UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

ICT International Doctoral School

A PHYLOGENETIC FRAMEWORK FOR LARGE-SCALE ANALYSIS OF MICROBIAL COMMUNITIES

Francesco Asnicar

Advisor

Prof. Enrico Blanzieri

Università degli Studi di Trento

Co-Advisor

Prof. Nicola Segata

Università degli Studi di Trento

Contents

Table of Contents

Doctoral program in Information and Communication Technology

Doctoral candidate

Francesco Asnicar

Cycle	**30**
Thesis	**A phylogenetic framework for large-scale analysis of microbial communities**
Advisor	**Enrico Blanzieri (University of Trento / DISI)**
Co-advisor	**Nicola Segata (University of Trento / Department CIBIO)**

1. **List of publications**

Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle
Pasolli E, **Asnicar F***, Manara S*, Zolfo M*, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N (* equal contribution)
Cell (2019)

Genomic and metagenomic insights into the microbial community of a thermal spring
Pedron R*, Esposito A*, Bianconi A, Pasolli E, Tett A, **Asnicar F**, Cristofolini M, Segata N, Jousson O (* equal contribution)
Microbiome (2019)

Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of Staphylococcus aureus strains in a paediatric hospital
Manara S*, Pasolli E*, Dolce D, Ravenni N, Campana S, Armanini F, **Asnicar F**, Mengoni A, Galli L, Montagnani C, Venturini E, Rota-Stabelli O, Grandi G, Taccetti G, Segata N (* equal contribution)
Genome medicine (2018)

Discovering causal relationships in grapevine expression data to expand gene networks. A case study: four networks related to climate change
Malacarne G*, Pilati S*, Valentini S*, **Asnicar F**, Moretto M, Sonego P, Masera L, Cavecchia V, Blanzieri E, Moser C (* equal contribution)
Frontiers in Plant Science (2018)

Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome
Ferretti P, Pasolli E*, Tett A*, **Asnicar F***, Gorfer V, Fedi S, Armanini F, Truong DT, Manara S, Zolfo M, Beghini F, Bertorelli R, De Sanctis V, Bariletti I, Canto R, Clementi R, Cologna M, Crifò T, Cusumano G, Gottardi S, Innamorati C, Masè C, Postai D, Savoi D, Duranti S, Lugli GA, Mancabelli L, Turroni F, Ferrario C, Milani C, Mangifesta M, Anzalone R, Viappiani A, Yassour M, Vlamakis H, Xavier R, Collado CM, Koren O, Tateo S, Soffiati M, Pedrotti A, Ventura M, Huttenhower C, Bork P, Segata N (* equal contribution)
Cell Host & Microbe (2018)

NES^{2}RA: Network expansion by stratified variable subsetting and ranking aggregation
Asnicar F, Masera L, Coller E, Gallo C, Sella N, Tolio T, Morettin P, Erculiani L, Galante F, Semeniuta S, Malacarne G, Engelen K, Argentini A, Cavecchia V, Moser C, Blanzieri E
International Journal of High Performance Computing Applications (2018)

Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life
Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, Selvenius J, Oikarinen S, Hyöty H, Virtanen SM, Ilonen J, Ferretti P, Pasolli E, Tett A, **Asnicar F**, Segata N, Vlamakis H, Lander ES, Huttenhower C, Knip M, Xavier RJ
Cell Host & Microbe (2018)

Draft Genome Sequences of Novel *Pseudomonas*, *Flavobacterium*, and *Sediminibacterium* Species Strains from a Freshwater Ecosystem
Pinto F, Tett A, Armanini F, **Asnicar F**, Boscaini A, Pasolli E, Zolfo M, Donati C, Salmaso N, Segata N
Genome Announcements (2018)

Profiling microbial strains in urban environments using metagenomic sequencing data
Zolfo M, **Asnicar F**, Manghi P, Pasolli E, Tett A, Segata N
Biology direct (2018)

Draft Genome Sequence of the Planktic Cyanobacterium *Tychonema bourrellyi*, Isolated from Alpine Lentic Freshwater
Pinto F, Tett A, Armanini F, **Asnicar F**, Boscaini A, Pasolli E, Zolfo M, Donati C, Salmaso N, Segata N
Genome announcements (2017)

Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis
Tett A, Pasolli E, Farina S, Truong DT, **Asnicar F**, Zolfo M, Beghini F, Armanini F, Jousson O, De Sanctis V, Bertorelli R, Girolomoni G, Cristofolini M, Segata N
Nature Biofilms and Microbiomes (2017)

Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling
Asnicar F*, Manara S*, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, Segata N (* equal contribution)
mSystems (2017)

Uncovering oral Neisseria tropism and persistence using metagenomic sequencing
Donati C, Zolfo M, Albanese D, Truong DT, **Asnicar F**, Iebba V, Cavalieri D, Jousson O, De Filippo C, Huttenhower C, Segata N
Nature Microbiology (2016)

Strain-level microbial epidemiology and population genomics from shotgun metagenomics
Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, **Asnicar F**, Truong DT, Tett A, Ardythe L Morrow, Segata N
Nature Methods (2016)

Discovering Candidates for Gene Network Expansion by Distributed Volunteer Computing
Asnicar F, Erculiani L, Galante F, Gallo C, Masera L, Morettin P, Sella N, Semeniuta S, Tolio T, Malacarne G, Engelen K, Argentini A, Cavecchia V, Moser C, Blanzieri E
Trustcom/BigDataSE/ISPA, IEEE (2015)

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

TN-Grid and gene@home project: Volunteer Computing for Bioinformatics
Asnicar F, Sella N, Masera L, Morettin P, Tolio T, Semeniuta S, Moser C, Blanzieri E, Cavecchia V
BOINC:FAST 2015, CEUR-WS (2015)

– **Paper under review**

Draft genome sequence of the new species "*Candidatus* Cibiobacter qucibialis" metagenomically assembled from the human gut microbiome
Nigro E*, Mazzoni C*, Alvari G, Baldi G, Calia G, Cantore T, Ciciani M, Dalfovo D, Fabbri L, Flor S, Golzato D, Lattanzi C, Marangoni S, Marianini G, Minardi G, Piccinno R, Pirrotta S, Tebaldi M, Tonazzolli A, Vannuccini F, Manara S, Zolfo M, Karcher N, **Asnicar F**, Tett A^, Edoardo Pasolli E^, Segata N^ (* equal contribution, ^ co-senior authors)
Currently in revision at *Microbiology Resource Announcements*

Combined metagenomic analysis of colorectal cancer datasets defines cross-cohort microbial diagnostic signatures and a link with choline degradation
Thomas AM*, Manghi P*, **Asnicar F**, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, and Segata N (* equal contribution)
Currently in revision at *Nature Medicine*

QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science
Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, **Asnicar F**, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Titus Brown C, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B,

Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG
Currently in revision at *Nature Biotechnology*, *PeerJ Preprint* available

A reference phylogeny of 10,575 genomes redefines major clades of bacteria and archaea
Zhu Q*, Mai U*, Pfeiffer W, Janssen S, **Asnicar F**, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, Kosciolek T, Yin JB, Huang S, Salam N, Jiao JY, Wu Z, Xu ZZ, Sayyari E, Morton JT, Podell S, Knights D, Li WJ, Huttenhower C, Segata N, Smarr L, Mirarab S, and Knight R (* equal contribution)
In revision

OneGenE: Regulatory Gene Network Expansion via Distributed Volunteer Computing on BOINC
Asnicar F*, Masera L*, Pistore D, Valentini S, Cavecchia V, and Blanzieri E (* equal contribution)
Accepted paper at the *27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing* (2019)

2. **Research/study activities**

During my years as a Ph.D. candidate at the University of Trento I had the opportunity to study and contribute to two lines of research: gene networks and metagenomics, both related to the Bioinformatic field. Under the supervision of Prof. Enrico Blanzieri (Department of Engineering and Computer Science, DISI), I worked on the gene networks contributing the field with the proposal of novel computational approaches for dealing with the network expansion problem. Under the supervision of Prof. Nicola Segata (Department CIBIO), I mainly focused on the development of novel computational tools for microbiome analysis, with the opportunity to participate and contribute in other works of the Computational Metagenomics laboratory. During these years I also had the opportunity to participate to many international conferences where I presented my research work, and to collaborate with several laboratories and researchers working on Bioinformatics and computational metagenomics around the world. Some of these collaborations allowed me to spend a period abroad of 4 months, during the end of my third year as Ph.D. candidate. In March 2017, I first visited the laboratory of Prof. Curtis Huttenhower at Harvard T.H. Chan School of Public Health in Boston, and then from April till July 2017, I moved at the University of California San Diego, under the supervision of Prof. Siavash Mir Arabbaygi. During this period I deepened my knowledge about phylogenetic and phylogenomic analysis that was of paramount importance for the development of PhyloPhlAn version 2. Extra to the research

work on the Bioinformatic field, during my years as Ph.D. candidate I also had the opportunity to participate to several tutoring and teaching activities.

– **Awards**

Travel grant for the **BITS 2018** conference, Torino, Italy, June 27-29, 2018.

Doctor Darwin Prize for best contribution in Evolutionary Medicine for the "Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling" paper, issued by the Italian Society for Evolutionary Biology (SIBE) during the Evoluzione 2017 conference, Roma, August 28-31, 2017.

Summer school grant for the **Second European Summer School on Nutrigenomics**, September 5-9, 2016, University of Camerino, Italy.

Travel fellowships for the **ISMB/ECCB 2015**, July 10-14, 2015, Dublin.

– **Research and training periods spent abroad**

From April to July of 2017 I have been visiting the laboratory of **Prof. Siavash Mir Arabbaygi** at the **University of California, San Diego** in San Diego (CA).

In March 2017 I have been visiting the laboratory of **Prof. Curtis Huttenhower** at the **Harvard T.H. Chan School of Public Health** in Boston (MA).

– **Teaching activities**

June 25-29 2018 - **Lecture** "Shotgun metagenomics with strain-level resolution" for the summer school in "Integrated 'Omics' Technologies into Aquatic Ecology", held by Dr. Claudio Donati and Dr. Nico Salmaso, and organized by the Edmund Mach Foundation (San Michele all'Adige, Trento, Italy) and International Research School in Applied Ecology (IRSAE).

April 21st 2018 - **Lecture** "Introduzione sul microbiota" (in Italian) for "Master di II livello Idrologia Medica e Medicina Termale", held by Prof. Plinio Richelmi, Facoltà di Medicina e Chirurgia, University of Pavia, Italy.

June 15-16, 2017 - **Workshop** "Microbiome Analysis in the Cloud" organized by the University of Maryland, Baltimore, MD. During this workshop I presented the following software: GraPhlAn, StrainPhlAn, and PanPhlAn, with a practical application focused on the Human Microbiome Project (HMP) dataset.

A.Y. 2017/2018 - **Teacher assistant** for the course Bioinformatics (mod. 2), held by Dr. Toma Tebaldi, Master Degree in Quantitative and Computational Biology, University of Trento, Italy.

A.Y. 2017/2018 - **Teacher assistant** for the course Bioinformatics (mod. 1), held by Prof. Enrico Blanzieri, Master Degree in Quantitative and Computational Biology, University of Trento, Italy.

A.Y. 2016/2017 - **Teacher assistant** for the course Bioinformatics (mod. 1), held by Prof. Enrico Blanzieri, Master Degree in Quantitative and Computational Biology, University of Trento, Italy.

A.Y. 2015/2016 - **Teacher assistant** for the course Concurrency, held by Prof. Paola Quaglia, Master Degree in Computer Science, University of Trento, Italy.

A.Y. 2015/2016 - **Tutor** for the course Fondamenti di Informatica, held by Prof. Alessandro Moschitti, Bachelor Degree in Civil, Environmental, and Mechanical Engineering, University of Trento, Italy.

A.Y. 2015/2016 - **Tutor** for the course Informatica, held by Prof. Andrea Passerini, Bachelor Degree in Biomolecular Sciences and Technologies, University of Trento, Italy.

A.Y. 2015/2016 - **Tutor** for the course Biologia Computazionale, held by Prof. Nicola Segata, Bachelor Degree in Biomolecular Sciences and Technologies, University of Trento, Italy.

A.Y. 2014/2015 - **Teacher assistant** for the course Computability and Computational Complexity, held by Prof. Gabriel M. Kuper, Master Degree in Computer Science, University of Trento, Italy.

Abstract

The human microbiome represents the community of archaea, bacteria, micro-eukaryotes, and viruses present in and on the human body. Metagenomics is the most recent and advanced tool that allows the study of the microbiome at high resolution by sequencing the whole genetic content of a biological sample. The computational side of the metagenomic pipeline is recognized as the most challenging one as it needs to process large amounts of data coming from next-generation sequencing technologies to obtain accurate profiles of the microbiomes. Among all the analyses that can be performed, phylogenetics allows researchers to study microbial evolution, resolve strain-level relationships between microbes, and also taxonomically place and characterize novel and unknown microbial genomes. This thesis presents a novel computational phylogenetic approach implemented during my doctoral studies. The aims of the work range from the high-quality visualization of large phylogenies to the reconstruction of phylogenetic trees at unprecedented scale and resolution. Large-scale and accurate phylogeny reconstruction is crucial in tracking species at strain-level resolution across samples and phylogenetically characterizing unknown microbes by placing their genomes reconstructed via metagenomic assembly into a large reference phylogeny. The proposed computational phylogenetic framework has been used in several different metagenomic analyses, improving our understanding of the complexity of microbial communities. It proved, for example, to be crucial in the detection of vertical transmission events from mothers to infants and for the placement of thousands of unknown metagenome-reconstructed genomes leading to the definition of many new candidate species. This poses the basis for large-scale and more accurate analysis of the microbiome.

1. Introduction

In this first chapter, I will review the main biological field (section **1.1 The human microbiome**) and the two computational topics (in sections **1.2 A primer on Computational metagenomics** and **1.3 A primer on computational phylogenetics**) that are the basis of the work presented in this thesis. The aims of the thesis and the structure of the following chapters are reported in the last section, **1.4 Aims and main contributions of the thesis**.

1.1 The human microbiome

The microbiome is the totality of bacteria, archaea, viruses, and micro-eukaryotes that inhabit a specific environment. Microbiomes have been studied in several different ecological niches, which range from the ocean (Sunagawa et al., 2015), to soil (Brown et al., 2015), permafrost (Mackelprang et al., 2011), alpine lakes (Monchamp et al., 2016), but also urbanized environments like hospitals (Shin et al., 2015) and metro (Mason et al., 2016). Microbiome studies looked also at the microbial communities present in the food chain, with the aim of improving food quality and safety (De Filippis et al., 2018). Even microbiomes associated with living hosts, including plants, animals like cattle (Wallace et al., 2015), poultry (Yeoman et al., 2012), and mice (Xiao et al., 2015) have been studied. Particular focus has been reserved for the microbiomes associated with the human body; as humans, we are carrying several different and at the same time specific microbiomes in and on our bodies. Two of the largest microbiome initiatives - the Human Microbiome Project in the US (HMP et al., 2012) and MetaHIT in Europe (Qin et al., 2010) - paved the way toward understanding the diversity of the human microbiome, by producing a large set of publicly available human microbiome data from many different body sites, like nasal passages, oral cavity, skin, gastrointestinal tract, and the urogenital tract.

One of the main interests about the human microbiome was the characterization of the microbial composition in different body locations, such as the oral cavity (Donati et al., 2016; HMP et al., 2012), the airways (HMP et al., 2012), the skin (Grice and Segre, 2011; HMP et al., 2012; Tett et al., 2017), and the urogenital (Aagaard et al., 2012; HMP et al., 2012) and gastrointestinal tracts (HMP et al., 2012; Qin et al., 2010). The latter is arguably one of the most studied human microbiomes. The human gut microbiome has been extensively studied in the last decade and, in particular, it has been shown to be associated with geography, age, diet, and health and disease.

It is particularly difficult to define what a healthy microbiome is, given the high variability of the microbial composition we can observe even in a cohort of only healthy individuals. However, we know that if we want to modify the microbial composition of an individual, the diet can play a pivotal role (Carmody et al., 2015; David et al., 2014). To give an example, it has been shown that a low-gluten diet shows enrichment of Bacteroidaceae family, compared to a high-gluten diet that shows an increase of Lachnospiraceae family (Hansen et al., 2018). Another aspect extensively studied and related to the diet is the effect of probiotics intake on the oral and gut microbiome (Suez et al., 2018; Zmora et al., 2018). Given this interesting link between microbiome and diet, to date, there is an increasing interest in personalized nutrition based on the microbiome (Zeevi et al., 2015), not only by the scientific community but also from private companies. Microbiome studies so far were geographically limited to a number of highly-sampled countries, especially the US, Europe, and China (HMP et al., 2012; Qin et al., 2010, 2014), which are representatives of limited

dietary habits. Recent works have focused on the study of rural cohorts (Brito et al., 2016; Obregon-Tito et al., 2015; Rampelli et al., 2015), with the goal of characterizing the microbiome of non-Westernized communities that were not subjected to urbanization and have more traditional lifestyles and diets. The survey of non-Westernized microbiomes has shown an increase in microbial richness when compared to microbiomes of people living in urbanized areas (De Filippo et al., 2010; Yatsunenko et al., 2012) and is leading to the discovery of novel and previously not-characterized microbial species (Pasolli et al., 2019).

Other than diet and geography, another recently studied microbiome association is with age, and in particular the debate on the onset of the infant gut microbiome (Dominguez-Bello et al., 2010; La Rosa et al., 2014; Palmer et al., 2007). There are studies that support the existence of the placenta microbiome (Aagaard et al., 2014), and others instead that are supporting the concept of a sterile placenta (Leiby et al., 2018; Perez-Muñoz et al., 2017). Some other studies focused instead on the perinatal period, trying to understand if there are differences in the microbiome development of vaginally and Cesarean-section delivered babies, and whether these can have an impact on their future health (Azad et al., 2013; Dominguez-Bello et al., 2010). Other groups have focused on identifying microbial species vertically transmitted from mothers to infants by longitudinal microbiome sampling of mothers and infants (Asnicar et al., 2017; Duranti et al., 2017; Ferretti et al., 2018; Jost et al., 2014; Milani et al., 2015; Yassour et al., 2018). Understanding how the infant gut colonization starts and how it leads to the development of a more complex adult-like microbial composition is crucial also in relation to the future health of the babies. Looking not only at the taxonomic composition but also at the transcriptional patterns as in (Asnicar et al., 2017) allows to study also the activity of vertically transmitted strains in the infant's gut. The identification of a panel of species that are vertically transmitted from mothers to infants opens new venues in understanding how the microbiome is spread and maintained within a population.

In recent years, the human microbiome has been convincingly associated with a number of diseases in humans. Specific biomedical efforts then explicitly focused on using the microbiome as a therapeutic target. The set of diseases that have been investigated in connection with the microbiome is now relatively large and includes: irritable bowel syndrome (Durbán et al., 2013; Saulnier et al., 2011) and inflammatory bowel disease (Franzosa et al., 2019; Greenblum et al., 2012; Morgan et al., 2012; Nielsen et al., 2014), type 1 (Heintz-Buschart et al., 2016) and type 2 diabetes (Forslund et al., 2015; Karlsson et al., 2013; Qin et al., 2012), Crohn's disease (Erickson et al., 2012; Gevers et al., 2014; Lewis et al., 2015; Quince et al., 2015), colorectal cancer (Feng et al., 2015; Vogtmann et al., 2016; Yu et al., 2017; Zeller et al., 2014), rheumatoid arthritis (Scher et al., 2013; Zhang et al., 2015), and necrotizing enterocolitis (Claud et al., 2013; Ward et al., 2016), just to name a few. This field of research is very promising because the microbiome can be used as a non-invasive diagnostic marker (Yang et al., 2012; Yu et al., 2017).

An increasing amount of work is trying to elucidate the fundamental properties of microbial communities and characterize the role of the microbiome in the conditions and settings described above. Advances in the field are driven by improvements in sequencing technologies, development of accurate analysis techniques, adoption of appropriate study design and sample sizes, and availability of publicly available annotated data that can be reused in large meta-analyses. Still, a lot has to be uncovered. Among the aspects that are

receiving more attention, there is the description of the microbiome at the level of it single constitutive strains (rather than species) (Quince et al., 2017; Segata, 2018) and the identification of species without any available isolate data (Pasolli et al., 2019).

Within the human microbiome research, I will now introduce the two fields of computational metagenomics and computational phylogenetics, as they will be the basis of this thesis work.

1.2 A primer on Computational Metagenomics

Metagenomics is the study of microbial communities associated with a given environment by sequencing the whole genetic content of a sample and without the need for cultivation. This allows the study not only of well-characterized microbes but also of species that are recalcitrant to isolation. High-throughput sequencing technologies were and are key to the development of the field, with increasing sequencing depths enabling the study of a growing fraction of microbiome members. A shotgun metagenomic study includes as first steps sample collection, storage, and sequencing. For these first steps, we can rely on well-established protocols that have been used for years. The analysis of the sequenced data, instead, is still posing challenges as we aim at higher resolution analyses of the microbiome and have to deal with the increasing amount of data. Computational analyses of metagenomics data are today essential for elucidating and characterizing the microbiome members and their interactions. We can further split the computational approaches into two categories, as suggested in (Quince et al., 2017): the sequence analysis that deals with the first taxonomic and functional characterization, and post-processing analysis that aims at applying statistical and machine learning approaches to interpret the results and link them back to biological aspects.

While performing sequence analysis, there are two main questions that we can ask with respect to a microbiome sample: “who is there?” and “what can they (potentially) do?”. To answer the first question “who is there?”, we need to uncover the microbial diversity by figuring out which species are present and at what relative abundance. In the literature, there are several software tools that deal with this “taxonomic profiling” challenge. Some of them utilize a reference-based approach, while others use a k-mer based approach. Reference-based methods determine microbial species by exploiting publicly available genomic databases in different ways. Genomes deposited in publicly available databases can be either used to identify species by directly mapping the metagenomic reads against the whole genomes or to pre-compute species-specific markers, which are smaller than the whole genome making the mapping step faster. A whole-genome reference-based approach like SLIMM (Dadi et al., 2017) that produces taxonomic profiles based on reads mapping classification, can be more sensitive to low abundant species but it is likely not computationally feasible given the always increasing number of genomes deposited in public databases. Marker-based approaches instead select a maximally-informative fraction of the available genomic information, hence result in a much faster computation. Marker-based approaches are generally very accurate with a low false-positive rate (Freitas et al., 2015; Sczyrba et al., 2017; Truong et al., 2015), with the downside that it is not possible to profile species for which markers cannot be extracted - which in most cases means that their genomic data is not available - or for low-abundant species for which there is not enough reads data when mapping their markers. K-mer based approaches instead exploit statistics (like frequency) computed on all substrings of a certain length k (Lu et al., 2017; Popic et al., 2018; Wood and Salzberg, 2014). The upside in using a k-mer based taxonomic profile is

that it can profile very-low abundant species. A downside of the same approach is that to increase sensitivity we should consider large values for k, but this will result in a memory-intensive computation. Other than reference and k-mer based methods, another approach for taxonomic profiling exploits instead a reference phylogenetic tree. Taxonomic profilers that exploit a reference phylogeny, place the metagenomic reads into the reference phylogeny and infer the taxonomic profile according to their placement into the tree (Nguyen et al., 2014). To answer the "who is there?" question, many taxonomic profilers have been proposed and are available in the literature, with many of them exploiting complementary approaches. As there is no a clear winner, the choice today seems to tend to marker-based profilers as they are able to accurately profile in a quick way the increasing amount of data that we are able to generate with current high-throughput sequencing technologies.

The second question - "what can they potentially do?" - focuses on the ability to retrieve the overall gene repertoire and thus the functional potential of a microbiome directly from metagenomics data. This is generally referred to as functional potential analysis when done on metagenomics data, whereas in the presence of paired metagenomes and metatranscriptomes (sequencing of the RNA content of a community) it can be referred to as functional expression analysis. There are a number of tools that are dealing with the functional analysis from metagenomics and metatranscriptomics data like HUMAnN1 (Abubucker et al., 2012), HUMAnN2 (Franzosa et al., 2018), COGNIZER (Bose et al., 2015), MEGAN (Huson et al., 2016), and ShotMAP (Nayfach et al., 2015).

Other than on the two classical questions discussed above, computational metagenomics has recently focused its efforts on the accurate genome reconstruction from metagenomics data, to study and characterize previously unseen microbial diversity. Accurate genome reconstruction from metagenomic data starts with metagenomic assembly, whose goal is to reconstruct longer consensus DNA stretches, named contigs, by assembling together short reads. For example, metaSPAdes (Nurk et al., 2017) and MEGAHIT (Li et al., 2015) are two metagenomic assemblers that are routinely used for contigs reconstruction. The computed contigs from the assembly analysis can be further organized into genome bins by using contig binners like METABAT2 (Kang et al., 2015). A bin of contigs represents the sequences of a specific microbial genome present in the microbiome. This allows to retrieve bacterial genomes of uncultivable species and new genetic variants (strains) of known species and thus shedding new light on microbial diversity.

All the computational metagenomics tools presented so far directly analyze the raw shotgun metagenomic sample and aim at extracting higher-level profiles like its taxonomic composition, its functional potential profile with its set of genes, or the possibility to recover *quasi*-complete microbial genomes. Such microbiome profiles are typically further processed by the application of statistical and machine learning approaches that allow interpreting the results, even from a biological point of view. Post-processing is very wide and is not a methodological focus of the present thesis, so we refer elsewhere (Quince et al., 2017) for an introduction on this part.

1.3 A primer on Computational Phylogenetics

Phylogenetics is the study of the relationships between organisms that allows us to understand the evolutionary patterns of organisms. These relationships are generally inferred starting from genomic data, either genes or proteins, and are derived according to

evolutionary models. With the increased availability of large sets of microbial genomes from isolate sequencing, the last decade witnessed a dramatic increase of phylogenetic analyses applied to microbial organisms. Phylogenetic analysis was indeed extensively used in different contexts (Aanensen et al., 2016; Anzai et al., 1997; Holt et al., 2015; Khan et al., 2008; Manara et al., 2018; Musser and Kapur, 1992), considering different microbial species studied in isolation like *Pseudomonas aeruginosa*, *Escherichia coli*, and *Klebsiella pneumoniae*. Most of these results were obtained thank to computational advances that allowed scaling phylogenetic profiling to many genes or genomic regions. However, phylogenetic methods are still struggling to cope with the increased availability of microbial genomes, need advanced computational skills to be correctly used, and have been only scarcely applied so far in shotgun metagenomics. My thesis aims at filling these gaps.

Similarly to isolate sequencing datasets, phylogenetics can provide many crucial insights in metagenomics. For example, it can help understanding how new genomes reconstructed via metagenomic assembly are related with the known and characterized microbial reference genomes deposited in publicly available databases. Also, taxonomic assignments of genomes from metagenomes can be based on their phylogenetic placement. Moreover, phylogenetic differences in strains within the same microbial species in different microbiomes can pinpoint important bio-geographical associations of phenotypic traits (e.g. resistance to antibiotics) in large microbial communities. Through phylogenetic analysis, we can also study and characterize the within diversity of a microbial organism and hence how its strains evolved in time.

The basic concept in phylogenetics is a tree, which is the main output of a phylogenetic analysis. A tree structure is used to organize and represent the inferred relationships between organisms, where the leaves of the tree represent the genomes and the internal nodes are the branching point that represents a split in the evolution of the organism, where new lineages start diverging. The sum of the length of the branches between two leaf nodes in a phylogenetic tree represents their phylogenetic distance.

Phylogenetics can be subdivided into two main approaches: phenetic and cladistics. A phenetic approach computes a tree (called dendrogram) based on a concept of distance. On the other hand, a cladistic approach computes a tree (named cladogram) considering many different evolutionary pathways, and then choosing one based either on parsimony or a likelihood strategy. A classical phylogenomic analysis focuses on the whole genome of very few species to reconstruct their phylogenetic tree; with phylogenetics, instead, we are exploiting a single gene for inferring the phylogenetic signal. The latter approach is often applied in the bacterial world, where the ubiquitously conserved 16S rRNA gene (or one of its nine variable regions) is used for building even very large phylogenies (DeSantis et al., 2006; Parks et al., 2018; Quast et al., 2013). However, thanks to the increasing amount of publicly available data, the improvement of software analysis tools, and the novel genomes reconstructed from metagenomes (Brown et al., 2015; Hildebrand et al., 2019; Parks et al., 2017; Pasolli et al., 2019), we can now deepen the study of previously unexplored microbiome members. To be able to discriminate this uninvestigated microbial diversity, though, we cannot use a single gene, but we need to exploit larger sets of genes to maintain a high resolution. This requires to deal with a very large number of genomes and with *quasi*-complete genome sequences, or a large number of genes. Computational phylogenetics deals with this increasing amount of data and the need to use larger sets of genes for

accurate phylogeny reconstruction thanks to computational advances and the ability to reduce and speed up the multiple-sequence alignment step. Indeed, computational phylogenetics is now fundamental to unravel the within-species diversity and for the characterization of unknown species.

Common steps in computational phylogenetics are the identification of the genes from the inputs, the multiple-sequence alignment of the extracted sequences, and then the inference of the phylogeny. The identification of the genes or proteins from the genomes of interest is done by searching or aligning, using BLAST (Altschul et al., 1990), USEARCH (Edgar, 2010), or Diamond (Buchfink et al., 2015), for instance, the reference gene sequences against the inputs. The identified sequences of each gene extracted from the genomes are then all aligned together using a multiple-sequence aligner like MUSCLE (Edgar, 2004), PASTA (Mirarab et al., 2015), MAFFT (Katoh and Standley, 2013), T-coffee (Notredame et al., 2000), OPAL (Wheeler and Kececioglu, 2007), SATé-II (Liu et al., 2012), CLUSTAL W (Thompson et al., 1994) to produce what is called a multiple-sequence alignment (or MSA). After having computed the MSAs of the extracted sequence of each marker gene, there are two schools of thought in computational phylogenetics: the first one is the so-called supermatrix (or concatenation) approach and the second one is the supertree (or gene trees) approach. A supermatrix approach (or concatenation) concatenates one after the other all the MSAs marker gene sequences of each input genome into one long MSA sequence. The concatenated MSA is then provided as input for the phylogenetic reconstruction step, using software tools like RAxML (Stamatakis, 2014), IQ-TREE (Nguyen et al., 2015a), or FastTree (Price et al., 2009, 2010). A supertree (or gene trees) approach instead independently reconstructs a phylogenetic tree for each of the multiple-sequence aligned marker genes, using the same software tools as just described in the concatenation approach, and then exploits summary approaches like ASTRAL (Mirarab et al., 2014), ASTRAL-II (Mirarab and Warnow, 2015), ASTRAL-III (Zhang et al., 2018), or ASTRID (Vachaspati and Warnow, 2015) to infer a consensus phylogenetic tree that reflects the phylogenies of each single marker gene computed.

Challenges in computational phylogenetics today are many and are mainly related to the difficulty of accurately reconstructing very large phylogenies. When dealing with tens of thousands of input genomes, the required computational time for the phylogeny inference step can increase exponentially. This is largely due to the fact that there are tens of thousands of genomes and for each of them an MSA up to hundreds of thousands of positions. Removing genomes can be one solution that can help in building very large phylogenies, with the drawback of losing the removed genomes from the inferred phylogeny. The other way to tackle this problem is by reducing in length the MSA. There are several approaches that aim at reducing in length the MSA trying to reduce at minimum the loss of the phylogenetic signal, to make possible the reconstruction of very large phylogenies (Capella-Gutiérrez et al., 2009; Castresana, 2000; Chang et al., 2014; Dress et al., 2008; Edgar, 2009; Penn et al., 2010; Sela et al., 2015; Talavera and Castresana, 2007; Valdar, 2002; Webb et al., 2017). More details about these approaches are reported and discussed in **Chapter 3**.

Giving the increasing availability of data, we can now analyze at the same time hundreds of thousands of genomes. This is pushing the community to ask for a phylogenetic framework

able to analyze this amount of data in a reasonable time while keeping the highest possible accuracy.

1.4 Aims and main contributions of the thesis

The works presented in this thesis aim to tackle the challenges arising in both the computational metagenomics and computational phylogenetics fields introduced above. The main contribution of this thesis to the computational metagenomics and computational phylogenetics fields are described in **Chapter 2**, **3** and **4**, and their respective aims are:

- the visualization of large-scale hierarchical and phylogenetic trees;
- the development of a novel phylogenetic framework for the analysis of microbiome data from metagenomics;
- the use of phylogenetics in metagenomics to solve the "vertical transmission" problem.

Several applications of the developed tools are then presented and motivated in **Chapter 5**. Other than the contributions presented in this thesis, I report below other computational contributions I made into several software tools I developed and/or maintained:

- I developed GraPhlAn[1], presented in **Chapter 2**;
- I developed export2graphlan[2], a framework that integrates results coming from other analysis tools and provides GraPhlAn-like input files, to ease and automatize the use of GraPhlAn;
- I maintained and improved the first version of PhyloPhlAn[3] and I implemented PhyloPhlAn 2 presented in **Chapter 3**;
- I maintained and added features to MetaPhlAn2[4], a marker-based taxonomic profiler for shotgun metagenomics data, and I developed the q2-metaphlan2 QIIME 2 plugin for MetaPhlAn2;
- I maintained hclust2[5] that is used for the visualization of taxonomic and functional profiles as heatmaps with clustering possibilities;
- I developed the metagenomics pre-processing pipeline[6] used for the quality-control screening of the raw sequence data and to generate the "cleaned" metagenomes to be used for downstream analysis;
- I wrote the recipes that allow the packaging of the tools listed above to be integrated into Conda[7] and Bioconda[8];

---

[1] GraPhlAn repository: https://bitbucket.org/nsegata/graphlan
[2] export2graphlan repository: https://bitbucket.org/CibioCM/export2graphlan
[3] PhyloPhlAn repository: https://bitbucket.org/nsegata/phylophlan, PhyloPhlAn 2 is available in the "dev" branch of the PhyloPhlAn repository
[4] MetaPhlAn2 repository: https://bitbucket.org/biobakery/metaphlan2, and q2-metaphlan2 repository: https://bitbucket.org/biobakery/metaphlan2-install
[5] hclust2 repository: https://bitbucket.org/nsegata/hclust2
[6] Preprocessing repository: https://bitbucket.org/CibioCM/preprocessing
[7] My Anaconda personal page containing packages and environments: https://anaconda.org/fasnicar
[8] Packages available on Bioconda:
GraPhlAn https://bioconda.github.io/recipes/graphlan/README.html,
export2graphlan https://bioconda.github.io/recipes/export2graphlan/README.html,
MetaPhlAn2 https://bioconda.github.io/recipes/metaphlan2/README.html, and
hclust2 https://bioconda.github.io/recipes/hclust2/README.html

- I implemented the PC++ algorithm[9] that is the core algorithm of NESRA, NES^{2}RA, and the OneGenE algorithms;
- I developed and implemented the gene@home project, NESRA, NES^{2}RA, and OneGenE algorithms, and the post-processing pipeline[10], all presented in **Chapter 6**.

Most of the following chapters are based on published articles that are fully reported in **Chapter 2**, **4**, and **6**, while only the main parts are reported in **Chapter 5**.

More specifically, **Chapter 2** will present GraPhlAn (Asnicar et al., 2015a), a Python package software specifically developed for the high-quality visualization of large-scale hierarchical and phylogenetic trees, with the possibility of visualizing associated metadata in and on the tree for explorative analysis.

Then, in **Chapter 3** I will discuss PhyloPhlAn 2, the new and improved implementation of PhyloPhlAn that allows to accurately phylogenetically analyze microbes derived from microbial communities. In this chapter, I'll present the details of the implementation showing several examples that will support and demonstrate the PhyloPhlAn 2 capabilities.

Chapter 4 will present a recent work focused on the study of microbial species vertically transmitted from the mother to the infant (Asnicar et al., 2017). Even though the cohort used in this work is small (five mother-infant couples), we were able to perform an extensive panel of analyses, ranging from the taxonomic profiling to several strain-level analysis, concluding with the study of functional potential and expression, thanks to the availability of metatranscriptomic data for two of the five couples.

Then, **Chapter 5** expands the set of analyses where PhyloPhlAn 2 played a pivotal role in the reconstruction of large phylogenies at multiple distinct diversity levels, from strain-level to tree-of-life size phylogenies (Donati et al., 2016; Ferretti et al., 2018; Pasolli et al., 2019; Tett et al., 2017). This chapter contains portions of several published and recently submitted works that comprise a larger cohort for studying the vertical microbiome transmission from mothers to infants, and the phylogenetic characterization of unknown reconstructed genomes with an extreme genetic diversity.

In addition to the contributions to the computational metagenomics and phylogenetics fields, in **Chapter 6** I'll present the second line of research I carried on during my doctoral studies, as a continuation of a course project started during the last year of my M.Sc. in Computer Science with Prof. Enrico Blanzieri. This chapter is reporting the two initial works done on the gene network expansion problem (Asnicar et al., 2015b, 2015c) and reports at the end only the abstracts of two more recent works (Asnicar et al., 2016, 2019).

Finally, in the Conclusions (**Chapter 7**) I summarize the overall results of each chapter, presenting also future application and potential developments of PhyloPhlAn.

[9] PC++, BOINC (Anderson, 2004) implementation of the PC algorithm (Spirtes and Glymour, 1991): https://bitbucket.org/francesco-asnicar/pc-boinc

[10] Repository containing the scripts and code running in the gene@home BOINC project and for the post-processing pipeline: https://bitbucket.org/francesco-asnicar/gene_network_expansion

2. Compact graphical representation of phylogenetic data and metadata with GraPhlAn

This chapter introduces GraPhlAn, a novel framework I developed and implemented during the first part of my doctoral work for the integrated and compact visualization of phylogenetic tree structures and relevant quantitative non-phylogenetic data. Although several software packages are available for the visualization of phylogenies, none of them can incorporate metadata with a variety of visualization choices or can be produced programmatically. GraPhlAn indeed allows also to display in and on the tree several metadata of different types, ranging from heatmaps to bar plots to external markers to highlight specific features, aiding explorative analysis through data visualization. GraPhlAn is a visualization Python package that directly allows the generation of high-quality figures form metagenomic analysis tools like MetaPhlAn2, HUMAnN2, LEfSe, and a combination of them. This integration is made possible through the export2graphlan package I developed, which allows the additional integration with the BIOM (BIological Observation Matrix) file format (McDonald et al., 2012), generally used in the 16S/QIIME analysis framework. GraPhlAn is becoming a popular tool for the visualization of phylogenetic and microbiome data, with more than 150 citations since its publication. I am maintaining and regularly updating GraPhlAn and I am committed to further developing and supporting the software.

This chapter is reporting the following article:

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

Abstract

The increased availability of genomic and metagenomic data poses challenges at multiple analysis levels, including visualization of very large-scale microbial and microbial community data paired with rich metadata. We developed GraPhlAn (Graphical Phylogenetic Analysis), a computational tool that produces high-quality, compact visualizations of microbial genomes and metagenomes. This includes phylogenies spanning up to thousands of taxa, annotated with metadata ranging from microbial community abundances to microbial physiology or host and environmental phenotypes. GraPhlAn has been developed as an open-source command-driven tool in order to be easily integrated into complex, publication-quality bioinformatics pipelines. It can be executed either locally or through an online Galaxy web application. We present several examples including taxonomic and phylogenetic visualization of microbial communities, metabolic functions, and biomarker discovery that illustrate GraPhlAn's potential for modern microbial and community genomics.

2.1 Introduction

Modern high-throughput sequencing technologies provide comprehensive, large-scale datasets that have enabled a variety of novel genomic and metagenomic studies. A large number of statistical and computational tools have been developed specifically to tackle the complexity and high-dimensionality of such datasets and to provide robust and interpretable results. Visualizing data including thousands of microbial genomes or metagenomes,

however, remains a challenging task that is often crucial to driving exploratory data mining and to compactly summarizing quantitative conclusions.

In the specific context of microbial genomics and metagenomics, next-generation sequencing in particular produces datasets of unprecedented size, including thousands of newly sequenced microbial genomes per month and a tremendous increase in genetic diversity sampled by isolates or culture-free assays. Displaying phylogenies with thousands of microbial taxa in hundreds of samples is infeasible with most available tools. This is especially true when sequencing profiles need to be placed in the context of sample metadata (e.g. clinical information). Among recently developed tools, iTOL (Letunic and Bork, 2007, 2011) targets interactive analyses of large-scale phylogenies with a moderate amount of overlaid metadata, whereas ETE (Huerta-Cepas et al., 2010) is a Python programming toolkit focusing on tree exploration and visualization that is targeted for scientific programmers, and Krona (Ondov et al., 2011) emphasizes hierarchical quantitative information typically derived from metagenomic taxonomic profiles. Neither of these tools provides an automatable environment for non-computationally expert users in which very large phylogenies can be combined with high-dimensional metadata such as microbial community abundances, host or environmental phenotypes, or microbial physiological properties.

In particular, a successful high-throughput genomic visualization environment for modern microbial informatics must satisfy two criteria. First, software releases must be free and open-source to allow other researchers to verify and to adapt the software to their specific needs and to cope with the quick evolution of data types and datasets size. Second, visualization tools must be command-driven in order to be embedded in computational pipelines. This allows for a higher degree of analysis reproducibility, but the software must correspondingly be available for local installation and callable through a convenient interface (e.g. API or general scripting language). Local installations have also the advantage of avoiding the transfer of large or sensitive data to remote servers, preventing potential issues with the confidentiality of unpublished biological data. Neither of these criteria, of course, prevent tools from also being embeddable in web-based interfaces in order to facilitate use by users with limited computational expertise (Blankenberg et al., 2010; Giardine et al., 2005; Goecks et al., 2010; Oinn et al., 2004) and all such tools must regardless produce informative, clear, detailed, and publication-ready visualizations.

2.2 Materials & Methods

GraPhlAn is a new tool for compact and publication-quality representation of circular taxonomic and phylogenetic trees with potentially rich sets of associated metadata. It was developed primarily for microbial genomic and microbiome-related studies in which the complex phylogenetic/taxonomic structure of microbial communities needs to be complemented with quantitative and qualitative sample-associated metadata. GraPhlAn is available at http://cibiocm.bitbucket.org/tools/graphlan.html.

2.2.1 Implementation strategy

GraPhlAn is composed of two Python modules: one for drawing the image and one for adding annotations to the tree. GraPhlAn exploits the annotation file to highlight and personalize the appearance of the tree and of the associated information. The annotation file

does not perform any modifications to the structure of the tree, but it just changes the way in which nodes and branches are displayed. Internally, GraPhlAn uses the matplotlib library (Hunter, 2007) to perform the drawing functions.

2.2.2 The export2graphlan module

Export2graphlan is a framework to easily integrate GraPhlAn into already existing bioinformatics pipelines. Export2graphlan makes use of two external libraries: the pandas python library (McKinney, 2012) and the BIOM library, only when BIOM files are given as input.

Export2graphlan can take as input two files: the result of the analysis of MetaPhlAn (either version 1 or 2) or HUMAnN, and the result of the analysis of LEfSe. At least one of these two input files is mandatory. Export2graphlan will then produce a tree file and an annotation file that can be used with GraPhlAn. In addition, export2graphlan can take as input a BIOM file (either version 1 or 2).

Export2graphlan performs an analysis on the abundance values and, if present, on the LDA score assigned by LEfSe, to annotate and highlight the most abundant clades and the ones found to be biomarkers. Through a number of parameters the user can control the annotations produced by export2graphlan.

2.3 Results and Discussion

2.3.1 Plotting taxonomic trees with clade annotations

The simplest structures visualizable by GraPhlAn include taxonomic trees (i.e. those without variable branch lengths) with simple clade or taxon nomenclature labels. These can be combined with quantitative information such as taxon abundances, phenotypes, or genomic properties. GraPhlAn provides separate visualization options for trees (thus potentially unannotated) and their annotations, the latter of which (the annotation module) attaches metadata properties using the PhyloXML format (Han and Zmasek, 2009). This annotation and subsequent metadata visualization process (**Fig. 1**) can be repeatedly applied to the same tree.

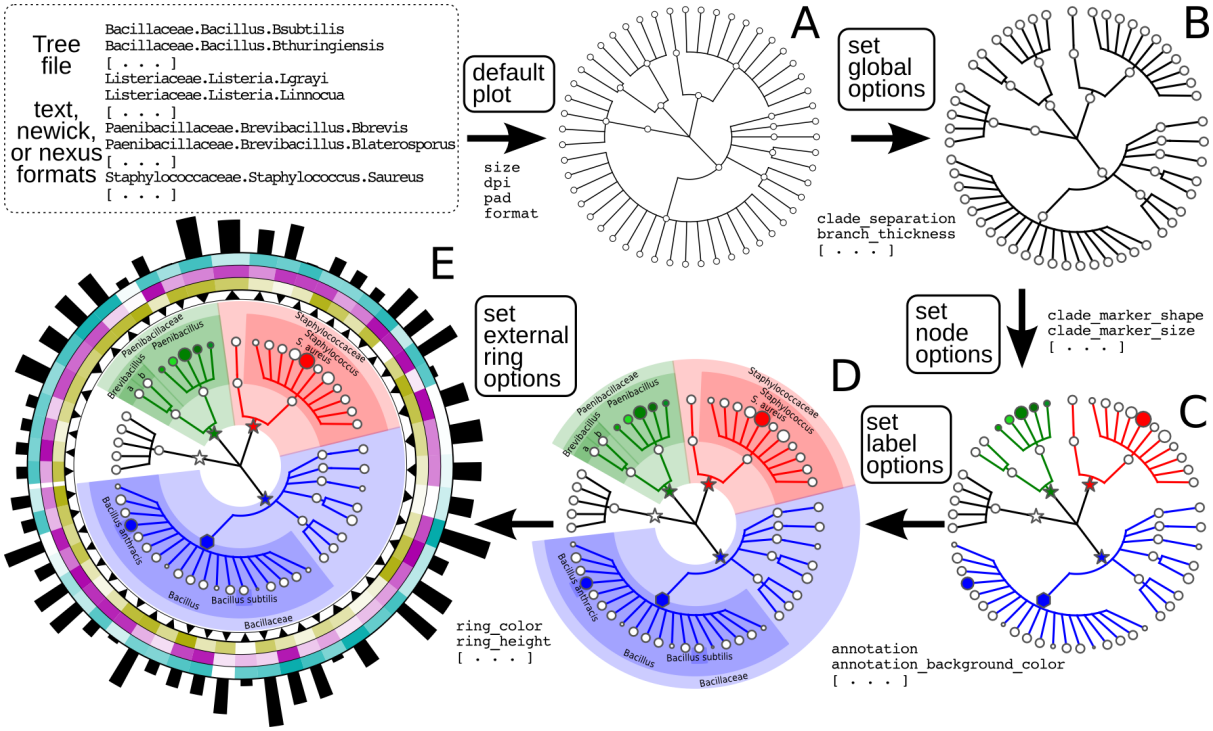


Figure 1: Schematic and simplified example of GraPhlAn visualization of annotated phylogenies and taxonomies.

The software can start from a tree in Newick, Nexus, PhyloXML, or plain text formats. The "default plot" (A) produces a basic visualization of the tree's hierarchical structure. Through an annotation file, it is possible to configure a number of options that affect the appearance of the tree. For instance, some global parameters will affect the whole tree structure, such as the color and thickness of branches ("set global options," B). The same annotation file can act on specific nodes, customizing their shape, size, and color ("set node options," C). Labels and background colors for specific branches in the tree can also be configured ("set label options," D). External to the circular area of the tree, the annotation file can include directives for plotting different shapes, heatmap colors, or bar-plots representing quantitative taxon traits ("set external ring options," E).

The GraPhlAn tree visualization (plotting module) takes as input a tree represented in any one of the most common data formats: Newick, Nexus (Maddison et al., 1997), PhyloXML (Han and Zmasek, 2009), or plain text. Without annotations, the plotting module generates a simple version of the tree (**Fig. 1A**), but the process can then continue by adding a diverse set of visualization annotations. Annotations can affect the appearance of the tree at different levels, including its global appearance ("global options" e.g. the size of the image, **Fig. 1B**), the properties of subsets of nodes and branches ("node options" e.g. the color of a taxon, **Fig. 1C**), and the background features used to highlight sub-trees ("label options" e.g. the name of a species containing multiple taxa, **Fig. 1D**). A subset of the available configurable options includes the thickness of tree branches, their colors, highlighting background colors and labels of specific sub-trees, and the sizes and shapes of individual nodes. Wild cards are supported to share graphical and annotation details among sub-trees by affecting all the descendants of a clade or its terminal nodes only. These features in combination aim to conveniently highlight specific sub-trees and metadata patterns of interest.

Additional taxon-specific features can be plotted as so-called external rings when not directly embedded into the tree. External rings are drawn just outside the area of the tree and can be used to display specific information about leaf taxa, such as abundances of each species in different conditions/environments or their genome sizes. The shapes and forms of these rings are also configurable; for example, in **Fig. 1E** ("set external ring options"), the elements of the innermost external ring are triangular, indicating the directional sign of a genomic property. The second, third, and fourth external rings show leaf-specific features, using a heatmap gradient from blank to full color. Finally, the last external ring is a bar-plot representing a continuous property of leaf nodes of the tree.

2.3.2 Compact representations of phylogenetic trees with associated metadata

Visualizing phylogenetic structures and their relation to external metadata is particularly challenging when the dimension of the internal structure is large. Mainly as a consequence of the low cost of sequencing, current research in microbial genomics and metagenomics needs indeed to visualize a considerable amount of phylogenetic data. GraPhlAn can easily handle such cases, as illustrated here in an example of a large phylogenetic tree (3,737 taxa, provided as a PhyloXML file in the software repository, see Availability section) with multiple types of associated metadata (**Fig. 2**).

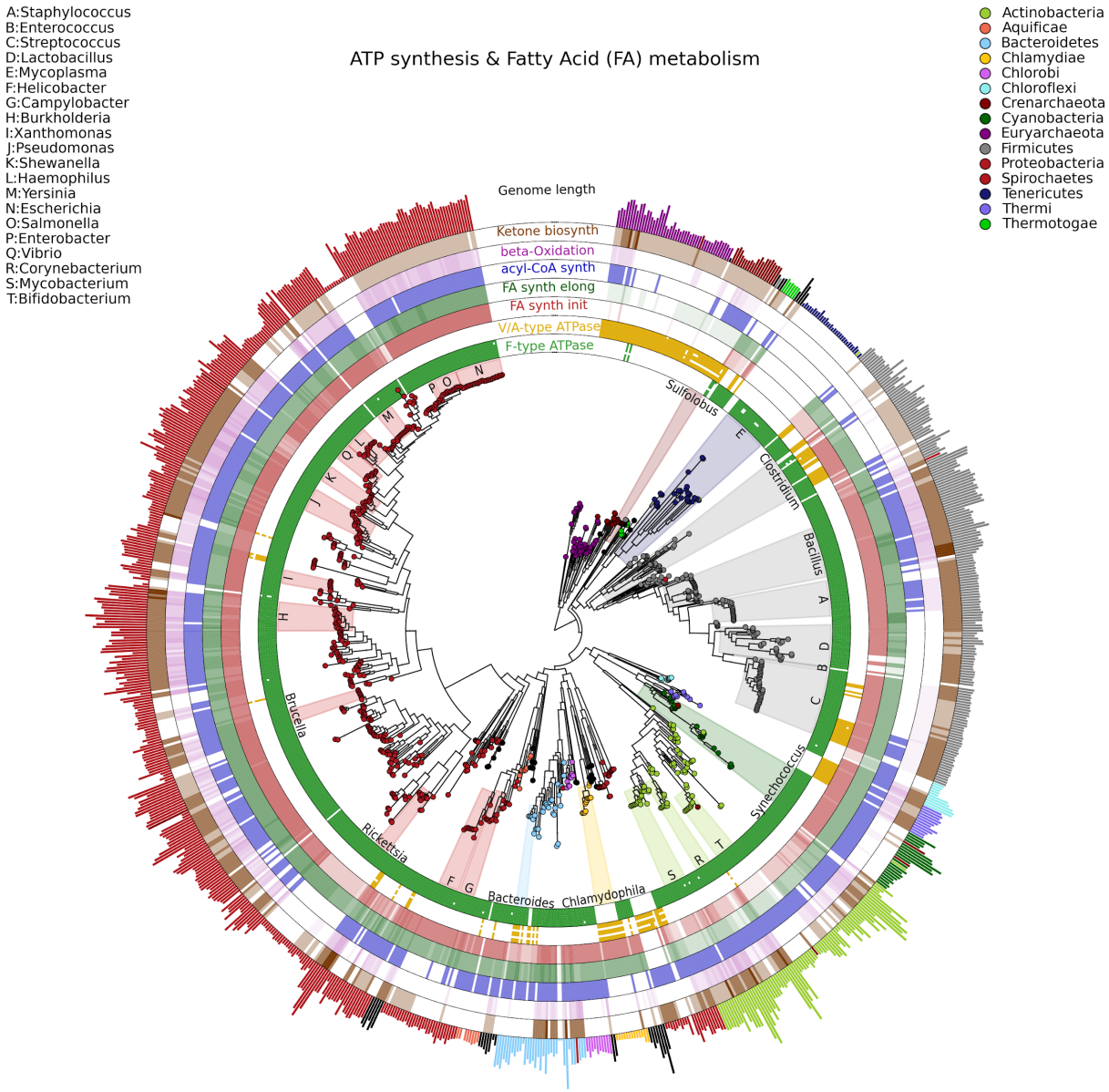


Figure 2: A large, 3,737 genome phylogeny annotated with functional genomic properties.

We used the phylogenetic tree built using PhyloPhlAn (Segata et al., 2013) on all available microbial genomes as of 2013 and annotated the presence of ATP synthesis and Fatty Acid metabolism functional modules (as annotated in KEGG) and the genome length for all genomes. Colors and background annotation highlight bacterial phyla, and the functional information is reported in external rings. ATP synthesis rings visualize the presence (or absence) of each module, while Fatty Acid metabolism capability is represented with a gradient color. Data used in this image are available as indicated in the “Datasets used” paragraph, under “Materials and Methods” section.

Specifically, we used GraPhlAn to display the microbial tree of life as inferred by PhyloPhlAn (Segata et al., 2013), annotating this evolutionary information with genome-specific metadata (**Fig. 2**). In particular, we annotated the genome contents related to seven functional modules from the KEGG database (Kanehisa et al., 2012), specifically two different ATP synthesis machineries (M00157: F-type ATPase and M00159: V/A-type ATPase) and five modules for bacterial fatty acid metabolism (M00082: Fatty acid

biosynthesis, initiation, M00083: Fatty acid biosynthesis elongation, M00086: acyl-CoA synthesis, M00087: beta-Oxidation, and M00088: Ketone body biosynthesis). We then also annotated genome size as an external circular bar plot.

As expected, it is immediately visually apparent that the two types of ATPase are almost mutually exclusive within available genome annotations, with the V/A-type ATPase (module M00159) present mainly in *Archaea* and the F-type ATPase (module M000157) mostly characterizing *Bacteria*. Some exceptions are easily identifiable: *Thermi* and *Clamydophilia*, for instance, completely lack the F-type ATPase, presenting only the typically archaea-specific V/A-type ATPase. As previously discussed in the literature (Cross and Müller, 2004; Mulkidjanian et al., 2007), this may due to the acquisition of V/A-type ATPase by horizontal gene transfer and the subsequent loss of the F-type ATPase capability. Interestingly, some species such as those in the *Streptococcus* genus and some *Clostridia* still show both ATPase systems in their genomes.

With respect to fatty acid metabolism, some clades - including organisms such as *Mycoplasmas* - completely lack any of the targeted pathways. Indeed, *Mycoplasmas* are the smallest living cells yet discovered, lacking a cell wall (Razin, 1992) and demonstrating an obligate parasitic lifestyle. Since they primarily exploit host molecular capabilities, *Mycoplasmas* do not need to be able to fulfill all typical cell functions, and this is also indicated by the plotted very short genome sizes. *Escherichia*, on the other hand, has a much longer genome, and all the considered fatty acid metabolism capabilities are present. These evolutionary aspects are well known in the literature, GraPhlAn permits them and other phylogeny-wide genomic patterns to be easily visualized for further hypothesis generation.

2.3.3 Visualizing microbiome biomarkers

GraPhlAn provides a means for displaying either phylogenetic (trees with branch lengths) or taxonomic (trees without branch length) data generated by other metagenomic analysis tools. For instance, we show here examples of GraPhlAn plots for taxonomic profiles (**Fig. 3**), functional profiles (**Fig. 4**), and specific features identified as biomarkers (**Fig. 3** and **4**). In these plots, GraPhlAn highlights microbial sub-trees that are found to be significantly differentially abundant by LEfSe (Segata et al., 2011), along with their effect sizes as estimated by linear discriminant analysis (LDA). To enhance biomarker visualization, we annotated them in the tree with a shaded background color and with clade names as labels, with decreasing font sizes for internal levels. To represent the effect size, we scaled the node color from black (low LDA score) to full color (high LDA score).

