UNIVERSITY OF TRENTO - Italy
Department CIBIO

International PhD Program in Biomolecular Sciences
Department of Cellular, Computational
and Integrative Biology – CIBIO
XXXIV Cycle

Metagenomics-based strain-resolved bacterial genomics and transmission dynamics of the human microbiome

Tutors

Prof. Nicola Segata

CIBIO, University of Trento

External Advisor

Dr. Georg Zeller

EMBL Heidelberg

Co-Advisor

Dr. Mireia Valles-Colomer

CIBIO, University of Trento

Ph.D. Thesis of

Nicolai Karcher

Academic Year 2020 - 2021

Declaration

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

“*If you don't like bacteria, you're on the wrong planet.*”

- Stewart Brand

Abstract

The human gut microbiome is home to many hundreds of different microbes which play a crucial role in human physiology. For most of them, little is known about how their genetic diversity translates into functional traits and how they interact with their host, which is to some extent due to the lack of isolate genomes. Cultivation-free metagenomic approaches yield extensive amounts of bacterial genetic data, and recently developed algorithms allow strain-level resolution and reconstruction of bacterial genomes from metagenomes, yet bacterial within-species diversity and transmission dynamics after fecal microbiota transplantation remain largely unexplored over cohorts and using these technological advances. To investigate bacterial within-species diversity I first undertook large-scale exploratory studies to characterize the population-level genomic makeup of the two key human gut microbes *Eubacterium rectale* and *Akkermansia muciniphila*, leveraging many hundreds of bacterial draft genomes reconstructed from short-read shotgun metagenomics datasets from all around the planet. For *E. rectale*, I extended previous observations about clustering of subspecies with geography, which suggested isolation by distance and the putative ancestral loss of four distinct motility operons, rendering a subspecies specifically found in Europe immotile. For *A. muciniphila,* I found that there are several closely related but undescribed *Akkermansia* spp. in the human gut that are all likely human-specific but are differentially associated with host body mass index, showcasing metabolic differences and distinct co-abundance patterns with putative cognate phages. For both species, I discovered distinct subspecies-level genetic variation in structural polysaccharide synthesis operons. Next, utilizing a complementary strain-resolved approach to track strains between individuals, I undertook a fecal microbiota transplantation (FMT) meta-analysis integrating 24 distinct clinical metagenomic datasets. I found that patients with an infectious disease or those who underwent antibiotic treatment displayed increased donor strain uptake and that some bacterial clades engraft more consistently than others. Furthermore, I developed a machine-learning framework that allows optimizing microbial parameters - such as bacterial richness - in the recipient after FMT based on donor microbiome features, representing first steps towards making a rational donor choice. Taken together, in my work I extended the strain-level understanding of human gut commensals and showcased that genomes from metagenomes can be suitable to conduct large-scale bacterial population genetics studies on other understudied human gut commensals. I further confirmed that strain-resolved metagenomics allows tracking of strains and thus inference of strain engraftment characteristics in an FMT meta-analysis, revealing important differences in engraftment over cohorts and species and paving the way towards better designed FMTs. I believe that my work is an important contribution to the field of microbiome research, showcasing the power of shotgun metagenomics, modern algorithms and large-scale data analysis to reveal previously unattainable insights about the human gut microbiome.

Index

List of Abbreviations

AUROC: area under the receiver-operating curve
CARD: comprehensive antibiotics resistance database
CAZy: carbohydrate-active enzyme
CDS: coding sequence
CRC: colorectal cancer
CRISPR: clustered regularly interspaced short palindromic repeats
EPS: exopolysaccharide
FMT: fecal microbiota transplantation
GI: genomic island
IBD: inflammatory bowel disease
KEGG: Kyoto encyclopedia of Genes and Genomes
KO: KEGG orthology
kSGB: known SGB
LODO: leave-one-dataset-out
LPS: lipopolysaccharide
MAG: metagenome-assembled genome
MDRB: multidrug resistant bacteria
OLS: ordinary least squares
PAM: partitioning around medoids
PCoA: principle coordinate analysis
rCDI: recurrent C. difficile infection
RF: random forest
SGB: species-level group
SNV: single nucleotide variant
SQ: sulfoquinovose
TLR4: Toll-like receptor 4
t-SNE: t-distributed stochastic neighbor embedding
uSGB: unknown SGB
VC: viral cluster

Chapter 1 | Introduction and aims

1.1 | Introduction

Prokaryotes are thought to be the oldest and most ubiquitous group of life on earth. They can be found virtually everywhere, thriving in the crushing pressures and heat of hydrothermal vents (Dick 2019), in hypersaline lakes (Naghoni et al. 2017), and in arctic ice sheets (Anesio et al. 2017). They are also found in abundance in soil (Fierer 2017), in all organs of plants (Trivedi et al. 2020), and presumably in association with all domains of macroscopic animals: in sponges and corals (Thomas et al. 2016), insects (Engel and Moran 2013) and other invertebrates. Finally, they live in and on mammals, including humans where they play pivotal physiological roles: among other things, bacteria aid their host in metabolism, synthesizing amino-acids and fermenting resistant carbohydrates to feed host enterocytes (Oliphant and Allen-Vercoe 2019). Host-associated bacteria protect from invading gastrointestinal pathogens (Wu and Wu 2012) by inhabiting ecological niches, while at the same time regulating the immune system to maintain immune homeostasis by helping it distinguish between commensal and pathogenic bacteria (Wu and Wu 2012; Zheng, Liwinski, and Elinav 2020; Levy, Thaiss, and Elinav 2016).

Typically, prokaryotes do not live in isolation. Instead, they interact with each other and their environment in a multitude of ways; sustaining each other in intricate nutritional networks or, conversely, competing with or even preying on each other. In this way, they form interconnected and diverse ecological communities: for example, the total number of distinct microbial species associated with the human body ecosystem has been estimated to be almost 5,000 (Pasolli et al. 2019). Similarly, the number of bacterial species detected in the ocean exceeds 35,000 (Sunagawa et al. 2015), the true number most probably greatly surpassing this estimate. These communities of prokaryotes and other microbes have been termed microbiomes or microbiota.

Until only a few decades ago, examination of microbiomes remained confined largely to characterization of single species in isolation because the technological means to study these communities holistically were not available. At the end of the 1970s, Carl Woese discovered that the sequence of the bacterial 16S rRNA gene can be used to determine the relatedness of bacteria. This was a seminal leap in microbial ecology since it allowed researchers to contextualize the relatedness of *any* microbe (including previously unknown ones) while also revolutionizing our understanding of the tree of life by showing the existence of the three domains (Bacteria, Archaea, Eukaryote) that we are familiar with nowadays (Woese and Fox 1977). Yet, researchers were still not able to study the genetic diversity of microbiomes as a whole because this requires

determining the sequence and abundance of hundreds of different kinds of 16S variants at the same time. High-throughput DNA sequencing technologies developed towards the end of the 20th century filled this gap, allowing researchers to sequence massive amounts of DNA at once. By sequencing all 16S rRNA molecules in a sample simultaneously, researchers were then able to survey the genetic diversity of microbial communities which lead to important insights about the microbiome's response to antibiotics treatment and the association of microbial composition with diet and age (Dethlefsen et al. 2008; Ley et al. 2005; Hopkins, Sharp, and Macfarlane 2002), to name only a few examples.

While 16S rRNA gene sequencing gives insight into the genetic makeup of microbial communities and in theory can contain enough information to delineate microbial species, its phylogenetic resolution is limited in practice due to only partial coverage of the gene using short-read sequencing technology, which can be problematic given the phenotypic differences observed in strains of some bacterial species: most *Escherichia coli* strains are harmless members of the human gut, but certain strains are enteropathogens with severe clinical manifestation, and strains can become virulent with only very minor genetic alterations (Proença, Barral, and Gordo 2017). Other gastrointestinal *E. coli* strains were shown to become genotoxic, a feature encoded by a genomic island merely 50.000 bases long (Pleguezuelos-Manzano et al. 2020). Similarly, *Salmonella enterica* strains have been shown to differ in their host range, their antibiotic resistance profile and their pathogenicity (Fricke et al. 2011). Strain-level phenotypic differences are slowly beginning to emerge also for many non-pathogenic, human-associated gut bacteria: comprehensive *in vitro* bacterial screens have shown that *Bacteroides* spp. strains of the same species differ in their antibiotic susceptibility (Maier et al. 2021). While these studies provide detailed understanding of strain-level phenotypic differences, they are labour-intensive and thus do not scale well to the number of species and strains in the human gut.

The limited resolution of 16S rRNA sequencing was overcome with the development of shotgun metagenomics, which, instead of targeting hypervariable regions of the 16S rRNA gene for sequencing, enables sequencing of all DNA molecules in a community indiscriminately. Naturally, shotgun metagenomics yields a lot more genetic information which triggered the development of shotgun metagenomic microbial profiling tools that leveraged either marker genes such as MetaPhlAn, mOTUs and MetaPhyler (Truong et al. 2015; Milanese et al. 2019; Liu et al. 2011) or whole-genome read mapping/k-mer matching such as Kraken/Bracken and CLARK (Wood and Salzberg 2014; Lu et al. 2017; Ounit et al. 2015) allowing for high-resolution, species-level microbial profiling. Furthermore, tools such as Humann2 and PanPhlan (Franzosa et al. 2018; Scholz et al. 2016) were developed to analyze the entirety of the genetic material in microbial

communities from a functional perspective, mapping it to protein databases to delineate the functional repertoire of communities and species therein. These developments led to discoveries such as the virtually limitless bacterial genetic repertoire in the human gut as well as the large fraction of functionally uncharacterized proteins (Human Microbiome Project Consortium 2012; J. Qin et al. 2010; N. Qin et al. 2014), the association of distinct bacterial species with diseases such as colorectal cancer (Zeller et al. 2014) and inflammatory bowel disease (Morgan et al. 2012; Rooks et al. 2014) as well as the enrichment of a *Ruminococcus gnavus* subclade in inflammatory bowel disease (Hall et al. 2017).

Soon after establishing shotgun metagenomics as the go-to approach to study microbial communities at high resolution, researchers started re-exploring the diversity of bacterial communities including the human gut microbiome with this new technology. They soon realized that the wealth of sequencing information generated is large enough to understand bacterial genetic diversity beyond the species-level: in one of the first studies aiming to understand the subspecies genetic variability in the human gut from newly emerged shotgun sequencing technology, Schloissnig et al. mapped metagenomic reads to bacterial reference genomes of around 100 species to quantify their within-species genetic diversity, structural variation landscape, and the selective pressures acting on them (Schloissnig et al. 2013). More recently, other studies and methods were developed that allowed subspecies-level delineation of strain core gene sequences (those genes present in nearly all strains of a species (Truong et al. 2017; Luo et al. 2015; Albanese and Donati 2017; Costea et al. 2017). Since these studies also typically utilized thousands of publicly available metagenomes, they yielded high-level insights into the community ecology and population genetics of human gut bacteria: for each species, single strains typically strongly dominate gut communities and they differ in their genetic coherence with some species showing distinct clusters in subspecies genetic space, while others are more continuous. However, these approaches are fundamentally limited to well-characterized bacteria since they rely on a comprehensive reference genome set, and even if reference information is sufficiently available, understanding genomic structure (gene neighborhood/synteny) and larger structural variations is difficult.

Bacterial genomes can be used to conduct comparative genomics studies, and previous works have capitalized on them to get a better understanding of strain-level genetic structure and genome-wide gene content differences, antibiotic- and virulence gene profiles and evolutionary trajectories of specific human-associated bacteria such as *Escherichia coli* (Touchon et al. 2009; Rasko et al. 2008), *Staphylococcus aureus* (Manara et al. 2018; Booth et al. 2001; Suzuki et al. 2012; Bosi et al. 2016), *Streptococcus pneumoniae* (Donati et al. 2010), *Klebsiella pneumoniae* (Holt et al. 2015) and *Salmonella enterica* (Fricke et al. 2011; Thomson et al. 2008). These studies

have utilized previously developed ideas from the then relatively young field of pangenomics, observing the distribution of gene families over strains (core- and accessory genes for common and rare genes, respectively) and the openness of a species' entire genetic pool when more and more strains are considered (Medini et al. 2005). These studies mostly focused on pathogens since they are naturally well represented in reference databases due to their clinical relevance, as opposed to isolate genomes of gut commensals which are, despite recent 'culturomics' efforts (Browne et al. 2016; Forster et al. 2019), notoriously underrepresented in reference databases (Truong et al. 2017).

Metagenomics also allows reconstructing genomes from complex bacterial communities, building on algorithms to assemble short reads into longer DNA sequences (Nurk et al. 2017; Kang et al. 2019; Bishara et al. 2018). These genome-like groups of DNA sequences have been termed metagenome-assembled genomes (MAGs). The release of an ever-growing number of publicly available shotgun metagenomes has sparked an explosive rise in the number of available MAGs: in 2019, three large-scale assembly efforts yielded a total of more than 200,000 MAGs assembled from many thousands of gut shotgun metagenomes (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019), dwarfing the number of isolate genomes available for most human gut commensals (Browne et al. 2016; Forster et al. 2019). These MAGs represent a large and hitherto untapped resource to conduct comparative genomics and strain-level analysis of important but understudied human gut commensals.

Strain-level delineation of bacteria is not only a powerful approach to better functionally characterize and genetically contrast strains of bacterial species, but also a means to understand transmission patterns of bacteria: metagenomic studies have shown that the human gut is most probably initially sterile (Perez-Muñoz et al. 2017) and is being 'inoculated' mostly by vaginal and skin-residing strains from the mother during natural birth (Ferretti et al. 2018; Reyman et al. 2019; Dominguez-Bello et al. 2010; Podlesny and Fricke 2021). From then on, the strain repertoire of the human gut is not static, but can change over time as we retain, lose and take up strains from the environment, typically from co-housing or otherwise physically close individuals (Koo et al. 2019; Brito et al. 2019).

Conceptually, eradication of disease-driving strains and introduction of health-promoting strains through external intervention is a promising idea in the medical context. Hence, understanding strain transmission dynamics is also of growing interest in the context of fecal microbiota transplantations (FMT): FMT is a medical procedure referring to the transfer of stool from a healthy individual into the gastrointestinal tract of a patient. It is indicated primarily for recurrent *Clostridium difficile* infections not responding to

antibiotic treatment. Recently, FMT is also being explored for other gastrointestinal conditions such as inflammatory bowel disease (Suskind et al. 2015; Vaughn et al. 2016; Damman et al. 2015; Kong et al. 2020), multi-drug resistant bacterial infection (Bar-Yoseph et al. 2021; Leo et al. 2020) and others (non-response to immune-checkpoint inhibition in cancer (Baruch et al. 2021; Davar et al. 2021), irritable bowel syndrome (Goll et al. 2020), metabolic syndrome (Li et al. 2016)).

Despite the evident medical relevance, understanding how microbial communities in the recipient develop after FMT is still not well understood and current donor selection protocols do not take into account the wide variety of gut microbial differences that are exhibited among healthy individuals that could potentially affect the medical outcome of FMTs. Many studies including metagenomic shotgun sequencing of microbial DNA in stool from donors and recipients have been conducted in the context of FMT to examine strain engraftment into the recipient (Suskind et al. 2015; Vaughn et al. 2016; Damman et al. 2015; Kong et al. 2020; Bar-Yoseph et al. 2021; Leo et al. 2020; Baruch et al. 2021; Davar et al. 2021; Goll et al. 2020; Li et al. 2016; Moss et al. 2017; Aggarwala et al. 2021; Hourigan et al. 2019; Koopen et al. 2021; Kumar et al. 2017; Podlesny and Fricke 2020; Smillie et al. 2018; Verma et al. 2021; Watson et al. 2021; Zhao et al. 2020; Ianiro et al. 2020). However, all these studies were small and encompassed only one medical condition, limiting their power to investigate differences between cohorts and conditions, links between microbial features and clinical success and the association of donor features with desirable microbiome traits post-FMT.

1.2 | Aims

In this work, I set out to contribute to a detailed understanding of the genetic and functional diversity and transmission characteristics of human gut commensal strains. To this end, I relied on both recent technological advances to obtain high-quality metagenome-assembled genomes (MAGs) as well as methods to identify and track strains in metagenomic samples.

My aims were as follows:

1. To assess whether MAGs can be used to conduct comparative genomics studies similarly to what has previously been done using isolate genomes.
2. To delineate the genetic, genomic, functional, and evolutionary landscape of select species of the human gut microbiome that could not be studied previously due to only few isolate genomes being available.
3. To assess how strains engraft and persist after fecal microbiota transplantation (FMT) in different conditions.

Human gut commensals are severely underrepresented in isolate databases, but comprehensive human gut MAG collections were produced recently (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019, 2021) and benchmarks suggest that MAGs from human gut microbiomes can approach the quality of isolate genomes (Sczyrba et al. 2017). Owing to the lack of (isolate) genomes until very recently, human gut commensals have not been examined in large-scale, genome-based comparative genomics studies despite strain-level differences in bacteria potentially being crucial. Hence, I conducted two directed, MAG-based comparative genomics studies on the two key human gut microbiome members *Eubacterium rectale* and *Akkermansia muciniphila* (**Aims 1 and 2**).

Furthermore, several studies using shotgun metagenomic sequencing have been performed in the context of FMT. While some of those were employed to compare strains between donor and recipient microbiomes and understand how strains from the donor engraft in the recipient and persist after FMT, many studies were only sufficiently powered to assess clinical outcomes and did not focus on microbiome composition assessment, limiting the conclusions that can be drawn from strain-based analysis. Thus, I conducted an FMT meta-analysis comprising a total of 24 distinct datasets over several disease conditions to study how strains engraft and persist in the recipient after allogeneic FMT in different conditions (**Aim 3**).

1.3 | Structure

Chapter 1 contains the introduction to the field of strain-resolved microbiome research as well as the aims of the work.

Chapter 2 contains the manuscript "**Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations**" (N. Karcher *et al.*, Genome Biology), which represents the first large-scale comparative genomics and population genetics study conducted on a human gut microbe based on MAGs.

Chapter 3 contains the manuscript "**Genomic diversity and ecology of human-associated *Akkermansia* species in the gut microbiome revealed by extensive metagenomic assembly**" (N. Karcher, E. Nigro *et al.*, Genome Biology), which sheds light onto uncharacterized *Akkermansia* spp. in the human gut.

Chapter 4 contains the manuscript "**Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases**" (G. Ianiro, M. Punčochář, N. Karcher *et al.,* in review at

Nature Medicine), which explores differences in strain engraftment and persistence after FMT.

Chapter 5 describes contributions I have made to research projects not directly related to my thesis.

Chapter 6 recapitulates my work, contextualizes its findings in the field of microbiome research and illustrates potential future research directions.

1.4 | Contributions

In Chapters 2, 3 and 4 I was involved to varying degrees in conceptualization, data gathering, analysis, interpretation and writing. I have not been involved in cohort recruitment, sampling and data generation (i.e. sequencing).

1.5 | References for the introduction

Albanese, Davide, and Claudio Donati. 2017. “Strain Profiling and Epidemiology of Bacterial Species from Metagenomic Sequencing.” *Nature Communications* 8 (1): 1–14.

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*, February. https://doi.org/10.1038/s41586-019-0965-1.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. “A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome.” *Nature Biotechnology* 39 (1): 105–14.

Anesio, Alexandre M., Stefanie Lutz, Nathan A. M. Chrismas, and Liane G. Benning. 2017. “The Microbiome of Glaciers and Ice Sheets.” *Npj Biofilms and Microbiomes* 3 (1): 1–11.

Baruch, Erez N., Ilan Youngster, Guy Ben-Betzalel, Rona Ortenberg, Adi Lahat, Lior Katz, Katerina Adler, et al. 2021. “Fecal Microbiota Transplant Promotes Response in Immunotherapy-Refractory Melanoma Patients.” *Science* 371 (6529): 602–9.

Bar-Yoseph, Haggai, Shaqed Carasso, Shlomit Shklar, Alexander Korytny, Razi Even Dar, Haneen Daoud, Roni Nassar, et al. 2021. “Oral Capsulized Fecal Microbiota Transplantation for Eradication of Carbapenemase-Producing Enterobacteriaceae Colonization With a Metagenomic Perspective.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 73 (1): e166–75.

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.

Booth, M. C., L. M. Pence, P. Mahasreshti, M. C. Callegan, and M. S. Gilmore. 2001. “Clonal Associations among Staphylococcus Aureus Isolates from Various Sites of Infection.” *Infection and Immunity* 69 (1): 345–52.
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.
Brito, Ilana L., Thomas Gurry, Shijie Zhao, Katherine Huang, Sarah K. Young, Terrence P. Shea, Waisea Naisilisili, et al. 2019. “Transmission of Human-Associated Microbiota along Family and Social Networks.” *Nature Microbiology* 4 (6): 964–71.
Browne, Hilary P., Samuel C. Forster, Blessing O. Anonye, Nitin Kumar, B. Anne Neville, Mark D. Stares, David Goulding, and Trevor D. Lawley. 2016. “Culturing of ‘Unculturable’ Human Microbiota Reveals Novel Taxa and Extensive Sporulation.” *Nature* 533 (7604): 543–46.
Costea, Paul I., Luis Pedro Coelho, Shinichi Sunagawa, Robin Munch, Jaime Huerta-Cepas, Kristoffer Forslund, Falk Hildebrand, Almagul Kushugulova, Georg Zeller, and Peer Bork. 2017. “Subspecies in the Global Human Gut Microbiome.” *Molecular Systems Biology* 13 (12): 960.
Damman, Christopher J., Mitchell J. Brittnacher, Maria Westerhoff, Hillary S. Hayden, Matthew Radey, Kyle R. Hager, Sara R. Marquis, Samuel I. Miller, and Timothy L. Zisman. 2015. “Low Level Engraftment and Improvement Following a Single Colonoscopic Administration of Fecal Microbiota to Patients with Ulcerative Colitis.” *PloS One* 10 (8): e0133925.
Davar, Diwakar, Amiran K. Dzutsev, John A. McCulloch, Richard R. Rodrigues, Joe-Marc Chauvin, Robert M. Morrison, Richelle N. Deblasio, et al. 2021. “Fecal Microbiota Transplant Overcomes Resistance to Anti-PD-1 Therapy in Melanoma Patients.” *Science* 371 (6529): 595–602.
Dethlefsen, Les, Sue Huse, Mitchell L. Sogin, and David A. Relman. 2008. “The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing.” *PLoS Biology* 6 (11): e280.
Dick, Gregory J. 2019. “The Microbiomes of Deep-Sea Hydrothermal Vents: Distributed Globally, Shaped Locally.” *Nature Reviews. Microbiology* 17 (5): 271–83.
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.
Engel, Philipp, and Nancy A. Moran. 2013. “The Gut Microbiota of Insects - Diversity in Structure and Function.” *FEMS Microbiology Reviews* 37 (5): 699–735.
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.
Fierer, Noah. 2017. “Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome.” *Nature Reviews. Microbiology* 15 (10): 579–90.

Forster, Samuel C., Nitin Kumar, Blessing O. Anonye, Alexandre Almeida, Elisa Viciani, Mark D. Stares, Matthew Dunn, et al. 2019. “A Human Gut Bacterial Genome and Culture Collection for Improved Metagenomic Analyses.” *Nature Biotechnology* 37 (2): 186–92.

Fricke, W. Florian, Mark K. Mammel, Patrick F. McDermott, Carmen Tartera, David G. White, J. Eugene Leclerc, Jacques Ravel, and Thomas A. Cebula. 2011. “Comparative Genomics of 28 Salmonella Enterica Isolates: Evidence for CRISPR-Mediated Adaptive Sublineage Evolution.” *Journal of Bacteriology* 193 (14): 3556–68.

Goll, Rasmus, Peter Holger Johnsen, Erik Hjerde, Joseph Diab, Per Christian Valle, Frank Hilpusch, and Jorunn Pauline Cavanagh. 2020. “Effects of Fecal Microbiota Transplantation in Subjects with Irritable Bowel Syndrome Are Mirrored by Changes in Gut Microbiome.” *Gut Microbes* 12 (1): 1794263.

Holt, Kathryn E., Heiman Wertheim, Ruth N. Zadoks, Stephen Baker, Chris A. Whitehouse, David Dance, Adam Jenney, et al. 2015. “Genomic Analysis of Diversity, Population Structure, Virulence, and Antimicrobial Resistance in Klebsiella Pneumoniae, an Urgent Threat to Public Health.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (27): E3574–81.

Hopkins, M. J., R. Sharp, and G. T. Macfarlane. 2002. “Variation in Human Intestinal Microbiota with Age.” *Digestive and Liver Disease: Official Journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver* 34 Suppl 2 (September): S12–18.

Human Microbiome Project Consortium. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486 (7402): 207–14.

Kang, Dongwan, Feng Li, Edward S. Kirton, Ashleigh Thomas, Rob S. Egan, Hong An, and Zhong Wang. 2019. “MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies.” e27522v1. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27522v1.

Kong, Lingjia, Jason Lloyd-Price, Tommi Vatanen, Philippe Seksik, Laurent Beaugerie, Tabassome Simon, Hera Vlamakis, Harry Sokol, and Ramnik J. Xavier. 2020. “Linking Strain Engraftment in Fecal Microbiota Transplantation With Maintenance of Remission in Crohn’s Disease.” *Gastroenterology* 159 (6): 2193–2202.e5.

Koo, Hyunmin, Joseph A. Hakim, David K. Crossman, Elliot J. Lefkowitz, and Casey D. Morrow. 2019. “Sharing of Gut Microbial Strains between Selected Individual Sets of Twins Cohabitating for Decades.” *PloS One* 14 (12): e0226111.

Leo, Stefano, Vladimir Lazarevic, Myriam Girard, Nadia Gaïa, Jacques Schrenzel, Victoire de Lastours, Bruno Fantin, et al. 2020. “Metagenomic Characterization of Gut Microbiota of Carriers of Extended-Spectrum Beta-Lactamase or Carbapenemase-Producing Enterobacteriaceae Following Treatment with Oral Antibiotics and Fecal Microbiota Transplantation: Results from a Multicenter Randomized Trial.” *Microorganisms* 8 (6). https://doi.org/10.3390/microorganisms8060941.

Levy, Maayan, Christoph A. Thaiss, and Eran Elinav. 2016. “Metabolites: Messengers between the Microbiota and the Immune System.” *Genes & Development* 30 (14): 1589–97.

Ley, Ruth E., Fredrik Bäckhed, Peter Turnbaugh, Catherine A. Lozupone, Robin D. Knight, and Jeffrey I. Gordon. 2005. “Obesity Alters Gut Microbial Ecology.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (31): 11070–75.

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.

Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J. Xavier, and Dirk Gevers. 2015. “ConStrains Identifies Microbial Strains in Metagenomic Datasets.” *Nature Biotechnology* 33 (10): 1045–52.

Maier, Lisa, Camille V. Goemans, Jakob Wirbel, Michael Kuhn, Claudia Eberl, Mihaela Pruteanu, Patrick Müller, et al. 2021. “Unravelling the Collateral Damage of Antibiotics on Gut Bacteria.” *Nature* 599 (7883): 120–24.

Manara, Serena, Edoardo Pasolli, Daniela Dolce, Novella Ravenni, Silvia Campana, Federica Armanini, Francesco Asnicar, et al. 2018. “Whole-Genome Epidemiology, Characterisation, and Phylogenetic Reconstruction of Staphylococcus Aureus Strains in a Paediatric Hospital.” *Genome Medicine* 10 (1): 82.

Morgan, Xochitl C., Timothy L. Tickle, Harry Sokol, Dirk Gevers, Kathryn L. Devaney, Doyle V. Ward, Joshua A. Reyes, et al. 2012. “Dysfunction of the Intestinal Microbiome in Inflammatory Bowel Disease and Treatment.” *Genome Biology* 13 (9): R79.

Naghoni, Ali, Giti Emtiazi, Mohammad Ali Amoozegar, Mariana Silvia Cretoiu, Lucas J. Stal, Zahra Etemadifar, Seyed Abolhassan Shahzadeh Fazeli, and Henk Bolhuis. 2017. “Microbial Diversity in the Hypersaline Lake Meyghan, Iran.” *Scientific Reports* 7 (1): 11522.

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*, March. https://doi.org/10.1038/s41586-019-1058-x.

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

Oliphant, Kaitlyn, and Emma Allen-Vercoe. 2019. “Macronutrient Metabolism by the Human Gut Microbiome: Major Fermentation by-Products and Their Impact on Host Health.” *Microbiome* 7 (1): 91.

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.

Perez-Muñoz, Maria Elisa, Marie-Claire Arrieta, Amanda E. Ramer-Tait, and Jens Walter. 2017. “A Critical Assessment of the ‘Sterile Womb’ and ‘in Utero Colonization’ Hypotheses: Implications for Research on the Pioneer Infant Microbiome.” *Microbiome* 5 (1): 48.

Pleguezuelos-Manzano, Cayetano, Jens Puschhof, Axel Rosendahl Huber, Arne van Hoeck, Henry M. Wood, Jason Nomburg, Carino Gurjao, et al. 2020. “Mutational Signature in Colorectal Cancer Caused by Genotoxic Pks+ E. Coli.” *Nature* 580 (7802): 269–73.

Podlesny, Daniel, and W. Florian Fricke. 2021. “Strain Inheritance and Neonatal Gut Microbiota Development: A Meta-Analysis.” *International Journal of Medical Microbiology: IJMM* 311 (3): 151483.
Proença, João T., Duarte C. Barral, and Isabel Gordo. 2017. “Commensal-to-Pathogen Transition: One-Single Transposon Insertion Results in Two Pathoadaptive Traits in Escherichia Coli -Macrophage Interaction.” *Scientific Reports* 7 (1): 4504.
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, Nan, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, et al. 2014. “Alterations of the Human Gut Microbiome in Liver Cirrhosis.” *Nature* 513 (7516): 59–64.
Reyman, Marta, Marlies A. van Houten, Debbie van Baarle, Astrid A. T. M. Bosch, Wing Ho Man, Mei Ling J. N. Chu, Kayleigh Arp, et al. 2019. “Impact of Delivery Mode-Associated Gut Microbiota Dynamics on Health in the First Year of Life.” *Nature Communications* 10 (1): 4997.
Rooks, Michelle G., Patrick Veiga, Leslie H. Wardwell-Scott, Timothy Tickle, Nicola Segata, Monia Michaud, Carey Ann Gallini, et al. 2014. “Gut Microbiome Composition and Function in Experimental Colitis during Active Disease and Treatment-Induced Remission.” *The ISME Journal* 8 (7): 1403–17.
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.
Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. “Critical Assessment of Metagenome Interpretation-a Benchmark of Metagenomics Software.” *Nature Methods* 14 (11): 1063–71.
Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. “Ocean Plankton. Structure and Function of the Global Ocean Microbiome.” *Science* 348 (6237): 1261359.
Suskind, David L., Mitchell J. Brittnacher, Ghassan Wahbeh, Michele L. Shaffer, Hillary S. Hayden, Xuan Qin, Namita Singh, et al. 2015. “Fecal Microbial Transplant Effect on Clinical Outcomes and Fecal Microbiome in Active Crohn’s Disease.” *Inflammatory Bowel Diseases* 21 (3): 556–63.
Suzuki, Haruo, Tristan Lefébure, Paulina Pavinski Bitar, and Michael J. Stanhope. 2012. “Comparative Genomic Analysis of the Genus Staphylococcus Including Staphylococcus Aureus and Its Newly Described Sister Species Staphylococcus Simiae.” *BMC Genomics* 13 (January): 38.
Thomas, Torsten, Lucas Moitinho-Silva, Miguel Lurgi, Johannes R. Björk, Cole Easson, Carmen Astudillo-García, Julie B. Olson, et al. 2016. “Diversity, Structure and

Convergent Evolution of the Global Sponge Microbiome.” *Nature Communications* 7 (June): 11870.
Thomson, Nicholas R., Debra J. Clayton, Daniel Windhorst, Georgios Vernikos, Susanne Davidson, Carol Churcher, Michael A. Quail, et al. 2008. “Comparative Genome Analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 Provides Insights into Evolutionary and Host Adaptation Pathways.” *Genome Research* 18 (10): 1624–37.
Trivedi, Pankaj, Jan E. Leach, Susannah G. Tringe, Tongmin Sa, and Brajesh K. Singh. 2020. “Plant–microbiome Interactions: From Community Assembly to Plant Health.” *Nature Reviews. Microbiology* 18 (11): 607–21.
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.
Vaughn, Byron P., Tommi Vatanen, Jessica R. Allegretti, Aiping Bai, Ramnik J. Xavier, Joshua Korzenik, Dirk Gevers, Amanda Ting, Simon C. Robson, and Alan C. Moss. 2016. “Increased Intestinal Microbial Diversity Following Fecal Microbiota Transplant for Active Crohn’s Disease.” *Inflammatory Bowel Diseases* 22 (9): 2182–90.
Woese, Carl R., and George E. Fox. 1977. “Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (11): 5088–90.
Wu, Hsin-Jung, and Eric Wu. 2012. “The Role of Gut Microbiota in Immune Homeostasis and Autoimmunity.” *Gut Microbes* 3 (1): 4–14.
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 (November): 766.
Zheng, Danping, Timur Liwinski, and Eran Elinav. 2020. “Interaction between Microbiota and Immunity in Health and Disease.” *Cell Research* 30 (6): 492–506.

Chapter 2 | Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations

2.1 | Context and contribution

The first project of my thesis was intended as a proof of concept to understand if quality-controlled MAGs can be used to conduct detailed comparative genomics studies on human gut commensal species (aim 1). We chose *Eubacterium rectale* as a target species because it is frequently and abundantly encountered in the human gastrointestinal tract and communities are regularly dominated by a single strain of this species (Truong et al. 2017), which we anticipated are good conditions to produce a high number of high quality MAGs from thousands of publicly available shotgun metagenomes.

Thus, I generated more than 1,300 high-quality MAGs from 6,500 shotgun metagenomic assemblies using a reference-based binning approach developed for this work. Combining these MAGs with the few available isolate genomes, I delineated the strain-level phylogeny of this species, confirming previous reports of strong biogeographic clustering and extending this observation with a novel subspecies found predominantly in African countries. More detailed analysis of the bacterial phylogeny revealed tentative evidence of isolation by distance in this bacterial species, in turn suggesting human-microbe co-dispersal. Functional characterization of *E. rectale* revealed absence of motility operons due to ancestral loss in a subspecies specific in europe as well as the presence of a large exopolysaccharide synthesis operon of unknown function specifically in some members of this subspecies.

This work represents not only an advance in our understanding of the strain-level population genetics and functional understanding of *E. rectale*, but it shows that MAGs can be suitable for high-resolution, strain-level comparative genomics analysis in the human gut, opening up new avenues for in-depth characterization of other bacterial taxa in the human gut.

For this work, I was involved in conceptualization and have conducted the analysis, interpreted and visualized the results and wrote the manuscript.

2.1 | Manuscript

Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations

Karcher, N.[1], Pasolli[2], Asnicar, F.[1], Huang, K.D.[1,3], Tett, A.[1], Manara, S.[1], Armanini, F.[1], Bain, D.[4], Duncan, S.H.[4], Louis, P.[4], Zolfo, M.[1], Manghi, P.[1], Valles-Colomer, M.[1], Raffaetà, R.[5], Rota-Stabelli, O.[3], Collado, M.C.[6], Zeller, G.[7], Falush, D.[8], Maixner, F.[9], Walker, A. W.[4], Huttenhower, C.[10,11], Segata, N.[12]

Genome Biol. 2020 Jun 8;21(1):138. doi: 10.1186/s13059-020-02042-y.

Affiliations
1 Department CIBIO, University of Trento, Trento, Italy.
2 Department of Agriculture, University of Naples, Naples, Italy.
3 Fondazione Edmund Mach, S. Michele all'Adige, Italy.
4 Rowett Institute, University of Aberdeen, Aberdeen, UK.
5 Free University of Bozen-Bolzano, Bolzano, Italy.
6 IATA-CSIC, Valencia, Spain.
7 EMBL, Heidelberg, Germany.
8 University of Bath, Bath, UK.
9 Institute for Mummy studies, Eurac Research, Bolzano, Italy.
10 Harvard T.H. Chan School of Public Health, Boston, MA, USA.
11 The Broad Institute, Cambridge, MA, USA.
12 Department CIBIO, University of Trento, Trento, Italy. nicola.segata@unitn.it.

Note: The version below is the ahead-of-print version of the manuscript, prior to editorial edits.

Abstract

Background

Eubacterium rectale is one of the most prevalent human gut bacteria, but its diversity and population genetics are not well understood because large-scale whole-genome investigations of this microbe have not been carried out.

Results

Here, we leverage metagenomic assembly followed by a reference-based binning strategy to screen over 6,500 gut metagenomes spanning geography and lifestyle and reconstruct over 1,300 *E. rectale* high-quality genomes from metagenomes. We extend previous results of biogeographic stratification, identifying a new subspecies predominantly found in African individuals and showing that closely related non-human primates do not harbor *E. rectale*. Comparison of pairwise genetic and geographic distances between subspecies suggests that isolation by distance and co-dispersal with human populations might have contributed to shaping the contemporary population structure of *E. rectale*. We confirm that a relatively recently diverged *E. rectale* subspecies specific to Europe consistently lacks motility operons and that it is immotile *in-vitro*, probably due to ancestral genetic loss. The same subspecies exhibits expansion of its carbohydrate metabolism gene repertoire including the acquisition of a genomic island strongly enriched in glycosyltransferase genes involved in exopolysaccharide synthesis.

Conclusions

Our study provides new insights into the population structure and ecology of *E. rectale* and shows that shotgun metagenomes can enable population genomics studies of microbiota members at a resolution and scale previously attainable only by extensive isolate sequencing.

Introduction

The composition of the human gut microbiota is variable across individuals and only few bacterial species are consistently present in populations of different geographic origin and lifestyle. Current large-scale metagenomic surveys (Pasolli et al. 2017) reported that merely three species (*Eubacterium rectale*, *Faecalibacterium prausnitzii*, *Ruminococcus torques*) and few other poorly characterized microbes are detected at >0.1% relative abundance in more than 90% of adult healthy individuals (Pasolli et al. 2017). In a recent study using metagenomic assembly and reference-independent binning, *E. rectale* was the species for which the most genomes from metagenomes could be reconstructed (Pasolli et al. 2019). The large number of publicly available metagenomic cohorts and accurate methods for genome reconstruction from metagenomes thus provide an unprecedented opportunity to gain insights into this otherwise relatively poorly investigated bacterial species using metagenomic data at a global scale.

E. rectale is a member of the *Firmicutes* phylum, belonging to the *Lachnospiraceae* family. The proposed type strain of *E. rectale* (A1-86) is rod-shaped, Gram-positive, strictly anaerobic and motile (Duncan and Flint 2008). *E. rectale* produces butyrate and other short chain fatty acids (SCFAs) from carbohydrates not directly accessible by the host, which play a role in promoting intestinal health in the host (Ríos-Covián et al. 2016). The relative abundance of *E. rectale* in the gut has been reported to be reduced compared to controls in diseases such as cystic fibrosis (Bruzzese et al. 2014), Crohn's disease (Kabeerdoss et al. 2015), ulcerative colitis (Fite et al. 2013) and colorectal cancer (Zeller et al. 2014), suggesting that it is replaced or outcompeted in certain disease states. *E. rectale* is an important gut anaerobe, and it is thus crucial to study its population genetics and strain-level epidemiology.

The population structure of *E. rectale* has been investigated in previous studies (Truong et al. 2017; Scholz et al. 2016; Costea et al. 2017), which have used read-mapping based approaches to study the population level genetics of bacterial commensals from metagenomes. Although these approaches provided valuable insights such as the variable degree of intra-species biogeographic stratification in different species including *E. rectale*, they were not conducted at the resolution of whole genomes. Metagenomic assembly together with reference-free binning has recently been employed in meta-analyses showing that microbial genomes can be consistently reconstructed from metagenomes (Pasolli et al. 2019; Almeida, Mitchell, et al. 2019; Nayfach et al. 2019). However these reference-free binning approaches could miss genomic regions with divergent tetranucleotide frequencies.

In this work, we extracted more than 1,300 high-quality *E. rectale* genomes from more than 6,500 gut metagenomic assemblies using a targeted, reference-based binning

approach that is applicable when at least a few (isolate) genomes are available. This pipeline produced genomes that compare favourably to genomes from a reference-free binning approach. The genomes that were assembled from metagenomes were used for the first large-scale genome-based population-level genomic analysis of *E. rectale* exemplifying how studies typically performed with cultured isolate sequencing data can be performed on carefully quality-controlled genomes from metagenomes. We extended the number of subspecies identified in previous investigations (Truong et al. 2017; Scholz et al. 2016; Costea et al. 2017) by identifying a subspecies predominantly found in African individuals. Comparing median genetic distances to estimated geographic distances between pairs of subspecies indicated that pairs of subspecies are isolated by distance, in turn suggesting host-microbe co-dispersal. Whole-genome functional analysis confirmed the presence of a uniquely non-motile subspecies exhibiting loss of motility associated with a shift in carbohydrate metabolism gene repertoire.

Results and discussion

Reconstruction of >1,300 high-quality *Eubacterium rectale* genomes from >6,500 metagenomes

Metagenomic samples are a rich source for microbial genomes, but reconstructing bacterial genomes from metagenomes with sufficient completeness and accuracy remains challenging. To extract *E. rectale* genomes from metagenomes, we developed a three-step procedure consisting of (i) single sample metagenomic assembly, (ii) compilation of high-quality *E. rectale* reference sequences, and (iii) use of these references to bin the metagenomic assemblies (**Additional File 1: Fig. S1, Methods**). We applied this pipeline on a collection of 6,775 gut metagenomic assemblies obtained from our previous studies (Pasolli et al. 2019; Tett et al. 2019b). These assemblies were generated using metaSPAdes (Nurk et al. 2017) if paired-end reads were available or MegaHIT otherwise (Li et al. 2015). We produced 47 manually-curated reference (MCR) *E. rectale* genomes from 170 assembled metagenomes of diverse geographic origin in which *E. rectale* was particularly highly abundant (**Methods**). These genomes are smaller than genomes obtained from isolate sequencing due to prioritization of specificity over sensitivity in the manually-curated binning process (**Methods**). For the last step of the pipeline, we used the 47 MCR genomes (**Additional File 2: Tab. S1**) together with seven *E. rectale* isolate genomes from NCBI available at the time (**Additional File 3: Tab. S2**) as references for the reference-based binning that was applied to all 6,775 assembled metagenomes (**Methods**). Semi-simulated metagenomic assemblies (**Methods**) allowed us to set optimal parameter values for the binning procedure (**Fig. 1A**).

We found that this pipeline reconstructs *E. rectale* genomes with high fidelity, outperforming reference-free metagenomic binning in terms of completeness (Pasolli et

al. 2019; D. D. Kang et al. 2015) while slightly increasing contamination (1,7% median increase in completeness, 0.5% median increase in contamination) (**Fig. 1D, Fig. 1E**). The pan-genome characteristics of the reconstructed *E. rectale* genomes more closely resemble those of isolate *E. rectale* genomes than the *E. rectale* genomes coming from reference-free binning (**Fig. 1 F, G**), further suggesting that they generally are of high quality.

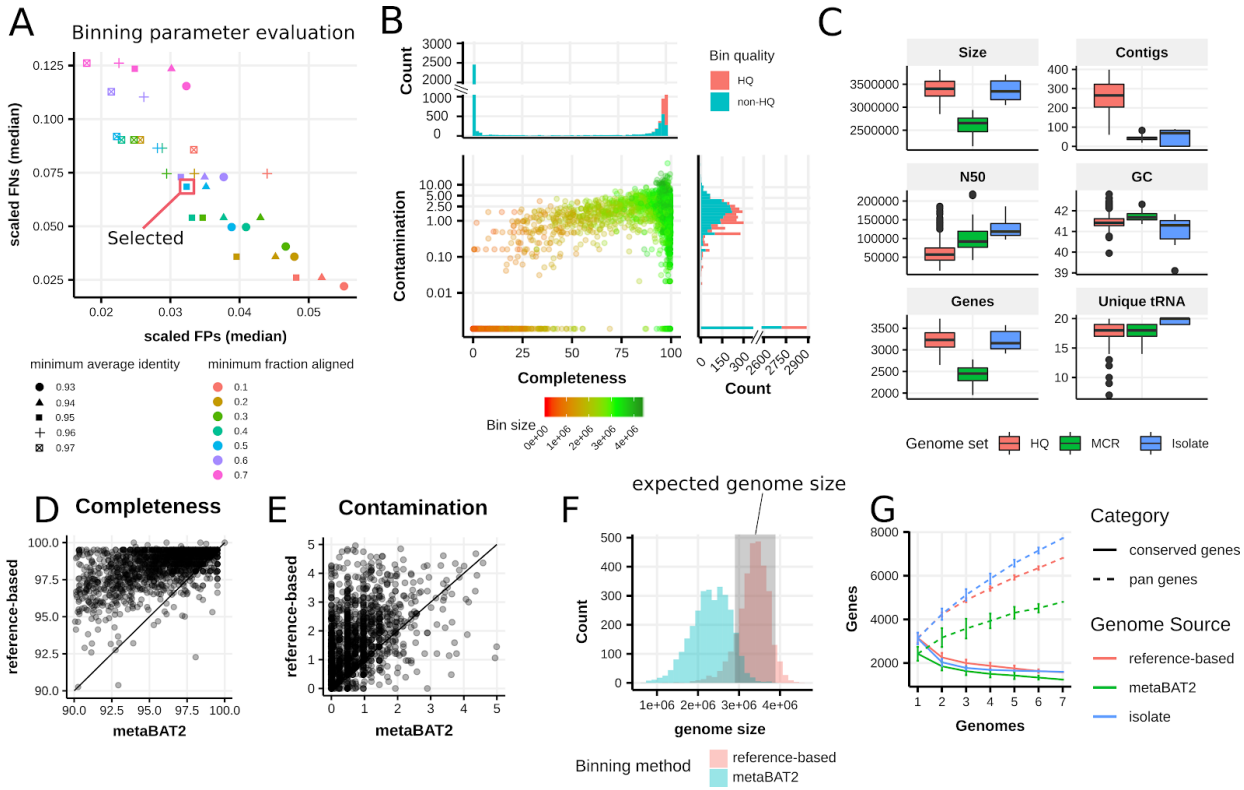


Figure 1: Reconstruction of 1,321 high-quality (HQ) *E. rectale* genomes from 6,775 fecal metagenomes. (**A**) The parameters for the binning step of our reference-based workflow (average identity and fraction of contig aligned) were chosen using *E. rectale*-free metagenomic assemblies spiked with *E. rectale* sequences obtained from isolate genomes (**Methods**). We report the median number of false positive (FP) bases (binned contigs not coming from spike-in) and false negative (FN) bases (contigs coming from spike-in that were not binned). FP and FN values are scaled with respect to the average *E. rectale* isolate genome size. The red square indicates the parameter value combination used in this study. (**B**) Estimation of completeness and contamination for all extracted genomes using CheckM (Parks et al. 2015). (**C**) Comparison of genome characteristics for *E. rectale* isolate genomes, genomes from metagenomes reconstructed with a semi-supervised approach (MCR), and the large set of automatically reconstructed genomes (HQ). **(D, E)** Completeness and contamination estimates for bins extracted using the reference-based binning approach used in this study and bins produced by a reference-independent pipeline using metaBAT2 (Pasolli et al. 2019; D. D. Kang et al. 2015). Only genomes with >90% completeness and <5% contamination in both approaches are shown. **(F)** The sizes of the *E. rectale* genomes reconstructed with the reference-based pipeline are very consistent with the genome sizes (gray area) from cultured isolate sequencing (gray shading) while the reference-independent pipeline produces genomes of smaller size. **(G)** Pangenome characteristics for seven *E.*

rectale isolate genomes from NCBI available at the time of processing (**Additional File 3: Tab. S2**) as well as seven genomes from the reference-based binning and from Pasolli et al. (Pasolli et al. 2019). For both binning methods, we considered the same seven, randomly selected European metagenomes as well as all seven cultured isolate genomes originating from studies in Europe/North America.

We obtained a total of 1,321 high quality (HQ) *E. rectale* genomes by applying our pipeline to a set of 6,613 publicly available gut metagenomic assemblies as well as 162 gut metagenomic assemblies from rural populations in Madagascar and Ethiopia we recently sequenced (Pasolli et al. 2019; Tett et al. 2019a) (**Fig. 1B, Additional File 4: Tab. S3**). The combined cohort of 6,775 gut samples encompasses 38 datasets from 30 countries with samples collected from individuals ranging in age from infants to elderly, and spanning different health conditions and lifestyles (**Additional File 5: Tab. S4**). The 1,321 HQ *E. rectale* genomes contain less than 400 contigs and passed recently proposed completeness and contamination cutoffs (90% and 5% respectively) for high-quality metagenome-assembled genomes (Bowers et al. 2017). In line with recent large-scale metagenomic assembly efforts (Pasolli et al. 2019; Almeida, Mitchell, et al. 2019; Nayfach et al. 2019), we did not consider the presence of tRNA- and rRNA genes as criteria for high quality metagenome-assembled genomes because of the inherent difficulty of reconstructing genes that are conserved across related species (Bowers et al. 2017). The genomes were however further required to pass an additional quality measure we developed based on polymorphic site rates across core genes to flag genomes that are likely to incorporate strain-level variation from more than one strain (**Methods**). The HQ genomes had an average length of 3.39 M bases (s.d. 0.22 M) and an average GC content of 41.47% (s.d. 0.27%), which was consistent with the genomes from isolate sequencing available for this species (**Fig. 1C**). The quality, number, and diverse nature of this combined cohort enabled us to undertake a large-scale genomic investigation of this currently under-characterized gut anaerobe species.

A large-scale phylogeny refines *E. rectale* population structure and association with geography

To get an overview of the *E. rectale* population structure, we first performed a phylogenetic analysis of the 1,321 HQ genomes in combination with eight publicly available cultured isolate genomes and two additional *E. rectale* isolates we sequenced for this work (**Methods**, **Additional File 3: Tab. S2**). The core gene concatenation approach we used (**Methods**) yielded 1,071 core genes and a total alignment length of 1.02M nucleotides. The maximum likelihood phylogeny and the ordination based on this alignment (**Fig. 2A, Fig. 2B)** confirmed previous observations that *E. rectale* strains fall into discrete groups (Truong et al. 2017; Scholz et al. 2016; Costea et al. 2017).

Clustering of core gene genetic distances using Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw 1990) supported the existence of four subspecies (Prediction Strength consistently over 0.8 for k = 4, **Additional File 1: Fig. S2**, **Fig. 2D**, **Methods**), one of which was not observed before (Truong et al. 2017; Scholz et al. 2016; Costea et al. 2017). Three of these four subspecies are large and well-defined monophyletic subtrees in the phylogeny and only a minority of strains of the largest subspecies is falling in divergent paraphyletic subtrees (**Fig. 2A**).

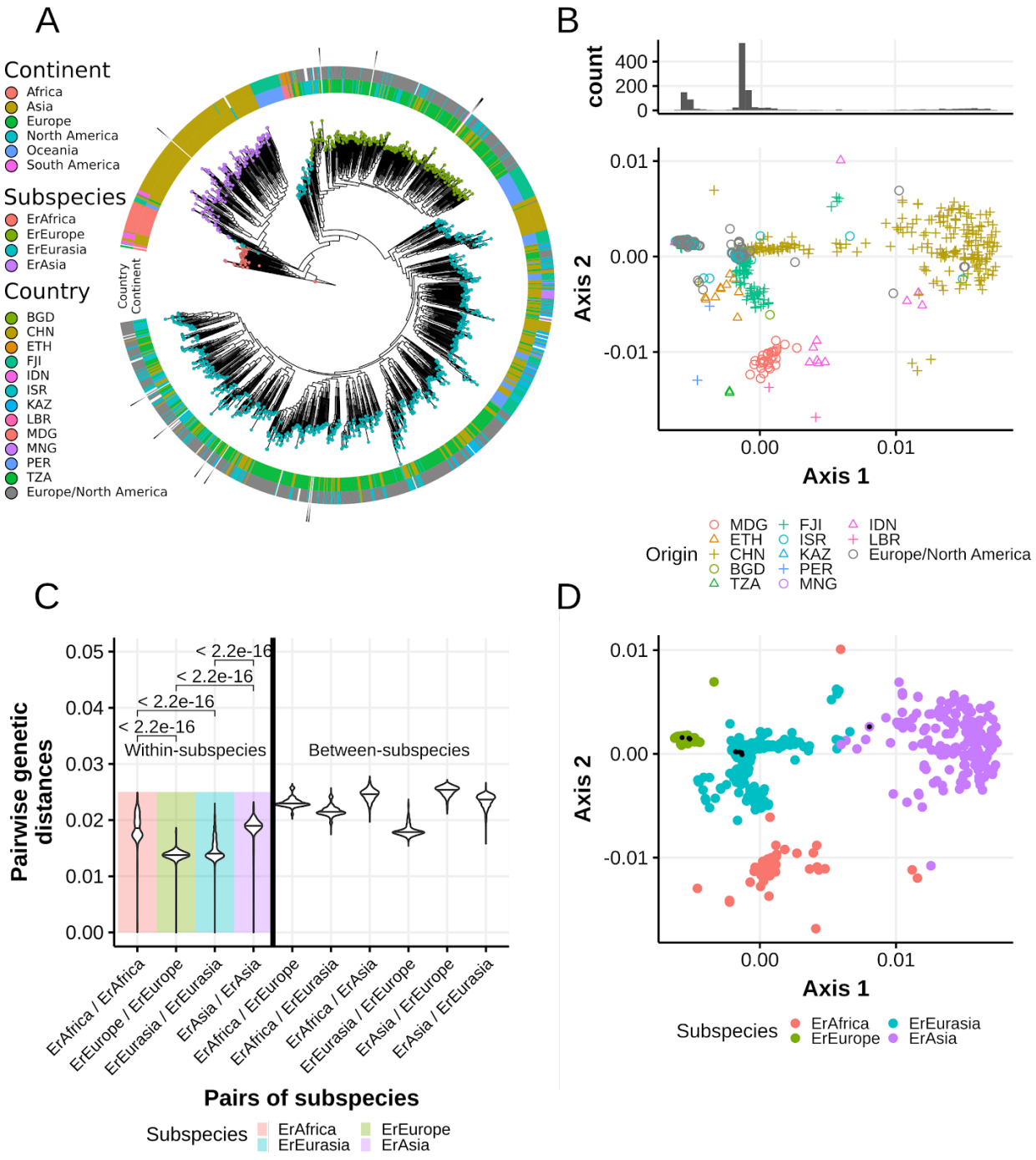


Figure 2: *E. rectale* consists of four geographically stratified subspecies. **(A)** Maximum-likelihood phylogenetic tree of all *E. rectale* genomes, built from a concatenated core gene alignment using PhyloPhlAn2 (**Methods**) and rooted based on a phylogenetic tree including *E. rectale* sister species. **(B)** Non-metric Multidimensional Scaling plot of pairwise genetic distances between all *E. rectale* genomes. **(C)** Distribution of intra- and inter-subspecies core gene genetic distances. P-values were obtained using bidirectional Wilcoxon rank-sum tests. **(D)** Subspecies assignment using PAM clustering with k = 4 (**Methods**). Black points indicate genomes obtained from cultured isolate sequencing.

The four *E. rectale* subspecies showed very strong geographic enrichment and were named accordingly. The three most represented subspecies correspond to what we designated as ErEurasia, ErEurope and ErAsia as they predominantly comprised strains from these regions. ErAfrica, the fourth and previously unobserved subspecies included strains derived mostly from sub-Saharan African countries (Madagascar, Tanzania, Liberia) but also contains strains from Peru and Indonesia (**Fig. 2A, Fig. 2D**). While ErAfrica, ErEurope and ErAsia are geographically relatively well contained, ErEurasia appears to be comparatively widespread (**Additional File 1: Fig. S3**) with strains retrieved from gut metagenomes in Ethiopia and Fiji also belonging to this subspecies, albeit with divergent genetic makeup (**Fig. 2B, Additional File 1: Fig. S4**, **Additional File 1: Fig. S5**). Nonetheless, ErEurasia appears specifically enriched in central/northern Asian countries, with individuals from Kazakhstan, Mongolia and Russia almost exclusively harboring genetically representative ErEurasia strains (**Fig. 3A**, **Additional File 1: Fig. S4**, **Additional File 1: Fig. S6**). While subspecies-specific SNV analysis confirmed that ErEurope and ErEurasia occasionally co-exist, the other subspecies almost never co-colonize (**Additional File 1: Fig. S7, Additional File 1: Fig. S8, Methods**) and thus the geographic distribution inferred from our reconstructed *E. rectale* genomes does not obscure lowly abundant strains. Subspecies membership of the ten strains for which we had isolate genomes is congruent with their putative geographic origin (**Additional File 3: Tab. S2**), and while discrepancies between subspecies assignment and the geographical association of some strains exist (**Additional File 1: Fig. S9**), our data strengthens the notion of geographic stratification in *E. rectale* and provides a first approximation of the population structure of *E. rectale* on a global scale.

Genetic divergence between subspecies confirmed that they should be considered part of the same species as their pairwise genetic dissimilarities are well below 5%, which is the threshold typically used to define bacterial species (Pasolli et al. 2019; Konstantinidis and Tiedje 2005). Indeed, the two most divergent subspecies are ErAsia and ErEurope which are at ~2.5% median genetic distance and no pair of strains ever exceeds 3% genetic distance (**Fig. 2C, Additional File 1: Fig. S10**). Nonetheless, the four subspecies have different intra- and inter-clade genetic variability. Strains belonging to ErEurope and ErEurasia have smaller intra-subspecies genetic variability (1.38% and

1.4% median variability, respectively) compared to ErAsia and ErAfrica (median 1.9% and 1.95%, respectively, **Fig. 2C**). ErEurope and ErEurasia are both the individually least genetically diverse and most closely related pair of subspecies.

