UNIVERSITY OF TRENTO - Italy

**International PhD Program in Biomolecular Sciences**

**Department of Cellular, Computational and Integrative Biology – CIBIO**

**XXXI Cycle**

# Strain-level (meta)genomic profiling of bacteria from hospital pathogens to non-human primate commensals


**Tutor**

Prof. Nicola Segata

*Department of Cellular, Computational and Integrative Biology - CIBIO*

*University of Trento, Italy*

**Ph.D. Thesis of**

Serena Manara

*Department of Cellular, Computational and Integrative Biology - CIBIO*

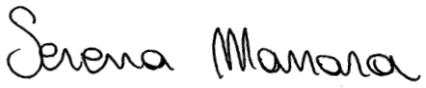
*University of Trento, Italy*



Academic Year 2017-2018

**Declaration**

I Serena Manara confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.



# Table of Contents

## Abstract

Studying microbial organisms at the level of single genetic variants (strains) is key not only for human pathogens but also for commensal members of the human microbiome. However, several limitations make isolation-based methods only partially effective in surveying the complexity of host-associated microbial communities. Novel biotechnological advances are revolutionizing the study of host-associated microbes, enabling the transition from low-resolution cultivation-based typing to cultivation-free metagenomic characterizations. In my doctoral work, I tested the hypothesis that appropriate analytical tools applied to genomic and metagenomic data can provide information about microbes at a resolution comparable to that of cultivation-based methods. To this end, I employed a set of integrated methods to reconstruct the genome and analyse the functional and transmission patterns of pathogenic and commensal microbes across human and non-human hosts in different contexts.

We initially focused on the whole-genome sequencing of a cohort of 184 *Staphylococcus aureus* infections from patients with a set of diverse diseases at multiple departments of Meyer's Children Hospital in Florence, Italy. By applying a combination of isolation-based techniques and computational analysis, we surveyed the epidemiology, transmission patterns, and genomic features associated with both highly-studied and under-investigated *S. aureus* clones. We identified new infective clones and two novel variants of the beta-lactam resistance cassette. We moreover profiled the virulence and resistance factors typically associated with this opportunistic pathogen, observed the dispensability of genes previously considered as putative targets for vaccine development, and tracked the transmission of a newly-emerging epidemic clone.

We then focused on the challenging task of extracting strain-level genomic information from cultivation-free metagenomic sequencing of stool samples obtained from mothers and their infants during the first year of life. By applying genetic and pangenomic profiling tools, we showed that the spread of microbiome members can be inferred from metagenomes directly, and we tracked the vertical transmission of microbial strains from mother to her infant and their corresponding transcriptional profiles. This pilot study laid the foundations for larger cohort studies investigating microbiome transmission via metagenomic sequencing.

The next step was the application of cultivation-free approaches to identify and survey currently neglected host-associated microbes. In order to explore those species lacking relatively close already sequenced genomes, we performed a large-scale assembly-based analysis to reconstruct high-quality microbial genomes for species and strains in the under-investigated microbiome of non-human primates (NHPs). Overall, less than one-quarter of the recovered genomes were assigned to known species or species previously observed in the human microbiome. The remaining genomes were assigned to over 1,000 new species, which improved the mappable fraction of NHP metagenomes by over 600%.

The analysis of this newly-established catalog of NHP-associated species in the context of available human-associated microbial genomes further exposed the loss of biodiversity from wild and captive NHPs to non-Westernized and Westernized human populations, showing that microbiome members shared between NHPs and humans mostly belong to uncharacterized species that are heavily lifestyle-dependent.

Through characterization of a cohort of *Staphylococcus aureus* isolates, tracking of transmission of commensals from mother to infant, and recovering of microbial dark matter associated with non-human primates, we showed that cultivation-free profiling of known and unknown host-associated species can achieve a resolution for comparative genomics that is close to that available for isolate sequencing.

# 1. Introduction to the thesis

Microbes living in and on the human body play important roles in host physiology and metabolism (HMP et al. 2012; Qin et al. 2010; Clemente et al. 2012). They are for example responsible for degrading otherwise indigestible food components (Bäckhed et al. 2005), preventing pathogen colonization (Stecher and Hardt 2011) and modulating host immune system (Palm, de Zoete, and Flavell 2015). The set of bacteria, archaea, viruses, and microeukaryotes associated with a specific host are defined as its microbiome. The human microbiome has been extensively studied in the past fifteen years by a few large-scale investigation initiatives (HMP et al. 2012; Qin et al. 2010) and a large number of smaller studies focusing on specific body sites and conditions, which together uncovered a substantial fraction of the overall human microbiome diversity. Most research has however focused on characterizing the relative abundances of genera (16S rRNA amplicon sequencing) and species (shotgun metagenomics), without reaching the resolution of single microbial organisms and their specific genetic variants usually called strains.

Characterizing and discovering host-associated microbes at the level of single strains is indeed key for both human pathogens and commensals, mainly because of the substantial phenotypic and genotypic differences between strains belonging to the same species (Schloissnig et al. 2013; Truong et al. 2017; Segata 2018). One illuminating example of this within-species diversity is *Escherichia coli*, a common gut commensal that is however also well-known for its pathogenic strains linked with gastroenteritis and the haemolytic–uremic syndrome (Frank et al. 2011), necrotizing enterocolitis in preterm infants (Ward et al. 2016a) and cancer onset (Cuevas-Ramos et al. 2010). Other less-investigated examples are *Prevotella copri*, whose different strains are linked with low-fat high-fiber healthy diets (De Filippis et al. 2019) or with increased risk of rheumatoid arthritis (Scher et al. 2013), and *Eggerthella lenta*, whose specific strains are able to degrade and therefore inactivate the cardiac drug digoxin with largely unpredictable negative therapeutic outcomes (Haiser et al. 2013). Being able to characterize microbial communities at the level of single strains to perform comparative genomics analysis is thus crucial to resolve host-microbe relationships, and has relevant implications for human health (Segata 2018).

Cultivation-based studies have been successful in deeply characterizing specific human-associated microbes, especially after whole-genome sequencing (WGS) allowed deep mining of their genome. Early investigations based on WGS focused mainly on pathogens (Cole et al. 1998; Fraser et al. 1998; Parkhill et al. 2000; Nelson et al. 2003) or opportunistic pathogens. This is, for instance, the case of methicillin-resistant *Staphylococcus aureus* and its large repertoire of previously unknown virulence genes (Kuroda et al. 2001), and of *Pseudomonas aeruginosa*, with its large genome full of regulatory genes explaining plasticity and intrinsic drug resistance (Stover et al. 2000). In many cases, whole-genome isolate sequencing allowed to tell apart less harmful strains and those associated with enhanced risks of life-threatening diseases, as in the case of *Helicobacter pylori* (Blaser 1997; Atherton et al. 1997; Alm et al. 1999). In the following

decade, cultivation-based studies allowed the investigation of most of the relevant human and non-human pathogens, as well as some commensals or bacteria of particular interest (Pallen and Wren 2007; Ventura et al. 2009). However, despite the high resolution that genome sequencing now allows, cultivation-based approaches can only partially survey the diversity of host-associated microbial communities.

The limitations of cultivation-based techniques in surveying the strain-level epidemiology of known and unknown microbes are mainly technical, with many bacterial species being slow-growers, requiring co-cultivation, or specific conditions that cannot be easily foreseen. Nevertheless, cultivation-recalcitrant species are not the only reason for which cultivation-based methods are not enough to survey host-associated microbial diversity. Even with the new methods that expanded the fraction of microbes that can be successfully grown *in vitro* (Stewart 2012), isolation is a labour-intensive task and would be impossible to apply it on a large scale to assess the entire diversity of a species and potentially to decipher the yet-to-be-characterized fraction of the microbial world.

In recent years, new opportunities to access cultivation-recalcitrant species have been offered by metagenomics, which studies microbial communities by sequencing the total genetic content of a sample of interest. Many metagenomic studies have focused on characterizing the associations of specific microbiome compositions or species with health and disease (HMP et al. 2012; Qin et al. 2012; Durbán et al. 2013; Zeller et al. 2014; Nielsen et al. 2014; Zhang et al. 2015; Ward et al. 2016b; Vogtmann and Goedert 2016). With the cost of next-generation sequencing continuing to decrease, shotgun metagenomics data have become increasingly available enabling the meta-analysis of large datasets and the recovery of even greater microbial diversity. However, analytical limitations in the ability to mine metagenomes at large scale with strain-level resolution are an obstacle to really uncover the strain diversity and dynamics of the worldwide human microbiome. When I started my doctoral studies, it was indeed unclear whether metagenomics could reach the required precision to study microbes as single strains, track them in complex communities and confidently discover new species from metagenomic data.

## 1.1 Aims and main contribution of the thesis

In this thesis, I employed several experimental and computational tools applied on different clinically-relevant settings to **verify the main hypothesis that metagenomics can be empowered with the needed strain-level resolution to complement and extend cultivation-based methods to characterize, track and discover new microbial organisms**. The studies reported in this thesis investigated pathogenic and commensal host-associated microbial strains through the application of different cultivation-based and cultivation-free methods, ranging from epidemiological analysis from whole-genome isolate sequencing to the reconstruction of potentially uncharacterized strains from non-human metagenomes. Particular focus has been put on the assessment of the functional and transmission patterns, as well as on the microbe-host relationship.

More specifically, I worked on three main methodological aspects:

1. assess the depth of genomic resolution that can be reached with whole-genome isolate sequencing of a relatively well-known opportunistic pathogen (*Staphylococcus aureus*);

2. test the feasibility of tracking multiple potentially uncultivable members of the microbiome at once by extracting strain-level genomic information from cultivation-free metagenomic data and matching them across samples;

3. reconstruct genomes of microbial strains associated with under-investigated hosts, for the purpose of discovering new uncharacterized microbes.

Simultaneously, I aimed at answering questions of biomedical relevance:

1. *How complex is the epidemiology of Staphylococcus aureus infections in a hospital?* Are we assessing the entire genetic variability of this opportunistic pathogen? Are there new infective clones and how can we effectively survey them? Is there an epidemiological reason for the current lack of a vaccine against *S. aureus*, such as dispensability of genes encoding for putative vaccine targets?

2. *How is the infant microbiome established and which are the most relevant sources?* What is the role of vertical transmission of microbes from the mother in the establishment of the infant gut microbiome? How many and which species are vertically transmitted, and are they active in the gut of the infant?

3. *What is the extent of host-specific microbiome in primates?* How similar are the human and non-human gut microbiomes? Can we find potential indication of coevolution or species-loss linked with industrialized lifestyles?

The works reported in this thesis contributed to the study of host-associated microbial strains by proving the possibilities related to in-depth analysis of whole-genome isolate sequencing for pathogens surveillance and by showing how strain-level metagenomics can reach a comparable resolution and even access still uncharacterized microbial species associated with under-investigated hosts. Overall, we showed that strain-level metagenomic approaches can complement and expand isolation-based comparative genomic analysis and drive the discovery of new microbial clades. As discussed in the corresponding sections in the introduction of each chapter, our framework contributed to the current wave of metagenomic analyses that are answering to questions previously addressable only by cultivation-based assays.

## 1.2 Organization of the thesis

This thesis is structured into chapters, each one reporting published (**Chapters 2**, **3** and **5**) or unpublished (**Chapter 4**) manuscripts describing the work I performed during my doctoral studies. A brief introduction and discussion to each article and the link with the overall thesis rationale are reported at the beginning of each chapter, together with a statement about my specific contribution to each study.

- **Chapter 2** reports the article "*Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of Staphylococcus aureus strains in a paediatric hospital*" published in Genome Medicine in 2018. In this work, we applied whole-genome isolate sequencing to survey the epidemiology, the genetic traits and the transmission patterns of a cohort of *S. aureus* infections.

- **Chapter 3** reports the article "*Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling*" published in mSystems in 2017. In this work, we applied a cultivation-free approach to study the transmission of microbiome members from mother to infant at the strain-level and to extract relevant genomic and functional information.

- **Chapter 4** reports the article "*Microbial genomes from gut metagenomes of non-human primates expand the primate-associated bacterial tree-of-life with over 1,000 novel species*", which is ready to be submitted to a scientific journal. In this work, we applied a metagenomic assembly and binning approach to reconstruct microbial genomes from non-human primate metagenomes, which included a large number of yet-to-be-characterized species and functions.

- **Chapter 5** reports the abstracts and a brief introduction to four articles I contributed to during collaborative work I performed during my studies. How these articles I co-authored are linked with those reported in the previous chapters is reported in the introduction to each specific study.

# Chapter 2. Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of *Staphylococcus aureus* strains in a paediatric hospital

## 2.1 Introduction to the chapter

This chapter reports the work I performed during the first period of my doctoral studies, which aimed at understanding and characterize at the molecular level the epidemiology of *Staphylococcus aureus* infecting patients in multiple departments at Meyer's Children Hospital in Florence. In this first work, we applied a clinical microbiology approach based on the isolation of this relatively well-known opportunistic pathogen followed by whole-genome sequencing and analysis. Although many studies focused on *Staphylococcus aureus* in the past, our more general goal was to understand at which level of resolution this bacterium can be studied using whole-genome sequencing, with specific focus on the transmission pattern, intra-hospital evolutionary history, and novel antibiotic-resistance variants. We thus performed a large body of analysis on the sequenced isolate genomes, which allowed us to deeply characterize our cohort, identify novel strains and variants, and to explore the genetic variability and pathogenic potential of *S. aureus*. Overall, we conclude that whole-genome sequencing is very effective for easy-to-isolate pathogens for which many reference genomes are already available in public repositories.

We focused on *Staphylococcus aureus* because this microbe is one of the most dangerous pathogens in hospital settings. *S. aureus* is, in fact, a common inhabitant of the skin and upper airways (Wertheim et al. 2005; Mainous et al. 2006; Tong, Davis, et al. 2015) but also an opportunistic pathogen causing mild to life-threatening infections (Tong, Davis, et al. 2015; Esposito, Noviello, and Leone 2016; Rhee et al. 2015). Invasive *S. aureus* infections have an extremely high mortality in absence of effective treatments such as antibiotics (Peacock and Paterson 2015), to which *S. aureus* is particularly prone to acquire resistances (Rammelkamp and Maxon 1942; Bondi and Dietz 1945). The most well-known and studied is the resistance to methicillin (MRSA) (Peacock and Paterson 2015), but *S. aureus* has proven able to develop resistance to all antibiotics that entered the clinics so far (Pantosti, Sanchini, and Monaco 2007; Monaco et al. 2017; Raad et al. 2007; Hiramatsu 2001), and its vancomycin-intermediate methicillin-resistant variant has been indicated by World Health Organization as high priority for research and development of new antibiotics (Tacconelli et al. 2018). Many studies have investigated the global and local epidemiology of *S. aureus*, but have often focused on the most virulent MRSA strains (Voss and Doebbeling 1995; Stefani et al. 2012; Chambers and Deleo 2009; Harris et al. 2010), despite most lethal nosocomial infections are caused by methicillin-sensitive clones (MSSA) (Sievert et al. 2013; Monaco et al. 2017). Studies unbiasedly addressing both the highly virulent or resistant population and the more mild variants of *S. aureus* are thus crucial for global surveillance and for the identification of new infective clones. However, such research is still currently limited (Copin, Shopsin, and Torres 2018).

The lack of integrated large-scale analysis for the epidemiology of *S. aureus* in a hospital setting thus motivated our work.

In the article reported in this Chapter and published in Genome Medicine in 2018, we thus studied the epidemiology and genetics of *S. aureus* by whole-genome sequencing 234 isolates obtained from 160 patients under treatment in different departments of Anna Meyer Children's University Hospital (Florence, Italy). Overall, we limited downstream analysis to investigate only high-quality genomes, which covered 135 patients affected by different diseases. Reconstruction of the whole cohort phylogeny and the combination of four different *in-silico* typing methods (Multi-Locus Sequence Typing and typing of the Staphylococcal Cassette Chromosome *mec*, of the *spa* gene, and of the Panton-Valentine Leukocidin) highlighted the high diversity of the *S. aureus* community, with 80 different lineages covering local and global clones. Among these, we identified five new sequence types lacking reference genomes in public databases and unusual clones associated with livestock or typical of the Middle East, supporting the changing epidemiology of *S. aureus* in the clinics. Through the screening of the most relevant resistance and virulence genes, we exposed the variability and complexity of these pathogenicity factors in the different departments, and we reported an increased prevalence of highly-resistant and lowly-virulent clones in chronic infections and the opposite pattern of lowly-resistant and highly-virulent variants in acute infections. An high degree of variability was observed also for the Staphylococcal Cassette Chromosome *mec* responsible for resistance to beta-lactams, for which we described two novel variants carrying extra antibiotic resistance genes. Given the high genetic variability observed, we expanded our analysis to survey the dispensability of antigens previously clinically tested for vaccine development, and highlighted their uneven conservation that may play a critical role in the difficult development of full-coverage vaccines for *S. aureus* infections. Lastly, we reconstructed the timed phylogeny of a specific clone to exclude a potential hospital-specific outbreak. Results suggested that the one we identified was not a hospital-specific clone, but a newly emerging infective variant spreading in the nosocomial environment of different European countries.

**Outlook.** Overall, our results highlight the under-investigated complexity of *S. aureus* epidemiology and advocate the need for wider genome-based analysis. The application of the approach presented in this study to include other hospitals and countries would help the unbiased identification of newly arising infective clones to reassess the efforts for the development of new therapies and update the clinical practice. An analysis of the current costs for bacterial whole-genome sequencing and the level of standardization that some of the analytical tool can reach, let us also conclude that NGS-based surveillance of *S. aureus* will be soon ready to be applied routinely in a clinical setting.

**Contribution.** For the work reported in this chapter, I coordinated the pre-sequencing sample processing, performed some of the post-sequencing computational steps (e.g. gene-based profiling and annotation), carried out the statistical association analysis, and I

led the data interpretation and writing of the manuscript. Sample and clinical data collection and *S. aureus* isolation was performed by the personnel at the Meyer hospital.

This chapter reports the following article:

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

Serena Manara^, Edoardo Pasolli^, Daniela Dolce^, Novella Ravenni, Silvia Campana, Federica Armanini, Francesco Asnicar, Alessio Mengoni, Luisa Galli, Carlotta Montagnani, Elisabetta Venturini, Omar Rota-Stabelli, Guido Grandi, Giovanni Taccetti and Nicola Segata

^ these authors contributed equally

*Genome Medicine* 2018

## 2.2 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 characterization, 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 three years, including both methicillin-resistant (MRSA) and methicillin-sensitive (MSSA) *S. aureus* from different infection types. We typed known and unknown clones from their genomes by Multilocus sequence typing (MLST), Staphylococcal Cassette Chromosome *mec* (SCC*mec*), 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 SCC*mec* 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-SCC*mec*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.

## Keywords

*Staphylococcus aureus*, microbial genomics, microbial epidemiology, bacterial pathogens

## 2.3 Background

*Staphylococcus aureus* is a bacterium commonly found on the skin (15%), in the nostrils (27%), and in the pharynx (10-20%) of healthy adults (Wertheim et al. 2005; Mainous et al. 2006; Tong, Davis, et al. 2015), but it is also the cause of a number of diseases, whose severity ranges from common community-associated skin infections to fatal bacteraemia (Tong, Davis, et al. 2015; Esposito, Noviello, and Leone 2016; Rhee et al. 2015). *S. aureus* is a leading cause of surgical, device-related, and pleuropulmonary infections, which can result into life-threatening infective endocarditis or even sepsis (Peacock and Paterson 2015). The mortality of *S. aureus* invasive infections was extremely high (>80%) in the pre-antibiotic era (Peacock and Paterson 2015; Skinner and Keefer 1941), and only the introduction of penicillin at the beginning of the 1940s was able to contain it. However, resistant strains carrying a penicillinase/beta-lactamase quickly emerged (Rammelkamp and Maxon 1942; Kirby 1944; Bondi and Dietz 1945), and more than 90% of current human-associated isolates are resistant to penicillin (Peacock and Paterson 2015). Similarly, the introduction of the penicillinase-resistant antibiotic methicillin was quickly followed by the emergence of methicillin-resistant (MRSA) clones (Barber 1961; Parker and Jevons 1964; Jevons, Coe, and Parker 1963). *S. aureus* is capable of acquiring resistance to virtually every antibiotic that has entered clinical use (Pantosti, Sanchini, and Monaco 2007; Monaco et al. 2017), including recently developed agents like daptomycin and linezolid (Raad et al. 2007; Liu et al. 2011) and the last resort antibiotic vancomycin (Kos et al. 2012; Hiramatsu 2001). In 2017 the World Health Organization has listed

vancomycin-intermediate and resistant MRSA among the high priority pathogens for research and development of new antibiotics (Tacconelli et al. 2018).

*S. aureus* ability to spread worldwide and to cause outbreaks in both hospitals and the community (Tosas Auguet et al. 2018; Coll et al. 2017) has fostered the study of its global epidemiology (Tong, Davis, et al. 2015; Monaco et al. 2017; Voss and Doebbeling 1995; Stefani et al. 2012; Chambers and Deleo 2009). Some lineages are very prevalent worldwide (e.g. CC5 and CC8) (Stefani et al. 2012), whereas others have a more localized spreading range, like the CC5 ST612 clone, which has been found only in South Africa and Australia (Stefani et al. 2012; Chambers and Deleo 2009; Jansen van Rensburg et al. 2011). MRSA prevalence is also highly geographically variable, ranging from <1% in some Northern European countries to >50% in some American and Asian countries, with livestock-associated MRSA disseminating in the last two decades (Stefani et al. 2012). Newly emerging highly pathogenic and pandemic clones have also been globally characterized (Harris et al. 2010; Mediavilla et al. 2012) and are often the results of recombination events as in the case of the ST239-SCC*mec*III clone (Chambers and Deleo 2009; Harris et al. 2010; Deurenberg and Stobberingh 2008). *S. aureus* investigations have however often underestimated the importance of non-MRSA clones, usually considering only hypervirulent or specifically relevant methicillin-sensitive (MSSA) lineages (Monaco et al. 2017), even though MSSA is the most common cause of surgical site infection (Sievert et al. 2013) and one of the major nosocomial pathogens (Monaco et al. 2017).

Untargeted profiling of the entire *S. aureus* population in a given site or area is as important as its global epidemiology and it is crucial for surveillance and prevention of local outbreaks. Some studies have for instance unbiasedly assessed the local epidemiology of nosocomial *S. aureus*, suggesting that this pathogen is only rarely transmitted from nurses to hospitalised patients in presence of adequate infection prevention measures (Price et al. 2017), and that the community acts as major source of nosocomial MRSA (Prosperi et al. 2013). Studies surveying the whole *S. aureus* population in hospitals have however focused on single aspects, like the diversity of the population, its virulence and resistance traits, and its transmission in presence of an outbreak (Harris et al. 2013; Stein et al. 2006; Bertin et al. 2006; Coombs et al. 2007; Wang et al. 2001; Saiman et al. 2003) or in non-emergency conditions (Givney et al. 1997; Blok et al. 2003; Tong, Holden, et al. 2015). Despite the large body of researches on *S. aureus*, studies addressing a whole *S. aureus* infective population at a given site through whole genome sequencing to simultaneously look at the epidemiology, phylogenetic reconstruction, genomic characterization, and transmission pathways of infective clones are currently limited (Copin, Shopsin, and Torres 2018). Expanding these types of studies will be crucial for an in-depth global monitoring of *S. aureus*.

Here we report an in-depth epidemiological and genomic investigation of *S. aureus* infections in a paediatric hospital in Italy. With a whole-genome sequencing approach, we

reconstructed the phylogenies of the clones in the cohort, characterized known clones and variants, screened for resistance and virulence genes, and tested for the presence of an outbreak. This allowed us to appreciate the high diversity of the *S. aureus* community, with 80 different lineages, variability of the resistance cassettes, and uneven conservation of various antigens previously clinically tested for vaccine development. We further report an increased prevalence of highly resistant and lowly virulent clones in chronic infections, and the rise of a newly emerging clone already reported in other hospitals. Overall, our results highlight the complexity of *S. aureus* epidemiology and advocate the need for wider genome-based analysis.

## 2.4 Materials and Methods

### Sample collection and *S. aureus* isolation

Samples were collected at Anna Meyer Children's University Hospital (Florence, Italy) from 160 patients from January 2013 to December 2015. Metadata were also collected (**Additional file 1: Table S1**). We analysed samples obtained from the most common sites of infection for *S. aureus*, namely airways (bronchial aspirates, sputum or oropharyngeal and nasal swabs) or from soft-tissue and skin lesions. All samples were processed for the detection of bacteria using selective (Mannitol Salt Agar 2, bioMérieux) and chromogenic culture media for MRSA (BBL™ CHROMagar™ MRSA II, Becton Dickinson). In order to confirm species-level identification, mass spectrometry analysis was performed using Matrix Assisted Laser Desorption/Ionization Time of Flight (MALDI-TOF) (VITEK® MS, bioMérieux). Antibiotic susceptibility was evaluated using the automated system VITEK®2 (bioMerieux) with the card AST- P632 (see **Additional file 1: Table S1** for antibiograms). All identified strains were stored at -80 ° C for the following molecular analysis.

### Molecular characterization of *S. aureus* and MRSA isolates

DNA extraction was performed from pure *S. aureus* cultures after 24 hours of incubation at 37 ° C on Columbia agar + 5% sheep blood (bioMérieux) using QIAamp DNA Mini Kit (cat. num. 51306, QIAGEN, Netherlands) according to manufacturer's specifications. DNA was purified using Agencourt AMPure XP (Beckman Coulter, California, USA) according to manufacturer's specifications. Extracted DNA was stored at - 20° C for further analyses.

In order to determine the potential virulence of SA/MRSA strains, a specific PCR assay for the presence of the gene (*lukS-lukF*) encoding for the Panton-Valentine Leukocidin (PVL), was set up following a previously published protocol (Lina et al. 1999). The *mecA* gene and other loci of the SCC*mec* cassette were analyzed using different multiplex PCR. The protocol suggested by Milheirico *et al* (Milheirico, Oliveira, and de Lencastre 2007) has been used as a screening test for most frequent SCC*mec* cassettes types (type I, II, III, IV, V and VI) and then confirmed with other methods in equivocal cases (Milheirico, Oliveira, and de Lencastre 2007; Oliveira and de Lencastre 2002; Oliveira, Milheiriço, and de Lencastre 2006; Milheiriço, Oliveira, and de Lencastre 2007).

PCR-based MLST typing was carried out with 25 µl reaction volumes containing 2 µl of chromosomal DNA, 20 µM of each primer, 1 U of Taq DNA polymerase (Super AB Taq, AB analitica), 2.5 µl of 10x PCR buffer (supplied with the Taq polymerase), 1.5 µM $MgCl_2$ and 250 µM each deoxynucleoside triphosphates. The PCR was performed with an initial 5-min denaturation at 95°C, followed by 30 cycles of annealing at 55°C for 1 min, extension at 72°C for 1 min, and denaturation at 95°C for 1 min, followed by a final extension step of 72°C for 5 min. The amplified products were purified and then amplified with the BigDye® Terminator v3.1 Cycle Sequencing Kit's (Applied Biosystem) with the primers used in the initial PCR amplification. The sequences of both strands were determined with an ABI Prism 310 DNA sequencer. Isolates with the same ST have identical sequences at all seven MLST loci.

### Isolates sequencing and data pre-processing

DNA libraries were prepared with Nextera XT DNA Library Preparation Kit (Illumina, California, USA). Quality control was performed with Caliper LabChip GX (Perkin Elmer) prior to shotgun sequencing with MiSeq (Illumina, California, USA), with an expected sequencing depth of 260 Mb/library (expected coverage >80X). 129M reads were generated (704K reads/sample s.d. 349K).

Sequences were pre-processed by removing low quality (mean quality lower than 25) or low complexity reads, reads mapping to human genome or to large and small ribosomal units of bacteria, fungi and human, and known contaminants (e.g. phiX174, Illumina spike-in). All genomes are available at the NCBI Sequence Read Archive (BioProject accession number PRJNA400143).

### Genomes Assembly and Annotation

Pre-processed reads were *de novo* assembled using SPAdes version 3.6.1 (Bankevich et al. 2012) and discarding contigs shorter than 1,000 nt. We selected for our analysis only reconstructed genomes with an N50 > 50,000. We obtained high-quality genomes (N50>50,000 and less than 250 contigs) for 135 of the 160 patients enrolled. Genomes belonging to the remaining 25 patients were excluded from further analyses. Genomes were annotated with Prokka version 1.11 (Seemann 2014) using default parameters and adding --addgenes and --usegenus options.

### Genomes Alignment / Phylogenetic analysis

The sets of 1,464 concatenated genes used as input for constructing whole cohort (**Fig. 1**) and strain (**Fig. 2**) phylogenetic trees were generated using Roary version 3.4.2 (Page et al. 2015). Maximum likelihood trees were inferred with RAxML version 8.0.26 (Stamatakis 2014) using a GTR replacement model with four discrete categories of Gamma. Support at nodes was estimated using 100 bootstrap pseudo-replicates (option "-f a"). The phylogenetic tree in **Additional file 2: Fig. S1** was inferred using the presence-absence binary matrix of the core and accessory genes computed with Roary version 3.4.2 (Page et al. 2015)] in RAxML version 8.0.26 (Stamatakis 2014) with option “-m BINGAMMA”.

Phylogenetic analyses were conducted using only one single isolate per patient; when multiple isolates from different timepoints of the same patient were available, the reconstructed genome with the highest N50 and the lowest number of contigs was selected. In most cases (n = 30) patients maintained the same ST over time; in discrepant cases (n = 2) we selected the most prevalent clone.

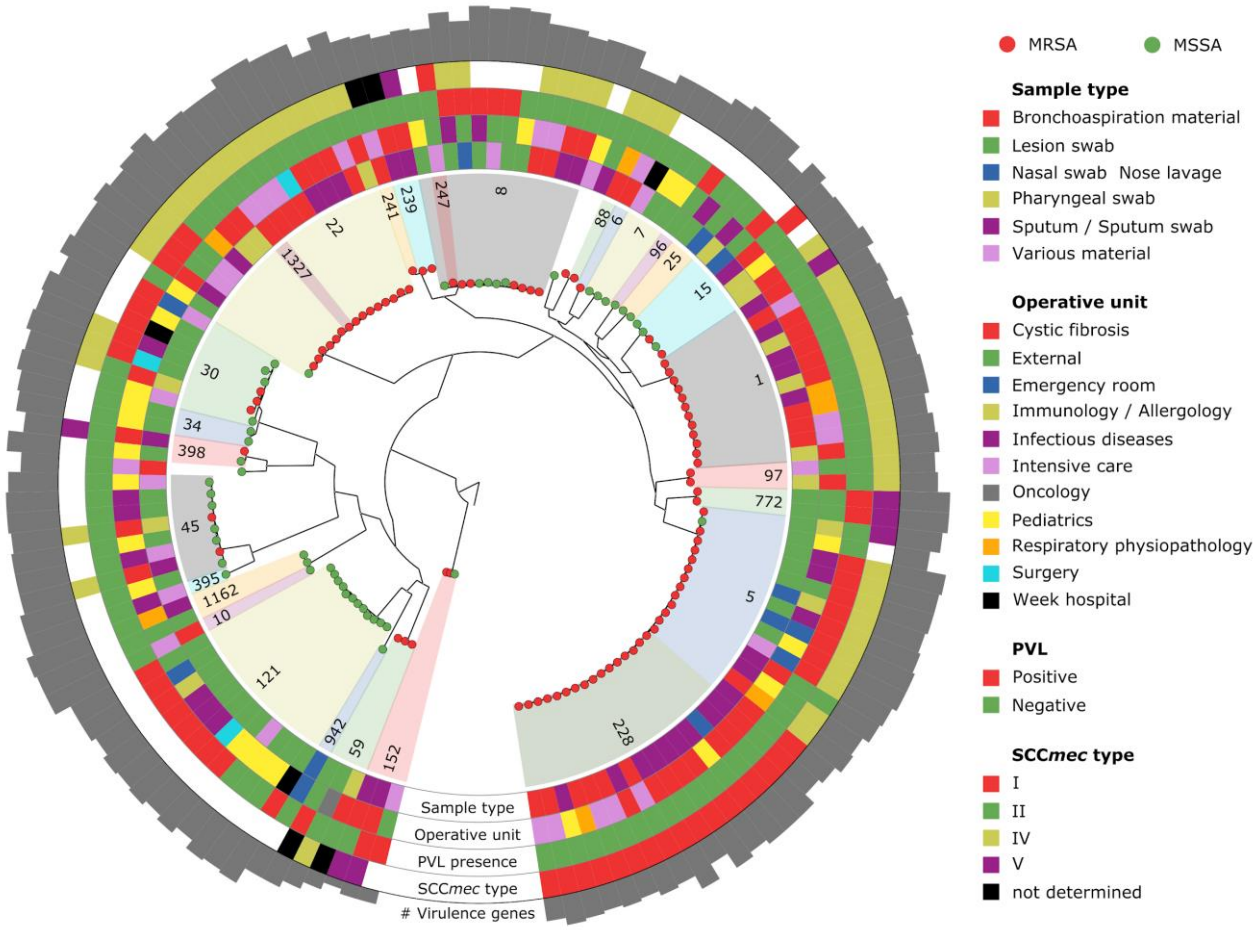



**Figure 1. Phylogenetic tree of the whole cohort.** Phylogenetic tree based on the 1,464 core genes (1,194,183 bases) of the 135 single-patient *S. aureus* isolates. STs are distinguished by means of numbers and background colours in the inner ring. Sample type, operative unit, PVL presence, and SCC*mec* type are colour-coded in the following rings. On the outermost ring, the number of virulence genes is reported as bar plot (total considered = 79).

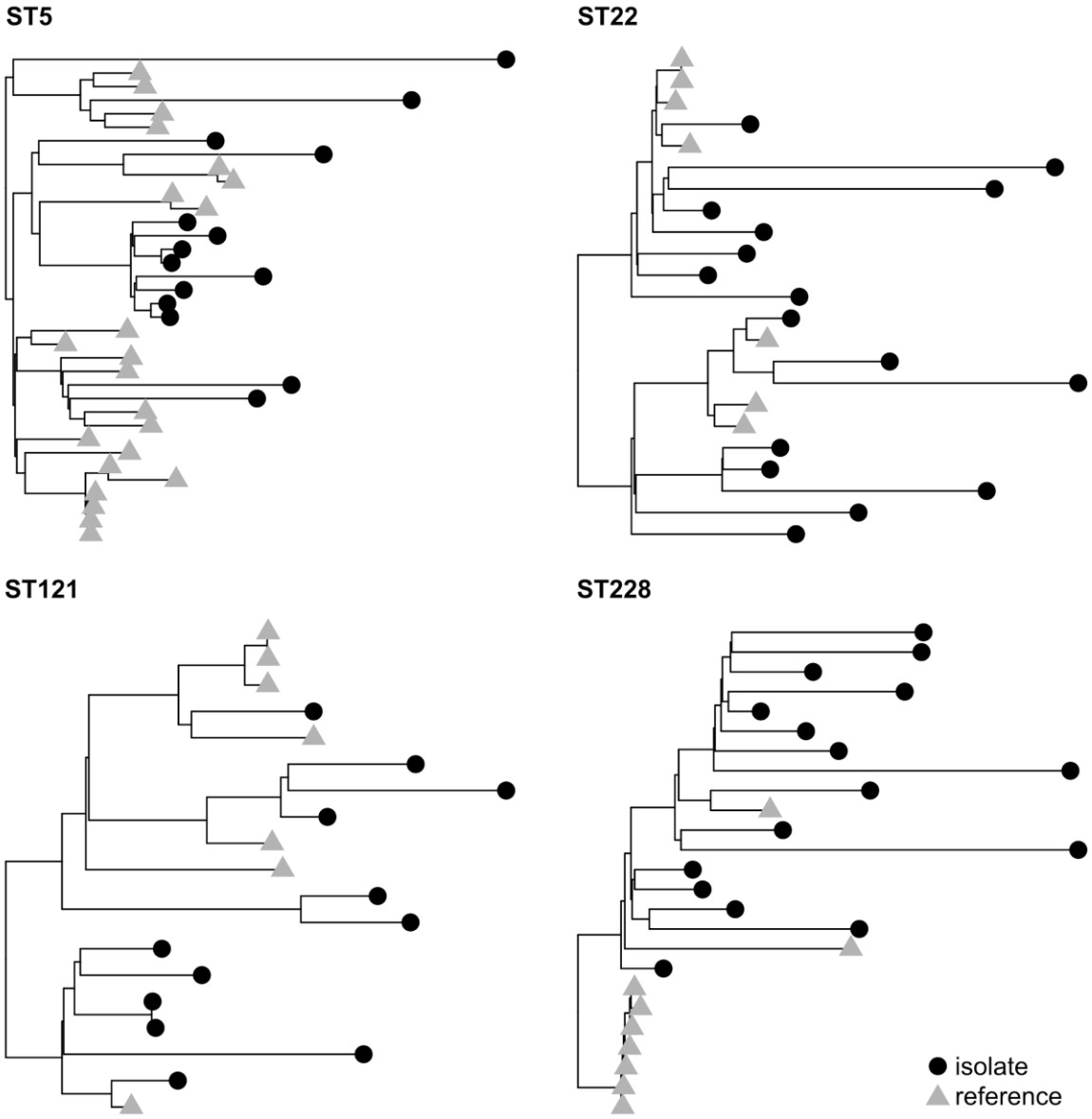



**Figure 2.** Whole-genome maximum likelihood phylogenetic trees of the four most relevant STs. All available reference genomes for ST22, ST121, and ST228 have been included. For ST5, 1,478 reference genomes were available, but only 24 were included for the sake of clarity. The phylogenetic tree of ST1 and available reference genomes was also produced, but it is not reported here to avoid overlapping with **Figure 5**.

### *In-silico* Sequence Type (ST), SCC*mec*, and *spa*-type identification

In order to assign SCC*mec* type also to equivocal cases and to confirm PCR-based SCC*mec* typing, the same set of primers (Milheirico, Oliveira, and de Lencastre 2007) and other primer sets (Zhang et al. 2005; Boye et al. 2007) were mapped to reconstructed genomes by BLAST (Altschul 1990). In most cases the two methods were consistent. In discordant cases, PCR was repeated. Sequence typing and *spa*-typing were conducted using MetaMLST (Zolfo et al. 2017) and the DNAGear software (AL-Tam et al. 2012)

respectively. Many isolates were not assigned a *spa*-type because of the limitations of short-read shotgun sequencing in repeated regions, which cause problems in genome assembly.

### Virulence factors and Resistance genes analysis

Selected virulence factors and resistance genes (as in (Gordon et al. 2014)) were searched for by mapping reference genes (**Additional file 3: Table S2**) to all reconstructed genomes with BLAST (Altschul 1990) with the following parameters [-evalue 1e-10 -perc_identity 90 -gapopen 5 -gapextend 5] with a match >75%. Virulence genes to be searched for were selected on the basis of a careful literature review for their clinical relevance (Inoshima et al. 2011; Dinges, Orwin, and Schlievert 2000; Smith et al. 2011; Haupt et al. 2008; Shannon and Flock 2004; Shannon, Uekötter, and Flock 2005; Courjon et al. 2015; Diep et al. 2008; Ellington et al. 2008; Kim et al. 1994; Miethke et al. 1993; Rooijakkers and van Strijp 2007; Thammavongsa et al. 2015; Udo, Boswihi, and Al-Sweih 2016; Al Laham et al. 2015; Biber et al. 2012; D. M. Geraci et al. 2014; Daniela M. Geraci et al. 2014; Amagai et al. 2000; Hanakawa et al. 2002; Ladhani 2003; Tung et al. 2000; Jongerius et al. 2012; Sanchez et al. 2013; Paharik and Horswill 2016; Schwab et al. 1993).

### Analysis of genes with available vaccine targets

Genes of interest were identified as those *S. aureus* vaccine candidates that had already entered clinical trials (according to http://clinicaltrials.gov as of January 2018), and those candidates that showed promising results in preclinical trials. For each genome, we extracted the reference sequences using BLAST (Altschul 1990) with default parameters. Extracted genes were pairwise globally aligned with the reference, and evaluated for synonymous and non-synonymous SNVs, insertions and/or deletions.

### Bayesian divergence estimates

We estimated divergence times of ST1 SCC*mec*IV t127 PVL- clones using Beast2 (Drummond et al. 2012) and the core genome (core genes = 1,464). We defined the best fitting model priors by testing the combination of three clock models (uncorrelated relaxed exponential; uncorrelated relaxed lognormal; strict), three demographic models (birth-death; coalescent Bayesian skyline; constant), and two substitution models (HKY - Hasegawa, Kishino, Yano; generalised time reversible). Bayesian Markov chain Monte Carlo was run for 500 Mio. generations and sampled every 1,000 generations. We chose the combination of models that resulted in the highest Bayes factor after parameter correction using AICM in Tracer (see **Additional file 4: Table S3**).

### Statistical tests

Associations between STs/virulence genes/antibiotic resistance markers and sample/operative unit types were found by performing a Fisher's exact test between the class of interest and the remaining set of samples.

## 2.5 Results and discussion

We investigated the epidemiology and the whole-genome genetics of *Staphylococcus aureus* isolated from multiple operative units of the same paediatric hospital in Italy (Meyer's Children Hospital, Florence). Two hundred thirty-four *S. aureus* isolates from 160 patients were retrieved from diverse clinical specimens, tested for antibiotic susceptibility, and subjected to whole-genome sequencing (see **Methods**). The study produced 184 high-quality reconstructed *S. aureus* genomes with a N50 larger than 50,000 and less than 250 contigs (**Additional file 1: Table S1**). Downstream analyses are focused on the 135 high-quality strains recovered from distinct patients.

### Genome sequencing highlights the presence of common clonal complexes and five newly sequenced clones

We first performed a whole-genome phylogenetic analysis to investigate the population structure of *S. aureus* in our cohort. The phylogeny was built using one isolate for each patient (n=135) and using the 1,464 core genes representing a core genome of >1.19M bases (see **Methods** and **Fig. 1**). The genomic diversity of *S. aureus* is highlighted by the relatively large number of accessory genes even in a limited cohort of clinical isolates (n=6,909 from a pangenome of 8,373, **Additional file 2: Fig. S2**), in concordance with a recent study based on the pangenome of 64 strains from different ecological niches (Bosi et al. 2016). The gene presence/absence phylogenetic model considering both core and genes confirmed the structure of the one built on the core genome alone, with however a slightly higher strain-diversity for isolates belonging to the same ST (**Additional file 2: Fig. S1**). Despite this diversity, we found the presence of a reduced set of closely related strains in the cohort (**Fig. 1**) mostly associated with distinct Multiple Locus Sequence Typing clones (STs) (Maiden et al. 1998) (see **Methods**). We identified a total of 29 different STs, with five of them - ST228, ST22, ST5, ST121, and ST1 – found in at least 12 patients (**Table 1** and **Additional file 1: Table S1**) with evidence of ST replacement in only one patient (Patient 091 switching from ST228 to ST22) of the 32 patients sampled at multiple timepoints. This longitudinal strain consistency was confirmed by whole-genome analysis (mean intra-patient variability = 56.42 SNVs), for which the replacing event in Patient 091 accounted for 6,238 SNVs between the 2013 and 2016 isolates, 0.22% of the genome. The 29 identified STs belong to 14 clonal complexes (CCs), with the five most prevalent CCs (CC5, CC22, CC8, CC1, and CC121) comprising more than 60% of the isolates. *Spa*-typing (AL-Tam et al. 2012) further refined the typing resolution: we found 44 distinct *spa*-types (**Additional file 1: Table S1**), with t001, t002, t008, and t127 being the most prevalent (i.e. present in >4 isolates, **Table 1**). We also investigated the presence of the Panton-Valentine Leukocidin (PVL), a two-component prophage virulence factor allowing *S. aureus* escape from the host immune system, that was found in 27.4% of the samples (**Additional file 1: Table S1**).

| ST | CC | # isolates (MRSA) | Predominant SCC*mec* type (# isolates) | Predominant *spa*-type (# isolates) | # PVL+ | Avg. genome length (bp) | Avg. # contigs | Avg. N50 | Avg. # CDS | Avg. # genes |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 12 (12) | IV (11) | t127 (3) | 0 | 2814074.3 | 29.4 | 326193.0 | 2601.3 | 2666.8 |
| 772 | | 2 (2) | V (2) | t657 (2) | 2 | 2768135.0 | 46.0 | 208282.5 | 2538.0 | 2605.0 |
| 5 | 5 | 14 (13) | IV (10) | t002 (5) | 8 | 2785946.1 | 39.4 | 250660.3 | 2580.1 | 2640.5 |
| 228 | | 16 (16) | I (16) | t001 (5) | 0 | 2837918.4 | 81.9 | 87688.4 | 2639.8 | 2700.7 |
| 6 | 6 | 1 (1) | IV (1) | t5238 (1) | 0 | 2796820.0 | 40.0 | 150271.0 | 2584.0 | 2648.0 |
| 7 | 7 | 3 (0) | n.a. | t1743 (1) | 0 | 2747478.7 | 66.0 | 147717.0 | 2521.3 | 2588.0 |
| 8 | 8 | 11 (6) | IV (6) | t008(6) | 5 | 2821267.1 | 52.7 | 259331.5 | 2625.2 | 2681.8 |
| 239 | | 2 (2) | V (1) | t037 (1) | 0 | 2900431.5 | 90.5 | 90036.5 | 2697.5 | 2762.0 |
| 241 | | 1 (1) | n.d. | t030 (1) | 0 | 2884707.0 | 87.0 | 105325.0 | 2707.0 | 2768.0 |
| 247 | | 1 (1) | I (1) | t197 (1) | 0 | 2776359.0 | 76.0 | 74230.0 | 2567.0 | 2630.0 |
| 10 | 10 | 1 (0) | n.a. | n.a. | 0 | 2799287.0 | 110.0 | 52819.0 | 2634.0 | 2698.0 |
| 1162 | | 2 (0) | n.a. | n.a. | 0 | 2867105.0 | 58.0 | 184821.5 | 2702.0 | 2767.0 |
| 15 | 15 | 5 (2) | IV (1); I (1) | t084 (1); t853 (1) | 1 | 2719481.8 | 45.6 | 226394.0 | 2496.2 | 2556.6 |
| 22 | 22 | 15 (14) | IV (13) | t852 (1); t1977 (1); t223 (1); t005 (1) | 3 | 2793443.3 | 56.0 | 124918.3 | 2599.5 | 2662.4 |
| 1327 | | 1 (1) | IV (1) | n.a. | 0 | 2758892.0 | 42.0 | 164290.0 | 2547.0 | 2612.0 |
| 25 | 25 | 2 (0) | n.a. | t258 (1); t2242 (1) | 0 | 2758786.5 | 16.5 | 697459.0 | 2554.5 | 2617.5 |
| 30 | 30 | 7 (3) | IV (3) | t019 (2) | 5 | 2792108.9 | 58.1 | 139913.6 | 2603.3 | 2666.1 |
| 34 | | 2 (0) | n.a. | t3905 (1) | 0 | 2821562.0 | 57.5 | 140057.0 | 2665.5 | 2730.5 |
| 45 | 45 | 8 (2) | IV (2) | t015 (2) | 0 | 2762203.4 | 34.4 | 390455.1 | 2591.5 | 2654.6 |
| 59 | 59 | 3 (3) | IV (1) | t216 (1); t437 (1) | 1 | 2799567.0 | 53.0 | 130817.3 | 2595.7 | 2662.0 |
| 88 | 88 | 1 (1) | IV (1) | t4701 (1) | 0 | 2791324.0 | 36.0 | 206283.0 | 2575.0 | 2642.0 |
| 96 | 96 | 1 (0) | n.a. | n.a. | 1 | 2783146.0 | 39.0 | 141877.0 | 2591.0 | 2652.0 |
| 97 | 97 | 2 (2) | IV (2) | t359 (1) | 0 | 2756222.0 | 27.0 | 401302.0 | 2570.0 | 2635.5 |
| 121 | 121 | 12 (0) | n.a. | t3274 (1); t314 (1); t2530 (1) | 9 | 2814764.5 | 48.0 | 146041.3 | 2631.3 | 2694.3 |
| 152 | 152 | 3 (2) | V (2) | t355 (1) | 2 | 2753826.7 | 31.0 | 267208.7 | 2551.0 | 2608.0 |
| 395 | 395 | 1 (0) | n.a. | n.a. | 0 | 2759659.0 | 22.0 | 574357.0 | 2574.0 | 2640.0 |
| 398 | 398 | 2 (1) | V (1) | t011 (1) | 0 | 2754012.0 | 57.5 | 206459.0 | 2524.0 | 2589.5 |
| 942 | 942 | 1 (0) | n.a. | n.a. | 0 | 2813978.0 | 82.0 | 61174.0 | 2654.0 | 2718.0 |
| - | n.a. | 3 (1) | IV (1) | n.a. | 0 | 2739763.7 | 46.0 | 157754.0 | 2535.0 | 2599.3 |

**Table 1.** Genomic characteristics of the different STs, including SCC*mec* and *spa*-type, presence of PVL, genome length, N50 (shortest sequence length at 50% of the genome), and number of contigs, coding DNA sequences (CDS), and genes. The combination of the four methods (MLST, SCC*mec*-, and *spa*-typing, and PVL presence) yielded 80 different lineages. Three isolates were not assigned to any specific ST and are reported in the last row of the table.