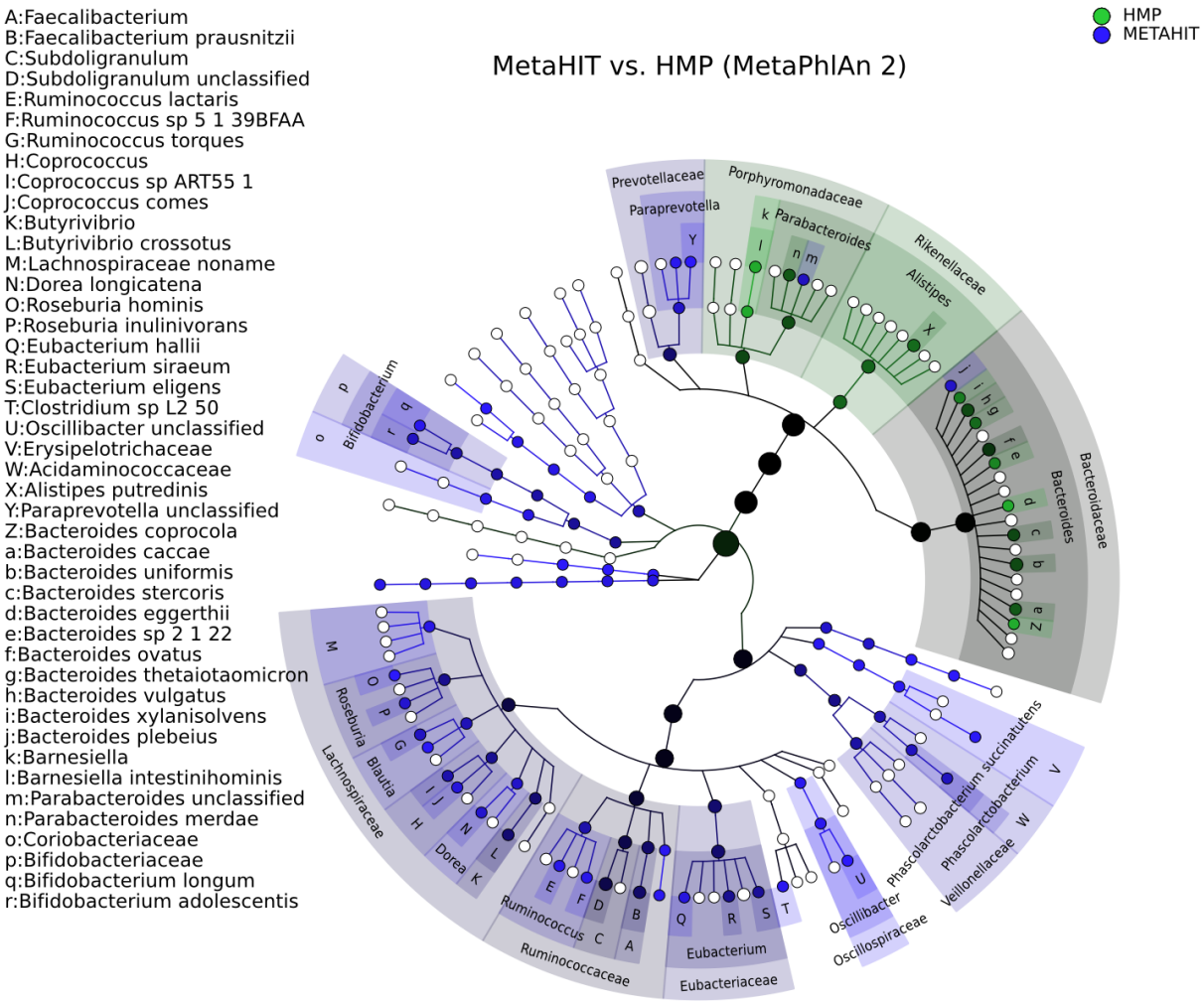


Figure 3: Taxonomic comparison between HMP and MetaHIT stool samples.

The taxonomic cladogram shows a comparison between the MetaHIT and HMP studies limited to samples from the gut (for the latter) and from healthy subjects (for the former). This image has been generated by GraPhlAn using input files from the supporting "export2graphlan" script (see "Materials and Methods") applied on the output of MetaPhlAn2 (Segata et al., 2012a) and LEfSe (Segata et al., 2011). Colors distinguish between HMP (green) and MetaHIT (blue), while the intensity reflects the LDA score, an indicator of the effect sizes of the significant differences. The size of the nodes correlates with their relative and logarithmically scaled abundances. Data used for this image is available as indicated under "Datasets used" paragraph in the "Materials and Methods" section.

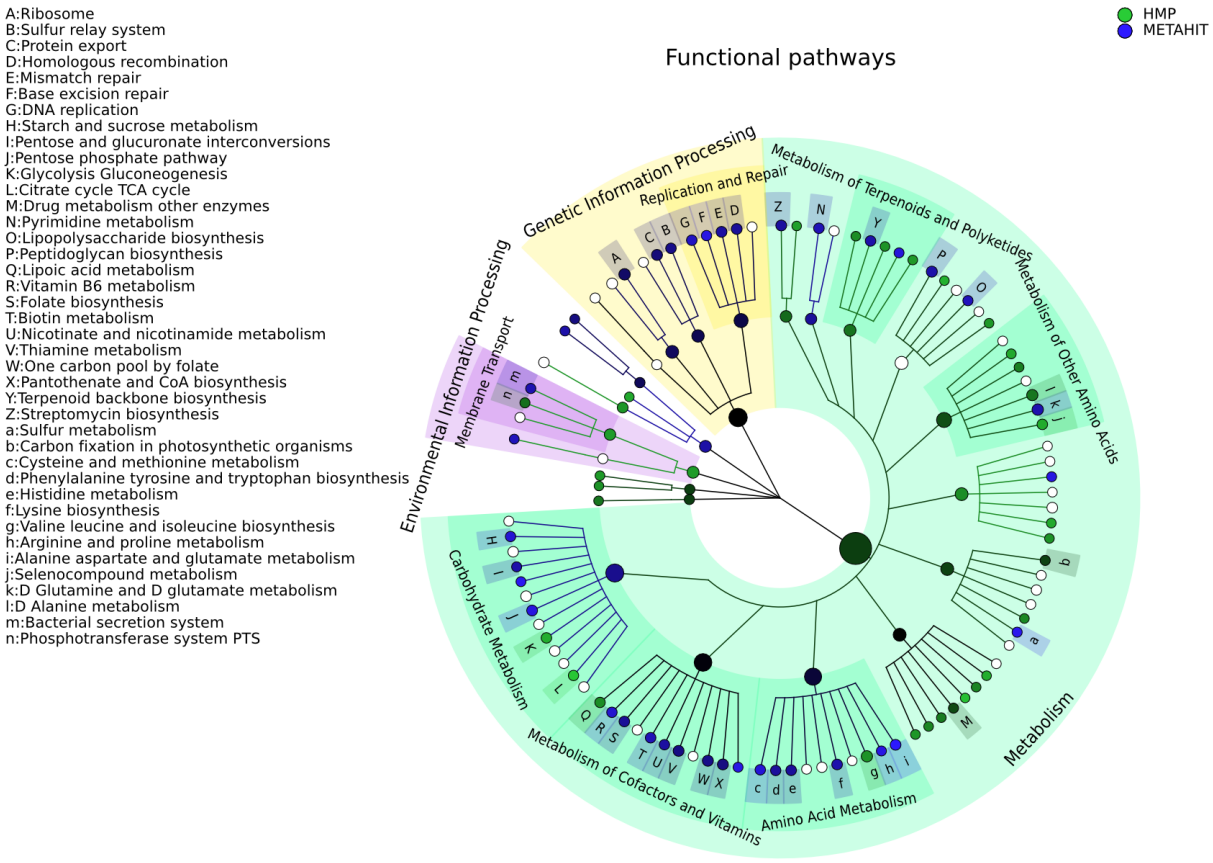


Figure 4: Comparison of microbial community metabolic pathway abundances between HMP and MetaHIT.

Comparison of functional pathway abundances from the HMP (green) and MetaHIT (blue). This is the functional counterpart of the plot in **Fig. 3** and was obtained applying GraPhlAn on HUMAnN (Abubucker et al., 2012) metabolic profiling. The intensity of the color represents the LDA score, and the sizes of the nodes are proportional to the pathway relative abundance estimated by HUMAnN. Three major groups are automatically highlighted by specifying them to the export2graphlan script: Environmental Information Processing, Genetic Information Processing, and Metabolism. Data used for this image is available as indicated under “Datasets used” paragraph in “Materials and Methods” section.

Fig. 3 shows the taxonomic tree of biomarkers (significantly differential clades) resulting from a contrast gut metagenome profiles from the Human Microbiome Project (HMP) (HMP et al., 2012) and MetaHIT samples (Qin et al., 2010). Only samples from healthy individuals in the latter cohort were included. The filtered dataset was analyzed using LEfSe (Segata et al., 2011) and the cladogram obtained using the *export2graphlan* script provided with GraPhlAn and discussed in the following section. As expected, the image highlights that *Firmicutes* and *Bacteroides* are the two most abundant taxa in the healthy gut microbiome (David et al., 2014; Wu et al., 2011). The *Bacteroidetes* phylum contains many clades enriched in the HMP dataset, while *Firmicutes* show higher abundances for MetaHIT samples. GraPhlAn can thus serve as a visual tool for inspecting specific significant differences between conditions or cohorts.

Functional ontologies can be represented by GraPhlAn in a similar way and provide complementary features to the types of taxonomic analyses shown above. Metabolic profiles quantified by HUMAnN (Abubucker et al., 2012) using KEGG (Kanehisa et al., 2014) from

the same set of HMP and MetaHIT samples are again contrasted on multiple functional levels in **Fig. 4**. The tree highlights three different broad sets of metabolic pathways: Environmental Information Processing, Genetic Information Processing, and Metabolism, with the last being the largest subtree. More specific metabolic functions are specifically enriched in the HMP cohort, such as Glycolysis and the Citrate cycle, or in the MetaHIT cohort, such as Sulfur Metabolism and Vitamin B6 Metabolism. This illustrates GraPhlAn's use with different types of data, such as functional trees in addition to taxonomies or phylogenies. By properly configuring input parameters of *export2graphlan*, we automatically obtained both **Fig. 3** and **Fig. 4** (bash scripts used for these operations are available in the GraPhlAn software repository).

2.3.4 Reproducible integration with existing analysis tools and pipelines

Graphical representations are usually a near-final step in the complex computational and metagenomic pipelines, and automating their production is crucial for convenient but reproducible analyses. To this end, GraPhlAn has been developed with command-driven automation in mind, as well as flexibility in the input “annotation file” so as to be easily generated by automated scripts. Depending on the specific analysis, these scripts can focus on a diverse set of commands to highlight the features of interest. Despite this flexibility, we further tried to ease the integration of GraPhlAn by providing automatic offline conversions for some of the available metagenomic pipelines and by embedding it into the well-established Galaxy web framework (Blankenberg et al., 2010; Giardine et al., 2005; Goecks et al., 2010).

In order to automatically generate GraPhlAn plots from a subset of available shotgun metagenomic tools comprising MetaPhlAn (for taxonomic profiling), HUMAnN (for metabolic profiling), and LEfSe (for biomarker discovery), we developed a script named “export2graphlan” able to convert the outputs of these tools into GraPhlAn input files as schematized in **Fig. 5**. This conversion software is also meant to help biologists by providing initial, automated input files for GraPhlAn that can then be manually tweaked for specific needs such as highlighting clades of particular interest. The export2graphlan framework can further accept the widely adopted BIOM format, both versions 1 and 2 (McDonald et al., 2012). This makes it possible to readily produce GraPhlAn outputs from other frameworks such as QIIME (Caporaso et al., 2010) and mothur (Schloss et al., 2009) for 16S rRNA sequencing studies.

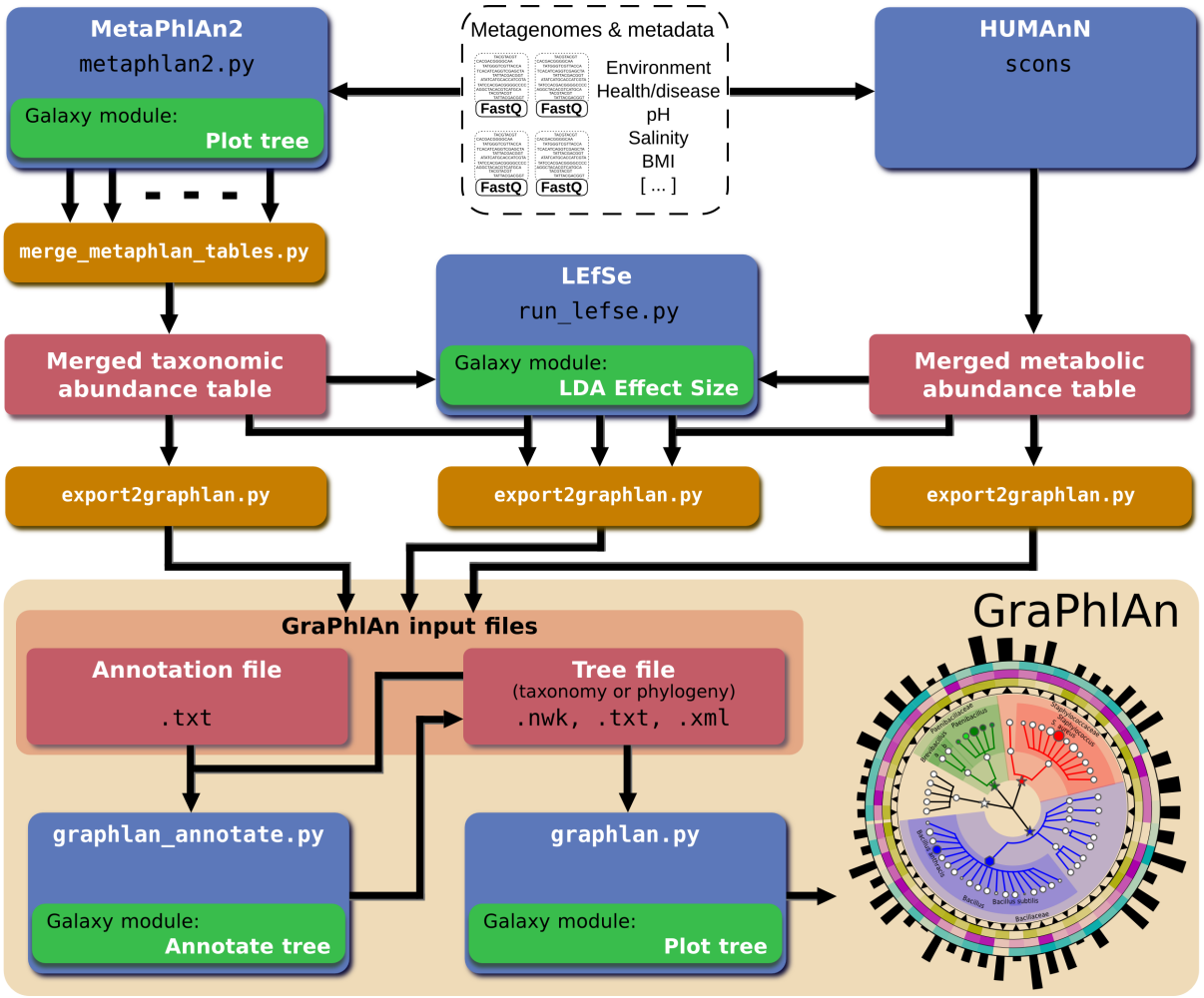


Figure 5: Integration of GraPhlAn into existing analyses pipelines.
We developed a conversion framework called "export2graphlan" that can deal with several output formats from different analysis pipelines, generating the necessary input files for GraPhlAn. Export2graphlan directly supports MetaPhlAn2, LEfSe, and HUMAnN output files. In addition, it can also accept BIOM files (both version 1 and 2), making GraPhlAn available for tools supporting this format including the QIIME and mothur systems. The tools can be ran on local machine as well as through the Galaxy web system using the modules reported in green boxes.

A web-based deployment of the GraPhlAn application is available to the public via Galaxy at http://huttenhower.sph.harvard.edu/galaxy/. The Galaxy interface of GraPhlAn consists of four processing modules: (1) *Upload file*, that manages the upload of the input data into Galaxy; (2) *GraPhlAn Annotate Tree*, which allows the user to specify the annotations that will be applied to the final image; (3) *Add Rings to tree*, an optional step to select an already uploaded file in Galaxy that will be used as an annotation file for the external rings; and (4) *Plot tree*, that sets some image parameters such as the size, the resolution, and the output format.

2.4 Conclusions

We present GraPhlAn, a new method for generating high-quality circular phylogenies potentially integrated with diverse, high-dimensional metadata. We provided several

examples showing the application of GraPhlAn to phylogenetic, functional, and taxonomic summaries. The system has already been used for a variety of additional visualization tasks, including highlighting the taxonomic origins of metagenomic biomarkers (Segata et al., 2011, 2012b; Shogan et al., 2014; Xu et al., 2014), exposing specific microbiome metabolic enrichments within a functional ontology (Abubucker et al., 2012; Sczesnak et al., 2011), and representing 16S rRNA sequencing results (Ramirez et al., 2014). GraPhlAn is, however, not limited to microbiome data and has additionally been applied to animal and plant taxonomies (The Tree of Sex Consortium, 2014) and to large prokaryotic phylogenies built using reference genomes (Baldini et al., 2014; Chai et al., 2014; Langille et al., 2013; Segata et al., 2013).

Compared to the other existing state-of-the-art approaches such as Krona (Ondov et al., 2011) and iTOL (Letunic and Bork, 2007, 2011), GraPhlAn provides greater flexibility, configuration, customization, and automation for publication reproducibility. It is both easily integrable into automated computational pipelines and can be used conveniently online through the Galaxy-based web interface. The software is available open-source, and the features highlighted here illustrate a number of ways in which its visualization capabilities can be integrated into microbial and community genomics to display large tree structures and corresponding metadata.

2.5 Data and software availability

2.5.1 Description of the datasets and figure generation

The data of the taxonomic trees presented in **Fig. 1** is available in the guide folder, inside the examples directory of the GraPhlAn repository (https://bitbucket.org/nsegata/graphlan). This same image is thoroughly described under the “A step-by-step example” section, in the GraPhlAn wiki included in the repository.

The genomic data used for the Tree of Life in **Fig. 2** was obtained from the Integrated Microbial Genomes (IMG) data management system of the U.S. Department of Energy Joint Genome Institute (DOE JGI) 2.0 dataset (http://jgi.doe.gov/news_12_1_06/). From the KEGG database (Kanehisa and Goto, 2000; Kanehisa et al., 2014) we focused on the following modules: M00082, M00083, M00086, M00087, M00088, M00157, and M00159. The input data for drawing **Fig. 2** is available in the *PhyloPhlAn* folder under the *examples* directory of the GraPhlAn repository.

In **Fig. 3**, to comprehensively characterize the asymptomatic human gut microbiota, we combined 224 fecal samples (>17 million reads) from the Human Microbiome Project (HMP) (HMP et al., 2012; Human Microbiome Project Consortium, 2012) and the MetaHIT (Qin et al., 2010) projects, two of the largest gut metagenomic collections available. The taxonomic profiles were obtained by applying MetaPhlAn2. The 139 fecal samples from the HMP can be accessed at http://hmpdacc.org/HMASM/, whereas the 85 fecal samples from MetaHIT were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/, study accession number ERP000108). The input files for obtaining this image with GraPhlAn are present into the examples folder of the repository, inside the *hmp_metahit* directory. The two input files represent the merge result of the MetaPhlAn analysis (*hmp_metahit.txt*) and the LEfSe result on the first file (*hmp_metahit.lefse.txt*). The bash script provided exploits the export2graphlan capabilities to generate the annotation file.

The functional profiles used in **Fig. 4** are the reconstruction of the metabolic activities of microbiome communities. The HUMAnN pipeline (Abubucker et al., 2012)infers community function directly from short metagenomic reads, using the KEGG ortholog (KO) groups. HUMAnN was run on the same samples of **Fig. 3**. The dataset is available on-line at http://www.hmpdacc.org/HMMRC/. As for the previous figure, the input files for obtaining **Fig. 4** are uploaded in the *hmp_metahit_functional* folder, inside the examples directory of the repository. The two files (*hmp_metahit_functional.txt* and *hmp_metahit_functional.lefse.txt*) represent the result of HUMAnN on the HMP and MetaHIT datasets and the result of LEfSe executed on the former file. The bash script provided executes export2graphlan for generating the annotation file and then invoking GraPhlAn for plotting the functional tree.

The dataset of supplementary **Fig. S1** refers to a 16S rRNA amplicon experiment. Specifically, it consists of 454 FLX Titanium sequences spanning the V3 to V5 variable regions, obtained from 24 healthy samples (12 male and 12 female) for a total of 301 samples. Detailed protocols used for enrollment, sampling, DNA extraction, 16S amplification and sequencing are available on the Human Microbiome Project Data Analysis and Coordination Center website HMP Data Analysis and Coordination Center (http://www.hmpdacc.org/tools_protocols/tools_protocols.php). This data are pilot samples from the HMP project (Segata et al., 2011). The input files for obtaining this image is available in the *examples* folder of the export2graphlan repository (https://bitbucket.org/CibioCM/export2graphlan), inside the *hmp_aerobiosis* directory. The two files represent the taxonomic tree of the HMP project and the results of LEfSe executed on the same data.

In the supplementary **Fig. S2** we used the saliva microbiome profiles obtained by 16S rRNA sequencing on the IonTorrent platform (amplifying the hypervariable region V3). The dataset comprises a total of 13 saliva samples from healthy subjects as described in (Dassi et al., 2014) and it is available in the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sra). The input BIOM file for drawing this image is available in the *saliva_microbiome* directory inside the *examples* folder of the GraPhlAn repository.

For the supplementary **Fig. S3** data represent the temporal dynamics of the human vaginal microbiota, and were taken from the study of (Gajer et al., 2012). Data were obtained by 16S rRNA using the 454 pyrosequencing technology (sequencing the V1 and V2 hypervariable regions). The dataset is composed of samples from 32 women that self-collected samples twice a week for 16 weeks. The input file, provided in BIOM format, is present in the *vaginal_microbiota* folder inside the *examples* directory of the GraPhlAn repository.

2.5.2 Software repository, dependences, and user support

GraPhlAn is freely available (http://cibiocm.bitbucket.org/tools/graphlan.html) and released open-source in Bitbucket (https://bitbucket.org/nsegata/graphlan) with a set of working examples and a complete tutorial that guides users throughout its functionality. GraPhlAn uses the matplotlib library (Hunter, 2007). GraPhlAn is also available via a public Galaxy instance at http://huttenhower.sph.harvard.edu/galaxy/.

Export2graphlan is freely available and released open-source in Bitbucket (https://bitbucket.org/CibioCM/export2graphlan) along with a number of examples helpful for

testing if everything is correctly configured and installed. The export2graphlan repository is also present as a sub-repository inside the GraPhlAn repository. The export2graphlan module exploits the pandas library (McKinney, 2012) and the BIOM library (McDonald et al., 2012).

Both GraPhlAn and export2graphlan are supported through the Google group "GraPhlAn-users" (https://groups.google.com/forum/#!forum/graphlan-users), available also as a mailing list at: graphlan-users@googlegroups.com.

Acknowledgments

We would like to thank the members of the Segata and Huttenhower labs for helpful suggestions, the WebValley team and participants for inspiring comments and tests, and the users that tried the alpha version of GraPhlAn providing invaluable feedback to improve the software.

Supplemental Information

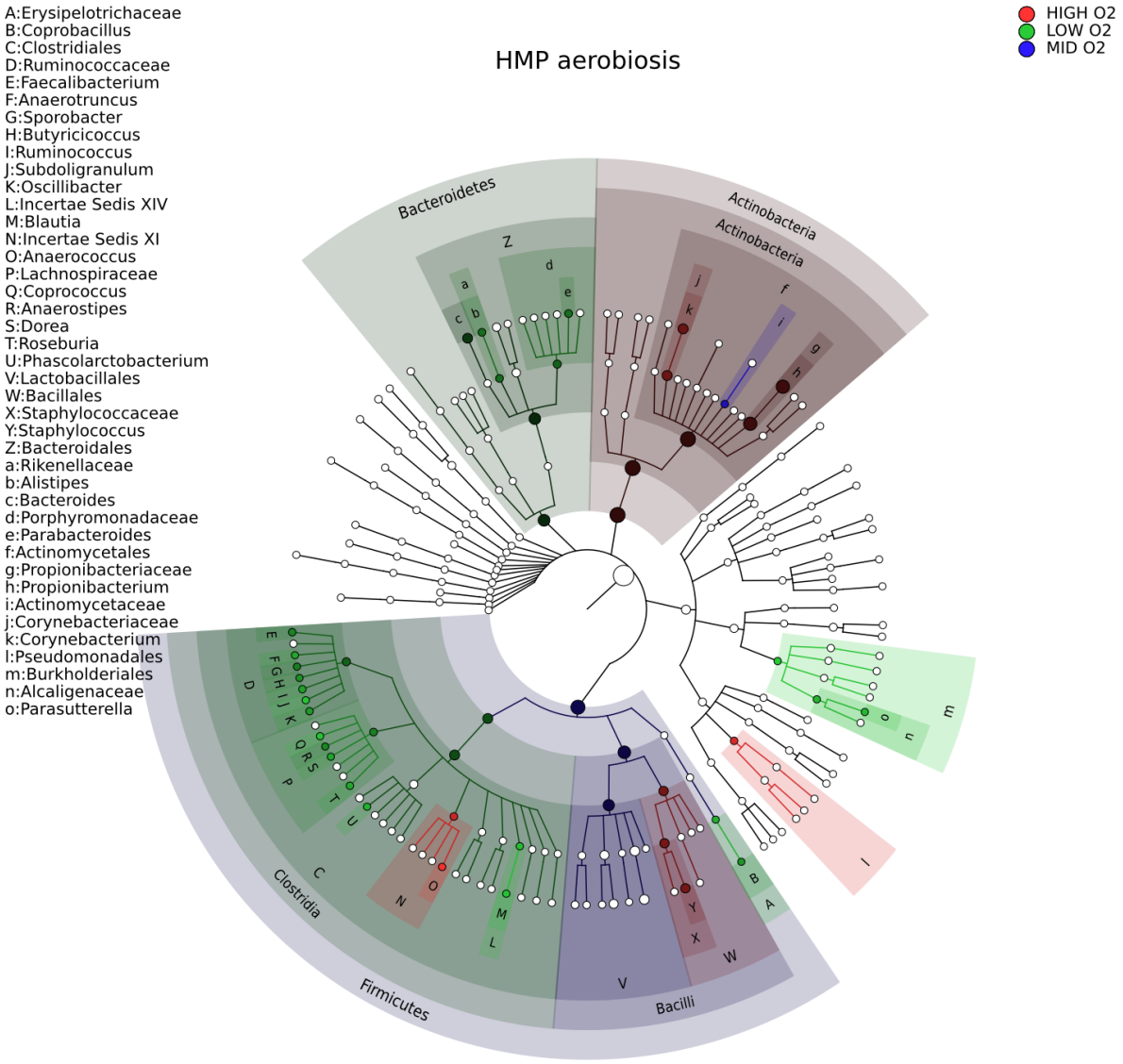


Figure S1: Aerobiosis analysis under aerobic, anaerobic, and microaerobic conditions

The cladogram shows the aerobiosis analysis of the HMP data in three O_2-dependent classes: aerobic (red), anaerobic (blue), and microaerobic (green). The node size reflects the abundance level of each clade, colors are assigned accordingly to one of the three classes, while the lightness intensity of colors respect the LDA score assigned by LEfSe to biomarkers. Data used for this image is available as indicated under “Datasets used” paragraph in “Materials and Methods” section.

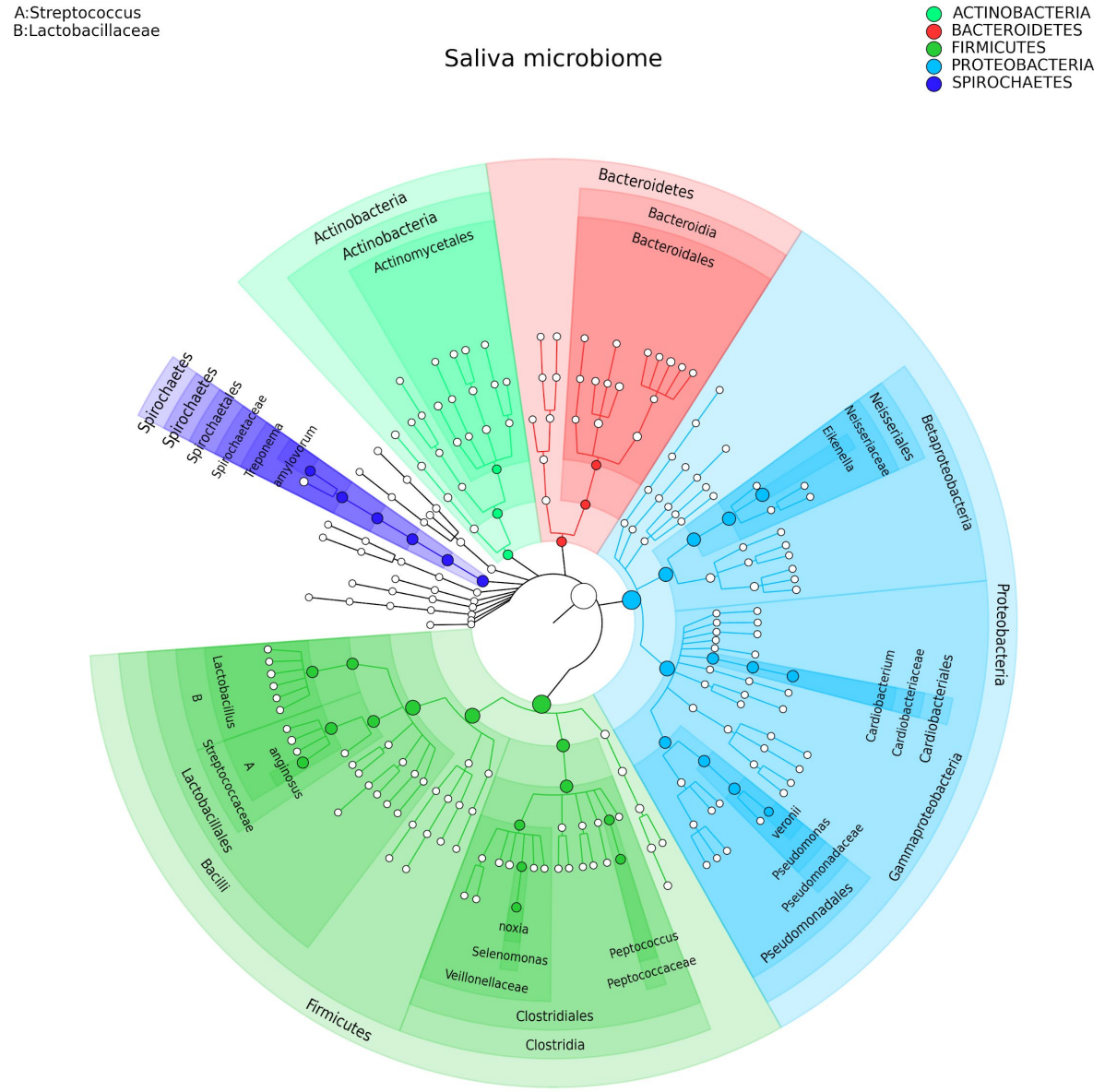


Figure S2: Characterization of the saliva microbiome

This image shows the taxonomic enrichment of the first saliva microbiome sequenced using IonTorrent PGM technology. We exploit export2graphlan capability of handle BIOM files to generate the annotation and tree files for GraPhlAn. Data used for this image is available as indicated under “Datasets used” paragraph in “Materials and Methods” section.

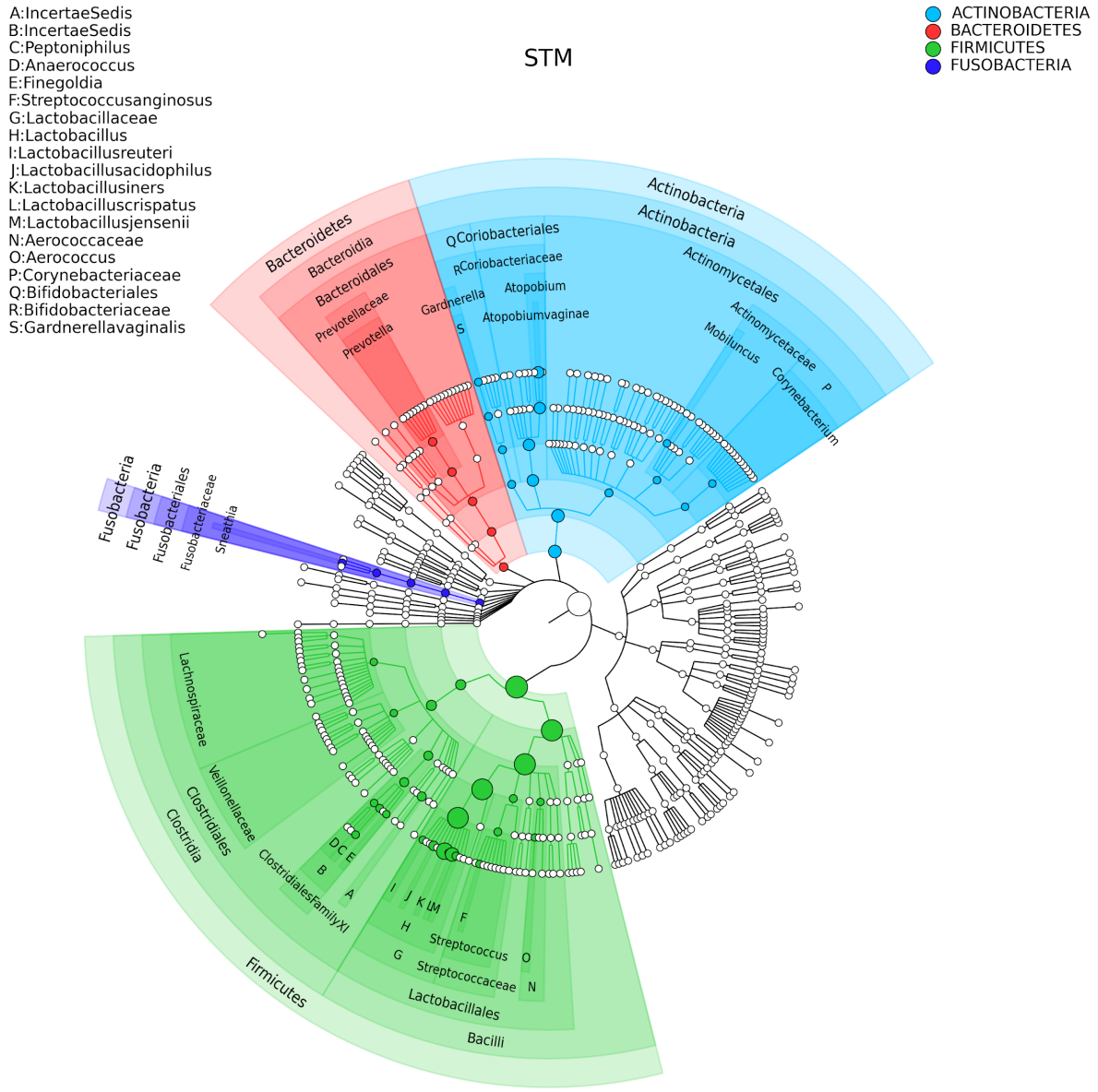


Figure S3: Characterization of temporal dynamics of the human vaginal microbiota

We take the data as a BIOM file from the (Gajer et al., 2012) study. We use export2graphlan to generate the needed files for plotting the circular tree with GraPhlAn. Data used for this image is available as indicated under “Datasets used” paragraph in “Materials and Methods” section.

3. Precise phylogenetic placement of microbial isolates and partial genomes from metagenomes using PhyloPhlAn 2

In this chapter, I present PhyloPhlAn 2 that is a novel phylogenetic framework able to accurately reconstruct from strain-level to tree-of-life size phylogenies. PhyloPhlAn 2 has been designed and developed to be easy-to-use, fast, modular, and flexible, and represents a step towards filling the gap for the use of large-scale phylogenetic analysis in metagenomics and microbial genomics. In the chapter, we will illustrate how the automatic strain-level phylogeny reconstruction with PhyloPhlAn 2 can be as accurate as other customs phylogenetic analysis that requires manual supervision and deep knowledge of the problem at hand. Importantly, PhyloPhlAn 2 implements an efficient pipeline to automatically retrieve from publicly available resources both reference genomes and species-specific sets of phylogenetically informative proteins. This eases both the integration of reference genomes with genomes of a known species to study and the possibility of using a species-specific set of markers for extracting the phylogenetic signal to resolve strain-level phylogenies. In addition, PhyloPhlAn 2 can find the closest species-level genome bin (SGB) to phylogenetically characterize genomes reconstructed from metagenomes. Finally, the ability of scaling-up to tens of thousands genomes allows PhyloPhlAn 2 to reliably reconstruct tree of life phylogenies in a reasonable amount of time. This work is not published yet, but we are submitting it for consideration for publication in a scientific journal.

Asnicar F, Beghini F, Bolzan M, Cumbo F, Manara S, Pasolli E, Knight R, Mirarab S, Huttenhower C, Segata N

Precise phylogenetic placement of microbial isolates and partial genomes from metagenomes using PhyloPhlAn 2

In preparation

Abstract

The continuously increasing availability of genomic data for microbes and microbiomes provides new opportunities to unravel new microbial diversity. Phylogenetic analysis for novel and available microbial genomes is crucial to organize such sequence diversity, and recapitulating all available microbial genomics in an updated tree-of-life phylogeny is key for elucidating evolutionary patterns and relationships between species. However, these tasks are currently hampered by the lack of comprehensive phylogenetic frameworks able to automatically reconstruct phylogenies from isolate genomes and to characterize genomes from metagenomics data. In this work, we propose PhyloPhlAn 2 that is a phylogenetic software framework providing an easy-to-use, fast, modular, customizable, and flexible pipeline for accurate large-scale phylogenetic analysis at multiple resolution levels. PhyloPhlAn 2 allows to automatically retrieve reference genomes to help in elucidating phylogenetic relationships of newly reconstructed and not-yet-characterized genomes and automatically retrieves the maximally phylogenetically informative protein families to be used for accurate reconstruction of strain-level phylogenies. The new framework is also specifically tailored at profiling genomes reconstructed from metagenomes, by using the recently developed catalog of species-level genome bin to first assess the novelty of input genome bins and then place the new genomes into the phylogenetic context of existing genomes and species. Several real-world examples demonstrate the ability of PhyloPhlAn 2

to perform such tasks enabling deeper and more accessible investigations in microbial genomics.

3.1 Introduction

Genomes from both isolate sequencing and metagenomic assembly are continuously generated and made available through public resources. This increasing amount of microbial data is filling gaps into the overall characterization of the microbial diversity in the human body and on earth. Phylogenetic analysis is crucial in this context to both evaluate the degree of novelty of new microbial sequences and to characterize genomes that are obtained without phenotypic information. Moreover, reconstructing a complete microbial tree-of-life is fundamental in understanding evolutionary relations at very large scale, and in metagenomics, this can provide crucial insights about the relations between members of the microbiome. To date, however, there are no scalable and automatic phylogenetic methods that can tackle these challenges.

In the literature, there are few pipelines implementing a full phylogenetic analysis, including the first implementation of PhyloPhlAn (Segata et al., 2013), PhyloSift (Darling et al., 2014), ezTree (Wu, 2018), and GToTree (Lee, 2019). However, these pipelines are limited in terms of modularity of the phylogenetic analysis, of flexibility in the choice of the computational approaches for each module, and of the selection of the sequence markers to use to extract the phylogenetic signal. None of the above tools, for instance, are flexible with respect to the set of markers used in the analysis, making it difficult to adapt the phylogeny estimation to the input (i.e., strain-level or tree-of-life). Only recently proposed ones like GToTree are trying to automatically retrieve reference genomes from public resources, but none of them allows to automatically retrieve species-specific sets of core genes to be used for accurate strain-level phylogeny inference.

Several successful tools are available that implement single specific steps of a phylogenetic pipeline. These include the steps for the multiple-sequence alignment estimation (MUSCLE (Edgar, 2004), MAFFT (Katoh and Standley, 2013), T-Coffee (Notredame et al., 2000), OPAL (Wheeler and Kececioglu, 2007), PASTA (Mirarab et al., 2015), and UPP (Nguyen et al., 2015b)) or the phylogeny reconstruction (FastTree (Price et al., 2009, 2010), RAxML (Stamatakis, 2014), ASTRAL (Mirarab and Warnow, 2015; Mirarab et al., 2014; Zhang et al., 2018), ASTRID (Vachaspati and Warnow, 2015), and IQ-TREE (Nguyen et al., 2015a)). However, none of them can be used automatically, and linking them in a manner that is appropriate for the specific phylogenetic tasks is not simple and require substantial expertise in computational phylogenetics.

Finally, another key aspect is the computational efficiency of phylogenetic pipelines, especially when dealing with tens of thousands of genomes. Multi-locus sequence type (MLST) typing, for instance, can be an alternative and quick way to assign a sequence type to a genome based on the SNPs profile of as few as five to ten loci for each species. However, strain-level resolved phylogenies integrating thousands of reference genomes result in a more accurate characterization of species subtypes and their characteristics, at a higher resolution than MLST profiling. Whole genome large-scale microbial phylogenies are thus an open computational challenge.

Here we present PhyloPhlAn 2, a fully automatic and complete phylogenetic pipeline that retrieves and integrates thousands of reference genomes from public resources. PhyloPhlAn 2 automatically retrieves species-specific sets of core proteins from UniRef90 to build accurate strain-level phylogenies and is able to scale-up to tens of thousands of genomes for the inference of tree-of-life size phylogenies. Additionally, PhyloPhlAn 2 phylogenetically places both known or yet-to-be-characterized species reconstructed from metagenomes.

3.2 Results

3.2.1 Fully automated, precise phylogenetic placement of genomes and metagenomes

PhyloPhlAn 2 has been designed to provide an easy-to-use and fully automatic pipeline for accurate phylogenetic analysis (**Figure 1**). Key features in PhyloPhlAn 2 are the automatic retrieval of thousands of reference genomes from public resources and of species-specific sets of UniRef90 proteins to allow for accurate reconstruction of strain-level phylogenies. In PhyloPhlAn 2, the complete, accessible and automatic phylogenetic pipeline is achieved through configuration files (described in the **Methods**, **Figure 1**). Additionally, PhyloPhlAn 2 has been designed to build very large phylogenies, scaling up to tens of thousands of input genomes thanks to the implementation of several approaches that allow to reduce in size the multiple-sequence alignments (MSAs) without losing the phylogenetic signal (see **Methods**). Further details of the internal steps of PhyloPhlAn 2, such as the automatic download of reference genomes and core sets of proteins families, the phylogenetic analysis steps, and the choice between a concatenation or a gene tree summary approach, are detailed in the **Methods** section. PhyloPhlAn 2 provides as outputs the reconstructed phylogeny and the generated MSA. Optionally, it can also provide a table of the estimated mutation rates for the inputs (**Figure 1**). Both the MSA and the estimated mutation rates can be used for downstream analysis like a phylogenetic bootstrapping analysis (see **Methods**).

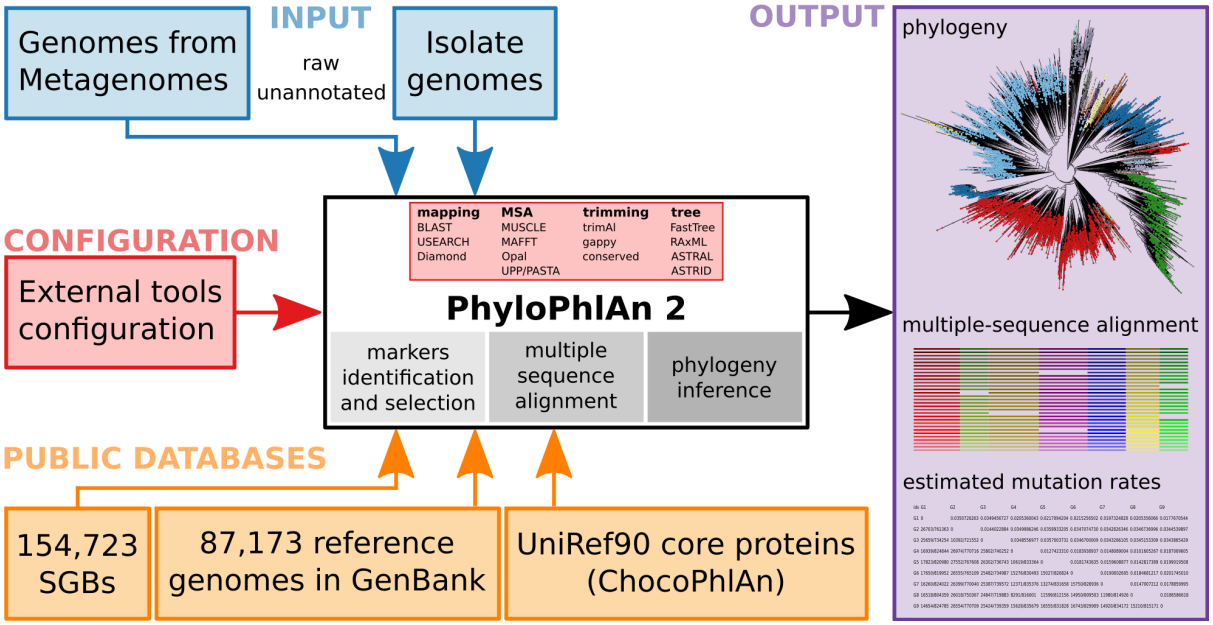


Figure 1. PhyloPhlAn 2 pipeline overview. The overview of the PhyloPhlAn 2 pipeline, showing the different input types accepted (unannotated genomes, proteomes, and metagenome-reconstructed genomes), the configuration file that contains the details for running the needed external tools and also specify whether to run a concatenation or a gene

trees analysis, the possibility to include information like the closest SGB assignment, the automatic retrieval of reference genomes from Genbank, and the automatic download of species-specific sets of core proteins from the UniRef90 database. The output panel is showing the three different outputs of PhyloPhlAn 2: the phylogenetic tree, the concatenated MSA, and, if specified, the estimated mutation rates for the input genomes that can be used for downstream evaluations.

3.2.2 PhyloPhlAn 2 on automating and facilitating phylogenetic analysis of new isolate genomes from extant species

The ease with which microbial isolate genomes can be sequenced and assembled is not paralleled by the analysis part. Newly sequenced genomes usually require some standard operations to phylogenetically characterize them in the context of available reference genomes deposited in public databases like NCBI and ENA. PhyloPhlAn 2 has been developed to completely automate this task. The pipeline starts from the newly sequenced genomes and uses pre-computed UniRef90 core genes for the species of interest and builds a whole-genome phylogenetic tree. In the process, PhyloPhlAn 2 can integrate all (or a specified number) of the available reference genomes present in NCBI for the species of interest, facilitating the downstream analysis, without the need to manually search and retrieve the genomes.

To illustrate the automatic pipeline for this tasks, we applied PhyloPhlAn 2 on a set of *Staphylococcus aureus* isolates genomes we analyzed elsewhere (Manara et al., 2018) and present the resulting strain-level phylogeny of the 135 *S. aureus* genomes based on the identified core set of 2,128 (of which 1,658 met the requirement to be present in at least 99% of the inputs) UniRef90 (**Figure 2A**). In the original work of (Manara et al., 2018) the whole-genome phylogeny has been built by computing first the set of core genes (1,464 core genes with a coreness of at least 99%) using Roary (Page et al., 2015). To directly compare the two *S. aureus* phylogenies we computed the normalized patristic distances of the two trees and report them as a scatterplot in **Figure 2B**. It is immediately clear that the two different phylogenetic approaches are retrieving the same phylogenetic structure, with the advantage that in PhyloPhlAn 2 there is no need to compute the pangenome (that can be incomplete if there are not enough genomes from that species) to identify the right set of core genes to be used, but that is instead automatic. Still in a fully automated way, PhyloPhlAn 2 can also retrieve and analyze other publicly available genomes, and to illustrate this we present an ordination plot (**Figure 2C**) of the normalized patristic distances of the 135 *S. aureus* isolates integrated with a 1,000 reference genomes automatically retrieved from the Genbank public database, **Supplementary Figure S1** shows the unrooted phylogeny. It can be appreciated how the phylogenetic differences in the large strain-level phylogeny of 1,135 genomes resemble the sequence type as assigned by the MLST[11,12] profiling and their clonal complex. We thus showed how PhyloPhlAn 2 can automate the full phylogenetic analysis pipeline for isolate genomes, only taking as input the raw genomes, and that the generated phylogeny is at least as accurate as the one resulting from ad hoc pipelines that required expert supervision and weeks of computational resources.

[11] Seemann T, mlst, https://github.com/tseemann/mlst

[12] This publication made use of the PubMLST website (https://pubmlst.org/) developed by Keith Jolley (Jolley and Maiden, 2010) and sited at the University of Oxford. The development of that website was funded by the Wellcome Trust.

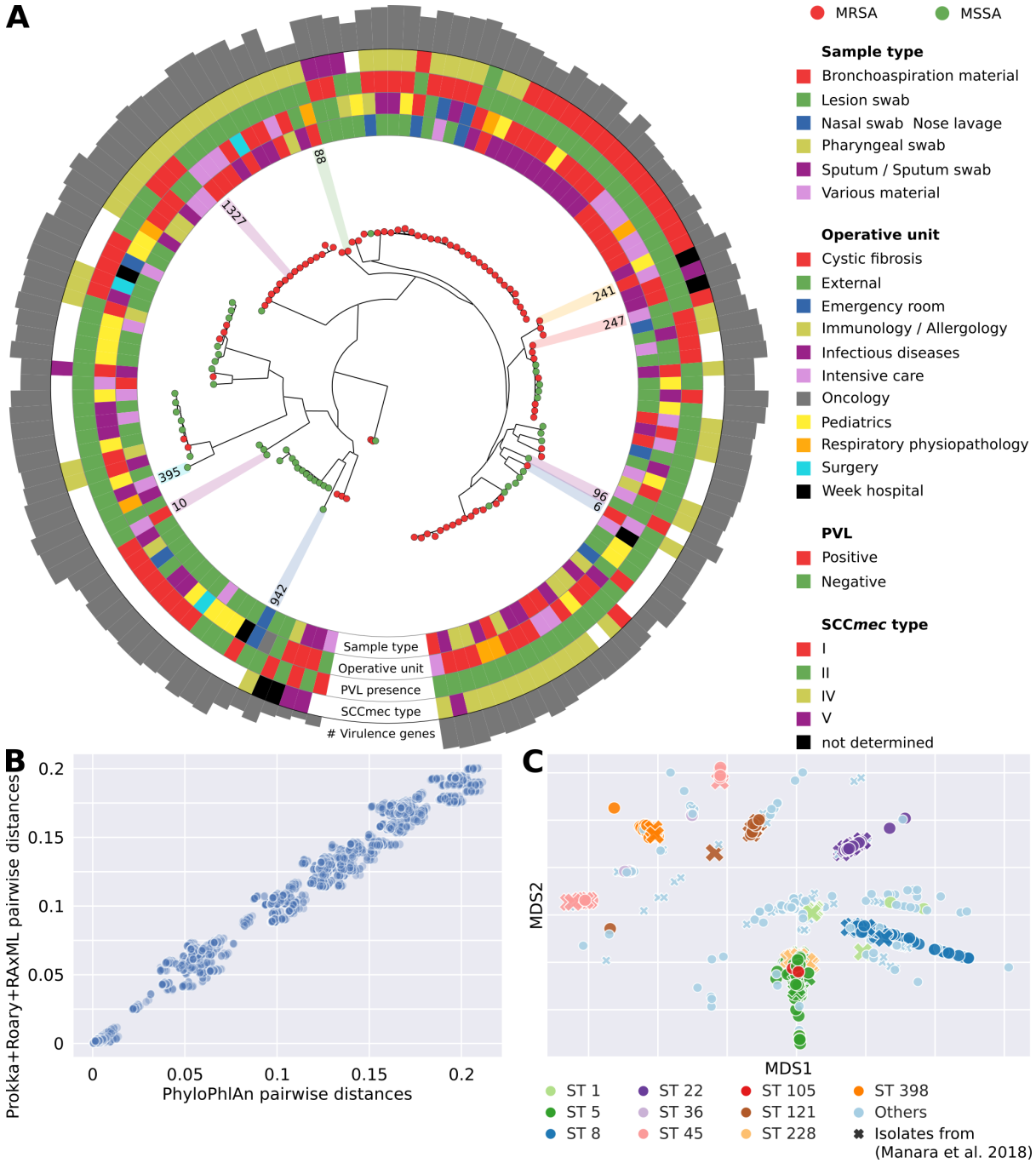
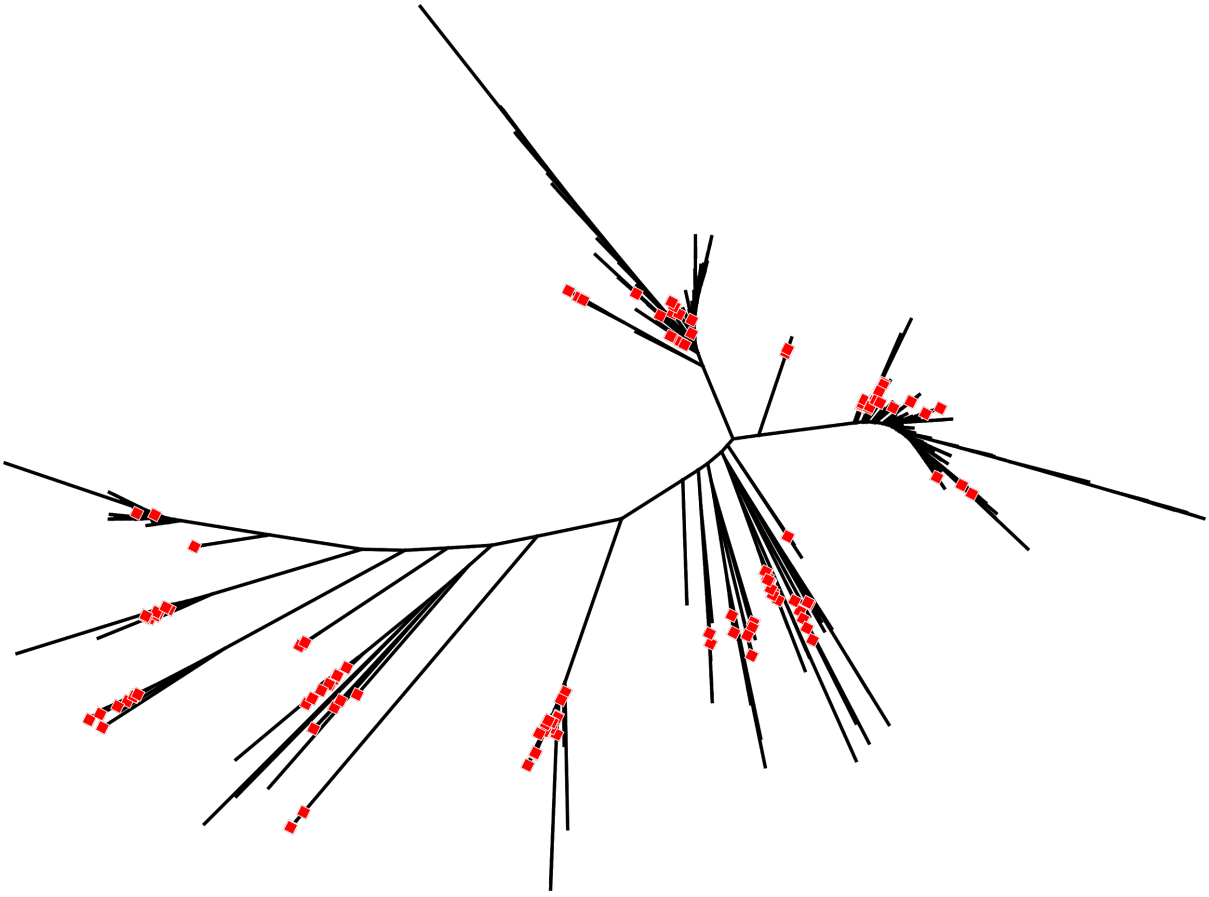


Figure 2. Accurate automatic reconstruction of *Staphylococcus aureus* strain-level phylogeny. (A) Phylogenetic tree of our 135 *S. aureus* strains reconstructed by PhyloPhlAn 2 using the pre-identified and automatically retrieved species-specific core genes, and displayed with GraPhlAn using the same annotations as in Figure 1 of (Manara et al., 2018). **(B)** Scatterplot of the normalized patristic distances in the PhyloPhlAn-reconstructed tree and in the original phylogeny in (Manara et al., 2018). **(C)** Ordination plot of the phylogenomic distances of the phylogenetic tree built integrating the 135 *S. aureus* isolates with 1,000 reference genomes (**Supplementary Figure 1**), coloring the ten most prevalent STs and highlighting the 135 isolate genomes.



Supplementary Figure 1. Unrooted phylogeny of *Staphylococcus aureus* including the 135 isolates and a 1,000 reference genomes. Highlighted in red the 135 isolates we previously identified in (Manara et al., 2018). The phylogeny reconstructed with PhyloPhlAn 2 and based on 1,658 core UniRef90 proteins, is showing how the 135 *S. aureus* isolate genomes are representing a good species diversity, being well distributed and placed in the phylogeny along with the 1,000 reference genomes.