E. rectale consists of at least four geographically stratified subspecies, exhibits differential within- and between subspecies genetic variability (**Fig. 2C**) and is found in almost all adult control samples regardless of origin and lifestyle conditions (comprising differential levels of urbanization and sanitation as well as different diets) (**Additional File 1: Fig. S11, Methods**). Altogether, this showed that *Eubacterium rectale* is a globally spread human gut commensal and that the population genetics of *E. rectale* should be studied in light of the evolutionary relationship to its host.

Correlation between subspecies' geographic and genetic distances suggests isolation by distance

An important aspect in investigating the evolutionary relationship between a microbe and its host is the level of host specificity and its transmission patterns. We screened for the presence of *E. rectale* in 146 publicly available metagenomes from wild non-hominid primates as well as 29 metagenomes from wild non-human hominids (**Methods**). We found no evidence for the presence of *E. rectale* in any of these metagenomes using MetaPhlAn2 (**Methods**). Similarly, none of the genomes assembled from non-human metagenomes is closely related (i.e. within 5% genetic distance) to any of the available *E. rectale* genomes. To assess the possibility of inter-individual *E. rectale* strain transmission in human populations, we further analyzed metagenomic data from mother-infant pairs in multiple cohorts (N=532 samples; **Methods**) and found evidence of vertical transmission (25% transmission rate within the first year of infant's life, **Additional File 1: Fig. S12**). Overall, these analyses suggest that *E. rectale* is specific to humans and that it can be transmitted within populations.

Considering the reported specificity of *E. rectale* to humans, the differential degree of relatedness of *E. rectale* subspecies might be due to the effects of isolation by distance (Wright 1943) and we thus tested whether *E. rectale* genetics supports this hypothesis. To this end, we compared median pairwise genetic distances with geographic distances between pairs of subspecies (Diniz-Filho et al. 2013). Owing to sparse sampling outside Europe and the occurrence of ErEurasia and ErAfrica strains outside their ascribed geographic areas, we assigned representative point locations to each subspecies that do not take these outlying strains into account (**Discussion**) (**Methods, Additional File 1: Fig. S14**). Under these approximations, we found a statistically significant correlation (p-value 0.041) between pairwise geographic and median genetic distances of subspecies (**Fig. 3B**) that is confirmed when directly considering pairwise distances

between samples (p-value < 1e-16), suggesting that *E. rectale* genetic stratification could have been to some extent shaped by physical isolation of strains over time.

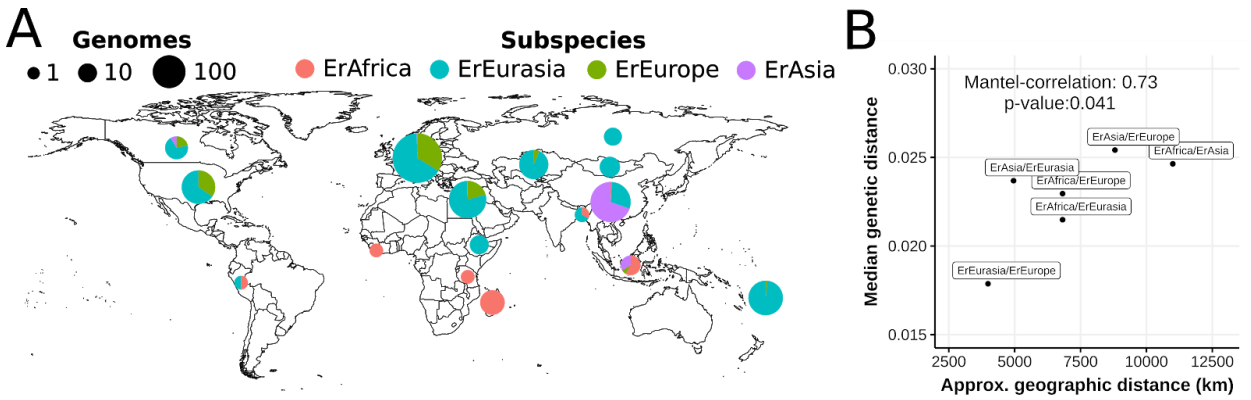


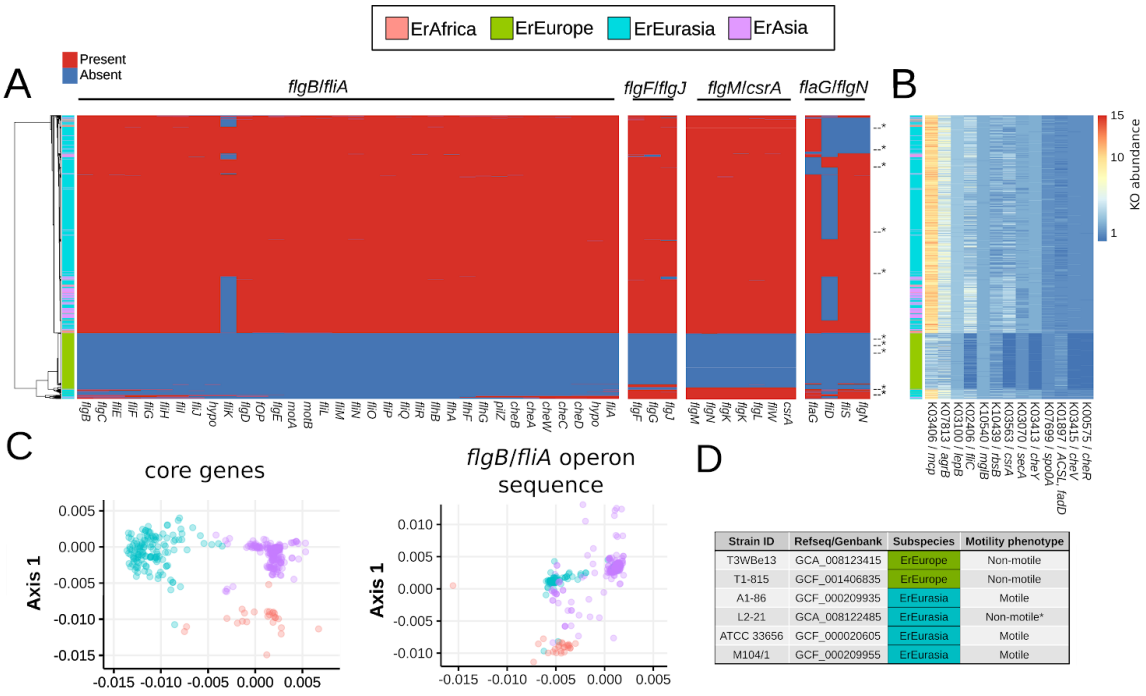
Figure 3: *Eubacterium rectale* subspecies distribution suggests subspecies are isolated by distance. (A) Relative prevalence of *E. rectale* subspecies per country (European countries are aggregated). The size of the pie charts is proportional to the total number of genomes obtained per region/country. For a map of Europe, see **Additional File 1: Fig. S13**. **(B)** Pairwise approximated geographic distances between subspecies (considering representative locations) correlate with their median genetic distances (**Methods** for details). A Mantel test between pairwise genetic and geographic distances using the Pearson correlation coefficient yielded a correlation of 0.73 and a p-value of 0.041.

ErEurope strains are immotile due to loss of motility operons

To analyze the functional repertoires of the *E. rectale* subspecies, we compared the presence and absence of functionally annotated gene clusters across all *E. rectale* genomes. ErEurope was found to be much more functionally divergent from the other subspecies than what genetic data would suggest (**Additional File 1: Fig. S15**). We computed differentially prevalent gene families (KEGG Orthology gene families, KOs) and found that the most distinguishing feature of ErEurope genomes is the absence of a large number of motility-related genes, some of which are part of an operon previously reported to be absent in a group of *E. rectale* strains corresponding to what we called ErEurope (Costea et al. 2017; Neville et al. 2013). Our analysis confirmed that many motility-related genes in *E. rectale* and in closely related species are organized in four operons (Neville et al. 2013) and showed that ErEurope strains consistently and specifically lack all genes of these four motility operons (**Fig. 4A**), whereas the remaining *E. rectale* subspecies largely possess these operons.

To support the hypothesis that these operons are necessary for motility, we performed *in vitro* motility characterization tests in anaerobic conditions on six cultured *E. rectale* strains, two of which were not described before (two ErEurope and four ErEurasia isolates, **Additional File 3: Tab. S2**) and showed that the absence of these motility

operons renders *E. rectale* strains immotile *in-vitro* with a microscopy-based assay of motility (**Fig. 4D**, **Methods**). The vast majority of non-ErEurope strains possess the motility operons, although there are a few exceptions (verified with contig-based analysis, **Additional File 1: Fig. S16**, **Methods**) with 41 non-ErEurope genomes (3.1% of all non-ErEurope strains) lacking >20% of these motility genes (**Fig. 4A**, **Additional File 1: Fig. S17)** and 16 non-ErEurope genomes (1.2%, **Additional File 1: Fig. S18**) specifically missing the largest of the four operons (*flgB*/*fliA*, **Fig. 4A**). Within non-ErEurope strains, the genetic distances inferred from the *flgB*/*fliA* operon are highly correlated with those from the core genome (Pearson correlation of 0.8, Mantel test p-value < 0.001, **Fig. 4C, Additional File 1: Fig. S19**). This suggests past operon/core-genome co-diversification and thus that the common ancestor of all *E. rectale* strains possessed these operons. Motility operons show high structural consistency among *E. rectale* and closely related species, providing additional support for their homologous nature (Neville et al. 2013). Together, we take this as evidence that operon motility loss in *E. rectale* is a stochastic event that can lead to viable, immotile *E. rectale* strains and that the subspeciation of ErEurope might be connected to one of such stochastic operon losses in the common ancestor of ErEurope strains.



Strain ID	Refseq/Genbank	Subspecies	Motility phenotype
T3WBe13	GCA_008123415	ErEurope	Non-motile
T1-815	GCF_001406835	ErEurope	Non-motile
A1-86	GCF_000209935	ErEurasia	Motile
L2-21	GCA_008122485	ErEurasia	Non-motile*
ATCC 33656	GCF_000020605	ErEurasia	Motile
M104/1	GCF_000209955	ErEurasia	Motile

Figure 4: **ErEurope is consistently immotile due to loss of motility operons**. **(A)** No genes from the four motility operons of *E. rectale* (Neville et al. 2013) are detected in ErEurope strains, and only a very small fraction of non-ErEurope genomes are lacking some or all of these genes (**Additional File 1: Fig. S18**). Asterisks denotes cultured isolate genomes. **(B)** Differentially abundant, non-operon potentially motility-associated KOs between ErEurope and the remaining subspecies. *csrA* was added despite being present in the

flgM/*csrA* operon because it can be found elsewhere in some *E. rectale* genomes as well. We annotated genes using eggNOG-mapper (Huerta-Cepas et al. 2017) and only KOs of the *E. rectale* reference genome annotated by KEGG (Kanehisa and Goto 2000) are considered. Potentially motility-associated KOs were defined as being part of at least one of the following KEGG pathways: quorum sensing, bacterial chemotaxis, flagellar assembly and two-component system. P-values were calculated using a two-sided Wilcoxon test and corrected for multiple testing at 5% FDR using the Benjamini-Hochberg method. **(C)** Core gene sequence- and *flgB/fliA* operon sequence genetic clustering for all motile strains (those belonging to either ErAfrica, ErEurasia or ErAsia). **(D)** *In-vitro* motility characterization via phase-contrast microscopy of six *E. rectale* isolates (**Methods**). Asterisk marks Strain L2-21, which is the only immotile ErEurasia strain, presumably as a consequence of the specific lack of the *flgB*/*fliA* motility operon we found in its genomes.

Reduced genome size and increased functional divergence is associated with the loss of motility

Comparison of genome sizes between subspecies suggested that ErEurope strains have lost a considerable amount of genetic material since their split with ErEurasia, its most closely related subspecies. The median genome size of ErEurope is smaller than the median genome size of all other subspecies and 353k bases smaller than that of ErEurasia (**Additional File 1: Fig. S20**). This difference far exceeds the cumulative length of the lost motility operons (mean cumulative size 54.5 kbps, sd 13 kbps., **Additional File 1: Fig. S21**), suggesting a gradual loss of genetic material.

We further investigated the evolutionary trajectories of subspecies by studying the differentiation of their gene repertoire in light of their genetic divergence. The gene distances between ErEurope and ErEurasia were similar to those between other pairs of subspecies (excluding motility operon genes) (**Additional File 1: Fig. S22**), but when normalized by their respective genetic distances, the resulting measure of the rate of functional divergence between ErEurope and ErEurasia strains clearly exceeded that of other pairs of subspecies (**Additional File 1: Fig. S23**), indicating that ErEurope and ErEurasia diverged functionally at an accelerated rate compared to other pairs of subspecies. This could represent adaptive processes triggered by the loss of motility in ErEurope strains.

ErEurope genomes have reduced copy numbers of motility-associated genes that are not part of the four motility operons

We investigated the specific functions that are differentiating ErEurope and ErEurasia strains and found a total of 170 differentially abundant KEGG orthologous families (KOs) (**Additional File 6: Tab. S5**). Among them, we identified 13 KOs that were potentially motility-associated but were not found on any of the four motility operons (except for *csrA*, which can be found on a motility operon but also elsewhere in some genomes). 12 of these 13 KOs are underrepresented in ErEurope strains (**Fig. 4B)**.

The 12 out-of-operon, potentially motility-associated genes with reduced copy numbers in ErEurope comprised genes coding for proteins such as methyl-accepting chemotaxis protein (Mcp), flagellin (FliC) as well as the chemotaxis proteins CheR, CheY and CheV (several other chemotaxis genes can be found on the *flgb*/*fliA* operon), which are directly involved in motility. This group also contained genes coding for proteins involved in cellular mechanisms that are indirectly related to motility, such as the accessory gene regulator B (*agrB*) that was shown to be involved in quorum sensing in *Staphylococcus aureus* (Zhang et al. 2002), and the carbon storage regulator A (*csrA*) that is involved in biofilm formation in *E. coli (Liu et al. 1997; Jackson et al. 2002)* as well as quorum sensing in *Pseudomonas aeruginosa (Sonnleitner, Romeo, and Bläsi 2012)*. The signal peptidase I (*lepB*) gene as well as the protein translocase subunit *secA* gene are both coding for proteins required for protein export, a process crucial for flagellum assembly. We therefore speculate that the underrepresentation of these genes is the consequence of a gradual loss of these functionally redundant genes in early ErEurope strains.

ErEurope strains have a distinct carbohydrate metabolism gene repertoire

To investigate whether carbohydrate metabolism gene repertoires differ between subspecies, we annotated all genomes using the CAZy database (Cantarel et al. 2009) (**Methods**). We found that strains belonging to ErEurope harbor significantly more carbohydrate-active enzymes (all p-values < 1e-9, **Fig. 5A**) compared to the three remaining subspecies despite their smaller genome size. Consequently, ErEurope exhibits a much larger density of carbohydrate active genes (p-value < 2.2e-16, **Fig. 5B**), and it clusters separately and distantly from the remaining subspecies also based on genome-wide CAZy gene content differences (**Fig. 5C).**

To understand in what way the carbohydrate metabolism of ErEurope strains has diverged from the other subspecies, we computed differentially abundant CAZy families between ErEurope and ErEurasia: in total, there were 43 differentially abundant CAZy families separating the two subspecies (**Fig. 5D**). ErEurope is enriched in putatively catabolic CAZy families (glycoside hydrolases, carbohydrate esterases, carbohydrate-binding module) targeting either hemicelluloses (xylans, arabinans, arabinoxylans) or pectins (galactans, arabinogalactans) (**Fig. 5D**). We performed *in vitro* carbohydrate utilization tests using six cultured *E. rectale* isolates (two of them belonging to ErEurope and four to ErEurasia) to understand on which carbohydrate substrates ErEurope strains grew better (optical density measured after 48 hours of growth, **Additional File 2: Tab. 1**, **Methods**). We found that, compared to ErEurasia strains, ErEurope strains grew better on xylan and inulin (both representing complex, plant-associated carbohydrates) and worse on sucrose. Furthermore, one of the two ErEurope strains was specifically able to grow on arabinan (**Additional File 2: Tab. 1, Methods**). Together, these results indicate that ErEurope strains tend to be better at

utilizing certain classes of complex, plant-associated carbohydrates compared to ErEurasia strains. These genomic differences might represent adaptive changes due to the loss of motility.

Subspecies	Strain	Negative control	Glucose	Raffinose	Sucrose	SPS	L-Arabinose	D-Arabinose	Inulin (Chicory)	Inulin (Dahlia)	Beta-Glucan	Arabinan	Xylan
ErEurope	T3WBE13	-	++	++	+	+++	++	-	++	++	-	++	+
ErEurope	T1-815	-	++	++	-	++	++	-	++	++	-	-	+
ErEurasia	A1-86	-	+	++	++	+++	++	-	+	+	-	-	+
ErEurasia	L2-21	-	++	++	+++	+++	++	-	+	+	-	-	-
ErEurasia	ATCC 33656	-	++	++	++	++	++	-	-	+	-	-	-
ErEurasia	M104/1	-	++	++	++	+++	++	-	+	+	-	-	-

Table 1: *In-vitro* carbohydrate growth assays (**Methods**). The symbols represent growth (measured by OD at 650 nm after 48 hours) as follows: “-”: OD less than 0.1, “+”: OD between 0.1 and 0.3, “++”: OD between 0.3 and 0.7, “+++”: OD > 0.7.

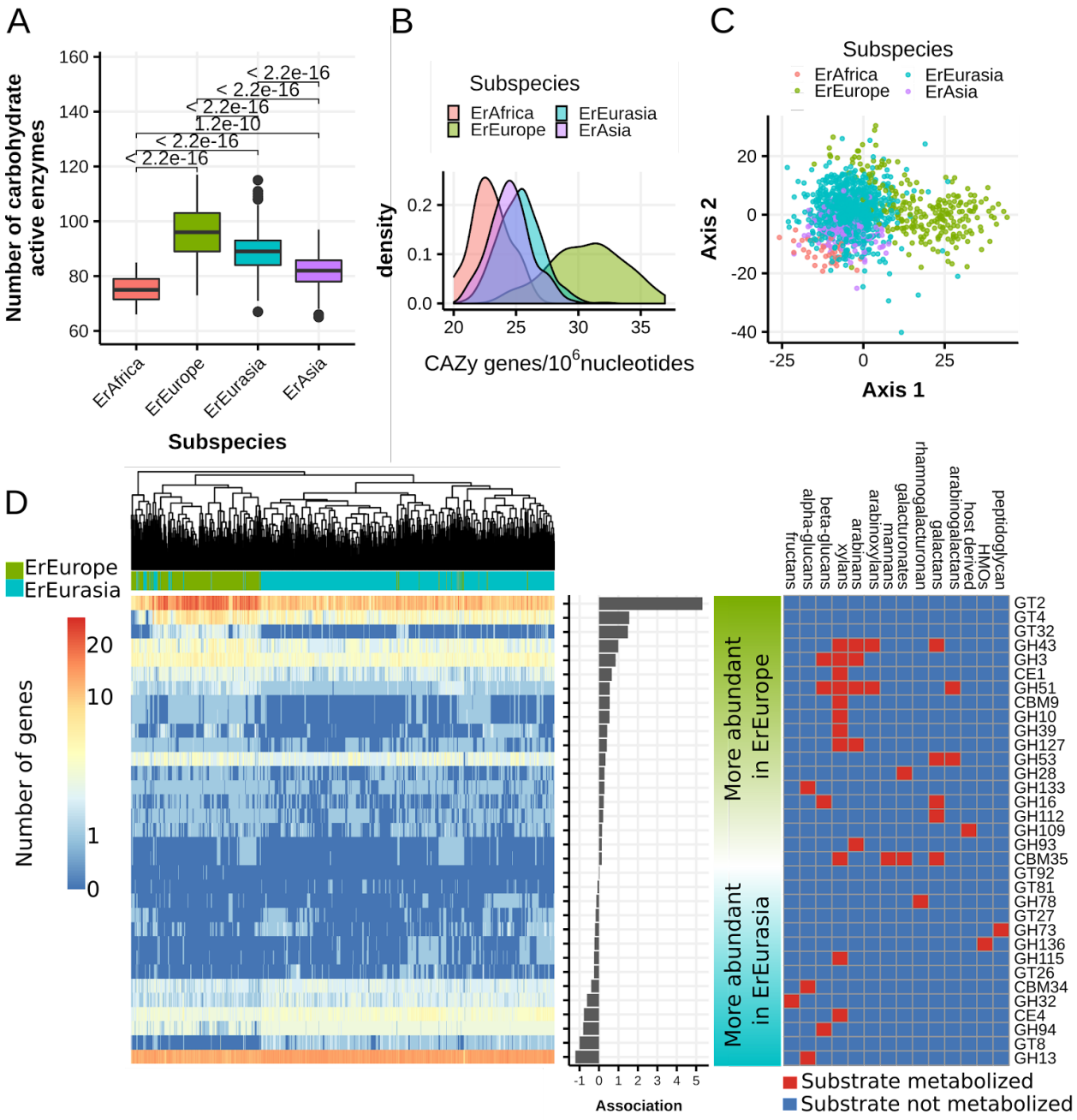


Figure 5: **The immotile subspecies ErEurope exhibits a comparatively strong shift in Carbohydrate-Active Enzyme (CAZy) gene repertoire. (A)** ErEurope exhibits higher Carbohydrate-Active Enzyme (CAZy) family counts than the other subspecies. **(B)** Density estimates of the number of CAZy genes per 10^6 nucleotides in the genome for each subspecies. **(C)** Non-metric Multidimensional scaling plot based on pairwise Manhattan distances between CAZy gene family abundances. **(D) Left:** Differentially abundant carbohydrate-active gene families between genomes of ErEurope and ErEurasia. P-values were corrected at 5% family-wise error rate using the Bonferroni method. Color-scale is logarithmic. **Middle**: Effect size and direction of association (difference in mean copy number between ErEurope and ErEurasia). **Right**: Putative links between catabolic carbohydrate

active enzyme families (CBM, CE, GH) and their substrates. CBM = carbohydrate-binding module, CE = carbohydrate esterase, GH = glycoside hydrolase, GT = glycosyltransferase.

A novel genomic island specific to ErEurope contains a battery of glycosyltransferase genes

Profiling the carbohydrate-related gene repertoire of *E. rectale* revealed another defining feature of ErEurope genomes. This subspecies is strongly enriched in genes coding for some glycosyltransferase (GT) gene families (**Fig. 5D**). Specifically, ErEurope strains possess more GT genes (from the families GT2, GT4 and GT32) compared to other subspecies, with GT2 being particularly strongly overrepresented (p-value < 1e-12, **Fig. 6A**). We found that the cultured *E. rectale* isolate genome T1-815 and several other ErEurope genomes derived from metagenomes contain a genomic region enriched in GT2, GT4 and GT32 genes (**Fig. 6B)** that is part of a genomic island (GI). This GI (when present) is consistently located in the same genomic position (**Fig. 6C**), and its GC content is clearly distinct from the remaining part of the genome (average GC content 37.7% vs 42.3%, **Fig. 6D**). The GI has a length of ~50k bps (average 49,668 bps, s.d. 2,176 bps) and is found in its entirety on the same contig in 56 ErEurope strains (corresponding to ~21% of all ErEurope genomes) with prevalence rates of up to 50% in ErEurope when partial detection of the GT-enriched region is considered sufficient to call the GI present (**Additional File 1: Fig. S24**). No traces of this GI were detected in any other subspecies.

The GT-enriched region of the GI is responsible for most of the enrichment of GT2, GT4 and GT32 gene families in ErEurope strains (**Additional File 1: Fig. S25**). While the non GT-enriched part of the GI remains largely functionally unannotated (**Additional File 7: Tab. S6**), many of the GT genes are associated with synthesizing exopolysaccharides in the context of biofilm formation, protein glycosylation or cell wall polysaccharide synthesis. This may represent an adaptation of ErEurope strains to synthesize exopolysaccharides or other structural carbohydrates, a change that might prove advantageous for an immotile gut commensal.

In order to investigate the origin of the GI, we checked for signals of co-diversification between the core genome and operon sequences. The sequence of the GI is more conserved (<1% pairwise genetic distance) than the rest of the core genome (**Fig. 6E**) and core gene distances and genomic island gene distances are significantly but very weakly correlated (Mantel-test Pearson correlation 0.16, p-value: 0.015). We screened the metagenomic assemblies of the human microbiome in search of homologous sequences of the GI, but found no evidence of any other human-associated microbe with the sequence of this GI (**Methods**). This suggests that the GI originated from a microbe which is not a common current member of the human gut microbiota.

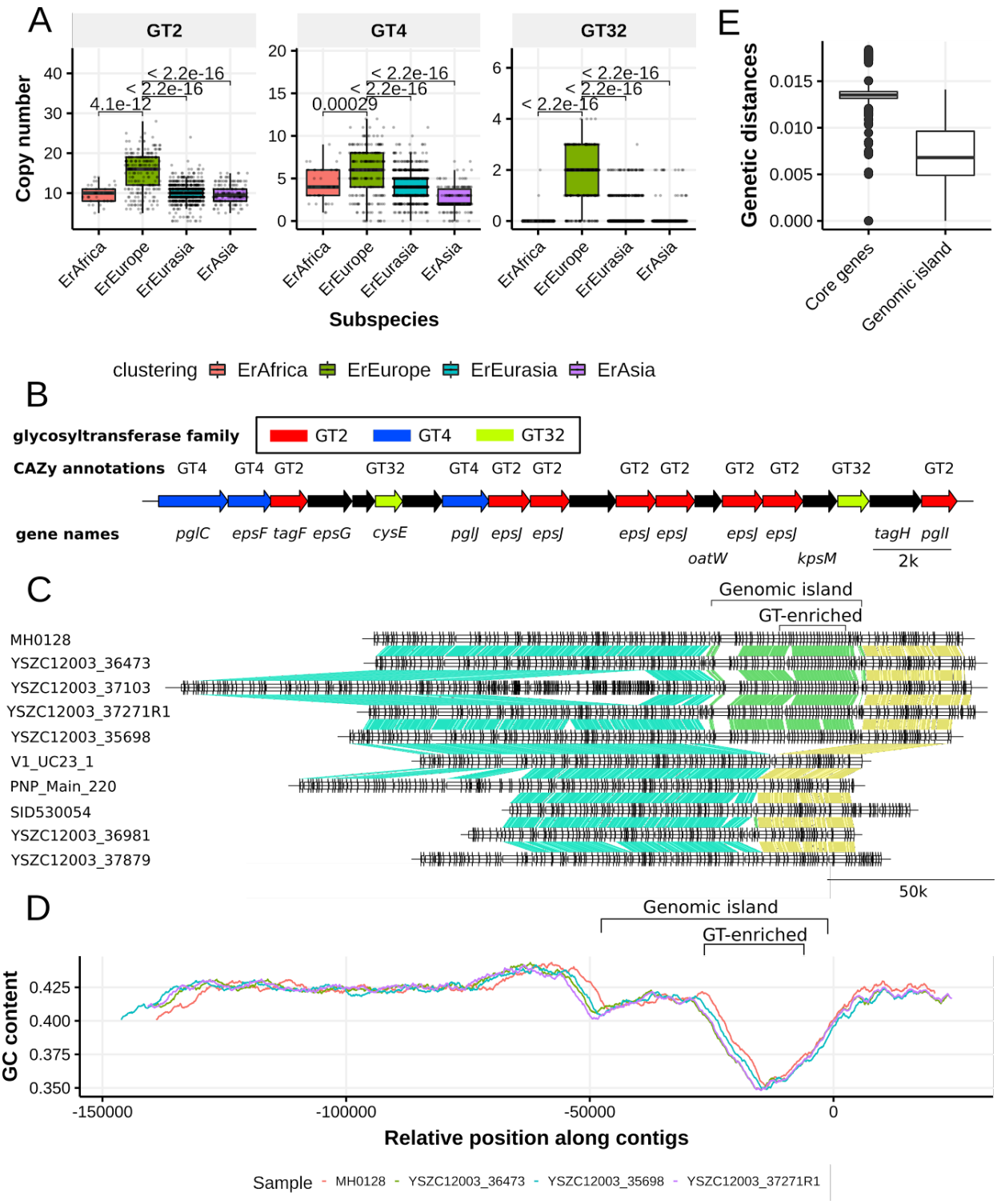


Figure 6: A newly discovered genomic island enriched for glycosyltransferase genes in ErEurope. (**A**) Genome-wide counts of the GT2, GT4 and GT32 families by subspecies. (**B**) Annotated open reading frames of the GT-enriched part of a representative example of the genomic island specific to ErEurope. (**C**) Comparative genomic analysis of the genomic island (**Methods**). The top five ErEurope strains contain the genomic island, whereas the bottom five

do not. Colored segments connecting pairs of genes indicate orthologous genes inferred using progressiveMauve (Darling, Mau, and Perna 2010). (**D**) GC content along the four contigs from ErEurope strains containing the ErEurope genomic island (**Methods**). YSZC12003_37103 is not shown here because another genomic insertion would misalign the sequences. (**E**) Pairwise genetic distances between strains using orthologous genes from the genomic island are lower than those based on core genes. All 56 ErEurope strains with fully extracted genomic island are considered here.

Discussion

New technologies and computational tools are generating an unprecedented amount of strain-specific genomic information that can be the foundation of a new generation of microbiome studies (Pasolli et al. 2019; Almeida, Mitchell, et al. 2019; Almeida, Nayfach, et al. 2019; Nayfach et al. 2019; Manara et al. 2019). Large scale species-specific whole-genome investigations can now be performed without cultivation (Tett et al. 2019b), and - using metagenomic assembly combined with a reference-dependent binning approach - be applied on many thousands of single metagenomes. We demonstrated this with *Eubacterium rectale*, one of the most prevalent human gut species.

Our analysis of *E. rectale* population structure revealed an extreme degree of biogeographic stratification and specificity to the human host. Our data largely supports the hypothesis that the observed stratification (**Fig. 2B, Fig. 3A**) is at least in part the consequence of isolation by distance (**Fig. 3B**) brought about by host-microbe co-dispersal, possibly due to migration movements of early humans. While population structure shaped by isolation by distance has previously been described for the (opportunistic) human pathogen *H. pylori* (Falush et al. 2003; Linz et al. 2007; Moodley et al. 2012), here we report for the first time similar evolutionary signatures in a human gut commensal. Interestingly, vertical transmission rates were found to be low in both *H. pylori (Schwarz et al. 2008; Delport et al. 2006)* and *E. rectale*. The estimated transmission rate of 25% observed between mother-infant pairs for *E. rectale* (**Additional File 1: Fig. S12**) suggests that strain seeding from the local (social) environment contributes to the observed biogeographic stratification.

However, isolation by distance is likely not the only force acting on the genetics of *E. rectale.* Most ErAfrica strains happen to originate from individuals living a traditional lifestyle. It is possible that selection effects by host lifestyle as is the case for *P. copri* (Tett et al. 2019b) influence the genetic structure of *E. rectale* strains as well. Since there are no large datasets that contrast individuals from the same population living different lifestyles, it is difficult to quantify the effect of host lifestyle on the population structure *E. rectale.* Nonetheless, we have tested for subspecies association with age (**Additional File 1: Fig. S26**) and BMI (**Additional File 1: Fig. S27**) as well as diet

(**Additional File 1: Fig. S28**) and found no significant differences. Furthermore, ErAfrica strains are sometimes found in countries outside of Africa, and ErEurasia strains - despite being genetically distinct - are unexpectedly found in Fiji, observations that are not easily explained by isolation by distance. More comprehensive and better georeferenced metagenomic sampling of currently undersampled populations in South America, Africa and Oceania that explicitly contrasts modern and traditional lifestyles will provide more conclusive answers. Powered by such data, our approach of large-scale genome reconstruction from metagenomes will open up new avenues to more broadly study the patterns of host-microbe co-evolution and co-differentiation.

E. rectale is consistently found in all cohorts used in this study and never found in wild non-human primates. This can suggest that the common ancestor of *E. rectale* was part of the gut microbiome of early humans prior to their expansion out of Africa. Bayesian phylogeny rooting did not support this hypothesis (**Additional File 1: Fig. S29**), but future studies exploring ancient DNA pools retrieved from prehistoric human gut content and sampling of undersampled populations (especially those from Africa) could shed further light on the issue. Even without clocked phylogenies, key aspects in the genetic events that shaped the current dispersion of *E. rectale* strains could be found. Perhaps the most intriguing case is the evolutionary history of ErEurope that can be parsimoniously explained assuming a single operon loss event prior to its separation from early ErEurasia strains. This event must have occurred relatively recently compared to the other *E. rectale* subspeciation events as ErEurope has a comparatively low genetic diversity, is closely related to ErEurasia, and is geographically extremely well confined.

Studying the loss of major motility operons in EuEurope provided a detailed example of how large-scale strain-level metagenomics combined with experimental testing can reveal evolutionary and ecological patterns. ErEurope and its closest sister subspecies ErEurasia have functionally diverged at an accelerated pace after the loss of motility, and this is exemplified by the reduced number of extra-operon motility genes and the divergent carbohydrate metabolism gene repertoire in ErEurope. We speculate that, with the lack of motility, ErEurope strains might have been forced to change and extend their repertoire of catabolic carbohydrate active enzymes to be able to metabolize a wider range of energetically unfavourable carbohydrates such as Inulin and Xylan (**Additional File 2: Tab. 1**) due to the inability to scavenge for energetically more favourable carbohydrates. A large genomic island specific to ErEurope was also identified that harbours many genes implied in exopolysaccharide synthesis in the context of biofilm synthesis. The loss of motility operons might have triggered a change in ecological niche in ErEurope strain, which in turn lead to adaptive processes in ErEurope with a combination of genome reduction and horizontal gene transfer events.

We provide an accurate, targeted approach to reconstruct genomes from metagenomes - based on a high-quality set of species-specific genomes (**Additional File 1: Fig. S1**) - which in our evaluation on *E. rectale* (**Fig. 1D, Fig. 1E**) compares favourably to a state of the art reference-independent tool. The merit of such an approach needs to be further validated on other species and could then be useful for exploring diverse microbiomes including those from non-human environments. In the future, such efforts could be improved by technological advances including long read technologies (Bishara et al. 2018; Nicholls et al. 2019) and single-cell sequencing (Marcy et al. 2007), by even larger meta-analyses, and by in-depth phenotypic characterization that could pave the way to a deeper understanding of the complexity of the human microbiome on a subspecies-level.

Materials and Methods

Description of public and newly sequenced metagenomic datasets

We considered a total of 6,775 human gut shotgun metagenomes from 38 datasets spanning 30 countries (**Additional File 4: Tab. S3, Additional File 5: Tab. S4**). Most samples were obtained from publically available datasets; a total of 163 samples came from new cohorts we recently sequenced: We included 113 samples from Madagasy individuals (Pasolli et al. 2019) and 50 samples from Ethiopian individuals (Tett et al. 2019a). The datasets we used are composed of individuals with different diets, exposure to environmental stressors (including antibiotics) and sanitary conditions. As such, some of those individuals can be described as ‘westernised’ and others as ‘non-westernised’ (Segata 2015).

Furthermore, in this study we used four publically available datasets containing a total of 175 shotgun metagenomes coming from wild, non-human primates (Amato et al. 2019; Hicks et al. 2018; Tung et al. 2015; Orkin et al. 2019) (**Additional File 8: Tab. S7**).

Prevalence testing of *E. rectale* in human and great ape metagenomes

Based on taxonomic profiling using MetaPhlAn2

All human- and non-human great ape samples were profiled using MetaPhlAn2 (version 2.7) (Truong et al. 2015) with default parameters. Reads were mapped to markers using Bowtie2 (version 2.3.4, parameters --very-sensitive, --no_unal) (Langmead and Salzberg 2012). *E. rectale* was determined to be present in a sample if its relative abundance exceeded 0.1% and at least 20% of all *E. rectale* marker genes were hit.

Based on metagenomic assembly and binning

In order to find *E. rectale* genomes assembled from wild non-human primate metagenomes, we assembled and binned as described elsewhere (Pasolli et al. 2019; 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, n.d.) a total of 2,895 metagenomic high-quality genomes obtained from 175 publicly available metagenomes from wild, non-human primates (**Additional File 8: Tab. S7**). These 175 metagenomes come from four different datasets spanning 22 non-human primate species including Chimpanzees and Gorillas from 14 different countries on five continents (Amato et al. 2019; Hicks et al. 2018; Tung et al. 2015; Orkin et al. 2019). We then estimated genetic distances between each of the reconstructed genomes and the set of *E. rectale* isolate genomes using MASH (Ondov et al. 2016), and found that not a single bin generated from the non-human primates was within 23% genetic distance of any *E. rectale* isolate. To confirm that this result is not dependent on the binning method, we also applied the

reference-based binning procedure we proposed in this work to these assemblies. We found that not a single bin was more than 5% complete, confirming our previous result that the metagenomic assemblies of wild non-human primates used in this study do not contain *E. rectale* genomes.

Determining Vertical transmission rates of *E. rectale*

Vertical transmission of *E. rectale* was assessed in three publicly-available longitudinally-sampled mother-infant datasets: Backhed et al (N=398 samples; 96 mothers-infant pairs) (Bäckhed et al. 2015), Asnicar et al (N=18, 5 mother-infant pairs) (Asnicar et al. 2017), and Ferretti et al (N=116 samples, 21 mothers and 25 infants) (Ferretti et al. 2018). Strain-level single-nucleotide variant profiling was performed with StrainPhlAn2 (Truong et al. 2017) with database version mpa_v294_CHOCOPhlAn_201901 and options sample_with_n_markers=10 and marker_in_n_samples=10. Pairwise genetic distances normalized by median branch length (nGD) were created using PyPhlAn (https://bitbucket.org/nsegata/pyphlan).

Strain transmission was assumed when two individuals harboured identical strains, with strain identity inferred when the pairwise normalize genetic distances are below the first percentile of the nGD distribution of samples of unrelated individuals, thus allowing a 1% false discovery rate. *E. rectale* transmission rates were defined as the proportion of mother-infant pairs harbouring *E. rectale* that carried the same strain at ≥1 time points.

The assembly and reference-based binning of *E. rectale* genomes from 6,775 gut metagenomes

The reference-based binning approach employed here consists of three principal steps (**Fig. 1A**): Individual assembly of all 6775 gut metagenomes, compilation of a high-quality *E. rectale* genome set consisting of both isolate genomes and manually-curated reference genomes from metagenomes and reference-based binning of all 6775 gut metagenomic assemblies using the high-quality *E. rectale* genome set as a reference for binning contigs.

First, we assembled each gut metagenome individually using metaSPAdes (version 3.10.1) with standard parameter settings (Nurk et al. 2017) as described by Pasolli et al. (Pasolli et al. 2019). We used MEGAHIT (version 1.1.1) (Li et al. 2015) instead of metaSPAdes for those metagenomes with only unpaired reads.

Next, we compiled a set of *E. rectale* reference genomes consisting of manually curated metagenomic bins obtained using anvi'o (version 2.3.2) (Eren et al. 2015) as well as genomes from isolate sequencing. Anvi'o visually integrates information about depth, tetranucleotide frequency and taxonomy of metagenomic assemblies on a contig-by-contig level, facilitating human-aided binning. We followed the author's

recommended workflow for preparation of metagenomic assemblies for manual inspection (http://merenlab.org/2016/06/22/anvio-tutorial-v2/). We complemented the taxonomic assignment provided by centrifuge (version 1.0.4) (Kim et al. 2016) with an ad-hoc approach, mapping the assembled contigs against the bacterial RefSeq database using BLAST (version 2.6.0) (Altschul et al. 1990). Based on the results of this mapping, we assigned taxonomic labels to each contig of that species against which the largest fraction of the contig mapped with a mean identity score of at least 75%. Manually curated bins were used only when hierarchical clustering of tetranucleotide frequency and coverage as well as taxonomic assignments indicated an *E. rectale* bin of high quality. We maximised precision of manually-curated reference genomes by excluding contigs that were not clearly belonging to *E. rectale*. A total of 170 metagenomes with high depth and coverage over *E. rectale* isolate genomes were queried in this manual binning process. From these 170 metagenomic assemblies, we reconstructed 47 manually-curated reference genomes (MCR) with an average length of 2.61 Mbps (s.d. 0.20 Mbps), an average number of contigs of 41.1 (s.d. 11.14), an average N50 of 102,000 bps (s.d. 36,000 bps) and average CheckM (Parks et al. 2015) completeness and contamination estimates of 96.6% (s.d. 3.5%) and 0.2% (s.d. 0.3%), respectively (**Additional File 2: Tab. S1**). These MCR genomes have very good assembly characteristics (N50, nr. of contigs) but are shorter due to the maximization of precision during the manual curation step, which we expected to improve reference-based binning performance since the chance of faulty binning of small contigs from closely related species due to propagation of contamination in the reference is reduced.

The final step consisted of mapping all 6,775 assembled metagenomes against the set of genomes consisting of the 47 manually-curated reference *E. rectale* genomes (**Additional File 2: Tab. S1**) as well as seven isolate genomes from NCBI (**Additional File 3: Tab. S2**). We considered a contig to come from *E. rectale* if it mapped with a mean identity score of at least 95% over at least 50% of its length against the set of reference genomes. We determined optimal thresholds for minimum mean identity score / fraction mapping based on simulations with semi-synthetic data (see section below).

The *E. rectale* metagenomic assemblies resulting from the procedure in this section were quality controlled and compared favourably against reference-free binning (see below).

Parameter selection for the reference-based binning using semi-synthetic metagenomes

We used semi-synthetic data to select optimal parameter values in the reference-based binning approach. We spiked in sequences originating from *E. rectale* isolate genomes

into metagenomic assemblies where *E. rectale* was undetectable using MetaPhlAn2 (version 2.7) (Truong et al. 2015). We applied reference-based binning as outlined above and evaluated performance over a grid of parameter values. The parameter values are 1) the mean identity score of the query contig against the reference database and 2) the fraction of the query contig mapping against the database. False positives are defined as those nucleotides binned that originated from the originally *E. rectale*-free metagenomic assembly; false negatives are defined as those spiked-in nucleotides that were not binned. The reference genomes which were not completely scaffolded were spiked-in as they are, whereas the completely scaffolded reference genomes were sliced into uniformly distributed pieces between 1000 and 50000 in length. We tested performance using all combinations of isolate genomes and 50 metagenomes without detectable levels of *E. rectale* (MetaPhlAn2) randomly chosen among all 6775 metagenomes.

Comparison of reference-based- against reference-free binning

We compared genomes extracted by the reference-based binning method described above with those from a large-scale, reference-free binning effort (Pasolli et al. 2019). Briefly, the study by Pasolli et al. used metaBAT2 (D. Kang et al. 2019), a state of the art reference-free binning software, on single-sample metagenomic assemblies to produce more than 150,000 genomes from metagenomes. The extracted genomes along with 80,990 reference genomes were clustered into species-level groups using pairwise genetic distances using MASH (Ondov et al. 2016)). These groups were taxonomically labeled with the species associated with the reference genome(s) present in the group, considering the most common species label if multiple reference genomes with different assigned species were present. We selected the species-level group corresponding to *E. rectale* and compared those genomes that were more than 90% complete and less than 5% contaminated in both approaches. Very rarely, the reference-free binning by Pasolli et al. produced more than one bin assigned to *E. rectale* in a given metagenome. In these cases, only the more complete bin was evaluated. No longitudinal samples were considered.

Quality control of the genomes

Filtering of genomes for downstream analysis consisted of removing lowly covered contigs in bins (those that are below 20% of the median genome-wide coverage) followed by further quality checks. High quality (HQ) *E. rectale* genomes were defined as those with CheckM (version 1.0.12) (Parks et al. 2015) completeness > 90% and contamination < 5%, a total size larger than 2.9 Mbps and smaller than 3.89 Mbps (calculated as the 95% and 105% of the size of the smallest/largest *E. rectale* isolate genome), less than 400 contigs and an estimate of within-sample strain heterogeneity of less than 0.3% (see below). As expected, the HQ genomes generally miss rRNA genes,

containing on average 0.56 of them (sd 0.75). In total, we reconstructed 1,321 HQ genomes that passed all these quality criteria and were used for further analysis.

Polymorphism-based strain heterogeneity assessment for additional quality control

We developed a method for *ad hoc* estimation of within-metagenome strain heterogeneity for each genome based on the number of polymorphic sites over *E. rectale* protein-coding genes. After gene calling performed using Prodigal (version 2.6.3) (Hyatt et al. 2010), we mapped reads back to protein-coding genes using Bowtie2 (version 2.3.4, parameters --very-sensitive-local and -a) (Langmead and Salzberg 2012) and determined dominant and second-dominant alleles over all protein coding nucleotides. For this, we only considered base calls with a PHRED quality score of at least 30 and only those positions with a coverage of at least 10. We considered a position non-variant if the dominant allele constituted more than 80% of the total number of nucleotides mapped to that given position. In order to calculate the polymorphism rate, we translated dominant and second-dominant nucleotide sequences into protein sequences and divided the total number of non-synonymous mutations between the two by the total number of positions.

E. rectale genome annotation

We used Prokka (version 1.12) (Seemann 2014) for gene calling and functional annotation of bacterial genomes. Roary (version 3.8.2) (Page et al. 2015) with settings '-i 95 -cd 95 -e -z --mafft' was used for core and pan-genome clustering as well as for generating core gene alignments (Katoh and Standley 2013). Core genes were defined as those genes present in at least 95% of genomes. All gene clusters were annotated with KO information using eggNOG-mapper (Huerta-Cepas et al. 2017)) using representative gene sequences obtained from Roary. CAZy annotations (Cantarel et al. 2009) were obtained using a local dbcan distribution (release 6.0) (Yin et al. 2012), which uses HMMER (version 3.1b2) (Eddy 1998) to identify carbohydrate-active enzyme families in protein sequences. We used dbcan on translated protein-coding genes (Prodigal) and filtered hits for E-value < 1e-18 and coverage > 0.3 as suggested by the authors. Only one randomly selected sample per individual was considered for this analysis. Assignment of substrates to carbohydrate active enzymes was based on the information provided in the CAZy database (www.cazy.org, (Lombard et al. 2014)), CAZypedia (www.cazypedia.org, (CAZypedia Consortium 2018)) and dbCAN (Yin et al. 2012).

Functional divergence rate

Genomic distances were calculated based on the Roary gene presence/absence matrix. Motility operon genes were identified by blasting representative operon gene sequences

against representative gene sequences from roary and subsequently removed from the gene presence/absence matrix. Pairwise Jaccard distances between genomes were then computed using the 'vegdist' function in the 'vegan' R package. The genetic distances were defined as the hamming distance on a core gene alignment produced by roary. The rate of functional divergence was calculated by dividing pairwise inter-subspecies genomic distances by their corresponding genetic distance.

Phylogenetic analyses

If not stated otherwise, the phylogenetic analyses were performed with PhyloPhlAn 3.0 (Segata et al. 2013) (https://github.com/biobakery/PhyloPhlAn)
The phylogeny in **Fig. 2** was built using the 1,071 core genes extracted as described above. PhyloPhlAn was run with the following options: “--diversity low --fast”. For the internal steps the following tools with their set of parameters were used:

- blastn (version 2.6.0+), (Altschul et al. 1990) with parameters: “-outfmt 6 -max_target_seqs 1000000”;
- mafft (version 7.310), (Katoh and Standley 2013) with the “--anysymbol” option;
- trimal (version 1.2rev59), (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) with the “-gappyout” option;
- RAxML (version 8.1.15), (Stamatakis 2014) with parameters: “-p 1989 -m GTRCAT”.

To infer the bayesian phylogeny, we built a core-gene alignment (using an in-house script (https://bitbucket.org/CibioCM/genomealnbuilder)) from 46 metagenomes randomly selected to represent the four subspecies with following parameters: “contigs_based -minqual 30 -minlen 50 -maxsnps 0.03 -mincov 5 -aln_len 500 -pid 95.0”. We used trimAl to remove gappy columns from the alignment. (Capella-Gutiérrez et al., 2009). After filtering, the alignment included 1,356,039 nucleotide positions. BEAST v2.5.1 (Bouckaert et al., 2014) was used to infer a phylogeny, using a GTR model of nucleotide substitution (with 4 gamma categories). To choose the best clock and demographic models we performed a model selection comparing coalescent constant, coalescent exponential, coalescent bayesian skyline, and coalescent extended bayesian skyline models (for the demographic priors) and a strict molecular clock. Convergence of the posterior probability distribution was assessed by visualising log files with Tracer v1.7 (Rambaut et al., 2018). The most fitting combination of models was a coalescent constant population with a strict molecular clock: this analysis was run longer for >12,000,000 iterations with an effective sample size (ESS) of key parameters of over 200.

E. rectale subspecies definition

To define subspecies, we used the Partitioning around Medoids algorithm (Kaufman and Rousseeuw 1990) on the hamming distances (not considering gaps) computed on the concatenated nucleotide core-gene alignment (produced by Roary, see above). In order to determine the optimal number of clusters, we used the Prediction Strength metric (Tibshirani and Walther 2005). The PAM clustering algorithm is minimizing the sum of distances of each sample to the closest centroid, which is why it is prone to over-separate dense clouds of points. In order to produce more even sample densities, we subsampled all Eurasian/North American datasets to 50% and applied the PAM algorithm on this subset. We calculated Prediction Strength values on 50 random subsamples in order to obtain information regarding the variation of Prediction Strength values with respect to the subsamples. After having determined the optimal number of clusters (k = 4) following the standard procedure (Tibshirani and Walther 2005), we assigned cluster membership to all genomes based on the distance of each genome to the cluster corresponding to the closest centroid. We chose random cluster centroids from the 50 iterations, as the cluster assignment was very stable over subsamples for k = 4. We used the ‘pam’ function in the ‘cluster’ package and the ‘prediction.strength’ function of the ‘fpc’ package in R with 20 internal divisions.

E. rectale subspecies abundance estimation

We have used subspecies-specific Single Nucleotide Variants (SNVs) (defined using the core gene alignment as those nucleotides that are present in more than 90% of a subspecies but absent in more than 90% of the remaining ones) to estimate subspecies abundances in the samples. We mapped reads to consensus core gene sequences and - for each subspecies - calculated the median of the coverage ratios between the subspecies specific alleles and the respective total coverages. We have restricted this analysis to only those metagenomes where the mean depth over all subspecies-specific positions was at least 5 and where at least 75% of the set of subspecies-specific positions was covered at least 3 times. We have removed samples where the sum of estimated relative abundances is bigger than 1.25 or smaller than 0.75. For metagenomes passing these filters, we have scaled the estimated relative abundances to sum up to 1.

Analysis of the *E. rectale* motility operons

The identification of the motility operons of *E. rectale* used in our analysis is based on the work by Neville et al. (Neville et al. 2013). Briefly, Neville et al. determined and characterized motility operons in *E. rectale* and closely related species using isolate genomes. We annotated the operons in our genomes based on the *E. rectale* strain A1-86 used as a reference for the annotation by Neville et al. (**Additional File 9: Tab.**

S8). Differently from the original analysis reporting the presence of three motility operons for *E. rectale*, we separated the *flgM/csrA* and *flaG/flgN* operons since we did not find them to be in close vicinity in both the genome of *E. rectale* strain A1-86 as well as in the genomes we reconstructed from metagenomes. In *E. rectale* strain A1-86, the largest operon (termed '*flgB*/*fliA*') has a total length of 30,520 nucleotides and 34 coding sequences. The three remaining operons have a length of 1,984, 6,764 and 4,152 nucleotides and contain three, seven and four coding sequences, respectively (Neville et al. 2013).

We determined the presence and absence of motility operon sequences using two different strategies. In the first, we extracted operon gene sequences from *E. rectale* strain A1-86 and blasted them against our genomes. We removed all hits that were shorter than 75% of query gene as well as redundant blast hits. After this, all hits had an identity score of at least 95% and all E-values were smaller than 1E-44 and were thus used to determine operon gene presence/absence. The second strategy involved extracting the genes immediately upstream and downstream of all motility operons (bordering gene sequences were taken from *E. rectale* strain A1-86), blasting them against all genomes and finding contigs on which both bordering genes of an operon could be found. We considered only those cases in which an operon could be identified well (exactly two blast hits per contig with an E-value < E-30). When a motility operon could be identified on a contig, we extracted all protein-coding genes between bordering genes and annotated them by mapping against the motility gene sequences of *E. rectale* strain A1-86.

Analysis of exopolysaccharide genomic island

Detection of GT-enriched genetic element

We noticed a pronounced physical enrichment of genes annotated with glycosyltransferase activity in *E. rectale* isolate genome T1-815. We blasted this genetic stretch against all *E. rectale* genomes and extracted and aligned sequences (using mafft (version v7.310) (Katoh and Standley 2013) and standard parameter settings) in case there was a single blast hit with a length of at least 95% of the length of the genetic element of T1-815 (E-value cutoff of E-30), which was the case in a total of 56 ErEurope genomes. We further annotated the fully extracted sequences with Uniprot information (The UniProt Consortium 2017) as provided by prokka (version 1.12).

Determining total size of genetic island

In order to determine the boundaries of this genetic island, we first blasted the GT-enriched sequence discovered in T1-815 against all *E. rectale* genomes and extracted exceptionally long contigs (contigs at least 100k nucleotides long) with a single blast hit of at least 30k in size and an E-value of less than 1E-30. We then

blasted these contigs back against all HQ *E. rectale* genomes, targeting strains of ErEurope with long contigs not enriched for GT genes since those represented strains without the genetic element. This two-step approach is necessary since the contig containing the genetic element from T1-815 is comparatively short (around 50k bps) and was unable to attract contigs from ErEurope strains where the genetic element is absent. We aligned contigs with progressiveMauve (Darling, Mau, and Perna 2010) (build date Feb 13 2015) and used only contigs that spanned the operon completely. When the genetic element borders were visualized, we used the first gene upstream/downstream of the genetic element that was inferred to be orthologous among all ten ErEurope genomes as the bordering genes. We calculated GC content along the contigs using a rolling window with window size 20000 and a step size of 10. Using this approach, each contigs' 10k positions to either end were not queried.

Search for possible donor organism in gut metagenomic assemblies

In order to find a possible source microbe for the GT-enriched genomic island, we screened >9500 human gut metagenomic assemblies (Pasolli et al. 2019) for a similar sequence. We blasted (blastn with parameters '-word_size 7') a representative sequence of the genomic island against all bins as well as all unbinned fractions of the metagenomic assemblies. Among the bins, we found several hits with a mean identity score > 98% across the entire length of genomic island in non-*E. rectale* bins. Yet, we noticed that in all those samples, a complete quality *E. rectale* genome was binned as well, which suggests that these contigs truly belong to *E. rectale*. We observed the same pattern in the unbinned fraction of contigs and concluded that this sequence might be unique among contemporary human gut commensals.

Physical distances between subspecies

We estimated geographic distances between subspecies in order to look for a correlation between pairwise geographic and genetic distances of subspecies. We associated ErAfrica with Tanzania, ErEurope with Germany and ErAsia with Eastern China based on evident geographic enrichment (**Fig. 3A)**. The geographic association of ErEurasia is less clear, with strains being found in Europe and central/northern Asia, but also in Fiji and Ethiopia, although strains from these two countries are genetically distinct from ErEurasia strains found in Eurasian countries (**Fig. 3B, Additional File 1: Fig. S4, Additional File 1: Fig. S5**). We associated ErEurasia with Kazakhstan because individuals from central/northern Asian countries (Kazakhstan, Mongolia, Russia) almost exclusively harboured genetically representative ErEurasia strains (**Additional File 1: Fig. S30, Additional File 1: Fig. S31**). Distances were approximated with the distm function of the geosphere package in R (Hijmans 2017). Physical distances between subspecies were defined as the shortest path between geographic locations associated with subspecies with the exception of the distance

between ErAfrica and ErAsia, which was determined as the shortest path across the Arabian peninsula.

Bacterial strains, isolation and growth media

The bacterial strains L2-21 and T3WBe13 were isolated from human faecal samples from a healthy adult male consuming a vegetarian diet who had not taken any antibiotics or other medication known to influence the human colonic microbiota for a period of more than three months prior to providing the samples. Strain L2-21 was isolated in 1995 as described previously (Barcenilla et al. 2000). T3WBe13 was isolated from another faecal sample from the same donor 22 years later. It was grown on clarified rumen fluid based M2 medium (Miyazaki et al. 1997) containing a range of soluble sugars (M2GSC containing glucose, cellobiose and soluble starch, 0.2% final concentration of each) following a 10-fold serial dilution in basal M2 medium (containing 10% clarified rumen fluid) with 100 µl of slurry inoculated into 10 ml volumes of either M2 medium containing 0.2% wheat arabinoxylan (Megazyme) or 0.2% pre-treated bran as described previously (Duncan et al. 2016). Following 48 h incubation at 37 ºC the samples were enriched for a total of three times on the same medium prior to preparing a 10-fold serial dilution and inoculating roll tubes (M2GSC medium). Single colonies were picked into M2GSC broth. Strains A1-86, T1-815, ATCC 33656 and M104/1 have been described previously.

Genome sequencing

The two genomes isolated and sequenced in this work (L2-21 and T3WBe13) were grown on M2GSC broths. Genomic DNA was extracted using the FastDNA SPIN Kit for Soil (MP Biomedicals). The sequencing libraries were prepared using the NexteraXT DNA Library Preparation Kit (Illumina, California, USA), following manufacturers guidelines. Library quality was assessed using the Caliper LabChip GX (High-Throughput Bioanalyzer) according to the manufacturer's instructions. The sequencing was performed on a HiSeq2500 machine (Illumina, California, USA).