According to both antibiotic susceptibility testing (oxacillin and cefoxitin susceptibility, **Additional file 1: Table S1**) and genome analysis (presence of the SCC*mec* cassette, see **Methods**), 63.7% of the isolates were classified as methicillin-resistant *S. aureu*s (MRSA). Most strains (n=54) belonged to SCC*mec*IV; type I cassettes were also abundant (n=19), whereas cassettes type V (n=8) and II (n=1) were less represented. Methicillin-resistance was unevenly distributed across the phylogenetic tree (**Fig. 1**) and partially independent from the STs. All CC1 isolates (n=14, ST1 and ST772) were MRSA, and so were the isolates belonging to CC5 (n=30, ST5 and ST228) and CC22 (n=16, ST22 and ST1327). All CC121 (n=12, ST121) and CC10 (n=3, ST10 and ST1162) isolates were instead methicillin-sensitive (MSSA), and other clonal complexes (CC8, CC30, CC45) showed balanced proportions of sensitive and resistant strains. SCC*mec*I (n=19) was the most CC-specific cassette, as it was found almost exclusively in CC5 isolates (ST5 and ST228), with the exception of one ST15 and one ST8 isolates, while neither SCC*mec*IV nor SCC*mec*V were associated with specific STs.

For five of the recovered STs, namely ST241, ST942, ST1162, ST1327, and ST1866, no sequenced genome is publicly available (as genomes of *S. aureus* in RefSeq (Pruitt, Tatusova, and Maglott 2007) version 2017 (Haft et al. 2018). Although a large number of *S. aureus* genome sequences are available in NCBI, these are biased toward a limited set of clinically relevant STs (Planet et al. 2017; Copin, Shopsin, and Torres 2018), with many others being neglected. This underrepresentation of less pathogenic or less known strains may lead to a poor understanding of the host–pathogen interactions at the genomic level, and to an underestimation of emerging or re-emerging pathogenic strains (Planet et al. 2017; Chambers and Deleo 2009).

### Co-presence of local, global, animal-associated, and hypervirulent clones

We combined the four characterization methods (MLST, SCC*mec*-, *spa*-, and PVL-typing) to identify specific known clones in the cohort, yielding 80 different lineages. The most prevalent were the South German/Italian ST228-SCC*mec*I clone (n=16, 11.85%) and the E-MRSA-15 ST22-SCC*mec*IV clone (n=13, 9.63%), followed by the USA400 ST1-SCC*mec*IV t127 (n=11, 8.15%) clone, the USA800 paediatric clone ST5-SCC*mec*IV t002 (n=10, 7.41%), and the USA500 E-MRSA-2/6 clone ST8-SCC*mec*IV t008 PVL- (n=4, 2.96%) (**Additional file 5: Table S4**). Several other clones, including the highly virulent USA300 ST8-SCC*mec*IV PVL+ clone (n=2, 1.48%), were also found, confirming a heterogeneous clone composition in Italian hospitals (Mato et al. 2004; Floriana Campanile et al. 2015). Surprisingly, we did not isolate any ST80, the most prevalent community-associated MRSA clone in Europe (Vandenesch et al. 2003).

We moreover identified two isolates (1.48%) belonging to the livestock-associated MRSA (LA-MRSA) ST398 clone (Stefani et al. 2012; Witte et al. 2007) (**Table 1**). This clone has already been reported in patients that had regular exposure to livestock in several countries (Stefani et al. 2012; van Cleef, Graveland, et al. 2011; Cuny, Köck, and Witte 2013) including Italy(Soavi et al. 2010; Pan et al. 2009; Monaco et al. 2013), but our

results and other reports(Mammina et al. 2010; Cuny, Köck, and Witte 2013; van Cleef, Monnet, et al. 2011; Verkade et al. 2012) of infections in non-exposed subjects suggest that the within-subject transmission for these clones is not rare. Similar conclusions can be drawn for another LA-MRSA, namely ST97 (n=2, 1.48%, **Table 1**), that is the leading causes of bovine mastitis, but is only rarely reported in humans (Sung, Lloyd, and Lindsay 2008; Spoor et al. 2013; Cuny, Wieler, and Witte 2015; Feltrin et al. 2016). This growing incidence of LA-MRSA strains (n=4, 2.96% in our cohort) causing zoonotic infections highlights the existence of underestimated reservoirs of *S. aureus* strains that could become epidemic (Mediavilla et al. 2012; Fitzgerald 2012; Harrison et al. 2017).

One isolate was assigned to ST395, which is an unusual strain unable to exchange DNA via bacteriophages with other *S. aureus* strains because of a modification in the wall teichoic acid (WTA) (Winstel et al. 2013; Larsen et al. 2017). The same modification, however, enables ST395 to exchange DNA with coagulase-negative Staphylococci (CoNS) (Larsen et al. 2017), making it particularly prone to exchange SCC*mec* elements and others with other commonly found staphylococci, e.g. *S. epidermidis*.

### Genomic signatures of chronic versus acute *S. aureus* infections

In order to investigate the potential association of clones and antibiotic resistance with specific hospital operative units, we cross-checked the prevalence of SCC*mec* types, STs and PVL+ clones with both OUs and sample types (see **Methods**). Strains from the cystic fibrosis (CF, n = 76) unit were positively associated with the presence of SCC*mec*I (n = 19, ten from CF unit; p-value = 0.03), a cassette known to be hospital-associated (Molina et al. 2008; Asghar 2014). Strains from the same unit were also associated with ST1 (n = 12, seven from CF unit; p-value = 0.04), whereas we noted a reduced prevalence of the PVL genes (n = 37, only two from CF unit; p-value = 0.0002) and of ST121 (n = 12, none from CF unit; p-value = 0.02). This reflects the relatively attenuated virulence which is a well-known phenomenon in long-term *S. aureus* infections (McAdam et al. 2011; Goerke and Wolz 2010; Kahl 2010; Cullen and McClean 2015). Similarly, sputum samples (n = 33; 88.7% from CF unit) was associated with ST228 (n = 16, nine from sputum; p-value = 0.004) and SCC*mec*I (n = 19, 11 from sputum; p-value = 0.0008), and negatively correlated with PVL (n = 37, only two from sputum; p-value = 0.001). The high correlation of ST228 with lung isolates and specifically with CF has already been observed in Spain (Molina et al. 2008). A similar pattern of increased resistance and lowered virulence has been observed for another sample type linked with long-term lung infections, namely broncho-aspiration material (n = 23; 78.2% from intensive care unit). Strains from this sample type were associated to SCC*mec*IV (n = 54, 14 from broncho-aspiration material; p-value = 0.008), and with PVL- (n = 98, 23 from broncho-aspiration material; p-value = 0.0005) and MRSA clones (n = 83, 21 from broncho-aspiration material; p-value = 0.002), highlighting once again the loss of virulence and the acquisition of resistance in long-term lung infections (McAdam et al. 2011; Goerke and Wolz 2010; Kahl 2010; Cullen and McClean 2015).

On the contrary, patients from both emergency room (n = 5) and the infectious diseases unit (n = 15) show an overrepresentation of PVL+ clones (n = 37, four from emergency room and nine from infectious diseases; p-values = 0.02 and 0.005, respectively) indicative of acute rather than chronic infections. Lesion swabs (n = 31) are strongly associated with MSSA (n = 49, 31 from lesion swabs; p-value = 3e-08). This sample type was also associated to the hypervirulent ST121 clone (Rao et al. 2015; Goering et al. 2008) (n = 12, 11 from lesion swabs; p-value = 2e-05) and to the presence of the PVL (n = 37, 14 from lesion swabs; p-value = 3e-07), suggesting that in our cohort skin and soft tissue infections (SSTIs) are predominantly caused by hypervirulent MSSA strains. Lesion swabs from children in care at the infectious diseases unit (n = 12, 80% of the samples from this operative unit) are also characterised by high prevalence of the virulent ST45 clone (Roberts 2014; Moore et al. 2010) (n = 8, three from lesion swabs; p-value = 0.04) that is known to be associated with SSTIs (M. Z. David et al. 2011; Tinelli et al. 2009; Baranovich et al. 2010; Wu et al. 2010). The expected (Rasigade et al. 2010) association between PVL (n = 37) and ST121 (n = 12, nine PVL+; p-value = 0.001) and ST30 (n = 7, five PVL+; p-value = 0.003) supports once again the observed increased virulence of these STs (Isobe et al. 2012; Fernandez et al. 2017; Rao et al. 2015; Goering et al. 2008), which is partially in conflict with the hypothesis of lesion colonization by commensal strains present in the skin microbiome (Byrd, Belkaid, and Segre 2018; Tett et al. 2017).

### Discovery of novel variants of SCC*mec*IV with kanamycin, trimethoprim, and bleomycin resistance

We next investigated the specific genetic variants of the four types of SCC*mec* cassettes identified and discussed above. This is relevant because the epidemiology of this genetic element is disentangled by that of the rest of the genome by virtue of its high horizontal mobility (Hiramatsu et al. 2001; Hanssen and Ericson Sollid 2006). Moreover, the SCC*mec* can host genes encoding not only for resistance to beta-lactams (Hartman and Tomasz 1981; Archer et al. 1994), but also for other antibiotic resistances or virulence factors (Hartman and Tomasz 1981).

More than a half of the MRSA isolates in our collection (n=86) carried SCC*mec*IV (62.8%). This cassette type has spread widely in the last decades, often substituting the previously more prevalent nosocomial SCC*mec* types I and II (F. Campanile et al. 2012; Stefani et al. 2012), and it is now common especially in European clinical isolates (Stefani et al. 2012; Floriana Campanile et al. 2015). Another cassette that has spread in recent years following a similar path is SCC*mec*V (F. Campanile et al. 2012; Valsesia et al. 2010), the third most prevalent cassette type in our cohort (10.5% of the MRSA isolates) after the more traditionally hospital-associated SCC*mec*I (Stefani et al. 2012; Asghar 2014) (22.1% of the MRSA isolates). We moreover isolated one MRSA carrying SCC*mec*II, which is widely diffused in the US but only rarely found in Italy/Europe (Chambers and Deleo 2009; Tenover and Goering 2009). Consistently, the SCC*mec*II isolate was recovered from Patient 115, which is consistent with the personal history of the patient. For two isolates, it

was not possible to classify the cassette neither with PCR nor with *in-silico* PCR using standard primers (Milheirico, Oliveira, and de Lencastre 2007).

By aligning reconstructed SCC*mec* with reference cassettes (see **Methods**), we observed a certain degree of variability inside the same cassette type, specifically in type IV (**Fig. 3**). Subtypes IVa, IVb, and IVc were identified, with some SCC*mec* elements showing insertions. Two cassettes in particular were not consistent with the already described subtypes: the SCC*mec* type IVc carried by MF062, which was enriched with genes for kanamycin (Pedersen, Benning, and Holden 1995) and bleomycin (Gennimata, Davies, and Tsiftsoglou 1996; Dortet et al. 2017) resistance, and the type IVa carried by MR090 that showed insertion of genes involved in resistance to trimethoprim (Burdeska et al. 1990; Rouch et al. 1989) (**Fig. 3**).


**Figure 3.** Overview of the SCC*mec*IV cassette variability in our cohort, compared with available reference cassettes for the recovered subtypes IVa, IVb, and IVc. Genes are marked as arrows in the direction of transcription. To avoid biases due to misassemble of the region of interest, only cassettes found on a single contig are reported. Annotated SCC*mec* are grouped together with the closest reference cassette subtype. Some genomes showed insertions of genes involved in resistance to trimethoprim (MR090) and to kanamycin and bleomycin (MF062).

## Non-SCC*mec* resistance profiles show different patterns in chronic and acute infections

*S. aureus* can easily acquire a number of resistances, including those to the last resort antibiotics vancomycin (Whitener et al. 2004; Tenover et al. 2004) and daptomycin (Stefani et al. 2015). According to results presented in previous paragraphs and elsewhere (Noto et al. 2008), resistances can occur by gene acquisition in the SCC*mec* cassette. Most resistances are however encoded by genes that are found in other parts of the genome or that have been horizontally transferred through different genetic elements (Chambers and Deleo 2009). Given the high importance of multi-drug resistance in *S. aureus (Tacconelli et al. 2018)*, we therefore tested the presence or absence of specific resistance genes in our cohort (Wright 2007) (**Fig. 4** and **Additional file 3: Table S2**). Consistently with previous literature (Peacock and Paterson 2015), most of the isolates tested positive for *blaZ* (81.5%), responsible for penicillin resistance (96.3% concordance with antibiotic susceptibility test, as per presence of the *pbp* and/or *mecA* genes). No isolates were found positive for genes encoding resistance to vancomycin (*van*, 100% concordance with antibiotic susceptibility test) and to fusidic acid (*fusB* and *far*, 94.1% concordance with antibiotic susceptibility test). Antibiotic resistances were sometimes associated with specific CCs, as for the increased representation of *aacA.aphD* (gentamicin resistance, 92.6% concordance with antibiotic susceptibility test) and *ermA* (erythromycin resistance, phenotypic resistance not tested) in CC5 isolates, whose genomes tended to lack instead the *blaZ* gene (penicillin resistance) (**Fig. 4**). Overall, two isolates from acute skin infections were negative for all the resistance genes tested, while six CF and intensive care unit isolates were positive for six (33.3%) of them. This pattern of increased resistance in long-term infections, together with their observed reduced virulence, completes the scenario of reduced virulence and increased resistance that has been observed in this and previous studies (McAdam et al. 2011; Goerke and Wolz 2010; Kahl 2010; Cullen and McClean 2015).

## Emergence and disease-associated diversity of clinically relevant virulence factors

*S. aureus* has a large repertoire of virulence genes, and it is able to evade the host immune system through a variety of strategies. Some of the genes usually involved in immune evasion were present in almost all our isolates (**Fig. 4** and **Additional file 3: Table S2**). These include genes encoding: the phenol-soluble modulin alpha and beta and the delta-haemolysin Hld, responsible for leukocytes and erythrocytes lysis respectively (Dinges, Orwin, and Schlievert 2000); the immunoglobulin-binding protein Sbi that inhibits IgG and IgA (Smith et al. 2011; Haupt et al. 2008); and some genes part of the GIα genomic island (*ssl6* and *ssl9*).

Other genes belonging to the immune evasion island IEC2 were present in many but not all isolates, for example the one encoding for the antiplatelet extracellular fibrinogen binding protein Efb (Shannon and Flock 2004; Shannon, Uekötter, and Flock 2005) and those encoding various haemolysins (*hla*, *hlg*) (Inoshima et al. 2011; Dinges, Orwin, and Schlievert 2000) (**Fig. 4** and **Additional file 3: Table S2**). In addition to the 27.4% prevalence of the *lukF* and *lukS* PVL-genes discussed above, one sample (MR029, from

emergency room) was positive for the epidermal cell differentiation inhibitor Edin, which has been found to promote the translocation of *S. aureus* into the bloodstream (Courjon et al. 2015). One of the two USA300 isolates (MR047, from nasal swab) tested positive for the arginine catabolic mobile element (ACME), another important virulence factor (gene *arcA*) that has been shown to be responsible for the increased pathogenicity of *S. aureus* and specifically of USA300 clones (Diep et al. 2008; Ellington et al. 2008).

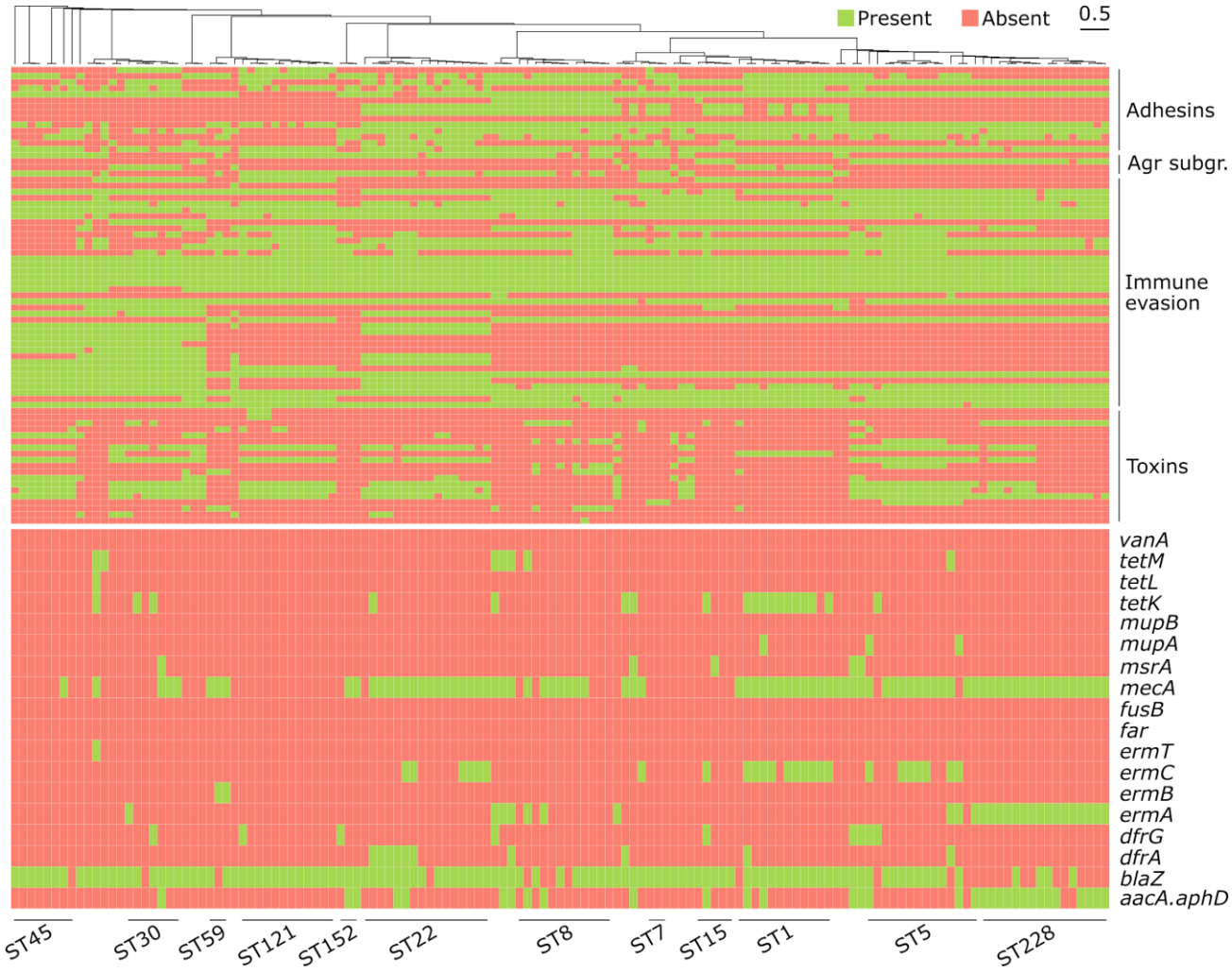



**Figure 4.** Presence/absence profile of 79 genes encoding for virulence factors (upper part of the heatmap) and 18 genes encoding for resistance (bottom part). Some virulence and resistance factors were more represented in specific STs (only STs found in >1 samples are specifically mentioned), as in the case of gentamicin resistance that is more prevalent in the ST228 isolates. For a more detailed overview of the single genomes' profiles, see **Additional file 3: Table S2**.

Many virulence genes were associated to specific STs (**Fig. 4** and **Additional file 3: Table S2**). ST22 (n = 15), for instance, was associated with the toxic shock syndrome toxin TSST-1 (n = 8, three from ST22; p-value = 0.04; present in 20% of the ST22 clones) (Dinges, Orwin, and Schlievert 2000; Kim et al. 1994; Miethke et al. 1993), other pyrogenic

toxin superantigens known as staphylococcal enterotoxins (SEs, mean n = 34.9 ± 28.1 s.d.; p-value < 0.02 for *seg*, *sei*, *sem*, *sen*, *seo*, present on average in 86.7% of ST22 and 49.5% of non-ST22), and various *ssl* immune evasion genes (mean n = 57.4 ± 40.2 s.d.; p-value < 0.01 for *ssl1*, *ssl3*, *ssl4*, *ssl7*, *ssl11*, *ssl12*, present on average in 93.3% of ST22 and 21.1 of non-ST22) (Rooijakkers and van Strijp 2007; Thammavongsa et al. 2015). ST22-IV EMRSA-15 clones positive for *tst1* are usually described as "Middle Eastern variant" (Udo, Boswihi, and Al-Sweih 2016; Al Laham et al. 2015; Biber et al. 2012), but a high prevalence in an Italian neonatal intensive care unit (Daniela M. Geraci et al. 2014) and pre-school children living in Palermo, Italy (D. M. Geraci et al. 2014), has been observed. Authors suggested that the Middle Eastern clone might be more widely spread than estimated and might have diffused in the Mediterranean populations as a community-acquired MRSA (D. M. Geraci et al. 2014; Daniela M. Geraci et al. 2014), as suggested by our analysis. TSST-1 is responsible for an increased pyrogenic, emetic, and superantigen activity, together with SEs (10627489;11544350). SEs (mean n = 34.9 ± 28.1 s.d.) were associated with all "virulent" STs, such as ST5 (n = 15, p-value < 0.04 for *sed*, *seg*, *sei*, *sej*, *sem*, *sen*, *seo*, *sep*, present on average in 79.2% of ST5 and 32.3% of non-ST5), ST45 (n = 8; p-value < 0.02 for *sec*, *seg*, *sei*, *sel*, *sem*, *seo*, present on average in 98.2% of ST45 and 38.9% of non-ST45), ST121 (n = 12; p-value < 0.02 for *seb*, *seg*, *sei*, *sem*, *sen*, *seo*, present on average in 86.1% of ST121 and 41.7% of non-ST121), and -to a lower extent- ST30 (n = 7; p-value < 0.02 for *sei*, *sem*, *sen*, present on average in 100% of ST30 and 50% of non-ST30).

The hypervirulent ST121 MSSA isolates obtained from lesion swabs (n = 12) were instead associated with the genes encoding for the exfoliative toxins Eta and Etb (n = 3 from ST121 swabs, and n = 0 for non-ST121, p-value = 0.0006 for both genes), responsible for the skin manifestations of bullous impetigo and Staphylococcal scalded skin syndrome (Amagai et al. 2000; Hanakawa et al. 2002; Ladhani 2003), the gene *bbp* (n = 12 from ST121, n = 7 from non-ST121; p-value = 1.35e-08) that interacts with the extracellular matrix bone sialoprotein and contributes to staphylococcal arthritis and osteomyelitis (Tung et al. 2000), and the immune evasion gene *ecb* (n = 12 from ST121, n = 36 from non-ST121; p-value = 1.51e06), which is required for the persistence of *S. aureus* in host tissues and the formation of abscesses (Jongerius et al. 2012). The latter was also present in all and only the isolates belonging to ST1, ST7, ST10, ST15, ST30, ST34, and ST398, suggesting a strong dependence on ST (**Fig. 4** and **Additional file 3: Table S2**).

Isolates retrieved from sputum samples of CF patients (n = 38) showed a positive association with the adhesin-encoding genes *sdrD* (n = 34 from CF, n = 69 from non-CF; p-value = 0.03) and *sdrE* (n = 27 from CF, n = 48 from non-CF; p-value = 0.03), and a negative association with *bbp* (n = 1 from CF, n = 18 from non-CF; p-value = 0.01), contrary to samples from infectious diseases unit (n = 15, four positive for *bbp* gene). This finding is consistent with the increased need for adhesins in chronic lung infections (Sanchez et al. 2013; Paharik and Horswill 2016; Cullen and McClean 2015), including in CF (Schwab et al. 1993).

## Conservation of genes encoding vaccine candidates

Unlike other bacterial infections, prior exposure to *S. aureus* does not seem to provide protective immunity (Giersing et al. 2016), therefore vaccines are an attractive yet challenging option to prevent disease. Researchers have long attempted to produce an effective vaccine against *S. aureus*, but even though few have proved promising in animal models, the two vaccines so far tested in efficacy clinical trials have failed (Giersing et al. 2016; Verkaik, van Wamel, and van Belkum 2011; Salgado-Pabón and Schlievert 2014; A. I. Fattom et al. 2004). Since the main issue is the polymorphic expression of *S. aureus* surface antigens and the redundancy of its virulence proteins (Giersing et al. 2016; Golubchik et al. 2013; Dreisbach et al. 2011), we tested the prevalence and conservation of a number of genes encoding vaccine candidates described in the literature (**Table 2**).

Among antigens that have been proposed as targets for vaccine development, the alpha haemolysin toxin gene *hla* (Giersing et al. 2016; Hua et al. 2015; Bagnoli 2017) and the genes coding for capsular biosynthesis *cap5* and *cap8 (A. I. Fattom et al. 2004; A. Fattom et al. 2015)* are highly prevalent in our cohort (91.9% and 97.8 % of the isolates respectively). Nevertheless, these genes showed a larger degree of variability compared to the others we considered, which may explain the poor results obtained in clinical trials (Giersing et al. 2016; Hua et al. 2015; Bagnoli 2017; A. I. Fattom et al. 2004; A. Fattom et al. 2015). Other genes that code for proteins used alone or in combination in vaccine formulations, such as the virulence determinant SpA (Yang et al. 2018) and the fibronectin binding protein ClfA (Frenck et al. 2017; Begier et al. 2017; Anderson et al. 2012), are present in most of our strain collection. In some of these genes indels are prevalent (>90%, **Table 2**), but they are frequently found in repeated regions that may not critically impact the protein structure, as in the case of the *spa* gene.

Vaccines have also been proposed for *S. aureus* strains with specific characteristics. For instance, targeting the toxicity determinant TSST-1 (5.9% prevalence of *tst1*) (Roetzer, Jilma, and Eibl 2017; Narita et al. 2015) or the PVL proteins LukF-LukS (27.4% prevalence of *lukF-lukS*) (Rouha et al. 2015; Badarau et al. 2016) aims at selectively preventing the most virulent or lethal infections. In our cohort, despite their low prevalence, both *tst1* and the PVL genes were conserved at 99%, except for a few isolates that had indels in the latter (**Table 2**). The gamma-haemolysins HlgAB and HlgCB genes (Rouha et al. 2015; Badarau et al. 2016) were instead highly prevalent (97.8-100%) and quite conserved (69.6-94.8%). The opposite approach is targeting genes with a lower virulence profile, which may be more prevalent and conserved than those coding for highly toxic factors. Among them, the genes encoding for the manganese uptake receptor (*mntC*) (Frenck et al. 2017; Begier et al. 2017; Anderson et al. 2012) and for the iron acquisition factor (*isdB*) (Fowler et al. 2013; Moustafa et al. 2012), which are indeed present in all or all but one the isolates of our cohort. Non-synonymous mutations are rare in *mntC* (20.7% of the isolates, with only one non-synonymous SNV), and, whenever not affected by indels that may or may not affect the protein structure, also the *isdB* gene is highly conserved (>99% identity, **Table 2**).

| gene | # positive isolates (%) | Distribution of non-syn SNVs w.r.t. reference seq. | | | | | | Latest trials | ClinicalTrials identifier | Reference |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | <1% | <2% | <5% | >=5% | indels | | | |
| *clfA* | 95 (70.4%) | 0% | 9.5% | 0% | 0% | 0% | 90.5% | Phase I-II | NCT01643941 NCT01364571 | (Frenck et al. 2017; Begier et al. 2017; Creech et al. 2017) |
| *csa1a* | 70 (51.9%) | 41.4% | 4.3% | 0% | 0% | 0% | 54.3% | Preclinical | | (Bagnoli et al. 2015; Torre et al. 2015) |
| *csa1b* | 36 (26.7%) | 19.4% | 47.2% | 0% | 2.8% | 0% | 30.6% | | | |
| *esxA* | 134 (99.3%) | 85.1% | 14.9% | 0% | 0% | 0% | 0% | Preclinical | | (Bagnoli et al. 2015; Torre et al. 2015) |
| *esxB* | 89 (65.9%) | 0% | 98.9% | 1.1% | 0% | 0% | 0% | Preclinical | | (Bagnoli et al. 2015; Torre et al. 2015) |
| *esxC* | 89 (65.9%) | 25.8% | 40.4% | 33.7% | 0% | 0% | 0% | | | |
| *esxD* | 89 (65.9%) | 58.4% | 41.6% | 0% | 0% | 0% | 0% | | | |
| *fhuD2* | 135 (100%) | 31.1% | 68.9% | 0% | 0% | 0% | 0% | Preclinical | | (Bagnoli et al. 2015; Torre et al. 2015) |
| *hla* | 124 (91.9%) | 6.5% | 0% | 9.7% | 78.2% | 2.4% | 3.2% | Phase II | NCT02296320 | (Bagnoli et al. 2015; Torre et al. 2015) |
| *hlgA* | 135 (100%) | 65.2% | 29.6% | 2.2% | 0% | 0% | 3% | | | |
| *hlgB* | 132 (97.8%) | 11.4% | 65.2% | 22.7% | 0% | 0% | 0.8% | Preclinical | | (Delfani et al. 2016) |
| *hlgC* | 135 (100%) | 61.5% | 8.1% | 28.1% | 2.2% | 0% | 0% | | | |
| *isdB* | 134 (99.3%) | 11.9% | 21.6% | 0% | 0% | 0% | 66.4% | Phase III | NCT00518687 | (Moustafa et al. 2012; Fowler et al. 2013) |
| *lukF* | 37 (27.4%) | 0% | 97.3% | 0% | 0% | 0% | 2.7% | | | |
| *lukS* | 37 (27.4%) | 56.8% | 40.5% | 0% | 0% | 0% | 2.7% | Phase I-II | NCT01011335 | (Landrum et al. 2017) |
| *mntC* | 135 (100%) | 79.3% | 20.7% | 0% | 0% | 0% | 0% | Phase I-II | NCT01643941 NCT01364571 | (Frenck et al. 2017; Begier et al. 2017; Creech et al. 2017) |
| *tst* | 8 (5.9%) | 0% | 100% | 0% | 0% | 0% | 0% | Phase I | NCT02340338 | (Schwameis et al. 2016) |

**Table 2.** Sequence variability of genes of interest for vaccine development. Number (and relative abundance) of isolates positive for the gene, followed by the percentage of positive isolates carrying 0 or less than 1%, 2%, 5%, or more/equal to 5% of non-synonymous SNVs or insertions-deletions (Indels) with respect to reference gene. Both clinical trial IDs (ClinicalTrials.gov database identifiers, http://clinicaltrials.gov) and reference studies refer to the latest available trials.

Finally, we also analysed the conservation of *csa1A, csa1B, fhuD2 and esxA*, genes recently described as being promising vaccine candidates in preclinical studies (Schluepen et al. 2013; Bagnoli et al. 2015). The two genes encoding for the conserved antigen Csa (*csa1A* and *csa1B*) are present in 51.9% and 26.7% of the isolates respectively, and are conserved in only a fraction of the cases (**Table 2**). By contrast, the iron uptake gene *fhuD2* is present in all isolates, with a maximum of 1% non-synonymous variation in sequence (**Table 2**). Also the genes encoding for the ESAT-6-like secretion system (*esxA*, *esxB*, *esxC*, *esxD*) are well represented in the cohort, but only *esxA* is present in all but one isolates and has no non-synonymous mutations in 85.1% of the isolates (**Table 2**). Therefore, on the basis of their conservation, both FhuD2 and EsxA appear to be promising targets for vaccine formulations.

### Phylogenetics of specific STs highlights the aggressive spread of a novel independently acquired ST1 clone

We investigated the hypothesis that some of the prevalent STs could be hospital-associated clones. We estimated the ST phylogenies using a whole-genome maximum likelihood approach (see **Methods**). In most cases we observed that isolates in our cohort, despite sharing the same ST, SCC*mec*, and *spa* types, were not monophyletic subtrees when considering external reference genomes for the same STs. This is the case, for example, of the ST228 and ST5 clones (**Fig. 2**). This suggests independent acquisition of the clones and no evidence of transmission among the selected hospitalised patients, while person-to-person transmission from healthy carriers or non-selected patients cannot be ruled out (Tosas Auguet et al. 2018; Coll et al. 2017). Only two ST121 MSSA isolates were found to be almost identical and both were retrieved in the same time window from patients 096 and 098 (8 SNVs). For ST1, instead, all but two isolates belonged to the same sub-lineage, typed as SCC*mec*IV t127 PVL-.

We further estimated divergence times for all the 16 isolates belonging to the ST1 SCC*mec*IV t127 PVL- clone, including those obtained from earlier or later time points of the same patients. We used a Bayesian approach (Drummond et al. 2012) (see **Methods**) integrating all the reference genomes publicly available for ST1 and the two ST1 SCC*mec*V isolates from our cohort (**Additional file 4: Table S3**). These analyses were performed to test the hypothesis that all ST1 SCC*mec*IV t127 belong to a clone specific of Meyer's hospital. The relaxed exponential clock model with constant coalescent prior and GTR substitution model resulted the most appropriate model (**Additional file 6: Table S5**). This model estimated that the Meyer's clone has emerged approximately 6 to 28 years ago as a specific branch of the ST1 tree, which has been estimated to be 26-160 years old (**Fig. 5**). However, age of the Meyer's clone does not match with the time of emergence of the clone in the hospital. Moreover, an isolate obtained in a recent study investigating the spread of a ST1 SCC*mec*IV t127 clone in Irish hospitals (Earls et al. 2017) and carrying a virulence and resistance profile very close to the one of our cohort (differences in gene presence: 2/79 and 0/18 respectively) is phylogenetically rooted inside the Meyer's cluster (161 SNVs intra-cluster; 412 SNVs inter-cluster). These two findings suggest that ST1

SCC*mec*IV t127 is not specific of the Meyer Children's hospital but might represent a newly arising community clone that is now spreading in the nosocomial environment of different countries (Earls et al. 2017; M. D. David et al. 2006).

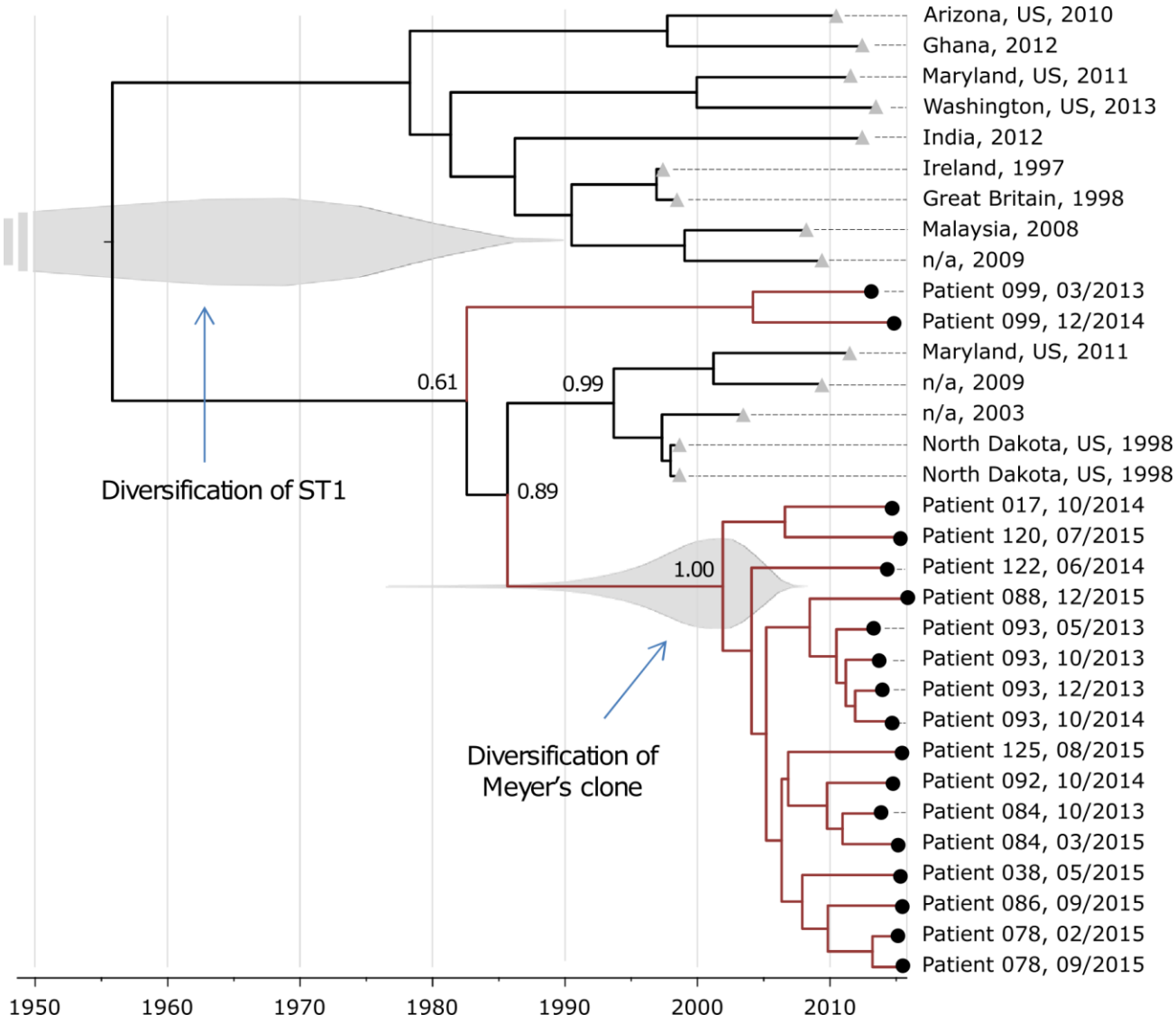



**Figure 5.** Bayesian timed tree of ST1 isolates, including reference genomes. Location and date of sample collection is reported for each isolate. For samples collected at Meyer 's Children Hospital (black circles), patient code is reported instead of location. The two North Dakota samples were collected from the same subject. "n/a" indicates that no information is available for location of sample collection. Numbers at selected nodes are posterior probabilities. Grey areas are the distributions of Marginal Posterior Probabilities for the diversification of ST1 and the diversification of Mayer-specific clone.

## 2.6 Conclusions

In this study we investigated the epidemiology of *S. aureus* in different operative units of Anne Meyer's Children's University Hospital (Florence, Italy) over a timespan of three years by whole genome isolate sequencing. Our analyses highlighted a high diversity of STs, SCC*mec*, and *spa*-types, resulting into a wide number of clones. Some of these clones had been previously described in the literature as livestock-associated, and we described them in non-exposed children thus supporting the spreading of such clones in the non-at-risk community. We moreover described the presence of hypervirulent and geographically-unusual clones, and of five STs for which no sequenced genome was available in public databases. Our refined analysis of the SCC*mec* cassettes highlighted the presence of further resistances and diversity within the same cassette type. On the contrary, when considering single infection-types or specific STs or clones as it is usual in *S. aureus* epidemiological studies, the genomic diversity was limited, with an increased pattern of resistance genes in chronic patients and a larger number of virulence factors in acute infections. Altogether, these observations shed more light on the complexity of *S. aureus* epidemiology and on the need for a more unbiased survey of the commensal and pathogenic *S. aureus* community, to avoid the misrepresentation of specific genomic traits.

Whole-genome-based routine surveillance of *S. aureus* and other hospital-related pathogens would further allow to get a more unbiased idea of the rising clones and better informing clinical practices, which usually focused on the most dangerous or well-known strains. Performing such epidemiological studies as soon as a new putative nosocomial clone arises could allow us to conclude whether the new clone has arisen in that very hospital or it is a recent sub-clone spreading also in the non-hospitalised population and therefore more frequently isolated also in the clinics. These wider-focus studies would not only allow the assessment of the epidemiology of specific pathogens and clones in the hospital setting, but also the survey of the prevalence and conservation of their virulence and resistance traits. This could lead to the identification of antigens of interest for vaccine development and of specific sub-clones representing the main burden of infection, and therefore reassessing the efforts for the discovery of new treatments.

Whole genome sequencing studies are crucial to survey the global epidemiology of infectious agents, including *S. aureus*, as genome-based data are reproducible and can be easily meta-analysed without the confounding of batch effects. The meta-analysis of pathogenic, commensal, and environmental *S. aureus* isolates could lead to a deeper knowledge of the epidemiology of this bacterium and may help in understanding how to prevent and treat infections without boosting antibiotic resistance.

## List of abbreviations

CC: clonal complex

CDS: coding sequence

CF: cystic fibrosis

CoNS: Coagulase-negative Staphylococci

indels: insertions and deletions

MLST: Multilocus Sequence typing

MRSA: methicillin-resistant *S. aureus*

MSSA: methicillin-sensitive *S. aureus*

PVL: Panton-Valentine Leukocidin

SCCmec: Staphylococcal Cassette Chromosome *mec*

SNV: single-nucleotide variation

Spa: Staphylococcal protein A

ST: sequence type

WGS: whole-genome sequencing

## 2.7 Declarations

### Ethics approval and consent to participate

The study received ethical approval from the Paediatric Ethics Committee, Autonomous Section of the Regional Ethics Committee for Clinical Trials at the Children's Meyer Hospital, Florence on July 1st, 2014. Data collection has been performed in accordance with the Declaration of Helsinki. All patients gave written informed consent.

### Consent for publication

Not applicable

### Availability of data and material

The datasets generated and analysed during the current study are available in the NCBI Sequence Read Archive (BioProject accession number PRJNA400143, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA400143/).

### Competing interests

The authors declare that they have no competing interests

### Funding

This work was supported in part by the Italian Ministry of Health “Ricerca finalizzata” RF-2010-2316179 to GT. This work was also partially supported by the European Union H2020 Marie-curie grant (707345) to E.P. and by a European Union FP7 Marie-Curie grant (PCIG13-618833), H2020-ERC-STG grant (MetaPG-716575), and by MIUR project FIR-RBFR13EWWI to NS.

## Authors' contributions

GT, SC, and NS conceived and supervised the study. DD, NR, and FA performed DNA extractions and library preparation. SM, FA, and EP conducted the computational and statistical analyses. SM, DD, GG, AM, GT, and NS analyzed and interpreted the results. ORS performed the phylogenetic analysis. SM, EP, DD, and NS wrote the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

The authors would like to acknowledge the laboratory of Computational Metagenomics for the valuable input. We would also like to thank Sander Wuyts and Stijn Wittouck for assistance in the plotting of the SCC*mec*IV graphs, and Dr. Megan Earls and Prof. David Coleman for sharing the ST1 SCC*mec*IV t127 genome.

