25, and 'External_genomes_NR50' less than 20. Markers for the species having only one genome included in the pan-proteome, for which the definition of coreness is trivial, were classified as tier 'U', provided that they have zero 'Uniqueness'.

The definition of specific tiers allows the retrieval of the maximum number of unique markers. Marker discovery procedure was performed iteratively for each tier. Candidate markers that meet the tier-defined thresholds were ranked using a score function defined as follows:

$$Score = S_{coreness} * S_{uniqueness50} * S_{uniqueness90}$$

Where

$$S_{coreness} = \sqrt{coreness_{\%}}$$

$$S_{uniqueness90} = -log(1 - \frac{10^4 - min(10^4, uniqueness_{90})}{10^4 - 10^{-4}}) * \frac{1}{5}$$

$$S_{uniqueness50} = -log(1 - \frac{10^4 - min(10^4, uniqueness_{50})}{10^4 - 10^{-4}}) * \frac{1}{5}$$

The score function as defined will prioritize the selection of candidate markers highly conserved in the clade (high 'coreness' value) but shared with the smallest possible number of other species (low values of 'uniqueness'). Tier type is assigned to each candidate marker, and if more than 50 candidate markers were identified, we selected up to 150 markers from the ranked list. If not enough markers were identified (less than 50), the procedure was repeated using the subsequent tier's thresholds. If no markers were identified using tier C thresholds, the species was discarded.

Nucleotide sequences for each marker selected with this procedure are then considered as entries for the MetaPhlAn database. To refine the number of species estimated by the 'uniqueness' parameter, marker sequences were split into non-overlapping chunks of 150bp and mapped versus an index built using all the reference genomes used for the marker identification process using bowtie2 (version 2.3.4.3, parameters '-a --very-sensitive --no-unal --no-hq --no-sq'). We accounted for a newly identified species based on the 'uniqueness' parameter if at least 150 consecutive nucleotides of the marker sequence were found in the identified target reference genome.

We performed an additional step of curation for markers for species with genomes obtained with Co-Abundance gene Groups (CAGs) (Nielsen *et al*, 2014). To reduce the number of false positives, we removed the CAG species if more than 50% of its markers were shared with the species that gave the taxonomy to the CAG genome.

Each marker has associated an entry in the MetaPhlAn database which includes the species for which the sequence is a marker, the list of species sharing the marker, the sequence length, and the taxonomy of the species. Viral markers were taken from the v20_m200 MetaPhlAn2 database.

Altogether, this identified a total of 1.1M markers for 13,475 species (**Table S1.2.8**).

MetaPhlAn 3 taxonomic profiling

The raw reads in a metagenomic sample are mapped inside MetaPhlAn 3 to the database of 1.1M markers using bowtie2 (Langmead & Salzberg, 2012). The default bowtie2 mapping parameters are those of the 'very-sensitive' preset but are customizable via the MetaPhlAn 3 settings. In MetaPhlAn 3 the input can be provided as a single fastq file (compressed with several algorithms), multiple fastq files archived in a single file, or as a pre-performed mapping.

Internally, MetaPhlAn 3 estimates the coverage of each marker and computes the clade's coverage as the robust average of the coverage across the markers of the same clade. The clade's coverages are then normalized across all detected clades to obtain the relative abundance of each taxa as previously described (Truong *et al*, 2015; Segata *et al*, 2012b). In version 3, we further optimized the parameter of the robust average which excludes the top and bottom quantiles of the marker abundances ("stat_q" parameter) which is now set by default to 0.2 (i.e. excludes the 20% of markers with the highest abundance as well as the 20% of markers with the lowest abundance). To further improve the quality of the read-mapping we adopted further controls before and after mapping by discarding low-quality sequences and alignments (reads shorter than 70bp and alignment with a MAPQ value less than 5 are discarded). We also introduced a new feature for estimating the portion of the metagenome that would map against taxa not present in current databases; this estimation is computed by subtracting from the total number of reads the average read depth of each taxa normalized by the taxon-specific average genome length. Additionally, the new output format for MetaPhlAn 3 includes by default the NCBI taxonomy ID of each profiled clade allowing for better comparisons between tools and tracking of the species name in case of taxonomic reassignment. Alongside the default MetaPhlAn output format, profiles can be now reported using the CAMI output format defined by (Belmann *et al*, 2015; BioBoxes) that can be used for performing benchmarks with the OPAL framework (Meyer *et al*, 2019). To support post-profiling analyses, a convenience R script for computing Weighted and Unweighted UniFrac distances (Lozupone & Knight, 2005) from MetaPhlAn profiles is now available in the software repository (metaphlan/utils/calculate_unifrac.R), alongside the phylogeny (in Newick format) comprising all MetaPhlAn 3 taxa. The improvements and addition in MetaPhlAn 3 compared to the previous MetaPhlAn 2 version are summarized in **Supplementary Table 1.2.2**.

StrainPhlAn 3 strain profiling

StrainPhlAn performs genotyping at the strain level by reconstructing sample-specific consensus sequences of MetaPhlAn markers and using them for multiple-sequence alignment and phylogenetic modeling (Truong *et al*, 2017). StrainPhlAn 3 improves the original implementation in several aspects: (i) the integration of an improved and validated pipeline for consensus sequence generation (Zolfo *et al*, 2019), (ii) the integration of PhyloPhlAn 3 (Asnicar *et al*, 2020) which improves the quality of the phylogenetic modeling and the flexibility of the analysis, and (iii) a refined algorithm for filtering samples not supported by enough species' markers and markers not enough conserved across strains and samples.

StrainPhlAn 3 takes as input the alignment results from the MetaPhlAn 3 profiling (i.e. the mapping of the metagenomic samples against the MetaPhlAn species-specific markers) as

well as the MetaPhlAn 3 markers' database. For each sample, StrainPhlAn 3 reconstructs high-quality consensus sequences of the species-specific markers by considering, at each position of the marker, the nucleotide with the highest frequency among the reads mapping against the marker and covering that position. By default, consensus markers reconstructed with less than 8 reads or with a breadth of coverage (i.e. fraction of the marker covered by reads) lower than 80% are discarded ("--breadth_threshold" parameter). Ambiguous bases are defined as positions in the alignment with quality lower than 30 or high polymorphisms (major allele dominance lower than 80%) and are considered for the threshold on the breadth of coverage as unmapped positions.

After marker reconstruction, the filtering algorithm discards samples with less than 20 markers, as well as markers present in less than 80% of the samples ("--sample_with_n_markers" and "--marker_in_n_samples" parameters, respectively). Then, markers are trimmed by removing the leading and trailing 50 bases ("--trim_sequences" parameter), since these are usually supported by lower coverage due to the boundary effect during mapping, and a polymorphic rates report is generated for optional inspection by the user. Finally, filtered samples and markers are processed by PhyloPhlAn 3 for phylogenetic reconstruction. By default, reconstructed sequences are mapped against the markers database using BLASTn (Altschul *et al*, 1990), multiple sequence alignment is performed by MAFFT (Katoh & Standley, 2013) and phylogenetic trees are produced by RAxML (Stamatakis, 2014) Due to the reconstruction of a strain-level phylogeny, PhyloPhlAn was set to run with "--diversity low" parameter.

Phylogenetic trees produced by StrainPhlan 3 can also be used to identify identical strains across samples, which can be exploited, for example, in strain transmission analyses (Ferretti *et al*, 2018; Shao *et al*, 2019). This is now supported by the newly-added "strain_transmission.py" script. This script processes the phylogenetic tree produced by StrainPhAn together with metadata describing relations between the samples (e.g. longitudinal samples or samples with a relation of interest such as mother/infant pairings) to infer strain transmission events. First, using the phylogenetic tree, a pairwise distance matrix is generated and normalized by the total branch length of the tree. Using the distance matrix and the associated metadata, a threshold defining identical strains is inferred selecting the first percentile of the non-related-samples distances distribution (i.e. setting an upper bound on the theoretical false-discovery rate at 1%). If longitudinal samples are provided, only one is considered per subject, and samples not included in the metadata are considered as non-related. Finally, related sample pairs with a distance smaller than the inferred threshold are reported as potential transmission events.

HUMAnN 3 data and algorithm updates

Functional potential profiling of microbial communities is performed by HUMAnN using pangenomes annotated with UniRef90 on all species detectable per sample with MetaPhlAn. ChocoPhlAn pangenomes used by HUMAnN for functional profiling are directly available as the species pan-proteomes annotated with the UniRef90 clusters. To obtain a nucleotide representation of each pan-proteome, we identified a representative of the cluster for each pan-protein by selecting a UniProtKB or UniParc entry taxonomically assigned to the desired species. Each cluster representative was used for extracting the nucleotide sequence from the source reference genome and the several functional annotations from different systems (GO terms (Ashburner *et al*, 2000), KEGG modules (Kanehisa *et al*, 2014), KO identifiers, Pfam accessions (Finn *et al*, 2014), EC numbers (Bairoch, 2000), and eggNOG accessions (Powell *et al*, 2014)) associated with the UniProtKB entry. Alongside the functional annotations, we associated each UniRef90 cluster with its corresponding UniRef50 cluster in order to provide multiple levels of functional resolution.

HUMAnN 3 implements a number of new options for fine-tuning the steps in its tiered search (e.g. passing custom search parameters to bowtie2 (Langmead & Salzberg, 2012) and DIAMOND (Buchfink *et al*, 2015) in the pangenome and translated search steps, respectively). We performed a round of additional accuracy and performance tuning on these new parameters prior to the main evaluations of the paper. To minimize overfitting potential, we conducted initial tuning of HUMAnN 3 on the above-described human-like synthetic metagenome, which featured a structure and species composition that were distinct from those of the CAMI and nonhuman synthetic metagenomes used in downstream inter-method comparisons (**Fig. 1.2.1**).

We first considered two new options when assigning reads to species pangenomes: 1) requiring pangene sequences to be covered above a threshold fraction of sites before any alignments to those sequences were accepted (“database sequence coverage filtering”) and 2) allowing a read to align to multiple pangenes instead of the single target favored by bowtie2’s default settings (as used in HUMAnN 2). Coverage filtering (new option 1) was already implemented in HUMAnN 2 for post-processing translated search results, where it was shown to increase UniRef90-level specificity considerably at a small cost to sensitivity (Franzosa *et al*, 2018). We observed similar results here in the context of pangenome search; as a result, HUMAnN 3 now imposes (separately tunable) database-sequence coverage filters during its pangenome and translated search steps (both default to 50%; **Fig. S1.2.4**). Conversely, allowing a read to hit up to 5 pangenes (new option 2, as implemented via

bowtie2's "-k 5" setting) had very little impact on accuracy and is not enabled by default in HUMAnN 3.

We additionally considered new options to tune the stringency and memory usage of DIAMOND 0.9 during translated search. The most impactful of these was reducing the identity threshold for per-read alignment to UniRef90 from 90% (the HUMAnN 2 default) to 80% (the new default for HUMAnN 3; **Fig. S1.2.4**). While the former value was chosen to respect the average identity among UniRef90 family members, the 80% threshold is more forgiving of variation within read-length windows of a protein-level UniRef90 alignment. Coupled with HUMAnN's database sequence coverage filter, the 80% threshold correctly aligns considerably more reads during translated search without compromising specificity.

While HUMAnN 2 accepted DIAMOND's (default) top-20 database targets per query read, we newly evaluated the top 1 and top 5 targets, as well as any targets within 1, 2, or 10% of the best hit's score. We selected the "within 1% score of the best hit" filter (DIAMOND's "--top 1" option) as a new default for HUMAnN 3 on the basis of a marked increase in UniRef90 specificity with minimal loss of sensitivity. Finally, we explored tuning DIAMOND's memory via the "--block-size (-b)" and "--index-chunks (-c)" flags. We found the achievable increases in speed to be small relative to their corresponding memory requirements, and so HUMAnN 3 continues to favor DIAMOND's default, lower-memory configuration.

PanPhlAn 3 with expanded databases and functional annotations

PanPhlAn performs strain-level metagenomic profiling by identifying the species-specific gene repertoire composition inside individual metagenomic samples (Scholz *et al*, 2016). It maps metagenomes against the pangenome of a species of interest using bowtie2 (Langmead & Salzberg, 2012). After coverage normalization (by summing the gene coverage of all genes in a gene family and dividing it by the average gene length of that family), PanPhlAn builds a coverage curve of genes families across each sample and assesses which of these gene families are present or absent. This leads to the creation of a binary matrix of gene family presence/absence across all samples.

Compared to the previous versions, in PanPhlAn 3 we adopt a new ChocoPhlAn 3 pre-computed pangenome database of 2,298 species built from species included in MetaPhlAn 3 for which at least 2 reference genomes are available. For species having more than 200 reference genomes available, the pangenome is made using a representative subset of 200 genomes maximizing the Mash distances between them (Ondov *et al*, 2016). PanPhlAn pangenomes from the database are composed of a FASTA file of all contigs, pre-computed bowtie2 indexes and a tab-separated values file containing the UniRef90 ID of the gene family

as well as gene name, position in genomes,on contigs, and functional and structural annotations

Moreover, new functionalities include a script for quick visualization of the presence/absence matrix with functionalities for clustering of gene family's profiles across samples. An empirical p-value can be computed for each cluster based on the ratio between the sum of the genes lengths of one group and its total span along the contig. Thus a significantly "close" genes group can be identified and computation of empirical p-values assessing whether or not the genetic proximity of these families along the contigs could be considered significant. This eases the detection and identification of mobile elements in metagenomic samples.

PhyloPhlAn 3

PhyloPhlAn 3 is an easy-to-use method to perform taxonomic contextualization and phylogenetic analysis of microbial genomes and of metagenome-assembled genomes (MAGs). PhyloPhlAn among its databases exploits both the set of core genes and of reference genomes identified by ChocoPhlAn 3 and extracted from the 111,825 UniProt Proteomes for each taxonomic species. The methods, performance, and examples of PhyloPhlAn are described elsewhere (Asnicar *et al*, 2020) and refers to the same version incorporated into bioBakery 3. In brief, the core genes included in the PhyloPhlAn 3 database are used to identify sequence homologs in the input genomes and MAGs that are then aligned, concatenated, and used for phylogeny reconstruction. A set of MAGs previously analyzed (Pasolli *et al*, 2019) can also be included to provide phylogenetic contextualization of newly assembled MAGs. PhyloPhlAn 3 thus provides the methodology to integrate assembly-based methods and phylogenetic analysis into the bioBakery 3 analysis framework.

Synthetic metagenomes and gold standards for bioBakery 3 evaluations

We tuned and evaluated MetaPhlAn 3 and HUMAnN 3 using multiple different synthetic metagenomes of known species and gene content. The first set included synthetic metagenomes and gold-standard taxonomic profiles from the CAMI challenge representing five human body site-specific microbiomes and the murine gut microbiome (Sczyrba *et al*, 2017; Fritz *et al*, 2019). All such CAMI metagenomes were used for the evaluation of taxonomic profiling methods (including MetaPhlAn 3) while the first five lexically ordered metagenomes from each environment (human body sites and mouse gut) were used for the evaluation of functional profiling methods (including HUMAnN 3).

Second, because gold standard functional profiles were not provided for the CAMI metagenomes, we generated them ourselves by 1) functionally annotating the genomes sampled to build the CAMI metagenomes (and then 2) weighting their functional contributions

according to mean coverage depth per “sample”. Notably, this approach to gold-standard construction does not account for gene-to-gene variation in read sampling along the length of community genomes. As a result, comparing the gold standards with functional profiles derived directly from the metagenome underestimates the profiles’ accuracy (by ~0.1 units of Bray-Curtis distance at the UniRef90 level).

We applied procedures for community genome annotation developed during HUMAnN2 benchmarking to aid in gold-standard construction (Franzosa *et al*, 2018). Briefly, we first identified and translated open reading frames (ORFs) within the CAMI genomes using Prodigal (Hyatt *et al*, 2010), and then aligned the translated ORFs against the v3 UniRef90 and UniRef50 sequence databases using DIAMOND (Buchfink *et al*, 2015). Each ORF was assigned to the best-scoring UniRef90 family to which it aligned with at least 90% identity and 80% mutual coverage (if any); similarly, ORFs were assigned to the best-scoring UniRef50 family to which they aligned with at least 50% identity. Functional annotations were then transferred from UniRef90 and UniRef50 representatives to the corresponding ORFs, with UniRef90-derived, enzyme commission (EC) annotations forming the basis for the main functional profiling evaluation (**Fig. 1.2.1** and **Fig. S1.2.3**).

We constructed additional synthetic metagenomes by sampling sequencing reads from curated microbial genome sets using ART (Huang *et al*, 2012) with an Illumina HiSeq 2500 error model. One such group of metagenomes (abbreviated synphlan-nonhuman) was designed to mirror the sequencing depth and community structure of the CAMI metagenomes: i.e. inclusive of 30-million, 150-nt paired-end sequencing reads sampled from species genomes with a log-normal abundance distribution. However, the synphlan-nonhuman metagenomes are distinct from the CAMI metagenomes in that they exclude genomes of human-associated microbial species (defined as species detected in MetaPhlAn 3 profiles of metagenomes from the Expanded Human Microbiome Project, HMP1-II (Lloyd-Price *et al*, 2017)). In addition, 50% of species sampled for the synphlan-nonhuman metagenomes were associated with at least two sequenced isolate genomes and 50% of species pairs were congeneric sisters. We constructed an additional synthetic metagenome (synphlan-humanoid) based on the top-50 most abundant species detected from HMP1-II metagenomes to use for initial tuning of HUMAnN 3 (**Fig. S1.2.4**). This metagenome contained 10-million, 100-nt paired-end reads sampled evenly from underlying species genomes. We constructed gold standard taxonomic profiles for these metagenomes based on the sampled genomes’ taxonomic annotations and target sampling coverage; we constructed gold standard functional profiles based on UniProt-derived annotations of the species’ protein-coding genes.

Evaluation of MetaPhlAn 3 and HUMAnN 3 on synthetic data

To assess the performance of MetaPhlAn 3, we compared it with its previous version, MetaPhlAn 2 (Truong *et al*, 2015), alongside mOTUs2 (Milanese *et al*, 2019) and Bracken (Lu *et al*, 2017; Wood *et al*, 2019). We profiled a total of 118 synthetic metagenomes spanning different ecosystems: (i) 49 synthetic metagenomes (10 Airways, 10 Gastrointestinal Tract, 10 Oral, 10 Skin, 9 Urogenital tract) provided by the 2nd CAMI challenge (Sczyrba *et al*, 2017) resemble the composition of the Human Microbiome as described by the Human Microbiome Project (Turnbaugh *et al*, 2007); (ii) 64 synthetic metagenomes generated by CAMISIM and modeled after the murine gut microbiome (Fritz *et al*, 2019); (iii) 5 synthetic metagenomes including non-human associated species (see above).

Each software was run using default parameters as described in their respective user manuals. Additionally, mOTUs2 was run with parameters "-C recall" and "-C precision" in order to increase precision and recall, respectively. When not directly available from the tool (MetaPhlAn 2 and Bracken), output profiles were converted into the CAMI output format as described by the BioBoxes RFC (Belmann *et al*, 2015; BioBoxes) in order to benchmark with the OPAL framework (Meyer *et al*, 2019) (version 1.0.5).

From the panel of measures computed by OPAL, we selected a subset (precision, recall, F1 score) for comparisons (**Table S1.2.9**). Additionally to these measures, we computed the Pearson Correlation Coefficient between the predicted and expected relative abundance and the Bray-Curtis similarity index using arcsin square-root normalized relative abundances (**Table S1.2.3**).

MetaPhlAn 3 includes markers describing species groups, a case is not taken into account by OPAL. To perform the evaluation, we expanded the species group to represent all contained species and considered a true positive if the expected species matches one species taxonomically placed under the species group. In case of no matches, we consider as false positive only one species.

We also assessed the performance in terms of run-time and memory usage. We profiled five HMP samples (SRS014235, SRS011271, SRS064645, SRS023346, SRS048870) with all the aforementioned software (using only one thread) and tracked every second of the execution till the end of process the resident set size (RSS) memory usage using ps.

We evaluated HUMAnN 3, HUMAnN 2 (Franzosa *et al*, 2018), and Carnelian (Nazeen *et al*, 2020) on 30 CAMI metagenomes and the 5 synphlan-nonhuman metagenomes. Evaluations of HUMAnN 3 were carried out using version 3.0.0-alpha of the software, MetaPhlAn 3,

bowtie2 version 2.3.5.1, and DIAMOND version 0.9.24. Evaluations on HUMAnN 2 were carried out using version 0.11.1 of the software, MetaPhlAn version 2.7.5, bowtie2 version 2.3.5.1, and DIAMOND version 0.8.36 (HUMAnN 2 is not compatible with DIAMOND version 0.9). HUMAnN 3 and 2 were run with their default settings and full-size databases alongside the "--threads 6" option. UniRef90 abundance profiles were converted to EC abundance profiles (to facilitate comparisons with Carnelian) using the "uniref90_level4ec" option of the humann_regroup_table script.

We evaluated Carnelian version 1.0.0 following installation and usage instructions given at http://cb.csail.mit.edu/cb/carnelian/ and https://github.com/snz20/carnelian. Specifically, we first converted synthetic metagenome reads to FASTA format (this step was not counted toward the total runtime of the Carnelian method). Reads were then scanned for peptide fragments using "carnelian.py translate" wrapping FragGeneScan (Rho *et al*, 2010) version 1.31 with the "-n 3" option. Peptides were then assigned to EC categories using "carnelian.py predict" wrapping Vowpal Wabbit 8.1.1 and the EC-2010-DB model supplied at the above URLs. Finally, adjusted EC abundances were estimated using "carnelian.py abundance" and the average EC family gene lengths supplied with the software and a fragment size of 150 (to match the reads of the CAMI and synphlan-nonhuman metagenomes).

All method calls were made with the humann_benchmark utility script to track total runtime and memory usage (maximum resident set size, MaxRSS). Runtimes were converted to equivalent CPU-hours. For multi-step computations, CPU-hours were summed while the overall maximum MaxRSS was retained. Predicted EC abundances were sum-normalized to 1.0 at the community and per-species levels prior to Bray-Curtis dissimilarity computations.

Colorectal cancer microbiome meta-analysis

We applied the new MetaPhlAn 3 and HUMAnN 3 on a set of human gut metagenomes profiling colorectal cancer patients and controls, updating our previous meta-analyses performed with MetaPhlAn 2 and HUMAnN 2 (Thomas *et al*, 2019; Wirbel *et al*, 2019). To the previous meta-analysis, we added two more datasets that became available afterward (Gupta *et al*, 2019; Yachida *et al*, 2019). In total, we analyzed 1,262 metagenomes from 10 datasets (for a total of CRC metagenomes and 600 controls, **Table S1.2.10**). The dataset was stratified by country of origin with the exception of the two Italian cohorts published in (Thomas *et al*, 2019) which were kept separate due to differences in the DNA extraction protocols. Results were thus computed on nine distinct sub-cohorts.

MetaPhlAn 3 and HUMAnN 3 were used for the taxonomic and functional profiling of all sub-cohorts. Meta-analysis on the species-level, pathways, UniRef90 gene families, and enzyme

commission (EC) categories relative abundances were performed on the sub-cohorts as previously described (Thomas *et al*, 2019). In brief, relative abundances were arcsine-square-root transformed, Cohen's D was computed by the escalc function (metafor R package (Viechtbauer, 2010) to model random effects, and I^2 estimates and Cochran's Q-test were used for quantifying study-heterogeneity and assessing their statistical significance. Multidimensional scaling analysis was performed on the Weighted UniFrac distance (vegan "cmdscale" and rbiom "unifrac" function (Oksanen *et al*, 2008) computed on the relative abundance data adjusted for study batch effect with MMUPHin (Ma, 2019) and normalized using arcsin-square root. Alpha-diversity analysis was performed on the data after being rarefied to the 10th percentile of the read depth in each cohort.

We used MetAML (Pasolli *et al*, 2016) to feed species-level and pathway-level relative abundances to a Random Forest model (Breiman, 2001). Age was also added to the feature-set, as this covariate has been shown to improve microbiome predictions in CRC (Ghosh *et al*, 2020). MetAML executed the Random-Forest implementation by Scikit-Learn v.0.22.2 with the following parameters: 10,000 estimator trees, square-root as the proportion of feature sampled in entrance to each estimator, no-maximum depth for the trees, 1 sample as the minimum amount for each leaf of each tree, "gini" as impurity criterion. Considering each cohort, we tested the taxonomical and the functional potential profiles in the CRC prediction problem in a standard cohort-specific cross-validation as well as on the more reproducible leave-one-dataset-out (LODO) setting (Thomas *et al*, 2019; Wirbel *et al*, 2019).

UniRef90 *cutC* gene family IDs were selected from the UniRef90 database included in HUMAnN 3. Species richness was calculated by tallying species with non zero relative abundance. Differential species richness and *cutC* abundance tests were performed using the Wilcoxon rank-sum test, wilcox.test, as implemented in the 'stats' R package.

HMP2 IBD metagenome and metatranscriptome profiling

We applied MetaPhlAn 3 and HUMAnN 3 to 1,635 metagenomes and 817 metatranscriptomes from the HMP2 Inflammatory Bowel Disease (IBD) Multi-omics Database (IBDMDB) (Lloyd-Price *et al*, 2019). We took advantage of previously quality-controlled sequencing data from this cohort as downloaded from http://ibdmdb.org (June 2020). Following the standard bioBakery workflow (McIver *et al*, 2018) for combined meta-omic sequencing data, we processed the HMP2 metagenomes using HUMAnN 3.0.0.alpha.1 (including taxonomic prescreening performed by MetaPhlAn 3). We then processed the paired HMP2 metatranscriptomes using their corresponding metagenomic taxonomic profiles as guides for pangenome selection. To quantify improved performance in bioBakery 3, we compared the HUMAnN logs produced during the runs described above with logs downloaded from

http://ibdmdb.org describing analyses of the same samples using MetaPhlAn 2.6.0 and HUMAnN 2.11.0.

To identify expression-level microbial metabolic biomarkers of IBD activity from the HMP2 dataset, we sum-normalized UniRef90 gene family abundance profiles to “copies per million” (CPM) units and then summed UniRef90 CPMs according to enzyme commission (EC) annotations using HUMAnN utility scripts. We then compared community-level EC expression with other sample properties using a mixed effects model implemented in R’s lmerTest package (Kuznetsova *et al*, 2017) (using subject as a random effect to account for repeated longitudinal sampling):

$$log(RNA) \sim log(DNA) + diagnosis + diagnosis{:}active + age + antibiotics + (1|subject)$$

For a given EC, we evaluated the above model over paired meta-omes in which the EC’s metatranscriptomic abundance (RNA) and metagenomic abundance (DNA) were both non-zero; ECs were excluded if they failed to satisfy this condition in at least 10% of paired meta-omes. This approach avoids interpreting RNA non-detection as strong evidence of “down-regulation” (relative to DNA abundance, identifying zero RNA reads for a feature is more common due to the wide dynamic range of gene expression values and the large fraction of sequencing depth absorbed by non-coding RNAs).

The inclusion of DNA abundance as a covariate in the above model accounts for the strong dependence between a function’s gene (metagenomic) copy number and its metatranscriptomic abundance. Thus, associations between EC RNA and other covariates can be interpreted as associations with “residual expression” (potentially reflecting up- or down-regulation of community genes independent of changes in metagenome structure). Subject age at study enrollment and per-sample antibiotics exposure were included as additional clinical covariates. The statistical significance of model covariates was assessed after performing Benjamini-Hochberg FDR correction on model p-values batched by covariate and level.

We focused on associations between residual EC expression and subject diagnosis and disease activity with diagnosis. Here, subject diagnosis was divided broadly into Crohn’s disease (CD; n=49), ulcerative colitis (UC; n=30), and non-IBD controls (n=27). Due to the longitudinal nature of the HMP2 dataset, subjects diagnosed with CD and UC experienced variation in disease severity over the course of the study. The effects of disease activity on the microbiome were previously quantified as a “dysbiosis score” (Lloyd-Price *et al*, 2019) measuring ecological deviation from the control microbiome population. Samples from CD and

UC patients that deviated most strongly by this measure were classified as “active.” Of 788 paired meta-omes considered here, 363 were from CD patients (76 with “active” CD), 227 were from UC patients (23 with “active” UC”), and 198 were from non-IBD controls. Consistent with earlier analyses of the HMP2 dataset (Lloyd-Price *et al*, 2019), we did not detect significant differences in EC expression as a function of diagnosis alone (i.e. independent of disease activity), as non-active IBD meta-omes tend to be similar to those from control patients.

Strain-level analysis of *Ruminococcus bromii*

For *Ruminococcus bromii* population genetic analysis, from the 9,316 metagenomes spanning 46 datasets considered by Pasolli et al. (Pasolli *et al*, 2019), we selected 4,077 samples in which *R. bromii* was found present with a relative abundance above 0.05%. Strain-level profiling with StrainPhlAn 3 was performed using default parameters. 702 samples were discarded due to the low number and/or poor quality of the reconstructed markers (samples having less than 20 markers and markers present in less than the 80% of the samples are excluded). 124 *R. bromii* MetaPhlAn 3 markers were used to generate a multiple sequence alignment. A phylogenetic distance matrix was produced by the dismat function from the EMBOSS package (Rice *et al*, 2000) (Kimura 2-parameter distance correction) using the multiple sequence alignment file produced by StrainPhlAn. Prediction strength analysis performed on the phylogenetic distance matrix using the prediction.strength function included in the “fpc” R package (Hennig, 2010) version 2.2 revealed the presence of 4 optimal clusters (strength threshold 0.8). PAM clustering was subsequently applied on the phylogenetic distance matrix using the “cluster” R package (Kaufman & Rousseeuw, 2009) version 2.1. The phylogenetic tree generated by PhyloPhlAn was plotted with GraPhlAn (Asnicar *et al*, 2015). For visualization purposes, European countries were grouped into the EUR group. Tree cluster colors were assigned by considering the most common cluster assigned to leaves, and clusters 3 and 4 were joined into the “Others” group for the sake of discussion. In order to detect possible events of vertical transmission of R. bromii, we executed the “strain_transmission.py” script using as input the phylogenetic tree produced by StrainPhlAn.

Pangenome-based strain-level analysis was performed on the same selected set of samples using PanPhlAn 3 with a *R. bromii* pangenome composed of 8 reference genomes available on NCBI (GCA_002834165, GCA_002834225, GCA_002834235, GCA_003466165, GCA_003466205, GCA_003466225, GCA_900101355 and GCA_900291485). After mapping the metagenomic samples to the pangenome, a binary matrix of presence/absence was built using the PanPhlAn profiling script with default options for strain detection and filtering (--min_coverage 2 --left_max 1.25 --right_min 0.75). The resulting matrix describes the

presence/absence of 6,151 UniRef90 families across 2,679 metagenomics samples and 8 reference genomes.

In order to simplify the visualization of these results, we first discarded the genes families present in less than 2 samples or absent in 5 or less samples. Then, the Jaccard distance based on presence/absence fingerprint was computed for both genes families and samples. Hierarchical clustering was built using the Ward criterion ("ward.D2" in R "hclust" function). A second more stringent filtering removed all genes families present in more than 95% or less than 5% of the remaining samples.

For assessing the correlation between the strain-level genomics and pangenomics results, we compared the phylogenetic distance distributions retrieved from the StrainPhlAn and PanPhlAn analyses. We used RAxML version 8.2.4 (Stamatakis, 2014) to generate phylogenetic distances between samples from PanPhlAN results. PanPhlAn information was coded as the presence-absence fingerprint of each sample and distances were computed using the substitution model based on these two states (argument -m MULTICAT of RAxML). One outlier sample was discarded due to mislabelled genomes. The StrainPhlAn phylogenetic distances were produced during the execution of the "strain_transmission.py" script. Correlation between PanPhlAn and StrainPhlAn pairwise distances was calculated using the Pearson correlation Coefficient.

Data Availability

Human and murine synthetic metagenomes and gold standards provided by the CAMI Challenge are available at https://data.cami-challenge.org/participate.

Non-human synthetic metagenomes and gold standards are available at https://www.dropbox.com/s/scax4jzwghfx8gu/synphlan_nonhuman_feb2020.tar?dl=1. CRC metagenomic datasets analyzed in the meta-analysis are available in the Sequence Read Archive under accession numbers PRJEB7774, PRJNA531273, PRJNA447983, PRJDB4176, PRJEB12449, PRJEB27928, PRJDB4176, PRJEB10878, and PRJEB6070. Sequences and data for the Integrative Human Microbiome Project are available at the IBDMDB website (https://ibdmdb.org/) and deposited in SRA under accession number PRJNA398089.

Taxonomic profiles, functional profiles, and sample metadata of the CRC datasets are available as **Table S1.2.5** and **Table S1.2.10**. Taxonomic profiles and functional profiles of the HMP IBDMDB dataset are newly available at https://ibdmdb.org/.

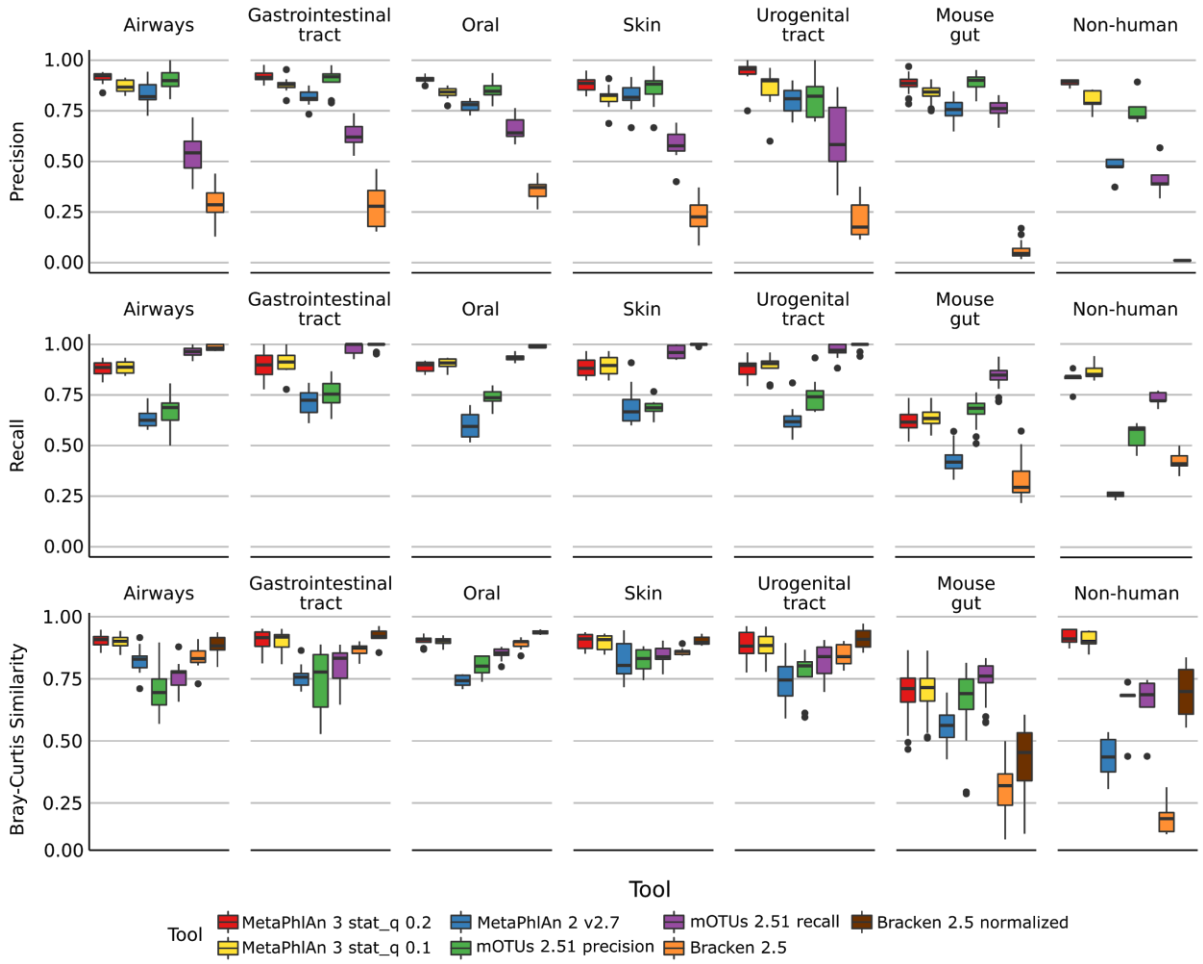
Profiles are also available through the curatedMetagenomicData R package (Pasolli *et al*, 2017). The full list of metagenomic datasets and samples used for the strain-level analysis of

Ruminococcus bromii is reported in **Table S1** from (Pasolli *et al*, 2019). *Ruminococcus bromii* reference genomes are deposited in GenBank under accession GCA_002834165, GCA_002834225, GCA_002834235, GCA_003466165, GCA_003466205, GCA_003466225, GCA_900101355 and GCA_900291485.

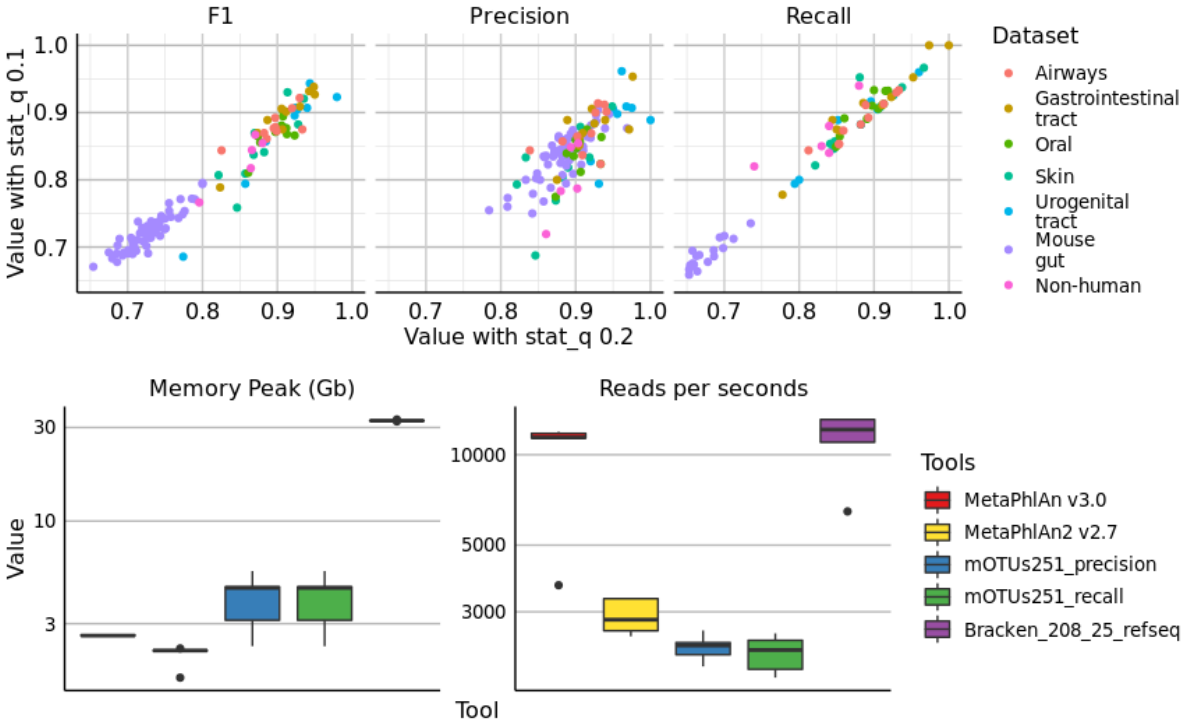
Funding

The work was supported by the European Research Council (ERC-STG project MetaPG) to NS; by MIUR 'Futuro in Ricerca' (grant No. RBFR13EWWI_001) to NS; by the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project) to NS; by the National Cancer Institute of the National Institutes of Health (1U01CA230551) to NS; and by the Premio Internazionale Lombardia e Ricerca 2019 to N.S.