Experimental assessment of carbohydrate metabolism

The *E. rectale* strains were pre-grown overnight on M2GSC medium and inoculated into basal YCFA medium (Lopez-Siles et al. 2012) containing individual carbohydrate substrates added at 0.2% w/v concentration. The carbohydrate sources tested were Glucose (Sigma Aldrich), Raffinose (Sigma Aldrich), Arabinan - Sugar Beet (Megazyme), Soluble Potato Starch (Sigma Aldrich), D-Arabinose (Sigma Aldrich), L-Arabinose (Sigma Aldrich), Beta-Glucan (Megazyme), Inulin - Chicory (Sigma Aldrich), Xylan – Oat Spelt (Sigma Aldrich), Inulin - Dahlia (Sigma Aldrich) and Sucrose (Fisher Scientific). As negative controls, cells were grown in basal YCFA with no added carbon sources. 100 µL of each culture was then inoculated from its M2GSC growth

medium into single carbohydrate or basal YCFA medium in triplicate under anaerobic conditions using oxygen-free CO_2 and incubated at 37^oC. Optical density measurements were taken spectrophotometrically after 48 hours at wavelength 650 nm (Amersham Pharmacia Biotech, UK).

Experimental validation of motility

In vitro screening for motility was tested using cultures grown to exponential phase (optical density 0.35-0.55) in M2GSC medium, then one drop was added to a dimpled glass slide anaerobically and covered with a glass cover slip. The wet mount was examined using phase-contrast to screen for motility. If individual cells were seen to be moving across the field of view, they were classified as motile.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by NIH NHGRI grant R01HG005220, NIDDK grant R24DK110499, NIDDK grant U54DE023798, CMIT grant 6935956 to C.H., and by the European Research Council (ERC-STG project MetaPG-716575), MIUR ‘‘Futuro in Ricerca’’ RBFR13EWWI_001, the European Union (H2020-SFS-2018-1 project MASTER-818368 and H2020-SC1-BHC project ONCOBIOME-825410) to N.S. Further support was provided by the Programma Ricerca Budget prestazioni Eurac 2017 of the Province of Bolzano, Italy to F.M., and by the EU-H2020 (DiMeTrack-707345) to E.P. and N.S.

Availability of Data and Materials

All datasets used in this study are publically available and matched with their respective PMID (**Additional File 5**). The High-Quality *E. rectale* MAGs in fasta format and a metadata file are available at http://segatalab.cibio.unitn.it/data/Erectale_Karcher_et_al.html and in the following Zenodo repository: doi:10.5281/zenodo.3763191. The two new isolate genomes L2-21 and T3BWe13 have been uploaded to NCBI and can be found in RefSeq under the accession numbers GCF_008122485.1 and GCF_008123415.1, respectively.

Contributions

NS and NK conceived and supervised the study. EP, FeA, FrA, SM, PM, MCC performed data acquisition. NK, EP, AT, MVC, RR, ORS, MZ, DF, FrA, GZ, FM, CH, KH, performed data analysis. DB, SHD, PL, AW designed and performed *in-vitro* experiments. NK and NS performed data interpretation and wrote the manuscript. All authors read and approved the final manuscript.

Supplementary Figures

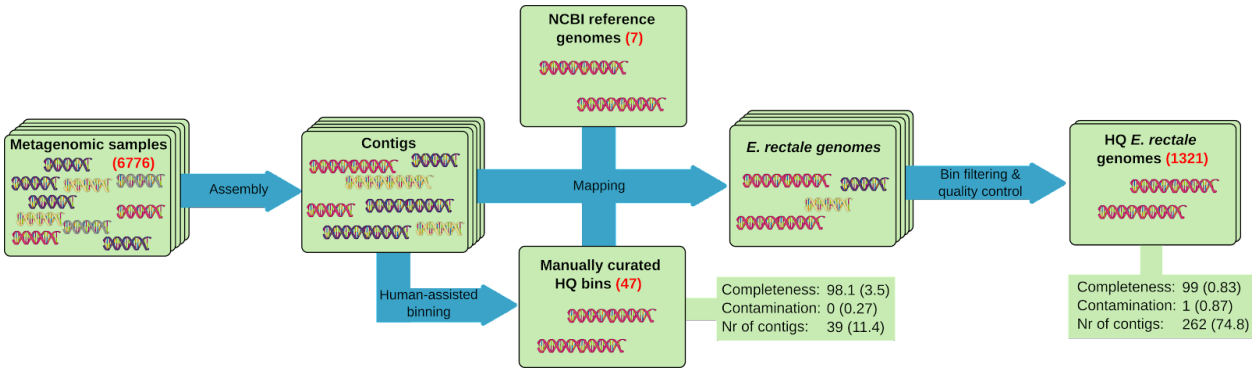


Fig S1: An integrated, reference-based workflow for genome reconstruction from metagenomes. Numbers in red parentheses correspond to the set size. Numbers for completeness, contamination and the number of contigs correspond to the mean and standard deviation (in parenthesis).

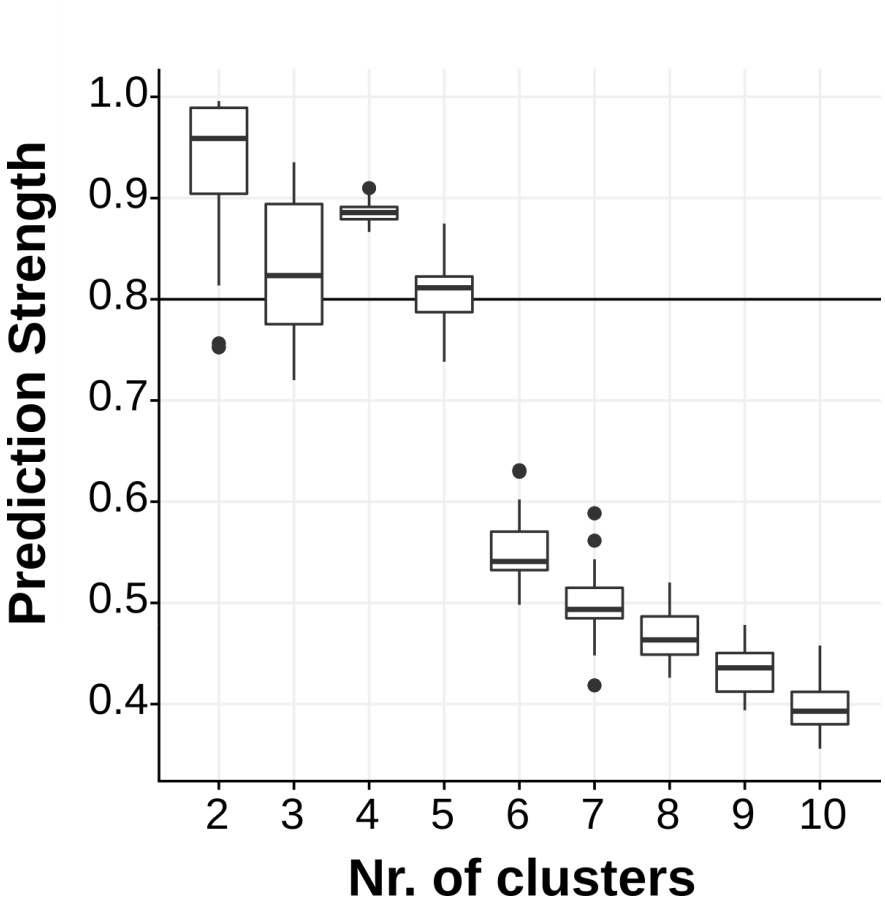


Fig S2: Prediction Strength values for varying numbers of clusters obtained by Partitioning Around Medoids (PAM) clustering (**Methods**). The horizontal line corresponds to a Prediction Strength value of 0.8, suggested by Tibshirani et al. as a cutoff for adequate clustering strength (Tibshirani and Walther 2005).

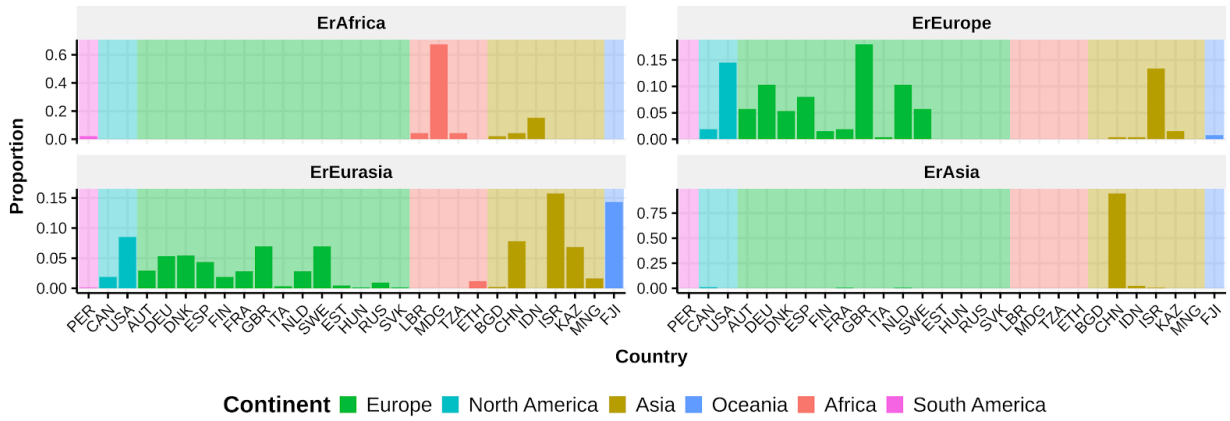


Fig S3: Proportions of *E. rectale* subspecies over countries and continents. Proportions sum up to 1 for each subspecies.

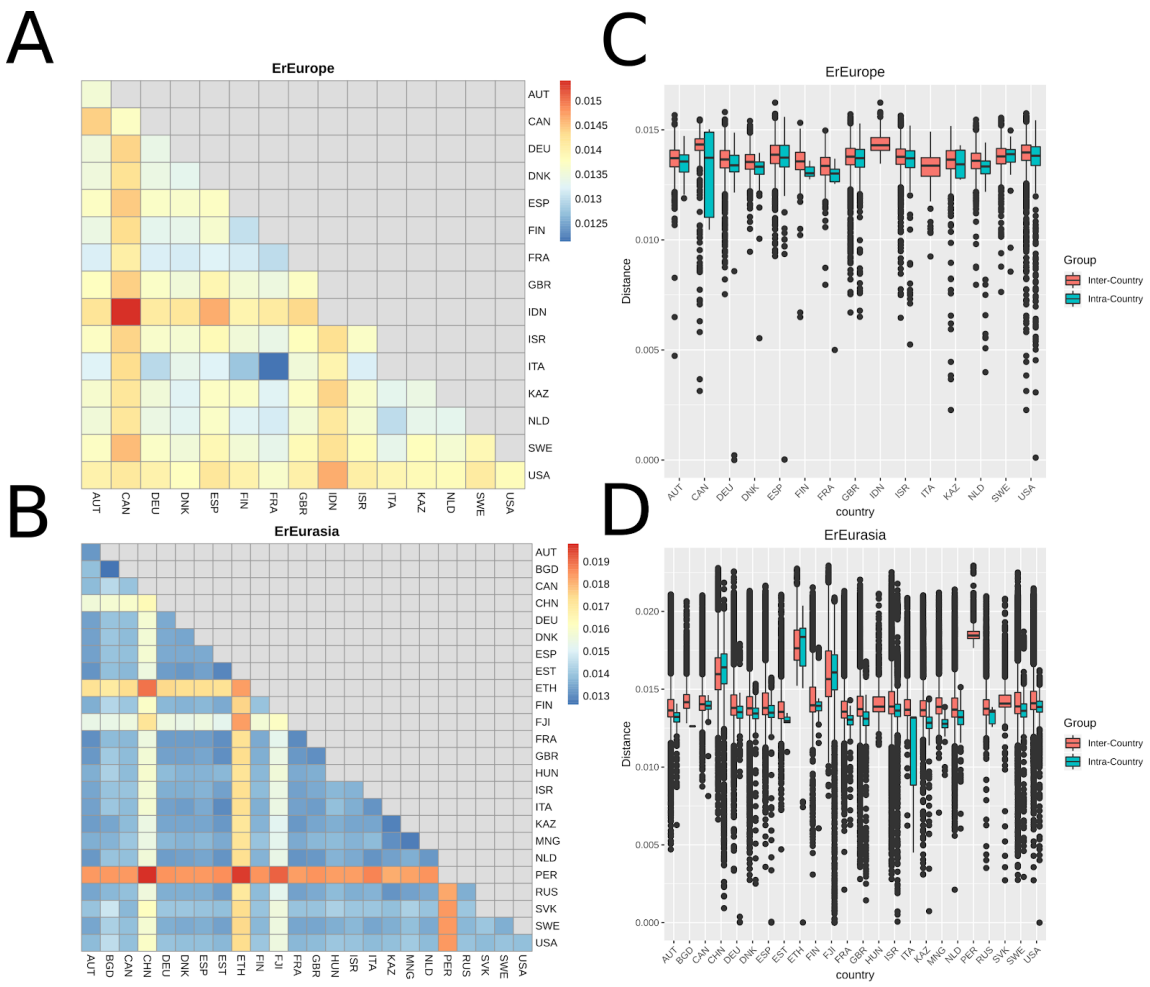


Fig S4: (**A**, **B**) Heatmaps of median pairwise genetic distances between countries considering ErEurope (**A)** or ErEurasia (**B**) individually. (**C**, **D**) Boxplots of within- and between country genetic distances considering ErEurope (**C**) and ErEurasia (**D**) individually.

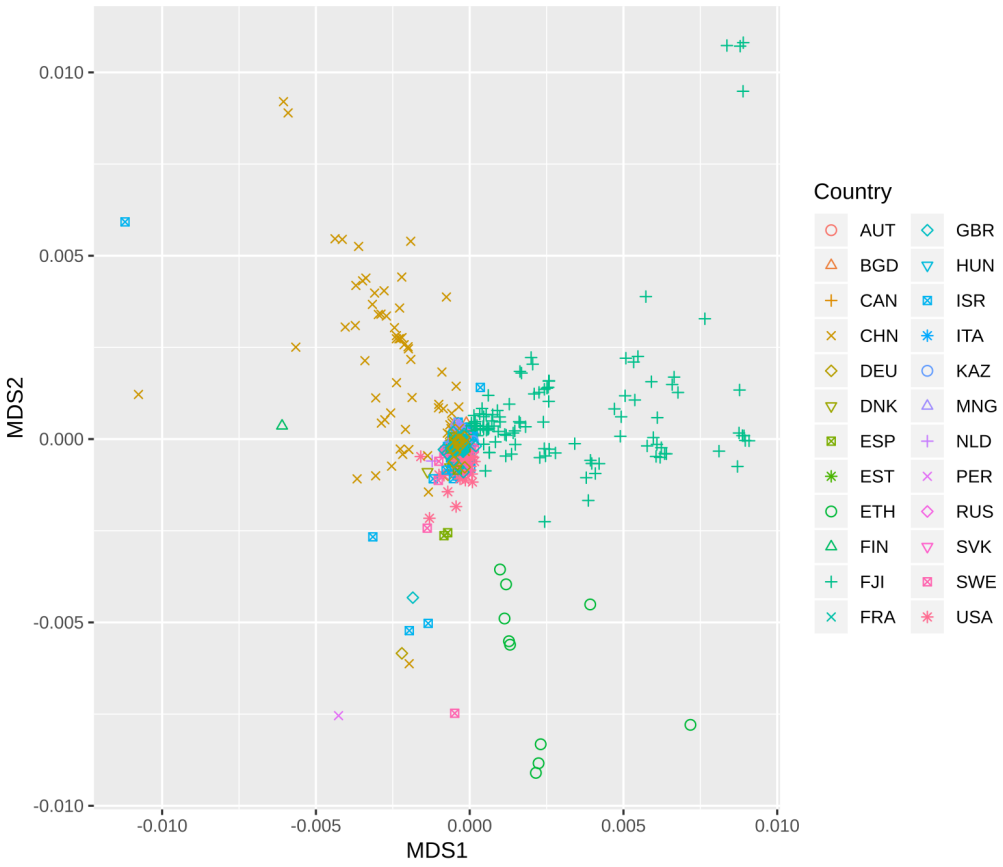


Fig S5: nMDS plot of ErEurasia. See **Additional File 1: Fig. S6** for comparison.

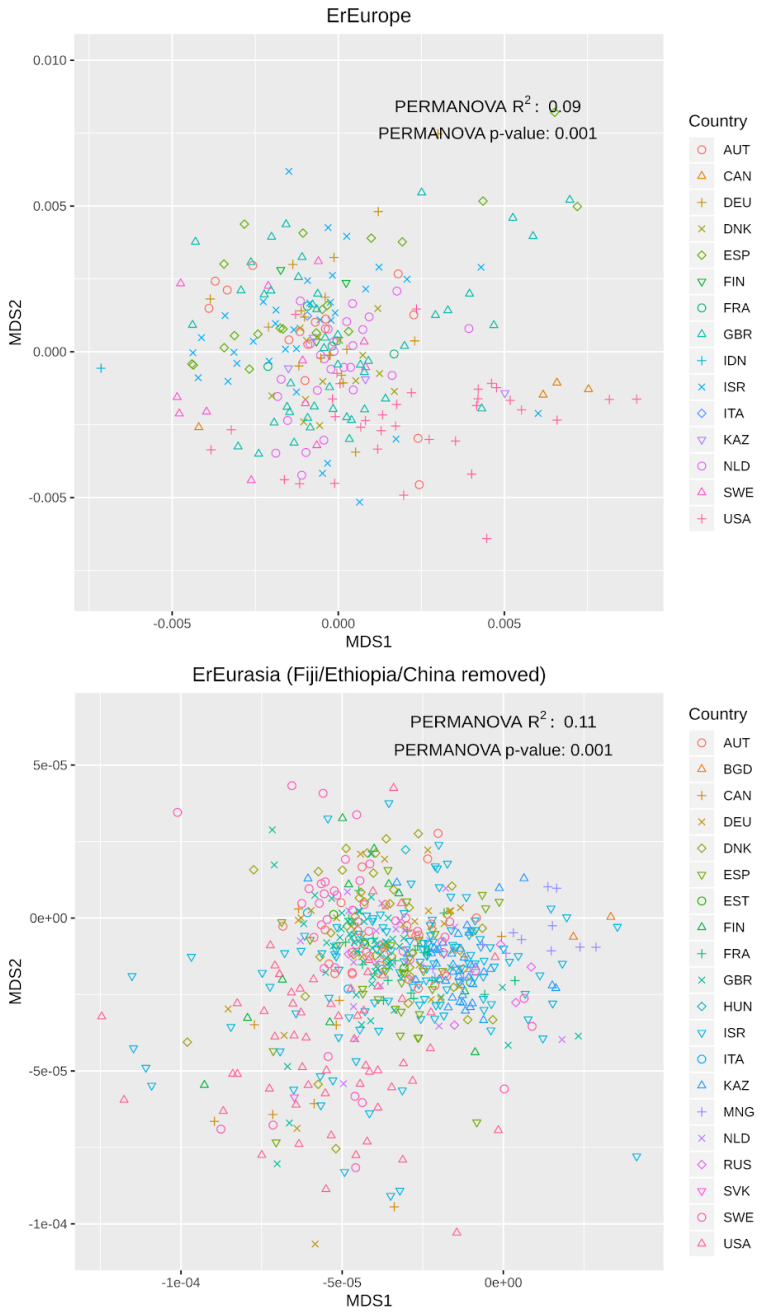


Fig S6: nMDS plots of ErEurope and ErEurasia based on pairwise genetic distances. PERMANOVA was calculated using Country membership. For ErEurasia, we removed Fijian, Chinese, Ethiopian and Peruvian strains for PERMANOVA calculation and from the ordination plot. 24 outlier samples were further removed to facilitate visualization; most of these came from the USA or Israel. See **Additional File 1: Fig. S5** for comparison.

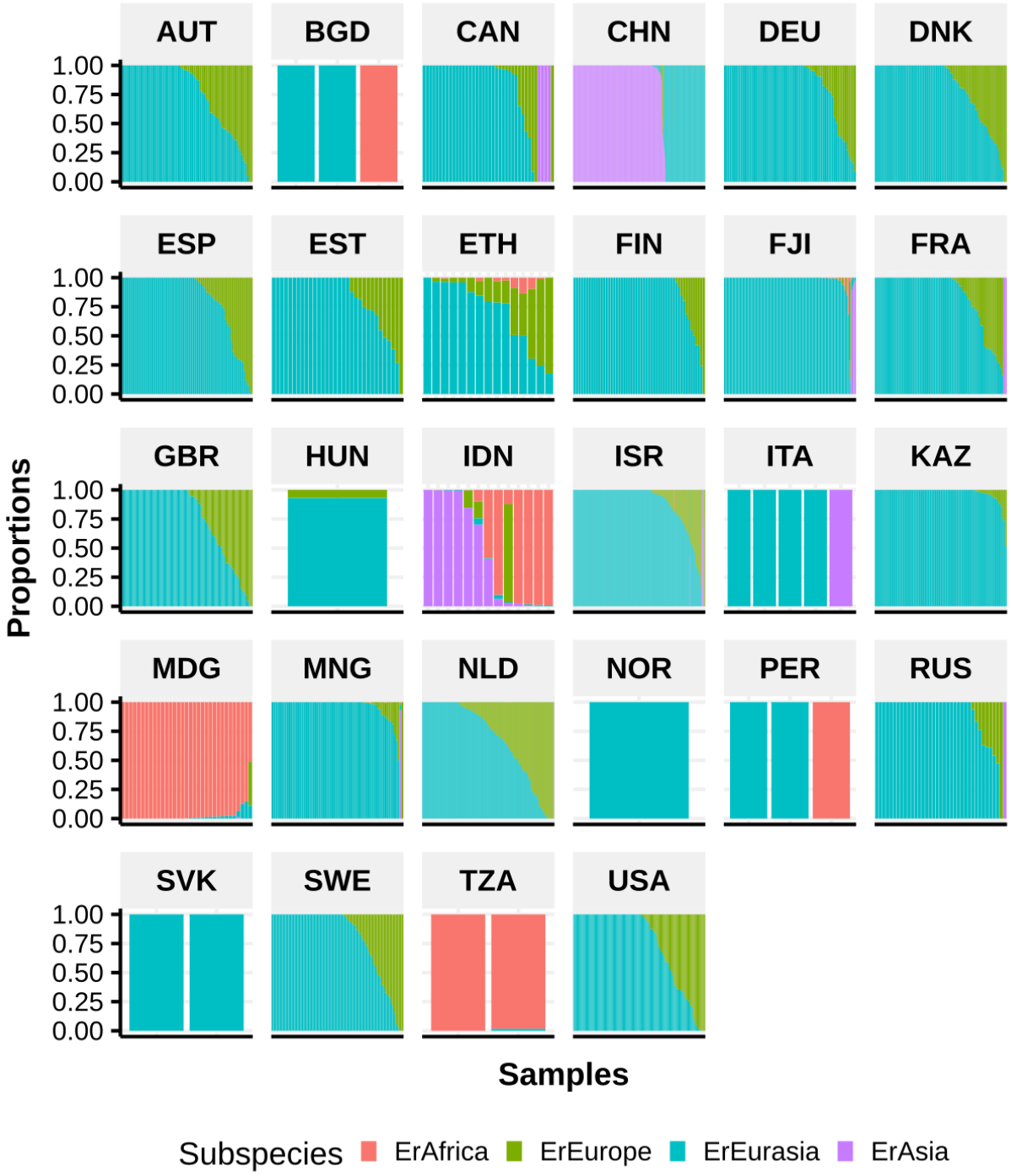


Fig S7: Barplots of subspecies relative abundances over all metagenomic samples that had sufficient coverage over subspecies-specific SNVs (**Methods**).

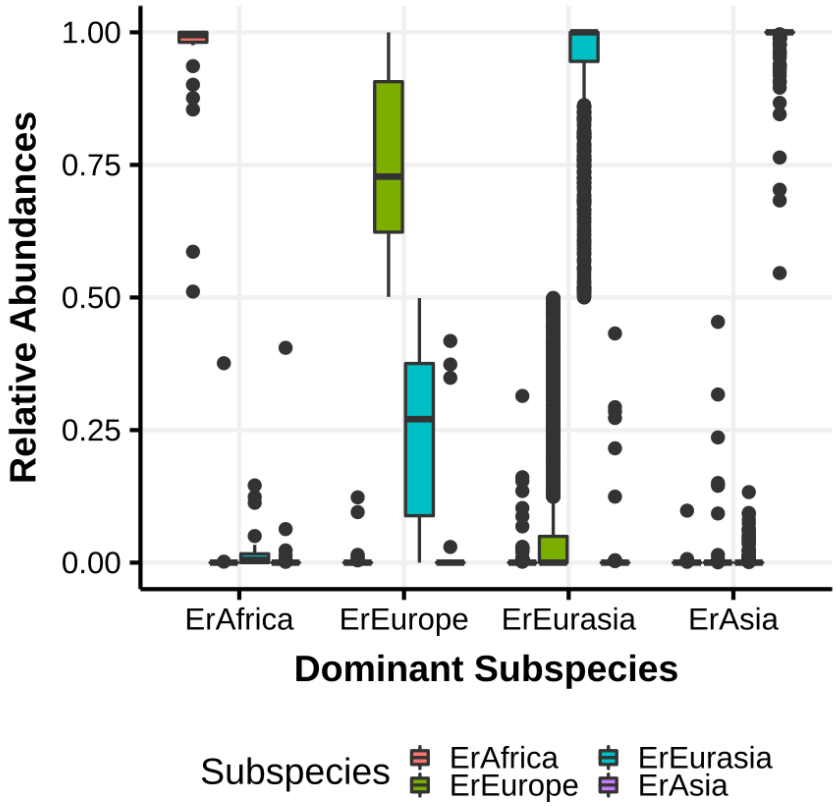


Fig S8: Boxplots of subspecies relative abundances over all metagenomic samples that had sufficient coverage over subspecies-specific SNVs (**Methods**).

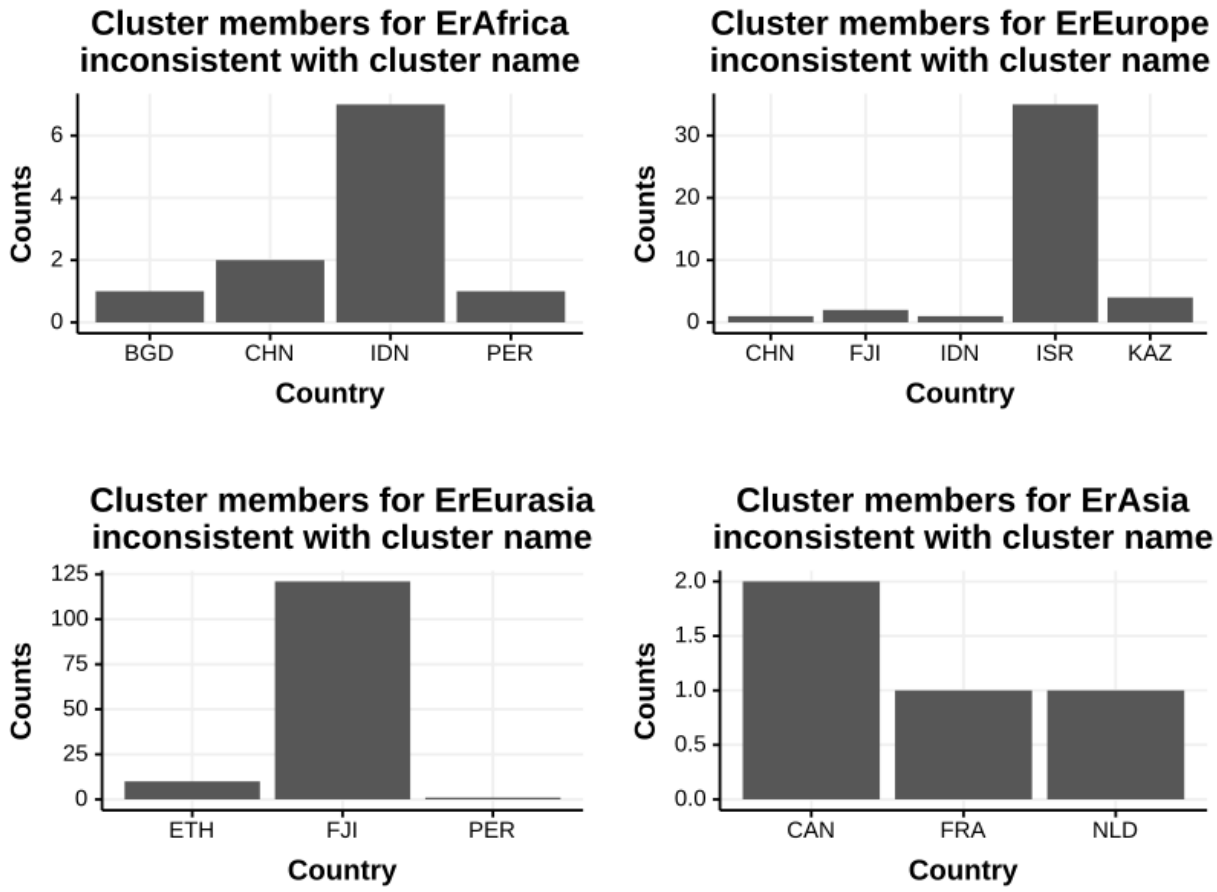


Fig S9: Barplots showing the country of origin of strains showing inconsistent membership (with respect to the subspecies name). For ErEurope and ErEurasia, strains coming from North America are not considered inconsistent.

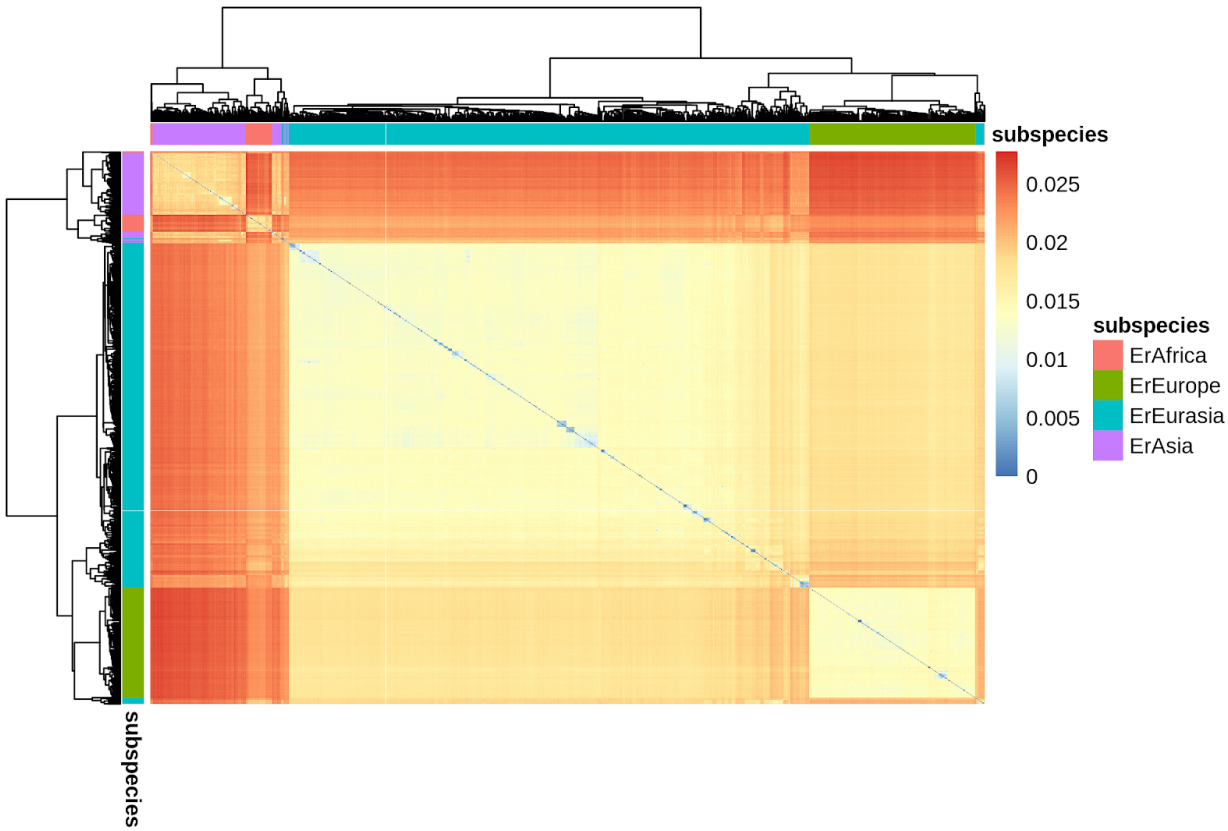


Fig S10: Heatmap of pairwise genetic distances between all high quality *E. rectale* genomes extracted from metagenomes.

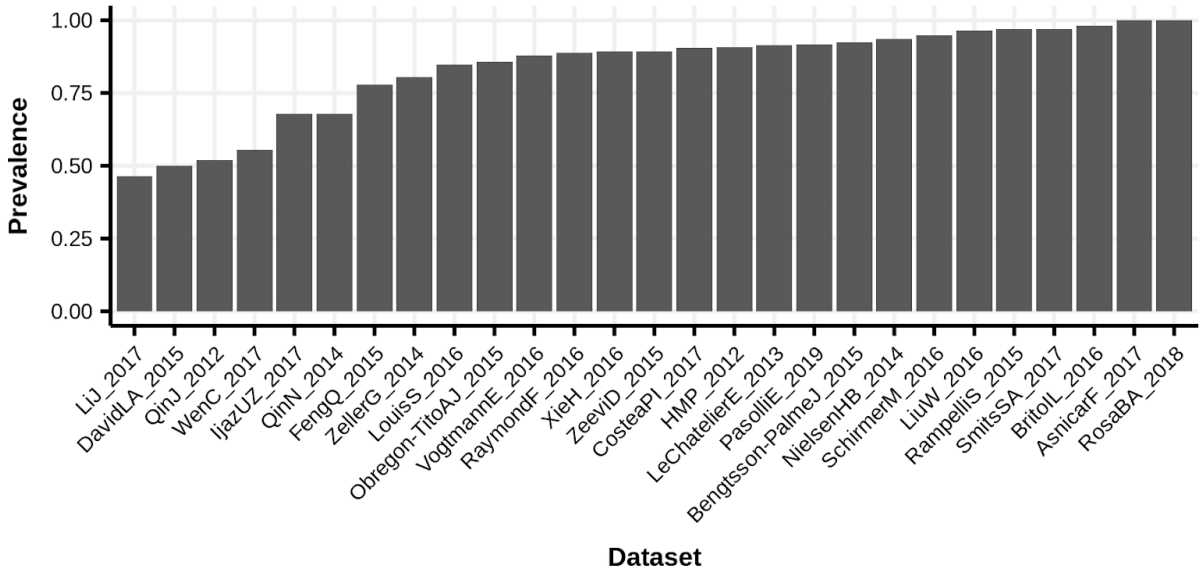


Fig S11: Prevalence proportions of *Eubacterium rectale* (Prevalence defined as relative abundance > 0.1%) in adult control samples. Some datasets did not contain any adult control samples and are thus not shown here. Relative abundances were inferred using MetaPhlAn2 (Truong et al. 2015) (**Methods**).

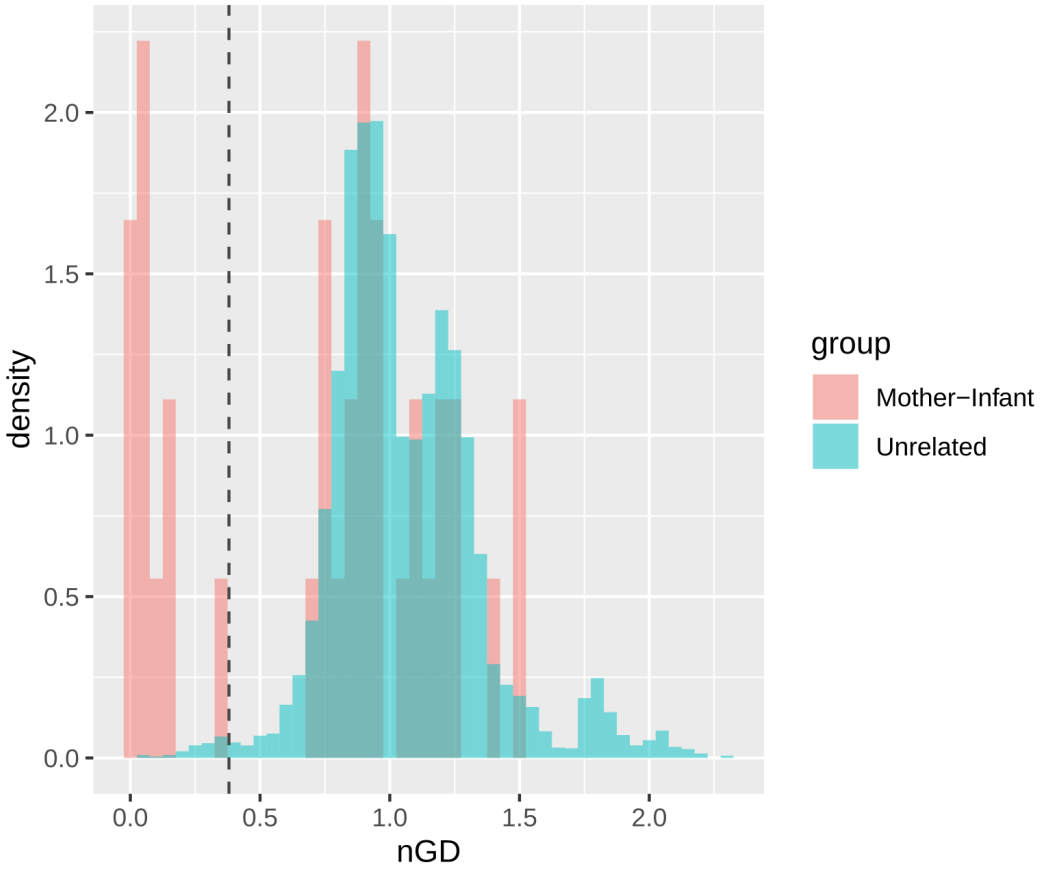


Fig S12: Distribution of genetic distances (as normalized branch lengths) between genomes isolated from Mother-infant and unrelated pairs of individuals. Datasets used were Backhed et al (N=398 samples; 96 mothers-infant pairs) (Bäckhed et al. 2015), Asnicar et al (N=18, 5 mother-infant pairs) (Asnicar et al. 2017), and Ferretti et al (N=116 samples, 21 mothers and 25 infants) (Ferretti et al. 2018). Dashed line indicates the 1-percentile of the distribution of unrelated individuals, used as a conservative cutoff to call a pair of strains identical (see **Methods**).

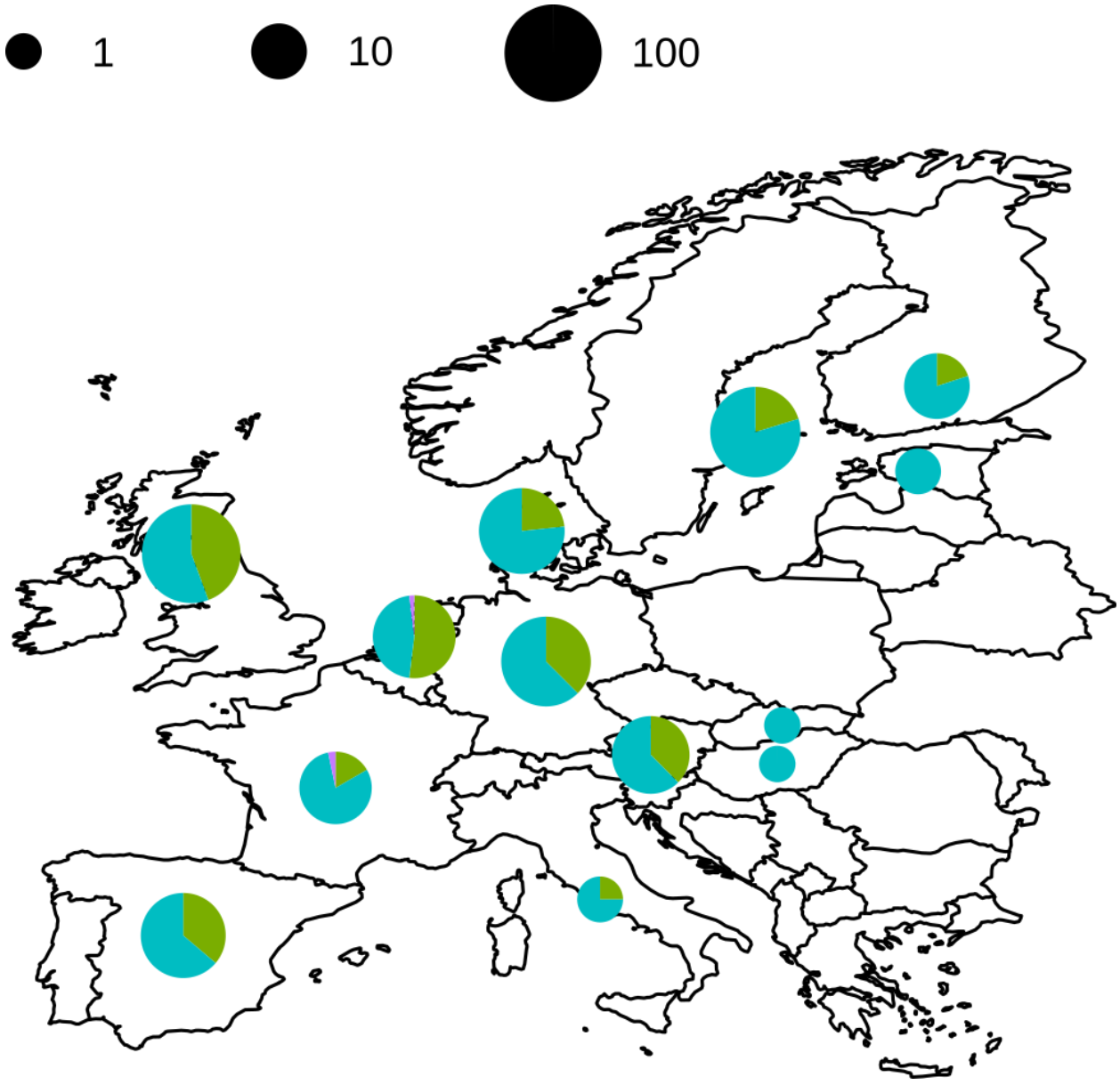


Fig S13: Ratio of subspecies prevalence within European countries. The circle size corresponds to the sample size. Cyan = ErEurasia, Green = ErEurope, Purple = ErChina.

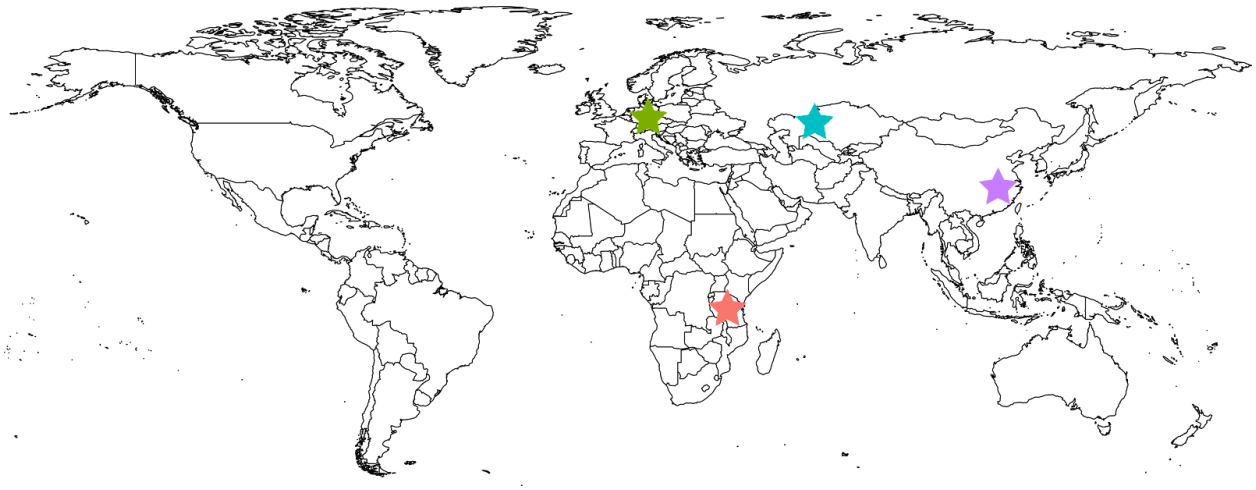


Fig S14: Point locations chosen for the four *E. rectale* subspecies. Compare with **Fig. 3**. Red = ErAfrica, Green = ErEurope, Blue = ErEurasia, Purple = ErAsia.

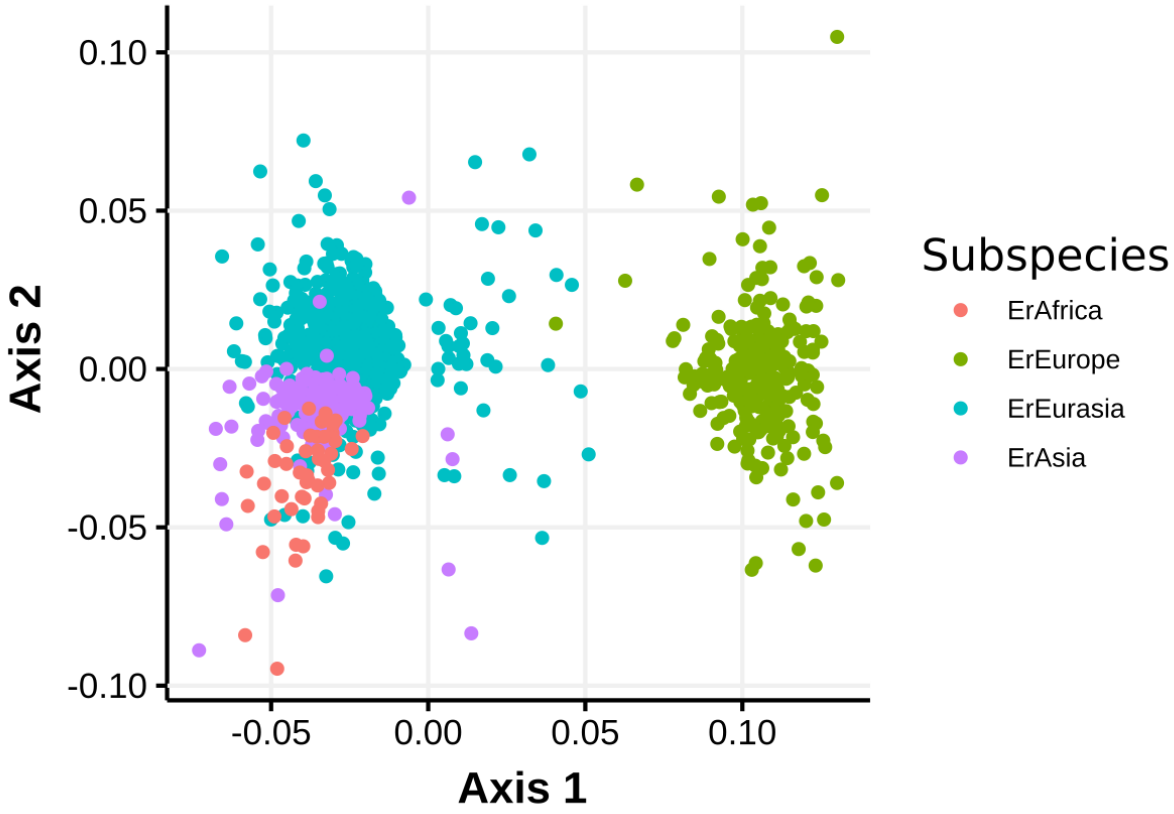


Fig S15: Ordination based on pairwise Jaccard distances computed on KO profiles (**Methods**).

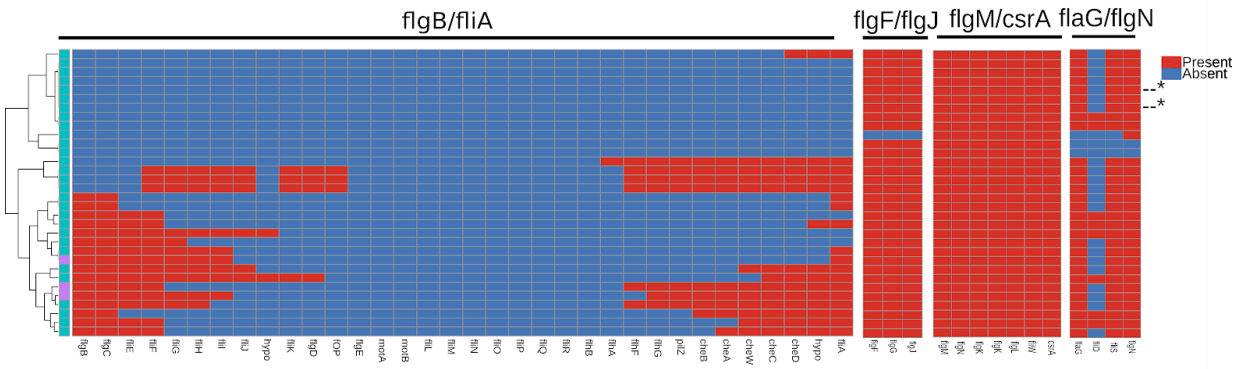


Fig S16: Heatmap of motility gene presence/absence in all those non-ErEurope strains that have some part of the *flgB*/*fliA* operon missing and that had their motility operons fully spanned on single contigs. Gene presence/absence inferred by extracting operons with bordering operon genes present (operon border genes determined using (Neville et al. 2013)). Asterisks mark isolate genomes. Row colors: Cyan = ErEurasia, Purple = ErAsia.

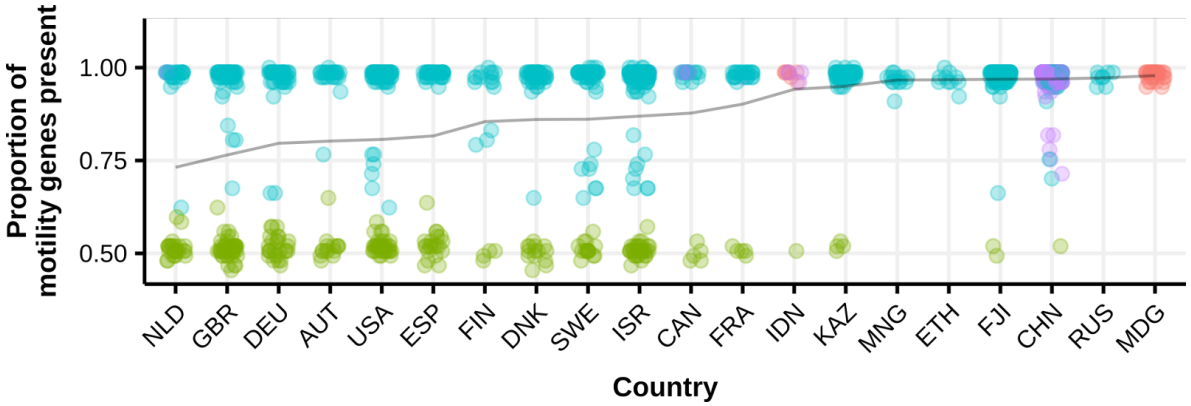


Fig S17: Proportion of motility KOs present in the HQ *E. rectale* genomes, stratified according to country and subspecies membership. Motility-association is defined as described above. Line corresponds to mean proportion per country. Only countries with at least 5 genomes are shown.

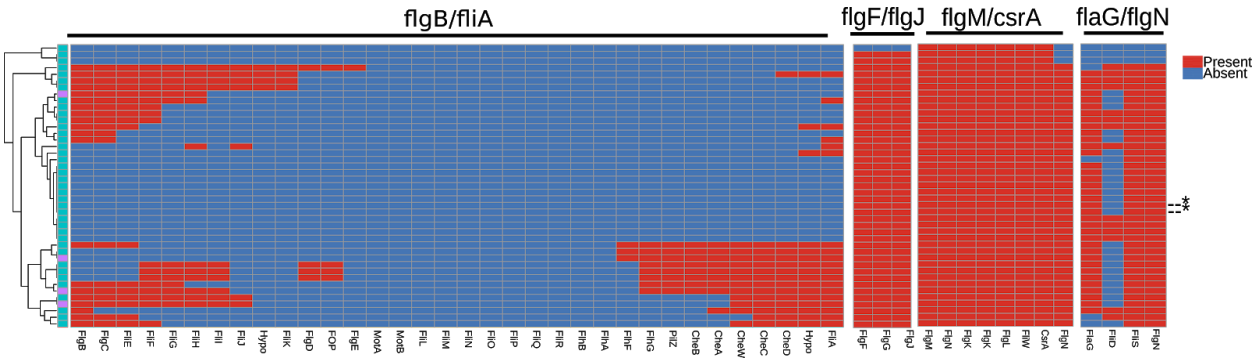


Fig S18: Heatmap of motility gene presence/absence in all those non-ErEurope strains that have some part of the *flgB*/*fliA* operon missing. Gene presence/absence inferred by mapping operon genes (sequences taken from (Neville et al. 2013)) against all extracted genomes. Asterisks mark isolate genomes. Row colors: Cyan = ErEurasia, Purple = ErAsia.

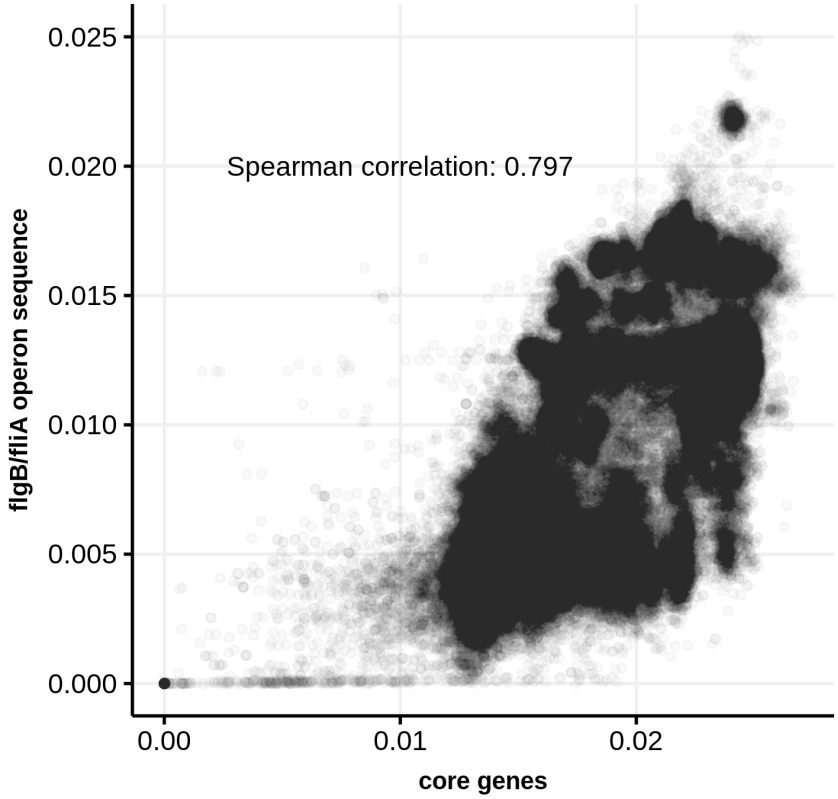


Fig S19: Scatter plot between pairwise genetic distances inferred from core genes and the largest motility operon (*flgB*/*fliA*) gene sequences for all non-ErEurope strains for which the full *flgB*/*fliA* operon could be extracted.

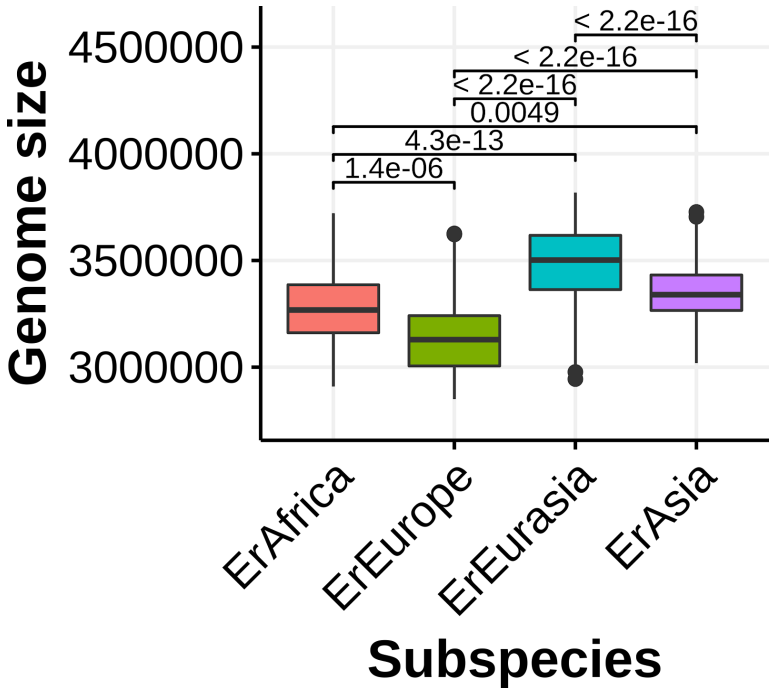


Fig S20: Boxplot of genome sizes by subspecies. P-values were calculated using a two-sided Wilcoxon test.