3.2.3 Robust phylogenetic and taxonomic placement for known and unknown metagenome-assembled genomes

A different task that can be performed by PhyloPhlAn 2 is the assignment of a putative taxonomic label and the phylogenetic placement of genomes reconstructed from metagenomes. In this setting, it is not known *a priori* to what species the reconstructed genomes belong to. To taxonomically label these reconstructed genomes, PhyloPhlAn 2 takes as input a set of genomes reconstructed from metagenomes and finds their closest species-level genome bin (SGB) (Pasolli et al., 2019) based on the Mash distance (Ondov et al., 2016) as an approximation of the ANI distance. The default threshold is 5% on the Mash distance, as suggested elsewhere (Bowers et al., 2017; Jain et al., 2018) to be a good compromise for defining bacterial species. The result here can have two outputs where the closest SGB is either a known SGB (kSGB) or an unknown SGB (uSGB). In the first case, the putative taxonomic label can be assigned directly, given that a subset of the input reconstructed genomes is at most 5% genetically distant from a reference genome. In the latter case, where the closest SGB is a uSGB, the putative taxonomic label can be assigned at different levels, depending on whether the uSGB falls into a bin that has been assigned a

taxonomy at the genus level (GGB), at the family level (FGB). If neither a GGB or an FGB have been assigned to the uSGB, the taxonomy is assigned at the phylum level.

Regardless of the level of known taxonomy associated with the genomes, PhyloPhlAn 2 further characterizes the input genomes by phylogenetically analyze them in the context of an automatically downloaded set of closest reference genomes. To this end, the software retrieves the *n* isolates or previously metagenomically assembled genomes and builds a phylogeny merging them with the input genomes. In **Figure 3A**, we analyzed 369 genomes reconstructed from 50 metagenomes of the Ethiopian cohort using the 5% Mash distance threshold. The top heatmap in the **Figure 3A** is showing the presence/absence profile of the 20 most prevalent SGBs in the Ethiopian cohort for each of the 50 metagenomes, while the bottom heatmap is reporting the total number of uSGBs and kSGBs found in each sample.

We then decided to focus on the typical human gut colonizer *Escherichia coli* (kSGB ID 10068) and two very prevalent uSGBs, which were assigned to the Chlamydiae phylum (IDs 19435 and 19436) to illustrate the second-tier phylogeny construction for these input MAGs. For the *E. coli*, we reconstructed eight genomes from the Ethiopia metagenomes, and we used PhyloPhlAn 2 to automatically download the set of core UniRef90 proteins (13,838 in total of which 6,220 were retained for the phylogenetic) and 200 reference genomes. We then reconstructed the phylogeny and annotated it with the *E.coli* phylotypes, annotated using the Python package EzClermont (Waters et al., 2018). We show the ordination based on the patristic distances in **Figure 3B** which is showing good clustering according to the phylotypes and where the eight metagenome-reconstructed genomes are placed.

We reconstructed ten genome bins falling into the two uSGBs IDs 19435 and 19436, and we used PhyloPhlAn 2 for downloading up to two reference genomes for each species in the Chlamydiae phylum. We also downloaded two reference genomes for the two most abundant species in the Actinobacteria phylum: *Mycobacterium abscessus* (with 1,385 reference genomes) and *Mycobacterium tuberculosis* (with 3,988 reference genomes). Actinobacteria is the closest phylum to Chlamydiae, according to the tree-of-life phylogeny in **Figure 4**, and the four Actinobacteria reference genomes are used as the outgroup to root the phylogeny. Since the lowest taxonomic rank is at the phylum level, the phylogeny shown in **Figure 3C** has been built using the 400 universal marker genes proposed in (Segata et al., 2013).

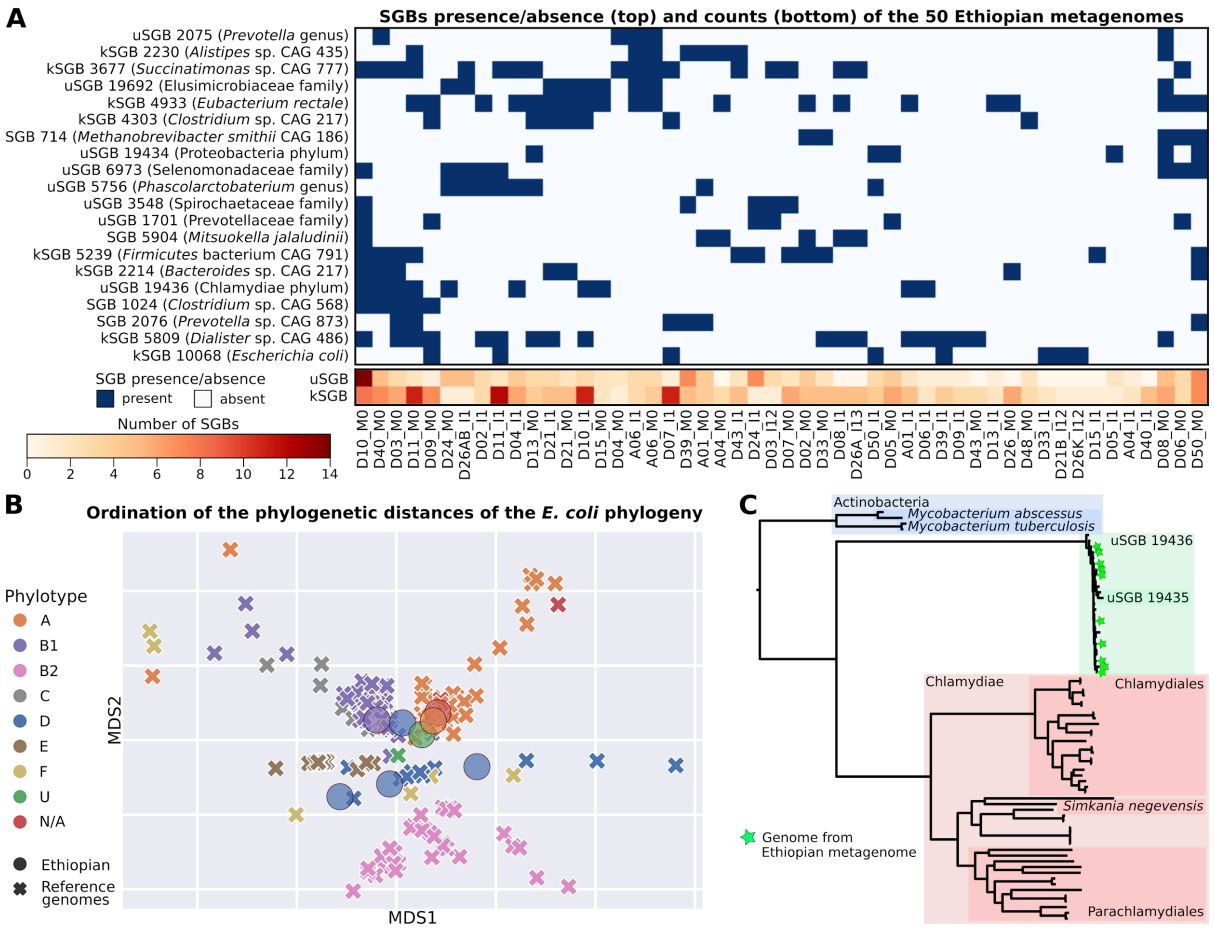


Figure 3. Metagenomic analysis of the Ethiopian cohort using PhyloPhlAn 2. (A) Heatmap of the 20 SGBs most prevalent in the 50 Ethiopian metagenomes. The top part reports the presence/absence of each of the top 20 SGBs, displayed with their identifier and the closest assigned taxonomic label. The bottom part summarizes the total number of uSGBs and kSGBs found in each metagenome. **(B)** Ordination of the patristic distances of the *E. coli* phylogeny built using the core set of UniRef90 of *E.coli* (6,220 proteins) and including the eight Ethiopian bins falling into the *E. coli* kSGB (ID 10068) integrated with a 1,000 *E. coli* reference genomes. **(C)** Phylogenetic tree of the ten Ethiopian bins phylogenetically close to two uSGBs (IDs 19435 and 19436) and assigned to the Chlamydiae phylum. The phylogeny has been integrated with up to two reference genomes for each Chlamydiae species and two genomes for the two most prevalent Actinobacteria species used to root the phylogeny. In this analysis, we used the 400 PhyloPhlAn universal markers.

3.2.4 PhyloPhlAn 2 reconstruction of the largest available tree-of-life

The increasing availability of microbial genomes as well as of publicly available shotgun metagenomics datasets allows for always more accurate genomes reconstruction through computational approaches and is driving the need to put these large amounts of genomic data into phylogenetic relationships. Tree-of-life phylogenies are fundamental to place and characterize both genomes for which a reference is available and novel organisms that are too genetically distant from all known organisms. In this context, we retrieved genomes from the largest set of microbial species currently available, considering all the microbial genomes in NCBI and the additional 154,000 genomes from our recent work (Pasolli et al., 2019). The resulting set of 19,607 genomes (after dereplication at the species level) is the largest dataset available and no methods have ever been applied to such a large genome set. In

Figure 4 we present a novel microbial tree of life built with PhyloPhlAn 2 applied on the 19,607 genomes. The number of genomes in the phylogeny has been reduced to 17,672 after discarding those that did not meet the requirements of the quality control filters implemented in PhyloPhlAn 2, such as the minimum number of universal markers that have to be present in each genome. Genomes, when possible, have been colored according to their phylum label. The accuracy of the phylogeny built using the 400 PhyloPhlAn universal markers proposed in (Segata et al., 2013) is supported by the concordance between the phylogenetic placement of the genomes and their phylum label assigned independently from PhyloPhlAn 2 (**Figure 5**). The reconstruction of this large microbial tree of life took in total ten days and 15 hours using 100 CPUs, of which five days and three hours were used by IQ-TREE (Nguyen et al., 2015a) for inferring the phylogeny. This very-large phylogenetic analysis is based on the 400 PhyloPhlAn universal markers for which we retained in the final MSA a different number of significant positions for each marker, according to its conservation. The concatenated MSA contains 4,522 amino acids aligned positions for a total of 17,672 genomes. This is showing that PhyloPhlAn 2 is able to reconstruct in a reasonable amount of time very-large phylogenies, scaling up to tens of thousands of input genomes and it is promising also for future phylogenetic analysis, with a potentially larger number of available genomes.

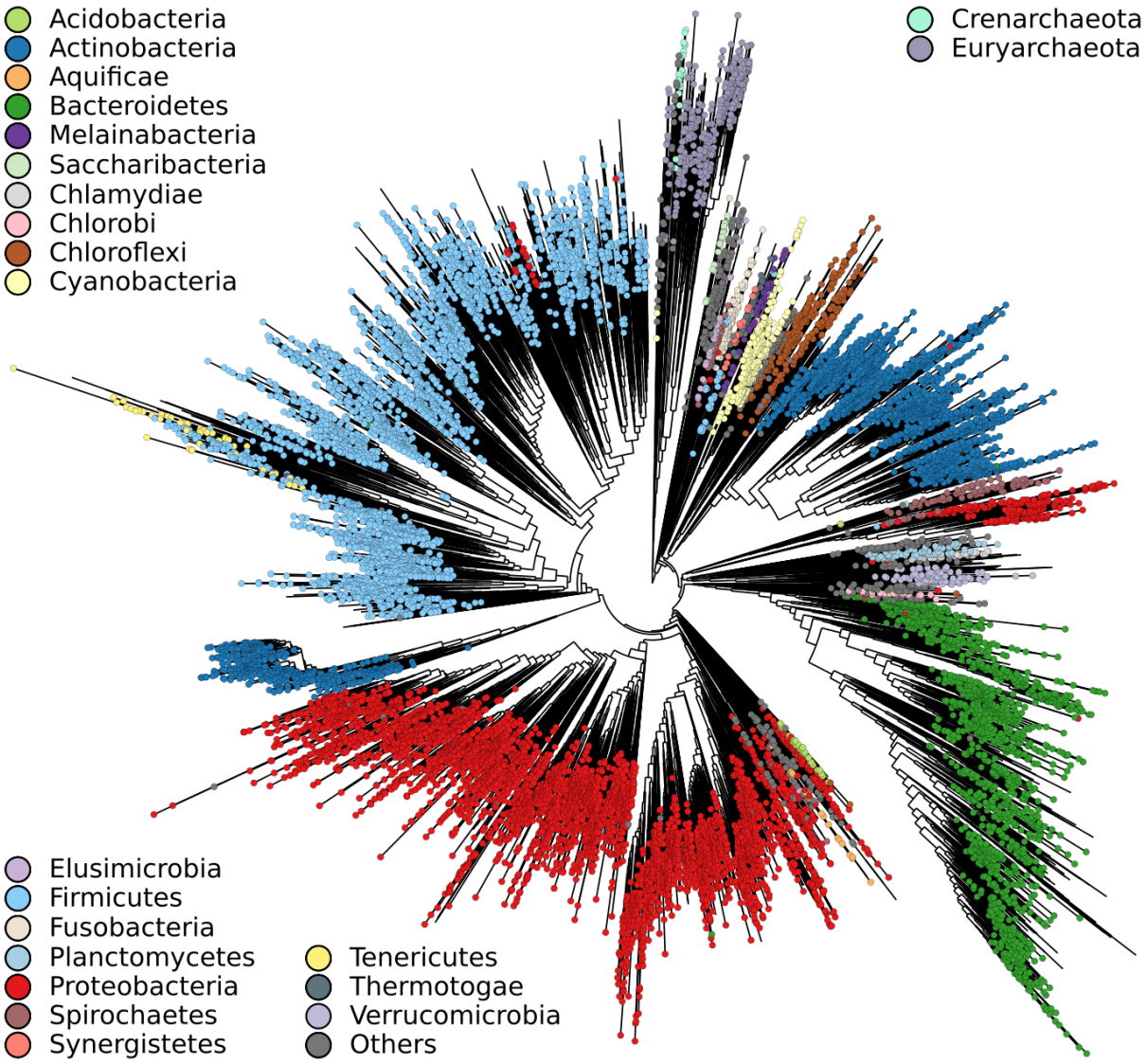


Figure 4. Microbial tree of life built with 17,672 genomes. Maximum-likelihood phylogeny representing the known microbial tree of life built with a starting set of nearly 20,000 genomes, including the SGB representatives from our recent proposed set of genomes reconstructed from metagenomes (Pasolli et al., 2019) and the representatives of the almost 8,000 genomes reconstructed from environmental metagenomes by (Parks et al., 2017). The phylogeny, based on the 400 PhyloPhlAn markers, counts 4,522 amino acids aligned positions for each of the 17,672 genomes and it has been built in 10 days and 15 hours in total.

3.3 Conclusions

We presented PhyloPhlAn 2 and show that it is an effective tool to automatically build accurate strain-level whole-genome phylogenies. These strain-level phylogenies can be automatically enriched including as many reference genomes as available in public databases, and this allows to put newly sequenced isolate genomes into a phylogenetic context. Moreover, a unique feature present in PhyloPhlAn 2 is the assignment of a taxonomic label and phylogenetically characterize novel genomes reconstructed from metagenomes based on their closest species-level genome bin assignment. Finally, we show how PhyloPhlAn 2 is able to manage very-large sets of genomes to build tree-of-life size phylogenies. We believe that PhyloPhlAn 2 can serve as an instrument to recapitulate present and future microbial diversity for both single isolate genomes and metagenomes.

3.4 Methods

3.4.1 Configuration files

The new version of PhyloPhlAn 2 has been designed to work with configuration files that specify both the type of phylogenetic pipeline that will be executed (concatenation or gene trees) and also which external tools and parameters to use. With PhyloPhlAn 2 we also distribute a script named “phylophlan_write_default_configs.sh” that exploits the “phylophlan_write_config_file.py” script for correctly generating four default configuration files, two for the concatenation pipeline and two for the gene trees pipeline. For each pipeline, the two configuration files are needed to distinguish the type of markers database that will be used: a gene markers database needs a configuration file for dealing with nucleotides, while a protein markers database needs one for dealing with amino acids.

3.4.2 Automatic download of reference genomes and core UniRef90 as markers database

In PhyloPhlAn 2, we provide two scripts able to automatically retrieve from public resources both reference genomes and species-specific core sets of UniRef90 proteins, for easing the phylogenetic placement. The script “phylophlan_get_reference.py” lists and downloads the available species: at the time of writing we count 647 archaea species with 828 reference genomes, 16,960 bacterial species with 86,192 reference genomes, and 14 eukaryotic species relevant for the human microbiome analysis with 153 reference genomes. Using the value “all” for the “--get” parameter, the user can download a specified number of reference genomes for all the available species (regulated by the parameter “--how_many”, set to four by default). This allows to build microbial tree-of-life phylogenies in an easy way. The “phylophlan_setup_database.py” script, instead, allows to either format (on multiple files or in a multi-fasta file) a given set of genes to be readable by PhyloPhlAn or to automatically download a pre-identified set of core UniRef90 proteins for a specific species. The latter

option allows to build strain-level resolution phylogenies without the need to compute the pangenome and identify the set of core genes for a given species, which can be computationally intensive otherwise.

3.4.3 Metagenomic pipeline

A novel addition in PhyloPhlAn 2 is the assignment of the closest species-level genome bins (SGBs), a concept and framework we recently introduced (Pasolli et al., 2019), to the set of genome bins from metagenomic assemblies provided as input. This is achieved by using the "phylophlan_metagenomic.py" script that groups the bins based on their closest assigned SGB (configurable using the "--threshold" param, set to 0.05 by default). The user can then decide to select subsets of inputs and use the "phylophlan_get_reference.py" script to download the needed reference genomes, and "phylophlan_setup_database.py" script in the case of a known SGB (kSGB) to download the core set of UniRef90.

3.4.4 Phylogenetic inference pipeline

A standard computational phylogenetics pipeline can be divided into four main steps: marker genes identification, multiple sequence alignment, concatenation or gene trees inference, and phylogeny reconstruction.

The marker genes identification step requires a mapping of the database of markers against the input genomes to extract their homologous to be aligned. Since both markers and inputs can be a mix of genes (genomes) and proteins (proteomes), this step requires a tool able to perform a search in a translated sequence space. PhyloPhlAn 2 currently supports blast (Altschul et al., 1990), USEARCH (Edgar, 2010), and Diamond (Buchfink et al., 2015). Depending on the type of markers, PhyloPhlAn 2 will continue the phylogenetic analysis on the nucleotide space if both markers and inputs are nucleotides, whereas it will proceed on the amino acid space if markers are proteins and inputs a mix of genomes and proteomes. The result of this part in PhyloPhlAn 2 is the set of marker genes containing the unaligned sequences found in the inputs.

At this point, each multi-fasta of each marker is aligned using one of the MSA software available. In PhyloPhlAn 2 we included and tested the following tools: MUSCLE (Edgar, 2004), MAFFT (Katoh and Standley, 2013), Opal (Wheeler and Kececioglu, 2007), UPP (Nguyen et al., 2015b), and PASTA (Mirarab et al., 2015). However, PhyloPhlAn 2 is not limited to the software listed above, as the settings for other MSA tools can be manually inserted into the configuration file. The results from this step are the multiple-sequence alignment for the set of markers.

3.4.5 Choice of concatenation versus gene trees approach

This is the crucial point where a concatenation- or a gene trees-based phylogenetic analysis will be performed. For the concatenation pipeline, all the computed MSAs are concatenated into one large MSA that will be used for the final phylogeny reconstruction. For the gene trees pipeline, instead, each single MSA is used to compute one phylogeny and, through a summary method, all the generated phylogenies are used to derive the final phylogeny.

3.4.6 Large-scale phylogenies

The main challenge when building very large phylogenies is to limit the length of the MSA that will be provided to the inference phylogeny tool. To reduce the length of an MSA, a

number of methods (Capella-Gutiérrez et al., 2009; Castresana, 2000; Dress et al., 2008; Sela et al., 2015; Webb et al., 2017) or ways of scoring each position in the MSA (Chang et al., 2014; Edgar, 2009; Penn et al., 2010; Talavera and Castresana, 2007; Valdar, 2002) have been proposed and a recent comparison work (Tan et al., 2015) suggests that Noisy and trimAl are the best approaches. However, when comparing the execution time, trimAl is faster (seconds compared to hours required by Noisy) and for this reason is the one we decided to use as default in PhyloPhlAn 2.

Other approaches for shortening the MSA are the removal of gappy regions (determined based on gaps distribution) or the removal of single gaps, the removal of the conserved regions with a limited phylogenetic signal, and the removal of extremely variable positions, probably representing lowly-conserved or noisy regions. Several different scoring measures have been proposed for evaluating MSA quality. In PhyloPhlAn 2 we exploit a scoring measure to retain a limited number of phylogenetically-relevant positions.

In PhyloPhlAn 2 it is possible to use a combination of the above approaches. For instance, as default settings for the reconstruction of very-large phylogenies (parameters: “--diversity high --fast”), PhyloPhlAn 2 applies trimAl ((Capella-Gutiérrez et al., 2009) with “-gappyout” param) for the removal of gappy regions, then the removal of conserved regions by considering all position that do not vary more than 95% (param “--not_variant_threshold 0.95”, set automatically by the previous params). Finally, to avoid wrongly placed genomes in the phylogeny due to too many missing positions, a final check on the aligned sequence for each genome removes the regions that are still more than 65% gaps (“--fragmentary_threshold 0.65” param, set automatically by the first two params).

Moreover, in PhyloPhlAn 2 we implemented three different scoring functions (“trident”, “muscle”, and “random”) that assign a phylogenetic score to each position in the MSA and, in combination with a subsample function, retain only a certain number of positions. The “random” function simply assigns a random number to each column of the MSA. The “trident” score, as proposed in (Valdar, 2002), is a weighted combination of three different measures: symbol diversity, stereochemical diversity, and gaps frequency. Gaps frequency counts the number of gaps in each column. Symbol diversity measures the entropy of the column by weighting the frequency of the different occurring symbols. Stereochemical diversity, instead, is a score based on a substitution matrix. In PhyloPhlAn 2 we provide four substitution matrices: MIQS (Yamada and Tomii, 2014), PFASUM60 (Keul et al., 2017), VTML200 (Edgar, 2009), and VTML240 as implemented in MUSCLE (Edgar, 2004), along with scripts for generating custom ones. The “muscle” scoring function re-implements the scoring function available in MUSCLE ((Edgar, 2004), using the “-scorefile” param). After having scored each position of each MSA, PhyloPhlAn uses one of the implemented subsample functions: "phylophlan", “onethousand", “sevenhundred", “fivehundred", “threehundred", “onehundred", “fifty", “twentyfive", “tenpercent", “twentyfivepercent", and "fiftypercent", to retain only a certain number of positions. While it is clear how many positions will be retained for each MSA using one of the following subsamples functions: “onethousand", “sevenhundred", “fivehundred", “threehundred", “onehundred", “fifty", “twentyfive", “tenpercent", “twentyfivepercent", and "fiftypercent", the “phylophlan” one is instead based on the formula in (Segata et al., 2013) and it is specific for the set of 400 universal markers proposed in the same work.

3.4.7 Phylogeny post-processing

PhyloPhlAn 2 provides as output the reconstructed phylogeny, its MSA, and if specified the estimated mutation rates. The phylogenetic tree can be further use in downstream analysis. One evaluates the phylogenetic distances distribution to detect if a clade is an outlier with respect to the other clades in the phylogeny. Tools like TreeShrink (Mai and Mirarab, 2017) directly analyze the phylogeny to identify outlier branches. Additionally, PhyloPhlAn 2 can estimate a sequence-based similarity measure based on the MSAs between all input genomes. This can be an alternative approach in detecting outlier clades in the phylogeny. Finally, the availability of the MSA and the reconstructed phylogenetic tree can be used for further phylogenetic analysis like bootstrapping.

3.4.8 Software and Data availability

PhyloPhlAn 2 is released open-source and available in Bitbucket at https://bitbucket.org/nsegata/phylophlan, currently, the new implementation is present in the "dev" branch of the repository. The Ethiopian cohort is not yet fully available, as right now only the raw sequencing data for five of the 50 Ethiopian metagenomes are available in NCBI-SRA under the BioProject PRJNA504891.

4. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling

In this chapter, I expand the phylogenetic framework introduced in **Chapter 3** to the task of identifying the presence of the same microbial strains across microbiome samples, i.e. strain tracking. This is an important task as it can open new venues for performing epidemiology and population genomics of microorganisms directly from metagenomics. In this chapter, the strain-tracking approach I developed was applied to the problem of inferring vertical transmission of microbial organisms from mothers to their infants during the first few months of life. Characterizing the vertical mother-to-infant transmission phenomenon is crucial in microbiome research to shed light on the dynamics of microbiome colonization and development in the infant. In the study presented in this chapter, we sampled five mother-infant couples, with two couples having a longitudinal sampling up to one year after birth. The final goal of the study was to demonstrate that the strain-level analysis of microbial species from shotgun metagenomic data is feasible and potentially informative. This goal was indeed achieved and this paper is referred to as the methodological foundation for more recent and larger cohorts vertical transmission studies (Cabral et al., 2017; Davenport et al., 2017; Ferretti et al., 2018; Korpela and de Vos, 2018; Miyoshi et al., 2017; Vatanen et al., 2018; Wampach et al., 2017, 2018; Ximenez and Torres, 2017; Yassour et al., 2018). In this chapter, I also dig into the biological problem and interacted closely with the colleagues responsible for sampling, DNA extraction, library preparation, and sequencing in order to consider any potential problem of false positive or false negatives and model the experimental noise in the analysis pipeline. Importantly, a novel - and still unique - addition of this work to the vertical microbiome transmission literature is the availability of metatranscriptomics data that allowed to investigate the functional expression of the members of the microbiome that were transmitted from the mother to the infant.

The chapter is based on the following article:

Asnicar F*, Manara S*, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, and Segata N (* equal contribution)

Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling

mSystems (2017)

Abstract

The gut microbiome becomes shaped in the first days of life and continues to increase its diversity during the first months. Links between the configuration of the infant gut microbiome and infant health are being shown, but a comprehensive strain-level assessment of microbes vertically transmitted from mother to infant is still missing. We collected fecal and breast milk samples from multiple mother-infant pairs during the first year of life and applied shotgun metagenomic sequencing followed by computational strain-level profiling. We observed that several specific strains, including those of *Bifidobacterium bifidum*, *Coprococcus comes*, and *Ruminococcus bromii*, were present in samples from the same mother-infant pair, while being clearly distinct from those carried by other pairs, which is indicative of vertical transmission. We further applied metatranscriptomics to study the in vivo gene expression of vertically transmitted microbes and found that transmitted strains of

Bacteroides and *Bifidobacterium* species were transcriptionally active in the guts of both adult and infant. By combining longitudinal microbiome sampling and newly developed computational tools for strain-level microbiome analysis, we demonstrated that it is possible to track the vertical transmission of microbial strains from mother to infants and to characterize their transcriptional activity. Our work provides the foundation for larger-scale surveys to identify the routes of vertical microbial transmission and its influence on postinfancy microbiome development.

Importance

Early infant exposure is important in the acquisition and ultimate development of a healthy infant microbiome. There is increasing support for the idea that the maternal microbial reservoir is a key route of microbial transmission, and yet much is inferred from the observation of shared species in mother and infant. The presence of common species, *per se*, does not necessarily equate to vertical transmission, as species exhibit considerable strain heterogeneity. It is therefore imperative to assess whether shared microbes belong to the same genetic variant (i.e., strain) to support the hypothesis of vertical transmission. Here we demonstrate the potential of shotgun metagenomics and strain-level profiling to identify vertical transmission events. Combining these data with metatranscriptomics, we show that it is possible not only to identify and track the fate of microbes in the early infant microbiome but also to investigate the actively transcribing members of the community. These approaches will ultimately provide important insights into the acquisition, development, and community dynamics of the infant microbiome.

4.1 Introduction

The community of microorganisms that dwell in the human gut has been shown to play an integral role in human health (Clemente et al., 2012; HMP et al., 2012; Qin et al., 2010; Tamburini et al., 2016), facilitating, for instance, the harvesting of nutrients that would otherwise be inaccessible (Bäckhed et al., 2005), modulating the host metabolism and immune system (Palm et al., 2015), and preventing infections by occupying the ecological niches that could otherwise be exploited by pathogens (Stecher and Hardt, 2011). The essential role of the intestinal microbiome is probably best exemplified by the successful treatment of dysbiotic states, such as chronic life-threatening *Clostridium difficile* infections, using microbiome transplantation therapies (Britton and Young, 2014; Fuentes et al., 2014; Khoruts et al., 2010).

The gut microbiome is a dynamic community shaped by multiple factors throughout an individual's life, possibly including prebirth microbial exposure. The early development of the infant microbiome has been proposed to be particularly crucial for longer-term health (Bäckhed et al., 2015; Palmer et al., 2007; Yatsunenko et al., 2012), and a few studies have investigated the factors that are important in defining its early structure (Azad et al., 2013; Dominguez-Bello et al., 2010; La Rosa et al., 2014; Milani et al., 2015). In particular, gestational age at birth (La Rosa et al., 2014), mode of delivery (Azad et al., 2013; Dominguez-Bello et al., 2010), and early antibiotic treatments (Greenwood et al., 2014) have all been shown to influence the gut microbial composition in the short term and the pace of its development in the longer term.

Vertical transmission of bacteria from the body and breast milk of the mother to her infant has gained attention as an important source of microbial colonization (Aagaard et al., 2012; Cabrera-Rubio et al., 2012; Dominguez-Bello et al., 2010; Hunt et al., 2011) in addition to the microbial organisms obtained from the wider environment (Flores et al., 2014; Song et al., 2013), including the delivery room (Shin et al., 2015). Results from early cultivation-based and cultivation-free methods (16S rRNA community profiling and a single metagenomic study) have indeed suggested that the mother could transfer microbes to the infant by breastfeeding (Jost et al., 2014) and that a vaginal delivery has the potential of seeding the infant gut with members of the mother's vaginal community (Bäckhed et al., 2015; Biasucci et al., 2010; Dominguez-Bello et al., 2010, 2016) that would not be available via caesarean section. However, a more in-depth analysis is required to elucidate the role of vertical transmission in the acquisition and development of the infant gut microbiome.

Current knowledge of the vertical transmission of microbes from mothers to infants has hitherto focused on the cultivable fraction of the community (Makino et al., 2011) or lacked strain-level resolution (Bäckhed et al., 2015). Many microbial species are common among unrelated individuals (Lozupone et al., 2012); therefore, in instances where a species is identified in both mother and infant (Palmer et al., 2007; Turnbaugh et al., 2009), it remains inconclusive if this is due to vertical transmission. Strain-level analysis has shown that different individuals are associated with different strains of common species (Schloissnig et al., 2013; Scholz et al., 2016), and it is therefore crucial to profile microbes at the strain level to ascertain the most probable route of transfer. This has been performed only for specific microbes by cultivation methods (Makino et al., 2011; Milani et al., 2015), but many vertically transmitted microorganisms remain hard to cultivate (Milani et al., 2015); thus, the true extent of microbial transmission remains unknown. A further crucial aspect, still largely unexplored, is the fate of vertically acquired strains: if they are transcriptionally active rather than merely transient, that may suggest possible colonization of the infant intestine. Although studies have described the transcriptional activity of intestinal microbes under different conditions (Bao et al., 2015; Gosalbes et al., 2012; Maurice et al., 2013; Turnbaugh et al., 2010), no studies have applied metatranscriptomics to characterize the activity of vertically transmitted microbes *in vivo*.

In this work, we present and validate a shotgun metagenomic pipeline to track mother-to-infant vertical transmission of microbes by applying strain-level profiling to members of the mother and infant microbiomes. Moreover, we assessed the transcriptional activity of vertically transmitted microbes to elucidate if transferred strains are not only present but also transcriptionally active in the infant gut.

4.2 Results and Discussion

We analyzed the vertical transmission of microbes from mother to infant by enrolling 5 mother-infant pairs and collecting fecal samples and breast milk (see Materials and Methods) when each infant was 3 months of age (time point 1). Two mother-infant pairs (pair 4 and pair 5) were additionally sampled at 10 months postbirth (time point 2), and one pair (pair 5) was sampled at 16 months postbirth (time point 3; see **Fig. S1** in the supplemental material). We applied shotgun metagenomic sequencing to all 24 microbiome samples (8 mother fecal samples, 8 infant fecal samples, and 8 milk samples), generating 1.2 G reads (average, 39.6 M reads/sample; standard deviation [SD], 28.7 M reads/sample) (see **Table S1** in the supplemental material). Metatranscriptomics (average, 90.55 M reads/sample; SD,

46.86 M reads/sample) was also applied on fecal samples of two pairs (pair 4 at time point 2 and pair 5 at time point 3) to investigate the differential expression profiles of the bacterial strains in the gut of mothers and their infants.

Shared mother-infant microbial species. In our cohort, the infant intestinal microbiome was dominated by *Escherichia coli* and *Bifidobacterium* spp., such as *B. longum*, *B. breve*, and *B. bifidum* (**Fig. 1A** and **S2**). These species in some cases reached abundances higher than 75% (e.g., *E. coli* at 85.2% in infant pair 3 at time point 1 and *B. breve* at 78.8% in infant pair 5 at time point 1), which is consistent with previous observations (Koenig et al., 2011; Kurokawa et al., 2007; Yatsunenko et al., 2012). As expected, the intestines of the mothers had a greater microbial diversity than those of the infants, with high abundances of *Prevotella copri*, *Clostridiales* (e.g., *Coprococcus* spp. and *Faecalibacterium prausnitzii*), and *Bacteroidales* (e.g., *Parabacteroides merdae* and *Alistipes putredinis*). Interestingly, the postweaning microbiome of infant of pair 5 (time point 3, 16 months postbirth) had already shifted toward a more "mother-like" composition (**Fig. 1B**), with an increase in diversity and the appearance of *Parabacteroides merdae*, *Coprococcus* spp., and *Faecalibacterium prausnitzii* (Koenig et al., 2011; Palmer et al., 2007). Nevertheless, this 16-month-old infant still retained some infant microbiome signatures, such as a high abundance of bifidobacteria that were present at only low levels in the mothers' samples (**Fig. 1A** and **C**).

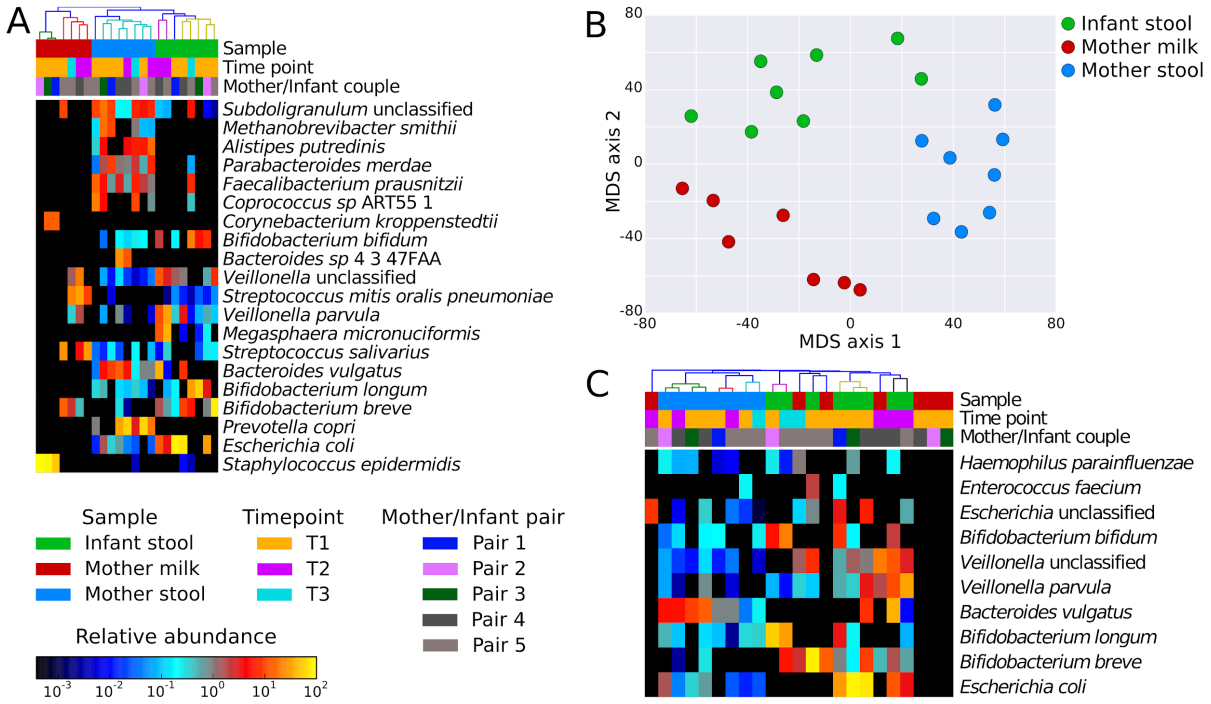


Fig 1. Microbial composition of mother and infant samples and shared bacteria within mother-infant pairs. (A) Quantitative microbial taxonomic composition of the metagenomic samples from milk and fecal samples of mothers and infants as estimated by MetaPhlAn2 analysis (Truong et al., 2015) (only the 20 most abundant species are indicated). Milk samples present low microbial richness compared to fecal samples. (B) Ordination plot of microbiome composition showing clustering of the three different sample types: mother feces, infant feces, and breast milk samples. The two infant samples close to the cluster of mother feces and in between the clusters of mothers and infants are from later time points, denoting the convergence of the infant microbiome toward an adult-like one. (C) The abundances of the 10 microbial species detected (>0.1% abundance) in at least one infant

and the respective mother (shared species have been identified on the basis of samples from time point 1 [T1] only).

We extracted and successfully sequenced microbial DNA from 7 of 8 milk samples. Microbial profiling of milk samples was hindered by a high abundance of interfering molecules (proteins, fats, proteases—e.g., plasmin—and calcium ions) (Bickley et al., 1996; Cremonesi et al., 2006; Schrader et al., 2012) that affected the efficiency of the extraction and amplification steps. Even so, we obtained an average of 3.08 Gb (SD, 1.5 Gb) per sample, of which 26 Mb (SD, 56 Mb) were from nonhuman reads (a level higher than that seen in the only other metagenomic study) (Ward et al., 2013) (see **Table S1**).

Milk samples had limited microbial diversity at the first sampling time (time point 1, 3 months postbirth) and included skin-associated bacteria such as *Corynebacterium kroppenstedtii* and *Staphylococcus epidermidis*. Cutaneous taxa, however, were observed in only low abundances in the gut microbiome of infants, confirming that skin microbes are not colonizers of the human gut (**Fig. 1A**). At later time points, the milk samples were enriched in *B. breve* and in bacteria usually found in the oral cavity, such as *Streptococcus* and *Veillonella* spp. The presence of oral taxa in milk has been previously observed by 16S rRNA sequencing (Cabrera-Rubio et al., 2012; Dominguez-Bello et al., 2010; Hunt et al., 2011; Jost et al., 2014) and shotgun metagenomics (Ward et al., 2013). This could be caused by retrograde flux into the mammary gland during breastfeeding (Ramsay et al., 2004) whereby cutaneous microbes of the breast and from the infant oral cavity are transmitted to the breast glands (Jeurink et al., 2013). However, this remains a hypothesis because no oral samples were collected in this study. These observations are summarized in the ordination analysis (**Fig. 1B**), in which the different samples (infant feces, mother feces, and milk) clustered by type, with weaning representing a key factor in the shift from an infant to an adult-like microbiome structure (Costello et al., 2012; Koenig et al., 2011; Palmer et al., 2007).

Comparing the species present in both the mother and infant pairs (**Fig. 1C**), we observed that many shared species (e.g., *Escherichia*, *Bifidobacterium*, and *Veillonella* spp.) occurred at a much higher abundance in the infant than in the mother, possibly due to the lower level of species diversity and therefore to competition in the gut. *Bacteroides vulgatus* was found at relatively high abundance (average, 16.3%; SD, 13%) in both the infant and the mother of pair 4 at both time point 1 and time point 2. The presence of shared species in mother-infant pairs observed here and elsewhere (Dominguez-Bello et al., 2010; Faith et al., 2013; Jost et al., 2014; La Rosa et al., 2014; Milani et al., 2015) confirms that mothers are a potential reservoir of microbes vertically transmissible to infants, but it remains unproven whether the same strain is transmitted to the infant from the mother or if an alternative transmission route is involved.

Strains shared between mothers and infants are indicative of vertical transmission. While different individuals have a core of shared microbial species, it has been shown that these common species consist of distinct strains (Schloissnig et al., 2013; Scholz et al., 2016). To analyze microbial transmission, it is therefore crucial to assess whether a mother and her infant harbor the same strain. To this end, we further analyzed the metagenomic samples at a finer strain-level resolution. This was achieved by applying a recent strain-specific pangenome-based method called PanPhlAn (Scholz et al., 2016), as well as a

genetics-based method called StrainPhlAn (D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, and N. Segata, submitted for publication) (see Materials and Methods), which identifies single-nucleotide variants (SNVs) in species-specific marker genes.

Using the SNV-based analysis, we observed considerable strain-level heterogeneity in the species present in the intestines of the mothers also with respect to available reference genomes (**Fig. 2**; see also **Fig. S3** in the supplemental material). This heterogeneity was not observed within the mother-infant pairings, as in the case of *Bifidobacterium* spp., *Ruminococcus bromii*, and *Coprococcus comes*. The infant of pair 4 at time point 2, for example, harbored a strain of *B. bifidum* that matched his mother's at 99.96% sequence identity and yet was clearly distinct from the *B. bifidum* strains of other infants in the cohort (**Fig. 2A**), which differed by at least 0.6% of the nucleotides. The observation that the *B. bifidum* strains from the mother and the infant of pair 4 were too similar to be consistent with the observed strain-level variation across subjects in the cohorts was highly statistically significant (*P* value, 4.7e−40) (see **Fig. S4**). This was also true for the *C. comes* (*P* value, 1.9e−3) (99.87% intrapair similarity and 1.6% and 1.61% divergence compared to the closest strain and the average value, respectively) (**Fig. 2B**) and *R. bromii* (*P* value, 4.9e−8) (99.93% similarity and 1.53% and 2.63% diversity—same as described above) (**Fig. 2C**) strains that were shared by pair 5. Mother-infant sharing of the same strain was also confirmed by strain-level pangenome analysis (Scholz et al., 2016) that showed that the strains from the same pair carried the same unique gene repertoire (see **Fig. S5**). It is accepted that, while the possibility of independent acquisition of strains from a shared environmental source cannot be excluded, the finding that mother-infant pairs have shared strains represents strong evidence of vertical microbiome transmission. On average, we could reconstruct and observe vertical transmission from mother to infant for 14% of the species found to be shared within mother and infant pairings.

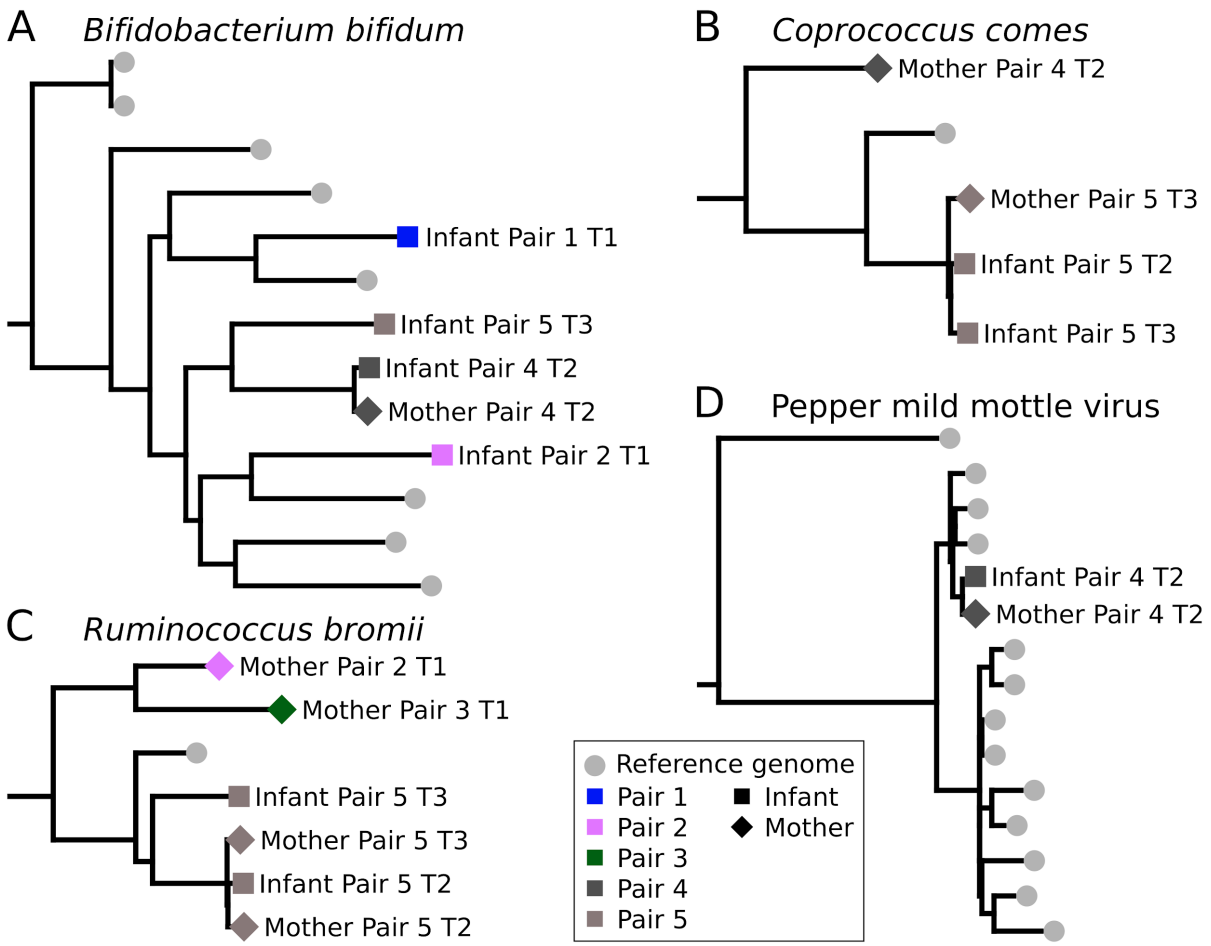


Fig 2. Strain-level phylogenetic trees for microbes present in both the mother and infant. Phylogenetic trees were built by the StrainPhlAn method using species-specific markers confirming the presence of the same strain in the mother and infant intestinal microbiomes, thus suggesting vertical transmission. Available reference genomes were included in the phylogenetic trees. Here we report three bacterial species, namely, (A) *Bifidobacterium bifidum*, (B) *Coprococcus comes*, and (C) *Ruminococcus bromii*, and the most abundant viral species found in pair 4, (D) pepper mild mottle virus. Other species-specific phylogenetic trees (*B. adolescentis*, *B. breve*, and *B. longum*) are reported in **Fig. S3**.

Strain transmission does not, however, exclude later replacement of the vertically acquired organisms, as we highlighted by looking at the postweaning time point in our cohort (pair 5 at time point 3) which harbored the highest number of shared species, with 70.4% present in the infant and mother (at a relative abundance of >0.1%, according to the MetaPhlAn2 profiles). A proportion (11%) of these common species were shown to be the same strain (**Fig. 2**; see also **Fig. S3** in the supplemental material), according to both PanPhlAn and StrainPhlAn analyses (see, for example, the data from *B. adolescentis* and *C. comes*) (**Fig. 2**; see also **Fig. S3** and **S5**). However, some strains that were shared at earlier time points were replaced at time point 3. Of note, the *R. bromii* strain found in an infant at time point 3 was different from that found at time point 2, and both strains were distinct from the strain observed in the mother at both time points (**Fig. 2C**). This was also observed for the latter infant time point for *B. breve* (see **Fig. S3B**) and *B. longum* (see **Fig. S3C**). Although it is not possible to generalize these results because of the small sample size, these replacement

events suggest that originally acquired maternal strains can subsequently be replaced (Morowitz et al., 2011; Sharon et al., 2013).

We then extended our analysis to the viral organisms detectable from metagenomes and metatranscriptomes, as viruses have the potential to be vertically transmitted also. The DNA viruses identified from our metagenome samples largely consisted of bacteriophages of the *Caudovirales* order, a common order of tailed bacteriophages found in the intestine (Ogilvie and Jones, 2015; Tamburini et al., 2016). We identified *Enterobacter* and *Shigella* phages as the most prevalent phages among the tested samples, in agreement with the high prevalence of members of the *Enterobacteriaceae* family and particularly of members of the *Escherichia* genus (see **Fig. 1A** and **Table S3**). We also identified crAssphage at high breadth of coverage (Dutilh et al., 2014) and provided further evidence for the hypothesis that the *Bacteroides* genus is the host for this virus (Dutilh et al., 2014), as the microbiome of crAssphage-positive mothers was enriched in *B. vulgatus* (see **Fig. 1A** and **Table S3**). However, the low breadth of coverage for many of the DNA viruses made it difficult to identify pair-specific phage variants (see **Table S3**). Analysis of the RNA viruses from the metatranscriptomic samples identified instead the presence of an abundant pepper mild mottle virus (PMMoV), a single-stranded positive-sense RNA virus of the genus *Tobamovirus*, in all of the four metatranscriptomes from pairs 4 and 5. Surprisingly, transcripts from the PMMoV were found in greater abundance than all the other microbial transcripts found for the mother of pair 4. PMMoV has already been reported in the gut microbiome (Reyes et al., 2010; Victoria et al., 2009; Zhang et al., 2006), and other related viruses of the same family have been shown to be able to enter and persist in eukaryotic cells (Balique et al., 2013; de Medeiros et al., 2005). The high abundance of PMMoV in mother-infant pair 4 allowed us to reconstruct its full genome (99.9%) and to perform a phylogenetic analysis demonstrating that the mother and the infant shared identical PMMoV strains, which were clearly distinct from the PMMoV reference genomes (27 SNVs in total; **Fig. 2D**). Although the coverage was lower, the same evidence of a shared PMMoV strain was observed within pair 5. The analysis of PMMoV polymorphisms within each sample also suggests the coexistence of different PMMoV haplotypes in the same host (**Fig. S6**). Although vertical transmission of RNA viruses and PMMoV specifically would be intriguing, because of the age and dietary habits of the infants (see **Table S1**) this finding could be related to the exposure to a common food source (Colson et al., 2010). Our analysis of the virome characterized directly from shotgun metagenomics thus highlighted that viruses can be tracked across mother-infant microbiomes also and that experimental virome enrichment protocols (Reyes et al., 2012; Thurber et al., 2009) have the potential to provide an even clearer snapshot of viral vertical transmission.

Differences in the overall levels of functional potential and expression in mothers and infants. The physiology of the mammary gland (milk) as well as the adult and infant intestine is reflected by niche-specific microbial communities as reported above and in previous studies (Azad et al., 2013; Bäckhed et al., 2015; Cabrera-Rubio et al., 2012; Costello et al., 2012; Hunt et al., 2011; Jeurink et al., 2013; Palmer et al., 2007). To characterize the overall functional potential of the microbial communities inhabiting these niches, we complemented the taxonomic analysis above by employing HUMAnN2 (see Materials and Methods). As expected, there was considerable overlap in the functionality of the gut microbiomes of the mothers and infants (**Fig. 3A**), with 87% of pathways present in mother and infant, 50% of which were significantly different in abundance (at an alpha value of 0.05). Nevertheless,

there were notable differences. For instance, the microbiomes of the infants showed a higher potential for utilization of intestinal mucin as a carbon source (*P* value, 0.016) and for folate biosynthesis (*P* value, 1.8e−6) while displaying a lower potential for starch degradation (*P* value, 9.8e−6), consistent with previous observations (LeBlanc et al., 2013; Marcobal et al., 2011; Tailford et al., 2015; Turroni et al., 2011a; Yatsunenko et al., 2012). Mucin utilization, specifically by infant gut microbial communities, is reflective of the higher abundance of mucin-degrading bifidobacteria observed from the taxonomic analyses described above (Marcobal et al., 2011; Tailford et al., 2015; Turroni et al., 2011a; Yatsunenko et al., 2012), whereas increased folate biosynthesis (LeBlanc et al., 2013; Marcobal et al., 2011; Tailford et al., 2015; Yatsunenko et al., 2012) and decreased starch degradation (Bäckhed et al., 2005) have been purported to represent responses to the limited dietary intake in infants compared to adults. Interestingly, the intestinal samples from the postweaning infant of pair 5 (16 months postbirth) clustered together with the adults' intestinal samples (**Fig. 3B**), suggesting that the shift toward an adult-like microbiome observed in the taxonomic profiling (**Fig. 1B**) is also reflected by or is a consequence of a change in community functioning. Among the most prevalent pathways in the milk microbiomes that we observed were those involved in galactose and lactose degradation (Flint et al., 2012), as well as in biosynthesis of aromatic compounds (**Fig. S7A**). This was specifically true for production of chorismate, a key intermediate for the biosynthesis of essential amino acids and vitamins found in milk (LeBlanc et al., 2013) (**Fig. 3C** and **S7A**).

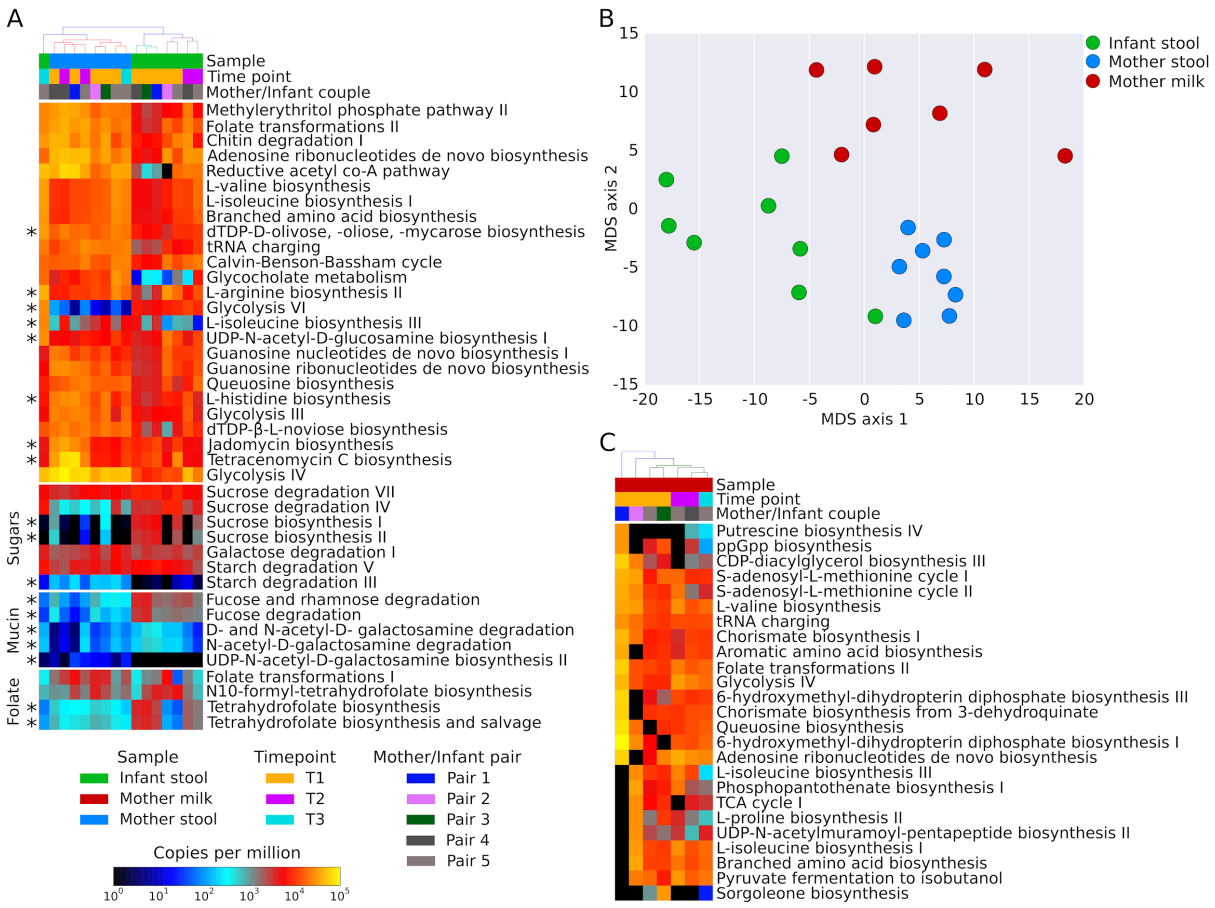


Fig 3. Functional potential analyses. (A) HUMAnN2 heat map reporting the 25 most abundant pathways in the fecal samples of mothers and infants. Specific pathways of interest (sugars, mucin, and folate metabolism) are added at the bottom. The asterisk (*)

near the heat map highlights statistically significant pathways. (B) Multidimensional scaling (MDS) result from functional potential profiles, showing the differences between fecal samples of mothers and infants and milk samples. In particular, the infant feces point in the mother feces cluster corresponds to time point 3 of pair 5, showing a shift from the infant microbiome toward an adult-like microbiome. (C) HUMAnN2 results for the 25 most abundant pathways found only in the milk samples. TCA, tricarboxylic acid.