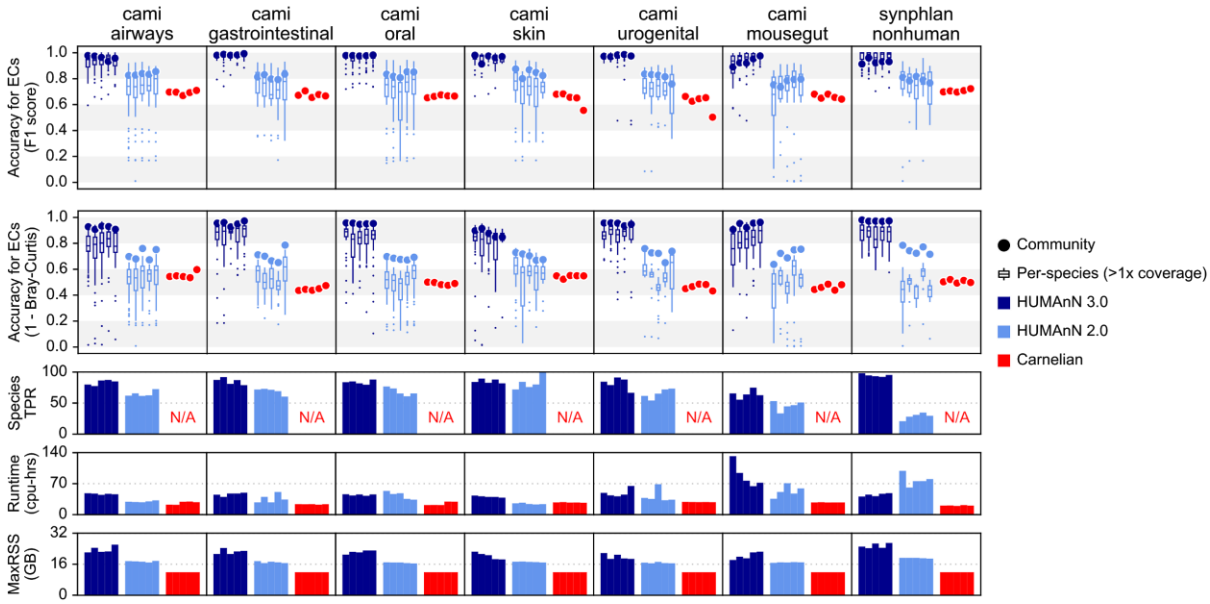
Supplementary Figures



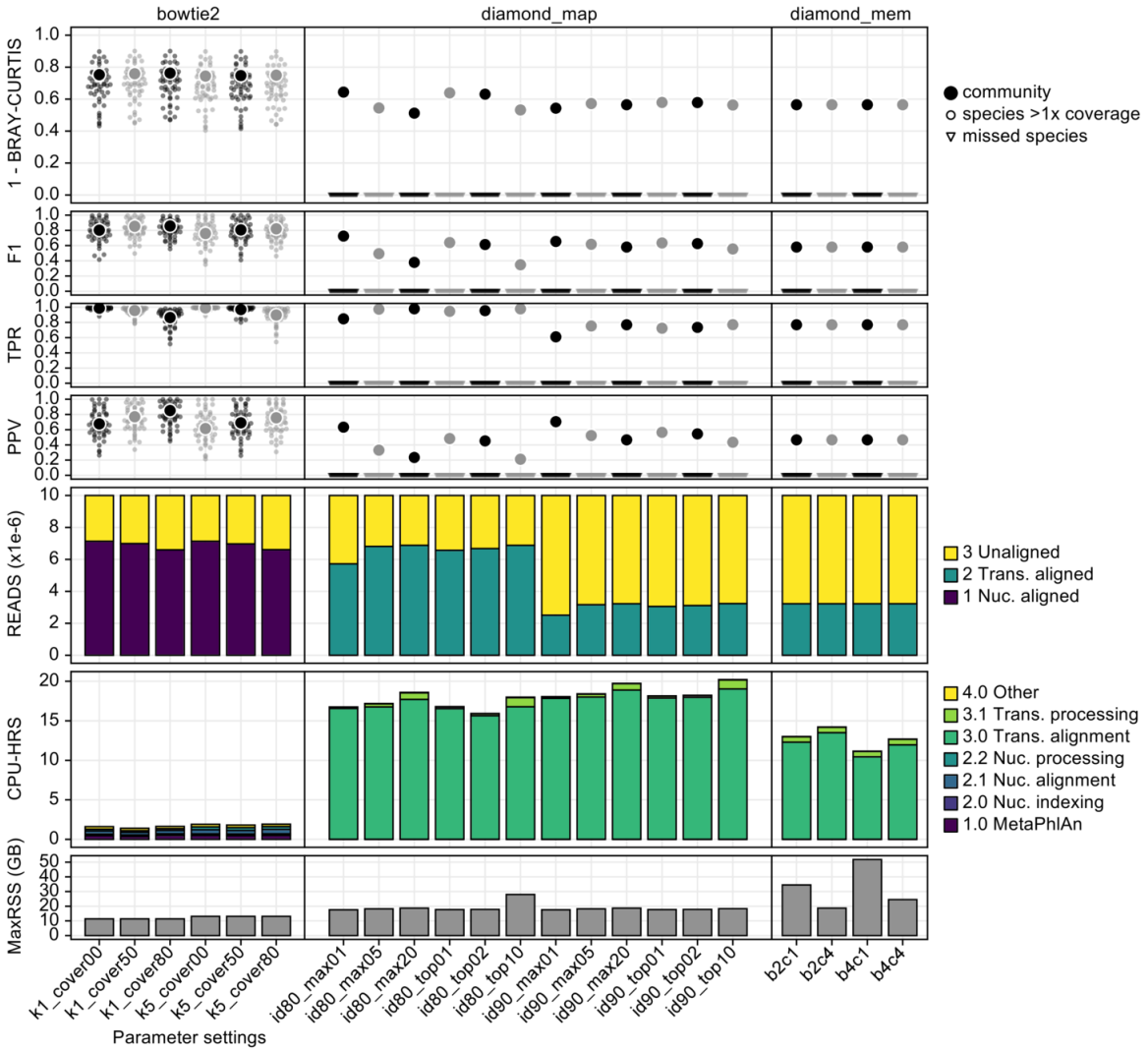
Supplementary Figure 1.2.1: Performance metrics (Precision, Recall, Bray-Curtis similarity) of MetaPhlAn 3.0, MetaPhlAn2, mOTU, and Bracken species-level profiling of the CAMI human-associated, CAMI mouse gut, and non-human datasets. Bray-Curtis similarity index is calculated on arcsine-square-root transformed relative abundances



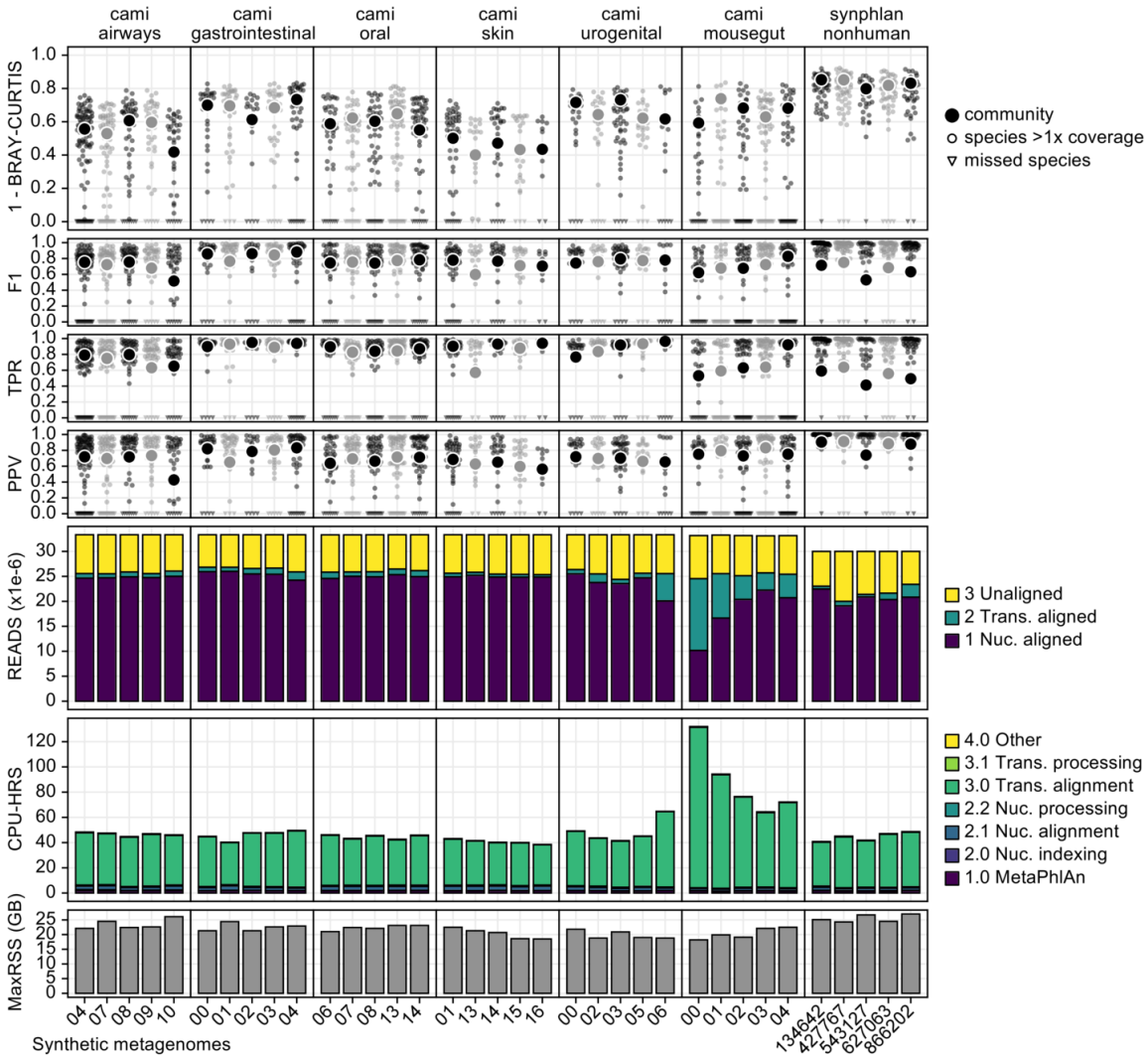
Supplementary Figure 1.2.2: (top) Scatter plots of precision, recall, and F1 score, of all the synthetic metagenomes profiled with MetaPhlAn 3 using stat_q=0.2 (default value for MetaPhlAn 3) and stat_q=0.1 (rho = 0.97). **(bottom)** Comparison of memory usage (maxRSS) and speed of taxonomic profilers included in the evaluation. Each tool was run on 5 HMP metagenomes using 1 thread.



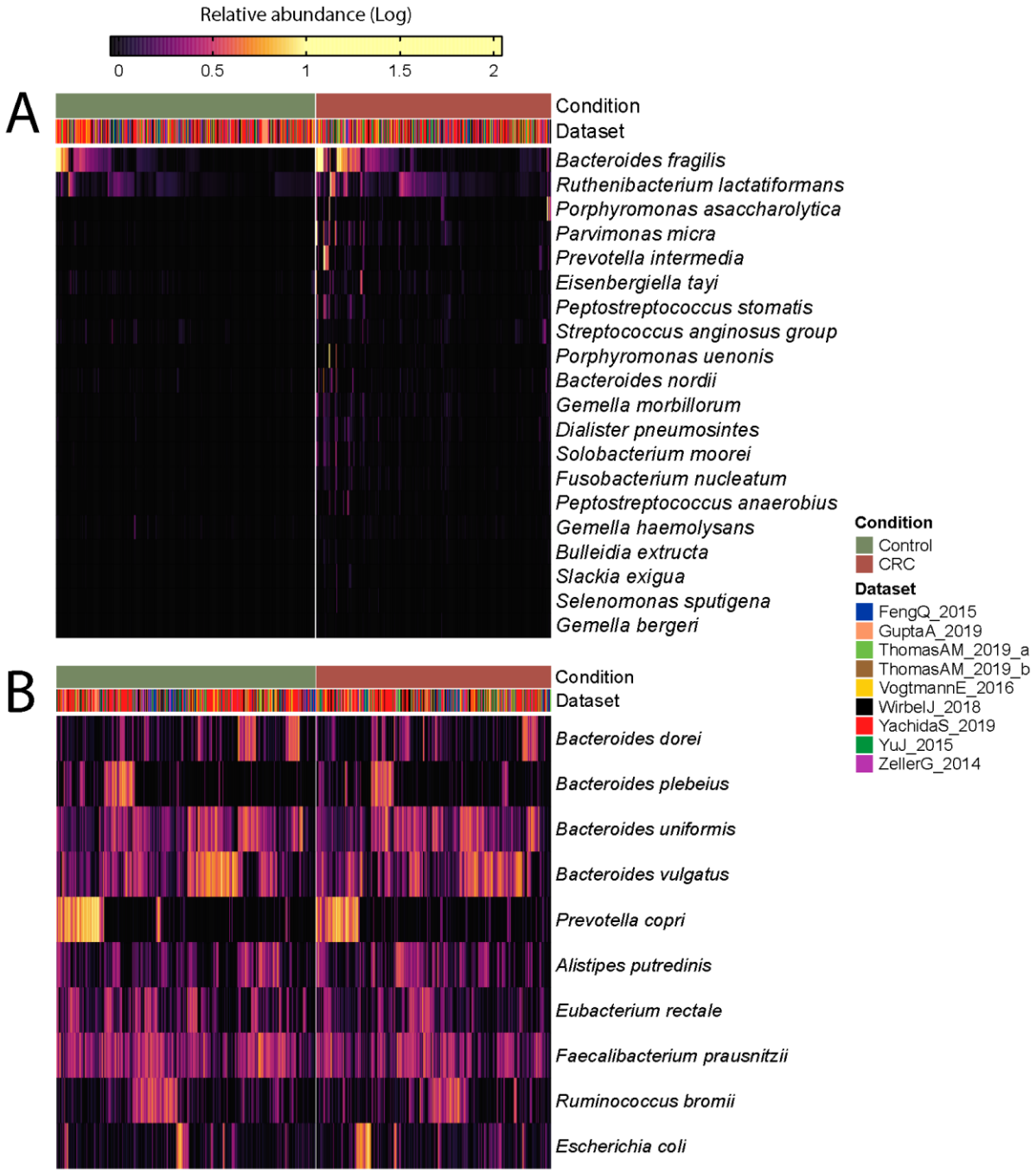
Supplementary Figure 1.2.3: This figure expands Fig. 1.2.1D from the main text to further compare HUMAnN 3, HUMAnN 2, and Carnelian on the basis of F1 score for accuracy of enzyme commission (EC) family detection, runtime (cpu-hrs), and peak memory usage (MaxRSS).



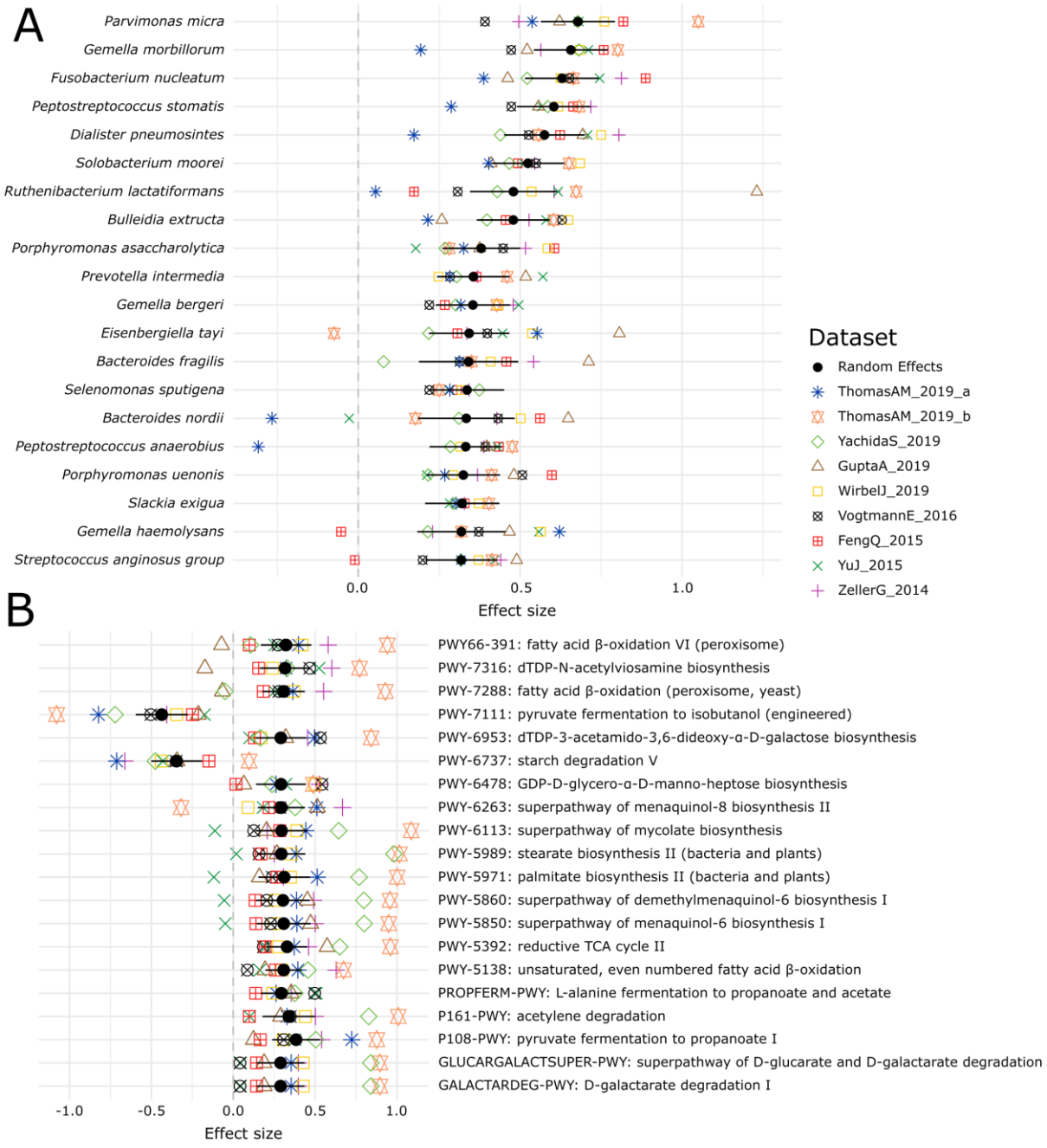
Supplementary Figure 1.2.4: This figure summarizes our initial optimization of HUMAnN 3 based on the synphlan-humanoid metagenome with a UniRef90 gold standard. Pangenome search (bowtie2 phase) was evaluated in "--bypass-translated-search" mode and translated search (diamond phase) was evaluated in "--bypass-nucleotide-search" mode. Left ("bowtie2") column: We compared accuracy and performance requesting 1 vs. 5 hits from Bowtie 2 and performing post hoc filtering of target sequences requiring 0% (i.e. no filtering), 50%, and 80% of sites to be hit. HUMAnN 3 defaults to a single hit (unchanged from HUMAnN 2) but requires 50% coverage of database sequences (similar to HUMAnN 2's translated search filter). Center ("diamond_map") column: We compared a variety of DIAMOND stringency filters during translated search. HUMAnN 3 uses alignments with >80% identity within 1% score of the best alignment (id80_top01), which is more sensitive (but otherwise similar) to the HUMAnN 2 default (id90_max20). Right column ("diamond_mem"): We evaluated different memory utilization settings in DIAMOND, but kept the DIAMOND default (b2c4) in both HUMAnN 2 and 3.



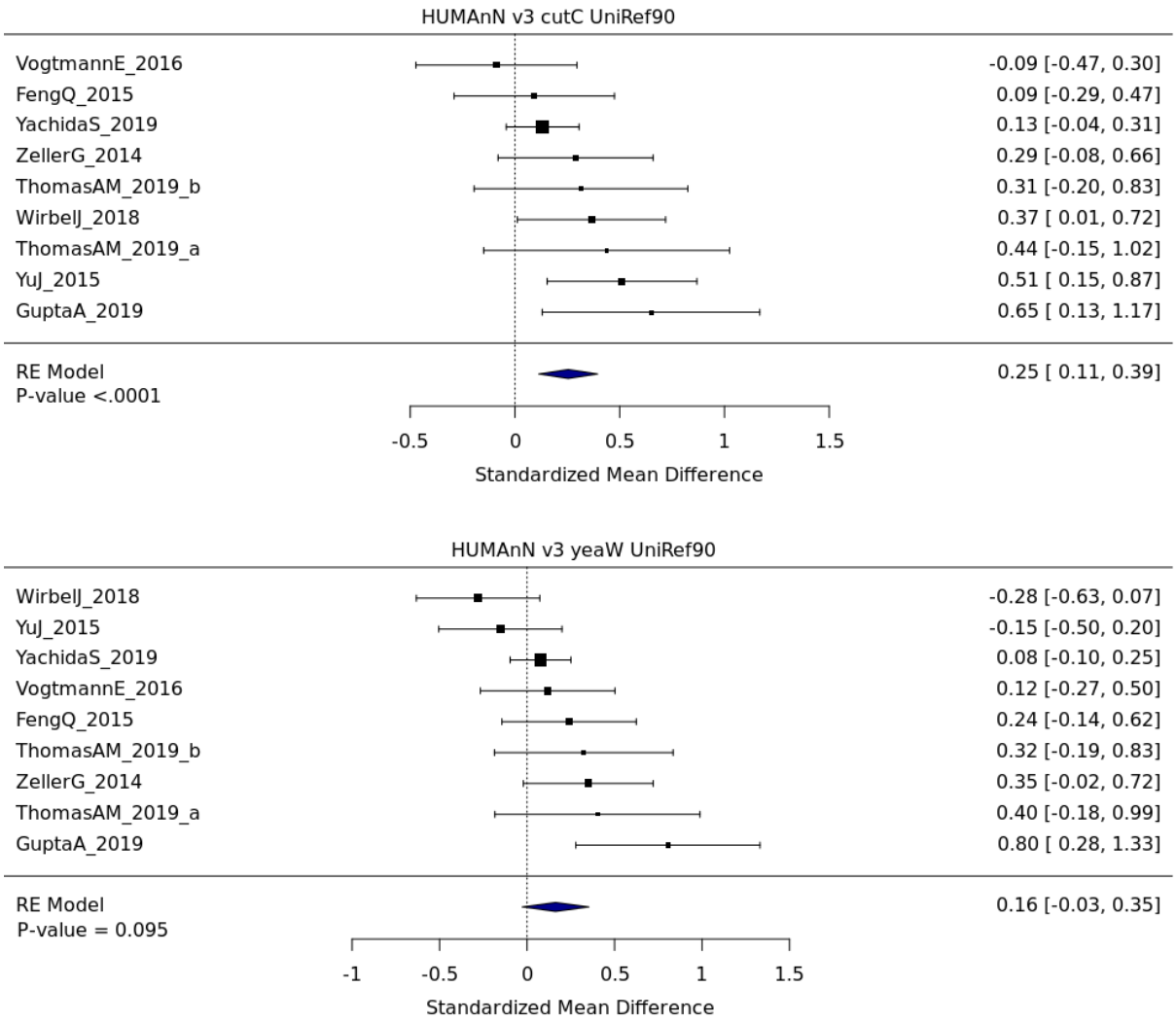
Supplementary Figure 1.2.5: This figure provides a high-resolution view of HUMAnN 3's performance in the evaluations of main-text Fig. 1.2.1D (accuracy and performance on CAMI and non-human-associated metagenomes). The top four rows (1 - BC, F1, TPR, and PPV) detail measures of accuracy for UniRef90-level protein families at the community (large dot) and well-covered-species (small dots) levels. The "READS" row indicates the stage of HUMAnN 3's tiered search where sample reads were aligned; ~75% of most samples' reads were explained, with the vast majority of the reads assigned by known pangenomes outside of the CAMI mousegut samples (which relied more heavily on translated search for explanations). The "CPU-HRS" row indicates the time spent in various phases of HUMAnN 3's tiered search, with the translated search step dominating overall runtime. The MaxRSS row indicates the peak memory usage (in GBs) for each sample, and was consistently in the 20-25 GB range.



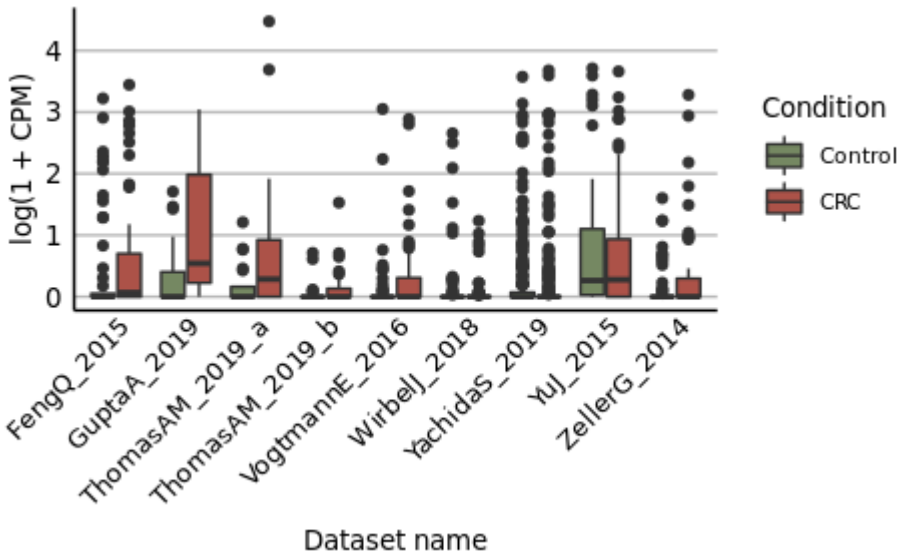
Supplementary Figure 1.2.6: Log-transformed relative abundances of the top 20 MetaPhlAn 3 species associated with colorectal cancer (A) and top 10 most abundant species (B) identified with a meta-analysis on 1,262 samples.



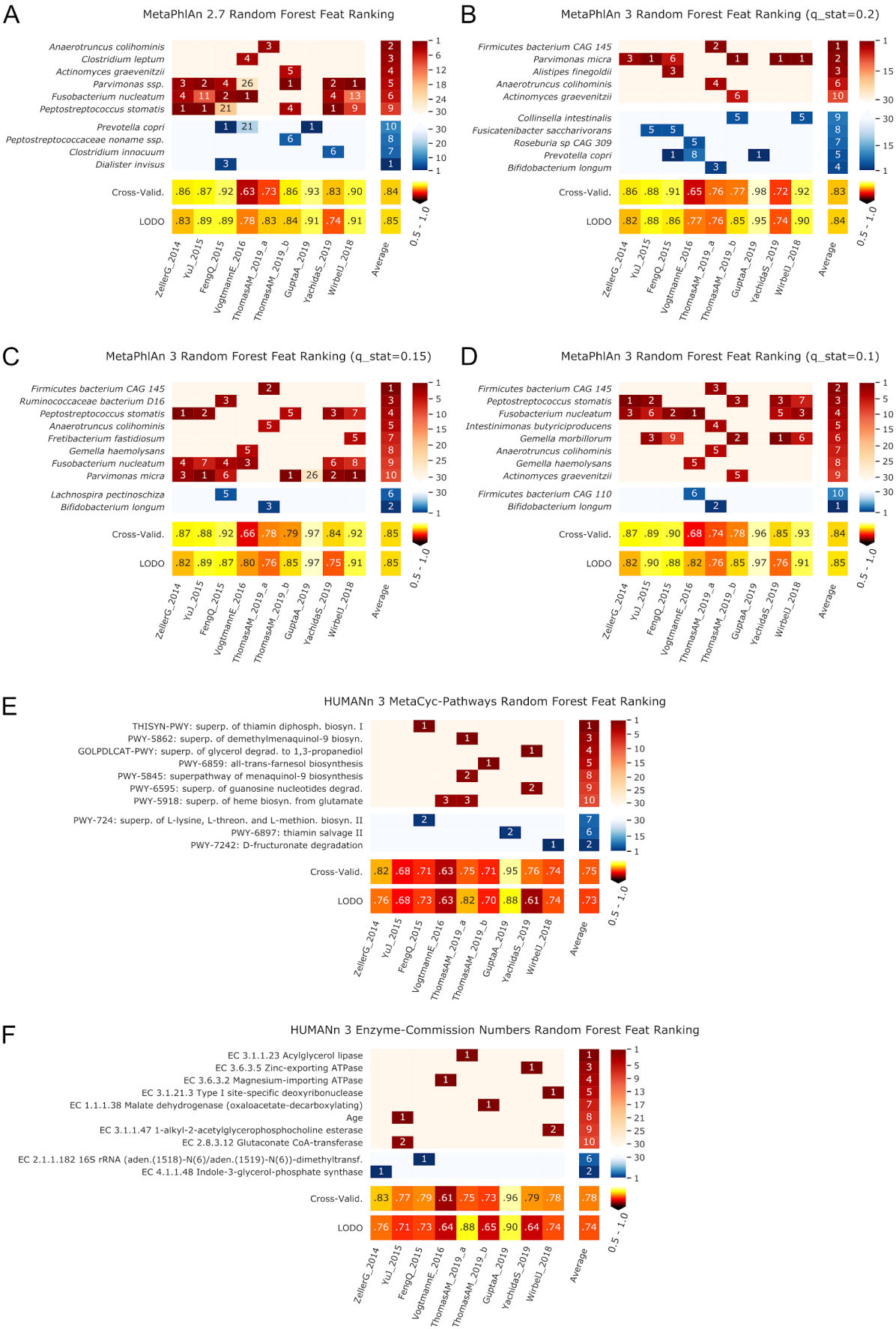
Supplementary Figure 1.2.7: Meta-analysis of the CRC datasets on the MetaPhlAn 3.0 species-level relative abundances **(A)** and relative abundance of MetaCyc pathway profiles generated with HUMAnN 3 **(B)**.



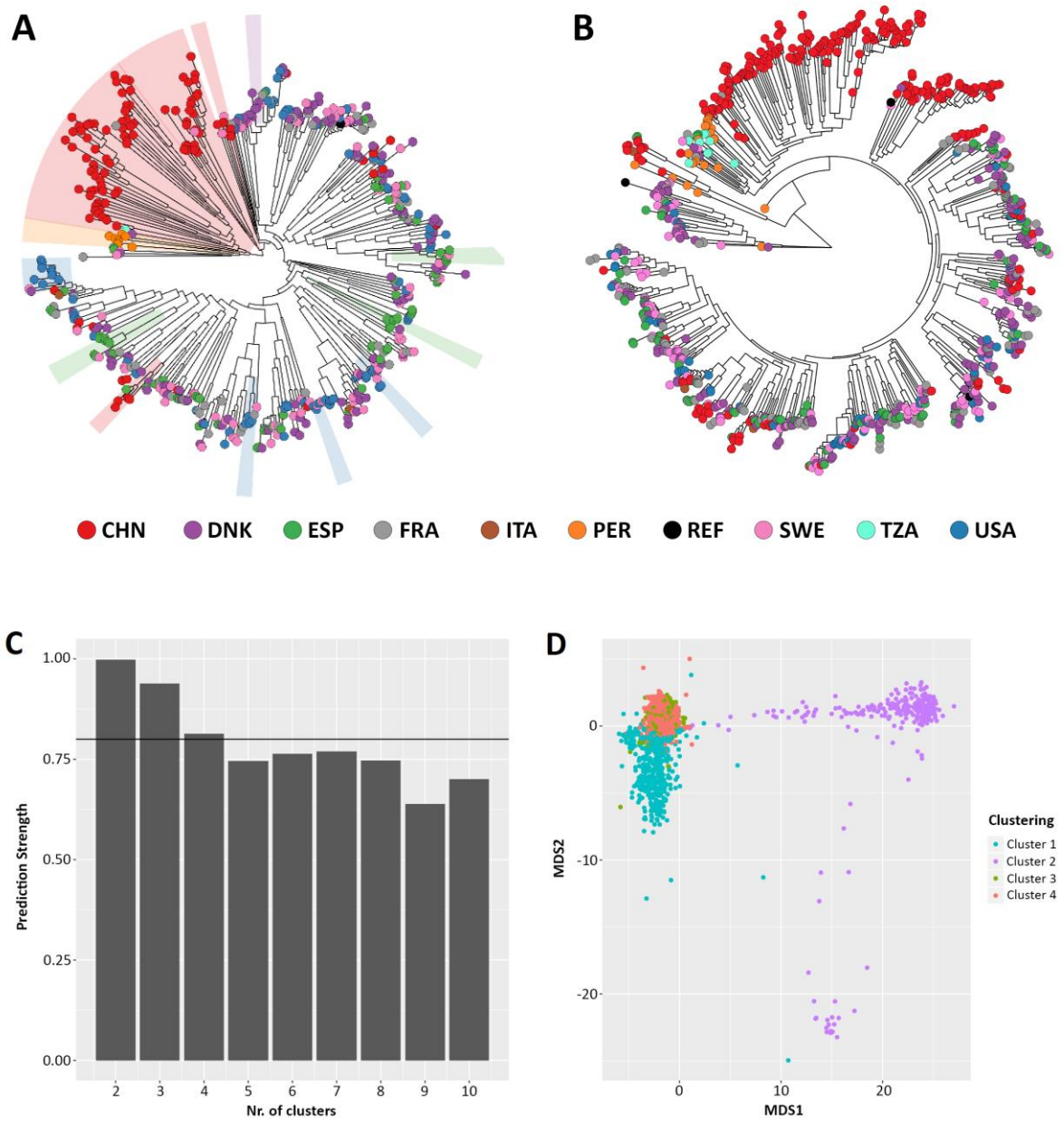
Supplementary Figure 1.2.8: Forest plot reporting effect sizes calculated using a meta-analysis of standardized mean differences and a random effects model on cutC and yeaW relative abundances between CRC and control samples.



Supplementary Figure 1.2.9: Distribution of yeaW gene relative abundance (log10 count-per-million normalized) extracted from HUMAnN gene family profiles.


A
MetaPhlAn 2.7 Random Forest Feat Ranking
Anaerotruncus colihominis
Clostridium leptum
Actinomyces graevenitzii
Parvimonas ssp.
Fusobacterium nucleatum
Peptostreptococcus stomatis
Prevotella copri
Peptostreptococcaceae noname ssp.
Clostridium innocuum
Dialister invisus
Cross-Valid.
LODO
ZellerG_2014
YuJ_2015
FengQ_2015
VogtmannE_2016
ThomasAM_2019_a
ThomasAM_2019_b
GuptaA_2019
YachidaS_2019
WirbelJ_2018
Average
0.5 - 1.0
B
MetaPhlAn 3 Random Forest Feat Ranking (q_stat=0.2)
Firmicutes bacterium CAG 145
Parvimonas micra
Alistipes finegoldii
Anaerotruncus colihominis
Actinomyces graevenitzii
Collinsella intestinalis
Fusicatenibacter saccharivorans
Roseburia sp CAG 309
Prevotella copri
Bifidobacterium longum
Cross-Valid.
LODO
C
MetaPhlAn 3 Random Forest Feat Ranking (q_stat=0.15)
Firmicutes bacterium CAG 145
Ruminococcaceae bacterium D16
Peptostreptococcus stomatis
Anaerotruncus colihominis
Fretibacterium fastidiosum
Gemella haemolysans
Fusobacterium nucleatum
Parvimonas micra
Lachnospira pectinoschiza
Bifidobacterium longum
Cross-Valid.
LODO
D
MetaPhlAn 3 Random Forest Feat Ranking (q_stat=0.1)
Firmicutes bacterium CAG 145
Peptostreptococcus stomatis
Fusobacterium nucleatum
Intestinimonas butyriciproducens
Gemella morbillorum
Anaerotruncus colihominis
Gemella haemolysans
Actinomyces graevenitzii
Firmicutes bacterium CAG 110
Bifidobacterium longum
Cross-Valid.
LODO
E
HUMANn 3 MetaCyc-Pathways Random Forest Feat Ranking
THISYN-PWY: superp. of thiamin diphosph. biosyn. I
PWY-5862: superp. of demethylmenaquinol-9 biosyn.
GOLPDLCAT-PWY: superp. of glycerol degrad. to 1,3-propanediol
PWY-6859: all-trans-farnesol biosynthesis
PWY-5845: superpathway of menaquinol-9 biosynthesis
PWY-6595: superp. of guanosine nucleotides degrad.
PWY-5918: superp. of heme biosyn. from glutamate
PWY-724: superp. of L-lysine, L-threon. and L-methion. biosyn. II
PWY-6897: thiamin salvage II
PWY-7242: D-fructuronate degradation
Cross-Valid.
LODO
F
HUMANn 3 Enzyme-Commission Numbers Random Forest Feat Ranking
EC 3.1.1.23 Acylglycerol lipase
EC 3.6.3.5 Zinc-exporting ATPase
EC 3.6.3.2 Magnesium-importing ATPase
EC 3.1.21.3 Type I site-specific deoxyribonuclease
EC 1.1.1.38 Malate dehydrogenase (oxaloacetate-decarboxylating)
Age
EC 3.1.1.47 1-alkyl-2-acetylglycerophosphocholine esterase
EC 2.8.3.12 Glutaconate CoA-transferase
EC 2.1.1.182 16S rRNA (aden.(1518)-N(6)/aden.(1519)-N(6))-dimethyltransf.
EC 4.1.1.48 Indole-3-glycerol-phosphate synthase
Cross-Valid.
LODO

Supplementary Figure 1.2.10: Features identified by the random-forest analysis on the species profiled with MetaPhlAn2 and MetaPhlAn 3 using different values of q_stat, and by HUMANn 3 grouping UniRef90 in MetaCyc pathways and Enzyme Commission numbers.



Supplementary Figure 1.2.11: Comparison between StrainPhlAn **(A)** and StrainPhlAn 3 **(B)** strain level profiling capabilities. *Ruminococcus bromii* species was profiled on 1,590 metagenomes. **(C)** Prediction strength at different cluster numbers and **(D)** PAM clustering results on the StrainPhlAn 3 phylogenetic distance matrix expose four optimal clusters of *Ruminococcus bromii* strains.

Supplementary Tables

Supplementary Table 1.2.1: Average values of F1 scores of MetaPhlAn 3, MetaPhlAn2, mOTUs2, and Kraken species-level profiles computed on the 123 synthetic metagenomes.