## Additional files

**Additional file 1: Table S1.** Characteristics of the single isolates, including collection details, genome assembly statistics, genomic features, and results of antibiotic susceptibility testing. (XLSX 65 KB)

**Additional file 2: Figure S1.** Pangenome analysis statistics. **Figure S2.** Phylogenetic model based on gene presence/absence. (PDF 581 KB)

**Additional file 3: Table S2.** Presence / absence profile of virulence and antibiotic resistance genes in the cohort. (XLSX 5,861 KB)

**Additional file 4: Table S3.** Characteristics of the genomes included in the ST1 analyses. (XLSX 12 KB)

**Additional file 5: Table S4.** Most relevant clones represented in the cohort and their abundance. (XLSX 10 KB)

**Additional file 6: Table S5.** Bayesian clock models tested and their results. (XLSX 12 KB)

## 2.8 Supplementary Figures

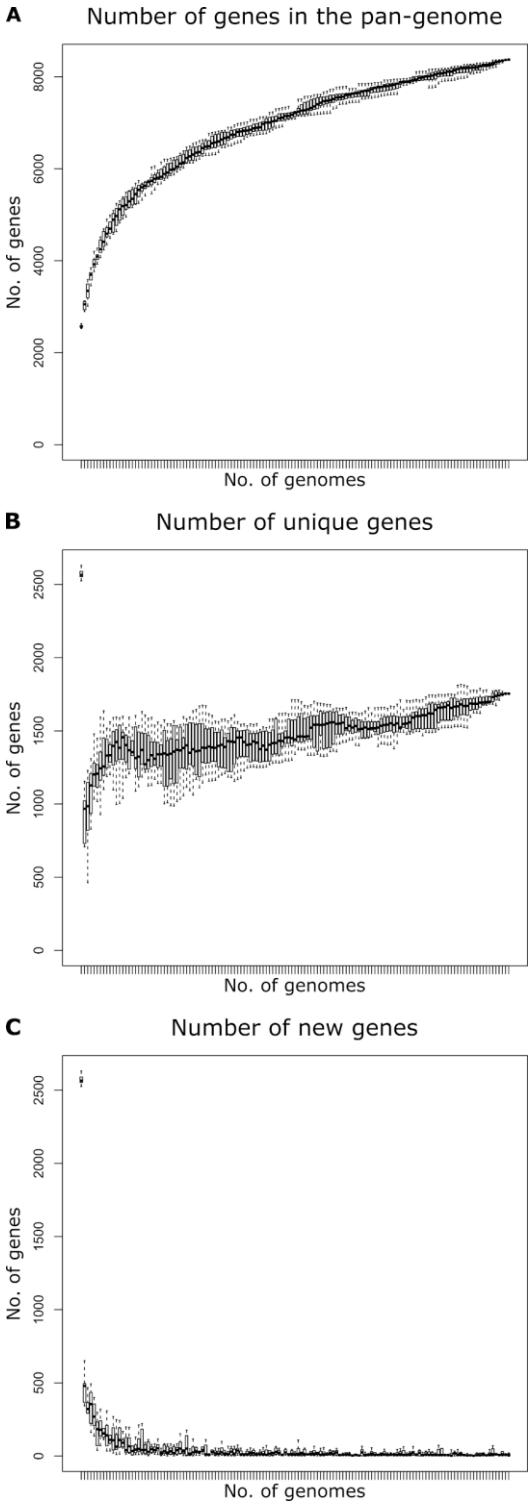



**Fig S1.** Pangenome analysis statistics. The Roary-computed Pangenome of the cohort consists of 8,373 genes (A), displaying a relatively high variability with a large proportion of unique genes (1,754) (B). However, the number of newly added genes per genome quickly drops to very small number (C).

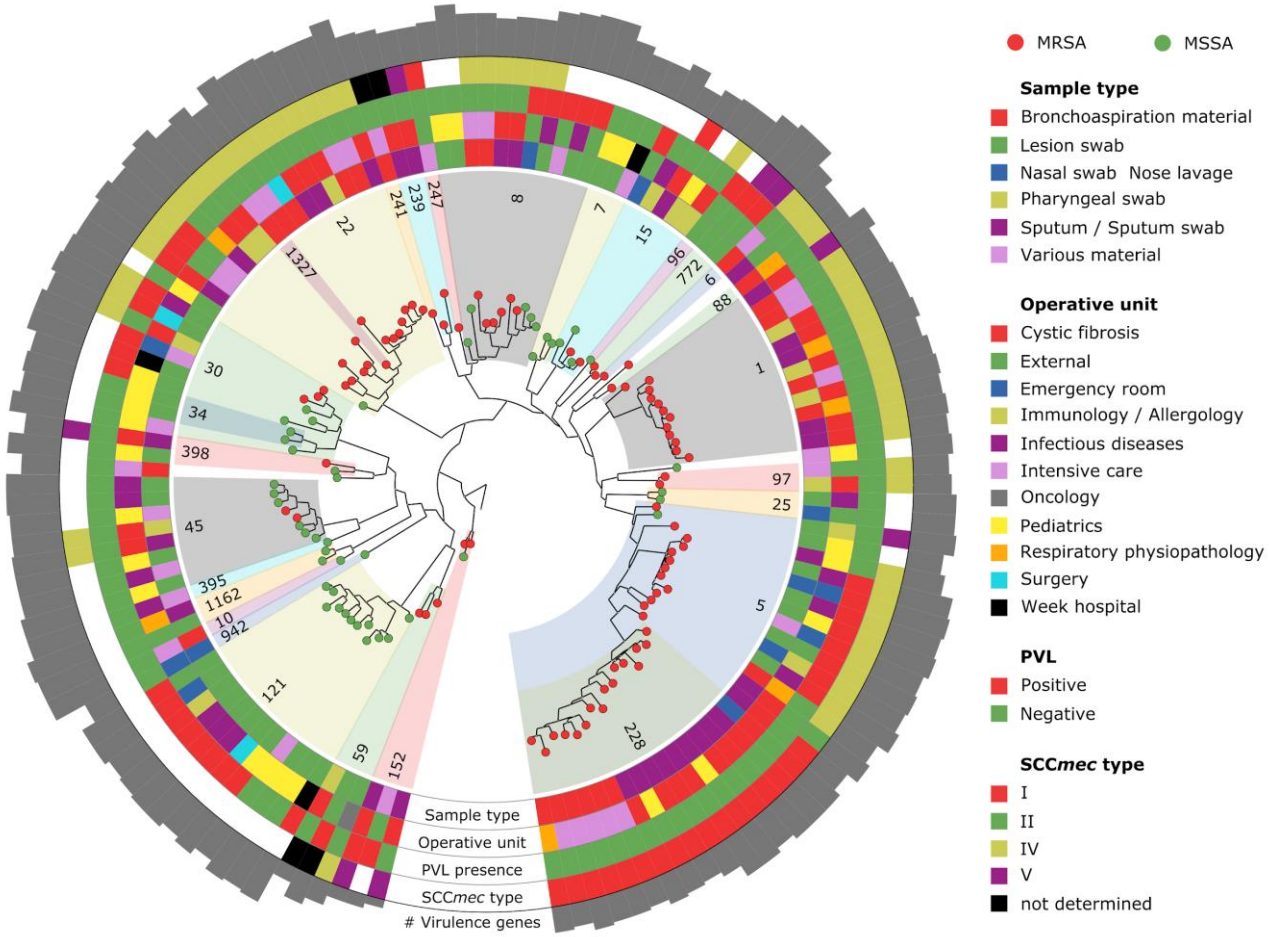



**Fig S2.** Phylogenetic model based on gene presence/absence (1,464 core and 6,909 accessory genes in total) in the 135 single-patient *S. aureus* isolates. STs are distinguished by means of numbers and background colours in the inner ring. Sample type, operative unit, PVL presence, and SCC*mec* type are colour-coded in the following rings. On the outermost ring, the number of virulence genes is reported as bar plot (total considered = 79).

## 2.9 Supplementary Tables

Captions of supplementary tables are reported below. Tables are available for download on the online version of the paper: https://doi.org/10.1186/s13073-018-0593-7

**Supplementary Table 1.** Characteristics of the single isolates, including collection details (patient code, type of sample, collection date and operative unit), genome assembly statistics (total length of the genome, number of contigs reconstructed, their median and mean length, N50, CG percentage number of CDS, genes, rRNA, sig_peptide, tmRNA, and tRNA), genomic features (ST and CC, SCC*mec* and *spa*-type, PVL presence, total number of virulence genes), and results of antibiotic susceptibility testing (POS = positive; NEG = negative; R = resistant; I = intermediate; S = sensitive; - = not available). Isolates presented in this study are reported as "selected" (n=135), whereas further timepoints of the same patient not analysed in this study are reported as "further timepoint" (n=43). ST1 isolates only used for the outbreak analysis (**Fig. 5**) are reported as "selected as further timepoint" (n=6).

**Supplementary Table 2.** Presence/absence (1/0) profile of 86 genes encoding for virulence factors (sheet "Virulence") and 18 genes encoding for resistances to antibiotics (sheet "Resistance"). A brief description of the function of the gene is reported in the first row.

**Supplementary Table 3.** Characteristics of the genomes included in the ST1 analyses. n.d. = not declared; [1] sample name is reported for samples obtained in the present study, reference genomes are named according to their accession number; [2] codes of patients enrolled in the present study are reported, * indicates patient for which more than one reference genome was available; [3] collection dates for reference genomes as reported in NCBI; some cases have no collection date specified, the date was assigned according to the reference publications.

**Supplementary Table 4.** Most relevant clones represented in the cohort and their abundance. Data regarding ST, SCC*mec*-, and *spa*-type for all 184 isolates are reported in Supplementary Table 1

**Supplementary Table 5.** Bayesian clock models tested and their results with respect to the one best fitting the cohort data (relaxed exponential constant generalised time reversible clock). Estimation of tree and subtree age is reported.

## 2.10 References


Al Laham, Nahed, José R. Mediavilla, Liang Chen, Nahed Abdelateef, Farid Abu Elamreen, Christine C. Ginocchio, Denis Pierard, Karsten Becker, and Barry N. Kreiswirth. 2015. “MRSA Clonal Complex 22 Strains Harboring Toxic Shock Syndrome Toxin (TSST-1) Are Endemic in the Primary Hospital in Gaza, Palestine.” *PloS One* 10 (3): e0120008.

AL-Tam, Faroq, Anne-Sophie Brunel, Nicolas Bouzinbi, Philippe Corne, Anne-Laure Bañuls, and Hamid Reza Shahbazkia. 2012. “DNAGear--a Free Software for Spa Type Identification in Staphylococcus Aureus.” *BMC Research Notes* 5 (November): 642.

Altschul, S. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1990.9999.

Amagai, M., N. Matsuyoshi, Z. H. Wang, C. Andl, and J. R. Stanley. 2000. “Toxin in Bullous Impetigo and Staphylococcal Scalded-Skin Syndrome Targets Desmoglein 1.” *Nature Medicine* 6 (11): 1275–77.

Anderson, Annaliesa S., Alita A. Miller, Robert G. K. Donald, Ingrid L. Scully, Jasdeep S. Nanra, David Cooper, and Kathrin U. Jansen. 2012. “Development of a Multicomponent Staphylococcus Aureus Vaccine Designed to Counter Multiple Bacterial Virulence Factors.” *Human Vaccines & Immunotherapeutics* 8 (11): 1585–94.

Archer, G. L., D. M. Niemeyer, J. A. Thanassi, and M. J. Pucci. 1994. “Dissemination among Staphylococci of DNA Sequences Associated with Methicillin Resistance.” *Antimicrobial Agents and Chemotherapy* 38 (3): 447–54.

Asghar, Atif H. 2014. “Molecular Characterization of Methicillin-Resistant Staphylococcus Aureus Isolated from Tertiary Care Hospitals.” *Pakistan Journal of Medical Sciences Quarterly* 30 (4): 698–702.

Badarau, Adriana, Harald Rouha, Stefan Malafa, Michael B. Battles, Laura Walker, Nels Nielson, Ivana Dolezilkova, et al. 2016. “Context Matters: The Importance of Dimerization-Induced Conformation of the LukGH Leukocidin of Staphylococcus Aureus for the Generation of Neutralizing Antibodies.” *mAbs* 8 (7): 1347–60.

Bagnoli, Fabio. 2017. “Staphylococcus Aureus Toxin Antibodies: Good Companions of Antibiotics and Vaccines.” *Virulence* 8 (7): 1037–42.

Bagnoli, Fabio, Maria Rita Fontana, Elisabetta Soldaini, Ravi P. N. Mishra, Luigi Fiaschi, Elena Cartocci, Vincenzo Nardi-Dei, et al. 2015. “Vaccine Composition Formulated with a Novel TLR7-Dependent Adjuvant Induces High and Broad Protection against Staphylococcus Aureus.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (12): 3680–85.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (5): 455–77.

Baranovich, T., H. Zaraket, I. I. Shabana, V. Nevzorova, V. Turcutyuicov, and H. Suzuki. 2010. “Molecular Characterization and Susceptibility of Methicillin-Resistant and Methicillin-Susceptible Staphylococcus Aureus Isolates from Hospitals and the Community in Vladivostok, Russia.” *Clinical Microbiology and Infection*. https://doi.org/10.1111/j.1469-0691.2009.02891.x.

Barber, M. 1961. “Methicillin-Resistant Staphylococci.” *Journal of Clinical Pathology* 14 (July): 385–93.

Begier, Elizabeth, David Joshua Seiden, Michael Patton, Edward Zito, Joseph Severs, David Cooper, Joseph Eiden, et al. 2017. “SA4Ag, a 4-Antigen Staphylococcus Aureus Vaccine, Rapidly Induces High Levels of Bacteria-Killing Antibodies.” *Vaccine* 35 (8): 1132–39.

Bertin, Mary L., Joan Vinski, Steven Schmitt, Camille Sabella, Lara Danziger-Isakov, Michael McHugh, Gary W. Procop, Geraldine Hall, Steven M. Gordon, and Johanna Goldfarb. 2006. “Outbreak of Methicillin-Resistant Staphylococcus Aureus Colonization and Infection in a Neonatal Intensive Care Unit Epidemiologically Linked to a Healthcare Worker With Chronic Otitis.” *Infection Control & Hospital Epidemiology*. https://doi.org/10.1086/504933.

Biber, Asaf, Izeldeen Abuelaish, Galia Rahav, Meir Raz, Liran Cohen, Lea Valinsky, Dianna Taran, et al. 2012. “A Typical Hospital-Acquired Methicillin-Resistant Staphylococcus Aureus Clone Is Widespread in the Community in the Gaza Strip.” *PloS One* 7 (8): e42864.

Blok, Hetty E. M., Annet Troelstra, Titia E. M. Kamp-Hopmans, Ada C. M. Gigengack-Baars, Christina M. J. E. Vandenbroucke-Grauls, Annemarie J. L. Weersink, Jan Verhoef, and Ellen M. Mascini. 2003. “Role of Healthcare Workers in Outbreaks of Methicillin-Resistant Staphylococcus Aureus: A 10-Year Evaluation from a Dutch University Hospital.” *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America* 24 (9): 679–85.

Bondi, A., Jr, and C. C. Dietz. 1945. “Penicillin Resistant Staphylococci.” *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine* 60 (October): 55–58.

Bosi, Emanuele, Jonathan M. Monk, Ramy K. Aziz, Marco Fondi, Victor Nizet, and Bernhard Ø. Palsson. 2016. “Comparative Genome-Scale Modelling of Staphylococcus Aureus Strains Identifies Strain-Specific Metabolic Capabilities Linked to Pathogenicity.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (26): E3801–9.

Boye, K., M. D. Bartels, I. S. Andersen, J. A. Møller, and H. Westh. 2007. “A New Multiplex PCR for Easy Screening of Methicillin-Resistant Staphylococcus Aureus SCCmec Types I-V.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 13 (7): 725–27.

Burdeska, A., M. Ott, W. Bannwarth, and R. L. Then. 1990. “Identical Genes for Trimethoprim-Resistant Dihydrofolate Reductase from Staphylococcus Aureus in Australia and Central Europe.” *FEBS Letters* 266 (1-2): 159–62.

Byrd, Allyson L., Yasmine Belkaid, and Julia A. Segre. 2018. “The Human Skin Microbiome.” *Nature Reviews. Microbiology* 16 (3): 143–55.

Campanile, F., D. Bongiorno, M. Falcone, F. Vailati, M. B. Pasticci, M. Perez, A. Raglio, et al. 2012. “Changing Italian Nosocomial-Community Trends and Heteroresistance in Staphylococcus Aureus from Bacteremia and Endocarditis.” *European Journal of Clinical Microbiology & Infectious Diseases*. https://doi.org/10.1007/s10096-011-1367-y.

Campanile, Floriana, Dafne Bongiorno, Marianna Perez, Gino Mongelli, Laura Sessa, Sabrina Benvenuto, Floriana Gona, AMCLI – S. aureus Survey Participants, Pietro E. Varaldo, and Stefania Stefani. 2015. “Epidemiology of Staphylococcus Aureus in Italy: First Nationwide Survey, 2012.” *Journal of Global Antimicrobial Resistance* 3 (4): 247–54.

Chambers, Henry F., and Frank R. Deleo. 2009. “Waves of Resistance: Staphylococcus Aureus in

the Antibiotic Era." *Nature Reviews. Microbiology* 7 (9): 629–41.

Cleef, Brigitte A. G. L. van, Haitske Graveland, Anja P. J. Haenen, Arjen W. van de Giessen, Dick Heederik, Jaap A. Wagenaar, and Jan A. J. W. Kluytmans. 2011. "Persistence of Livestock-Associated Methicillin-Resistant Staphylococcus Aureus in Field Workers after Short-Term Occupational Exposure to Pigs and Veal Calves." *Journal of Clinical Microbiology* 49 (3): 1030–33.

Cleef, Brigitte A. G. L. van, Dominique L. Monnet, Andreas Voss, Karina Krziwanek, Franz Allerberger, Marc Struelens, Helena Zemlickova, et al. 2011. "Livestock-Associated Methicillin-Resistant Staphylococcus Aureus in Humans, Europe." *Emerging Infectious Diseases* 17 (3): 502–5.

Coll, Francesc, Ewan M. Harrison, Michelle S. Toleman, Sandra Reuter, Kathy E. Raven, Beth Blane, Beverley Palmer, et al. 2017. "Longitudinal Genomic Surveillance of MRSA in the UK Reveals Transmission Patterns in Hospitals and the Community." *Science Translational Medicine* 9 (413). https://doi.org/10.1126/scitranslmed.aak9745.

Coombs, G. W., H. Van Gessel, J. C. Pearson, M-R Godsell, F. G. O'Brien, and K. J. Christiansen. 2007. "Controlling a Multicenter Outbreak Involving the New York/Japan Methicillin-Resistant Staphylococcus Aureus Clone." *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America* 28 (7): 845–52.

Copin, Richard, Bo Shopsin, and Victor J. Torres. 2018. "After the Deluge: Mining Staphylococcus Aureus Genomic Data for Clinical Associations and Host-Pathogen Interactions." *Current Opinion in Microbiology* 41 (February): 43–50.

Courjon, Johan, Patrick Munro, Yvonne Benito, Orane Visvikis, Coralie Bouchiat, Laurent Boyer, Anne Doye, et al. 2015. "EDIN-B Promotes the Translocation of Staphylococcus Aureus to the Bloodstream in the Course of Pneumonia." *Toxins* 7 (10): 4131–42.

Creech, C. Buddy, Robert W. Frenck Jr, Eric A. Sheldon, David J. Seiden, Martin K. Kankam, Edward T. Zito, Douglas Girgenti, et al. 2017. "Safety, Tolerability, and Immunogenicity of a Single Dose 4-Antigen or 3-Antigen Staphylococcus Aureus Vaccine in Healthy Older Adults: Results of a Randomised Trial." *Vaccine* 35 (2): 385–94.

Cullen, Louise, and Siobhán McClean. 2015. "Bacterial Adaptation during Chronic Respiratory Infections." *Pathogens* 4 (1): 66–89.

Cuny, Christiane, Robin Köck, and Wolfgang Witte. 2013. "Livestock Associated MRSA (LA-MRSA) and Its Relevance for Humans in Germany." *International Journal of Medical Microbiology: IJMM* 303 (6-7): 331–37.

Cuny, Christiane, Lothar H. Wieler, and Wolfgang Witte. 2015. "Livestock-Associated MRSA: The Impact on Humans." *Antibiotics (Basel, Switzerland)* 4 (4): 521–43.

David, M. D., A. M. Kearns, S. Gossain, M. Ganner, and A. Holmes. 2006. "Community-Associated Meticillin-Resistant Staphylococcus Aureus: Nosocomial Transmission in a Neonatal Unit." *The Journal of Hospital Infection* 64 (3): 244–50.

David, Michael Z., Susan Boyle-Vavra, Diana L. Zychowski, and Robert S. Daum. 2011. "Methicillin-Susceptible Staphylococcus Aureus as a Predominantly Healthcare-Associated Pathogen: A Possible Reversal of Roles?" *PloS One* 6 (4): e18217.

Delfani, Somayeh, Ashraf Mohabati Mobarez, Abbas Ali Imani Fooladi, Jafar Amani, and Mohammad Emaneini. 2016. "Protection of Mice against Staphylococcus Aureus Infection by

a Recombinant Protein ClfA-IsdB-Hlg as a Vaccine Candidate." *Medical Microbiology and Immunology* 205 (1): 47–55.

Deurenberg, Ruud H., and Ellen E. Stobberingh. 2008. "The Evolution of Staphylococcus Aureus." *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 8 (6): 747–63.

Diep, Binh An, Gregory G. Stone, Li Basuino, Christopher J. Graber, Alita Miller, Shelley-Ann des Etages, Alison Jones, et al. 2008. "The Arginine Catabolic Mobile Element and Staphylococcal Chromosomal Cassette Mec Linkage: Convergence of Virulence and Resistance in the USA300 Clone of Methicillin-Resistant Staphylococcus Aureus." *The Journal of Infectious Diseases* 197 (11): 1523–30.

Dinges, M. M., P. M. Orwin, and P. M. Schlievert. 2000. "Exotoxins of Staphylococcus Aureus." *Clinical Microbiology Reviews* 13 (1): 16–34, table of contents.

Dortet, Laurent, Delphine Girlich, Anne-Laure Virlouvet, Laurent Poirel, Patrice Nordmann, Bogdan I. Iorga, and Thierry Naas. 2017. "Characterization of BRPMBL, the Bleomycin Resistance Protein Associated with the Carbapenemase NDM." *Antimicrobial Agents and Chemotherapy* 61 (3). https://doi.org/10.1128/AAC.02413-16.

Dreisbach, Annette, Magdalena M. van der Kooi-Pol, Andreas Otto, Katrin Gronau, Hendrik P. J. Bonarius, Hans Westra, Herman Groen, Dörte Becher, Michael Hecker, and Jan M. van Dijl. 2011. "Surface Shaving as a Versatile Tool to Profile Global Interactions between Human Serum Proteins and the Staphylococcus Aureus Cell Surface." *Proteomics* 11 (14): 2921–30.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. "Bayesian Phylogenetics with BEAUti and the BEAST 1.7." *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/mss075.

Earls, Megan R., Peter M. Kinnevey, Gráinne I. Brennan, Alexandros Lazaris, Mairead Skally, Brian O'Connell, Hilary Humphreys, Anna C. Shore, and David C. Coleman. 2017. "The Recent Emergence in Hospitals of Multidrug-Resistant Community-Associated Sequence Type 1 and Spa Type t127 Methicillin-Resistant Staphylococcus Aureus Investigated by Whole-Genome Sequencing: Implications for Screening." *PloS One* 12 (4): e0175542.

Ellington, Matthew J., Lianne Yearwood, Mark Ganner, Claire East, and Angela M. Kearns. 2008. "Distribution of the ACME-arcA Gene among Methicillin-Resistant Staphylococcus Aureus from England and Wales." *The Journal of Antimicrobial Chemotherapy* 61 (1): 73–77.

Esposito, Silvano, Silvana Noviello, and Sebastiano Leone. 2016. "Epidemiology and Microbiology of Skin and Soft Tissue Infections." *Current Opinion in Infectious Diseases* 29 (2): 109–15.

Fattom, Ali I., Gary Horwith, Steve Fuller, Myra Propst, and Robert Naso. 2004. "Development of StaphVAX, a Polysaccharide Conjugate Vaccine against S. Aureus Infection: From the Lab Bench to Phase III Clinical Trials." *Vaccine* 22 (7): 880–87.

Fattom, Ali, Albert Matalon, John Buerkert, Kimberly Taylor, Silvia Damaso, and Dominique Boutriau. 2015. "Efficacy Profile of a Bivalent Staphylococcus Aureus Glycoconjugated Vaccine in Adults on Hemodialysis: Phase III Randomized Study." *Human Vaccines & Immunotherapeutics* 11 (3): 632–41.

Feltrin, Fabiola, Patricia Alba, Britta Kraushaar, Angela Ianzano, María Angeles Argudín, Paola Di Matteo, María Concepción Porrero, et al. 2016. "A Livestock-Associated, Multidrug-Resistant, Methicillin-Resistant Staphylococcus Aureus Clonal Complex 97 Lineage Spreading in Dairy Cattle and Pigs in Italy." *Applied and Environmental Microbiology* 82 (3): 816–21.

Fernandez, Silvina, Camila Ledo, Santiago Lattar, Mariángeles Noto Llana, Andrea Mendoza Bertelli, Sabrina Di Gregorio, Daniel O. Sordelli, Marisa I. Gómez, and Marta E. Mollerach. 2017. “High Virulence of Methicillin Resistant Staphylococcus Aureus ST30-SCCmecIVc-spat019, the Dominant Community-Associated Clone in Argentina.” *International Journal of Medical Microbiology: IJMM* 307 (4-5): 191–99.

Fitzgerald, J. Ross. 2012. “Livestock-Associated Staphylococcus Aureus: Origin, Evolution and Public Health Threat.” *Trends in Microbiology* 20 (4): 192–98.

Fowler, Vance G., Keith B. Allen, Edson D. Moreira, Moustafa Moustafa, Frank Isgro, Helen W. Boucher, G. Ralph Corey, et al. 2013. “Effect of an Investigational Vaccine for Preventing Staphylococcus Aureus Infections after Cardiothoracic Surgery: A Randomized Trial.” *JAMA: The Journal of the American Medical Association* 309 (13): 1368–78.

Frenck, Robert W., Jr, C. Buddy Creech, Eric A. Sheldon, David J. Seiden, Martin K. Kankam, James Baber, Edward Zito, et al. 2017. “Safety, Tolerability, and Immunogenicity of a 4-Antigen Staphylococcus Aureus Vaccine (SA4Ag): Results from a First-in-Human Randomised, Placebo-Controlled Phase 1/2 Study.” *Vaccine* 35 (2): 375–84.

Gennimata, D., J. Davies, and A. S. Tsiftsoglou. 1996. “Bleomycin Resistance in Staphylococcus Aureus Clinical Isolates.” *The Journal of Antimicrobial Chemotherapy* 37 (1): 65–75.

Geraci, Daniela M., Mario Giuffrè, Celestino Bonura, Domenica Matranga, Aurora Aleo, Laura Saporito, Giovanni Corsello, Anders Rhod Larsen, and Caterina Mammina. 2014. “Methicillin-Resistant Staphylococcus Aureus Colonization: A Three-Year Prospective Study in a Neonatal Intensive Care Unit in Italy.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0087760.

Geraci, D. M., C. Bonura, M. Giuffrè, A. Aleo, L. Saporito, G. Graziano, R. M. Valenti, and C. Mammina. 2014. “tst1-Positive ST22-MRSA-IVa in Healthy Italian Preschool Children.” *Infection* 42 (3): 535–38.

Giersing, Birgitte K., Sana S. Dastgheyb, Kayvon Modjarrad, and Vasee Moorthy. 2016. “Status of Vaccine Research and Development of Vaccines for Staphylococcus Aureus.” *Vaccine* 34 (26): 2962–66.

Givney, R., A. Vickery, A. Holliday, M. Pegler, and R. Benn. 1997. “Methicillin-Resistant Staphylococcus Aureus in a Cystic Fibrosis Unit.” *Journal of Hospital Infection*. https://doi.org/10.1016/s0195-6701(97)90165-1.

Goering, Richard V., Ribhi M. Shawar, Nicole E. Scangarella, F. Patrick O’Hara, Heather Amrine-Madsen, Joshua M. West, Marybeth Dalessandro, et al. 2008. “Molecular Epidemiology of Methicillin-Resistant and Methicillin-Susceptible Staphylococcus Aureus Isolates from Global Clinical Trials.” *Journal of Clinical Microbiology* 46 (9): 2842–47.

Goerke, Christiane, and Christiane Wolz. 2010. “Adaptation of Staphylococcus Aureus to the Cystic Fibrosis Lung.” *International Journal of Medical Microbiology: IJMM* 300 (8): 520–25.

Golubchik, Tanya, Elizabeth M. Batty, Ruth R. Miller, Helen Farr, Bernadette C. Young, Hanna Larner-Svensson, Rowena Fung, et al. 2013. “Within-Host Evolution of Staphylococcus Aureus during Asymptomatic Carriage.” *PloS One* 8 (5): e61319.

Gordon, N. C., J. R. Price, K. Cole, R. Everitt, M. Morgan, J. Finney, A. M. Kearns, et al. 2014. “Prediction of Staphylococcus Aureus Antimicrobial Resistance by Whole-Genome Sequencing.” *Journal of Clinical Microbiology* 52 (4): 1182–91.

Haft, Daniel H., Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin,

Kathleen O'Neill, Wenjun Li, et al. 2018. "RefSeq: An Update on Prokaryotic Genome Annotation and Curation." *Nucleic Acids Research* 46 (D1): D851–60.

Hanakawa, Yasushi, Norman M. Schechter, Chenyan Lin, Luis Garza, Hong Li, Takayuki Yamaguchi, Yasuyuki Fudaba, et al. 2002. "Molecular Mechanisms of Blister Formation in Bullous Impetigo and Staphylococcal Scalded Skin Syndrome." *The Journal of Clinical Investigation* 110 (1): 53–60.

Hanssen, Anne-Merethe, and Johanna U. Ericson Sollid. 2006. "SCCmec in Staphylococci: Genes on the Move." *FEMS Immunology and Medical Microbiology* 46 (1): 8–20.

Harrison, Ewan M., Francesc Coll, Michelle S. Toleman, Beth Blane, Nicholas M. Brown, M. Estee Török, Julian Parkhill, and Sharon J. Peacock. 2017. "Genomic Surveillance Reveals Low Prevalence of Livestock-Associated Methicillin-Resistant Staphylococcus Aureus in the East of England." *Scientific Reports*. https://doi.org/10.1038/s41598-017-07662-2.

Harris, Simon R., Edward J. P. Cartwright, M. Estée Török, Matthew T. G. Holden, Nicholas M. Brown, Amanda L. Ogilvy-Stuart, Matthew J. Ellington, et al. 2013. "Whole-Genome Sequencing for Analysis of an Outbreak of Meticillin-Resistant Staphylococcus Aureus: A Descriptive Study." *The Lancet Infectious Diseases* 13 (2): 130–36.

Harris, Simon R., Edward J. Feil, Matthew T. G. Holden, Michael A. Quail, Emma K. Nickerson, Narisara Chantratita, Susana Gardete, et al. 2010. "Evolution of MRSA during Hospital Transmission and Intercontinental Spread." *Science* 327 (5964): 469–74.

Hartman, B., and A. Tomasz. 1981. "Altered Penicillin-Binding Proteins in Methicillin-Resistant Strains of Staphylococcus Aureus." *Antimicrobial Agents and Chemotherapy* 19 (5): 726–35.

Haupt, Katrin, Michael Reuter, Jean van den Elsen, Julia Burman, Steffi Hälbich, Julia Richter, Christine Skerka, and Peter F. Zipfel. 2008. "The Staphylococcus Aureus Protein Sbi Acts as a Complement Inhibitor and Forms a Tripartite Complex with Host Complement Factor H and C3b." *PLoS Pathogens* 4 (12): e1000250.

Hiramatsu, K. 2001. "Vancomycin-Resistant Staphylococcus Aureus: A New Model of Antibiotic Resistance." *The Lancet Infectious Diseases* 1 (3): 147–55.

Hiramatsu, K., L. Cui, M. Kuroda, and T. Ito. 2001. "The Emergence and Evolution of Methicillin-Resistant Staphylococcus Aureus." *Trends in Microbiology* 9 (10): 486–93.

Hua, L., T. S. Cohen, Y. Shi, V. Datta, J. J. Hilliard, C. Tkaczyk, J. Suzich, C. K. Stover, and B. R. Sellman. 2015. "MEDI4893* Promotes Survival and Extends the Antibiotic Treatment Window in a Staphylococcus Aureus Immunocompromised Pneumonia Model." *Antimicrobial Agents and Chemotherapy* 59 (8): 4526–32.

Inoshima, Ichiro, Naoko Inoshima, Georgia A. Wilke, Michael E. Powers, Karen M. Frank, Yang Wang, and Juliane Bubeck Wardenburg. 2011. "A Staphylococcus Aureus Pore-Forming Toxin Subverts the Activity of ADAM10 to Cause Lethal Infection in Mice." *Nature Medicine* 17 (10): 1310–14.

Isobe, Hirokazu, Tomomi Takano, Akihito Nishiyama, Wei-Chun Hung, Shuichi Kuniyuki, Yasuhiro Shibuya, Ivan Reva, et al. 2012. "Evolution and Virulence of Panton-Valentine Leukocidin-Positive ST30 Methicillin-Resistant Staphylococcus Aureus in the Past 30 Years in Japan." *Biomedical Research* 33 (2): 97–109.

Jansen van Rensburg, M. J., V. Eliya Madikane, A. Whitelaw, M. Chachage, S. Haffejee, and B. Gay Elisha. 2011. "The Dominant Methicillin-resistant Staphylococcus Aureus Clone from

Hospitals in Cape Town Has an Unusual Genotype: ST612." *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 17 (5): 785–92.

Jevons, M. P., A. W. Coe, and M. T. Parker. 1963. "Methicillin Resistance in Staphylococci." *The Lancet* 1 (7287): 904–7.

Jongerius, Ilse, Maren von Köckritz-Blickwede, Malcolm J. Horsburgh, Maartje Ruyken, Victor Nizet, and Suzan H. M. Rooijakkers. 2012. "Staphylococcus Aureus Virulence Is Enhanced by Secreted Factors That Block Innate Immune Defenses." *Journal of Innate Immunity* 4 (3): 301–11.

Kahl, Barbara C. 2010. "Impact of Staphylococcus Aureus on the Pathogenesis of Chronic Cystic Fibrosis Lung Disease." *International Journal of Medical Microbiology: IJMM* 300 (8): 514–19.

Kim, J., R. G. Urban, J. L. Strominger, and D. C. Wiley. 1994. "Toxic Shock Syndrome Toxin-1 Complexed with a Class II Major Histocompatibility Molecule HLA-DR1." *Science* 266 (5192): 1870–74.

Kirby, W. M. 1944. "EXTRACTION OF A HIGHLY POTENT PENICILLIN INACTIVATOR FROM PENICILLIN RESISTANT STAPHYLOCOCCI." *Science* 99 (2579): 452–53.

Kos, Veronica N., Christopher A. Desjardins, Allison Griggs, Gustavo Cerqueira, Andries Van Tonder, Matthew T. G. Holden, Paul Godfrey, et al. 2012. "Comparative Genomics of Vancomycin-Resistant Staphylococcus Aureus Strains and Their Positions within the Clade Most Commonly Associated with Methicillin-Resistant S. Aureus Hospital-Acquired Infection in the United States." *mBio* 3 (3). https://doi.org/10.1128/mBio.00112-12.

Ladhani, Shamez. 2003. "Understanding the Mechanism of Action of the Exfoliative Toxins of Staphylococcus Aureus." *FEMS Immunology and Medical Microbiology* 39 (2): 181–89.

Landrum, Michael L., Tahaniyat Lalani, Minoo Niknian, Jason D. Maguire, Duane R. Hospenthal, Ali Fattom, Kimberly Taylor, et al. 2017. "Safety and Immunogenicity of a Recombinant Staphylococcus Aureus α-Toxoid and a Recombinant Panton-Valentine Leukocidin Subunit, in Healthy Adults." *Human Vaccines & Immunotherapeutics*. https://doi.org/10.1080/21645515.2016.1248326.

Larsen, Jesper, Paal S. Andersen, Volker Winstel, and Andreas Peschel. 2017. "Staphylococcus Aureus CC395 Harbours a Novel Composite Staphylococcal Cassette Chromosome Mec Element." *The Journal of Antimicrobial Chemotherapy* 72 (4): 1002–5.

Lina, G., Y. Piemont, F. Godail-Gamot, M. Bes, M. -O. Peter, V. Gauduchon, F. Vandenesch, and J. Etienne. 1999. "Involvement of Panton-Valentine Leukocidin--Producing Staphylococcus Aureus in Primary Skin Infections and Pneumonia." *Clinical Infectious Diseases*. https://doi.org/10.1086/313461.

Liu, Catherine, Arnold Bayer, Sara E. Cosgrove, Robert S. Daum, Scott K. Fridkin, Rachel J. Gorwitz, Sheldon L. Kaplan, et al. 2011. "Clinical Practice Guidelines by the Infectious Diseases Society of America for the Treatment of Methicillin-Resistant Staphylococcus Aureus Infections in Adults and Children: Executive Summary." *Clinical Infectious Diseases*. https://doi.org/10.1093/cid/cir034.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, et al. 1998. "Multilocus Sequence Typing: A Portable Approach to the Identification of Clones within Populations of Pathogenic Microorganisms." *Proceedings of the National Academy of Sciences of the United States of America* 95 (6): 3140–45.

Mainous, Arch G., 3rd, William J. Hueston, Charles J. Everett, and Vanessa A. Diaz. 2006. “Nasal Carriage of Staphylococcus Aureus and Methicillin-Resistant S Aureus in the United States, 2001-2002.” *Annals of Family Medicine* 4 (2): 132–37.

Mammina, Caterina, Cinzia Calà, Maria R. A. Plano, Celestino Bonura, Antonietta Vella, Rachele Monastero, and Daniela M. Palma. 2010. “Ventilator-Associated Pneumonia and MRSA ST398, Italy.” *Emerging Infectious Diseases* 16 (4): 730–31.

Mato, R., F. Campanile, S. Stefani, M. I. Crisóstomo, M. Santagati, Santos I. Sanches, and H. de Lencastre. 2004. “Clonal Types and Multidrug Resistance Patterns of Methicillin-Resistant Staphylococcus Aureus (MRSA) Recovered in Italy during the 1990s.” *Microbial Drug Resistance* 10 (2): 106–13.

McAdam, Paul R., Anne Holmes, Kate E. Templeton, and J. Ross Fitzgerald. 2011. “Adaptive Evolution of Staphylococcus Aureus during Chronic Endobronchial Infection of a Cystic Fibrosis Patient.” *PloS One* 6 (9): e24301.

Mediavilla, José R., Liang Chen, Barun Mathema, and Barry N. Kreiswirth. 2012. “Global Epidemiology of Community-Associated Methicillin Resistant Staphylococcus Aureus (CA-MRSA).” *Current Opinion in Microbiology* 15 (5): 588–95.

Miethke, T., K. Duschek, C. Wahl, K. Heeg, and H. Wagner. 1993. “Pathogenesis of the Toxic Shock Syndrome: T Cell Mediated Lethal Shock Caused by the Superantigen TSST-1.” *European Journal of Immunology* 23 (7): 1494–1500.

Milheiriço, Catarina, Duarte C. Oliveira, and Hermínia de Lencastre. 2007. “Multiplex PCR Strategy for Subtyping the Staphylococcal Cassette Chromosome Mec Type IV in Methicillin-Resistant Staphylococcus Aureus: ‘SCCmec IV Multiplex.’” *The Journal of Antimicrobial Chemotherapy* 60 (1): 42–48.

Milheirico, C., D. C. Oliveira, and H. de Lencastre. 2007. “Update to the Multiplex PCR Strategy for Assignment of Mec Element Types in Staphylococcus Aureus.” *Antimicrobial Agents and Chemotherapy*. https://doi.org/10.1128/aac.01362-07.

Molina, Auxiliadora, Rosa Del Campo, Luis Máiz, María-Isabel Morosini, Adelaida Lamas, Fernando Baquero, and Rafael Cantón. 2008. “High Prevalence in Cystic Fibrosis Patients of Multiresistant Hospital-Acquired Methicillin-Resistant Staphylococcus Aureus ST228-SCCmecI Capable of Biofilm Formation.” *The Journal of Antimicrobial Chemotherapy* 62 (5): 961–67.

Monaco, Monica, Palmino Pedroni, Andrea Sanchini, Annalisa Bonomini, Annamaria Indelicato, and Annalisa Pantosti. 2013. “Livestock-Associated Methicillin-Resistant Staphylococcus Aureus Responsible for Human Colonization and Infection in an Area of Italy with High Density of Pig Farming.” *BMC Infectious Diseases*. https://doi.org/10.1186/1471-2334-13-258.

Monaco, Monica, Fernanda Pimentel de Araujo, Melania Cruciani, Eliana M. Coccia, and Annalisa Pantosti. 2017. “Worldwide Epidemiology and Antibiotic Resistance of Staphylococcus Aureus.” *Current Topics in Microbiology and Immunology* 409: 21–56.

Moore, Carol L., Paola Osaki-Kiyan, Marybeth Perri, Susan Donabedian, Nadia Z. Haque, Anne Chen, and Marcus J. Zervos. 2010. “USA600 (ST45) Methicillin-Resistant Staphylococcus Aureus Bloodstream Infections in Urban Detroit.” *Journal of Clinical Microbiology* 48 (6): 2307–10.

Moustafa, Moustafa, George R. Aronoff, Chandra Chandran, Jonathan S. Hartzel, Steven S. Smugar, Claude M. Galphin, Lionel U. Mailloux, et al. 2012. “Phase IIa Study of the Immunogenicity and Safety of the Novel Staphylococcus Aureus Vaccine V710 in Adults with

End-Stage Renal Disease Receiving Hemodialysis." *Clinical and Vaccine Immunology: CVI* 19 (9): 1509–16.

Narita, Kouji, Dong-Liang Hu, Krisana Asano, and Akio Nakane. 2015. "Vaccination with Non-Toxic Mutant Toxic Shock Syndrome Toxin-1 Induces IL-17-Dependent Protection against Staphylococcus Aureus Infection." *Pathogens and Disease* 73 (4). https://doi.org/10.1093/femspd/ftv023.

Noto, Michael J., Barry N. Kreiswirth, Alastair B. Monk, and Gordon L. Archer. 2008. "Gene Acquisition at the Insertion Site for SCCmec, the Genomic Island Conferring Methicillin Resistance in Staphylococcus Aureus." *Journal of Bacteriology* 190 (4): 1276–83.

Oliveira, Duarte C., and Hermínia de Lencastre. 2002. "Multiplex PCR Strategy for Rapid Identification of Structural Types and Variants of the Mec Element in Methicillin-Resistant Staphylococcus Aureus." *Antimicrobial Agents and Chemotherapy* 46 (7): 2155–61.

Oliveira, Duarte C., Catarina Milheiriço, and Hermínia de Lencastre. 2006. "Redefining a Structural Variant of Staphylococcal Cassette Chromosome Mec, SCCmec Type VI." *Antimicrobial Agents and Chemotherapy* 50 (10): 3457–59.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. "Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv421.

Paharik, Alexandra E., and Alexander R. Horswill. 2016. "The Staphylococcal Biofilm: Adhesins, Regulation, and Host Response." *Microbiology Spectrum* 4 (2). https://doi.org/10.1128/microbiolspec.VMBF-0022-2015.

Pan, Angelo, Antonio Battisti, Alessia Zoncada, Francesco Bernieri, Massimo Boldini, Alessia Franco, Maurilio Giorgi, et al. 2009. "Community-Acquired Methicillin-Resistant Staphylococcus Aureus ST398 Infection, Italy." *Emerging Infectious Diseases* 15 (5): 845–47.

Pantosti, Annalisa, Andrea Sanchini, and Monica Monaco. 2007. "Mechanisms of Antibiotic Resistance in Staphylococcus Aureus." *Future Microbiology* 2 (3): 323–34.

Parker, M. T., and M. P. Jevons. 1964. "A SURVEY OF METHICILLIN RESISTANCE IN STAPHYLOCOCCUS AUREUS." *Postgraduate Medical Journal* 40 (December): SUPPL:170–78.

Peacock, Sharon J., and Gavin K. Paterson. 2015. "Mechanisms of Methicillin Resistance in Staphylococcus Aureus." *Annual Review of Biochemistry* 84: 577–601.

Pedersen, L. C., M. M. Benning, and H. M. Holden. 1995. "Structural Investigation of the Antibiotic and ATP-Binding Sites in Kanamycin Nucleotidyltransferase." *Biochemistry* 34 (41): 13305–11.

Planet, Paul J., Apurva Narechania, Liang Chen, Barun Mathema, Sam Boundy, Gordon Archer, and Barry Kreiswirth. 2017. "Architecture of a Species: Phylogenomics of Staphylococcus Aureus." *Trends in Microbiology* 25 (2): 153–66.

Price, James R., Kevin Cole, Andrew Bexley, Vasiliki Kostiou, David W. Eyre, Tanya Golubchik, Daniel J. Wilson, et al. 2017. "Transmission of Staphylococcus Aureus between Health-Care Workers, the Environment, and Patients in an Intensive Care Unit: A Longitudinal Cohort Study Based on Whole-Genome Sequencing." *The Lancet Infectious Diseases*. https://doi.org/10.1016/s1473-3099(16)30413-3.