To further evaluate the functional capacity of the gut-associated microbiomes and analyze the *in vivo* transcription, we performed metatranscriptomics analyses of the feces of two mother-infant pairs (see Materials and Methods). HUMAnN2 was used to identify differences in the transcriptional levels of pathways in the gut of the mothers and infants. The most notable global difference was that fermentation pathways were highly transcribed in the mother compared to that of the infant. This reflects the transition of the gut from an aerobic to an anaerobic state and the associated shift from facultative anaerobes to obligate anaerobes over the first few months of life (Houghteling and Walker, 2015; Turroni et al., 2012). The same is true for pathways involved in starch degradation, which were not only poorly represented in the metagenomes but also negligibly expressed in the infants' transcriptomes. What is evident is that the transcriptional patterns for different members differed considerably, as illustrated for pair 4 and pair 5 (**Fig. 4A** and **S7B**, respectively). For example, we observed in the infant of pair 4 that *B. vulgatus* was more transcriptionally active (average of 2.7 [SD, 2.5] normalized transcript abundance [NTA]; see Materials and Methods) than both *E. coli* (245-fold change [average, 0.4 SD and 0.6 NTA]) and *Bifidobacterium* spp. (6.6-fold change [average, 0.01 SD and 0.01 NTA]). Although these differences were statistically significant (*P* values were lower than 1e−50 in both cases), their physiological significance remains unclear.

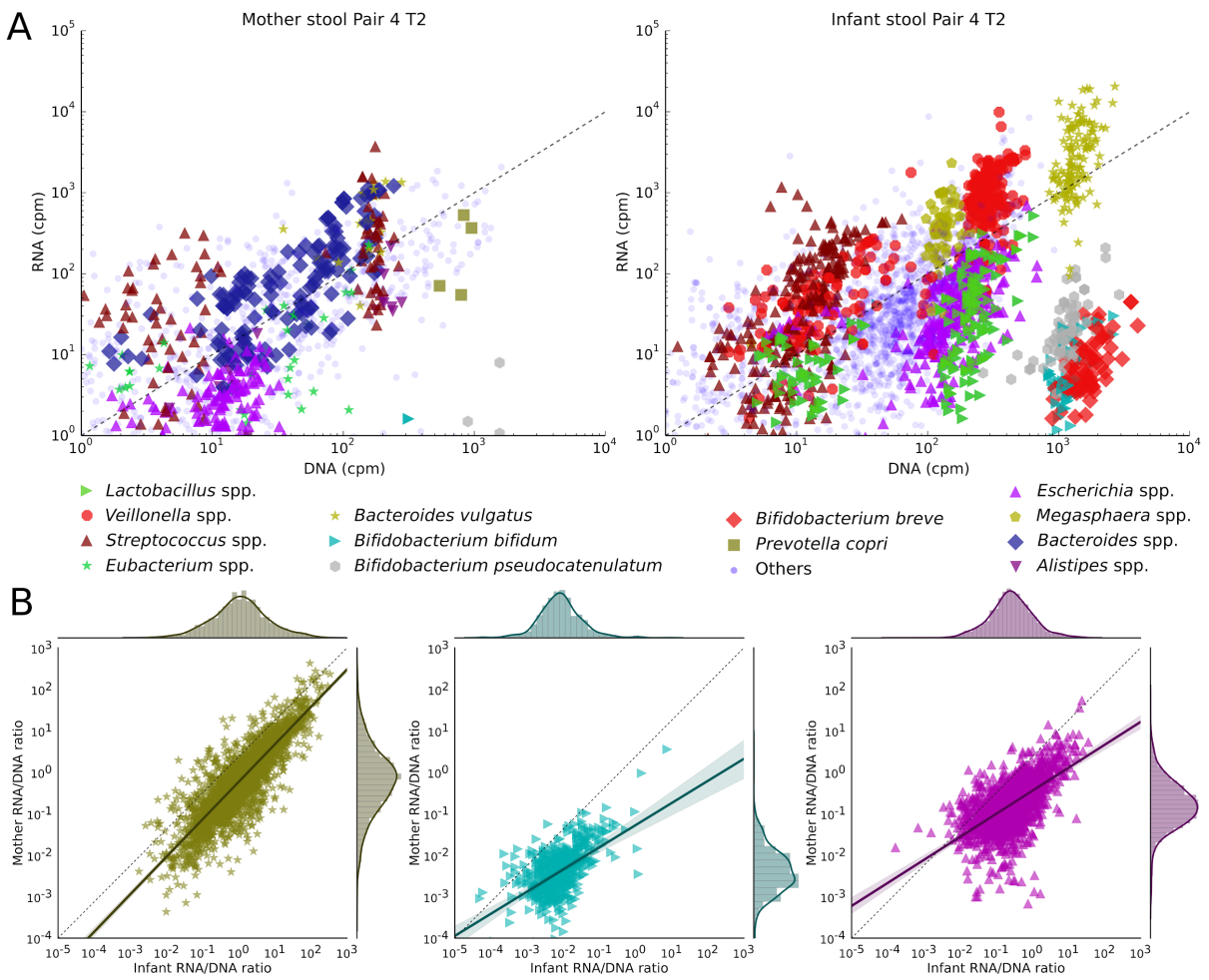


Fig 4. Transcription levels of metabolic pathways and genes in mother and infant pair 4 at time point 2. (A) Scatterplots showing the transcription rates of metabolic pathways of shared and nonshared species and genera of interest for both the mother and infant of pair 4 at time point 2. (B) Comparison between transcription rates of gene families in mother and infant gut microbiomes.

Strain-specific transcriptional differences in mothers and infants. To further explore the transcriptional activity of the intestinal microbiomes and, more specifically, to ascertain which individual microbial members are transcriptionally active in the gut, we employed the strain-specific metatranscriptomic approach implemented in PanPhlAn (Scholz et al., 2016) (see Materials and Methods). Of particular interest is the transcriptional activity of the shared mother-infant strains that, based on our strain-level analyses, are likely to have been vertically acquired by the infant by the maternal route. Such transcriptional analyses can clarify whether these transmitted strains were not only present in the infant gut but also functioning, therefore suggesting that the transmitted strains could have potentially colonized. For three transmitted species in pair 4 (*B. vulgatus*, *E. coli*, and *B. bifidum*), we show that they were active in both the mother intestine and the infant intestine (**Fig. 4B**). Of note is that *B. bifidum* was more active in the infant than in the mother (2.5-fold change; **Fig. 4B**), which was expected as this species is a known early colonizer of the infant gut (Koenig et al., 2011; Kurokawa et al., 2007; Yatsunenko et al., 2012). Interestingly, the *B. bifidum* strain of pair 5 showed the opposite behavior (**Fig. S7C**). We postulate that this was because the infant of pair 5 was of postweaning age (10% breast milk diet) compared to the infant of pair 4 (90% breast milk diet) and that the difference reflects the change in substrate

availability from breast milk to solid food, which might have a detrimental effect on the bifidobacterial population (Koenig et al., 2011; Turroni et al., 2011b, 2012). Moreover, in support of our metagenomics analyses indicating that the microbiome of infant of pair 5 was shifting toward a more adult-like structure (**Fig. 1B**), we observed high transcriptional activity for *R. bromii*, a species commonly associated with adults, which could be seen as a hallmark of this transition (Scott et al., 2015; Walker et al., 2011).

It is well established that metatranscriptomic profiling provides a more accurate account of the actual community functioning than metagenomics alone. Here we show that the combination of the two approaches affords the exploration of which members not only are transmitted but also are actively participating in the community and therefore offers a more detailed account of the microbial community dynamics.

Conclusions. Human-associated microbiomes are complex and dynamic communities that are continuously interacting with the host and are under the influence of environmental sources of microbial diversity. Identifying and understanding the transmission from these external sources are crucial to understanding how the infant gut is colonized and ultimately develops an adult-like composition. However, detecting direct transmission is not a trivial task: many species are ubiquitous in host-associated environments and in the wider environment alike, and yet they comprise a myriad of different strains and phenotypic capabilities. Therefore, detection of microbial transmission events requires the ability to characterize microbes at the strain level. The epidemiological tracking of pathogens by cultivation-based isolate sequencing has proven successful (Gardy et al., 2011; Loman et al., 2013), but it relies on time-consuming protocols and can focus on only a limited number of species. In contrast, while there have been some examples of strain-level tracking from metagenomic data (Li et al., 2016; Loman et al., 2013), this remains challenging. In this study, we developed methods for identifying the vertical flow of microorganisms from mothers to their infants and showed that mothers are sources of microbes that might be important in the development of the infant gut microbiome.

We demonstrated that high-resolution computational methods applied to shotgun metagenomic and metatranscriptomic data enable the tracking of strains and strain-specific transcriptional patterns across mother-infant pairs. In our cohort of five mother-infant pairs, we detected several species with substantial genetic diversity between different pairs but identical genetic profiles in the mother and her infant, indicative of vertical transmission. These include some bifidobacteria typical of the infant gut (i.e., *B. longum*, *B. breve*, *B. bifidum*, and *B. adolescentis*) but also *Clostridiales* species usually found in the adult intestine (i.e., *R. bromii* and *C. comes*) and viral organisms. These results confirm that the infant receives a maternal microbial imprinting that might play an important role in the development of the gut microbiome in the first years of life.

The strain-level investigation of vertically transmitted microbes was followed by characterization of the transcriptional activity of the transmitted strains in the mother and infant environments. We found that the transcriptional patterns of strains shared within the single pairs were different between mother and infant, suggesting successful adaptation of maternally transmitted microbes to the infant gut.

Taking the results together, our work provides preliminary results and methodology to expand our knowledge of how microbial strains are transmitted across microbiomes.

Expanding the cohort size and considering other potential microbial sources of transmission, such as additional mother and infant body sites, as well as other family members (i.e., fathers and siblings) and environments (hospital and house surfaces), will likely shed light on the key determinants in early infant exposure and the seeding and development of the infant gut microbiome.

4.3 Materials and methods

Sample collection and storage. In total, five mother-infant pairs were enrolled. Fecal samples and breast milk were collected for all pairs at 3 months (time point 1); additional samples were collected for pair 4 and pair 5 at 10 months (time point 2) and for pair 5 only at 16 months (time point 3) (see **Table S1** and **Fig. S1** in the supplemental material). All aspects of recruitment and sample and data processing were approved by the local ethics committee. Fecal samples were collected from mothers and infants in sterile feces tubes (Sarstedt, Nümbrecht, Germany) and immediately stored at −20°C. In those cases where metatranscriptomics was applied, a fecal aliquot was removed prior to freezing the remaining feces. This aliquot was stored at 4°C, and the RNA was extracted within 2 h of sampling to preserve RNA integrity. Milk was expressed and collected midflow by mothers into 15-ml centrifuge tubes (VWR, Milan, Italy) and immediately stored at −20°C. Within 48 h of collection, all milk samples and feces samples were moved to storage at −80°C until processed.

Extraction of nucleic acids for metagenomic analysis. DNA was extracted from feces using a QIAamp DNA stool minikit (Qiagen, Netherlands). Milk DNA was extracted using a PowerFood microbial DNA isolation kit (Mo Bio, Inc., CA). Both procedures were performed according to the specifications of the manufacturers. Extracted DNA was purified using an Agencourt AMPure XP kit (Beckman Coulter, Inc., CA). Metagenomic libraries were constructed using a Nextera XT DNA library preparation kit (Illumina, CA, USA) according to manufacturer instructions and were sequenced on a HiSeq 2500 platform (Illumina, CA, USA) at an expected sequencing depth of 6 Gb/library.

Extraction of nucleic acids for metatranscriptomic analysis. Fecal samples for metatranscriptomic profiling were pretreated as described previously (Giannoukos et al., 2012). Briefly, 110 µl of lysis buffer (30 mM Tris·Cl, 1 mM EDTA [pH 8.0], 1.5 mg/ml of proteinase K, and 15 mg/ml of lysozyme) was added to 100 mg of feces and incubated at room temperature for 10 min. After pretreatment, samples were treated with 1,200 µl of Qiagen RLT Plus buffer (from an AllPrep DNA/RNA minikit [Qiagen, Netherlands]) containing 1% (vol) beta-mercaptoethanol and were transferred into 2-ml sterile screw-cap tubes (Starstedt, Germany) filled with 1 ml of zirconia-silica beads (BioSpec Products, OK, USA) (<0.1 mm in diameter). Tubes were placed on a Vortex-Genie 2 mixer with a 13000-V1-24 Vortex adapter (Mo Bio, Inc., CA) and shaken at maximum speed for 15 min. Lysed fecal samples were homogenized using QIAshredder spin columns (Qiagen, Netherlands), and homogenized sample lysates were then extracted with an AllPrep DNA/RNA minikit (Qiagen, Netherlands) according to the manufacturer's specifications. Extracted RNA and DNA were purified using Agencourt RNAClean XP and Agencourt AMPure XP (Beckman Coulter, Inc., CA) kits, respectively. Total RNA samples were subjected to rRNA depletion, and metatranscriptomic libraries were prepared using a ScriptSeq Complete Gold kit (epidemiology)-low input (Illumina, CA, USA). Metagenomic libraries were prepared with a

Nextera XT DNA library preparation kit (Illumina, CA, USA). All libraries were sequenced on a HiSeq 2500 platform (Illumina, CA, USA) at an expected depth of 6 Gb/library.

Sequencing data preprocessing. The metagenomes and metatranscriptomes were preprocessed by removing low-quality reads (mean quality value of less than 25), trimming low-quality positions (quality less than 15), and removing reads less than 90 nucleotides in length using FastqMcf (Aronesty, 2013). Further quality control steps involved the removal of human reads and the reads from the Illumina spike-in (bacteriophage Phi-X174) by mapping the reads against the corresponding genomes with Bowtie 2 (Langmead and Salzberg, 2012). Metatranscriptomes were additionally processed to remove rRNA by mapping the reads against 16S and 23S rRNA gene databases (SILVA_119.1_SSURef_Nr99_tax_silva and SILVA_119_LSURef_tax_silva (Quast et al., 2013)) and to remove contaminant adapters using trim_galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the following parameters: -q 0, –nextera, and –stringency 5. The milk sample of mother-infant pair 4 at time point 1 was discarded from further analyses because of the low number of microbial reads (less than 400,000 bp) obtained after the quality control steps (see **Table S1**). All metagenomes and metatranscriptomes have been deposited in and are available at the NCBI Sequence Read Archive.

Taxonomic and strain-level analysis. Taxonomic profiling was performed with MetaPhlAn2 (Truong et al., 2015) (with default parameters) on the 23 metagenomic samples that passed the quality control. MetaPhlAn2 uses clade-specific markers for taxonomically profiling shotgun metagenomic data and to quantify the clades present in the microbiome with species-level resolution.

Strain-level profiling was performed with PanPhlAn (Scholz et al., 2016) and a novel strain-level profiling method called StrainPhlAn (Truong et al., submitted). PanPhlAn is a pangenome-based approach that profiles the presence/absence pattern of species-specific genes in the metagenomes. The presence/absence profiles of the genes are then used to characterize the strain-specific gene repertoire of the members of the microbiome. PanPhlAn has been executed using the following parameters: --min_coverage 1, --left_max 1.70, and --right_min 0.30. PanPhlAn is available with supporting documentation at http://segatalab.cibio.unitn.it/tools/panphlan. StrainPhlAn is a complementary method based on analysis of SNVs that reconstructs the genomic sequence of species-specific markers. StrainPhlAn builds the strain-level phylogeny of microbial species by reconstructing the consensus marker sequences of the dominant strain for each detected species. The extracted consensus sequences are multiply aligned using MUSCLE version v3.8.1551 (Edgar, 2004) (default parameters), and the phylogeny is reconstructed using RAxML version 8.1.15 (Stamatakis, 2014) (parameters: -m GTRCAT and -p 1234). StrainPhlAn is available with supporting documentation at http://segatalab.cibio.unitn.it/tools/strainphlan.

Functional profiling from metagenomes and metatranscriptomes. The functional potential and transcriptomic analyses were performed with both HUMAnN2 (Franzosa et al., 2014) and PanPhlAn (Scholz et al., 2016). HUMAnN2 selects the most representative species from a metagenome and then builds a custom database of pathways and genes that is used as a mapping reference for the coupled metatranscriptomic sample to quantify transcript abundances. We computed the normalized transcript abundance (NTA), which we define as the average coverage of a genomic region in the metatranscriptomic versus that in the corresponding metagenomic sample normalized by the total number of reads in each

sample. PanPhlAn infers the expression of the strain-specific gene families by extracting them from the metagenome and matching them in the metatranscriptome. PanPhlAn has been executed using the following parameters: --rna_norm_percentile 90 and --rna_max_zeros 90.

Profiling of DNA and RNA viruses. We investigated the presence of viral and phage genomes by mapping the reads present in the metagenomes and metatranscriptomes against 7,194 viral genomes available in RefSeq (release 77). The average coverage and average sequencing depth were computed with SAMtools (Li et al., 2009) and BEDTools (Quinlan and Hall, 2010).

The presence of the pepper mild mottle virus (PMMoV) was confirmed by mapping the reference genome (NC_003630) against the metatranscriptomic samples from the mother and infant of pair 4 and pair 5. In the mother and infant of pair 4, 424,510 and 119 reads were mapped, respectively, while in the mother and infant of pair 5, 1,444 and 61 of the reads were mapped, respectively. In the two mothers (pair 4 and pair 5), the values for breadth of coverage were 0.99 and 0.98 and for average coverage were 6,562 and 22, respectively. In the two infants (pair 4 and pair 5), the values for breadth of coverage were 0.6 and 0.5 and for average coverage were 1.81 and 0.95, respectively. Additionally, we extracted the shared fractions of the PMMoV genome present in both the mother and the infant of pair 4, together with the same regions of all the available reference genomes (n = 13 [specifically, accession no. LC082100.1, KJ631123.1, AB550911.1, AY859497.1, KU312319.1, KP345899.1, NC_003630.1, M81413.1, KR108207.1, KR108206.1, AB276030.1, AB254821.1, and LC082099.1]). The resulting sequences were aligned using MUSCLE version v3.8.1551 (default parameters), and the resulting alignment was used to build a phylogenetic tree with RAxML v. 8.1.15 (parameters: -m GTRCAT and -p 1234).

Statistical analyses and data visualization. The taxonomic and functional heat maps were generated using hclust2 (parameters: --f_dist_f Euclidean, --s_dist_f braycurtis, and -l) available at https://bitbucket.org/nsegata/hclust2. The multidimensional scaling plots were computed with the sklearn Python package (Pedregosa et al., 2011).

Biomarker discovery (**Fig. S7A**) was performed by applying the linear discriminant analysis effect size (LEfSe) algorithm (Segata et al., 2011) (parameter: -l 3.0) on HUMAnN2 profiles. The two functional trees (**Fig. S7A**) have been automatically annotated with export2graphlan.py (GraPhlAn package) and displayed with GraPhlAn (Asnicar et al., 2015a) using default parameters.

Accession number(s). All metagenomes and metatranscriptomes have been deposited and are available at the NCBI Sequence Read Archive under BioProject accession number PRJNA339914.

Acknowledgements

We thank Marco Ventura and his group for performing the DNA extraction from the milk samples.

This work was supported by Fondazione CARITRO fellowship Rif.Int.2013.0239 to N.S. The work was also partially supported by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant

agreement no. PCIG13-GA-2013-618833 (N.S.), by startup funds from the Centre for Integrative Biology, University of Trento (N.S.), by MIUR Futuro in Ricerca RBFR13EWWI_001 (N.S.), by Leo Pharma Foundation (N.S.), and by Fondazione CARITRO fellowship Rif.int.2014.0325 (A.T.).

Supplementary Material

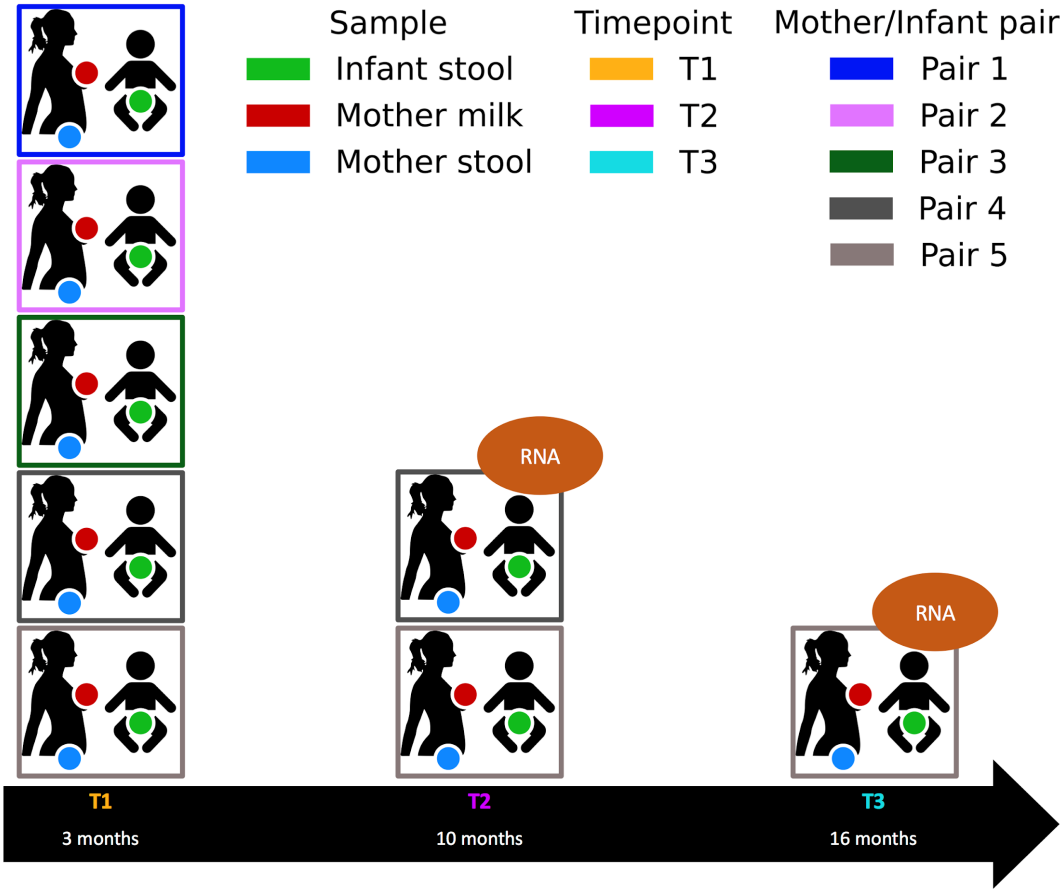


Fig S1. Study design. A schematic representation of the mother-infant pairs involved in the study, the sample types, and the time points considered is presented. Marked with the "RNA" label, the mother-infant pairs for which stool metatranscriptomes were produced are indicated.

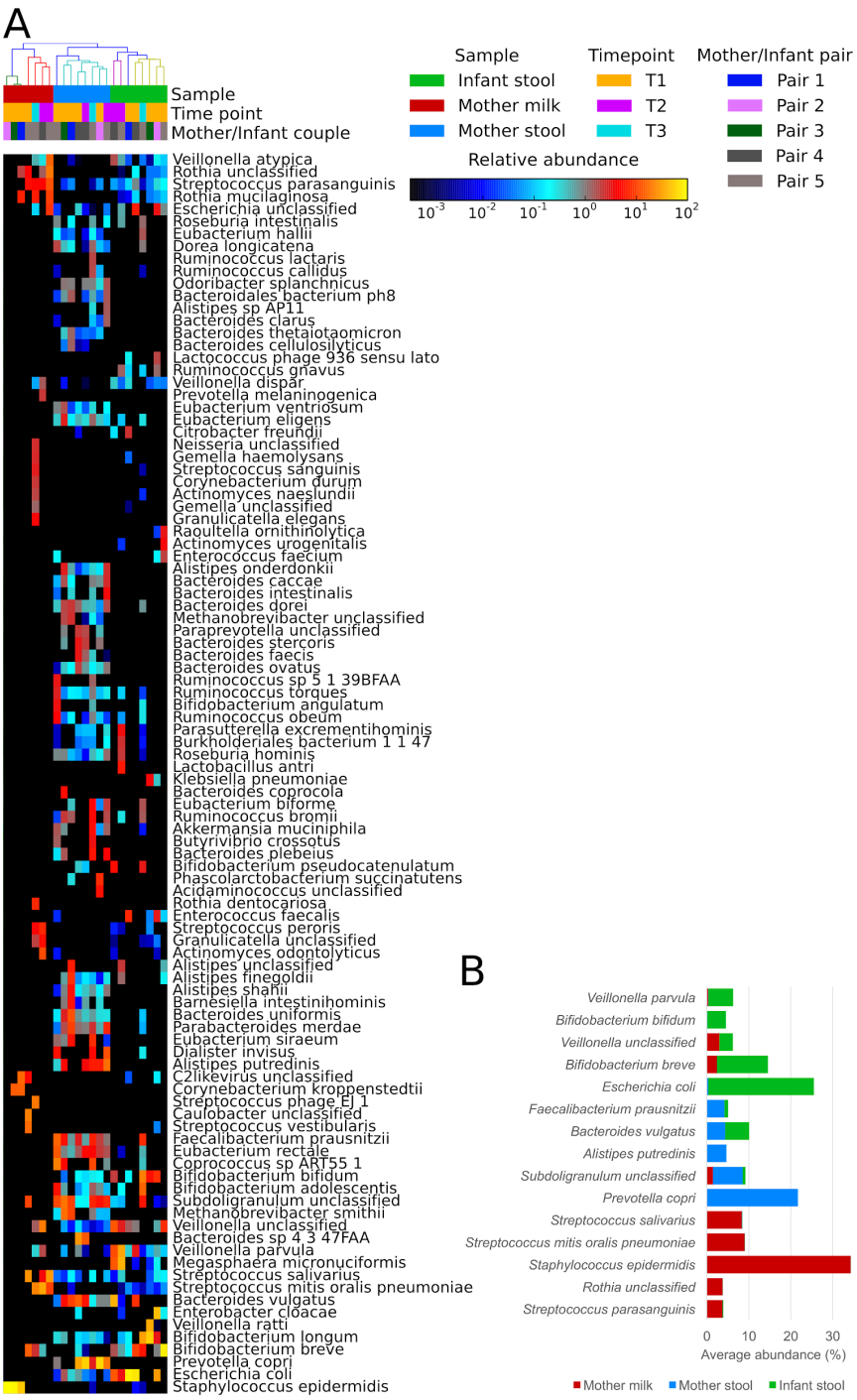


Fig S2. Extensive taxonomic profiling of the top 100 species from MetaPhlAn2 analysis and the five most highly represented niche-specific species. (A) The heat map shows differences in terms of species richness between mother, infant, and milk metagenomes. In particular, the milk samples have very low microbial diversity, especially at time point 1. The microbiomes of the mothers have instead higher diversity than both the milk microbiomes and the infant microbiomes. (B) We selected the five most highly represented species on average for each sample type (mother milk, mother stool, and infant stool) and plotted their average abundances in each niche. Each sample type is dominated by its five most highly represented species that are, in general, underrepresented in the other niches.

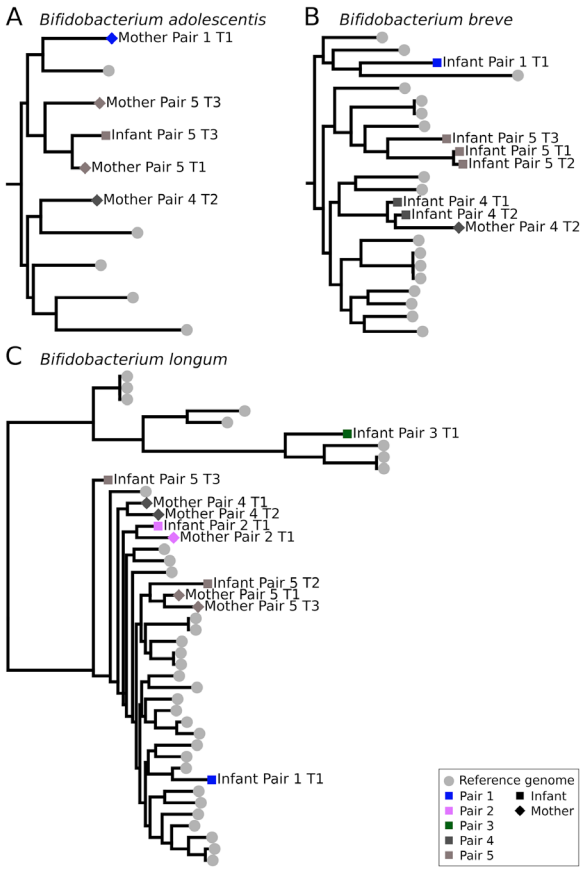


Fig S3. Strain-level analysis showing vertical transmission from mother to infant of bifidobacterium species. The phylogenetic trees were produced by applying StrainPhlAn for the following species: (A) *Bifidobacterium adolescentis*, (B) *Bifidobacterium breve*, and (C) *Bifidobacterium longum*. In each tree, a clade containing one (or more) samples of the mother and infant of the same pair is observed. This suggests that the strain is shared between mother and infant, hence suggesting vertical transmission.

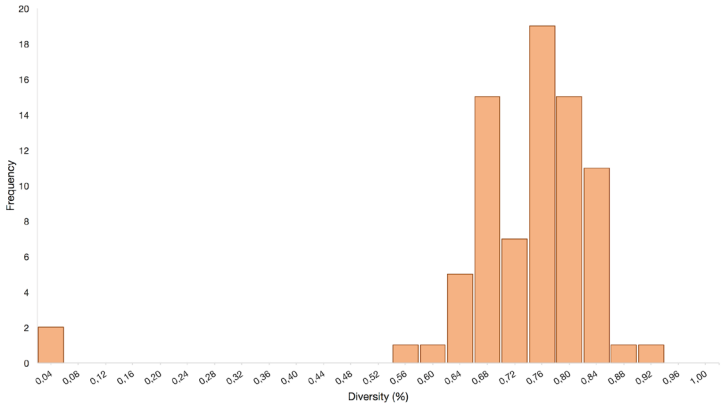


Fig S4. Distribution of SNV rates of *Bifidobacterium bifidum*. We computed the SNV rates of the strains of *B. bifidum* reconstructed with StrainPhlAn (the phylogenetic tree is presented in **Fig. 2A**). The two strains of the mother and the infant of pair 4 at time point 2 have an SNV rate of 0.04. The first bin has a frequency of two because it comprises not only the SNV rate of pair 4 at time point 2 but also the SNV rate of the two reference genomes reported in the upper part of the phylogenetic tree in **Fig. 2A**. The two reference genomes have an SNV rate of 0, meaning that they are identical.

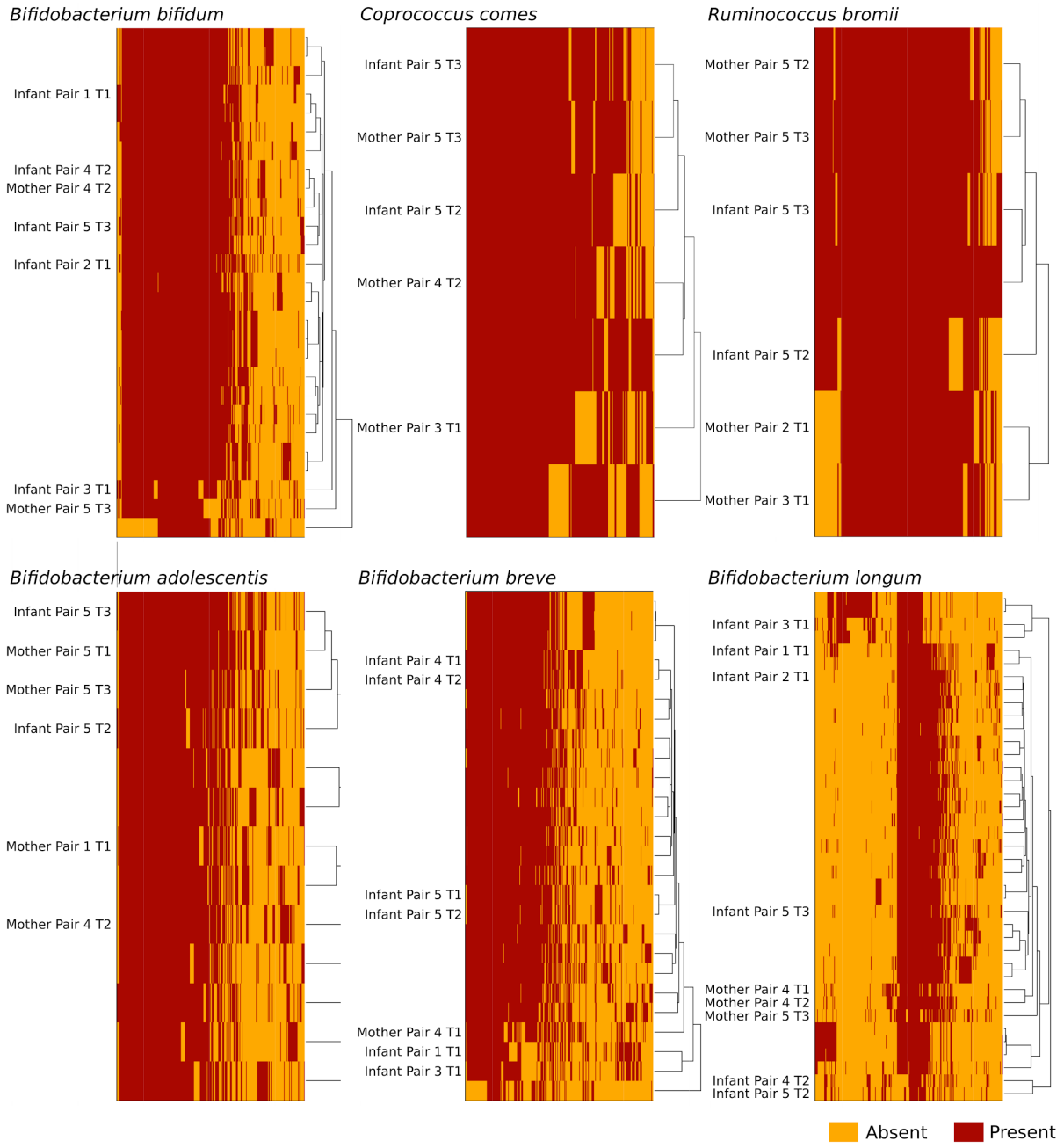


Fig S5. Strain-level analysis by applying PanPhlAn confirms vertical transmission. We applied PanPhlAn to validate the results obtained with StrainPhlAn (**Fig. 2** and **S3**). The pangenome-based strain-level analysis shows the presence and absence (in red and yellow, respectively) of the species-specific gene families of the following species: *B. bifidum*, *C. comes*, *R. bromii*, *B. adolescentis*, *B. breve*, and *B. longum*. Samples are clustered according to hierarchical clustering based on the Euclidean distance of the samples' pangenome profiles.

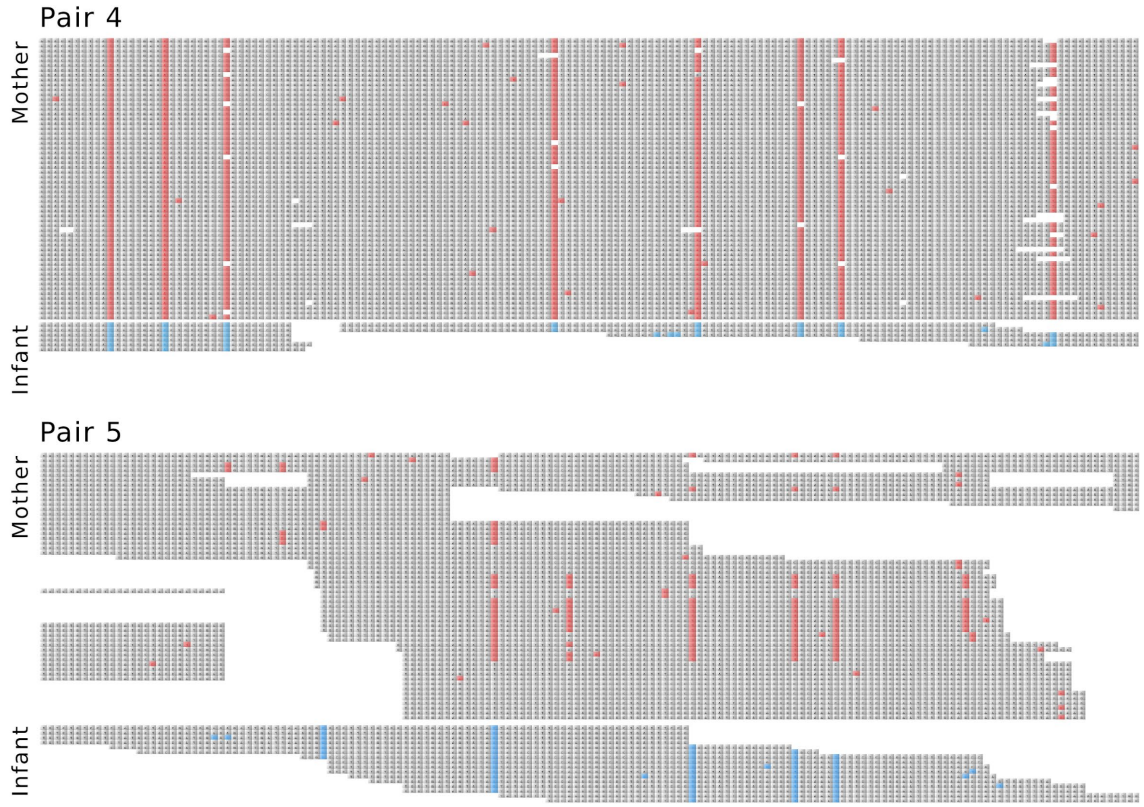


Fig S6. Read alignment of pepper mild mottle virus (PMMoV) for both pair 4 and pair 5. Alignments of mother and infant of both pair 4 and pair 5 against the PMMoV reference genome are presented, showing variations highlighted in red (mother) and blue (infant) for a window of 160 bp. Pair 4 data (from position 3216 to position 3376 in the PMMoV genome) show the agreement between the mother and infant variations, suggesting that they share the same strain of the PMMoV. Pair 5 data (from position 4450 to position 4610 in the PMMoV genome) show the presence of more than one viral strain in the mother. Variations in the infant data are coherent with data from the mother, with the former harboring only a subset of the mother's strains.

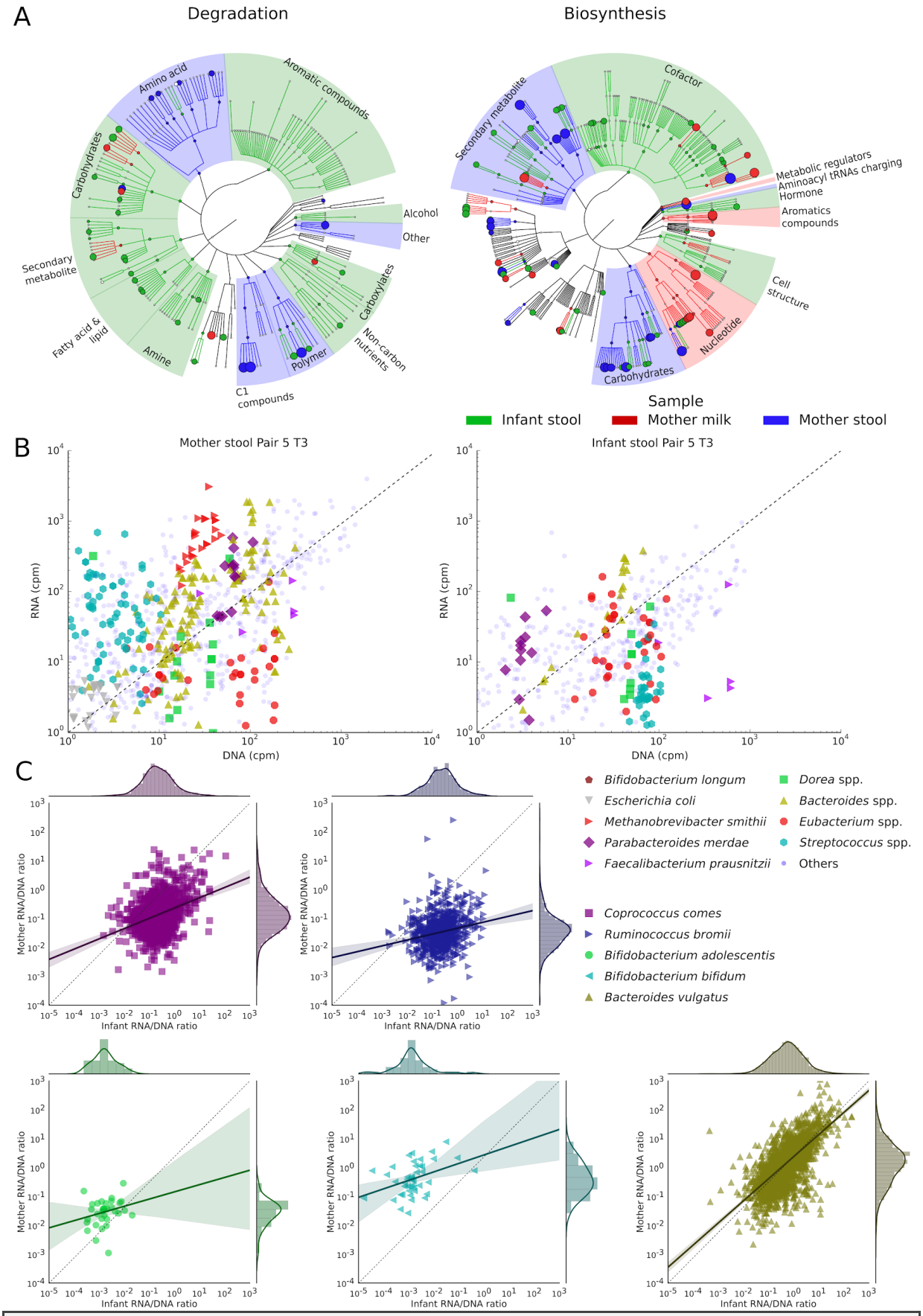


Fig S7. Functional potential biomarker analysis and metabolic pathway expression in mother and infant of pair 5 at time point 3. (A) Degradation and biosynthesis pathways

revealed by HUMAnN2 results processed with LEfSe to investigate differentially expressed pathways and functions. Biomarkers for the three classes are reported in different colors as follows: green, infant feces; red, mother milk; blue, mother feces. The sizes of the clades represent the linear discriminant analysis (LDA) effect sizes assigned by LEfSe (see Materials and Methods). Infants were harboring mainly sugar degraders and showed a higher potential for degradation of aromatic compounds and biosynthesis of cofactors. The microbial communities from the mothers showed instead higher representation of pathways involved in the biosynthesis of carbohydrates and antibiotics and in the degradation of C1 compounds and amino acids. (B and C) Metatranscriptomic analysis of samples from the mother and infant of pair 5 at time point 3 performed with both HUMAnN2 and PanPhlAn. (B) Scatterplots showing the transcription rates of metabolic pathways of different species and genera of interest obtained from HUMAnN2. (C) Comparison between transcription rates of gene families from PanPhlAn data.

For the supplementary tables, I report below only their captions, the tables are available for download on the online version of the paper: https://doi.org/10.1128/MSYSTEMS.00164-16.

Table S1. Sample metadata and raw data. The table reports the sample metadata, the efficiency of extraction, and information about the raw reads.

Table S2. MetaPhlAn2 abundance profiles. The table reports relative abundances of different microbes in metagenomic samples, as profiled with MetaPhlAn2.

Table S3. DNA virus abundance data. The table shows the breadth of coverage and the average depth of coverage for the DNA viruses found in the metagenomes.

5. Applications of GraPhlAn and PhyloPhlAn 2 in other works

In this chapter, I present several applications of the phylogenetic framework from microbiome research that I introduced in the previous chapters. These applications were extracted from a number of works I have co-authored and for which I was responsible for the phylogenetic analysis. Each research article I consider here represents a different application and begins with a brief introduction that explains my role in the research. Then I report the abstract and the main parts related to my contributions.

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

The work by (Ferretti et al., 2018) has as its primary goal the study and characterization of the species that are vertically transmitted from mother to infant. This article can be seen as an extension of the work reported in **Chapter 4** (Asnicar et al., 2017), supported by a larger cohort of 25 mother-infant couples followed in time up to 4 months postpartum. To be able to determine when a vertical transmission event from mother to infant happens, this work heavily relies on the ability to resolve at the strain-level resolution the relationships between microbial genomes. Other than the detectable species that we are able to retrieve using the standard metagenomics analysis tools, in the last part of this work we also used a metagenomic assembly strategy to reconstruct and recover potentially novel species that cannot be detected using reference-based profiling tools. We then used the new version of PhyloPhlAn 2 I developed to accurately place into the microbial tree of life the uncharacterized reconstructed genomes, and GraPhlAn for visualizing the tree with the annotation to highlight the assigned phylum and the eight additional mother-to-infant vertical transmission events we were able to discover through the phylogenetic analysis.

Ferretti P, Pasolli E*, Tett A*, Asnicar F*, Gorfer V, Fedi S, Armanini F, Truong DT, Manara S, Zolfo M, Beghini F, Bertorelli R, De Sanctis V, Bariletti I, Canto R, Clementi R, Cologna M, Crifò T, Cusumano G, Gottardi S, Innamorati C, Masè C, Postai D, Savoi D, Duranti S, Lugli GA, Mancabelli L, Turroni F, Ferrario C, Milani C, Mangifesta M, Anzalone R, Viappiani A, Yassour M, Vlamakis H, Xavier R, Collado CM, Koren O, Tateo S, Soffiati M, Pedrotti A, Ventura M, Huttenhower C, Bork P, and Segata N (* equal contribution)

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

Cell Host & Microbe (2018)

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.

5.1.1 Vertically Transmitted Microbes Are More Likely to Be Stable Colonizers

Vertical microbial transmission from the mother to the infant can either be transient or lead to longer-lasting colonization of the infant gut (Korpela et al., 2018). Of the vertically transferred strains, 17 were identified at more than one time point in the infant (**Figures 4A** and **S6**). In 12 of these 17 cases, after the first occurrence of the maternally acquired strain we found no subsequent replacement by another conspecific strain, i.e., 70.5% of the strains were retained and 29.5% replaced. In contrast, the 163 strains present in the infant at more than one time point, but without evidence that the mother was the source, were found to be replaced in 119 cases (73% replacement) and retained in 44 (27% retention). Vertically transmitted strains therefore seem to have a better fitness for colonization than strains without evidence of acquisition from the mother (70.5% versus 27% stable colonization, Fisher test, $p < 0.001$). This supports the intriguing hypothesis that maternal strains are likely to be more ecologically adaptable in the infant compared with non-maternal strains.

5.1.2 Conspecific Strain Diversity within Fecal Species Is Higher in the Infant Than in the Mother

We next investigated the total strain heterogeneity for each species in the microbiome of the infants compared with that of the mothers. To estimate conspecies strain heterogeneity and dominance, we analyzed the number of polymorphic nucleotide positions in the single-copy marker genes of each detected strain, as well as the average frequency of the dominant allelic variant in polymorphic positions. The analysis of maternal gut samples confirmed that the adult human gut tends to harbor only one strain of a given species (Truong et al., 2017), with an average fraction of polymorphic sites of 0.31% (**Figure 4B**). The infant gut microbiome at day 1 (T1) instead has a very high conspecific strain heterogeneity with 6.1-fold more polymorphisms than the mother ($p = 1 \times 10^{-7}$). As observed, the early infant microbiome at day 1 postpartum is characterized by a high species diversity (**Figure 1B**), which is thus also accompanied by a high strain diversity, further suggesting that the pioneering microbiome is a complex community of microbes shaped by the process of ecological selection over time. Correspondingly, at later time points there is a decrease in the intra-species polymorphic rates up to 1 month (T4), to levels comparable with those of the mothers (no significant difference at 1 month compared with the mother). Simultaneously, a higher relative frequency of the dominant strain is observed (**Figure 4B**). Samples collected from the infant at 4 months of age (T5) then suggest that the strain diversity is increasing and remains significantly higher than the diversity in the mothers ($p = 0.0014$), potentially as a consequence of the increased exposure of the infant to other possible sources of microbial seeding from the environment. Comparing the conspecific strain diversity of the infant over time with the other maternal body sites (**Figure 4B**), we identified markedly different levels of heterogeneity (**Table S4**), with the maternal tongue dorsum significantly more strain diverse than the infant gut ($p < 1 \times 10^{-10}$ for all time points). The maternal skin and vaginal microbiomes have instead a strain diversity in line with that of the infant stool (**Table S4**). Interestingly, and in contrast to the stool, the maternal oral strain diversity compared with infants is significantly higher ($p = 2 \times 10^{-9}$ at T2, t test). Nevertheless, in the infant oral cavity, we identified the same pattern observed in the gut, namely a high species and conspecific strain diversity (**Figures 1B** and **4B**) followed by a

rapid decline in species and strain heterogeneity due to selection, which is observed to start after a few days postpartum.

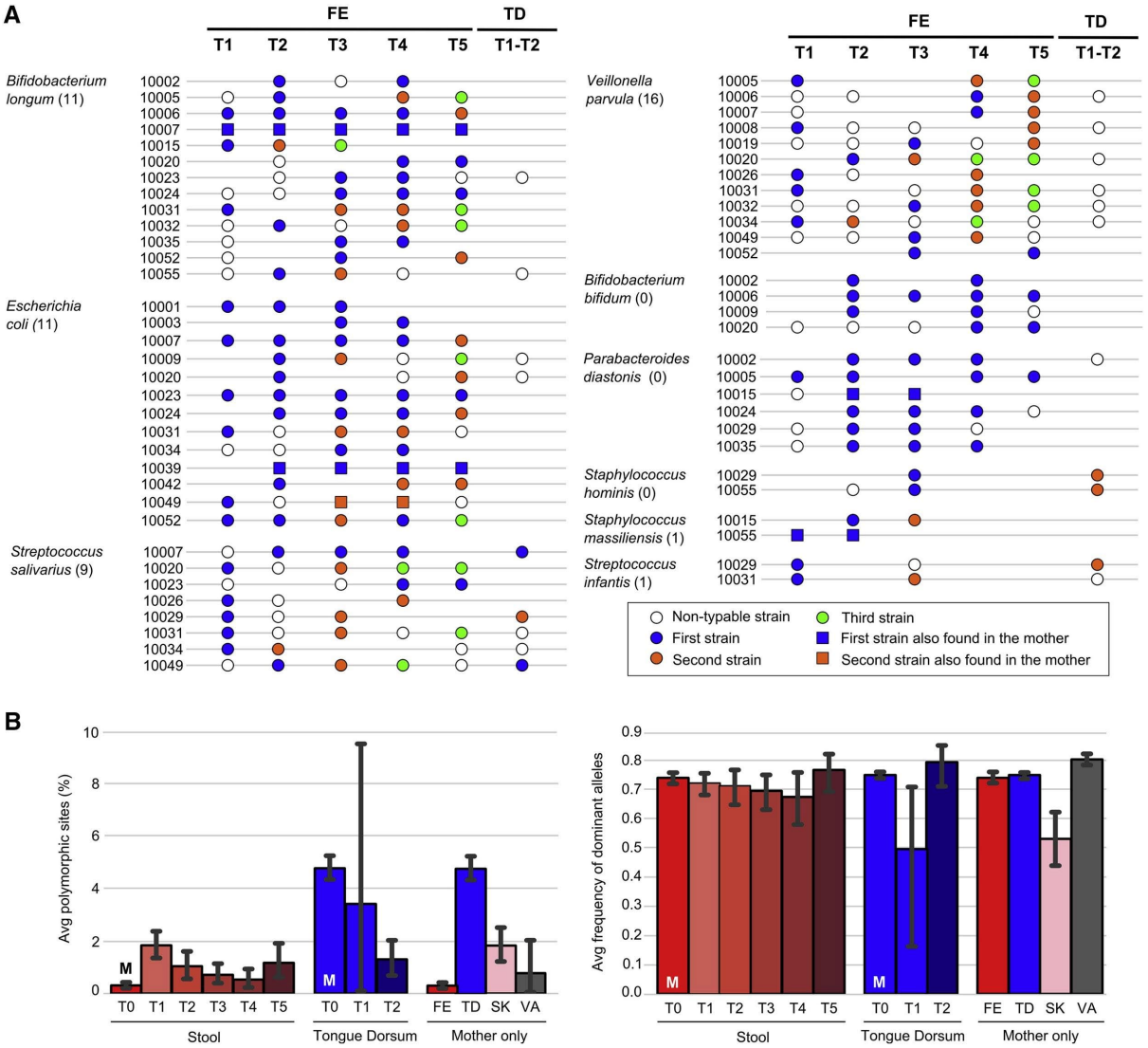


Figure 4 Strain Persistence, Strain Replacement Events, and Strain Heterogeneity. (A) Map of the strain dynamics in longitudinal infant stool (FE) samples for selected species (for full map, see **Figure S6**). The tongue dorsum (TD) column shows the species for which at least one of the strains found in stool was also present on the tongue dorsum. Blue circles represent the first strain of the species identified in the infant, whereas orange and green circles denote the second and third longitudinally identified strain, respectively. Empty circles refer to species for which strain profiling was not possible in the specific sample. Missing samples and samples lacking the species are not reported. The total number of infant replacement events observed in each species is shown in parentheses. **(B)** Mean percentages of polymorphic sites and average frequency of the dominant alleles in polymorphic sites for each body site and time point ("M" indicates maternal samples). Color coding is as per **Figure 1**. p values are reported in **Table S4**. Error bars refer to 95% confidence intervals.

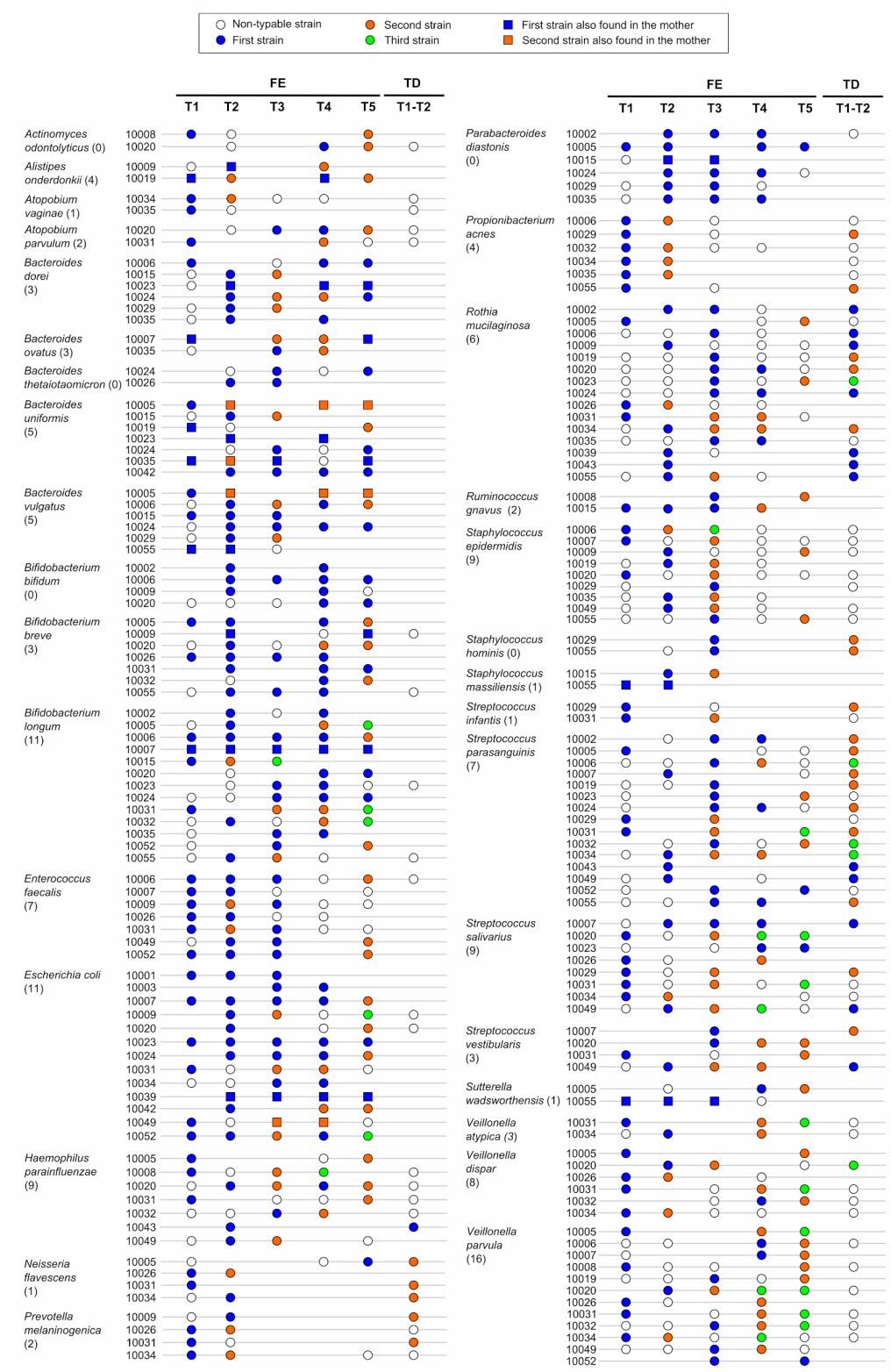


Figure S6. Related to Figure 4. Full list of strain replacement events in the infant body sites. Species with at least two pairs are shown. Empty circle present when only the identification at the species-level was available, i.e. the strain reconstruction was not possible (non-typable strain). Missing circle when the species is not present in the sample. In brackets, the number of replacement events per species (only infants stool samples are considered). In total, we identified 136 replacement events (on average 4.7 events per species and 5.4 per pair).