Tool	Airways	Gastrointestinal tract	Oral	Skin	Urogenital tract	Mouse gut	Non-human
MetaPhlAn v3.0 stat_q 0.2	0.880	0.894	0.869	0.853	0.869	0.722	0.830
MetaPhlAn v3.0 stat_q 0.1	0.896	0.908	0.898	0.883	0.906	0.728	0.855
MetaPhlAn2 v2.7	0.723	0.761	0.672	0.751	0.701	0.544	0.330
mOTUs251_precision	0.768	0.824	0.787	0.761	0.779	0.770	0.632
mOTUs251_recall	0.683	0.768	0.772	0.720	0.727	0.800	0.529
Bracken_208_25_refseq	0.440	0.426	0.525	0.359	0.339	0.091	0.021

Supplementary Table 1.2.2: bioBakery 3 software improvements.

https://www.dropbox.com/s/sbar6dggsrohz1m/Supplementary_table_2_biobakery_comparision.xlsx?dl=0

Supplementary Table 1.2.3: Mean and ranked values of Bray-Curtis dissimilarity and arcsine-square-root normalized Bray-Curtis dissimilarity obtained by MetaPhlAn 3, MetaPhlAn2, mOTUs2, and Kraken on the synthetic metagenomes considered in the evaluation.

https://www.dropbox.com/s/ptcmml1mkri002x/Supplementary_table_3_taxonomic_profiling_bray_curtis.xlsx?dl=0

Supplementary Table 1.2.4: Comparison of runtime and memory consumption of MetaPhlAn 3, MetaPhlAn2, mOTUs2, and Kraken+Bracken on the 5 HMP metagenomes.

Tool	Elapsed time (mean h)	Elapsed time (sd h)	Memory Peak (mean Gb)	Memory Peak (sd Gb)	Reads per second (mean)	Reads per second (sd)
MetaPhlAn v3.0	3.1120	2.65	2.614	0.01	10031.2	3,560.25
MetaPhlAn2 v2.7	8.2183	3.58	2.079	0.27	2911.2	400.08
mOTUs251_precision	10.6094	4.92	4.034	1.30	2283	234.26
mOTUs251_recall	11.5189	6.18	4.035	1.30	2186	310.86
Bracken_208_25_refseq	2.3504	1.40	32.529	0.27	11163	2,759.93

Supplementary Table 1.2.5: MetaPhlAn 3 taxonomic profiles and HUMAnN 3 functional profiles of the 1,262 CRC samples.

https://www.dropbox.com/s/3jn5tgdx7ssw5v2/Supplementary_table_5_CRC_metaphlan_humann_profiles.xlsx?dl=0

Supplementary Table 1.2.6: MetaPhlAn 3 species-level and HUMAnN 3 pathway abundances CRC meta-analysis results.

https://www.dropbox.com/s/s4zlop6jz2f35ek/Supplementary_table_6_CRC_metaanalysis_metaphlan_results.xlsx?dl=0

Supplementary Table 1.2.7: MetaPhlAn 3 species merged according to the species-level genome bin (SGB) system.

https://www.dropbox.com/s/r2fv71jxw81y6h4/Supplementary_table_7_metaphlan3_merged_species_with_SGB.xlsx?dl=0

Supplementary Table 1.2.8: Number of distinct MetaPhlAn 3 markers per species.

https://www.dropbox.com/s/yl7638cps97i72f/Supplementary_table_8_metaphlan3_markers_per_species.xlsx?dl=0

Supplementary Table 1.2.9: Per-sample OPAL binary measures (true positive, false positive, false negative, precision, recall, F1 score) computed on MetaPhlAn 3, MetaPhlAn2, mOTUs2, and Kraken species-level profiles computed on the 123 synthetic metagenomes.

https://www.dropbox.com/s/eokxfegceb6b7t4/Supplementary_table_9_evaluation.xlsx?dl=0

Supplementary Table 1.2.10: Metadata of all the 1,262 samples from the 10 CRC datasets.

https://www.dropbox.com/s/kwvdzcjoni0kug7/Supplementary_table_10_CRC_metaanalysis_datasets.xlsx?dl=0

Section references

Aagaard K, Riehle K, Ma J, Segata N, Mistretta T-A, Coarfa C, Raza S, Rosenbaum S, Van den Veyver I, Milosavljevic A, *et al* (2012) A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One* 7: e36466

Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, *et al* (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 8: e1002358

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, *et al* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46: W537–W544

Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz JM, Reeves D, Gandara J, Chhangawala S, *et al* (2015) Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst* 1: 72–87

Alivisatos AP, Blaser MJ, Brodie EL, Chun M, Dangl JL, Donohue TJ, Dorrestein PC, Gilbert JA, Green JL, Jansson JK, *et al* (2015) MICROBIOME. A unified initiative to harness Earth's microbiomes. *Science* 350: 507–508

Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD & Finn RD (2019) A new genomic blueprint of the human gut microbiota. *Nature* 568: 499–504

Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, *et al* (2020) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* doi:10.1038/s41587-020-0603-3 [PREPRINT]

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF & Quince C (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* 11: 1144–1146

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410

Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, Kartashov A, Leehr D, Ménager H, Nedeljkovich M, *et al* (2016) Common Workflow Language, v1.0. doi:10.6084/M9.FIGSHARE.3115156.V2 [DATASET]

Andrews S & Others (2010) FastQC: a quality control tool for high throughput sequence data. [PREPRINT]

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29

Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, *et al* (2017) Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* 2

Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, *et al* (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* 11: 2500

Asnicar F, Weingart G, Tickle TL, Huttenhower C & Segata N (2015) Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* 3: e1029

Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Res* 28: 304–305

Bangayan NJ, Shi B, Trinh J, Barnard E, Kasimatis G, Curd E & Li H (2020) MG-MLST: Characterizing the Microbiome at the Strain Level in Metagenomic Data. *Microorganisms* 8

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, *et al* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19: 455–477

Barone M, Turroni S, Rampelli S, Soverini M, D'Amico F, Biagi E, Brigidi P, Troiani E & Candela M (2019) Gut microbiome response to a modern Paleolithic diet in a Western lifestyle context. *PLoS One* 14: e0220619

Beghini F, Pasolli E, Truong TD, Putignani L, Cacciò SM & Segata N (2017) Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. *ISME J*

Belmann P, Dröge J, Bremges A, McHardy AC, Sczyrba A & Barton MD (2015) Bioboxes: standardised containers for interchangeable bioinformatics software. *Gigascience* 4: 47

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27: 573–580

Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, Chen X, Cocolin L, Eversole K, Corral GH, *et al* (2020) Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8: 103

BioBoxes RFC Github

Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL, Knight R, Maxon ME, Northen TR, Pollard KS, *et al* (2016) Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *MBio* 7

Bloom BH (1970) Space/time trade-offs in hash coding with allowable errors. *Commun ACM* 13: 422–426

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

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, *et al* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37: 852–857

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, *et al* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 35: 725–731

Brady A & Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 6: 673–676

Breiman L (2001) Random Forests. *Mach Learn* 45: 5–32

Breitwieser FP, Baker DN & Salzberg SL (2018) KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol* 19: 198

Breitwieser FP, Lu J & Salzberg SL (2019a) A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 20: 1125–1136

Breitwieser FP, Pertea M, Zimin AV & Salzberg SL (2019b) Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* 29: 954–960

Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, Jenkins AP, Naisilisili W, Tamminen M, Smillie CS, Wortman JR, *et al* (2016) Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535: 435–439

Broder AZ (1997) On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)* pp 21–29. IEEE

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH & Banfield JF (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523: 208–211

Buchfink B, Xie C & Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12: 59–60

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA & Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13: 581–583

Cameron SJS, Lewis KE, Huws SA, Hegarty MJ, Lewis PD, Pachebat JA & Mur LAJ (2017) A pilot study using metagenomic sequencing of the sputum microbiome suggests potential bacterial biomarkers for lung cancer. *PLoS One* 12: e0177062

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, *et al* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 42: D459-71

Caussy C & Loomba R (2018) Gut microbiome, microbial metabolites and the development of NAFLD. *Nat Rev Gastroenterol Hepatol* 15: 719–720

Chaumeil P-A, Mussig AJ, Hugenholtz P & Parks DH (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*

Chen B, Zhao Y, Li S, Yang L, Wang H, Wang T, Bin Shi, Gai Z, Heng X, Zhang C, *et al* (2018) Variations in oral microbiome profiles in rheumatoid arthritis and osteoarthritis with potential biomarkers for arthritis screening. *Sci Rep* 8: 17126

Chen L-X, Anantharaman K, Shaiber A, Eren AM & Banfield JF (2020) Accurate and complete genomes from metagenomes. *Genome Res* 30: 315–333

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J & Sayers EW (2016) GenBank. *Nucleic Acids Res* 44: D67-72

Cochrane G, Karsch-Mizrachi I, Takagi T & Sequence Database Collaboration IN (2016) The international nucleotide sequence database collaboration. *Nucleic Acids Res* 44: D48–D50

Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S & Bork P (2017a) metaSNV: A tool for metagenomic strain level analysis. *PLoS One* 12: e0182392

Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung F-E, *et al* (2017b) Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotechnol* 35: 1069–1076

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, *et al* (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331: 430–434

Dayhoff MO (1976) The origin and evolution of protein superfamilies. *Fed Proc* 35: 2132–2138

Delcher AL, Bratke KA, Powers EC & Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23: 673–679

Delmont TO, Gaia M, Hinsinger DD, Fremont P, Guerra AF, Murat Eren A, Vanni C, Kourlaiev A, d'Agata L, Clayssen Q, *et al* (2020) Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *Cold Spring Harbor Laboratory*: 2020.10.15.341214

Devkota S (2020) The gut microbiome during acute lifestyle transition. *Nat Med* 26: 1013–1015

Donati C, Zolfo M, Albanese D, Tin Truong D, Asnicar F, Iebba V, Cavalieri D, Jousson O, De Filippo C, Huttenhower C, *et al* (2016) Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. *Nat Microbiol* 1: 16070

Dröge J & McHardy AC (2012) Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform* 13: 646–655

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7: e1002195

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, *et al* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* 47: D427–D432

Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, *et al* (2021) Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 6: 3–6

Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, *et al* (2015) Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* 6: 6528

Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, Armanini F, Truong DT, Manara S, Zolfo M, *et al* (2018) Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* 24: 133-145.e5

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, *et al* (2008) The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 26: 541–547

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, *et al* (2014) Pfam: the protein families database. *Nucleic Acids Res* 42: D222-30

Flint HJ, Scott KP, Duncan SH, Louis P & Forano E (2012) Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* 3: 289–306

Flores GE, Bates ST, Knights D, Lauber CL, Stombaugh J, Knight R & Fierer N (2011) Microbial biogeography of public restroom surfaces. *PLoS One* 6: e28132

Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, Dunn M, Mkandawire TT, Zhu A, Shao Y, *et al* (2019) A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* 37: 186–192

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, *et al* (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* 15: 962–968

Fritz A, Hofmann P, Majda S, Dahms E, Dröge J, Fiedler J, Lesker TR, Belmann P, DeMaere MZ, Darling AE, *et al* (2019) CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 7: 17

Ghensi P, Manghi P, Zolfo M, Armanini F, Pasolli E, Bolzan M, Bertelle A, Dell'Acqua F, Dellasega E, Waldner R, *et al* (2020) Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics. *npj Biofilms and Microbiomes* 6: 47

Ghosh TS, Das M, Jeffery IB & O'Toole PW (2020) Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *Elife* 9

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM & Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312: 1355–1359

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, *et al* (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345: 1369–1372

Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, Prieto PA, Vicente D, Hoffman K, Wei SC, *et al* (2018) Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* 359: 97–103

Gregor I, Dröge J, Schirmer M, Quince C & McHardy AC (2016) PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* 4: e1603

Grice EA & Segre JA (2011) The skin microbiome. *Nat Rev Microbiol* 9: 244–253

Gupta A, Dhakan DB, Maji A, Saxena R, P K VP, Mahajan S, Pulikkan J, Kurian J, Gomez AM, Scaria J, *et al* (2019) Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *mSystems* 4

Guthrie JL & Gardy JL (2017) A brief primer on genomic epidemiology: lessons learned from Mycobacterium tuberculosis. *Ann N Y Acad Sci* 1388: 59–77

Heinken A, Ravcheev DA, Baldini F, Heirendt L, Fleming RMT & Thiele I (2019) Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome* 7: 75

Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, *et al* (2016) Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* 2: 16180

Hennig C (2010) fpc: Flexible procedures for clustering. *R package version* 2: 0–3

Huang W, Li L, Myers JR & Marth GT (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* 28: 593–594

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, *et al* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44: D286-93

Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207–214

Huson DH, Auch AF, Qi J & Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17: 377–386

Huson DH, Mitra S, Ruscheweyh H-J, Weber N & Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21: 1552–1560

Huttenhower C, Knight R, Brown CT, Caporaso JG, Clemente JC, Gevers D, Franzosa EA, Kelley ST, Knights D, Ley RE, *et al* (2014) Advancing the microbiome research community. *Cell* 159: 227–230

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW & Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT & Aluru S (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9: 5114

Jansson JK & Hofmockel KS (2018) The soil microbiome — from metagenomics to metaphenomics. *Current Opinion in Microbiology* 43: 162–168 doi:10.1016/j.mib.2018.01.013 [PREPRINT]

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T & Bork P (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 36: D250-4

Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G & Huttenhower C (2015) High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput Biol* 11: e1004557

Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M & Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42: D199-205

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H & Wang Z (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7: e7359

Karcher N, Pasolli E, Asnicar F, Huang KD, Tett A, Manara S, Armanini F, Bain D, Duncan SH, Louis P, *et al* (2020) Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biol* 21: 138

Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, Nielsen J & Backhed F (2013) Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498: 99–103

Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, *et al* (2019) The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 20: 1085–1093

Katoh K & Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30: 772–780

Kaufman L & Rousseeuw PJ (2009) Finding Groups in Data: An Introduction to Cluster Analysis John Wiley & Sons

Keohane DM, Ghosh TS, Jeffery IB, Molloy MG, O'Toole PW & Shanahan F (2020) Microbiome and health implications for ethnic minorities after enforced lifestyle changes. *Nat Med* 26: 1089–1095

Kim D, Song L, Breitwieser FP & Salzberg SL (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 26: 1721–1729

Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S, *et al* (2016) Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB* pp 87–90.

Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, *et al* (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* 16: 410–422

Koren O, Spor A, Felin J, Fak F, Stombaugh J, Tremaroli V, Behre CJ, Knight R, Fagerberg B, Ley RE, *et al* (2011) Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc Natl Acad Sci U S A* 108 Suppl 1: 4592–4598

Korpela K, Costea P, Coelho LP, Kandels-Lewis S, Willemsen G, Boomsma DI, Segata N & Bork P (2018) Selective maternal seeding and environment shape the human gut microbiome. *Genome Res* 28: 561–568

Köster J & Rahmann S (2012) Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 28: 2520–2522

Kuznetsova A, Brockhoff PB, Christensen RHB & Others (2017) lmerTest package: tests in linear mixed effects models. *J Stat Softw* 82: 1–26

Laforest-Lapointe I & Arrieta M-C (2018) Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *mSystems* 3

Langdon A, Crook N & Dantas G (2016) The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med* 8: 39

Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357–359

LaPierre N, Alser M, Eskin E, Koslicki D & Mangul S (2020) Metalign: efficient alignment-based metagenomic profiling via containment min hash. *Genome Biol* 21: 242

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, *et al* (2013) Richness of human gut microbiome correlates with metabolic markers. *Nature* 500: 541–546

Lee CS, Kim M, Lee C, Yu Z & Lee J (2016) The Microbiota of Recreational Freshwaters and the Implications for Environmental and Public Health. *Front Microbiol* 7: 1826

Leimbach A, Hacker J & Dobrindt U (2013) E. coli as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity. In *Between Pathogenicity and Commensalism*, Dobrindt U Hacker JH & Svanborg C (eds) pp 3–32. Berlin, Heidelberg: Springer Berlin Heidelberg

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, *et al* (2011) The European Nucleotide Archive. *Nucleic Acids Res* 39: D28-31

Lesker TR, Durairaj AC, Gálvez EJC, Lagkouvardos I, Baines JF, Clavel T, Sczyrba A, McHardy AC & Strowig T (2020) An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome. *Cell Rep* 30: 2909-2922.e6

Levy M, Blacher E & Elinav E (2017) Microbiome, metabolites and host immunity. *Curr Opin Microbiol* 35: 8–15

Ley RE, Turnbaugh PJ, Klein S & Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444: 1022–1023

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: 1674–1676

Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760

Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659

Lin H-H & Liao Y-C (2016) Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* 6: 24175

Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, *et al* (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569: 655–662

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, *et al* (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550: 61–66

Loeffler C, Karlsberg A, Martin LS, Eskin E, Koslicki D & Mangul S (2020) Improving the usability and comprehensiveness of microbial databases. *BMC Biol* 18: 37

Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71: 8228–8235

Lu J, Breitwieser FP, Thielen P & Salzberg SL (2017) Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci* 3: e104

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, *et al* (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395: 565–574

Luo C, Knight R, Siljander H, Knip M, Xavier RJ & Gevers D (2015) ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* 33: 1045–1052

Ma S (2019) MMUPHin Bioconductor

Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM & Jansson JK (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480: 368–371

Manara S, Asnicar F, Beghini F, Bazzani D, Cumbo F, Zolfo M, Nigro E, Karcher N, Manghi P, Metzger MI, *et al* (2019) Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biol* 20: 299

Manara S, Pasolli E, Dolce D, Ravenni N, Campana S, Armanini F, Asnicar F, Mengoni A, Galli L, Montagnani C, *et al* (2018) Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of Staphylococcus aureus strains in a paediatric hospital. *Genome Med* 10: 82

McFarland LV & Bernasconi P (1993) Saccharomyces boulardii'. A Review of an Innovative Biotherapeutic Agent. *Microb Ecol Health Dis* 6: 157–171

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P & Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 4: 63–72

McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, *et al* (2017) Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol* 18: 182

McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, Segata N & Huttenhower C (2018) bioBakery: a meta'omic analysis environment. *Bioinformatics* 34: 1235–1237

MetaSUB International Consortium (2016) The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome* 4: 24

Meyer F, Bremges A, Belmann P, Janssen S, McHardy AC & Koslicki D (2019) Assessing taxonomic metagenome profilers with OPAL. *Genome Biol* 20: 51

Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A & McHardy AC (2018) AMBER: Assessment of Metagenome BinnERs. *Gigascience* 7

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, *et al* (2008) The metagenomics RAST server--a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 1–8

Mikheenko A, Saveliev V & Gurevich A (2016) MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32: 1088–1090

Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, *et al* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 10: 1014

Mitra S, Stärk M & Huson DH (2011) Analysis of 16S rRNA environmental sequences using MEGAN. *BMC Genomics* 12 Suppl 3: S17

Morgan XC, Segata N & Huttenhower C (2013) Biodiversity and functional genomics in the human microbiome. *Trends Genet* 29: 51–58

Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, *et al* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 13: R79

Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, Stewart CJ, Metcalf GA, Muzny DM, Gibbs RA, *et al* (2017) The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome* 5 doi:10.1186/s40168-017-0373-4 [PREPRINT]

Nasko DJ, Koren S, Phillippy AM & Treangen TJ (2018) RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol* 19: 165

Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, Pollard KS & Sharpton TJ (2015) Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. *PLoS Comput Biol* 11: e1004573

Nayfach S, Rodriguez-Mueller B, Garud N & Pollard KS (2016) An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res* 26: 1612–1625

Nayfach S, Shi ZJ, Seshadri R, Pollard KS & Kyrpides NC (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568: 505–510

Nazeen S, Yu YW & Berger B (2020) Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. *Genome Biol* 21: 47

NCBI (2020) Genome List - NCBI. *https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/*

NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 42: D7-17

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, *et al* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 32: 822–828

Nurk S, Meleshko D, Korobeynikov A & Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27: 824–834

Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, *et al* (2015) Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat Commun* 6: 6505

Oh J, Byrd AL, Deming C, Conlan S, NISC Comparative Sequencing Program, Kong HH & Segre JA (2014) Biogeography and individuality shape function in the human skin metagenome. *Nature* 514: 59–64

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens HH, Wagner H, Oksanen MJ & Suggests M (2008) The vegan package. *Community ecology package* 10

Olm MR, West PT, Brooks B, Firek BA, Baker R, Morowitz MJ & Banfield JF (2019) Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* 7: 26

Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S & Pop M (2019) Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief Bioinform* 20: 1140–1150

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S & Phillippy AM (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 17: 1–14

Ounit R, Wanamaker S, Close TJ & Lonardi S (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16: 236

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, *et al* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33: 5691–5702

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, *et al* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42: D206-14

Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, Kodira C, Mohiuddin M, Brunelle J, Driscoll M, *et al* (2014) Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Front Microbiol* 5: 298

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P & Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25: 1043–1055

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P & Tyson GW (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2: 1533–1542

Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, *et al* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176: 649-662.e20

Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, *et al* (2017) Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 14: 1023–1024

Pasolli E, Truong DT, Malik F, Waldron L & Segata N (2016) Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput Biol* 12: e1004977

Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T & McHardy AC (2011) Taxonomic metagenome sequence assignment with structured output models. *Nat Methods* 8: 191–192

Patwa LG, Fan T-J, Tchaptchet S, Liu Y, Lussier YA, Sartor RB & Hansen JJ (2011) Chronic intestinal inflammation induces stress-response genes in commensal Escherichia coli. *Gastroenterology* 141: 1842-51.e1–10

Peabody MA, Van Rossum T, Lo R & Brinkman FSL (2015) Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics* 16: 363

Peng Y, Leung HCM, Yiu SM & Chin FYL (2011) Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 27: i94-101

Petschow B, Doré J, Hibberd P, Dinan T, Reid G, Blaser M, Cani PD, Degnan FH, Foster J, Gibson G, *et al* (2013) Probiotics, prebiotics, and the host microbiome: the science of translation. *Ann N Y Acad Sci* 1306: 1–17

Pevzner PA, Tang H & Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* 98: 9748–9753

Piro VC, Dadi TH, Seiler E, Reinert K & Renard BY (2020) ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. 406017

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, *et al* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* 42: D231-9

Power RA, Parkhill J & de Oliveira T (2017) Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet* 18: 41–50

Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, Kearney SM, Perrotta AR, Berdy B, Zhao S, Lieberman TD, *et al* (2019) A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* 25: 1442–1452

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, *et al* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464: 59–65

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, *et al* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490: 55–60

Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, *et al* (2014) Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513: 59–64

Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G & Eren AM (2017a) DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol* 18: 181

Quince C, Walker AW, Simpson JT, Loman NJ & Segata N (2017b) Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 35: 833–844

Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, Brigidi P, Crittenden AN, Henry AG & Candela M (2015) Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr Biol* 25: 1682–1693

Rath S, Rud T, Pieper DH & Vital M (2019) Potential TMA-Producing Bacteria Are Ubiquitously Found in Mammalia. *Front Microbiol* 10: 2966

Relman DA (2013) Metagenomics, infectious disease diagnostics, and outbreak investigations: sequence first, ask questions later? *JAMA* 309: 1531–1532

Rho M, Tang H & Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 38: e191

Rice P, Longden I & Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16: 276–277

Saary P, Mitchell AL & Finn RD (2020) Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol* 21: 244

Schaubeck M, Clavel T, Calasan J, Lagkouvardos I, Haange SB, Jehmlich N, Basic M, Dupont A, Hornef M, von Bergen M, *et al* (2016) Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut* 65: 225–237

Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL & Segata N (2016) Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* 13: 435–438

Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, *et al* (2017) Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat Methods* 14: 1063–1071

Segata N (2015) Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. *Curr Biol* 25: R611–R613

Segata N (2018) On the Road to Strain-Resolved Comparative Metagenomics. *mSystems* 3

Segata N, Baldini F, Pompon J, Garrett WS, Truong DT, Dabiré RK, Diabaté A, Levashina EA & Catteruccia F (2016) The reproductive tracts of two malaria vectors are populated by a core microbiome and by gender- and swarm-enriched microbial biomarkers. *Scientific Reports* 6 doi:10.1038/srep24207 [PREPRINT]

Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS & Huttenhower C (2013) Computational meta'omics for microbial community studies. *Mol Syst Biol* 9: 666

Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, Huttenhower C & Izard J (2012a) Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol* 13: R42

Segata N & Huttenhower C (2011) Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One* 6: e24704

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O & Huttenhower C (2012b) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 9: 811–814

Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, Kumar N, Stares MD, Rodger A, Brocklehurst P, *et al* (2019) Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* 574: 117–121

Silva GGZ, Cuevas DA, Dutilh BE & Edwards RA (2014) FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2: e425

Silva GGZ, Green KT, Dutilh BE & Edwards RA (2016) SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics* 32: 354–361

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV & Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212

Sivan A, Corrales L, Hubert N, Williams JB, Aquino-Michaels K, Earley ZM, Benyamin FW, Lei YM, Jabri B, Alegre M-L, *et al* (2015) Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science* 350: 1084–1089

Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, Hohmann EL, Staley C, Khoruts A, Sadowsky MJ, *et al* (2018) Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe* 23: 229-240.e5

Sommer F & Bäckhed F (2013) The gut microbiota--masters of host development and physiology. *Nat Rev Microbiol* 11: 227–238

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313

Steinegger M & Söding J (2018) Clustering huge protein sequence sets in linear time. *Nat Commun* 9: 2542

Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R & Watson M (2019) Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol* 37: 953–961

Strous M, Kraft B, Bisdorf R & Tegetmeyer HE (2012) The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol* 3: 410

Sun L, Xie C, Wang G, Wu Y, Wu Q, Wang X, Liu J, Deng Y, Xia J, Chen B, *et al* (2018) Gut microbiota and intestinal FXR mediate the clinical benefits of metformin. *Nat Med* 24: 1919–1929

Sunagawa S, Coelho LP, Chaffron S & Kultima JR (2015) Structure and function of the global ocean microbiome.

Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, *et al* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* 10: 1196–1199

Suzek BE, Huang H, McGarvey P, Mazumder R & Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282–1288

Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH & UniProt Consortium (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31: 926–932

Tang WHW, Wang Z, Levison BS, Koeth RA, Britt EB, Fu X, Wu Y & Hazen SL (2013) Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N Engl J Med* 368: 1575–1584

Tanoue T, Morita S, Plichta DR, Skelly AN, Suda W, Sugiura Y, Narushima S, Vlamakis H, Motoo I, Sugita K, *et al* (2019) A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* 565: 600–605

Tatusov RL, Galperin MY, Natale DA & Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28: 33–36

Teeling H, Waldmann J, Lombardot T, Bauer M & Glöckner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5: 163

Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, Armanini F, Manghi P, Bonham K, Zolfo M, *et al* (2019) The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* 26: 666-679.e7

The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 47: D330–D338

Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, *et al* (2019) Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med* 25: 667–678

Thomas AM & Segata N (2019) Multiple levels of the unknown in microbiome research. *BMC Biol* 17: 48

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, *et al* (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551: 457–463

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C & Segata N (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 12: 902–903

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

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R & Gordon JI (2007) The human microbiome project. *Nature* 449: 804–810

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS & Banfield JF (2004) Community structure and metabolism

through reconstruction of microbial genomes from the environment. *Nature* 428: 37–43

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47: D506–D515

Van Rossum T, Peabody MA, Uyaguari-Diaz MI, Cronin KI, Chan M, Slobodan JR, Nesbitt MJ, Suttle CA, Hsiao WWL, Tang PKC, *et al* (2015) Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. *Front Microbiol* 6: 1405

Vandeputte D, Tito RY, Vanleeuwen R, Falony G & Raes J (2017) Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol Rev* 41: S154–S167

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66–74

Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. *J Stat Softw* 36: 1–48

Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, Goedert JJ, Shi J, Bork P & Sinha R (2016) Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One* 11: e0155362

Voorhies AA, Mark Ott C, Mehta S, Pierson DL, Crucian BE, Feiveson A, Oubre CM, Torralba M, Moncera K, Zhang Y, *et al* (2019) Study of the impact of long-duration space missions at the International Space Station on the astronaut microbiome. *Sci Rep* 9: 9911

Wallace RJ, Rooke JA, McKain N, Duthie C-A, Hyslop JJ, Ross DW, Waterhouse A, Watson M & Roehe R (2015) The rumen microbial metagenome associated with high methane production in cattle. *BMC Genomics* 16: 839

van der Walt AJ, van Goethem MW, Ramond J-B, Makhalanyane TP, Reva O & Cowan DA (2017) Assembling metagenomes, one community at a time. *BMC Genomics* 18

Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, Keddy KH, Salje H, Moore S, Mukhopadhyay AK, *et al* (2017) Genomic history of the seventh pandemic of cholera in Africa. *Science* 358: 785–789

West PT, Probst AJ, Grigoriev IV, Thomas BC & Banfield JF (2018) Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* 28: 569–580

Wetterstrand KA (2020) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data*

What are proteomes? *UniProt*

Wilson MR, O'Donovan BD, Gelfand JM, Sample HA, Chow FC, Betjemann JP, Shah MP, Richie MB, Gorman MP, Hajj-Ali RA, *et al* (2018) Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. *JAMA Neurol* 75: 947–955

Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, *et al* (2019) Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 25: 679–689

Wood DE, Lu J & Langmead B (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol* 20: 257

Wood DE & Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15: R46

Wu Y-W, Simmons BA & Singer SW (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32: 605–607

Xiong W, Giannone RJ, Morowitz MJ, Banfield JF & Hettich RL (2015) Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. *J Proteome Res* 14: 133–141

Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, *et al* (2019) Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med* 25: 968–976

Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, Selvenius J, Oikarinen S, Hyöty H, Virtanen SM, *et al* (2018) Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* 24: 146-154.e4

Ye SH, Siddle KJ, Park DJ & Sabeti PC (2019) Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* 178: 779–794

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W & Glöckner FO (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 42: D643-8

Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, Tang L, Zhao H, Stenvang J, Li Y, *et al* (2017) Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66: 70–78

Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA & Koonin EV (2018) Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* 3: 38–46

Ze X, Duncan SH, Louis P & Flint HJ (2012) Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME J* 6: 1535–1543

Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, *et al* (2014) Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 10: 766

Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, Wu X, Li J, Tang L, Li Y, *et al* (2015) The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* 21: 895–905

Zhao Y, Tang H & Ye Y (2012) RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28: 125–126

Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, *et al* (2019) Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun* 10: 5477

Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, Bushman FD & Segata N (2019) Detecting contamination in viromes using ViromeQC. *Nat Biotechnol* 37: 1408–1412

Zolfo M, Tett A, Jousson O, Donati C & Segata N (2017) MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res* 45: e7

Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, *et al* (2019) 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* 37: 179–185

Zynda G (2020) ncbi_genomes. *https://github.com/zyndagj/ncbi_genomes*

Section 2

2.1. Large-scale comparative metagenomics of *Blastocystis*, a common member of the human gut microbiome

Introduction to the chapter

The gut microbiome is not composed only by bacteria. And (micro) eukaryotes in the gut are not all parasites. A small portion of the microbiome, the so-called “eukaryome” (Andersen *et al*, 2013), is composed of eukaryotic organisms having interaction with the host that is not only limited to parasitism but also mutualism and commensalism (Lukeš *et al*, 2015). Thanks to recent studies that showed that some species are harmless or even beneficial to the gut , we observed a reversal of the trend of removal of eukaryotic species in the gut by treatments with drugs (Coyle *et al*, 2011). But still, their role in human gut health is largely unexplored. One of these species is *Blastocystis* spp., a microeukaryote living in the human gut but its role is unclear as previous studies report *Blastocystis* spp. both as a pathogen and as harmless.

By leveraging a couple of thousand metagenomes and the available 8 *Blastocystis* spp. subtypes genomes, we investigated its epidemiology and ecology and provided an in-depth genomic analysis of the several available *Blastocystis* spp. subtypes reference genomes and genomes reconstructed from the metagenomes. In this work, we expanded the current knowledge by searching for its presence in more than 2000 metagenomic samples spanning all the continents except Australia and Antarctica. This is the first large-scale investigation of *Blastocystis* spp. conducted using a metagenomic approach, another work by (Andersen *et al*, 2015) investigated its presence using only a small subset of metagenomes considered in this work. We observed Blastocystis presence in around 15% of the subjects and showed that its prevalence is higher in healthy subjects, suggesting that *Blastocystis* may play a positive role in the gut ecology. In particular, by reconstructing 43 *Blastocystis* genomes, we showed that performing metagenomic assembly of microbial eukaryotes is possible, a particular task that enables further integration of these organisms in metagenomic profilers.

For this article, I retrieved all the reference genomes and built the custom pipeline for the calculation of coverage values, the pipeline for the assessment of the limit of detection, the gene prediction and annotation of the retrieved metagenome-assembled genomes and the custom pipeline for the phylogenetic placement. I performed most of the data interpretation and of the writing of the manuscript.

This chapter contains a work published in the following article:

Francesco Beghini, Edoardo Pasolli, Tin Duy Truong, Lorenza Putignani, Simone M Cacciò, Nicola Segata

Large-scale comparative metagenomics of *Blastocystis*, a common member of the human gut microbiome

The ISME Journal (2017) - https://doi.org/10.1038/ismej.2017.139

Abstract

The influence of unicellular eukaryotic microorganisms on human gut health and disease is still largely unexplored. *Blastocystis* spp. commonly colonize the gut, but its clinical significance and ecological role are currently unsettled. We have developed a high-sensitivity bioinformatic pipeline to detect *Blastocystis* subtypes (STs) from shotgun metagenomics, and applied it to 12 large datasets, comprising 1,689 subjects of different geographic origin, disease status and lifestyle. We confirmed and extended previous observations on the high prevalence the microrganism in the population (14.9%), its non-random and ST-specific distribution, and its ability to cause persistent (asymptomatic) colonization. These findings, along with the higher prevalence observed in non-westernized individuals, the lack of positive association with any of the disease considered, and decreased presence in individuals with dysbiosis associated with colorectal cancer and Crohn's disease, strongly suggest that *Blastocystis* is a component of the healthy gut microbiome. Further, we found an inverse association between body mass index and *Blastocystis*, and strong co-occurrence with archaeal organisms (*Methanobrevibacter smithii*) and several bacterial species. The association of specific microbial community structures with *Blastocystis* was confirmed by the high predictability (up to 0.91 Area Under the Curve) of the microorganism colonization based on the species-level composition of the microbiome. Finally, we reconstructed and functionally profiled 43 new draft *Blastocystis* genomes and discovered a higher intra subtype variability of ST1 and ST2 compared to ST3 and ST4. Altogether, we provide an in-depth epidemiologic, ecological, and genomic analysis of *Blastocystis*, and show how metagenomics can be crucial to advance population genomics of human parasites.

Introduction

Blastocystis spp. (referred to as *Blastocystis* in the manuscript) is a unicellular eukaryotic microorganism that belongs to the Stramenopile phylum. This phylum encompasses an extremely large diversity of organisms including free-living flagellates, parasites of plants (e.g., *Peronospora*) and animals (e.g., *Phytium insidiosum*), organisms resembling fungi in terms of cytology and ecology, and a myriad of photosynthetic lineages that range from single-cell diatoms to giant multicellular brown algae (Derelle *et al*, 2016). *Blastocystis* is a common inhabitant of the gut of humans and other animals (Clark *et al*, 2013). Its prevalence in humans varies within and between populations, but it is higher in underdeveloped countries, where it can reach 100% (El Safadi *et al*, 2014). This is likely the result of poor hygiene conditions, contact with animal reservoirs, and consumption of contaminated water or food (Tan, 2008). Isolates of *Blastocystis* from different hosts are morphologically very similar, yet display substantial genetic variability: based on nucleotide differences in the small subunit ribosomal DNA gene, 17 different subtypes (STs) are recognized, nine of which (ST1 to ST9) are associated with human colonization (Alfellani *et al*, 2013b; Tan, 2008). Previous studies reported the presence of *Blastocystis* in all continents (Alfellani *et al*, 2013a), attesting its ubiquitous distribution, but the overall epidemiological picture is still incomplete.

Whether *Blastocystis* is to be considered a pathogen, a commensal or even a beneficial member of the human gut microbiome is still unclear (Lukeš *et al*, 2015). Indeed, some studies have implicated it in intestinal diseases, including Inflammatory Bowel Disease (IBD) and Irritable Bowel Syndrome (IBS), thus supporting a pathogenic potential (Tan *et al*, 2010). Further, genome analysis of an ST7 isolate revealed the presence of genes encoding potential virulence factors, notably hydrolases and serine and cysteine proteases (Denoeud *et al*, 2011). On the other hand, studies of healthy, randomly sampled individuals have shown a high carriage of *Blastocystis* and a prolonged colonization of the gut (Scanlan & Marchesi, 2008; Scanlan *et al*, 2014). Therefore, unbiased large-scale investigations are needed to clarify its role as an aetiological agent of disease, but targeted epidemiological investigations of *Blastocystis* at a global scale are impractical.

Cultivation-free, sequencing-based metagenomic technologies (Morgan *et al*, 2013; Segata *et al*, 2013a; Tyson *et al*, 2004; Venter *et al*, 2004) can potentially overcome some of these issues. Many large-scale metagenomic studies have been performed to characterize the complex consortium of organisms constituting the human gut microbiome, and recent strain-level analyses started to unravel the population structure of bacterial species (Scholz *et al*, 2016; Truong *et al*, 2017) but little attention has been devoted to intestinal parasites (Andersen *et al*, 2013). Indeed, until now, only one investigation (Andersen *et al*, 2015) used a

metagenomic approach to study *Blastocystis* within 316 samples of the MetaHIT dataset (Qin *et al*, 2010) and there is thus the unmet opportunity to exploit larger sets of metagenomes for parasite profiling and epidemiology.

In order to expand the size, genetic depth, and host population diversity of epidemiologic investigation, we developed a bioinformatic pipeline to detect the presence of *Blastocystis* from metagenomes and applied it on 12 published large metagenomic datasets of the human gut microbiome. Overall, 1,689 subjects from 18 different countries and 4 continents (Europe, Africa, Asia and North/South America) were studied, allowing us to survey the prevalence, ST distribution, and genome characteristics of the microorganism, and to investigate its association with disease conditions and the structure of the resident gut bacterial population.

Materials and Methods

Metagenomic datasets and data pre-processing

We analysed 2,154 publicly available gut metagenomic samples from twelve studies. We considered the nine largest metagenomics studies we were aware of and were available as of July 2015 to which we added three additional studies to expand the geographical span of our analysis (**Table 2.1.1**). The selected raw metagenomes were processed with FastqMcf (Aronesty, 2013) by trimming positions with quality < 15, removing low-quality reads (mean quality < 25), and discarding reads shorter than 90 nt. Human DNA and Illumina spike-in DNA (Bacteriophage phiX174) were then removed by using BowTie2 (Langmead & Salzberg, 2012) to map the reads against the reference genomes.

Eight of the considered studies aimed at characterizing the human gut in different health conditions whereas four considered subjects not affected by documented medical conditions (**Table 2.1.1**). We collected and manually curated the main available metadata associated with the samples (Pasolli *et al*, 2016). The metadata fields considered here are BMI, age, gender and disease status. We made the complete metadata table associated with the samples publicly available at https://bitbucket.org/CibioCM/metaml/src.

We performed the analysis using the 9 available genomes of *Blastocystis* subtypes as reference. These include the complete genome sequence of one isolate from ST7 (Denoeud *et al*, 2011) and one from ST4 (accession codes CABX01000000 and JPUL02000000, respectively). Additionally, we used the draft genomes of other STs (ST1, ST2, ST3, ST4, ST6, ST8 and ST9) isolated from humans that have been recently deposited in public databases (accession codes LXWW00000000, JZRJ00000000, JZRK00000000, JZRL00000000, JZRM00000000, JZRN00000000, JZRO00000000). Before using these genomes in our analysis, and because *Blastocystis* sequencing projects are likely to contain DNA from other organisms, we screened all contigs of all assemblies for potential bacterial and archaeal contamination. We did this by mapping with BLASTN the *Blastocystis* assemblies against the set of ~55,000 publicly available archaeal and bacterial genomes. By considering matches over at least 500 nucleotides and a nucleotide identity of at least 90%, we removed all contigs with bacterial or archaeal matches over more than 3% of the length of the contig. Overall, we removed 613 contigs after screening out a minimum of 246,384 nucleotides for ST4 and a maximum of over 4.5 M nucleotides for ST6 (see **Supplementary Table 2.1.1**). We notice that our procedure was set to be quite aggressive in avoiding potential contamination, but this is a safe strategy for our investigation as more than 10 M nucleotides

remained available for all ST and these are largely sufficient to assess the presence of *Blastocystis* in metagenomes as reported below.