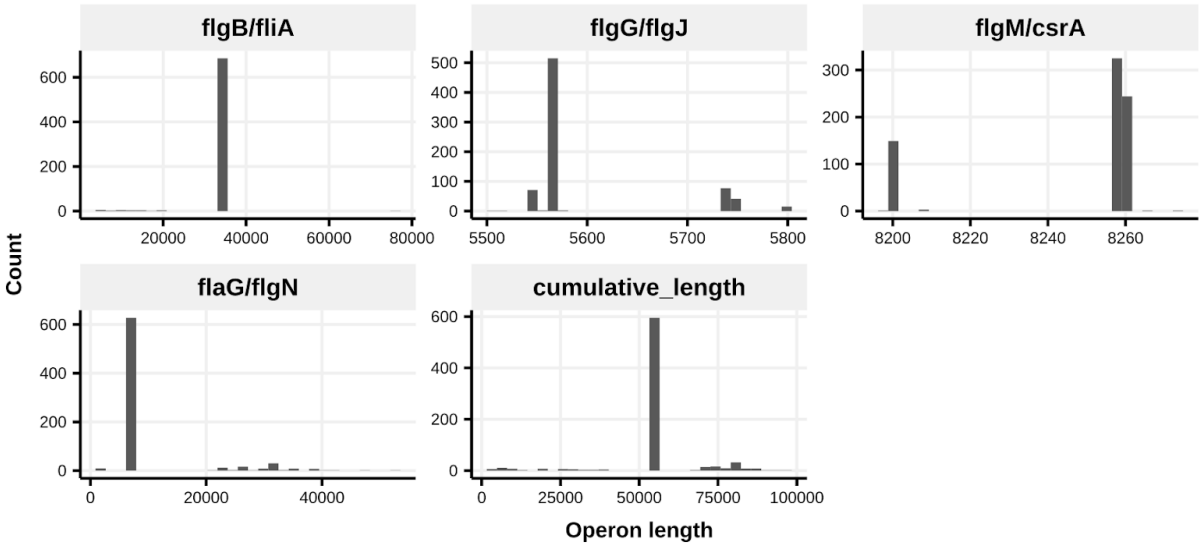


Fig S21: Histograms of operon lengths and cumulative operon length for all HQ genomes. Operon sequences and their length are generally very well conserved, with the exception of some *flaG*/*flgN* operons.

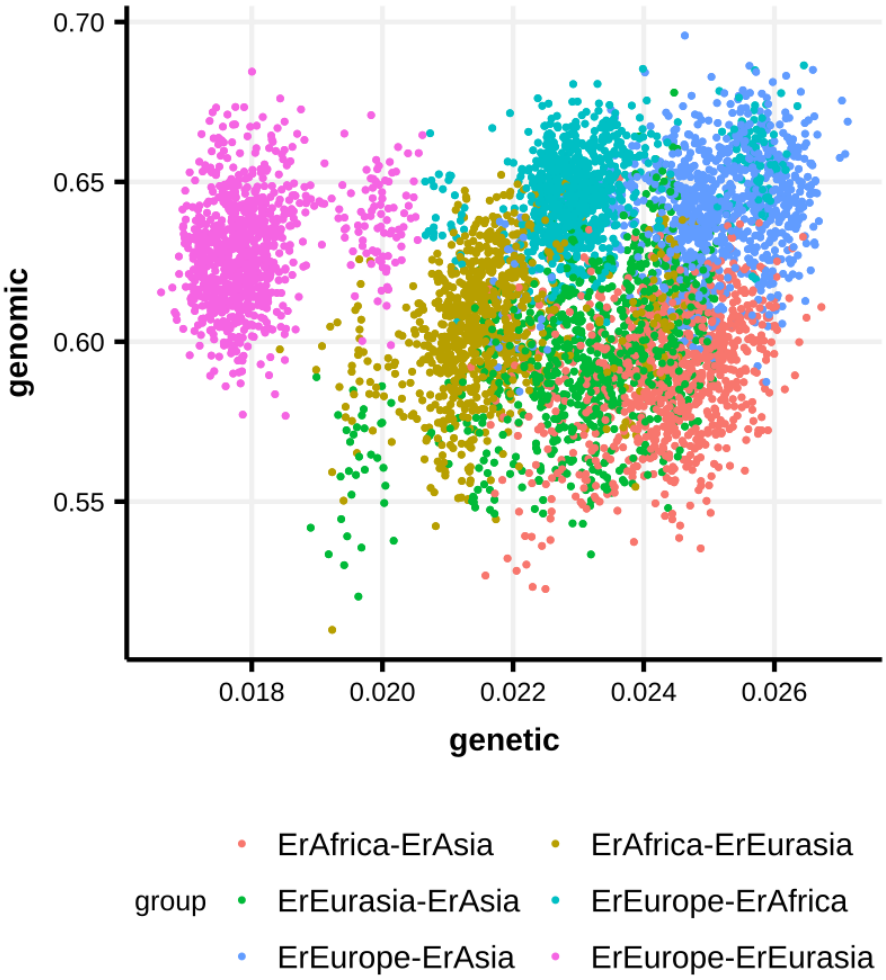


Fig S22: Scatterplot of pairwise genetic (Hamming distance on core gene alignment) and genomic (Jaccard distance on gene presence/absence, excluding gene clusters corresponding to motility operon genes) distances. For visualization purposes, genomes were subsampled to 30 random samples per subspecies.

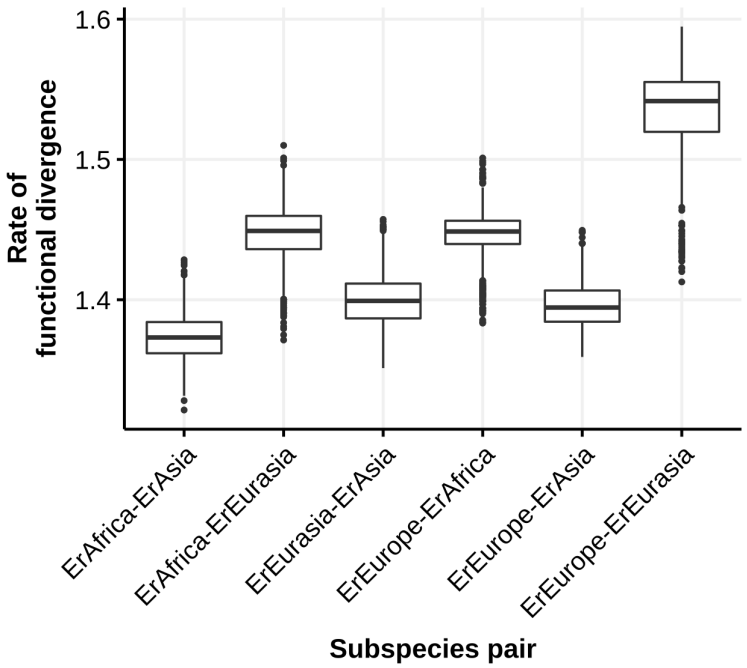


Fig S23: Functional divergence rates of pairs of subspecies, calculated by dividing pairwise inter-subspecies genomic distances by their corresponding genetic distance (**Methods**).

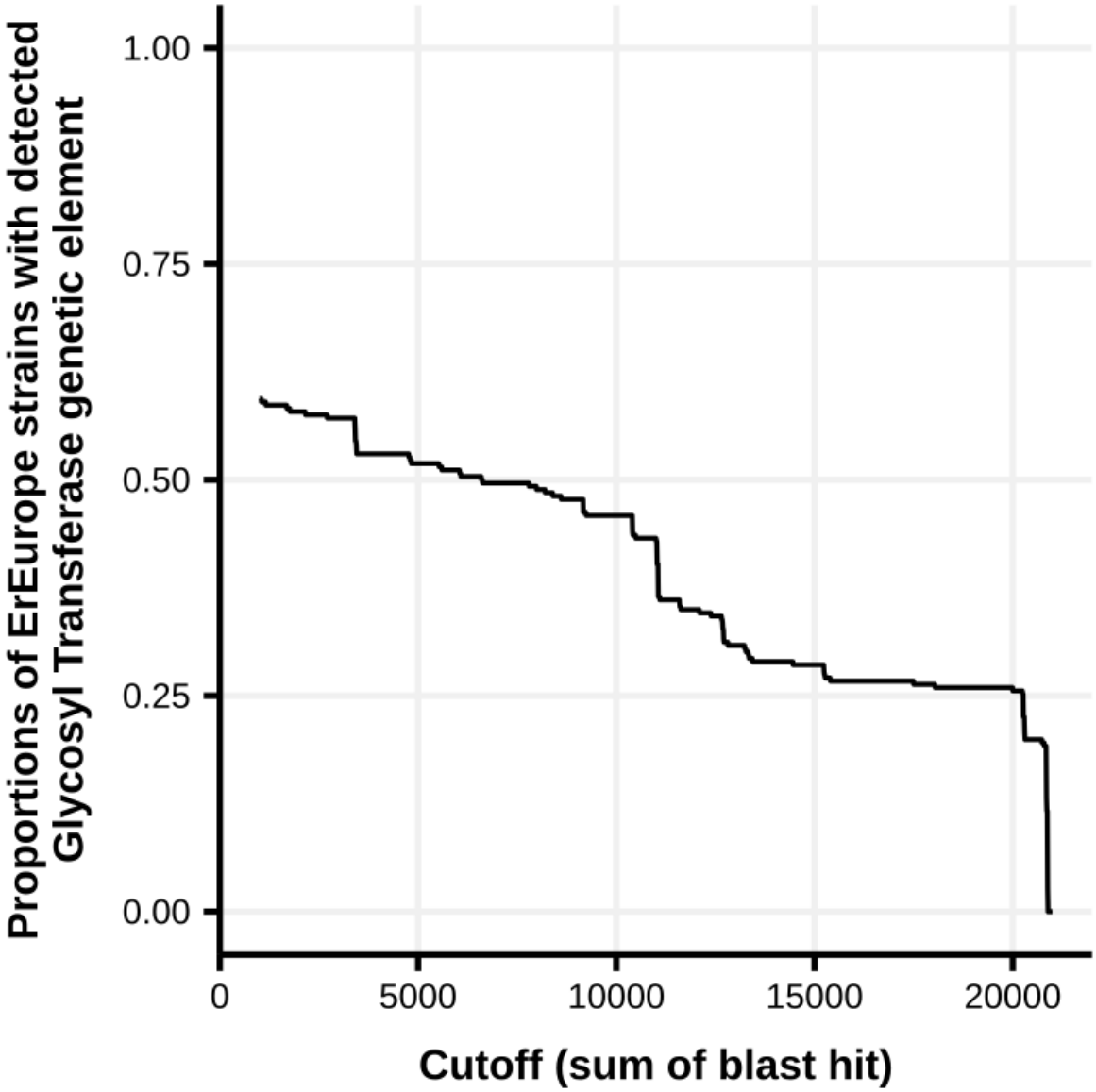


Fig S24. Line plot showing the proportion of detected GT-enriched genomic islands as a function of the total length of blast hits. This plot suggests that the true fraction of ErEurope strains possessing the genetic element is distinctly higher than ~21% (corresponding to the proportion of ErEurope strains where the full genetic element could be detected), probably due to partial assembly.

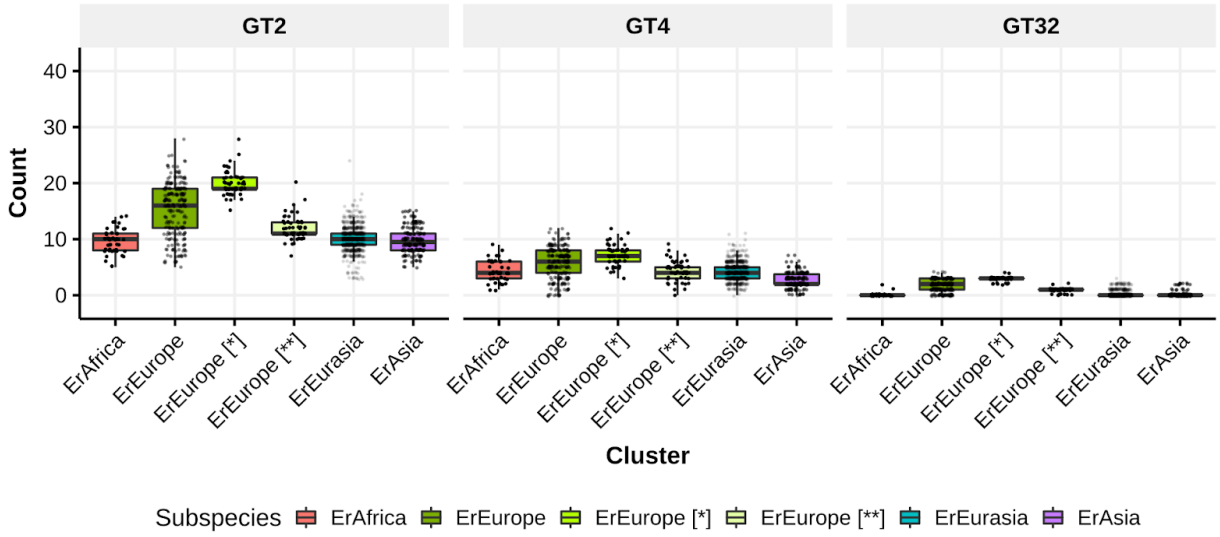


Fig S25: Boxplots of genome-wide GT counts stratified by subspecies. [*] corresponds to counts for those ErEurope strains with completely extracted GT-enriched genomic islands, [**] corresponds to those ErEurope but with counts corresponding to the GT-enriched genomics island removed. See **Fig. 6A**.

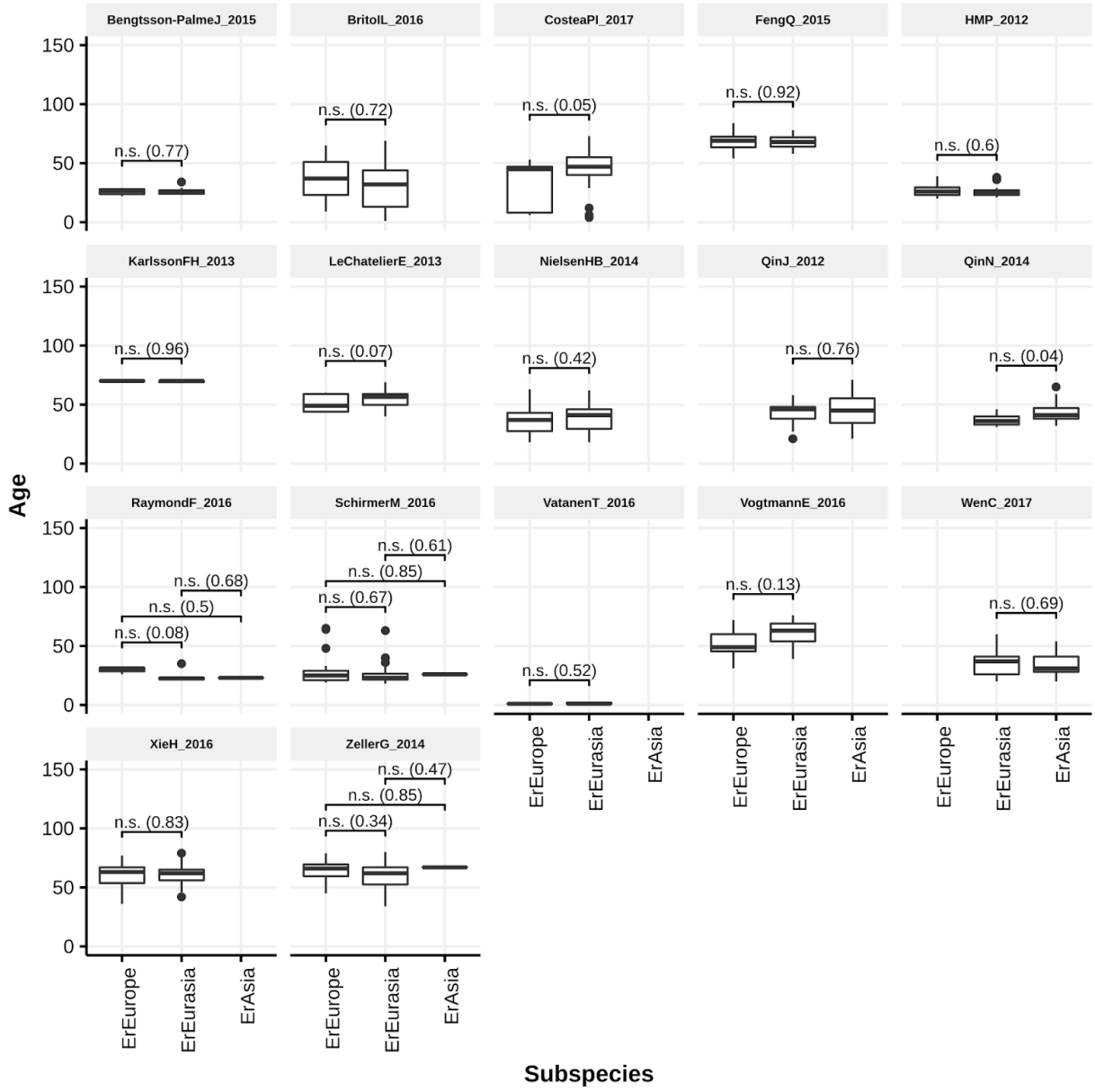


Fig S26: Boxplots of age grouped by subspecies. Label corresponds to significance level at 5% FDR (FDR-correction using Benjamini-Hochberg), numbers in parenthesis correspond to uncorrected p-values. P-values calculated using two-sided Wilcoxon tests.

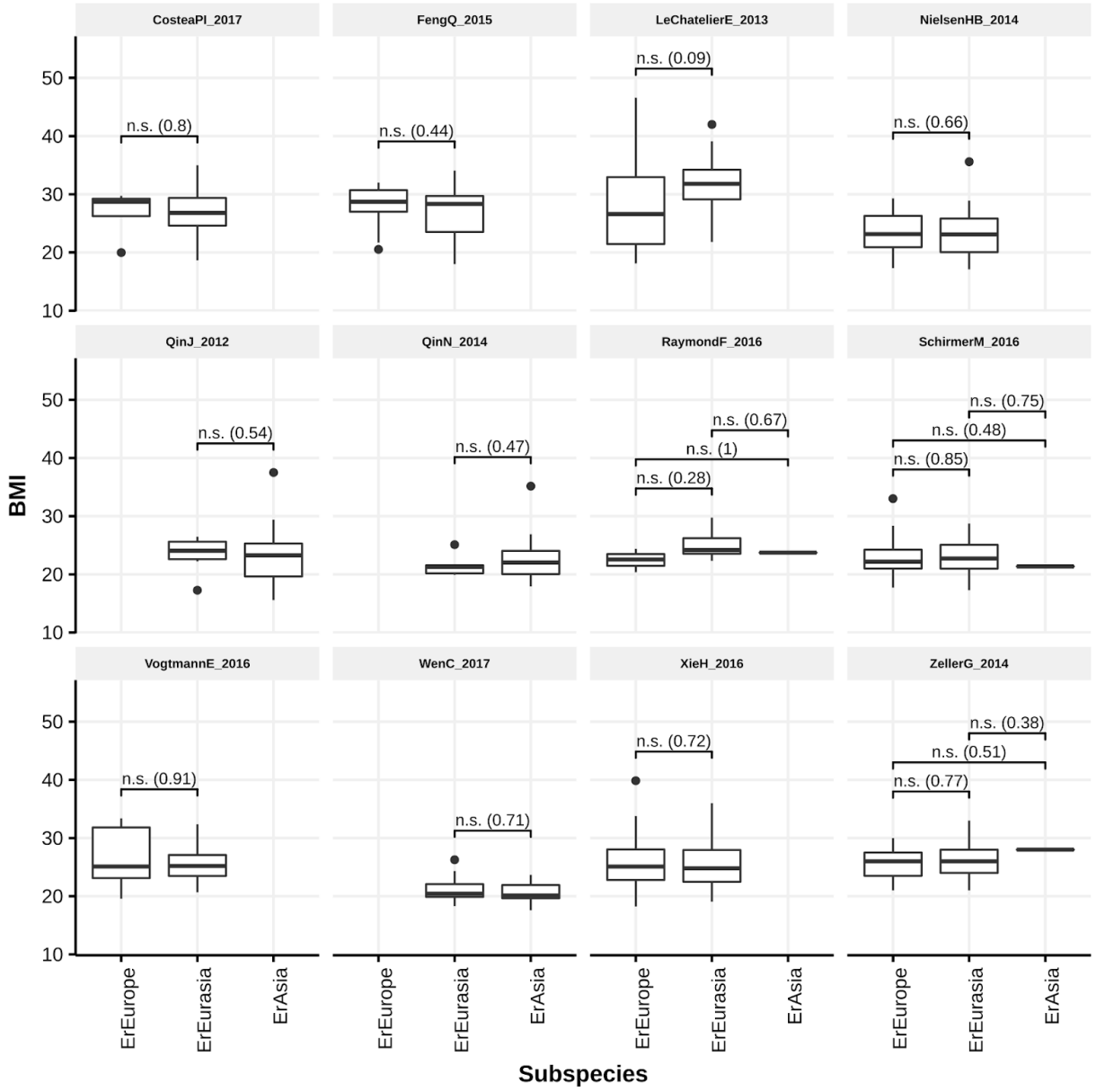


Fig S27: Boxplots of BMI grouped by subspecies. Label corresponds to significance level at 5% FDR (FDR-correction using Benjamini-Hochberg), numbers in parenthesis correspond to uncorrected p-values. P-values calculated using two-sided Wilcoxon tests.

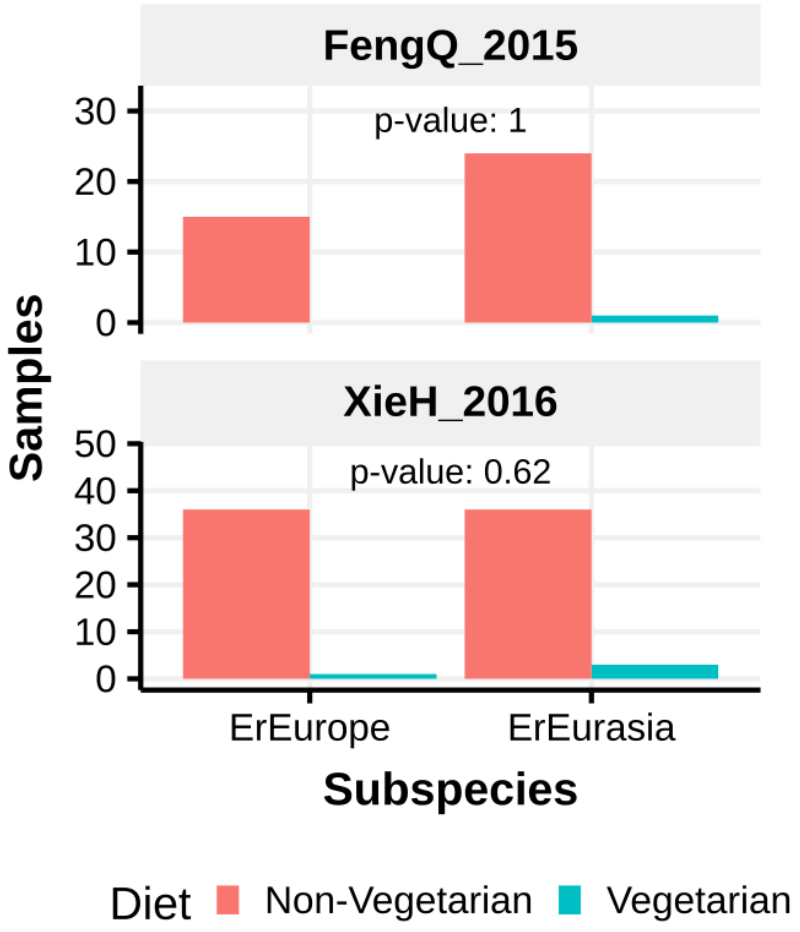


Fig S28: Bar Plots showing the distribution of ErEurope and ErEurasia in two datasets where qualitative diet information (vegetarian/non-vegetarian) was available. P-values were calculated using a two-sided Fisher test.

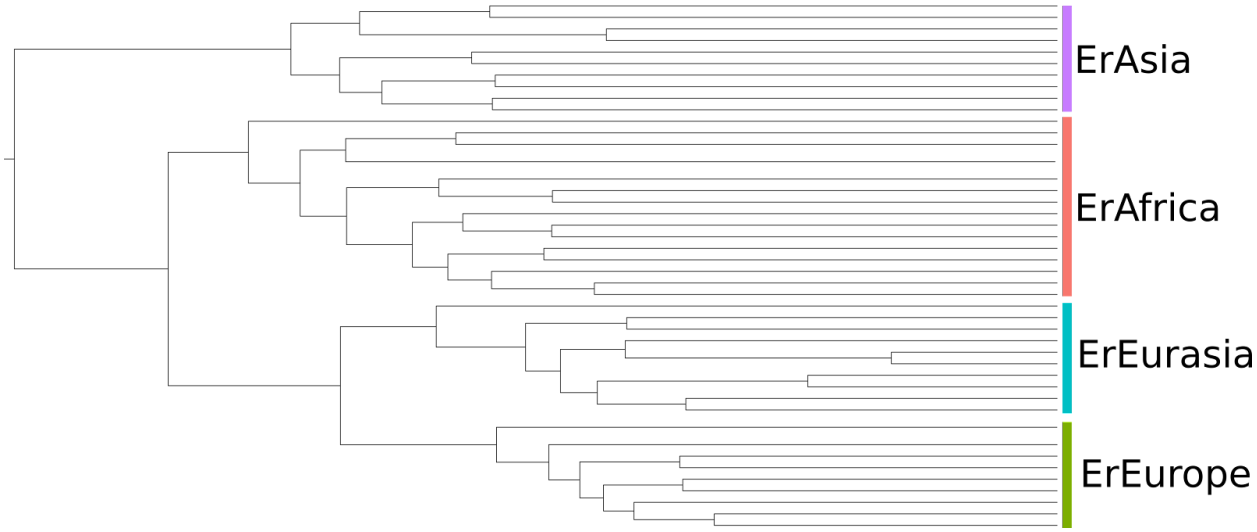


Fig S29: Rooted bayesian phylogeny (Bouckaert et al. 2014) built on a randomly chosen, representative subset of samples per subspecies (**Methods**). Clade topology is consistent across subsamples.

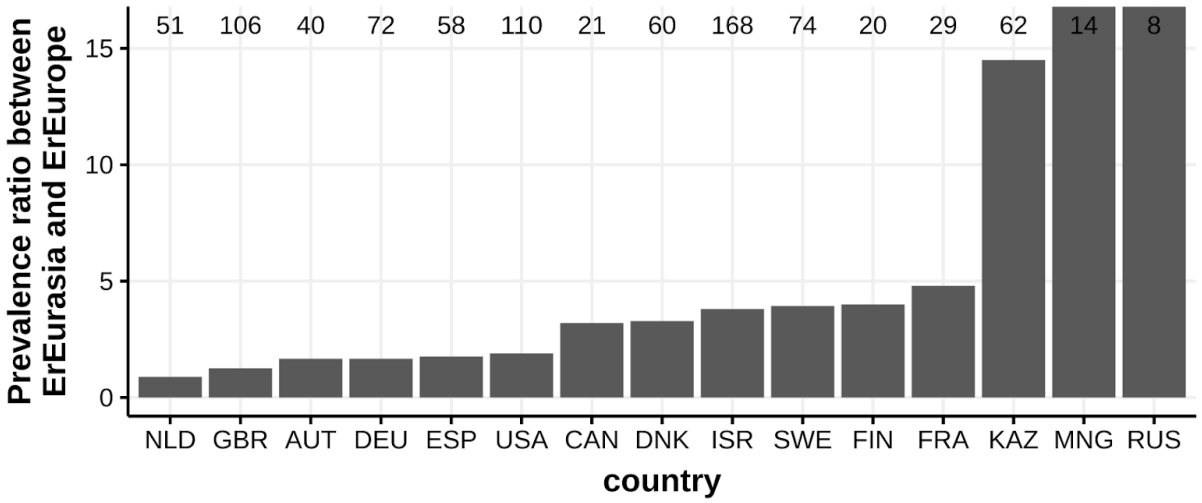


Fig S30: Ratios of prevalence between ErEurasia and ErEurope for Eurasian/North American countries. Mongolia and Russia have undefined ratios, since no ErEurope genomes were reconstructed from samples originating in these countries. The number above the bars indicates the number of genomes reconstructed from each country. Only countries with at least 5 genomes are shown.

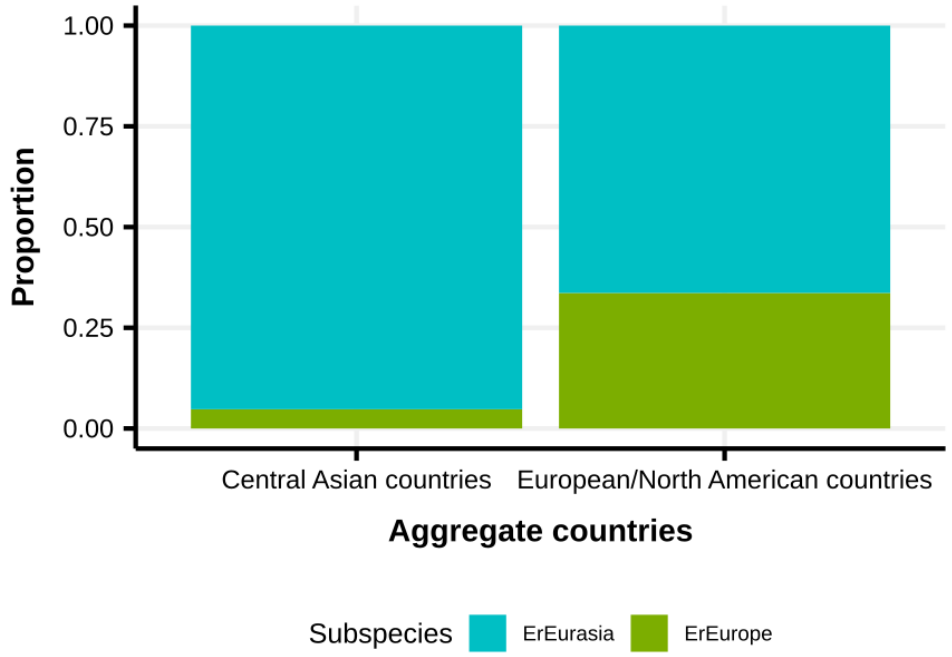


Fig S31: Prevalence ratios of ErEurasia and ErEurope between Kazakhstan, Mongolia and Russia (in aggregate) against the remaining countries in Europe and North America. P-value for differential prevalence is 9.8E-09 (Fisher test).

Additional Files

For Supplementary Tables see the online version of this publication.

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*, February. https://doi.org/10.1038/s41586-019-0965-1.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2019. “A Unified Sequence Catalogue of over 280,000 Genomes Obtained from the Human Gut Microbiome.” *bioRxiv*. https://doi.org/10.1101/762682.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Amato, Katherine R., Jon G Sanders, Se Jin Song, Michael Nute, Jessica L. Metcalf, Luke R. Thompson, James T. Morton, et al. 2019. “Evolutionary Trends in Host Physiology Outweigh Dietary Niche in Structuring Primate Gut Microbiomes.” *The ISME Journal* 13 (3): 576–87.

Asnicar, Francesco, Serena Manara, Moreno Zolfo, Duy Tin Truong, Matthias Scholz, Federica Armanini, Pamela Ferretti, et al. 2017. “Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling.” *mSystems* 2 (1). https://doi.org/10.1128/mSystems.00164-16.

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.

Barcenilla, A., S. E. Pryde, J. C. Martin, S. H. Duncan, C. S. Stewart, C. Henderson, and H. J. Flint. 2000. “Phylogenetic Relationships of Butyrate-Producing Bacteria from the Human Gut.” *Applied and Environmental Microbiology* 66 (4): 1654–61.

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.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. 2014. “BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.” *PLoS Computational Biology* 10 (4): e1003537.

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 (August): 725.

Bruzzese, Eugenia, Maria Luisa Callegari, Valeria Raia, Sara Viscovo, Riccardo Scotto, Susanna Ferrari, Lorenzo Morelli, et al. 2014. “Disrupted Intestinal Microbiota and Intestinal Inflammation in Children with Cystic Fibrosis and Its Restoration with Lactobacillus GG: A Randomised Clinical Trial.” *PloS One* 9 (2): e87796.

Cantarel, Brandi L., Pedro M. Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. 2009. “The Carbohydrate-Active EnZymes

Database (CAZy): An Expert Resource for Glycogenomics." *Nucleic Acids Research* 37 (Database issue): D233–38.
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.
CAZypedia Consortium. 2018. "Ten Years of CAZypedia: A Living Encyclopedia of Carbohydrate-Active Enzymes." *Glycobiology* 28 (1): 3–8.
Costea, Paul I., Luis Pedro Coelho, Shinichi Sunagawa, Robin Munch, Jaime Huerta-Cepas, Kristoffer Forslund, Falk Hildebrand, Almagul Kushugulova, Georg Zeller, and Peer Bork. 2017. "Subspecies in the Global Human Gut Microbiome." *Molecular Systems Biology* 13 (12): 960.
Darling, Aaron E., Bob Mau, and Nicole T. Perna. 2010. "progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement." *PloS One* 5 (6): e11147.
Delport, Wayne, Michael Cunningham, Brenda Olivier, Oliver Preisig, and Schalk W. van der Merwe. 2006. "A Population Genetics Pedigree Perspective on the Transmission of Helicobacter Pylori." *Genetics* 174 (4): 2107–18.
Diniz-Filho, José Alexandre F., Thannya N. Soares, Jacqueline S. Lima, Ricardo Dobrovolski, Victor Lemes Landeiro, Mariana Pires de Campos Telles, Thiago F. Rangel, and Luis Mauricio Bini. 2013. "Mantel Test in Population Genetics." *Genetics and Molecular Biology* 36 (4): 475–85.
Duncan, Sylvia H., and Harry J. Flint. 2008. "Proposal of a Neotype Strain (A1-86) for Eubacterium Rectale. Request for an Opinion." *International Journal of Systematic and Evolutionary Microbiology* 58 (Pt 7): 1735–36.
Duncan, Sylvia H., Wendy R. Russell, Andrea Quartieri, Maddalena Rossi, Julian Parkhill, Alan W. Walker, and Harry J. Flint. 2016. "Wheat Bran Promotes Enrichment within the Human Colonic Microbiota of Butyrate-Producing Bacteria That Release Ferulic Acid." *Environmental Microbiology* 18 (7): 2214–25.
Eddy, S. R. 1998. "Profile Hidden Markov Models." *Bioinformatics* 14 (9): 755–63.
Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. "Anvi'o: An Advanced Analysis and Visualization Platform for 'Omics Data." *PeerJ* 3 (October): e1319.
Falush, Daniel, Thierry Wirth, Bodo Linz, Jonathan K. Pritchard, Matthew Stephens, Mark Kidd, Martin J. Blaser, et al. 2003. "Traces of Human Migrations in Helicobacter Pylori Populations." *Science* 299 (5612): 1582–85.
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.
Fite, Alemu, Sandra Macfarlane, Elizabeth Furrie, Bahram Bahrami, John H. Cummings, Douglas T. Steinke, and George T. Macfarlane. 2013. "Longitudinal Analyses of Gut Mucosal Microbiotas in Ulcerative Colitis in Relation to Patient Age and Disease Severity and Duration." *Journal of Clinical Microbiology* 51 (3): 849–56.
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.
Hijmans, Robert J. 2017. “Geosphere: Spherical Trigonometry.” https://CRAN.R-project.org/package=geosphere.
Huerta-Cepas, Jaime, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. 2017. “Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper.” *Molecular Biology and Evolution* 34 (8): 2115–22.
Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (March): 119.
Jackson, Debra W., Kazushi Suzuki, Lawrence Oakford, Jerry W. Simecka, Mark E. Hart, and Tony Romeo. 2002. “Biofilm Formation and Dispersal under the Influence of the Global Regulator CsrA of Escherichia Coli.” *Journal of Bacteriology* 184 (1): 290–301.
Kabeerdoss, Jayakanthan, Prabavathi Jayakanthan, Srinivasan Pugazhendhi, and Balakrishnan S. Ramakrishna. 2015. “Alterations of Mucosal Microbiota in the Colon of Patients with Inflammatory Bowel Disease Revealed by Real Time Polymerase Chain Reaction Amplification of 16S Ribosomal Ribonucleic Acid.” *The Indian Journal of Medical Research* 142 (1): 23–32.
Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.
Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities.” *PeerJ* 3 (August): e1165.
Kang, Dongwan, Feng Li, Edward S. Kirton, Ashleigh Thomas, Rob S. Egan, Hong An, and Zhong Wang. 2019. “MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies.” e27522v1. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27522v1.
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.
Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data. an Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics.
Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. “Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences.” *Genome Research* 26 (12): 1721–29.
Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. “Genomic Insights That Advance the Species Definition for Prokaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (7): 2567–72.
Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.
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.

Linz, Bodo, François Balloux, Yoshan Moodley, Andrea Manica, Hua Liu, Philippe Roumagnac, Daniel Falush, et al. 2007. “An African Origin for the Intimate Association between Humans and Helicobacter Pylori.” *Nature* 445 (7130): 915–18.
Liu, M. Y., G. Gui, B. Wei, J. F. Preston 3rd, L. Oakford, U. Yüksel, D. P. Giedroc, and T. Romeo. 1997. “The RNA Molecule CsrB Binds to the Global Regulatory Protein CsrA and Antagonizes Its Activity in Escherichia Coli.” *The Journal of Biological Chemistry* 272 (28): 17502–10.
Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. 2014. “The Carbohydrate-Active Enzymes Database (CAZy) in 2013.” *Nucleic Acids Research* 42 (Database issue): D490–95.
Lopez-Siles, Mireia, Tanweer M. Khan, Sylvia H. Duncan, Hermie J. M. Harmsen, L. Jesús Garcia-Gil, and Harry J. Flint. 2012. “Cultured Representatives of Two Major Phylogroups of Human Colonic Faecalibacterium Prausnitzii Can Utilize Pectin, Uronic Acids, and Host-Derived Substrates for Growth.” *Applied and Environmental Microbiology* 78 (2): 420–28.
Manara, Serena, Francesco Asnicar, Francesco Beghini, Davide Bazzani, Fabio Cumbo, Moreno Zolfo, Eleonora Nigro, et al. 2019. “Microbial Genomes from Non-Human Primate Gut Metagenomes Expand the Primate-Associated Bacterial Tree of Life with over 1000 Novel Species.” *Genome Biology* 20 (1): 299.
Marcy, Yann, Cleber Ouverney, Elisabeth M. Bik, Tina Lösekann, Natalia Ivanova, Hector Garcia Martin, Ernest Szeto, et al. 2007. “Dissecting Biological ‘dark Matter’ with Single-Cell Genetic Analysis of Rare and Uncultivated TM7 Microbes from the Human Mouth.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (29): 11889–94.
Miyazaki, K., J. C. Martin, R. Marinsek-Logar, and H. J. Flint. 1997. “Degradation and Utilization of Xylans by the Rumen Anaerobe Prevotella Bryantii (formerly P. Ruminicola Subsp. Brevis) B(1)4.” *Anaerobe* 3 (6): 373–81.
Moodley, Yoshan, Bodo Linz, Robert P. Bond, Martin Nieuwoudt, Himla Soodyall, Carina M. Schlebusch, Steffi Bernhöft, et al. 2012. “Age of the Association between Helicobacter Pylori and Man.” *PLoS Pathogens* 8 (5): e1002693.
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*, March. https://doi.org/10.1038/s41586-019-1058-x.
Neville, B. Anne, Paul O. Sheridan, Hugh M. B. Harris, Simone Coughlan, Harry J. Flint, Sylvia H. Duncan, Ian B. Jeffery, et al. 2013. “Pro-Inflammatory Flagellin Proteins of Prevalent Motile Commensal Bacteria Are Variably Abundant in the Intestinal Microbiome of Elderly Humans.” *PloS One* 8 (7): e68919.
Nicholls, Samuel M., Joshua C. Quick, Shuiquan Tang, and Nicholas J. Loman. 2019. “Ultra-Deep, Long-Read Nanopore Sequencing of Mock Microbial Community Standards.” *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz043.
Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. “metaSPAdes: A New Versatile Metagenomic Assembler.” *Genome Research* 27 (5): 824–34.
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.

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* 31 (22): 3691–93.

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: 1–14.

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.

Ríos-Covián, David, Patricia Ruas-Madiedo, Abelardo Margolles, Miguel Gueimonde, Clara G. de Los Reyes-Gavilán, and Nuria Salazar. 2016. “Intestinal Short Chain Fatty Acids and Their Link with Diet and Human Health.” *Frontiers in Microbiology* 7 (February): 185.

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.

Schwarz, Sandra, Giovanna Morelli, Barica Kusecek, Andrea Manica, Francois Balloux, Robert J. Owen, David Y. Graham, Schalk van der Merwe, Mark Achtman, and Sebastian Suerbaum. 2008. “Horizontal versus Familial Transmission of Helicobacter Pylori.” *PLoS Pathogens* 4 (10): e1000180.

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*.

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* 4: 2304.

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. n.d. “Microbial Genomes from Gut Metagenomes of Non-Human Primates Expand the Primate-Associated Bacterial Tree-of-Life with over 1,000 Novel Species.”

Sonnleitner, Elisabeth, Alessandra Romeo, and Udo Bläsi. 2012. “Small Regulatory RNAs in Pseudomonas Aeruginosa.” *RNA Biology* 9 (4): 364–71.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and

Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.
Tett, Adrian, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, et al. 2019a. “The Prevotella Copri Complex Comprises Four Distinct Clades That Are Underrepresented in Westernised Populations.” *bioRxiv*. https://doi.org/10.1101/600593.
———. 2019b. “The Prevotella Copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations.” *Cell Host & Microbe*, September. https://doi.org/10.1016/j.chom.2019.08.018.
The UniProt Consortium. 2017. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 45 (D1): D158–69.
Tibshirani, Robert, and Guenther Walther. 2005. “Cluster Validation by Prediction Strength.” *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 14 (3): 511–28.
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.
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.
Wright, S. 1943. “Isolation by Distance.” *Genetics* 28 (2): 114–38.
Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. “dbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme Annotation.” *Nucleic Acids Research* 40 (Web Server issue): W445–51.
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 (November): 766.
Zhang, Linsheng, Lillian Gray, Richard P. Novick, and Guangyong Ji. 2002. “Transmembrane Topology of AgrB, the Protein Involved in the Post-Translational Modification of AgrD in Staphylococcus Aureus.” *The Journal of Biological Chemistry* 277 (38): 34736–42.

Chapter 3 | Genomic diversity and ecology of human-associated *Akkermansia* species in the gut microbiome revealed by extensive metagenomic assembly

3.1 | Context and contribution

After having confirmed that MAGs of human gut bacteria can be used to conduct comparative genomics on *E. rectale* (Chapter 2), we next wanted to extend this idea to a human gut bacterium that has medical relevance: we decided for *Akkermansia muciniphila* since this species was previously found associated with leanness in humans and mice. In further studies, investigators supplemented obese humans with *Akkermansia muciniphila* and could show weight reduction and improvement of insulin- and overweight blood parameters upon supplementation compared to placebo, suggesting mechanistic involvement in controlling diabetes. Furthermore, *Akkermansia muciniphila* is the only known species of the *Akkermansia* genus and indeed the only species in the *Verrucomicrobium* phylum in the human gut, showcasing its phylogenetic peculiarity, and only the type strain has been investigated so far. Thus, *Akkermansia muciniphila* is a very interesting but understudied human gut bacterium.

I obtained *Akkermansia* spp. genomes generated by Pasolli *et al.* (Pasolli et al. 2019) and contextualized them with *Akkermansia* spp. isolate genomes. My work revealed that *Akkermansia glycaniphila*, the only other *Akkermansia* species described to date, cannot be found in the human gut, but that a total of four undescribed *Akkermansia* species-level groups, next to the known *A. muciniphila* species, can be delineated in the human gut, all of which are likely specific to humans. Of those, only *A. muciniphila* was confirmed to be significantly negatively associated with host BMI. Further analysis revealed co-exclusion of *Akkermansia* spp. within individuals and differences in B12 synthesis capacities. Cross-referencing of CRISPR spacer inserts with a viral database revealed putative cognate *Akkermansia* sp. phage pairs, differential presence of subspecies in humans and mice and two distinct putative lipopolysaccharide synthesis operons that drive a large fraction of intra-species gene-content differences in *Akkermansia muciniphila*. Overall, this work led to the discovery and differentiation of hitherto undescribed *Akkermansia* spp. in the human gut and also further shows that MAG-based comparative genomics is possible (See also Chapter 2).

In this work I was involved in conceptualization and took over parts of the analysis (comparison of whole-genome to 16S genetic distances, corrin ring gene analysis, CRISPR analysis and the discovery/analysis of the lipopolysaccharide operon variants), interpreted the results and wrote the manuscript.

3.2 | Manuscript

Genomic diversity and ecology of human-associated *Akkermansia* species in the gut microbiome revealed by extensive metagenomic assembly

Karcher, N.[1,*], Nigro, E.[2,*], Punčochář, M.[1], Blanco-Míguez, A.[1], Ciciani, M.[1], Manghi, P.[1], Zolfo, M.[1], Cumbo, F.[1], Manara, S.[1], Golzato, D.[1], Cereseto, A.[1], Arumugam, M.[2], Nam Bui, T. P.[3], Tytgat, H.L.P.[3,4], Valles-Colomerm, M.[1,*], de Vos, W.M.[3,5,*], Segata, N.[6,7,*]

Genome Biol. 2021 Jul 14;22(1):209. doi: 10.1186/s13059-021-02427-7.

Affiliations
1 Department CIBIO, University of Trento, Trento, Italy.
2 Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
3 Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands.
4 Current address: Nestlé Institute of Health Sciences, Nestlé Research, Société des Produits Nestlé S.A., Lausanne, Switzerland.
5 Human Microbiome Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland.
6 Department CIBIO, University of Trento, Trento, Italy. nicola.segata@unitn.it.
7 IEO, European Institute of Oncology IRCCS, Milan, Italy. nicola.segata@unitn.it.
* Contributed equally.

Note: The version below is the ahead-of-print version of the manuscript, prior to editorial edits.

Abstract

Background

Akkermansia muciniphila is a human gut microbe with a key role in the physiology of the intestinal mucus layer and reported associations with decreased body mass and increased gut barrier function and health. Despite its biomedical relevance, the genomic diversity of *A. muciniphila* remains understudied and that of closely related species, except for *A. glycaniphila*, unexplored.

Results

We present a large-scale population genomics analysis of the *Akkermansia* genus using 188 isolate genomes and 2,226 genomes assembled from 18,600 metagenomes from humans and other animals. While we do not detect *A. glycaniphila*, the *Akkermansia*

strains in the human gut can be grouped into five distinct candidate species, including *A. muciniphila*, that show remarkable whole-genome divergence despite surprisingly similar 16S rRNA gene sequences. These candidate species are likely human-specific, as they are detected in mice and non-human primates almost exclusively when kept in captivity. In humans, *Akkermansia* candidate species display ecological co-exclusion, diversified functional capabilities, and distinct patterns of associations with host body mass. Analysis of CRISPR-Cas loci reveals new variants and spacers targeting newly discovered putative bacteriophages. Remarkably, we observe an increased relative abundance of *Akkermansia* when cognate predicted bacteriophages are present, suggesting ecological interactions. *A. muciniphila* further exhibits subspecies-level genetic stratification with associated functional differences such as a putative exo/lipopolysaccharide operon.

Conclusions

We uncover a large phylogenetic and functional diversity of the *Akkermansia* genus in humans. This variability should be considered in the ongoing experimental and metagenomic efforts to characterize the health-associated properties of *A. muciniphila* and related bacteria.

Introduction

The human body is home to several distinct microbiomes which represent functionally and phylogenetically diverse microbial ecosystems that are key for human health (Human Microbiome Project Consortium 2012; Qin et al. 2010; Pasolli et al. 2019). A frequent and abundant inhabitant of the gut microbiome is *Akkermansia muciniphila*, a Gram-negative, non-motile anaerobic bacterium specialized in the degradation of mucin (Derrien et al. 2004). *A. muciniphila* can utilize mucin as its sole carbon and nitrogen source (Derrien et al. 2004), thus growth in its natural habitat is not directly dependent on the influx of dietary compounds. *A. muciniphila* continues to attract attention due to its association with host health: the relative abundance of *A. muciniphila* is inversely correlated with obesity in humans (M. Yang et al. 2020; de Vos 2017) and it was shown to alleviate insulin resistance and obesity while increasing gut barrier function in a mouse model of diet-induced obesity (Everard et al. 2013). Its potential as a next-generation probiotic in the battle against metabolic disorders was confirmed in a first intervention trial targeting humans with metabolic syndrome and obesity (Depommier et al. 2019).

The human microbiome hosts a vast bacterial diversity at the level of distinct strains belonging to the same species (i.e. conspecific strains) (Truong et al. 2017; Scholz et al. 2016; Pasolli et al. 2019; Paul I. Costea et al. 2017; Schloissnig et al. 2013; Van Rossum et al. 2020). The genomic variation of conspecific strains often exceeds 3% nucleotide variation in the core genes, and when comparing pairs of conspecific strains it is frequently observed that 25% of genes are present in only one of the two, causing each human microbiome to be unique at the strain level (Pasolli et al. 2019). Importantly, this subspecies genomic variability translates into phenotypic variability, for example, in connection with host lifestyle (De Filippis et al. 2019; Fehlner-Peach et al. 2019; Maier et al. 2018) and at the immunological level (C. Yang et al. 2020; Geva-Zatorsky et al. 2017). However, experimental *Akkermansia* research still heavily relies on the type strain *A. muciniphila* Muc^T (ATCC BAA-835), and on a few more genomes of newly isolated strains that became available recently (Guo et al. 2017; Xing et al. 2019; Kirmiz et al. 2020). Furthermore, only a single other species in the *Akkermansia* genus – *A. glycaniphila* (Pyt^T, DSM 100705) – has so far been described and genomically characterized (Ouwerkerk et al. 2016, 2017). There is thus the urgent need to expand our understanding of the genomic variation and (sub)species diversity of *Akkermansia* for improving both the interpretation of its functions and its potential use in biomedicine.

Recently, a large number of publicly available metagenomes of human-associated microbial communities have been mined to produce hundreds of thousands of metagenome-assembled genomes (MAGs) (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019, 2020) and methods to profile and investigate strains directly in

metagenomes have become increasingly effective (Truong et al. 2017; Scholz et al. 2016; Paul Igor Costea et al. 2017; Luo et al. 2015; Quince et al. 2017). While these tools offer the opportunity to characterize population genomics of important but poorly characterized human-associated bacteria, only a few species have been investigated so far at high genomic resolution and global scale (Tett et al. 2019; Karcher et al. 2020; Hall et al. 2017; De Filippis, Pasolli, and Ercolini 2020).

Here, we present a comprehensive genomic characterization of *Akkermansia muciniphila* and closely related *Akkermansia* spp., using a total of 2,226 MAGs belonging to the *Akkermansia* genus, 188 publicly available isolate genomes and 6 newly sequenced isolate genomes. The *Akkermansia* MAGs were obtained by expanding our recent catalog of human associated MAGs (Pasolli et al. 2019) with 166,518 additional MAGs from 45 different datasets comprising samples also from mice and non-human primates for an integrated catalog of 321,241 MAGs (see **Methods**). Next to the species-level clade with the *A. muciniphila* type strain, we show the existence of four other *Akkermansia* candidate species that colonize the human gut. These five candidate species display strong co-exclusion within a given host, are phylogenetically stratified at the subspecies level and are at the same time widely distributed across hosts, age, and geography. Comparison of candidate species shows differential association with BMI in humans and vitamin B12 synthesis capabilities. We also analyzed the genomic organization of CRISPR-Cas loci (providing adaptive immunity against foreign DNA (Barrangou and Marraffini 2014)) in *Akkermansia* candidate species and found these to differ in their locus architecture and spacer numbers. We furthermore identified *de novo* assembled putative bacteriophages with spacer hits from *Akkermansia* candidate species, and found that viral detectability correlates strongly with the relative abundance of cognate *Akkermansia* candidate species, suggesting an intimate ecological interplay. These, and other genomic analyses in this work, provide a solid basis for future mechanistic explorations and biomedical applications of *Akkermansia*.

Results and discussion

A large-scale metagenomics-based analysis of *Akkermansia* candidate species

In order to study the diversity of *Akkermansia* species in the human microbiome, we collected all genomes available from isolate sequencing as well as MAGs from large collections of metagenomic datasets and unified them into a single genomic resource. Specifically, we gathered and quality controlled 119 isolate genomes from NCBI that were taxonomically annotated as *A. muciniphila,* as well as 69 labelled as *Akkermansia* sp. (Derrien et al. 2004; Ogata et al. 2020; M. Yang et al. 2020; H. Li et al. 2015; Garzetti et al. 2017; Guo et al. 2017; Medvecky et al. 2018; Poyet et al. 2019; Jiang et al. 2019; Liu et al. 2020; Ouwerkerk et al. 2016). We further obtained 2,226 MAGs

taxonomically annotated to the genus *Akkermansia* from a total of 18,600 shotgun metagenomes (see **Methods**) sampled from multiple hosts including humans, non-human primates, and mice. Only high-quality MAGs - defined as those with at least 90% estimated genomic completeness and at most 5% estimated genomic contamination (Parks et al. 2015) - were included in the analysis. We further enhanced our genome set with 6 isolate genomes (Ouwerkerk 2016). The integrated *Akkermansia* genome resource we consider for downstream analysis thus consists of a total of 2,420 genomes (**Additional file 1**: **Table S1**).

Multiple under-characterized *Akkermansia* candidate species are present in the human microbiome

We reconstructed the phylogeny of all genomes in our set using the 400 universal marker genes adopted in PhyloPhlAn 3 (Segata et al. 2013; Zhu et al. 2019; Asnicar et al. 2020) (**Fig. 1A, Methods**) including *Verrucomicrobium spinosum* as an outgroup (Derrien et al. 2004). This revealed the presence of several well-defined monophyletic clades (**Fig. 1A**). In addition to the previously described *A. glycaniphila* species (Ouwerkerk et al. 2016), these clades – following the validated species-level genome bins (SGBs) approach based on whole-genome genetic distances (Pasolli et al. 2019) (see **Methods**) – delineate candidate species (**Fig. 1B**). The candidate species are genetically distinct, with inter-candidate species genome-wide average estimated nucleotide identities generally below 90% (except between a single pair of candidate species, **Fig. 1B**). We confirmed those results using genome similarity estimates obtained using PhyloPhlAn 3 markers (**Additional file 2**: **Fig. S1**). One of the five delineated candidate species (henceforth "*A. muciniphila*") includes the type strain of *A. muciniphila* (Muc^T) (Derrien et al. 2004) as well as 108 isolate genomes. The remaining four candidate species (SGB9223, SGB9224, SGB9227 and SGB9228) comprise not only MAGs but also isolate genomes that were, however, taxonomically described as *A. muciniphila* or *Akkermansia* spp. in NCBI. Cultivated members of the candidate species were retrieved not only from humans, but also from mice, non-human primates and – very rarely – other mammals, such as elephants, horses and reindeers (Xing et al. 2019). *A. glycaniphila* was originally isolated from a python (Ouwerkerk et al. 2016), and we did not uncover new diversity for this species in the available datasets, suggesting that *A. glycaniphila* is not found in mammals. A reason for the taxonomic assignment of cultivated strains to the *A. muciniphila* species and the generally underestimated diversity of the genus is probably the surprisingly high similarity displayed by 16S rRNA gene sequences of these strains, with 16S rRNA gene sequences of strains in different candidate species never diverging by more than 2% (**Fig. 1B, Methods**). Taken together, these data show that a total of five *Akkermansia* candidate species exist in the human, mouse, and non-human primate gut microbiomes, four of them remaining under-investigated and uncharacterized.