5.1.3 Strains Belonging to as yet Uncharacterized Species Are Also Vertically Transmitted

To perform strain profiling for microbes belonging to poorly characterized species without available genomes, we expanded our analysis by performing metagenomic assembly (Nurk et al., 2017) on each sample followed by contig binning (Kang et al., 2015), phylogenetic profiling (Segata et al., 2013), and whole-genome strain identity inference (**STAR Methods**). Overall, we reconstructed 1,132 metagenome-assembled genomes (on average five per sample; **Table S6A**) with sufficient quality (≥50% completeness, ≤5% contamination) to be amenable for strain tracking. Of these, 763 genomes could be assigned to a known species by applying a 95% percent identity threshold on the whole sequence (**STAR Methods**) and were therefore not further processed because these strains were captured by the reference-based methods already considered above. However, the remaining 369 genomes (**Figure 6**; **Table S6A**) did not belong to any of the 13,575 microbial species for which at least one reference genome is available (**STAR Methods**), including 36 that could not even be assigned below the level of family. The genera containing most of the unknown species were *Streptococcus* (32 genomes), *Clostridium* (31), and *Prevotella* (31).

Next, we compared the 369 taxonomically uncharacterized genomes against each other to identify the presence of the same strain in different metagenomic samples. Using a strict threshold of 99.5% identity over the full length of the genomes, we identified eight vertical transmission events (**Figure 6**; **Table S6B**). In six cases the strain sharing was between the mother and infant gut microbiomes. Two of these strains belonged to uncharacterized species in the *Akkermansia* and *Bacteroides* genera (less than 89% identity with the closest available genomes over less than 75% of the length), while for the other four strains classification was even more challenging and we could only infer that they belonged to four different phyla (Verrucomicrobia, Proteobacteria, Bacteroidetes, and Firmicutes; **Table S6B**). In addition to the fecal transmission routes, two vertical transmission events were also observed from other body sites with an uncharacterized Clostridiales strain (99.9% of similarity) shared by the maternal vaginal community and the stool of the infant, and an unknown *Leptospira* strain (99.9% of similarity) shared by the skin microbiome of the mother and the saliva microbiome of the infant. There was only one case of a strain from an unknown species with 99.9% similarity within an unrelated mother-infant pair, strongly confirming the occurrence of vertical transmission for the eight genomes above (Fisher test, $p < 1\times10-9$) and confirming that uncharacterized species have a role in the mother-to-infant microbial seeding.

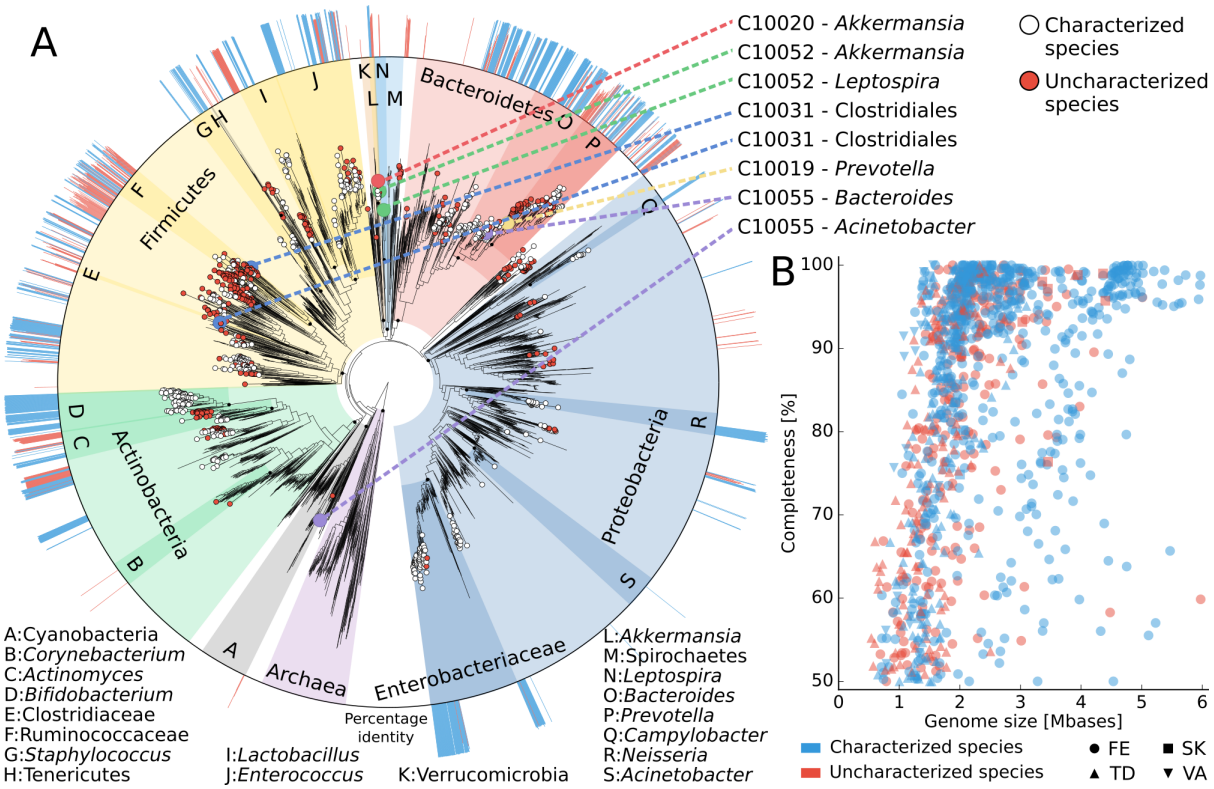


Figure 6 Phylogenetic Placement of 1,132 Metagenomically Reconstructed Genomes and Mother-to-Infant Transmission of Taxonomically Uncharacterized Strains. (A) We used PhyloPhlAn2 (Segata et al., 2013) to place the genomes reconstructed with metaSPAdes (Nurk et al., 2017) and binned with MetaBAT2 (Kang et al., 2015) (**STAR Methods**) on the microbial "Tree of Life" (Ciccarelli et al., 2006; Segata et al., 2013), which encompasses 4,000 species with available reference genomes. Leaf nodes without circles refer to reference genomes from known species, white circles indicate reconstructed genomes that are close (>95% identity) to a known species, and red circles show reconstructed genomes that cannot be assigned (<95% identity) to known species. The eight events of mother-to-infant transmission of strains from species yet to be described are called out on the top right, and the external ring of the phylogeny reports the percent identity of each leaf node against the closest genomes from known species (values below 95% are shown in red). **(B)** The reconstructed genomes with completeness >50% from each body site are plotted with the corresponding completeness and genome size.

5.2. A reference phylogeny of 10,575 genomes redefines major clades of bacteria and archaea

Very-large phylogenies reconstruction is challenging because of the large number of genomes that constitute the tree and the number of positions in the concatenated MSA for building the phylogeny, in the case of a concatenation approach. The latter issue can be slightly ameliorated by employing a gene trees approach, where many phylogenies based on shorter MSAs can be inferred and the final tree can be built by using a summary method. In this work, the 400 marker genes of PhyloPhlAn have been validated with respect to the classical set of ribosomal proteins (Wu and Eisen, 2008; Wu and Scott, 2012) and used to build a tree of life phylogenies of 10,575 genomes using a gene trees approach. This is one of the results of my research period abroad I spent at the University of California, San Diego in the laboratory of Prof. Siavash Mirarab. The main focus of the Siavash lab is phylogeny reconstruction using gene trees and summary methods, and multiple sequence alignments. The main goals of my research visit period were the understanding and application of the gene trees approach to phylogenetic analysis in the bacterial domain. In the same campus, I collaborated also with the Knight lab lead by Prof. Rob Knight, which together with Prof. Siavash resulted in the work presented below. My contribution to this work was mainly in helping with retrieving the reference genomes using the RepoPhlAn[13] tool, mapping the 400 PhyloPhlAn marker genes against the identified set of 10,575 reference genomes, helping in performing the MSA for each marker, annotating with UniRef50 the 400 markers, and finally producing several version of the tree of life phylogeny proposed in this work by using the concatenation approach. This work is submitted and currently in review.

Zhu Q*, Mai U*, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, Kosciolek T, Yin JB, Huang S, Salam N, Jiao JY, Wu Z, Xu ZZ, Sayyari E, Morton JT, Podell S, Knights D, Li WJ, Huttenhower C, Segata N, Smarr L, Mirarab S, and Knight R (* equal contribution)

A reference phylogeny of 10,575 genomes redefines major clades of bacteria and archaea

In revision

Abstract

Rapid growth of genome data provides opportunities in updating microbial evolutionary relationships. This is challenged by the discordant evolution of individual genes. We built a reference phylogeny of 10,575 evenly-sampled bacterial and archaeal genomes, based on a comprehensive set of 381 markers, using a gene tree summary method alongside conventional strategies. High resolution was achieved for deep branches, recovering Archaea, the candidate phyla radiation (CPR), and non-CPR bacteria as three major clades. Our tree indicates remarkably closer evolutionary proximity between Archaea and Bacteria than previous estimates that were limited to fewer “core” genes. The estimated timeline suggests a continuous pattern of diversification after the Archaea-Bacteria split. We released the tree and genome catalog as a database that will benefit the microbial research community.

[13] The RepoPhlAn repository in Bitbucket: https://bitbucket.org/nsegata/repophlan

5.2.1 Introduction

In this work, we built a reference phylogeny (**Fig. 1**) of 10,575 bacterial and archaeal genomes. They were sampled from all 86,200 non-redundant genomes available from NCBI GenBank and RefSeq 13 as of March 7, 2017 (**Fig. S1**), using a purpose-built statistical approach which maximizes the covered biodiversity. Our phylogenetic reconstruction used 381 marker genes, selected from whole genomes solely by sufficient sequence conservation to identify homology, without a priori assumptions on their functional roles. The whole dataset totals 1.16 trillion non-gap amino acids, making it among the largest single datasets upon which de novo phylogenetic trees have been built, and represents a remarkable expansion from previous efforts of the kind considering both taxon and gene sampling (**Table S1**). To account for the discrepancy among the evolutionary histories of individual genes, we used a summary approach that accounts for gene tree discrepancy to infer a species tree, and we compared this to the conventional gene alignment concatenation approach. The resulting species tree provides high resolution of the basal relationships among microbial clades, revealing a unified pattern of bacterial and archaeal evolution, in which the two domains are in closer proximity compared to previous estimations. It further enabled us to revisit timings of important evolutionary events and to make corrections to previously established taxonomic hierarchies. The whole workflow from genome sampling to tree building was fully automated, without the need for manual curation, enabling convenient future upgrades. We made the trees, the genome catalog, and the extended database files publicly available at: https://github.com/biocore/wol, under an open source license.

5.2.2 Improved resolution of deep phylogeny achieved by gene tree summary

The ASTRAL (**Figs. 1** and **S5**) and CONCAT trees (**Figs. S6** and **S7**) have high levels of overall congruence in topology when compared to trees derived from implicit (e.g., distance-based) analyses (**Fig. S4A**). The congruence is higher at shallow branches, but generally decreases as phylogenetic depth increases (**Fig. S8**). Nevertheless, the trees are consistent in supporting multiple recently defined above-phylum taxonomic groups, including Asgard, TACK, Terrabacteria, FCB, PVC, Parcubacteria and Microgenomates (Castelle and Banfield, 2018; Rinke et al., 2013), while revealing the para/polyphyletic properties of large but promiscuous taxonomic groups such as phyla Proteobacteria and Firmicutes. A detailed interpretation of the phylogenies in reference to taxonomy is provided in **Supplementary Text 3**. When considering branch support statistics, the ASTRAL tree has notably higher confidence than the CONCAT trees in untangling the relationships among the early branching clades (**Fig. S9**, also see **Figs. S5-S7**). This high resolution is directly related to the large number of gene trees used in the inference, as using fewer loci notably decreased the branch support of the species tree (**Fig. S4B**). These observations, in addition to the fact that ASTRAL trees use all the data but CONCAT is limited only to a selection of 100 sites per gene, motivated us to use the ASTRAL tree as our reference.

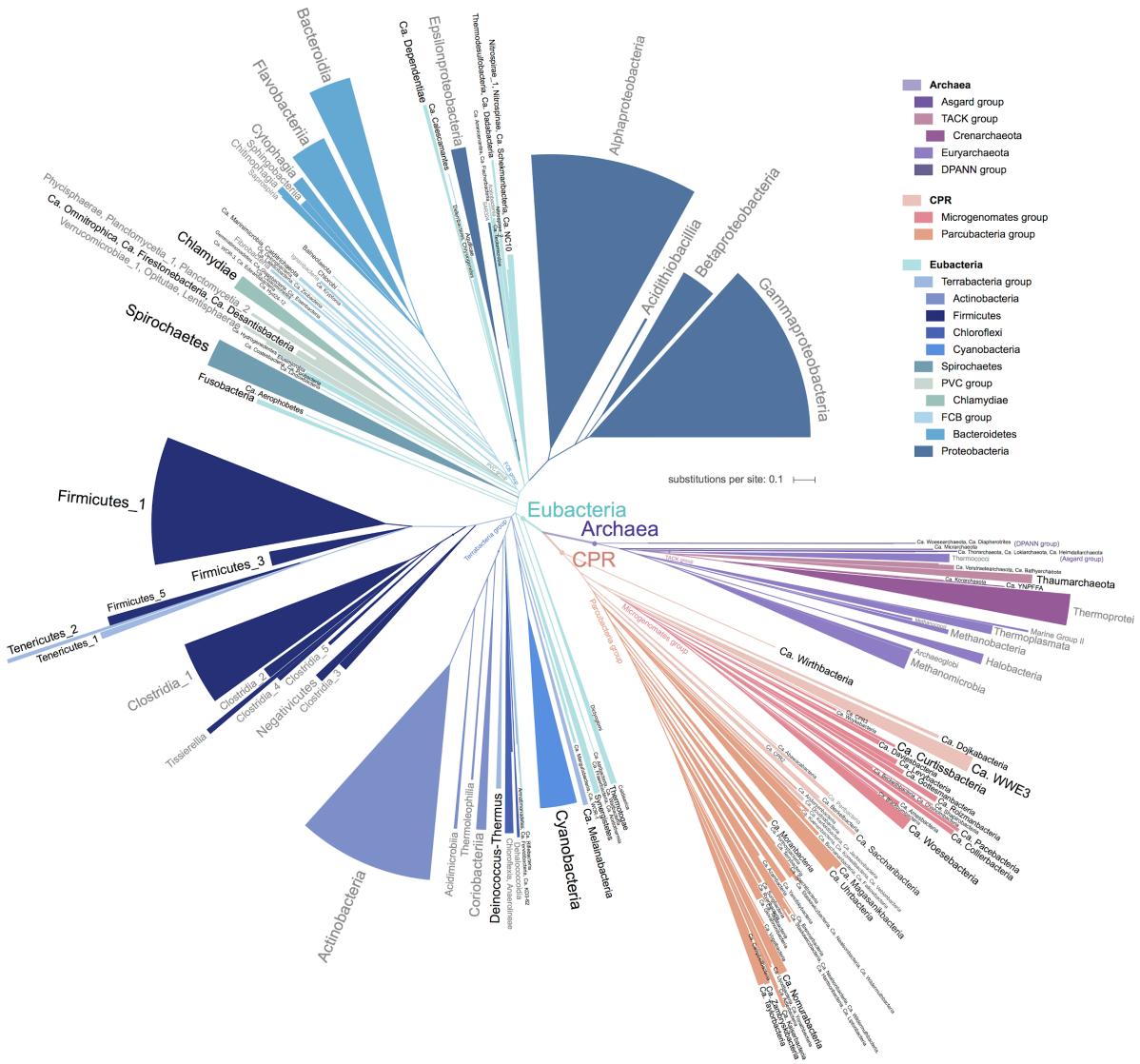


Fig. 1. A new view of the bacterial and archaeal tree of life. The tree contains 10,575 evenly distributed bacterial and archaeal genomes, with topology reconstructed using ASTRAL based on individual trees of 381 globally sampled marker genes, and branch lengths estimated based on 100 most conserved sites per gene. Branches with effective number of genes (en) ≤ 5 and local posterior probability (lpp) ≤ 0.5 were collapsed into polytomies. Taxonomic labels at internal nodes and tips reflect the tax2tree curation result. Color codes were assigned to above-phylum groups and phyla with 100 or more representatives. To realistically display the tree in a page, it was collapsed to clades (sectors) representing phyla with at least one taxon (black), and classes with at least 10 taxa (grey). The radius of a sector indicates the median distance to all descending taxa of the clade, and the angle is proportional to the number of descendants. For polyphyletic taxonomic groups, minor clades with less than 5% descendants of that of the most specious clade were omitted, while the remaining clades were appended a numeric suffix sorted by the number of descendants from high to low.

5.2.3 Archaea, CPR, and Eubacteria are three major clades

Phylogenetic trees built by both strategies (**Figs. 1, S5-S7**) recapitulated clear separation between Archaea (669 taxa) and Bacteria (9,906 taxa) at the base. Meanwhile, the candidate phyla radiation (CPR), a recently discovered group of mainly uncultivated

microorganisms (Brown et al., 2015), forms a strict monophyletic group (1,445 taxa) located at the base of the bacterial lineage. We repurposed the classical term “Eubacteria” to describe its sister clade, which comprises all the remaining bacteria (8,461 taxa). The three basal clades, unlike in results from ribosomal proteins, show similar levels of evolutionary divergence both within each clade and with respect to their shared parental node (**Figs. 2A**, **S10** and **Table S3**).

The separation into these three clades is also evident in a Principal Coordinates Analysis (PCoA) of a simple measure of genome distance that solely uses marker gene presence / absence (**Figs. 2B** and **S11**), as well as the PERMANOVA test (p-value = 0.001 between each pair of clades), and the Random Forests classification (accuracy = 1.0 for correct predictions and 0.0 for incorrect ones) (**Table S4**). Consistent with recent studies on the distinct biological capacities of CPR (Meheust et al., 2018; Wrighton et al., 2012), our results further strengthen the view of CPR as a unique, early branching clade of microorganisms.

5.2.4 Evolutionary proximity between Archaea and Bacteria

ASTRAL and CONCAT trees both reveal a relatively short branch connecting the most recent common ancestors of Archaea and Bacteria (**Figs. 1**, **2A** and **S10**). Its length is fractional comparing to the dimensions of both clades (appr. 0.13-0.14 by conserved sites, 0.09-0.11 by random sites) (**Table S3**). This pattern is in contrast to previous trees built using fewer marker genes, all or most of which are ribosomal proteins which were considered to be effective markers for assessing global microbial evolution (Ramulu et al., 2014) (e.g., (Castelle and Banfield, 2018; Ciccarelli et al., 2006; Yutin et al., 2012)). To test whether the choice of marker genes is the main reason for the differential inter-domain distance, we re-estimated branch lengths of the ASTRAL tree using 30 ribosomal proteins extracted from the genomes. Consistent with these studies, we observed an elongated branch connecting Bacteria and Archaea. Its length relative to clade dimensions is about 10-fold as the estimate using the 381 global marker genes (**Table S3**). We also calculated the overall phylogenetic distance between taxa of the two domains, as relative to the intra-domain distances. This relative distance based on the ribosomal proteins (4.5-5.0) is around three times that of the distance by the global marker genes (1.5-1.6) (**Table S5**).

We tested whether the potential saturation of amino acid substitution could cause an underestimation of the domain separation. The ratio between phylogenetic distance and sequence distance is similar between pairs of taxa selected both from Bacteria, both from Archaea, or one from each domain (**Fig. S12**). This indicates that the relative length of the branch connecting the two domains compared to the intra-domain branches is not substantially impacted by saturation.

We further evaluated how individual gene trees impact the observed proximity between Bacteria and Archaea. Except for a few outliers, which include several “core” genes like *rpoC* (RNA polymerase subunit β’, 18.27), *tuf* (elongation factor Tu, 12.18) and *fusA1* (elongation factor G, 9.54), most gene trees have the relative Archaea-Bacteria distance between 1 and 3 (mean: 2.00) (**Figs. 2C**, **S13A** and **File S1**), which is consistent with that of the species tree summarizing the global marker genes, and in contrast to that by only the ribosomal proteins (**Fig. 2C** and **Table S5**).

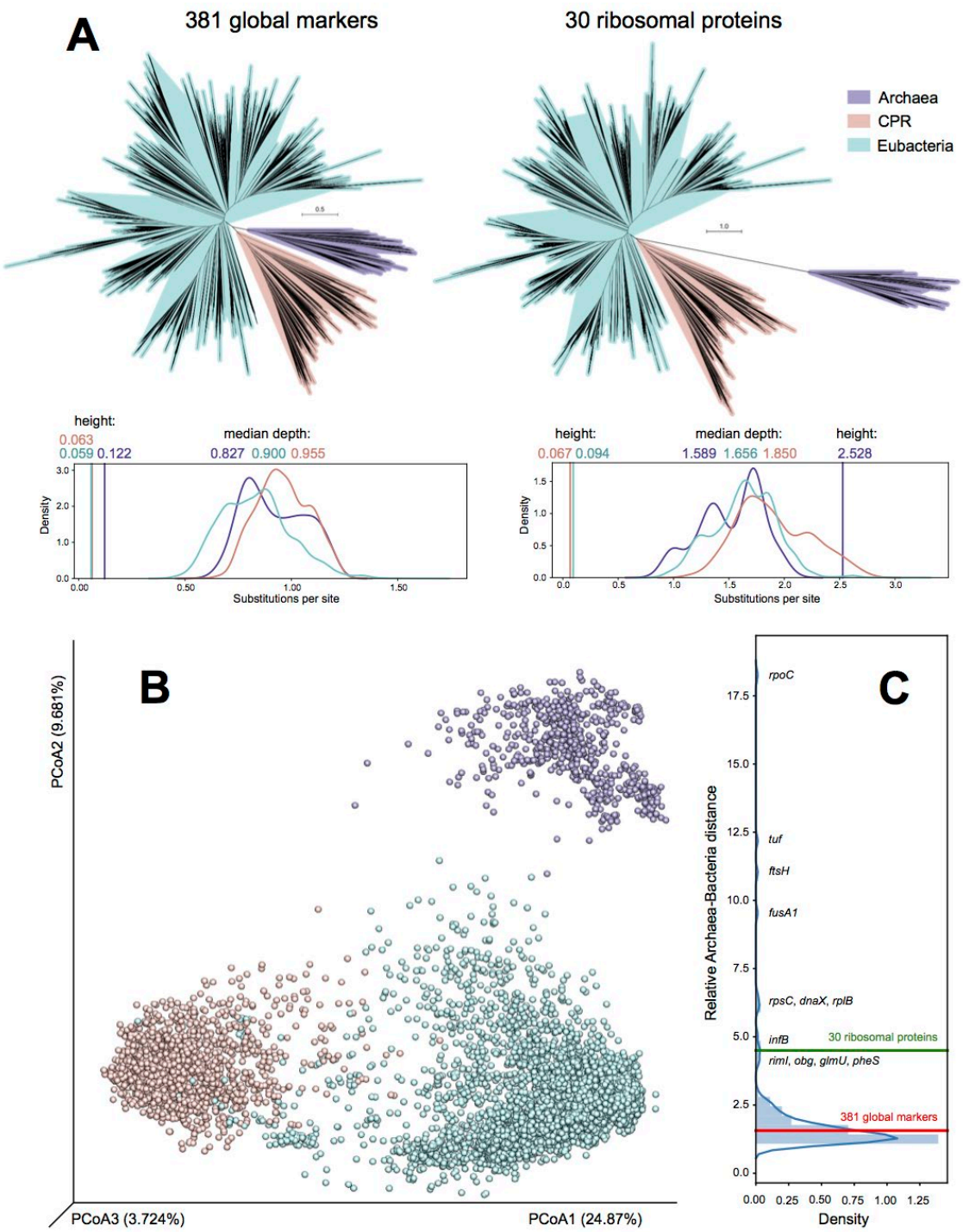


Fig. 2. Dimensions and separation of the three major clades: Archaea, CPR and Eubacteria. A. The unrooted, drawn-to-scale ASTRAL tree with branch lengths re-estimated using the 381 global marker genes (conserved site sampling) (left) or using the 30 ribosomal proteins (right). The Gaussian kernel density function of the depths by all descendants (sums of branch lengths from a tip to the lowest common ancestor (LCA) of the clade) was plotted; the height (length of the branch connecting the LCA to the parental node shared by all three clades) was marked as a vertical line. **B.** PCoA of the Jaccard distance matrix of the marker gene profiles in the sampled genomes. The proportion of variance explained was marked at each axis. Note that axis 1 which separates CPR from Eubacteria explains 2.6 times as much variance as axis 2 does which separates Archaea from Bacteria. **C.** Distribution of relative Archaea-Bacteria distances of each of 161 gene trees containing at

least half of the taxa of both domains. A histogram with Gaussian kernel density function is plotted. The red and yellow lines indicate the values of the ASTRAL tree with branch lengths re-estimated using the global markers and the ribosomal proteins, respectively.

5.3. Combined metagenomic analysis of colorectal cancer datasets defines cross-cohort microbial diagnostic signatures and a link with choline degradation

In this work, we focused on a meta-analysis of colorectal cancer (CRC) cohorts and made publicly available two new CRC cohorts sampled in Italy. The goal of the paper is to determine through the use of a machine learning approach a microbial signature that can be used as a non-invasive diagnostic marker. Phylogenetic analysis in this work focused on identifying directly from metagenomic data new variants of the choline TMA-lyase (*cutC*) gene that we found relevant in a novel potential mechanism for carcinogenesis. These new *cutC* gene variants we reconstructed are partially belonging to known species, but are for the most part reconstructed from genomes of yet-to-be-characterized species we extracted directly from the metagenomes. We also identified differential abundances of the single-copy *cutC* gene in CRC-associated samples. Thanks to the phylogenetic analysis, we identified four variants associated with the phylogenetic structure, and some *cutC* variants that belong to *Hungatella hathewayi*, *Clostridium asparagiforme*, *Klebsiella oxytoca*, and *Escherichia coli* were significantly associated to CRC samples.

Thomas AM*, Manghi P*, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, and Segata N (* equal contribution)

Combined metagenomic analysis of colorectal cancer datasets defines cross-cohort microbial diagnostic signatures and a link with choline degradation

Currently in revision at *Nature Medicine*

Abstract

Several studies have investigated links between the gut microbiome and colorectal cancer (CRC), but questions remain about existing and novel biomarkers and their validity across cohorts and populations. We performed a meta-analysis of 969 fecal metagenomes (413 carcinomas, 143 adenomas, and 413 controls), including five publicly available datasets, two new cohorts, and two additional validation datasets. Unlike microbiome shifts associated with gastrointestinal syndromes, the gut microbiome in CRC showed reproducibly higher species and pathway richness than controls (P < 0.01), partially due to expansions of species typically from the oral cavity and of newly associated species such as *Streptococcus tigurinus* and *Streptococcus dysgalactiae*. Meta-analysis of the microbiome functional potential identified gluconeogenesis and the putrefaction and fermentation pathways to be associated with CRC, whereas the stachyose and starch degradation pathways were more abundant in healthy controls. Predictive microbiome signatures trained on multiple datasets showed consistently high accuracy in successively held-out and independent validation cohorts (average AUC 0.84, minimum 0.81). Pooled analysis of raw sequencing data showed that the choline trimethylamine-lyase (*cutC*) gene was over-abundant in CRC (P = 0.001) with the strength of association differing between the four *cutC* variants we identified and the variation confirmed in independent validations (P < 1e-6) as well as at transcriptional level (P = 0.035). The combined analysis of heterogeneous CRC cohorts and independent validation cohorts thus identify reproducible microbiome biomarkers and high-accuracy

predictive models, that can be the basis for clinical prognostic tests and hypothesis-driven mechanistic studies.

5.3.1 Increased abundance of choline TMA-lyase enzymes in CRC

Microbiome-derived metabolites have been implicated in carcinogenesis (Di Martino et al., 2013; Ou et al., 2012). We chose to focus on trimethylamine (TMA), an amine produced by bacteria from choline and carnitine, because it has been recently shown to play a role in complex diseases such as atherosclerosis and primary sclerosing cholangitis (Jie et al., 2017; Kummen et al., 2017). Since dietary components have been shown to be linked with CRC risk (Huxley et al., 2009; Johnson et al., 2013; Wei et al., 2004), we hypothesized that the TMA-producing potential of the human gut microbiome could also be associated to CRC (Oellgaard et al., 2017). To test this hypothesis, we built a database of genes belonging to the main TMA-synthesis pathways and used it to reconstruct and quantify the presence of such genes in the 764 CRC-associated metagenomes considered here. The main genes associated with TMA-synthesis are those encoding the choline TMA-lyase (*cutC*), the L-carnitine dioxygenase (*yeaW*) and the L-carnitine/gamma-butyrobetaine antiporter (*caiT*) and we identified them in 923, 5,185 and 5,709 available bacterial genomes, respectively. Putative *cutC* sequences belonged mainly to *Proteobacteria* (mostly *Gamma-* and a few *Deltaproteobacteria*) and *Firmicutes* (mainly from *Clostridia* and a few *Bacilli*), with few *Actinobacteria* as reported previously (Rath et al., 2017).

Screening the 7 CRC-associated metagenomic datasets, we found that only one of them had a significant increase of *caiT* in CRC samples compared to controls, whereas no significant differences were detected for *yeaW* (**Suppl. Fig. 21**). However, we found increased abundance of *cutC* in CRC samples compared to controls in all seven datasets ($P < 0.05$ by Wilcoxon Rank Sum test on RPKM abundances for five datasets, **Figure 4A**). Meta-analytical synthesis indicated an overall strong association with no evidence of heterogeneity ($P = 0.001$, $\mu = 0.27$, 95% CI [0.1, 0.42], $I^2 = 4.2\%$, Q-test = 0.65, **Figure 4B**). We also analyzed the abundance of the activating choline trimethylamine-lyase enzyme (*cutD*), finding a significant increase in CRC (meta analysis results P = 0.001, $\mu = 0.32$, 95% CI [0.16, 0.47], $I^2 = 0\%$, Q-test = 0.96, **Suppl. Fig. 22**). These results indicate that TMA production might happen preferentially via choline degradation, and not via carnitine, and could substantially affect the amounts of TMA and trimethylamine oxide (TMAO) in an individual (Kalnins et al., 2015). Intermediate levels of *cutC* in adenomas (**Figure 4A**) is further suggestive of a TMA action along the adenoma-carcinoma axis. We validated the increased *cutC* gene abundance in CRC by qPCR (Rath et al., 2017) on a subset of samples from Cohort1 with enough DNA left after sequencing, and found that the metagenomic findings were confirmed (one-tailed Wilcoxon signed rank test $P = 0.024$, **Figure 4D**). Further quantification of *cutC* transcript abundance from the co-extracted RNA in the same dataset also pointed to an over-expression of this gene in CRC ($P = 0.035$, **Figure 4E**).

We further explored the role of *cutC* in the gut microbiome by reconstructing sample-specific sequence variants using a reference-aided targeted assembly approach (see **Methods**). We found a large sequence divergence for the gene encoding this enzyme that is known to occur in single copies in the genomes (Rath et al., 2017) and we identified four main sequence variants that are associated with the taxonomic structure (**Figure 4B**, **Suppl. Figs. 23-24**). Interestingly, the most prevalent (46.5%) *cutC* sequence type belonged to an unknown uncultured *Eubacterium* species with only 95% sequence identity to the closest

known and taxonomically characterized variant. This *cutC* variant was associated with non-CRC samples (OR 0.38, 95% CI [0.25, 0.57], *P* = 0.0001, Fisher Test), whereas *cutC* sequence types mostly belonging to *Hungatella hathewayi* and *Clostridium asparagiforme* (*Firmicutes*) were significantly CRC-associated (OR 2.14, 95% CI [1.29, 3.56], *P* = 0.004, Fisher test), as were sequence types belonging to *Klebsiella oxytoca* and *Escherichia coli* (OR 1.85, 95% CI [1.13, 3], *P* = 0.02, Fisher Test - **Figure 4B**). Altogether, these novel findings highlight that sequence variants of this enzyme can be strongly associated with disease, potentially because of corresponding differences in the efficacy of choline degradation and TMA production (Jameson et al., 2016; Romano et al., 2015).

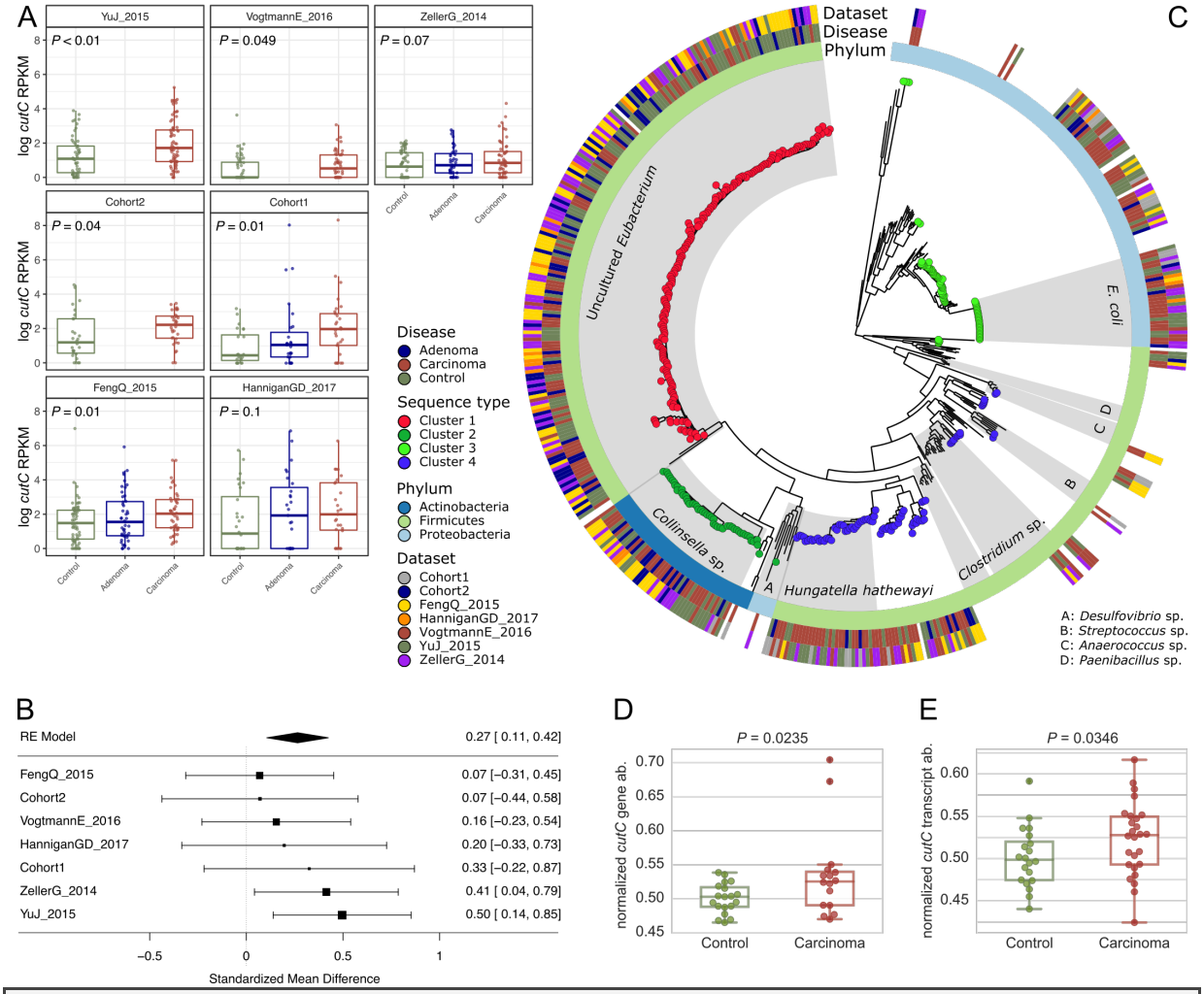


Figure 4. Choline TMA-lyase cutC and its genetic variants are strong biomarkers for CRC-associated stool samples. (A) Boxplots showing the log of reads per kilobase million (RKPM) abundances obtained using ShortBRED for the choline TMA-lyase enzyme cutC. P-values were computed by Wilcoxon Signed-Rank tests comparing values between controls and carcinomas for each dataset. **(B)** Forest plot reporting effect sizes calculated using a meta-analysis of standardized mean differences and a random effects model on cutC RPKM abundances between carcinomas and controls. **(C)** Phylogenetic tree of sample-specific cutC sequence variants identified four main sequence variants. Tips with no circles represent cutC sequence variants from genomes absent from the datasets. Taxonomy was assigned based on mapping against existing cutC sequences (criteria of 80% coverage, >97% identity and minimum 2,000nt alignment length). **(D)** qPCR validation of cutC gene

abundance and **(E)** cutC transcript abundance (normalized by total 16S rRNA gene/transcript abundance) on a subset of DNA samples from Cohort1. qPCR validation P-values are obtained by 1-tail Wilcoxon Signed-Rank test.

5.4. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis

The skin microbiome is a difficult human body site to study because of the low microbial biomass (Byrd et al., 2018; Kong et al., 2017). There is a panel of species that have been defined as typically present in a skin sample like *Staphylococcus epidermidis*, *Propionibacterium acnes*, and *S. caprae*/*S. capitis*. However, the role that the skin microbiome plays in health or disease state is not yet elucidated. The goal of the following work is to study and characterize the skin microbiome members in the presence of psoriatic lesions compared to unaffected locations. In the last section, we focused on the phylogenetic characterization of unknown assemblies coming from the *Malassezia* spp., the *Anaerococcus* spp. of the *Peptostreptococcaceae* family, and novel unknown members placed between the *Chromobacteriaceae* and *Neisseriaceae* families. In this work, I thus expanded the phylogenetic framework I developed to the study of micro-Eukaryotic organisms, specifically from the genus *Malassezia*.

Tett A, Pasolli E, Farina S, Truong DT, Asnicar F, Zolfo M, Beghini F, Armanini F, Jousson O, De Sanctis V, Bertorelli R, Girolomoni G, Cristofolini M, and Segata N

Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis

Nature Biofilms and Microbiomes (2017)

Abstract

Psoriasis is an immune-mediated inflammatory skin disease that has been associated with cutaneous microbial dysbiosis by culture-dependent investigations and rRNA community profiling. We applied, for the first time, high-resolution shotgun metagenomics to characterise the microbiome of psoriatic and unaffected skin from 28 individuals. We demonstrate psoriatic ear sites have a decreased diversity and psoriasis is associated with an increase in *Staphylococcus*, but overall the microbiomes of psoriatic and unaffected sites display few discriminative features at the species level. Finer strain-level analysis reveals strain heterogeneity colonisation and functional variability providing the intriguing hypothesis of psoriatic niche-specific strain adaptation or selection. Furthermore, we accessed the poorly characterised, but abundant, clades with limited sequence information in public databases, including uncharacterised *Malassezia* spp. These results highlight the skins hidden diversity and suggests strain-level variations could be key determinants of the psoriatic microbiome. This illustrates the need for high-resolution analyses, particularly when identifying therapeutic targets. This work provides a baseline for microbiome studies in relation to the pathogenesis of psoriasis.

5.4.1 Psoriatic microbial niches comprise a large proportion of unknown microbes

The skin is inhabited by diverse taxa and intra-species variability that is poorly characterised. This “dark matter” includes species, genera, or higher-level taxonomic ranks with either no or only a few representative references genomes. By employing an assembly-based genome reconstruction approach for each skin metagenome (see **Methods**), we identified contig clusters with little or no homology to the reference data sets (**Fig. 4**), which, therefore, represent taxa without any closely related sequenced strains. To explore these “unknowns” further, we first compared the assemblies to the closest available references based on

sensitive mapping capturing even at low sequence similarities. This enabled us to identify a number of uncharacterised but abundant eukaryotic and bacterial organisms from the cutaneous microbiomes. A common, eukaryotic inhabitant of the skin microbial community is the fungus *Malassezia globosa* (Gaitanis et al., 2012). Where sufficient coverage permitted, we identified 18 "unknown" clusters with either weak or divergent genome content compared to *M. globosa*. Their reconstructed genomes averaged in length 4.4 Mb (s.d. 2.8 Mb), which means a large fraction of the genome was reconstructed given the draft genomes obtained from pure culture of *M. globosa* CBS 796653 and *M. restricta* CBS 8742 are 8.96 and 7.26-Mb long, respectively (Wu et al., 2015). Recently, representative genomes for all 14 accepted species of the *Malassezia* genus have been sequenced (Wu et al., 2015), whereby the authors report phylogenetically the *Malassezia* genus supports three main clusters. Phylogenetic comparison of our *Malassezia* reconstructed genomes (**Fig. 5a**) finds most fit in cluster 2 and are closest to *M. globosa* and *M. restricta*, the two most common *Malassezia* spp. found on human skin (Wu et al., 2015).

Unexplored diversity is still, however, present within the metagenomically retrieved *Malassezia* genomes in cluster 2 (**Fig. 5a**). Only four of the reconstructed genomes are placed close enough (Average Nucleotide Identity (ANI, (Goris et al., 2007)) > 97.5%) to *M. globosa* (patient 19 and one strain from patient 102) or *M. restricta* (Patients 105 and 106) to be confidently assigned to these two species. For the other *Malassezia* genomes, e.g., the highest ANI of the strains from Patient 9 compared to the most phylogenetically related *M. restricta* 8742 is 92.84% (s.d. 2.52%), which is suggestive that they may represent a distinct and unsequenced species. More strikingly, patient 16's ear is instead inhabited by a more distantly related strain, which cannot be assigned to any of the three clusters, and therefore could be an unknown more ancestral *Malassezia* spp. (**Fig. 5a**); intriguingly, this ear was diseased and may therefore be colonised by a hitherto uncharacterised fungal species, which may be of relevance to psoriatic disease.

Focusing on unknown bacterial clusters, we detected a cluster in Patient 9 that likely represents an uncharacterised *Anaerococcus* spp. (**Fig. 5b**) as its closest reference is *Anaerococcus* spp. 9402080, but the two genomes only share an ANI value of 80.6% (s.d. 5.28%). Both ears of Patient 9 are also inhabited by an unknown bacterial taxon that is related but cannot be placed in either the *Chromobacteriaceae* or *Neisseriaceae* families (**Fig. 5c**). What emerges from the analysis of the "microbial dark matter" is that the skin microbiome, both unaffected and in relation to disease, is much more complex and diverse than taxonomic profiling based on references genomes alone permits. Thus, such hidden diversity highlights limitations in our reference data sets and the potential role of these unknown taxa in skin health and disease could be overlooked.

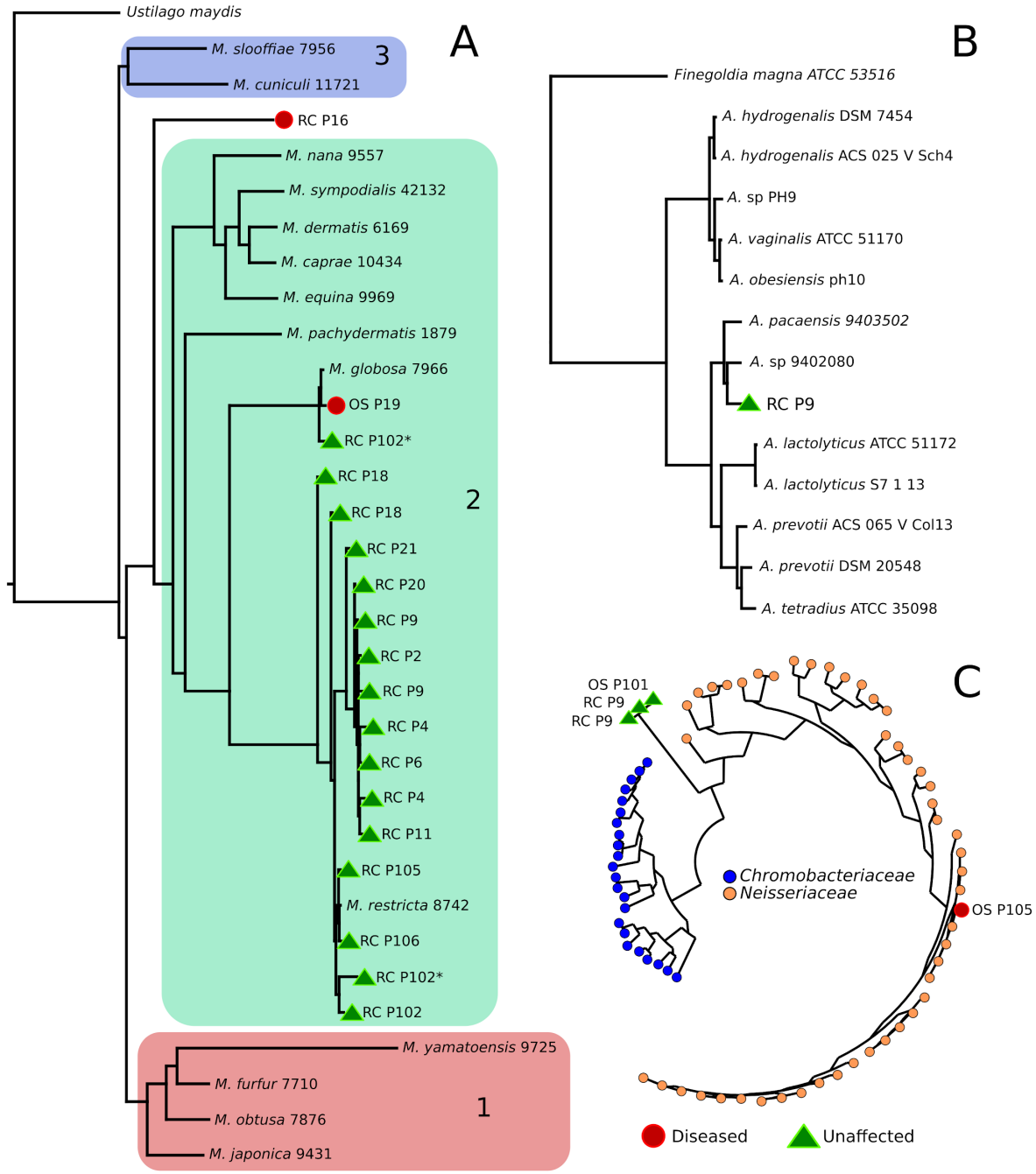


Fig. 5 Phylogenetic analysis of taxonomically uncharacterised metagenomic assemblies (unknowns) compared to the closest representative reference genomes. a Phylogenetic tree of "unknown" eukaryotic assemblies compared to reference *Malassezia* genomes. The inclusion of the *Malassezia* spp. and *Ustilago maydis* available reference genomes in the tree shows that unaffected and diseased skin is colonised by uncharacterised *Malassezia* and *Malassezia* spp. Marked with asterisk indicate the two *Malassezia* genomes reconstructed from the patient's 102 left ear. *Malassezia* clusters, 1–3, are congruent with those reported previously (Wu et al., 2015), with most of the *Malassezia* reconstructed genomes falling within cluster 2. **b** Phylogenetic tree of "unknown" bacterial assemblies in the *Peptostreptococcaceae* family. *Anaerococcus* spp. and *Finegoldia magna* reveal a novel *Anaerococcus* spp. on the ear of patient 9. **c** Phylogenetic tree of members of the *Chromobacteriaceae* (23 species) and *Neisseriaceae* (47 species) and "unknown" assemblies from Patients 9 and 101 are unable to be placed in either family.

5.5. Uncovering oral *Neisseria* tropism and persistence using metagenomic sequencing

This work focuses on the identification and characterization of Neisseriaceae strains from oral metagenomics samples. We included in the phylogenetic analysis the information about the eight DNA uptake sequences (DUSs) dialects, which are short DNA sequences highly repeated in the genome of species in this family. From the visualization of the DUSs copy numbers on resulting phylogenetic tree, it is immediately clear the association of the DUSs with the *Neisseria* subclades. In this work, I focused on the phylogenetic reconstruction and visualization of the Neisseriaceae family genomes used then as reference for the metagenomic analysis.

Donati C, Zolfo M, Albanese D, Truong DT, Asnicar F, Iebba V, Cavalieri D, Jousson O, De Filippo C, Huttenhower C, and Segata N

Uncovering oral Neisseria tropism and persistence using metagenomic sequencing

Nature Microbiology (2016)

Abstract

Microbial epidemiology and population genomics have previously been carried out near-exclusively for organisms grown in vitro. Metagenomics helps to overcome this limitation, but it is still challenging to achieve strain-level characterization of microorganisms from culture-independent data with sufficient resolution for epidemiological modelling. Here, we have developed multiple complementary approaches that can be combined to profile and track individual microbial strains. To specifically profile highly recombinant neisseriae from oral metagenomes, we integrated four metagenomic analysis techniques: single nucleotide polymorphisms in the clade's core genome, DNA uptake sequence signatures, metagenomic multilocus sequence typing and strain-specific marker genes. We applied these tools to 520 oral metagenomes from the Human Microbiome Project, finding evidence of site tropism and temporal intra-subject strain retention. Although the opportunistic pathogen *Neisseria meningitidis* is enriched for colonization in the throat, *N. flavescens* and *N. subflava* populate the tongue dorsum, and *N. sicca*, *N. mucosa* and *N. elongata* the gingival plaque. The buccal mucosa appeared as an intermediate ecological niche between the plaque and the tongue. The resulting approaches to metagenomic strain profiling are generalizable and can be extended to other organisms and microbiomes across environments.

5.5.1 Genome-wide phylogenetic analysis of neisseriae identifies a group of closely related species that colonize humans.

Several species of neisseriae are known to colonize the mucosa of the oropharynx of healthy individuals (Knapp and Hook, 1988). Although 16S rRNA-based taxonomy is unable to distinguish closely related species (Bennett et al., 2012), whole genome-based taxonomy has been shown to clearly classify the different species (Marri et al., 2010; Muzzi et al., 2013). We retrieved all the 241 draft and final genomes belonging to the Neisseriaceae family (**Supplementary Table 1**) and reconstructed their phylogeny using a concatenated alignment of 400 conserved proteins (Segata et al., 2013). The resulting phylogeny (rooted using *Chromobacterium violaceum*, a microorganism in the Neisseriales order but not in the Neisseriaceae family, **Fig. 1**) shows the occurrence of three major clusters of closely related species that are common colonizers of the oral mucosae.

The most basal cluster includes *N. sicca*, *N. mucosa* and *N. macacae*, the second cluster with intermediate branching includes *N. flavescens* and *N. subflava*, and the third most derived cluster includes *N. cinerea*, N. polysaccharea, *N. lactamica* and the opportunistic pathogen *N. meningitidis*. The latter cluster also includes *N. Gonorrhoeae*, which is known to be closely related to the meningococcus at the genomic level (Tinsley and Nassif, 1996). More basally branching is a monophyletic clade including *N. elongata* and *N. bacilliformis*, which are also known to colonize humans. Overall, our phylogeny is consistent with other studies (Muzzi et al., 2013), although the order of some rather deep branches can still be further investigated (Bennett et al., 2014). However, the tree's most external clades (from *N. sicca* to *N. meningitidis*) have high statistical support and are highly consistent with other studies. For this reason, and because oral neisseriae are almost exclusively in this subtree, we restrict our metagenomic study to this set of strains (see **Methods**).

Other species such as *Kingella* and *Eikenella* have been identified to be related to the human neisseriae at the genus or family level and can sporadically be found in humans. The genomes of these groups are characterized by specific forms of the DNA uptake sequences (DUSs), defined as 12 bp sequences that are repeated thousands of times in the genomes of neisseriae (**Supplementary Table 2**), with higher frequency in the core regions (Treangen et al., 2008), and that regulate the genomic integration of exogenous DNA by transformation (Frye et al., 2013) (**Fig. 1**). The available neisseriae spp. genomes and their reconstructed phylogeny constitute the base of our metagenomic strain-level investigation performed on the set of 520 shotgun metagenomic samples from the oral cavity sequenced by the Human Microbiome Project (HMP, (HMP et al., 2012; Human Microbiome Project Consortium, 2012)) and its recent follow-up phase.

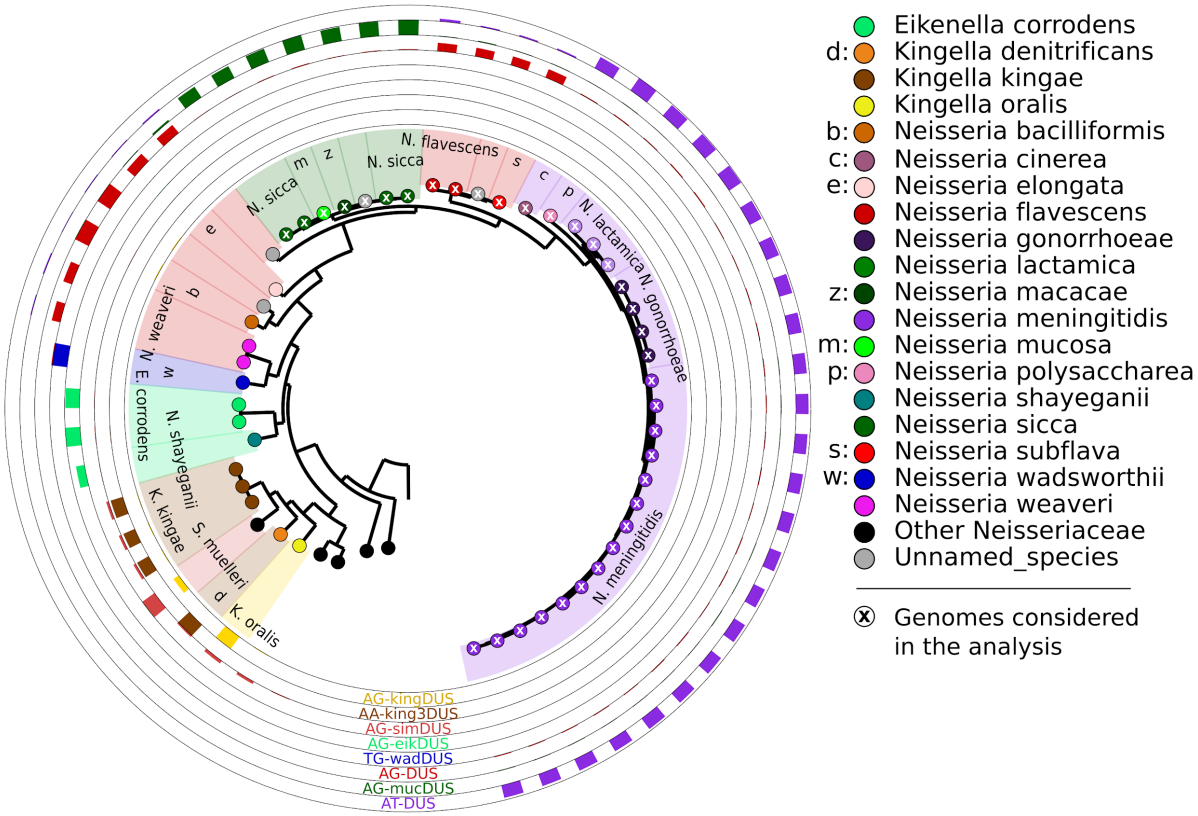


Figure 1 | The phylogenetic structure of the family Neisseriaceae identifies well-defined subtrees of closely related species. The phylogeny is constructed using all available genomes (non-draft genomes only for *N. meningitidis* and *N. gonorrhoeae*) by concatenating and aligning 400 conserved proteins automatically identified in the considered genomes (Segata et al., 2013). *Neisseria* strains have been shown to encode in their genomes many copies of short genetic features (12mers called DUSs) that regulate genomic recombination and are subclade- or species-specific. The relative copy numbers of the eight previously identified (Frye et al., 2013) DUS dialects are reported as external circular bar plots, confirming the univocal association between them and specific subclades of this family. An 'X' in the node indicates a genome used for the core genome-based metagenomic analysis applied for the strain-level characterization. Single letters are used for species names that could not be overlaid to the tree because of lack of space. The rings represent the different DUSs detected. GraPhlAn was used to visualize the tree and associated information with a circular layout (Asnicar et al., 2015a).

5.6. Large-scale metagenomic assembly reveals potential phylogeography and niche functional adaptations of *Eubacterium rectale* subspecies

Eubacterium rectale is an anaerobic, spore-forming, and short-chain fatty acid producing bacterium of the Lachnospiraceae family, and is among the most prevalent commensal species in the human gut. Previous short-read mapping-based analyses have provided evidence that *E. rectale* exhibits distinct genetic clustering as well as biogeographic stratification, with one cluster being exclusive to Chinese individuals and two others being exclusive to European and North American cohorts. Here we present a large-scale metagenomic assembly and binning approach on more than 6,000 gut metagenomes to reconstruct more than 1,300 high-quality *E. rectale* genomes (<400 contigs, estimated completeness >90%, and contamination <5% according to CheckM (Parks et al., 2015), and CMSeq[14] strain heterogeneity <0.3%). Importantly, we have added metagenomes from previously unobserved populations, mostly from non-Westernized cohorts from Oceania, Africa, and South America. In this work, I used PhyloPhlAn 2 to infer a large maximum likelihood phylogeny based on the specific set of core genes (1,071) of the 1,321 high-quality *E. rectale* genomes reconstructed from metagenomes. The resulting phylogeny was crucial for determining the biogeographic associations of the different *E. rectale* subtypes, highlighting the existence of a basal African subspecies and of an immotile European-specific clade, which is the only one lacking motility operons, probably as an indirect consequence of a change in its ecological niche. This work is not published yet, but we are submitting it for consideration for publication in a scientific journal.