Prosperi, Mattia, Nazle Veras, Taj Azarian, Mobeen Rathore, David Nolan, Kenneth Rand, Robert L. Cook, Judy Johnson, J. Glenn Morris, and Marco Salemi. 2013. “Molecular Epidemiology of Community-Associated Methicillin-Resistant Staphylococcus Aureus in the Genomic Era: A Cross-Sectional Study.” *Scientific Reports*. https://doi.org/10.1038/srep01902.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2007. “NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins.” *Nucleic Acids Research* 35 (Database issue): D61–65.

Raad, Issam, Hend Hanna, Ying Jiang, Tanya Dvorak, Ruth Reitzel, Gassan Chaiban, Robert Sherertz, and Ray Hachem. 2007. “Comparative Activities of Daptomycin, Linezolid, and Tigecycline against Catheter-Related Methicillin-Resistant Staphylococcus Bacteremic Isolates Embedded in Biofilm.” *Antimicrobial Agents and Chemotherapy* 51 (5): 1656–60.

Rammelkamp, Charles H., and Thelma Maxon. 1942. “Resistance of Staphylococcus Aureus to the Action of Penicillin.” *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine* 51 (3): 386–89.

Rao, Qing, Weilong Shang, Xiaomei Hu, and Xiancai Rao. 2015. “Staphylococcus Aureus ST121: A Globally Disseminated Hypervirulent Clone.” *Journal of Medical Microbiology* 64 (12): 1462–73.

Rasigade, Jean-Philippe, Frederic Laurent, Gerard Lina, Helene Meugnier, Michele Bes, François Vandenesch, Jerome Etienne, and Anne Tristan. 2010. “Global Distribution and Evolution of Panton-Valentine Leukocidin-Positive Methicillin-Susceptible Staphylococcus Aureus, 1981-2007.” *The Journal of Infectious Diseases* 201 (10): 1589–97.

Rhee, Yoona, Alla Aroutcheva, Bala Hota, Robert A. Weinstein, and Kyle J. Popovich. 2015. “Evolving Epidemiology of Staphylococcus Aureus Bacteremia.” *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America* 36 (12): 1417–22.

Roberts, Jill C. 2014. “Classification of Epidemic Community-Acquired Methicillin-Resistant Staphylococcus Aureus by Anatomical Site of Isolation.” *BioMed Research International* 2014 (May): 904283.

Roetzer, Andreas, Bernd Jilma, and Martha M. Eibl. 2017. “Vaccine against Toxic Shock Syndrome in a First-in-Man Clinical Trial.” *Expert Review of Vaccines* 16 (2): 81–83.

Rooijakkers, Suzan H. M., and Jos A. G. van Strijp. 2007. “Bacterial Complement Evasion.” *Molecular Immunology* 44 (1-3): 23–32.

Rouch, D. A., L. J. Messerotti, L. S. L. Loo, C. A. Jackson, and R. A. Skurray. 1989. “Trimethoprim Resistance Transposon Tn4003 from Staphylococcus Aureus Encodes Genes for a Dihydrofolate Reductase and Thymidylate Synthetase Flanked by Three Copies of IS257.” *Molecular Microbiology* 3 (2): 161–75.

Rouha, Harald, Adriana Badarau, Zehra C. Visram, Michael B. Battles, Bianka Prinz, Zoltán Magyarics, Gábor Nagy, et al. 2015. “Five Birds, One Stone: Neutralization of α-Hemolysin and 4 Bi-Component Leukocidins of Staphylococcus Aureus with a Single Human Monoclonal Antibody.” *mAbs* 7 (1): 243–54.

Saiman, Lisa, Alicia Cronquist, Fann Wu, Juyan Zhou, David Rubenstein, William Eisner, Barry N. Kreiswirth, and Phyllis Della-Latta. 2003. “An Outbreak of Methicillin-Resistant Staphylococcus Aureus in a Neonatal Intensive Care Unit.” *Infection Control & Hospital Epidemiology*. https://doi.org/10.1086/502217.

Salgado-Pabón, Wilmara, and Patrick M. Schlievert. 2014. “Models Matter: The Search for an Effective Staphylococcus Aureus Vaccine.” *Nature Reviews. Microbiology* 12 (8): 585–91.

Sanchez, Carlos J., Jr, Katrin Mende, Miriam L. Beckius, Kevin S. Akers, Desiree R. Romano, Joseph C. Wenke, and Clinton K. Murray. 2013. “Biofilm Formation by Clinical Isolates and the Implications in Chronic Infections.” *BMC Infectious Diseases* 13 (January): 47.

Schluepen, Christina, Enrico Malito, Ambra Marongiu, Markus Schirle, Elisabeth McWhinnie, Paola Lo Surdo, Marco Biancucci, et al. 2013. “Mining the Bacterial Unknown Proteome: Identification and Characterization of a Novel Family of Highly Conserved Protective Antigens in Staphylococcus Aureus.” *Biochemical Journal* 455 (3): 273–84.

Schwab, U. E., A. E. Wold, J. L. Carson, M. W. Leigh, P. W. Cheng, P. H. Gilligan, and T. F. Boat. 1993. “Increased Adherence of Staphylococcus Aureus from Cystic Fibrosis Lungs to Airway Epithelial Cells.” *The American Review of Respiratory Disease* 148 (2): 365–69.

Schwameis, Michael, Bernhard Roppenser, Christa Firbas, Corina S. Gruener, Nina Model, Norbert Stich, Andreas Roetzer, Nina Buchtele, Bernd Jilma, and Martha M. Eibl. 2016. “Safety, Tolerability, and Immunogenicity of a Recombinant Toxic Shock Syndrome Toxin (rTSST)-1 Variant Vaccine: A Randomised, Double-Blind, Adjuvant-Controlled, Dose Escalation First-in-Man Trial.” *The Lancet Infectious Diseases* 16 (9): 1036–44.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.

Shannon, Oonagh, and Jan-Ingmar Flock. 2004. “Extracellular Fibrinogen Binding Protein, Efb, from Staphylococcus Aureus Binds to Platelets and Inhibits Platelet Aggregation.” *Thrombosis and Haemostasis* 91 (4): 779–89.

Shannon, Oonagh, Andreas Uekötter, and Jan-Ingmar Flock. 2005. “Extracellular Fibrinogen Binding Protein, Efb, from Staphylococcus Aureus as an Antiplatelet Agent in Vivo.” *Thrombosis and Haemostasis* 93 (5): 927–31.

Sievert, Dawn M., Philip Ricks, Jonathan R. Edwards, Amy Schneider, Jean Patel, Arjun Srinivasan, Alex Kallen, Brandi Limbago, Scott Fridkin, and National Healthcare Safety Network (NHSN) Team and Participating NHSN Facilities. 2013. “Antimicrobial-Resistant Pathogens Associated with Healthcare-Associated Infections: Summary of Data Reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2009-2010.” *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America* 34 (1): 1–14.

Skinner, David, and Chester S. Keefer. 1941. “SIGNIFICANCE OF BACTEREMIA CAUSED BY STAPHYLOCOCCUS AUREUS: A STUDY OF ONE HUNDRED AND TWENTY-TWO CASES AND A REVIEW OF THE LITERATURE CONCERNED WITH EXPERIMENTAL INFECTION IN ANIMALS.” *Archives of Internal Medicine* 68 (5): 851–75.

Smith, Emma Jane, Livia Visai, Steven W. Kerrigan, Pietro Speziale, and Timothy J. Foster. 2011. “The Sbi Protein Is a Multifunctional Immune Evasion Factor of Staphylococcus Aureus.” *Infection and Immunity* 79 (9): 3801–9.

Soavi, Laura, Roberto Stellini, Liana Signorini, Benvenuto Antonini, Palmino Pedroni, Livio Zanetti, Bruno Milanesi, et al. 2010. “Methicillin-Resistant Staphylococcus Aureus ST398, Italy.” *Emerging Infectious Diseases* 16 (2): 346–48.

Spoor, Laura E., Paul R. McAdam, Lucy A. Weinert, Andrew Rambaut, Henrik Hasman, Frank M. Aarestrup, Angela M. Kearns, Anders R. Larsen, Robert L. Skov, and J. Ross Fitzgerald.

2013. “Livestock Origin for a Human Pandemic Clone of Community-Associated Methicillin-Resistant Staphylococcus Aureus.” *mBio* 4 (4). https://doi.org/10.1128/mBio.00356-13.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.

Stefani, Stefania, Floriana Campanile, Maria Santagati, Maria Lina Mezzatesta, Viviana Cafiso, and Giovanni Pacini. 2015. “Insights and Clinical Perspectives of Daptomycin Resistance in Staphylococcus Aureus: A Review of the Available Evidence.” *International Journal of Antimicrobial Agents* 46 (3): 278–89.

Stefani, Stefania, Doo Ryeon Chung, Jodi A. Lindsay, Alex W. Friedrich, Angela M. Kearns, Henrik Westh, and Fiona M. Mackenzie. 2012. “Meticillin-Resistant Staphylococcus Aureus (MRSA): Global Epidemiology and Harmonisation of Typing Methods.” *International Journal of Antimicrobial Agents* 39 (4): 273–82.

Stein, Michal, Shiri Navon-Venezia, Inna Chmelnitsky, David Kohelet, Orna Schwartz, Orly Agmon, and Eli Somekh. 2006. “AN OUTBREAK OF NEW, NONMULTIDRUG-RESISTANT, METHICILLIN-RESISTANT STAPHYLOCOCCUS AUREUS STRAIN (SCCMEC TYPE IIIA VARIANT-1) IN THE NEONATAL INTENSIVE CARE UNIT TRANSMITTED BY A STAFF MEMBER.” *The Pediatric Infectious Disease Journal*. https://doi.org/10.1097/01.inf.0000219407.31195.44.

Sung, Julia M-L, David H. Lloyd, and Jodi A. Lindsay. 2008. “Staphylococcus Aureus Host Specificity: Comparative Genomics of Human versus Animal Isolates by Multi-Strain Microarray.” *Microbiology* 154 (Pt 7): 1949–59.

Tacconelli, Evelina, Elena Carrara, Alessia Savoldi, Stephan Harbarth, Marc Mendelson, Dominique L. Monnet, Céline Pulcini, et al. 2018. “Discovery, Research, and Development of New Antibiotics: The WHO Priority List of Antibiotic-Resistant Bacteria and Tuberculosis.” *The Lancet Infectious Diseases* 18 (3): 318–27.

Tenover, Fred C., and Richard V. Goering. 2009. “Methicillin-Resistant Staphylococcus Aureus Strain USA300: Origin and Epidemiology.” *The Journal of Antimicrobial Chemotherapy* 64 (3): 441–46.

Tenover, Fred C., Linda M. Weigel, Peter C. Appelbaum, Linda K. McDougal, Jasmine Chaitram, Sigrid McAllister, Nancye Clark, et al. 2004. “Vancomycin-Resistant Staphylococcus Aureus Isolate from a Patient in Pennsylvania.” *Antimicrobial Agents and Chemotherapy* 48 (1): 275–80.

Tett, Adrian, Edoardo Pasolli, Stefania Farina, Duy Tin Truong, Francesco Asnicar, Moreno Zolfo, Francesco Beghini, et al. 2017. “Unexplored Diversity and Strain-Level Structure of the Skin Microbiome Associated with Psoriasis.” *NPJ Biofilms and Microbiomes* 3 (June): 14.

Thammavongsa, Vilasack, Hwan Keun Kim, Dominique Missiakas, and Olaf Schneewind. 2015. “Staphylococcal Manipulation of Host Immune Responses.” *Nature Reviews. Microbiology* 13 (9): 529–43.

Tinelli, Marco, Monica Monaco, Maurizio Vimercati, Antonio Ceraminiello, and Annalisa Pantosti. 2009. “Methicillin-Susceptible Staphylococcus Aureus in Skin and Soft Tissue Infections, Northern Italy.” *Emerging Infectious Diseases* 15 (2): 250–57.

Tong, Steven Y. C., Joshua S. Davis, Emily Eichenberger, Thomas L. Holland, and Vance G. Fowler Jr. 2015. “Staphylococcus Aureus Infections: Epidemiology, Pathophysiology, Clinical Manifestations, and Management.” *Clinical Microbiology Reviews* 28 (3): 603–61.

Tong, Steven Y. C., Matthew T. G. Holden, Emma K. Nickerson, Ben S. Cooper, Claudio U. Köser, Anne Cori, Thibaut Jombart, et al. 2015. “Genome Sequencing Defines Phylogeny and Spread of Methicillin-resistantStaphylococcus Aureusin a High Transmission Setting.” *Genome Research*. https://doi.org/10.1101/gr.174730.114.

Torre, Antonina, Marta Bacconi, Chiara Sammicheli, Bruno Galletti, Donatello Laera, Maria Rita Fontana, Guido Grandi, et al. 2015. “Four-Component Staphylococcus Aureus Vaccine 4C-Staph Enhances Fcγ Receptor Expression in Neutrophils and Monocytes and Mitigates S. Aureus Infection in Neutropenic Mice.” *Infection and Immunity* 83 (8): 3157–63.

Tosas Auguet, Olga, Richard A. Stabler, Jason Betley, Mark D. Preston, Mandeep Dhaliwal, Michael Gaunt, Avgousta Ioannou, et al. 2018. “Frequent Undetected Ward-Based Methicillin-Resistant Staphylococcus Aureus Transmission Linked to Patient Sharing Between Hospitals.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 66 (6): 840–48.

Tung, H. s., B. Guss, U. Hellman, L. Persson, K. Rubin, and C. Rydén. 2000. “A Bone Sialoprotein-Binding Protein from Staphylococcus Aureus: A Member of the Staphylococcal Sdr Family.” *Biochemical Journal* 345 Pt 3 (February): 611–19.

Udo, E. E., S. S. Boswihi, and N. Al-Sweih. 2016. “High Prevalence of Toxic Shock Syndrome Toxin-Producing Epidemic Methicillin-Resistant Staphylococcus Aureus 15 (EMRSA-15) Strains in Kuwait Hospitals.” *New Microbes and New Infections* 12 (July): 24–30.

Valsesia, Giorgia, Marco Rossi, Sonja Bertschy, and Gaby E. Pfyffer. 2010. “Emergence of SCCmec Type IV and SCCmec Type V Methicillin-Resistant Staphylococcus Aureus Containing the Panton-Valentine Leukocidin Genes in a Large Academic Teaching Hospital in Central Switzerland: External Invaders or Persisting Circulators?” *Journal of Clinical Microbiology* 48 (3): 720–27.

Vandenesch, Francois, Timothy Naimi, Mark C. Enright, Gerard Lina, Graeme R. Nimmo, Helen Heffernan, Nadia Liassine, et al. 2003. “Community-Acquired Methicillin-Resistant Staphylococcus Aureus Carrying Panton-Valentine Leukocidin Genes: Worldwide Emergence.” *Emerging Infectious Diseases* 9 (8): 978–84.

Verkade, Erwin, Anneke M. C. Bergmans, Andries E. Budding, Alex van Belkum, Paul Savelkoul, Anton G. Buiting, and Jan Kluytmans. 2012. “Recent Emergence of Staphylococcus Aureus Clonal Complex 398 in Human Blood Cultures.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0041855.

Verkaik, Nelianne J., Willem J. B. van Wamel, and Alex van Belkum. 2011. “Immunotherapeutic Approaches against Staphylococcus Aureus.” *Immunotherapy* 3 (9): 1063–73.

Voss, A., and B. N. Doebbeling. 1995. “The Worldwide Prevalence of Methicillin-Resistant Staphylococcus Aureus.” *International Journal of Antimicrobial Agents* 5 (2): 101–6.

Wang, J. T., S. C. Chang, W. J. Ko, Y. Y. Chang, M. L. Chen, H. J. Pan, and K. T. Luh. 2001. “A Hospital-Acquired Outbreak of Methicillin-Resistant Staphylococcus Aureus Infection Initiated by a Surgeon Carrier.” *The Journal of Hospital Infection* 47 (2): 104–9.

Wertheim, Heiman F. L., Damian C. Melles, Margreet C. Vos, Willem van Leeuwen, Alex van Belkum, Henri A. Verbrugh, and Jan L. Nouwen. 2005. “The Role of Nasal Carriage in Staphylococcus Aureus Infections.” *The Lancet Infectious Diseases* 5 (12): 751–62.

Whitener, Cynthia J., Sarah Y. Park, Fred A. Browne, Leslie J. Parent, Kathleen Julian, Bulent Bozdogan, Peter C. Appelbaum, et al. 2004. “Vancomycin-Resistant Staphylococcus Aureus

in the Absence of Vancomycin Exposure." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 38 (8): 1049–55.

Winstel, Volker, Chunguang Liang, Patricia Sanchez-Carballo, Matthias Steglich, Marta Munar, Barbara M. Bröker, Jose R. Penadés, et al. 2013. "Wall Teichoic Acid Structure Governs Horizontal Gene Transfer between Major Bacterial Pathogens." *Nature Communications* 4: 2345.

Witte, Wolfgang, Birgit Strommenger, Christian Stanek, and Christiane Cuny. 2007. "Methicillin-Resistant Staphylococcus Aureus ST398 in Humans and Animals, Central Europe." *Emerging Infectious Diseases* 13 (2): 255–58.

Wright, Gerard D. 2007. "The Antibiotic Resistome: The Nexus of Chemical and Genetic Diversity." *Nature Reviews. Microbiology* 5 (3): 175–86.

Wu, Dejing, Qun Wang, Yonghong Yang, Wenjing Geng, Qiang Wang, Sangjie Yu, Kaihu Yao, Lin Yuan, and Xuzhuang Shen. 2010. "Epidemiology and Molecular Characteristics of Community-Associated Methicillin-Resistant and Methicillin-Susceptible Staphylococcus Aureus from Skin/soft Tissue Infections in a Children's Hospital in Beijing, China." *Diagnostic Microbiology and Infectious Disease*. https://doi.org/10.1016/j.diagmicrobio.2009.12.006.

Yang, Liuyang, Heng Zhou, Ping Cheng, Yun Yang, Yanan Tong, Qianfei Zuo, Qiang Feng, Quanming Zou, and Hao Zeng. 2018. "A Novel Bivalent Fusion Vaccine Induces Broad Immunoprotection against Staphylococcus Aureus Infection in Different Murine Models." *Clinical Immunology* 188 (March): 85–93.

Zhang, K., J. -A. McClure, S. Elsayed, T. Louie, and J. M. Conly. 2005. "Novel Multiplex PCR Assay for Characterization and Concomitant Subtyping of Staphylococcal Cassette Chromosome Mec Types I to V in Methicillin-Resistant Staphylococcus Aureus." *Journal of Clinical Microbiology*. https://doi.org/10.1128/jcm.43.10.5026-5033.2005.

Zolfo, Moreno, Adrian Tett, Olivier Jousson, Claudio Donati, and Nicola Segata. 2017. "MetaMLST: Multi-Locus Strain-Level Bacterial Typing from Metagenomic Samples." *Nucleic Acids Research* 45 (2): e7.

# Chapter 3. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling

## 3.1 Introduction to the chapter

In **Chapter 2** I showed how cultivation-based investigation of a known opportunistic pathogen allows us to deeply investigate its genetic traits and to conduct comparative genomic analysis. In this Chapter, I will introduce an experimental and analytic framework that we proposed and allowed us to extract similar genomic information from bacterial genomes using cultivation-free metagenomic sequencing. This is a key result that can enable the simultaneous study of hundreds of members of the human microbiome in a single experiment without the need of cultivating each target organism separately. The framework was applied on the characterization of stool and breast milk microbiome samples with the final goal of investigating the occurrence of microbial transmission from mothers to their infants. This pilot study has laid the methodological basis for strain-level analysis of microbes from shotgun metagenomic data, and has paved the way for larger cohort studies on the "vertical" transmission of the microbiome from mother to infant and on the dynamics of microbial development in infants.

The gut microbiome plays important roles in human physiology and metabolism (Clemente et al. 2012; HMP et al. 2012; Bäckhed et al. 2005; Palm, de Zoete, and Flavell 2015) and adapts to changes in diet, environment and antibiotic use throughout the life of a person. Early-life development of the gut microbiome has been shown to be key for future health of the infant (Bäckhed et al. 2015; Yatsunenko et al. 2012; Palmer et al. 2007) and to be influenced, among others, by mode of delivery (Dominguez-Bello et al. 2010; Azad et al. 2013), perinatal antibiotic use (Greenwood et al. 2014) and gestational age at birth (La Rosa et al. 2014). The infant gut microbiome is established in the first days of life and rapidly evolves during the first few months. Elucidating how this process occurs and what are the primary sources of the colonizing microbiome could unravel how dysbiotic conditions arise and possibly how to prevent them. Transmission of bacteria from the mother to the infant during vaginal delivery (Dominguez-Bello et al. 2010; Biasucci et al. 2010) and breastfeeding (Jost et al. 2014) has been proposed as one of the main routes of microbial seeding of the gut. Despite preliminary studies aimed at elucidating which microbial clades are vertically acquired from the mother, these were either limited to the cultivable fraction of the microbiome (Makino et al. 2011) or lacked strain-level resolution, as in 16S rRNA amplicon sequencing studies (Bäckhed et al. 2015; Dominguez-Bello et al. 2010; Biasucci et al. 2010). The need for strain-level resolution is explained by the fact that the same species can be detected in unrelated people (Lozupone et al. 2012) that are however carrying different strains (Scholz et al. 2016; Schloissnig et al. 2013). Therefore, previous identification of the same species in mother and infant (Palmer et al. 2007; Turnbaugh et al. 2009) does not necessarily represent vertical transmission events, and a comprehensive strain-level understanding of which infant-associated microbes are acquired from the mother was lacking. It is however unlikely that cultivation-based

approaches can scale to the throughput necessary for surveying the overall diversity of the human microbiome and its transmission between hosts.

Metagenomic approaches sequencing the total DNA content of the sample allow to access uncultivable microbes, but a validated metagenomic pipeline to track microbes at the strain-level was still lacking at the time of writing of this study. Even less explored was whether the transmitted strains would be transcriptionally active in the infant gut, and if so, how. Despite different studies had already reported the transcriptional activity of members of the gut microbiome (Turnbaugh et al. 2010; Maurice, Haiser, and Turnbaugh 2013), no studies had investigated the activity of vertically transmitted strains *in vivo*.

In the article reported in this Chapter and published in mSystems in 2017, we identified and functionally characterized commensal strains transmitted from the mother to the infant during the first months of life by detecting their presence across microbiome samples obtained from different individuals. Strain-tracking was performed without isolation of the single microbes, but through the use of computational tools able to track microbes at the strain level directly from shotgun metagenomics data, through the application of both a SNP-based (Truong et al. 2017) and a pangenome-based (Scholz et al. 2016) approach. This allows the identification and tracking also of cultivation-recalcitrant species, shedding light on the contribution of different and potentially unexpected microbes in the development of the infant gut microbiome during the first year of life. To this end, we collected and shotgun sequenced stool and breast milk samples obtained from five mother-infant pairs sampled longitudinally. Metatranscriptomics was applied on two selected mother-infant pairs to investigate the differential expression profiles of the vertically transmitted bacterial strains in the gut of mothers and their children. Overall, we identified a number of species shared within the mother-infant pair, and for many of them the very same strain was present while being different from those carried by other mother-infant pairs. This finding highlights the importance of strain-level analysis for detecting vertical transmission events. At early time points, shared strains consisted mainly of *Bifidobacteria* and *Escherichia coli*. We observed the development of the post-weaning microbiome toward a more adult-like composition, with additional sharing of several Lachnospiraceae. This shift was evident also in the functional potential of the microbiome, with pathways highly represented in infants approaching lower adult-like levels, as in the case of folate biosynthesis and intestinal mucin utilization. Metatranscriptomics further highlighted different transcriptional patterns of vertically transmitted strains in mothers and infants, as in the case of *Bacteroides vulgatus* that was highly transcribing in the infant but not in the mother's gut. Post-weaning metatranscriptomic data highlighted an increase in the expression levels of genes and pathways involved in starch metabolism and fermentation, consistently with the change in the diet of the infant, suggesting intriguing adaptation patterns.

Overall, this pilot study presented and validated a methodological approach for strain-level investigation of microbiome members from shotgun metagenomics data, paving the way for larger cohort studies on vertical transmission (Miyoshi et al. 2017; Wampach et al.

2017; Cabral et al. 2017; Ximenez and Torres 2017; Davenport et al. 2017; Yassour et al. 2018; Wampach et al. 2018; Vatanen et al. 2018; Korpela and de Vos 2018), including the study that is partially reported in **Chapter 5.1** (Ferretti et al. 2018). Availability of meta-transcriptomic data for two of the mother-infant pairs allowed us to survey also the differential activity of the transmitted strains in the adult and infant gut, and represents one of the main novelty points contributed by this article to the vertical transmission literature available at the time of publication.

**Outlook.** Our work and related methodologies published in the same period were thus crucial in raising the awareness that metagenomic sequencing can be the needed tool to characterize and track microbial strains with a comparable level of resolution provided by whole-genome isolate sequencing. More work is needed to show how to identify species that are still hidden in a metagenomic sample (**Chapter 4**), but the article in this chapter showed for the first time that cultivation-free metagenomic approaches coupled with appropriate computational methods can survey the influx of microbial organisms in the infant gut microbiome.

**Contribution.** For this article, I performed DNA and RNA co-extraction from stool and milk samples (including protocols for rRNA depletion), prepared metagenomic and metatranscriptomic libraries for Illumina HiSeq sequencing, contributed on the application of the computational metagenomic tools with the other first co-author, led the data interpretation, and wrote the majority of the manuscript. The bioinformatic analysis was done in collaboration with the other co-first author (Francesco Asnicar) who performed the analysis based on the hypotheses and assumptions we made together and under the supervision of the corresponding author.

This chapter reports the following article:

**Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling**

Francesco Asnicar^, Serena Manara^, Moreno Zolfo, Duy Tin Truong, Matthias Scholz, Federica Armanini, Pamela Ferretti, Valentina Gorfer, Anna Pedrotti, Adrian Tett, and Nicola Segata

^ these authors contributed equally

*mSystems* 2017

## 3.2 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.

## 3.3 Introduction

The community of microorganisms that dwell in the human gut has been shown to play an integral role in human health (Qin et al. 2010; Clemente et al. 2012; Tamburini et al. 2016; HMP et al. 2012), facilitating, for instance, the harvesting of nutrients that would otherwise be inaccessible (Bäckhed et al. 2005), modulating the host metabolism and immune system (Palm, de Zoete, and Flavell 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 (Fuentes et al. 2014; Khoruts et al. 2010; Britton and Young 2014).

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; Yatsunenko et al. 2012; Palmer et al. 2007), and a few studies have investigated the factors that are important in defining its early structure (Dominguez-Bello et al. 2010; Azad et al. 2013; Milani et al. 2015; La Rosa et al. 2014). In particular, gestational age at birth (La Rosa et al. 2014), mode of delivery (Dominguez-Bello et al. 2010; Azad et al. 2013), 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 (Dominguez-Bello et al. 2010; Aagaard et al. 2012; Hunt et al. 2011; Cabrera-Rubio et al. 2012) 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; Dominguez-Bello et al. 2010; Biasucci et al. 2010; Dominguez-Bello et al. 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 (Scholz et al. 2016; Schloissnig et al. 2013), 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 (Milani et al. 2015; Makino et al. 2011), 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 (Turnbaugh et al. 2010; Maurice, Haiser, and Turnbaugh 2013; Gosalbes et al. 2012; Bao et al. 2015), 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.

## 3.4 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 (Yatsunenko et al. 2012; Kurokawa et al. 2007; Koenig et al. 2011). 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* (Palmer et al. 2007; Koenig et al. 2011). 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**).

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**).

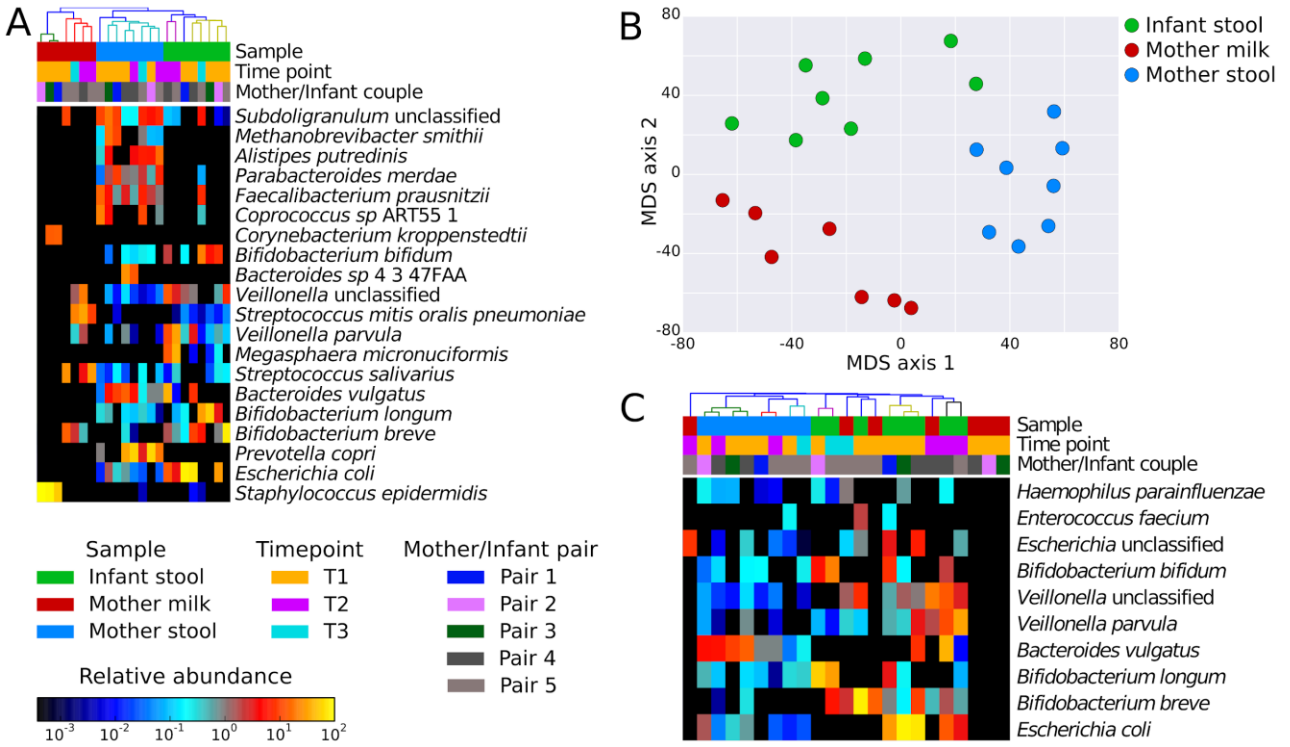



**Fig 1. Microbial composition of mother and infant samples and shared bacteria within mother-infant pairs.** (A) Quantitative microbial taxonomic composition of the metagenomic samples from milk and fecal samples of mothers and infants as estimated by MetaPhlAn2 analysis (Truong et al. 2015) (only the 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).

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 (Dominguez-Bello et al. 2010; Hunt et al. 2011; Cabrera-Rubio et al. 2012; 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 (Palmer et al. 2007; Koenig et al. 2011; Costello et al. 2012).

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; Milani et al. 2015; La Rosa et al. 2014; Jost et al. 2014; Faith et al. 2013) 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 (Scholz et al. 2016; Schloissnig et al. 2013). 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.

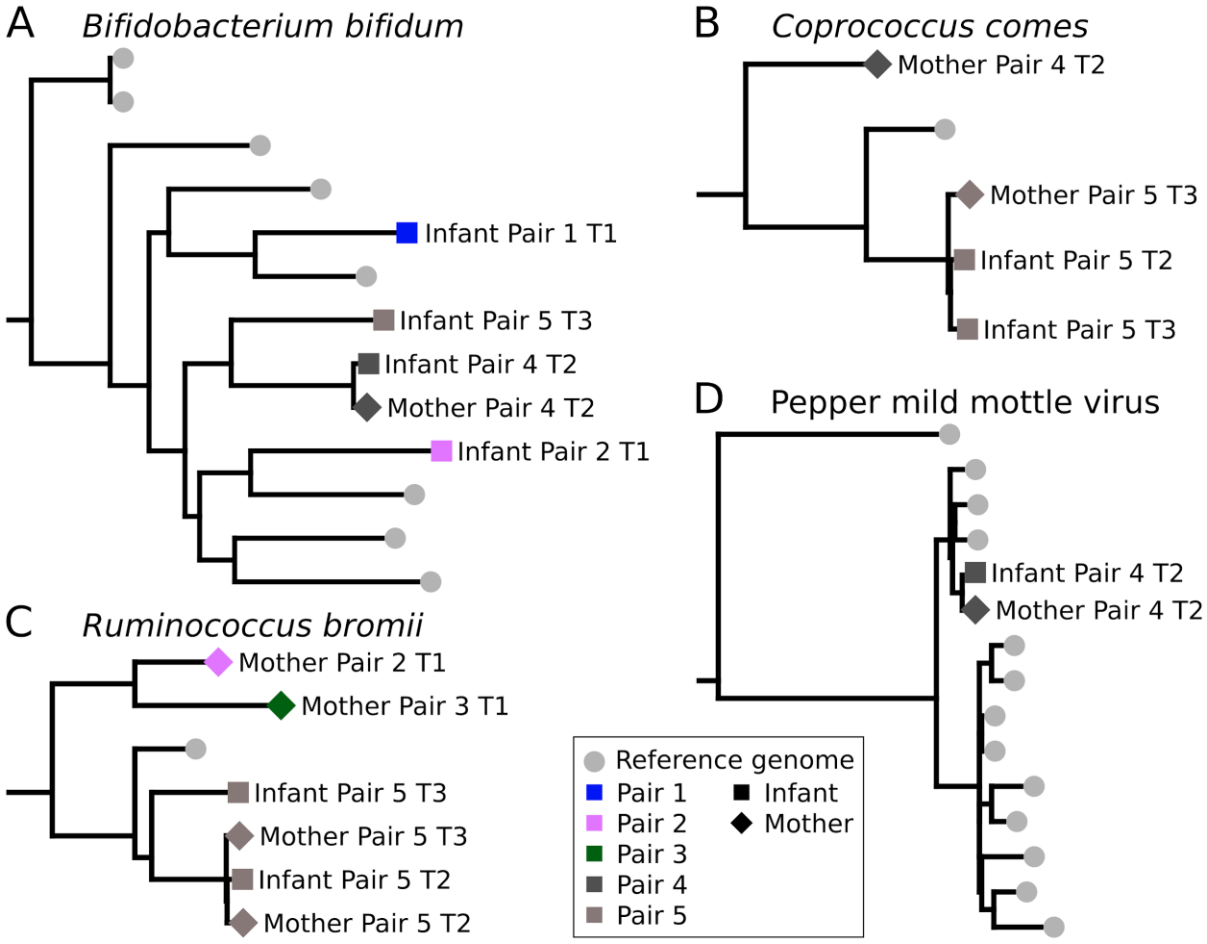



**Fig 2. Strain-level phylogenetic trees for microbes present in both the mother and infant.** Phylogenetic trees were built by the StrainPhlAn method using species-specific markers confirming the presence of the same strain in the mother and infant intestinal microbiomes, thus suggesting vertical transmission. 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 (Tamburini et al. 2016; Ogilvie and Jones 2015). 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 (Victoria et al. 2009; Reyes et al. 2010; 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 (de Medeiros et al. 2005; Balique et al. 2013). 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 (Bäckhed et al. 2015; Palmer et al. 2007; Azad et al. 2013; Hunt et al. 2011; Cabrera-Rubio et al. 2012; Jeurink et al. 2013; Costello et al. 2012). 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 (Yatsunenko et al. 2012; Marcobal et al. 2011; Tailford et al. 2015; Turroni, Milani, et al. 2011; LeBlanc et al. 2013). 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 (Yatsunenko et al. 2012; Marcobal et al. 2011; Tailford et al. 2015; Turroni, Milani, et al. 2011), whereas increased folate biosynthesis (Yatsunenko et al. 2012; Marcobal et al. 2011; Tailford et al. 2015; LeBlanc et al. 2013) 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**).

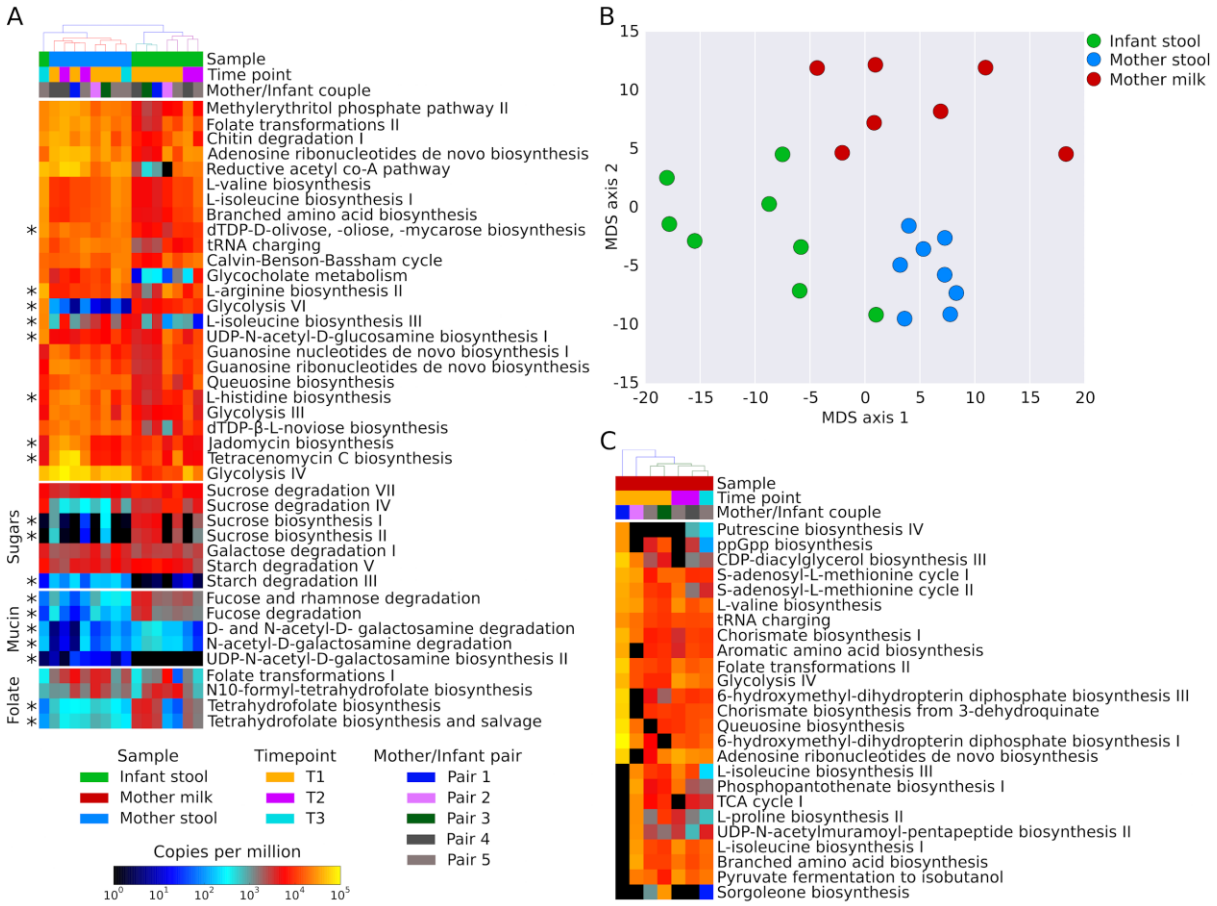



**Fig 3. Functional potential analyses.** (A) HUMAnN2 heat map reporting the 25 most abundant pathways in the fecal samples of mothers and infants. Specific pathways of interest (sugars, mucin, and folate metabolism) are added at the bottom. The asterisk (*) near the heat map highlights statistically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

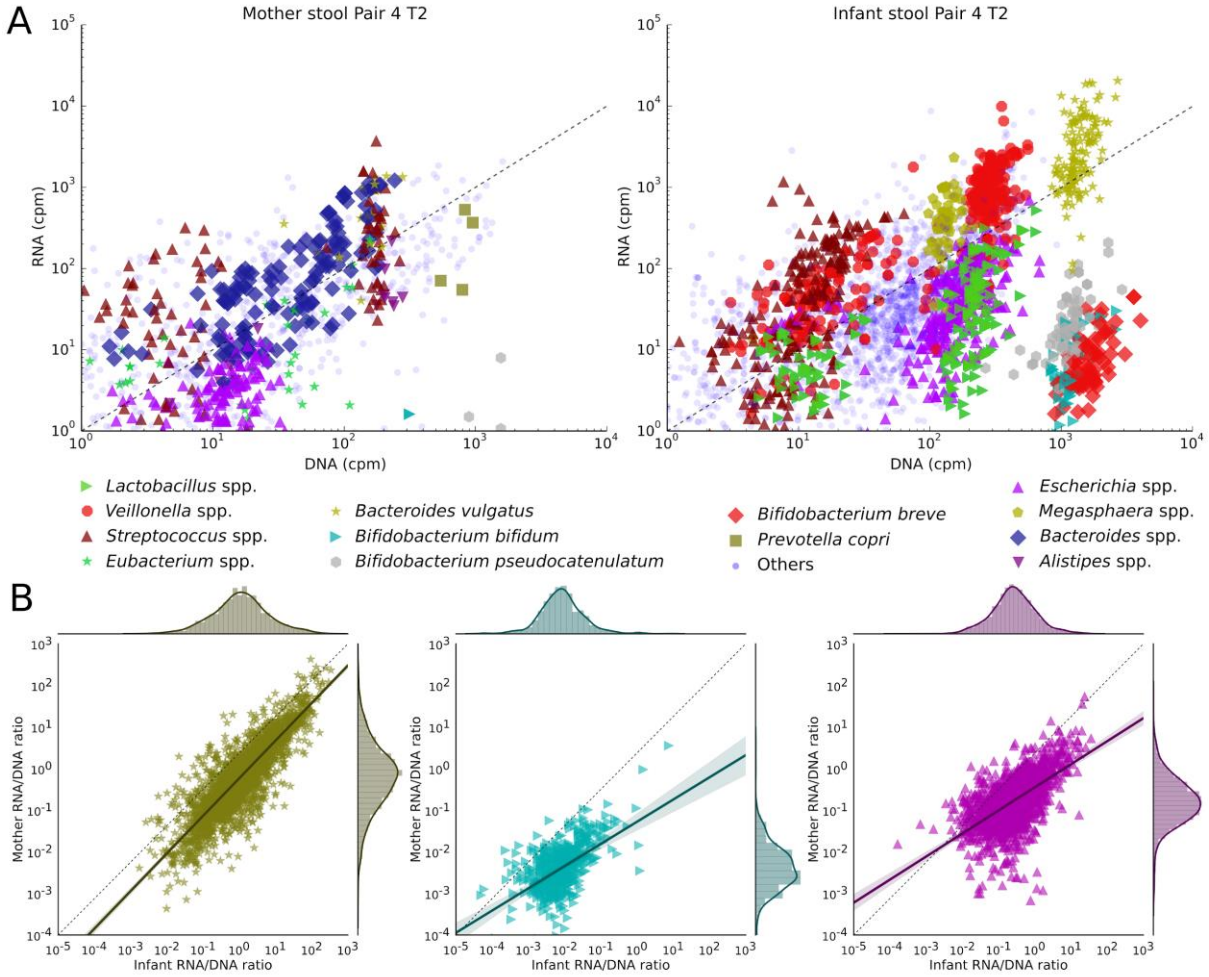



**Fig 4. Transcription levels of metabolic pathways and genes in mother and infant pair 4 at time point 2.** (A) Scatterplots showing the transcription rates of metabolic pathways of shared and nonshared species and genera of interest for both the mother and infant of pair 4 at time point 2. (B) Comparison 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 (Yatsunenko et al. 2012; Kurokawa et al. 2007; Koenig et al. 2011). Interestingly, the *B. bifidu*m 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. 2012; Turroni, Foroni, et al. 2011). 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 (Walker et al. 2011; Scott et al. 2015).

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.

## 3.5 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 (Loman et al. 2013; S. S. Li et al. 2016), 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.

## 3.6 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 (H. 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. 2015) 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.).

## 3.7 Supplementary Figures

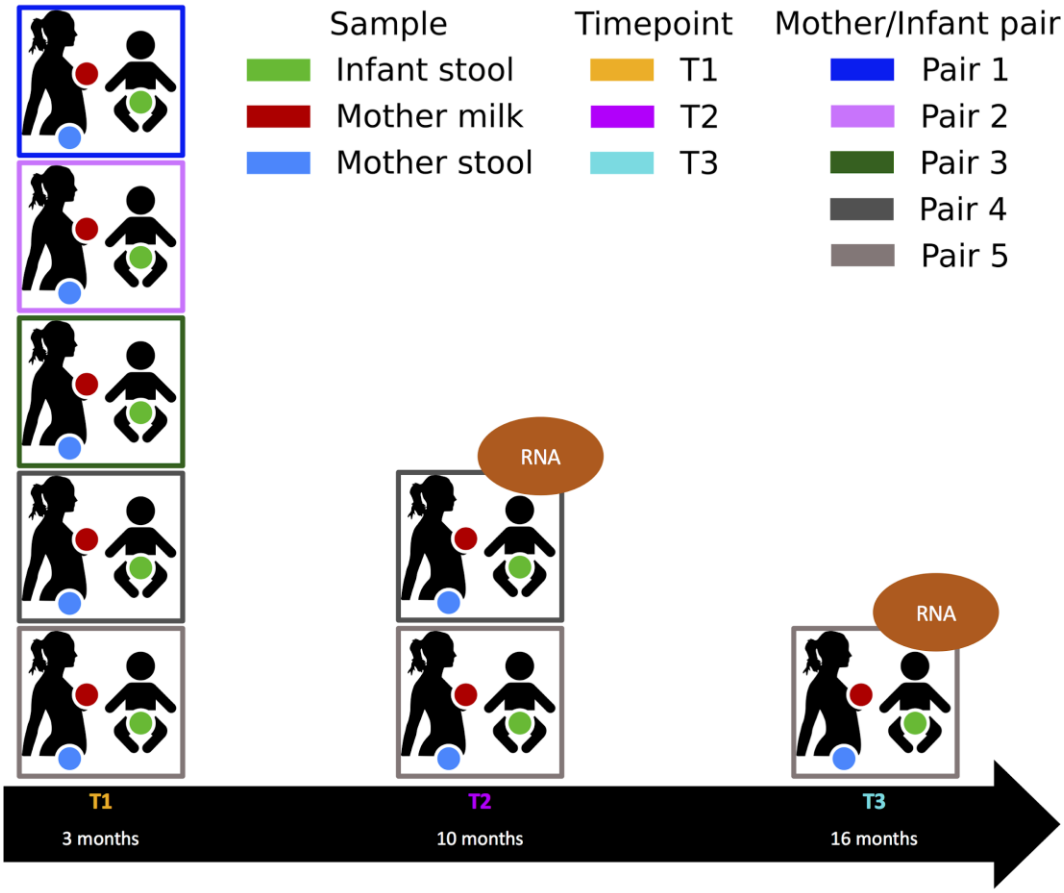



**Fig S1. Study design.** A schematic representation of the mother-infant pairs involved in the study, the sample types, and the time points considered is presented. Marked with the "RNA" label, the mother-infant pairs for which stool metatranscriptomes were produced are indicated.