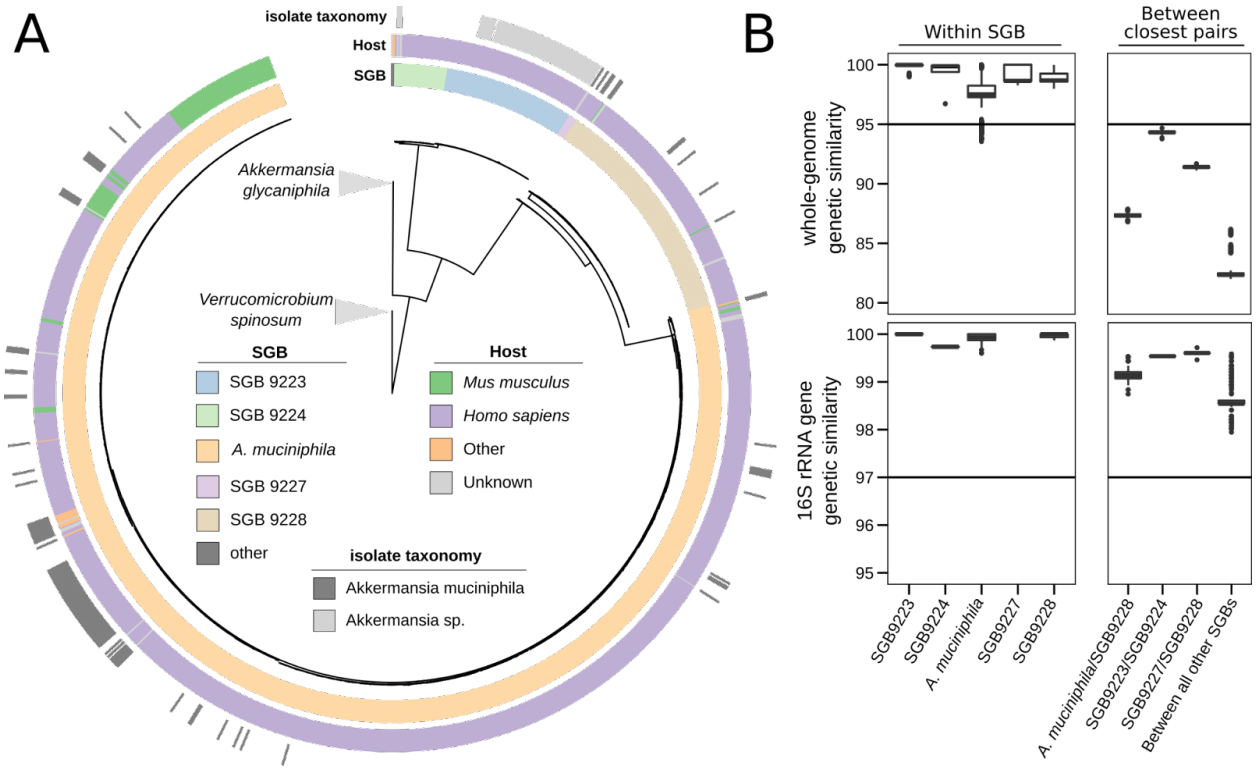


Figure 1. **The *Akkermansia* genus comprises four additional candidate species phylogenetically rooted between the already characterized *A. glycaniphila* and *A. muciniphila*** (**A**) Whole genome phylogeny of the 2,420 metagenome-assembled genomes (MAGs) reconstructed here and the genomes from isolate sequencing available in NCBI taxonomically annotated as *A. muciniphila* or *Akkermansia spp*. The phylogenetic tree is rooted using *Verrucomicrobium spinosum* as an outgroup and was built using PhyloPhlAn 3 (Asnicar et al. 2020) with 400 universal markers (see **Methods**). SGB: species-level genome bin (see **Methods**). (**B**) Within- and between-clade whole-genome average estimated nucleotide identity (fastANI (Jain et al. 2018), top panels) and full-length 16S sequence distances (bottom panels) among *Akkermansia* SGBs provide evidence that these are candidate species.

Akkermansia candidate species are enriched in humans and co-exclude within a host

We next set out to assess host specificity, co-abundance patterns and metadata associations of the *Akkermansia* candidate species. To this end, we first developed a marker-based method with increased sensitivity compared to metagenomic assembly to detect the presence and relative abundance of *Akkermansia* candidate species in metagenomes. In brief, this was done by (1) identifying genes that were core to each of the *Akkermansia* candidate species and at the same time never detected in other *Akkermansia* or non-*Akkermansia* species-level groups (*marker genes*), and (2) using these marker genes as targets for read-mapping inside MetaPhlAn 3.0 (Segata et al.

2012; Beghini et al. 2020) to estimate their coverage and relative abundance (see **Methods**). By profiling the 13,237 metagenomic samples from 98 publicly available datasets with sufficient metadata (**Additional file 1**: **Table S2**), we found that *Akkermansia* candidate species differed strongly in their prevalence across hosts. *A. muciniphila* is by far the most prevalent candidate species across all hosts, being detected in 34% of adult humans and reaching a maximum prevalence of 54% in laboratory-held mice (**Fig. 2A**). The other candidate species were detected at lower prevalence (< 25%) across all hosts (**Fig. 2A**). Interestingly, *Akkermansia* candidate species were found to be much more often present in captive animals than in free-living ones: while laboratory mice and mice humanized via microbiome transplantation are fairly likely to host *A. muciniphila,* SGB9224, or SGB9228 (up to 54% prevalence), in wild mice solely SGB9228 was detected in only 5 mice from a single study (out of 60 mice from 3 different studies being assessed; 8% prevalence). Similarly, merely two out of 162 samples from wild non-human primates tested positive for any *Akkermansia* candidate species (1.2% prevalence). Despite potential biases due to uneven sampling and effects of diet, these data suggest a marked specificity of *Akkermansia* candidate species for the human gut (with the exception of *A. glycaniphila*), and while strains from these candidate species can colonize mice and non-human primates, such colonization appears to be predominantly a consequence of man-made environments, suggesting colonization from care-taking humans as a plausible mechanism.

While *Akkermansia* candidate species are found in almost half of all human samples, the presence of one is strongly anti-correlated with the others (**Fig. 2B, C**): it is extremely rare to detect more than one candidate species present in the same host, with only 46 instances of two co-occurring candidate species in human metagenomes out of 4,171 cases in which at least one was present (corresponding to a co-occurence rate of ~1%), and no instance of more than two co-occurring candidate species. These five closely-related candidate species thus show a mutual exclusion pattern suggestive of complex and possibly host-mediated ecological interactions that remain to be explored.

A. muciniphila but not the other candidate species is associated with lower BMI

The presence and abundance of *A. muciniphila* in the gut microbiome has been negatively associated with body mass index in previous studies based on 16S rRNA gene sequencing (Dao et al. 2016, 2019), and the link was shown to be causal in both mice and humans by supplementation with cells of *A. muciniphila* Muc^T (Everard et al. 2013; Depommier et al. 2019). Because of the limitations of 16S rRNA gene amplicon sequencing in distinguishing *Akkermansia* candidate species (**Fig. 1B**), we performed an analysis on the association between relative abundances of individual *Akkermansia* candidate species and BMI. We considered 3,311 samples in 22 different metagenomic datasets from five continents and adjusted for age and sex in a random effect model

meta-analysis (**Additional file 1**: **Table S3**). Interestingly, only the relative abundances of *A. muciniphila* were found to be significantly negatively associated with BMI, while associations of other candidate species were not statistically significant (**Fig. 2D**), suggesting that *A. muciniphila* should be regarded as the primary candidate for microbiota-based therapeutic interventions aimed at improving host metabolic health as a recent proof-of-concept trial also reported (Depommier et al. 2019).

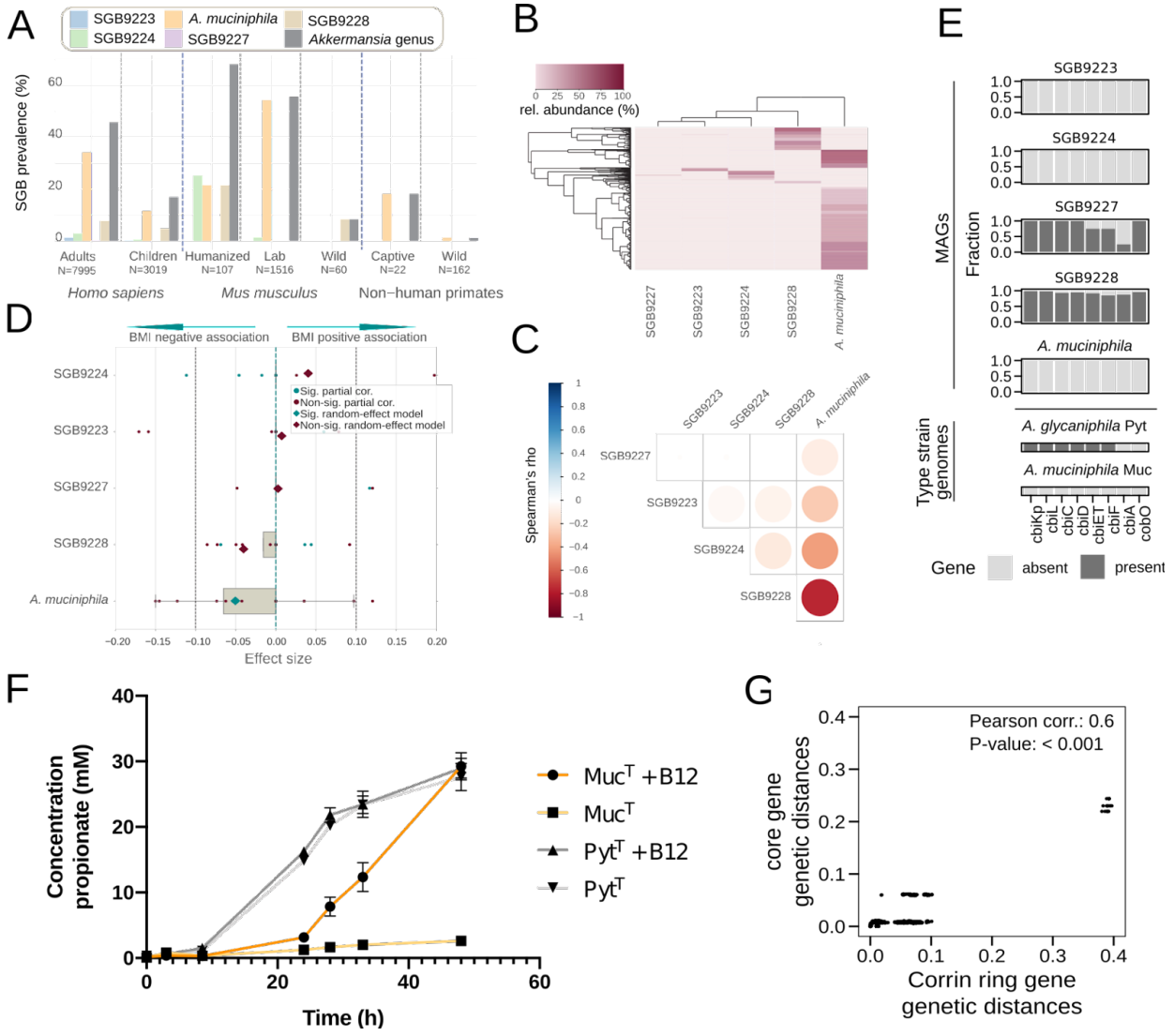


Figure 2. Prevalence and insights into the ecological and functional characteristics of *Akkermansia* candidate species. (A) *Akkermansia* candidate species have variable prevalence across hosts and wild versus captive mice and non-human primates. We computed prevalences using species-specific marker genes (see **Methods**) applied on a total of 13,237 metagenomic samples. **(B, C)** *Akkermansia* candidate species are strongly mutually exclusive (analysis based on 4,171 *Akkermansia*-positive human metagenomes). (**D**) *A. muciniphila* but not the other *Akkermansia* candidate species is associated with decreased host body mass index (BMI) according to a meta-analysis random effect model of partial

correlations adjusted for age and sex (see **Methods**) comprising 3,311 human metagenomic samples from 22 datasets (**Additional file 1**: **Table S2**). (**E**) Corrin ring biosynthesis operon genes are consistently present only in candidate species SGB9227 and SGB9228 (see **Methods**). (**F**) Growth analysis of the *A. muciniphila* and *A. glycaniphila* type strains shows propionate production by Pyt^T but not Muc^T in the absence of vitamin B12. This is indicative of endogenous production of vitamin B12 (acting as a cofactor for the methyl-malonyl CoA synthase reaction) by Pyt^T but not Muc^T. (**G**) Core gene genetic distances are correlated with corrin ring biosynthesis gene genetic distances. Pairwise distances were computed only for strains in which all genes were found together on the same contig.

We next tested whether available host characteristics other than BMI were associated with *Akkermansia* candidate species relative abundances and also examined whether genetic stratification by host parameters could be detected within candidate species. At the candidate species-level, no association with age was detected, while sex (as self-declared by the individuals) was associated with the relative abundance of *A. muciniphila* (after adjusting for age and BMI), with women harboring comparatively higher relative abundances (P-value = 4.8e-05), as observed elsewhere (X. Zhang et al. 2021). To test for associations of candidate species with host metadata at the level of their internal phylogenomic structure, we subsequently computed PERMANOVA statistics for all combinations of single candidate species and host, age, geography and westernization status. While some significant associations were identified (especially for *A. muciniphila* and SGB9228), the largest effect size among the significant (P-value < 0.05) tests was a PERMANOVA R^2 of 0.10 for SGB9228 with continent (**Additional file 2**: **Fig. S2**, **Additional file 2**: **Fig. S3**, **Additional file 2**: **Fig. S4**), suggesting that no strong associations of strain-level structure with host metadata is detectable.

With *Akkermansia* supplementation becoming available (Depommier et al. 2019; Zhou 2017), it is relevant to verify that such interventions are not potentially causing microbial anti-drug resistances to spread in the human gut. To this end, we first screened all *Akkermansia* genomes for antibiotic resistance genes using the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al. 2020). Overall, we found only 4 genes known to be involved in antibiotic resistance present in more than 1% of all genomes. Among those, only *adeF* (encoding a membrane protein of a drug efflux complex) is consistently found in most genomes (prevalence of 81% over all genomes), but still never present in SGB9223, SGB9224, nor SGB9227 (**Additional file 2**: **Fig. S5**). In addition to these well-cataloged resistance genes, a recent study reported the presence of 8 genes (including 3 antibiotic resistance genes) in *A. muciniphila* strain GP36 derived from the broad-range plasmid RSF1010 that is found in many gram-negative bacteria including *E. coli* (Guo et al. 2017). We queried all genomes for the presence of this plasmid-derived sequence and found 55 genomes (2.2% overall prevalence) in which we could detect at least 50% of the sequence of RSF1010 at 70%

average sequence identity or higher. A total of 49 of the 55 instances were found in *A. muciniphila* (2.5% prevalence in *A. muciniphila*). In all 55 positive cases, these genes were found on contigs larger than the plasmid (~8 kb), suggesting that they may be integrated into the bacterial genome (as also reported in (Guo et al. 2017)). Of note, the *A. muciniphila* type strain Muc^T carries no antibiotic resistance genes and its use does not raise any antibiotic resistance concern as also indirectly confirmed by dose scaling pilot studies in humans and toxicological studies in rabbit and other model organisms (Depommier et al. 2019; Druart et al. 2021); however, ongoing and future human trials with strains different from the type strain should carefully consider their antibiotic resistance potential. In conclusion, although the rare occurrence of antibiotic resistance genes from plasmid RSF1010 in some *A. muciniphila* genomes has evident safety implications for their use in therapeutic applications, our findings indicate that *Akkermansia* candidate species mostly lack genetic means to defend themselves against currently used antibiotics.

Vitamin B12 synthesis capabilities were independently lost by two *Akkermansia* candidate species

Due to its essential nature and limited availability in the human gut, vitamin B12 (cobalamin) is regarded as a key element in host-microbe interactions (Degnan, Taga, and Goodman 2014). In a recent study, 75 *Akkermansia* strains were reported to differ in their potential to produce this important cofactor (Kirmiz et al. 2020). We set out to characterize the vitamin B12 synthesis capabilities of the *Akkermansia* candidate species as well as *A. glycaniphila*. By identifying corrin ring biosynthesis genes as a proxy for vitamin B12 synthesis capability (Degnan, Taga, and Goodman 2014), we confirm that the large majority of MAGs from candidate species SGB9227 and SGB9228 encode most proteins involved in producing vitamin B12 (75% of SGB9227 MAGs encode all proteins except CbiA; 92% of SGB9928 MAGs encode all proteins except CbiF), while those genes were never found in *A. muciniphila,* SGB9223, nor SGB9224 (**Fig. 2E**). Interestingly, the more phylogenetically distant *A. glycaniphila* Pyt^T (Ouwerkerk et al. 2016) also contains 6 out of 8 corrin ring biosynthesis genes (**Fig. 2E**). The differential vitamin B12 synthesis capabilities of *Akkermansia* spp*.* were successfully validated by *in vitro* assays: propionate production (a proxy for vitamin B12 production, as the pathway includes the B12-dependent methyl-malonyl CoA synthase reaction (Ottman et al. 2017)) was detected when growing *A. glycaniphila* Pyt^T but not *A. muciniphila* Muc^T in the absence of vitamin B12 (**Fig. 2F**, **Additional file 2**: **Fig. S6**). The *cbiA* gene that we did not detect in the majority of SGB9227 MAGs is not found in *A. glycaniphila* Pyt^T either, suggesting that this gene may not be necessary for B12 production in *Akkermansia* spp. Furthermore, we detected a strong correlation between pairwise genetic distances of corrin ring biosynthesis genes and core genes between strains of SGB9227, SGB9228, and the singular *A. glycaniphila* genome (Spearman rho

= 0.6, P-value < 0.001; **Fig. 2G**), suggesting that the B12 biosynthesis genes are ancestral to all *Akkermansia* candidate species and were lost by *A. muciniphila,* SGB9223, and SGB9224 candidate species in the human gut. The most likely evolutionary scenario would consist of two independent loss events: one after the most recent common ancestor of SGB9223/SGB9224 separated from the ancestor of the remaining candidate species, and another after the ancestor *A. muciniphila* separated from the ancestor of SGB9228 (**Additional file 2**: **Fig. S7**). Taken together, these results reveal two independent B12 synthesis loss events in *Akkermansia* candidate species and indicate that new *Akkermansia* strains should be studied for their potential to increase colonic vitamin B12 biosynthesis.

Akkermansia candidate species encode a novel variant of type I-C CRISPR-Cas loci

CRISPR-Cas systems are widely used by prokaryotes to fend off foreign DNA (Mojica and Rodriguez-Valera 2016), and can be exploited to alter the microbiome makeup (Hamilton et al. 2019). However, they have only been studied in detail for a limited number of bacteria, and strain-level variations have been documented (Hamilton et al. 2019). We thus screened our catalog of *Akkermansia* genomes and MAGs for the presence of CRISPR-Cas loci. A great majority of genomes (68%, **Fig. 3A**) harboured at least one CRISPR-Cas locus, and while type I-C loci (Makarova et al. 2020) were detected in all *Akkermansia* candidate species, *A. muciniphila* is the only species in the genus sometimes carrying a type II-C locus (33%, **Fig. 3A**). In 9% of the cases, *A. muciniphila* strains carried both the type II-C locus and the type I-C locus (**Fig. 3A**).

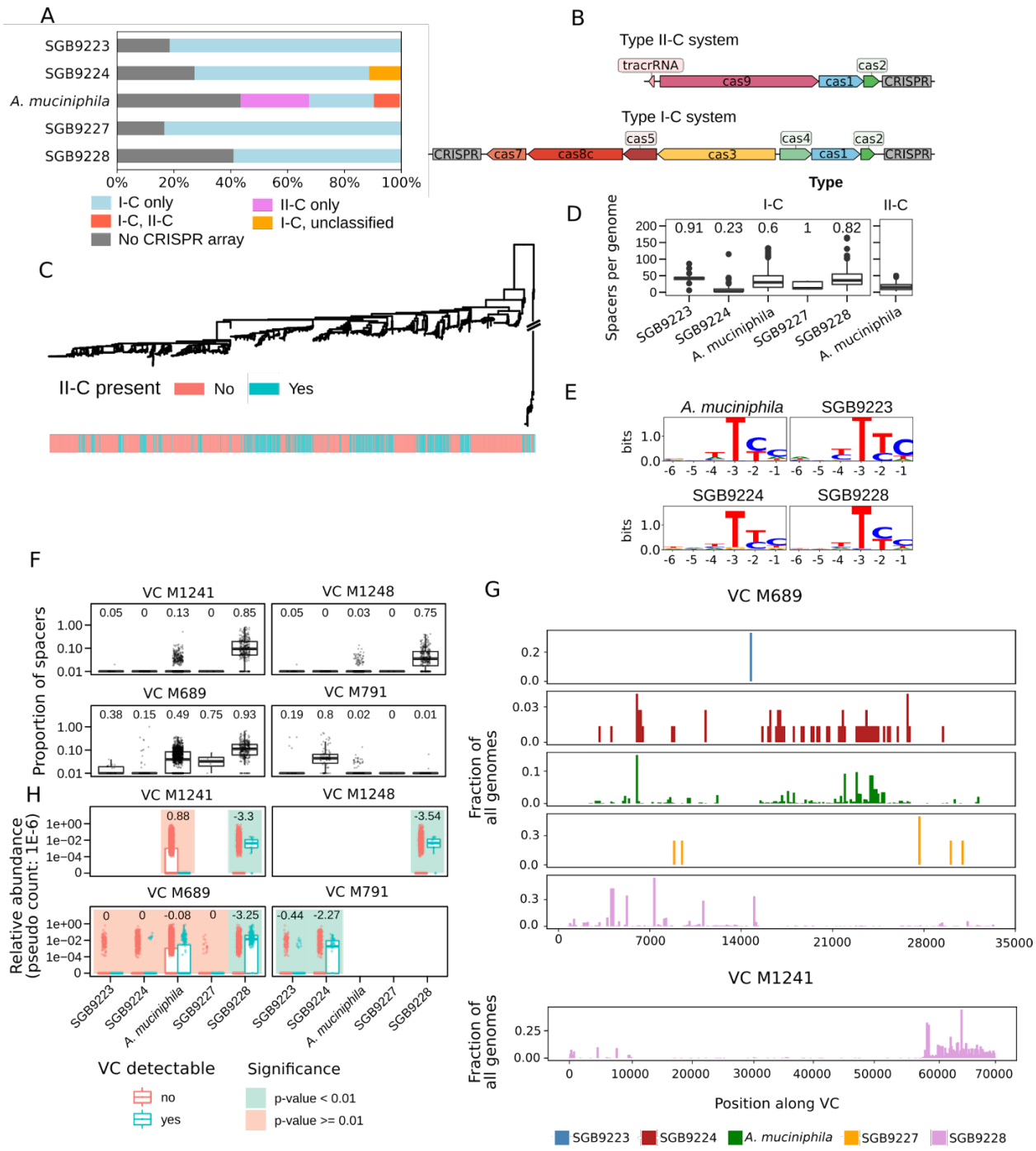


Figure 3. The CRISPR-Cas system of *Akkermansia* candidate species and their viral targets. (A) CRISPR locus type composition of *Akkermansia* candidate species. All candidate species possess CRISPR locus type I-C, with the exception of *A. muciniphila* in which type II-C is present in more than 30% of the genomes. **(B)** Representative locus organization of CRISPR loci over *Akkermansia* candidate species. Some type I-C loci contain only one CRISPR array. Gene and CRISPR array lengths are scaled to correspond to the median

length over all loci. **(C)** Phylogenetic tree of *A. muciniphila* subspecies colored by type II-C presence. **(D)** The total number of spacer sequences for the genomes in each *Akkermansia* candidate species. Type II-C loci were only found in *A. muciniphila*. Numbers above the boxplots correspond to the fraction of type I-C loci with two CRISPR arrays. **(E)** Logo plots of predicted PAM sequences in putative (phage) Viral Clusters (VCs, see **Methods**) upstream of sequences with perfect matches against CRISPR spacer sequences from type I-C loci. **(F)** Proportion of CRISPR spacers within candidate species genomes with a near-perfect match (at most 2 mismatched nucleotides) for four VCs. The number above the box plots corresponds to the fraction of genomes with at least one spacer hit against a given VC (see **Methods**). (**G**) Mapping of spacers from *Akkermansia* genomes against two representative VCs, visualized with a sliding window of 150 nt. See **Additional file 2**: **Fig. S8** for the remaining VCs. (**H**) Distribution of the relative abundances of the *Akkermansia* candidate species based on the presence or absence of each cognate VC in the metagenome (**Additional file 1**: **Table S2**, see **Methods**). P-values for differential abundance were determined via two-sided Wilcoxon rank-sum tests. P-values of < 0.01 were considered significant. The numbers above the box plots correspond to the generalized fold-change, with negative numbers indicating a higher bacterial abundance when a VC is detected (Wirbel et al. 2019).

The structure of type I-C loci in *Akkermansia* candidate species differs notably from the canonical organization (Makarova et al. 2020): Cas3, Cas5, Cas8c and Cas7 genes are encoded in the opposite direction of Cas4, Cas1 and Cas2, thus representing a novel variant of type I-C CRISPR-Cas loci. The majority of loci (62.4%) contain two CRISPR arrays, one upstream and one downstream of the Cas gene cassette. In contrast, the type II-C loci of *A. muciniphila* have the canonical structure (Russel et al. 2020) in 95% of the strains in which the locus was detected (**Fig. 3B**). The presence of a type II-C locus in *A. muciniphila* has no clear phylogenetic structure (**Fig. 3C**), highlighting a peculiar evolutionary history. *Akkermansia* candidate species also differ in the total number of spacer sequences encoded in each genome as well as the fraction of loci that contain two (instead of one) CRISPR arrays (**Fig. 3B, D**). SGB9223 and SGB9228 on average contain more spacer sequences (median 43 s.d. 15.4 and median 36 s.d. 29.5) compared to SGB9224, which has the lowest number of spacers (median 3, s.d. 16.4). Similarly, 91% of all genomes from SGB9223 contain two CRISPR arrays (one upstream, one downstream of the Cas gene cassette), whereas only 23% of genomes in SGB9224 do so (**Fig. 3D**). *Akkermansia* candidate species thus generally contain CRISPR-Cas systems, with distinct loci structure and spacer composition, indicating considerable divergence in their exposure to foreign DNA over their evolutionary trajectory.

Newly discovered putative phages are recognized by *Akkermansia* CRISPR-Cas systems and tend to co-occur with cognate candidate species

We next identified *de novo* assembled, putative intestinal bacteriophages in shotgun gut viromes defining Viral Clusters (VCs, see **Methods**) and screened them for the presence of *Akkermansia* CRISPR-Cas spacers. We found no spacer hits against any of the known intestinal phages currently in RefSeq (Brister et al. 2015), but we instead detected a total of eight VCs with spacer hits (**Additional file 2**: **Fig. S8**), four of which consistently attracted spacer sequences from at least one *Akkermansia* candidate species (**Fig. 3F,** See **Methods**) which we considered for further analysis. While some VCs exhibited hits from spacers from only one of the candidate species (SGB9228 for M1241 or M1248), other VCs (M689) were found to attract spacer sequences from all candidate species. Regardless of VCs, SGB9228 genomes on average have the highest total fraction of spacer sequences hit (**Additional file 2**: **Fig. S9**).

The mapping of *Akkermansia* spacer sequences against VCs revealed that spacer sequences tend to cluster locally in the phage genome, and that different locations on the viral genome attract spacers in a species-dependent fashion (**Fig. 3G, Additional file 2**: **Fig. S8**). Furthermore, identification of the sequences directly upstream of all spacer sequence hits allowed reconstruction of the canonical type I-C protospacer adjacent motif (PAM) “TTC” (**Fig. 3E**) found in *Bacillus halodurans* (Leenay et al. 2016). The presence of multiple, distinct hits for some species-VC combination (**Fig. 3G, Additional file 2**: **Fig. S8**) suggests that these matches are not spurious and that many combinations of *Akkermansia* candidate species and viral clusters reflect multiple bacterium-phage interactions in the intestinal environment. To further investigate potential ecological interactions between *Akkermansia* candidate species and phages, we assessed the co-occurrence between candidate species and the matching VCs across 13,237 metagenomes (see **Methods**). For 5 out of 10 putatively interacting pairs of VCs and candidate species (defined as those pairs where more than 10% of genomes of the candidate species have at least one VC-matching spacer), we found that a candidate species is significantly more abundant (P-value < 0.01) whenever the cognate VC is detectable (**Fig. 3F, Fig. 3H, see Methods**). Taken together, our analysis showed that CRISPR spacer sequences found in the genome of *Akkermansia* candidate species can be frequently mapped to four putative phages that co-occur with their cognate candidate species, suggesting that they are ecologically interacting in the human gut.

A. muciniphila is stratified in four subspecies with different host preferences and functional profiles

In all bacterial species a large fraction of the phenotypic variability is encoded at the subspecies level (De Filippis et al. 2019; Maier et al. 2018; M. Yang et al. 2020; Fehlner-Peach et al. 2019). We thus further focused on the intra-species genetic variation of *A. muciniphila* given its prevalence and relevance due to its association with lower host BMI (**Fig. 2D**). We found *A. muciniphila* to have four monophyletic subclades

that we labeled Amuc1 to Amuc4 (**Fig. 4A**). We left strains unassigned that are not part of one of these monophyletic subclades (accounting for 29% of all *A. muciniphila* strains). The subspecies were found to have similar within-subspecies genetic similarities - always exceeding 98% identity - while between-subspecies genetic similarities range from 95.4% genetic similarity between Amuc1 and Amuc4 to 98.6% between the more closely related Amuc2 and Amuc3 (**Fig. 4B**). This inter-subspecies genetic divergence was coupled also with a diversification of the functional profiles of the strains (**Fig. 5A**).

Amuc1 is the most prevalent subspecies in humans (47%), followed by Amuc2 and Amuc3 (27% and 24% respectively, **Fig. 4C**). To investigate whether these global prevalences were driven by particular host factors, we studied the distribution of *A. muciniphila* subspecies across host metadata (**Additional file 2**: **Fig. S10**). In addition to a significantly higher prevalence of Amuc4 in non-westernized human populations compared to non-Amuc4 (Fisher-test P-value < 0.001), we found that subspecies were differentially distributed across hosts. In particular, Amuc2 and Amuc3 are specific to humans and never found in mice and non-human primates, whereas Amuc1 and Amuc4 can be found in both humans and mice (**Fig. 4C**) but in different proportions, suggesting differential fitness of *A. muciniphila* subspecies in mice compared to humans. Notably, all *A. muciniphila* genomes we obtained from mice came from laboratory-held mice. Due to the lack of subspecies-specific marker genes we were unable to extend prevalence analysis to samples lacking successfully reconstructed *Akkermansia* MAGs, but our data nonetheless suggests *Akkermansia* in mice may be acquired from humans and that there is a strong preference of laboratory mice to acquire only the Amuc1 (to which Muc^T belongs) and Amuc4 *A. muciniphila* subspecies, which might have important implications for pre-clinical mice models.

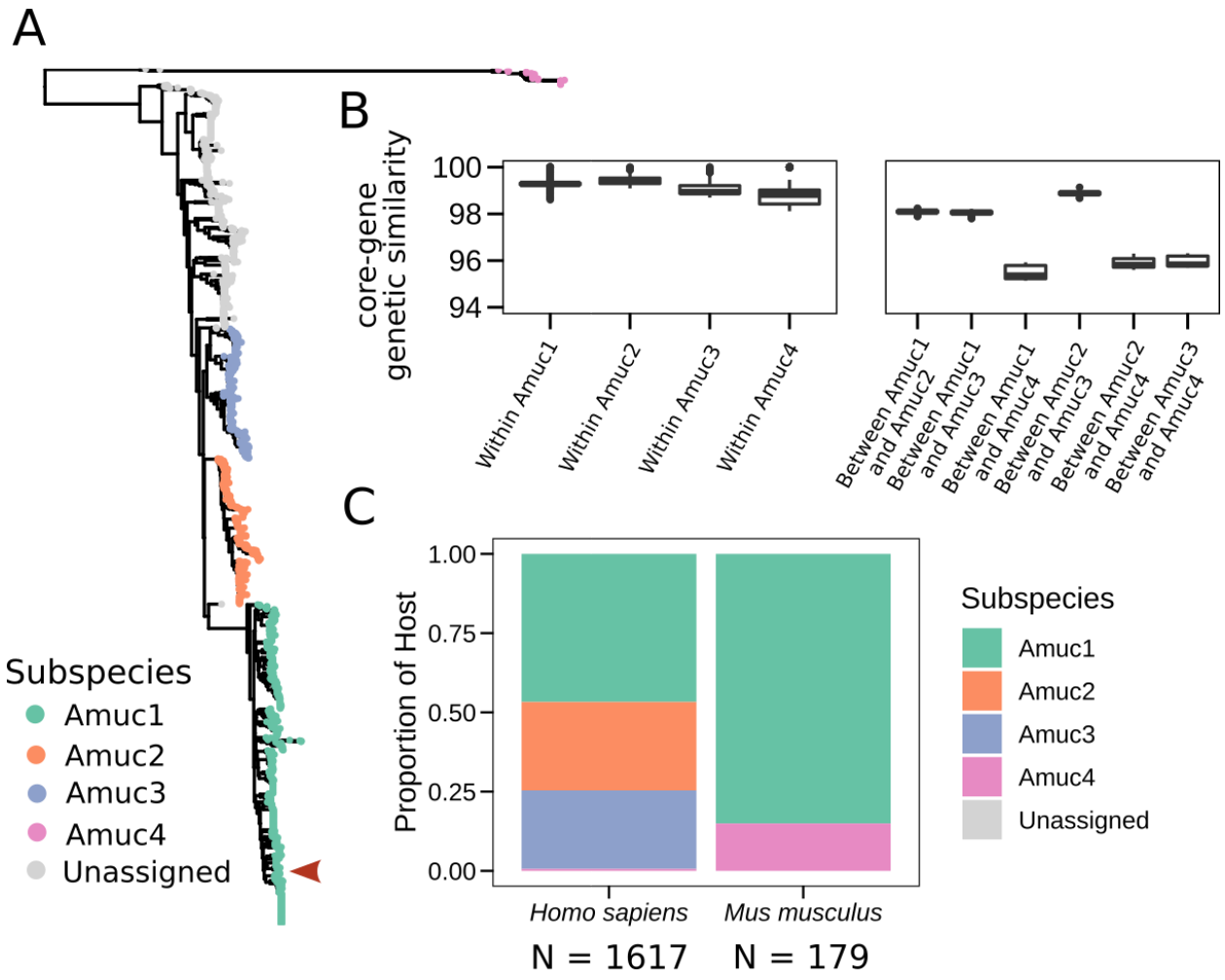


Figure 4. *A. muciniphila* is stratified in multiple subspecies with distinct host preferences. (**A**) Phylogenetic tree of *A. muciniphila* based on a core-gene alignment built using 169 clade specific core genes (see **Methods**). The red arrow indicates the Muc^{T} type strain. (**B**) Within- and between-subspecies core-gene nucleotide identities confirm the subspecies diversification defined on the phylogeny. (**C**) Per-host frequency of *A. muciniphila* subspecies assembled from metagenomes. All 174 mouse *A. muciniphila* genomes were reconstructed from stool metagenomes of laboratory-mice.

Two functionally related but distinct glycosyltransferase-rich operons are found in the *A. muciniphila* subspecies

Surface glycoconjugates are known to form a species- and sometimes even strain-specific glycan barcode, conferring bacteria with unique interaction properties (Tytgat et al. 2016). Two well-conserved archetypes of a glycosyltransferase-rich operon were detected in the same genomic location in different *A. muciniphila* strains (**Fig. 5B,**

Additional file 2: **Fig. S11**). Both operon archetypes predominantly contain genes annotated as glycosyltransferases (GTs) belonging to two different CAZyme families (GT2 and GT4), albeit in different proportions: while archetype 1 contains five GT2 and four GT4 copies, archetype 2 contains three GT2 and six GT4 copies. GTs belonging to these families are typically involved in lipo- and/or exopolysaccharide biosynthesis, which are key in microbiota-host interactions (Tytgat et al. 2016). However, despite both operon archetypes being mostly composed of functionally related GTs, only a few pairs of proteins displayed detectable but very remote sequence similarity (**Fig. 5B**). The two operon archetypes were notably differentially distributed among *A. muciniphila* subspecies: subspecies Amuc2 and Amuc4 always possessed archetype 2 (whenever detectable), whereas strains belonging to Amuc1 and Amuc3 had either archetype. *A. muciniphila* thus encodes one of two possibly very distantly related operons that are putatively involved in lipo/exopolysaccharide (LPS/EPS) biosynthesis functions, hinting at a possible divergence of their surface glycoconjugates as well as host-specific selective advantages.

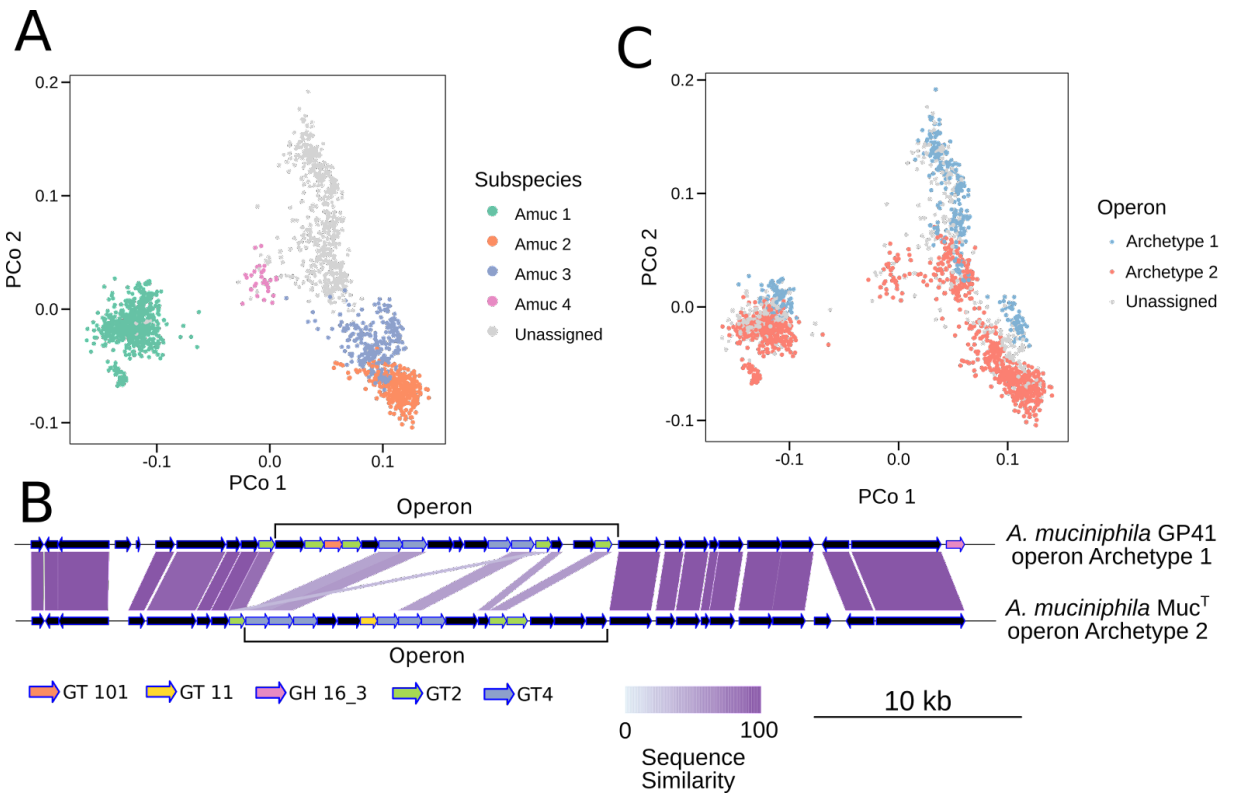


Figure 5. Functional diversification of *A. muciniphila* subspecies and cognate Exopolysaccharide/LipidA synthesis operon. (A) Ordination analysis (Jaccard-distance based PCoA using gene presence and absence information) reveals a diversification of gene repertoires of *A. muciniphila* subspecies. Genes found in less than 3% of strains were excluded. Subspecies designation is derived from the *A. muciniphila* phylogenetic tree in **Fig.**

4. **(B)** Operon archetypes putatively involved in Exopolysaccharide/LipidA synthesis in *A. muciniphila* GP41 (operon archetype 1) and *A. muciniphila* Muc^T (operon archetype 2). (**C**) PCoA (same as in (A)) colored by operon archetype membership. Genomes in which neither operon could be found were labeled 'Unassigned'.

Discussion

The possibility of extracting whole (draft) microbial genomes of sufficient quality directly from metagenomic sequences (Pasolli et al. 2019; Almeida et al. 2019, 2020; Nayfach et al. 2019; Manara et al. 2019; Truong et al. 2017) together with the quickly increasing availability of metagenomes from diverse populations (Pasolli et al. 2017) is revolutionizing the way human-associated microbes can be studied and characterized (Tett et al. 2019; Karcher et al. 2020; Hall et al. 2017; De Filippis, Pasolli, and Ercolini 2020). Exploiting a combined set of over 18,600 metagenomic samples from multiple hosts, we studied the population genomics and genetic characteristics of bacterial strains belonging to the *Akkermansia* genus. While *A. muciniphila* is recognized as a keystone species of the human microbiome, current biomedical and translational research is still driven by the type strain Muc^T (Derrien et al. 2004), thus neglecting the genomic and phenotypic variability of conspecific strains as well as of closely related species. Previous comparative genomic efforts were able to survey only a fraction of the diversity in the *Akkermansia* genus we describe here due to limited availability of isolate genomes (Guo et al. 2017; Poyet et al. 2019; Xing et al. 2019). At the same time, we extended similar ongoing work using MAGs for this genus (Lv et al. 2020) with a larger genome set and more diverse metagenomic sample set including non-human hosts, allowing us to explore aspects such as the association of *Akkermansia* abundances with phenotypes (particularly with respect to BMI), the in-depth analysis of some of its genetic features such as the machinery for vitamin B12 synthesis and a novel LPS/EPS operon, and the discovery of bacteriophages likely interacting with *Akkermansia* in the human gut.

Our analysis of 2,420 *Akkermansia* genomes delineates four candidate species in addition to the well-defined *A. muciniphila* species. The five candidate species are prevalent in the human gut microbiome and are found in other mammals such as mice and non-human primates almost exclusively when living in man-made environments, suggesting that all *Akkermansia* candidate species are specifically adapted to the human gut. All candidate species have very high pairwise sequence similarity of the full-length 16S rRNA gene (> 98%) and substantially lower whole-genome similarity (< 90% for all pairs except SGB9223 and SGB9224). These unusual genomic characteristics are likely the reason why the diversity of the *Akkermansia* genus has been overlooked by extensive 16S rRNA gene amplicon sequencing surveys in the past. Most bacterial species at < 95% genomic similarity have > 3% divergence of the

16S rRNA gene (Pasolli et al. 2019), and the substantially different pattern observed in *Akkermansia* might suggest rapid genomic diversification of these clades in humans.

A potential instance of adaptive evolution in *Akkermansia* emerging from our analysis could be the loss of vitamin B12 synthesis capabilities that likely occurred independently in the ancestors of two candidate species. Vitamin B12 promotes symbiotic metabolic relationships between gut microbes (Degnan, Taga, and Goodman 2014). Notably, a bidirectional syntrophy has been described between *A. muciniphila* Muc^T and *Anaerobutyricum soehngenii* (formerly known as *Eubacterium hallii (Shetty et al. 2018)*), with Muc^T converting mucin into oligosaccharides and acetate which are used by the butyrate-producer *Anaerobutyricum soehngenii*, in turn providing (pseudo)vitamin B12 to enable propionate production by Muc^T (Belzer et al. 2017). Hence, loss of vitamin B12 synthesis genes might have been metabolically favourable for *Akkermansia* candidate species SGB9223/9224 and *A. muciniphila* given the potential to syntrophically interact in this way with other species. Our results warrant future investigations also at the level of subspecies clades: for example, the presence of one of the two putative LPS/EPS operons we described in *A. muciniphila* may be driven by host-microbe interactions and host-specific factors such as diet or lifestyle.

Experimental efforts to investigate *Akkermansia*-host interactions that are currently fueled by findings of their potential role in biomedical settings (ranging from obesity (M. Yang et al. 2020; Everard et al. 2013; Depommier et al. 2019; de Vos 2017) to cancer treatments (Routy et al. 2018; Matson et al. 2018)) should consider some aspects of the genus-wide genomic diversity we are reporting here. For example, only *A. muciniphila* was confirmed in our analysis to be associated with decreased BMI and it is possible that the *A. muciniphila* subspecies might also display different strengths of association. Moreover, the limitations of animal-based experimental approaches should be particularly considered for *Akkermansia*: our finding that no *Akkermansia* candidate species is consistently detected in wild mice and primates may indeed suggest that these animals are not natural hosts for *Akkermansia* and raises the question whether host-*Akkermansia* interactions can be meaningfully recapitulated in mice. Similarly, we obtained MAGs from only two out of four *A. muciniphila* subspecies from mice, suggesting that not all subspecies may be well adapted to the mouse gut, which has important implications for *in vivo* experiments. Further delineation of subspecies through bacterial isolation or single-cell sequencing will be required to answer this question conclusively.

The ecology of *Akkermansia* may however be driven not only by aspects of host fitness, as interaction with bacteriophages also potentially contribute to shaping the population structure and diversity of this microbe. While no known phages have been so far linked with *Akkermansia* as a host, we identified at least four putative phages from gut viromes and gut metagenomes that display genomic regions recognized by spacer sequences in

CRISPR-Cas loci in a species-specific manner. These putative phages also tend to co-occur with their cognate candidate species. Understanding the mechanisms of interaction between these phages and their targets could be an important experimental step in order to develop intervention strategies to modulate the presence and abundance of *Akkermansia* candidate species in the gut.

Our work provides a large-scale strain-level analysis of *Akkermansia* that can be the basis for future further investigations of this genus. We also further highlighted the potential of metagenomics-based investigations of bacteria of the human microbiome, which is particularly important given that most bacterial species have very little genomic information available from cultivation efforts. In our work we also introduced new analysis types for MAG-based investigations complementing those already performed on other bacteria such as *Eubacterium rectale* (Karcher et al. 2020), *Prevotella copri* (Tett et al. 2019), *Ruminococcus gnavus* (Hall et al. 2017), and *Faecalibacterium prausnitzii* (De Filippis, Pasolli, and Ercolini 2020). Further extending and applying this approach to the hundreds of species in the human microbiomes will be crucial to better understand the bacterial constituents of human-associated microbial ecosystems.

Material and Methods

Collection and taxonomic annotation of *Akkermansia* sp. genomes

The *Akkermansia* genomic sequences used in this work were retrieved from four sources: (i) newly sequenced *Akkermansia* genomes from cultivated strains (Ouwerkerk 2016), (ii) publicly available isolated genomes from NCBI (downloaded as of March 2020) that were labeled as *Akkermansia muciniphila* or *Akkermansia* sp., (iii) metagenome-assembled genomes (MAGs) coming from a collection of metagenomes from human microbiome by Pasolli et al. (Pasolli et al. 2019)**,** and (iv) 166,518 additional MAGs reconstructed from 9,172 metagenomes (**Additional file 1**: **Table S4**) obtained with a validated assembly-based pipeline similarly to Pasolli et al. (Pasolli et al. 2019).

For the 166,518 additional MAGs reconstructed specifically for this work, the metagenomes were assembled using metaSPAdes (Nurk et al. 2017) if paired-end metagenomes were available, and MEGAHIT (D. Li et al. 2016) otherwise. In both cases default parameters were used. Contigs longer than 1,500 nucleotides were binned into MAGs using MetaBAT2 (Kang et al. 2019). We assigned MAGs to previously defined species-level genome bins (SGB) (Pasolli et al. 2019) based on whole-genome nucleotide similarity estimation using Mash (Ondov et al. 2016) and only MAGs falling in the SGBs belonging to the *Akkermansiaceae* family were further considered. We then quality controlled the MAGs and genomes using checkM (version 1.1.3) (Parks et al. 2015) and kept genomes estimated to be high-quality according to genomic completeness > 90% and genomic contamination <5%.

The above procedure resulted in a total of 2,420 *Akkermansia* genomes being considered in this work (http://segatalab.cibio.unitn.it/data/Akkermansia_Karcher_et_al.html): 188 isolate genomes from NCBI (119 labeled as *Akkermansia muciniphila* and 69 labeled as *Akkermansia* sp.), 2,226 MAGs and 6 novel genomes coming from strains isolated from the human gut. The 2,420 genomes were assigned to a total of five candidate species which includes the already recognized *Akkermansia muciniphila* species and four additional SGBs: SGB9223, SGB9224, SGB9227, and SGB9228 as summarized in Table 1.

	SGB9223	SGB9224	*A. muciniphila*	SGB 9227	SGB9228	Total
MAGs	29	93	1802	4	298	2226
Isolate	66	3	108	2	9	188

genomes (NCBI)						
Isolate genomes (generated)	0	0	6	0	0	6
Total	95	96	1916	6	307	2420

Table 1: Summary of the number of genomes per candidate species.

Identification and comparison of the 16S rRNA genes from genomes and MAGs

16S rRNA genes were identified using Barrnap (version 0.9) with default parameters. We only considered extracted 16S rRNA genes sequences longer than 1000 nucleotides. We retained a total of 445 16S rRNA sequences (255 from isolate genomes and 190 from MAGs). Mapping all these sequences against the NCBI's 16S rRNA gene set identified 11 outlying 16S rRNA genes that had > 98% whole-gene identity to a 16S rRNA gene of a family other than *Akkermansiaceae,* which we removed. We then aligned the sequences using mafft (version v7.471, (Katoh and Standley 2014)) with parameters: *--quiet --anysymbol --localpair --maxiterate 1000)* and computed pairwise edit distances between all sequences.

Genome annotation and gene clustering

We detected and annotated ORFs on all genomes using Prokka (version 1.14) (Seemann 2014). Coding sequences (CDS) were then assigned to a UniRef90 cluster (Suzek et al. 2015) by performing a Diamond search (version 0.9.24) (Buchfink, Xie, and Huson 2015) of the CDS against the UniRef90 database (version 201906) and assigning a Uniref90-ID if the mean sequence identity to the centroid sequence is over 90% and if it covers more than 80% of the centroid sequence. Protein sequences that could not be assigned to any UniRef90 cluster were de-novo clustered using MMseqs2 (Steinegger and Söding 2017) following the Uniclust90 criteria (Mirdita et al. 2017).

Whole-genome phylogenetic analysis

The phylogenetic analyses were performed with PhyloPhlAn3 (Asnicar et al. 2020), using either 400 universal marker genes when applied on the 2,420 *Akkermansia* genomes, or core genes when applied to each separate candidate species. Core genes of an *Akkermansia* candidate species were those ORFs whose assigned UniRef90 annotation (or *de-novo* clustered gene family) was present in at least 80% of the genomes of the candidate species. The number of core genes varied across candidate species, with 1,131 for SGB9223, 799 for SGB9224, 996 for SGB9228, and 169 for *A.*

muciniphila. The phylogenies were obtained using PhyloPhlAn 3.0 using the following flags, in both cases, universal markers and specific core genes: "*--force_nucleotides --trim greedy --fast --diversity low".* The following tools with their specific parameter are used inside the PhyloPhlAn3 framework, diamond was used over blast to generate the database when the database sequences were in proteins:

- diamond (version v2.0.2.140, (Buchfink, Xie, and Huson 2015)) with parameters: *makedb* (to generate the database), *"blastx --quiet --threads 1 --outfmt 6 --more-sensitive --id 50 --max-hsps 35 -k 0"* (to map the dna) and "*blastp --quiet --threads 1 --outfmt 6 --more-sensitive --id 50 --max-hsps 35 -k 0 "*
- blast (version 2.10.1+, (McGinnis and Madden 2004; Camacho et al. 2009)) with parameters: "*makeblastdb -parse_seqids -dbtype nucl* and *blastn -outmft 6 -max_target_seqs 1000000"*
- Mafft (version v7.471, (Katoh and Standley 2014)) with parameters : "*--quiet --anysymbol --localpair --maxiterate 1000"*
- trimal (v1.4.rev15 build[2013-12-17], (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009)) with parameters: ":*-gappyout"*
- RAxML (version 8.2.12), (Stamatakis 2014)) with parameters: "*-p 1989 -m GTRCAT -x 1989 -# 100 -f a"*

Relative abundance estimation of candidate species

In order to estimate the presence and relative abundance of the *Akkermansia* candidate species we extended the database of unique marker genes of MetaPhlAn 3.0 (Beghini et al. 2020; Truong et al. 2015) with those of the newly defined *Akkermansia* candidate species: Unique marker genes were defined starting from the core genes of each of the 5 *Akkermansia* candidate species identified on the clustered gene families described above. Core genes of each *Akkermansia* candidate species were divided into 150 nucleotide fragments and then aligned against the genomes of all SGBs including both the other *Akkermansia* candidate species as well as the whole set of bacterial and archaeal SGBs defined in Pasolli et al (Pasolli et al. 2019) using bowtie2 (version 2.3.5.1; --sensitive option) (Longmead and Salzberg 2012). A core gene was considered present in a genome if at least one of the gene's fragments was mapping against it. Core genes never found in more than 1% of the sequences included in any other SGBs were selected as marker genes, obtaining 39, 22, 115, 100 and 135 species-specific unique markers for SGB9223, SGB9224, SGB9227, SGB9228 and *A. muciniphila*, respectively. MetaPhlan 3 was then used with default parameters. Prevalence of candidate species was defined as the percentage of samples in which the candidate species was detected. Similarly, the prevalence of the *Akkermansia* genus was defined as the percentage of samples in which at least one of the candidate species could be detected.

Covariation among candidate species

Covariation among relative abundances of *Akkermansia* candidate species was assessed in 4,171 human metagenome samples in which at least one of the candidate species was detected (out of the 11,014 metagenomes from humans, **Additional file 1**: **Table S2**) by performing pairwise Spearman's correlations (cor.test in the stats R package (Computing 2013)). We corrected for multiple testing using the Benjamini-Hochberg procedure at 10% FDR.

Association between candidate species and metadata parameters

The association between relative abundances of *Akkermansia* candidate species and host BMI, age, and gender was analysed in 3,311 human metagenomic samples from 22 datasets in which this information was available (**Additional file 1**: **Table S3**). For continuous variables (age and BMI), Spearman's correlations were computed using the *pcor.test* function from the *ppcor* R package (Kim 2015) controlling for the remaining covariates. Resulting correlations were used as input in the *metacor* function from the *meta* R package (Balduzzi, Rücker, and Schwarzer 2019) using Fisher's Z transformation of correlations and the Paule-Mandel estimator of between-study variance in the random effects model. For categorical variables (sex), an ordinary least squares (OLS) model was first used to adjust for age and BMI. Statistical significance (Wald test) and effect sizes (standardized mean difference) of the associations were extracted from the sex beta coefficients. Resulting effect sizes were inverse-variance averaged using the Paule-Mandel estimator of between-study variance as implemented in the statsmodels python library (Seabold and Perktold 2010) and custom code. We corrected for multiple testing using the Benjamini-Hochberg procedure at 10% FDR.

Identification of Corrin ring biosynthesis genes

Anaerobic corrin ring biosynthesis gene names were obtained from (Shelton et al. 2019). Corresponding KEGG Orthologs (KOs) were then identified in the clustered gene sequences (see above) using KOFAM (Aramaki et al. 2020). Only those hits that passed the optimized bit-score cutoffs from KOFAM were considered. We found a total of 316 genomes with at least one significant hit.

Determination of vitamin B12 utilization and production by *A. muciniphila* and *A. glycaniphila*

The type strains *A. muciniphila* Muc^T (ATCC BAA-835) and *A. glycaniphila* Pyt^T (DSM 100705) were grown in minimal bicarbonate buffered medium supplemented with 0.6% threonine, 30 mM 3:1 Glc:GlcNAc and a vitamin mixture with and without added vitamin B12 (van der Ark et al., 2018). Cultures were inoculated with a preculture produced on mucin-supplemented medium. At several time points (0, 3, 8.5, 21, 28, 33, 48 h) a 1 mL sample was collected to measure cell density (OD 600nm) and determine propionate

concentration as a proxy for vitamin B12 production. Substrate utilization and metabolite production were quantified on a Thermo Scientific HPLC system equipped with an Agilent Metacarb 67H 300 x 6.5 mm column. The column was kept at 45°C, running 0.005 M H_2SO_4 eluent at a flow rate of 1 mL/min. Detection was performed using a refractive index detector. All measurements were performed in duplicate.

LPS Operon identification

The pangenome of *A. muciniphila* was reconstructed using the UniRef90 assignments and complemented with the *de-novo* clustered gene families (see above). Pan-genes were then also annotated with CAZy using a local dbCAN distribution (Yin et al. 2012) (database version V9 with suggested E-value and HMM coverage cutoffs of 1E-18 and 0.35, respectively). We specifically focused on the differential copy number and distribution of the glycosyltransferase enzymes class 2 and 4 (GT2/GT4) in the *A. muciniphila* genomes. We observed two groups within this set of genes that were co-present and mutually exclusive in genomes, suggesting a large structural variation and operon-type distribution of genes. We then determined the two putative archetypes by manual inspection of gene distribution and order on isolate genomes. Finally, the detected dichotomy was confirmed by performing BLAST on operon genes (including bordering genes from the isolate genomes) against all genomes and observing their presence/absence (**Additional file 2**: **Fig. S11**).

CAZy annotation and gene clustering

dbCAN2 ((H. Zhang et al. 2018), database version 07312020) was used to annotate centroid sequences of gene clusters (see above) with Carbohydrate-Active enZYmes (CAZY) information (Cantarel, Coutinho, and Henrissat 2012). dbCAN2 was used with default parameters, and hits with an E-value > 10E-15 and those that covered less than 35% of a given dbCAN2-HMM were removed.