Karcher N, Asnicar F, [..], Zeller G C, Segata N

Large-scale metagenomic assembly reveals potential phylogeography and niche functional adaptations of *Eubacterium rectale* subspecies

In preparation

5.6.1 A large-scale phylogeny refines *Eubacterium rectale* population genetics and biogeography

To study the global *Eubacterium rectale* population structure, we inferred a gene alignment of the 1,321 high-quality genomes we reconstructed from publicly available metagenomes, in combination with 8 publicly available isolate genomes and 3 additional isolates we sequenced for this work. The resulting concatenated alignment consists of 1,071 core genes for a total of 1.02M nucleotides. The maximum likelihood phylogeny based on a core gene alignment (**Figure 2A**) confirmed previous observations that *E. rectale* strains fall into genetically discrete groups (Costea et al., 2017; Scholz et al., 2016; Truong et al., 2017). Clustering using Partitioning Around Medoids on a random subset of samples with equal population densities (see **Methods**) supports the existence of four subspecies (Prediction Strength values of over 0.8 for k = 4, **Figures 2B-C**, **Methods**), one of which was not identified before (Costea et al., 2017; Scholz et al., 2016; Truong et al., 2017). The three previously reported subspecies consist predominantly of Eurasian, European, and East Asian strains, which we henceforth call ErEurasia, ErEurope, and ErAsia. The fourth, previously unobserved subspecies comprises strains coming from predominantly sub-Saharan African individuals (Madagascar, Tanzania, and Liberia) and was thus named ErAfrica (**Figures 2A-D**). All subspecies are monophyletic, except for one ErAsia strain in

[14] CMSeq repository: https://bitbucket.org/CibioCM/cmseq

ErAfrica and one ErEurasia genome in ErAsia. ErEurasia does not represent a well-defined phylogenetic clade, as there are seven ErEurasia strains within ErAfrica or ErAsia clades. ErEurope is instead an early-branching clade contained into the most recent common ancestor of the ErEurasia clade. The putative geographic origin of all isolate genomes, ten belonging to ErEurope or ErEurasia and one belonging to ErAsia, corresponds to their subspecies assignment. In summary, we are able to confirm and extend the notion about *E. rectale* geographic stratification using large-scale metagenomic assembly and binning.

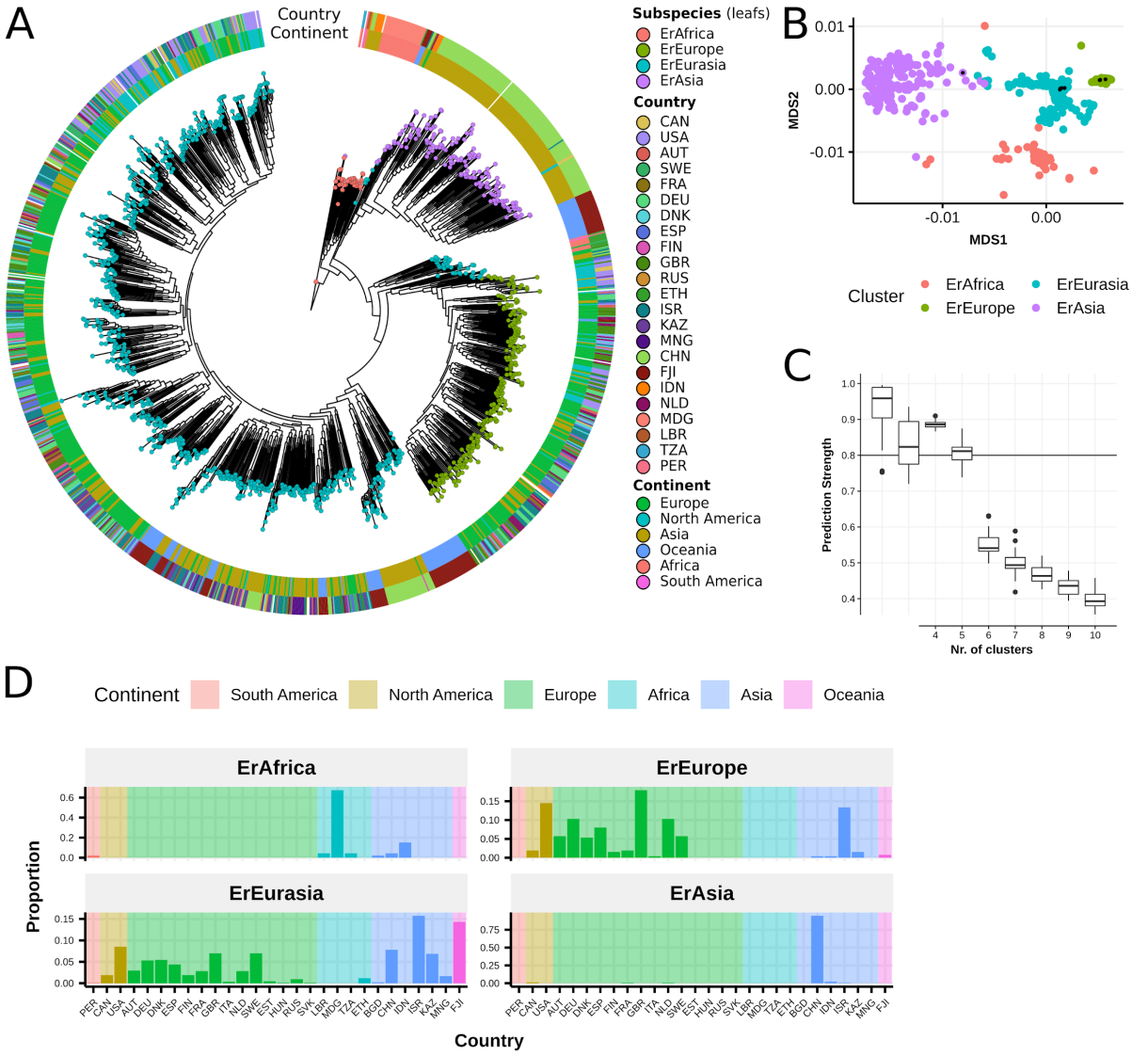


Figure 2: *Eubacterium rectale* can be divided into four geographically stratified subspecies. (A) Maximum-likelihood phylogenetic tree of all HQ genomes, built on a concatenated core gene alignment using PhyloPhlAn 2 (see **Methods**). The tree is rooted based on a phylogenetic tree built including E. rectale sister species (see **Methods**). **(B)** Subspecies assignment using Partitioning Around Medoids (PAM) clustering with k = 4 on data with equalized population densities (see **Methods**). Point colors correspond to leaf node colors in (A). Black points indicate genomes obtained from isolate sequencing. **(C)** Prediction Strength values for the different predesignated number of clusters (k, see **Methods**). **(D)** Subspecies composition with regards to country and continent.

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

In this work, we presented a very large-scale metagenomic assembly and binning approach to study unexplored microbial diversity in human microbiomes. By exploiting 9,428 publicly available metagenomes we were able to reconstruct 154,723 microbial genomes. All the reconstructed genomes were clustered together, at 5% average nucleotide identity, and recapitulated into 4,930 species-level genome bins (SGBs). We then divided the SGBs into two sets: 1,134 known SGBs (kSGBs) that are SGBs containing at least a reference genome, and 3,796 unknown SGBs (uSGBs) that do not contain any reference genome. My phylogenetic analysis pipeline has been extensively used in this work to place, characterize, and study the uSGBs with respect to the kSGBs for which we have a taxonomic label. We then focused on the most prevalent uSGB that we named "*Candidatus* Cibiobacter qucibialis" and contains 1,813 reconstructed genomes phylogenetically placed between *Faecalibacterium* and *Ruminococcus* genera. Phylogenetic analysis of this uSGB allowed us to observe geographical specificity of some strains, which were associated with the non-Westernized lifestyle. We used a phylogenetic approach to study both a kSGB and two uSGBs. In the first case, we focused on a kSGB belonging to the phylum *Succinatimonas*, including the only available reference genome. In the latter case, we phylogenetically characterized two uSGBs enriched in non-Westernized populations and assigned to the phylum *Elusimicrobia*, including also the available reference genomes belonging to the same phylum. From these phylogenetic analyses, we were able to observe that the clustering approach used is phylogenetically consistent and also that in some cases, intra-SGB diversity is associated with the non-Westernized lifestyle. Finally, we employed a phylogenetic approach to compare and evaluate at the strain-level resolution the genomes reconstructed with our metagenomic assembly and binning approach with the co-assembly and co-binning approach proposed by other methods. This work thus again confirms that the phylogenetic framework I developed is highly relevant for present and future metagenomic analyses.

Pasolli E, Asnicar F*, Manara S*, Zolfo M*, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N (* equal contribution)

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

Cell (2019)

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 (unknown SGBs [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.

5.7.1 Human Microbiome Genomes Belong to ~5,000 Functionally Annotated SGBs

To organize the 154,723 genomes into species-level genome bins (SGBs), we employed an all-versus-all genetic distance quantification followed by clustering and identification of genome bins spanning a 5% genetic diversity, which is consistent with the definition of known species (see **STAR Methods**) and with other reports (Jain et al., 2018). We obtained 4,930 SGBs from 22 known phyla (**Figure 1A**; **Table S4**). This is likely an underestimate of the total phylum-level diversity, because some SGBs are very divergent from all previously available reference genomes and cannot be confidently assigned to a taxonomic family (**Table S4**): 345 SGBs (58% of which with HQ or multiple reconstructed genomes) display more than 30% Mash-estimated genetic distance (Ondov et al., 2016) from the closest isolate with a phylum assignment (**Figure S2A**). The SGB genomic catalog spans on average 3.0%, SD 1.8% intra-SGB nucleotide genetic variability, and each SGB contains up to 3,457 genomes from different individuals (average 31.4, SD 147.6; **Figures 1C** and **S2B**).

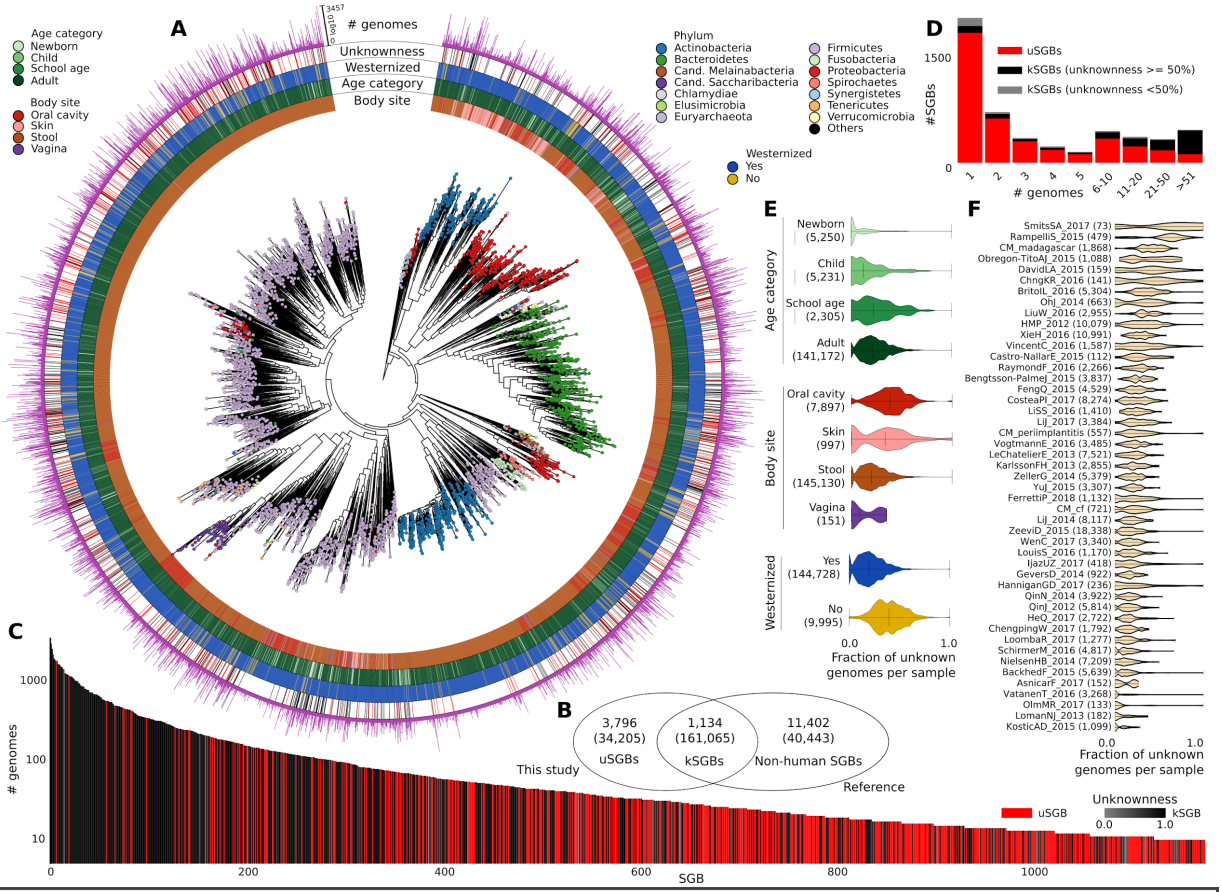


Figure 1. 4,930 species-level genome bins (SGBs) assembled from 9,428 meta-analyzed body-wide metagenomes. (A) A human-associated microbial phylogeny of representative genomes from each SGB. **Figure S3A** reports the same phylogeny but including isolate genomes not found in the human-associated metagenomes. **(B)** Overlap of

SGBs containing both existing microbial genomes (including other metagenomic assemblies) and genomes reconstructed here (kSGBs), SGBs with only genomes reconstructed here and without existing isolate or metagenomically-assembled genomes (uSGBs), and SGBs with only existing genomes and no genomes from our metagenomic assembly of human microbiomes (non-human SGBs). **(C)** Many SGBs contain no genomes from sequenced isolates or publicly available metagenomic assemblies (uSGBs). Only SGBs containing >10 genomes are shown. **(D)** Fraction of uSGBs and kSGBs as a function of the size of the SGBs (i.e. number of genomes in the SGB). **(E)** Distribution of the fraction of uSGBs in each sample by age category, body site, and lifestyle. **(F)** Distribution of the fraction of uSGBs in each study.

5.7.2 The Reconstructed Genomes and SGBs Increase the Diversity and Mappability of the Human Microbiome

We identified 3,796 SGBs (i.e., 77.0% of the total) covering unexplored microbial diversity as they represent species without any publicly available genomes from isolate sequencing or previous metagenomic assemblies (**Figures 1B** and **S3A**). These SGBs, that we named unknown SGBs (uSGBs), include on average 9.0, SD 45.4 reconstructed genomes, and 1,693 of them (45%) had at least one HQ genome. Recursive clustering of SGBs' representatives at genus- and family-level genetic divergence (see **STAR Methods**) provided taxonomic context for 75.2% of the uSGBs with 1,472 assignments to genera and 1,383 more to families (**Table S4**). The 941 uSGBs that were left unplaced at family level remained unassigned for limitations of whole-genome similarity estimates, but we report the similarity and taxonomy of the closest matching strain (**Table S4**).

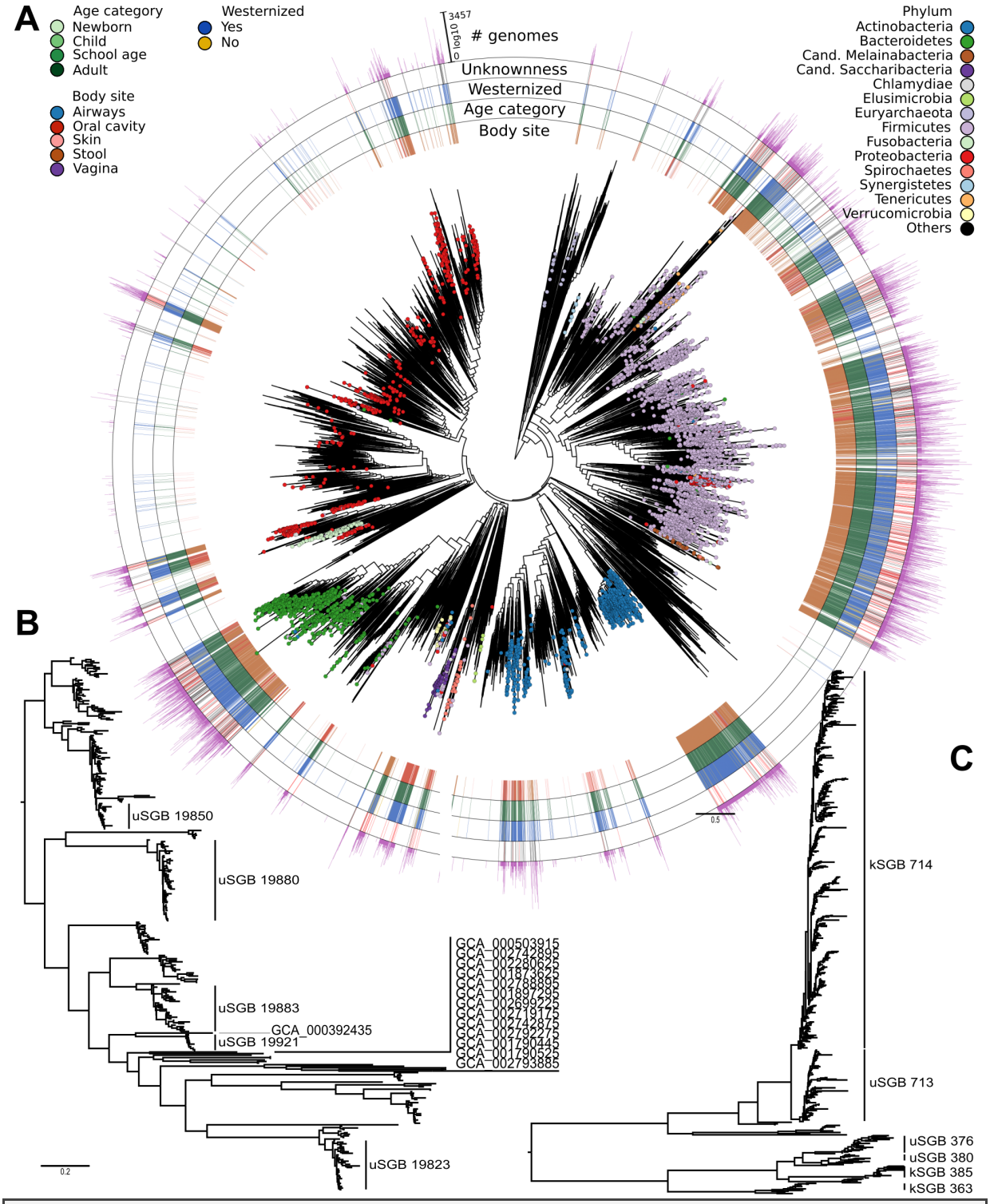


Figure S3. Related to Figure 1. Phylogenetic trees for all SGBs and reference genomes, and subtrees of Saccharibacteria and Archaea. (A) Phylogenetic tree that includes the representatives of the SGBs presented in **Figure 1A** together with all the non-human bins (represented in white in the external rings), for a total of 16,332 genomes (15,299 after the internal quality control in PhyloPhlAn). **(B)** Phylogenetic tree of the 337 reconstructed genomes taxonomically assigned to the candidate phylum Saccharibacteria present in the 108 SGBs, including available reference genomes (publicly available reference genomes are labelled with the "GCA" prefix). **(C)** Phylogenetic tree of the 675

archaeal genomes reconstructed in this study. 487 genomes belong to the *Methanobrevibacter smithii* kSGB (ID 714).

5.7.3 Several Prevalent Uncharacterized Intestinal Clostridiales Clades Occur Phylogenetically between Ruminococcus and Faecalibacterium

Some of the uSGBs with the largest number of reconstructed genomes are also highly abundant in the gut microbiome, with 1,153 uSGBs totaling >13,000 genomes each present in the sample where it has been reconstructed at an average abundance >1% (and 172 uSGBs at >5% average abundance). Among them, uSGB ID 15286, that we named "*Candidatus* Cibiobacter qucibialis", is the most prevalent uSGB, comprising 1,813 reconstructed genomes. This species is phylogenetically placed between *Faecalibacterium* and *Ruminococcus* (**Figures 3A** and **S5A**), key members of the gut microbiome that are typically present at comparably lower abundances (1.84% *Faecalibacterium* kSGB and 1.29% *Ruminococcus* kSBG in contrast to 2.47% *Ca*. Cibiobacter qucibialis). Six other prevalent (1,563 total genomes) and abundant (1.14% average abundance) SGBs occurred monophyletically in the same subtree between faecalibacteria and ruminococci (**Figure 3A**). Only one of these seven total SGBs contains an isolate genome, which is the recently sequenced *Gemmiger formicilis* genome (Gossling and Moore, 1975) included in kSGB ID 15300 (1,212 genomes, **Figures 3A** and **3B**). A genome from the *Subdoligranulum variabile* species, itself not found in any of the study's assemblies, was the only other reference phylogenetically close to this clade, explaining the previous identification of an unknown *Subdoligranulum* ("*Subdoligranulum* unclassified") as the most prevalent single taxon in reference-based profiles of the gut microbiome (Pasolli et al., 2017). This prevalent 7-SGBs clade comprising 3,370 reconstructed genomes that can be very abundant (>5% relative abundance in >200 samples) is thus an important but so far neglected genus-level lineage in the human microbiome.

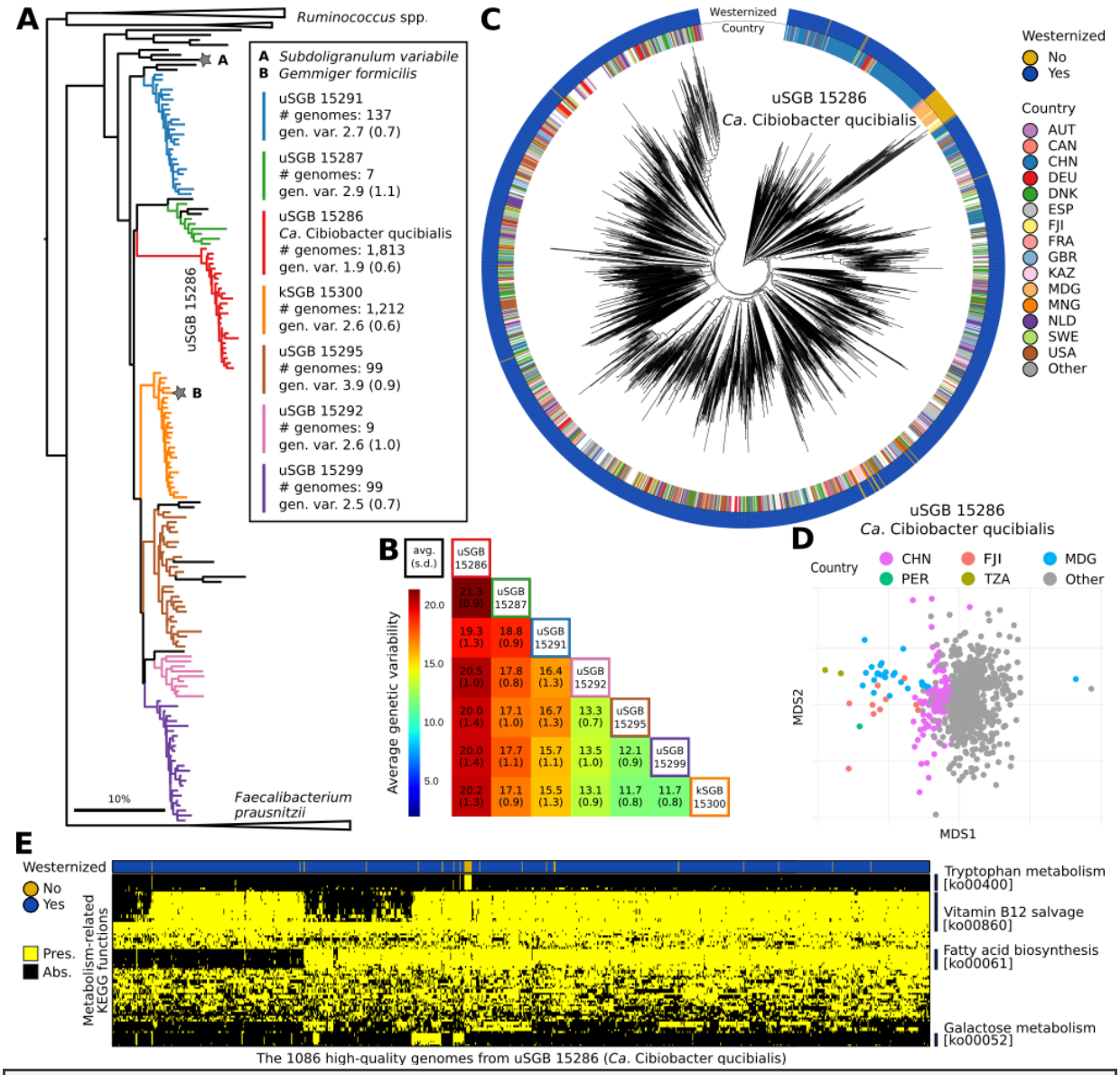


Figure 3. Several prevalent intestinal uSGBs are found within the Clostridiales order related to *Ruminococcus* and *Faecalibacterium*. (A) All SGBs in the assembled phylogeny (**Figure 1A**) placed between reference genomes for *Ruminococcus* and *Faecalibacterium* species that are reported as collapsed trees. A maximum of 25 HQ genomes from each SGB are displayed, and SGBs with less than 3 genomes are left black. **(B)** The monophyletic clade with the 6 uSGBs and the kSGB containing *Gemmiger formicilis* represent clearly divergent species with inter-species genetic distance typical of genus-level divergence (average 16.6 s.d. 3.1 nucleotide distance). **(C)** A whole-genome phylogeny for the 1,806 genomes in *Ca*. Cibiobacter qucibialis (**STAR Methods**). Some subtrees associate with geography and non-Westernized populations, while others seem to be geography- and lifestyle-independent (see text). **(D)** Multidimensional scaling of genetic distances among genomes of *Ca*. Cibiobacter qucibialis highlights the divergence of strains carried by non-Westernized populations, with Chinese populations subclustering within the large cluster of Westernized populations. **(E)** Madagascar-associated strains of *Ca*. Cibiobacter qucibialis (uSGB 15286) uniquely possess the *trp* operon for tryptophan metabolism (**Table S7**). Other functional clusters in Westernized strains from geographically heterogeneous populations include vitamin B12 and fatty acid biosynthesis, and galactose metabolism. The KEGG functions present in >80% or in <20% of the samples were discarded except for significant associations with lifestyle.

In an estimated maximum-likelihood whole-genome phylogeny of the 1,813 genomes belonging to *Ca*. Cibiobacter qucibialis (**Figure 3C**), genomes of non-Westernized populations were placed together in a monophyletic subtree (**Figure 3C**). This subtree included 26 strains from the Madagascar microbiomes we sequenced in this work, in addition to strains from three other populations with traditional lifestyles but differing geographic locations (**Figure 3D**). Although the non-Westernized subtree includes few genomes (2% of the total), this is a consequence of limited sampling from these population types because the prevalence of this SGB in Westernized populations is comparable (23% against 15% in non-Westernized populations). No clear internal clustering was evident for Westernized samples (**Figure 3C**), except for a large set of 222 samples retrieved from the seven Chinese cohorts that are monophyletically placed in the same subtree despite widely different pre-sequencing protocols (**Table S6**) and resemble non-Westernized genomes (**Figures 3C** and **3D**). This suggests a complex process of gut microbial ecological establishment in which both host lifestyle and biogeography play roles with comparable effect sizes.

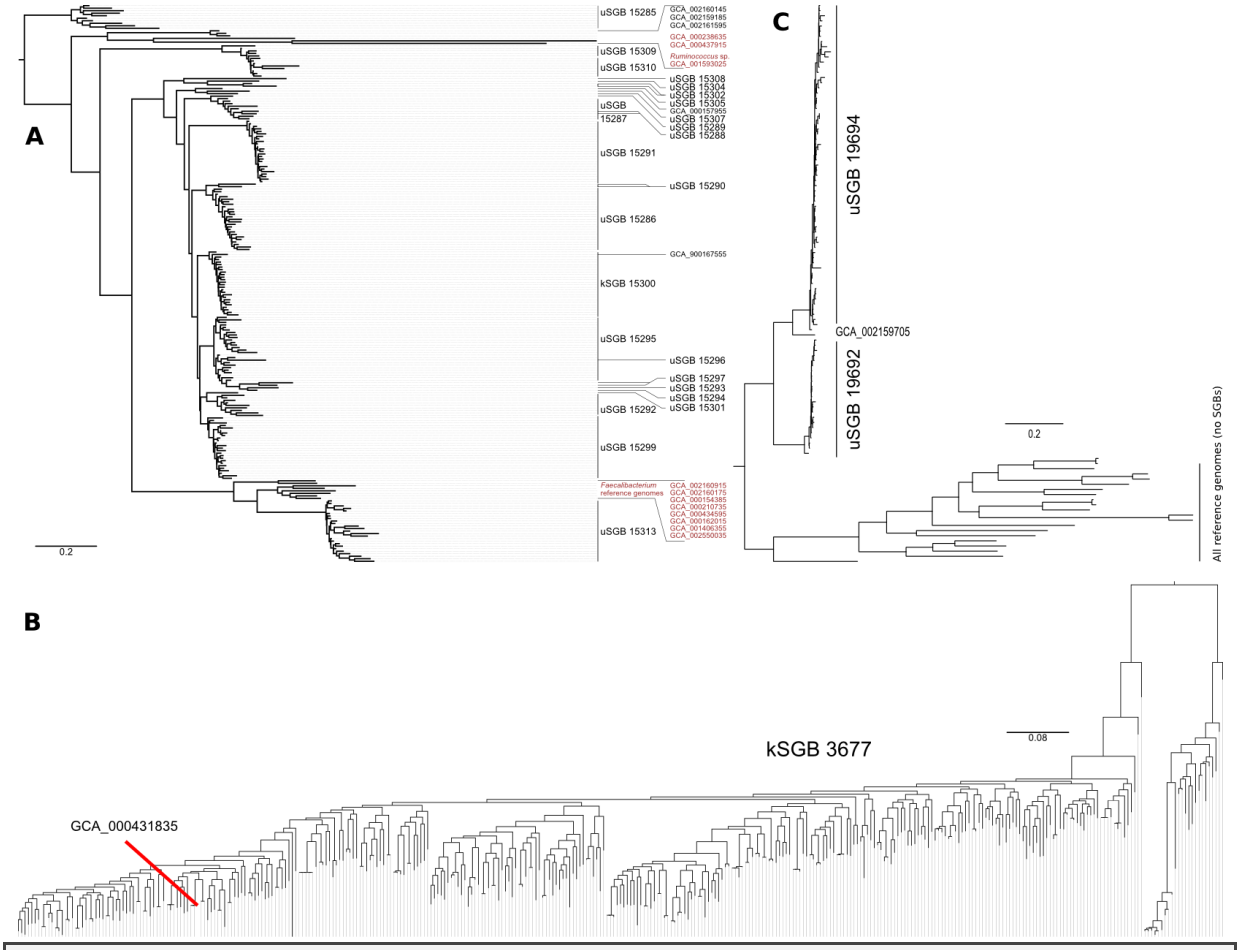


Figure S5. Related to Figure 3 and 5. Phylogenetic trees for SGBs placed between Ruminococcus and Faecalibacterium, Succinatimonas kSGB (ID 3677), and two Elusimicrobia uSGBs. (A) Phylogenetic tree of SGBs placed between reference genomes for *Ruminococcus* and *Faecalibacterium* species in **Figure 1A** (highlighted in red), as already reported in **Figure 3A** but without collapsed branches and including the two reference genomes GCA_000238635 and GCA_000437915 (also highlighted), originally labelled as *Subdoligranulum* sp. 4_3_54A2FAA and *Subdoligranulum* sp. CAG:314,

respectively. **(B)** Phylogenetic tree of the Succinatimonas kSGB (ID 3677) including the only available reference genome. **(C)** Phylogenetic tree of the two Elusimicrobia uSGBs enriched in non-Westernized populations and of all the available Elusimicrobia reference genomes.

5.7.4 Discussion

This work expands the collection of microbial genomes associated with the human microbiome by more than doubling the current collections with over 150,000 newly reconstructed genomes, in the process recovering hidden functional and phylogenetic diversity associated with global populations (particularly those that are undersampled from non-Western lifestyles and non-gut areas, **Figure 1E**). More than 94% of metagenomic reads can now be mapped to the expanded genome catalog for half of the gut microbiomes, enabling a much more comprehensive profiling of these communities. The metagenomic-assembly strategies employed here (Li et al., 2015; Nurk et al., 2017) represent a scalable methodology for very large-scale integration of metagenomes (**Figure 6**) that we extensively validated (**STAR Methods**; **Figures 7** and **S7**) and could be fruitfully applied to additional or non-human-associated metagenomes. The methods are also compatible with emerging technologies such as synthetic (Kuleshov et al., 2016) or single-molecule (Brown et al., 2017) long-read sequencing, which will further add to the diversity of microbial genomes. Finally, the study's results themselves emphasize the phylogenetic and functional diversity that remains to be captured from rare organisms, especially for sample types other than stool, global human populations, and varied lifestyles for the human microbiome.

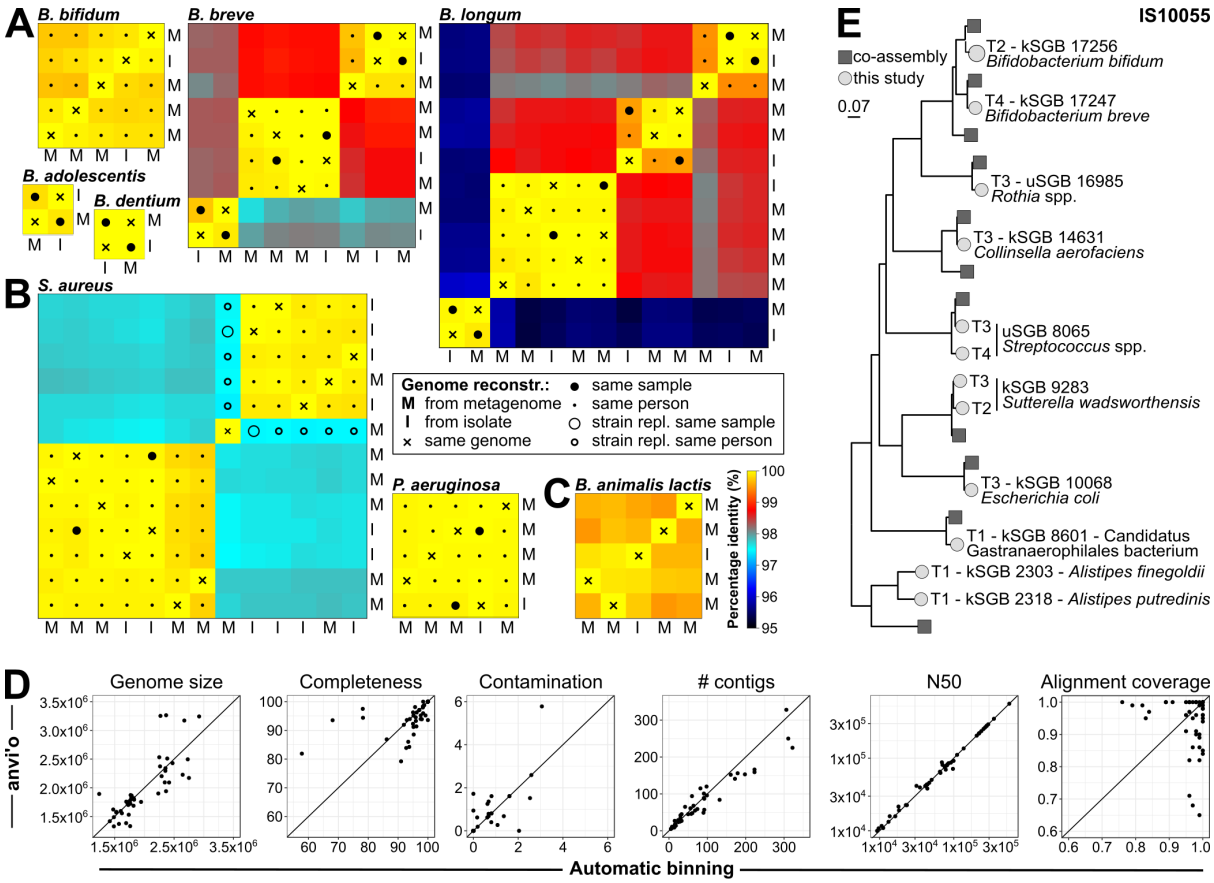


Figure 7. Quality of the single-sample assembled genomes against multiple alternative genome reconstruction approaches. (A) Percentage identity between genomes from isolates (I) and genomes we reconstructed from metagenomes (M) for 5

Bifidobacterium species from the FerrettiP_2018 dataset (Ferretti et al., 2018). We mark isolates and metagenomes coming from the same specimen (big filled circles), and coming from specimens of the same mother-infant pair (small filled circles). In all cases, our automatic pipeline reconstructs genomes from metagenomes that are almost identical to the genomes of the expected isolated strains. **(B)** The strains of *S. aureus* and *P. aeruginosa* isolated from three patients are almost perfectly matching the genomes reconstructed from sputum metagenomes sequenced at multiple time points. In the only case in which an *S. aureus* genome from a metagenome is not matching the strain isolated from a previous time point in the same patient, we verified with MLST typing that a clinical event of strain-replacement from ST45 to ST273 occurred. **(C)** In the dataset by (Nielsen et al., 2014), we successfully recover at >99.5% identity the strain of a *B. animalis* subspecies lactis present in a commercial probiotic product that was consumed by the enrolled subjects, even if the probiotic strain was at low relative abundance in the stool microbiome (<0.3% on average (Nielsen et al., 2014)). **(D)** Comparison of the 46 manually curated genomes (using anvi'o) with automatically assembled (using metaSPAdes) and binned (using MetaBAT2) genomes. **(E)** Example comparison between the set of single-sample assembled genomes and co-assembled genomes for a time series (n=5) of gut metagenomes from a newborn. Several genomes reconstructed with the two approaches have the same phylogenetic placement, with single-sample assembly retrieving the same (or a very closely related) genome at multiple timepoints, and both methods retrieving some unique genomes. This is an example of the comprehensive comparison performed in the **STAR Methods** and reported in **Table S2** and **Figure S7B**.

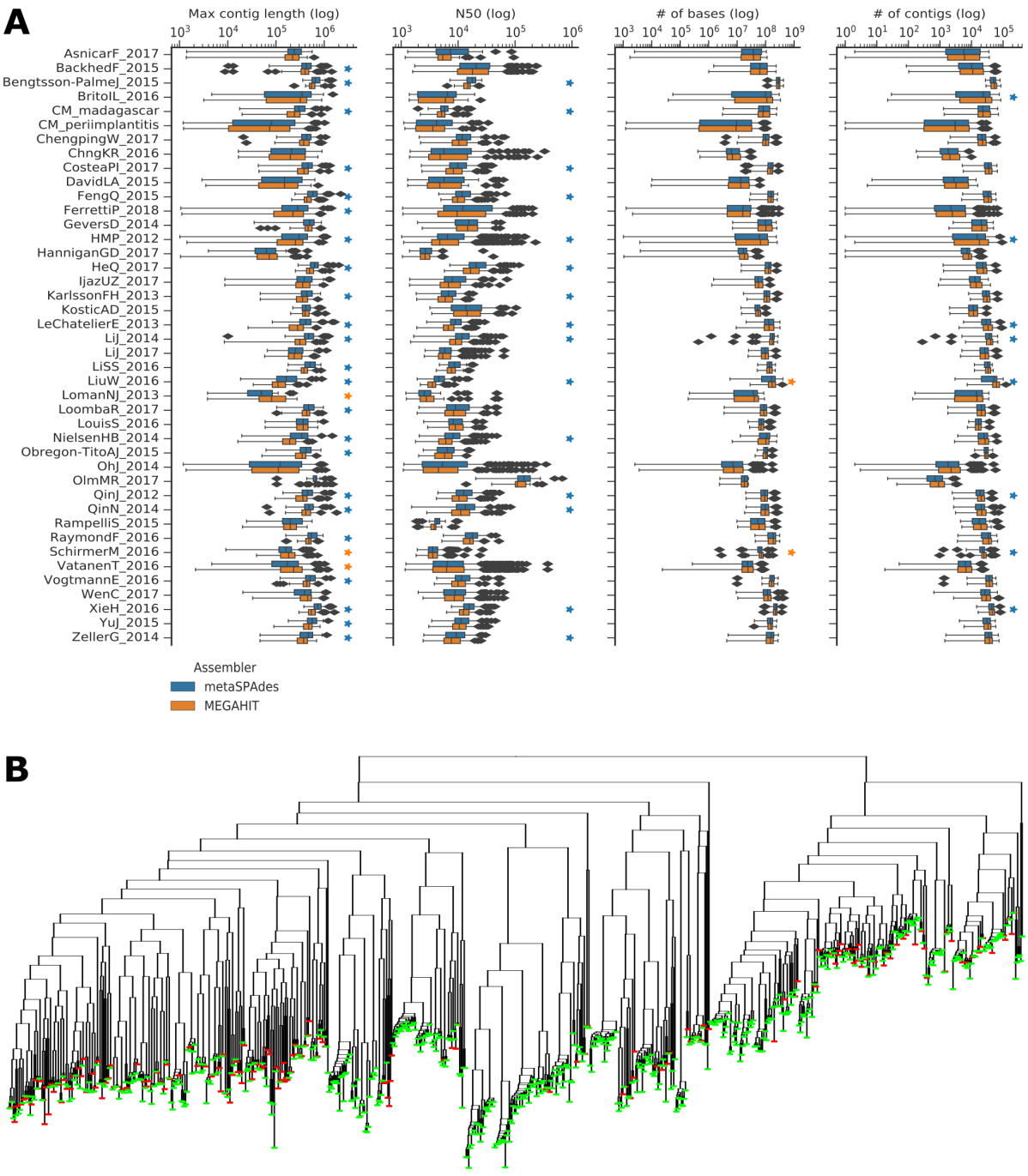


Figure S7. Related to Figure 7. Comparison between MEGAHIT and metaSPAdes assemblies and between assembly and co-assembly. (A) Comparison between metaSPAdes and MEGAHIT assemblers across all the considered datasets confirms that metaSPAdes performs consistently better especially in recovering long contigs. Stars indicate statistically significance (t-test, p-value <0.05). **(B)** Phylogenetic tree built on the genomes of gut adult metagenomes from 25 women from the FerrettiP_2018 dataset showing comparison between the set of single-sample assembled genomes (in green) and co-assembled genomes (in red). Several genomes reconstructed with the two approaches have the same phylogenetic placement, with single-sample assembly retrieving a total of 605 genomes spanning 257 SGBs, while co-assembly retrieved 172 genomes.

6. Other computational biology research

This chapter briefly introduces a slightly different panel of bioinformatics works, related to the gene network analysis that I started during my master and continued during my doctoral studies, supervised by Prof. Enrico Blanzieri at the Department of Engineering and Computer Science of the University of Trento.

The focus of this chapter is on the ability to identify novel putative interactions between genes belonging to a gene network and on the set-up of the computational infrastructure that exploits the distributed volunteer computing.

This chapter is organized as follow: I fully report two articles, the first one discusses the computational infrastructure for distributed volunteer-based computing (Asnicar et al., 2015b) and the second introduces a first gene network expansion algorithm (named NESRA) we proposed (Asnicar et al., 2015c). The last part of the chapter reports only the abstract of two other works (Asnicar et al., 2016, 2019) that are an extension of the NESRA (Asnicar et al., 2015c) algorithm.

6.1. TN-Grid and gene@home project: Volunteer Computing for Bioinformatics

This article introduces the TN-Grid BOINC platform. TN-Grid exploits the BOINC (Berkeley Open Infrastructure for Network Computing) framework for distributing the computations to the volunteers. The actual only BOINC project running in TN-Grid to date is named gene@home and its goal is a very large-scale gene network expansion analysis.

The section is based on the following article:

Asnicar F, Sella N, Masera L, Morettin P, Tolio T, Semeniuta S, Moser C, Blanzieri E, and Cavecchia V

TN-Grid and gene@home project: Volunteer Computing for Bioinformatics

BOINC:FAST 2015 (2015)

Abstract

The ability to reconstruct and find genes that belong or are connected to a gene regulatory network is of essential importance in biology, in order to understand how the biological processes of an organism work. The main limitation in performing gene network expansion is related to the huge amount of computations needed to discover new candidate genes. Given these premises we decided to adopt the BOINC platform that allows us to use the very powerful computational resources of the volunteers. We set up a BOINC server in which we developed a specific work generator that implements our gene network expansion algorithm. Furthermore, we developed an *ad hoc* version of the PC algorithm (PC++) able to run in the BOINC environment, on the client computers. We present and discuss some statistics and preliminary scientific results of the gene@home BOINC project, the first one hosted by the TN-Grid infrastructure.

6.1.1 Introduction

TN-Grid[15] has been thought and developed as a service platform, as a way to give to local research groups, in search of powerful computing infrastructures, a guided access to the power of the world-wide, volunteers based, distributed BOINC computing network. The idea was to use TN-Grid for informing people (researcher, technicians, and students) about BOINC, uncovering its strengths and weaknesses, discussing about integrating their own algorithms and their scientific pipelines into the BOINC framework, and eventually to help them doing this task. Providing access to such a big computational power may also help scientists to broaden their investigation outlooks, going to areas that would have been unfeasible to approach without it.
TN-Grid is the result of a joint effort made by two institutions of the Italian National Research Council (CNR), namely the Institute of Materials for Electronics and Magnetism (IMEM) and the Institute of Cognitive Sciences and Technologies (ISTC), both having local branches in Trento, Italy.
TN-Grid is a so-called *umbrella* project, which means that it is open to host different scientific projects even belonging to distant scientific areas. The first scientific project that we hosted is gene@home, that is a collaboration with Edmund Mach Foundation (FEM) and the Department of Information Engineering and Computer Science (DISI) of the University of Trento. We plan to add other projects to the system in the near future.
At the time of writing TN-Grid is the only public, BOINC based, active project in Italy.

6.1.2 Gene@home

The gene@home project is the first one hosted by the TN-Grid infrastructure. The project was born in the fall of 2013 with the collaboration of the students of the Laboratory of Biological Data Mining course. The gene@home project is multi-disciplinary that spans different disciplines: Computer Science, Biology, Statistics, and Data Analysis and can be also defined as a distributed computational biology project. The final goal is the creation of an automatic system for performing Gene Network Expansion in such a way that could be easily used by biologists through a web interface.
As described in the following Section 2.1, network expansion is a complex task aimed to discover relations among genes involved in a particular biological process. In our study, the task is performed by the PC-Iterative Method (PC-IM), using the PC algorithm (Spirtes et al., 2002) for inferring causal relations among genes. The biological species we studied so far are *Arabidopsis thaliana* and *Escherichia coli*, and we expanded 2 different local gene networks for the former and 13 for the latter.

6.1.2.1 Gene Network Expansion

Gene Network Expansion (GNE) is a research topic in Computational Systems Biology that deals with the discovery of functional dependencies within genes of a species, and genes that take part in the specific biological process to be studied. In biological processes, genes can act as enhancers or inhibitors of the activity of other genes, through a process named Regulation of Gene Expression. Regulation processes are represented by Biological pathways. Nowadays, we have an incomplete knowledge about pathways: discovering new genes is hence important for completing biological pathways, and therefore for gene-specific medical studies, fostering novel methods for pharmaceutical treatment (Arroyo et al., 2015).

[15] http://gene.disi.unitn.it/test/

Inputs of the network expansion algorithm are Omics data[16]. GNE differentiates from the most commonly used Network Inference (NI). NI reconstructs the complete set of gene interactions without the restriction of finding the ones that take part in a specific process, but with a not completely satisfying accuracy and sensitivity when analyzing single biological pathways. Our method for GNE, on the other hand, improves NI's results returning a ranked set of genes interacting with the local gene network of interest.

6.1.2.2 PC Algorithm

The PC algorithm (Spirtes et al., 2002) is a causal structure discovery method, that can be applied to find causal relations among variables of a system, when an input quantity data matrix representing the system entities is available. As scale-free networks, biological networks are characterized by a power law function on the degree of the nodes (Albert, 2005), and PC algorithm showed to be a valid method to test causal relations in sparse networks, as the biological ones (Maathuis et al., 2010).

A pseudo-code of the essential part of the PC algorithm is reported in **Algorithm 1**. At first, the PC algorithm creates a complete graph, assuming that all the variables are correlated with each other. Nodes of the graph correspond to data matrix variables, hence genes. Once created the complete graph, the algorithm tests the direct correlation between each pair of variables, removing non-correlating edges that do not present a statistically significant correlation, computed using Pearson's correlation coefficient. Then, the algorithm starts to condition all couples of variables X_i, X_j, with $i \neq j$ to all the sets S of neighbors of X_i, such that $S \in Neigh(X_i)$ and $|S| = l$, removing non-correlated edges when conditioned to the set S. This part is inserted in a loop, where the cardinality of S, called level l, increases at each cycle, up to $|Neigh(X_i)|$. This conditioning cycle is the most computationally expensive part of the algorithm. The number of sets of n elements, over a set of k elements (k-*Subset*) is given by the binomial formula $\frac{n}{k}$, that gives a factorial complexity to the algorithm.

Because the removal of edges depends on both input data and variables order (see Section 2.4), it is not possible to know in advance at which level the algorithm will halt: this means that it is not possible to exactly predict the execution time. In our experiments, however, it has never taken more than a few hours run-time on an ordinary laptop.

```
Data: T, Set of transcripts, E expression data
Input: Significance level α
Result: An undirected graph with causal relationship between transcripts Graph
G ← complete undirected graph with nodes in T;
l ← −1;
while l < |G| do
  l ← l + 1;
    foreach ∃u, v ∈ G s.t. |AdjG(u) ∖ {v}| ≥ l do // AdjG(u) adjacent nodes of u in G
        if v∈AdjG(u) then
            foreach A ⊆ AdjG(u) ∖ {v} s.t. |A| = l do
                if u,v are conditionally independent given A w.r.t. E with significance level
α then
                    remove edge {u, v} from G;
                end
            end
```

[16] Omics refers to many fields in Biology: Genomics, Transcriptomics, Metagenomics, Proteomics, Metabolomics. Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.

```
        end
    end
end
return G;
```

Algorithm 1: PC Algorithm: skeleton procedure (Kalisch and Buhlmann, 2007).

6.1.2.3 PC-IM Algorithm

The PC algorithm is just the core part of the method we used to discover candidate genes for expanding a gene regulatory network. The GNE task is performed by the PC-IM algorithm, which requires as input an already characterized Local Gene Network (LGN) and gene expression data (Coller, 2013). The PC-IM algorithm is said to be iterative because the analysis of the whole set of genes is computed multiple times, a parameter that we refer to as *iterations*. Each iteration is performed over a random permutation of the input variables, mitigating in this way the order-dependency issue of the PC algorithm.
Given a LGN, an observation data set, and the size of the graphs into which divide the set of genes of the organism, the PC-IM algorithm generates non- overlapping blocks of extra-LGN genes. To each of these extra blocks, the LGN genes are added. An additional extra block with partially overlapping genes may be added in the case that the data set is not a multiple of graph size minus the number of gene in the LGN. At this point, a single PC algorithm is executed for each block. This process is repeated for the number of iterations. The final output of this process is an ordered list of candidate genes found to be connected with the input LGN.

6.1.2.4 PC* Algorithm

The PC algorithm analyzes pairs of variables following the arbitrary order of the variables, in our case genes or microarray probes. If the variables are permuted, the output will change because when an edge in the graph is removed, its absence affects the future conditioning sets. In fact, when the execution removes an edge with conditioning sets of dimension l, it cuts away some conditional dependency to check with conditioning sets of the same dimension.
The PC* algorithm solves the order-dependent problem of the input, postponing the edge removal at the end of each loop, just before increasing the size of the conditioning sets. In more detail, at each level l edges are not removed from the graph, but instead they are marked as “to remove”. This allows the algorithm to check a larger space of possible conditional dependencies for a given size of the conditional set S. Since PC* algorithm checks many more conditions, its execution time takes much longer than a single PC run. From the tests we did, PC* returns a subset of the union set of outputs of multiple PCs.

6.1.3. BOINC

An execution of PC-IM requires, depending on the parameters, up to thousands of runs of the PC algorithm on input data of relevant size. This setting is particularly suitable for a BOINC project, for this reason we decided to implement the PC-IM method using the BOINC infrastructure (Anderson, 2004).

6.1.3.1 PC++ Algorithm, BOINC Version

Starting from the R implementation of the PC algorithm, included in the *“pcalg”* package (Hauser and Bühlmann, 2012; Kalisch et al., 2012), we implemented a C++ version of the

"skeleton" procedure of the PC algorithm, since we did not need the final DAG (Direct Acyclic Graph) estimation phase. We chose C++ because we needed high computational performances.
Our implementation showed an impressive speed-up of 240 in execution time, and conversely reducing the RAM consumption of about 10 times. We carefully optimized the most CPU demanding subroutines, i.e. the creation of all the $\frac{n}{k}$ subsets and the evolution of the causal dependency-testing formula, when the conditioning set is big. To solve these problems, we used more efficient data structures and switched from recursion to dynamic programming.

6.1.3.2 BOINC Server

The BOINC server is the part of the BOINC infrastructure that performs distribution management, data maintenance and project information visualization. Our BOINC server is running on a Virtual machine with 2 GB of RAM and two cores AMD Opteron™ Processor 6276. The basic components included in a BOINC server are:

- a database server (MySQL);
- the BOINC daemons (to name few of them: *scheduler*, *feeder*, *work generator*, *transitioner*, *validator*, *assimilator*, and *file deleter*);
- a web server (Apache).

The MySQL database stores the data related to the BOINC part of the project (e.g. users, computers, workunits, statistics). We made use of MySQL also to store the data regarding the dispatch and management of all PC-IM executions. This allows us to keep track of which workunits, input observation data and relative output files are related to a specific GNE task. Among all the BOINC daemons, the only one that we modified is the work generator. All the other daemons that are running on our BOINC server have not been modified.

6.1.3.3 Work Generator

The work generator generates the workunits. To easily manage and keep track of the PC-IM executions, we first designed and implemented a database (we will refer to it from now on as the *gene database*) using the already running MySQL daemon for BOINC. We also use the *gene database* to manage the input data, users, notifications, and it will be also the middle layer between the work generator and the future web-interface where biologists will schedule new PC-IM executions.
The work generator was implemented using the Python programming language, that allowed us a fast and high-level coding. We collected some measures about the performance of the work generator, such as the single workunit creation time and the overall workunits creation time necessary to complete a single PC-IM. Since we did not find any bottlenecks, we decided to not implement the work generator using more efficient languages. The work generator exploits our *gene database* to know and keep track of the work that will be generated or that has been generated, as well as the possibility to notify the user that submits the specific PC-IM when it is almost finished.
Since BOINC APIs are accessible only through C++ functions and not Python scripts directly, we built two simple C++ programs that wrap all the necessary BOINC functions for the work generator.
Since there is not a direct relation between the execution time of a single PC and the dimension of the input, PC time execution can largely vary. So, it's hard for the work generator to exactly estimate a workunit time execution. To overcome this issue, we are

planning to build a complete benchmark machine that will execute a few instances of PCs, eventually with different parameters, and use it to estimate the running time of the workunits.

6.1.3.4 Post Processing

The processing of the partial results of a PC-IM in order to get the candidate lists starts in the client application, as soon as a single workunit finishes. Indeed, since each workunit cannot contain a PC execution of a different PC-IM, we were able to insert a first partial counting of the arcs found by the PC executions of that workunit, reducing also in this way the size of the output file that the volunteers return back to the server.
The gene@home project has two issues that, in general, creates difficulties related to the BOINC distribution of work. The size of our input data files is generally in the order of one hundred MBs, and the size of our results averages a dozen MBs. Thanks to the developers of BOINC we got an update of the BOINC server that now implements the distribution and receiving of workunits and results in a gzip compressed format.
Since a single run of PC-IM can produce a very large number of workunits, we developed an *ad hoc* program that is in charge of moving the results of the workunits of a PC-IM, when all of them are returned. The script that moves the results exploits the *gene database*, where for every PC-IM executed by the work generator, we store the number of workunits that has been produced.
On the server, a validation step of the returned workunits is performed, and then the canonical result is moved on a dedicated server, that has a large store capacity. Currently we are using a double validation method, that consists in sending each workunit to two diverse volunteers in order to be able to validate the results later. The returned results then must be equal bitwise. Because of the nature of our project, we have not find a way to internally validate a result of a single workunit, without requiring the double validation phase.
Externally to the TN-Grid infrastructure, we have a pipeline of Python programs that complete the processing of the partial results of each PC-IM.

6.1.4. Educational and Social Aspects

The gene@home project was born from a conversation between Prof. Enrico Blanzieri (University of Trento), Dr. Valter Cavecchia (CNR), and Dr. Claudio Moser (FEM). Its realization and running involved students and BOINC volunteers.