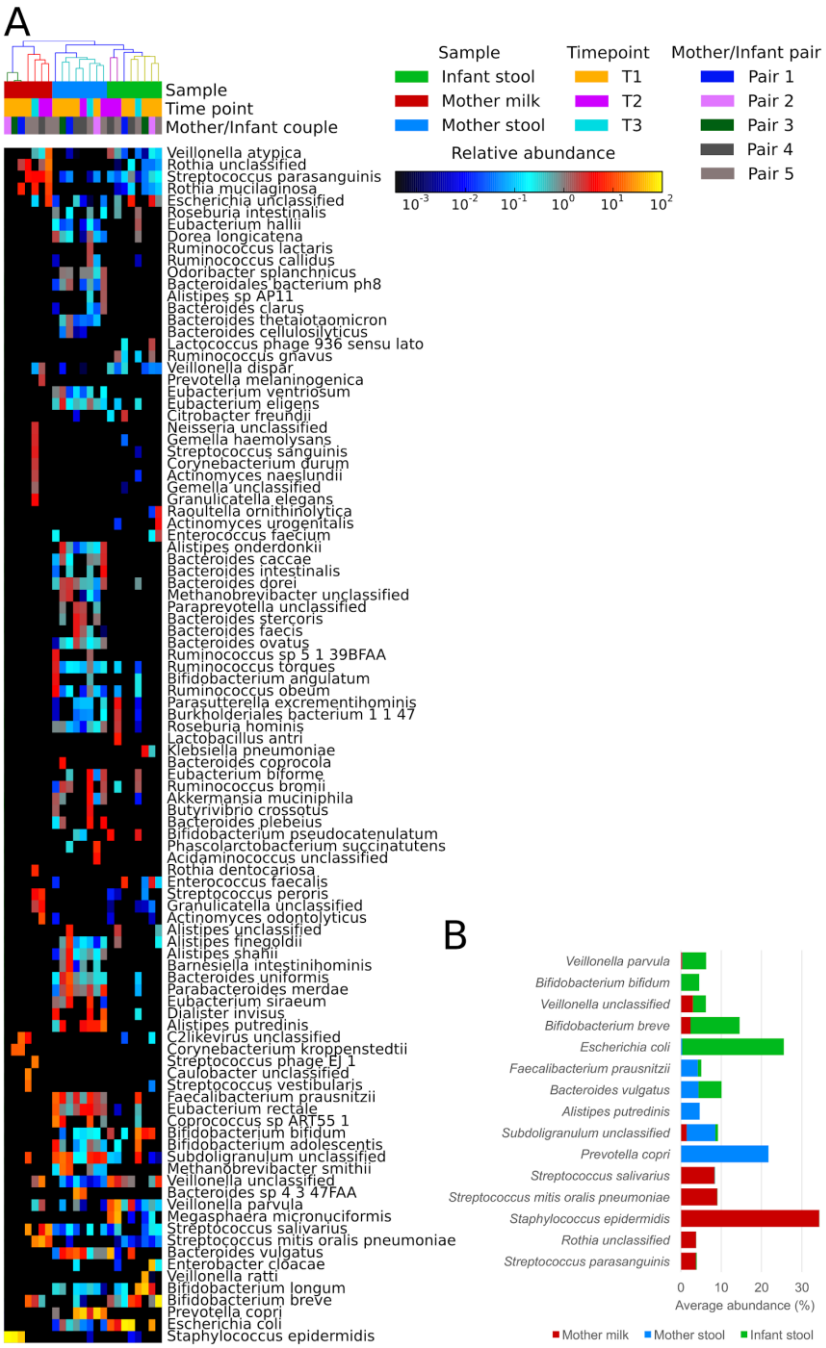



**Fig S2. Extensive taxonomic profiling of the top 100 species from MetaPhlAn2 analysis and the five most highly represented niche-specific species.** (A) The heat map shows differences in terms of species richness between mother, infant, and milk metagenomes. In particular, the milk samples have very 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.

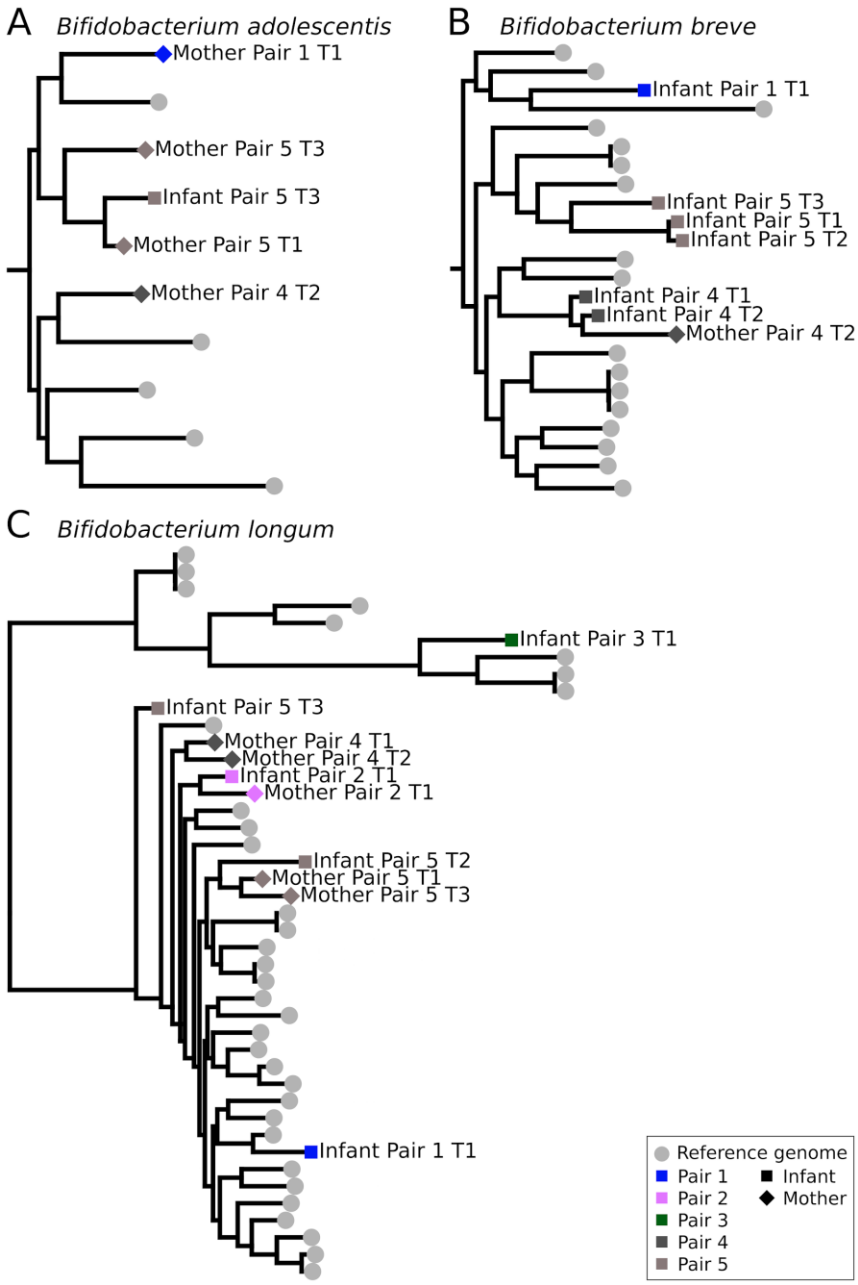



**Fig S3. Strain-level analysis showing vertical transmission from mother to infant of bifidobacterium species.** The phylogenetic trees were produced by applying StrainPhlAn for the following species: (A) *Bifidobacterium adolescentis*, (B) *Bifidobacterium breve*, and (C) *Bifidobacterium longum*. 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.

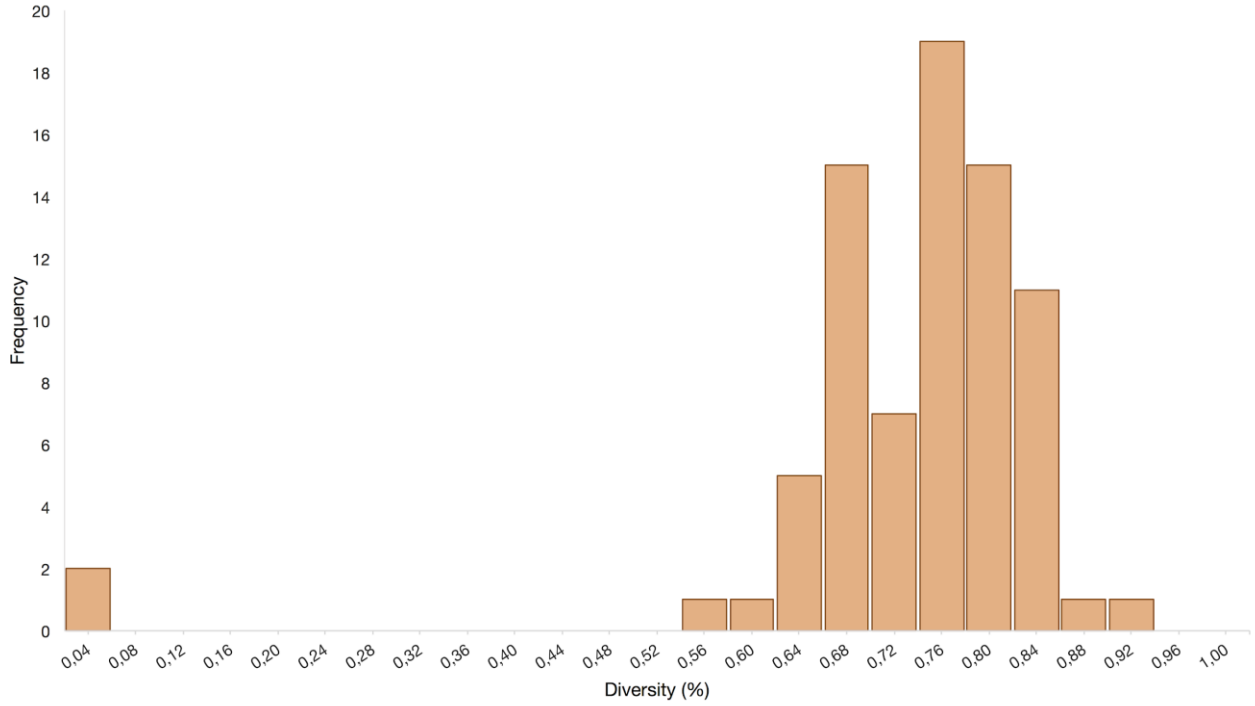



**Fig S4. Distribution of SNV rates of *Bifidobacterium bifidum*.** We computed the SNV rates of the strains of *B. bifidum* reconstructed with StrainPhlAn (the phylogenetic tree is presented in **Fig. 2A**). The two strains of the mother and the infant of pair 4 at time point 2 have an SNV rate of 0.04. 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.

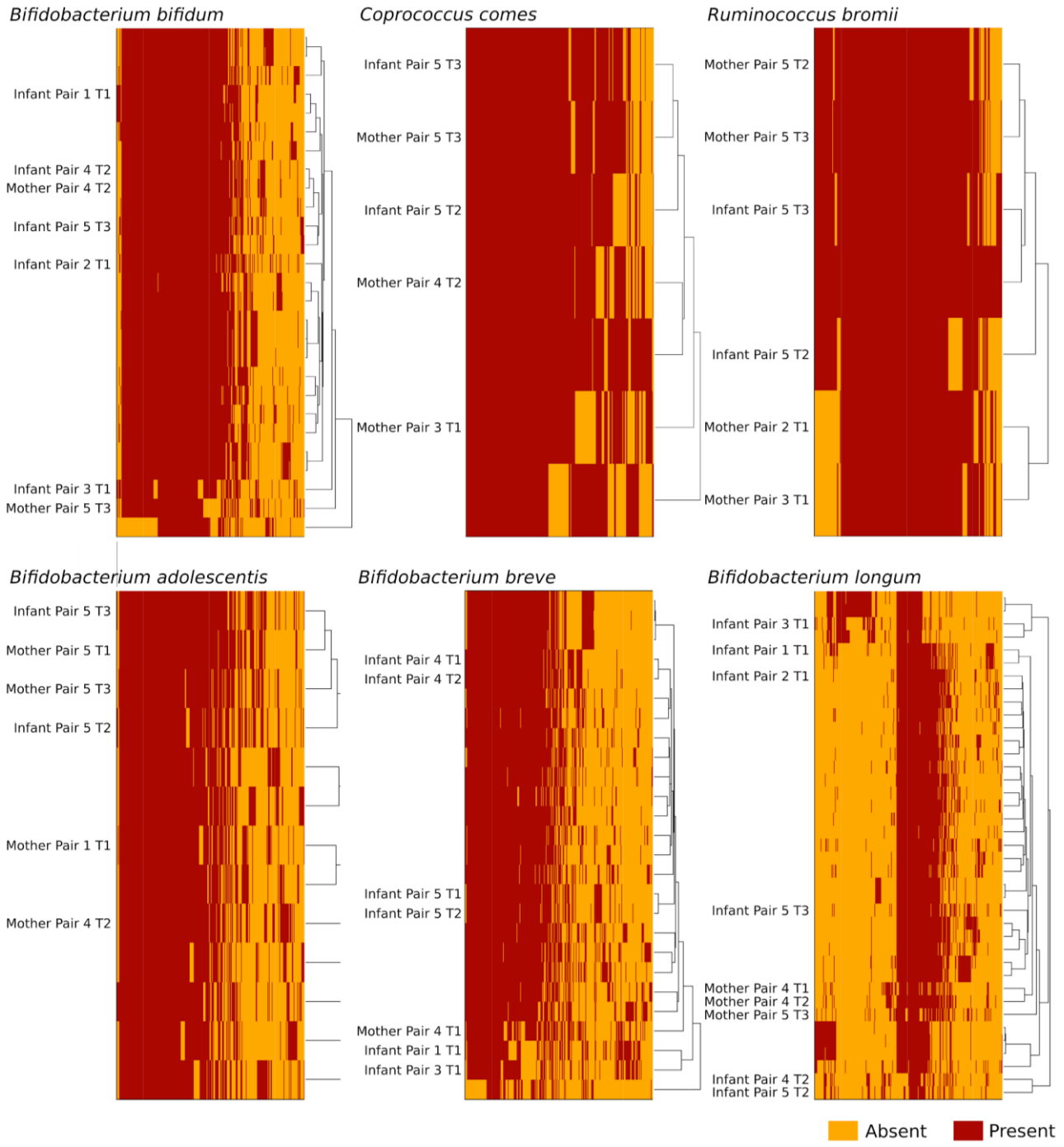



**Fig S5. Strain-level analysis by applying PanPhlAn confirms vertical transmission.** We applied PanPhlAn to validate the results obtained with StrainPhlAn (**Fig. 2** and **S3**). The pangenome-based strain-level analysis shows the presence and absence (in red and yellow, respectively) of the species-specific 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.

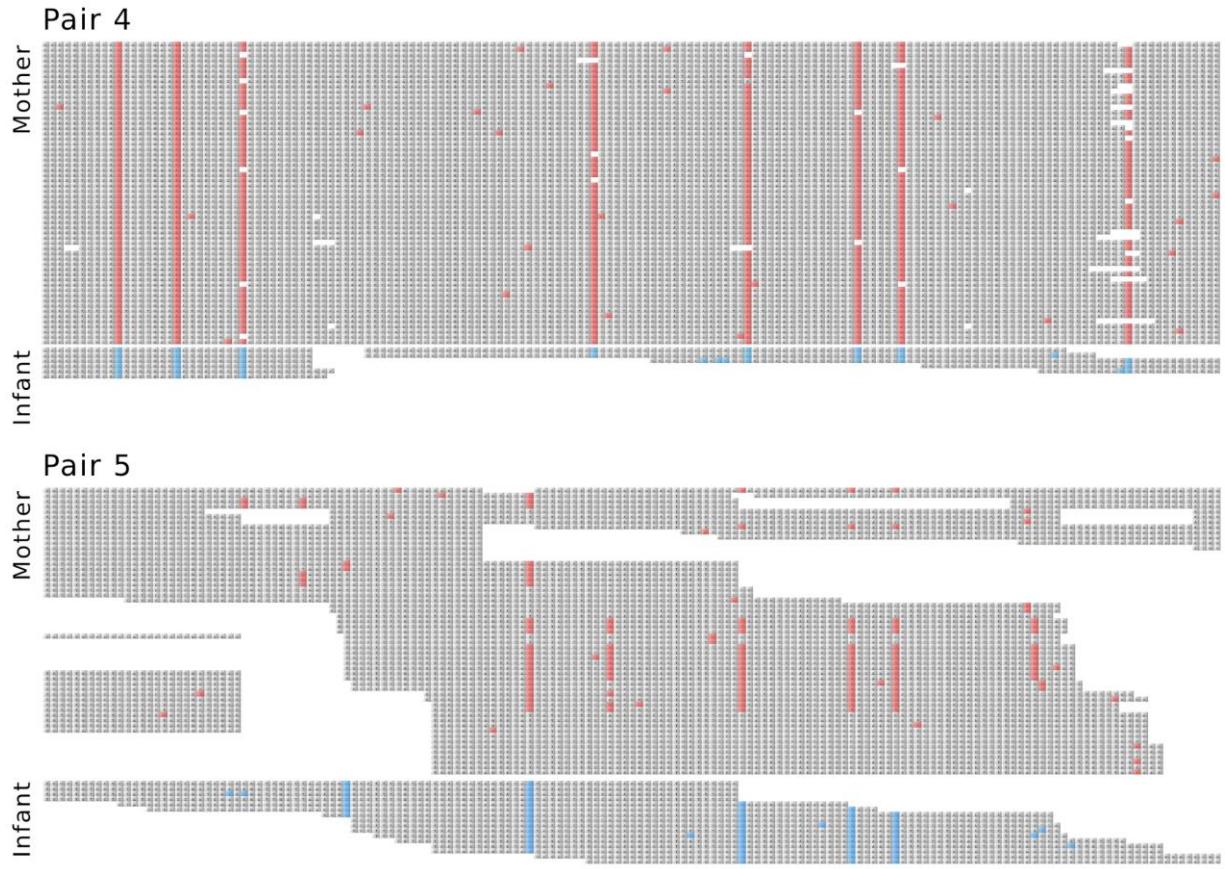



**Fig S6. Read alignment of pepper mild mottle virus (PMMoV) for both pair 4 and pair 5.** Alignments of mother and infant of both pair 4 and pair 5 against the PMMoV reference genome are presented, showing variations highlighted in red (mother) and blue (infant) for a window of 160 bp. Pair 4 data (from 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.

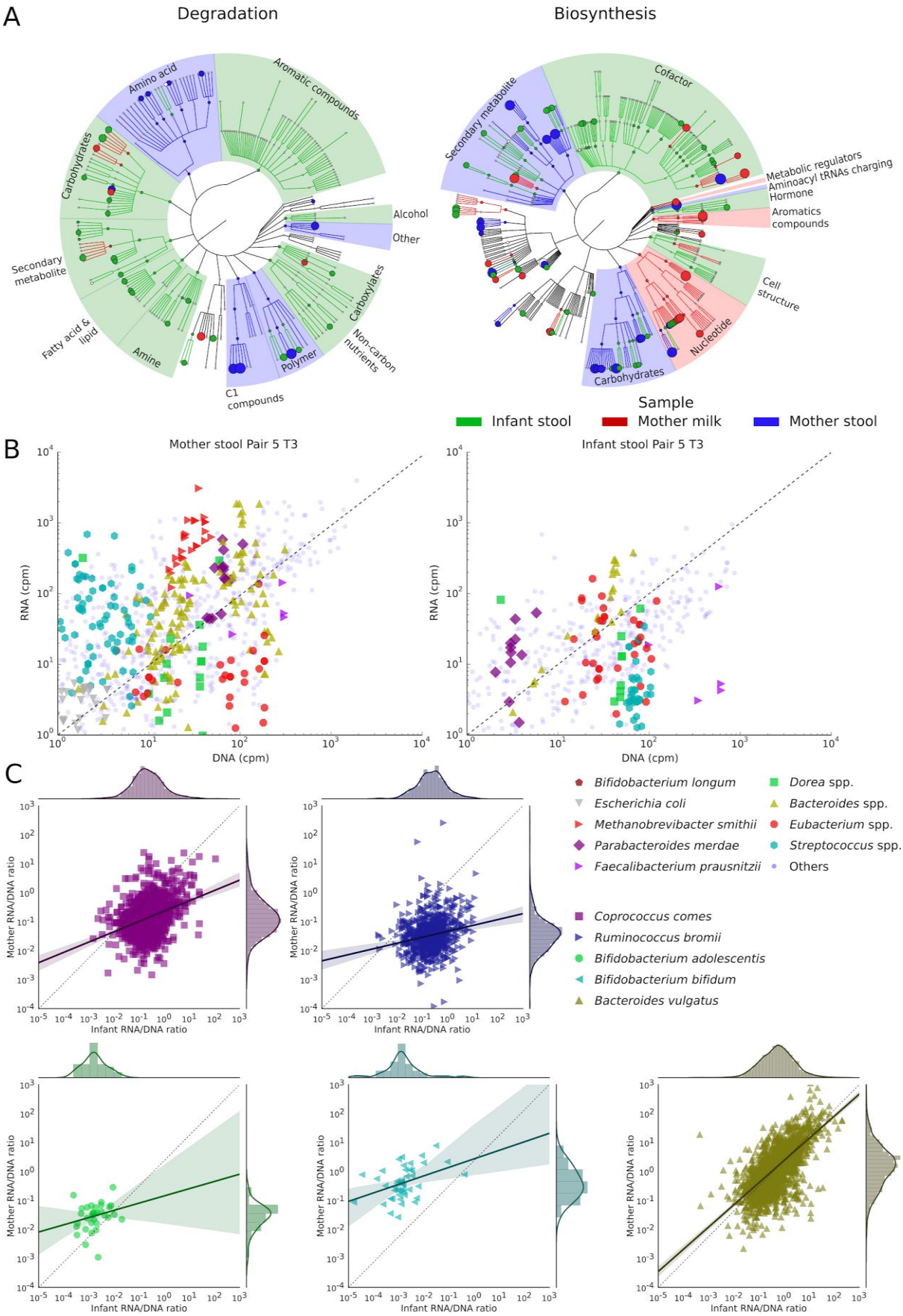

A
Degradation
Biosynthesis
Amino acid
Aromatic compounds
Carbohydrates
Alcohol
Other
Secondary metabolite
Carboxylates
Fatty acid & lipid
Amine
Non-carbon nutrients
Polymer
C1 compounds
Secondary metabolite
Cofactor
Metabolic regulators
Aminoacyl tRNAs charging
Hormone
Aromatics compounds
Cell structure
Nucleotide
Carbohydrates
Sample
Infant stool
Mother milk
Mother stool
B
Mother stool Pair 5 T3
Infant stool Pair 5 T3
RNA (cpm)
DNA (cpm)
C
Mother RNA/DNA ratio
Infant RNA/DNA ratio
Bifidobacterium longum
Escherichia coli
Methanobrevibacter smithii
Parabacteroides merdae
Faecalibacterium prausnitzii
Dorea spp.
Bacteroides spp.
Eubacterium spp.
Streptococcus spp.
Others
Coprococcus comes
Ruminococcus bromii
Bifidobacterium adolescentis
Bifidobacterium bifidum
Bacteroides vulgatus

**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.

## 3.8 Supplementary Tables

Captions of supplementary tables are reported below. 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.

## 3.9 References


Aagaard, K., K. Riehle, J. Ma, N. Segata, T. A. Mistretta, C. Coarfa, S. Raza, et al. 2012. “A Metagenomic Approach to Characterization of the Vaginal Microbiome Signature in Pregnancy.” *PloS One* 7 (6): e36466.

Aronesty, Erik. 2013. “Comparison of Sequencing Utility Programs.” *The Open Bioinformatics Journal* 7 (1).

Asnicar, Francesco, George Weingart, Timothy L. Tickle, Curtis Huttenhower, and Nicola Segata. 2015. “Compact Graphical Representation of Phylogenetic Data and Metadata with GraPhlAn.” Edited by Jaume Bacardit. *PeerJ* 3 (June): e1029.

Azad, Meghan B., Theodore Konya, Heather Maughan, David S. Guttman, Catherine J. Field, Radha S. Chari, Malcolm R. Sears, et al. 2013. “Gut Microbiota of Healthy Canadian Infants: Profiles by Mode of Delivery and Infant Diet at 4 Months.” *CMAJ: Canadian Medical Association Journal = Journal de l’Association Medicale Canadienne* 185 (5): 385–94.

Bäckhed, Fredrik, Ruth E. Ley, Justin L. Sonnenburg, Daniel A. Peterson, and Jeffrey I. Gordon. 2005. “Host-Bacterial Mutualism in the Human Intestine.” *Science* 307 (5717): 1915–20.

Bäckhed, Fredrik, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, et al. 2015. “Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life.” *Cell Host & Microbe* 17 (5): 690–703.

Balique, Fanny, Philippe Colson, Abdoulaye Oury Barry, Claude Nappez, Audrey Ferretti, Khatoun Al Moussawi, Tatsiana Ngounga, et al. 2013. “Tobacco Mosaic Virus in the Lungs of Mice Following Intra-Tracheal Inoculation.” *PloS One* 8 (1): e54993.

Bao, Guanhui, Mingjie Wang, Thomas G. Doak, and Yuzhen Ye. 2015. “Strand-Specific Community RNA-Seq Reveals Prevalent and Dynamic Antisense Transcription in Human Gut Microbiota.” *Frontiers in Microbiology* 6 (September): 896.

Biasucci, Giacomo, Monica Rubini, Sara Riboni, Lorenzo Morelli, Elena Bessi, and Cristiana Retetangos. 2010. “Mode of Delivery Affects the Bacterial Community in the Newborn Gut.” *Early Human Development* 86 Suppl 1 (1, Supplement): 13–15.

Bickley, J., J. K. Short, D. G. McDowell, and H. C. Parkes. 1996. “Polymerase Chain Reaction (PCR) Detection of Listeria Monocytogenes in Diluted Milk and Reversal of PCR Inhibition Caused by Calcium Ions.” *Letters in Applied Microbiology* 22 (2): 153–58.

Britton, Robert A., and Vincent B. Young. 2014. “Role of the Intestinal Microbiota in Resistance to Colonization by Clostridium Difficile.” *Gastroenterology* 146 (6): 1547–53.

Cabral, Damien J., Jenna I. Wurster, Myrto E. Flokas, Michail Alevizakos, Michelle Zabat, Benjamin J. Korry, Aislinn D. Rowan, et al. 2017. “The Salivary Microbiome Is Consistent between Subjects and Resistant to Impacts of Short-Term Hospitalization.” *Scientific Reports* 7 (1): 11040.

Cabrera-Rubio, Raul, M. Carmen Collado, Kirsi Laitinen, Seppo Salminen, Erika Isolauri, and Alex Mira. 2012. “The Human Milk Microbiome Changes over Lactation and Is Shaped by Maternal Weight and Mode of Delivery.” *The American Journal of Clinical Nutrition* 96 (3): 544–51.

Clemente, Jose C., Luke K. Ursell, Laura Wegener Parfrey, and Rob Knight. 2012. “The Impact of the Gut Microbiota on Human Health: An Integrative View.” *Cell* 148 (6): 1258–70.

Colson, Philippe, Hervé Richet, Christelle Desnues, Fanny Balique, Valérie Moal, Jean-Jacques Grob, Philippe Berbis, 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 (4): e10041.

Costello, Elizabeth K., Keaton Stagaman, Les Dethlefsen, Brendan J. M. Bohannan, and David A. Relman. 2012. “The Application of Ecological Theory toward an Understanding of the Human Microbiome.” *Science* 336 (6086): 1255–62.

Cremonesi, P., B. Castiglioni, G. Malferrari, I. Biunno, C. Vimercati, P. Moroni, S. Morandi, and M. Luzzana. 2006. “Technical Note: Improved Method for Rapid DNA Extraction of Mastitis Pathogens Directly from Milk.” *Journal of Dairy Science* 89 (1): 163–69.

Davenport, Emily R., Jon G. Sanders, Se Jin Song, Katherine R. Amato, Andrew G. Clark, and Rob Knight. 2017. “The Human Microbiome in Evolution.” *BMC Biology* 15 (1): 127.

Dominguez-Bello, Maria G., Elizabeth K. Costello, Monica Contreras, Magda Magris, Glida Hidalgo, Noah Fierer, and Rob Knight. 2010. “Delivery Mode Shapes the Acquisition and Structure of the Initial Microbiota across Multiple Body Habitats in Newborns.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (26): 11971–75.

Dominguez-Bello, Maria G., Kassandra M. De Jesus-Laboy, Nan Shen, Laura M. Cox, Amnon Amir, Antonio Gonzalez, Nicholas A. Bokulich, et al. 2016. “Partial Restoration of the Microbiota of Cesarean-Born Infants via Vaginal Microbial Transfer.” *Nature Medicine* 22 (3): 250–53.

Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, et al. 2014. “A Highly Abundant Bacteriophage Discovered in the Unknown Sequences of Human Faecal Metagenomes.” *Nature Communications* 5 (July): 4498.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97.

Faith, Jeremiah J., Janaki L. Guruge, Mark Charbonneau, Sathish Subramanian, Henning Seedorf, Andrew L. Goodman, Jose C. Clemente, et al. 2013. “The Long-Term Stability of the Human Gut Microbiota.” *Science* 341 (6141): 1237439.

Ferretti, Pamela, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Valentina Gorfer, Sabina Fedi, Federica Armanini, et al. 2018. “Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome.” *Cell Host & Microbe* 24 (1): 133–45.e5.

Flint, Harry J., Karen P. Scott, Sylvia H. Duncan, Petra Louis, and Evelyne Forano. 2012. “Microbial Degradation of Complex Carbohydrates in the Gut.” *Gut Microbes* 3 (4): 289–306.

Flores, Gilberto E., J. Gregory Caporaso, Jessica B. Henley, Jai Ram Rideout, Daniel Domogala, John Chase, Jonathan W. Leff, et al. 2014. “Temporal Variability Is a Personalized Feature of the Human Microbiome.” *Genome Biology* 15 (12): 531.

Franzosa, Eric A., Xochitl C. Morgan, Nicola Segata, Levi Waldron, Joshua Reyes, Ashlee M. Earl, Georgia Giannoukos, et al. 2014. “Relating the Metatranscriptome and Metagenome of the Human Gut.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (22): E2329–38.

Fuentes, Susana, Els van Nood, Sebastian Tims, Ineke Heikamp-de Jong, Cajo J. F. ter Braak, Josbert J. Keller, Erwin G. Zoetendal, and Willem M. de Vos. 2014. “Reset of a Critically

Disturbed Microbial Ecosystem: Faecal Transplant in Recurrent Clostridium Difficile Infection." *The ISME Journal* 8 (8): 1621–33.

Gardy, Jennifer L., James C. Johnston, Shannan J. Ho Sui, Victoria J. Cook, Lena Shah, Elizabeth Brodkin, Shirley Rempel, et al. 2011. "Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak." *The New England Journal of Medicine* 364 (8): 730–39.

Giannoukos, Georgia, Dawn M. Ciulla, Katherine Huang, Brian J. Haas, Jacques Izard, Joshua Z. Levin, Jonathan Livny, et al. 2012. "Efficient and Robust RNA-Seq Process for Cultured Bacteria and Complex Community Transcriptomes." *Genome Biology* 13 (3): R23.

Gosalbes, M. J., J. J. Abellan, A. Durbán, A. E. Pérez-Cobas, A. Latorre, and A. Moya. 2012. "Metagenomics of Human Microbiome: Beyond 16s rDNA." *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 18 Suppl 4 (July): 47–49.

Greenwood, Corryn, Ardythe L. Morrow, Anne J. Lagomarcino, Mekibib Altaye, Diana H. Taft, Zhuoteng Yu, David S. Newburg, Doyle V. Ward, and Kurt R. Schibler. 2014. "Early Empiric Antibiotic Use in Preterm Infants Is Associated with Lower Bacterial Diversity and Higher Relative Abundance of Enterobacter." *The Journal of Pediatrics* 165 (1): 23–29.

HMP, Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, et al. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (June): 207.

Houghteling, Pearl D., and W. Allan Walker. 2015. "Why Is Initial Bacterial Colonization of the Intestine Important to Infants' and Children's Health?" *Journal of Pediatric Gastroenterology and Nutrition* 60 (3): 294–307.

Hunt, Katherine M., James A. Foster, Larry J. Forney, Ursel M. E. Schütte, Daniel L. Beck, Zaid Abdo, Lawrence K. Fox, Janet E. Williams, Michelle K. McGuire, and Mark A. McGuire. 2011. "Characterization of the Diversity and Temporal Stability of Bacterial Communities in Human Milk." *PloS One* 6 (6): e21313.

Jeurink, P. V., J. van Bergenhenegouwen, E. Jiménez, L. M. J. Knippels, L. Fernández, J. Garssen, J. Knol, J. M. Rodríguez, and R. Martín. 2013. "Human Milk: A Source of More Life than We Imagine." *Beneficial Microbes* 4 (1): 17–30.

Jost, Ted, Christophe Lacroix, Christian P. Braegger, Florence Rochat, and Christophe Chassard. 2014. "Vertical Mother-Neonate Transfer of Maternal Gut Bacteria via Breastfeeding." *Environmental Microbiology* 16 (9): 2891–2904.

Khoruts, Alexander, Johan Dicksved, Janet K. Jansson, and Michael J. Sadowsky. 2010. "Changes in the Composition of the Human Fecal Microbiome after Bacteriotherapy for Recurrent Clostridium Difficile-Associated Diarrhea." *Journal of Clinical Gastroenterology* 44 (5): 354–60.

Koenig, Jeremy E., Aymé Spor, Nicholas Scalfone, Ashwana D. Fricker, Jesse Stombaugh, Rob Knight, Largus T. Angenent, and Ruth E. Ley. 2011. "Succession of Microbial Consortia in the Developing Infant Gut Microbiome." *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl 1 (Supplement 1): 4578–85.

Korpela, Katri, and Willem M. de Vos. 2018. "Early Life Colonization of the Human Gut: Microbes Matter Everywhere." *Current Opinion in Microbiology* 44 (August): 70–78.

Kurokawa, Ken, Takehiko Itoh, Tomomi Kuwahara, Kenshiro Oshima, Hidehiro Toh, Atsushi

Toyoda, Hideto Takami, et al. 2007. “Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes.” *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 14 (4): 169–81.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

La Rosa, Patricio S., Barbara B. Warner, Yanjiao Zhou, George M. Weinstock, Erica Sodergren, Carla M. Hall-Moore, Harold J. Stevens, et al. 2014. “Patterned Progression of Bacterial Populations in the Premature Infant Gut.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (34): 12522–27.

LeBlanc, Jean Guy, Christian Milani, Graciela Savoy de Giori, Fernando Sesma, Douwe van Sinderen, and Marco Ventura. 2013. “Bacteria as Vitamin Suppliers to Their Host: A Gut Microbiota Perspective.” *Current Opinion in Biotechnology* 24 (2): 160–68.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Li, Simone S., Ana Zhu, Vladimir Benes, Paul I. Costea, Rajna Hercog, Falk Hildebrand, Jaime Huerta-Cepas, et al. 2016. “Durable Coexistence of Donor and Recipient Strains after Fecal Microbiota Transplantation.” *Science* 352 (6285): 586–89.

Loman, Nicholas J., Chrystala Constantinidou, Martin Christner, Holger Rohde, Jacqueline Z-M Chan, Joshua Quick, Jacqueline C. Weir, et al. 2013. “A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia Coli O104:H4.” *JAMA: The Journal of the American Medical Association* 309 (14): 1502–10.

Lozupone, Catherine A., Jesse I. Stombaugh, Jeffrey I. Gordon, Janet K. Jansson, and Rob Knight. 2012. “Diversity, Stability and Resilience of the Human Gut Microbiota.” *Nature* 489 (7415): 220–30.

Makino, Hiroshi, Akira Kushiro, Eiji Ishikawa, Delphine Muylaert, Hiroyuki Kubota, Takafumi Sakai, Kenji Oishi, 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.” *Applied and Environmental Microbiology* 77 (19): 6788–93.

Marcobal, Angela, Mariana Barboza, Erica D. Sonnenburg, Nicholas Pudlo, Eric C. Martens, Prerak Desai, Carlito B. Lebrilla, et al. 2011. “Bacteroides in the Infant Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways.” *Cell Host & Microbe* 10 (5): 507–14.

Maurice, Corinne Ferrier, Henry Joseph Haiser, and Peter James Turnbaugh. 2013. “Xenobiotics Shape the Physiology and Gene Expression of the Active Human Gut Microbiome.” *Cell* 152 (1-2): 39–50.

Medeiros, Ricardo B. de, Juliana Figueiredo, Renato de O. Resende, and Antonio C. De Avila. 2005. “Expression of a Viral Polymerase-Bound Host Factor Turns Human Cell Lines Permissive to a Plant- and Insect-Infecting Virus.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (4): 1175–80.

Milani, Christian, Leonardo Mancabelli, Gabriele Andrea Lugli, Sabrina Duranti, Francesca Turroni, Chiara Ferrario, Marta Mangifesta, et al. 2015. “Exploring Vertical Transmission of Bifidobacteria from Mother to Child.” *Applied and Environmental Microbiology* 81 (20): 7078–

87.

Miyoshi, Jun, Alexandria M. Bobe, Sawako Miyoshi, Yong Huang, Nathaniel Hubert, Tom O. Delmont, A. Murat Eren, Vanessa Leone, and Eugene B. Chang. 2017. “Peripartum Antibiotics Promote Gut Dysbiosis, Loss of Immune Tolerance, and Inflammatory Bowel Disease in Genetically Prone Offspring.” *Cell Reports* 20 (2): 491–504.

Morowitz, M. J., V. J. Denef, E. K. Costello, B. C. Thomas, V. Poroyko, D. A. Relman, and J. F. Banfield. 2011. “Strain-Resolved Community Genomic Analysis of Gut Microbial Colonization in a Premature Infant.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (3): 1128–33.

Ogilvie, Lesley A., and Brian V. Jones. 2015. “The Human Gut Virome: A Multifaceted Majority.” *Frontiers in Microbiology* 6 (September): 918.

Palmer, Chana, Elisabeth M. Bik, Daniel B. DiGiulio, David A. Relman, and Patrick O. Brown. 2007. “Development of the Human Infant Intestinal Microbiota.” *PLoS Biology* 5 (7): e177.

Palm, Noah W., Marcel R. de Zoete, and Richard A. Flavell. 2015. “Immune-Microbiota Interactions in Health and Disease.” *Clinical Immunology* 159 (2): 122–27.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–30.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464 (7285): 59–65.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools.” *Nucleic Acids Research* 41 (Database issue): D590–96.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.

Ramsay, Donna T., Jacqueline C. Kent, Robyn A. Owens, and Peter E. Hartmann. 2004. “Ultrasound Imaging of Milk Ejection in the Breast of Lactating Women.” *Pediatrics* 113 (2): 361–67.

Reyes, Alejandro, Matthew Haynes, Nicole Hanson, Florent E. Angly, Andrew C. Heath, Forest Rohwer, and Jeffrey I. Gordon. 2010. “Viruses in the Faecal Microbiota of Monozygotic Twins and Their Mothers.” *Nature* 466 (7304): 334–38.

Reyes, Alejandro, Nicholas P. Semenkovich, Katrine Whiteson, Forest Rohwer, and Jeffrey I. Gordon. 2012. “Going Viral: Next-Generation Sequencing Applied to Phage Populations in the Human Gut.” *Nature Reviews. Microbiology* 10 (9): 607–17.

Schloissnig, Siegfried, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, et al. 2013. “Genomic Variation Landscape of the Human Gut Microbiome.” *Nature* 493 (7430): 45–50.

Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. 2016. “Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics.” *Nature*

*Methods* 13 (5): 435–38.

Schrader, C., A. Schielke, L. Ellerbroek, and R. Johne. 2012. “PCR Inhibitors - Occurrence, Properties and Removal.” *Journal of Applied Microbiology* 113 (5): 1014–26.

Scott, Karen P., Jean-Michel Antoine, Tore Midtvedt, and Saskia van Hemert. 2015. “Manipulating the Gut Microbiota to Maintain Health and Treat Disease.” *Microbial Ecology in Health and Disease* 26 (February): 25877.

Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S. Garrett, and Curtis Huttenhower. 2011. “Metagenomic Biomarker Discovery and Explanation.” *Genome Biology* 12 (6): R60.

Sharon, Itai, Michael J. Morowitz, Brian C. Thomas, Elizabeth K. Costello, David A. Relman, and Jillian F. Banfield. 2013. “Time Series Community Genomics Analysis Reveals Rapid Shifts in Bacterial Species, Strains, and Phage during Infant Gut Colonization.” *Genome Research* 23 (1): 111–20.

Shin, Hakdong, Zhiheng Pei, Keith A. Martinez 2nd, Juana I. Rivera-Vinas, Keimari Mendez, Humberto Cavallin, and Maria G. Dominguez-Bello. 2015. “The First Microbial Environment of Infants Born by C-Section: The Operating Room Microbes.” *Microbiome* 3 (1): 59.

Song, Se Jin, Christian Lauber, Elizabeth K. Costello, Catherine A. Lozupone, Gregory Humphrey, Donna Berg-Lyons, J. Gregory Caporaso, et al. 2013. “Cohabiting Family Members Share Microbiota with One Another and with Their Dogs.” Edited by Detlef Weigel. *eLife* 2 (April): e00458.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.

Stecher, Bärbel, and Wolf-Dietrich Hardt. 2011. “Mechanisms Controlling Pathogen Colonization of the Gut.” *Current Opinion in Microbiology* 14 (1): 82–91.

Tailford, Louise E., Emmanuelle H. Crost, Devon Kavanaugh, and Nathalie Juge. 2015. “Mucin Glycan Foraging in the Human Gut Microbiome.” *Frontiers in Genetics* 6 (March): 81.

Tamburini, Sabrina, Nan Shen, Han Chih Wu, and Jose C. Clemente. 2016. “The Microbiome in Early Life: Implications for Health Outcomes.” *Nature Medicine* 22 (7): 713–22.

Thurber, Rebecca V., Matthew Haynes, Mya Breitbart, Linda Wegley, and Forest Rohwer. 2009. “Laboratory Procedures to Generate Viral Metagenomes.” *Nature Protocols* 4 (4): 470–83.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. “MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling.” *Nature Methods* 12 (10): 902–3.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. “Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes.” *Genome Research* 27 (4): 626–38.

Turnbaugh, Peter J., Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, et al. 2009. “A Core Gut Microbiome in Obese and Lean Twins.” *Nature* 457 (7228): 480–84.

Turnbaugh, Peter J., Christopher Quince, Jeremiah J. Faith, Alice C. McHardy, Tanya Yatsunenko, Faheem Niazi, Jason Affourtit, et al. 2010. “Organismal, Genetic, and Transcriptional Variation in the Deeply Sequenced Gut Microbiomes of Identical Twins.” *Proceedings of the National*

*Academy of Sciences of the United States of America* 107 (16): 7503–8.

Turroni, Francesca, Elena Foroni, Fausta Serafini, Alice Viappiani, Barbara Montanini, Francesca Bottacini, Alberto Ferrarini, et al. 2011. “Ability of Bifidobacterium Breve to Grow on Different Types of Milk: Exploring the Metabolism of Milk through Genome Analysis.” *Applied and Environmental Microbiology* 77 (20): 7408–17.

Turroni, Francesca, Christian Milani, Douwe van Sinderen, and Marco Ventura. 2011. “Genetic Strategies for Mucin Metabolism in Bifidobacterium Bifidum PRL2010: An Example of Possible Human-Microbe Co-Evolution.” *Gut Microbes* 2 (3): 183–89.

Turroni, Francesca, Clelia Peano, Daniel A. Pass, Elena Foroni, Marco Severgnini, Marcus J. Claesson, Colm Kerr, et al. 2012. “Diversity of Bifidobacteria within the Infant Gut Microbiota.” *PloS One* 7 (5): e36957.

Vatanen, Tommi, Damian R. Plichta, Juhi Somani, Philipp C. Münch, Timothy D. Arthur, Andrew Brantley Hall, Sabine Rudolf, et al. 2018. “Genomic Variation and Strain-Specific Functional Adaptation in the Human Gut Microbiome during Early Life.” *Nature Microbiology*, December. https://doi.org/10.1038/s41564-018-0321-5.