Retrieval of CRISPR spacers in viruses from metagenomes and viromes

Metagenomes enriched for virus-like particles (i.e. viromes) were retrieved through SRA (Leinonen et al. 2011) from 708 samples from five studies (Minot et al. 2013; Ly et al. 2016; Norman et al. 2015; Hannigan et al. 2018; Liang et al. 2020). Samples were uniformly preprocessed with TrimGalore version 0.4.4 (Krueger 2015) to remove low quality and short reads (Phred quality < 20, read length < 75; parameters: --stringency 5 --length 75 --quality 20 --max_n 2 --trim-n). Reads aligning to the human genome (hg19) were identified and subsequently removed via mapping with Bowtie2 version 2.4.1 (Longmead and Salzberg 2012) in global mode. Raw reads were assembled with metaSPAdes (Nurk et al. 2017) version 3.10.1 (k-mer sizes: -k 21,33,55,77,99,127). The efficacy of viral enrichment of each virome was evaluated with ViromeQC (Zolfo et al. 2019), and 126 out of 708 samples had an enrichment higher than 50-fold. Contigs a)

longer than 1500bp; b) originating from highly-enriched viromes (i.e. enrichment >= 50x); c) found binned in the same Species-level Genome Bin (Pasolli et al. 2019) in less than 30 metagenomes; and d) found in the unbinned fraction of more than 20 metagenomes (Pasolli et al. 2019) were retained as putative viral contigs. After this, contigs originating from non-highly-enriched viromes with a high sequence similarity were added to the collection (BLAST identity >= 80%, length >= 1000 nucleotides, by using BLAST, version 2.6.0 (Altschul et al. 1990)). Sequences homologous to the virome-derived contigs were searched in unbinned contigs of Pasolli *et al.* with mash version 2.0 (Ondov et al. 2016), and contigs with a distance lower than 10% (p-value <= 0.05) to any viral contig were added to the collection. Finally, we added 699 full genomes of taxonomically annotated gut bacteriophages from RefSeq, release 99 (Brister et al. 2015) that were also found in at least 20 metagenomes of the unbinned fraction of Pasolli *et al*. (Pasolli et al. 2019).
Putative viral contigs were then clustered at 70% identity with VSearch version 2.14.2 (Rognes et al. 2016) (parameters --cluster_fast --id 0.7 --strand both) and further grouped if they shared more than one third of their sequence at 90% sequence identity or more to produce 1345 “Viral Clusters” (VCs) that were further analyzed.

CRISPR arrays and Cas genes were predicted using CRISPRCasTyper version 1.2.1 (default parameters) (Russel et al. 2020). In order to understand potential interaction of candidate species and VCs, we aligned spacer sequences against VCs with BLAST version 2.2.31 (parameters -task blastn-short -gapopen 1 -gapextend 2 -penalty -1 -reward 1 -evalue 1 -word_size 10). Near-perfect matches were defined as matches with an edit distance <= 2. CRISPR-Cas loci structures were plotted using DNA Features Viewer version 3.0.3 (Zulkower and Rosser 2020). Sequence logos were generated using Logomaker version 0.8 (Tareen and Kinney 2020). We used spacers from orphan as well as non-orphan CRISPR arrays for all spacer-based analyses (**Fig. 3D**, **F-H**). For subsequent analysis, we considered only those four VCs where at least 5% of the genomes of a given candidate species had at least one spacer sequence with a hit.

In order to detect the presence of a VC in a metagenome, we mapped a total of 13,381 gut metagenomes against VC contigs with Bowtie2 (Longmead and Salzberg 2012) version 2.4.1 in global mode. Breadth and depth of coverage were evaluated for each VC with bedtools version 2.29.1 (Quinlan and Hall 2010) (genomecov command, default parameters). Only alignments with a Bowtie2 alignment score (AS:i tag) greater than -50 were considered. A VC was considered detected if at least one sequence in the cluster had a breadth of coverage of at least 50%. Differential abundance of VCs in subspecies were assessed with two-sided Wilcoxon rank-sum tests. P-values of < 0.01 were considered significant.

Declarations

Authors' contributions. NS, WdV, MVC, and NK conceived and supervised the study. NK, EN, MP, ABM, PM, MZ, FC, and MVC performed the data acquisition. NK, EN, MP, ABM, MC, PM, MZ, FC, DG, SM, and MVC performed the data analysis. TPNB, HLPT, and WdV designed and performed the in vitro experiments. NK, EN, MC, AC, MA, MVC, WdV, and NS performed the data interpretation and wrote the manuscript. All authors read and approved the final manuscript.

Funding. This work was supported by the European Research Council (ERC-STG project MetaPG-716575) to NS; by MIUR 'Futuro in Ricerca' (grant No. RBFR13EWWI_001) to NS; by the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project) to NS; by the National Cancer Institute of the National Institutes of Health (1U01CA230551) to NS; and by the Premio Internazionale Lombardia e Ricerca 2019 to NS. This work was partly supported by the SIAM Gravitation Grant 024.002.002 and the 2008 Spinoza Award of the Netherlands Organization for Scientific Research to WMdV. We thank the Institute of Biotechnology, University of Helsinki (Finland) for providing both Illumina and PacBio sequences of new isolates.

Availability of data and materials. The 2,420 *Akkermansia* genomes and MAGs considered in this work are available at http://segatalab.cibio.unitn.it/data/Akkermansia_Karcher_et_al.html as well as in Zenodo under the following accession 0.5281/zenodo.5018705 [111].

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable

Competing interests. WMdV is co-founder and holds stock in A-mansia Biotech Belgium. All the other authors declare that they have no competing interests.

Supplementary Figures

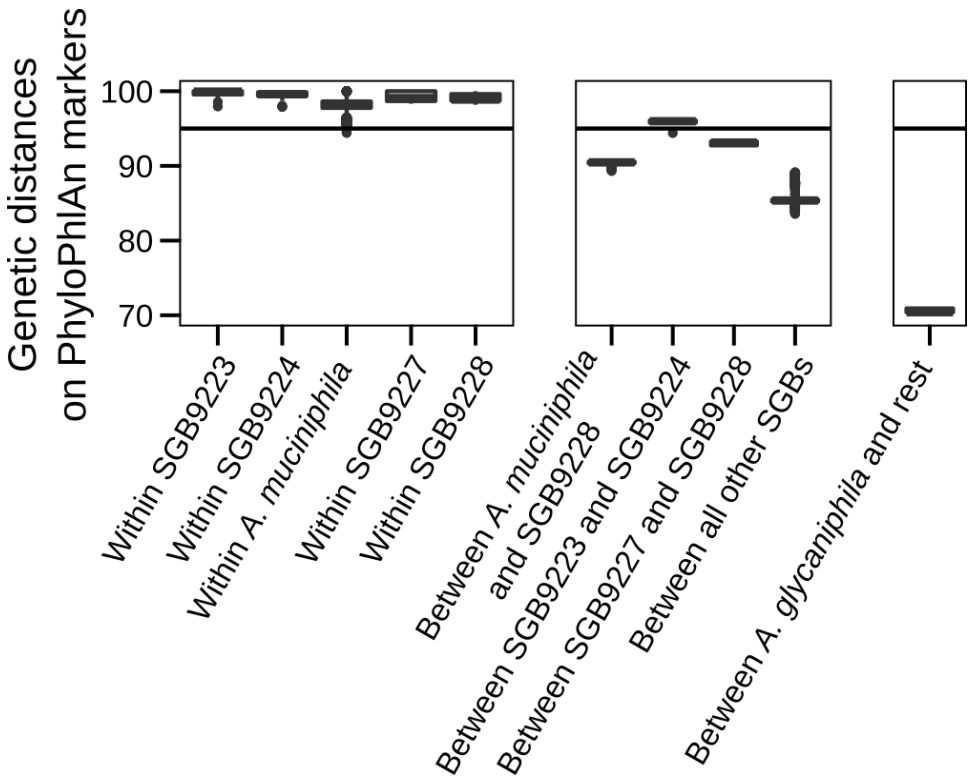


Fig. S1: Within- and between-clade whole-genome genetic distance estimates obtained using PhyloPhlAn 3 (Jain et al. 2018). Related to **Figure 1**.

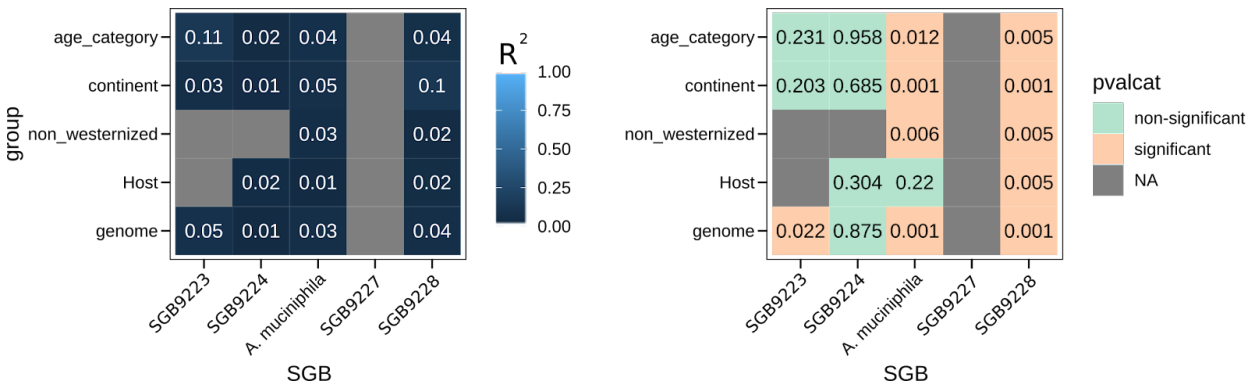


Fig. S2: For each SGB and each considered metadata, we performed a PERMANOVA test on pairwise distances. PERMANOVA tests were performed using the Adonis function from the vegan R package. SGB9227 was omitted because of its small size.

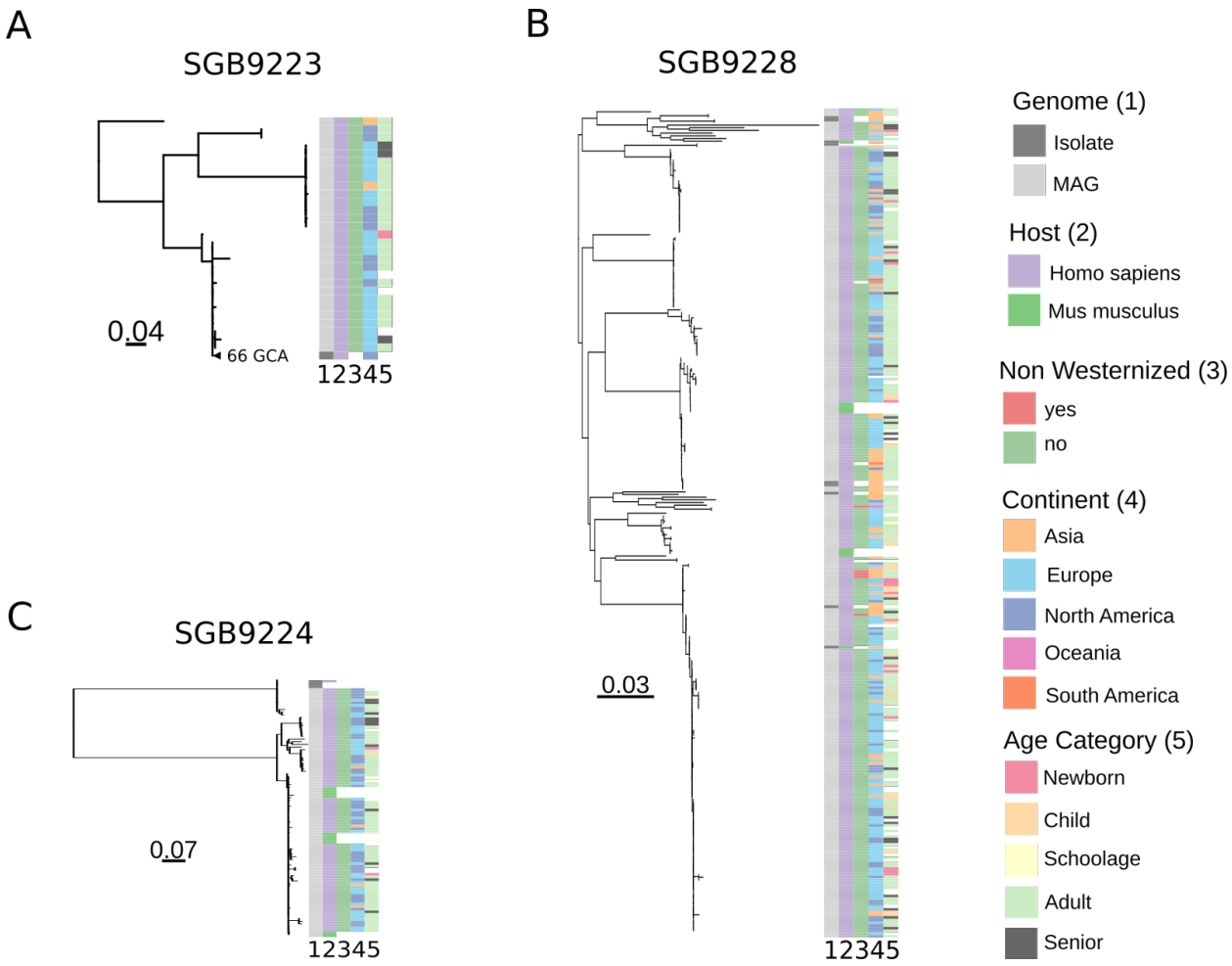


Fig. S3: Phylogenetic trees of *Akkermansia* candidate species built using core genes (see **Methods**). SG9227 is not shown due to its small size. The core genes used for each tree are in at least 80% of the genomes in candidate species, respectively 1131, 799 and 996 for SGB9223, SGB9224 and SGB9228.

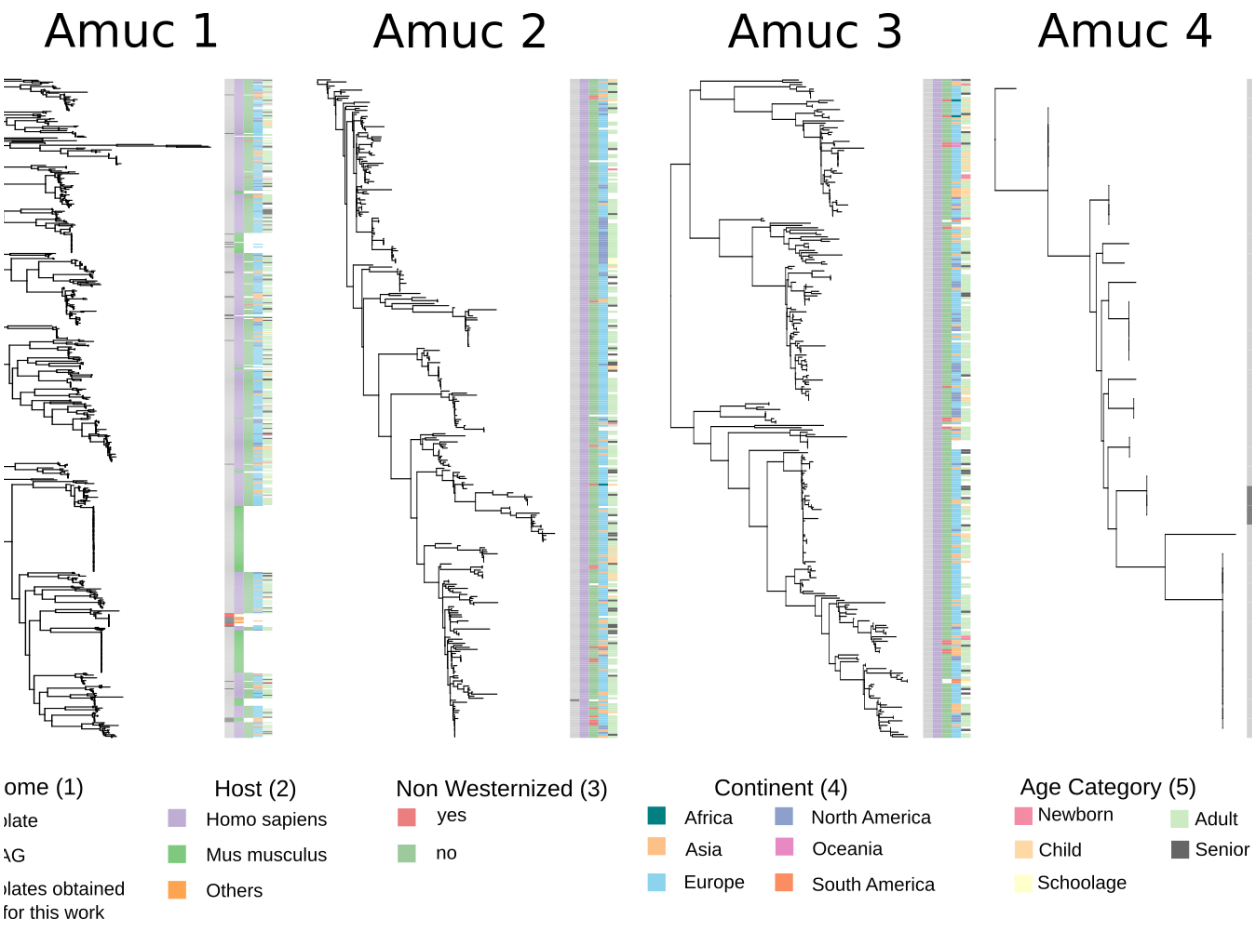


Fig. S4: Phylogenetic tree of *A. muciniphila* subspecies built using 169 core genes (see **Figure 4, Methods**).

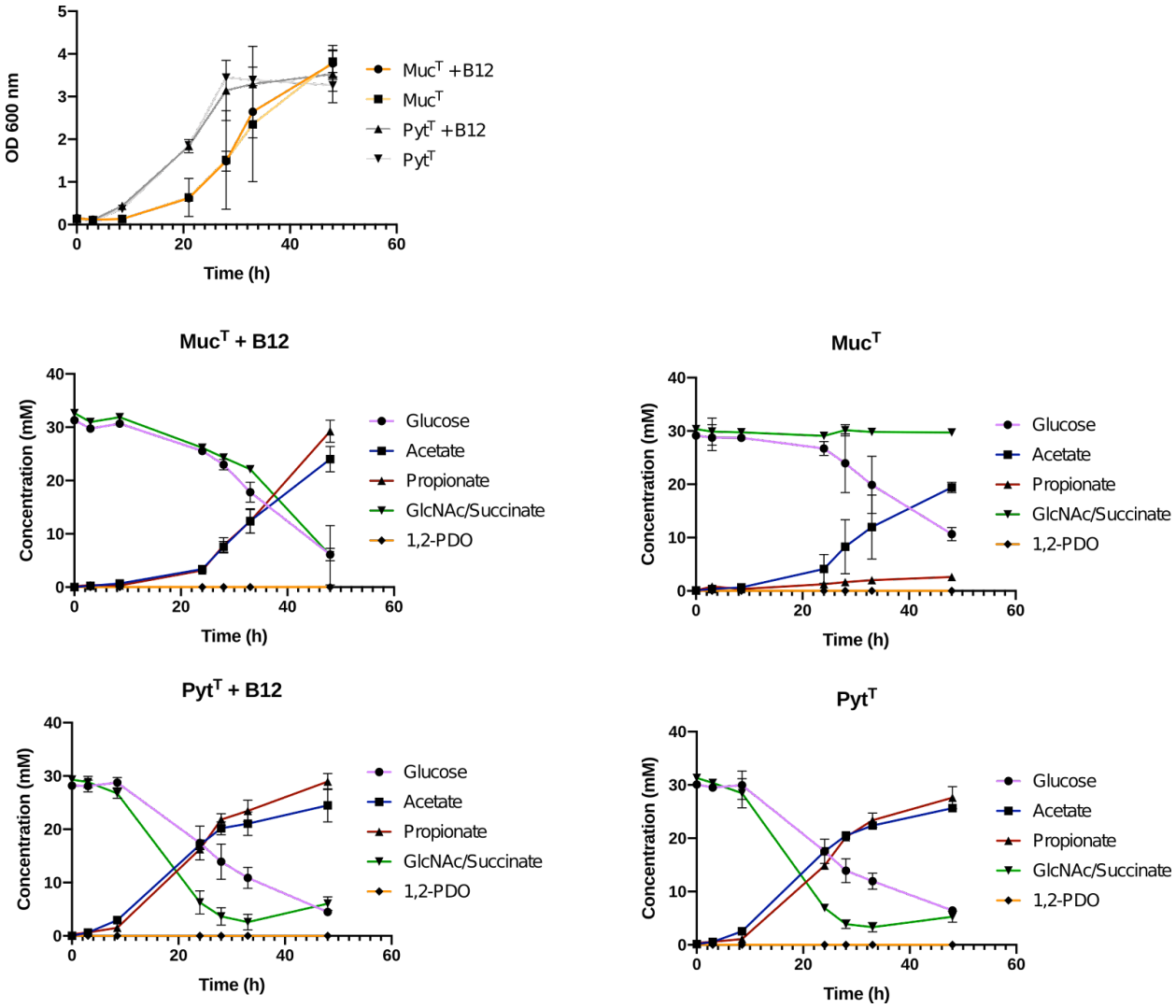


Fig. S5: (A) Growth curve of *A. muciniphila* Muc^T and *A. glycaniphila* Pyt^T in minimal medium in the presence and absence of vitamin B12. Metabolites produced and substrates utilized by both Muc^T (B) and Pyt^T(C) both in the presence and absence of vitamin B12 were measured using HPLC, with propionate being a proxy for B12 production and utilization as its production is dependent on the B12-dependent methyl-malonyl CoA synthase reaction (Ottman et al. 2017).

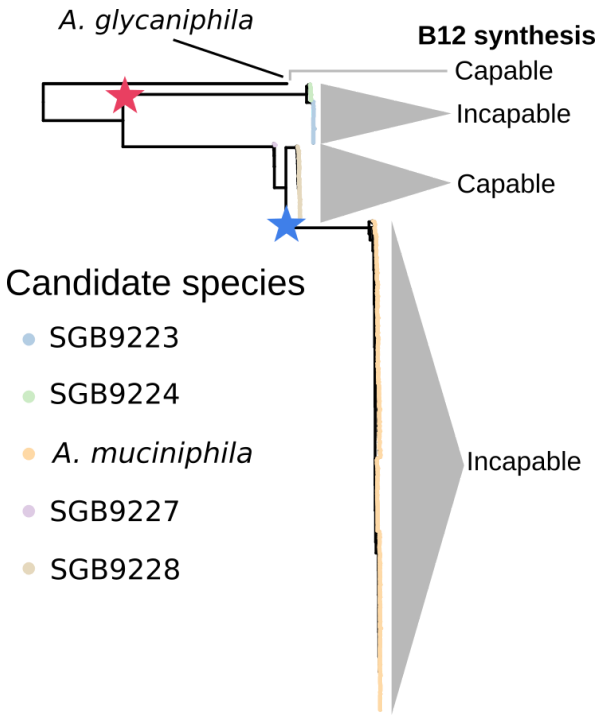


Fig. S6: Phylogenetic tree of all candidate species (including *A. glycaniphila*) annotated with B12 synthesis capabilities. Stars indicate two putative loss of vitamin B12 biosynthesis capability events in the most recent common ancestor of SGB9223/SGB9224 (red star) and *A. muciniphila* (blue star).

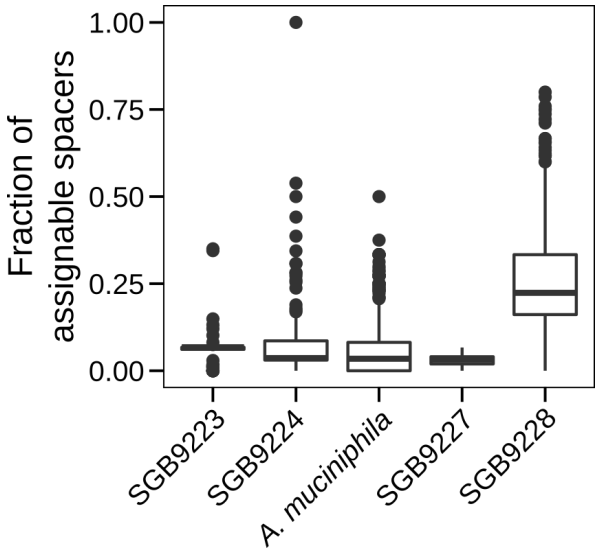


Fig. S7: The total fraction of assignable spacer sequences (those with a near-perfect match against a gut phage, see **Methods**) per *Akkermansia* candidate species.

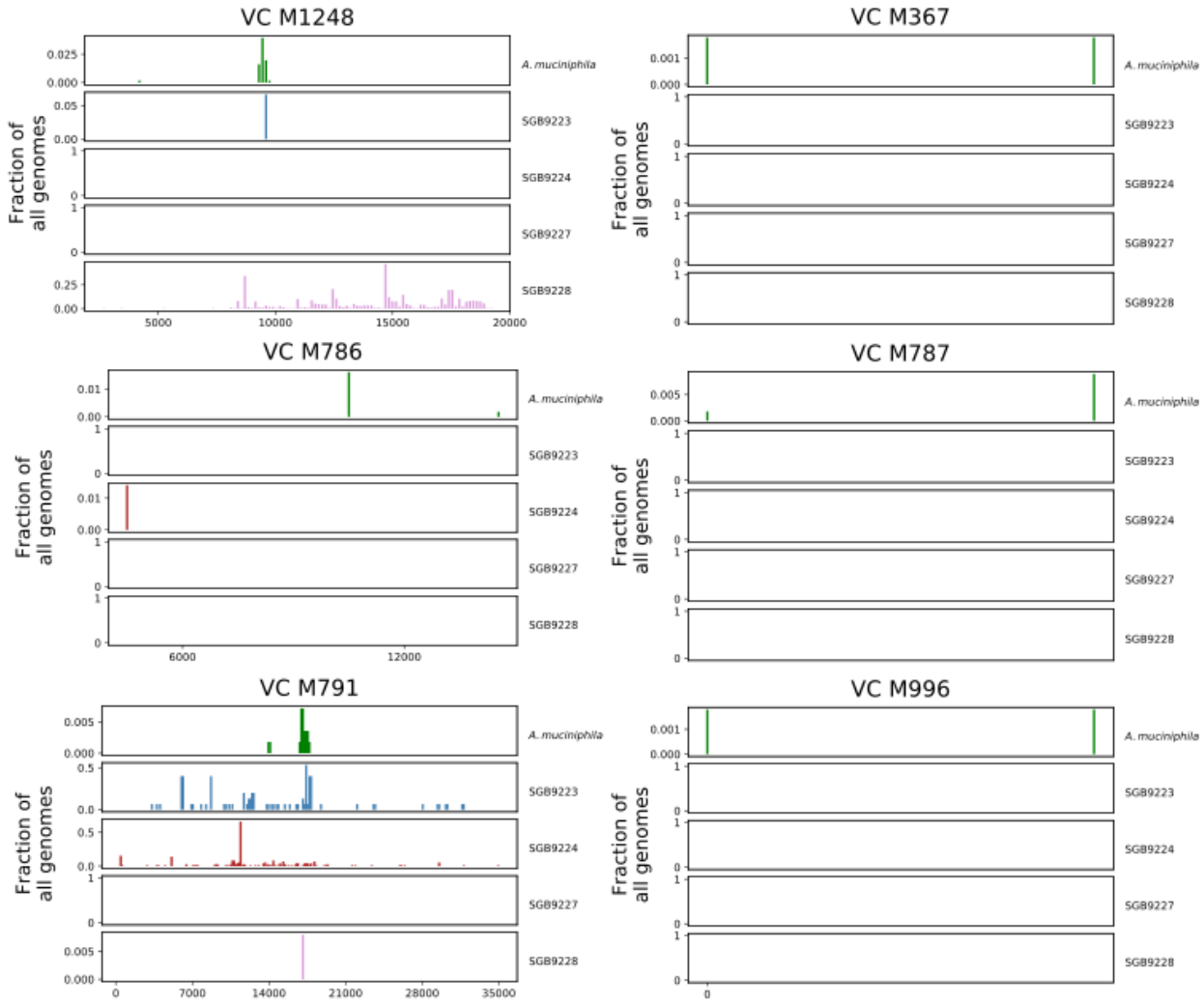


Fig. S8: Maps of spacers from *Akkermansia* genomes against six VCs, visualized with a sliding window of 150 nt. See **Fig. 3G**.

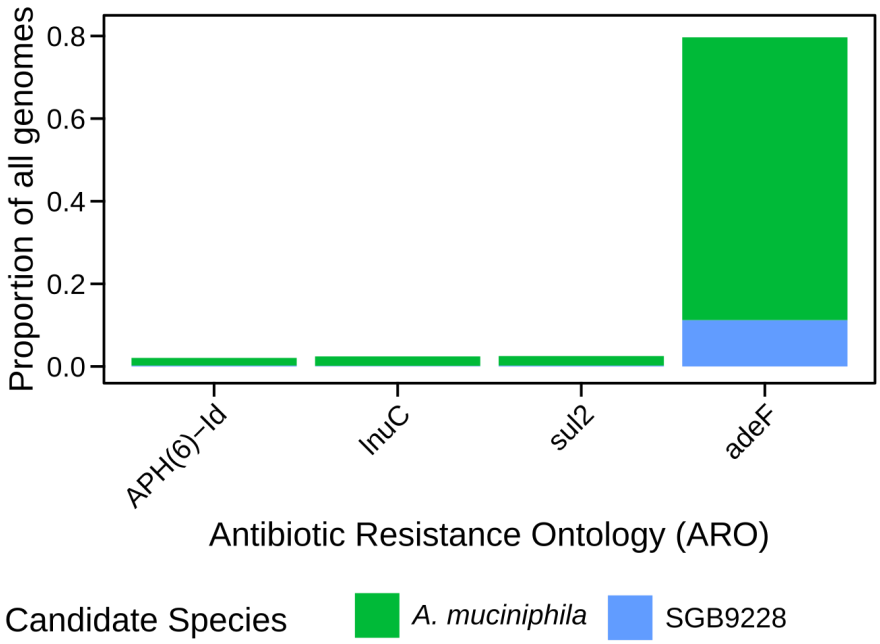


Fig. S9: We annotated all genomes using the resistance gene identifier (RGI) software (https://github.com/arpcard/rgi). Hits above the bit score threshold that covered at least 70% of the reference sequence in CARD were kept. Only hits in the ARO that are found in at least 1% of genomes are shown.

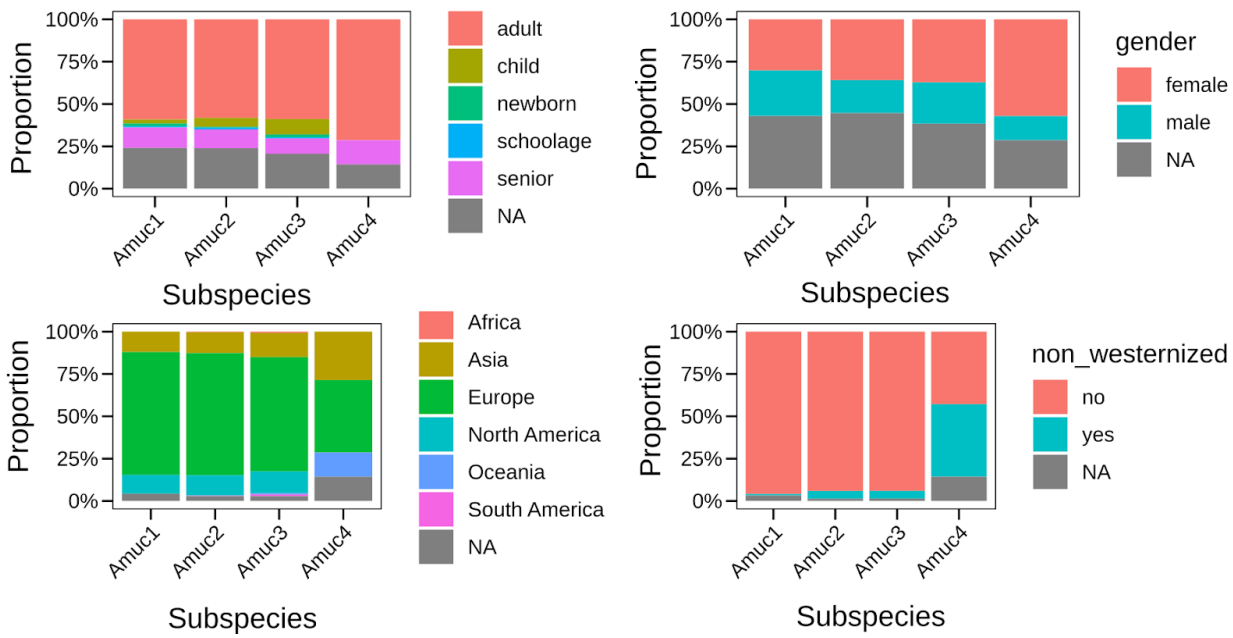


Fig. S10: Barplots showing the distribution of *A. muciniphila* subspecies for host age, gender, continent (origin) and westernization status.

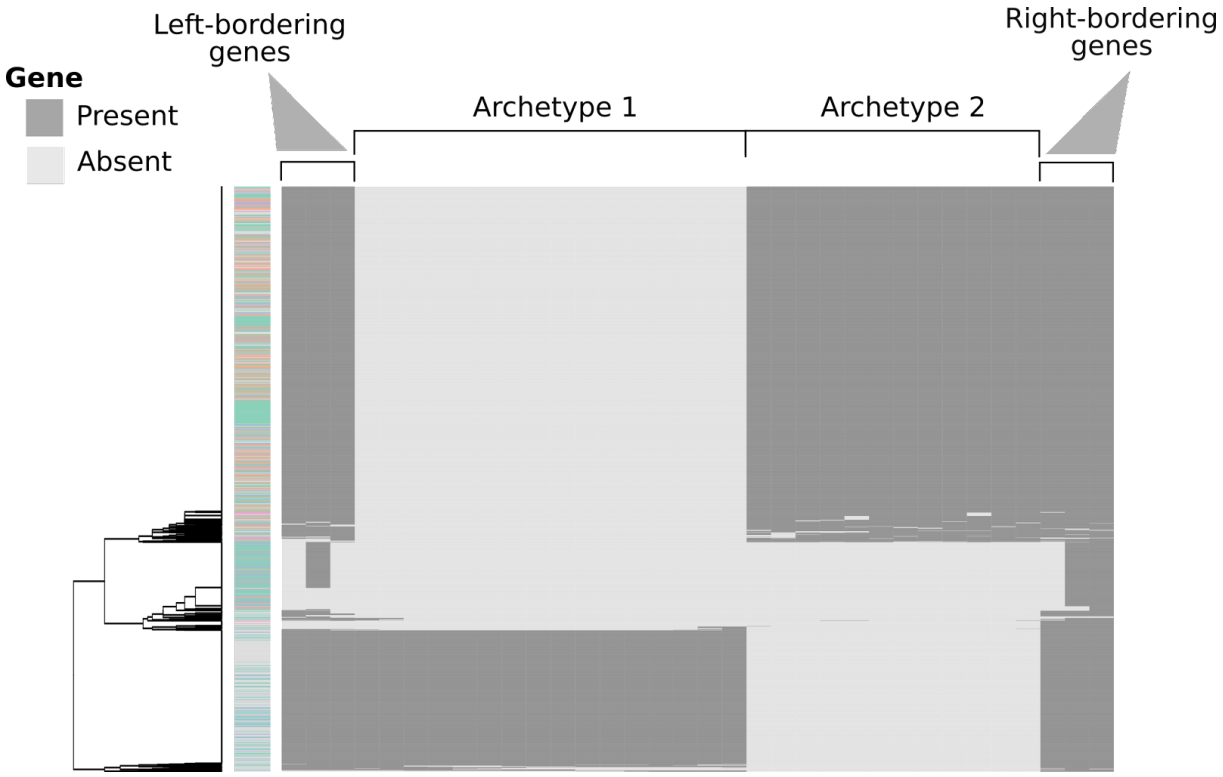


Fig. S11: Gene presence/absence heatmap for both operon archetypes as well as three left- and right-bordering genes. Most MAGs have either one or the other archetype as well as both left- and right-bordering genes, whereas a small fraction of MAGs has neither operon and also only some bordering genes. Related to **Fig. 5**.

Supplementary Tables

For Supplementary Tables see the online version of this publication.

References

Alcock, Brian P., Amogelang R. Raphenya, Tammy T. Y. Lau, Kara K. Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, et al. 2020. “CARD 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database.” *Nucleic Acids Research* 48 (D1): D517–25.

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.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2020. “A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome.” *Nature Biotechnology*, July. https://doi.org/10.1038/s41587-020-0603-3.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Aramaki, Takuya, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. 2020. “KofamKOALA: KEGG Ortholog Assignment Based on Profile HMM and Adaptive Score Threshold.” *Bioinformatics* 36 (7): 2251–52.

Asnicar, Francesco, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, et al. 2020. “Precise Phylogenetic Analysis of Microbial Isolates and Genomes from Metagenomes Using PhyloPhlAn 3.0.” *Nature Communications* 11 (1): 2500.

Balduzzi, Sara, Gerta Rücker, and Guido Schwarzer. 2019. “How to Perform a Meta-Analysis with R: A Practical Tutorial.” *Evidence-Based Mental Health* 22 (4): 153–60.

Barrangou, Rodolphe, and Luciano A. Marraffini. 2014. “CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity.” *Molecular Cell* 54 (2): 234–44.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Miguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2020. “Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.11.19.388223.

Belzer, Clara, Loo Wee Chia, Steven Aalvink, Bhawani Chamlagain, Vieno Piironen, Jan Knol, and Willem M. de Vos. 2017. “Microbial Metabolic Networks at the Mucus Layer Lead to Diet-Independent Butyrate and Vitamin B12 Production by Intestinal Symbionts.” *mBio* 8 (5). https://doi.org/10.1128/mBio.00770-17.

Brister, J. Rodney, Danso Ako-Adjei, Yiming Bao, and Olga Blinkova. 2015. “NCBI Viral Genomes Resource.” *Nucleic Acids Research* 43 (Database issue): D571–77.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods*. https://doi.org/10.1038/nmeth.3176.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (December): 421.

Cantarel, Brandi, Pedro Coutinho, and Bernard Henrissat. 2012. “Carbohydrate-Active Enzymes Database, Metagenomic Expert Resource.” *Encyclopedia of*

Metagenomics. https://doi.org/10.1007/978-1-4614-6418-1_25-10.
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.
Computing, R. F. S. 2013. “R: A Language and Environment for Statistical Computing.” *Vienna: R Core Team*. https://www.yumpu.com/en/document/view/6853895/r-a-language-and-environment-for-statistical-computing.
Costea, Paul I., Luis Pedro Coelho, Shinichi Sunagawa, Robin Munch, Jaime Huerta-Cepas, Kristoffer Forslund, Falk Hildebrand, Almagul Kushugulova, Georg Zeller, and Peer Bork. 2017. “Subspecies in the Global Human Gut Microbiome.” *Molecular Systems Biology* 13 (12): 960.
Costea, Paul Igor, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork. 2017. “metaSNV: A Tool for Metagenomic Strain Level Analysis.” *PloS One* 12 (7): e0182392.
Dao, Maria Carlota, Eugeni Belda, Edi Prifti, Amandine Everard, Brandon D. Kayser, Jean-Luc Bouillot, Jean-Marc Chevallier, et al. 2019. “Akkermansia Muciniphila Abundance Is Lower in Severe Obesity, but Its Increased Level after Bariatric Surgery Is Not Associated with Metabolic Health Improvement.” *American Journal of Physiology. Endocrinology and Metabolism* 317 (3): E446–59.
Dao, Maria Carlota, Amandine Everard, Judith Aron-Wisnewsky, Nataliya Sokolovska, Edi Prifti, Eric O. Verger, Brandon D. Kayser, et al. 2016. “Akkermansia Muciniphila and Improved Metabolic Health during a Dietary Intervention in Obesity: Relationship with Gut Microbiome Richness and Ecology.” *Gut* 65 (3): 426–36.
De Filippis, Francesca, Edoardo Pasolli, and Danilo Ercolini. 2020. “Newly Explored Faecalibacterium Diversity Is Connected to Age, Lifestyle, Geography, and Disease.” *Current Biology: CB*, October. https://doi.org/10.1016/j.cub.2020.09.063.
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.
Degnan, Patrick H., Michiko E. Taga, and Andrew L. Goodman. 2014. “Vitamin B12 as a Modulator of Gut Microbial Ecology.” *Cell Metabolism* 20 (5): 769–78.
Depommier, Clara, Amandine Everard, Céline Druart, Hubert Plovier, Matthias Van Hul, Sara Vieira-Silva, Gwen Falony, et al. 2019. “Supplementation with Akkermansia Muciniphila in Overweight and Obese Human Volunteers: A Proof-of-Concept Exploratory Study.” *Nature Medicine* 25 (7): 1096–1103.
Derrien, Muriel, Elaine E. Vaughan, Caroline M. Plugge, and Willem M. de Vos. 2004. “Akkermansia Muciniphila Gen. Nov., Sp. Nov., a Human Intestinal Mucin-Degrading Bacterium.” *International Journal of Systematic and Evolutionary Microbiology*. https://doi.org/10.1099/ijs.0.02873-0.
Druart, Céline, Hubert Plovier, Matthias Van Hul, Alizée Brient, Kirt R. Phipps, Willem M. de Vos, and Patrice D. Cani. 2021. “Toxicological Safety Evaluation of Pasteurized Akkermansia Muciniphila.” *Journal of Applied Toxicology: JAT* 41 (2): 276–90.
Everard, Amandine, Clara Belzer, Lucie Geurts, Janneke P. Ouwerkerk, Céline Druart, Laure B. Bindels, Yves Guiot, et al. 2013. “Cross-Talk between Akkermansia

Muciniphila and Intestinal Epithelium Controls Diet-Induced Obesity." *Proceedings of the National Academy of Sciences of the United States of America* 110 (22): 9066–71.
Fehlner-Peach, Hannah, Cara Magnabosco, Varsha Raghavan, Jose U. Scher, Adrian Tett, Laura M. Cox, Claire Gottsegen, et al. 2019. "Distinct Polysaccharide Utilization Profiles of Human Intestinal Prevotella Copri Isolates." *Cell Host & Microbe* 26 (5): 680–90.e5.
Garzetti, Debora, Sandrine Brugiroux, Boyke Bunk, Rüdiger Pukall, Kathy D. McCoy, Andrew J. Macpherson, and Bärbel Stecher. 2017. "High-Quality Whole-Genome Sequences of the Oligo-Mouse-Microbiota Bacterial Community." *Genome Announcements* 5 (42). https://doi.org/10.1128/genomeA.00758-17.
Geva-Zatorsky, Naama, Esen Sefik, Lindsay Kua, Lesley Pasman, Tze Guan Tan, Adriana Ortiz-Lopez, Tsering Bakto Yanortsang, et al. 2017. "Mining the Human Gut Microbiota for Immunomodulatory Organisms." *Cell* 168 (5): 928–43.e11.
Guo, Xianfeng, Shenghui Li, Jiachun Zhang, Feifan Wu, Xiangchun Li, Dan Wu, Min Zhang, et al. 2017. "Genome Sequencing of 39 Akkermansia Muciniphila Isolates Reveals Its Population Structure, Genomic and Functional Diverisity, and Global Distribution in Mammalian Gut Microbiotas." *BMC Genomics* 18 (1): 800.
Hall, Andrew Brantley, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K. Lagoudas, et al. 2017. "A Novel Ruminococcus Gnavus Clade Enriched in Inflammatory Bowel Disease Patients." *Genome Medicine* 9 (1): 103.
Hamilton, Thomas A., Gregory M. Pellegrino, Jasmine A. Therrien, Dalton T. Ham, Peter C. Bartlett, Bogumil J. Karas, Gregory B. Gloor, and David R. Edgell. 2019. "Efficient Inter-Species Conjugative Transfer of a CRISPR Nuclease for Targeted Bacterial Killing." *Nature Communications* 10 (1): 4544.
Hannigan, Geoffrey D., Melissa B. Duhaime, Mack T. Ruffin 4th, Charlie C. Koumpouras, and Patrick D. Schloss. 2018. "Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome." *mBio* 9 (6). https://doi.org/10.1128/mBio.02248-18.
Human Microbiome Project Consortium. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (7402): 207–14.
Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries." *Nature Communications* 9 (1): 5114.
Jiang, Xiaofang, A. Brantley Hall, Timothy D. Arthur, Damian R. Plichta, Christian T. Covington, Mathilde Poyet, Jessica Crothers, et al. 2019. "Invertible Promoters Mediate Bacterial Phase Variation, Antibiotic Resistance, and Host Adaptation in the Gut." *Science* 363 (6423): 181–87.
Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. "MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies." *PeerJ* 7 (July): e7359.
Karcher, Nicolai, Edoardo Pasolli, Francesco Asnicar, Kun D. Huang, Adrian Tett, Serena Manara, Federica Armanini, et al. 2020. "Analysis of 1321 Eubacterium Rectale Genomes from Metagenomes Uncovers Complex Phylogeographic

Population Structure and Subspecies Functional Adaptations.” *Genome Biology*. https://doi.org/10.1186/s13059-020-02042-y.
Katoh, Kazutaka, and Daron M. Standley. 2014. “MAFFT: Iterative Refinement and Additional Methods.” *Methods in Molecular Biology* 1079: 131–46.
Kim, Seongho. 2015. “Ppcor: An R Package for a Fast Calculation to Semi-Partial Correlation Coefficients.” *Communications for Statistical Applications and Methods* 22 (6): 665–74.
Kirmiz, Nina, Kadir Galindo, Karissa L. Cross, Estefani Luna, Nicholas Rhoades, Mircea Podar, and Gilberto E. Flores. 2020. “Comparative Genomics Guides Elucidation of Vitamin B12 Biosynthesis in Novel Human-Associated Akkermansia Strains.” *Applied and Environmental Microbiology* 86 (3). https://doi.org/10.1128/AEM.02117-19.
Krueger, Felix. 2015. “Trim Galore.” *A Wrapper Tool around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files* 516: 517.
Leenay, Ryan T., Kenneth R. Maksimchuk, Rebecca A. Slotkowski, Roma N. Agrawal, Ahmed A. Gomaa, Alexandra E. Briner, Rodolphe Barrangou, and Chase L. Beisel. 2016. “Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems.” *Molecular Cell* 62 (1): 137–47.
Leinonen, Rasko, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. 2011. “The Sequence Read Archive.” *Nucleic Acids Research* 39 (Database issue): D19–21.
Liang, Guanxiang, Maire A. Conrad, Judith R. Kelsen, Lyanna R. Kessler, Jessica Breton, Lindsey G. Albenberg, Sarah Marakos, et al. 2020. “The Dynamics of the Stool Virome in Very Early Onset Inflammatory Bowel Disease.” *Journal of Crohn’s & Colitis*, May. https://doi.org/10.1093/ecco-jcc/jjaa094.
Li, Dinghua, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. 2016. “MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler Driven by Advanced Methodologies and Community Practices.” *Methods* 102 (June): 3–11.
Li, Hai, Julien P. Limenitakis, Tobias Fuhrer, Markus B. Geuking, Melissa A. Lawson, Madeleine Wyss, Sandrine Brugiroux, et al. 2015. “The Outer Mucus Layer Hosts a Distinct Intestinal Microbial Niche.” *Nature Communications* 6 (September): 8292.
Liu, Chang, Nan Zhou, Meng-Xuan Du, Yu-Tong Sun, Kai Wang, Yu-Jing Wang, Dan-Hua Li, et al. 2020. “The Mouse Gut Microbial Biobank Expands the Coverage of Cultured Bacteria.” *Nature Communications* 11 (1): 79.
Longmead, B., and S. L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie2.” https://www.sid.ir/en/journal/ViewPaper.aspx?ID=436196.
Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J. Xavier, and Dirk Gevers. 2015. “ConStrains Identifies Microbial Strains in Metagenomic Datasets.” *Nature Biotechnology* 33 (10): 1045–52.
Lv, Q. B., S. H. Li, Y. Zhang, Y. C. Wang, Y. Z. Peng, and X. X. Zhang. 2020. “A Thousand Metagenome-Assembled Genomes of Akkermansia Reveal New Phylogroups and Geographical and Functional Variations in Human Gut.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.09.10.292292v1.abstract.
Ly, Melissa, Marcus B. Jones, Shira R. Abeles, Tasha M. Santiago-Rodriguez, Jonathan Gao, Ivan C. Chan, Chandrabali Ghose, and David T. Pride. 2016. “Transmission of

Viruses via Our Microbiomes.” *Microbiome* 4 (1): 64.
Maier, Lisa, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, Exene Erin Anderson, Ana Rita Brochado, et al. 2018. “Extensive Impact of Non-Antibiotic Drugs on Human Gut Bacteria.” *Nature* 555 (7698): 623–28.
Makarova, Kira S., Yuri I. Wolf, Jaime Iranzo, Sergey A. Shmakov, Omer S. Alkhnbashi, Stan J. J. Brouns, Emmanuelle Charpentier, et al. 2020. “Evolutionary Classification of CRISPR-Cas Systems: A Burst of Class 2 and Derived Variants.” *Nature Reviews. Microbiology* 18 (2): 67–83.
Manara, Serena, Francesco Asnicar, Francesco Beghini, Davide Bazzani, Fabio Cumbo, Moreno Zolfo, Eleonora Nigro, et al. 2019. “Microbial Genomes from Non-Human Primate Gut Metagenomes Expand the Primate-Associated Bacterial Tree of Life with over 1000 Novel Species.” *Genome Biology* 20 (1): 299.
Matson, Vyara, Jessica Fessler, Riyue Bao, Tara Chongsuwat, Yuanyuan Zha, Maria-Luisa Alegre, Jason J. Luke, and Thomas F. Gajewski. 2018. “The Commensal Microbiome Is Associated with anti–PD-1 Efficacy in Metastatic Melanoma Patients.” *Science* 359 (6371): 104–8.
McGinnis, Scott, and Thomas L. Madden. 2004. “BLAST: At the Core of a Powerful and Diverse Set of Sequence Analysis Tools.” *Nucleic Acids Research* 32 (Web Server issue): W20–25.
Medvecky, Matej, Darina Cejkova, Ondrej Polansky, Daniela Karasova, Tereza Kubasova, Alois Cizek, and Ivan Rychlik. 2018. “Whole Genome Sequencing and Function Prediction of 133 Gut Anaerobes Isolated from Chicken Caecum in Pure Cultures.” *BMC Genomics* 19 (1): 561.
Minot, Samuel, Alexandra Bryson, Christel Chehoud, Gary D. Wu, James D. Lewis, and Frederic D. Bushman. 2013. “Rapid Evolution of the Human Gut Virome.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (30): 12450–55.
Mirdita, Milot, Lars von den Driesch, Clovis Galiez, Maria J. Martin, Johannes Söding, and Martin Steinegger. 2017. “Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments.” *Nucleic Acids Research* 45 (D1): D170–76.
Mojica, Francisco J. M., and Francisco Rodriguez-Valera. 2016. “The Discovery of CRISPR in Archaea and Bacteria.” *The FEBS Journal* 283 (17): 3162–69.
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.
Norman, Jason M., Scott A. Handley, Megan T. Baldridge, Lindsay Droit, Catherine Y. Liu, Brian C. Keller, Amal Kambal, et al. 2015. “Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease.” *Cell* 160 (3): 447–60.
Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. “metaSPAdes: A New Versatile Metagenomic Assembler.” *Genome Research* 27 (5): 824–34.
Ogata, Yusuke, Mitsuo Sakamoto, Moriya Ohkuma, Masahira Hattori, and Wataru Suda. 2020. “Complete Genome Sequence of Akkermansia Muciniphila JCM 30893, Isolated from Feces of a Healthy Japanese Male.” *Microbiology Resource Announcements* 9 (7). https://doi.org/10.1128/MRA.01543-19.

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.
Ottman, Noora, Mark Davids, Maria Suarez-Diez, Sjef Boeren, Peter J. Schaap, Vitor A. P. Martins Dos Santos, Hauke Smidt, Clara Belzer, and Willem M. de Vos. 2017. “Genome-Scale Model and Omics Analysis of Metabolic Capacities of Akkermansia Muciniphila Reveal a Preferential Mucin-Degrading Lifestyle.” *Applied and Environmental Microbiology* 83 (18). https://doi.org/10.1128/AEM.01014-17.
Ouwerkerk, Janneke P. 2016. “Akkermansia Species: Phylogeny, Physiology and Comparative Genomics.” https://research.wur.nl/en/publications/akkermansia-species-phylogeny-physiology-and-comparative-genomics.
Ouwerkerk, Janneke P., Steven Aalvink, Clara Belzer, and Willem M. de Vos. 2016. “Akkermansia Glycaniphila Sp. Nov., an Anaerobic Mucin-Degrading Bacterium Isolated from Reticulated Python Faeces.” *International Journal of Systematic and Evolutionary Microbiology* 66 (11): 4614–20.
Ouwerkerk, Janneke P., Jasper J. Koehorst, Peter J. Schaap, Jarmo Ritari, Lars Paulin, Clara Belzer, and Willem M. de Vos. 2017. “Complete Genome Sequence of Akkermansia Glycaniphila Strain PytT, a Mucin-Degrading Specialist of the Reticulated Python Gut.” *Genome Announcements* 5 (1). https://doi.org/10.1128/genomeA.01098-16.
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.
Poyet, M., M. Groussin, S. M. Gibbons, J. Avila-Pacheco, X. Jiang, S. M. Kearney, A. R. Perrotta, et al. 2019. “A Library of Human Gut Bacterial Isolates Paired with Longitudinal Multiomics Data Enables Mechanistic Microbiome Research.” *Nature Medicine* 25 (9): 1442–52.
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.
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.
Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.
Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé.

2016. “VSEARCH: A Versatile Open Source Tool for Metagenomics.” *PeerJ* 4 (October): e2584.
Routy, Bertrand, Emmanuelle Le Chatelier, Lisa Derosa, Connie P. M. Duong, Maryam Tidjani Alou, Romain Daillère, Aurélie Fluckiger, et al. 2018. “Gut Microbiome Influences Efficacy of PD-1-Based Immunotherapy against Epithelial Tumors.” *Science* 359 (6371): 91–97.
Russel, Jakob, Rafael Pinilla-Redondo, David Mayo-Muñoz, Shiraz A. Shah, and Søren J. Sørensen. 2020. “CRISPRCasTyper: An Automated Tool for the Identification, Annotation and Classification of CRISPR-Cas Loci.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.05.15.097824.
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.
Seabold, Skipper, and Josef Perktold. 2010. “Statsmodels: Econometric and Statistical Modeling with Python.” *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-011.
Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.
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* 4: 2304.
Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. 2012. “Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes.” *Nature Methods* 9 (8): 811–14.
Shelton, Amanda N., Erica C. Seth, Kenny C. Mok, Andrew W. Han, Samantha N. Jackson, David R. Haft, and Michiko E. Taga. 2019. “Uneven Distribution of Cobamide Biosynthesis and Dependence in Bacteria Predicted by Comparative Genomics.” *The ISME Journal* 13 (3): 789–804.
Shetty, Sudarshan A., Simone Zuffa, Thi Phuong Nam Bui, Steven Aalvink, Hauke Smidt, and Willem M. De Vos. 2018. “Reclassification of Eubacterium Hallii as Anaerobutyricum Hallii Gen. Nov., Comb. Nov., and Description of Anaerobutyricum Soehngenii Sp. Nov., a Butyrate and Propionate-Producing Bacterium from Infant Faeces.” *International Journal of Systematic and Evolutionary Microbiology* 68 (12): 3741–46.
Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.
Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology* 35 (11): 1026–28.
Suzek, Baris E., Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and UniProt Consortium. 2015. “UniRef Clusters: A Comprehensive and Scalable

Alternative for Improving Sequence Similarity Searches.” *Bioinformatics* 31 (6): 926–32.
Tareen, Ammar, and Justin B. Kinney. 2020. “Logomaker: Beautiful Sequence Logos in Python.” *Bioinformatics* 36 (7): 2272–74.
Tett, Adrian, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, et al. 2019. “The Prevotella Copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations.” *Cell Host & Microbe*, September. https://doi.org/10.1016/j.chom.2019.08.018.
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.
Tytgat, Hanne L. P., François P. Douillard, Justus Reunanen, Pia Rasinkangas, Antoni P. A. Hendrickx, Pia K. Laine, Lars Paulin, Reetta Satokari, and Willem M. de Vos. 2016. “Lactobacillus Rhamnosus GG Outcompetes Enterococcus Faecium via Mucus-Binding Pili: Evidence for a Novel and Heterospecific Probiotic Mechanism.” *Applied and Environmental Microbiology* 82 (19): 5756–62.
Van Rossum, Thea, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. 2020. “Diversity within Species: Interpreting Strains in Microbiomes.” *Nature Reviews. Microbiology* 18 (9): 491–506.
Vos, Willem M. de. 2017. “Microbe Profile: Akkermansia Muciniphila: A Conserved Intestinal Symbiont That Acts as the Gatekeeper of Our Mucosa.” *Microbiology* 163 (5): 646–48.
Wirbel, Jakob, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, et al. 2019. “Meta-Analysis of Fecal Metagenomes Reveals Global Microbial Signatures That Are Specific for Colorectal Cancer.” *Nature Medicine* 25 (4): 679–89.
Xing, Juyuan, Xiaobo Li, Yingjiao Sun, Juanjuan Zhao, Shaohua Miao, Qin Xiong, Yonggang Zhang, and Guishan Zhang. 2019. “Comparative Genomic and Functional Analysis of Akkermansia Muciniphila and Closely Related Species.” *Genes & Genomics* 41 (11): 1253–64.
Yang, Chao, Ilaria Mogno, Eduardo J. Contijoch, Joshua N. Borgerding, Varun Aggarwala, Zhihua Li, Sophia Siu, et al. 2020. “Fecal IgA Levels Are Determined by Strain-Level Differences in Bacteroides Ovatus and Are Modifiable by Gut Microbiota Manipulation.” *Cell Host & Microbe* 27 (3): 467–75.e6.
Yang, Meng, Shambhunath Bose, Sookyoung Lim, Jaegu Seo, Joohyun Shin, Dokyung Lee, Won-Hyong Chung, Eun-Ji Song, Young-Do Nam, and Hojun Kim. 2020. “Beneficial Effects of Newly Isolated Akkermansia Muciniphila Strains from the Human Gut on Obesity and Metabolic Dysregulation.” *Microorganisms*. https://doi.org/10.3390/microorganisms8091413.
Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. “dbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme

Annotation.” *Nucleic Acids Research* 40 (Web Server issue): W445–51.
Zhang, Han, Tanner Yohe, Le Huang, Sarah Entwistle, Peizhi Wu, Zhenglu Yang, Peter K. Busk, Ying Xu, and Yanbin Yin. 2018. “dbCAN2: A Meta Server for Automated Carbohydrate-Active Enzyme Annotation.” *Nucleic Acids Research* 46 (W1): W95–101.
Zhang, Xiuying, Huanzi Zhong, Yufeng Li, Zhun Shi, Huahui Ren, Zhe Zhang, Xianghai Zhou, et al. 2021. “Sex- and Age-Related Trajectories of the Adult Human Gut Microbiota Shared across Populations of Different Ethnicities.” *Nature Aging*. https://doi.org/10.1038/s43587-020-00014-2.
Zhou, Kequan. 2017. “Strategies to Promote Abundance of Akkermansia Muciniphila, an Emerging Probiotics in the Gut, Evidence from Dietary Intervention Studies.” *Journal of Functional Foods* 33 (June): 194–201.
Zhu, Qiyun, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G. Sanders, Pedro Belda-Ferre, et al. 2019. “Phylogenomics of 10,575 Genomes Reveals Evolutionary Proximity between Domains Bacteria and Archaea.” *Nature Communications*. https://doi.org/10.1038/s41467-019-13443-4.
Zolfo, Moreno, Federica Pinto, Francesco Asnicar, Paolo Manghi, Adrian Tett, Frederic D. Bushman, and Nicola Segata. 2019. “Detecting Contamination in Viromes Using ViromeQC.” *Nature Biotechnology* 37 (12): 1408–12.
Zulkower, Valentin, and Susan Rosser. 2020. “DNA Features Viewer: A Sequence Annotation Formatting and Plotting Library for Python.” *Bioinformatics* 36 (15): 4350–52.