Table 2.1.1: List and characteristics of the metagenomic datasets used in this study. Abbreviations: CD Crohn's disease; STEC Shiga-toxigenic Escherichia coli; T2D Type 2 diabetes; UC Ulcerative colitis.

Dataset name	Conditior	Country	#subjects	# total samples	# samples with condition [5]	#total reads (10^9)	#reads per sample (10^6) mean ± std	Age [yrs] median (IQR)	Reference
Candela	Healthy	Italy, Tanzania	38	38	-	0.85	22.3 ± 19.3	30 (23 - 38)	(Rampelli *et al*, 2015)
HMP	Healthy	USA	111	191	-	20.50	108.5 ± 31.7	26 (23 - 28)	(Human Microbiome Proj Consortium, 2012)
Karlsson	T2D	Denmark 10 EU countr	145	145	53	4.49	31.0 ± 17.6	70 (69 - 71)	(Karlsson *et al*, 2013)
LeChatelier	Obesity	Denmark	292	292	169	20.14	69.0 ± 23.2	56 (50 - 61)	(Le Chatelier *et al*, 2013
Liu	Healthy	China, Mongolia	110	110	-	6.41	58.2 ± 26.8	-	(Liu *et al*, 2016)
Loman	STEC infection	Germany	37	44	44	0.39	9.0 ± 12.0	-	(Loman *et al*, 2013)
MetaHIT	CD, UC	Denmark, Spain	124	124	25	5.60	45.1 ± 18.4	54 (49 - 60)	(Qin *et al*, 2010)
Nielsen	CD, UC	Denmark, Spain	318	396	148	21.40	53.9 ± 20.2	49 (40 - 59)	(Nielsen *et al*, 2014)
Obregon-Tito	Healthy	Peru, USA	58	58	-	2.73	47.1 ± 20.8	26 (17 - 35)	(Obregon-Tito *et al*, 201
Qin	Liver cirrhosis	China	237	237	123	12.24	51.6 ± 30.9	45 (38 - 54)	(Qin *et al*, 2014)
T2D	T2D	China	363	363	170	14.60	40.2 ± 11.8	48 (38 - 57)	(Qin *et al*, 2012)
Zeller	Colorecta cancer	France	156	156	53	9.37	60.0 ± 25.4	63 (58 - 70)	(Zeller *et al*, 2014)
Total			1689	2154	785	118.72	55.12 ± 29.0	49 (36 - 62)	

[5] Except for condition "healthy"

Detection of *Blastocystis* STs from metagenomes

Metagenomic reads were mapped to reference genomes using the Bowtie2 aligner (Langmead & Salzberg, 2012) and an end-to-end alignment for paired ends reads. The Bowtie2 output was processed by Samtools (Li *et al*, 2009) and the sorted and indexed BAM file was processed with BEDtools (Quinlan, 2014) to compute the breadth of coverage for each subtype ("genomecov –bg" parameter), which represents the fraction of the target genome covered by at least one metagenomic read (Molnar & Ilie, 2015). The relative abundance in subjects colonized over two timepoints was estimated by counting the number of reads mapped to the *Blastocystis* reference genome normalized by the total number of reads in the sample.

In this work, we define a sample as positive for a *Blastocystis* ST if its genome has a breadth of coverage of at least 10%. This value was chosen based both on (i) the similarity between the genomes of different *Blastocystis* STs and (ii) on the false positive detection rate for the presence of a second *Blastocystis* ST when another one is present. For the first criteria, we quantified the average fraction of the genome of a *Blastocystis* ST shared at a sequence similarity higher than 80% with a distinct *Blastocystis* ST genome using LAST (Kiełbasa *et al*, 2011) ("-l 100 -f BlastTab" parameters). The maximum fraction of matching genome was 3%, with the only exceptions of ST4-ST8 and ST6-ST9 which share more than 15% of the genome. However, this value substantially decreases at percentage identity thresholds >80% which is a very conservative threshold considering that the maximum identity at which a read of 100 nt can be mapped against a reference genome is 95%. Additionally, at the 10% breadth of coverage threshold, we did not find any co-occurrence of ST4 and ST8 in the samples, and for the cases in which ST6 and ST9 co-occurred we manually confirmed that most of the reads outside the shared genomic regions mapped only against the ST with the highest breadth of coverage. For the second criteria, we looked at the distribution of the number of additional STs in addition to the one with the largest breadth of coverage detected when varying the threshold (**Supplementary Figure 2.1.9**). This distribution goes from seven (all the STs in addition to the dominant one) to one (only the dominant ST detected), but it is already plateauing at 10% breadth of coverage confirming that such value does not produce false positives. Multiple lines of evidence thus support the 10% breadth of coverage value to be safe in avoiding false positives. False negatives would be minimized at lower threshold value, but false negatives are arguably less problematic than false positives, and false negatives are an intrinsic and unavoidable problem in metagenomics.

Assessing the limit of detection for Blastocystis in metagenomes

To assess the sensitivity of our procedure in detecting *Blastocystis*, we performed semi-synthetic experiments by spiking-in known amounts of synthetic reads from known *Blastocystis* genomes into real *Blastocystis*-negative gut metagenomes. For each ST, the synthetic reads were obtained with an Illumina-based sequencing simulator with typical sequencing error rates and noise (McElroy *et al*, 2012). As real *Blastocystis*-negative gut metagenomes we considered metagenomes from the HMP, Karlsson, LeChatelier, and Obregon-Tito datasets subsampled after QC to the typical metagenome size of 50M reads. The procedure was repeated at multiple fractions of *Blastocystis* relative abundance from 0.001% to 1% (for a total of 30 abundance values) and considering seven distinct real gut microbiomes for each simulation and ST. With this analysis (**Supplementary Figure 2.1.1**), we empirically found that the chosen detection threshold (10% breadth of coverage) corresponds to a limit of detection slightly below 0.03% abundance. ST7 has an even lower limit of detection which is due to the length of its genome (about 50% larger than the other STs). As mentioned above, our *Blastocystis* detection pipeline aims at minimizing the false positive rate, so even though thresholds lower than 10% breadth of coverage would positively impact the limit of detection, we again preferred to avoid calling the presence of *Blastocystis* without strong quantitative evidence. The limit of the detection of our procedure is higher than what can be achieved with PCR-based approaches, that are however limited in the amount of genomic information that they can provide.

Metagenomic assembly and *Blastocystis* contig binning

The 43 metagenomic samples in which we detected a breadth of coverage higher than 66.6% for at least one *Blastocystis* genome, were selected for *de novo* metagenomic assembly. This was performed using SPAdes version 3.9.0 (Bankevich *et al*, 2012). Contigs shorter than 1,000 nt were discarded, and contigs from *Blastocystis* identified by mapping with BLASTN the screened contigs against the *Blastocystis* reference genomes. Specifically, we assigned a contig to a *Blastocystis* subtype if it had at least 90% identity over at least half of its length against the available reference genome.

Whole genome phylogenetic analysis

To infer the phylogeny of the newly assembled genomes we adopted a core-gene based strategy (Page *et al*, 2015; Segata *et al*, 2013b). The core gene set was generated by aligning all the annotated genes of the *Blastocystis* ST4 WR1 genome against all 8 available reference genomes and the 43 genomes we newly assembled using BLASTN (Evalue: 1e-50, word size:9). To be included in the core gene set, a gene was required to be present in at least 75%

of the analysed genomes with an identity higher than 65% over at least 600bp. The identified core gene sequences were then aligned using MUSCLE (Edgar, 2004), concatenated in a single alignment, and processed with trimAL ("-gappyout" parameter) (Capella-Gutiérrez *et al*, 2009) to remove excessively gapped sub alignments and poorly aligned regions. The phylogeny was then built using RAxML version 8.1.15 (Stamatakis, 2014) with the GTRGAMMA model and 100 bootstrap steps.

Using this approach, we identified a core gene set of 9 genes (average alignment length of each gene of 2,443 bp and standard deviation of 1,374 bp) for a total concatenated alignment length of 21,984 bp. To improve the resolution at a lower phylogenetic level, we repeated the process within the genomes of ST1, ST2, ST3 and ST4 separately and reconstructed their intra subtype phylogeny on which a larger shared core genome can be identified. Subtype-specific trees were generated by fragmenting each genome in portions of 2,000bp and treating them as genes because no genome annotation was available and de-novo annotation would have introduced biases. Criteria for the inclusion in the core gene adapted to the intra-ST case included the requirement that a sequence was present in all the genomes with an identity higher than 95%. Single nucleotide variant distribution within every subtype was calculated using nucmer (Kurtz *et al*, 2004) pipeline for computing pairwise alignment and SNV reporting. For ST1 the average pairwise alignment was 3,431,933 bp (s.d. 2,323,310 bp), for ST2 3,876,563 bp (s.d. 1,985,719 bp), for ST3 2,318,013 bp (s.d. 2,917,467 bp), and for ST4 3,432,010 bp (s.d. 3,131,603 bp).

Functional prediction and annotation

We considered the 19 reconstructed genomes accounting >5 Mbp for gene prediction and annotation. *Ad-initio* gene prediction was performed using SNAP (Korf, 2004) to generate HMM models for all the STs using the available annotations to build the HMM reference profile. Genome annotation was performed using MAKER (Cantarel *et al*, 2008) with default parameters using the HMM models previously generated. Newly predicted proteins were then functionally annotated with eggnog-mapper (Huerta-Cepas *et al*, 2016a) using the eggNOG (Huerta-Cepas *et al*, 2016b) Eukaryotic dataset. ST-specific KOG functions were determined by performing a Fisher's exact test between the genomes of a particular ST and the other ones. Adjusted p-values were computed through the false-discovery rate correction.

Microbiome profiling and co-occurrence analysis

All samples were processed with MetaPhlAn2 (Segata *et al*, 2012b; Truong *et al*, 2015) to quantitatively profile the whole microbial population exploiting the properties of species-specific markers (Huang *et al*, 2014). We used the obtained abundance profiles to investigate

the co-occurrence or co-exclusion (Faust *et al*, 2012) of *Blastocystis* with other members of the microbiome. In particular, the Wilcoxon rank-sum test was used to identify the microbial features that were associated with the presence or absence of *Blastocystis*. In computing this test, duplicates from the same subject were discarded and a threshold of 0.05 was considered as significance level. Additional analysis for finding bacterial clades associated with *Blastocystis* presence was performed using the LEfSe (LDA effect size) tool (Segata *et al*, 2011). Finally, a machine learning-based approach was applied to further investigate if the microbiome signature is predictive for the presence of *Blastocystis*. The species abundances generated by MetaPhlAn2 were used to discriminate between *Blastocystis* positive and negative samples. For this purpose we considered a random forest (RF) classifier (Breiman, 2001) implemented in the MetAML tool (Pasolli *et al*, 2016). First, prediction accuracies were assessed by an unbiased 10-fold cross validation procedure, repeated and averaged over 20 independent runs. Then, we applied a leave-one-dataset-out approach, in which the presence of *Blastocystis* in a given dataset is predicted by training the model on the samples from the other independent studies. Prediction accuracies were evaluated in terms of area under the ROC curve (AUC) statistics, which can be interpreted as the probability that the classifier ranks a randomly chosen positive sample higher than a randomly chosen negative one, assuming that the positive sample ranks higher than the negative one. The free parameters of the classifiers were set as follows: i) the number of decision trees was equal to 500; ii) the number of features to consider when looking for the best split was equal to the root of the number of original features; iii) the quality of a decision tree split was measured using the Gini impurity criterion. The software framework used for this experiment is open-source and available online at http://segatalab.cibio.unitn.it/tools/metaml. Alpha diversity was computed for each dataset by considering Gini-Simpson and Shannon indexes under the condition of presence or absence of *Blastocystis* and the Student's t-test (significance level set to 0.05) was used to test significance between the two conditions.

Results

Meta-analysis for *Blastocystis* in large metagenomic datasets

We screened large-scale intestinal metagenomic datasets to assess the prevalence of *Blastocystis* and its STs, infer epidemiologic characteristics, and examine the characteristics of their genomes. Overall, we processed 2,154 faecal microbiome samples from 1,689 subjects from 12 datasets (**Table 2.1.1**). These datasets span diverse disease conditions including colorectal cancer (Zeller *et al*, 2014), type 2 diabetes (Karlsson *et al*, 2013; Qin *et*

al, 2012), liver cirrhosis (Qin *et al*, 2010), obesity (Le Chatelier *et al*, 2013), and IBD (Nielsen *et al*, 2014; Qin *et al*, 2010). All these studies almost exclusively focused on the bacterial components of the microbiome and did not report the presence of microbial Eukaryotes, with the above mentioned exception that focused on a single metagenomic dataset (Andersen *et al*, 2015). The wide range of distinct health conditions and geographic origins of the hosts we considered here are thus a key factor in this study.

To assess the presence of *Blastocystis,* we used a sequence mapping based approach aided by the availability of draft genome sequences from eight subtypes (ST1, ST2, ST3, ST4, ST6, ST7, ST8 and ST9), all known to be associated with human colonization. After removing potential bacterial sequences contaminating these genomes (see **Methods** and **Supplementary Table 2.1.1**), we estimated the fraction of each target genome covered by metagenomic reads (i.e., the breadth of coverage) and we considered samples positive for *Blastocystis* when the breadth of coverage was higher than 10% (see **Methods**). Using this approach, *Blastocystis* is detected when present at a concentration as low as 0.03% in typical metagenomic samples of 50M reads (**Supplementary Figure 2.1.1**). Downstream analyses detailed in the rest of the work are based on this detection threshold.

Blastocystis prevalence and subtype dominance is biogeographically variable

We first determined the prevalence of *Blastocystis* in the overall dataset, which included 2,154 faecal samples from 1,689 subjects. The microorganism was detected in 321 samples, originating from subjects in ten countries (China, Denmark, France, Mongolia, Norway, Peru, Spain, Sweden, Tanzania and USA) from four continents, with an overall prevalence of 14.9%. The prevalence was higher in European subjects (243 of 1084 samples, 22.4%) and lower in Chinese ones (24 of 600, 4.0%). Despite the relatively small size of the dataset, 15 (55.6%) of the 27 Tanzanian subjects (Rampelli *et al*, 2015) were positive for *Blastocystis*, whereas all the Italian subjects (n=11) from the same study were negative. *Blastocystis* was not detected in the Shiga toxin-producing *E. coli* (STEC)-infection dataset (Loman *et al*, 2013).

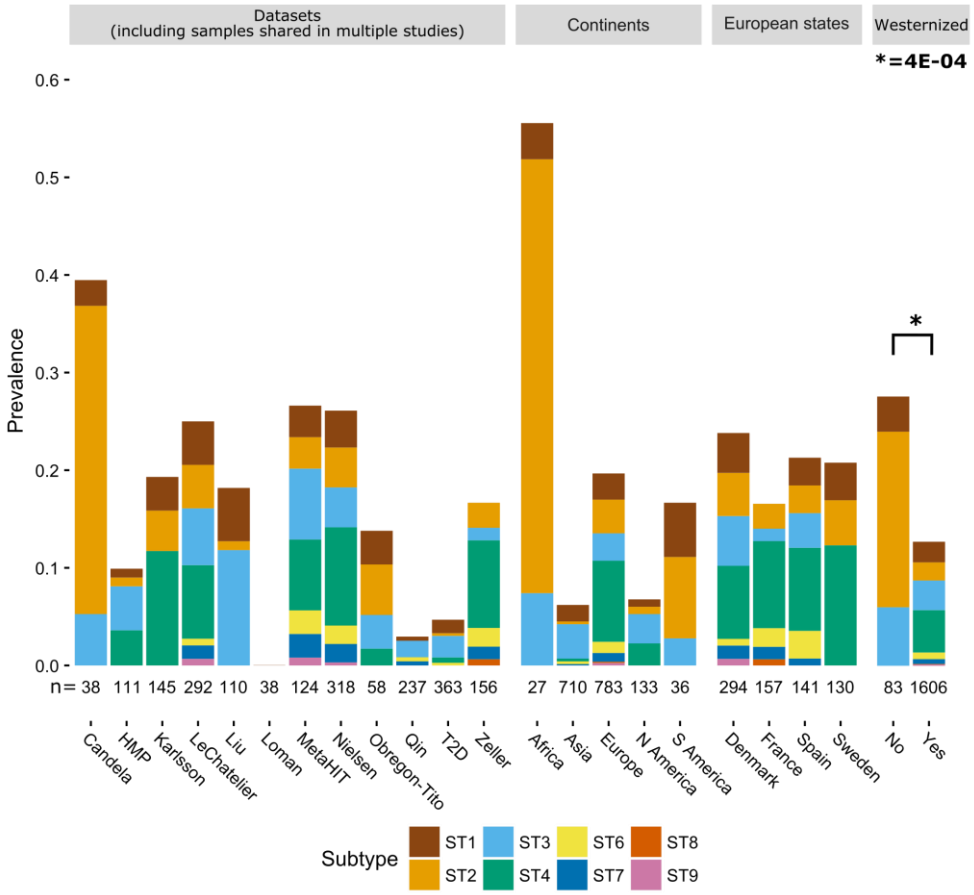


Figure 2.1.1: Prevalence of *Blastocystis* and *Blastocystis* subtypes in the different datasets (A), different continents (B), different European states (C), and between westernized and non-westernized subjects (D) (see **Supplementary Table 2** for more details)**.** Stacked barplots show the prevalence in each category; numbers below the bars refer to the number of samples in the corresponding category, where duplicates from the same subject are eventually discarded. Statistical significance was assessed by Fisher's exact test.

The prevalence of *Blastocystis* appears to be influenced by the DNA extraction procedure used in the different studies, being higher when methods combining mechanical and chemical lysis steps are used (**Supplementary Figure 2.1.2**). This suggests that efficient DNA extraction from lysis-resistant microorganism cysts requires appropriate procedures, and that comparison across studies should consider this factor (Yoshikawa *et al*, 2011). On the other hand, cohort differences may have a larger impact on prevalence than methodological aspects, as exemplified by large differences in prevalence between three European datasets (LeChatelier, MetaHIT and Nielsen) and the Chinese T2D dataset, despite the use of the same DNA extraction procedure.

We then examined the prevalence of the different *Blastocystis* STs among individuals colonized with single STs (**Figure 2.1.1, Supplementary Table 2.1.2**). While some aspects such as the wide geographic distribution of ST3 (detected in 10 of 12 datasets), and the overall low prevalence of ST6, ST7 and ST9, are in agreement with the current global epidemiologic information (Clark *et al*, 2013), two new points of particular relevance emerged. First, ST2

appears to predominate in the non-industrialized cohorts analysed, which are hunter-gatherer populations from Tanzania (Rampelli *et al*, 2015) and Peru (Obregon-Tito *et al*, 2015)(**Figure 2.1.1**). The difference in ST2 prevalence between non-westernized (including data from (Liu *et al*, 2016)) and westernized individuals is highly statistically significant (p = 5.75E-10). This raises the hypothesis that ST2 is one of the members of the gut microbiome that have been affected by westernization processes (Segata, 2015). Second, the prevalence of ST4 is very high among European subjects (**Figure 2.1.1**), which is in sharp contrast with the absence, or extreme rarity, of this ST in other regions of the world (e.g., p = 6.57E-16 for the difference in prevalence between Europe and Asia, **Supplementary Figure 2.1.3**), except the US. The difference in ST4 prevalence between westernized and non-westernized individuals is also statistically significant ($p < 0.046$). These data confirm and extend previous observations on the peculiar geographical distribution of ST4 (Forsell *et al*, 2012). Overall, ST2 and ST4 thus appear to be the *Blastocystis* subtypes most influenced by geography and lifestyles.

Blastocystis prevalence is higher in subjects with low BMI and in healthy controls for Crohn's disease and colorectal cancer

We tested the association between the presence of *Blastocystis* and available parameters of interest (see **Methods**), and found that body mass index (BMI) is strongly negatively correlated with *Blastocystis* prevalence. In the metagenomic study that specifically targeted the obesity phenotype (Le Chatelier *et al*, 2013), we detected *Blastocystis* in 39.4% normal weight individuals, compared to 15.4% obese subjects (p = 2E-05, **Figure 2.1.2A**). This is consistent with findings from a study of Danish subjects (Andersen *et al*, 2015). The other datasets include a smaller number of obese subjects, thus providing less statistical power to test the association. Nonetheless, a higher *Blastocystis* prevalence in normal weight individuals compared to overweight and obese ones was evident in six of the eight datasets, two of which supported by statistical significance (**Figure 2.1.2A**).

Interestingly, when considering all the European datasets that used the same collection and processing protocols (n=715, 126% more samples than (Andersen *et al*, 2015)), the difference in *Blastocystis* prevalence between normal weight and obese subjects was again strongly significant (p = 5E-03), as it was between normal weight and overweight (p = 0.011), and between non-overweight and overweight (p = 0.015). At the level of specific subtypes, only ST4 reached statistical significance (p = 0.03 between normal weight and obese), suggesting that association between *Blastocystis* and BMI is probably not subtype-specific.

We did not find an increased prevalence of *Blastocystis* in subjects affected by any of the considered diseases (lowest one-side p = 0.4 for T2D in the Karlsson dataset). Conversely,

Blastocystis was positively associated (p = 0.008) with the control group in the colorectal cancer dataset (Zeller *et al*, 2014), with only 3 of the 53 (5.7%) patients positive for the microorganism compared to 15 of the 61 (24.6%) healthy controls (**Figure 2.1.2B**). This trend was also confirmed in the same dataset when considering patients with large adenomas (n=14), as only one was positive for *Blastocystis*. Crohn's disease, but not ulcerative colitis, was also negatively associated with the presence of *Blastocystis* (p = 0.04). Our findings seem to contrast other reports especially for colorectal cancer (Kumarasamy *et al*, 2014), whereas for IBD existing data already associated ulcerative colitis rather than Crohn's disease with decreased *Blastocystis* prevalence (Petersen *et al*, 2013) although different conclusions were reached in other reports (Cekin *et al*, 2012).

Previous data on this association are however sparse, debated in clinical settings, and potentially affected by publication bias. More independent investigations are needed to elucidate these relations, but our results suggest that the ecological niche of *Blastocystis* is independent from disease-associated microbiome dysbiosis features. A further hypothesis supported by the above associations and the absence of *Blastocystis* in STEC-positive subjects, is that *Blastocystis* is actually less common in individuals with gastro-intestinal symptoms and other microbiome-associated disease conditions (Scanlan *et al*, 2014).

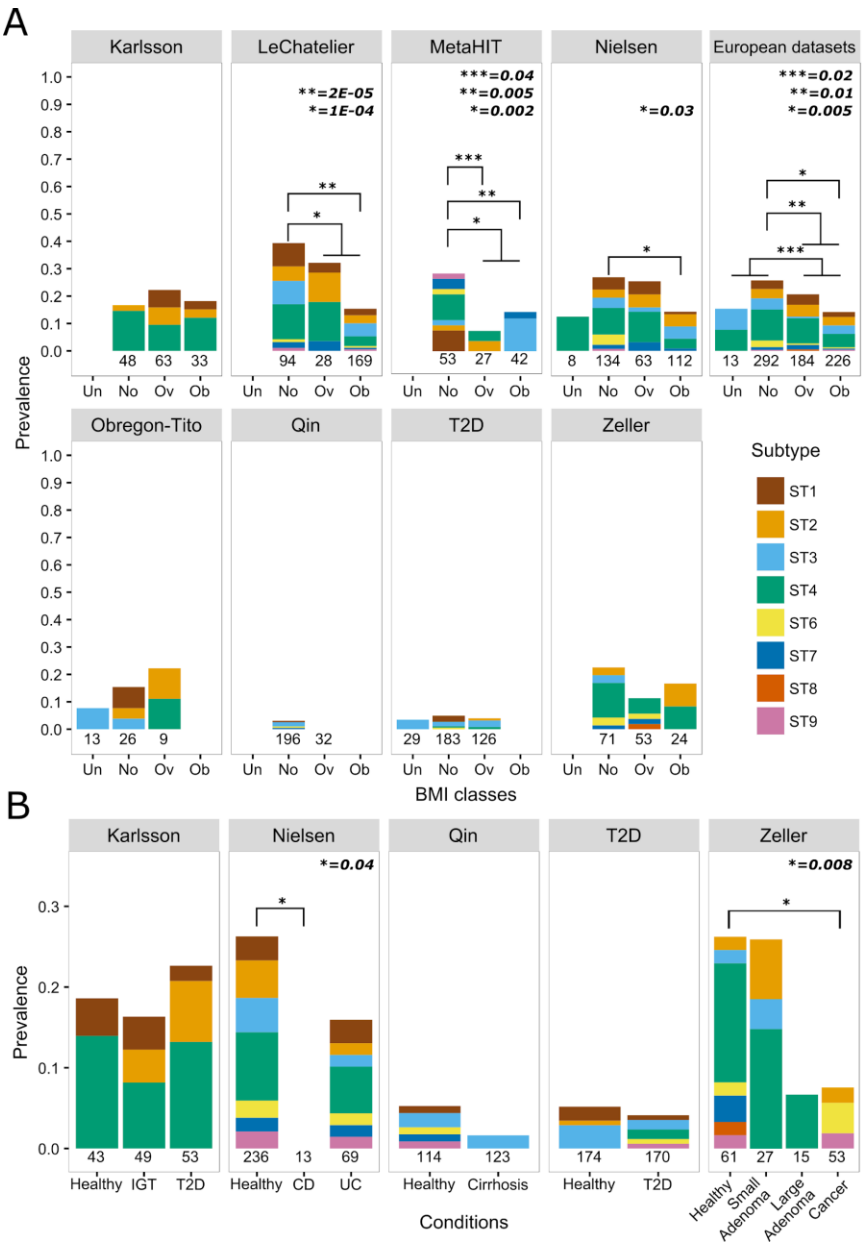


Figure 2.1.2. Blastocystis prevalence in BMI classes (A) and different health conditions (B) for the considered datasets. Barplots show the prevalence of Blastocystis in different health conditions reported in the analysed datasets. BMI classes considered were underweight (Un), normal (No), overweight (Ov) and obese (Ob). The total number of samples in each class and dataset is reported below the bars. Bars associated with a total number of samples less than four are not shown. Note that scales in panels A and B are different. Abbreviations: CD Crohn's disease; IGT Impaired Glucose Tolerance; T2D Type 2 diabetes; UC Ulcerative colitis. Fisher's exact test was used as statistical significance test.

Stable Blastocystis colonization is subtype-independent

To study the persistence of *Blastocystis* colonization and determine the subtypes involved in chronic colonization, we analysed the metagenomic dataset of subjects who provided stool samples at multiple timepoints. A total of 121 subjects, 43 from the HMP dataset (Human Microbiome Project Consortium, 2012) and 78 from the Nielsen dataset (Nielsen *et al*, 2014), were sampled at two timepoints (mean 219 and 163 days after first sampling, respectively).

Blastocystis was identified above the detection threshold in 22 subjects (7 from HMP and 15 from Nielsen) in at least one of the timepoints considered (**Supplementary Table 2.1.3**). Of the 22 positive subjects, 14 (64%) maintained the colonization over the two timepoints, whereas 5 subjects acquired and 3 subjects lost the colonization between the two timepoints (**Figure 2.1.3A, Supplementary Table 2.1.3**). For the cases of colonization acquisition/loss and accordingly with our detection limit of 0.03% relative abundance, *Blastocystis* is indeed absent in the subject or may be present at very low abundance which is still indicative of variations in the ecological relation of *Blastocystis* with the resident microbiome. In subjects with stable colonization, the relative abundance of *Blastocystis* changed only slightly in the majority of the cases (**Figure 2.1.3B**) and we did not observe variations higher than three folds.

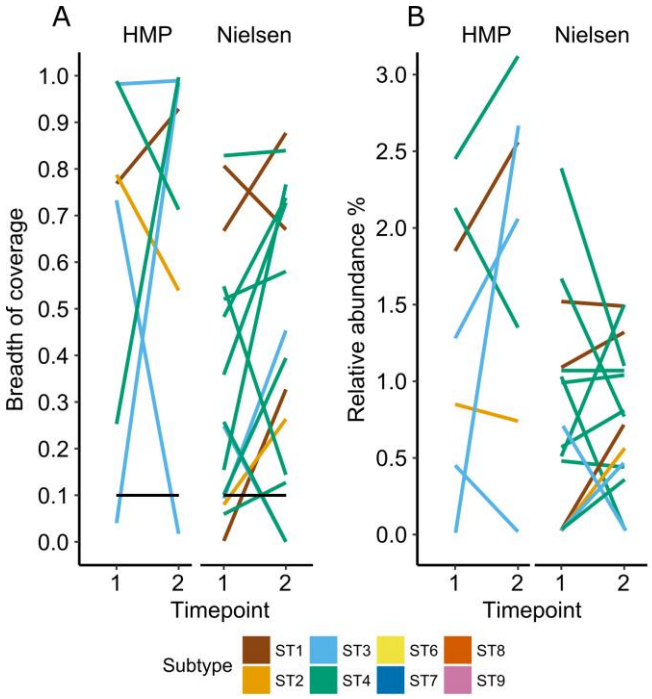


Figure 2.1.3: Breadth of coverage (A) and relative abundance (B) of Blastocystis in subjects colonized over two timepoints (see **Supplementary Table 3** for more details). In the breadth of coverage plots, samples below the threshold of detection are also indicated. The breadth of coverage represents the fraction of the reference genome covered by at least one metagenomic read. The relative abundance is estimated by dividing the number of reads mapped to the Blastocystis reference genome with the total number of reads in the sample.

In the 14 subjects with stable colonization, we always found the same ST at the two timepoints, suggesting that ST replacement is not a frequent event in the healthy human gut, at least over the relatively short timeframes considered in the datasets (**Figure 2.1.3**). The subtypes commonly found in humans (ST1-ST4) all appeared as stable colonizers, suggesting that this phenomenon is not subtype-dependent.

Whole genome genetic analysis of *Blastocystis* subtypes

Isolates belonging to the same *Blastocystis* subtype display some genetic variability, as highlighted by studies of ribosomal markers (Yoshikawa *et al*, 2016) and a few housekeeping genes (Stensvold *et al*, 2012; Yoshikawa *et al*, 2016). However, the extent of polymorphism at the genome level and variability in gene content within different STs is unknown. To this end, we reconstructed draft *Blastocystis* genomes from the metagenomes and performed comparative genomic analysis. In total, 43 assemblies were obtained using metagenomic assembly with SPAdes (Bankevich *et al*, 2012) followed by binning and taxonomic assignment (see **Methods**, **Supplementary Table 2.1.4**) from the samples with very high *Blastocystis* abundance. Specifically, 16 new genomes were very closely related to the available genome of ST4, 7 to ST2 and ST1, 9 to ST3, whereas only 4 genomes were assembled from the phylogenetically related ST6, ST8, and ST9. A simple genetic feature such as the average GC content (**Supplementary Figure 2.1.4**) was already distinctive across STs, in that ST1, ST2, and ST3 that have a genome much richer in GC (average 52.6%, 52.0%, 51.5% respectively) than ST4, ST8, and ST9 (average 40.0%, 42.3%, 41.5% respectively), whereas ST6 is in between these two groups (average 44.9%).

We then integrated the nine available genomes with the 43 new assemblies to reconstruct the genome-scale phylogeny of the *Blastocystis* genus using the concatenation of aligned core genomic fragments. This reconstruction relies upon the substantial fraction of the genome that is conserved across strains and have been performed for assemblies with an average of 4.4 Mb of reconstructed genome (**Supplementary Table 2.1.4**). While the overall structure of the tree (**Figure 2.1.4A**) confirms previous phylogenetic analyses based on single marker genes (Yoshikawa *et al*, 2016) and ST-specific phylogenies consistently place the reconstructed genomes from multiple sample of the same patient (**Figure 2.1.4D**), substantial genetic diversity is detected within each ST (**Figure 2.1.4B-E**). Strains belonging to ST1 show the highest genetic diversity with, on average, 1.5% (s.d. 0.10%) single nucleotide substitutions in the genomic regions conserved between pairs of strains (**Figure 2.1.4F**). ST4 shows instead an overall much higher sequence conservation (average 0.27% s.d. 0.14% divergence), in agreement with findings from single marker genes (Stensvold *et al*, 2012). ST2 and ST3 display intermediate genetic diversity compared to ST1 and ST4.

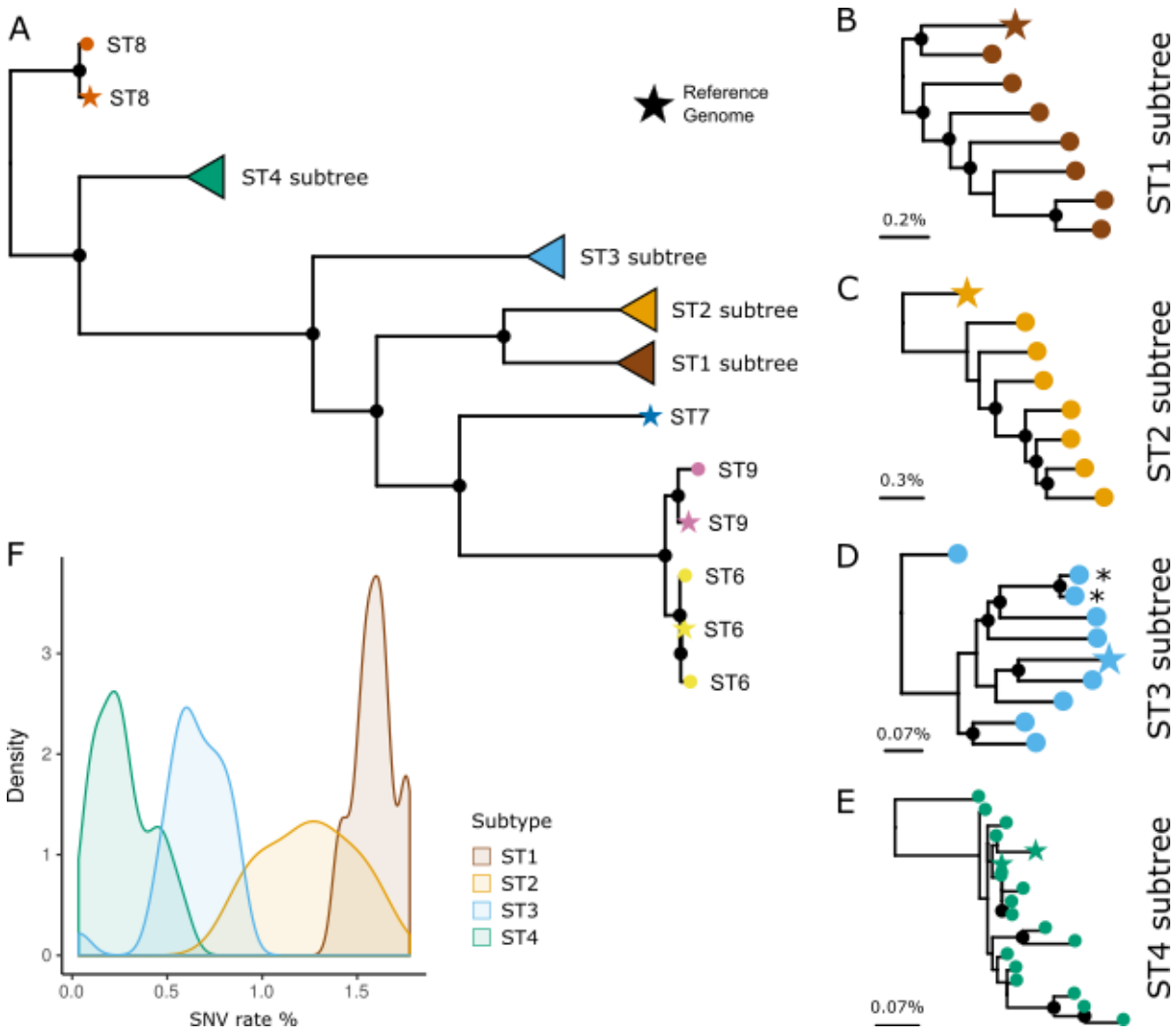


Figure 2.1.4: Phylogenetic relation between the 9 available *Blastocystis* reference genomes and 43 newly reconstructed genome assemblies from metagenomes. From the overall phylogenetic tree (**A**) we also report the subtrees of the four subtypes with more than 3 genomes (**B-E**) and compare the sequence diversity they span (**F**). Maximum likelihood phylogenetic trees were inferred using concatenated aligned shared genomic regions identified in reference genomes and assemblies (see **Methods**). The asterisk highlights samples acquired from the same subject at two different timepoints. Black filled circles denote bootstrap support greater than 80%. The scale bar represents the average SNV rate calculated on the pairwise alignment.

We then restricted the genomic analysis to the 19 genomes for which at least 5Mb have been reconstructed and performed a functional annotation and characterization of these high-quality assemblies (4 for ST1, 4 for ST2, 4 for ST3, and 7 for ST4, see **Supplementary Table 2.1.4**) by using the eggNOG (Huerta-Cepas *et al*, 2016b) database (see **Methods**). Unsurprisingly, less than half of the genes identified were assigned to known COG functional categories (from 42.7% of ST4 to 46.7% of ST3, **Figure 2.1.5A**). Only few categories were not represented in *Blastocystis* (e.g., as expected, the cellular machinery for cell motility) and the four STs generally contained a very similar number of proteins in these broad categories (**Figure 2.1.5B**). The only exceptions are category J (Translation, ribosomal structure and biogenesis) and category A (RNA processing and modification) that are overrepresented in the genomes

of ST3 and underrepresented in those of ST2 (p < 1E-04), as well as categories D (Cell cycle control and mitosis) and T (Signal Transduction, all p < 1E-04).