Victoria, Joseph G., Amit Kapoor, Linlin Li, Olga Blinkova, Beth Slikas, Chunlin Wang, Asif Naeem, Sohail Zaidi, and Eric Delwart. 2009. “Metagenomic Analyses of Viruses in Stool Samples from Children with Acute Flaccid Paralysis.” *Journal of Virology* 83 (9): 4642–51.

Walker, Alan W., Jennifer Ince, Sylvia H. Duncan, Lucy M. Webster, Grietje Holtrop, Xiaolei Ze, David Brown, et al. 2011. “Dominant and Diet-Responsive Groups of Bacteria within the Human Colonic Microbiota.” *The ISME Journal* 5 (2): 220–30.

Wampach, Linda, Anna Heintz-Buschart, Joëlle V. Fritz, Javier Ramiro-Garcia, Janine Habier, Malte Herold, Shaman Narayanasamy, et al. 2018. “Birth Mode Is Associated with Earliest Strain-Conferred Gut Microbiome Functions and Immunostimulatory Potential.” *Nature Communications* 9 (1): 5091.

Wampach, Linda, Anna Heintz-Buschart, Angela Hogan, Emilie E. L. Muller, Shaman Narayanasamy, Cedric C. Laczny, Luisa W. Hugerth, et al. 2017. “Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life.” *Frontiers in Microbiology* 8 (May): 434.

Ward, Tonya L., Sergey Hosid, Ilya Ioshikhes, and Illimar Altosaar. 2013. “Human Milk Metagenome: A Functional Capacity Analysis.” *BMC Microbiology* 13 (1): 116.

Ximenez, Cecilia, and Javier Torres. 2017. “Development of Microbiota in Infants and Its Role in Maturation of Gut Mucosa and Immune System.” *Archives of Medical Research* 48 (8): 666–80.

Yassour, Moran, Eeva Jason, Larson J. Hogstrom, Timothy D. Arthur, Surya Tripathi, Heli Siljander, Jenni Selvenius, et al. 2018. “Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life.” *Cell Host & Microbe* 24 (1): 146–54.e4.

Yatsunenko, Tanya, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, et al. 2012. “Human Gut Microbiome Viewed across Age and Geography.” *Nature* 486 (7402): 222–27.

Zhang, Tao, Mya Breitbart, Wah Heng Lee, Jin-Quan Run, Chia Lin Wei, Shirlena Wee Ling Soh, Martin L. Hibberd, Edison T. Liu, Forest Rohwer, and Yijun Ruan. 2006. “RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses.” *PLoS Biology* 4 (1)

# Chapter 4. Microbial genomes from gut metagenomes of non-human primates expand the primate-associated bacterial tree-of-life with over 1,000 novel species

## 4.1 Introduction to the chapter

In **Chapter 3** I introduced a cultivation-free metagenomic framework to track microbes at the strain-level across samples, that targets microorganisms from virtually all known species. However, this approach cannot survey the so-called “microbial dark matter” present in metagenomic samples, that is defined as the fraction of microbes in a microbiome that cannot be identified because the species they belong to lack representative genomes and were never characterized nor cataloged before. In this Chapter, I use an assembly-based metagenomic approach we recently proposed ((Pasolli et al. 2019), partially reported in **Chapter 5.2**) for the discovery and characterization of unknown microbial species in the gut microbiome of non-human primates. I will show how the previously unexplored microbial diversity in the gut microbiome of non-human primates can be used for strain-level comparative genomic analysis also in the context of the microbes in the human gut. In this work, we aimed at reconstructing genomes from metagenomes of these under-investigated non-human hosts as a first necessary step to expand our understanding of the primate microbiome and its co-evolution within primates.

A full understanding of the human microbiome cannot be reached without clarifying first its patterns and trajectories of co-evolution with humans and other primates. Some lines of research are directly investigating ancient microbiome samples to reconstruct the microbiome evolutionary history, but technical limitations due to ancient DNA degradation make this a challenging operation (Cano et al. 2000; Raúl Y. Tito et al. 2008; Raul Y. Tito et al. 2012; Rasmussen et al. 2015; Maixner et al. 2016; Sonnenburg and Sonnenburg 2019). Moreover, obtaining ancient gut metagenomes from more than few thousand years ago seems completely infeasible (Maixner et al. 2016). Recently proposed alternatives to ancient microbiome sampling include the study of co-evolving microbes through comparative genomic analysis of humans and non-human primates (Amato 2019), our closest evolutionary relatives. A few studies surveyed the composition of microbiomes associated with NHPs and their overlap with the human one, but they were either hindered by the low-resolution method applied (Yildirim et al. 2010; Ochman et al. 2010; Degnan et al. 2012; Moeller et al. 2013; Gomez et al. 2016; Moeller et al. 2016; Hicks et al. 2018; Cabana et al. 2019; Greene et al. 2019; Moeller et al. 2012, 2014; Clayton et al. 2016) or by the limitations of reference-based approaches that allow the identification only of already characterized species (Tung et al. 2015; Srivathsan et al. 2015; Hicks et al. 2018; X. Li et al. 2018; Orkin et al. 2019; Amato et al. 2018). Indeed, the large majority of reads in NHP metagenomes cannot map against any known species for which reference genomes are available, thus hindering a complete overview of the NHP microbiome. However, recent advances in metagenomics now provide the basis for *de novo* assembly

of microbial genomes from metagenomes, hence providing the possibility to expand the catalog of species associated with the NHP microbiome.

In the unpublished and currently under submission article reported in this chapter, we reconstructed a large number of previously unknown microbial taxa associated with non-human primate (NHP) microbiomes by applying metagenomic assembly tools to reconstruct microbial genomes from publicly available NHP metagenomic data. These newly-reconstructed genomes greatly expand our understanding of the microbial diversity associated with NHPs and increased the mappability of NHP metagenomes by over 600% with respect to the sole collection of reference genomes available in NCBI. We identified over 1,000 new species, 760 new genera, and 265 new families, showing that almost 90% of the microbial diversity has been overlooked in previous studies. A meta-analysis of this expanded catalog of NHP-associated microbes in the context of a large-scale human microbiome assembly effort (Pasolli et al. 2019) showed that species overlap between human and non-human microbiomes is scarce. Captive NHPs exposed to human environment and diet represent an exception, and show microbial signatures more similar to the human ones, with a larger fraction of known species that is reflected in a much higher metagenome mappability (average 61.8% w.r.t. 39.2% of datasets of wild NHPs). Overall, species sharing occurs mostly between NHPs and populations with a non-Westernized lifestyle, and mainly consists of uncharacterized clades only recently discovered in human metagenomes, thus supporting the loss of lifestyle-dependent microbiome members.

**Outlook.** Our work greatly expanded the number of microbial species identified in NHP microbiomes and the mappability of related metagenomes, thus enabling more reliable and comprehensive comparative genomic analysis with humans to expose host-microbiome co-evolution patterns. Although a considerable fraction of NHP metagenomes remains unmapped, this study posed the basis for further assembly-based efforts to retrieve more and more unknown taxa from these under-investigated hosts as soon as new metagenomic data will be available. Altogether, this would enable a more comprehensive investigation of the NHP microbiomes and possibly further studies on co-evolution of microbial communities with their primate hosts.

**Contribution.** For this article, I performed most of the computational analysis, data interpretation, and writing of the manuscript. The automatic pipeline for genome reconstruction from metagenomes and part of the statistical analysis (the analysis on functional profiles) were performed by co-authors.

This chapter reports the following unpublished article:

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

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

^ these authors contributed equally

*In submission*

## 4.2 Abstract

Humans have coevolved with their microbial communities to establish a mutually advantageous relation that is still poorly characterized and should be studied for a more comprehensive understanding of the human microbiome. Comparative (meta)genomic analysis of human and non-human primate (NHP) microbiomes offers a promising approach to study this symbiosis. However, despite few NHP metagenomic investigations, only very few species in NHP microbiomes can be characterized due to their poor representation in the available cataloged microbial diversity, thus limiting the potentialities of such comparative approaches. In this study, we reconstructed >1,000 previously uncharacterized microbial species from six available NHP metagenomic cohorts, resulting in an increase of the mappable fraction of metagenomic reads by 600%. These novel species highlight that almost 90% of the microbial diversity associated with NHPs has been previously overlooked. Comparative analysis of this new catalog of NHP taxa with the collection of >150,000 genomes from human metagenomes pointed at a very limited species-level overlap between human and NHP microbiomes (~10%). This overlap occurs mainly between NHPs and non-Westernized human populations and for NHPs living in captivity, suggesting that host lifestyle plays a role comparable to that of host speciation in shaping the intestinal microbiome. Several NHP-specific candidate species are phylogenetically related to human-associated microbes (e.g. Elusimicrobia) and could be the consequence of host-dependent evolutionary trajectories. The newly reconstructed species greatly expand the microbial diversity associated with NHPs, thus enabling better interrogation of the primate microbiome and empowering in-depth human and non-human comparative and coevolution studies.

## 4.3 Introduction

The human microbiome is a complex ecosystem, consisting of diverse microbial communities that have important functions in host physiology and metabolism (Sommer and Bäckhed 2013). The gut microbiome is influenced by several factors including diet (David et al. 2014), physical activity (Bressa et al. 2017), use of antibiotics (Langdon, Crook, and Dantas 2016) and other lifestyle-related aspects. Studies comparing the microbiome of rural and industrialized communities have also shown that dietary and lifestyle changes linked to Westernization have played a pivotal role in the loss of many

microbial taxa and in the rise of others (Segata 2015; Brito et al. 2016; Obregon-Tito et al. 2015; Rampelli et al. 2015; Smits et al. 2017; Liu et al. 2016; Pasolli et al. 2019). Although it is difficult to establish causality and mechanisms for these links (Blaser 2017; Hold 2014), recent studies have extended the identifiable members of the human microbiome to now cover >90% of its overall diversity (Pasolli et al. 2019), which is a prerequisite for advancing the understanding of the role of microbes in human physiology and metabolism.

A comprehensive understanding of the current structure of the human microbiome needs to consider the study of how the microbiome has coevolved with humans. Ancient intestinal microbiome samples (i.e. coprolites) can give some insights on the gut microbial composition of pre-industrialized and prehistoric humans and date back to a few thousand years (Cano et al. 2000; Raúl Y. Tito et al. 2008; Raul Y. Tito et al. 2012; Rasmussen et al. 2015; Maixner et al. 2016), but the time-dependent degradation issues of microbial DNA limits the possibility of profiling more ancient samples (Sonnenburg and Sonnenburg 2019). Some patterns of coevolution between humans and their microbiomes can however be in principle investigated by comparative and phylogenetic analysis of genomes and metagenomes in non-human primates (NHPs), the closest evolutionary relatives of humans (Amato 2019). However, a very substantial fraction of the microbiome in NHPs is currently uncharacterized and a comprehensive comparative sequence-level analysis against human microbiomes is thus unfeasible.

Recent studies of NHPs uncovered part of their hidden microbial diversity but only very partially contributed to the extension of the genetic blueprint of the microbiome in these hosts. Several 16S rRNA amplicon sequencing studies investigated the microbiome composition of NHPs (Yildirim et al. 2010; Ochman et al. 2010; Degnan et al. 2012; Moeller et al. 2013; Gomez et al. 2016; Moeller et al. 2016; Hicks et al. 2018; Cabana et al. 2019; Greene et al. 2019), and some, including a meta-analysis (Nishida and Ochman 2019), investigated the overlap and specificity of microbial communities associated with humans and NHPs (Moeller et al. 2012, 2014; Clayton et al. 2016). Yet, because this approach is low-resolution and lacks functional characterization, many coevolution aspects cannot be studied. Some studies have also applied shotgun metagenomics on NHP microbiomes (Tung et al. 2015; Srivathsan et al. 2015; Hicks et al. 2018; X. Li et al. 2018; Orkin et al. 2019; Amato et al. 2018), but all of them have applied a reference-based computational profiling approach, which solely allows the identification of the very few known microbial species present in NHPs, disregarding those that have not been characterized yet. However, because of the advances in metagenomic assembly (D. Li et al. 2015; Nurk et al. 2017) and its application on large cohorts (Pasolli et al. 2019), there is now the possibility to compile a more complete catalog of species and genomes in NHP microbiomes and thus enable accurate coevolution and comparative analyses.

In this study, we meta-analyzed 203 available shotgun-sequenced NHPs metagenomes and performed a large-scale assembly-based analysis uncovering over 1,000 yet-to-be-described species associated with NHP hosts, improving NHP gut metagenomes mappability by over 600%. We moreover compared the newly established catalog of NHP-

associated species in the context of a large-scale human microbiome assembly project (Pasolli et al. 2019) to expose the overlap and divergence between the NHP and human gut microbiome, showing that captive NHPs harbor microbial compositions and even strains more similar to the human ones and that the extent of microbiome overlap is strongly lifestyle-dependent. Through comparative microbiome analysis we reconstruct the loss of biodiversity from wild and captive NHP to non-Westernized and Westernized human populations.

## 4.4 Results and discussion

To investigate the extent to which the composition of the gut microbiome overlaps across different primates for both known and currently uncharacterized microbes, we meta-analyzed a large set of gut microbiomes from humans and non-human primates (NHPs) that are publically available. Six datasets were available for NHPs (Tung et al. 2015; Srivathsan et al. 2015; X. Li et al. 2018; Hicks et al. 2018; Orkin et al. 2019; Amato et al. 2018) spanning 22 host species from 14 different countries in five continents (**Supplementary Table 1** and **Supplementary Figure 1**), totaling 203 metagenomic samples that we retrieved and curated for this work. Microbiome samples from adult human healthy individuals were retrieved from 47 datasets considered in a recent meta-analysis (Pasolli et al. 2019) on 9,428 human gut metagenomes and used as comparative resource. Human samples include both Westernized and non-Westernized populations from different countries, whereas NHPs datasets cover four primate clades, including Old and New World monkeys, apes and lemurs (**Supplementary Table 1**, **Figure 1A**). Two datasets (LiX_2018 and SrivathsanA_2015) surveyed NHPs in captivity, which were fed a specific human-like diet (X. Li et al. 2018) or a diet similar to the one of wild NHPs (Srivathsan et al. 2015), respectively.

### The newly metagenome-assembled genomes (MAGs) greatly increase the mappable diversity of NHP microbiomes

Reference-based taxonomic profiling of all the 203 samples (see **Methods** and **Supplementary Table 2**) confirmed that a very large fraction of NHP metagenomes remains unmapped and uncharacterized (average estimated mapped reads 2.1%±3.64% st. dev., **Supplementary Table 3**), indicating a paucity of microbial genomes representative for members of the gut microbiome of NHPs. We thus employed an assembly-based approach (see **Methods**) to reconstruct microbial genomes *de novo* in the whole set of available NHP metagenomic samples. After single-sample assembly and contig binning of the 203 NHP metagenomes considered, we retrieved a total of 2,985 metagenome-assembled genomes (MAGs) (**Supplementary Table 4**) that exceeded the threshold for being considered of medium quality (completeness >50% and contamination <5%) according to recent guidelines (Bowers et al. 2017)). A large fraction of these genomes (34.6%) could additionally be considered of high quality (completeness >90% and contamination <5%) and provide the basis for assessing the diversity of NHP microbiomes. Functional annotation of all MAGs (see **Methods**, (“UniProt: The Universal Protein Knowledgebase” 2016)) showed low levels of functional characterization in NHPs,

with only 1,049±482 UniRef50 assigned per MAG, in contrast with the 1,426±591 assigned to MAGs from non-Westernized samples and 1,840±847 assigned to those obtained from Westernized populations.

We first mapped the 2,985 obtained MAGs against the previously described species-level genome bins (SGBs, i.e. clusters of MAGs spanning 5% genetic diversity, see **Methods**) that recapitulate the >150,000 MAGs from the human microbiome and the >80,000 reference microbial genomes from public repositories. In total, 310 MAGs (10.39%) fell into 99 SGBs containing at least one known reference genome (called kSGBs), whereas 489 (16.38%) belonged to 200 unknown species (called uSGBs) lacking reference genomes but previously identified in the human microbiome (**Figure 1C** and **Table 1**). The large majority of the MAGs remained however unassigned, with 2,186 MAGs (73.23%) showing >5% genetic distance to any SGB. These completely unknown MAGs firstly reconstructed in this work from NHPs' gut metagenomes were *de novo* clustered into 1,009 NHP-specific SGBs (here defined as primate SGBs or pSGBs) with the same procedure that defines SGBs at 5% genetic diversity we previously employed and validated (Pasolli et al. 2019) (**Figure 1C** and **Table 1**). Overall, NHP microbiomes comprised 1,308 SGBs covering 22 phyla (**Figure 1B**) that expanded the known NHP microbiome diversity with new candidate species mostly expanding the Firmicutes, Bacteroidetes, Euryarchaeota and Elusimicrobia phyla. On the contrary, Actinobacteria were generally underrepresented among NHP SGBs (**Figure 1B**). Although some species were shared between NHPs and humans, our analysis highlighted extensive microbial diversity specifically associated with primates other than humans.

This expanded set of genomes improved the fraction of metagenomic reads in each metagenome that could be mapped by over 6 folds (612%) with respect to the sole reference genomes available in public repositories (>80,000, see **Methods**), and by 2 folds (206.5%) with respect to the catalog of genomes expanded with the MAGs from over 9,500 human metagenomes (Pasolli et al. 2019) (**Figure 1E**). Overall, the average metagenome mappability reached 38.2%, with however uneven increase across datasets (**Figure 1E**). The LiX_2018 dataset of NHPs in captivity reached a mappability of 77.6%, whereas the AmatoKR_2018 dataset of wild NHPs reached merely 17.4% mappability (**Figure 1E**). The fact that LiX_2018 was already highly mapped even when using the available reference genomes alone (22.2% w.r.t. 1% of AmatoKR_2018) and that the human SGBs database was responsible for the largest increase in mappability (reaching 60.7%, w.r.t. 3% of AmatoKR_2018) further confirms that microbiomes from NHPs in captivity are more similar to human ones (**Figure 1E**) than those from wild hosts. Also, the TungJ_2015 dataset reached high mappability levels (63.9%), but this was expected as this is the largest dataset in our meta-analysis (23.6% of the samples considered in this study), with all samples (n=48) from the same host. The AmatoKR_2018 cohort, on the contrary, surveyed many different wild hosts (n=18, 95 samples) that are not covered by other datasets and that have therefore a limited sample size, explaining the limited gain in mappability (14.4% with respect to the human catalog). Overall, the almost 3,000 MAGs

provide the basis for a deeper understanding of the composition and structure of the primate's gut microbiome.

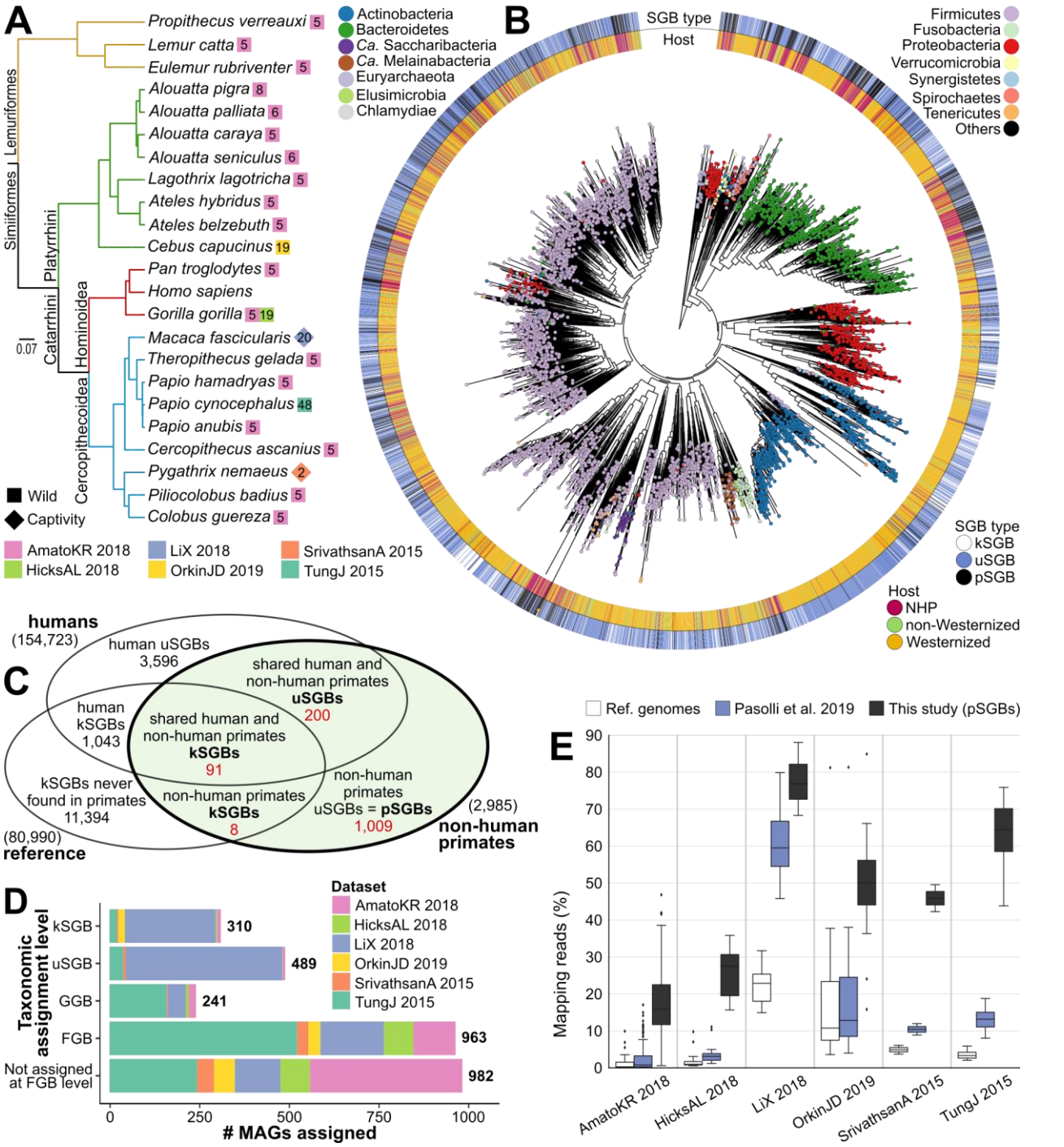



**Figure 1. The expanded set of microbial genomes and species from the gut microbiomes of NHPs. A)** Phylogenetic tree of the primate species considered in this study (adapted from (Springer et al. 2012)), reporting the dataset and number of samples per species; **B)** Microbial phylogeny of the 4,930 species-level genome bins (SGBs, using single representative genomes, see **Methods**) and the 1,009 SGBs that are specific to NHPs and newly retrieved in this study. **C)** Overlap between the sets of SGBs

reconstructed from NHP metagenomes and at least one reference microbial genome (kSGBs), between SGBs reconstructed both from NHP and human metagenomes but lacking a reference genome (uSGBs), and identification of newly assembled SGBs from NHPs metagenomes only (pSGBs); **D)** Fraction of MAGs assigned to clades at different taxonomic levels; samples unassigned at the species-level (kSGB or uSGB) could be assigned to known genus-level genome bins (GGBs) or family-level genome bins (FGBs), or remained unassigned at the family level (not assigned at FGB level); **E)** Statistics of NHP metagenomic read mappability before and after the addition of MAGs from human and NHP metagenomes. We observed an average increase of 612% with respect to reference genomes alone and 206% with respect to the catalog of human MAGs.

| | **# of MAGs in** | | | | **% of MAGs in** | | |
|---|---|---|---|---|---|---|---|
| **dataset** | **SGBs** | **kSGBs** | **uSGBs** | **pSGBs** | **kSGBs** | **uSGBs** | **pSGBs** |
| SrivathsanA_2015 | 92 | 4 | 7 | 81 | 4.3 | 7.6 | 88.0 |
| TungJ_2015 | 985 | 21 | 39 | 925 | 2.1 | 4.0 | 93.9 |
| AmatoKR_2018 | 578 | 10 | 7 | 561 | 1.7 | 1.2 | 97.1 |
| HicksAL_2018 | 177 | 4 | 1 | 172 | 2.3 | 0.6 | 97.2 |
| LiX_2018 | 1043 | 253 | 435 | 355 | 24.3 | 41.7 | 34.0 |
| OrkinJD_2019 | 110 | 18 | 0 | 92 | 16.4 | 0.0 | 83.6 |
| **Total** | 2985 | 310 | 489 | 2186 | 10.4 | 16.4 | 73.2 |

**Table 1.** Number (and percentage) of bins assigned to kSGBs, uSGBs and pSGBs for each dataset.

## Only few and mostly unexplored gut microbes are in common between humans and NHPs

We first investigated how many of the microbial species identified in NHPs were also detected at least once in the human gut microbiome, finding only about 10% overlap (291 of the 2,985 SGBs) between NHP and human gut microbial species. Many of the species found both in NHPs and humans (200 MAGs, 68%) are currently unexplored species without reference genomes (uSGBs). In addition, very few of the newly recovered MAGs belonged to species previously isolated from NHPs but never found in human microbiome 102

samples. This set of eight known species includes *Helicobacter macacae*, which can cause chronic colitis in macaques (Marini et al. 2010; Fox et al. 2007), and *Bifidobacterium moukalabense*, whose type strain was originally isolated from *Gorilla gorilla gorilla* samples (Tsuchida et al. 2014) and we reconstructed from two samples of the same host (**Supplementary Table 5**). The other six known species (*Fibrobacter* sp. UWS1, *Caryophanon tenue*, *Staphylococcus nepalensis*, *Staphylococcus cohnii*, *Enterococcus thailandicus*, *Serratia* sp. FGI94) comprise one MAG only from our dataset and confirm the paucity of isolated and characterized taxa specifically associated to NHPs.

When looking at species with previously-assigned taxonomic labels identified in NHPs, we found a total of 91 species with sequenced representatives (kSGBs) that can also be found in the human microbiome. However, many of them (64.65%) are still rather uncharacterized species as they represent sequenced genomes assigned to genus-level clades without an official species name (e.g. with species names labelled as “sp.” or “bacterium”, **Supplementary Table 6**). Most of such relatively unknown kSGBs were from the *Clostridium* genus (15 kSGBs), and several others belonged to the *Prevotella* (9) and *Ruminococcus* (6) genera. However, both the two most represented human kSGBs assigned to the *Prevotella* genus (13 and 11 MAGs recovered respectively, **Figure 2A** and **Supplementary Table 7**), were retrieved from *Macaca fascicularis* in captivity from the LiX_2018 dataset, consistently with previous literature (Amato et al. 2015; Ma et al. 2014). Among those kSGBs with an unambiguously assigned taxonomy, two highly prevalent *Treponema* species, *T. berlinense* and *T. succinifaciens*, were reconstructed from 14 and 11 samples respectively from different studies and host species (**Figure 2A** and **Supplementary Table 7**). These two species were previously found to be enriched in non-Westernized populations (Pasolli et al. 2019), with 45 genomes reconstructed from different countries. *T. berlinense* and *T. succinifaciens* may thus represent known taxa that are common to primate hosts but that are under negative selective pressure in modern Westernized lifestyles.

The majority (68.7%) of the 291 species shared between humans and NHPs are SGBs without available reference genomes and taxonomic definition (i.e. uSGBs, **Figures 1C** and **1D**). Many of these uSGBs remain unassigned at higher taxonomic levels, with only 25 of them assigned to known genera and 102 to known families. Overall, more than one-third (36.5%) of the uSGBs shared with humans were highly uncharacterized and unassigned at the family level (**Supplementary Table 7**). Among these, also five out of the ten most prevalent shared uSGBs (accounting for 61 MAGs in total) were assigned to the Bacteroidetes phylum (**Figure 2A**) but remained unassigned at lower taxonomic levels (**Supplementary Table 7**). Even among uSGBs, the *Treponema* genus was highly represented, with 9 genomes reconstructed from different samples of *Papio cynocephalus* from the TungJ_2015 dataset (**Supplementary Table 7**). Common human-NHP taxa thus represent only a small fraction of their microbiome and these taxa generally belong to very poorly characterized taxonomic clades.

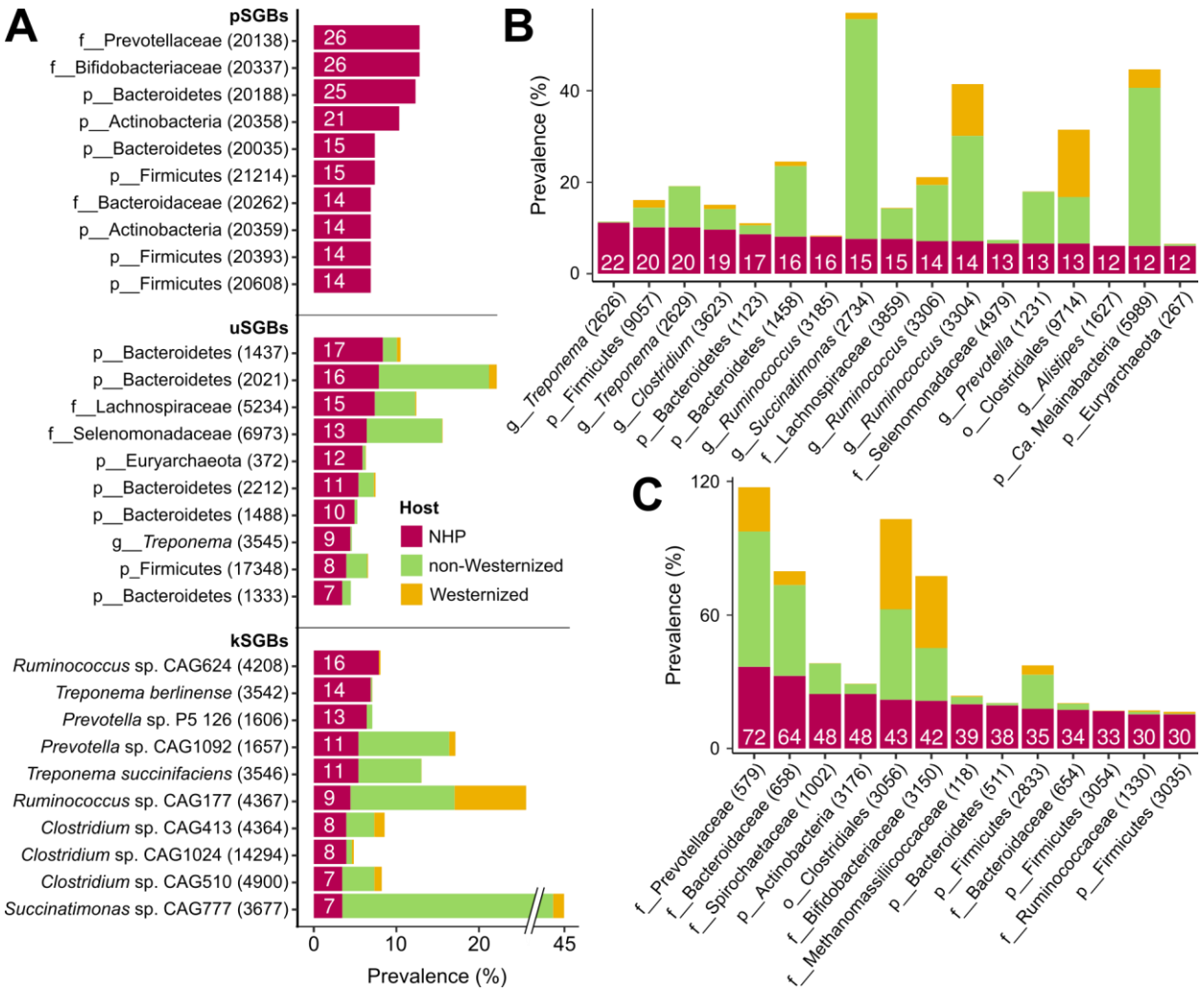



**Figure 2. Most prevalent NHP genome bins from species-level to family-level and their prevalence in Westernized and non-Westernized human populations. A)** Most prevalent pSGBs, uSGBs and kSGBs in NHPs and their prevalence in Westernized and non-Westernized humans; **B)** most prevalent GGBs in NHPs (>11 NHP samples) and their prevalence in Westernized and non-Westernized humans; **C)** most prevalent FGBs in NHPs (>=30 NHP samples) and their prevalence in Westernized and non-Westernized humans. Numbers inside the bars represent the number of NHP samples in which the specific SGB, GGB or FGB has been found. Full list of SGBs, GGBs, and FGBs in **Supplementary Tables 7** and **9**.

## Species overlap between human and NHP microbiomes is heavily lifestyle-dependent

Microbiomes of NHPs in captivity showed reduced numbers of previously unseen microbial diversity (pSGBs) and a larger set of strains from species also found in humans (kSGBs and uSGBs) when compared to wild NHPs. Indeed, eight of the ten most prevalent human-associated SGBs found in at least five NHP samples (**Supplementary Table 7**) were recovered from the LiX_2018 and SrivathsanA_2015 datasets, the only two studies which surveyed the microbiome of NHPs in captivity. Accordingly, a high fraction of genomes

reconstructed from the LiX_2018 captive dataset matches previously described species (64.2%), in contrast with an average of 7.0% ± 6.0% for the MAGs in wild datasets (**Supplementary Table 7**). Overall, these numbers suggest that the microbiome of captive animals is probably a poor representation of the real diversity of their microbiome in the wild, and that exposure of NHPs to the human-associated environment and somehow human-like diet and sanitary procedures can inflate the similarity between human and NHP microbiomes.

The overlap in microbiome composition between wild NHPs and humans is mostly due to the sharing of SGBs characteristic of microbiomes of non-Westernized rather than Westernized human hosts. This is clear when observing that only three SGBs present in NHPs are enriched in prevalence in stool samples from Westernized populations (Fisher's test, Bonferroni-corrected p-values <0.05), with respect to 41 SGBs enriched in non-Westernized datasets (**Figure 3** and **Supplementary Table 8**). Even for those three SGBs associated with Westernized populations, the average prevalence in Westernized datasets was only 0.42%. The SGB found in NHPs that is most strongly associated with non-Westernized populations is *Succinatimonas* sp. (kSGB 3677, prevalence 41.6% in non-Westernized datasets, 1.3% in Westernized datasets; Fisher's test, Bonferroni-corrected p-value 2.74E-223, **Figure 3**), from a genus able to degrade plant sugars such as D-xylose, a monosaccharide present in hemicellulose and enriched in diets rich in plant products. Consistently, the broader *Succinatimonas* genus-level cluster had a prevalence of 48.05% in non-Westernized datasets and of 1.4% in Westernized ones (**Figure 2B**), in agreement both with the folivore diet of most NHPs considered here and with previous observations of enriched D-xylose degradation pathways in non-Westernized populations (De Filippo et al. 2010). Overall, the three most prevalent genus-level genome bins in NHPs (two from the *Treponema* genus, and one from the Firmicutes, all >10% prevalence in NHPs) had an average prevalence of 4.5% in non-Westernized and of 0.6% in Westernized populations (**Figure 2B**).

At the family-level, many *Prevotella* SGBs are both very prevalent in NHPs and in non-Westernized human populations. The overall *Prevotellaceae* family is the most prevalent in NHPs (36.55%), and its prevalence is even higher in non-Westernized human microbiomes (60.55%), while not reaching 20% in Westernized ones (**Figure 2C**). Consistently, four out of the 20 SGBs most associated with non-Westernized human populations belonged to the *Prevotella* genus (SGBs 1680, 1657, 1613, 1614, **Figure 3**), and were however retrieved only from the LiX_2018 dataset of captive *Macaca fascicularis*. Similarly, the only shared SGB assigned at the species level was *Treponema succinifaciens* (kSGB 3546), which was present in 8.22% of non-Westernized samples and in only 0.02% Westernized microbiomes (**Figure 3** and **Supplementary Table 8**), but all of the samples were from the two datasets of NHPs in captivity (LiX_2018 and SrivathsanA_2015), supporting once again the observation that when well-characterized species are found in NHPs, these are usually from captive hosts. The family Spirochaetaceae, to which the genus *Treponema* belongs, was however prevalent also in

wild NHPs (24.37%) and non-Westernized samples (13.67%), while being almost absent in Westernized ones (0.13%, **Figure 2C**). These data thus suggest that the level of similarity between human and NHP microbiomes depends not only on the host species but also on lifestyle variables that could be at least partially assessed both in NHPs (wild versus captive animals) and humans (Westernized vs non-Westernized populations).

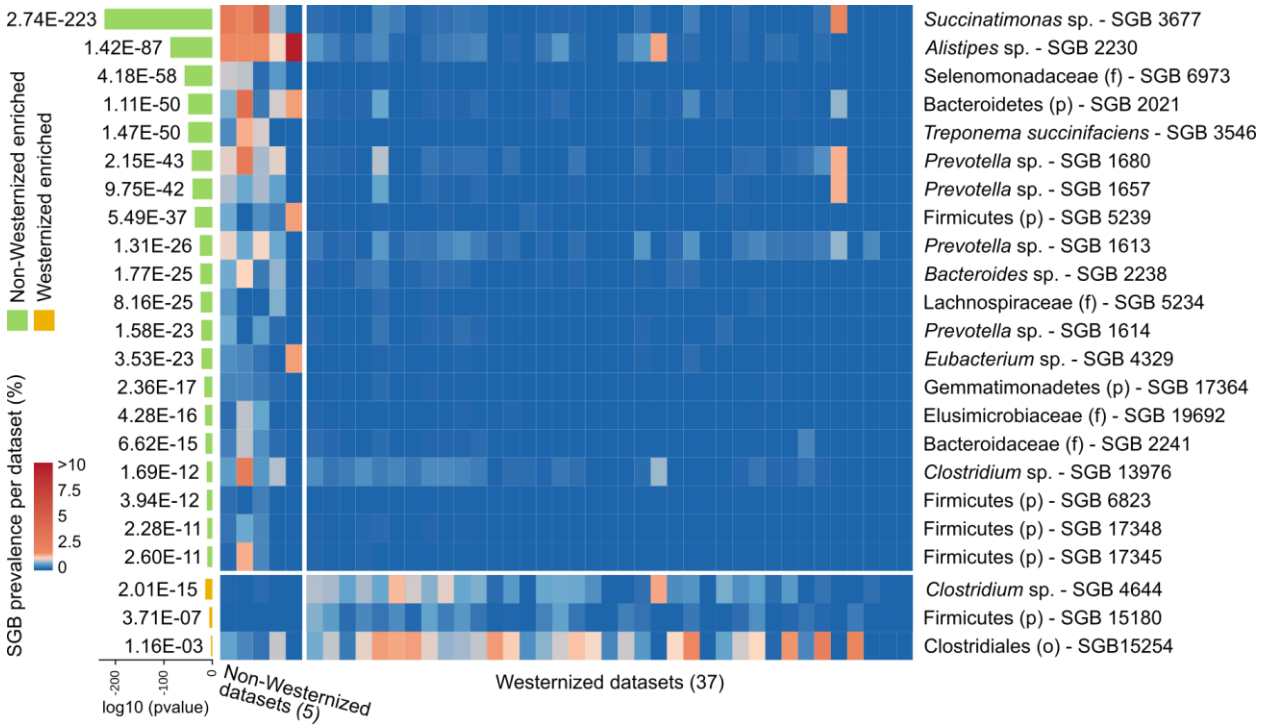



**Figure 3. Prevalences of the NHP SGBs found in humans differentially present in Westernized or non-Westernized human populations.** Association of SGBs found in at least three NHP metagenomes with the gut microbiome of Westernized or non-Westernized populations, together with their prevalence in the different datasets (Fisher's test Bonferroni-corrected p-values, full results in **Supplementary Table 8**).

## Most microbial genomes from NHP metagenomes belong to novel species

More than two thirds (2,186) of the MAGs recovered from NHPs (2,985) belonged to the 1,009 newly defined and previously unexplored SGBs (pSGBs) never found in human microbiomes so far. Some of these pSGBs seem to be key components of the NHP microbiome, with six of them (recapitulating 128 MAGs) within the 10 most prevalent SGBs in NHP microbiomes (**Figure 2A** and **Supplementary Table 7**). The distribution of pSGBs was however not homogeneous among datasets, with the LiX_2018 dataset being the one with the highest fraction of bins assigned to known species (23.5% of the MAGs assigned to kSGBs) and AmatoKR_2018 having 97.23% of the MAGs unassigned at the species level (56.57% unassigned at the family level, **Figure 1D**). This again reflects the different composition of the two datasets, with the captive *Macaca fascicularis* of the LiX_2018 dataset fed with specific human-like diets (X. Li et al. 2018) and the AmatoKR_2018

dataset spanning 18 NHP species living in the wild, which explains its high diversity (**Figure 1A**).

Many of the 1,009 pSGBs were taxonomically unplaced even at higher taxonomic levels, with only 109 pSGBs assigned to a known microbial genus (10.8%, 241 MAGs, see **Methods**), and 386 pSGBs to a known microbial family (38.3%, 963 MAGs, **Figure 1D**). The 514 pSGBs (50.9%, 982 MAGs) that remained unassigned may represent new microbial clades above the level of the bacterial families (**Figure 1D**). The majority of these pSGBs unassigned even at genus level or above was placed, based on genome similarity, into the two highly abundant human gut microbiome phyla of the Firmicutes (44.2% of the unassigned pSGBs, 514 total MAGs) and Bacteroidetes (30.9% of the unassigned pSGBs, 458 MAGs) with smaller fractions assigned to Proteobacteria (9.7%, 125 MAGs), Actinobacteria (5.5%, 108 MAGs), and Spirochaetes (2.8%, 37 MAGs). Although phylum-level composition with the dominance of Bacteroides and Firmicutes is quite consistent among primates, it is thus at the species and genus level that most of the inter-host diversity is occurring, possibly as a consequence of host co-speciation evolutionary dynamics.

To better taxonomically characterize these unassigned pSGBs, we grouped them into clusters spanning a genetic distance consistent with that of known genera and families (Pasolli et al. 2019) generating genus-level genome bins (GGBs) and family-level genome bins (FGBs). This resulted in the definition of 760 novel GGBs (73.6% of the total number of GGBs in NHP) and 265 novel FGBs (65.6% of all FGBs in NHP), with an increase of about 6% of the total GGBs and FGBs previously defined on reference genomes and >154,000 human MAGs. Eight of the ten most prevalent GGBs in NHP samples were part of this novel set of GGBs and were assigned to Coriobacteriales (36 MAGs), Bacteroidaceae (36 MAGs) and Prevotellaceae (33 MAGs) families. Among the most prevalent, only the two *Treponema* GGBs (42 MAGs from NHPs) were known and shared with humans (52 MAGs), mainly from non-Westernized populations (38 MAGs, **Figure 2B** and **Supplementary Table 9**). On the contrary, all of the ten most prevalent families were previously known and shared with humans (**Supplementary Table 9**). In the study of the overall diversity of the primate gut microbiome, it is thus key to consider the new sets of NHP gut microbes defined here that are largely belonging to novel microbial clades.

### Strain-level analysis highlights both host-specific and shared evolutionary trajectories

Despite the low overall degree of microbial sharing between human and non-human hosts at the species level, some bacterial families were common among primate hosts (**Figure 2C**) and motivated a deeper phylogenetic analysis of their internal genomic structure. By using a phylogenetic modelling based on 400 single-copy universal markers (Segata et al. 2013), we therefore reconstructed the phylogeny and the corresponding genetic ordination analysis of the five most relevant shared FGBs (**Figure 2C**), which included three known families (Prevotellaceae, Bacteroidaceae, Spirochaetaceae), and two unexplored FGBs assigned to the Actinobacteria phylum and the Clostridiales order. We observed the

presence of both intra-family host-specific clusters (**Figure 4A**) and clusters comprising genomes spanning human and non-human hosts. The phylogeny of the Bacteroidetes reconstructed to include all of the MAGs and reference genomes for the ten most prevalent characterized (kSGBs), uncharacterized (uSGBs) and newly-reconstructed NHP-specific (pSGBs) species assigned to this phylum (**Figure 4B** and **Supplementary Figure 2**), further confirms the presence of closely related sister clades one of which is specific to wild NHPs and the other spanning multiple hosts, including NHPs in captivity. This likely reflects a complex evolutionary pattern in which vertical coevolution, independent niche selection, and between-host species transmission are likely all simultaneously shaping the members of the gut microbiome of primates.

We also analyzed the under-investigated phylum of the Elusimicrobia as species in this clade were already shown to span a wide range of host environments ranging from aquatic sites to termite guts (Herlemann, Geissinger, and Brune 2007), and were recently found relatively prevalent in non-Westernized human populations (15.4% prevalence) while almost absent in Westernized populations (0.31% prevalence) (Pasolli et al. 2019). The phylum was clearly divided in two main clades (**Supplementary Figure 3**), with one including strains mostly from environmental sources or non-mammalian hosts, and the other (already reported in **Figure 4C**) comprising all the MAGs from humans, NHPs, rumen, and the type strain of *Elusimicrobium minutum (Geissinger et al. 2009)*. The genomes from wild NHPs belonged to an unknown SGB detected also in humans (uSGB 19690) and to two pSGBs (e.g. pSGBs 20223 and 20224) not found in human hosts. These two NHP-specific Elusimicrobia are sister clades of a relatively prevalent human-associated SGB (SGB 19694 comprising 64 MAGs from humans, **Figure 4C**). Such closely related but host-specific sister clades might reflect the evolutionary divergence of the hosts, while the presence of Elusimicrobia strains from macaques in captivity inside human-associated SGBs (**Figure 4C**) also confirms that these microbes can colonize different primate hosts.