6.1.4.1 Gene@home as a Course Project

In the first semester of academic year 2013-2014, the project was proposed to the students of the Biological Data Mining Laboratory course held in the master Computer Science program of the University of Trento. Claudio Moser and his collaborators at Foundation Edmund Mach provided biological annotated data and a preliminary version of the method implemented in R. Claudio Moser also covered the relevant biological topics during the course. The initial ambitious goal set was to systematically expand networks of interest of *Arabidopsis thaliana*. The attendance of the course increased steadily and eventually 22 students formed four groups devoted respectively to: 1) write the BOINC application; 2) manage the BOINC server; 3) preprocess the input data and post-process the results and 4) take care of communication and of the web site.
Students, now the first five authors of this paper, developed a C++ application for directly communicating with the BOINC client through the provided BOINC API. The executables were built for different operating systems and architectures, Windows (both 32 and 64 bit),

Linux (both 32 and 64 bit), and Mac OS X. After having a first version of the server and the client applications, the students tested the whole system, finally concluding the pre-alpha stage. The same students continued, in the form of a Research Project the activity in the second semester and gained extra academic credits.
The continuation of the BOINC project was proposed also during the first semester of the academic year 2014-2015 and one student, Stanislau Semeiuta, with the CUDA implementation of PC* while others worked on the application of gene@home to *E. coli*.
Overall, the class reached almost all the technical goals, with a large part of the students really engaged and who expressed a positive evaluation of the course with the exception of a small minority. Introducing BOINC in the teaching activity involved the students on several topics of distributed systems and it proved to be a good way of gaining technical and collaborative competences in a medium-size software project. Moreover, the students shared the general research goals and many of them worked beyond expectation.

6.1.4.2 BOINC Community

We contacted the administrators of the largest Italian BOINC users community (BOINC.Italy[17]) announcing the second, alpha phase of our project and asking them and their users to join us using the BOINC invitation code mechanism. This procedure implies registering the user on the projects web page before attaching the client to the project, also passing through CAPTCHA verification. This will filter the server from spammers and bots, minimizing the burden of the maintenance tasks.
After some time, some of the most active users in the BOINC world contacted us asking information about our project. We decided then to send the invitation code to anyone interested, explicit allowing them to re-distribute the invitation code to other people, but not to publishing it in public posts. Until today this rule was fully respected. Some statistics sites, e.g., BOINCstats[18] and Free-DC[19] also requested permission to collect and manage statistics data from our server. So, by providing a continuous flow of workunits, we started to see an increasing number of participants (see **Figure 1A**).
The computational power provided by the volunteers increased, reaching a peak of 1.5 gigaflops during December 2014 (see **Figure 1B**).
However, the credit per day, which is a good estimate of the “instant” power of the system, is recently (at mid February 2015) decreasing (see **Figure 1C**). From this chart, we may also notice that the average power (recent average credit, averaged over a week period, RAC) is also decreasing. There are many possible reasons for this:

- Most of the users are power users, that are users which provide high computational power. However, power users also like to frequently switch their computational power to different projects, pursuing their own interests, challenges, credit milestones, and badges. At the time of writing we count 175 registered users and 821 registered computers, with an average of 4.7 computers per user, with powerful machines running 24/7;
- We were not very informative about the status of our project and also news were issued rarely, losing our own appeal. Now, we need to keep the interest high, providing more frequent status updates and general information about our progresses.

[17] http://www.boincitaly.org
[18] http://boincstats.com
[19] http://stats.free-dc.org

In the forthcoming future, we will move to a new phase, switching to a semi-public stage, posting the invitation code to the project's home page. We are also designing credit badges, that can attract volunteers interested in collecting achievements.

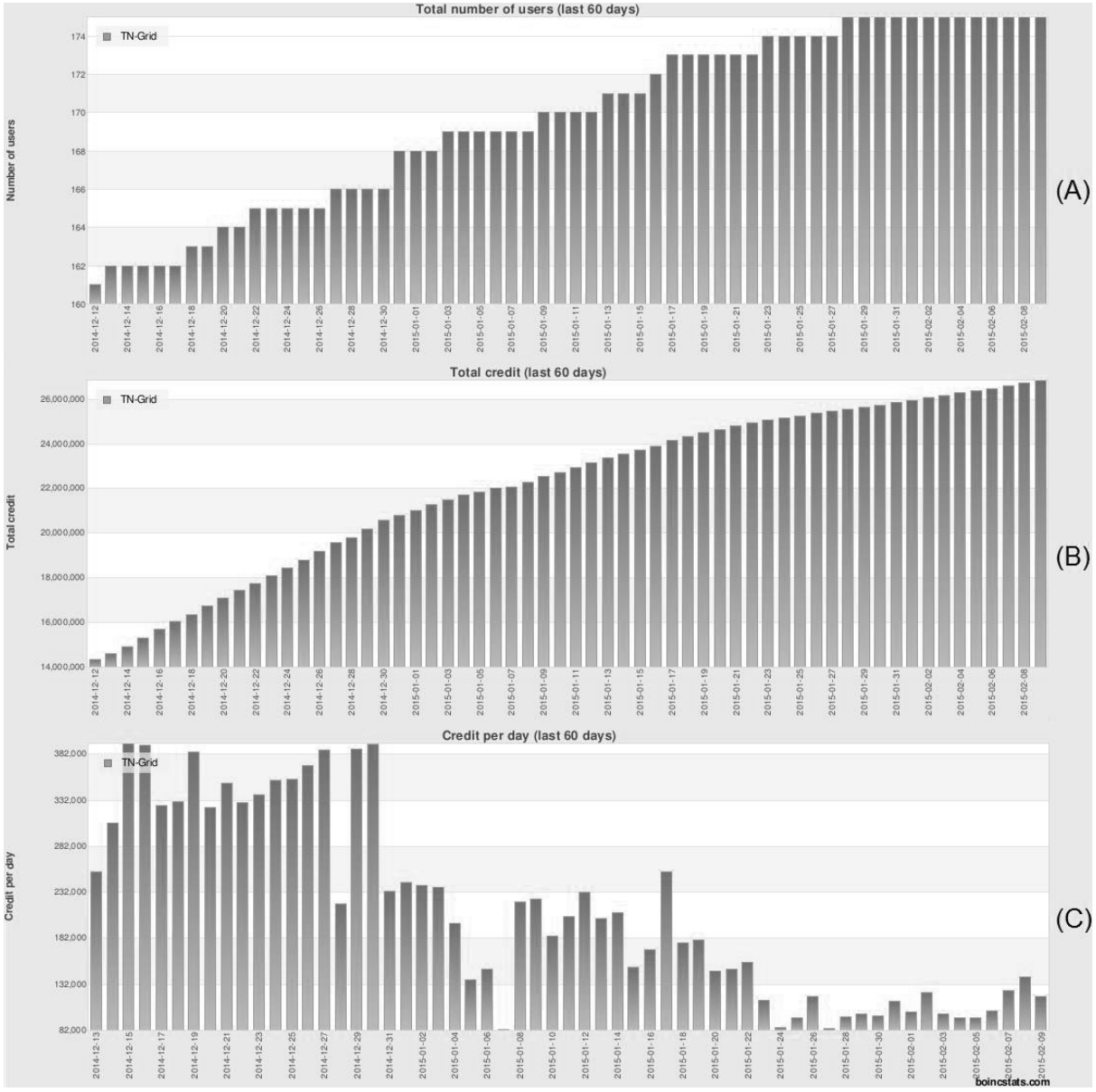


Figure 1: Snapshots were taken on February 10th (source BOINCstats). (A) Total number of users on gene@home during the last 60 days. (B) Total credit production (cumulative) on gene@home during the last 60 days. The credit trend depicted above is proportional to the flops values (1 gigaflop machine, running 24 hours a day, produces 200 units of credit in 1 day). (C) Amount of assigned credits per day on gene@home during the last 60 days.

6.1.5. Results

During the execution of the gene@home project we collected two types of data: statistical data about the performance of the BOINC server and application, and scientific results obtained after having analyzed the results of the workunits computed by the volunteers.

6.1.5.1 BOINC Results

After running the server for some months, we were able to collect some statistics about the reliability of the gene@home application, i.e. the ratio between valid and invalid results. This impacts both the scientific quality and the user experience of our project. Users really dislike running an application who ends up producing errors, thus not getting credits for the work done and wasting time and electric power: they could leave the project.
Statistics are automatically prepared by the BOINC server every day (see **Figure 2**), the data are taken as snapshots of the BOINC workunits database. In the summary **Table 1**, errors count includes only the computational errors: compute errors, validated errors, and invalid results. Results returned after the deadline or aborted by users, or download errors were not considered as errors.

Table 1: BOINC statistics of the gene@home project taken on four different days during the year 2014 (reference period: the previous seven days). Total number of returned results (#Over), successfully computed (Success), validated (Valid), pending validation (#Initial), and faulty (#Errors).

Date	#Over	Success (%)	Valid (%)	#Initial	#Errors
22 April	15543	15392 (99.0%)	15340 (98.7%)	19	85
20 May	69536	68621 (98.7%)	67096 (97.7%)	1450	89
16 December	33232	31798 (96.2%)	29525 (88.8%)	2147	38
24 December	91315	89536 (98.1%)	87584 (95.9%)	1716	61

Results presented in **Table 1** prove that our application is very reliable, although we still have some validation issues. We are still having, although rarely, validation problems (see **Figure 2**). Computers running the same task may return different results due to incorrect client software configuration. Otherwise, the problem could arise from a small bug inside the application checkpoint mechanism linked to a stop-and-restart after the very first seconds of running. We still need to further investigate this issue.
We are distributing executables built for various operating systems and architectures; namely Windows (both 32 and 64 bit), Linux (both 32 and 64 bit), and Mac OSX. We need to build statically linked executable for Linux for users running very old Linux kernels.
Handling of so-called *leftovers*. We distribute work in “batches”, i.e. sets of workunits belonging to the same sub-problem (a single run of a PC-IM). Sometimes it happens that almost all the workunits of the same set are returned and very few of them are not processed and waiting for timeouts before being re-distributed. We would like to find a way to compute this kind of workunits on a dedicated server in order to reduce the time needed to complete a PC-IM. One possible solution would be to use BOINC `restrict_wu_to_user()` mechanism to send all such workunits to a reliable, dedicated, and active user.

TN-Grid: Result summary

95988 results

Server state	# results
Inactive	0
Unsent	1080
Unknown	0
In progress	3593
Over	91315

'Over' results

Outcome	# results
---	0
Success	89536
Couldn't send	0
Computation error	1104
No reply	570
Didn't need	12
Validate error	0
Abandoned	93

'Success' results

Validate state	# results
Initial	1716
Valid	87584
Invalid	51
Workunit error - check skipped	0
Checked, but no consensus yet	10
Task was reported too late to validate	175

File Delete state	# results
Initial	1726
Ready to delete	0
Deleted	87810
Delete Error	0
Total files deleted	87810

'Client error' results

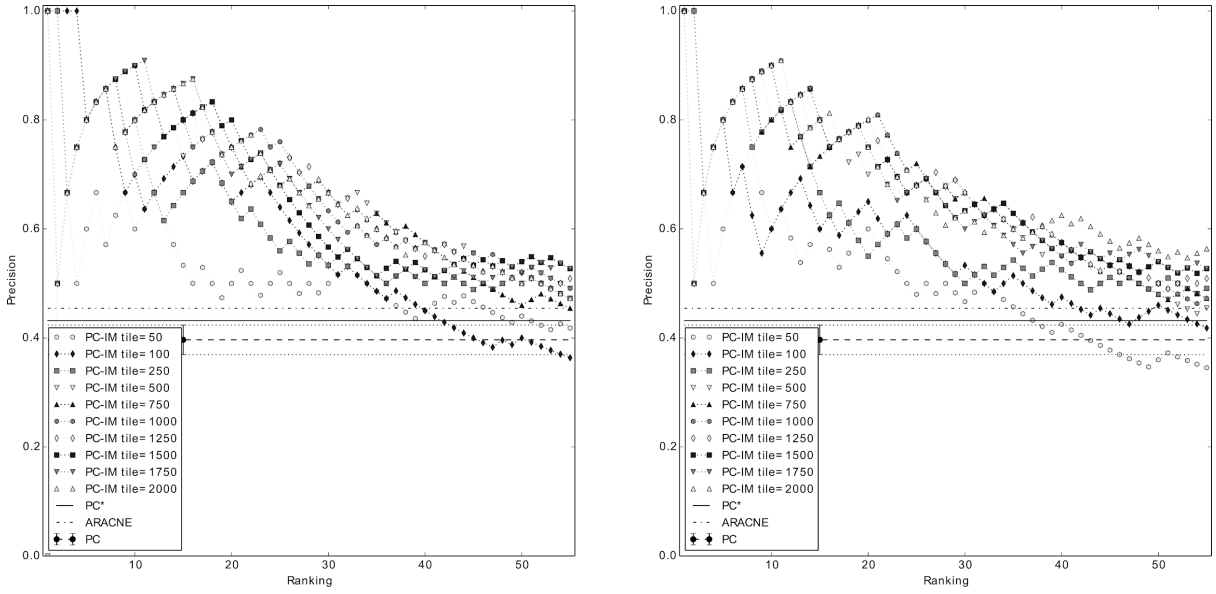
Client state	# results
Downloading	21
Processing	0
Compute error	10
Uploading	0
Done	10
Aborted by user	1063

Figure 2: Summary of the server statistics of the TN-Grid infrastructure running the gene@home project. Seven days snapshot, taken December 24th, 2014.

6.1.5.2 Preliminary Report on the Scientific Results

The experiments in **Figure 3** were conducted using the Flower Organ Specification Gene Regulatory Network (FOS) of the model plant *A. thaliana*, composed of 15 genes connected with 54 edges (Espinosa-Soto et al., 2004; Sánchez-Corrales et al., 2010). The data is composed of gene expression values available in the PLEXdb database (Dash et al., 2012) and consists of 393 hybridizations of 22,810 microarray probes[20]. Plots in **Figure 3** represent the trend of the precision of several PC-IM executions, ran with different values of tile size, and $\alpha = 0.05$. The precision is computed by comparison with a manually curated classification of the probes of *A. thaliana*. The comparison between the two plots permits to appreciate the change in precision by varying the iteration parameter i. The two plots show also, as a comparison reference, the precision computed on the results of the three major competitors: PC, PC*, and ARACNE (Margolin et al., 2006a, 2006b) (using the default parameters), the complete scientific results are in the process of being published.

[20] Probe is a general term for a piece of DNA or RNA that corresponds to a gene or a sequence of interest that has been labelled by biologists.



(a) i = 20 (b) i = 2000

Figure 3: Precision comparison of *A. thaliana* on FOS network. PC (average and variance), PC*, ARACNE, and PC-IMs with different tile size with a fixed $\alpha = 0.05$. We considered the first 55 genes of the result lists of each experiment and for each result lists we computed 55 precision values by considering an increasing list that initially contains the first gene found. All PC-IMs in Figure 3a have a number of iterations $i = 20$, in Figure 3b the number of iterations $i = 2000$.

6.1.6. Ongoing Developments

PC-IM performs a lot of computation, and even if our C++ implementation of the PC algorithm is really fast, we tried to achieve even better performances. For this reason we tried also solutions exploiting multithreading and Graphics processing unit (GPU) computing.

6.1.6.1 From Multithreading to GPU Computing

One of the goals is the performance improvement of a single PC run: this would help the PC-IM to be runnable on standard local machines. We opted for code parallelization. After some analysis, we concluded that the *skeleton* procedure is not trivially parallelizable, due to the edge removal-associated consequences that require complicated synchronization strategies to create a parallel version equivalent to the single threaded one. Instead, PC* is trivially parallelizable.

Initially, we introduced multithreaded processing using the Intel Threading Blocks (TBB) library[21] to parallelize PC*. In our implementation, we launch a number of threads, each taking as input one gene and the separation set size, that search among all genes for those ones that are conditionally independent given a size-specific separation set. Once checked all the pairs for one gene, it goes to the next unprocessed one. As edge removal is postponed, there is no synchronization between processing parts of threads. The only place that needs synchronization is the mutually exclusive list of unprocessed genes, whose access time is negligible with respect to the time to process gene expression data.

[21] https://www.threadingbuildingblocks.org

Table 2: Time (reported in milliseconds) to process one level of PC*. Columns marked with **CPU** report timings of 4-threads execution of the algorithm. Organisms: **At** stand for *Arabidopsis thaliana* and **Ec** for *Escherichia coli*, respectively.

	CPU	GPU	CPU	GPU	CPU	GPU	CPU	GPU
Tile size	**1000**	**1000**	**2000**	**2000**	**100**	**100**	**200**	**200**
Organism	**At**	**At**	**At**	**At**	**Ec**	**Ec**	**Ec**	**Ec**
Separation set size								
0	47	5	300	9	<1	<1	<1	<1
1	2940	600	18000	1350	<1	<1	80	3
2	1180	3100	9000	8000	90	40	890	320
3	68	100	600	220	320	100	2950	980
4	10	44	100	90	580	190	5330	1630
5	15	44	15	100	500	220	5515	2380
6	10	44	15	74	490	290	4437	3170
7			10	74	390	390	3690	4975
8			10	74	230	490	2500	6630
9					93	430	1800	8380
10							650	7020
11							230	5840
12							80	3950
13							15	2260

The final implementation produces exactly the same results as the single threaded PC*, but much faster. In our experiments, we have observed that it take less time by a factor of T to get the results, where T is a number of processing threads.
Then we decided to move to GPU computing, choosing to use NVIDIA CUDA for its better development tools with respect to OpenCL. The algorithm is conceptually the same as the TBB-based, but it has to take into account the need of transfer data between CPU and GPU memories, the differences in computation models of GPUs and CPUs, and the fact that single GPU threads are slower than single CPU ones.
Our NVIDIA GPU-based implementation showed to be coherent with the previous ones, so we evaluated its pros and cons. The most important timings are presented in **Table 2**. It can be seen from both tables that the GPU version significantly outperforms the CPU one on small sizes of separations sets. As we were using the parallel implementation of PC* on a 4 core machine, that gives approximately a speed-up of 50 for separation set size of 0 and 1 with respect to the initial single core version. We also observe that the performance boost depends on the nature of data being processed. **Table 2** shows that, for this particular data, it is beneficial to run the GPU version up to a separation set size of 4-6, while it is not the case with the other data that we have tested.

6.1.7 Conclusion

The first project hosted on the TN-Grid infrastructure is gene@home that, involving volunteer computing, implements a gene network expansion algorithm. We presented our project from different points of view: the technical side in which we described the implementation and

setup of the BOINC platform, the educational aspect that involved students of the University of Trento, and the social part that involves the relationship with the BOINC community of volunteers participating in our project.
We gave some details of the way we setup our BOINC server, the server software, and the features of the client application that we developed. The gene@home project started during the Laboratory of Biological Data Mining course in 2013-2014 at the University of Trento and engaged a group of students to what is now a long-term project. The social aspect of the participation in the gene@home project by BOINC users is important, and we discussed in particular the trend of the points assigned to the volunteers. In fact, gene@home initially was very attractive thanks also to the novelty of the problem, now we realized that we need to communicate the results in a steadier way in the near future.
The empirical data on gene@home comprises both statistical results of the BOINC performance that we obtained during the last year of executions, and a preliminary report of the scientific result that shows the effectiveness of the method.

Acknowledgments

The authors wish to thank Giulia Malacarne and Emanuela Coller of Edmund Mach Foundation, Daniele Campana, Giulia Corn, Laura Escobar, Ahmed Fadhil, Marco Giglio, Davide Giovannini, Bhuvan Hrestha, Erinda Jaupaj, Paolo Leoni, Laura Malvaso, Trung Nguyen, Quiynh Nguyen, Eko Susilo, Daniele Tovazzi, Chau Tran, and all the volunteers, in particular the BOINC Italy group.

6.2. Discovering Candidates for Gene Network Expansion by Distributed Volunteer Computing

We have now introduced the gene@home BOINC project hosted inside the TN-Grid infrastructure. One of the algorithms for the gene network expansion that are distributed and runs on the volunteers' computers is named NESRA and has been presented in the following article.

The section is based on the following article:

Asnicar F, Erculiani L, Galante F, Gallo C, Masera L, Morettin P, Sella N, Semeniuta S, Tolio T, Malacarne G, Engelen K, Argentini A, Cavecchia V, Moser C, and Blanzieri E

Discovering Candidates for Gene Network Expansion by Distributed Volunteer Computing

2015 IEEE Trustcom/BigDataSE/ISPA. Vol. 3. IEEE (2015)

Abstract

Our group has recently developed gene@home, a BOINC project that permits to search for candidate genes for the expansion of a gene regulatory network using gene expression data. The gene@home project adopts intensive variable-subsetting strategies enabled by the computational power provided by the volunteers who have joined the project by means of the BOINC client. Our project exploits the PC algorithm (Spirtes and Glymour, 1991) in an iterative way, for discovering putative causal relationships within each subset of variables. This paper presents our infrastructure, called TN-Grid, that is hosting the gene@home project. Gene@home implements a novel method for Network Expansion by Subsetting and Ranking Aggregation (NESRA), producing a list of genes that are candidates for the gene network expansion task. NESRA is an algorithm that has: 1) a ranking procedure that systematically subsets the variables; the subsetting is iterated several times and a ranked list of candidates is produced by counting the number of times a relationship is found; 2) several ranking steps are executed with different values of the dimension of the subsets and with different number of iterations producing several ranked lists; 3) the ranked lists are aggregated by using a state-of-the-art ranking aggregator. In our experimental results, we show that NESRA outperforms both the PC algorithm and its order-independent version called PC*. Evaluations and experiments are done by means of the gene@home project on a real gene regulatory network of the model plant *Arabidopsis thaliana*.

6.2.1 Introduction

Gene expression data are accumulating at an increasing pace and also resources that integrate different data sources are now available, for example Colombos (Meysman et al., 2014). The characterization of these causal relationships between the gene expression levels are not yet well known, even when considering model organisms. These information can be organized in gene regulatory networks (Hasty et al., 2001). In biological research, it is common to take into consideration prior knowledge about the phenomenon under consideration. In this scenario, methods that can guide the research suggesting candidate genes which could regulate, or could be regulated within a given gene network, are of essential importance. An expansion method can guide the discovery of candidate genes that could be causally connected to a priori known network. This is particularly important when considering a gene network that by knowledge or by hypothesis biologists assume to be

relevant. For instance, we can consider the Genetic Network Expansion System (GENESYS (Tanay and Shamir, 2001)).
The PC algorithm (Spirtes and Glymour, 1991), which name derives from the initials of its authors, is an algorithm that discovers causal relationships among variables. In particular, the PC algorithm is based on the systematic testing for conditional independence of variables given subsets of other variables. It has been comprehensively presented and evaluated by (Kalisch and Buhlmann, 2007) who proposed it also for gene network reconstruction purposes (Maathuis et al., 2010). For this task, some modifications of the original formulation of the PC were also proposed (Tan et al., 2008, 2011; Wang et al., 2010; Zhang et al., 2012). Other methods used for gene network reconstruction comprises: the Algorithm for the Reconstruction of Accurate Cellular NEtworks (ARACNE (Margolin et al., 2006a, 2006b)), the Bayesian Network inference with Java Objects (BANJO (Hartemink, 2005)), and Network Inference by Reverse-engineering (NIR (Gardner et al., 2003)). Allen and colleagues (Allen et al., 2012) have recently compared ARACNE with other competitors in the task of large scale networks reconstruction and ARACNE proved to be a state-of-the-art method.
The task of gene network expansion is different and somehow more computationally demanding than performing a pure gene network reconstruction (Marbach et al., 2012). A gene network reconstruction task should be performed genome-wide with a considerable accuracy. In this case, it will be in principle possible to use the same results for deriving the expansion of a given subnetwork. The available reconstruction methods, when applied to genome-wide data, are computationally demanding and, as we will see here, not accurate enough for using the results to perform an expansion task. The gene network expansion task start with a Local Gene Network (LGN) of an organism that is a subset of genes known to be causally connected. We can informally define the gene network expansion as: given a LGN, find other candidate genes that are causally connected with the LGN.
In this paper, we explicitly define the task of finding candidates for gene network expansion and we propose a novel method that we called Network Expansion by Subsetting and Ranking Aggregation (NESRA). NESRA is based on the PC algorithm that we run on our gene@home project, developed on the BOINC platform (Anderson, 2004). We evaluate NESRA on real data of the model plant *Arabidopsis thaliana*.
The paper is organized as follows: in Section II we detailed both the TN-Grid platform and the gene@home project based on volunteer distributed computing and then Section III introduces the main ideas of our approach. Section IV presents the NESRA algorithm, whose evaluation is described in Section V. Finally, Section VI draws some conclusions providing future insights for the gene@home project and the proposed methods.

6.2.2 TN-Grid and the gene@home BOINC project

TN-Grid[22] is a BOINC server installation that has been thought and developed as an *umbrella* project, a service platform to give to local research groups a guided access to the power of the world-wide, volunteer-based, distributed BOINC (Anderson, 2004) computing network. TN-Grid is the result of a joint effort made by two institutions of the Italian National Research Council (CNR), namely the Institute of Materials for Electronics and Magnetism (IMEM) and the Institute of Cognitive Sciences and Technologies (ISTC), both having local branches in Trento, Italy. At the time of writing, TN-Grid is the only public, BOINC-based active project in Italy.

[22] http://gene.disi.unitn.it/test/

The gene@home project is the first one hosted in the TN-Grid framework. It started as a collaboration between the Edmund Mach Foundation (FEM) and the Department of Information Engineering and Computer Science (DISI) of the University of Trento, Italy. The actual development of the gene@home project began as a course project in the academic year 2013/2014 during the Laboratory of Biological Data Mining course at the University of Trento, Italy. Gene@home is a distributed computational biology project based on the computation of the PC algorithm for the Gene Network Expansion task. The final goal of the project is the possibility to automatically perform Gene Network Expansion tasks on demand. After having setup our BOINC server, we coded several scripts to customize it accordingly to the needs of our project. In particular, we designed and developed the work generator using Python. BOINC APIs however, are available only through C++ libraries. For this reason, we implemented two C++ programs that wrap the necessary BOINC functions for the work generator. Our work generator is also responsible for the creation of the workunits that are then distributed to the volunteers. Each workunit is a composition of several PC executions. Because of this, the work generator has to predict the duration of each workunit. The duration of a workunit is not related to the execution time of a single PC run, its input data, or its parameters. So far, we are using a function that we obtained from a regression analysis on several workunits. However, as soon as we change the organism, the LGN, or the input data, we should redo such analysis. To solve this issue we plan to develop a benchmarking system able to evaluate the duration of a workunit, making the estimates of the work generator more precise.

One of the most relevant parts of our implementation is the client application. The client application has been developed to be portable on a number of different architectures (32 and 64 bit) and operating systems (Linux, Windows, and Mac OS). Our client application is a C++ implementation of the *skeleton* function (**Algorithm 1**), functionally equivalent to the one present in the *pcalg* R package (Hauser and Bühlmann, 2012; Kalisch et al., 2012). The choice of implementing the PC algorithm in C++ led to a speed-up of 240 times in the execution, together with a reduced memory consumption of about 10 times, when compared to the original version present in the R package. During the implementation and testing of the initial version of the gene@home project, we had to face several issues, mainly related to the characteristics of our project. One of them, in particular, is the amount of data that needs to be exchanged between the server and the users. We solved this problem with the help of the BOINC core developers that implemented the possibility of compressing the data during the upload and download phases. Subsequently, we optimized our implementation to further reduce the amount of data exchanged. When using a volunteer distributed system, one should be concerned about the validity of the results returned by the volunteers. On the gene@home server, we perform a validation step on the returned workunits, available in all BOINC systems. Because of the nature of our project, we were not able to find a self-validation method to confirm a result of a single workunit. For this reason, we are currently using a double validation method that consists of sending each workunit to two different volunteers. We then required the returned results to be equal bit-wise.

A first step of the processing of the results is implemented in the client application. Just before a workunit finishes, a first aggregation of the results of the workunit is performed. This was also necessary in order to dramatically reduce the size of the output file that the volunteers need to upload in the gene@home server. The results collected with the gene@home project undergo further offline processing developed into a pipeline of Python and R scripts, that complete the analysis of the partial results of each workunit.

In **Table I**, we present some statistical results of the BOINC server collected in 5 different periods of time. It is worth to note the high percentage values of successfully computed workunits, as well as the very low number of workunits that reported an error.

Table I: BOINC statistics of the gene@home project taken on four different days during the year 2014 and one time point in the year 2015 (reference period: the previous seven days). Date is represented as "dd/mm/yy". *Over*: the total number of returned results, *Success*: successfully computed, *Valid*: validated results, *Initial*: pending validation, and *Error*: faulty results. The percentages are relative to the total number of returned results (column *Over*).

Date	Over	Success	Valid	Initial	Error
22/04/14	15543	15392 (99.0%)	15340 (98.7%)	19	85 (0.005%)
20/05/14	69536	68621 (98.7%)	67096 (97.7%)	1450	89 (0.001%)
16/12/14	33232	31798 (96.2%)	29525 (88.8%)	2147	38 (0.001%)
24/12/14	91315	89536 (98.1%)	87584 (95.9%)	1716	61 (0.0007%)
27/03/15	32062	30598 (95.4%)	29525 (92.1%)	865	34 (0.001%)

Algorithm 1: NESRA.

Data: S set of candidate transcripts, S_{LGN} set of LGN transcripts, E expression data
Input: I set of values of number of iterations, D set of values of the subset dimension, A set of values of the significance level α, k maximum length of the lists
Result: ordered list of candidate transcripts

```
L ← ∅; // L set of ordered lists
foreach α ∈ A do
    foreach d ∈ D do
        foreach i ∈ I do
    L ← L ∪ RP(S, S_LGN, E, i, d, α)   // call Algorithm 2
L ← top(L, k)   // cut each list in L to the first k elements
return Ranking_aggregation(L);
```

6.2.3 Gene Network Expansion

Given a set S of gene transcripts whose level of expression has been measured p times in different conditions, such that for each $s_i \in S$ there is a vector $x_i \in \mathbb{R}^p$ of expression levels, and let us assume that there exists a golden truth directed graph $\mathcal{G} = (\mathcal{S}, \mathcal{B})$ with $\mathcal{B} \subset \mathcal{S} \times \mathcal{S}$ that represents the real causal relationships between the gene transcripts, it is possible to define the following tasks.

Task 1, *Gene Network reconstruction*. Given a subset of transcripts $N \subseteq S$, find a (direct) graph $G = (N, B)$ where $B \subset N \times N$ is a relation between the elements of N, and G approximates the subgraph in $\mathcal{G}$ obtained considering just the transcripts in N.

Task 2, *Gene Network expansion*. Given a graph $G = (N, B)$ where $N \subseteq S$and $B \subset N \times N$ is a causal relation between the elements of N, find a graph $G' = (N', B')$ such that N' is a superset of N, B' is a superset of B, and G' approximates the subgraph $\mathcal{G}$ obtained considering just N'.

> *Task 3*, *Discovering candidate genes for Gene Network expansion*. Given a graph $G = (N, B)$ where N is a subset of the transcripts of S and $B \subset N \times N$ is a relation between the elements of N, find a ranked list of elements of $S \setminus N$ such that the elements of the list are connected or very near to the elements of N in $\mathcal{G}$.

In this paper, we will consider Task 3, motivated by the fact that in biological research the work is often guided by prior knowledge about the relevance of some genes. Moreover, a high-quality candidate short list would suffice because the actual validation of the possible interactions requires a complex mix of analytical and wet-lab techniques. It is worth to note that a perfect solution for Task 1 encompassing the whole genome would perfectly solve also Task 2 and Task 3, for all the possible networks. In the same way a perfect solution for Task 2 for a specific network would solve also Task 3. However, the state-of-the-art methods are far from perfect and a good solution for Task 3, in terms of precision of the candidate lists would be useful whenever the whole network and the interactions are still not known, and moreover Task 1 and Task 2 are not solved yet.

6.2.4 NESRA

The general approach used by NESRA is to systematically and iteratively apply subsetting on the whole dataset, in order to compute several ranked lists with varying iterated subsetting parameters. The lists are then aggregated by means of a ranking aggregator. The high-level structure of NESRA is described in **Algorithm 1**. NESRA calls the ranking procedure (RP, **Algorithm 2**) many times with different parameters producing several rankings that are then inputted to the ranking aggregation method for producing a final list.
The ranking procedure has three steps, which respectively create the subsets (Step 1), execute several calls (Step 2) of the skeleton procedure of the PC algorithm (**Algorithm 3**) that processes the expression data of different subsets of the overall transcripts, and finally, compute the transcripts frequency that defines the order of each ranking (Step 3). The ranking procedure takes as parameters the number of iterations i and the dimension of the subset t as well as the significance level α for the PC algorithm. The computational cost of the PC algorithm is exponential in the number of nodes, but it behaves reasonably in the case of sparse networks (Maathuis et al., 2010). Is therefore important to use relatively small values of t. The ranking procedure is partially computed on the BOINC platform with the exception of the frequencies calculation and the rankings aggregation, which are executed outside BOINC.

6.2.4.1 Variable Subsetting

Subsetting is a computational practice that has been used in many domains including recently genomics (Peternelli and Rosa). It consists in the selection from the data available a subset of it, to be processed by the successive steps of the analysis. The idea in itself is not new and it can be found, with different names, in the very core of techniques, such as bootstrapping or subsampling like in bagging (Breiman, 1996) or singling-out features like in random forest (Breiman, 2001) or in feature selection itself. We prefer here to call it subsetting for the sake of clarity because we will specifically focus on variable subsetting, namely different subsets of the variables will be used for gene network reconstruction using the PC algorithm. We avoid to call it subsampling because subsampling does not affect the presence of a variable but select the samples of the variable. On the other hand, we do not call it feature selection because, in this setting the gene is not a feature that describes

something, nor variable selection because we do not select variables in any way that is not purely random.
In NESRA, subsetting is applied to genes that have to be selected for the application of the PC algorithm. The subsetting is iterated and systematic, controlled by two parameters (iterations and tile size) that vary. From the results of these executions we compute a ranked list of genes for each pair of parameters values. Finally, we provide as a final result the aggregation of the ranked lists.

6.2.4.2 Aggregation of ranked lists

The method that we propose as a solution for the problem formulated in Task 3, NESRA, exploits variable subsetting on top of ranking aggregation.
We applied different ranking aggregation methods on the ranked lists. These methods are a simple technique, called the *number of appearances*, and less simple methods, namely Borda Count (Borda, 1781) and MC4 heuristic (Dwork et al., 2001; Lin, 2010). The baseline method that we considered is the *number of appearances* that counts how many rankings a certain gene is present in, i.e. the more a gene is present, the higher its position in the aggregated rank. The Borda Count method consists in constructing a matrix whose elements b_{ij} are for each gene s_i and ranking r_j the rank of the gene s_i in the ranking r_j. After that a statistic for every gene is computed on the rows of the matrix. The two statistical measures that we considered are the mean (BC-mean) and the minimum (BC-min) of the elements. MC4 heuristic is an aggregator based on Markov chains and it consists in computing a transition matrix such that the steady state of the chain assigns a higher probability to the elements with higher rank. MC4 has as parameter the significance level α_{MC4}.

Algorithm 2: NESRA ranking procedure (RP).

Data: S set of candidate transcripts, S_{LGN} set of LGN transcripts, E expression data
Input: i≥l number of iterations, t subset dimension, α significance level
Result: l, ordered list of candidate transcripts

```
N ← |S|;
n ← |S_LGN|;
L ← ∅;
foreach g ∈ S do
    p_g = i;
    f_g = 0;
foreach j, l ≤ j ≤ i do  // Step 1: tiles creation
    S_temp ← S;
    foreach h, l ≤ h ≤ floor(N/t) do
        while |T_h,j| < t do
            random select g ∈ S_temp;
            T_h,j ← T_h,j ∪ {g};
            S_temp ← S_temp \ {g};
    if remainder(N/t) ≠ 0 then
    h ← floor(N/t);
        while S_temp ≠ ∅ do
            random select g ∈ S_temp;
            T_h+1,j ← T_h+1,j ∪ {g};
            S_temp ← S_temp \ {g};
        while |T_h+1,j| < t do
```

```
                random select g ∈ S \ T_{h+1,j};
                T_{h+1,j} ← T_{h+1,j} ∪ {g};
                p_g ← p_g + 1;
foreach j, l ≤ j ≤ i do    // Step 2: PC application
    foreach h, l ≤ h ≤ ceil(N/t) do
        N_{h,j} = PC(T_{h,j}, E, α)    // call Algo 3
foreach g ∈ S do    // Step 3: Transcripts frequency computation
    foreach q ∈ S_LGN do
        foreach j, l ≤ j ≤ i do
            foreach h, l ≤ h ≤ ceil(N/t) do
                if g ∈ AdjN_{h,j}(q) then    // adjacent nodes of q in N_{h,j}
        l ← l ∪ {g};
                    f_g ← f_g + 1;
    f'_g = f_g / (p_g * n)    // Normalized frequency
}
return l ordered w.r.t. f'_g;
```

6.2.4.3 The use of the gene@home project

NESRA exploits the gene@home project for computing the first two steps of the **Algorithm 2**. In details, the tiles creation (Step 1) is implemented in the work generator of the gene@home, while the application of the PC (Step 2) is implemented in the client application, running on the volunteer computers. A first aggregation of the results is then performed on the volunteer's computers, just before the workunit finishes. The complete processing of the results is then performed offline and outside BOINC by means of Python and R scripts.

Algorithm 3: PC Algorithm: skeleton procedure (Kalisch and Buhlmann, 2007).

```
Data: T, Set of transcripts, E expression data
Input: Significance level α
Result: An undirected graph with causal relationship between transcripts Graph
G ← complete undirected graph with nodes in T;
l ← -1;
while l < |G| do
  l ← l + 1;
    foreach ∃ u,v ∈ G s.t. |Adj_G(u)\{v}| ≥ l do    // Adj_G(u) adjacent nodes of u in G
        if v ∈ Adj_G(u) then
            foreach A ⊆ Adj_G(u)\{v} s.t. |A| = l do
                if u,v are conditionally independent given A w.r.t. E with significance level
α then
                    remove edge {u,v} from G;
return G;
```

6.2.5 Evaluation of NESRA on *Arabidopsis thaliana*

In our evaluation of NESRA, we used the Flower Organ Specification Gene Regulatory Network (FOS) of the model plant *Arabidopsis thaliana*. The FOS gene network has been characterized and validated *in vivo* by the use of specific mutants (Espinosa-Soto et al., 2004), and it encompasses 15 genes (AT3G02310.1, AT1G69120, AT5G61850, AT1G30950, AT1G65480, AT5G15800, AT5G-60910, AT5G20240, AT4G36920, AT3G54340, AT2G17950, AT1G24260, AT5G11530, AT4G18960, AT5G03840.1) linked by 54 causal relationships (Sánchez-Corrales et al., 2010). Gene Expression Data for testing

the algorithms were selected from the *A. thaliana* microarray expression data publicly available in the Plex database (Dash et al., 2012). The dataset consists of 393 hybridization experiments of the GeneChip Arabidopsis ATH1 Genome Array that contains 22,810 probe sets.

NESRA was run on the *A. thaliana* data as well as three competitors: PC, PC*, and ARACNE. The quality of the output list of NESRA and of the competitors was assessed by comparison with the available literature. A bibliographic search and classification of the genes provided in output by NESRA and by the competitors led to four classes: *Class 1*: genes reported to be biologically or functionally related to the LGN; *Class 2*: genes not reported to be directly related with the input network, but reported to be related with genes of Class 1; *Class 3*: genes described in literature, but reported not to be related with the input network or with the genes of Class 1; *Class 4*: genes not described in the available literature. A gene falling in Class 1 or Class 2 is considered to be a true positive and a gene in Class 3 or Class 4 a false positive. Precision is defined as the ratio between the number of true positives and the sum of true positives and false positives.

PC, PC*, and ARACNE solve the task of gene network reconstruction. For obtaining list of candidate genes for the expansion we considered all the genes that are connected to FOS genes in the resulting overall network. ARACNE was run with default parameters and the list was ranked according to the p-values that ARACNE itself provides. The PC algorithm was repeated 20 times shuffling the order of the input probe sets, given its dependency on the order. The results of the PC and PC* are reported in **Table II**, note that PC had a mean length of the list of 54.2 and so we took 55 as a cut-off for ARACNE and NESRA for sake of comparison. PC* found 44 genes, and since it is order independent we could not retrieve a result with 55 probes. We reported the result of ARACNE in **Table IV** because we used the p-values to evaluate the list at different cut-off values.

For NESRA we tried five different ranking aggregators: one based on the *number of appearances* in the 55-long lists used as baseline, two based on Borda Counts, BC-mean and BC-min, and then two based on MC4 with two significance level values: $\alpha_{MC4} = 0.05$ and $\alpha_{MC4} = 0.01$.

The sets of parameters I, D, and A (see **Algorithm 1**) used by NESRA for numbers of iterations, subset dimensions, and the significance level are: $I = \{100,250,500,1000,1500,2000\}$, $D = \{50,100,250,500,750,1000,1250,1500,1750,2000\}$, and $A = \{0.05\}$, respectively. An example of the output list of a run of NESRA is shown in **Table III**, where we aggregated 60 different rankings. In order to assess the stability of NESRA we selected 6 combination of parameters, and for each of them we repeated the procedure 30 times. Mean and standard deviations of the results are presented in **Table IV**.

MC4 and BC-mean present in general very good results. BC-min instead, gives more variable outputs, sometimes showing better results ($k = 5$), but in other cases behaving as the baseline method ($k = 20$ or $k = 55$). The results in **Table IV** show also that, regardless of the aggregation method used, NESRA find more correct genes (genes belonging to either Class 1 or Class 2) in the first 20 positions ($k = 5,10,20$) compared to ARACNE. ARACNE instead, finds an appreciable amount of correct genes only when considering a longer list ($k = 55$). We performed a t-test between the results of NESRA and the result of the PC, at $k = 55$. The results of the t-test suggest that NESRA have better performances almost for every aggregator considered.

Table II: *A. thaliana*, FOS network. Lists length and precision of the competitors, PC and PC*. PC values are mean and standard deviation of the 20 runs.

	Lists Length	**Precision**
PC, 20 runs	54.20 ± 1.28	0.39 ± 0.03
PC*	44	0.43

Table III: *A. thaliana*, FOS network. Example of output list of NESRA with ranking aggregation method MC4 with $\alpha_{MC4} = 0.01$ with precision 0.90 with $k = 5$ and 0.80 with $k = 10$.

Rank	**AffyID**	**Gene**	**Annotation**	**Class**
1	259089_at	AT3G04960	similar to unknown protein	Class 1 (Lee et al., 2005)
2	255644_at	AT4G00870	basic helix-loop-helix (bHLH) family protein	Class 2 (Hu et al., 2003)
3	265441_at	AT2G20870	cell wall protein precursor	Class1 (Cai et al., 2007)
4	267528_at	AT2G45650	AGL6 (AGAMOUS LIKE-6)	Class 1 (Yoo et al., 2011)
5.5	245571_at	AT4G14695	unknown protein	Class 4
5.5	249939_at	AT5G22430	similar to unknown protein	Class 1 (Zik and Irish, 2003)
7	245842_at	AT1G58430	RXF26	Class 1 (Shi et al., 2011)
8	248496_at	AT5G50790	ATSWEET10	Class 3 (Chen et al., 2012)
9	264180_at	AT1G02190	CER1 protein	Class 1 (Gómez-Mena et al., 2005)
10	261375_at	AT1G53160	SPL4 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 4)	Class 1 (Lal et al., 2011)

Table IV: *A. thaliana*, FOS network. NESRA precision (mean and standard deviation) on 30 different runs with: values of iterations $I' = \{100,500,2000\}$ and subset dimensions $D' = \{1000,2000\}$.

Aggregation Method	**k=5**	**k=10**	**k=20**	**k=55**
N of appearances	0.54 ± 0.054	0.54 ± 0.054	0.53 ± 0.060	0.42 ± 0.015
BC-mean	0.90 ± 0.098	0.65 ± 0.049	0.63 ± 0.038	0.43 ± 0.016

BC-min	0.86 ± 0.098	0.68 ± 0.038	0.60 ± 0.053	0.43 ± 0.021
MC4 (aMC4 = 0.05)	0.88 ± 0.098	0.65 ± 0.049	0.63 ± 0.038	0.42 ± 0.012
MC4 (aMC4 = 0.01)	0.88 ± 0.098	0.65 ± 0.049	0.63 ± 0.038	0.42 ± 0.012
ARACNE	0.20	0.30	0.35	0.45

6.2.6 Conclusions

We have presented the TN-Grid platform that hosts the gene@home BOINC project. In particular, the gene@home project has been developed with the idea of automatically perform the Gene Network Expansion task. The gene@home project, so far, is running only on the CPUs of the volunteers' computers. As a future improvement of gene@home, we developed and tested a parallel version of PC* for execution on Graphics Processing Units (GPUs). The choice of implementing PC* instead of PC is due to its independence with respect to the order of the input.

We also presented NESRA that is a new method that exploits variable subsetting and ranking aggregation to find candidate genes for the expansion of gene networks. The method relies on the BOINC platform for running the PC algorithm while all the post-processing, ranking and aggregation analyses, are performed offline. The evaluation on the FOS gene network of the model plant *Arabidopsis thaliana* shows good results, and when the results are compared to the biological literature, NESRA outperforms the competitors. In general, NESRA can be used to find candidate variables that are causally connected to other variables and it has proved to work with more than 20,000 variables. We foresee the application of NESRA also in other biological domains.

6.3. NES²RA: Network expansion by stratified variable subsetting and ranking aggregation

The NESRA algorithm is the first one we developed and we implemented on the gene@home BOINC project. Thanks also to the support of the volunteers, we were indeed able to derive other two gene network expansion algorithms, namely NES²RA (introduced in this section) and OneGenE (introduced in **Section 6.4**). The NES²RA algorithm is an improved version of the NESRA algorithm that includes some *a priori* information, specifically, it allows to model the degree of confidence of the user about the presence of each gene in the local gene network to be expanded.

Asnicar F, Masera L, Coller E, Gallo C, Sella N, Tolio T, Morettin P, Erculiani L, Galante F, Semeniuta S, Malacarne G, Engelen K, Argentini A, Cavecchia V, Moser C, and Blanzieri E

NES²RA: Network expansion by stratified variable subsetting and ranking aggregation

The International Journal of High Performance Computing Applications, 32(3), 380-392 (2016)

Abstract - Gene network expansion is a task of the foremost importance in computational biology. Gene network expansion aims at finding new genes to expand a given known gene network. To this end, we developed gene@home, a BOINC-based project that finds candidate genes that expand known local gene networks using NESRA. In this paper, we present NES²RA, a novel approach that extends and improves NESRA by modeling, using a probability vector, the confidence of the presence of the genes belonging to the local gene network. NES²RA adopts intensive variable-subsetting strategies, enabled by the computational power provided by gene@home volunteers. In particular, we use the skeleton procedure of the PC-algorithm to discover candidate causal relationships within each subset of variables. Finally, we use state-of-the-art aggregators to combine the results into a single ranked candidate genes list. The resulting ranking guides the discovery of unknown relations between genes and a priori known local gene networks. Our experimental results show that NES²RA outperforms the PC-algorithm and its order-independent PC-stable version, ARACNE, and our previous approach, NESRA. In this paper we extensively discuss the computational aspects of the NES²RA approach and we also present and validate expansions performed on the model plant *Arabidopsis thaliana* and the model bacteria *Escherichia coli*.

6.4. OneGenE: Regulatory Gene Network Expansion via Distributed Volunteer Computing on BOINC

The OneGenE expansion algorithm leverages the possibility of pre-computing all the single gene expansions of an organism by exploiting the large computational resources available through the gene@home BOINC project. The idea is that all the single gene expansions are dynamically aggregated off-line and at run-time, depending on the set of genes of interest by the user.

Asnicar F*, Masera L*, Pistore D, Valentini S, Cavecchia V, and Blanzieri E (* equal contribution)

OneGenE: Regulatory Gene Network Expansion via Distributed Volunteer Computing on BOINC

Accepted paper at *27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing* (2019)

Abstract - Gene regulatory network expansion is a task of the foremost importance in computational biology that aims at finding new genes to expand a given known gene regulatory network. To this end, we present OneGenE, a novel framework for gene regulatory network expansion that relies on the BOINC platform. OneGenE is an evolution of the NES^{2}RA algorithm, with the aim to overcome its main criticality, i.e. long response time for the final user. To achieve this goal, candidate expansion lists are pre-computed for each gene in the organism and then aggregated at runtime to produce the final expansion list for a given known gene regulatory network. We validated OneGenE on the expression data of *Pseudomonas aeruginosa*, comparing its results with the one obtained by NES^{2}RA and through a biological literature review.

7. Conclusions

Reconstruction of very large-scale and strain-level phylogenies are of the foremost importance for characterizing previously unseen species. However, large-scale phylogeny reconstruction is now based on tens of thousands of genomes and requires to use hundreds of marker genes to retrieve accurate phylogenetic signal. These two dimensionality challenges are dictating the need for new computational phylogenetic approaches able to scale-up in dimensionality while maintaining high reconstruction precision. It is indeed more and more clear that microbial diversity has to be analyzed at the level of single strains, and it is thus necessary to maintain within-species diversity resolution when elucidating potential biological patterns in a phylogenetic analysis.

To overcome these challenges and empower metagenomics with the methodologies developed for single isolate sequencing, in this thesis I presented and applied a novel framework for accurate large-scale phylogenetic analysis, able to deal with strain-level and tree-of-life size phylogenies, and with the aim of characterizing novel microbial genomes reconstructed form metagenomes.

In the first part of this thesis, I focused on methodological advances to enable deeper phylogenetic analyses:

1. GraPhlAn, presented in **Chapter 2**, is a tool for high-quality visualization of both hierarchical and phylogenetic trees that can guide explorative analysis through visualization thanks to the ability to display several relevant quantitative non-phylogenetic data;
2. PhyloPhlAn 2, presented in **Chapter 3**, is the automatic, customizable, and flexible pipeline for accurate large-scale phylogeny reconstruction that can be fundamental for the characterization of previously unseen genomes from metagenomic assembly;
3. The approach for strain-tracking across microbiomes, presented in **Chapter 4**, is a methodological contribution on how phylogenetics on metagenomic data can be used to perform tasks of great biological relevance, such as the detection of microbial members vertically transmitted from mothers to their infants.

In the second part of the thesis (**Chapter 5**), I presented several works where the tools from **Chapter 2**, **3**, and **4** were applied and played an important role in the analysis. The panel of works ranges from the large-scale study of vertically transmitted novel species reconstructed from metagenomes (**Section 5.1**, (Ferretti et al., 2018)) to building the largest reference microbial phylogeny with the main aim to improve phylogeny-aware tools (**Section 5.2**). The phylogenetic analysis of microbiomes of colorectal cancer patients allowed the identification of four variants of the *cutC* gene, with some of its variant significantly associated with carcinoma samples (**Section 5.3**). A similar phylogenetic analysis allowed to characterize unknown metagenomically reconstructed skin-associated organisms in both unaffected and psoriatic lesions (**Section 5.4**, (Tett et al., 2017)). The reconstruction of the Neisseriaceae family phylogeny allowed to study the correlation of the Neisseriaceae species with short and highly-repeated DNA sequences (**Section 5.5**, (Donati et al., 2016)). We moreover uncovered the within-species diversity of *Eubacterium rectale* by phylogenetically characterizing over 1,300 newly metagenomically reconstructed genomes, detecting four subspecies strongly associated with geography (**Section 5.6**). Finally, with the phylogenetic characterization of the largest metagenomic assembly and binning effort to date (**Section**

5.7, (Pasolli et al., 2019)) we unraveled extensive unexplored human microbiome diversity and characterized known and unknown species.

In **Chapter 6** I introduced and presented a second line of research I started during my M.Sc. degree and continued during my doctoral studies, whose aim is the application of computational and statistical approaches for the discovery of novel gene interactions within known gene networks (Asnicar et al., 2015b, 2015c, 2016, 2019), based on the distributed and volunteer-based computational framework BOINC (Anderson, 2004).

The work presented in this thesis represents a contribution to the computational metagenomics and computational phylogenetics fields, by proposing a novel phylogenetic framework that allows both the high-quality visualization and the reconstruction of large-scale phylogenies. The framework has been extensively used in different contexts to fill the gap of the application of phylogenetic analysis to metagenomics.

To increase the accuracy in building very large phylogenies, novel sets of phylogenetic marker genes able to capture the diversity at different taxonomic levels are needed. The continuously increasing sequencing of novel genomes and the reconstruction of genomes from metagenomes will allow the discovery of new and robust phylogenetic markers. This, in turn, will allow us to characterize an enormous previously unexplored microbial diversity and expand the catalog of known organisms.

8. Appendix

8.1 Other Works

In this chapter, I report the other research works I was involved in during my doctoral studies and that I did not include in the main discussion of the thesis. For these works, I report the citation of the article and its abstract.

Nigro E*, Mazzoni C*, Alvari G, Baldi G, Calia G, Cantore T, Ciciani M, Dalfovo D, Fabbri L, Flor S, Golzato D, Lattanzi C, Marangoni S, Marianini G, Minardi G, Piccinno R, Pirrotta S, Tebaldi M, Tonazzolli A, Vannuccini F, Manara S, Zolfo M, Karcher N, Asnicar F, Tett A^, Edoardo Pasolli E^, Segata N^ (* equal contribution, ^ co-senior authors)

Draft genome sequence of the new species "*Candidatus* Cibiobacter qucibialis" metagenomically assembled from the human gut microbiome

Currently in revision at *Microbiology Resource Announcements*

Abstract - We report the metagenomically-assembled draft genome of a human intestinal strain representing a previously unknown species we name "*Candidatus* Cibiobacter qucibialis". The new species is phylogenetically placed between *Ruminococcus* and *Faecalibacterium* and it should be considered a relevant member of the human gut microbiome.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Titus Brown C, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG

QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science

Currently in revision at *Nature Biotechnology*, *PeerJ Preprint* available

Abstract - We present QIIME 2, an open-source microbiome data science platform accessible to users spanning the microbiome research ecosystem, from scientists and engineers to clinicians and policymakers. QIIME 2 provides new features that will drive the next generation of microbiome research. These include interactive spatial and temporal analysis and visualization tools, support for metabolomics and shotgun metagenomics analysis, and automated data provenance tracking to ensure reproducible, transparent microbiome data science.

Pedron R*, Esposito A*, Bianconi I, Pasolli E, Tett A, Asnicar F, Cristofolini M, Segata N, and Jousson O (* equal contribution)

Genomic and metagenomic insights into the microbial community of a thermal spring

Microbiome (2019)

Abstract - **Background.** Water springs provide important ecosystem services including drinking water supply, recreation, and balneotherapy, but their microbial communities remain largely unknown. In this study, we characterized the spring water microbiome of Comano Terme (Italy) at four sampling points of the thermal spa, including natural (spring and well) and human-built (storage tank, bathtubs) environments. We integrated large-scale culturing and metagenomic approaches, with the aim of comprehensively determining the spring water taxonomic composition and functional potential. **Results.** The groundwater feeding the spring hosted the most atypical microbiome, including many taxa known to be recalcitrant to cultivation. The core microbiome included the orders Sphingomonadales, Rhizobiales, and Caulobacterales, and the families Bradyrhizobiaceae and Moraxellaceae. A comparative genomic analysis of 72 isolates and 30 metagenome-assembled genomes (MAGs) revealed that most isolates and MAGs belonged to new species or higher taxonomic ranks widely distributed in the microbial tree of life. Average nucleotide identity (ANI) values calculated for each isolated or assembled genome showed that 10 genomes belonged to known bacterial species (> 95% ANI), 36 genomes (including 1 MAG) had ANI values ranging 85–92.5% and could be assigned as undescribed species belonging to known genera, while the remaining 55 genomes had lower ANI values (< 85%). A number of functional features were significantly over- or underrepresented in genomes derived from the four sampling sites. Functional specialization was found between sites, with for example methanogenesis being unique to groundwater whereas methanotrophy was found in all samples. **Conclusions.** Current knowledge of aquatic microbiomes is essentially based on surface or human-associated environments. We started uncovering the spring water microbiome, highlighting an unexpected diversity that should be further investigated. This study confirms that groundwater environments host highly adapted, stable microbial communities composed of many unknown taxa, even among the culturable fraction.

Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, Selvenius J, Oikarinen S, Hyöty H, Virtanen SM, Ilonen J, Ferretti P, Pasolli E, Tett A, Asnicar F, Segata N, Vlamakis H, Lander ES, Huttenhower C, Knip M, and Xavier RJ

Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life

Cell Host & Microbe (2018)

Abstract - Bacterial community acquisition in the infant gut impacts immune education and disease susceptibility. We compared bacterial strains across and within families in a prospective birth cohort of 44 infants and their mothers, sampled longitudinally in the first months of each child's life. We identified mother-to-child bacterial transmission events and describe the incidence of family-specific antibiotic resistance genes. We observed two inheritance patterns across multiple species, where often the mother's dominant strain is transmitted to the child, but occasionally her secondary strains colonize the infant gut. In families where the secondary strain of *B. uniformis* was inherited, a starch utilization gene cluster that was absent in the mother's dominant strain was identified in the child, suggesting

the selective advantage of a mother's secondary strain in the infant gut. Our findings reveal mother-to-child bacterial transmission events at high resolution and give insights into early colonization of the infant gut.