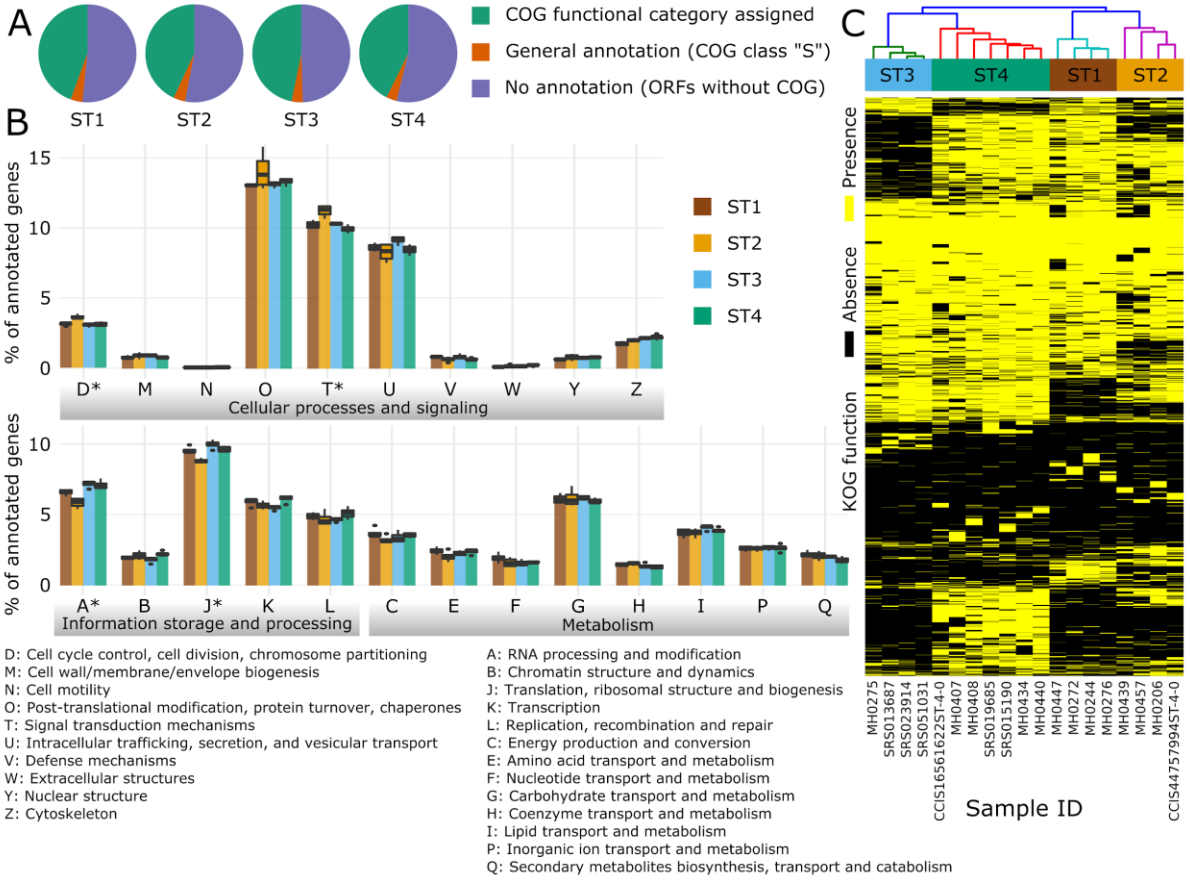


Figure 2.1.5: Functional annotation analysis of the 19 reconstructed genomes spanning four *Blastocystis* STs. Less than half of the genes predicted by MAKER (Cantarel *et al*, 2008) were assigned to known COG functional categories using the eggNOG (Huerta-Cepas et al 2016b) database (**A**, see Methods). These annotated genes can be grouped into 23 broad COG categories (**B**) that are a variable fraction of the total annotated genes. The asterisks denote categories for which one-way ANOVA statistical test gave p < 1E-04. Hierarchical clustering performed on the more specific KOG functions show that samples associated with the same ST cluster together (**C**).

More specific functional assignments based on the manually curated Clusters of Orthologous Groups for Eukaryotes (KOG, see **Supplementary Table 2.1.5**) (Huerta-Cepas *et al*, 2016b) further highlighted the differences in functional potential between STs and the substantial intra-ST functional consistency. This is clear from the hierarchical clustering analysis of the KOG profiles in each reconstructed genome (**Figure 2.1.5C**), in which the close phylogenetic relationship between strains in the same ST is recapitulated at the level of their functional potential. A total of 795 KOGs were found to be ST-specific (**Supplementary Table 2.1.5** and **Supplementary Figure 2.1.5**) after statistical significance testing with false-discovery rate correction (see **Methods**). For example, a cystatin (0IZK7 Cystatin B), that in ST7 has a potential role in parasitic cysteine protease and inhibition of host proteases (Denoeud *et al*, 2011; Wawrzyniak *et al*, 2012), is present in ST2 but not in ST1, ST3, and ST4. Likewise, we

found a glycoside hydrolase (hydrolase family 47) only in ST3, and this may be involved in the attack of the host intestinal epithelial cells (Denoeud *et al*, 2011). Finally, in ST4 genomes we found heat shock proteins (like 0PHA3 and KOG3047 - ubiquitously-expressed, prefoldin-like chaperone) and cytosolic Ca2+-dependent cysteine proteases (like KOG0045 - Calpain-like cysteine peptidase) that were not present in other ST genomes, and these may represent virulence factors unique to this ST. Altogether these data indicate that different *Blastocystis* STs have distinct functional potential niches that are currently only partially characterized. Further, we show for the first time that it is possible to characterize ST-specific functional repertoires that are conserved among strains of the same ST.

The presence of *Blastocystis* is highly correlated with gut microbiome composition

We found a very strong association between the presence of *Blastocystis* and the abundance of archaeal organisms ($p < 6.65E-37$). On the overall dataset, this association may be inflated because some DNA extraction procedures may favour non-bacterial organisms (Wesolowska-Andersen *et al*, 2014), but we observed strong statistical significance in all but two single datasets (**Figure 2.1.6** and **Supplementary Table 2.1.6**). Archaea in the human gut are represented primarily by *Methanobrevibacter smithii* (**Figure 2.1.6A**) which is in fact strongly associated with the presence of *Blastocystis*. Interestingly, several archaeal genes, likely acquired horizontally, are present in the *Blastocystis* genome (Denoeud *et al*, 2011), suggesting that the common ecological niche favours the interaction and the exchange of genetic material between the microorganism and the Archaea.

Several bacterial clades were also found to be strongly associated with *Blastocystis* presence, with a total of 68 significant associations with effect size larger than 3.3 as found by LEfSe analysis (Segata *et al*, 2011) (**Figure 2.1.6E** and **Supplementary Figure 2.1.6**). Bacteria in the *Firmicutes* phylum and in the *Clostridiales* order also appeared strongly enriched in samples positive for *Blastocystis* (**Supplementary Figure 2.1.7**). Species in this order included *Butyrivibrio crossotus* (significant in 7 datasets, **Figure 2.1.6B**), *Eubacterium siraeum* (significant in 6 datasets, **Figure 2.1.6C**), and *Coprococcus catus* (significant in 6 datasets, **Supplementary Figure 2.1.7**). In addition, the overall Clostridiales order is associated with the presence of *Blastocystis* (**Supplementary Table 2.1.6**). However, some clostridia tend to co-exclude with *Blastocystis*, such as *Ruminococcus gnavus* (significant in 6 datasets, **Figure 2.1.6D**) and *Clostridium bolteae* (significant in 5 datasets, **Supplementary Figure 2.1.7**). Therefore, while there is a general positive association between *Firmicutes*/*Clostridia* and *Blastocystis*, there are negative associations at the species-level, possibly due to competition for resources or different ecological niches.

In contrast, the most abundant intestinal bacterial genus, *Bacteroides*, is generally more abundant in *Blastocystis*-negative samples (**Supplementary Figure 2.1.7**), with five datasets in which this trend is significant. This association is also driving the general higher abundance of the Bacteroidetes phylum in *Blastocystis*-negative samples and possibly contrasting the opposite trend observed for the Firmicutes phylum (**Supplementary Figure 2.1.7**). Proteobacteria and Actinobacteria seem instead generally not influenced by the presence of *Blastocystis* with only one and two datasets, respectively, in which they appear significant. Specific species in these phyla can however still be strongly associated with *Blastocystis* presence (*e.g.*, the proteobacterium *Oxalobacter formigenes* significant in 5 datasets, **Supplementary Figure 2.1.7**) or absence (*e.g.*, the actinobacterium *Eggerthella* significant in 5 datasets, **Supplementary Figure 2.1.7**), suggesting that species-specific functional specialization has a higher ecological connection with *Blastocystis* than more general phylum-level characteristics.

We expanded the analysis on the association between *Blastocystis* and specific intestinal organisms by searching overall microbiome signatures predictive for the presence of the microorganism. Our previous work on such machine learning signatures (Pasolli *et al*, 2016) showed that all the diseases considered here can be associated, with a variable degree of accuracy, with their microbiome structures. In the case of *Blastocystis*, we found that microbiome signatures (**Figure 2.1.6F**) are always statistically significant and are even stronger than for the disease. Importantly, this is true not only when considering specific datasets with an unbiased cross-validation procedure (Pasolli *et al*, 2016), but also when predicting the presence of *Blastocystis* in a given dataset considering only the samples from the other independent studies. This confirms that *Blastocystis*-positive microbiomes have distinguishing features that are consistent across populations, geography, and batch effects.

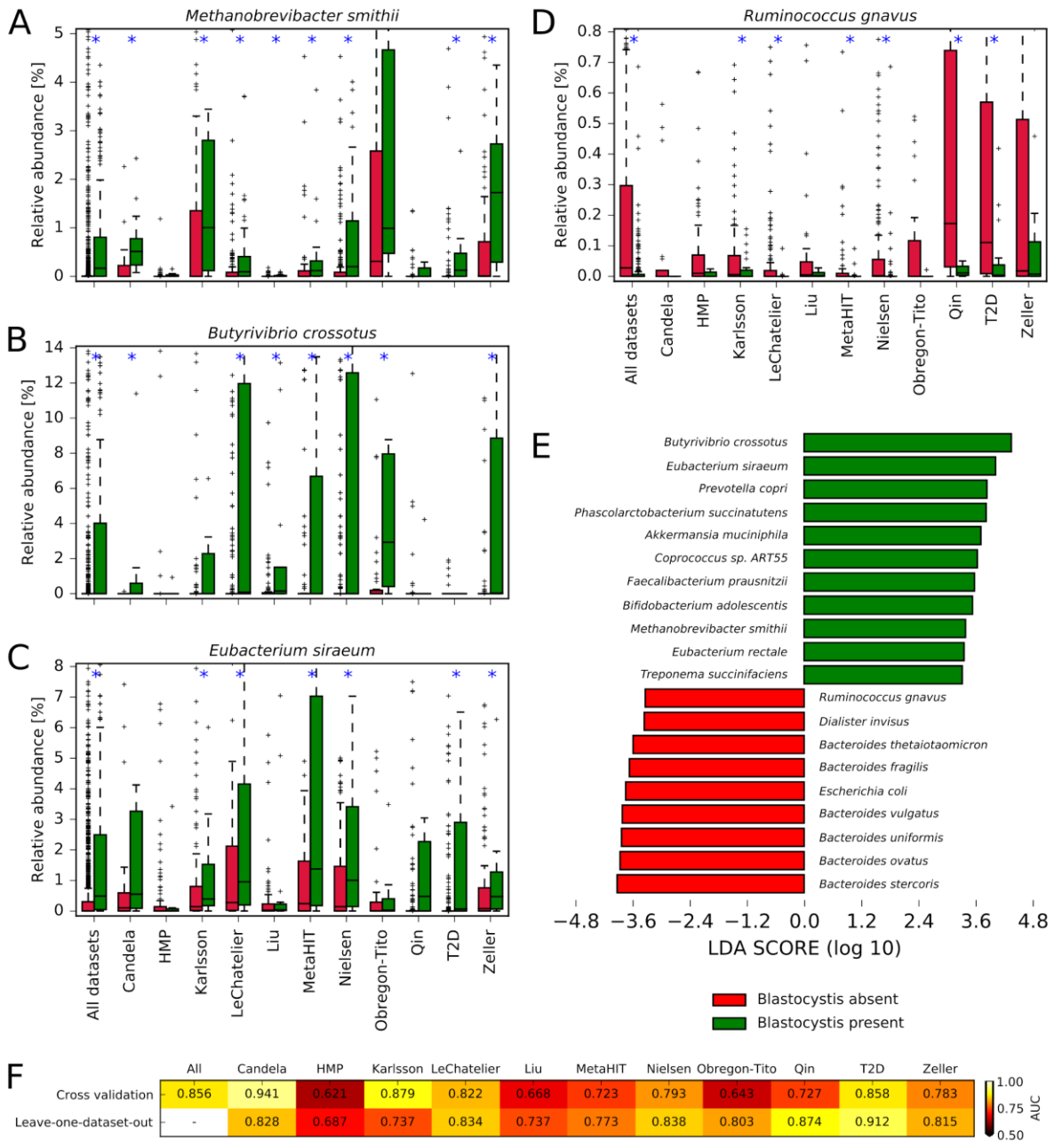


Figure 2.1.6. The presence (or absence) of *Blastocystis* is associated with major differences in the intestinal microbiome. Some species are strongly associated with the presence (**A-C**) or absence (**D**) of *Blastocystis* (plots for additional microbes are reported in **Supplementary Figure 7**). Boxplots report the distribution of abundances in samples with and without *Blastocystis*. Blue asterisks denote datasets where significant differences exist between the absence and presence of *Blastocystis*. LEfSe analysis (**E**) showed several other microorganisms statistically associated (α = 0.05) with *Blastocystis* presence at high effect size (threshold at 3.3). Machine learning-based approach reveals that the microbiome signature is predictive for the presence of *Blastocystis* (**F**). This is valid not only when considering specific datasets with an unbiased cross-validation procedure, but also when predicting the presence of the parasite in a given dataset considering only the samples from other independent studies (leave-one-dataset-out approach).

Overall, our analysis suggests that a consistent set of bacterial and archaeal organisms, and the overall composition of the microbiome, are associated with the presence (or absence) of *Blastocystis*. Interestingly, despite the many ecological associations found, microbiome diversity is instead not associated with the presence of *Blastocystis* (**Supplementary Figure**

2.1.8). Recent studies addressed the possible correlation between the presence of *Blastocystis* and other microbiome members that can of course also be influenced by other factors such as intestinal transit time. 16S rRNA amplicon sequencing revealed a higher abundance of Clostridia, Ruminococcaceae and Prevotellaceae, among *Blastocystis*-colonized individuals, while Enterobacteriaceae were enriched in *Blastocystis*-free patients (Audebert *et al*, 2016). Two other studies found that individuals with an intestinal microbiome dominated by *Bacteroides* had less *Blastocystis* than those with *Ruminococcus* and *Prevotella*-driven enterotypes (Andersen *et al*, 2015, 2016); this was interpreted in terms of a correlation between *Blastocystis* and species richness, since the *Bacteriodes*-driven enterotype has a lower species richness compared to the other enterotypes. The same authors, however, pointed out that species richness alone could not explain other observed trends, such as the correlation between *Blastocystis* carriage and BMI in Danish individuals. They thus argued that the presence of specific microbial species could influence the ability of the microorganism to thrive in the gut, but were unable to identify those species. Here we expand this concept on a much larger cohort size and higher taxonomic resolution, and provide a list of bacterial and archaeal organisms that should be prioritized in future experimental investigations (e.g. *in vitro*) aimed at understanding the ecology of *Blastocystis* in the human gut and its potential direct interaction with bacterial members of the microbiome.

Discussion

We have developed a computational pipeline to detect *Blastocystis* in human gut metagenomic samples and applied it to a collection of >2,000 metagenomes from subjects representing all continents except Australia and Antarctica. This is the largest investigation on the prevalence of *Blastocystis* and its subtypes in humans, overcoming in size and geographic diversity the single metagenomic study of an European cohort (Andersen *et al*, 2015) and the other more traditional investigations (Bart *et al*, 2013; Ramírez *et al*, 2014; Scanlan *et al*, 2014, 2016; Villalobos *et al*, 2014). Importantly, we also assessed the association between the presence of *Blastocystis* STs and a number of disease conditions, studied the co-occurrence (or co-exclusion) with other members of the gut microbiome, and reconstructed the genomes of strains belonging to different STs and used them for phylogenetic and functional potential analyses.

We detected *Blastocystis* in subjects from 11 of the 12 datasets, confirming its global distribution. In agreement with current literature, the geographic distribution of subtypes was not random: ST3 was widely distributed, ST4 was strongly underrepresented outside Europe and USA, and ST2 predominated in the non-industrialized cohorts. These findings illustrate how important epidemiologic aspects can be studied by mining appropriate metagenomics datasets.

We confirmed that the microorganism is able to persist for months (Scanlan *et al*, 2014), and that all the *Blastocystis* STs commonly associated with humans are able to stably colonize the gut. The presence of *Blastocystis* was strongly negatively correlated with BMI, but microbiome diversity was not statistically associated with its presence, suggesting that the low microorganism prevalence in obese subjects is independent from the documented decrease in overall diversity (Pareek *et al*, 2011). Importantly, *Blastocystis* was significantly more prevalent in the control groups for the investigations on colorectal cancer and ulcerative colitis, and was absent in individuals with STEC infection (although no controls are available for this study). While these associations require additional follow-up studies, they are consistent with the general trend we observed of higher *Blastocystis* prevalence in healthy individuals. If we also consider the increased detection rate of the microorganism in non-westernized populations, its stable colonization in healthy subjects, and the high global prevalence, our work provides multiple and robust evidence to consider *Blastocystis* as a common member of the healthy human gut microbiome and further expands the findings of clinical studies of chronic colonization (Roberts *et al*, 2014) and carriage among healthy individuals (Scanlan & Marchesi, 2008).

We completed the analysis by showing how the presence and abundance of *Blastocystis* were strongly correlated with those of Archaea; other bacterial species and phyla were similarly correlated (or anti-correlated) with *Blastocystis.* These analyses, which raise new hypotheses about potential ecological or direct interactions of *Blastocystis* with specific bacterial members of the gut microbiome, would have not been possible with purely cultivation-based approaches. Phylogenomic analyses are another essential tool for microbial population genomics, but are almost exclusively performed on genomes obtained by sequencing isolates (Budroni *et al*, 2011; Klemm & Dougan, 2016). *Blastocystis* can be cultivated *in vitro* (Tan, 2008), but establishing a collection of microorganism cultures from individuals of diverse geographic origin is very laborious and time-consuming. Here we show that full *Blastocystis* genomes can be reconstructed from metagenomes, and provide novel information on the diversity in the genus, the phylogenetic relation within subtypes, and functional traits.

With the collections of publicly available metagenomes quickly growing in number and size, there is an unprecedented opportunity to unravel the population genomics of *Blastocystis* at multiple levels of resolution without the need of targeted isolation work. Importantly, the computational pipeline we developed here is applicable to other parasites and fungi, if genome information is available and the target organism is present at a sufficient abundance. We thus anticipate that metagenomic analysis coupled with the opportunity of mining the vast collections of gut metagenomes will soon become an indispensable tool to explore the epidemiology, genetics, and diversity of Eukaryotic microorganisms in the human host.

Acknowledgements. This work was supported in part by the European Union FP7 Marie-Curie grant (PCIG13-618833), MIUR grant FIR RBFR13EWWI, Fondazione Caritro grant Rif.Int.2013.0239, European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (MetaPG project), and Terme di Comano grants to NS, by the European Union H2020 Marie-curie grant (707345) to EP, and by the European Commission H2020 programme under contract number 643476 (www.compare-europe.eu) to SMC.

Supplementary tables

The supplementary tables are available for download on the online version of the paper (http://dx.doi.org/10.1038/ismej.2017.139). Here reported down below the captions:

Supplementary Table 2.1.1. Statistics for the nine considered reference genomes coming from eight different STs. Numbers refer to original genomes and after-screening genomes. Screening was devoted to remove potential bacterial and archaeal contamination.

Supplementary Table 2.1.2. Prevalence for each *Blastocystis* subtype in every category considered in **Figure 1**.

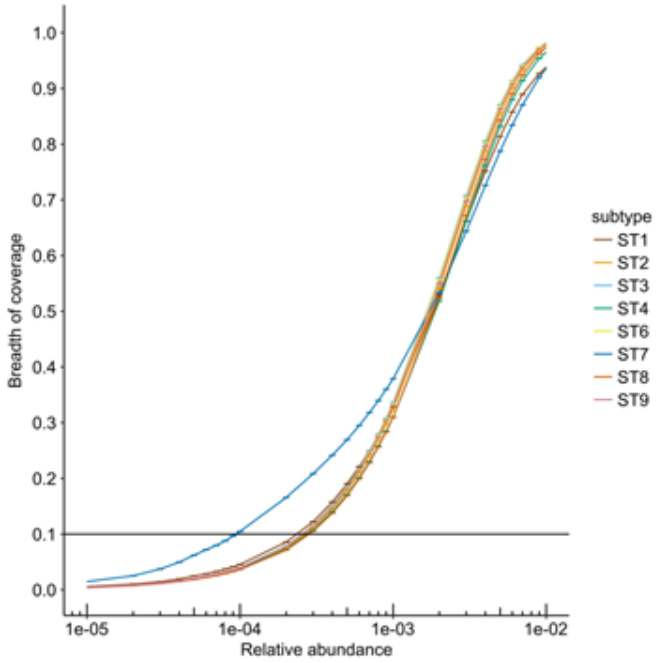
Supplementary Table 2.1.3. Breadth of coverage and relative abundance of *Blastocystis* in subjects infected over two timepoints.

Supplementary Table 2.1.4. Statistics for the 43 genomes associated with different STs reconstructed from the metagenomes.

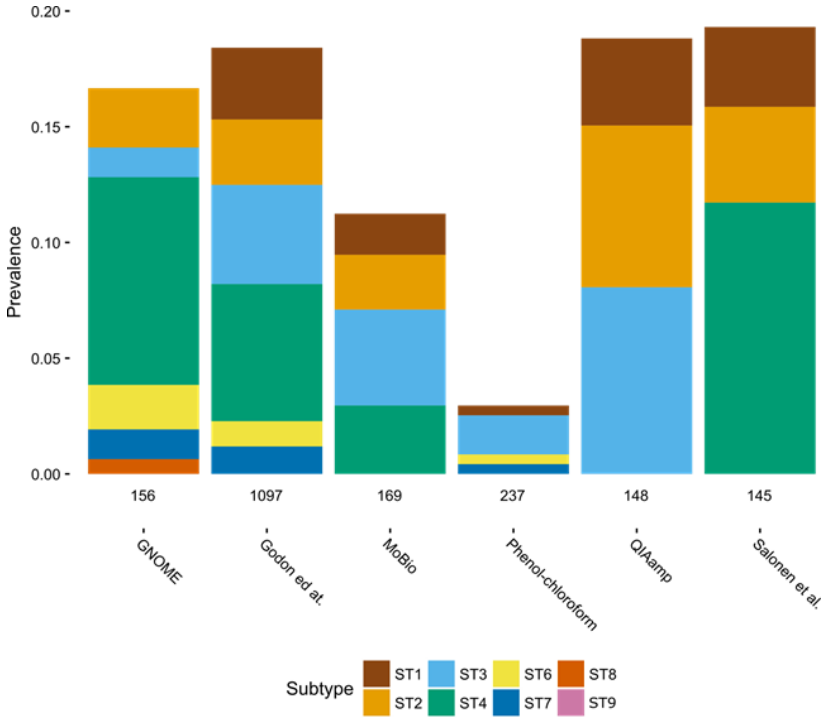
Supplementary Table 2.1.5. Description of the 795 ST-specific KOG functions, which were determined after statistical significance testing with false-discovery rate correction. The table reports the KOG functions with an adjusted p-value less than 0.2.

Supplementary Table 2.1.6. p-values associated with the statistical significance test aimed at finding association between *Blastocystis* presence and other organisms of the microbiome.

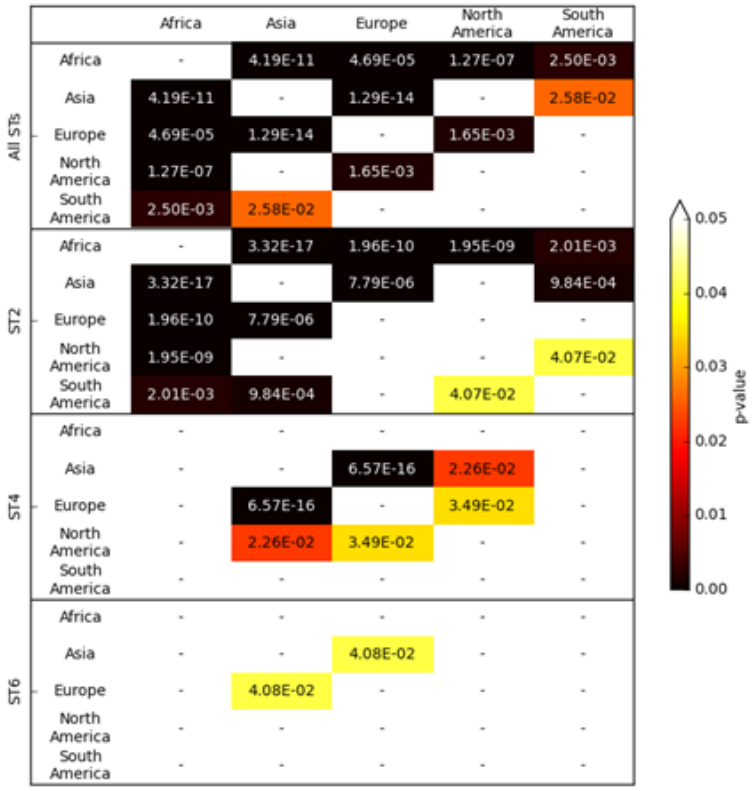
Supplementary figures



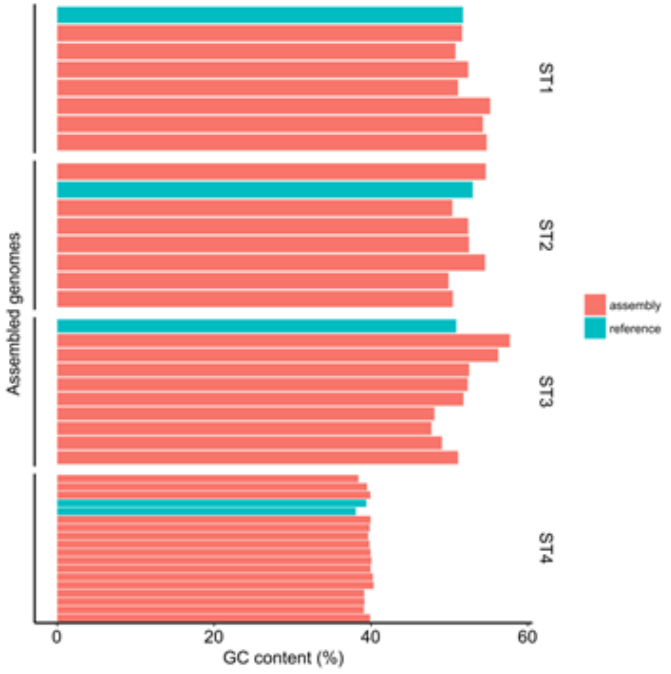
Supplementary Figure 2.1.1. Analysis on semi-synthetic data revealed that *Blastocystis* is detected through the developed methodology when present at a concentration as low as 0.03% in typical metagenomic samples of 50M reads.



Supplementary Figure 2.1.2. *Blastocystis* prevalence at varying DNA extraction procedures.

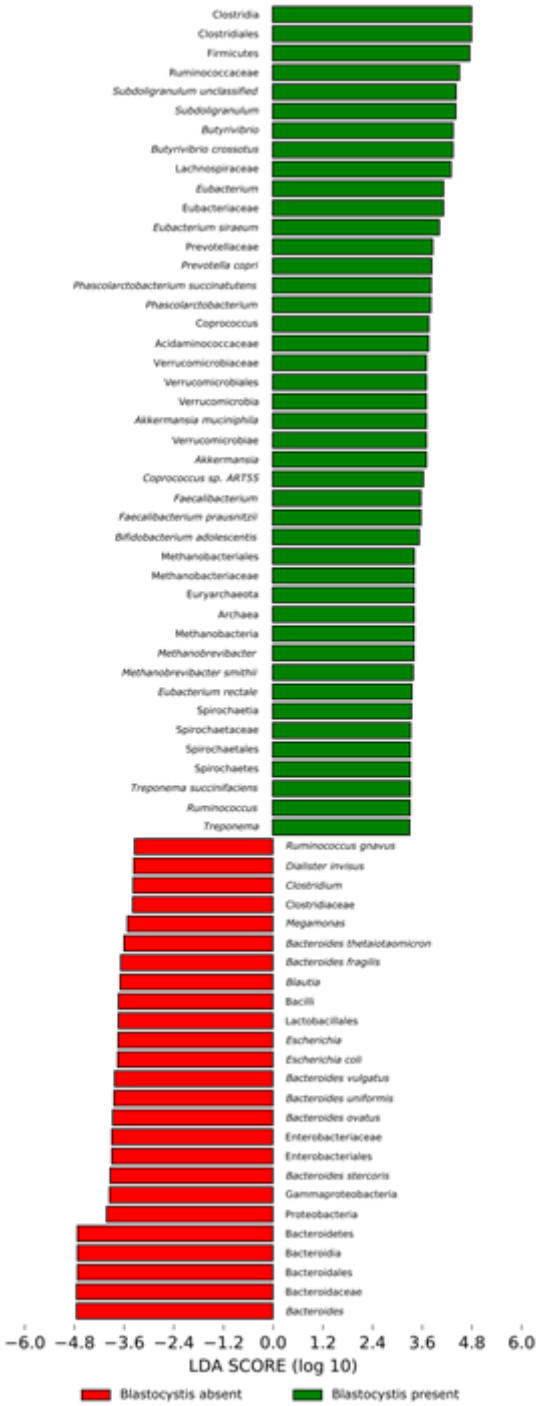


Supplementary Figure 2.1.3. p-values associated with the statistical significance test aimed at finding prevalence of specific STs in different continents. "-" denotes non-statistical significance.

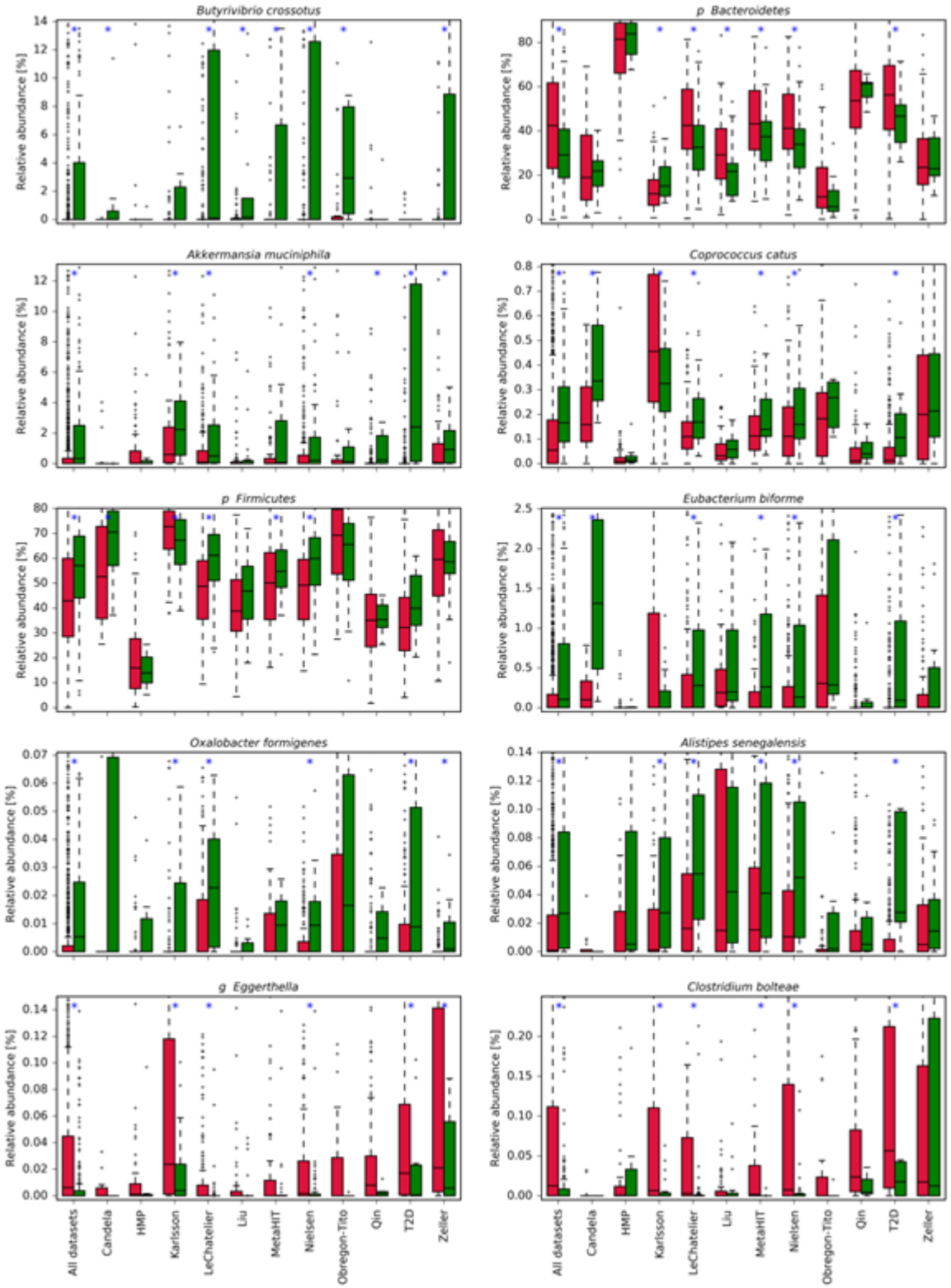


Supplementary Figure 2.1.4. GC-content of the reconstructed genomes associated with the four most prevalent ST types.

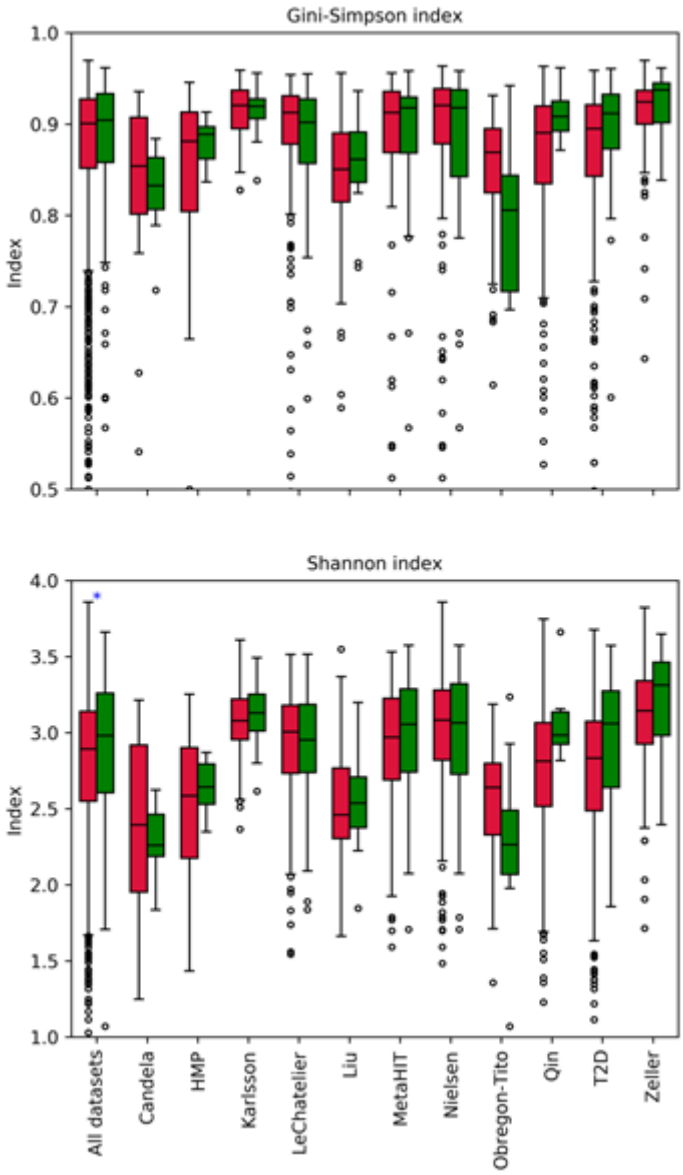
Supplementary Figure 2.1.5. Heatmap reporting the ST- specific KOG functions. Further details on these KOG functions are reported in **Supplementary Table 5**.



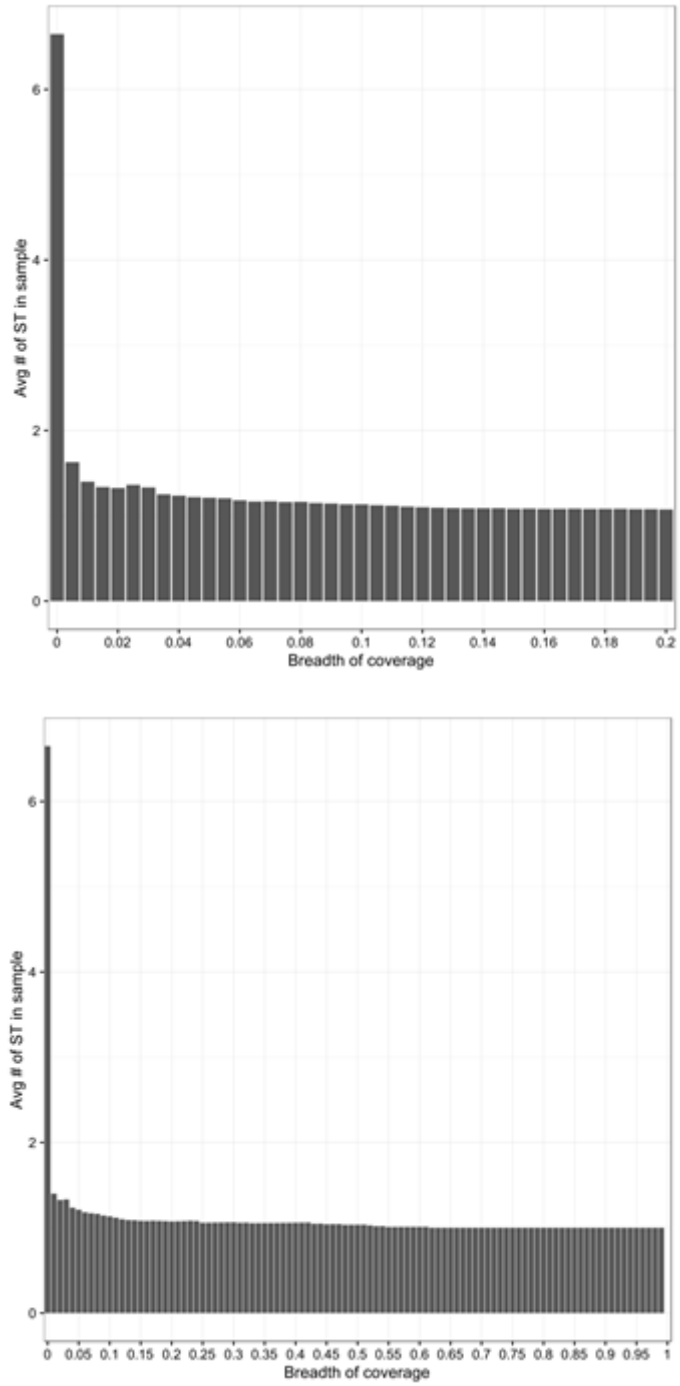
Supplementary Figure 2.1.6. The LEfSe analysis conducted on all the taxonomic levels extends the results reported in **Figure 2.1.6E**. Several microorganisms are statistically associated (α = 0.05) with *Blastocystis* presence at high effect size (threshold at 3.3). These include *Clostridiales* and Firmicutes that are associated with the presence of *Blastocystis* whereas *Bacteroides* and Proteobacteria tend to be more associated with its absence.



Supplementary Figure 2.1.7. Additional results than those reported in **Figure 2.1.6A-D** shows that the presence (or absence) of *Blastocystis* is associated with major differences in the intestinal microbiome.



Supplementary Figure 2.1.8. Gini-Simpson and Shannon indexes were considered to estimate the alpha diversity in each dataset under the condition of absence (in red) or presence (in green) of *Blastocystis*. Only in one case (reported with the blue asterisk) we observed statistical significance between the two conditions.



Supplementary Figure 2.1.9: False discovery rate plot. The two plots show the average subtypes detected at a given value of breadth of coverage. The distribution goes from seven (all the STs in addition to the dominant one) to one (only the dominant ST detected), but it is already plateauing at 10% breadth of coverage confirming that such value does not produce false positives.