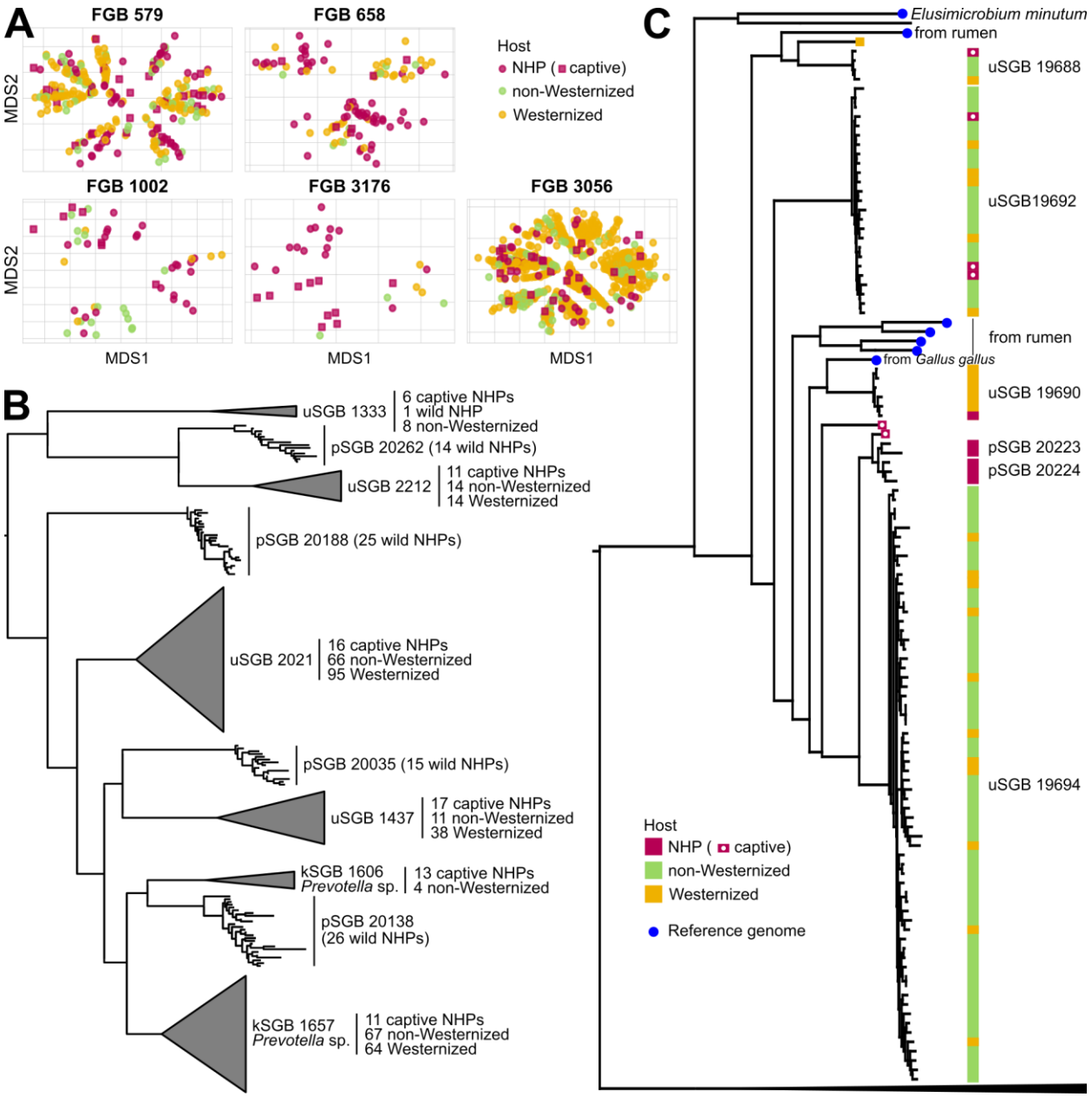



**Figure 4. Strain-level phylogenetic analysis of relevant microbial clades found both in NHPs and human microbiomes. A)** Ordination analysis using multidimensional scaling (MDS) on intra-FGB phylogenetic distances for the five most prevalent FGBs shared by NHPs and humans (**Figure 2C**), showing both host-specific and shared clusters; **B)** Phylogenetic tree of the ten most prevalent kSGBs, uSGBs and pSGBs assigned to the Bacteroidetes phylum reported in **Figure 2A**, with MAGs from wild NHPs in separate pSGB-subtrees and captive NHPs clustering into SGBs shared with humans (uncollapsed tree in **Supplementary Figure 2**); **C)** Phylogenetic tree of the Elusimicrobia phylum, with SGBs specifically associated with wild NHPs and others with humans and captive NHPs (uncollapsed tree in **Supplementary Figure 3**).

### Closely phylogenetically related *Treponema* species have different host-type preferences

The *Treponema* genus contains mostly non-pathogenic species commonly associated with the mammalian intestine and oral cavity (Norris et al. 2006). *Treponema* species seem to be under particular negative selection forces in Westernized populations as multiple studies found them at much higher abundance and prevalence in non-Westernized populations (De Filippo et al. 2010; Schnorr et al. 2014; Obregon-Tito et al. 2015; Pasolli et al. 2019; Angelakis et al. 2019), and they were also identified in ancient coprolites (Raul Y. Tito et al. 2012), and dental calculus of the Iceman mummy (Maixner et al. 2014). To better study its diversity and host-association, we investigated the phylogeny of this genus considering all the genomes from NHPs and humans currently available (**Figure 1B).** The 221 total genomes included 27 available reference genomes and 220 MAGs (96 oral and 124 intestinal) spanning 54 *Treponema* SGBs. These genomes are grouped into 34 distinct SGBs previously reconstructed from human metagenomes and 20 pSGBs newly reconstructed and uniquely associated with NHPs.

Phylogenetic analysis (**Figure 5A**) highlighted a clear and host-independent separation of oral and stool treponemas that is reflected at the functional level (**Figure 5B**), with oral species lacking several pathways encoded by SGBs recovered from stool samples. These included starch and sucrose metabolism, glycerolipid and glycerophospholipid metabolism, methane and sulfur metabolism, folate biosynthesis and phenylalanine, tyrosine and tryptophan biosynthesis, **Supplementary Figure 4**), consistently with the nutrients and carbon sources available in the two different body sites. Focusing on the intestinal species, the SGBs in this family were quite host-specific, with genomes recovered from different hosts clustering in specific subtrees (**Figure 5A**). This is for instance the case of uSGB 3548 and pSGB 21240 that, despite being phylogenetically related, were found only in humans and NHPs, respectively. *Treponema succinifaciens* (kSGB 3546) instead represented an exception, being reconstructed both from NHP (11 MAGs) and mostly non-Westernized human stool microbiomes (45 MAGs, **Figure 5A**). However, the closely related uSGB 3545 was recovered only from NHPs (*Papio cynocephalus*) and could represent a species specifically adapted to the gut of these NHPs or the consequence of the host speciation. It is quite striking that only 11 *Treponema* MAGs were available from Westernized stool samples despite the large number of gut metagenomes analyzed for this category (7,443 stool samples), whereas the same microbial genus was very prevalent in non-Westernized datasets (13.72% of non-Westernized samples, all but one non-Westernized datasets, **Figure 5A** and **Supplementary Table 4**). This raises the hypothesis that *Treponema* species might have been living within the gut of their primate hosts for a long time, and have remained with humans in the absence of lifestyle changes associated with urbanization (Sonnenburg and Sonnenburg 2019).

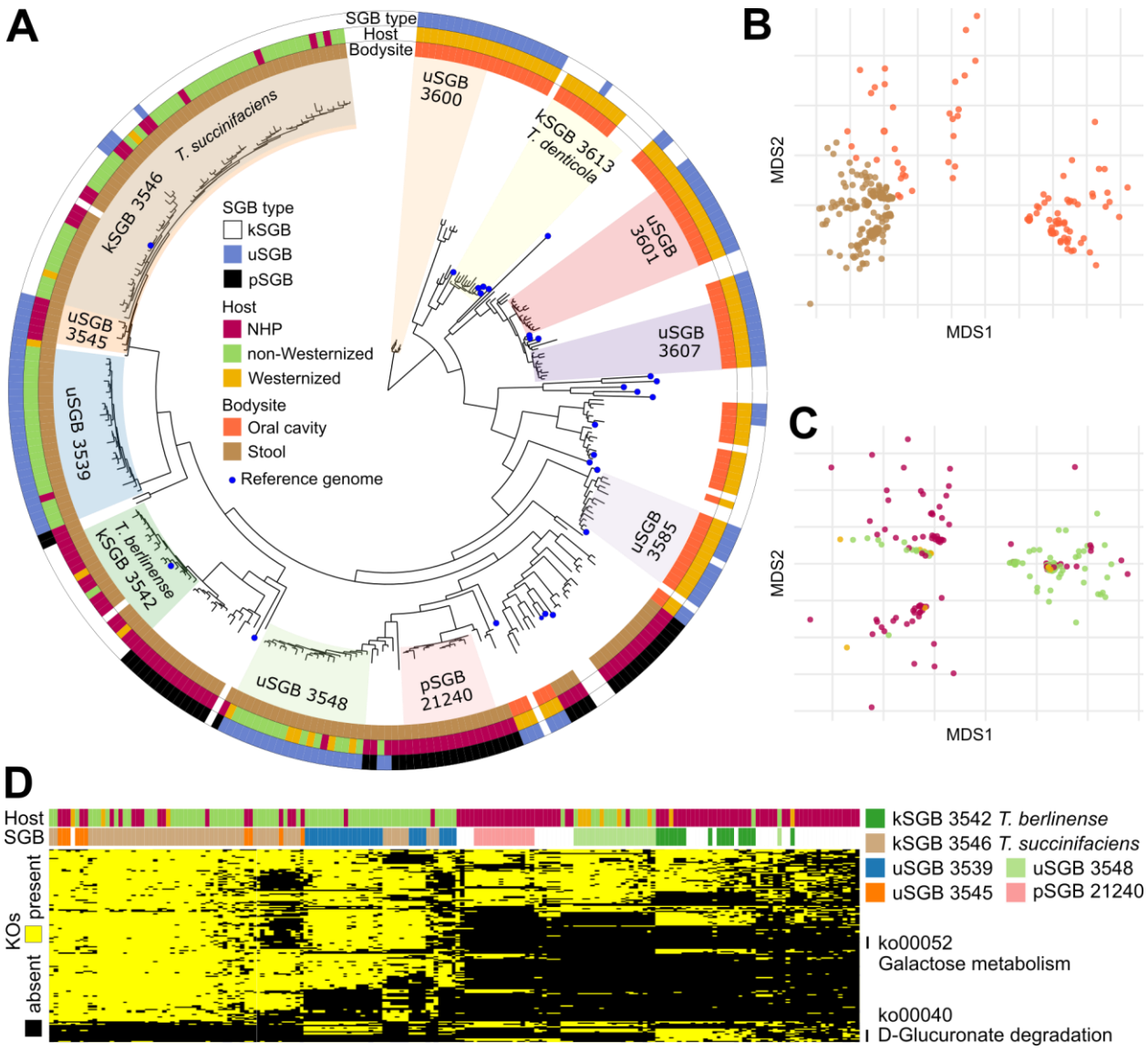



**Figure 5. The *Treponema* genus is the most prevalent among NHPs. A)** Phylogenetic tree of the *Treponema* genus, showing SGB host-specificity and a clear separation between oral and intestinal species (SGB annotation for >10 genomes); **B)** Ordination on functional annotations (UniREf50 clusters) of *Treponema* MAGs colored by bodysite showing separation of oral and intestinal MAGs at the functional level; **C)** Ordination on UniRef50 profiles of *Treponema* MAGs from stool samples only colored by host, showing host-specific functional profiles; **D)** Presence/absence profiles of KEGG orthology families (KOs) in *Treponema* MAGs recovered from stool samples (only KOs related to metabolism and present in at least 20% and less than 80% of samples are reported).

The host-specificity of related *Treponema* species is evident also at the functional level (**Figure 5C**) with several microbial pathways characterizing each species. When comparing the functional potential across hosts, we found for example that human strains were enriched for genes necessary for galactose metabolism (ko00052) and NHPs strains were instead encoding the pathway for the degradation of glucuronate-containing

polymers (ko00040), highly present in hemicellulose (**Figure 5D**), consistently with the different nutritional regimes of humans and NHPs. *Treponema* species enriched in NHPs were however including a substantially lower number of annotated functions (1,312±375 in NHPs w.r.t. 1,426±423 UniRef50 in Westernized samples), pointing to the need of future efforts in experimentally characterize the genes in under-investigated NHP species. The *Treponema* genus overall appears to be a key member of the primate-wide gut microbiome and for this reason its striking disappearance in human Westernized populations suggests that changes in recent lifestyle variables might be responsible for the disruption of intestinal microbes possibly coevolving with our body since the evolutionary era of primate host diversification.

## 4.5 Conclusions

In this study we expanded the fraction of characterized microbial diversity in the highly unexplored non-human primate metagenome, to enable species- and strain-level comparative genomics analysis of the human and non-human primate microbiome and generate hypotheses on relevant coevolutionary trajectories that shaped the current worldwide structure of the human microbiome. Through the application of strain-level single-sample *de novo* genome assembly on 203 NHP metagenomic samples, we uncovered over 1,000 new SGBs expanding the catalog of microbial species recovered from non-human primates by 77% and improving the mappability of NHP metagenomes by over 600%. These newly assembled genomes contributed to the identification of 760 new genus-level and 265 family-level genome bins that represent completely uncharacterized microbial clades never observed in humans. Compared to the over 150,000 MAGs available from human metagenomes (Pasolli et al. 2019) and because of multiple primate hosts that need to be studied, the NHP microbiome still remains undersampled. Given the efficiency of metagenomic assembly pipelines (Nayfach et al. 2019; Almeida et al. 2019) and the availability of complementary tools to explore the microbial diversity in a microbiome (Zou et al. 2019; Gawad, Koh, and Quake 2016), the limiting factor appears to be the technical difficulties in sampling primates in the wild.

The newly established collection of NHP microbial species showed that at fine-grained taxonomic resolution, there is little overlap between the gut microbiomes of humans and NHPs, with 6% of the overall species found in wild NHP that were identified at least once in human microbiomes. Captive NHPs exposed to more human-like environments and diets showed instead higher species sharing with humans (49%) and a higher degree of metagenome mappability. On the other hand, microbiomes from wild NHPs overlapped comparatively much more (163%) with human populations adopting non-Westernized rather than Westernized lifestyles. Because lifestyle patterns appear to have an impact on the structure of the gut microbiome comparable in effect size to that of the primate host species, NHP and potentially ancient microbiome samples (Cano et al. 2000; Raúl Y. Tito et al. 2008; Raul Y. Tito et al. 2012; Rasmussen et al. 2015; Maixner et al. 2016) are thus more suitable for host-microbe coevolutionary analyses as they are likely less confounded by recent lifestyle changes.

Our strain-level investigations of specific taxonomic clades (**Figure 4** and **Figure 5**) showed the presence of both species with strains spanning multiple hosts, and of sister species associated to different primates. While the former is suggestive of recent inter-host transmission or common acquisition from common sources, the second can be the basis to study microbial evolution as a consequence of host speciation, especially if phylogenies can be dated using ancient microbiome samples (Tett et al. 2019) or other time constraints (Lebreton et al. 2017). Our framework can thus be exploited to study inter-host species and zoonotic microbial transmission that is currently mostly limited to specific pathogens of interest (Chan et al. 2015; Han, Kramer, and Drake 2016; Peiris et al. 2016; Trung et al. 2017; Yan et al. 2017; Olea-Popelka et al. 2017). The catalog of primate-associated microbial genomes can thus serve as a basis for a better comprehension of the human microbiome in light of recent and ancient cross-primate transmission and environmental acquisition of microbial diversity.

## 4.6 Methods

### Analyzed datasets

In our meta-analysis, we considered and curated six publicly available gut metagenomic datasets (**Figure 1A** and **Supplementary Table 1**) spanning 22 non-human primate (NHP) species from 14 different countries in five continents (**Supplementary Figure 1**), and metagenomic samples from healthy individuals from 47 datasets included in the curatedMetagenomicData package (Pasolli et al. 2017). In total, our study considers 203 metagenomic samples from the gut of NHPs and 9,428 human metagenomes from different body sites.

The non-human primates datasets were retrieved from four studies considering wild animals and two studies surveying animals in captivity. All but one study produced gut metagenomes of one single species. One work (Amato et al. 2018) instead analyzed the gut microbiome of 18 species of wild NHPs from nine countries (**Figure 1A** and **Supplementary Table 1**) to test the influence of folivory on its composition and function, and highlighted that host phylogeny has a stronger influence than diet. With a similar approach, (Hicks et al. 2018) shotgun sequenced 19 wild western lowland gorillas (*Gorilla gorilla gorilla*) in the Republic of the Congo as part of a 16S rRNA study including sympatric chimpanzees and modern humans microbiomes that demonstrated the compositional divergence between the primates clades' microbiome and the seasonal shift in response to changing dietary habits throughout the year. (Orkin et al. 2019) exposed similar seasonal patterns linked with water and food availability by surveying the microbiome of 20 wild white-faced capuchin monkeys (*Cebus capucinus imitator*) in Costa Rica. (Tung et al. 2015) instead found that social group membership and networks are good predictors of the taxonomic and functional structure of the gut microbiome by surveying 48 wild baboons (*Papio cynocephalus*) in Kenya. Studies in captivity instead include (Srivathsan et al. 2015), who sequenced the gut microbiome of two red-shanked doucs langurs (*Pygathrix nemaeus*) in captivity that were fed a specific mix of plants to test for the ability of metabarcoding vs metagenomics to identify the plants eaten by the

primates from the feces, and (X. Li et al. 2018), who surveyed the change in microbiome composition and function in 20 cynomolgus macaques (*Macaca fascicularis*) fed either a high-fat and low-fiber or a low-fat and high-fiber diet and showed that the first provoked a change toward a more human-like microbiome. Despite the relevance of these six works, none of them attempted at reconstructing novel microbial genomes from NHPs.

### Available genomes used as reference

To define known species-level genome bins (kSGBs), we considered the 80,853 annotated genomes (here referred to as reference genomes) available as of March 2018 in the NCBI GenBank database (NCBI Resource Coordinators 2018). These comprise both complete (12%) and draft (88%) genomes. Draft genomes include also metagenome-assembled genomes (MAGs) and co-abundance gene groups (CAGs).

### Mapping-based taxonomic analysis

As a preliminary explorative test, taxonomic profiling was performed with MetaPhlAn2 (Truong et al. 2015) with default parameters. Additional profiling was performed by using the parameter “-t rel_ab_w_read_stats“ in order to estimate the read mappability for each profiled species.

### Genomes reconstruction and clustering

In order to reconstruct microbial genomes for both characterized and yet-to-be-characterized species, we applied a single sample metagenomic assembly and contig binning approach we described and validated elsewhere (Pasolli et al. 2019). Briefly, assemblies were produced with MEGAHIT (D. Li et al. 2015) and contigs longer than 1,000 nt were binned with MetaBAT2 (Kang et al. 2015) to produce 7,420 genome bins. Quality control with CheckM 1.0.7 (Parks et al. 2015) yielded 1,033 high-quality genome bins (completeness >90%, contamination <5% as described in (Pasolli et al. 2019)) and 1,952 medium quality genome bins (completeness >50% and contamination <5%). Bins were clustered at 5% genetic distance based on whole-genome nucleotide similarity estimation using Mash (version 2.0; option “-s 10000” for sketching) (Ondov et al. 2016). Overall, we obtained 99 kSGBs containing at least one reference genome retrieved from NCBI GenBank (NCBI Resource Coordinators 2018), 200 uSGBs lacking a reference genome but clustering together with genomes reconstructed in (Pasolli et al. 2019), and 1,009 pSGBs consisting only of genomes newly reconstructed in this study (**Figure 1C**). SGBs were further clustered into genus-level genome bins (GGBs) and family-level genome bins (FGBs) spanning 15% and 30% genetic distance, respectively.

### Phylogenetic analysis

Phylogenies were reconstructed using the newly developed version of PhyloPhlAn (Segata et al. 2013). The phylogenetic trees in **Figure 1B** and **Figure 4C** are based on the 400 universal markers as defined in PhyloPhlAn (Segata et al. 2013) and have been built using the following set of parameters: “--diversity high --fast --remove_fragmentary_entries --

fragmentary_threshold 0.67 --min_num_markers 50 --trim greedy" and "--diversity low --accurate --trim greedy --force_nucleotides", respectively.

From the reconstructed phylogeny in **Figure 1B**, we extracted the SGBs falling into the *Treponema* subtree, including also pSGBs. We then applied PhyloPhlAn 2 on all reference genomes and human and non-human primates microbial genomes belonging to the extracted SGBs to produce the phylogenetic tree reported in **Figure 5A** [with params --diversity low --trim greedy --min_num_marker 50].

External tools with their specific options as used in the PhyloPhlAn framework are:

- diamond (version v0.9.9.110, (Buchfink, Xie, and Huson 2015)) with parameters: "blastx --quiet --threads 1 --outfmt 6 --more-sensitive --id 50 --max-hsps 35 -k 0" and with parameters: "blastp --quiet --threads 1 --outfmt 6 --more-sensitive --id 50 --max-hsps 35 -k 0";
- mafft (version v7.310, (Katoh and Standley 2013)) with the "--anysymbol" option;
- trimal (version 1.2rev59, (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009)) with the "-gappyout" option;
- FastTree (version 2.1.9, (Price, Dehal, and Arkin 2010)) with "-mlacc 2 -slownni -spr 4 -fastest -mlnni 4 -no2nd -gtr -nt" options;
- RAxML (version 8.1.15, (Stamatakis 2014)) with parameters: "-m PROTCATLG -p 1989".

Trees in **Figures 1B** and **5A** were visualized with GraPhlAn (Asnicar et al. 2015). Phylogenetic tree of the primates was obtained from (Springer et al. 2012), manually pruned with iTOL (Letunic and Bork 2016) to report only species considered in this study, and visualized with FigTree v.1.4.3 ("FigTree" n.d.).

### Mappability

We estimated the percentage of raw reads in each sample that could align to known bacterial genomes, SGBs and pSGBs using a previously described method (Pasolli et. al, 2019). Briefly, each raw metagenome was subsampled at 1% to reduce the computational cost of mapping. Subsampled reads were filtered to remove alignments to the Human genome (hg19). Short (i.e. lower than 70 bp) and low-quality (mean sequencing quality < 20) reads were discarded.

Each sample was mapped against three groups of indexes: i) a set of 80,990 reference genomes used to define the set of known SGBs in (Pasolli et al. 2019); ii) the 154,753 known and unknown SGBs from (Pasolli et al. 2019); and iii) the 1,009 SGBs from NHPs reconstructed in this study. The mapping was performed with BowTie2 (Langmead and Salzberg 2012) v. 2.3.5 in end-to-end mode. The mapping was performed incrementally (i.e. reads that are reported to map against pSGBs do not map against any reference genome or human SGB). Additionally, BowTie2 alignments scoring less than -20 (tag AS:i) were excluded, to avoid to overestimate the number of mapping reads. The mappability

fraction was calculated by dividing the number of aligning reads by the number of high-quality reads within each sample.

### Functional analysis

Metagenome-assembled genomes reconstructed in this study were annotated with Prokka 1.12 (Seemann 2014) using default parameters. Proteins inferred with Prokka were then functionally annotated with UniRef90 and UniRef50 using diamond v0.9.9.110 (Buchfink, Xie, and Huson 2015).

KEGG Orthology (KO) for the UniRef50 annotations were retrieved from the UniProt website using the Retrieve/ID mapping tool. KOs related to metabolism were filtered and used to produce a presence/absence matrix for generating **Figure 5D** and **Supplementary Figure 4**. Non-metric multidimensional scaling plots were generated using the Jaccard distance with the metaMDS function in the vegan R package (Oksanen et al. 2008).

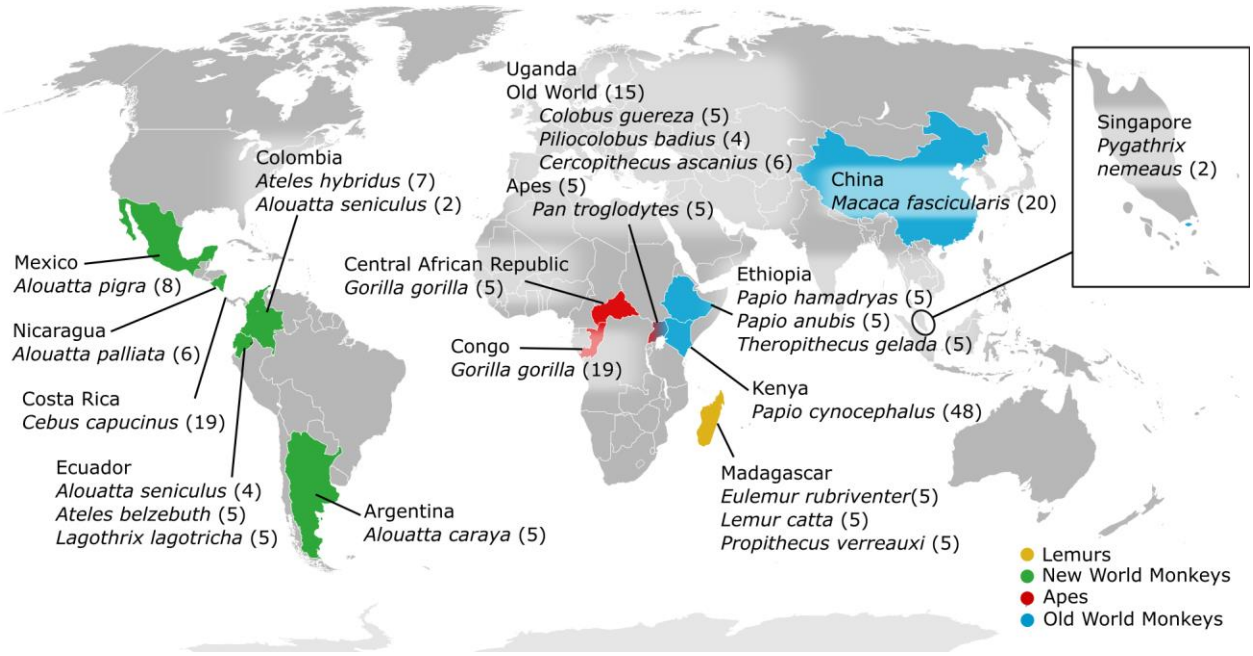
### Statistical analysis

Statistical significance was verified through Fisher's test with multiple hypothesis testing correction with either Bonferroni or FDR as reported in the text.

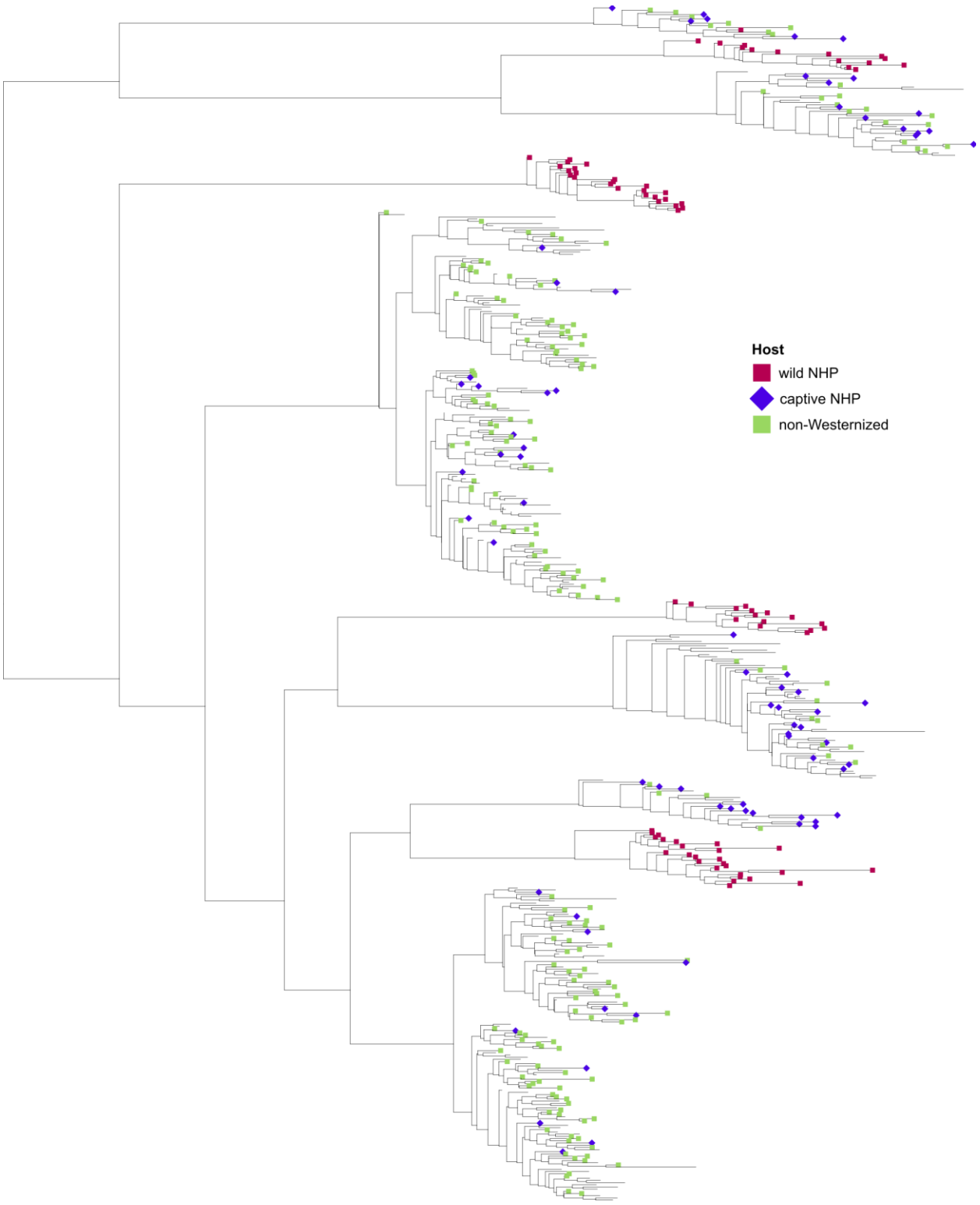
## Acknowledgements

This work was supported by the European Research Council (ERC-STG project MetaPG-716575), MIUR ‘‘Futuro in Ricerca’’ RBFR13EWWI_001, and the European Union (H2020-SFS-2018-1 project MASTER-818368) to N.S.

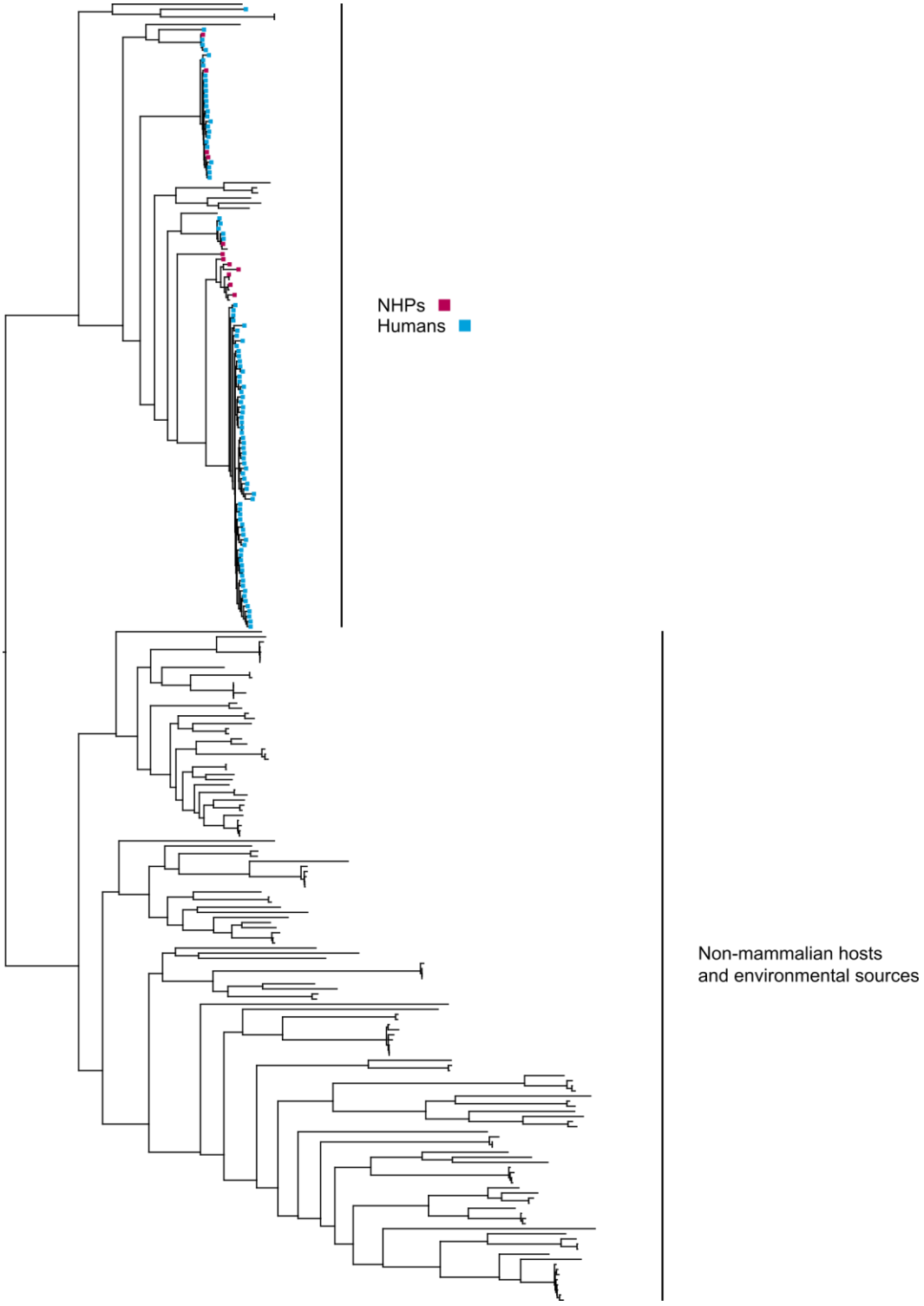
## 4.7 Supplementary Figures



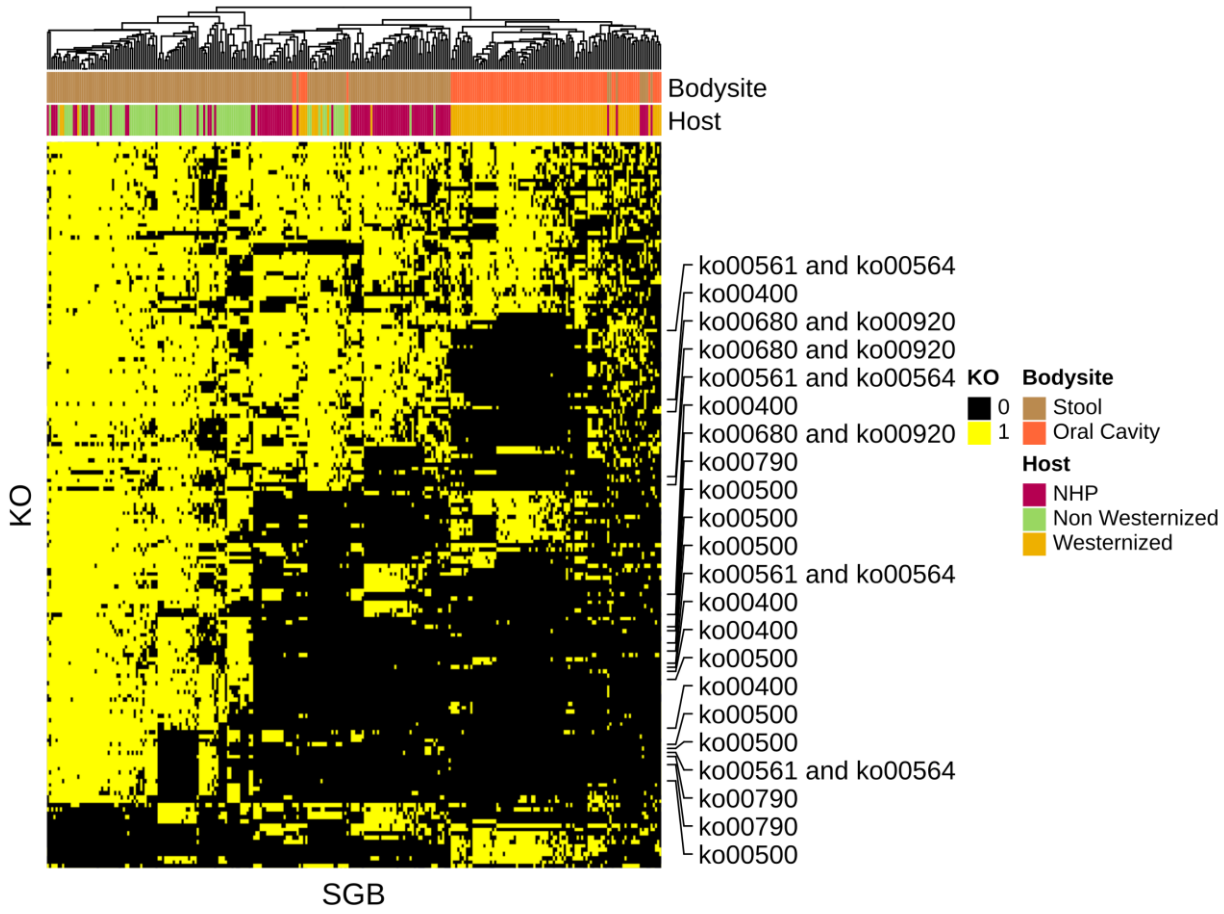

**Supplementary Figure 1.** World map reporting NHP metagenomic samples considered in this study together with host and country information.




**Supplementary Figure 2.** Phylogenetic tree of the Bacteroidetes phylum (uncollapsed version of the tree in **Figure 4B**).




**Supplementary Figure 3.** Phylogenetic tree of the Elusimicrobia phylum (uncollapsed version of the tree in **Figure 4C**).




**Supplementary Figure 4.** KO presence/absence profile in *Treponema* MAGs recovered from both stool and oral cavity samples. Only KOs related to metabolism and present in at least 20% and less than 80% of samples are reported.

## 4.8 Supplementary Tables

Captions of supplementary tables are reported below. Tables are available for download at this link:
https://www.dropbox.com/sh/ptd6jhx0neomnzn/AAAYHSgL8SwAiQrVPNEDMFP6a?dl=0 .

**Supplementary Table 1.** NHP gut metagenome datasets considered in this study, together with relevant information like Pubmed ID, host, wild or captivity status, country of sampling, number of samples.

**Supplementary Table 2.** MetaPhlAn2 profiles of all metagenomic samples from NHPs considered in this study.

**Supplementary Table 3.** Estimated mapped reads according to MetaPhlAn2, both per sample and per dataset.

**Supplementary Table 4.** Description of single-sample assembled genomes (host, wild or captivity status, country of sampling), their assembly and quality statistics, their assigned taxonomy, and SGB, GGB and FGB assignment.

**Supplementary Table 5.** SGB presence/absence in the considered metagenomic samples from NHPs.

**Supplementary Table 6.** Description of the 60 kSGBs retrieved from NHPs and shared with humans that are lacking an official species name (e.g. with species name labelled as “sp.” or “bacterium”), together with the number of reference genomes and MAGs retrieved for each host category.

**Supplementary Table 7.** Description of the SGBs (kSGBs, uSGBs and pSGBs) reported in this study, together with the assigned taxonomy and the number of reference genomes and MAGs retrieved from each host category and NHP dataset.

**Supplementary Table 8.** SGBs found in >3 NHP samples and their association with either Westernized or non-Westernized microbiomes (Fisher’s test Bonferroni-corrected p-values, comprehensive list of SGBs shown in **Figure 3A**).

**Supplementary Table 9. Tab1)** GGBs description, together with number of MAGs assigned and number of dataset and samples in which the GGB is found, divided by host and NHP dataset. **Tab2)** Same information for FGBs recovered in this study.

**Supplementary Table 10.** List of *Treponema* SGBs reported in the phylogeny in **Figure 5A** and their prevalence in NHPs and humans.

## 4.9 References


Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. “A New Genomic Blueprint of the Human Gut Microbiota.” *Nature* 568 (7753): 499–504.

Amato, Katherine R. 2019. “Missing Links: The Role of Primates in Understanding the Human Microbiome.” *mSystems*. https://doi.org/10.1128/msystems.00165-19.

Amato, Katherine R., Jon G Sanders, Se Jin Song, Michael Nute, Jessica L. Metcalf, Luke R. Thompson, James T. Morton, et al. 2018. “Evolutionary Trends in Host Physiology Outweigh Dietary Niche in Structuring Primate Gut Microbiomes.” *The ISME Journal* 13 (3): 576–87.

Amato, Katherine R., Carl J. Yeoman, Gabriela Cerda, Christopher A. Schmitt, Jennifer Danzy Cramer, Margret E. Berg Miller, Andres Gomez, et al. 2015. “Variable Responses of Human and Non-Human Primate Gut Microbiomes to a Western Diet.” *Microbiome* 3 (November): 53.

Angelakis, E., D. Bachar, M. Yasir, D. Musso, F. Djossou, B. Gaborit, S. Brah, et al. 2019. “Treponema Species Enrich the Gut Microbiota of Traditional Rural Populations but Are Absent from Urban Individuals.” *New Microbes and New Infections* 27 (January): 14–21.

Asnicar, Francesco, George Weingart, Timothy L. Tickle, Curtis Huttenhower, and Nicola Segata. 2015. “Compact Graphical Representation of Phylogenetic Data and Metadata with GraPhlAn.” *PeerJ* 3 (June): e1029.

Blaser, Martin J. 2017. “The Theory of Disappearing Microbiota and the Epidemics of Chronic Diseases.” *Nature Reviews. Immunology* 17 (8): 461–63.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. “Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea.” *Nature Biotechnology* 35 (8): 725–31.

Bressa, Carlo, María Bailén-Andrino, Jennifer Pérez-Santiago, Rocío González-Soltero, Margarita Pérez, Maria Gregoria Montalvo-Lominchar, Jose Luis Maté-Muñoz, Raúl Domínguez, Diego Moreno, and Mar Larrosa. 2017. “Differences in Gut Microbiota Profile between Women with Active Lifestyle and Sedentary Women.” *PloS One* 12 (2): e0171352.

Brito, I. L., S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, et al. 2016. “Mobile Genes in the Human Microbiome Are Structured from Global to Individual Scales.” *Nature* 535 (7612): 435–39.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.

Cabana, F., J. B. Clayton, K. A. I. Nekaris, W. Wirdateti, D. Knights, and H. Seedorf. 2019. “Nutrient-Based Diet Modifications Impact on the Gut Microbiome of the Javan Slow Loris (Nycticebus Javanicus).” *Scientific Reports* 9 (1): 4078.

Cano, R. J., F. Tiefenbrunner, M. Ubaldi, C. Del Cueto, S. Luciani, T. Cox, P. Orkand, K. H. Künzel, and F. Rollo. 2000. “Sequence Analysis of Bacterial DNA in the Colon and Stomach of the Tyrolean Iceman.” *American Journal of Physical Anthropology* 112 (3): 297–309.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73.

Chan, Jasper F. W., Susanna K. P. Lau, Kelvin K. W. To, Vincent C. C. Cheng, Patrick C. Y. Woo, and Kwok-Yung Yuen. 2015. “Middle East Respiratory Syndrome Coronavirus: Another Zoonotic Betacoronavirus Causing SARS-like Disease.” *Clinical Microbiology Reviews* 28 (2): 465–522.

Clayton, Jonathan B., Pajau Vangay, Hu Huang, Tonya Ward, Benjamin M. Hillmann, Gabriel A. Al-Ghalith, Dominic A. Travis, et al. 2016. “Captivity Humanizes the Primate Microbiome.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (37): 10376–81.

David, Lawrence A., Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, et al. 2014. “Diet Rapidly and Reproducibly Alters the Human Gut Microbiome.” *Nature* 505 (7484): 559–63.

De Filippo, Carlotta, Duccio Cavalieri, Monica Di Paola, Matteo Ramazzotti, Jean Baptiste Poullet, Sebastien Massart, Silvia Collini, Giuseppe Pieraccini, and Paolo Lionetti. 2010. “Impact of Diet in Shaping Gut Microbiota Revealed by a Comparative Study in Children from Europe and Rural Africa.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (33): 14691–96.

Degnan, Patrick H., Anne E. Pusey, Elizabeth V. Lonsdorf, Jane Goodall, Emily E. Wroblewski, Michael L. Wilson, Rebecca S. Rudicell, Beatrice H. Hahn, and Howard Ochman. 2012. “Factors Associated with the Diversification of the Gut Microbial Communities within Chimpanzees from Gombe National Park.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (32): 13034–39.

“FigTree.” n.d. Accessed March 6, 2019. http://tree.bio.ed.ac.uk/software/figtree/.

Fox, J. G., S. R. Boutin, L. K. Handt, N. S. Taylor, S. Xu, B. Rickman, R. P. Marini, et al. 2007. “Isolation and Characterization of a Novel Helicobacter Species, ‘Helicobacter Macacae,’ from Rhesus Monkeys with and without Chronic Idiopathic Colitis.” *Journal of Clinical Microbiology*. https://doi.org/10.1128/jcm.01100-07.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. “Single-Cell Genome Sequencing: Current State of the Science.” *Nature Reviews. Genetics* 17 (3): 175–88.

Geissinger, Oliver, Daniel P. R. Herlemann, Erhard Mörschel, Uwe G. Maier, and Andreas Brune. 2009. “The Ultramicrobacterium ‘Elusimicrobium Minutum’ Gen. Nov., Sp. Nov., the First Cultivated Representative of the Termite Group 1 Phylum.” *Applied and Environmental Microbiology* 75 (9): 2831–40.