Pinto F, Tett A, Armanini F, Asnicar F, Boscaini A, Pasolli E, Zolfo M, Donati C, Salmaso N, and Segata N

Draft Genome Sequences of Novel *Pseudomonas*, *Flavobacterium*, and *Sediminibacterium* Species Strains from a Freshwater Ecosystem

Genome Announcements (2018)

Abstract - Freshwater ecosystems represent 0.01% of the water on Earth, but they support 6% of global biodiversity that is still mostly uncharacterized. Here, we describe the genome sequences of three strains belonging to novel species in the *Pseudomonas*, *Flavobacterium*, and *Sediminibacterium* genera recovered from a water sample of Lake Garda, Italy.

Manara S*, Pasolli E*, Dolce D*, Ravenni N, Campana S, Armanini F, Asnicar F, Mengoni A, Galli L, Montagnani C, Venturini E, Rota-Stabelli O, Grandi G, Taccetti G, Segata N (* equal contribution)

Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of *Staphylococcus aureus* strains in a paediatric hospital

Genome medicine (2018)

Abstract - **Background.** *Staphylococcus aureus* is an opportunistic pathogen and a leading cause of nosocomial infections. It can acquire resistance to all the antibiotics that entered the clinics to date, and the World Health Organization defined it as a high-priority pathogen for research and development of new antibiotics. A deeper understanding of the genetic variability of *S. aureus* in clinical settings would lead to a better comprehension of its pathogenic potential and improved strategies to contrast its virulence and resistance. However, the number of comprehensive studies addressing clinical cohorts of *S. aureus* infections by simultaneously looking at the epidemiology, phylogenetic reconstruction, genomic characterisation, and transmission pathways of infective clones is currently low, thus limiting global surveillance and epidemiological monitoring. **Methods.** We applied whole-genome shotgun sequencing (WGS) to 184 *S. aureus* isolates from 135 patients treated in different operative units of an Italian paediatric hospital over a timespan of 3 years, including both methicillin-resistant *S. aureus* (MRSA) and methicillin-sensitive *S. aureus* (MSSA) from different infection types. We typed known and unknown clones from their genomes by multilocus sequence typing (MLST), Staphylococcal Cassette Chromosome *mec* (*SCCmec*), Staphylococcal protein A gene (*spa*), and Panton-Valentine Leukocidin (PVL), and we inferred their whole-genome phylogeny. We explored the prevalence of virulence and antibiotic resistance genes in our cohort, and the conservation of genes encoding vaccine candidates. We also performed a timed phylogenetic investigation for a potential outbreak of a newly emerging nosocomial clone. **Results.** The phylogeny of the 135 single-patient *S. aureus* isolates showed a high level of diversity, including 80 different lineages, and co-presence of local, global, livestock-associated, and hypervirulent clones. Five of these clones do not have representative genomes in public databases. Variability in the epidemiology is mirrored by variability in the SCCmec cassettes, with some novel variants of the type IV cassette carrying extra antibiotic resistances. Virulence and

resistance genes were unevenly distributed across different clones and infection types, with highly resistant and lowly virulent clones showing strong association with chronic diseases, and highly virulent strains only reported in acute infections. Antigens included in vaccine formulations undergoing clinical trials were conserved at different levels in our cohort, with only a few highly prevalent genes fully conserved, potentially explaining the difficulty of developing a vaccine against *S. aureus*. We also found a recently diverged ST1-*SCCmec*IV-*t127* PVL− clone suspected to be hospital-specific, but time-resolved integrative phylogenetic analysis refuted this hypothesis and suggested that this quickly emerging lineage was acquired independently by patients. **Conclusions.** Whole genome sequencing allowed us to study the epidemiology and genomic repertoire of *S. aureus* in a clinical setting and provided evidence of its often underestimated complexity. Some virulence factors and clones are specific of disease types, but the variability and dispensability of many antigens considered for vaccine development together with the quickly changing epidemiology of *S. aureus* makes it very challenging to develop full-coverage therapies and vaccines. Expanding WGS-based surveillance of *S. aureus* to many more hospitals would allow the identification of specific strains representing the main burden of infection and therefore reassessing the efforts for the discovery of new treatments and clinical practices.

Zolfo M, Asnicar F, Manghi P, Pasolli E, Tett A, Segata N

Profiling microbial strains in urban environments using metagenomic sequencing data

Biology direct (2018)

Abstract - **Background.** The microbial communities populating human and natural environments have been extensively characterized with shotgun metagenomics, which provides an in-depth representation of the microbial diversity within a sample. Microbes thriving in urban environments may be crucially important for human health, but have received less attention than those of other environments. Ongoing efforts started to target urban microbiomes at a large scale, but the most recent computational methods to profile these metagenomes have never been applied in this context. It is thus currently unclear whether such methods, that have proven successful at distinguishing even closely related strains in human microbiomes, are also effective in urban settings for tasks such as cultivation-free pathogen detection and microbial surveillance. Here, we aimed at a) testing the currently available metagenomic profiling tools on urban metagenomics; b) characterizing the organisms in urban environment at the resolution of single strain and c) discussing the biological insights that can be inferred from such methods. **Results.** We applied three complementary methods on the 1614 metagenomes of the CAMDA 2017 challenge. With MetaMLST we identified 121 known sequence-types from 15 species of clinical relevance. For instance, we identified several *Acinetobacter* strains that were close to the nosocomial opportunistic pathogen *A. nosocomialis*. With StrainPhlAn, a generalized version of the MetaMLST approach, we inferred the phylogenetic structure of *Pseudomonas stutzeri* strains and suggested that the strain-level heterogeneity in environmental samples is higher than in the human microbiome. Finally, we also probed the functional potential of the different strains with PanPhlAn. We further showed that SNV-based and pangenome-based profiling provide complementary information that can be combined to investigate the evolutionary trajectories of microbes and to identify specific genetic determinants of virulence and antibiotic resistances within closely related strains. **Conclusion.** We show that

strain-level methods developed primarily for the analysis of human microbiomes can be effective for city-associated microbiomes. In fact, (opportunistic) pathogens can be tracked and monitored across many hundreds of urban metagenomes. However, while more effort is needed to profile strains of currently uncharacterized species, this work poses the basis for high-resolution analyses of microbiomes sampled in city and mass transportation environments.

Malacarne G*, Pilati S*, Valentini S*, Asnicar F, Moretto M, Sonego P, Masera L, Cavecchia V, Blanzieri E, and Moser C (* equal contribution)

Discovering causal relationships in grapevine expression data to expand gene networks. A case study: four networks related to climate change

Frontiers in Plant Science (2018)

Abstract - In recent years the scientific community has been heavily engaged in studying the grapevine response to climate change. Final goal is the identification of key genetic traits to be used in grapevine breeding and the setting of agronomic practices to improve climatic resilience. The increasing availability of transcriptomic studies, describing gene expression in many tissues and developmental, or treatment conditions, have allowed the implementation of gene expression compendia, which enclose a huge amount of information. The mining of transcriptomic data represents an effective approach to expand a known local gene network (LGN) by finding new related genes. We recently published a pipeline based on the iterative application of the PC-algorithm, named NES^{2}RA, to expand gene networks in Escherichia coli and Arabidopsis thaliana. Here, we propose the application of this method to the grapevine transcriptomic compendium Vespucci, in order to expand four LGNs related to the grapevine response to climate change. Two networks are related to the secondary metabolic pathways for anthocyanin and stilbenoid synthesis, involved in the response to solar radiation, whereas the other two are signaling networks, related to the hormones abscisic acid and ethylene, possibly involved in the regulation of cell water balance and cuticle transpiration. The expansion networks produced by NES^{2}RA algorithm have been evaluated by comparison with experimental data and biological knowledge on the identified genes showing fairly good consistency of the results. In addition, the algorithm was effective in retaining only the most significant interactions among the genes providing a useful framework for experimental validation. The application of the NES^{2}RA to *Vitis vinifera* expression data by means of the BOINC-based implementation is available upon request (valter.cavecchia@cnr.it).

Pinto F, Tett A, Armanini F, Asnicar F, Boscaini A, Pasolli E, Zolfo M, Donati C, Salmaso N, and Segata N

Draft Genome Sequence of the Planktic Cyanobacterium *Tychonema bourrellyi*, Isolated from Alpine Lentic Freshwater

Genome Announcements (2017)

Abstract - We describe here the draft genome sequence of the cyanobacterium *Tychonema bourrellyi*, assembled from a metagenome of a nonaxenic culture. The strain (FEM_GT703) was isolated from a freshwater sample taken from Lake Garda, Italy. The draft genome sequence represents the first assembled *T. bourrellyi* strain.

Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Ardythe L Morrow, and Segata N

Strain-level microbial epidemiology and population genomics from shotgun metagenomics

Nature Methods (2016)

Abstract - Identifying microbial strains and characterizing their functional potential is essential for pathogen discovery, epidemiology and population genomics. We present pangenome-based phylogenomic analysis (PanPhlAn; http://segatalab.cibio.unitn.it/tools/panphlan), a tool that uses metagenomic data to achieve strain-level microbial profiling resolution. PanPhlAn recognized outbreak strains, produced the largest strain-level population genomic study of human-associated bacteria and, in combination with metatranscriptomics, profiled the transcriptional activity of strains in complex communities.

9. 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.

Aagaard, K., Ma, J., Antony, K.M., Ganu, R., Petrosino, J., and Versalovic, J. (2014). The placenta harbors a unique microbiome. Sci. Transl. Med. *6*, 237ra65.

Aanensen, D.M., Feil, E.J., Holden, M.T.G., Dordel, J., Yeats, C.A., Fedosejev, A., Goater, R., Castillo-Ramírez, S., Corander, J., Colijn, C., et al. (2016). Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive Staphylococcus aureus in Europe. MBio *7*.

Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., 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.

Albert, R. (2005). Scale-free networks in cell biology. J. Cell Sci. *118*, 4947–4957.

Allen, J.D., Xie, Y., Chen, M., Girard, L., and Xiao, G. (2012). Comparing statistical methods for constructing large scale gene networks. PLoS One *7*, e29348.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Anderson, D.P. (2004). BOINC: a system for public-resource computing and storage. In Fifth IEEE/ACM International Workshop on Grid Computing, pp. 4–10.

Anzai, Y., Kudo, Y., and Oyaizu, H. (1997). The phylogeny of the genera Chryseomonas, Flavimonas, and Pseudomonas supports synonymy of these three genera. Int. J. Syst. Bacteriol. *47*, 249–251.

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

Arroyo, R., Suñé, G., Zanzoni, A., Duran-Frigola, M., Alcalde, V., Stracker, T.H., Soler-López, M., and Aloy, P. (2015). Systematic identification of molecular links between core and candidate genes in breast cancer. J. Mol. Biol. *427*, 1436–1450.

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

Asnicar, F., Sella, N., Masera, L., Morettin, P., Tolio, T., Semeniuta, S., Moser, C., Blanzieri, E., and Cavecchia, V. (2015b). TN-Grid and gene@home project: Volunteer Computing for Bioinformatics.

Asnicar, F., Erculiani, L., Galante, F., Gallo, C., Masera, L., Morettin, P., Sella, N., Semeniuta, S., Tolio, T., Malacarne, G., et al. (2015c). Discovering Candidates for Gene Network Expansion by Distributed Volunteer Computing. In 2015 IEEE Trustcom/BigDataSE/ISPA, pp. 248–253.

Asnicar, F., Masera, L., Coller, E., Gallo, C., Sella, N., Tolio, T., Morettin, P., Erculiani, L., Galante, F., Semeniuta, S., et al. (2016). NES2RA: Network expansion by stratified variable subsetting and ranking aggregation. Int. J. High Perform. Comput. Appl.

Asnicar, F., Manara, S., Zolfo, M., Truong, D.T., 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., Masera, L., Pistore, D., Valtentini, S., Cavecchia, V., and Blanzieri, E. (2019). OneGene: Regulatory Gene Network Expansion via Distributed Volunteer Computing on BOINC. In 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing,.

Azad, M.B., Konya, T., Maughan, H., Guttman, D.S., Field, C.J., Chari, R.S., Sears, M.R., Becker, A.B., Scott, J.A., Kozyrskyj, A.L., et al. (2013). Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. CMAJ *185*, 385–394.

Bäckhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., and Gordon, J.I. (2005). Host-bacterial mutualism in the human intestine. Science *307*, 1915–1920.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 690–703.

Baldini, F., Segata, N., Pompon, J., Marcenac, P., Shaw, W.R., Dabiré, R.K., Diabaté, A., Levashina, E.A., and Catteruccia, F. (2014). Evidence of natural Wolbachia infections in field populations of Anopheles gambiae. Nat. Commun. *5*, 3985.

Balique, F., Colson, P., Barry, A.O., Nappez, C., Ferretti, A., Moussawi, K.A., Ngounga, T., Lepidi, H., Ghigo, E., Mege, J.-L., et al. (2013). Tobacco mosaic virus in the lungs of mice following intra-tracheal inoculation. PLoS One *8*, e54993.

Bao, G., Wang, M., Doak, T.G., and Ye, Y. (2015). Strand-specific community RNA-seq reveals prevalent and dynamic antisense transcription in human gut microbiota. Front. Microbiol. *6*, 896.

Bennett, J.S., Jolley, K.A., Earle, S.G., Corton, C., Bentley, S.D., Parkhill, J., and Maiden, M.C.J. (2012). A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria. Microbiology *158*, 1570–1580.

Bennett, J.S., Watkins, E.R., Jolley, K.A., Harrison, O.B., and Maiden, M.C.J. (2014). Identifying Neisseria species by use of the 50S ribosomal protein L6 (rplF) gene. J. Clin. Microbiol. *52*, 1375–1381.

Biasucci, G., Rubini, M., Riboni, S., Morelli, L., Bessi, E., and Retetangos, C. (2010). Mode of delivery affects the bacterial community in the newborn gut. Early Hum. Dev. *86 Suppl 1*, 13–15.

Bickley, J., Short, J.K., McDowell, D.G., and Parkes, H.C. (1996). Polymerase chain reaction (PCR) detection of Listeria monocytogenes in diluted milk and reversal of PCR inhibition caused by calcium ions. Lett. Appl. Microbiol. *22*, 153–158.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Curr. Protoc. Mol. Biol. *Chapter 19*, Unit 19.10.1–21.

Borda, J.-C. de (1781). Mémoire sur les élections au scrutin, Histoire de l'Académie Royale des Sciences, Paris. Cook WD (2006) Distance-Based and Ad Hoc Consensus Models in Ordinal Preference Ranking. Eur. J. Oper. Res *172*, 369–385.

Bose, T., Haque, M.M., Reddy, C., and Mande, S.S. (2015). COGNIZER: A Framework for Functional Annotation of Metagenomic Datasets. PLoS One *10*, e0142102.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., 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.

Breiman, L. (1996). Bagging predictors. Mach. Learn. *24*, 123–140.

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

Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. Nature *535*, 435–439.

Britton, R.A., and Young, V.B. (2014). Role of the intestinal microbiota in resistance to colonization by Clostridium difficile. Gastroenterology *146*, 1547–1553.

Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C., and Franklin, R.B. (2017). MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience *6*, 1–10.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature *523*, 208–211.

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

Byrd, A.L., Belkaid, Y., and Segre, J.A. (2018). The human skin microbiome. Nat. Rev. Microbiol. *16*, 143–155.

Cabral, D.J., Wurster, J.I., Flokas, M.E., Alevizakos, M., Zabat, M., Korry, B.J., Rowan, A.D., Sano, W.H., Andreatos, N., Ducharme, R.B., et al. (2017). The salivary microbiome is consistent between subjects and resistant to impacts of short-term hospitalization. Sci. Rep. *7*, 11040.

Cabrera-Rubio, R., Collado, M.C., Laitinen, K., Salminen, S., Isolauri, E., and Mira, A. (2012). The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. Am. J. Clin. Nutr. *96*, 544–551.

Cai, X., Ballif, J., Endo, S., Davis, E., Liang, M., Chen, D., DeWald, D., Kreps, J., Zhu, T., and Wu, Y. (2007). A putative CCAAT-binding transcription factor is a regulator of flowering timing in Arabidopsis. Plant Physiol. *145*, 98–105.

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

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Carmody, R.N., Gerber, G.K., Luevano, J.M., Jr, Gatti, D.M., Somes, L., Svenson, K.L., and Turnbaugh, P.J. (2015). Diet dominates host genotype in shaping the murine gut microbiota.

Cell Host Microbe *17*, 72–84.

Castelle, C.J., and Banfield, J.F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell *172*, 1181–1197.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552.

Chai, J., Kora, G., Ahn, T.-H., Hyatt, D., and Pan, C. (2014). Functional phylogenomics analysis of bacteria and archaea using consistent genome annotation with UniFam. BMC Evol. Biol. *14*, 207.

Chang, J.-M., Di Tommaso, P., and Notredame, C. (2014). TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol. Biol. Evol. *31*, 1625–1637.

Chen, L.-Q., Qu, X.-Q., Hou, B.-H., Sosso, D., Osorio, S., Fernie, A.R., and Frommer, W.B. (2012). Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. Science *335*, 207–211.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283–1287.

Claud, E.C., Keegan, K.P., Brulc, J.M., Lu, L., Bartels, D., Glass, E., Chang, E.B., Meyer, F., and Antonopoulos, D.A. (2013). Bacterial community structure and functional contributions to emergence of health or necrotizing enterocolitis in preterm infants. Microbiome *1*, 20.

Clemente, J.C., Ursell, L.K., Parfrey, L.W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. Cell *148*, 1258–1270.

Coller, E. (2013). Analysis of the PC algorithm as a tool for the inference of gene regulatory networks: evaluation of the performance, modification and application to selected case studies. phd. University of Trento.

Colson, P., Richet, H., Desnues, C., Balique, F., Moal, V., Grob, J.-J., Berbis, P., Lecoq, H., Harlé, J.-R., Berland, Y., et al. (2010). Pepper mild mottle virus, a plant virus associated with specific immune responses, Fever, abdominal pains, and pruritus in humans. PLoS One *5*, e10041.

Costea, P.I., Coelho, L.P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017). Subspecies in the global human gut microbiome. Mol. Syst. Biol. *13*, 960.

Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J.M., and Relman, D.A. (2012). The application of ecological theory toward an understanding of the human microbiome. Science *336*, 1255–1262.

Cremonesi, P., Castiglioni, B., Malferrari, G., Biunno, I., Vimercati, C., Moroni, P., Morandi, S., and Luzzana, M. (2006). Technical note: Improved method for rapid DNA extraction of mastitis pathogens directly from milk. J. Dairy Sci. *89*, 163–169.

Cross, R.L., and Müller, V. (2004). The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio. FEBS Lett. *576*, 1–4.

Dadi, T.H., Renard, B.Y., Wieler, L.H., Semmler, T., and Reinert, K. (2017). SLIMM: species

level identification of microorganisms from metagenomes. PeerJ *5*, e3138.

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., 4th, Bik, H.M., and Eisen, J.A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ *2*, e243.

Dash, S., Van Hemert, J., Hong, L., Wise, R.P., and Dickerson, J.A. (2012). PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Res. *40*, D1194–D1201.

Dassi, E., Ballarini, A., Covello, G., HTM-CMB2013, Quattrone, A., Jousson, O., De Sanctis, V., Bertorelli, R., Denti, M.A., and Segata, N. (2014). Enhanced microbial diversity in the saliva microbiome induced by short-term probiotic intake revealed by 16S rRNA sequencing on the IonTorrent PGM platform. J. Biotechnol. *190*, 30–39.

Davenport, E.R., Sanders, J.G., Song, S.J., Amato, K.R., Clark, A.G., and Knight, R. (2017). The human microbiome in evolution. BMC Biol. *15*, 127.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559–563.

De Filippis, F., Parente, E., and Ercolini, D. (2018). Recent Past, Present, and Future of the Food Microbiome. Annu. Rev. Food Sci. Technol. *9*, 589–608.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. U. S. A. *107*, 14691–14696.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

Di Martino, M.L., Campilongo, R., Casalino, M., Micheli, G., Colonna, B., and Prosseda, G. (2013). Polyamines: emerging players in bacteria-host interactions. Int. J. Med. Microbiol. *303*, 484–491.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc. Natl. Acad. Sci. U. S. A. *107*, 11971–11975.

Dominguez-Bello, M.G., De Jesus-Laboy, K.M., Shen, N., Cox, L.M., Amir, A., Gonzalez, A., Bokulich, N.A., Song, S.J., Hoashi, M., Rivera-Vinas, J.I., et al. (2016). Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nat. Med. *22*, 250–253.

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.

Dress, A.W.M., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S.J., and Stadler, P.F. (2008). Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol. Biol. *3*, 7.

Duranti, S., Lugli, G.A., Mancabelli, L., Armanini, F., Turroni, F., James, K., Ferretti, P.,

Gorfer, V., Ferrario, C., Milani, C., et al. (2017). Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. Microbiome *5*, 66.

Durbán, A., Abellán, J.J., Jiménez-Hernández, N., Artacho, A., Garrigues, V., Ortiz, V., Ponce, J., Latorre, A., and Moya, A. (2013). Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome. FEMS Microbiol. Ecol. *86*, 581–589.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. *5*, 4498.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the Web. In Proceedings of the 10th International Conference on World Wide Web, (ACM), pp. 613–622.

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

Edgar, R.C. (2009). Optimizing substitution matrix choice and gap parameters for sequence alignment. BMC Bioinformatics *10*, 396.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Erickson, A.R., Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One *7*, e49138.

Espinosa-Soto, C., Padilla-Longoria, P., and Alvarez-Buylla, E.R. (2004). A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell *16*, 2923–2939.

Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L., et al. (2013). The long-term stability of the human gut microbiota. Science *341*, 1237439.

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, D.T., 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.

Flint, H.J., Scott, K.P., Duncan, S.H., Louis, P., and Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. Gut Microbes *3*, 289–306.

Flores, G.E., Caporaso, J.G., Henley, J.B., Rideout, J.R., Domogala, D., Chase, J., Leff, J.W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., et al. (2014). Temporal variability is a personalized feature of the human microbiome. Genome Biol. *15*, 531.

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, S., Gudmundsdottir, V., Pedersen, H.K., et al. (2015). Disentangling type 2

diabetes and metformin treatment signatures in the human gut microbiota. Nature *528*, 262–266.

Franzosa, E.A., Morgan, X.C., Segata, N., Waldron, L., Reyes, J., Earl, A.M., Giannoukos, G., Boylan, M.R., Ciulla, D., Gevers, D., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. Proc. Natl. Acad. Sci. U. S. A. *111*, E2329–E2338.

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods *15*, 962–968.

Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol *4*, 293–305.

Freitas, T.A.K., Li, P.-E., Scholz, M.B., and Chain, P.S.G. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. *43*, e69.

Frye, S.A., Nilsen, M., Tønjum, T., and Ambur, O.H. (2013). Dialects of the DNA uptake sequence in Neisseriaceae. PLoS Genet. *9*, e1003458.

Fuentes, S., van Nood, E., Tims, S., Heikamp-de Jong, I., ter Braak, C.J.F., Keller, J.J., Zoetendal, E.G., and de Vos, W.M. (2014). Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection. ISME J. *8*, 1621–1633.

Gaitanis, G., Magiatis, P., Hantschke, M., Bassukas, I.D., and Velegraki, A. (2012). The Malassezia genus in skin and systemic diseases. Clin. Microbiol. Rev. *25*, 106–141.

Gajer, P., Brotman, R.M., Bai, G., Sakamoto, J., Schütte, U.M.E., Zhong, X., Koenig, S.S.K., Fu, L., Ma, Z.S., Zhou, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. Sci. Transl. Med. *4*, 132ra52.

Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. Science *301*, 102–105.

Gardy, J.L., Johnston, J.C., Ho Sui, S.J., Cook, V.J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., et al. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N. Engl. J. Med. *364*, 730–739.

Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe *15*, 382–392.

Giannoukos, G., Ciulla, D.M., Huang, K., Haas, B.J., Izard, J., Levin, J.Z., Livny, J., Earl, A.M., Gevers, D., Ward, D.V., et al. (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. *13*, R23.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451–1455.

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in

the life sciences. Genome Biol. *11*, R86.

Gómez-Mena, C., de Folter, S., Costa, M.M.R., Angenent, G.C., and Sablowski, R. (2005). Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. Development *132*, 429–438.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. *57*, 81–91.

Gosalbes, M.J., Abellan, J.J., Durbán, A., Pérez-Cobas, A.E., Latorre, A., and Moya, A. (2012). Metagenomics of human microbiome: beyond 16s rDNA. Clin. Microbiol. Infect. *18 Suppl 4*, 47–49.

Gossling, J., and Moore, W.E.C. (1975). Gemmiger formicilis, n.gen., n.sp., an Anaerobic Budding Bacterium from Intestines. Int. J. Syst. Evol. Microbiol. *25*, 202–207.

Greenblum, S., Turnbaugh, P.J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc. Natl. Acad. Sci. U. S. A. *109*, 594–599.

Greenwood, C., Morrow, A.L., Lagomarcino, A.J., Altaye, M., Taft, D.H., Yu, Z., Newburg, D.S., Ward, D.V., and Schibler, K.R. (2014). Early empiric antibiotic use in preterm infants is associated with lower bacterial diversity and higher relative abundance of Enterobacter. J. Pediatr. *165*, 23–29.

Grice, E.A., and Segre, J.A. (2011). The skin microbiome. Nat. Rev. Microbiol. *9*, 244–253.

Han, M.V., and Zmasek, C.M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics *10*, 356.

Hansen, L.B.S., Roager, H.M., Søndertoft, N.B., Gøbel, R.J., Kristensen, M., Vallès-Colomer, M., Vieira-Silva, S., Ibrügger, S., Lind, M.V., Mærkedahl, R.B., et al. (2018). A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. Nat. Commun. *9*, 4630.

Hartemink, A.J. (2005). Reverse engineering gene regulatory networks. Nat. Biotechnol. *23*, 554–555.

Hasty, J., McMillen, D., Isaacs, F., and Collins, J.J. (2001). Computational studies of gene regulatory networks: in numero molecular biology. Nat. Rev. Genet. *2*, 268–279.

Hauser, A., and Bühlmann, P. (2012). Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. J. Mach. Learn. Res. *13*, 2409–2464.

Heintz-Buschart, A., May, P., Laczny, C.C., Lebrun, L.A., Bellora, C., Krishna, A., Wampach, L., Schneider, J.G., 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.

Hildebrand, F., Moitinho-Silva, L., Blasche, S., Jahn, M.T., Gossmann, T.I., Heuerta-Cepas, J., Hercog, R., Luetge, M., Bahram, M., Pryszlak, A., et al. (2019). Antibiotics-induced monodominance of a novel gut bacterial order. Gut.

HMP, Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., et al. (2012). Structure, function and diversity of

the healthy human microbiome. Nature *486*, 207.

Holt, K.E., Wertheim, H., Zadoks, R.N., Baker, S., Whitehouse, C.A., Dance, D., Jenney, A., Connor, T.R., Hsu, L.Y., Severin, J., et al. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc. Natl. Acad. Sci. U. S. A. *112*, E3574–E3581.

Houghteling, P.D., and Walker, W.A. (2015). Why is initial bacterial colonization of the intestine important to infants' and children's health? J. Pediatr. Gastroenterol. Nutr. *60*, 294–307.

Hu, W., Wang, Y., Bowers, C., and Ma, H. (2003). Isolation, sequence analysis, and expression studies of florally expressed cDNAs in Arabidopsis. Plant Mol. Biol. *53*, 545–563.

Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python Environment for Tree Exploration. BMC Bioinformatics *11*, 24.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. Nature *486*, 215–221.

Hunt, K.M., Foster, J.A., Forney, L.J., Schütte, U.M.E., Beck, D.L., Abdo, Z., Fox, L.K., Williams, J.E., McGuire, M.K., and McGuire, M.A. (2011). Characterization of the diversity and temporal stability of bacterial communities in human milk. PLoS One *6*, e21313.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput. Biol. *12*, e1004957.

Huxley, R.R., Ansary-Moghaddam, A., Clifton, P., Czernichow, S., Parr, C.L., and Woodward, M. (2009). The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence. Int. J. Cancer *125*, 171–180.

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

Jameson, E., Doxey, A.C., Airs, R., Purdy, K.J., Murrell, J.C., and Chen, Y. (2016). Metagenomic data-mining reveals contrasting microbial populations responsible for trimethylamine formation in human gut and marine ecosystems. Microb Genom *2*, e000080.

Jeurink, P.V., van Bergenhenegouwen, J., Jiménez, E., Knippels, L.M.J., Fernández, L., Garssen, J., Knol, J., Rodríguez, J.M., and Martín, R. (2013). Human milk: a source of more life than we imagine. Benef. Microbes *4*, 17–30.

Jie, Z., Xia, H., Zhong, S.-L., Feng, Q., Li, S., Liang, S., Zhong, H., Liu, Z., Gao, Y., Zhao, H., et al. (2017). The gut microbiome in atherosclerotic cardiovascular disease. Nat. Commun. *8*, 845.

Johnson, C.M., Wei, C., Ensor, J.E., Smolenski, D.J., Amos, C.I., Levin, B., and Berry, D.A. (2013). Meta-analyses of colorectal cancer risk factors. Cancer Causes Control *24*, 1207–1222.

Jolley, K.A., and Maiden, M.C.J. (2010). BIGSdb: Scalable analysis of bacterial genome

variation at the population level. BMC Bioinformatics *11*, 595.

Jost, T., Lacroix, C., Braegger, C.P., Rochat, F., and Chassard, C. (2014). Vertical mother-neonate transfer of maternal gut bacteria via breastfeeding. Environ. Microbiol. *16*, 2891–2904.

Kalisch, M., and Buhlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J. Mach. Learn. Res. *8*, 613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P., and Others (2012). Causal inference using graphical models with the R package pcalg. J. Stat. Softw. *47*, 1–26.

Kalnins, G., Kuka, J., Grinberga, S., Makrecka-Kuka, M., Liepinsh, E., Dambrova, M., and Tars, K. (2015). Structure and Function of CutC Choline Lyase from Human Microbiota Bacterium Klebsiella pneumoniae. J. Biol. Chem. *290*, 21732–21740.

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

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. *40*, D109–D114.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. *42*, D199–D205.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ *3*, e1165.

Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99–103.

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

Keul, F., Hess, M., Goesele, M., and Hamacher, K. (2017). PFASUM: a substitution matrix from Pfam structural alignments. BMC Bioinformatics *18*, 293.

Khan, N.H., Ahsan, M., Yoshizawa, S., Hosoya, S., Yokota, A., and Kogure, K. (2008). Multilocus sequence typing and phylogenetic analyses of Pseudomonas aeruginosa Isolates from the ocean. Appl. Environ. Microbiol. *74*, 6194–6205.

Khoruts, A., Dicksved, J., Jansson, J.K., and Sadowsky, M.J. (2010). Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. J. Clin. Gastroenterol. *44*, 354–360.

Knapp, J.S., and Hook, E.W., 3rd (1988). Prevalence and persistence of Neisseria cinerea and other Neisseria spp. in adults. J. Clin. Microbiol. *26*, 896–900.

Koenig, J.E., Spor, A., Scalfone, N., Fricker, A.D., Stombaugh, J., Knight, R., Angenent, L.T., and Ley, R.E. (2011). Succession of microbial consortia in the developing infant gut microbiome. Proc. Natl. Acad. Sci. U. S. A. *108 Suppl 1*, 4578–4585.

Kong, H.H., Andersson, B., Clavel, T., Common, J.E., Jackson, S.A., Olson, N.D., Segre, J.A., and Traidl-Hoffmann, C. (2017). Performing Skin Microbiome Research: A Method to the Madness. J. Invest. Dermatol. *137*, 561–568.

Korpela, K., and de Vos, W.M. (2018). Early life colonization of the human gut: microbes matter everywhere. Curr. Opin. Microbiol. *44*, 70–78.

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

Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., and Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat. Biotechnol. *34*, 64–69.

Kummen, M., Vesterhus, M., Trøseid, M., Moum, B., Svardal, A., Boberg, K.M., Aukrust, P., Karlsen, T.H., Berge, R.K., and Hov, J.R. (2017). Elevated trimethylamine-N-oxide (TMAO) is associated with poor prognosis in primary sclerosing cholangitis patients with normal liver function. United European Gastroenterology Journal *5*, 532–541.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V.K., Srivastava, T.P., et al. (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. *14*, 169–181.

Lal, S., Pacis, L.B., and Smith, H.M.S. (2011). Regulation of the SQUAMOSA PROMOTER-BINDING PROTEIN-LIKE genes/microRNA156 module by the homeodomain proteins PENNYWISE and POUND-FOOLISH in Arabidopsis. Mol. Plant *4*, 1123–1132.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. *31*, 814–821.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

La Rosa, P.S., Warner, B.B., Zhou, Y., Weinstock, G.M., Sodergren, E., Hall-Moore, C.M., Stevens, H.J., Bennett, W.E., Jr, Shaikh, N., Linneman, L.A., et al. (2014). Patterned progression of bacterial populations in the premature infant gut. Proc. Natl. Acad. Sci. U. S. A. *111*, 12522–12527.

LeBlanc, J.G., Milani, C., de Giori, G.S., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. Curr. Opin. Biotechnol. *24*, 160–168.

Lee, M.D. (2019). GToTree: a user-friendly workflow for phylogenomics.

Lee, J.-Y., Baum, S.F., Alvarez, J., Patel, A., Chitwood, D.H., and Bowman, J.L. (2005). Activation of CRABS CLAW in the Nectaries and Carpels of Arabidopsis. Plant Cell *17*, 25–36.

Leiby, J.S., McCormick, K., Sherrill-Mix, S., Clarke, E.L., Kessler, L.R., Taylor, L.J., Hofstaedter, C.E., Roche, A.M., Mattei, L.M., Bittinger, K., et al. (2018). Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. Microbiome *6*, 196.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128.

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. *39*, W475–W478.

Lewis, J.D., Chen, E.Z., Baldassano, R.N., Otley, A.R., Griffiths, A.M., Lee, D., Bittinger, K., Bailey, A., Friedman, E.S., Hoffmann, C., et al. (2015). Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host Microbe *18*, 489–500.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and 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., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, S.S., Zhu, A., Benes, V., Costea, P.I., Hercog, R., Hildebrand, F., Huerta-Cepas, J., Nieuwdorp, M., Salojärvi, J., Voigt, A.Y., et al. (2016). Durable coexistence of donor and recipient strains after fecal microbiota transplantation. Science *352*, 586–589.

Lin, S. (2010). Rank aggregation methods. WIREs Comp Stat *2*, 555–570.

Liu, K., Warnow, T.J., Holder, M.T., Nelesen, S.M., Yu, J., Stamatakis, A.P., and Linder, C.R. (2012). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst. Biol. *61*, 90–106.

Loman, N.J., Constantinidou, C., Christner, M., Rohde, H., Chan, J.Z.-M., Quick, J., Weir, J.C., Quince, C., Smith, G.P., Betley, J.R., 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.

Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. Nature *489*, 220–230.

Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. *3*, e104.

Maathuis, M.H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. Nat. Methods *7*, 247–248.

Mackelprang, R., Waldrop, M.P., DeAngelis, K.M., David, M.M., Chavarria, K.L., Blazewicz, S.J., Rubin, E.M., and Jansson, J.K. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature *480*, 368–371.

Maddison, D.R., Swofford, D.L., and Maddison, W.P. (1997). NEXUS: an extensible file format for systematic information. Syst. Biol. *46*, 590–621.

Mai, U., and Mirarab, S. (2017). TreeShrink: Efficient Detection of Outlier Tree Leaves. In Comparative Genomics, (Springer International Publishing), pp. 116–140.

Makino, H., Kushiro, A., Ishikawa, E., Muylaert, D., Kubota, H., Sakai, T., Oishi, K., Martin, R., Ben Amor, K., Oozeer, R., et al. (2011). Transmission of intestinal Bifidobacterium longum subsp. longum strains from mother to infant, determined by multilocus sequencing

typing and amplified fragment length polymorphism. Appl. Environ. Microbiol. *77*, 6788–6793.

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.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., DREAM5 Consortium, Kellis, M., Collins, J.J., et al. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.

Marcobal, A., Barboza, M., Sonnenburg, E.D., Pudlo, N., Martens, E.C., Desai, P., Lebrilla, C.B., Weimer, B.C., Mills, D.A., German, J.B., et al. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. Cell Host Microbe *10*, 507–514.

Margolin, A.A., Wang, K., Lim, W.K., Kustagi, M., Nemenman, I., and Califano, A. (2006a). Reverse engineering cellular networks. Nat. Protoc. *1*, 662–671.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006b). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics *7 Suppl 1*, S7.

Marri, P.R., Paniscus, M., Weyand, N.J., Rendón, M.A., Calton, C.M., Hernández, D.R., Higashi, D.L., Sodergren, E., Weinstock, G.M., Rounsley, S.D., et al. (2010). Genome sequencing reveals widespread virulence gene exchange among human Neisseria species. PLoS One *5*, e11835.

Mason, C., Afshinnekoo, E., Ahsannudin, S., Ghedin, E., Read, T., Fraser, C., Dudley, J., Hernandez, M., Bowler, C., Stolovitzky, G., et al. (2016). The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. Microbiome *4*, 24.

Maurice, C.F., Haiser, H.J., and Turnbaugh, P.J. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell *152*, 39–50.

McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., et al. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience *1*, 7.

McKinney, W. (2012). pandas: a Foundational Python Library for Data Analysis and Statistics. O'Reilly Media, Inc.

de Medeiros, R.B., Figueiredo, J., Resende, R. de O., and De Avila, A.C. (2005). Expression of a viral polymerase-bound host factor turns human cell lines permissive to a plant- and insect-infecting virus. Proc. Natl. Acad. Sci. U. S. A. *102*, 1175–1180.

Meheust, R., Burstein, D., Castelle, C.J., and Banfield, J.F. (2018). Biological capacities clearly define a major subdivision in Domain Bacteria.

Meysman, P., Sonego, P., Bianco, L., Fu, Q., Ledezma-Tejeida, D., Gama-Castro, S., Liebens, V., Michiels, J., Laukens, K., Marchal, K., et al. (2014). COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. Nucleic Acids Res. *42*, D649–D653.

Milani, C., Mancabelli, L., Lugli, G.A., Duranti, S., Turroni, F., Ferrario, C., Mangifesta, M.,

Viappiani, A., Ferretti, P., Gorfer, V., et al. (2015). Exploring Vertical Transmission of Bifidobacteria from Mother to Child. Appl. Environ. Microbiol. *81*, 7078–7087.

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics *31*, i44–i52.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics *30*, i541–i548.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. J. Comput. Biol. *22*, 377–386.

Miyoshi, J., Bobe, A.M., Miyoshi, S., Huang, Y., Hubert, N., Delmont, T.O., Eren, A.M., Leone, V., and Chang, E.B. (2017). Peripartum Antibiotics Promote Gut Dysbiosis, Loss of Immune Tolerance, and Inflammatory Bowel Disease in Genetically Prone Offspring. Cell Rep. *20*, 491–504.

Monchamp, M.-E., Walser, J.-C., Pomati, F., and Spaak, P. (2016). Sedimentary DNA Reveals Cyanobacterial Community Diversity over 200 Years in Two Perialpine Lakes. Appl. Environ. Microbiol. *82*, 6472–6482.

Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. *13*, R79.

Morowitz, M.J., Denef, V.J., Costello, E.K., Thomas, B.C., Poroyko, V., Relman, D.A., and Banfield, J.F. (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proc. Natl. Acad. Sci. U. S. A. *108*, 1128–1133.

Mulkidjanian, A.Y., Makarova, K.S., Galperin, M.Y., and Koonin, E.V. (2007). Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. Nat. Rev. Microbiol. *5*, 892–899.

Musser, J.M., and Kapur, V. (1992). Clonal analysis of methicillin-resistant Staphylococcus aureus strains from intercontinental sources: association of the mec gene with divergent phylogenetic lineages implies dissemination by horizontal transfer and recombination. J. Clin. Microbiol. *30*, 2058–2063.

Muzzi, A., Mora, M., Pizza, M., Rappuoli, R., and Donati, C. (2013). Conservation of meningococcal antigens in the genus Neisseria. MBio *4*, e00163–13.

Nayfach, S., Bradley, P.H., Wyman, S.K., Laurent, T.J., Williams, A., Eisen, J.A., Pollard, K.S., and Sharpton, T.J. (2015). Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. PLoS Comput. Biol. *11*, e1004573.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015a). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

Nguyen, N.-P., Mirarab, S., Liu, B., Pop, M., and Warnow, T. (2014). TIPP: taxonomic identification and phylogenetic profiling. Bioinformatics *30*, 3548–3555.

Nguyen, N.-P.D., Mirarab, S., Kumar, K., and Warnow, T. (2015b). Ultra-large alignments using phylogeny-aware profiles. Genome Biol. *16*, 124.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., 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.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217.

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

Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 6505.

Oellgaard, J., Abitz Winther, S., Schmidt Hansen, T., Rossing, P., and Johan von Scholten, B. (2017). Trimethylamine N-oxide (TMAO) as a New Potential Therapeutic Target for Insulin Resistance and Cancer. Curr. Pharm. Des. *23*, 3699–3712.

Ogilvie, L.A., and Jones, B.V. (2015). The human gut virome: a multifaceted majority. Front. Microbiol. *6*, 918.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., et al. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics *20*, 3045–3054.

Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics *12*, 385.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132.

Ou, J., DeLany, J.P., Zhang, M., Sharma, S., and O'Keefe, S.J.D. (2012). Association between low colonic short-chain fatty acids and high bile acids in high colon cancer risk populations. Nutr. Cancer *64*, 34–40.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693.

Palm, N.W., de Zoete, M.R., and Flavell, R.A. (2015). Immune-microbiota interactions in health and disease. Clin. Immunol. *159*, 122–127.

Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A., and Brown, P.O. (2007). Development of the human infant intestinal microbiota. PLoS Biol. *5*, e177.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol *2*, 1533–1542.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and

Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D.T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nat. Methods *14*, 1023.

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 *0*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Penn, O., Privman, E., Landan, G., Graur, D., and Pupko, T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. Mol. Biol. Evol. *27*, 1759–1767.

Perez-Muñoz, M.E., Arrieta, M.-C., Ramer-Tait, A.E., and Walter, J. (2017). A critical assessment of the “sterile womb” and “in utero colonization” hypotheses: implications for research on the pioneer infant microbiome. Microbiome *5*, 48.

Peternelli, L.A., and Rosa, G.J.M. A novel approach for subset selection of SNP markers for cost-effective implementation of genomic selection.

Popic, V., Kuleshov, V., Snyder, M., and Batzoglou, S. (2018). Fast Metagenomic Binning via Hashing and Bayesian Clustering. J. Comput. Biol. *25*, 677–688.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., 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.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. *41*, D590–D596.

Quince, C., Ijaz, U.Z., Loman, N., Eren, A.M., Saulnier, D., Russell, J., Haig, S.J., Calus, S.T., Quick, J., Barclay, A., et al. (2015). Extensive Modulation of the Fecal Metagenome in Children With Crohn’s Disease During Exclusive Enteral Nutrition. Am. J. Gastroenterol. *110*, 1718–1729; quiz 1730.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833.

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

Ramirez, K.S., Leff, J.W., Barberán, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., et al. (2014). Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. Proc. Biol. Sci. *281*.

Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr. Biol. *25*, 1682–1693.

Ramsay, D.T., Kent, J.C., Owens, R.A., and Hartmann, P.E. (2004). Ultrasound imaging of milk ejection in the breast of lactating women. Pediatrics *113*, 361–367.

Ramulu, H.G., Groussin, M., Talla, E., Planel, R., Daubin, V., and Brochier-Armanet, C. (2014). Ribosomal proteins: toward a next generation standard for prokaryotic systematics? Mol. Phylogenet. Evol. *75*, 103–117.

Rath, S., Heidrich, B., Pieper, D.H., and Vital, M. (2017). Uncovering the trimethylamine-producing bacteria of the human gut microbiota. Microbiome *5*, 54.

Razin, S. (1992). Peculiar properties of mycoplasmas: the smallest self-replicating prokaryotes. FEMS Microbiol. Lett. *100*, 423–431.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature *466*, 334–338.

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., and Gordon, J.I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. Nat. Rev. Microbiol. *10*, 607–617.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431–437.

Romano, K.A., Vivas, E.I., Amador-Noguez, D., and Rey, F.E. (2015). Intestinal microbiota composition modulates choline bioavailability from diet and accumulation of the proatherogenic metabolite trimethylamine-N-oxide. MBio *6*, e02481.

Sánchez-Corrales, Y.-E., Álvarez-Buylla, E.R., and Mendoza, L. (2010). The Arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process. J. Theor. Biol. *264*, 971–983.

Saulnier, D.M., Riehle, K., Mistretta, T.-A., Diaz, M.-A., Mandal, D., Raza, S., Weidler, E.M., Qin, X., Coarfa, C., Milosavljevic, A., et al. (2011). Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology *141*, 1782–1791.

Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B., et al. (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. Elife *2*, e01202.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J., et al. (2013). Genomic variation landscape of the human gut microbiome. Nature *493*, 45–50.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. *75*, 7537–7541.

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

Schrader, C., Schielke, A., Ellerbroek, L., and Johne, R. (2012). PCR inhibitors - occurrence, properties and removal. J. Appl. Microbiol. *113*, 1014–1026.

Scott, K.P., Antoine, J.-M., Midtvedt, T., and van Hemert, S. (2015). Manipulating the gut microbiota to maintain health and treat disease. Microb. Ecol. Health Dis. *26*, 25877.

Sczesnak, A., Segata, N., Qin, X., Gevers, D., Petrosino, J.F., Huttenhower, C., Littman, D.R., and Ivanov, I.I. (2011). The genome of th17 cell-inducing segmented filamentous bacteria reveals extensive auxotrophy and adaptations to the intestinal environment. Cell Host Microbe *10*, 260–272.

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

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome Biol. *12*, R60.

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

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

Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. *4*, 2304.

Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. *43*, W7–W14.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. *23*, 111–120.

Shi, J.X., Malitsky, S., De Oliveira, S., Branigan, C., Franke, R.B., Schreiber, L., and

Aharoni, A. (2011). SHINE transcription factors act redundantly to pattern the archetypal surface of Arabidopsis flower organs. PLoS Genet. *7*, e1001388.

Shin, H., Pei, Z., Martinez, K.A., 2nd, Rivera-Vinas, J.I., Mendez, K., Cavallin, H., and Dominguez-Bello, M.G. (2015). The first microbial environment of infants born by C-section: the operating room microbes. Microbiome *3*, 59.

Shogan, B.D., Smith, D.P., Christley, S., Gilbert, J.A., Zaborina, O., and Alverdy, J.C. (2014). Intestinal anastomotic injury alters spatially defined microbiome composition and function. Microbiome *2*, 35.

Song, S.J., Lauber, C., Costello, E.K., Lozupone, C.A., Humphrey, G., Berg-Lyons, D., Caporaso, J.G., Knights, D., Clemente, J.C., Nakielny, S., et al. (2013). Cohabiting family members share microbiota with one another and with their dogs. Elife *2*, e00458.

Spirtes, P., and Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. Soc. Sci. Comput. Rev. *9*, 62–72.

Spirtes, P., Glymour, C.N., and Scheines, R. (2002). Cele-6rities (Systems Design Limited).

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

Stecher, B., and Hardt, W.-D. (2011). Mechanisms controlling pathogen colonization of the gut. Curr. Opin. Microbiol. *14*, 82–91.

Suez, J., Zmora, N., Zilberman-Schapira, G., Mor, U., Dori-Bachash, M., Bashiardes, S., Zur, M., Regev-Lehavi, D., Ben-Zeev Brik, R., Federici, S., et al. (2018). Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. Cell *174*, 1406–1423.e16.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science *348*, 1261359.

Tailford, L.E., Crost, E.H., Kavanaugh, D., and Juge, N. (2015). Mucin glycan foraging in the human gut microbiome. Front. Genet. *6*, 81.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.

Tamburini, S., Shen, N., Wu, H.C., and Clemente, J.C. (2016). The microbiome in early life: implications for health outcomes. Nat. Med. *22*, 713–722.

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., and Dessimoz, C. (2015). Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. Syst. Biol. *64*, 778–791.

Tan, M., AlShalalfa, M., Alhajj, R., and Polat, F. (2008). Combining multiple types of biological data in constraint-based learning of gene regulatory networks. In 2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 90–97.

Tan, M., Alshalalfa, M., Alhajj, R., and Polat, F. (2011). Influence of prior knowledge in constraint-based learning of gene regulatory networks. IEEE/ACM Trans. Comput. Biol.

Bioinform. *8*, 130–142.

Tanay, A., and Shamir, R. (2001). Computational expansion of genetic networks. Bioinformatics *17 Suppl 1*, S270–S278.

Tett, A., Pasolli, E., Farina, S., Truong, D.T., Asnicar, F., Zolfo, M., Beghini, F., Armanini, F., Jousson, O., De Sanctis, V., et al. (2017). Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. NPJ Biofilms Microbiomes *3*, 14.

The Tree of Sex Consortium (2014). Tree of Sex: A database of sexual systems. Scientific Data *1*, 140015.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. Nat. Protoc. *4*, 470–483.

Tinsley, C.R., and Nassif, X. (1996). Analysis of the genetic differences between Neisseria meningitidis and Neisseria gonorrhoeae: two closely related bacteria expressing two different pathogenicities. Proc. Natl. Acad. Sci. U. S. A. *93*, 11109–11114.

Treangen, T.J., Ambur, O.H., Tonjum, T., and Rocha, E.P.C. (2008). The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biol. *9*, R60.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

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

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Turnbaugh, P.J., Quince, C., Faith, J.J., McHardy, A.C., Yatsunenko, T., Niazi, F., Affourtit, J., Egholm, M., Henrissat, B., Knight, R., et al. (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc. Natl. Acad. Sci. U. S. A. *107*, 7503–7508.

Turroni, F., Milani, C., van Sinderen, D., and Ventura, M. (2011a). Genetic strategies for mucin metabolism in Bifidobacterium bifidum PRL2010: an example of possible human-microbe co-evolution. Gut Microbes *2*, 183–189.

Turroni, F., Foroni, E., Serafini, F., Viappiani, A., Montanini, B., Bottacini, F., Ferrarini, A., Bacchini, P.L., Rota, C., Delledonne, M., et al. (2011b). Ability of Bifidobacterium breve to grow on different types of milk: exploring the metabolism of milk through genome analysis. Appl. Environ. Microbiol. *77*, 7408–7417.

Turroni, F., Peano, C., Pass, D.A., Foroni, E., Severgnini, M., Claesson, M.J., Kerr, C., Hourihane, J., Murray, D., Fuligni, F., et al. (2012). Diversity of bifidobacteria within the infant gut microbiota. PLoS One *7*, e36957.

Vachaspati, P., and Warnow, T. (2015). ASTRID: Accurate Species TRees from Internode Distances. BMC Genomics *16 Suppl 10*, S3.

Valdar, W.S.J. (2002). Scoring residue conservation. Proteins *48*, 227–241.

Vatanen, T., Plichta, D.R., Somani, J., Münch, P.C., Arthur, T.D., Hall, A.B., Rudolf, S., Oakeley, E.J., Ke, X., Young, R.A., et al. (2018). Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nature Microbiology.

Victoria, J.G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S., and Delwart, E. (2009). Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. *83*, 4642–4651.

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

Walker, A.W., Ince, J., Duncan, S.H., Webster, L.M., Holtrop, G., Ze, X., Brown, D., Stares, M.D., Scott, P., Bergerat, A., et al. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J. *5*, 220–230.

Wallace, R.J., Rooke, J.A., McKain, N., Duthie, C.-A., Hyslop, J.J., Ross, D.W., Waterhouse, A., Watson, M., and Roehe, R. (2015). The rumen microbial metagenome associated with high methane production in cattle. BMC Genomics *16*, 839.

Wampach, L., Heintz-Buschart, A., Hogan, A., Muller, E.E.L., Narayanasamy, S., Laczny, C.C., Hugerth, L.W., Bindl, L., Bottu, J., Andersson, A.F., et al. (2017). Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life. Front. Microbiol. *8*, 434.

Wampach, L., Heintz-Buschart, A., Fritz, J.V., Ramiro-Garcia, J., Habier, J., Herold, M., Narayanasamy, S., Kaysen, A., Hogan, A.H., Bindl, L., et al. (2018). Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. Nat. Commun. *9*, 5091.

Wang, M., Augusto Benedito, V., Xuechun Zhao, P., and Udvardi, M. (2010). Inferring large-scale gene regulatory networks using a low-order constraint-based algorithm. Mol. Biosyst. *6*, 988–998.

Ward, D.V., Scholz, M., Zolfo, M., Taft, D.H., Schibler, K.R., Tett, A., Segata, N., and Morrow, A.L. (2016). Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. Cell Rep. *14*, 2912–2924.

Ward, T.L., Hosid, S., Ioshikhes, I., and Altosaar, I. (2013). Human milk metagenome: a functional capacity analysis. BMC Microbiol. *13*, 116.

Waters, N., Brennan, F., Holmes, A., Abram, F., and Pritchard, L. (2018). Easily phylotyping E. coli via the EzClermont web app and command-line tool.

Webb, A.E., Walsh, T.A., and O'Connell, M.J. (2017). VESPA: Very large-scale Evolutionary and Selective Pressure Analyses. PeerJ Comput. Sci. *3*, e118.

Wei, E.K., Giovannucci, E., Wu, K., Rosner, B., Fuchs, C.S., Willett, W.C., and Colditz, G.A. (2004). Comparison of risk factors for colon and rectal cancer. Int. J. Cancer *108*, 433–442.

Wheeler, T.J., and Kececioglu, J.D. (2007). Multiple alignment by aligning alignments. Bioinformatics *23*, i559–i568.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C., Wilkins, M.J., Hettich, R.L., Lipton, M.S., Williams, K.H., et al. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science *337*, 1661–1665.

Wu, Y.-W. (2018). ezTree: an automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among uncultivated prokaryotic draft genomes. BMC Genomics *19*, 921.

Wu, M., and Eisen, J.A. (2008). A simple, fast, and accurate method of phylogenomic inference. Genome Biol. *9*, R151.

Wu, M., and Scott, A.J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics *28*, 1033–1034.

Wu, G., Zhao, H., Li, C., Rajapakse, M.P., Wong, W.C., Xu, J., Saunders, C.W., Reeder, N.L., Reilman, R.A., Scheynius, A., et al. (2015). Genus-Wide Comparative Genomics of Malassezia Delineates Its Phylogeny, Physiology, and Niche Adaptation on Human Skin. PLoS Genet. *11*, e1005614.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science *334*, 105–108.

Xiao, L., Feng, Q., Liang, S., Sonne, S.B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D., et al. (2015). A catalog of the mouse gut metagenome. Nat. Biotechnol. *33*, 1103–1108.

Ximenez, C., and Torres, J. (2017). Development of Microbiota in Infants and its Role in Maturation of Gut Mucosa and Immune System. Arch. Med. Res. *48*, 666–680.

Xu, Z., Hansen, M.A., Hansen, L.H., Jacquiod, S., and Sørensen, S.J. (2014). Bioinformatic approaches reveal metagenomic characterization of soil microbial community. PLoS One *9*, e93445.

Yamada, K., and Tomii, K. (2014). Revisiting amino acid substitution matrices for identifying distantly related proteins. Bioinformatics *30*, 317–325.

Yang, F., Zeng, X., Ning, K., Liu, K.-L., Lo, C.-C., Wang, W., Chen, J., Wang, D., Huang, R., Chang, X., et al. (2012). Saliva microbiomes distinguish caries-active from healthy human populations. ISME J. *6*, 1–10.

Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S.M., 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.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.

Yeoman, C.J., Chia, N., Jeraldo, P., Sipos, M., Goldenfeld, N.D., and White, B.A. (2012). The microbiome of the chicken gastrointestinal tract. Anim. Health Res. Rev. *13*, 89–99.

Yoo, S.K., Wu, X., Lee, J.S., and Ahn, J.H. (2011). AGAMOUS-LIKE 6 is a floral promoter that negatively regulates the FLC/MAF clade genes and positively regulates FT in Arabidopsis. Plant J. *65*, 62–76.

Yu, J., Feng, Q., Wong, S.H., Zhang, D., Liang, Q.Y., 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., Puigbò, P., Koonin, E.V., and Wolf, Y.I. (2012). Phylogenomics of prokaryotic ribosomal proteins. PLoS One *7*, e36972.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized Nutrition by Prediction of Glycemic Responses. Cell *163*, 1079–1094.

Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., 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, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics *19*, 153.

Zhang, T., Breitbart, M., Lee, W.H., Run, J.-Q., Wei, C.L., Soh, S.W.L., Hibberd, M.L., Liu, E.T., Rohwer, F., and Ruan, Y. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. *4*, e3.

Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.-K., Liu, Z.-P., and Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics *28*, 98–104.

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.

Zik, M., and Irish, V.F. (2003). Global Identification of Target Genes Regulated by APETALA3 and PISTILLATA Floral Homeotic Gene Action. Plant Cell *15*, 207–222.

Zmora, N., Zilberman-Schapira, G., Suez, J., Mor, U., Dori-Bachash, M., Bashiardes, S., Kotler, E., Zur, M., Regev-Lehavi, D., Brik, R.B.-Z., et al. (2018). Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. Cell *174*, 1388–1405.e21.