2.2. Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study

Introduction to the chapter

The impact of cigarette smoke and tobacco products on human health has been widely established and linked to the rise of numerous adverse health issues and several systemic diseases (Law *et al*, 1997; Lee *et al*, 1997; Ho *et al*, 1993), including several types of cancers (Ang *et al*, 2010; Lortet-Tieulent *et al*, 2016). Tobacco assumption via the oral route makes the upper respiratory airways the first parts that make contact with the cigarette smoke. Recent studies have assessed the impact of cigarette smoke on the oral microbial communities (Wu *et al*, 2016a; Colman *et al*, 1976; Charlson *et al*, 2010) and a special focus has been made on the association with oral diseases, mostly in particular periodontal diseases (Haber, 1994; Mager *et al*, 2003; Socransky, 1977). Recent alternative ways of smoking like e-cigarettes or hookah or more “classical” like cigar smoking, gained popularity most in particular between young people (Department of Health and Human Services, 2016) and to date, little is known about the effect of these ways of smoking on the oral microbiome (Ganesan *et al*, 2020). It has also been observed that several socio-demographic aspects, like the education level, can influence smoking status (Chassin *et al*, 1996; Emmons *et al*, 1998; Pan *et al*, 2015).

In this chapter, I will present the result of an investigation we conducted on the NYC-HANES dataset, a population-based sample of New York City including extensive information on smoking habits, socio-demographics, oral health, and other health conditions and biomarkers (including serum cotinine levels). We tested the effect of several sociodemographic variables, such as education level or age, as confounding factors and tried to identify possible variables that could explain the predisposition of certain groups to smoke. Thanks to the availability of blood serum cotinine data, we assessed the effect of environmental exposure like secondhand smoke on the oral microbiome, an important aspect since it has been established its toxicity.

I performed all the analyses described in the paper and, with the help of the co-authors, I performed the data interpretation of the statistical tests and the manuscript writing. All the co-authors were involved in the creation of an R package to make the results reproducible, with all the code and functions used for analyzing the data. The package and the associated vignette are available at https://github.com/waldronlab/nychanesmicrobiome.

This chapter contains a work published in the following article:

Francesco Beghini, Audrey Renson, Christine P. Zolnik, Ludwig Geistlinger, Mykhaylo Usyk, Thomas U. Moody, Lorna Thorpe, Jennifer B. Dowd, Robert Burk, Nicola Segata, Heidi E. Jones, Levi Waldron

Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study

Annals of Epidemiology (2019) - https://doi.org/10.1016/j.annepidem.2019.03.005

Abstract

Purpose: The effect of tobacco exposure on the oral microbiome has not been established. **Methods**: We performed amplicon sequencing of the 16S ribosomal RNA gene V4 variable region to estimate bacterial community characteristics in 259 oral rinse samples, selected based on self-reported smoking and serum cotinine levels, from the 2013-14 New York City Health and Nutrition Examination Study. We identified differentially abundant operational taxonomic units (OTUs) by primary and secondhand tobacco exposure, and employed "microbe set enrichment analysis" to assess shifts in microbial oxygen utilization. **Results**: Cigarette smoking was associated with depletion of aerobic OTUs (Enrichment Score test statistic ES = -0.75, p = 0.002) with a minority (29%) of aerobic OTUs enriched in current smokers compared to never smokers. Consistent shifts in the microbiota were observed for current cigarette smokers as for non-smokers with secondhand exposure as measured by serum cotinine levels. Differential abundance findings were similar in crude and adjusted analyses. **Conclusion**: Results support a plausible link between tobacco exposure and shifts in the oral microbiome at the population level through three lines of evidence: 1) a shift in microbiota oxygen utilization associated with primary tobacco smoke exposure, 2) consistency of abundance fold-changes associated with current smoking and shifts along the gradient of secondhand smoke exposure among non-smokers, and 3) consistency after adjusting for a priori hypothesized confounders.

Keywords: microbiota; RNA, Ribosomal, 16S; human microbiome; oral health; tobacco; smoking

List of abbreviations

CI - Confidence Interval
CUNY - City University of New York
DNA - Deoxyribonucleic Acid
ES - Enrichment Score
FDR - False Discovery Rate
GSEA - Gene Set Enrichment Analysis
GSVA - Gene Set Variation Analysis
HMP - Human MIcrobiome Project
IRB - Institutional Review Board
NHANES - National Health Nutrition and Examination Survey
NYC DOHMH - New York City Department of Health and Mental Hygiene
NYC HANES - New York City Health Nutrition and Examination Survey
OR - Odds Ratio
ORA - Over-representation Analysis
OTU - Operational Taxonomic Unit
PCOA - Principal Coordinates Analysis
PERMANOVA - Permutational Multivariate Analysis of Variance
RNA - Ribonucleic Acid
rRNA - Ribosomal Ribonucleic Acid

Introduction

Dysbiosis of the dental plaque microbiome is a necessary step in the etiology of periodontitis and caries (Curtis *et al*, 2011), which have been linked to systemic illness, including cardiovascular diseases (Blaizot *et al*, 2009), type 2 diabetes mellitus (Borgnakke *et al*, 2013), obesity (Chaffee & Weston, 2010), low birth weight and preterm birth (Ide & Papapanou, 2013), rheumatoid arthritis (Detert *et al*, 2010), chronic obstructive pulmonary disease (Zeng *et al*, 2012), and oral and digestive cancers (Fitzpatrick & Katz, 2010). Tobacco exposure is a cause of these outcomes (Chang *et al*, 2014; Chang, 2012; Stallones, 2015; Laniado-Laborín, 2009; Zheng *et al*, 2016; Moura *et al*, 2014), but whether it causes them through shifts in the general oral microbiome is unknown (Wu *et al*, 2016a). If tobacco smoke causes harmful alterations of the oral microbiome, interventions targeting the oral microbiome could mitigate the impact of tobacco exposures. A key aspect of making this distinction lies in establishing whether a range of tobacco exposures, including cigarette smoking, secondhand smoke exposure, hookah and e-cigarette use, cause substantial changes in the structure and function of the general oral microbiome.

Evidence suggests that tobacco smoke exposure causes alterations to the oral microbiome, selecting a community enriched with opportunistic pathogens (Kumar *et al*, 2011; Mason *et al*, 2015) and negatively impacting the resilience and colonization resistance of the sub- and supragingival biofilms (Joshi *et al*, 2014). Such alterations may occur directly due to selective toxicity (Macgregor, 1989), or indirectly via alteration of the host immune system to produce both pro- and anti-inflammatory effects (Giannopoulou *et al*, 2003; Güntsch *et al*, 2006; Stämpfli & Anderson, 2009) which alter the oral biofilm and mucosal microbial habitats. Another potential mechanism by which tobacco smoke reconfigures the oral microbiome is via depletion of oxygen (Kenney *et al*, 1975), creating a hypoxic oral environment that favors anaerobiosis. Tobacco smoke may also favor anaerobiosis by increasing the amount of free iron (Ghio *et al*, 2008), and inhibiting oral peroxidase (Reznick *et al*, 2003). Anaerobic glycolysis in human salivary cells has been shown to dramatically increase after exposure to tobacco smoke (Eichel & Shahrik, 1969), and human experiments show reduction in periodontal pocket oxygen tension (Hanioka *et al*, 2000) and redox potential (Kenney *et al*, 1975) after smoking cigarettes. Low throughput studies of the oral microbiome have shown greater abundance of the anaerobes Prevotella intermedia (Eggert *et al*, 2001) and Lactobacillus spp. (Parvinen, 1984) in cigarette smokers. Ad hoc findings from high throughput studies have suggested that smokers have greater abundance of anaerobic microorganisms (Mason *et al*, 2015) and depletion of microbial functional pathways related to aerobic respiration (Wu *et al*, 2016a). Thus tobacco exposure could plausibly cause changes to the oral microbiome, but available results are limited to laboratory and small-sample studies.

This study suggests a causal link between tobacco exposure and alterations to the saliva microbiome among participants of the 2013-14 New York City Health and Nutrition Examination Study (NYC HANES). It contrasts current smokers to non-smokers with no recent secondhand exposure, and investigates a dose-response relationship among non-smokers with varying degrees of secondhand exposure assessed by quantitative serum cotinine level. It further investigates former smokers and smokers of e-cigarettes and hookah. Multiple lines of causal inference are used to test the hypothesis that tobacco smoke alters the saliva microbiome: controlling for hypothesized confounders, testing for a dose-response relationship, and testing for altered oxygen requirements of the microbial communities associated with tobacco exposure.

Materials and Methods

2013-2014 NYC-HANES

Data for the current study are sub-sampled from the 2013-14 NYC HANES, a population-based study of 1,575 non-institutionalized adults in New York City (Thorpe *et al*, 2015) modeled after the United States Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey (NHANES) (National Health and Nutrition Examination Survey, 2017). The NYC HANES sample was recruited using a three stage cluster household probabilistic design of all non-institutionalized adults 18 years of age or older. Consenting individuals provided information on smoking, including use of alternative tobacco products such as e-cigarettes and hookahs in the last 5 days, socio-demographic characteristics, and oral hygiene practices through face-to-face interviewing and audio computer assisted interviewing for sensitive questions. Participants also underwent physical examination and provided blood and oral rinse specimens for biomarker analysis. Serum specimens were analyzed for cotinine by liquid chromatography/tandem mass spectrometry (Perlman *et al*, 2016).

The study was conducted by the City University of New York (CUNY) School of Public Health in collaboration with the New York City Department of Health and Mental Hygiene (NYC DOHMH), with ethical approval from their respective institutional review boards (IRBs). The current sub-study received separate IRB approval from the CUNY School of Public Health.

Tobacco exposure outcome measures and selection of sub-sample

We selected a sub-sample of 297 participants for oral microbiome assessment based on self-reported tobacco use and serum blood cotinine level, classified into five mutually exclusive groups:

- Current smokers (n=90) were selected from participants who reported smoking more than 100 cigarettes in their lifetime, smoking a cigarette in the last 5 days, and not using any alternative tobacco product in the last 5 days (the 90 with highest measured serum cotinine were selected).
- Never smokers (n=45) were randomly selected from those reporting lifetime smoking of less than 100 cigarettes, no usage of any tobacco product in the last 5 days, and serum cotinine level less than 0.05 ng/mL.
- Former smokers (n=45) were randomly selected from participants reporting lifetime smoking of more than 100 cigarettes, but currently not smoking, no use of any tobacco product in the last 5 days, and serum cotinine level less than 0.05 ng/mL.
- Non-smokers with secondhand exposure (n=38) comprised all available former or never smokers with serum cotinine between 1 and 14 ng/mL (Jarvis *et al*, 1987). This group includes four individuals with serum cotinine > 10 ng/mL (Homa *et al*, 2015), who may have been light smokers; sensitivity analyses excluding these four had negligible effect (not shown, included in Reproducible Analysis code), so they are included in the secondhand exposure group.
- "Alternative" smokers (n=79) were participants with self-reported usage of hookah, cigar, cigarillo and/or e-cigarette in the last 5 days.

For quality control, 5% of samples were randomly selected and sequenced as technical replicates. Results from replicates were used instead of the original sample if the sequencing read count was greater than the original. An additional eight samples failed PCR amplification and were repeated. Fifteen specimens (n=4 current cigarette smokers, n=2 never smokers, n=2 former smokers, n=7 alternative smokers) were discarded for sequencing quality control (below). We excluded an additional 23 participants from the alternative smoker group who also reported smoking cigarettes in the last 5 days, in an attempt to isolate the effect of alternative tobacco exposures, resulting in an analytic sample of 259 participants.

Specimen collection, processing, and sequence analysis

Specimen collection, processing, and sequence analysis methods are described in detail in a companion paper (Renson *et al*, 2017). In brief, participants were asked to fast for 9 hours prior to oral rinse and blood specimen collection. A 20-second oral rinse was divided into two 5-second swish and 5-second gargle sessions using 15 mLs of Scope® mouthwash, which was transported on dry ice and stored at -80˚C. A modified protocol of the QIAamp DNA mini kit (QIAGEN) was used for DNA extraction (Renson *et al*, 2017). DNA was amplified for the V4 variable region of the 16S rRNA gene (Apprill *et al*, 2015; Parada *et al*, 2016). High-throughput amplicon sequencing was conducted on a MiSeq (Illumina, San Diego, CA) using 2x300 paired-end fragments. 16S read analysis was carried out using QIIME version 1.9.1

(Caporaso *et al*, 2010) and Phyloseq (McMurdie & Holmes, 2013). Paired-end reads were merged with fastq-join (Aronesty, 2013) and resulting low quality reads (PHRED score < 30) were discarded when joining the split reads (qiime split_libraries_fastq.py). Operational Taxonomic Unit (OTU) picking was performed with an open reference approach by clustering using UCLUST at 97% sequence similarity and taxonomy was assigned using the SILVA 123 (Pruesse *et al*, 2007) database as reference. The QIIME generated OTU table was converted for phyloseq processing. Samples with fewer than 1000 reads (n=15) were removed from the OTU table in the phyloseq preprocessing step. Genera present with a mean relative abundance of less than 2×10^{-4} were collapsed as "Other."

Unsupervised clustering

We explored differences in beta diversity measures via Principal Coordinates Analysis (PCoA) plots on Weighted UniFrac distances. The grouping of distances by smoking status was tested by PERMANOVA as implemented in the 'vegan' package (Oksanen *et al*, 2008), with 999 permutations.

Oral microbiome measures

We compared oral bacterial community characteristics by tobacco exposure group using four types of oral microbiome measures: 1. alpha (within-sample) diversity of the OTUs present; 2. beta (between-sample) diversity of OTUs; 3. OTU counts at the genus level; 4. enrichment of differential abundance categorized by oxygen requirement. For diversity measures, we estimated Chao1 Index, Shannon Index and observed OTUs, and weighted UniFrac (Lozupone & Knight, 2005) beta diversity using the estimate_richness and distance methods of the phyloseq Bioconductor package (McMurdie & Holmes, 2013).

Differential abundance analysis

We performed crude and adjusted negative binomial log-linear regression of tobacco exposure group to identify differentially abundant OTUs using edgeR (Robinson *et al*, 2010) Bioconductor package. Low-prevalence OTUs, those without 3 or more reads observed in at least 8 samples, were discarded. For adjusted models, a priori hypothesized confounders included age, sex, race, self-reported physical activity, education, diabetes status (based on serum HbA1c), and self-reported gum disease. Education (Giskes *et al*, 2005; Pierce *et al*, 1989), age, and sex (Emmons *et al*, 1998; Chassin *et al*, 1996; Pan *et al*, 2015) are known to be associated with smoking and could plausibly be associated with oral microbiome characteristics. Race was also treated as a possible confounder as studies suggest differences in nicotine metabolism by race/ethnicity (Benowitz *et al*, 2009b). A False Discovery

Rate (FDR, Benjamini & Hochberg, 1995) less than 0.05 was considered statistically significant. Results from edgeR were compared to results obtained from the application of DESeq2 (Love *et al*, 2014). Crude and adjusted coefficients were compared to assess which hypothesized confounders had the greatest effect on adjusted analyses.

Microbe set enrichment analysis for oxygen requirements

We categorized genera as aerobic, anaerobic, or facultative anaerobic (Bergey's Manual of Systematics of Archaea and Bacteria, 2015) integrating information from the IMG/MER database (Markowitz *et al*, 2012) and from the Bergey's Manual of Systematics of Archaea and Bacteria (Bergey's Manual of Systematics of Archaea and Bacteria, 2015). This resulted in three "microbe sets" of OTUs with common oxygen requirements. We applied two concurrent approaches to analyze whether the three microbe sets show coherent changes in abundance of the contained microbes for (i) smokers vs. never smokers (with no recent secondhand smoke exposure), and (ii) among non-smokers with exposure to secondhand smoke. First, over-representation of differentially abundant OTUs in each microbe set was tested based on the hypergeometric distribution (corresponds to a one-sided Fisher's exact test). Second, Gene Set Enrichment Analysis (GSEA, Subramanian *et al*, 2005) was used to test whether microbes of a particular microbe set accumulate at the top or bottom of the full OTU vector ordered by direction and magnitude of abundance change between the tested sample groups. Over-representation analysis (ORA) and GSEA were applied as implemented in the EnrichmentBrowser R/Bioconductor package (Geistlinger *et al*, 2016). Application of GSEA incorporated the voom-transformation (Ritchie *et al*, 2015) of OTU counts to concur with GSEA's assumption of roughly normally distributed data. As the implementations of GSEA and ORA required a binary outcome, serum cotinine levels were binned to contrast the upper tertile (> 4.42 ng/ml) against the lower tertile (< 1.76 ng/ml). We also analyzed serum cotinine level as a continuous measure using Gene Set Variation Analysis (GSVA, Hänzelmann *et al*, 2013).

Statement of reproducible research

Analyses were performed in QIIME version 1.9.1 and R version 3.5.1. All results presented in this manuscript are reproducible by installing the package and compiling its associated vignettes provided at https://github.com/waldronlab/nychanesmicrobiome.

Results

A total of 1.4 M reads (mean±sd: 4,758±3,463 reads/sample) were generated from 297 saliva mouthwash specimens (Thorpe *et al*, 2015) of NYC-HANES participants selected based on questionnaire and serum cotinine levels (Table 2.2.1, with serum cotinine levels by exposure group shown in Supplementary Figure 2.2.1). After quality control and filtering, we retained 91.7% of reads (5,007 mean, 3,491 s.d), which were then classified using the QIIME pipeline (Caporaso *et al*, 2010) into 1291 OTUs with more than 10 reads.

Taxonomic composition of the final analytic sample (n=259) was predominated by Streptococcus (36% average relative abundance) and Prevotella (17% average relative abundance), which were present in every sample. Other genera commonly associated with the oral cavity like Rothia, Neisseria, Veillonella and Gemella were also found with average relative abundances less than 10%.

Table 2.2.1. Demographics and characteristics of participants in the 2013-2014 NYC-HANES smoking and oral microbiome study

	Never smoker	Cigarette	Former smoker	Alternative smoker	Secondhand
n	43	86	43	49	38
Sex = Female (%)	28 (65.1)	45 (52.3)	25 (58.1)	22 (44.9)	22 (57.9)
Race/ethnicity (%)					
Non-Hispanic White	13 (30.2)	24 (27.9)	25 (58.1)	19 (38.8)	10 (26.3)
Non-Hispanic Black	13 (30.2)	33 (38.4)	4 (9.3)	9 (18.4)	11 (28.9)
Hispanic	10 (23.3)	19 (22.1)	10 (23.3)	12 (24.5)	14 (36.8)
Asian	3 (7.0)	9 (10.5)	3 (7.0)	3 (6.1)	2 (5.3)
Other	4 (9.3)	1 (1.2)	1 (2.3)	6 (12.2)	1 (2.6)
Educational achievement (%)					
College graduate or more	16 (37.2)	18 (20.9)	21 (48.8)	17 (34.7)	9 (23.7)
Less than High school diploma	8 (18.6)	24 (27.9)	4 (9.3)	10 (20.4)	14 (36.8)
High school graduate/GED	7 (16.3)	24 (27.9)	8 (18.6)	11 (22.4)	10 (26.3)
Some College or associate's degree	12 (27.9)	20 (23.3)	10 (23.3)	11 (22.4)	5 (13.2)
Age in years(mean (sd))	45.42 (16.50)	45.85 (13.07)	55.47 (18.00)	35.59 (16.44)	37.76 (14.70)
Age group (%)					
20-29	7 (16.3)	10 (11.6)	3 (7.0)	26 (53.1)	14 (36.8)
30-39	11 (25.6)	17 (19.8)	7 (16.3)	8 (16.3)	11 (28.9)
40-49	10 (23.3)	25 (29.1)	7 (16.3)	4 (8.2)	3 (7.9)
50-59	6 (14.0)	19 (22.1)	8 (18.6)	7 (14.3)	6 (15.8)
60 and over	9 (20.9)	15 (17.4)	18 (41.9)	4 (8.2)	4 (10.5)
Diabetes (%)					
Yes	5 (11.6)	5 (5.8)	7 (16.3)	3 (6.1)	2 (5.3)
No	38 (88.4)	81 (94.2)	36 (83.7)	38 (77.6)	36 (94.7)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	8 (16.3)	0 (0.0)
Physical activity (%)					
Very active	15 (34.9)	27 (31.4)	11 (25.6)	16 (32.7)	17 (44.7)
Somewhat active	20 (46.5)	37 (43.0)	20 (46.5)	25 (51.0)	15 (39.5)
Not very active/not active at all	8 (18.6)	22 (25.6)	12 (27.9)	8 (16.3)	6 (15.8)
Annual family income (%)					
Less Than $20,000	5 (11.6)	31 (36.0)	8 (18.6)	20 (40.8)	14 (36.8)
$20,000-$49,999	15 (34.9)	20 (23.3)	9 (20.9)	14 (28.6)	9 (23.7)
$50,000-$74,999	6 (14.0)	11 (12.8)	3 (7.0)	6 (12.2)	4 (10.5)
$75,000-$99,999	8 (18.6)	4 (4.7)	6 (14.0)	4 (8.2)	2 (5.3)
$100,000 or More	6 (14.0)	11 (12.8)	13 (30.2)	2 (4.1)	6 (15.8)
Missing	3 (7.0)	9 (10.5)	4 (9.3)	3 (6.1)	3 (7.9)
Serum Cotinine (median [IQR])	0.04 [0.04, 0.04]	271.49 [189.99, 360.99]	0.04 [0.04, 0.04]	10.54 [0.28, 55.36]	3.01 [1.39, 5.48]
Gum disease (self-reported) (%)					
Yes	4 (9.3)	9 (10.5)	5 (11.6)	4 (8.2)	4 (10.5)
No	39 (90.7)	76 (88.4)	38 (88.4)	45 (91.8)	34 (89.5)
Missing	0 (0.0)	1 (1.2)	0 (0.0)	0 (0.0)	0 (0.0)

Alpha and beta diversity of the oral microbiome by tobacco exposure

Alpha diversities were not significantly different between the five tobacco exposure groups (Shannon Index p=0.95, Observed OTUs p=0.08, Chao1 Index p=0.26 ANOVA test, Supplementary Figure 2.2.2). However, beta diversity differed between current cigarette smokers and never smokers, and was larger than differences by race/ethnicity, age, and other sociodemographic measures (Table 2.2.2). The overall microbiome composition and structure in these two classes differed and beta diversity was significantly explained by smoking status (R2=0.051, p<0.001, PERMANOVA test, Figure 2.2.1). Additionally, former smokers were significantly different from current smokers (p=0.001, R2=0.044, PERMANOVA test), but not from never smokers (p=0.16, R2=0.018, PERMANOVA test). Within former smokers, we found no evidence of differences between those who quit recently versus longer ago (Supplementary Figure 2.2.3).

Proteobacteria less abundant in the microbiome of smokers

In crude analyses, 46 differentially abundant OTUs, taxonomically assigned to 28 different genera, were identified between current cigarette smokers and never smokers (Supplementary Figure 2.2.4). Relative abundance of OTUs annotated as phylum Proteobacteria (*Neisseria*, *Lautropia*, *Haemophilus* and *Actinobacillus*) and Candidate division SR1 were found to be lower in current cigarette smokers compared to never smokers (Proteobacteria phylum t-test p-value=5e-07, logFC=-0.84, Supplementary Figure 2.2.5).

Adjusted differential abundance analysis, accounting for hypothesized confounders, identified fewer (n=21) differentially abundant OTUs between current and never smokers (Figure 2.2.2). The phylum Proteobacteria was still identified as less-abundant in current smokers in the adjusted model (t-test p-value=8e-07, logFC=-0.85). Adjusted coefficients were slightly attenuated toward the null compared to crude estimates (Figure 2.2.3). Addition of one hypothesized confounder at a time showed that age and education had the strongest impact, resulting in a median decrease in coefficient magnitude of 2 and 3 percent, respectively.

Table 2.2.2: PERMANOVA analysis on Weighted UniFrac distance measure. Model included smoking status (cigarette smokers/never smokers) and other sociodemographic measures. Df: degrees of freedom, R2: Coefficient of Determination

	Df	F.Model	R^2	Pr(>F)
Smoking status (Cigarette smokers vs Never Smokers)	1	7.0845	0.05137	0.002
Self reported gum disease	2	0.8649	0.01254	0.496
Race/ethnicity	4	1.6281	0.04723	0.05
Sex	1	2.2591	0.01638	0.06
Age groups	1	2.9418	0.02133	0.014
Physical activity	2	0.8085	0.01173	0.623
Education level	2	0.7274	0.01055	0.708
Diabetes	1	0.304	0.0022	0.937

Differences in oxygen utilization in the oral microbiome of smokers compared to never smokers

We functionally annotated the entire set of picked OTUs according to their oxygen requirement: 78 aerobic OTUs, 673 anaerobic OTUs and 395 facultative anaerobic OTUs. We failed to annotate 145 OTUs because their genera was annotated as uncultured bacteria or taxonomic resolution was higher than genus.

A minority of aerobic OTUs (29%) and a majority of anaerobic OTUs (60%) had higher mean abundance in current smokers as compared to never smokers. Facultative anaerobic OTUs were approximately evenly divided, with 51% having higher abundance in current smokers. We accordingly found differentially abundant OTUs between current smokers and never smokers to be over-represented in aerobic OTUs (Hypergeometric test, $p = 0.004$). Using Gene Set Enrichment Analysis (GSEA) to account for collinearity between OTUs and the direction of the abundance change (up / down), aerobic OTUs were significantly depleted among current smokers relative to never smokers (Enrichment Score test statistic $ES = -0.75$, $p = 0.002$, GSEA permutation test). Anaerobic OTUs were enriched in smokers relative to never smokers but the difference was not statistically significant ($ES = 0.36$, $p = 0.14$, GSEA permutation test). We also observed an enrichment of facultative anaerobic OTUs among never smokers compared to current smokers but this result was not statistically significant ($ES = -0.29$, $p = 0.48$, GSEA permutation test).

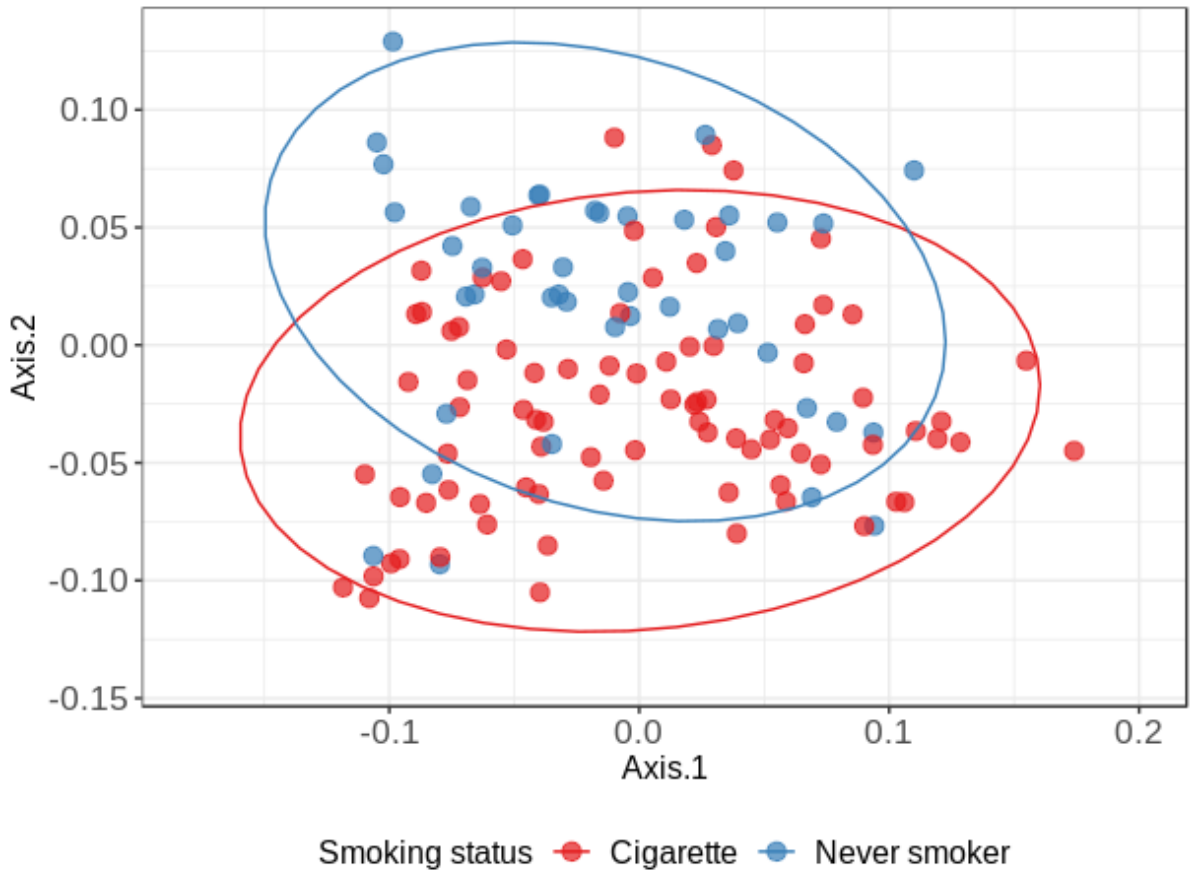


Figure 2.2.1: Principal coordinates analysis based on the weighted UniFrac distance. Dots in the ordination plot are samples from never smokers with negligible serum cotinine (blue, n=43) and current cigarette smokers (red, n=86); ellipsis indicating where 95% of observations are expected for each group. A separation between cigarette smokers and never smokers is present and is statistically significant (R^2 = 0.051, PERMANOVA p<0.001). A gradient also exists for the entire sample (n=259) by measured continuous serum cotinine level (R^2 =0.0485, PERMANOVA p = 0.001).

Comparison of those with secondhand smoke exposure to never smokers

To provide independent evidence for causal inference, we compared the coefficients estimated for the contrast of current smokers versus never smokers to the coefficients for serum cotinine level, estimated from a non-overlapping group of self-reported non-smokers (n=38) exposed to secondhand smoke. Consistency was assessed by calculating the Pearson Correlation of the two vectors of coefficients. This correlation is comparable to the Integrative Correlation Coefficient, which was originally proposed to assess the replicability of measurements from independent gene expression studies (Parmigiani *et al*, 2004; Cope *et al*, 2014). This correlation was estimated from the intersection of 121 OTUs passing the edgeR filter against low-variance features for both groups. Full differential abundance results for these 121 OTUs are reported in Supplemental File 1. We observed a positive correlation (Figure 2.2.4, Pearson's Correlation = 0.40, p = 5e-6); this correlation was slightly attenuated after adjusted for age and educational attainment (Supplementary Figure 2.2.6, Pearson's Correlation=0.28, p=0.002). This positive correlation identifies a similarity in the patterns of

differential abundance in smokers vs. never smokers when compared to the shifts associated with increasing exposure to secondhand smoke. However, the application of three concurrent approaches for microbe set enrichment analysis (GSEA, GSVA, and ORA) on samples from participants exposed to secondhand smoke, with continuous or dichotomous serum cotinine levels as the response variable, did not identify significant enrichment or depletion of aerobiosis or anaerobiosis, reflecting smaller shifts associated with secondhand smoke exposure.

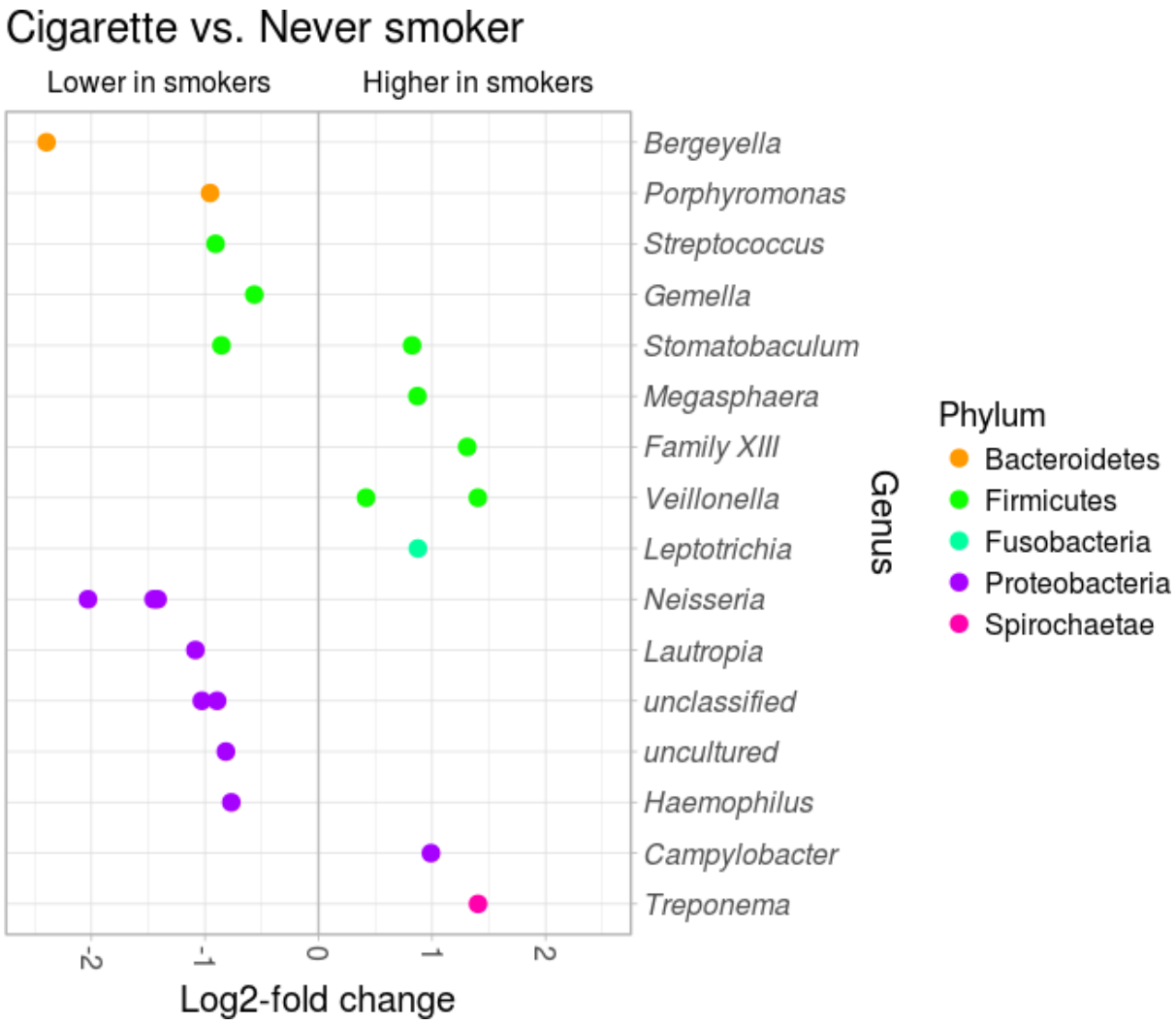


Figure 2.2.2: Adjusted multivariate differential analysis between current cigarette smokers (n=86) and never smokers (n=43). Starting from the 46 OTUs identified as differentially abundant from the crude model, adjusting for confounders OTU differentially abundant were reduced to 21.

Comparison of alternative tobacco exposures to never smokers

Differential abundance of OTUs from participants who used e-cigarettes, hookah, and/or cigar/cigarillo but not cigarettes (Table 2.2.1) were contrasted to the never smoker group. Phyla Actinobacteria, Firmicutes and Proteobacteria were more abundant in alternative smokers while Bacteroidetes and an uncultured bacterium from Saccharibacteria were more depleted. In those who only smoked hookah (n = 28), genera *Porphyromonas*, *Leptotrichia*, *Streptobacillus*, *Fusobacterium*, and an uncultured bacterium from Saccharibacteria were depleted. No OTUs were identified as differentially abundant among users of e-cigarette (n=11) or cigar/cigarillo (n=23) who did not use any other smoking products. GSEA identified a significant depletion of aerobic OTUs in cigar and cigarillo smokers (ES = -0.697, p = 0.04 GSEA permutation test) and depletion of facultative anaerobic OTUs in e-cigarette and hookah smokers compared to never smokers (e-cigarettes ES = -0.514 p = 0.03 GSEA permutation test; hookah ES = -0.489 p = 0.04 GSEA permutation test).

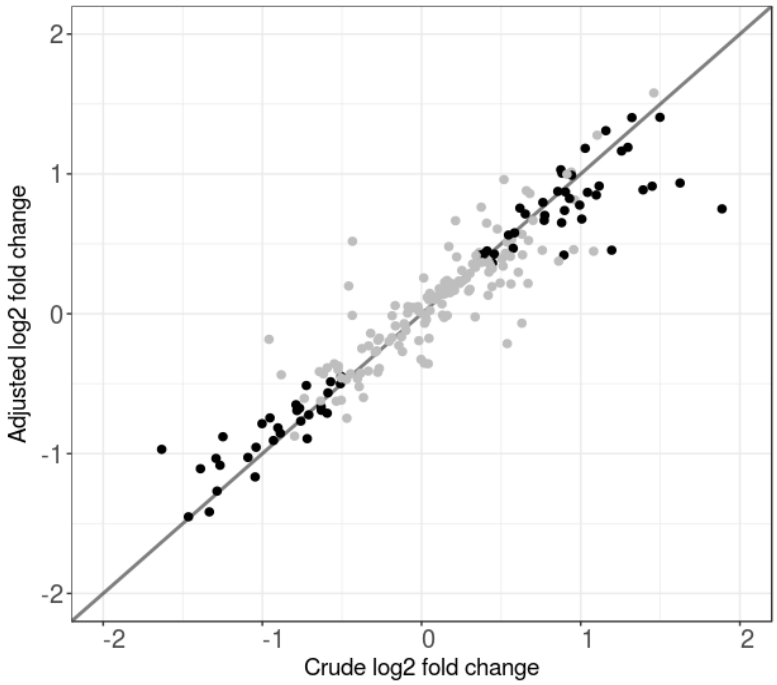


Figure 2.2.3: Comparison between 212 coefficients for current (n=86) vs. never smokers (n=43) from crude and adjusted negative binomial log-linear regression (adjusted for age, sex, race/ethnicity, self-reported physical activity, education, diabetes status, self-reported gum disease). Points in the scatter plot represent all differentially abundant OTUs, regardless of statistical significance, with black dot OTUs significant with the Wald test in crude analyses; coordinates are determined by the $\log_2$ fold change resulting from the crude analysis between current and never smokers (x axis) and the $\log_2$ fold change from the adjusted analysis between current and never smokers (y axis).