Gomez, Andres, Jessica M. Rothman, Klara Petrzelkova, Carl J. Yeoman, Klara Vlckova, Juan D. Umaña, Monica Carr, et al. 2016. “Temporal Variation Selects for Diet-Microbe Co-Metabolic Traits in the Gut of Gorilla Spp.” *The ISME Journal* 10 (2): 514–26.

Greene, Lydia K., Sally L. Bornbusch, Erin A. McKenney, Rachel L. Harris, Sarah R. Gorvetzian, Anne D. Yoder, and Christine M. Drea. 2019. “The Importance of Scale in Comparative Microbiome Research: New Insights from the Gut and Glands of Captive and Wild Lemurs.” *American Journal of Primatology*, April, e22974.

Han, Barbara A., Andrew M. Kramer, and John M. Drake. 2016. “Global Patterns of Zoonotic Disease in Mammals.” *Trends in Parasitology* 32 (7): 565–77.

Herlemann, Daniel P. R., Oliver Geissinger, and Andreas Brune. 2007. “The Termite Group I Phylum Is Highly Diverse and Widespread in the Environment.” *Applied and Environmental Microbiology* 73 (20): 6682–85.

Hicks, Allison L., Kerry Jo Lee, Mara Couto-Rodriguez, Juber Patel, Rohini Sinha, Cheng Guo, Sarah H. Olson, et al. 2018. “Gut Microbiomes of Wild Great Apes Fluctuate Seasonally in Response to Diet.” *Nature Communications* 9 (1): 1786.

Hold, G. L. 2014. “Western Lifestyle: A ‘Master’manipulator of the Intestinal Microbiota?” *Gut*. https://gut.bmj.com/content/63/1/5.short.

Kang, D. D., J. Froula, R. Egan, and Z. Wang. 2015. “MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities. PeerJ 3: e1165.”

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Langdon, Amy, Nathan Crook, and Gautam Dantas. 2016. “The Effects of Antibiotics on the Microbiome throughout Development and Alternative Approaches for Therapeutic Modulation.” *Genome Medicine* 8 (1): 39.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Lebreton, François, Abigail L. Manson, Jose T. Saavedra, Timothy J. Straub, Ashlee M. Earl, and Michael S. Gilmore. 2017. “Tracing the Enterococci from Paleozoic Origins to the Hospital.” *Cell* 169 (5): 849–61.e13.

Letunic, Ivica, and Peer Bork. 2016. “Interactive Tree of Life (iTOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees.” *Nucleic Acids Research* 44 (W1): W242–45.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31 (10): 1674–76.

Liu, Wenjun, Jiachao Zhang, Chunyan Wu, Shunfeng Cai, Weiqiang Huang, Jing Chen, Xiaoxia Xi, et al. 2016. “Unique Features of Ethnic Mongolian Gut Microbiome Revealed by Metagenomic Analysis.” *Scientific Reports* 6 (October): 34826.

Li, Xiaoping, Suisha Liang, Zhongkui Xia, Jing Qu, Huan Liu, Chuan Liu, Huanming Yang, et al. 2018. “Establishment of a Macaca Fascicularis Gut Microbiome Gene Catalog and Comparison with the Human, Pig, and Mouse Gut Microbiomes.” *GigaScience* 7 (9). https://doi.org/10.1093/gigascience/giy100.

Maixner, Frank, Ben Krause-Kyora, Dmitrij Turaev, Alexander Herbig, Michael R. Hoopmann, Janice L. Hallows, Ulrike Kusebauch, et al. 2016. “The 5300-Year-Old Helicobacter Pylori Genome of the Iceman.” *Science* 351 (6269): 162–65.

Maixner, Frank, Anton Thomma, Giovanna Cipollini, Stefanie Widder, Thomas Rattei, and Albert Zink. 2014. “Metagenomic Analysis Reveals Presence of Treponema Denticola in a Tissue Biopsy of the Iceman.” *PloS One* 9 (6): e99994.

Ma, Jun, Amanda L. Prince, David Bader, Min Hu, Radhika Ganu, Karalee Baquero, Peter Blundell, et al. 2014. “High-Fat Maternal Diet during Pregnancy Persistently Alters the Offspring Microbiome in a Primate Model.” *Nature Communications* 5 (May): 3889.

Marini, R. P., S. Muthupalani, Z. Shen, E. M. Buckley, C. Alvarado, N. S. Taylor, F. E. Dewhirst, M. T. Whary, M. M. Patterson, and J. G. Fox. 2010. “Persistent Infection of Rhesus Monkeys with

'Helicobacter Macacae' and Its Isolation from an Animal with Intestinal Adenocarcinoma." *Journal of Medical Microbiology*. https://doi.org/10.1099/jmm.0.019117-0.

Moeller, Andrew H., Patrick H. Degnan, Anne E. Pusey, Michael L. Wilson, Beatrice H. Hahn, and Howard Ochman. 2012. "Chimpanzees and Humans Harbour Compositionally Similar Gut Enterotypes." *Nature Communications* 3: 1179.

Moeller, Andrew H., Steffen Foerster, Michael L. Wilson, Anne E. Pusey, Beatrice H. Hahn, and Howard Ochman. 2016. "Social Behavior Shapes the Chimpanzee Pan-Microbiome." *Science Advances* 2 (1): e1500997.

Moeller, Andrew H., Yingying Li, Eitel Mpoudi Ngole, Steve Ahuka-Mundeke, Elizabeth V. Lonsdorf, Anne E. Pusey, Martine Peeters, Beatrice H. Hahn, and Howard Ochman. 2014. "Rapid Changes in the Gut Microbiome during Human Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 111 (46): 16431–35.

Moeller, Andrew H., Martine Peeters, Jean-Basco Ndjango, Yingying Li, Beatrice H. Hahn, and Howard Ochman. 2013. "Sympatric Chimpanzees and Gorillas Harbor Convergent Gut Microbial Communities." *Genome Research* 23 (10): 1715–20.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. 2019. "New Insights from Uncultivated Genomes of the Global Human Gut Microbiome." *Nature* 568 (7753): 505–10.

NCBI Resource Coordinators. 2018. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 46 (D1): D8–13.

Nishida, Alex H., and Howard Ochman. 2019. "A Great-Ape View of the Gut Microbiome." *Nature Reviews. Genetics* 20 (4): 195–206.

Norris, Steven J., Bruce J. Paster, Annette Moter, and Ulf B. Göbel. 2006. "The Genus Treponema." In *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass*, edited by Martin Dworkin, Stanley Falkow, Eugene Rosenberg, Karl-Heinz Schleifer, and Erko Stackebrandt, 211–34. New York, NY: Springer New York.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. "metaSPAdes: A New Versatile Metagenomic Assembler." *Genome Research* 27 (5): 824–34.

Obregon-Tito, Alexandra J., Raul Y. Tito, Jessica Metcalf, Krithivasan Sankaranarayanan, Jose C. Clemente, Luke K. Ursell, Zhenjiang Zech Xu, et al. 2015. "Subsistence Strategies in Traditional Societies Distinguish Gut Microbiomes." *Nature Communications* 6 (March): 6505.

Ochman, Howard, Michael Worobey, Chih-Horng Kuo, Jean-Bosco N. Ndjango, Martine Peeters, Beatrice H. Hahn, and Philip Hugenholtz. 2010. "Evolutionary Relationships of Wild Hominids Recapitulated by Gut Microbial Communities." *PLoS Biology* 8 (11): e1000546.

Oksanen, Jari, Roeland Kindt, Pierre Legendre, Bob O'Hara, Gavin L. Simpson, Peter Solymos, H. H. Stevens, Helene Wagner, Maintainer Jari Oksanen, and Mass Suggests. 2008. "The Vegan Package." *Community Ecology Package* 10. https://www.researchgate.net/profile/Gavin_Simpson/publication/228339454_The_vegan_Package/links/0912f50be86bc29a7f000000/The-vegan-Package.pdf.

Olea-Popelka, Francisco, Adrian Muwonge, Alejandro Perera, Anna S. Dean, Elizabeth Mumford, Elisabeth Erlacher-Vindel, Simona Forcella, et al. 2017. "Zoonotic Tuberculosis in Human Beings Caused by Mycobacterium Bovis—a Call for Action." *The Lancet Infectious Diseases* 17 (1): e21–25.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. “Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.” *Genome Biology* 17 (1): 132.

Orkin, Joseph D., Fernando A. Campos, Monica S. Myers, Saul E. Cheves Hernandez, Adrián Guadamuz, and Amanda D. Melin. 2019. “Seasonality of the Gut Microbiota of Free-Ranging White-Faced Capuchins in a Tropical Dry Forest.” *The ISME Journal* 13 (1): 183–96.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25 (7): 1043–55.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.” *Cell* 176 (3): 649–62.e20.

Pasolli, Edoardo, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, et al. 2017. “Accessible, Curated Metagenomic Data through ExperimentHub.” *Nature Methods* 14 (11): 1023–24.

Peiris, J. S. Malik, Benjamin J. Cowling, Joseph T. Wu, Luzhao Feng, Yi Guan, Hongjie Yu, and Gabriel M. Leung. 2016. “Interventions to Reduce Zoonotic and Pandemic Risks from Avian Influenza in Asia.” *The Lancet Infectious Diseases* 16 (2): 252–58.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments.” *PloS One* 5 (3): e9490.

Rampelli, Simone, Stephanie L. Schnorr, Clarissa Consolandi, Silvia Turroni, Marco Severgnini, Clelia Peano, Patrizia Brigidi, Alyssa N. Crittenden, Amanda G. Henry, and Marco Candela. 2015. “Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota.” *Current Biology: CB* 25 (13): 1682–93.

Rasmussen, Simon, Morten Erik Allentoft, Kasper Nielsen, Ludovic Orlando, Martin Sikora, Karl-Göran Sjögren, Anders Gorm Pedersen, et al. 2015. “Early Divergent Strains of Yersinia Pestis in Eurasia 5,000 Years Ago.” *Cell* 163 (3): 571–82.

Schnorr, Stephanie L., Marco Candela, Simone Rampelli, Manuela Centanni, Clarissa Consolandi, Giulia Basaglia, Silvia Turroni, et al. 2014. “Gut Microbiome of the Hadza Hunter-Gatherers.” *Nature Communications* 5 (April): 3654.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.

Segata, Nicola. 2015. “Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity.” *Current Biology: CB* 25 (14): R611–13.

Segata, Nicola, Daniela Börnigen, Xochitl C. Morgan, and Curtis Huttenhower. 2013. “PhyloPhlAn Is a New Method for Improved Phylogenetic and Taxonomic Placement of Microbes.” *Nature Communications*. https://doi.org/10.1038/ncomms3304.

Smits, Samuel A., Jeff Leach, Erica D. Sonnenburg, Carlos G. Gonzalez, Joshua S. Lichtman, Gregor Reid, Rob Knight, et al. 2017. “Seasonal Cycling in the Gut Microbiome of the Hadza Hunter-Gatherers of Tanzania.” *Science* 357 (6353): 802–6.

Sommer, F., and F. Bäckhed. 2013. “The Gut Microbiota—masters of Host Development and

Physiology." *Nature Reviews. Microbiology*. https://www.nature.com/articles/nrmicro2974.

Sonnenburg, Erica D., and Justin L. Sonnenburg. 2019. "The Ancestral and Industrialized Gut Microbiota and Implications for Human Health." *Nature Reviews. Microbiology* 17 (6): 383–90.

Springer, Mark S., Robert W. Meredith, John Gatesy, Christopher A. Emerling, Jong Park, Daniel L. Rabosky, Tanja Stadler, et al. 2012. "Macroevolutionary Dynamics and Historical Biogeography of Primate Diversification Inferred from a Species Supermatrix." *PloS One* 7 (11): e49521.

Srivathsan, Amrita, John C. M. Sha, Alfried P. Vogler, and Rudolf Meier. 2015. "Comparing the Effectiveness of Metagenomics and Metabarcoding for Diet Analysis of a Leaf-Feeding Monkey (P Ygathrix Nemaeus)." *Molecular Ecology Resources* 15 (2): 250–61.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu033.

Tett, A., K. D. Huang, F. Asnicar, and H. Fehlner-Peach. 2019. "The Prevotella Copri Complex Comprises Four Distinct Clades That Are Underrepresented in Westernised Populations." *BioRxiv*. https://www.biorxiv.org/content/10.1101/600593v1.abstract.

Tito, Raul Y., Dan Knights, Jessica Metcalf, Alexandra J. Obregon-Tito, Lauren Cleeland, Fares Najar, Bruce Roe, et al. 2012. "Insights from Characterizing Extinct Human Gut Microbiomes." *PloS One* 7 (12): e51146.

Tito, Raúl Y., Simone Macmil, Graham Wiley, Fares Najar, Lauren Cleeland, Chunmei Qu, Ping Wang, et al. 2008. "Phylotyping and Functional Analysis of Two Ancient Human Microbiomes." *PloS One* 3 (11): e3703.

Trung, Nguyen Vinh, Sébastien Matamoros, Juan J. Carrique-Mas, Nguyen Huu Nghia, Nguyen Thi Nhung, Tran Thi Bich Chieu, Ho Huynh Mai, et al. 2017. "Zoonotic Transmission of Mcr-1 Colistin Resistance Gene from Small-Scale Poultry Farms, Vietnam." *Emerging Infectious Diseases* 23 (3): 529–32.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 12 (10): 902–3.

Tsuchida, Sayaka, Shunsuke Takahashi, Pierre Philippe Mbehang Nguema, Shiho Fujita, Maki Kitahara, Juichi Yamagiwa, Alfred Ngomanda, Moriya Ohkuma, and Kazunari Ushida. 2014. "Bifidobacterium Moukalabense Sp. Nov., Isolated from the Faeces of Wild West Lowland Gorilla (Gorilla Gorilla Gorilla)." *International Journal of Systematic and Evolutionary Microbiology* 64 (Pt 2): 449–55.

Tung, Jenny, Luis B. Barreiro, Michael B. Burns, Jean-Christophe Grenier, Josh Lynch, Laura E. Grieneisen, Jeanne Altmann, Susan C. Alberts, Ran Blekhman, and Elizabeth A. Archie. 2015. "Social Networks Predict Gut Microbiome Composition in Wild Baboons." *eLife* 4 (March). https://doi.org/10.7554/eLife.05224.

"UniProt: The Universal Protein Knowledgebase." 2016. *Nucleic Acids Research* 45 (D1): D158–69.

Yan, Wenchao, Kerri Alderisio, Dawn M. Roellig, Kristin Elwin, Rachel M. Chalmers, Fengkun Yang, Yuanfei Wang, Yaoyu Feng, and Lihua Xiao. 2017. "Subtype Analysis of Zoonotic Pathogen Cryptosporidium Skunk Genotype." *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 55 (November): 20–

25.

Yildirim, Suleyman, Carl J. Yeoman, Maksim Sipos, Manolito Torralba, Brenda A. Wilson, Tony L. Goldberg, Rebecca M. Stumpf, Steven R. Leigh, Bryan A. White, and Karen E. Nelson. 2010. “Characterization of the Fecal Microbiome from Non-Human Wild Primates Reveals Species Specific Microbial Communities.” PloS One 5 (11): e13963.

Zou, Yuanqiang, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, et al. 2019. “1,520 Reference Genomes from Cultivated Human Gut Bacteria Enable Functional Microbiome Analyses.” Nature Biotechnology 37 (2): 179–85.

# Chapter 5. Additional published articles

In this Chapter, I report other published works that I contributed to during my doctoral studies and that included me as co-author. These articles are all related with the three works introduced in **Chapters 2**, **3**, and **4**; the article in **5.1** is a follow up on the mother-to-infant transmission investigation enabled by my work in **Chapter 3** but applied on a larger cohort; the article in **5.2** is a study investigating uncharacterized members of the human microbiome on which my investigation of the non-human primate microbiome in **Chapter 4** is based; in **5.3** I report two studies to which I contributed marginally and include a clinical description of the cohort subjected to whole-genome sequencing in **Chapter 2** and an article on the gut microbiome in colorectal cancer and its predictive power. Each article is introduced by a short paragraph on the rationale of the study, how it is linked with the main works I presented in the previous Chapters, and my role in the research. For each study, I report only the abstract, whereas the full article is linked at the end of the brief introduction.

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

This work published in Cell Host & Microbe in 2018 can be seen as a larger-scale application of the methodology presented and validated in the pilot study fully reported in **Chapter 3** on the vertical transmission of microbiome members from mother to infant. In this article, we enrolled a larger cohort of mothers and infants that were followed for four months after birth with samples collected from multiple body sites. Overall, we found evidence for vertical transmission of multiple strains and species, with skin and vaginal microbiome of the mother seeding the infant gut only transiently and gut strains being more prone to long-term colonization.

**Contribution.** In this study, I personally contributed with my expertise on metagenomic vertical microbiome transmission gained with the study in **Chapter 3** (both experimental and analytical side) and on the validation and interpretation of the data.

The abstract is reported below, the full-text article is available here https://doi.org/10.1016/j.chom.2018.06.005 .

**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

^ these authors contributed equally

*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.

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

This article for which I am co-second author has been published in Cell in 2019, and is the methodological basis for the work presented in **Chapter 4** about non-human primate metagenomes. This is to date the largest-scale assembly effort available and allowed the investigation of yet-to-be-characterized bacterial species. In this study, our group assembled over 150,000 genomes from 9,428 publicly available human metagenomes to reconstruct the unexplored human microbiome diversity. Reconstructed genomes were clustered together into 4,930 species-level genome bins (SGBs) spanning 5% genetic distance, with only a small fraction of these species being previously described. Over 70% of the SGBs had indeed no reference genome available, and were defined as “unknown species-level genome bins” (uSGBs), and represent previously undescribed microbial clades. The addition of uSGBs greatly expanded under investigated phyla and improved the mappability of human gut metagenomes to over 87%. In **Chapter 4**, I extended the framework and applied it on non-human primate to better characterized the microbiome of our closest evolutionary relatives.

**Contribution.** In this study, I contributed with part of the phylogenetic analysis, with the biological interpretation of the results and with the writing of the article. Among others, we reconstructed the phylogeny and the functional profile of the most prevalent uSGB (3,376 genomes reconstructed from many different geographical locations), which was phylogenetically placed between *Faecalibacterium* and *Ruminococcus*, which we named *Candidatus* Cibiobacter qucibialis. We highlighted an overall phylogenetic and functional similarity between genomes from geographically-distinct non-Westernized populations, which were generally different from those recovered from Westernized ones, suggesting that both geography and lifestyle have a role in shaping the human microbiome. We performed similar analysis also on other microbial taxa that were differentially present in Westernized and non-Westernized populations, and also in this case I contributed with the biological interpretation. I personally surveyed the functional traits associated with the investigated clades and found examples of bacterial functions associated with specific worldwide populations. Overall, this work was a very multidisciplinary collaboration in which I was involved from the beginning to the end.

The abstract is reported below, the full-text article is available here https://doi.org/10.1016/j.cell.2019.01.001 .

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

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

^ these authors contributed equally

*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 (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 2.85M genes in SGBs, many associated with conditions including infant development (94k) or Westernization (106k). 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.

## 5.3 Other works

In this chapter, I report two other published articles I was involved in my doctoral studies and to which I contributed marginally. Also in this case I report the abstract and a brief commentary highlighting my contribution.

**Methicillin-resistant *Staphylococcus aureus* eradication in cystic fibrosis patients: A randomized multicenter study**

Daniela Dolce, Stella Neri, Laura Grisotto, Silvia Campana, Novella Ravenni, Francesca Miselli, Erica Camera, Lucia Zavataro, Cesare Braggion, Ersilia V. Fiscarelli, Vincenzina Lucidi, Lisa Cariani, Daniela Girelli, Nadia Faelli, Carla Colombo, Cristina Lucanto, Mariangela Lombardo, Giuseppe Magazzù, Antonella Tosco, Valeria Raia, Serena Manara, Edoardo Pasolli, Federica Armanini, Nicola Segata, Annibale Biggeri, Giovanni Taccetti

*PloS one* 2019

This work published in PloS one in 2019 is linked with the study reported in **Chapter 2** regarding *S. aureus* epidemiology and characterization through whole-genome sequencing of a nosocomial cohort of patients. The article reported in this section is analyzing the efficacy of methicillin-resistant *S. aureus* (MRSA) eradication treatment in a set of Cystic Fibrosis (CF) patients enrolled from five CF Referral Centers from a purely clinical viewpoint. MRSA clearance rate was compared between patients subjected to early eradication treatment and controls. Although some cases of spontaneous clearance were observed in the control group, early eradication treatment showed an improvement in MRSA eradication in the treated group.

**Contribution.** In this work, I personally contributed with the *in silico* sequence typing (STs) and typing of the *spa* gene that were then used to characterize the specific clones and with *in silico* SCC*mec* and Panton-Valentine Leukocidin typing to confirm the results obtained by PCR analysis performed in the laboratory.

The abstract is reported below, the full-text article is available here https://doi.org/10.1371/journal.pone.0213497 .

### Abstract

**Background.** Few studies, based on a limited number of patients using non-uniform therapeutic protocols, have analyzed Methicillin-resistant *Staphylococcus aureus* (MRSA) eradication. **Methods.** In a randomized multicenter trial conducted on patients with new-onset MRSA infection, we evaluated the efficacy of an early eradication treatment (arm A) compared with an observational group (B). Arm A received oral rifampicin and trimethoprim/sulfamethoxazole (21 days). Patients' microbiological status, FEV1, BMI, pulmonary exacerbations and use of antibiotics were assessed. **Results.** Sixty-one

patients were randomized. Twenty-nine (47.5%) patients were assigned to active arm A and 32 (52.5%) patients to observational arm B. Twenty-nine (47.5%) patients, 10 patients in arm A and 19 in arm B, dropped out of the study. At 6 months MRSA was eradicated in 12 (63.2%) out of 19 patients in arm A while spontaneous clearance was observed in 5 (38.5%) out of 13 patients in arm B. A per-protocol analysis showed a 24.7% difference in the proportion of MRSA clearance between the two groups ($z = 1.37$, $P(Z>z) = 0.08$). Twenty-seven patients, 15 (78.9%) out of 19 in arm A and 12 (92.3%) out of 13 in arm B, were able to perform spirometry. The mean (±SD) FEV1 change from baseline was 7.13% (±14.92) in arm A and -1.16% (±5.25) in arm B ($p = 0.08$). In the same period the BMI change (mean ±SD) from baseline was 0.54 (±1.33) kg/m2 in arm A and -0.38 (±1.56) kg/m2 in arm B ($p = 0.08$). At 6 months no statistically significant differences regarding the number of pulmonary exacerbations, days spent in hospital and the use of antibiotics were observed between the two arms. **Conclusions.** Although the statistical power of the study is limited, we found a 24.7% higher clearance of MRSA in the active arm than in the observational arm at 6 months. Patients in the active arm A also had favorable FEV1 and BMI tendencies.

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

Andrew Maltez Thomas, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, Serena Manara, Nicolai Karcher, Chiara Pozzi, Sara Gandini, Davide Serrano, Sonia Tarallo, Antonio Francavilla, Gaetano Gallo, Mario Trompetto, Giulio Ferrero, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Jakob Wirbel, Petra Schrotz-King, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork, Georg Zeller, Francesca Cordero, Emmanuel Dias-Neto, João Carlos Setubal, Adrian Tett, Barbara Pardini, Maria Rescigno, Levi Waldron, Alessio Naccarati & Nicola Segata

*Nature Medicine* 2019

This article has been published in Nature Medicine in 2019. From the biological viewpoint, it is not directly linked with any of the studies reported in previous Chapters but is applying similar approaches like the ones used in the pilot study on vertical transmission of the microbiome presented in **Chapter 3**. One of the major contributions of this study is the development of machine-learning models able to distinguish patients with carcinomas from controls. Meta-analysis of publicly available and newly sequenced colorectal cancer (CRC) cohorts showed that microbes associated with the gut microbiome in CRC are consistent across different cohorts and studies and identified a link between CRC microbiome and choline degradation pathways.

**Contribution.** In this study, I contributed with the designing of the qPCR analysis for the quantification of the choline TMA-lyase gene *cutC* transcript abundance differences in

CRC and controls and with the interpretation of the results, especially on the study of the functional potential of the species of interest.

The abstract is reported below, the full-text article is available here https://doi.org/10.1038/s41591-019-0405-7 .


## Abstract

Several studies have investigated links between the gut microbiome and colorectal cancer (CRC), but questions remain about the replicability of biomarkers across cohorts and populations. We performed a meta-analysis of five publicly available datasets and two new cohorts and validated the findings on two additional cohorts, considering in total 969 fecal metagenomes. Unlike microbiome shifts associated with gastrointestinal syndromes, the gut microbiome in CRC showed reproducibly higher richness than controls ($P < 0.01$), partially due to expansions of species typically derived from the oral cavity. Meta-analysis of the microbiome functional potential identified gluconeogenesis and the putrefaction and fermentation pathways as being associated with CRC, whereas the stachyose and starch degradation pathways were associated with controls. Predictive microbiome signatures for CRC trained on multiple datasets showed consistently high accuracy in datasets not considered for model training and independent validation cohorts (average area under the curve, 0.84). Pooled analysis of raw metagenomes showed that the choline trimethylamine-lyase gene was overabundant in CRC ($P = 0.001$), identifying a relationship between microbiome choline metabolism and CRC. The combined analysis of heterogeneous CRC cohorts thus identified reproducible microbiome biomarkers and accurate disease-predictive models that can form the basis for clinical prognostic tests and hypothesis-driven mechanistic studies.

# Chapter 6. Conclusions of the Thesis

Characterization of microorganisms at the level of single strains is of the foremost importance for exposing relevant genomic traits of both host-associated pathogens and commensals. Given the limitations of isolation-based methods in surveying cultivation-recalcitrant microbes, novel cultivation-free methods are required to assess both known microbes that cannot be easily isolated and potentially yet-to-be-characterized ones. Strain-level metagenomics and metagenomic assembly are promising but yet under-explored approaches to expose the strain-level variability of characterized and uncharacterized species.

To verify the hypothesis that strain-level metagenomics can achieve the needed resolution to complement and possibly expand isolation-based approaches to characterize, track and discover host-associated microbes, I applied different strain-level analysis methods ranging from whole-genome isolate sequencing of a known opportunistic pathogen to the reconstruction of uncharacterized microbial strains from non-human primate metagenomes.

In **Chapter 2** I showed how whole-genome isolate sequencing can reach a very deep resolution, that coupled with appropriate computational analysis allows the identification of single strain variants, genetic plasticity and gene dispensability, and the tracking of potential outbreaks even in large cohorts. Our work pinpointed the underestimated epidemiological and genetic complexity of *S. aureus* and support the need for larger unbiased WGS-based studies to reconstruct the overall strain diversity of this opportunistic pathogen. We moreover proposed WGS to be used as a routine tool for pathogen surveillance and to inform clinical practice.

In **Chapter 3** I presented and validated a metagenomic approach to track the transmission of microbes from mother to infant, and more in general to characterize and follow single strains across individuals. In the same study, I also showed how metatranscriptomics can be extremely informative on the adaptation of single strains to new environments, as in the case of a strain seeded from the mother in the infant's gut. This study has greatly contributed to the field by introducing and validating a framework of analysis that has later been applied to larger-cohort studies.

In **Chapter 4** I showed how metagenomic assembly allows the study of microbial strains associated with non-human primates and the expansion of the catalog of species to include also previously uncharacterized ones. In the same work, I also demonstrated how this approach allows comparative genomic and functional analysis of specific taxa to expose host-microbe interactions. This work contributes to the field of strain-level metagenomics by expanding the catalog of species surveyed in non-human primate gut microbiomes, which greatly increases metagenomes mappability, therefore allowing more representative studies of their microbiome diversity.

In **Chapter 5** I reported other works in line with those presented in the previous chapters, all focused on the importance of studying microbes at the level of single strains to better understand their relationships with the host.

Some approaches on topics similar to those presented in this thesis have also been proposed very recently by other groups, for instance by contributing to complementary investigations of the mother-to-infant microbiome transmission (Miyoshi et al. 2017; Wampach et al. 2017; Cabral et al. 2017; Ximenez and Torres 2017; Davenport et al. 2017; Yassour et al. 2018; Wampach et al. 2018; Vatanen et al. 2018; Korpela and de Vos 2018; Ferretti et al. 2018) or by mining uncharacterized microbial diversity through metagenomic assembly applied on more limited datasets (Parks et al. 2017; Nayfach et al. 2019; Almeida et al. 2019). Although these studies were performed after we established the general framework (for the vertical microbiome transmission) on smaller datasets, they contributed in raising the awareness of the relevance of strain-level metagenomic profiling.

Overall, this thesis supports the crucial role of strain-level analysis and specifically of strain-level metagenomics for expanding our knowledge of host-associated microbial diversity. Through the investigation of both pathogenic and commensal microbial strains associated with human and non-human hosts, in the studies reported in this work we showed how different strain-level analysis methods result into a better understanding of microbe-host relationships. Although cultivation assays remain indispensable for a number of microbiology tasks including (but not limited to) high-throughput phenotype screening (Typas et al. 2008; Kritikos et al. 2017; Galardini et al. 2017; Zou et al. 2019), our work and related efforts are paving the way to scale multiple types of strain-level investigations to the size of many thousands samples for hundreds microbiome members.

## 6.1 Outlook and future works

The continuously decreasing costs of sequencing technologies coupled with their increasing depth are now providing a large amount of metagenomic data that could be meta-analysed for extracting previously unseen microbial diversity, toward a complete picture of human and non-human microbiomes. This, in turn, will foster the application of strain-level comparative genomics approaches to resolve host-microbe interactions at a previously inconceivable scale, with possible future biomedical implications.

Understanding how microbial strains are transmitted among human populations is for example extremely informative not only for pathogens but also for commensal species. If the relevance of tracking pathogenic microbes is more linked to public health concerns, a deeper understanding of how commensal microbiome members disseminate across populations with different lifestyles would expose relevant host-microbiome coevolution patterns, possibly suggesting new ways to hinder the loss of microbiome biodiversity in industrialized countries. Explaining whether this loss is linked with the lower exposure to non-human microbial sources, such as farmed co-habitant animals that are instead more common among non-Westernized populations, would first require characterizing the mostly unexplored microbiome diversity of these understudied non-human hosts.

Metagenomic assembly would therefore represent a first necessary step to solve the microbial dark matter in under-investigated host metagenomes to enable further comparative genomic analysis to reveal how direct and indirect transmission of microbes occurs across different hosts and give better insights into the role of host-to-host transmission barriers in fostering microbial adaptation and evolution.

The combination of large-scale strain-level metagenomics and de novo genome assembly efforts similar to those described in this thesis could hence expose a large microbial diversity inaccessible through cultivation-based methods and greatly expand our view of the human and non-human microbiome and its mechanistic relationship with the host. However, strain-level metagenomics is at its early steps, and some limitations still have to be addressed. Marker-based metagenomic approaches are indeed extremely powerful in tracking lowly abundant microbes, but have the caveat of not using genome-wide information and being limited to already characterized microbes. Viceversa, assembly-based metagenomic methods exploit the information obtained from the whole genome but are typically able to reconstruct only ten-to-twenty high-quality MAGs per sample, usually only for those microbes that are well represented in the microbiome. Additionally, the presence of multiple strains of the same species in the same metagenome can cause co-assembly of multiple strains and consequent incorrect genomes (Wu, Simmons, and Singer 2016; Pasolli et al. 2019). Moreover, the quality of the assembly is strongly dependent on the sequencing depth and errors (Alneberg et al. 2014; Quince et al. 2017). Nevertheless, improvements in the analysis algorithms could only partially solve these problems, because in most cases they are due to biological causes (e.g. long repeated regions or inter-species conserved regions). However, all of these limitations could be overcome by complementing the widely used short-reads technologies with the most recent and increasingly available long-reads sequencing methods (Kuleshov et al. 2016; Bishara et al. 2018) and by improving cultivation assays, in order to obtain full closed genomes of specific strains and possibly reconstruct multiple strains from the same metagenomic sample (Mukherjee et al. 2017; Parks et al. 2017; Zou et al. 2019).

After the conclusion of my doctoral studies, I would thus like to focus on some of these lines of research by applying both the biological and computational knowledge I acquired in the past four years as a PhD candidate to study under-investigated niches of the human and non-human microbiome and possibly contribute to solve biomedically-relevant host-microbe interactions.

# 7. References of the Thesis


Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. “A New Genomic Blueprint of the Human Gut Microbiota.” *Nature* 568 (7753): 499–504.

Alm, R. A., L. S. Ling, D. T. Moir, B. L. King, E. D. Brown, P. C. Doig, D. R. Smith, et al. 1999. “Genomic-Sequence Comparison of Two Unrelated Isolates of the Human Gastric Pathogen Helicobacter Pylori.” *Nature* 397 (6715): 176–80.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. “Binning Metagenomic Contigs by Coverage and Composition.” *Nature Methods* 11 (11): 1144–46.

Atherton, J. C., R. M. Peek Jr, K. T. Tham, T. L. Cover, and M. J. Blaser. 1997. “Clinical and Pathological Importance of Heterogeneity in vacA, the Vacuolating Cytotoxin Gene of Helicobacter Pylori.” *Gastroenterology* 112 (1): 92–99.

Bäckhed, Fredrik, Ruth E. Ley, Justin L. Sonnenburg, Daniel A. Peterson, and Jeffrey I. Gordon. 2005. “Host-Bacterial Mutualism in the Human Intestine.” *Science* 307 (5717): 1915–20.

Bishara, Alex, Eli L. Moss, Mikhail Kolmogorov, Alma E. Parada, Ziming Weng, Arend Sidow, Anne E. Dekas, Serafim Batzoglou, and Ami S. Bhatt. 2018. “High-Quality Genome Sequences of Uncultured Microbes by Assembly of Read Clouds.” *Nature Biotechnology*, October. https://doi.org/10.1038/nbt.4266.

Blaser, M. J. 1997. “Not All Helicobacter Pylori Strains Are Created Equal: Should All Be Eliminated?” *The Lancet* 349 (9057): 1020–22.

Cabral, Damien J., Jenna I. Wurster, Myrto E. Flokas, Michail Alevizakos, Michelle Zabat, Benjamin J. Korry, Aislinn D. Rowan, et al. 2017. “The Salivary Microbiome Is Consistent between Subjects and Resistant to Impacts of Short-Term Hospitalization.” *Scientific Reports* 7 (1): 11040.

Clemente, Jose C., Luke K. Ursell, Laura Wegener Parfrey, and Rob Knight. 2012. “The Impact of the Gut Microbiota on Human Health: An Integrative View.” *Cell* 148 (6): 1258–70.

Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, et al. 1998. “Deciphering the Biology of Mycobacterium Tuberculosis from the Complete Genome Sequence.” *Nature* 393 (6685): 537–44.

Cuevas-Ramos, Gabriel, Claude R. Petit, Ingrid Marcq, Michèle Boury, Eric Oswald, and Jean-Philippe Nougayrède. 2010. “Escherichia Coli Induces DNA Damage in Vivo and Triggers Genomic Instability in Mammalian Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (25): 11537–42.

Davenport, Emily R., Jon G. Sanders, Se Jin Song, Katherine R. Amato, Andrew G. Clark, and Rob Knight. 2017. “The Human Microbiome in Evolution.” *BMC Biology* 15 (1): 127.

De Filippis, Francesca, Edoardo Pasolli, Adrian Tett, Sonia Tarallo, Alessio Naccarati, Maria De Angelis, Erasmo Neviani, et al. 2019. “Distinct Genetic and Functional Traits of Human

Intestinal Prevotella Copri Strains Are Associated with Different Habitual Diets." *Cell Host & Microbe* 25 (3): 444–53.e3.

Durbán, Ana, Juan J. Abellán, Nuria Jiménez-Hernández, Alejandro Artacho, Vicente Garrigues, Vicente Ortiz, Julio Ponce, Amparo Latorre, and Andrés Moya. 2013. "Instability of the Faecal Microbiota in Diarrhoea-Predominant Irritable Bowel Syndrome." *FEMS Microbiology Ecology* 86 (3): 581–89.

Ferretti, Pamela, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Valentina Gorfer, Sabina Fedi, Federica Armanini, et al. 2018. "Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome." *Cell Host & Microbe* 24 (1): 133–45.e5.

Frank, Christina, Dirk Werber, Jakob P. Cramer, Mona Askar, Mirko Faber, Matthias an der Heiden, Helen Bernard, et al. 2011. "Epidemic Profile of Shiga-Toxin-Producing Escherichia Coli O104:H4 Outbreak in Germany." *The New England Journal of Medicine* 365 (19): 1771–80.

Fraser, C. M., S. J. Norris, G. M. Weinstock, O. White, G. G. Sutton, R. Dodson, M. Gwinn, et al. 1998. "Complete Genome Sequence of Treponema Pallidum, the Syphilis Spirochete." *Science* 281 (5375): 375–88.

Galardini, Marco, Alexandra Koumoutsi, Lucia Herrera-Dominguez, Juan Antonio Cordero Varela, Anja Telzerow, Omar Wagih, Morgane Wartel, et al. 2017. "Phenotype Inference in an Escherichia Coli Strain Panel." *eLife*. https://doi.org/10.7554/elife.31035.

Haiser, Henry J., David B. Gootenberg, Kelly Chatman, Gopal Sirasani, Emily P. Balskus, and Peter J. Turnbaugh. 2013. "Predicting and Manipulating Cardiac Drug Inactivation by the Human Gut Bacterium Eggerthella Lenta." *Science* 341 (6143): 295–98.

HMP, Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, et al. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (June): 207.

Korpela, Katri, and Willem M. de Vos. 2018. "Early Life Colonization of the Human Gut: Microbes Matter Everywhere." *Current Opinion in Microbiology* 44 (August): 70–78.

Kritikos, George, Manuel Banzhaf, Lucia Herrera-Dominguez, Alexandra Koumoutsi, Morgane Wartel, Matylda Zietek, and Athanasios Typas. 2017. "A Tool Named Iris for Versatile High-Throughput Phenotyping in Microorganisms." *Nature Microbiology* 2 (February): 17014.

Kuleshov, Volodymyr, Chao Jiang, Wenyu Zhou, Fereshteh Jahanbani, Serafim Batzoglou, and Michael Snyder. 2016. "Synthetic Long-Read Sequencing Reveals Intraspecies Diversity in the Human Microbiome." *Nature Biotechnology* 34 (1): 64–69.

Kuroda, M., T. Ohta, I. Uchiyama, T. Baba, H. Yuzawa, I. Kobayashi, L. Cui, et al. 2001. "Whole Genome Sequencing of Meticillin-Resistant Staphylococcus Aureus." *The Lancet* 357 (9264): 1225–40.

Miyoshi, Jun, Alexandria M. Bobe, Sawako Miyoshi, Yong Huang, Nathaniel Hubert, Tom O. Delmont, A. Murat Eren, Vanessa Leone, and Eugene B. Chang. 2017. "Peripartum Antibiotics Promote Gut Dysbiosis, Loss of Immune Tolerance, and Inflammatory Bowel Disease in Genetically Prone Offspring." *Cell Reports* 20 (2): 491–504.

Mukherjee, Supratim, Rekha Seshadri, Neha J. Varghese, Emiley A. Eloe-Fadrosh, Jan P. Meier-

Kolthoff, Markus Göker, R. Cameron Coates, et al. 2017. “1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life.” *Nature Biotechnology* 35 (7): 676–83.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. 2019. “New Insights from Uncultivated Genomes of the Global Human Gut Microbiome.” *Nature* 568 (7753): 505–10.

Nelson, Karen E., Robert D. Fleischmann, Robert T. DeBoy, Ian T. Paulsen, Derrick E. Fouts, Jonathan A. Eisen, Sean C. Daugherty, et al. 2003. “Complete Genome Sequence of the Oral Pathogenic Bacterium Porphyromonas Gingivalis Strain W83.” *Journal of Bacteriology* 185 (18): 5591–5601.

Nielsen, H. Bjørn, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, et al. 2014. “Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes.” *Nature Biotechnology* 32 (8): 822–28.

Pallen, Mark J., and Brendan W. Wren. 2007. “Bacterial Pathogenomics.” *Nature* 449 (7164): 835–42.

Palm, Noah W., Marcel R. de Zoete, and Richard A. Flavell. 2015. “Immune-Microbiota Interactions in Health and Disease.” *Clinical Immunology* 159 (2): 122–27.

Parkhill, J., B. W. Wren, K. Mungall, J. M. Ketley, C. Churcher, D. Basham, T. Chillingworth, et al. 2000. “The Genome Sequence of the Food-Borne Pathogen Campylobacter Jejuni Reveals Hypervariable Sequences.” *Nature* 403 (6770): 665–68.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. “Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life.” *Nature Microbiology* 2 (11): 1533–42.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.” *Cell* 176 (3): 649–62.e20.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464 (7285): 59–65.

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. “A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes.” *Nature* 490 (7418): 55–60.

Quince, Christopher, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. 2017. “DESMAN: A New Tool for de Novo Extraction of Strains from Metagenomes.” *Genome Biology* 18 (1): 181.

Scher, Jose U., Andrew Sczesnak, Randy S. Longman, Nicola Segata, Carles Ubeda, Craig Bielski, Tim Rostron, et al. 2013. “Expansion of Intestinal Prevotella Copri Correlates with Enhanced Susceptibility to Arthritis.” *eLife* 2 (November): e01202.

Schloissnig, Siegfried, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, et al. 2013. “Genomic Variation Landscape of the Human Gut Microbiome.” *Nature* 493 (7430): 45–50.

Segata, Nicola. 2018. “On the Road to Strain-Resolved Comparative Metagenomics.” *mSystems* 3 (2). https://doi.org/10.1128/mSystems.00190-17.

Stecher, Bärbel, and Wolf-Dietrich Hardt. 2011. “Mechanisms Controlling Pathogen Colonization of the Gut.” *Current Opinion in Microbiology* 14 (1): 82–91.

Stewart, Eric J. 2012. “Growing Unculturable Bacteria.” *Journal of Bacteriology* 194 (16): 4151–60.

Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. Brinkman, et al. 2000. “Complete Genome Sequence of Pseudomonas Aeruginosa PAO1, an Opportunistic Pathogen.” *Nature* 406 (6799): 959–64.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. “Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes.” *Genome Research* 27 (4): 626–38.

Typas, Athanasios, Robert J. Nichols, Deborah A. Siegele, Michael Shales, Sean R. Collins, Bentley Lim, Hannes Braberg, et al. 2008. “High-Throughput, Quantitative Analyses of Genetic Interactions in E. Coli.” *Nature Methods* 5 (9): 781–87.

Vatanen, Tommi, Damian R. Plichta, Juhi Somani, Philipp C. Münch, Timothy D. Arthur, Andrew Brantley Hall, Sabine Rudolf, et al. 2018. “Genomic Variation and Strain-Specific Functional Adaptation in the Human Gut Microbiome during Early Life.” *Nature Microbiology*, December. https://doi.org/10.1038/s41564-018-0321-5.

Ventura, Marco, Sarah O’Flaherty, Marcus J. Claesson, Francesca Turroni, Todd R. Klaenhammer, Douwe van Sinderen, and Paul W. O’Toole. 2009. “Genome-Scale Analyses of Health-Promoting Bacteria: Probiogenomics.” *Nature Reviews. Microbiology* 7 (1): 61–71.

Vogtmann, Emily, and James J. Goedert. 2016. “Epidemiologic Studies of the Human Microbiome and Cancer.” *British Journal of Cancer* 114 (3): 237–42.

Wampach, Linda, Anna Heintz-Buschart, Joëlle V. Fritz, Javier Ramiro-Garcia, Janine Habier, Malte Herold, Shaman Narayanasamy, et al. 2018. “Birth Mode Is Associated with Earliest Strain-Conferred Gut Microbiome Functions and Immunostimulatory Potential.” *Nature Communications* 9 (1): 5091.

Wampach, Linda, Anna Heintz-Buschart, Angela Hogan, Emilie E. L. Muller, Shaman Narayanasamy, Cedric C. Laczny, Luisa W. Hugerth, et al. 2017. “Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life.” *Frontiers in Microbiology* 8 (May): 434.

Ward, Doyle V., Matthias Scholz, Moreno Zolfo, Diana H. Taft, Kurt R. Schibler, Adrian Tett, Nicola Segata, and Ardythe L. Morrow. 2016a. “Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. Coli in Necrotizing Enterocolitis and Mortality in Preterm Infants.” *Cell Reports* 14 (12): 2912–24.

Wu, Yu-Wei, Blake A. Simmons, and Steven W. Singer. 2016. “MaxBin 2.0: An Automated Binning Algorithm to Recover Genomes from Multiple Metagenomic Datasets.” *Bioinformatics* 32 (4): 605–7.

Ximenez, Cecilia, and Javier Torres. 2017. “Development of Microbiota in Infants and Its Role in Maturation of Gut Mucosa and Immune System.” *Archives of Medical Research* 48 (8): 666–80.

Yassour, Moran, Eeva Jason, Larson J. Hogstrom, Timothy D. Arthur, Surya Tripathi, Heli Siljander, Jenni Selvenius, et al. 2018. “Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life.” *Cell Host & Microbe* 24 (1): 146–54.e4.

Zeller, Georg, Julien Tap, Anita Y. Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I. Costea, Aurélien Amiot, et al. 2014. “Potential of Fecal Microbiota for Early-stage Detection of Colorectal Cancer.” *Molecular Systems Biology* 10 (11): 766.

Zhang, Xuan, Dongya Zhang, Huijue Jia, Qiang Feng, Donghui Wang, Di Liang, Xiangni Wu, et al. 2015. “The Oral and Gut Microbiomes Are Perturbed in Rheumatoid Arthritis and Partly Normalized after Treatment.” *Nature Medicine* 21 (8): 895–905.

Zou, Yuanqiang, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, et al. 2019. “1,520 Reference Genomes from Cultivated Human Gut Bacteria Enable Functional Microbiome Analyses.” *Nature Biotechnology* 37 (2): 179–85.