Karcher N, Nigro E, Punčochář N, Blanco-Míguez A, Ciciani M, Manghi P, Zolfo M, Cumbo F, Manara S, Golzato D, Cereseto A, Arumugam M, Nam Bui TP, Tytgat H, Valles-Colomer M, de Vos W, Segata N. Supplementary Data for Genomic diversity and ecology of human-associated *Akkermansia* species in the gut microbiome revealed by extensive metagenomic assembly. Zenodo. Doi: 0.5281/zenodo.5018705 (2021).

Chapter 4 | Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases

4.1 | Context and contribution

The human gut is home to hundreds of bacterial strains, each with a unique genetic footprint. Deep shotgun metagenomic sequencing provides enough information to delineate bacterial strains in stool samples. This, together with the fact that strains are specific to individuals - which becomes evident by comparing the relatedness of strains from unrelated and related individuals (Truong et al. 2017; Albanese and Donati 2017; Lloyd-Price et al. 2017; Beghini et al. 2020) - has allowed researchers to utilize metagenomics-facilitated strain tracking to understand the strain sharing and transmission dynamics between individuals. Gut strain transmission is of particular interest in the context of fecal microbiota transplantation (FMT), yet previous studies tracking strains in the context of FMT were very small and consequently not well powered. We thus decided to conduct a meta-analysis of FMT datasets to describe strain transmission dynamics in the human gut (aim 3).

We gathered and analyzed a total of 24 shotgun metagenomics cohorts from 226 FMT instances and examined how strains engraft over cohorts and taxonomy. We found differences in how strains engraft over cohorts, which was driven by whether patients have an infectious disease or have taken antibiotics. We also found a statistically significant association between the strain engraftment rate and clinical success, suggesting that maximizing strain engraftment in the post-FMT recipient is a desirable outcome. Furthermore, we trained statistical models to predict the post-FMT microbiome and used these models in conjunction with a simulation framework to designate donor individuals that lead to an increase in predicted bacterial richness and other predefined bacterial clades in the recipient after FMT. Overall, this work provides insights into the variability of the strain engraftment landscape over cohorts and conditions and presents a first practical step towards making more informed donor choices for FMT.

For this project I was involved in dataset gathering and metadata cleaning. I was also responsible for the machine learning section and worked closely with Michal Punčochář, giving him feedback for his part of the analysis. Finally, as a shared first author of the manuscript, I was involved in writing the manuscript.

This manuscript is currently under review at Nature Medicine.

4.2 | Manuscript

Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases

Ianiro, G.[1,2,*], Punčochář, M.[3,*], Karcher, N.[3,*], Porcari, S.[1,2], Armanini, F.[3], Asnicar, F.[3], Beghini, F.[3], Blanco-Míguez, A.[3], Cumbo, F.[3], Manghi, P.[3], Pinto, F.[3], Masucci, L.[4,5], Quaranta, G.[4,5], De Giorgi, S.[1,2], Sciumè, G.D.[1,2], Bibbò, S.[1,2], Del Chierico, F.[6], Putignani, L.[6], Sanguinetti, M.[4,5], Gasbarrini, A.[1,2], Valles-Colomer, M.[^,3], Cammarota, G.[^,1,2], Segata, N.[^,3,7]

1. Digestive Disease Center, Fondazione Policlinico Universitario “A. Gemelli” IRCCS, Rome, Italy
2. Department of Translational Medicine and Surgery, Catholic University of Rome, Rome, Italy
3. Department CIBIO, University of Trento, Trento, Italy.
4. Microbiology Unit, Fondazione Policlinico Universitario “A. Gemelli” IRCCS, Rome, Italy
5. Department of Basic Biotechnological Sciences, Intensivological and Perioperative Clinics, Catholic University of Rome, Rome, Italy
6. Department of Diagnostic and Laboratory Medicine, Unit of Parasitology and Multimodal Laboratory Medicine Research Area, Unit of Human Microbiome, Bambino Gesù Children's Hospital IRCCS, Rome, Italy
7. IEO, Istituto Europeo di Oncologia IRCSS, Milan, Italy

* Contributed equally.

^ Co-corresponding author.

Note: The version below is the ahead-of-print version of the manuscript, prior to editorial edits.

Abstract

Fecal Microbiota Transplantation (FMT) is highly effective against recurrent *Clostridioides difficile* infection and is considered a promising treatment for other microbiome-related disorders, but a comprehensive understanding of microbial engraftment dynamics is lacking and prevents a more informed application of this therapeutic approach. Here we performed an integrated analysis of 1,364 novel and publicly-available metagenomic samples over 226 FMT recipients. By leveraging improved metagenomic strain-profiling to infer strain sharing, we found that recipients with higher donor strain engraftment had an increased likelihood of a positive clinical outcome. Across cohorts, increased engraftment was noted for antibiotic-treated recipients with infectious diseases compared to antibiotic naïve patients with noncommunicable disease. Bacteroidetes and Actinobacteria species (including *Bifidobacteria*) consistently engrafted better than Firmicutes except for six under-characterized Firmicutes species. Cross-dataset prediction of post-FMT species composition via machine learning proved to be accurate (0.77 average AUROC in leave-one-dataset-out evaluation), highlighted the relevance of microbial abundance, prevalence and taxonomy to infer post-FMT species presence, and was used in a simulation framework to pinpoint donors maximizing specific post-FMT microbiome characteristics. Exploring the FMT engraftment dynamics and their links with clinical variables, our study developed models with the potential to select optimal donors for disease-targeted FMT protocols.

Introduction

Fecal microbiota transplantation (FMT) is the medical procedure of transferring human feces from a healthy donor to a recipient in order to treat a disease associated with microbiome imbalance. FMT has shown success rates of nearly 90% for the treatment of recurrent *Clostridioides difficile* infection (rCDI) (Gianluca Ianiro, Murri, et al. 2019; Baunwall et al. 2020), for which it is an approved indication in clinical practice (Cammarota et al. 2019). FMT has more recently been explored in a variety of other diseases associated with microbiome alterations (De Groot et al. 2017; Rossen et al. 2015; Kootte et al. 2017) or in support of other therapies (Gianluca Ianiro et al. 2020; Davar et al. 2021; Baruch et al. 2021), but its efficacy is usually lower and less consistent over cohorts than for treatment of rCDI (Gianluca Ianiro, Eusebi, et al. 2019; Costello et al. 2017; Green et al. 2020). Some factors that may explain the variable clinical efficacy of FMT for diseases other than rCDI include the presence or abundance of single bacteria and the diversity of the patient microbiome at baseline (Kootte et al. 2017; Rossen et al. 2015), clinical characteristics of the disease (G. Ianiro et al. 2017; Moayyedi et al. 2015), the composition of the donor's gut microbiome (Moayyedi et al. 2015), specific aspects of the FMT working protocols (e.g. route of delivery, amount of infused feces) (Gianluca Ianiro et al. 2018), and differential engraftment among species (Kootte et al. 2017; Rossen et al. 2015). Yet, it is generally not known how strain engraftment might be linked with clinical remission after FMT.

The mechanisms and dynamics dictating which microbial taxa of the donor can engraft in the recipient are poorly understood. Initial studies able to track the successful transmission of donor strains to the recipient and their interaction with the resident strains were performed on very few donor/recipient pairs (S. S. Li et al. 2016). Availability of larger FMT trials and the advances in strain-resolved metagenomics enabled deeper analyses that started to unravel the engraftment efficiency of FMT across diseases and led to the development of statistical models to predict the post-FMT microbiome composition (Smillie et al. 2018), but such investigations remained confined to single cohorts (Smillie et al. 2018; Podlesny and Fricke 2020; Kumar et al. 2017; Aggarwala et al., n.d.; Lee et al. 2017; Wilson et al. 2021; Watson et al. 2021) without cross-cohort and cross-condition generalizability. As deeper strain-level metagenomics is possible (Truong et al. 2017; Beghini et al. 2021; Luo et al. 2015; Olm et al. 2021) and not limited to well characterized microbial taxa (Pasolli et al. 2019; Almeida et al. 2019; Nayfach et al. 2019), and as more metagenomic datasets are becoming available (Baruch et al. 2021; Bar-Yoseph et al. 2021; Damman et al. 2015; Davar et al. 2021; Goll et al. 2020; Hourigan et al. 2019; Gianluca Ianiro et al. 2020; Kong et al. 2020; Kumar et al. 2017; Leo et al. 2020; S. S. Li et al. 2016; Moss et al. 2017; Podlesny and Fricke 2020; Smillie et al. 2018; Suskind et al. 2015; Vaughn et al. 2016; Zhao et al. 2020), an integrative metagenomic analysis of existing datasets has the potential and power to uncover general patterns of microbial engraftment and of connected clinical aspects.

Here we present a systematic meta-analysis of 24 studies that investigated FMT in different clinical settings for which we employed novel strain-resolved metagenomic approaches targeting known microbial species as well as yet-to-be-described taxa to unravel the dynamics of FMT engraftment and its links with clinical outcomes. The considered medical conditions

include infectious diseases (rCDI, multi-drug resistant bacteria [MDRB]), non-infectious diseases (inflammatory bowel disease [IBD], irritable bowel syndrome [IBS], metabolic syndrome, chemotherapy-dependent diarrhea, metastatic melanoma refractory to immune checkpoint inhibitors, Tourette's syndrome), for a total of 21 cohorts including 200 recipients, 103 donors and 1248 samples (Table 1, Supplementary Table 1). Additionally, we included a total of 116 samples from 23 recipients and 8 donors from three novel cohorts enrolled for this study (Supplementary Table 2).

Results and discussion

A metagenomic meta-analysis of public and novel FMT microbiome datasets

In order to evaluate differences and similarities of donor microbiome engraftment after fecal microbiota transplantation (FMT) in different clinical conditions, we systematically retrieved all publicly-available FMT studies that assessed microbiome composition of donors and recipients (pre- and post-FMT) through shotgun metagenomics (see Methods), yielding a total of 21 datasets (Table 1, Supplementary Table 1) (Baruch et al. 2021; Bar-Yoseph et al. 2021; Damman et al. 2015; Davar et al. 2021; Goll et al. 2020; Hourigan et al. 2019; Gianluca Ianiro et al. 2020; Kong et al. 2020; Kumar et al. 2017; Leo et al. 2020; S. S. Li et al. 2016; Moss et al. 2017; Podlesny and Fricke 2020; Smillie et al. 2018; Suskind et al. 2015; Vaughn et al. 2016; Zhao et al. 2020). In each study, we removed samples that were not sequenced at sufficient depth (< 1 Gbp) and those for which we found evidence of mislabeling (see Methods), retaining a total of 1255 metagenomes. These metagenomes belong to a total of 203 FMT procedures for which at least one sample is available from the pre-FMT recipient, the post-FMT recipient, and the corresponding donor, which we refer to as “FMT triads”. As the recipient post-FMT samples were frequently collected at variable time points across cohorts, we selected - for each FMT triad - the sample collected closest to one month after FMT, as 30 days was the value that minimizes the overall time deviation across all FMT triads (Methods, Supplementary Figure 1).

Disease	# datasets (new datasets)	# recipients (new recipients)	# samples (new samples)	Number of post-FMT samples median [IQR]	Disease category	Countries
Clostridioides difficile infection	9 (1)	96 (16)	529 (94)	2.0 [3.0]	Infectious	Italy, Germany, Norway, USA, Canada
Inflammatory bowel disease	5 (1)	38 (2)	188 (8)	2.0 [1.0]	Chronic	France, Italy, USA
Multiple drug resistant bacteria colonization	3 (1)	21 (5)	109 (13)	1.0 [2.0]	Infectious	Italy, Israel, France, Netherlands, Switzerland
Melanoma	2	24	248	4 [7]	Oncological	Israel, USA
Tourette syndrome	1	5	25	2.0 [0.0]	Chronic	China
Metabolic syndrome	2	16	154	3 [0.2]	Chronic	Netherlands
Irritable bowel syndrome	1	20	91	2.0 [0.0]	Chronic	Norway
Tyrosine kinase inhibitor-dependent diarrhea	1	6	27	2.0 [1.5]	Oncological	Italy
Total	**24 (3)**	**226 (23)**	**1371 (116)**	**2 [2]**		

Table 1. Summary and main characteristics of the FMT datasets included in this meta-analysis. Numbers in parentheses correspond to data newly-collected in the present study.

We expanded and complemented the already available datasets with 116 additional metagenomic samples corresponding to 23 FMT triads from 3 novel cohorts of patients with recurrent *Clostridioides difficile* infection (rCDI), inflammatory bowel disease (IBD), and colonization from multi-drug resistant bacteria (MDRB, Table 1, Supplementary Table 2,

Methods), enrolled in Italy (Fondazione Policlinico Gemelli IRCCS and Bambino Gesù Children's Hospital), and sequenced at a higher read depth than most other existing FMT datasets (Supplementary Figure 2).

In total, 1,371 samples and 226 FMT triads from 24 different cohorts (Table 1) were thus included in the analysis, each dataset contributing on average 9.5 (sd: 5) FMT instances. While individual cohorts were limited in size, analyzing them together in a meta-analysis setting adds robustness and the possibility to assess cross-dataset consistency of FMT characteristics. Together, the datasets covered 9 different clinical conditions, including rCDI (n=9), IBD (n=5), MDRB colonization (n=3), melanoma (n=2), metabolic syndrome (n=2), and single cohorts of IBS, Tourette syndrome and diarrhea induced by tyrosine kinase inhibitors (Baruch et al. 2021; Bar-Yoseph et al. 2021; Damman et al. 2015; Davar et al. 2021; Goll et al. 2020; Hourigan et al. 2019; Gianluca Ianiro et al. 2020; Kong et al. 2020; Kumar et al. 2017; Leo et al. 2020; S. S. Li et al. 2016; Moss et al. 2017; Podlesny and Fricke 2020; Smillie et al. 2018; Suskind et al. 2015; Vaughn et al. 2016; Zhao et al. 2020). Studies enrolled adult participants with the exception of HouriganS_2019 (Hourigan et al. 2019), ZhaoH_2020 (Zhao et al. 2020), This_study_MDRB and This_study_IBD which included one 2 year old individual and otherwise children or teenagers. Included studies originated from countries in Europe (France, Germany, Italy, Netherlands, Norway), North America (USA), and Asia (China, Israel). All samples were processed following the same pipeline, subjected to quality-control and filtering steps (Methods), and analyzed by quantitative microbial taxonomic profiling and strain-level profiling including yet-to-be-characterized species based on the species-level genome bins (SGBs, see Methods) framework (Pasolli et al. 2019).

Strain-level metagenomics allows precise assessment of strain engraftment after FMT

To identify the transfer and engraftment of the donor microbiome in the FMT recipient and assess the transmission patterns across datasets and clinical variables, we exploited the high person-specificity of microbiome strains detectable by cultivation-free metagenomic approaches (Truong et al. 2017). While a consensus definition of "strain" in metagenomes is not available (Van Rossum et al. 2020; Segata 2018), we adopted an operational species-specific definition by comparing phylogenetic distance distributions of strains of a given species in the same individual sampled over multiple timepoints to those distributions obtained comparing strains from unrelated individuals (Valles-Colomer et al., n.d.) (Supplementary Table 3, Methods). Under the assumption that unrelated individuals who are not in contact hardly share any strains (Lloyd-Price et al. 2017; Truong et al. 2017; Albanese and Donati 2017; Beghini et al. 2021; Ferretti et al. 2018), we performed comparisons of intra-subject against inter-subject conspecific phylogenetic strain distance distributions and identified optimal species-specific phylogenetic cutoffs to define strain identity. Strain profiling and strain identity definitions are implemented in StrainPhlAn 3 (Beghini et al. 2021), which we leveraged here with a custom database of marker gene sequences from ~729,000 microbial genomes and metagenome-assembled genomes (MAGs) able to detect and model strains of a total of 4,992 yet-to-be characterized species (Pasolli et al. 2019) called unknown SGBs (uSGBs, see Methods).

After running the StrainPhlAn-based pipeline on all samples, we obtained an overview of strain transmission patterns in our meta-cohort by constructing undirected networks based on the number of common strains between samples (Methods, Fig. 1A, Supplementary Figure 3). These networks confirmed that samples from the same FMT triad tend to share many strains and thus cluster together while they are only weakly connected to samples of other FMT triads (Fig. 1A, PERMANOVA by FMT triad on network coordinates, R2=[0.16-0.85], q<0.05 in 14 of the 24 datasets, Supplementary Table 4). To account for different absolute numbers of strains that can be analyzed via StrainPhlAn over samples, we defined the strain sharing rate metric as the number of shared strains (i.e. identical strains) between two samples divided by the number of shared species with available strain profiles. A t-distributed stochastic neighbor embedding (t-SNE) projection of strain sharing rates between samples also supported the same FMT triad clustering (Fig. 1B), and K-medoids clustering on strain sharing rates yielded clusters of higher purity with respect to FMT triad membership than beta diversity measures (Fig. 1C, Supplementary Figure 4, Supplementary Table 5, Methods). Strain-level metagenomics can thus accurately describe strain sharing events within FMT triads.

Post-FMT engraftment is variable and influenced by donor-recipient relationship

While donors and recipients generally displayed minimal gut microbiome strain sharing before FMT (5% median strain sharing rate), strain sharing rates were much higher between post-FMT and donor samples and between pre-FMT and post-FMT samples (median 57% and 60% respectively). The substantial increase in donor-recipient strain sharing after FMT is also significantly stronger than the convergence of beta-diversity (Mann-Whitney U test, q=1e-39 vs q=2e-17; median strain sharing pre-FMT/donor 5%, median strain sharing rate post-FMT/donor 57%, median Bray-Curtis similarity pre-FMT/donor 31%, median Bray-Curtis similarity post-FMT/donor 46%; permutation test p<1e-4, Methods, Fig. 1D, Supplementary Table 6) confirming that the strain-identity-based profiling approach better captures the microbiome remodeling induced by FMT compared to species-level beta diversity measures.

The extent of donor-recipient strain sharing after FMT varied substantially across datasets and FMT triads (Supplementary Figure 3). 58.4% of post-FMT samples shared more strains with corresponding donor samples than with their pre-FMT, and the difference in shared strains between donor/post-FMT samples and pre-FMT/post-FMT samples also differed substantially across FMT triads (median: -3; range: -96 to 75, Supplementary Table 7, Supplementary Figure 5), showing that some recipients post-FMT are heavily dominated either by donor-engrafted or retained strains.

We also found that pre-FMT recipients shared more strains with related (usually cohabitating) donors than with unrelated donors (i.e. donors that in the original studies were specified as genetically unrelated, or recruited through public advertisement or hospitals' cohorts), which is in line with previous observations of cohabitation being linked with microbial strain sharing (Brito et al. 2019; Korpela et al. 2018) (Fig. 1E, related vs unrelated permutation test, permutation test p<1e-4, median strain sharing rate difference=0.18). This also holds In cases in which only a subset of the donors and recipients were related (Fig. 1E, mixed vs unrelated, permutation test p<1e-4, median strain sharing rate difference=0.15). Although pre-FMT donor/recipient strain

sharing had no detectable effect on post-FMT strain sharing (Spearman's rho=0.03, p=0.65), we nevertheless accounted for these potential strain sharing biases at baseline both when assessing strain engraftment and for its potential influence on FMT success. To this end, we subtracted baseline strain sharing between donor and recipient pre-FMT when calculating strain engraftment rates (Methods, Supplementary Figure 7). Together, these data confirm that strain-level metagenomics allows to assess engraftment of the donor microbial strains after FMT, and that the extent of donor microbiome engraftment is variable and influenced by pre-FMT donor-recipient direct interaction.

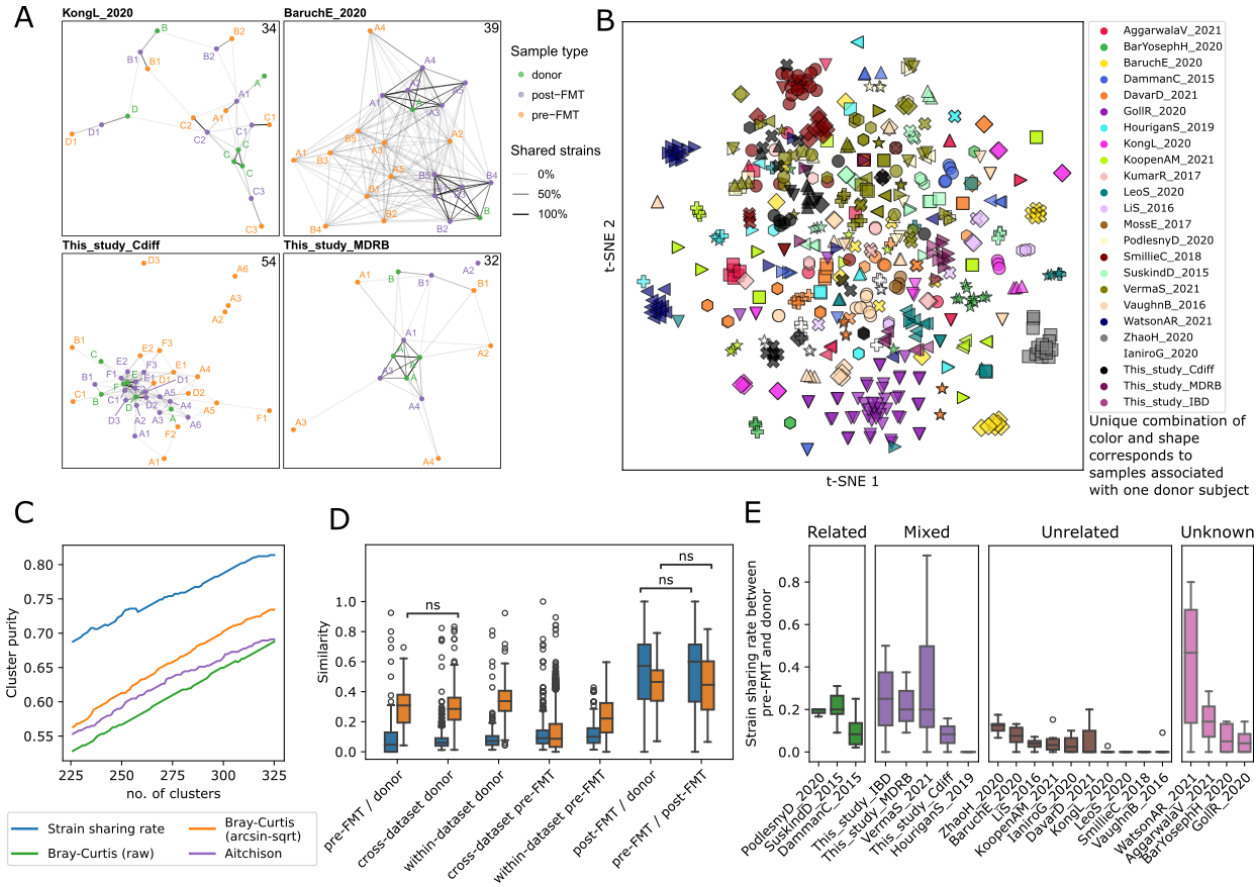


Figure 1. Overview of microbial strain sharing in FMT studies. (**A**) Strain sharing networks of the two novel FMT cohorts with *C. difficile* and MDRB colonization and of two previously published FMT cohorts (Kong et al. 2020; Baruch et al. 2021). Each node corresponds to a sample and is colored by its role in FMT triads (recipient pre-FMT sample, recipient post-FMT sample, and donor sample). The letters correspond to the donor subject and letter/number combinations indicate both associated donors (via the letter) as well as FMT instance membership (via the number) of pre-/post-FMT samples. Edge opacity is proportional to the number of shared strains (Methods) and scaled dataset-wise to the maximum number of shared strains (indicated in the top right corner). Supplementary Figure 3 reports the networks of all 24 datasets in the meta-cohort. Only edges corresponding to at least 2 shared strains are shown. (**B**) Ordination of samples from all cohorts based on strain sharing rates (t-SNE with perplexity=20). Samples from recipients (both pre-FMT and post-FMT) cluster around the corresponding donor samples while not exhibiting a strong cohort effect. See Supplementary Figure 8 for a PCoA ordination. (**C**) Strain-sharing enabled a more precise reconstruction of the true FMT triads compared to species-level beta-diversities (Supplementary Figure 4, Supplementary Table 5). We compare the K-medoids clustering purity (the fraction of samples belonging to a majority group of their assigned cluster) of FMT triads (starting at the true number of triads, i.e. 226) between strain sharing distances and on Bray-Curtis dissimilarities/Aitchison distances as a function of the number of clusters K. (**D**) Strain sharing rate and Bray-Curtis similarity between pairs of samples show that strain sharing rates increase much more after FMT compared to Bray-Curtis similarity (pre-FMT/donor median strain sharing rate: 5% and post-FMT/donor median starin sharing rate: 57%. pre-FMT/donor Bray-Curtis similarity: 0.31 and post-FMT/donor Bray-Curtis similarity: 0.46). Only pairs of samples with at least 5 species profiled at the strain level in both the donor and the recipient are displayed (see Methods). The significance was assessed by the Mann-Whitney U test and the p-values were FDR adjusted using the BH method. (**E**) Distribution of strain sharing rates between donor and corresponding recipient pre-FMT samples showing that donors share more strains with recipients pre-FMT when the individuals are “related” (same family/household or friends). Only pairs of samples with at least 5 species profiled at

the strain level in both the donor and the recipient were considered. The four categories describe the relationship between donor- and recipient individuals. Related: individuals are genetically related or cohabiting/friends; Unrelated: individuals are neither genetically related nor cohabiting/friends; Mixed: only some of the individuals are genetically-related or cohabiting/friends; Unknown: the relation of donors to recipients was not stated in the manuscript or metadata.

Donor strain engraftment tends to be higher in infectious disease patients pre-treated with antibiotics

To assess the extent of donor strain engraftment over strain retention in FMT recipients in different clinical settings, we compared the fraction of donor strains detectable in the post-FMT sample (fraction of donor strains) to the fraction of pre-FMT strains detectable in the post-FMT sample (fraction of retained strains; Fig. 2A). We found that patients who received antibiotics before FMT - as part of their therapy for underlying diseases or as pre-treatment before FMT - had a significantly higher fraction of donor strains compared to the fraction of retained strains (Wilcoxon signed-rank test, p=2e-16), while the opposite was true for recipients who did not receive antibiotics (Wilcoxon signed-rank test, p=1e-5, Fig. 2A). Accordingly, the median difference between the fraction of donor strains and the fraction of retained strains was also higher for recipients who received antibiotics compared to those who did not (difference of median difference=0.47, permutation test p<1e-4, Fig. 2A). Beside patients treated with antibiotics, we found that also recipients with infectious diseases had comparatively higher fractions of donor strains compared to the fraction of retained strains (Wilcoxon signed-rank test, p=8e-16), while the opposite was true for non-infectious disorders (Wilcoxon signed-rank test, p=6e-4), and also in that case the difference between the fraction of donor strains and the fraction of retained strains was higher for recipients with infectious disease compared to those with non-infectious disease (difference of median differences=-0.48, permutation test p=3e-4, Fig. 2A). The strain engraftment rate, which measures the fraction of donor strains that engrafted while taking into account the shared strains before FMT (Methods), is also (albeit non-significantly) higher for recipients receiving antibiotics or those with an infectious disease (permutation test antibiotics vs no antibiotics p=0.092, infectious vs non-infectious p=0.120, antibiotics & infectious disease vs no antibiotics & non-infectious disease p=0.052, Fig. 2B).

Indeed, patients with infectious disorders (rCDI and MDRB colonization in our set) often have a long history of repeated antibiotic courses and are pre-treated with specific antibiotics before FMT, while only two of the non-infectious disease cohorts we analyzed underwent treatment with antibiotics before FMT: SuskindD_2015 and BaruchE_2020. The SuskindD_2015 Crohn's disease cohort received rifaximin before FMT and exhibits strain sharing patterns similar to datasets with non-infectious disorders, which is consistent with previous results showing that rifaximin does not lead to substantial shifts in microbiome composition (Soldi et al. 2015). On the contrary, the BaruchE_2020 melanoma cohort, in which patients were pre-treated with neomycin and vancomycin, displayed strain sharing characteristics similar to cohorts with infectious diseases treated with antibiotics - possibly due to the disruptive effect of combined oral vancomycin (Jakobsson et al. 2010) and neomycin (Hu et al. 2016) treatment.

There were few exceptions to the stratification of datasets in high or low donor strain engraftment according to disease type and antibiotic use (Fig. 2A). The ZhaoH_2020 cohort, for

example, displayed exceptionally high strain sharing between donors and post-FMT recipients which is comparable with datasets of infectious diseases and pre-FMT antibiotics, even though FMT here is used as a potential treatment of a non-infectious disorder (Tourette syndrome) without the use of antibiotics. This cohort consists of children (8 years old on average) whose microbiome is likely less resistant to colonization from incoming strains as previous work has shown increased malleability of children's gut microbiomes upon external challenges such as probiotic strain supplementation (Lahti et al. 2013), dietary intervention (Burns et al. 2016) and lifestyle changes (Ruggles et al. 2018). Overall, the differential extent of strain engraftment and retention over cohorts can be well explained by antibiotic administration and the type of disease, and the few exceptions can still be partially reconciled by specific cohort characteristics.

These results show that both antibiotic administration and gastrointestinal infectious diseases are strongly associated with how many strains are taken up from the donor and how many strains are retained after FMT. However, since antibiotic treatment and infectious diseases are closely entangled variables in our meta-cohort, it is not possible with the data at hand to unravel their relative contribution to strain engraftment- and retention. Nonetheless, as both variables are known to lead to a decreased microbial diversity (Willmann et al. 2019; Zeng, Inohara, and Nuñez 2017; Chang et al. 2008) (Mann-Whitney U test comparing Shannon's diversity index, infectious vs. non-infectious p=3e-23, antibiotics vs. no antibiotics p=8e-14), and given that the substantially lower microbial alpha-diversity is likely making the recipient's gut more receptive to foreign strains from the donor, we hypothesize that these factors may have a combined effect on the overall engraftment.

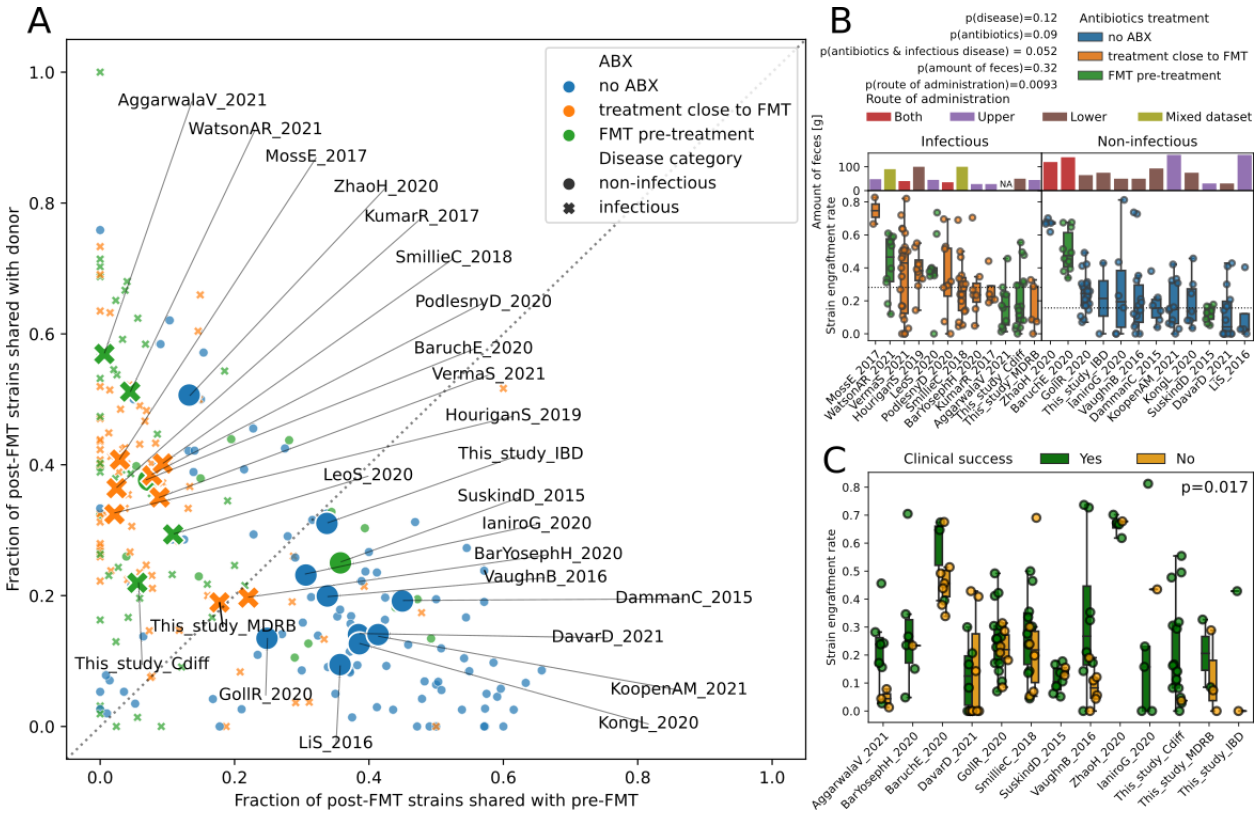


Figure 2. Variability of strain engraftment and retention across disease, antibiotic use, and clinical success. (A) Distribution of the fraction of donor strains and the fraction of retained strains present in the post-FMT samples for all FMT triads, showing a separation between higher fraction of donor strains and higher fraction of retained strains that associates with antibiotic administration and disease category. Small points represent individual FMT triads, large labeled marks represent dataset centroids, and the points/marks are colored by antibiotic administration. “FMT pre-treatment” indicates ABx treatment of recipients before FMT done specifically in support of the transplant whereas “treatment close to FMT” describes ABx intake before FMT without direct connection to the stool transplant. The shapes of the marks correspond to disease categories (infectious or not). **(B)** Variability of donor strain engraftment rate by disease category, antibiotics usage, and route and amount of administered feces highlighting the complex association between these variables and strain engraftment rates. The horizontal line is the median of per-dataset medians. The statistical tests are performed by permuting the variables associated with datasets (permutation test infectious vs. non-infectious p=0.120, antibiotics vs no antibiotics p=0.092, antibiotics & infectious disease vs no antibiotics & non-infectious disease p=0.052, amount of feces p=0.32, route of administration mixed vs. lower or upper p=0.0093, route of administration lower vs. upper p=0.93). **(C)** Association between clinical success of FMT and strain engraftment rates for the 13 studies in which the information on clinical success was available and for which at least 1 recipient was in each group. The definition of clinical success for each study is reported in Supplementary Table 1. Permutation tests with success labels permuted within each dataset pointed at an overall significant association of strain engraftment with clinical success (p=0.017), that was significant in only one of the datasets when considered individually (VaughnB_2016 with p=0.039).

Among the other clinical FMT variables we were able to consistently retrieve for most cohorts, we found an association of the route of stool administration with strain engraftment rates. Specifically, we found that a ‘mixed’ FMT protocol via both upper and lower gastrointestinal tract was strongly associated with higher engraftment rates (permutation test p=0.0093, Fig. 2B), suggesting that providing FMT to patients this way might be a way to enforce strain engraftment from the donor.

Association between strain engraftment and clinical success of FMT

Previous studies have suggested that strain engraftment might be associated with clinical success of FMT, but strong evidence is still lacking (Kootte et al. 2017; Rossen et al. 2015). We thus compared the strain engraftment rates across datasets with the clinical success of each FMT triad. We included cohorts for which a clinical outcome was defined and available in the metadata or in the publication for each FMT triad; moreover, we considered only those cohorts for which both the successful and the unsuccessful groups had at least one FMT triad (Supplementary Table 1). While we found that recipients experiencing clinical success showed significantly higher engraftment than the others for the VaughnB_2016 cohort (Mann-Whitney U-test, p=0.039), we did not detect a significant difference in any other cohort which have however limited statistical power (10.2 FMT triads per cohort on average, range=[2,20])). When jointly analyzing all datasets together we found a significant positive association between strain engraftment rate and clinical response to FMT (Fig. 2C, blocked permutation test p=0.017, see Methods) that was confirmed also via a random effects model meta-analysis (p=0.024, see Methods). The limited total sample size, the intrinsic approximations in the binary categorization of success of clinical treatments, and the heterogeneity of conditions tested limits our analyses, but the results overall suggest that higher microbiological engraftment is associated with a positive clinical outcome.

Post-FMT strain engraftment rates are phylum- and species-dependent

We then interrogated our dataset to understand whether microbial taxa coming from the donor have different likelihoods to engraft in the recipient. By leveraging the statistical power of our integrated cohorts, we calculated species-specific strain engraftment rates over all FMT triads for the 211 species for which the strain engraftment rate could be estimated with sufficient confidence (i.e. we could determine their engraftment in at least 15 FMT triads and 4 different datasets as defined by the strain engraftment rate, Methods, Fig. 3A, Supplementary Table 8). Overall, we found remarkable differences in engraftment rates over bacterial phyla (Kruskal-Wallis test, p=3e-11), with Bacteroidetes and Actinobacteria spp. (26 and 11 considered species respectively) displaying high average strain engraftment rates of 45%±12% and 46%±12%, respectively (Fig. 3B, Supplementary Table 8) compared to Firmicutes and Proteobacteria (averages 23%±14% and 29%±20%, respectively; Post-hoc Dunn tests, q<0.05, Figure 3B).

Six Firmicutes SGBs were among the set of the 20 most-engrafting species, including two understudied species with only a few isolate genomes available (*Dialister succinatiphilus, Phascolarctobacterium faecium*), two SGBs belonging to hitherto undescribed species (*Eubacterium* SGB6796, *Catenibacterium* SGB6783), and two others belonging to genera without cultured representatives (*Clostridia* SGB3957, *Ruminococcaceae* SGB15119). Of note, *Dialister succinatiphilus* - the SGB with the highest likelihood to engraft - and *Phascolarctobacterium faecium* are both members of the Negativicute class, characterized by a peculiar cell wall composition containing lipopolysaccharides, which results in a negative Gram stain (Rands, Brüssow, and Zdobnov 2019). As such, these Firmicutes species may have characteristics not completely in line with those of the most typical members of this phylum, possibly explaining their comparatively high engraftment rates. Among the top-engrafting non-Firmicutes species were several *Bacteroidales* members: *Bacteroides finegoldii* (60%)*, Bacteroides stercoris* (58%), *Alistipes putredinis* (54.2%), *Alistipes finegoldii*, *P. copri* clade A (Tett et al. 2021), and *Phocaeicola massiliensis* (62%). Furthermore the dysbiosis-associated species *Eggerthella lenta* (Gardiner et al. 2015) (strain engraftment rate=70%) and two bifidobacterial species (*B. bifidum, B. longum,* strain engraftment rates 62% and 57% respectively) engrafted regularly. In contrast, 19 out of the 20 least-engrafting species are Firmicutes and 16 of those are Clostridiales. *Acidaminoccus intestini, Streptococcus salivarius* and four other unnamed and uncharacterized Firmicutes species were never found to engraft in the FMT recipient despite being fairly prevalent in the donor (from 22% to 49% prevalence, Fig. 3A. Supplementary Table 8).

We also assessed the potential transmission of microbial Eukaryotic taxa, and found that only *Blastocystis* was detectable at enough coverage to infer transmission, as was expected (Beghini et al. 2017) (see Methods). Most FMT screening procedures exclude donors with *Blastocystis* in agreement with clinical consensus (Cammarota et al. 2019), and indeed this taxon has a lower prevalence in donors than in most Westernized populations (Asnicar et al. 2021; Tito et al. 2019): we detected five donors positive for *Blastocystis* in two cohorts (BarYosephH_2020, SmillieC_2018). No transmission to the recipient could be inferred as in four of the five cases *Blastocystis* was not detected at all in the recipient post-FMT and in the fifth case a different

subtype compared to the donor was present (Methods). Two putative retention events could in contrast be inferred in two recipients based on *Blastocystis* subtyping, confirming that this Eukaryotic microbe can be effectively profiled, in contrast to other eukaryotic species that can only occasionally be detected via metagenomics due to their low relative abundance. Of note, while *Blastocystis* is increasingly linked with favorable health conditions (Beghini et al. 2017), it does not seem to play a role in FMT possibly due to donor screening procedures, and it has been reported that even when transmission is occurring no associated symptoms were noted (Terveer et al. 2020).

Microbial engraftment is linked with bacterial phenotypic characteristics

We next assessed whether the taxonomic differences in strain engraftment we detected were associated with bacterial phenotypic properties. The more resistant Gram negative species tended to engraft better (Mahnert et al. 2019) (Mann-Whitney U-test q=3e-6, Supplementary Table 9), although exceptions such as the gram positive *Eggerthella lenta* displaying high engraftment rates were noted (Fig. 3A) and since most Firmicutes are Gram negative the association may be driven by characteristics of the Firmicutes phylum unrelated to cell wall structure. While spore formation and motility were not common traits among the species assessed, spore forming and motile species tended to display reduced engraftment rates (Mann-Whitney U-test q=0.007 and q=0.008, respectively, Fig. 3A). All of the above suggests that species engraftment is somewhat facilitated by specific microbial traits and that, in the future, more refined knowledge of phenotypic traits of under-characterized species could lead to mechanistic hypotheses underlying these associations.

Interestingly, bacterial species that were negatively associated with cardiometabolic health in the PREDICT 1 study (Asnicar et al. 2021) tended to engraft more frequently (Spearman's rho=0.36, p=4e-7), possibly due to more aggressive host colonization strategies or better fitness for a dysbiotic gut environment likely more affected by inflammation in FMT recipients. Although species prevalence in the gut of healthy individuals (assessed by surveying 9,120 gut metagenomic samples from 56 public studies, see Methods) did not significantly correlate with engraftment, the prevalence of bacteria in non-intestinal human body sites was associated with higher engraftment (prevalence at least 5% in any of the skin, vagina, oral or Airways samples, Mann-Whitney U-test p=8e-4) showing that species that are able to colonize non-intestinal environments have higher chance to engraft possibly due to increased oxygen tolerance. Together, these results show a remarkable variability in the overall engraftment rates among species in the human gut and suggest the possibility of screening donors to minimize the engraftment of species associated with unfavorable host conditions.

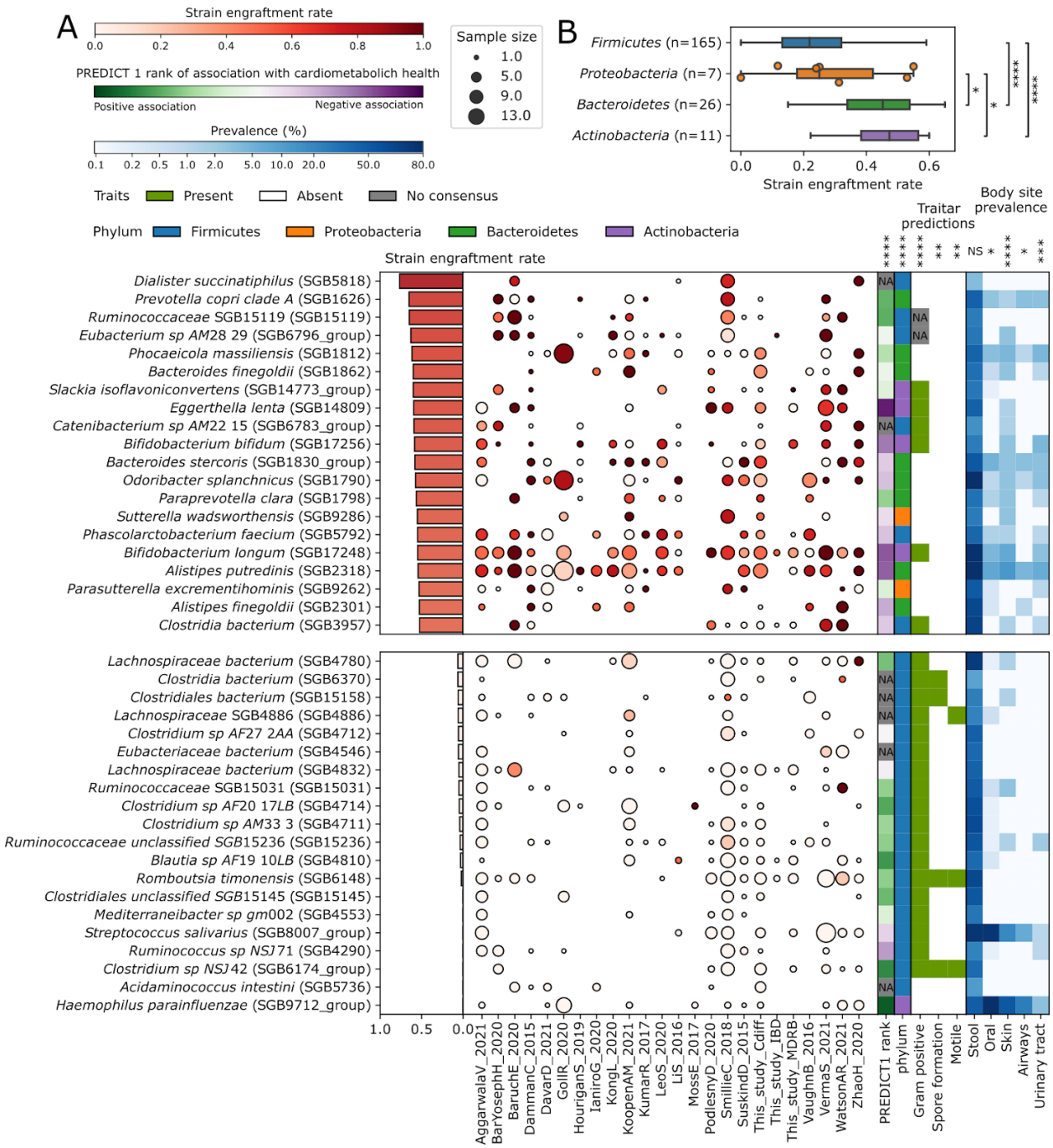


Figure 3. Bacterial strain engraftment rates and their associations with phenotypical properties, disease type, and ubiquity. (A) Overall and within-dataset strain engraftment rates and associations of species with predicted phenotypic properties (Weimann et al. 2016), cardiometabolic health (Asnicar et al. 2021), and prevalence (%) in different human body sites. Overall strain engraftment rate is computed over all triads. Out of 211 species assessed (Supplementary Table 8), the 20 species displaying highest and lowest engraftment rates are reported. Associations between continuous variables were tested with Spearman's rank correlation tests; those between binary categorical variables were tested with Mann-Whitney U test. The association with phylum was tested with the Kruskal-Wallis test. Tests were performed for all species including those not shown, and p-values were FDR-corrected using the BH method (Supplementary Table 9). Significance levels (*: q<0.05, **:

q<0.01, ***:q<0.001, ****:q<1e-4) are reported above each metadata column. Sample size is defined as the number of FMT triads in which the species could engraft as defined by the strain engraftment rate (see Methods). **(B)** Strain engraftment rates are significantly associated with bacterial phyla (Kruskal-Wallis test, p=3e-11; post-hoc Dunn tests FDR corrected using th BH metod, Firmicutes vs Bacteroides q=8.0e-9, Firmicutes vs. Actinobacteria q=3e-5, Proteobacteria vs. Bacteroidetes q=0.037, Proteobacteria vs. Actinobacteria q=0.037, the remaining pairs are non-significant, i.e. q>0.05). The Euryarchaeota and Verrucomicrobia phyla were omitted from the analysis as only one species in each of them were assessed in our analysis.

Machine learning models can predict post-FMT microbial composition

Understanding what are the donor and pre-FMT microbiome factors dictating the configuration of the microbial community in post-FMT patients could facilitate more advanced precision-medicine approaches for targeted modulation of the microbiome. Since strain engraftment from the donor only partially accounts for the post-FMT microbiome composition as strains can also persist or be taken up from the environment, we developed machine learning models to predict the microbiome composition post-FMT based on a set of quantitative features. Specifically, we trained Random Forest (RF) models to predict the presence or absence of species post-FMT using a total of 16 microbial and host features including taxonomy, bacterial abundances and alpha-diversity in pre-FMT and donor samples and bacterial prevalence in an unrelated cohort (see Methods). We found that these models are able to predict post-FMT species composition with an area under the receiver operating characteristic curve (AUROC) ranging from 0.781 to 0.911 (Avg.: 0.856, s.d.: 0.03) in a 5-fold cross-validation setting in which models for each dataset are repeatedly (5 times) trained on 80% of the data and evaluated on the remaining 20% (Fig. 4A, Methods). To understand how well these models generalize over different datasets and to exploit larger training sets, we also performed an analysis in which we predicted post-FMT species composition in a dataset by training the model on all the other datasets (Leave-One-Dataset-Out - LODO). In this setting, while AUROC values were lower than in the cross-validation setting, they were above 0.7 in all but one of the 24 cohorts (Avg.: 0.77 s.d.: 0.06). Finally, we have trained and evaluated - in a similar manner - Random Forest regression models to predict the post-FMT abundance of bacterial species (see Methods). These models were able to provide estimates of the abundance of bacterial species in the post-FMT microbiome that were correlated with those assessed by microbiome sequencing of the post-FMT samples (Spearman correlation 0.47, Fig. 4G, **Supplementary Figure 10**).

Analysis of the importance of each feature employed by the RF classifier highlighted that quantitative information on the abundance of the species in the donor and in the pre-FMT recipient as well as the overall prevalence are more relevant than characteristics such as the alpha-diversity of donor and recipient microbiomes, the beta-diversity between donor/recipient pairs or disease context (Fig. 4B). Surprisingly, single taxonomic features (i.e. the species or genus labels) proved not particularly important despite differences in strain engraftment rates over different clades (Fig. 3A). A more in-depth feature importance analysis evaluating all taxonomic levels simultaneously found that indeed the information content of the entire taxonomy is comparable to that of bacterial prevalence or abundance (**Supplementary Figure 11**) highlighting that feature importance scores of single taxonomic levels in the model were affected by the redundancy in the hierarchical taxonomic structure. Overall, we observed that

the composition of the post-FMT microbiome is generally predictable despite differences in cohort characteristics and host conditions and the presence of a species after the transplant is dictated primarily by the amount (or absence) in the donor and in the recipient as well as taxonomy and general prevalence.

Machine learning can predict the donor potential to shape the recipient microbiome and maximize specific microbiome features

To better understand to which extent the choice of the donor impacts the post-FMT gut microbiome composition, we set up a framework in which we substituted either the donor or the pre-FMT recipient of a triad with a randomly chosen donor or pre-FMT recipient from a different triad of the same dataset and then evaluated the decrease in AUROC upon this exchange. We found, as expected, a decrease in predictive performance upon exchange of either donors and recipients that was variable over cohorts (**Fig. 4C**). The performance decrease upon donor exchange was particularly pronounced in cohorts of infectious diseases patients pre-treated with antibiotics (Wilcoxon test p-values <0.001 when comparing infectious vs. non-infectious diseases and antibiotics pre-treatment vs. no pre-treatment, **Supplementary Figure 12**), consistent with a higher fraction of donor strains engrafting in the recipient in these conditions (**Fig. 3A**). The choice of donor has thus a higher influence on the post-FMT microbiome in infectious rather than non-infectious diseases.

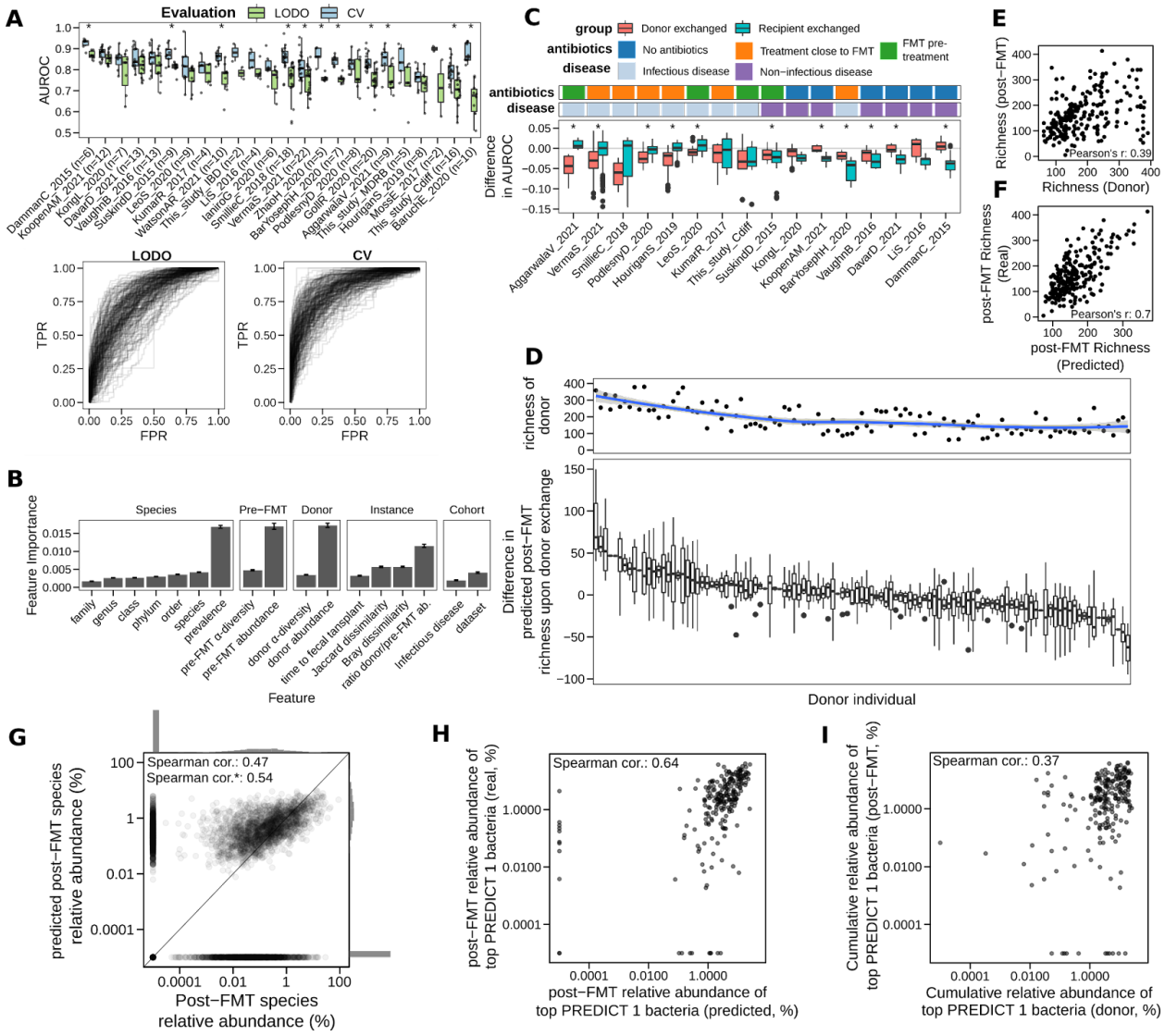


Figure 4. Random Forest models showcase the predictability of the post-FMT microbiome composition and the effect of different donors on the post-FMT microbial composition. (A) Post-FMT predictions of the presence or absence of species found in either the recipient or the donor. Each prediction refers to an FMT triad and corresponds to the AUROC of the predicted post-FMT species against the species actually detected in the post-FMT sample. We report both leave-one-dataset-out (LODO) and cross validation (CV) areas under the receiver operating curve (AUROC), and we calculate AUROCS per FMT triad. In the lower two panels, we show individual receiver operating curves. **(B)** The relative importance of each microbial feature used in the LODO model reported in (A) highlights that predictions are strongly dependent on the prevalence and pre-FMT/donor abundance of the species. While the feature importance of *individual* taxonomic levels shown to be relatively low, the combined information content of the taxonomy is comparable to that of bacterial abundances and prevalences (**Supplementary Figure 11**). **(C)** Distribution of the changes in AUROC values for the LODO models of panel (A) when the real donor/recipient pairing is modified by exchanging either donor or recipient with randomly chosen other donors or recipients from the same cohort. Only unique donor individuals are considered, and only datasets with more than two donors are reported. We removed 12 outlier data points out of 127 in the LeoS_2020 cohort for visualization purposes. **(D)** Top panel: distribution of species richness of FMT donors. Blue line is a LOESS fit. Bottom panel: Difference in post-FMT species richness upon donor exchange with respect to the

predicted post-FMT species richness of the unaltered triad (i.e. with the actual donor). **(E)** Donor species richness is positively correlated with recipient's post-FMT species richness (Pearson's test, r=0.39, p=2e-8). **(F)** Predicted post-FMT species richness is strongly correlated with the actual post-FMT richness (Pearson's test, r=0.7, p=1e-13). Bacterial richness is defined as the number of bacterial species detected. In order to circumvent the need to define an optimal decision threshold in our models, richness of predicted data is calculated as the sum of prediction probabilities over all species. **(G)** A Random Forest regression model is somewhat able to predict bacterial abundances in the post-FMT microbiome. As in the evaluation of the classifier, we did not evaluate species that are absent in both pre-FMT and donor individuals. The asterisk designates the Spearman correlation when omitting truly absent species predicted to be absent. For scatterplots of individual datasets see **Supplementary Figure 10**. **(H)** The cumulative abundance of the top 20% PREDICT 1 bacteria post-FMT can be predicted fairly accurately using the RF regression model. **(I)** The donor abundance is a worse predictor of the cumulative abundance of the top 20% PREDICT 1 bacteria than the RF regression model.