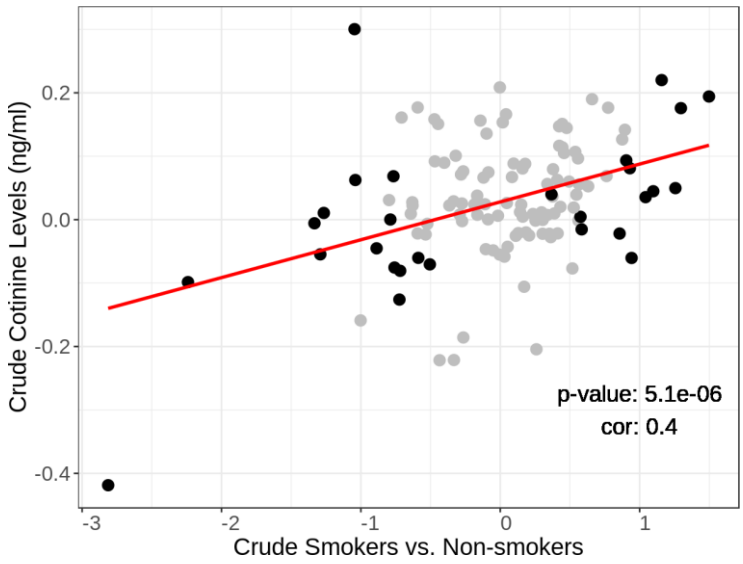


Figure 2.2.4: Comparison of $\log_2$ OTU fold changes between a) crude analyses of smokers (n=86) vs non-smokers with no detectable serum cotinine (n=43), and b) analysis of continuous cotinine levels among non-smokers exposed to secondhand smoke (n=38). Plot shows 121 OTUs that passed the edgeR filter for low-variance variables, including 28 OTUs that are differentially abundant in smokers (FDR <0.05, indicated in black). Coefficients for these two contrasts, involving different measures of exposure on different individuals, are positively correlated (Pearson Correlation = 0.58, p = 0.0013 for 28 OTUs; Pearson Correlation = 0.40, p = 5e-6 for all 121 OTUs)

Discussion

This study analyzes oral mouthwash specimens from a sub-sample of NYC HANES 2013-14 to provide multiple lines of causal inference supporting the hypothesis that tobacco smoke exposure alters the saliva microbiome. We found current smokers to harbour a different microbial composition compared to never smokers and the other tobacco exposure groups in terms of beta diversity, individual OTUs, and oxygen requirements. The microbiome of former smokers was more similar to never smokers than to current smokers. Phyla Candidate division SR1, *Bacteroides* and *Proteobacteria* were depleted in smokers with genera *Bergeyella*, *Porphyromonas*, *Prevotella*, *Haemophilus*, *Neisseria*, *Lautropia* and *Actinobacillus*, consistent with previous studies (Wu *et al*, 2016a; Morris *et al*, 2013; Charlson *et al*, 2010). The depletion of *Proteobacteria* may be especially important as this depletion has also been found among individuals with periodontal disease compared to healthy controls (Griffen *et al*, 2012). Further, *Proteobacteria* levels in the oral microbiome have been associated with insulin resistance and inflammation (Demmer *et al*, 2017). These shifts in genera largely remained with adjustment for hypothesized confounders.

The large microbiome shifts associated with current smokers compared to never smokers included significant depletion of oxygen-requiring bacteria, and corresponding (but not statistically significant) enrichment of anaerobic bacteria. This finding is consistent with a proposed mechanism (Kenney *et al*, 1975) by which smoking alters the oxygenation of the oral cavity, depleting oxygen and favouring anaerobic bacteria. Furthermore, the shifts in mean abundance occurring between current smokers and never smokers were positively correlated to those observed among non-smokers with varying levels of secondhand smoke exposure as measured by serum cotinine. This indicates a dose-response relationship for secondhand smoke exposure, and reduces the plausibility of residual confounding as an explanation for the shifts observed in these separate groups of participants.

Reduced aerobiosis and increased anaerobiosis in the oral cavity have implications for oral and systemic health. The Red Complex, a trio of anaerobic bacteria (*Treponema denticola*, *Porphyromonas gingivalis* and *Tannerella forsythia*) are linked with the development of periodontal disease (Rocas *et al*, 2001). While these primarily inhabit the dental plaque, an overall anaerobic oral environment may facilitate colonization. Although this study did not provide species-level resolution to observe the Red Complex, a previous study (Guglielmetti *et al*, 2014) found increased abundance of *Porphyromonas gingivalis* and *Tannerella forsythia* in the subgingival plaque of smokers compared to non-smokers. Furthermore, oral anaerobiosis could provide greater opportunity for movement of oral bacteria to distant anaerobic environments in the stomach and gut. This study demonstrates how oxygen

utilization provides a simplifying measure that can be used by future studies of the oral microbiome and health.

In a mixed group of users of alternative smoking products including e-cigarettes, hookahs, cigars and cigarillos, we found some alterations comparable to those in cigarette smokers (like *Lactococcus* and *Neisseria* genera), while others like *Porphyromonas* had an opposite trend. This small and heterogeneous group of alternative smoking products does not allow robust conclusions, but indicates the possibility that alternative products could alter the oral microbiome composition in ways similar to cigarette smoke. As e-cigarettes are gaining popularity in use, (Department of Health and Human Services, 2016) more research is needed to explore the effect of vaping on the oral microbiome.

This study has a number of limitations. As a cross-sectional study, changes to the oral microbiome in direct response to tobacco exposure were not measured; longitudinal data are needed to directly observe tobacco-induced changes to the oral microbiome. We defined secondhand exposure to smoke using a serum cotinine upper cut-off of 14 ng/mL (Benowitz *et al*, 2009a) Four of 38 participants in the secondhand exposure group had serum cotinine higher than the cut-off value of 10 ng/mL (Homa *et al*, 2015), and may have misreported recent light cigarette usage. However, removing these four individuals had negligible effect on reported results. We adjusted for self-reported gum disease as a measure of periodontal health; this measure is imperfect and residual confounding by periodontal health may remain. Additionally periodontal health may be a mediator rather than a confounder. Regardless, self-reported gum disease was not strongly associated with beta diversity in our analyses, so adjusting for it should not impact results. Additionally, we did not adjust for differences in dietary habits given the general lack of validity of self-reported diet data (Schoeller, 1995); we cannot rule out residual confounding by diet. However, our findings, which show a shift toward anaerobic bacteria among smokers make biological sense in response to smoke exposure, reducing the likelihood that these findings are the result of residual confounding. Finally, the current analysis is based on 16S rRNA gene analyses capturing only genus and higher-level taxonomic information; whole metagenomic sequencing may provide additional important information on shifts to the oral cavity caused by tobacco exposures and functional information.

The strengths of this study are that it included a racially/ethnically diverse group of participants and an array of tobacco exposure groups including self-reported non-smokers exposed to secondhand smoke. The study design allowed multiple, complementary comparisons, as well as biological analysis, to help distinguish causal associations from associations likely to be caused by residual confounding. This study introduces several analyses that are, to the best

of our knowledge, novel to epidemiological studies of the human microbiome. These include: 1) the application of Gene Set Enrichment Analysis methods for biological interpretation of gross microbiome shifts, 2) use of a scatter plot to visualize the comparison of crude vs. adjusted regression coefficients in high-dimensional data, and 3) application of the Integrative Correlation Coefficient (Parmigiani *et al*, 2004; Cope *et al*, 2014), a method introduced to assess reproducibility of gene expression studies, to make causal inference by comparing regression coefficients from different samples with different measures of tobacco exposure (smokers vs. non-smokers, and dose-response for continuous serum cotinine measurements).

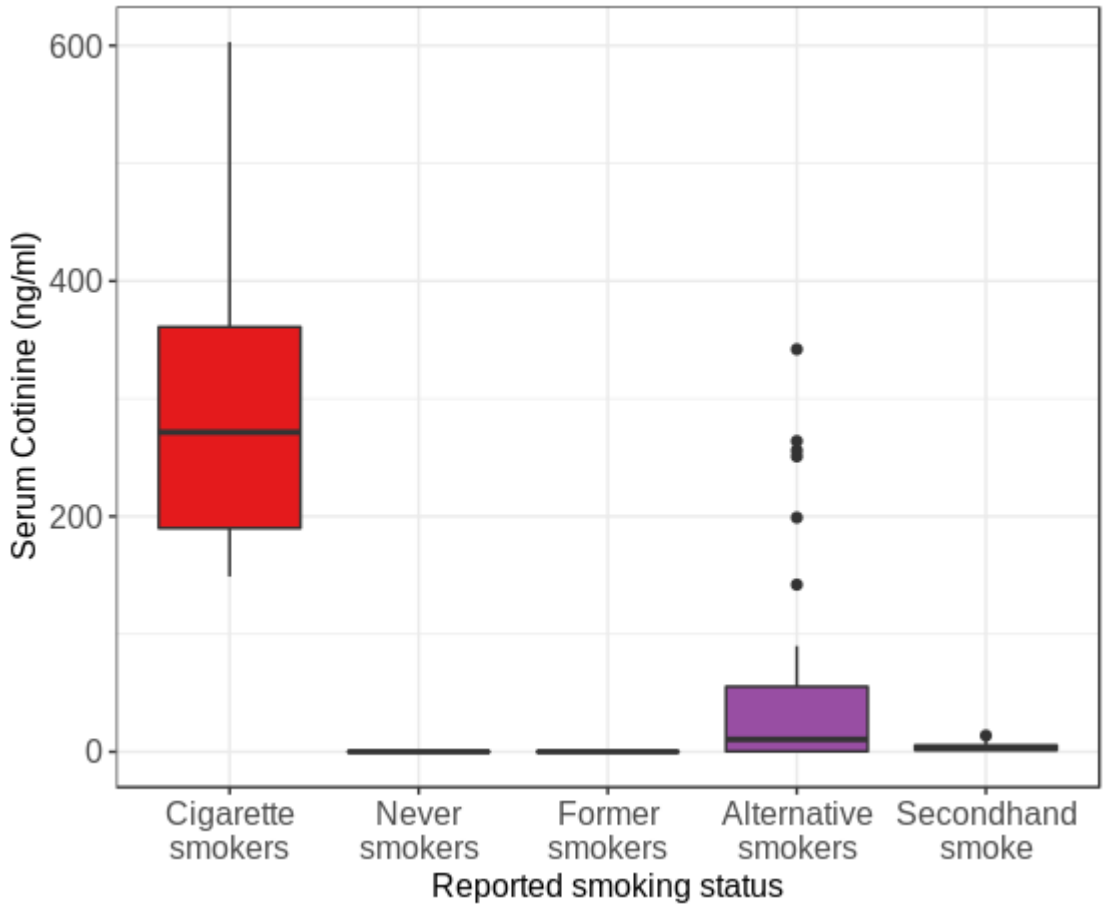
Conclusions

Overall shifts between aerobic and anaerobic microbiota is a relevant simplifying measure that should be considered in future health studies of the oral microbiome. These results support a plausible biological mechanism for population-level shifts in the oral microbiome caused by exposure to tobacco smoke, through three lines of observational evidence: 1) consistency of the microbiome shifts with reduced microbiota oxygen utilization as a biological mechanism for the shifts observed in smokers; 2) consistency of oral microbiome abundance fold-changes in current smokers versus non-smokers with abundance changes along the gradient of secondhand smoke exposure among non-smokers; and 3) tobacco-related associations that are stronger than associations with sociodemographic and health indicators, and that are not meaningfully affected by controlling for hypothesized confounders.

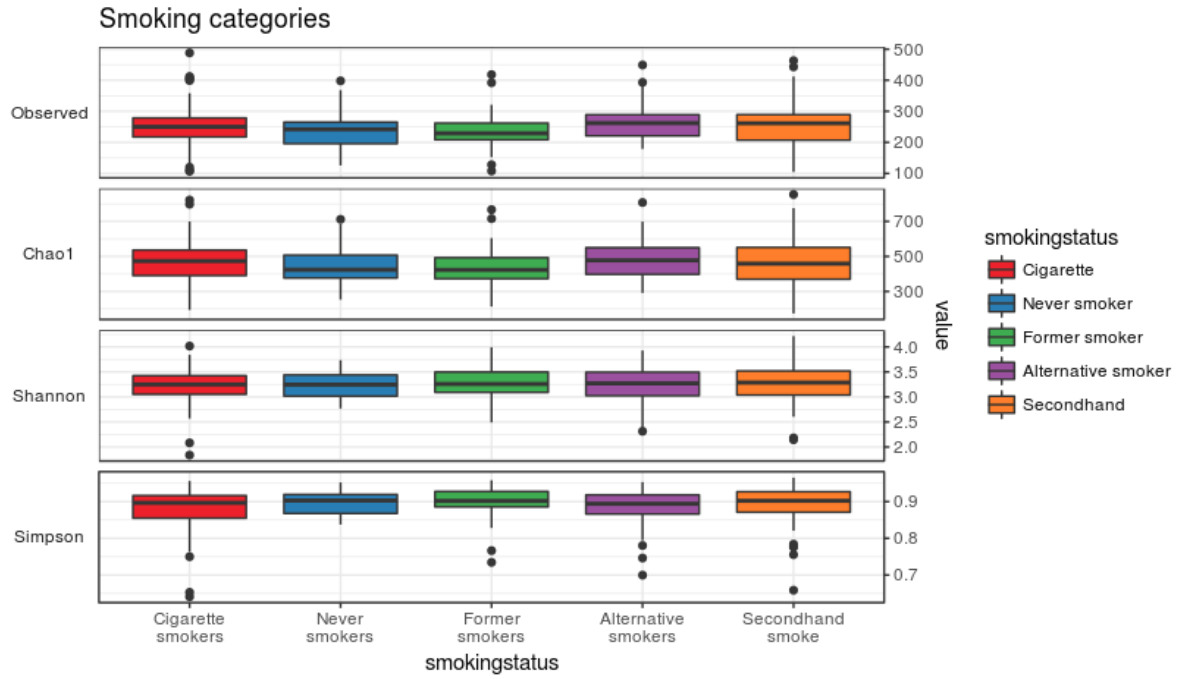
Acknowledgements: We would like to thank Sharon Perlman and Jennifer Rakeman-Cagno from the New York City Department of Health and Mental Hygiene for their collaboration during study implementation and design.

Funding: This study was supported by internal funds at the CUNY School of Public Health and Albert Einstein College of Medicine with salary support (JBD, AR, LW) from National Institute of Allergy and Infectious Diseases (1R21AI121784-01).

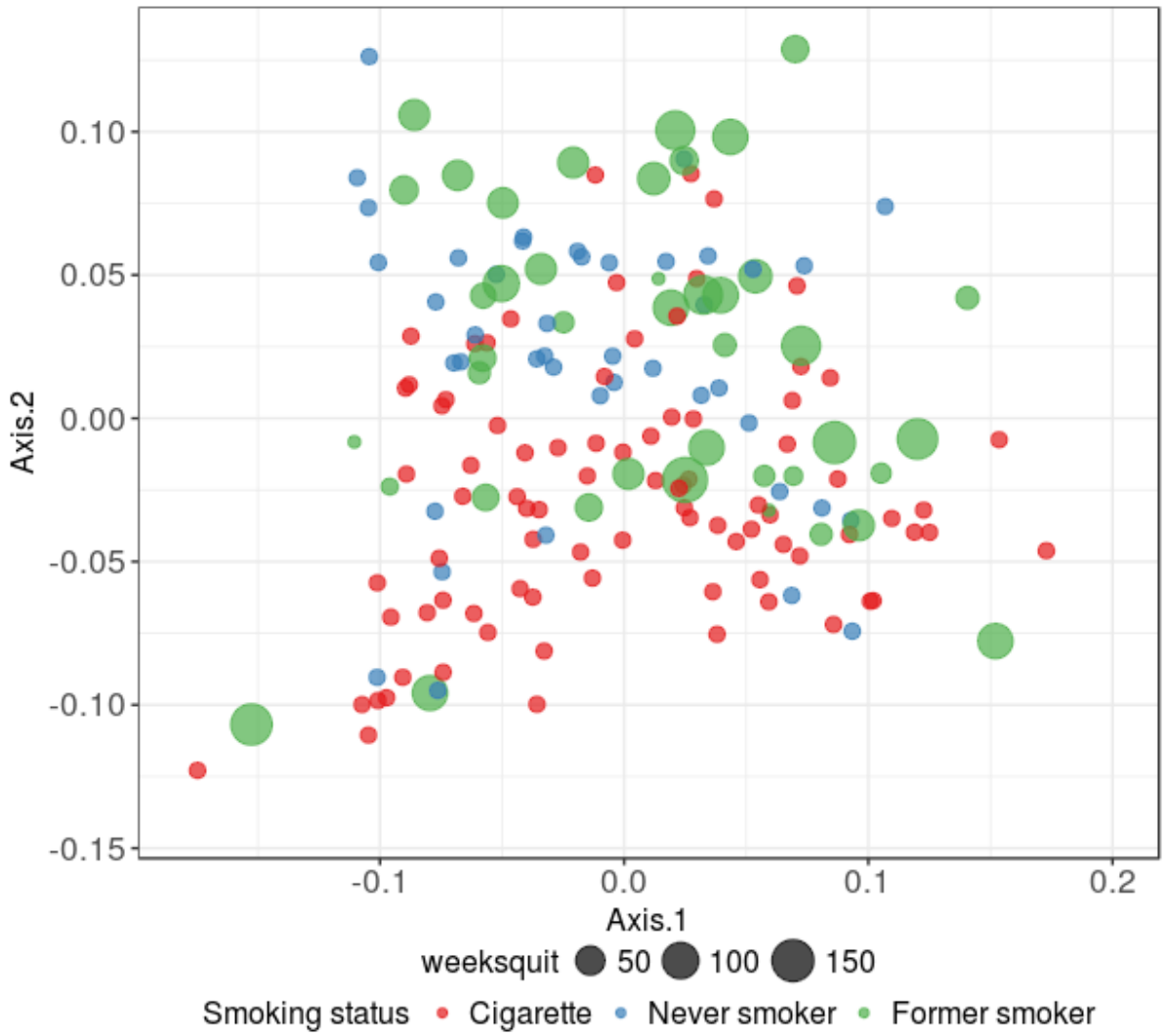
Supplementary figures



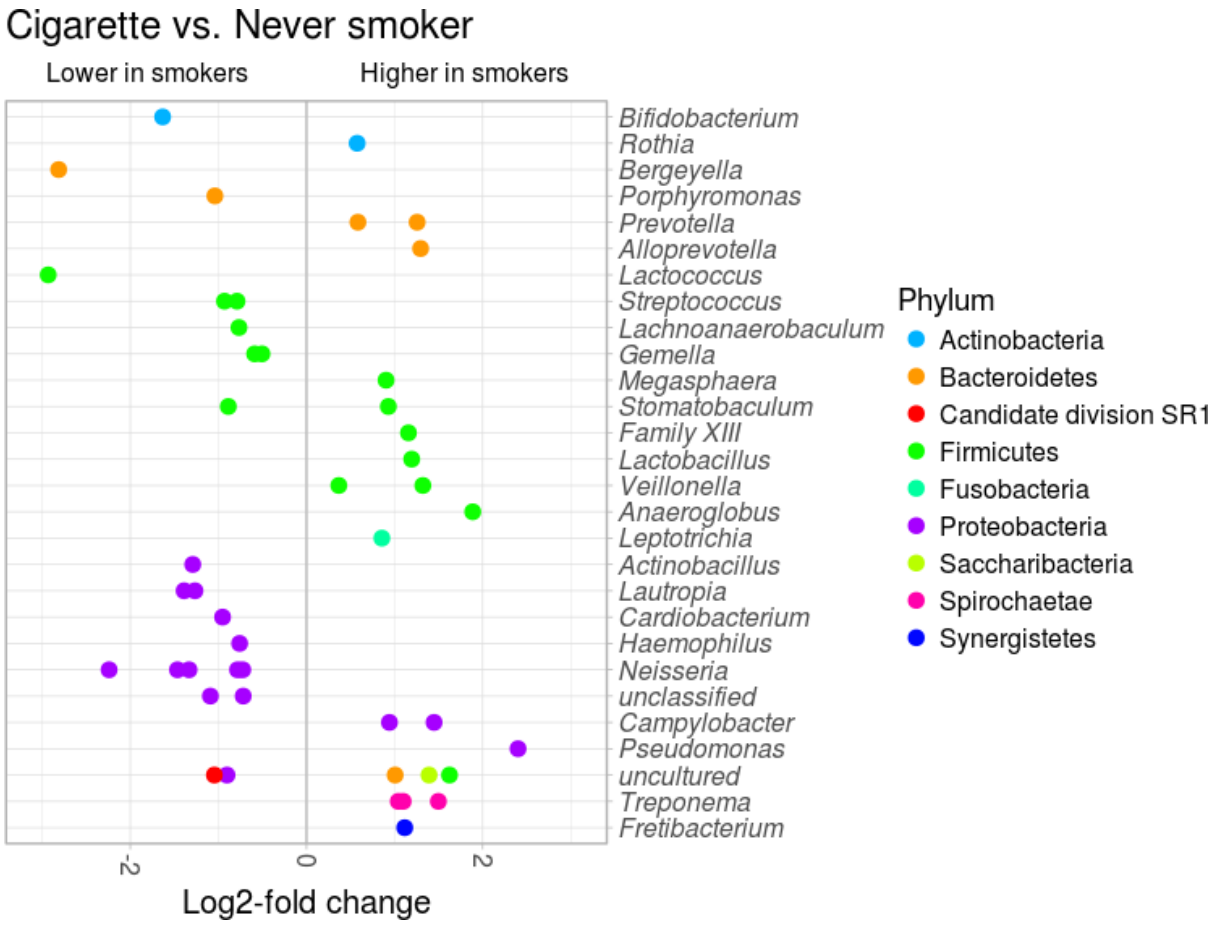
Supplementary Figure 2.2.1: Distribution of measured serum cotinine among the five smoke exposure groups with current cigarette smokers followed by smokers of alternative tobacco products showing the highest serum cotinine levels.



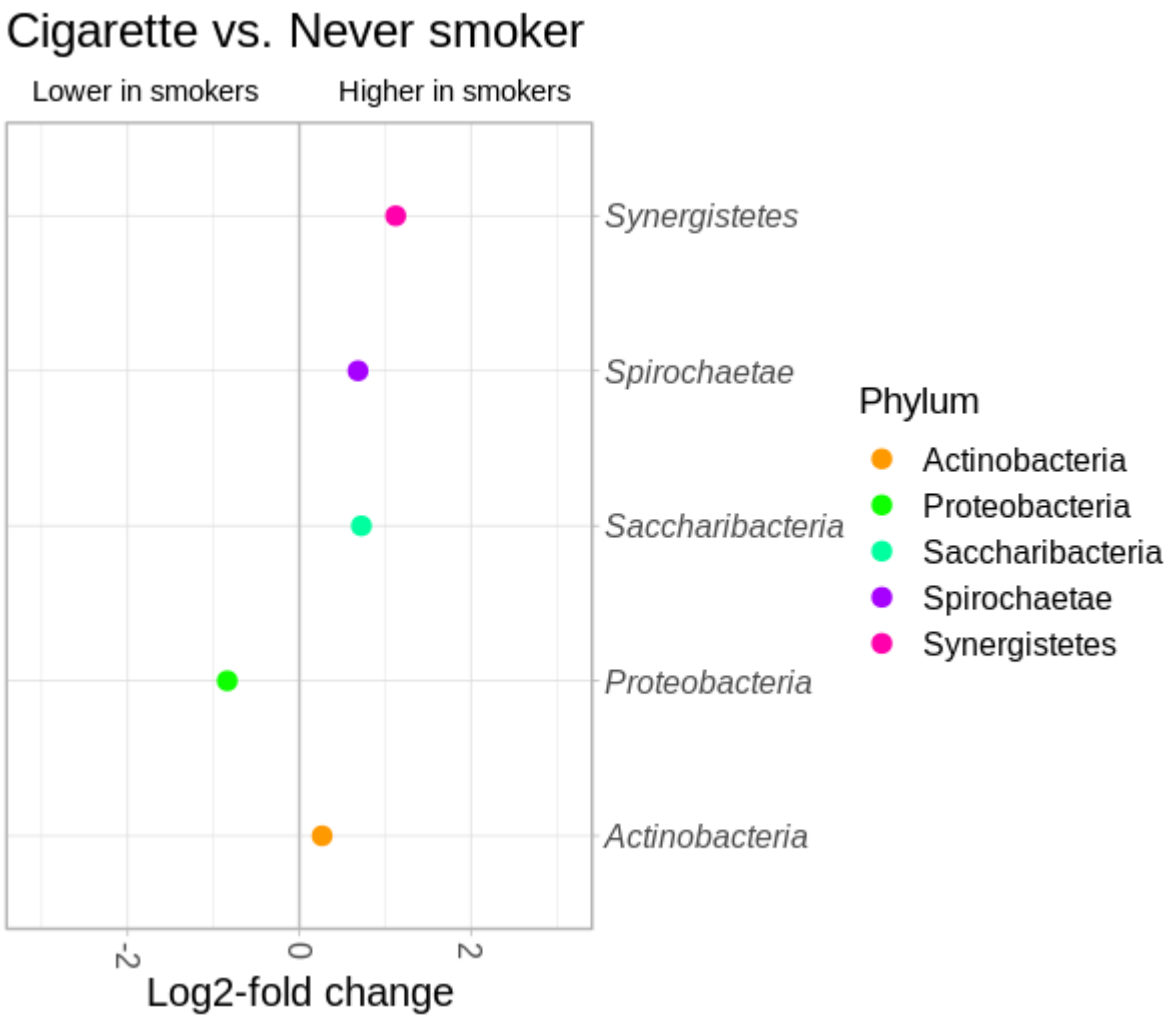
Supplementary Figure 2.2.2: Alpha diversity measures between the five tobacco exposure groups. No significant differences were observed between groups with four measures of richness and evenness.



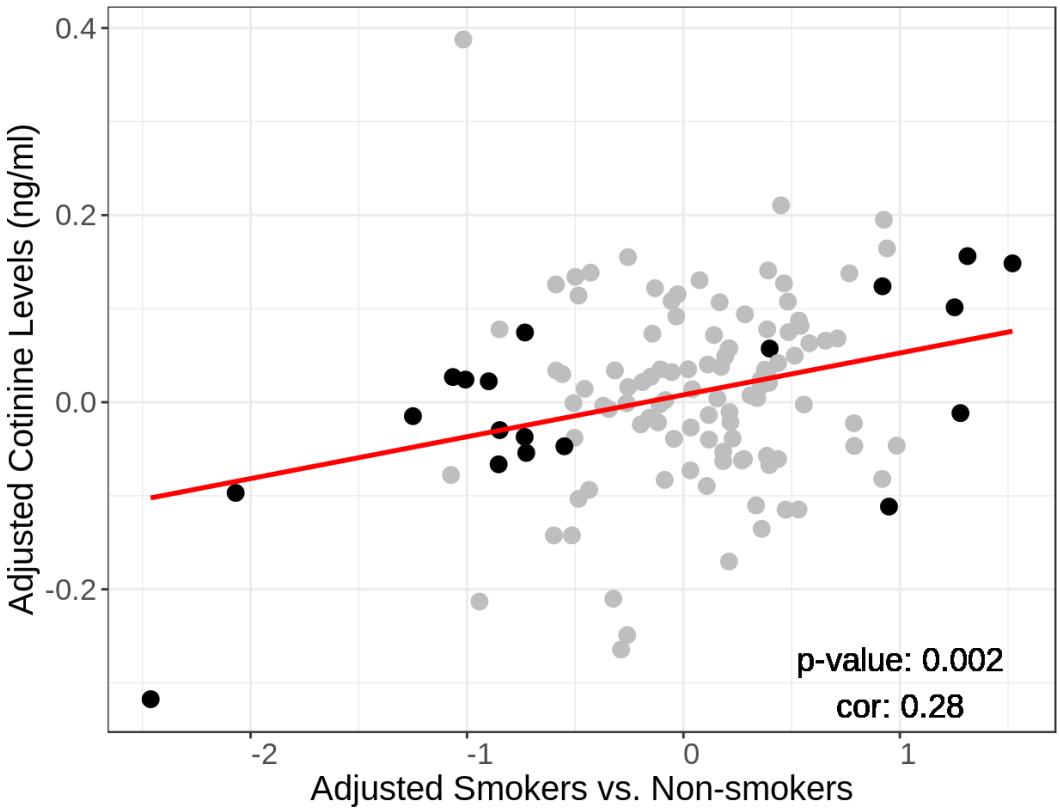
Supplementary Figure 2.2.3: Principal coordinates analysis based on the weighted UniFrac distance on samples from cigarette smokers (n=86), never smokers (n=43), and former smokers (n=43) with size of dot indicating how long ago they reported quitting.. A separation on the second axis between cigarette smokers and never smokers is present. We found former smokers were significantly different from current smokers (p=0.002, PERMANOVA test), but not from never smokers (p=0.16, PERMANOVA test).



Supplementary Figure 2.2.4: Crude differential analysis between current cigarette smokers (n=86) and never smokers (n=43). Dots in the plot are the 46 OTUs identified as differentially abundant without adjusting for hypothesized confounders and coloured according the taxonomy annotation at the phylum level. Position of the side of the plot is determined by the $\log_2$ fold change of abundance of the OTU.



Supplementary Figure 2.2.5: Crude differential analysis between current cigarette smokers (n=86) and never smokers (n=43) performed at the phylum level.



Supplementary Figure 2.2.6: Equivalent of Figure 2.2.4, with adjustment for age and educational attainment. Comparison of log_2 OTU fold changes between a) adjusted analyses of (x-axis) smokers (n=86) vs non-smokers with no detectable serum cotinine (n=43) and b) adjusted analysis of continuous cotinine levels among non-smokers exposed to secondhand smoke (n=38). Plot shows 121 OTUs that passed the edgeR filter for low-variance variables, including 19 OTUs differentially abundant in smokers (FDR<0.05, indicated in black). Similar to the crude model in Figure 2.2.4, a positive correlation was observed (Pearson Correlation = 0.68, p = 0.001 for 19 OTUs; Pearson Correlation = 0.28, p = 0.002 for all 121 OTUs).

2.3. Other contributions

In this chapter, I will report other works that I contributed to during my doctoral studies and that resulted in a published article including myself as co-author.

I will introduce each article with a short paragraph describing the study rationale and my contributions to the research. Each study will be reported only by the abstract and a link to the online full version of the article.

2.3.1. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0

(Asnicar *et al*, 2020) reports PhyloPhlAn 3, a novel phylogenetic framework able to accurately reconstruct fine-grained (species- and strain-level) and large size (tree of life) phylogenies. This novel method was designed to better handle a larger number of reference genomes when it comes to investigating their phylogenetic relationships. PhyloPhlAn 3 is a completely automatic pipeline able to retrieve species-specific markers and reference genomes for the species of interest, a novel feature compared to the previous version (Segata *et al*, 2013b). Leveraging the species-level genome bin (SGB) system described in (Pasolli *et al*, 2019) and reported below, PhyloPhlAn 3 can find the closest SGB given a reference genome.

My contribution to this work was the development of a module for ChocoPhlAn (see **Section 1**) that, starting from a species of interest, retrieves the UniRef90 gene families identified as core proteins and all the available reference genomes. This is described in the **Methods** section and in **Figure 1** of the paper. Moreover, the pipeline is also described in **Section 1** of this thesis.

Francesco Asnicar, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, Mattia Bolzan, Fabio Cumbo, Uyen May, Jon G Sanders, Moreno Zolfo, Evguenia Kopylova, Edoardo Pasolli, Rob Knight, Siavash Mirarab, Curtis Huttenhower, Nicola Segata

Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0

Nature Communications (2020) - https://doi.org/10.1038/s41467-020-16366-7

Abstract

Microbial genomes are available at an ever-increasing pace, as cultivation and sequencing become cheaper and obtaining metagenome-assembled genomes (MAGs) becomes more effective. Phylogenetic placement methods to contextualize hundreds of thousands of genomes must thus be efficiently scalable and sensitive from closely related strains to divergent phyla. We present PhyloPhlAn 3.0, an accurate, rapid, and easy-to-use method for large-scale microbial genome characterization and phylogenetic analysis at multiple levels of resolution. PhyloPhlAn 3.0 can assign genomes from isolate sequencing or MAGs to species-level genome bins built from >230,000 publically available sequences. For individual clades of interest, it reconstructs strain-level phylogenies from among the closest species using clade-specific maximally informative markers. At the other extreme of resolution, it scales to large phylogenies comprising >17,000 microbial species. Examples including *Staphylococcus aureus* isolates, gut metagenomes, and meta-analyses demonstrate the ability of PhyloPhlAn 3.0 to support genomic and metagenomic analyses.

2.3.2. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

This article by (Pasolli *et al*, 2019) presents a novel framework for retrieving metagenomic assembled genomes (MAGs) from public datasets after metagenomic assembly and binning. To date, this is the largest-scale effort made for retrieving novel MAGs, such novelty allowed to explain a large unexplained fraction of the human microbiome. A large number of reconstructed genomes were previously unknown and still-to-be characterized. Briefly, our group was able to reconstruct more than 154,000 microbial genomes from 9,428 metagenomic samples from mostly all the continents. In order to assign them a species-level-like label (spanning 5% genetic distance), a massive clustering effort was performed to identify 4,930 species-level genome bins (SGBs). Leveraging sample-associated metadata and the novel PhyloPhlAn 3 pipeline, we described the geographical association of several unknown species (uSGB) and explored novel species associated with non-Westernized populations.

In this work, I contributed by performing the massive functional and metabolic annotations of the whole set of 154,723 MAGs using UniRef90 and UniRef50. **Figure 3**, **Figure 5**, and **Figure S1** of the manuscript report some key results enabled by the functional characterization of these novel microbial genomes.

Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, Nicola Segata

Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle

Cell (2019) - https://doi.org/10.1016/j.cell.2019.01.001

Abstract

The body-wide human microbiome plays a role in health, but its full diversity remains uncharacterized, particularly outside of the gut and in international populations. We leveraged 9,428 metagenomes to reconstruct 154,723 microbial genomes (45% of high quality) spanning body sites, ages, countries, and lifestyles. We recapitulated 4,930 species-level genome bins (SGBs), 77% without genomes in public repositories (uSGBs). uSGBs are prevalent (in 93% of well-assembled samples), expand underrepresented phyla, and are enriched in non-Westernized populations (40% of the total SGBs). We annotated 2.85 M genes in SGBs, many associated with conditions including infant development (94,000) or Westernization (106,000). SGBs and uSGBs permit deeper microbiome analyses and increase the average mappability of metagenomic reads from 67.76% to 87.51% in the gut (median 94.26%) and 65.14% to 82.34% in the mouth. We thus identify thousands of microbial genomes from yet-to-be-named species, expand the pangenomes of human-associated microbes, and allow better exploitation of metagenomic technologies.

2.3.3. Microbial genomes from gut metagenomes of non-human primates expand the primate-associated bacterial tree-of-life with over 1,000 novel species

In the context of novel species' discovery, (Manara *et al*, 2019) describes the application of the framework presented by (Pasolli *et al*, 2019) (and previously reported) to cohorts of non-human primates (NHP). We performed metagenomic assembly and binning on 203 samples from 6 different studies aimed to characterize the gut microbiome of different non-human primates species. This effort resulted in the reconstruction of 2,985 MAGs, 2,186 of them describing 1,009 novel SGBs associated with non-human primates hosts (pSGB). Quantification of the mappability of the raw metagenomes showed that the novel identified pSGB increased the mappability by 600% and those novel species were accounting for more than 90% of the microbial diversity associated with NHP. Furthermore, using the MAG collection retrieved by (Pasolli *et al*, 2019), we observed that NHP held in captivity and exposed to the human environment were sharing a small number of species with humans, but non-westernized populations were the one that showed the higher number of shared species, most of them are still-to-be characterized.

In this work, I conducted the functional annotation of the MAGs with UniRef90 and UniRef50 for the successive functional analysis presented in **Figure 5** and **Figure S4**. I also performed all the statistical analyses on the functional profiles.

Serena Manara, Francesco Asnicar, Francesco Beghini, Davide Bazzani, Fabio Cumbo, Moreno Zolfo, Eleonora Nigro, Nicolai Karcher, Paolo Manghi, Marisa Isabell Metzger, Edoardo Pasolli, Nicola Segata

Microbial genomes from gut metagenomes of non-human primates expand the primate-associated bacterial tree-of-life with over 1,000 novel species

Genome Biology (2019) - https://doi.org/10.1186/s13059-019-1923-9

Abstract

Background

Humans have coevolved with microbial communities to establish a mutually advantageous relationship that is still poorly characterized and can provide a better understanding of the human microbiome. Comparative metagenomic analysis of human and non-human primate (NHP) microbiomes offer a promising approach to study this symbiosis. Very few microbial species have been characterized in NHP microbiomes due to their poor representation in the available cataloged microbial diversity, thus limiting the potential of such comparative approaches.

Results

We reconstruct over 1000 previously uncharacterized microbial species from 6 available NHP metagenomic cohorts, resulting in an increase of the mappable fraction of metagenomic reads by 600%. These novel species highlight that almost 90% of the microbial diversity associated with NHPs has been overlooked. Comparative analysis of this new catalog of taxa with the collection of over 150,000 genomes from human metagenomes points at a limited species-level overlap, with only 20% of microbial candidate species in NHPs also found in the human microbiome. This overlap occurs mainly between NHPs and non-Westernized human populations and NHPs living in captivity, suggesting that host lifestyle plays a role comparable to host speciation in shaping the primate intestinal microbiome. Several NHP-specific species are phylogenetically related to human-associated microbes, such as Elusimicrobia and Treponema, and could be the consequence of host-dependent evolutionary trajectories.

Conclusions

The newly reconstructed species greatly expand the microbial diversity associated with NHPs, thus enabling better interrogation of the primate microbiome and empowering in-depth human and non-human comparative and co-diversification studies.

2.3.4. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome

(Ferretti *et al*, 2018) presents a large-scale metagenomic investigation of a cohort of mothers and infants that were followed for four months after birth with samples collected from multiple body sites. The main goal of this study was to identify and characterize the species vertically transmitted from mother to infant. Characterization of the transmission of the bacteria was resolved at the strain level. We observed multiple strains from the skin, vaginal, and gut microbiome of the mother transmitted to the infant's gut; gut-associated strains were observed to be able to colonize the gut more stably whereas skin and vaginal associated strains were only colonizing the gut transiently.

In this work, I performed the screening of the 215 samples for the presence of microbial eukaryotes and performed the statistical analysis presented in **Figure 1C**.

Pamela Ferretti, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Valentina Gorfer, Sabina Fedi, Federica Armanini, Duy Tin Truong, Serena Manara, Moreno Zolfo, Francesco Beghini, Roberto Bertorelli, Veronica De Sanctis, Ilaria Bariletti, Rosarita Canto, Rosanna Clementi, Marina Cologna, Tiziana Crifò, Giuseppina Cusumano, Stefania Gottardi, Claudia Innamorati, Caterina Masè, Daniela Postai, Daniela Savoi, Sabrina Duranti, Gabriele Andrea Lugli, Leonardo Mancabelli, Francesca Turroni, Chiara Ferrario, Christian Milani, Marta Mangifesta, Rosaria Anzalone, Alice Viappiani, Moran Yassour, Hera Vlamakis, Ramnik Xavier, Carmen Maria Collado, Omry Koren, Saverio Tateo, Massimo Soffiati, Anna Pedrotti, Marco Ventura, Curtis Huttenhower, Peer Bork, Nicola Segata

Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome

Cell Host & Microbe (2018) - https://doi.org/10.1016/j.chom.2018.06.005

Abstract

The acquisition and development of the infant microbiome are key to establishing a healthy host-microbiome symbiosis. The maternal microbial reservoir is thought to play a crucial role in this process. However, the source and transmission routes of the infant pioneering microbes are poorly understood. To address this, we longitudinally sampled the microbiome of 25 mother-infant pairs across multiple body sites from birth up to 4 months postpartum. Strain-level metagenomic profiling showed a rapid influx of microbes at birth followed by strong selection during the first few days of life. Maternal skin and vaginal strains colonize only transiently, and the infant continues to acquire microbes from distinct maternal sources after birth. Maternal gut strains proved more persistent in the infant gut and ecologically better adapted than those acquired from other sources. Together, these data describe the mother-to-infant microbiome transmission routes that are integral in the development of the infant microbiome.

2.3.5. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation

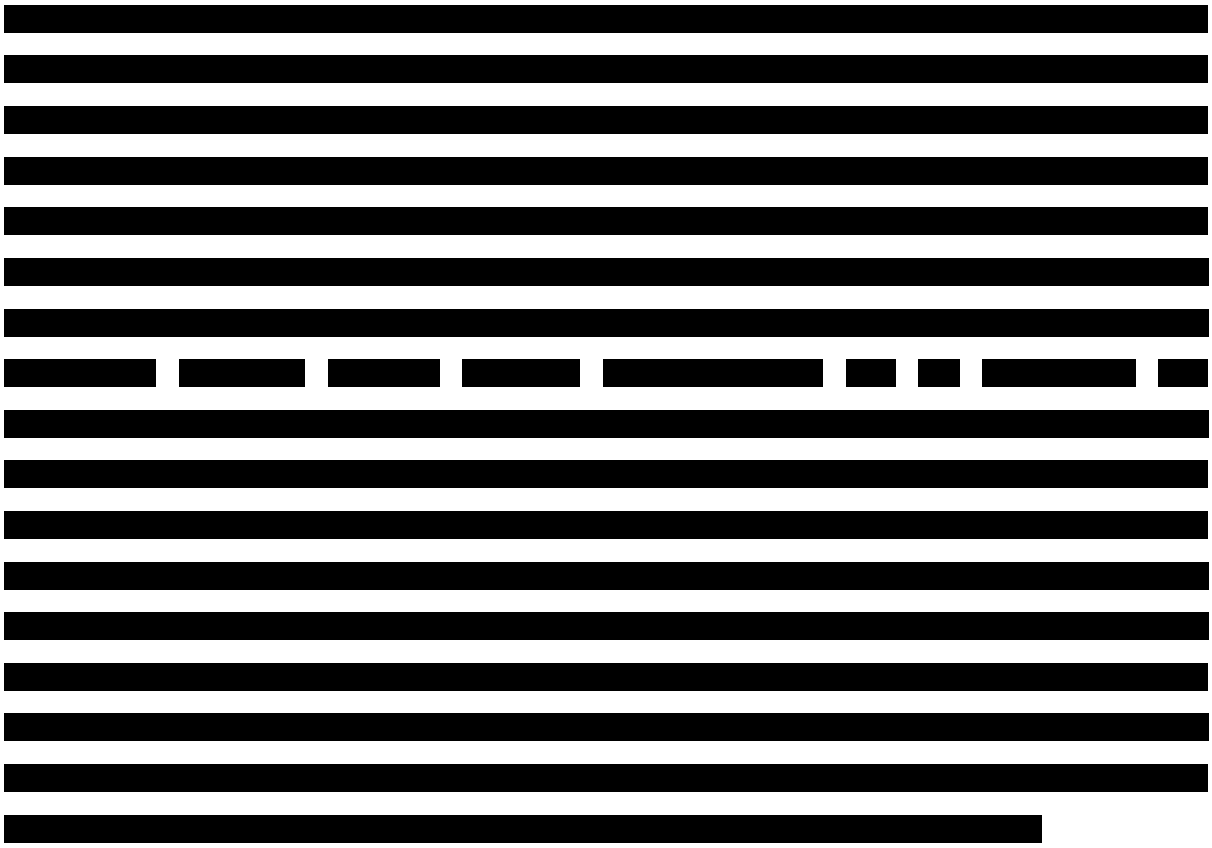
In this work, our group performed a meta-analysis of all the available datasets investigating a link between the gut microbiome and colorectal cancer (CRC). Besides, two new CRC cohorts from two Italian hospitals were included in the study. The main goal of the paper was to leverage machine learning approaches to identify microbial species that can be used as non-invasive biomarkers. The meta-analysis showed that the composition of the gut microbiome in CRC is consistent across different cohorts and we identified a link between the choline degradation pathway and CRC, suggesting a potential mechanism for carcinogenesis. We then identified a panel of species that could be possibly used as potential non-invasive biomarkers.

Andrew Thomas, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, Chiara Pozzi, Sara Gandini, Davide Serrano, Sonia Tarallo, Antonio Francavilla, Gaetano Gallo, Mario Trompetto, Francesca Cordero, Emmanuel Dias-Neto, João Setubal, Barbara Pardini, Maria Rescigno, Levi Waldron Alessio Naccarati, Nicola Segata

Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation

Nature Medicine (2019) - https://doi.org/10.1038/s41591-019-0405-7

Abstract



2.3.6. Sociodemographic variation in the oral microbiome

This paper, published in Annals of Epidemiology as a companion paper of the work previously reported in **Section 2.2** (Beghini *et al*, 2019) describes the oral microbiome of the subjects enrolled in the NYC-HANES population study by looking at a variety of sociodemographic measurements. We observed a total of 69 operational taxonomic units differently abundant in sociodemographic variables associated with health, race/ethnicity, and socioeconomic status, suggesting that the social disparities are reflected in the microbiome composition. This also highlighted the importance of considering social factors as confounders in microbiome studies.

As a companion paper of (Beghini *et al*, 2019) sharing the same data, in this work, I performed the 16S rRNA analysis with QIIME and contributed to the creation of the R package for future reproducibility of the sztudy through the creation of vignettes reporting the result of this study.

Audrey Renson, Heidi Jones, Francesco Beghini, Nicola Segata, Christine Zolnik, Mykhaylo Usyk, Thomas Moody, Lorna Thorpe, Robert Burk, Levi Waldron, Jennifer Dowd

Sociodemographic variation in the oral microbiome

Annals of Epidemiology (2019) - https://doi.org/10.1016/j.annepidem.2019.03.006

Abstract

Purpose: Variations in the oral microbiome are potentially implicated in social inequalities in oral disease, cancers, and metabolic disease. We describe sociodemographic variation of oral microbiomes in a diverse sample.

Methods: We performed 16S rRNA sequencing on mouthwash specimens in a subsample (n = 282) of the 2013–2014 population-based New York City Health and Nutrition Examination Study. We examined differential abundance of 216 operational taxonomic units, and alpha and beta diversity by age, sex, income, education, nativity, and race/ethnicity. For comparison, we examined differential abundance by diet, smoking status, and oral health behaviors.

Results: Sixty-nine operational taxonomic units were differentially abundant by any sociodemographic variable (false discovery rate < 0.01), including 27 by race/ethnicity, 21 by family income, 19 by education, 3 by sex. We found 49 differentially abundant by smoking status, 23 by diet, 12 by oral health behaviors. Genera differing for multiple sociodemographic characteristics included *Lactobacillus*, *Prevotella*, *Porphyromonas*, *Fusobacterium*.

Conclusions: We identified oral microbiome variation consistent with health inequalities, more taxa differing by race/ethnicity than diet, and more by SES variables than oral health behaviors. Investigation is warranted into possible mediating effects of the oral microbiome in social disparities in oral and metabolic diseases and cancers.

2.3.7. How inoculation affects the development and the performances of microalgal-bacterial consortia treating real municipal wastewater

Different from the works presented before that were prevalently human-associated studies, this work was conducted on municipal wastewater coming from the Trento's wastewater treatment plant. In collaboration with the Environmental Engineering department of the Università di Trento, we investigated the effects of different wastewater treatment using 16S rRNA sequencing. Two photobioreactors, one inoculated with a culture of *Chlorella vulgaris,* a microalga used for treating wastewater, and one without any treatment, were fed with wastewater and monitored for 100 days to characterize its microbial community and performances. We observed that the treatment with *Chlorella* was not affecting the photoreactor performance since photosynthetic microorganisms were able to grow spontaneously without any seeding, and the composition of the microbiome was similar after reaching the steady-state.

In this work, I contributed by performing the 16S rRNA data analysis with the QIIME pipeline.

Serena Petrini, Paola Foladori, Francesco Beghini, Federica Armanini, Nicola Segata, Gianni Andreottola

How inoculation affects the development and the performances of microalgal-bacterial consortia treating real municipal wastewater

Journal of Environmental Management (2020) - https://doi.org/10.1016/j.jenvman.2020.110427

Abstract

To date, little is known about the start-up of photobioreactors and the progressive development of stable microalgal-bacterial consortia with a view to the full-scale treatment of real wastewater. Two photo-sequencing bioreactors, one inoculated with *Chlorella vulgaris* (RC) and one with the absence of inoculum (RW), were fed with real municipal wastewater and run in parallel for 101 days. The influence of the inoculation was evaluated in terms of pollutant removal efficiency, excess sludge production, solids settleability and microbial community characteristics. No significant differences were observed in the removal of COD (89 ± 4%; 88 ± 3%) and ammonium (99 ± 1%; 99 ± 1%), mainly associated with bacteria activity. During the first weeks of acclimation, *Chlorella vulgaris* in RC promoted better P removal and very high variations of DO and pH. Conversely, under steady-state conditions, no significant differences were observed between the performances of RC and RW, showing good settleability and low effluent solids, 7 ± 8 and 13 ± 10 mg TSS/L respectively. Microbiome analysis via 16S rRNA gene sequencing showed that, despite a different evolution, the microbial community was quite similar in both reactors under steady state conditions. Overall, the results suggested that the inoculation of microalgae is not essential to engender a photobioreactor aimed at treating real municipal wastewater.

2.4. Conclusions

Development of novel computational metagenomics methods able to handle the pace of ever-growing availability of reference data and continuous integration of those into a curated database is of utmost importance for the generation of accurate profiles and performing cutting edge analyses. Despite the enormous amount of available data, researchers still face many challenges at leveraging them.

Here, in **Section 1**, I introduced ChocoPhlAn, an updated database of functionally annotated reference genomes and gene families that leverages the UniRef database. I also introduced bioBakery 3, an integrated environment for metagenomic analysis encompassing tools for data quality control, taxonomic, strain-level, functional, and phylogenetic profiling. Using the bioBakery 3 suite, we expanded the knowledge of microbial signatures associated with colorectal cancer by identifying novel biomarkers. By leveraging metatranscriptomics profiles, we identified enzyme families particularly over-regulated in IBD patients from the Integrative Human Microbiome Project (HMP2). Updated and improved taxonomic and functional profiles of the HMP2 study will be made available online and accessible for everyone interested in re-analyzing the data. An updated reference set of genomes allowed an in-depth understanding of the genomics and the biogeographic distribution of several reconstructed strains of *Ruminococcus bromii*, including the identification of two subclades associated with geography.

Thanks to the advancement of the methodology, in the coming years, the availability of new data will increase and automatic methods for data collection and handling are highly needed. Recent releases of metagenome-assembled genomes (MAGs) encompassing known and still-to-be characterized microbial species require an infrastructure able to process them and use its intrinsic information. The ChocoPhlAn pipeline developed as an integrated system for genomic organization can be a valuable tool for this purpose, although there are still challenges related to the taxonomy and the integration with current methods. MAGs reconstructed from metagenomes can be exploited for *de-novo* discovery of eukaryotes and viruses. Therefore, using metagenomic assembly and binning, in **Section 2.3.2** we showed how we expanded the catalog of known and unknown prokaryotic species by reconstructing more than 150,000 genomes from human hosts and further expanded our knowledge of non-human primates associated bacteria in **Section 2.3.6**. These works led to the release of novel bacterial genomes which greatly increased the metagenomes mappability and unraveled the missing microbial diversity.

Increasing awareness regarding the importance of non-bacterial microorganisms in the human microbiome has led researchers to investigate the presence of microbial eukaryotes and

viruses in the human microbiome. This is challenging for different reasons that range from the shortage of good reference genomes, both for eukaryotes and viruses, to the lack of studies that address this unexplored portion of the microbiome. In **Section 2.1** I presented a study conducted on more than 2,000 metagenomes to assess the presence of *Blastocystis* spp.: our goal was to determine the presence and prevalence in several cohorts targeting different populations all around the world. We observed a high prevalence of *Blastocystis*, averaging 15% of the subjects; in particular, its prevalence is higher in healthy subjects, and when present, it is able to long-term colonize the gut. We also observed that non-westernized populations tend to have higher values of prevalence. Taken together, these results suggest that *Blastocystis* may be a member of the healthy gut. We further demonstrated that metagenomic assembly of *Blastocystis* can be performed, allowing for in-depth phylogenetic and functional analyses. We have to shift from the paradigm that microbial eukaryotes are associated with negative outcomes and seeding the idea that they play a key role in gut ecology.

In **Section 2.2** we hypothesized that tobacco smoke exposure is able to alter the saliva microbiome and we provided multiple evidence for supporting causal inference. A subset of subjects enrolled in the NYC HANES 2013-14, a community-based health survey, was selected for assessment of the oral microbiome, self-reported tobacco use, and serum blood cotinine level and we showed with 16S rRNA that exposition to tobacco smoke can alter the composition of the oral microbiome, in particular by depleting aerobic genera by favoring microbial anaerobiosis. First of a kind, this study showed the impact of second-hand smoke on oral species by assessing the environmental exposure using blood serum cotinine levels and showed that the microbiome of subjects using alternative smoking products (e.g., e-cigarettes, hookahs, cigars, and cigarillos) exhibit alterations comparable to those in cigarette smokers. To investigate the enrichment of aerobic and anaerobic genera, we proposed a Microbe Set Enrichment Analysis for interpreting shifts in the microbiome, an analysis based on the Gene Set Enrichment Analysis (GSEA) (Subramanian *et al*, 2005). By leveraging epidemiological methods, we further explored in **Section 2.3.5** links between the oral microbiome and several socioeconomic variables collected in the NYC HANES questionnaire. We identified several genera differently abundant associated with health and socioeconomic disparities, suggesting a possible role of the microbiome as a mediator of social disparities in disease outcomes. Although several microbial signatures were identified at the genus level using 16S rRNA sequencing, the sequencing methodology is a limitation that does not provide higher-level taxonomic information. This can be overcome by using whole metagenomic sequencing which provides species- and strain-level resolution and would better elucidate the impact of cigarette smoke on the community's functional information. An ethnically diverse

group such as this cohort and the availability of several subject-associated metadata coupled with shotgun metagenomics may enable more comprehensive investigations, helping unravel the influence of socioeconomic statuses on the oral microbiome.

Future advancements in sequencing methods will induce a decrease in sequencing costs and an increase in newly available data, a pattern we observed in the last decade. More data means better methods able to handle them and, in this thesis, I provided ChocoPhlAn as a blueprint for large-scale data organization. Increased depth of sequencing provides a larger amount of metagenomic data, allowing for characterization of previously underlooked microbes and the discovery of new ones, smoothing the path towards a more complete understanding of the microbial diversity. Current investigations on microbial diversity are mainly limited to the westernized population, with the majority of metagenomes sampled from Europe, United States of America, and Europe, introducing a strong bias toward the understanding of the microbial diversity of non-westernized populations that still harbor a substantial unknown diversity. Metagenome-assembled genomes are a good tool for exploring the microbial dark matter in under-investigated metagenomes, an exploration currently limited by the inaccessibility to cultivation recalcitrant species. Refinishing of retrieved MAGs and improvement of methods for metagenomic assembly and binning will result in the generation of high-quality MAGs.

2.5. Future perspectives

The work presented in this thesis represents a contribution to the computational metagenomics field, but there is still a lot to be done. After the completion of my doctoral studies, it is my interest to use the knowledge I acquired during the three years as a PhD candidate to pursue some aspects I have addressed in this thesis.

Integration of MAGs into current methods and the adoption of the SGB system would allow for a better understanding of the microbiome. In particular, the SGB integration into MetaPhlAn is of extreme importance since marker-based approaches are a powerful tool for tracking lowly abundant microbes. The first level of integration was already done with PhyloPhlAn, which, using phylogenetic placement, is able to assign uncharacterized MAGs to an SGB. Increasing availability of new MAGs for known species would help define the species boundaries while unraveling novel pangenomic diversity. Taxonomic labeling is an issue when it comes to newly discovered species that is substantially addressed by the SGB system by assigning a cluster identifier but there is the need to adopt a consistent standard.

Uniformly processed metagenomes and associated metadata as currently implemented by curatedMetagenomicData (Pasolli *et al*, 2017) is a valid tool for performing large-scale meta-analyses to find associations between specific members of the microbiome and anthropometric measurements or health conditions. The continuous update and expansion of this resource with new profiles obtained with the recent bioBakery tools using updated references would allow for discovering novel associations.

Extensive application of metagenomic assembly aimed to unravel the missing unknown diversity in metagenomes would help identify novel viral and eukaryotic species, as in the case of the crAssphage (Dutilh *et al*, 2014). A higher gut microbiome diversity in non-Westernized populations can be attributed to the presence of a quite diverse community of eukaryotes and it is of extreme importance to focalize the attention on non-Westernized populations inasmuch such unknown diversity that could help understanding the evolution of the gut microbiome after adopting a Westernized lifestyle.

2.6. Section references

Alfellani MA, Stensvold CR, Vidal-Lapiedra A, Onuoha ESU, Fagbenro-Beyioku AF & Clark CG (2013a) Variable geographic distribution of Blastocystis subtypes and its potential implications. *Acta Trop* 126: 11–18

Alfellani MA, Taner-Mulla D, Jacob AS, Imeede CA, Yoshikawa H, Stensvold CR & Clark CG (2013b) Genetic diversity of blastocystis in livestock and zoo animals. *Protist* 164: 497–509

Andersen L, Karim AB, Roager HM, Vigsnæs LK, Krogfelt KA, Licht TR & Stensvold CR (2016) Associations between common intestinal parasites and bacteria in humans as revealed by qPCR. *Eur J Clin Microbiol Infect Dis* 35: 1427–1431

Andersen LO, Bonde I, Nielsen HB & Stensvold CR (2015) A retrospective metagenomics approach to studying Blastocystis. *FEMS Microbiol Ecol* 91: fiv072

Andersen LO, Vedel Nielsen H & Stensvold CR (2013) Waiting for the human intestinal Eukaryotome. *ISME J* 7: 1253–1255

Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, Westra WH, Chung CH, Jordan RC, Lu C, *et al* (2010) Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med* 363: 24–35

Apprill A, McNally S, Parsons R & Weber L (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol* 75: 129–137

Aronesty E (2013) Comparison of sequencing utility programs. *Open Bioinforma J* 7

Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, *et al* (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* 11: 2500

Audebert C, Even G, Cian A, Blastocystis Investigation Group, Loywick A, Merlin S, Viscogliosi E & Chabé M (2016) Colonization with the enteric protozoa Blastocystis is associated with increased diversity of human gut bacterial microbiota. *Sci Rep* 6: 25255

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, *et al* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19: 455–477

Bart A, Wentink-Bonnema EMS, Gilis H, Verhaar N, Wassenaar CJA, van Vugt M, Goorhuis A & van Gool T (2013) Diagnosis and subtype analysis of Blastocystis sp. in 442 patients in a hospital setting in the Netherlands. *BMC Infect Dis* 13: 389

Beghini F, Renson A, Zolnik CP, Geistlinger L, Usyk M, Moody TU, Thorpe L, Dowd JB, Burk R, Segata N, *et al* (2019) Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study. *Ann Epidemiol* 34: 18-25.e3

Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol* 57: 289–300

Benowitz NL, Bernert JT, Caraballo RS, Holiday DB & Wang J (2009a) Optimal serum cotinine levels for distinguishing cigarette smokers and nonsmokers within different racial/ethnic groups in the United States between 1999 and 2004. *Am J Epidemiol* 169: 236–248

Benowitz NL, Hukkanen J & Jacob P 3rd (2009b) Nicotine chemistry, metabolism, kinetics and biomarkers. *Handb Exp Pharmacol*: 29–60

Bergey's Manual of Systematics of Archaea and Bacteria (2015) Wiley

Blaizot A, Vergnes J-N, Nuwwareh S, Amar J & Sixou M (2009) Periodontal diseases and cardiovascular events: meta-analysis of observational studies. *Int Dent J* 59: 197–209

Borgnakke WS, Ylöstalo PV, Taylor GW & Genco RJ (2013) Effect of periodontal disease on diabetes: systematic review of epidemiologic observational evidence. *J Periodontol* 84: S135-52

Breiman L (2001) Random Forests. *Mach Learn* 45: 5–32

Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV, *et al* (2011) Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci U S A* 108: 4494–4499

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A & Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18: 188–196

Capella-Gutiérrez S, Silla-Martínez JM & Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, *et al* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7: 335–336

Cekin AH, Cekin Y, Adakan Y, Tasdemir E, Koclar FG & Yolcular BO (2012) Blastocystosis in patients with gastrointestinal symptoms: a case-control study. *BMC Gastroenterol* 12: 122

Chaffee BW & Weston SJ (2010) Association between chronic periodontal disease and obesity: a systematic review and meta-analysis. *J Periodontol* 81: 1708–1724

Chang K, Yang SM, Kim SH, Han KH, Park SJ & Shin JI (2014) Smoking and rheumatoid arthritis. *Int J Mol Sci* 15: 22279–22295

Chang SA (2012) Smoking and type 2 diabetes mellitus. *Diabetes Metab J* 36: 399–403

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, Hwang J, Bushman FD & Collman RG (2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS One* 5: e15216

Chassin L, Presson CC, Rose JS & Sherman SJ (1996) The natural history of cigarette smoking from adolescence to adulthood: demographic predictors of continuity and change. *Health Psychol* 15: 478–484

Clark CG, van der Giezen M, Alfellani MA & Stensvold CR (2013) Recent developments in Blastocystis research. *Adv Parasitol* 82: 1–32

Colman G, Beighton D, Chalk AJ & Wake S (1976) Cigarette smoking and the microbial flora of the mouth. *Aust Dent J* 21: 111–118

Cope L, Naiman DQ & Parmigiani G (2014) Integrative correlation: Properties and relation to canonical correlations. *J Multivar Anal* 123: 270–280

Coyle CM, Varughese J, Weiss LM & Tanowitz HB (2011) Blastocystis: To Treat or Not to Treat…. *Clin Infect Dis* 54: 105–110

Curtis MA, Zenobia C & Darveau RP (2011) The relationship of the oral microbiotia to periodontal health and disease. *Cell Host Microbe* 10: 302–306

Demmer RT, Breskin A, Rosenbaum M, Zuk A, LeDuc C, Leibel R, Paster B, Desvarieux M, Jacobs DR Jr & Papapanou PN (2017) The subgingival microbiome, systemic inflammation and insulin resistance: The Oral Infections, Glucose Intolerance and Insulin Resistance Study. *J Clin Periodontol* 44: 255–265

Denoeud F, Roussel M, Noel B, Wawrzyniak I, Da Silva C, Diogon M, Viscogliosi E, Brochier-Armanet C, Couloux A, Poulain J, *et al* (2011) Genome sequence of the stramenopile Blastocystis, a human anaerobic parasite. *Genome Biol* 12: R29

Department of Health and Human Services US (2016) E-Cigarette use among youth and young adults. A report of the Surgeon General. *of Health and Human Services …*

Derelle R, López-García P, Timpano H & Moreira D (2016) A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Mol Biol Evol* 33: 2890–2898

Detert J, Pischon N, Burmester GR & Buttgereit F (2010) The association between rheumatoid arthritis and periodontal disease. *Arthritis Res Ther* 12: 218

Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, *et al* (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 5: 4498

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797

Eggert FM, McLeod MH & Flowerdew G (2001) Effects of smoking and treatment status on periodontal bacteria: evidence that smoking influences control of periodontal bacteria at the mucosal surface of the gingival crevice. *J Periodontol* 72: 1210–1220

Eichel B & Shahrik HA (1969) Tobacco smoke toxicity: loss of human oral leukocyte function and fluid-cell metabolism. *Science* 166: 1424–1428

El Safadi D, Gaayeb L, Meloni D, Cian A, Poirier P, Wawrzyniak I, Delbac F, Dabboussi F, Delhaes L, Seck M, *et al* (2014) Children of Senegal River Basin show the highest prevalence of Blastocystis sp. ever observed worldwide. *BMC Infect Dis* 14: 164

Emmons KM, Wechsler H, Dowdall G & Abraham M (1998) Predictors of smoking among US college students. *Am J Public Health* 88: 104–107

Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J & Huttenhower C (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS*

Comput Biol 8: e1002606

Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, Armanini F, Truong DT, Manara S, Zolfo M, *et al* (2018) Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* 24: 133-145.e5

Fitzpatrick SG & Katz J (2010) The association between periodontal disease and cancer: a review of the literature. *J Dent* 38: 83–95

Forsell J, Granlund M, Stensvold CR, Clark CG & Evengård B (2012) Subtype analysis of Blastocystis isolates in Swedish patients. *Eur J Clin Microbiol Infect Dis* 31: 1689–1696

Ganesan SM, Dabdoub SM, Nagaraja HN, Scott ML, Pamulapati S, Berman ML, Shields PG, Wewers ME & Kumar PS (2020) Adverse effects of electronic cigarettes on the disease-naive oral microbiome. *Sci Adv* 6: eaaz0108

Geistlinger L, Csaba G & Zimmer R (2016) Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics* 17: 45

Ghio AJ, Hilborn ED, Stonehuerner JG, Dailey LA, Carter JD, Richards JH, Crissman KM, Foronjy RF, Uyeminami DL & Pinkerton KE (2008) Particulate matter in cigarette smoke alters iron homeostasis to produce a biological effect. *Am J Respir Crit Care Med* 178: 1130–1138

Giannopoulou C, Cappuyns I & Mombelli A (2003) Effect of smoking on gingival crevicular fluid cytokine profile during experimental gingivitis. *J Clin Periodontol* 30: 996–1002

Giskes K, Kunst AE, Benach J, Borrell C, Costa G, Dahl E, Dalstra JA, Federico B, Helmert U, Judge K, *et al* (2005) Trends in smoking behaviour between 1985 and 2000 in nine European countries by education. *J Epidemiol Community Health* 59: 395–401

Griffen AL, Beall CJ, Campbell JH, Firestone ND, Kumar PS, Yang ZK, Podar M & Leys EJ (2012) Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J* 6: 1176–1185

Guglielmetti MR, Rosa EF, Lourencao DS, Inoue G, Gomes EF, De Micheli G, Mendes FM, Hirata RD, Hirata MH & Pannuti CM (2014) Detection and quantification of periodontal pathogens in smokers and never-smokers with chronic periodontitis by real-time polymerase chain reaction. *J Periodontol* 85: 1450–1457

Güntsch A, Erler M, Preshaw PM, Sigusch BW, Klinger G & Glockmann E (2006) Effect of smoking on crevicular polymorphonuclear neutrophil function in periodontally healthy subjects. *J Periodontal Res* 41: 184–188

Haber J (1994) Smoking is a major risk factor for periodontitis. *Curr Opin Periodontol*: 12–18

Hanioka T, Tanaka M, Takaya K, Matsumori Y & Shizukuishi S (2000) Pocket Oxygen Tension in Smokers and Non-Smokers With Periodontal Disease. *J Periodontol* 71: 550–554

Hänzelmann S, Castelo R & Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14: 7

Ho KK, Pinsky JL, Kannel WB & Levy D (1993) The epidemiology of heart failure: the

Framingham Study. *J Am Coll Cardiol* 22: 6A-13A

Homa DM, Neff LJ, King BA, Caraballo RS, Bunnell RE, Babb SD, Garrett BE, Sosnoff CS, Wang L & Centers for Disease Control and Prevention (CDC) (2015) Vital signs: disparities in nonsmokers' exposure to secondhand smoke--United States, 1999-2012. *MMWR Morb Mortal Wkly Rep* 64: 103–108

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 Res* 42: D617-24

Huerta-Cepas J, Forslund K, Szklarczyk D, Jensen LJ, von Mering C & Bork P (2016a) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Cold Spring Harbor Laboratory*: 076331

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, *et al* (2016b) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44: D286-93

Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207–214

Ide M & Papapanou PN (2013) Epidemiology of association between maternal periodontal disease and adverse pregnancy outcomes--systematic review. *J Clin Periodontol* 40 Suppl 14: S181-94

Jarvis MJ, Tunstall-Pedoe H, Feyerabend C, Vesey C & Saloojee Y (1987) Comparison of tests used to distinguish smokers from nonsmokers. *Am J Public Health* 77: 1435–1438

Joshi V, Matthews C, Aspiras M, de Jager M, Ward M & Kumar P (2014) Smoking decreases structural and functional resilience in the subgingival ecosystem. *J Clin Periodontol* 41: 1037–1047

Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, Nielsen J & Backhed F (2013) Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498: 99–103

Kenney EB, Saxe SR & Bowles RD (1975) The effect of cigarette smoking on anaerobiosis in the oral cavity. *J Periodontol* 46: 82–85

Kiełbasa SM, Wan R, Sato K, Horton P & Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res* 21: 487–493

Klemm E & Dougan G (2016) Advances in Understanding Bacterial Pathogenesis Gained from Whole-Genome Sequencing and Phylogenetics. *Cell Host Microbe* 19: 599–610

Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5: 59

Kumar PS, Matthews CR, Joshi V, de Jager M & Aspiras M (2011) Tobacco smoking affects bacterial acquisition and colonization in oral biofilms. *Infect Immun* 79: 4730–4738

Kumarasamy V, Roslani AC, Rani KU & Kumar Govind S (2014) Advantage of using colonic washouts for Blastocystis detection in colorectal cancer patients. *Parasit Vectors* 7: 162

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C & Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5: R12

Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357–359

Laniado-Laborín R (2009) Smoking and chronic obstructive pulmonary disease (COPD). Parallel epidemics of the 21 century. *Int J Environ Res Public Health* 6: 209–224

Law MR, Morris JK & Wald NJ (1997) Environmental tobacco smoke exposure and ischaemic heart disease: an evaluation of the evidence. *BMJ* 315: 973–980

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, *et al* (2013) Richness of human gut microbiome correlates with metabolic markers. *Nature* 500: 541–546

Lee AJ, Fowkes FG, Carson MN, Leng GC & Allan PL (1997) Smoking, atherosclerosis and risk of abdominal aortic aneurysm. *Eur Heart J* 18: 671–676

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R & 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079

Liu W, Zhang J, Wu C, Cai S, Huang W, Chen J, Xi X, Liang Z, Hou Q, Zhou B, *et al* (2016) Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci Rep* 6: 34826

Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, Weir JC, Quince C, Smith GP, Betley JR, *et al* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *JAMA* 309: 1502–1510

Lortet-Tieulent J, Goding Sauer A, Siegel RL, Miller KD, Islami F, Fedewa SA, Jacobs EJ & Jemal A (2016) State-Level Cancer Mortality Attributable to Cigarette Smoking in the United States. *JAMA Intern Med* 176: 1792–1798

Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550

Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71: 8228–8235

Lukeš J, Stensvold CR, Jirků-Pomajbíková K & Wegener Parfrey L (2015) Are Human Intestinal Eukaryotes Beneficial or Commensals? *PLoS Pathog* 11: e1005039

Macgregor ID (1989) Effects of smoking on oral ecology. A review of the literature. *Clin Prev Dent* 11: 3–7

Mager DL, Haffajee AD & Socransky SS (2003) Effects of periodontitis and smoking on the microbiota of oral mucous membranes and saliva in systemically healthy subjects. *J Clin Periodontol* 30: 1031–1037

Manara S, Asnicar F, Beghini F, Bazzani D, Cumbo F, Zolfo M, Nigro E, Karcher N, Manghi P, Metzger MI, *et al* (2019) Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biol* 20: 299

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, *et al* (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 40: D115-22

Mason MR, Preshaw PM, Nagaraja HN, Dabdoub SM, Rahman A & Kumar PS (2015) The subgingival microbiome of clinically healthy current and never smokers. *ISME J* 9: 268–272

McElroy KE, Luciani F & Thomas T (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* 13: 74

McMurdie PJ & Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8: e61217

Molnar M & Ilie L (2015) Correcting Illumina data. *Brief Bioinform* 16: 588–599

Morgan XC, Segata N & Huttenhower C (2013) Biodiversity and functional genomics in the human microbiome. *Trends Genet* 29: 51–58

Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, Flores SC, Fontenot AP, Ghedin E, Huang L, *et al* (2013) Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am J Respir Crit Care Med* 187: 1067–1075

Moura MA de S, Bergmann A, Aguiar SS de & Thuler LCS (2014) The magnitude of the association between smoking and the risk of developing cancer in Brazil: a multicenter study. *BMJ Open* 4: e003736

National Health and Nutrition Examination Survey (2017) *https://www.cdc.gov/nchs/nhanes/about_nhanes.htm*

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, *et al* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 32: 822–828

Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, *et al* (2015) Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat Commun* 6: 6505

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens HH, Wagner H, Oksanen MJ & Suggests M (2008) The vegan package. *Community ecology package* 10

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA & Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31: 3691–3693

Pan Y, Wang W, Wang KS, Moore K, Dunn E, Huang S & Feaster DJ (2015) Age Differences in the Trends of Smoking Among California Adults: Results from the California Health Interview Survey 2001-2012. *J Community Health* 40: 1091–1098

Parada AE, Needham DM & Fuhrman JA (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 18: 1403–1414

Pareek CS, Smoczynski R & Tretyn A (2011) Sequencing technologies and genome sequencing. *J Appl Genet* 52: 413–435

Parmigiani G, Garrett-Mayer ES, Anbazhagan R & Gabrielson E (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res* 10: 2922–2927

Parvinen T (1984) Stimulated salivary flow rate, pH and lactobacillus and yeast concentrations in non-smokers and smokers. *Scand J Dent Res* 92: 315–318

Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, *et al* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176: 649-662.e20

Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, *et al* (2017) Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 14: 1023–1024

Pasolli E, Truong DT, Malik F, Waldron L & Segata N (2016) Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput Biol* 12: e1004977

Perlman SE, Chernov C, Farley SM, Greene CM, Aldous KM, Freeman A, Rodriguez-Lopez J & Thorpe LE (2016) Exposure to Secondhand Smoke Among Nonsmokers in New York City in the Context of Recent Tobacco Control Policies: Current Status, Changes Over the Past Decade, and National Comparisons. *Nicotine Tob Res* 18: 2065–2074

Petersen AM, Stensvold CR, Mirsepasi H, Engberg J, Friis-Møller A, Porsbo LJ, Hammerum AM, Nordgaard-Lassen I, Nielsen HV & Krogfelt KA (2013) Active ulcerative colitis associated with low prevalence of Blastocystis and Dientamoeba fragilis infection. *Scand J Gastroenterol* 48: 638–639

Pierce JP, Fiore MC, Novotny TE, Hatziandreu EJ & Davis RM (1989) Trends in cigarette smoking in the United States. Educational differences are increasing. *JAMA* 261: 56–60

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J & Glockner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35: 7188–7196

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, *et al* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464: 59–65

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, *et al* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490: 55–60

Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, *et al* (2014) Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513: 59–64

Quinlan AR (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* 47: 11.12.1-34

Ramírez JD, Sánchez LV, Bautista DC, Corredor AF, Flórez AC & Stensvold CR (2014) Blastocystis subtypes detected in humans and animals from Colombia. *Infect Genet Evol* 22: 223–228

Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, Brigidi P, Crittenden AN, Henry AG & Candela M (2015) Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr Biol* 25: 1682–1693

Renson A, Jones HE, Beghini F, Segata N, Zolnik CP, Usyk M, Moody TU, Thorpe L, Burk R, Waldron LD, *et al* (2017) Sociodemographic variation in the oral microbiome. *bioRxiv*: 189225

Reznick AZ, Klein I, Eiserich JP, Cross CE & Nagler RM (2003) Inhibition of oral peroxidase activity by cigarette smoke: in vivo and in vitro studies. *Free Radic Biol Med* 34: 377–384

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W & Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43: e47

Roberts T, Ellis J, Harkness J, Marriott D & Stark D (2014) Treatment failure in patients with chronic Blastocystis infection. *J Med Microbiol* 63: 252–257

Robinson MD, McCarthy DJ & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140

Rocas IN, Siqueira JF Jr, Santos KR & Coelho AM (2001) “Red complex” (Bacteroides forsythus, Porphyromonas gingivalis, and Treponema denticola) in endodontic infections: a molecular approach. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 91: 468–471

Scanlan PD, Knight R, Song SJ, Ackermann G & Cotter PD (2016) Prevalence and genetic diversity of Blastocystis in family units living in the United States. *Infect Genet Evol* 45: 95–97

Scanlan PD & Marchesi JR (2008) Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *ISME J* 2: 1183–1193

Scanlan PD, Stensvold CR, Rajilić-Stojanović M, Heilig HGHJ, De Vos WM, O’Toole PW & Cotter PD (2014) The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiol Ecol* 90: 326–330

Schoeller DA (1995) Limitations in the assessment of dietary energy intake by self-report. *Metabolism* 44: 18–22

Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL & Segata N (2016) Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* 13: 435–438

Segata N (2015) Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. *Curr Biol* 25: R611–R613

Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS & Huttenhower C (2013a) Computational meta’omics for microbial community studies. *Mol Syst Biol* 9: 666

Segata N, Börnigen D, Morgan XC & Huttenhower C (2013b) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 4: 2304

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS & Huttenhower C (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* 12: R60

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O & Huttenhower C (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 9: 811–814

Socransky SS (1977) Microbiology of periodontal disease -- present status and future considerations. *J Periodontol* 48: 497–504

Stallones RA (2015) The association between tobacco smoking and coronary heart disease. *Int J Epidemiol* 44: 735–743

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313

Stämpfli MR & Anderson GP (2009) How cigarette smoke skews immune responses to promote infection, lung disease and cancer. *Nat Rev Immunol* 9: 377–384

Stensvold CR, Alfellani M & Clark CG (2012) Levels of genetic diversity vary dramatically between Blastocystis subtypes. *Infect Genet Evol* 12: 263–273

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550

Tan KSW (2008) New insights on classification, identification, and clinical relevance of Blastocystis spp. *Clin Microbiol Rev* 21: 639–665

Tan KSW, Mirza H, Teo JDW, Wu B & Macary PA (2010) Current Views on the Clinical Relevance of Blastocystis spp. *Curr Infect Dis Rep* 12: 28–35

Thorpe LE, Greene C, Freeman A, Snell E, Rodriguez-Lopez JS, Frankel M, Punsalang A Jr, Chernov C, Lurie E, Friedman M, *et al* (2015) Rationale, design and respondent characteristics of the 2013-2014 New York City Health and Nutrition Examination Survey (NYC HANES 2013-2014). *Prev Med Rep* 2: 580–585

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C & Segata N (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 12: 902–903

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

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS & Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428: 37–43

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66–74

Villalobos G, Orozco-Mosqueda GE, Lopez-Perez M, Lopez-Escamilla E, Córdoba-Aguilar A, Rangel-Gamboa L, Olivo-Diaz A, Romero-Valdovinos M, Maravilla P & Martinez-

Hernandez F (2014) Suitability of internal transcribed spacers (ITS) as markers for the population genetic structure of Blastocystis spp. *Parasit Vectors* 7: 461

Wawrzyniak I, Texier C, Poirier P, Viscogliosi E, Tan KSW, Delbac F & El Alaoui H (2012) Characterization of two cysteine proteases secreted by Blastocystis ST7, a human intestinal parasite. *Parasitol Int* 61: 437–442

Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R & Licht TR (2014) Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* 2: 19

Wu J, Peters BA, Dominianni C, Zhang Y, Pei Z, Yang L, Ma Y, Purdue MP, Jacobs EJ, Gapstur SM, *et al* (2016) Cigarette smoking and the oral microbiome in a large study of American adults. *ISME J* 10: 2435–2446

Yoshikawa H, Dogruman-Al F, Turk S, Kustimur S, Balaban N & Sultan N (2011) Evaluation of DNA extraction kits for molecular diagnosis of human Blastocystis subtypes from fecal samples. *Parasitol Res* 109: 1045–1050

Yoshikawa H, Koyama Y, Tsuchiya E & Takami K (2016) Blastocystis phylogeny among various isolates from humans to insects. *Parasitol Int* 65: 750–759

Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, *et al* (2014) Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 10: 766

Zeng X-T, Tu M-L, Liu D-Y, Zheng D, Zhang J & Leng W (2012) Periodontal disease and risk of chronic obstructive pulmonary disease: a meta-analysis of observational studies. *PLoS One* 7: e46508

Zheng W, Suzuki K, Tanaka T, Kohama M, Yamagata Z & Okinawa Child Health Study Group (2016) Association between Maternal Smoking during Pregnancy and Low Birthweight: Effects by Maternal Age. *PLoS One* 11: e0146241