Finally, we investigated whether ML models have the potential to pinpoint particularly suitable donor individuals for improving microbiome features in recipients based on their individual microbiomes. Using the same simulation setting described in the previous paragraph, we first evaluated the donor effect in modulating post-FMT richness, a microbiome feature believed to be important for community stability and resilience (Lozupone et al. 2012) and also previously associated with clinical success in the context of ulcerative colitis (Rossen et al. 2015). Upon exchange of donors in triads, we found that some donors led to a consistent increase in predicted post-FMT richness compared to the original donor, whereas others led to a decreased predicted post-FMT richness (Fig. 4D). Ranking the donors within each dataset by their richness and comparing the top donors to bottom donors (one per dataset per group) shows that donor richness is much higher in donors predicted to induce high post-FMT richness (Fig. 4D, **Supplementary Figure 13**). Importantly such predictions were much more accurate than donor richness alone (Pearson’s r=0.735 vs r=0.41, p=1e-13 vs p=2e-8; Fig. 4E,F). In addition to richness, we have analogously exploited the framework to pinpoint donors that are predicted to maximize the cumulative prevalence of other pre-defined groups of bacteria, such as Firmicutes, species typically found in the oral cavity, or the set of taxa found positively linked with cardiometabolic health in the PREDICT 1 study (Asnicar et al. 2021) (**Supplementary Figure 14**). In all these situations, our models are accurate in predicting a given trait and are more predictive than using the quantitative microbial features of the donor as a direct estimator. We finally evaluated the regression model to predict the cumulative abundance of the same clades, finding that the model can predict the cumulative abundance of bacteria positively linked with cardiometabolic health better than the donor abundances alone (**Figure 4H, Figure 4I**), although the results are variable across different clades (**Supplementary Figure 15**) Taken together, these results illustrate that our machine learning framework provides predictive models of the composition of the post-FMT microbial communities that can be useful in the choice of a suitable donor given a specific post-FMT microbiome feature of clinical relevance, such as post-FMT bacterial richness.

Conclusion

In our meta-analysis of 1,371 metagenomic samples from 24 studies investigating FMT in different diseases, we built on improved strain-level profiling approaches to assess the extent of

microbial strain engraftment and retention upon FMT in relation to several clinical covariates. Donor strain engraftment varied substantially across cohorts, and we identified several parameters that influence its dynamics. Among them, we found that patients who received antibiotics before FMT (as part of their clinical therapy or as pre-conditioning to promote microbial colonization) and those with infectious diseases displayed a tendency for higher strain engraftment rates compared with those not treated with antibiotics or affected by noncommunicable diseases. This finding could explain the considerable discrepancies in the effectiveness of FMT between rCDI and chronic/non-infectious disorders (De Groot et al. 2017; Rossen et al. 2015; Kootte et al. 2017), and might pave the way to include antibiotic pre-conditioning in FMT working protocols to prime donor microbiome engraftment, even though the potential side effects of antibiotic treatments for non-infectious diseases (Ferri et al. 2017; Dancer 2004) will need to be properly considered with respect to the specific disease the FMT is proposed for. Nonetheless, while acknowledging the limitations induced by the small sample sizes of single cohorts and disease heterogeneity, the comparatively higher strain engraftment rates we identified through meta-analysis in patients who achieved clinical success upon FMT do support the recommendation for including antibiotic pre-conditioning in FMT protocols.

In the attempt to unravel the fitness of different microbial species in the FMT procedure, we found differential strain engraftment likelihoods associated with microbial taxonomy and phenotypic properties. Some species with immune modulation potential (e.g. Bifidobacteria spp.), Gram-negative bacteria, and some species with proinflammatory potential (*Bacteroides eggerthii, Eggerthella lenta*) displayed highest strain engraftment likelihoods, while several Firmicutes, including butyrate-producing bacteria, showed a low ability to engraft. As FMT is performed in patients with diseases, it remains to be elucidated whether the general higher engraftment rates of proinflammatory bacteria reflect intrinsic phenotypic traits that favor transmission and colonization in a new environment or rather a better fitness for an inflamed and dysbiotic environment (Browne et al. 2017). As FMT for healthy individuals poses ethical concerns, preclinical animal models could be employed in the future to address these questions. Additionally, the implementation of targeted, fine-tuned bacterial consortia as an alternative to traditional FMT would avoid the transfer of detrimental bacteria (including pathogens that could occasionally remain undetected after the screening of donor feces (Zellmer et al. 2020)) but it is still unclear whether such consortia can represent a suitable alternative to the complexity and diversity of FMT (Xiao et al. 2020; M. Li et al. 2015; Y. Li and Honda 2021).

Finally, we developed a machine learning model that was able to accurately predict the composition of the recipient's microbiome after FMT. Given that we trained this model on different datasets and over different diseases, it performed well in comparison to a previous, single-cohort study which also predicted the post-FMT presence/absence (Smillie et al. 2018). The model we trained can predict the donors with the highest potential to shape the recipient's microbial composition towards specific features such as increased species richness, a decreased proteobacterial richness or an increased cumulative abundance of bacteria associated with good cardiometabolic health. Together with a better identification of disease- and health-associated microbial features for each specific disease, this approach could lead to the development of personalized therapeutic FMT strategies based on the selection of the

recipient-specific optimal donor within a set of available donors, or the ad-hoc assembly of strain consortia.

Overall, using a sound strain tracking methodology, our meta-analysis identified the characteristics of donor strain engraftment upon FMT, described different dynamics associated with different disease types, supported the association between microbiome engraftment and clinical response, and proved the suitability of machine-learning models for donor selection. As more and larger datasets on the same diseases become available, this framework will allow to further stratify FMT strain dynamics of specific conditions and guide choices in clinical practice to exploit the full potential of FMT.

METHODS

Metagenomic dataset search strategy and selection

We systematically searched PubMed, Scopus and ISI Web of Knowledge as of 08/02/2021 for potentially eligible studies using the following search string: ((faecal microbiota suspension) OR (fecal microbiota suspension) OR (faecal microbiota transplant*) OR (fecal microbiota transplant*) OR (faecal microbiota donation) OR (fecal microbiota donation) OR (faecal microbiota transfer) OR (fecal microbiota transfer) OR (faecal microbiota infusion) OR (fecal microbiota infusion) OR (faecal microbial suspension) OR (fecal microbial suspension) OR (faecal microbial transplant*) OR (fecal microbial transplant*) OR (faecal microbial donation) OR (fecal microbial donation) OR (faecal microbial transfer) OR (fecal microbial transfer) OR (faecal microbial infusion) OR (fecal microbial infusion) OR (faecal suspension) OR (fecal suspension) OR (faecal transplant*) OR (fecal transplant*) OR (faecal donation) OR (fecal donation) OR (faecal transfer) OR (fecal transfer) OR (faecal infusion) OR (fecal infusion) OR (bacteriotherapy) OR (stool transplant*) OR (stool donation) OR (stool transfer) OR (stool infusion) OR (FMT)) AND ((Metagenom*) OR (shotgun) OR (engraft*) OR (whole genom*) OR (transkingdom) OR (WGS)). In addition, we manually searched the bibliographies of papers of interest to provide additional references. When needed, we contacted the authors to obtain additional data, metadata, or clarification of study methods.

We considered as eligible all original studies with the following characteristics: 1) human subjects of any age were treated with non-autologous FMT; 2) shotgun metagenomic analysis of donor feces and of recipient feces (before and after treatment) was performed. We excluded studies in which patients were treated with other therapeutic modulators of the gut microbiota (i.e. microbial consortia, antibiotics or probiotics), if less than 3 recipients were enrolled, and if raw sequencing data or metadata were not available or incomplete. In the case of randomized controlled trials that used autologous FMTs as placebo, we included only the patients treated with non-autologous FMT. If studies used stools from mixed donors for FMT (multidonor FMT), they were included only if sequencing of multidonor stool batches were available. Finally, we excluded animal model studies or non-original studies (reviews, meta-analyses, editorials, etc.). The eligibility of each study was assessed independently by two reviewers (NK and SP), and any disagreements were resolved by the opinion of a third reviewer (GI).

Sequencing data files and metadata were downloaded from public repositories as indicated in the original publications. If data were not publicly available, we contacted authors asking to provide it through private correspondence.

Metadata extraction and curation

Metadata extraction was carried out independently by two reviewers (NK and SP), using a data collection form. Discrepancies between the two reviewers were resolved by the opinion of a third investigator (GI). The following data were extracted from each study if available: authors' names, publication year, Bioproject Accession code, sequencing depth, study location, number of total samples, study disease, number of recipients and donors, donor type (i.e. whether donor individuals were related to the recipient, either family/household members or through friendship,

or whether they were unrelated), use of antibiotics before FMT, characteristics of infused feces (grams, volumes, use of frozen/fresh material), routes and number of infusions, follow-up, and clinical and microbiological outcomes.

Newly-collected metagenomic datasets

Three cohorts were newly-collected and sequenced in the context of this study. A first cohort (This_study_Cdiff) included 16 adult subjects with recurrent *C. difficile* infection who were treated with a single fecal transplant from 6 different donors, and their stool was collected just before FMT and at different time points (7, 15, 30, 60, 180 and 240 days) after FMT. FMT was performed with frozen faecal material. Donor selection and manipulation of faecal material were performed following international guidelines (Cammarota et al. 2019). All patients underwent FMT by colonoscopy, after bowel lavage and a 3-day vancomycin regimen, as previously described (Gianluca Ianiro, Murri, et al. 2019). A total of 94 stool samples were sequenced. A second cohort (This_study_IBD) included two pediatric patients with IBD who received single FMT from one of two related or unrelated donors. Stool samples were collected and sequenced at follow-up visits up to 30 days after treatment, yielding 8 metagenomic samples. A third cohort (This_study_MDRB) included five pediatric patients with large bowel colonization with Multi-Drug Resistant Bacteria (MDRB) and either acute leukemia (n=4 patients) or severe combined immunodeficiency (n=1 subject). Patients underwent single (n=4 subjects) or sequential (n=1 subjects, n=2 procedures) fecal transplant from one of two donors. Stool samples were collected and sequenced at follow-up visits up to 30 days after FMT (n=13 metagenomic samples in total). In both pediatric cohorts FMT was performed as previously described (Quagliariello et al. 2020). Consistent metadata of all 115 samples newly-collected in this study can be found in Supplementary Table 2.

Samples were stored at -80°C between 1 and 12 months before DNA extraction, which was performed using the DNeasy PowerSoil Pro Kit (QIAGEN, Germany) according to the manufacturer's procedures. DNA concentration was measured with Qubit (Thermo Fisher Scientific, USA), and DNA was then stored at -20°C. Sequencing libraries were prepared using the Illumina® DNA Prep (M) Tagmentation kit (Illumina, California, USA) following the manufacturer's guidelines. Sequencing was performed on the Illumina NovaSeq 6000 platform at a target sequencing depth of 7.5Gbp following manufacturer's protocols.

Newly-generated shotgun metagenomic sequences were pre-processed and quality controlled using the pipeline available at https://github.com/SegataLab/preprocessing. Shortly, reads were quality-controlled and those of low quality (quality score <Q20), fragmented (<75 bp), and with >2 ambiguous nucleotides were removed with Trim Galore (v0.6.6). Contaminant and host DNA was identified with Bowtie2 (v2.3.4.3) (Langmead and Salzberg 2012) using the parameter -sensitive-local, allowing confident removal of the phiX 174 Illumina spike-in and human reads (hg19 human genome release). Remaining high-quality reads were sorted and split to create forward, reverse and unpaired reads output files for each metagenome. Average sequencing depth after preprocessing was 7.3 s.d. 4.9 Gbp. Sequencing depth of each sample can be found in Supplementary Table 2.

Building the expanded SGB database

Species-level genome bins (SGBs) are clusters of microbial genomes and metagenome-assembled genomes (MAGs) defined to have no more than 5% pairwise genetic divergence (Pasolli et al. 2019). SGBs can contain taxonomically labeled microbial genomes from isolate sequencing (kSGBs) or can lack taxonomic contextualization from isolate sequencing (uSGBs). In this work we first extended the SGB database and then employed it to detect and profile the taxa present in metagenomes belonging to any kSGB or uSGB at species- and strain-level resolution.

The custom extended database was built starting from the 154,723 MAGs and 80,990 isolate sequencing genomes from Pasolli *et al* (Pasolli et al. 2019) and further expanded with 616,805 MAGs from different human body sites, animal hosts, and other environments, together with 155,767 isolate genomes in the NCBI Genbank database (Benson et al. 2012) available as of November 2020. We executed CheckM (version 1.1.4) (Parks et al. 2015) on the 1,008,148 genomes, filtering those with completeness below 50% or contamination above 5% to ensure high quality. Next, we minimized the redundancy among genomes by computing MASH distances (Ondov et al. 2016) on the quality-controlled sequences, and de-replicating sequences at 99,99% genetic identity. A total of 729,195 genomes [560,084 MAGs (Supplementary Table 11) and 169,111 reference genomes) were kept in the extended database..

Assembled contigs longer than >1,500 nucleotides were binned into MAGs with MetaBAT2 (Kang et al. 2019). Quality control was performed with CheckM (version 1.1.3) (Parks et al. 2015), and only medium- and high-quality genomes (completeness ≥50% and contamination ≤5%) were included in the custom extended database. Prokka (versions 1.12 and 1.13) (Seemann 2014) was used to annotate the genomes' open reading frames. Coding sequences (CDS) were assigned to a UniRef90 cluster (Suzek et al. 2015) by performing a Diamond search (version 0.9.24) (Buchfink, Xie, and Huson 2015) of the CDS on the UniRef90 database (version 201906) and assigning a UniRef90 identifier when the mean sequence identity to the centroid sequence was >90% and covered >80% of the centroid sequence. Sequences that could not be assigned to any UniRef90 cluster following this procedure were *de novo* clustered with MMseqs2 (Steinegger and Söding 2017) to SGBs following the Uniclust90 criteria (Mirdita et al. 2017).

Genomes were then clustered into species-level genome bins (SGBs) spanning ≤5% genetic diversity, and SGBs to genus-level genome bins (GGBs; 15% genetic diversity) and family-level genome bins (FGBs; 30% genetic diversity), following the procedure described in Pasolli *et al* (Pasolli et al. 2019). 'phylophlan_metagenomic', a subroutine of PhyloPhlAn 3 (Asnicar et al. 2020) that applies Mash (Ondov et al. 2016) to compute the whole-genome average nucleotide identity among genomes was used to assign MAGs to SGBs. When there was no SGB in the database with ≤5% genetic distance to a certain genome, new SGBs were created based on the average linkage assignment followed by hierarchical clustering (with a 5% genetic distance among genomes in the dendrogram allowed). Similarly, when no GGBs or FGBs below the

genetic distance threshold existed, the same was done to assign SGBs to novel GGBs and FGBs.

Definition of kSGBs and uSGBs and taxonomic assignment

SGBs containing at least one reference genome (kSGBs) were assigned the same species-level taxonomy of the reference genomes included in the kSGB following a majority rule. SGBs containing no reference genomes (uSGBs) were given the taxonomic annotation of its corresponding GGB (up to the genus level) if this included reference genomes, and of its FGB (up to the family level) if that included reference genomes. Alternatively, if no reference genomes were contained in the FGB, a phylum was assigned based on the majority rule of up to 100 closest reference genomes to the MAGs in the SGB as determined by 'phylophlan_metagenomic'. Taxonomic assignment of SGBs profiled in this study can be found in Supplementary Table 3.

Species-level profiling of metagenomic samples

Species-level profiling was performed on samples sequenced to a depth higher than 1 Gbp (n=1419; 100 samples being excluded from downstream analyses) using MetaPhlAn 3 (Beghini et al. 2020) with default parameters and the custom extended SGB database. uSGBs with less than 5 MAGs were discarded, as they are likely the result of assembly artifacts or chimeric sequences. Next, SGB core genes were defined as ORFs in a UniRef90 family or in a *de novo* clustered gene family (based on the Uniclust90 clustering procedure (Mirdita et al. 2017)) that were detected in at least half of the genomes of the SGB. Core genes were further filtered by selecting the highest threshold that allowed obtaining at least 800 core genes, and split into fragments of 150 nt. Core gene fragments were then aligned against the genomes of all SGBs using Bowtie2 (version 2.3.5.1; --sensitive option) (Langmead and Salzberg 2012). Marker genes of a SGB were defined as core genes with no fragments found in 99% of the genomes of any other SGB. When less than 10 marker genes were found for a SGB, conflicts were defined as occurrences of >200 of its core genes in more than 1% of the genomes of another SGB. All conflicts for each SGB were then retrieved to generate conflict graphs. Conflict graphs were processed iteratively, and SGBs were merged for each conflict to both minimize the number of merged SGBs and maximize the number of markers. Finally, a maximum of 200 marker genes were selected for each SGB, prioritizing first their uniqueness and next the larger sizes. SGBs with less than 10 markers were discarded at this point. Merged SGBs (*SGB_group*) profiled in this study can be found in Supplementary Table 3. The resulting 5.1M marker genes (average: 189 ± 34.25 sd marker genes/SGB) were used as a new reference database for MetaPhlAn 3 (species-level profiling) and StrainPhlAn 3 (strain-level profiling).

Strain-level profiling of metagenomic samples

Strain profiling was performed with a modified version of StrainPhlAn 3 (Beghini et al. 2020) using the custom SGB marker database described above. We modified the StrainPhlAn code to change the sample and marker filtering behavior. A sample was kept as long as it had at least 20 markers (parameter --sample_with_n_markers), a marker was kept as long as it was present in 50% of the samples (parameter --marker_in_n_samples) and a sample was kept if it had 10

markers after filtering of the markers (parameter --sample_with_n_markers_after_filt). All 2,576 SGBs profiled by MetaPhlAn were initially considered for the strain-level profiling. To improve stability of strain sharing detection and to more confidently define strain identity, we added 4,443 human gut metagenomic samples from 962 individuals older than six years from Westernized populations that were sampled longitudinally, obtained from 18 datasets available in the most up to date version of the curatedMetagenomicData R package (Pasolli et al. 2017) (Supplementary Table 12). For each individual, two samples being <= 6 months apart were selected. When more than 2 timepoints close in time were available, we selected the pair that maximized their chance to pass the filtering steps in StrainPhlAn, i.e. maximize the lower estimated coverage of the two. In case of ties, we took those with higher coverage. Coverage of an SGB was estimated in a sample as (sample sequencing depth) * (rel. abundance of the SGB) / (estimated genome length), with estimated genome length being extracted from the MetaPhlAn 3 enlarged database. For kSGBs this is determined using only the genome lengths of the reference genomes in the kSGB, while for uSGBs 7% is added to the average genome length (estimated to be the average difference between the genome sizes of reference genomes and MAGs within the same SGB).

We included in the strain analysis samples as primary (parameter --samples) if they had estimated coverage at least 2x of a given SGB genome, otherwise they were added as secondary samples (parameter --secondary_samples). 1033 SGBs that were detected in at least 20 primary samples were strain-profiled and thus considered in the strain-based analysis. In order to exclude strains likely coming from food sources, we included 216 MAGs in 19 SGBs (Supplementary Table 13) coming from food (Pasolli et al. 2020) and used them in the StrainPhlAn profiling with the --secondary_references parameters. Samples that had StrainPhlAn mutation rates < 0.0015 to any food MAG were discarded. SGBs in which we would discard more than 20% of the samples - constituting strains regularly found in food - were dropped altogether (n=3 SGBs: Bifidobacterium animalis SGB17278, Lactobacillus acidophilus SGB7044, Streptococcus thermophilus SGB8002)

Inference of strain transmission events

We obtained phylogenetic distances between strains as their leaf-to-leaf branch lengths along the trees produced by StrainPhlAn (built on marker genes alignments, retaining positions with at least 1% variability), normalized by dividing them by the median phylogenetic distance. To infer strain identity, we defined and applied operational species-specific definitions by identifying the threshold that optimally separated phylogenetic distance distributions of strains of a given species in the same individual sampled at two timepoints (*same strain*), to that in unrelated individuals (*different strains*). For all strain-profiled SGBs, we determined the phylogenetic distance threshold that best separates strains from the same subject (different post-FMT time-points of the same recipient or different samples of the same donor subject or different additional longitudinal samples of the same subject, always less than 6 months apart) from those of unrelated subjects with no possibility of direct transmission (subjects in different datasets) in the datasets we used in this study. For SGBs for which at least 50 same-individual and 50 unrelated comparisons were available, we determined the threshold that maximizes Youden’s index (defined as sensitivity + specificity - 1). If the calculated threshold is greater than

the 5th percentile of the distribution of subjects in different datasets, we adjusted the threshold to the 5th percentile as a bound on the false discovery rate. For SGBs for which less than 50 same-individual comparisons or less than 50 unrelated comparisons were available (in which optimal thresholds cannot reliably be estimated), we used the 3th percentile of the distribution of subjects in different datasets, which corresponded to the median of all the calculated percentiles in (Valles-Colomer et al., n.d.). All SGB-specific phylogenetic distance thresholds can be found in Supplementary Table 3. Finally, we defined strain identity for pairs of strains when their pairwise genetic distance fell below the SGB-specific thresholds.

Sample filtering

Strain-level profiling allows identification of miss-labeled samples (Podlesny and Fricke 2021). We identified and excluded post-FMT samples (n=21 out of 1419) that did not share any strain with neither their corresponding pre-FMT sample nor the donor's sample - something unexpected due to the high temporal stability of the gut microbiome (Lloyd-Price et al. 2017; Truong et al. 2017; Albanese and Donati 2017; Beghini et al. 2021) and thus potential cases of sample miss-labeling. We also identified outliers with > 20 shared strains between pre-FMT and donor samples while being from two supposedly unrelated individuals (n=2 cases, Supplementary Figure 16), most probably not representing true recipient-donor pairs. The third outlier with > 20 shared strains was coming from a dataset using both related and unrelated donors, but the Bray-Curtis dissimilarity between the donor and pre-FMT samples was close to zero (bc=0.019) suggesting they are the same biological sample and confirming the mislabeling. Finally, we excluded the ZouM_2019 cohort from the analysis because strain sharing sample clustering was heavily discordant from the grouping of FMT triads according to the metadata (Supplementary Figure 3) and ZouM_2019 was the only dataset with a median of only 1 strain shared between post-FMT and donor-samples (Supplementary Figure 17), further suggesting systematic errors in the metadata.

Inferring donor subject grouping

In three cohorts (BarYosephH_2020, DammanC_2015 and LeoS_2020) some donors provided stool material to multiple recipients, but we could not solve which donor samples were transferred to which patients neither from the metadata nor through private correspondence with the authors. For this reason we inferred grouping of donor samples into subjects using strain sharing: donor samples sharing more than 15 strains were grouped into one subject. This threshold allows confident matching of samples from the same subject, since unrelated samples virtually never share more than 3 strains, while longitudinal post-FMT samples frequently share more than 15(Supplementary Figure 18) as also reported elsewhere (Truong et al. 2017). Indeed, in these three datasets samples from the same assigned donor always shared at least 15 strains, while this was never observed among samples from different donor individuals.

Inferring donor-recipient matching

Donor-recipient matching was unavailable for DammanC_2015 and we were unable to obtain it through private correspondence with the authors. However, as at least one post-FMT sample of a recipient always shared ≥8 strains with one donor subject, while no post-FMT samples of the

same recipient shared ≥8 strains with any other donor subject (Supplementary Figure 19), we used the criterion of sharing ≥8 strains to infer donor-recipient matching in the dataset.

Definition of FMT triads

We considered only complete FMT triads, i.e. sets of at least one sample from the recipient pre-FMT, at least one from the donor, and at least one from the recipient post-FMT. In case of multiple sequential FMT transplants, we included only the first one. In case of multiple pre-FMT samples, we used the one collected closest to the FMT. When multiple donor samples were available, we picked one randomly (since donor samples from the same individual are reasonably stable in terms of species-level composition and strain identity, Supplementary Figure 20). Finally, when multiple post-FMT samples were available, we picked the one closest to 30 days post-FMT, which is the value that minimizes the sum of absolute deviations of timepoints (Supplementary Figure 1). This was done regardless of the total number of transplants received.

Assessing strain sharing, retention and engraftment

We defined *strain sharing rates* as the total number of shared strains between two samples divided by the number of species profiled by StrainPhlAn in common between the two samples. To quantify the fraction of post-FMT strains that were already present pre-FMT or that are shared with the donor, we used the fraction of post-FMT strains shared with pre-FMT (shared strains between post-FMT and pre-FMT divided by the number of strains profiled at post-FMT, i.e. the *fraction of retained strains*) and the fraction of post-FMT strains shared with the donor (shared strains between post-FMT and donor divided by the number of strains profiled at post-FMT, i.e. the *fraction of donor strains*).

Next, we determined the number of *engrafted strains* as the number of shared strains between post-FMT and the donor excluding the shared strains between pre-FMT and the donor samples. The number of *strains that could engraft* is defined as the number of cases in which StrainPhlAn can profile the strain in the donor sample while excluding both the shared strains between pre-FMT and donor and the cases where the species is present in the post-FMT, but no strain is profiled by StrainPhlAn (as in these cases it is not possible to determine the strain identity). Finally the *strain engraftment rate* was defined as the number of *engrafted strains* divided by the number of *strains that could engraft*. This measure was computed for each FMT triad (by aggregating over species) and also for each species (by aggregating over FMT triads). In the latter case, only species with at least 15 FMT triads from at least 4 datasets in which the *strain could engraft* were included in the analyses.

Visualization of strain sharing in cohorts

In order to analyze strain sharing in datasets, we computed networks as well as t-SNE plots based on the number of shared strains between pairs of samples. Within-vs-between FMT instance clustering of samples was assessed using PERMANOVA on euclidean distances between samples extracted from the networks. Unsupervised networks were visualized using the igraph package in R (version 1.2.6) (Csardi, Nepusz, and Others 2006) with the

Fruchterman-Reingold layout algorithm with squared edge weights, with edges being the number of shared strains and nodes representing samples. Only edges with more than 1 shared strain are shown. The t-SNE plot was generated using the scikit-learn package (Pedregosa et al. 2011) in Python (version 0.24.2) with perplexity = 20 and remaining parameters left default.

Comparing strain- and species-level measures for FMT triad clustering

In order to compare how well strain- and species-level information allow clustering of samples from the same FMT triads, we performed K-medoids clustering with Partitioning around Medoids (PAM) algorithm implemented in scikit-learn-extra Python package using strain sharing rates dissimilarities (defined as 1 - strain sharing rate) as compared to Aitchison distance and Bray-Curtis dissimilarity (on untransformed data, after arcsine square root transformation, and after logit transformation). In case of Aitchison distance, the zeros were replaced by the per-taxon minimal non-zero abundance and in case of logit transformation the zeros were replaced by the half of the minimal non-zero abundance globally. Clustering quality was assessed using the clustering purity, which is defined as the fraction of samples that belong to the majority class in their respective cluster. When calculating the purity of FMT triads with shared donor samples (donor samples having been administered to several recipients), we treated the single sample as multiple samples, each belonging to one of the associated FMT triads. This way the association was considered pure if the donor sample was clustered with any of the triads it belongs to.

Prevalence of the SGBs across different human body sites

We profiled 9,990 healthy human microbiome samples from 59 datasets spanning different body sites (airways, gastrointestinal tract, oral, skin and urogenital tract; Supplementary Table 12) using MetaPhlAn 3 (Beghini et al. 2020) with default parameters and the custom SGB database (see above). Only individuals older than 3 years and from cohorts involving industrialized non-rural populations were considered. Age, lifestyle and disease status were considered as reported in curatedMetagenomicData (Pasolli et al. 2017).

Annotation of SGB phenotypic traits

SGB phenotypes were predicted using Traitar (version 1.1.12) (Weimann et al. 2016) on the genes present in 50% of genomes available for each SGB in the custom SGB database. Only annotations for which the *phypat* and the *phypat+PGL* classifiers predictions were in agreement were used.

Statistical analysis

Total strain sharing variance explained by FMT triad membership assessed by PERMANOVA with the adonis function in vegan (version 2.5-7) (Oksanen et al. 2020). Differences between two groups were assessed with two-sided Mann Whitney U tests or Wilcoxon signed-rank tests if groups were paired. To compare differences between median strain sharing or engraftment measures in two groups of datasets against the null distribution, permutation tests were applied by randomly permuting the assignments between labels and dataset identifiers 9999 times. To compare median strain sharing rates of clinically successful FMT triads against failed ones, a

permutation test was applied by randomly permuting the success labels within each dataset 9999 times. Complementarily, we fitted a linear mixed model predicting strain engraftment rate with the clinical success as an indicator variable and the dataset identifier as a random effect using the R package lme4 (Bates et al. 2015); the significance was assessed by performing a likelihood-ratio test against a null model without the success indicator variable. The differences between strain engraftment rates among phyla were assessed by Kruskal-Wallis test, and the pairwise comparisons were compared with post-hoc Dunn's tests. Correlations between continuous variables were assessed using Pearson or Spearman tests. A Pearson's partial correlation was used to assess the association between donor richness and post-FMT richness while accounting for sequencing depth. Two linear regression models, the logarithm of donor sample depth against donor richness and the logarithm of post-FMT sample depth against post-FMT richness, were fitted with ordinary least squares. The residuals of the two models were correlated using Pearson's correlation coefficient. Correction for multiple testing (Benjamini–Hochberg procedure, q) was applied when appropriate with significance defined at $q<0.05$.

Machine learning

In order to predict the post-FMT microbiome, we organized the data such that each record represented a species in an FMT triad. As features associated with each record, we used information specific to each FMT triad (Jaccard distances and Bray-Curtis dissimilarities between pre-FMT and donor samples as estimates for their microbiome compositional similarity, ratio of pre-FMT and donor species abundances, time between FMT and sampling), species relative abundances and Shannon entropy values for pre-FMT and donor samples, information about species (taxonomy, prevalence in an unrelated set of metagenomic samples) and cohort-specific information (dataset, disease infectivity). We trained Random Forest models (Breiman 2001) both in a Leave-One-Dataset-Out (LODO) as well as in a 5-fold Cross-Validation (CV) fashion. In the CV setting, we repeated the entire training/evaluation with 5 resamplings and averaged the prediction probabilities. In order to not unfairly overestimate the model performance, we omitted species that were absent in both pre-FMT and donor samples in the evaluation step since those are easy to predict (Fig. 4A, Fig. 4B). Training and evaluation of Random Forest models was done using the classif.ranger learner (for the presence/absence classifier) and regr.ranger (for the abundance regressor) from the mlr3 package (version 0.10) in R (Lang et al. 2019) with parameter importance = 'permutation'. Feature importance values were then obtained directly from the trained Random Forest model. In the context of the pre-FMT/donor sample exchange experiments (Fig. 4C, Fig. 4D), we chose random pre-FMT/donor samples from a different FMT triad of the same dataset and exchanged all associated features. We ensured that donor samples came from a different FMT triad and from a different donor individual (since some donor individuals donated stool to more than one FMT triad).

Ethical compliance

Study procedures of the newly-collected datasets were performed in compliance with the Declaration of Helsinki. Ethical approval was granted by Ethics Committees of Fondazione

Policlinico Gemelli IRCCS (ID 3555/2021) and Ospedale Pediatrico Bambino Gesù IRCCS (1107_OPBG_2016). Written informed consent was obtained from all adult participants, and from parents of underaged participants.

Acknowledgements

We thank all study participants for their commitment and the following authors of included studies for their help in providing data and metadata: Haggai Bar-Yoseph, Rasmus Goll, Hyunmin Koo, Stefano Leo, Casey Morrow, Alan Moss, David Suskind, Faming Zhang. We also thank all the members of the HPC and NGS facilities at University of Trento, the whole FMT staff at the Fondazione Policlinico Gemelli IRCCS of Rome, all the FMT Ospedale Pediatrico Bambino Gesù Committee Collaborators (particularly Pietro Merli, Paola De Angelis, Giulia Angelino, Erminia Francesco Romeo and Livia Gargiullo, Stefania Pane, Sandra Martino, Lorenza Romani, Paola Bernaschi, Andrea Finocchi, Giulia Marucci, Francesca Rea, Simona Faraci, Patrizia D' Argenio, Luigi Dall'Oglio) and the biobank of the Ospedale Pediatrico Bambino Gesù.

Funding

This work was supported by the European Research Council (ERC-STG project MetaPG) to NS, by MIUR 'Futuro in Ricerca' (grant nr RBFR13EWWI_001) to NS; by the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project) to NS; by the National Cancer Institute of the National Institutes of Health (1U01CA230551) to NS; by the Premio Internazionale Lombardia e Ricerca 2019 to NS; by the EMBO ALTF 593-2020 to MV-C; by the Ricerca Finalizzata Giovani Ricercatori 2018 of the Italian Ministry of Health (project GR-2018-12365734) to GI.

Author contributions

GI and NS conceived and designed the study. MP and NK performed the analysis. GI, NK and SP performed the literature search. GI and GC supervised the sample collection and the clinical procedures. GI, MV-C, and NS supervised the analysis. MP, NK, FAr, FAs, FB, AB-M, FC, PM, and FP contributed to data acquisition, data analysis, or software developments. GI, SP, LM, GQ, SDG, GDS, SB, LP, FDC, MS, AG, and GC contributed to the clinical procedures and sample collection. GI, MP, NK, MV-C, and NS interpreted the data and wrote the manuscript. All authors provided critical revision of the manuscript and approved the final version for submission.

Competing interests

AG reports personal fees for consultancy for Eisai S.r.l., 3PSolutions, Real Time Meeting, Fondazione Istituto Danone, Sinergie S.r.l. Board MRGE, and SanofiS.p.A, personal fees for acting as a speaker for Takeda S.p.A, AbbVie, and Sandoz S.p.A, and personal fees for acting on advisory boards for VSL3 and Eisai. GC has received personal fees for acting as advisor for Ferring Therapeutics. GI has received personal fees for acting as speaker for Biocodex, Danone, Sofar, Malesci, Metagenics, and for acting as consultant/advisor for Ferring

Therapeutics, Giuliani, Metagenics. NS reports consultancy contracts with Zoe, Roche, Ysopia, and Freya and is co-founder of PreBiomics. The other authors have no potential competing interest to disclose.

Materials & Correspondence

Correspondence and requests for materials should be addressed to GI or NS.

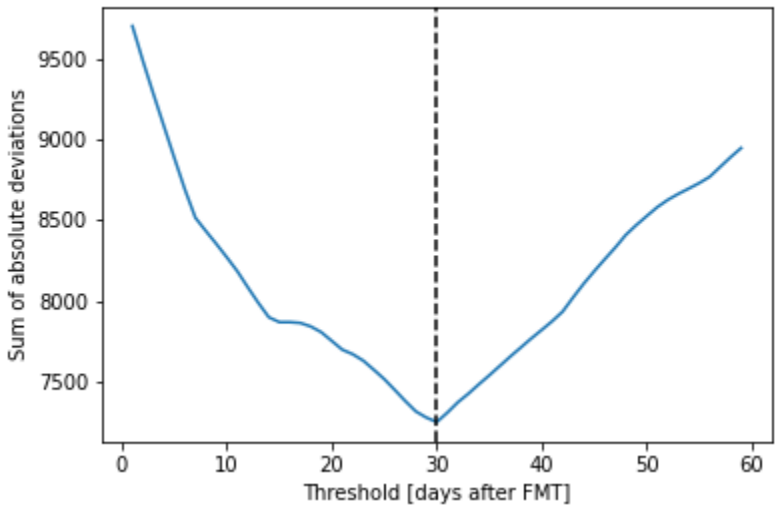
Data and code availability

Newly-generated shotgun metagenomics sequencing data are available at the European Nucleotide Archive under accession number PRJEB47909 [pending submission]. Metadata are available in Supplementary Table 2 and in curatedMetagenomicData (Pasolli et al. 2017) [pending submission]. All analyses were performed using open-source software.

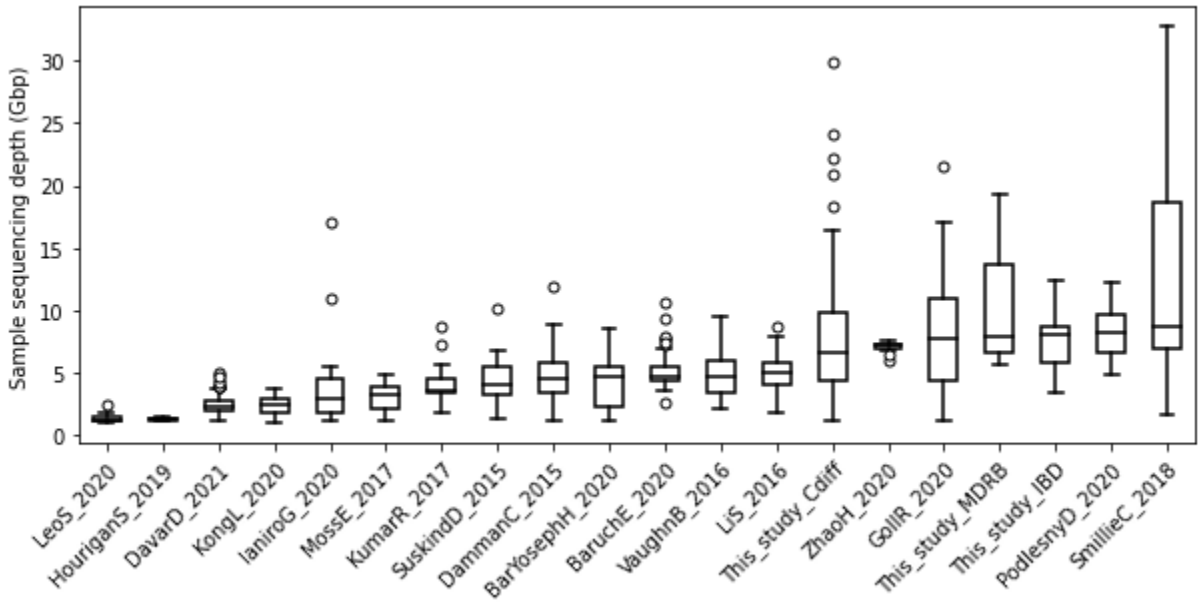
Supplementary Tables

https://docs.google.com/spreadsheets/d/1MulIdnPu2ZxH4ZVyASrwqV2u946CeWsiZpmL4pn6Qf4/edit#gid=939460804

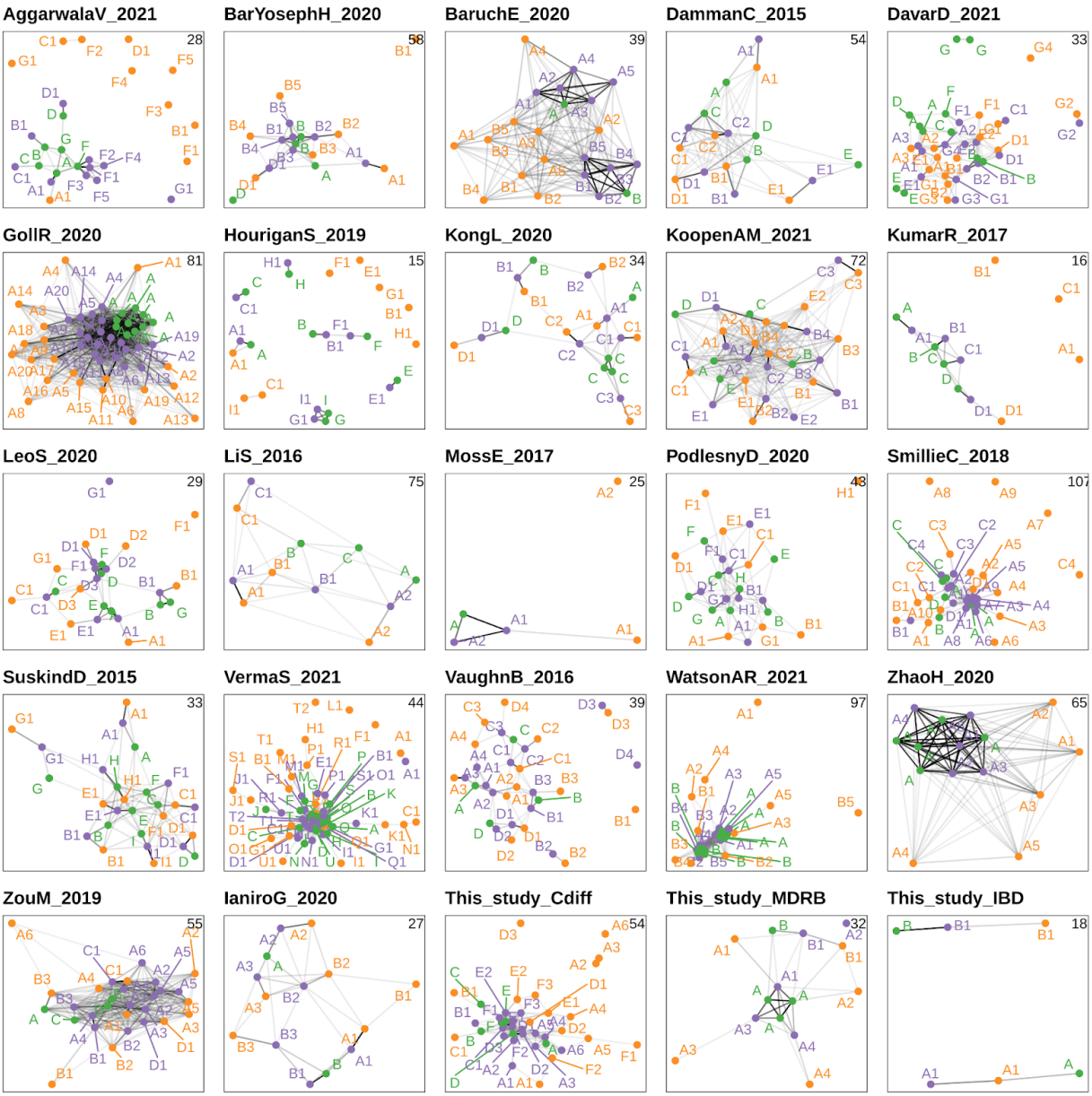
Supplementary Figures



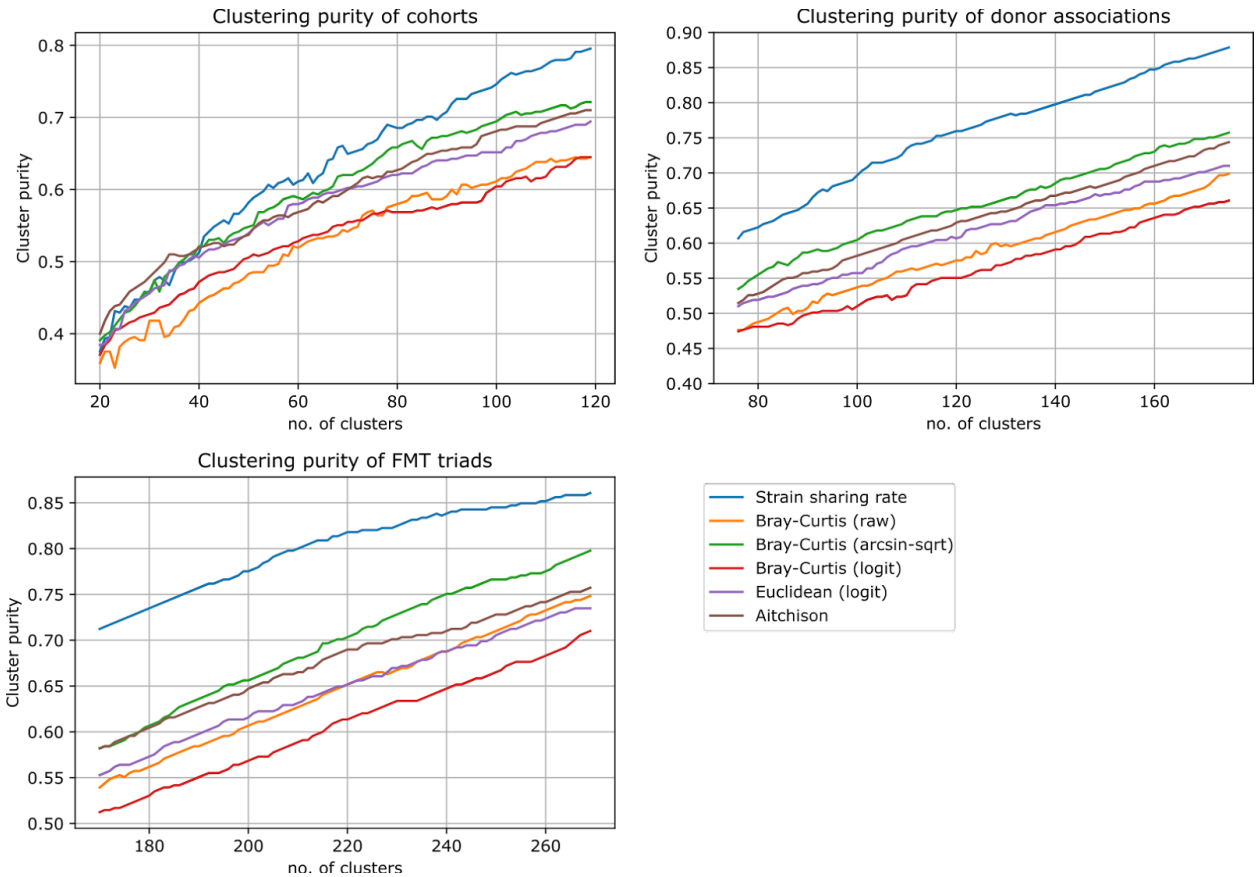
Supplementary Figure 1. Sum of absolute deviations (over all FMT instances) of the sampling time (days) of post-FMT samples. 30 days post-FMT is the value minimizing the sum of absolute deviations.



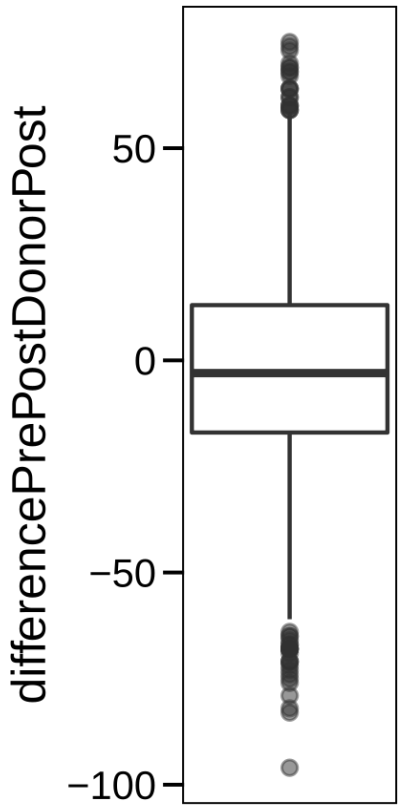
Supplementary Figure 2. Distribution of sequencing depths of all the 20 datasets included in this study.



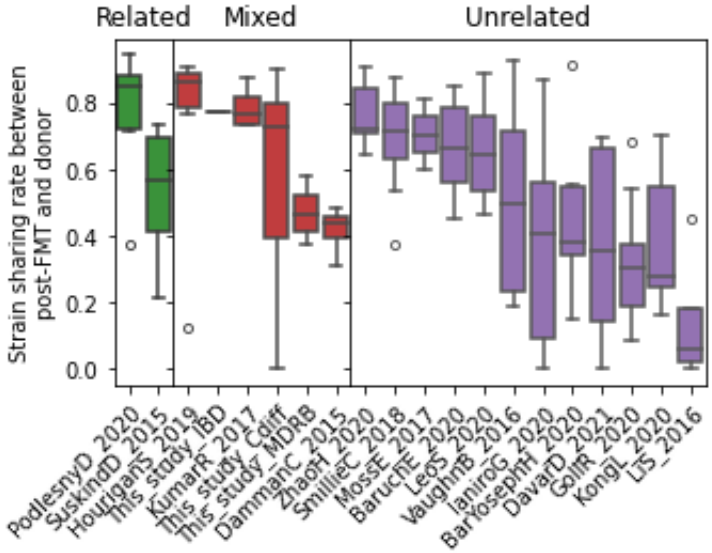
Supplementary Figure 3. Strain sharing networks of the datasets included in this study not shown in Fig. 1A as well as the ZouM_2019 cohort which wasn't analyzed further. Each node corresponds to a sample and is colored by its role in FMT triads (recipient pre-FMT sample, recipient post-FMT sample, and donor's sample). Edge opacity is proportional to the number of shared strains between two samples (*Methods)* and only edges corresponding to at least 2 shared strains are shown. The structure of the networks illustrates how FMT triads tend to cluster together but with different clustering characteristics across cohorts.



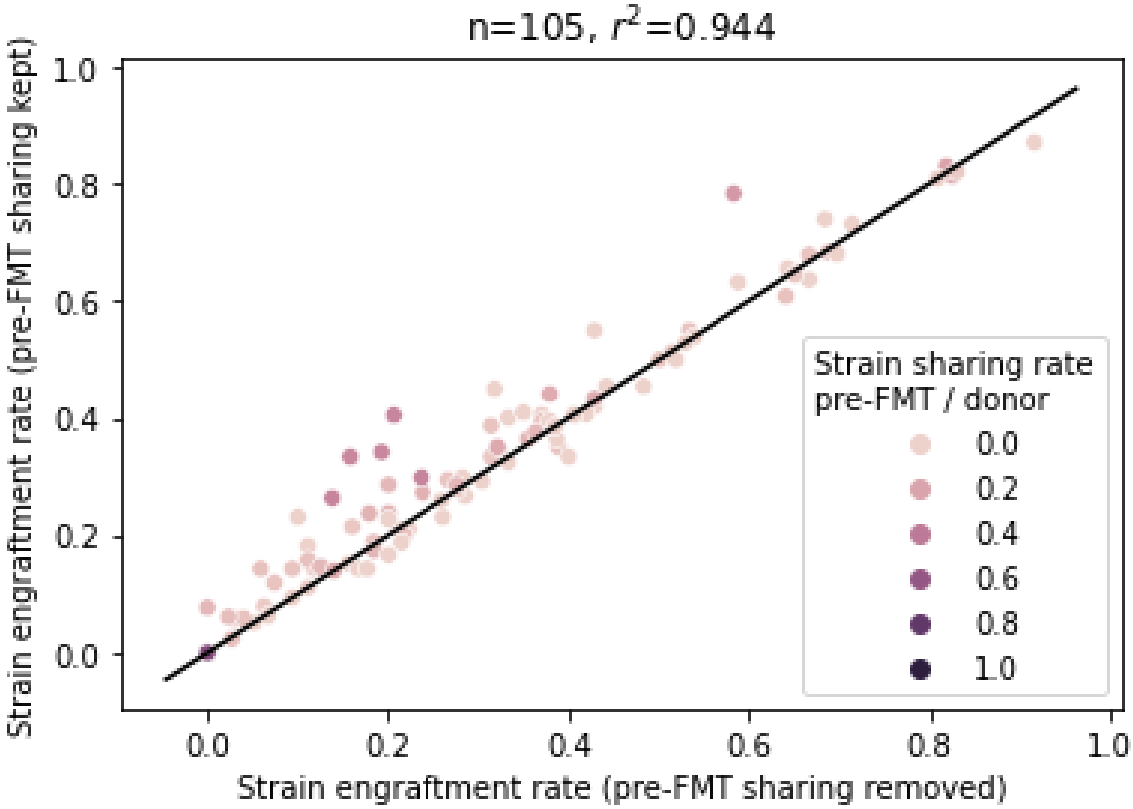
Supplementary Figure 4. The purity of K-medoids clustering with varying K shows that strain sharing rate outperforms beta diversity measures in clustering by donor associations and by FMT triads. In clustering by cohorts for the low number of clusters it gets outperformed by Aitchison distance, but catches up as the K increases.



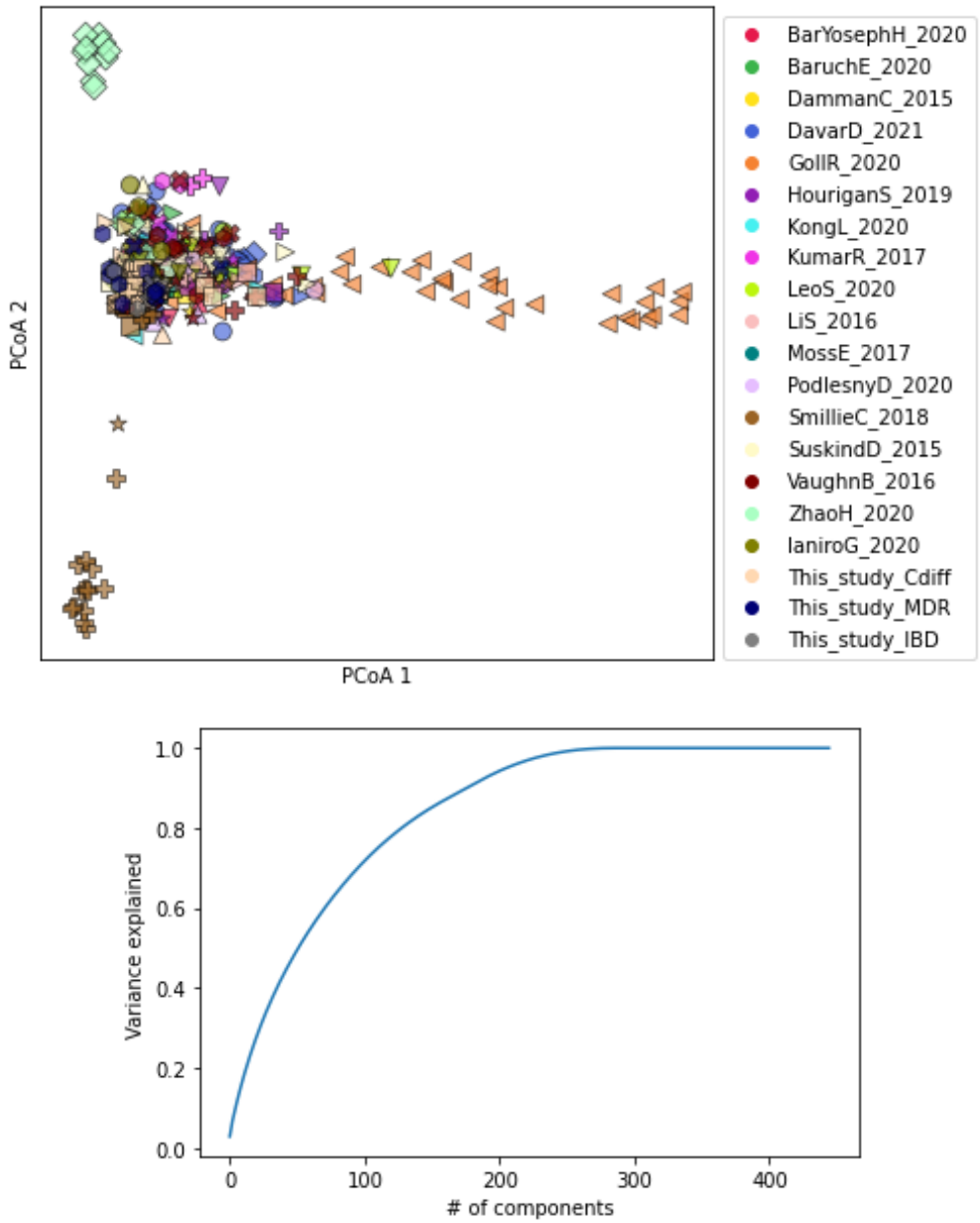
Supplementary Figure 5. Difference in the number of shared strains between pre-FMT/post-FMT and donor/post-FMT sample pairs over all FMT triads, showing the large variation with which recipients retain or take up strains. A positive number means that the pre-FMT sample shared more strains with the corresponding post-FMT sample than the donor with the corresponding post-FMT sample.



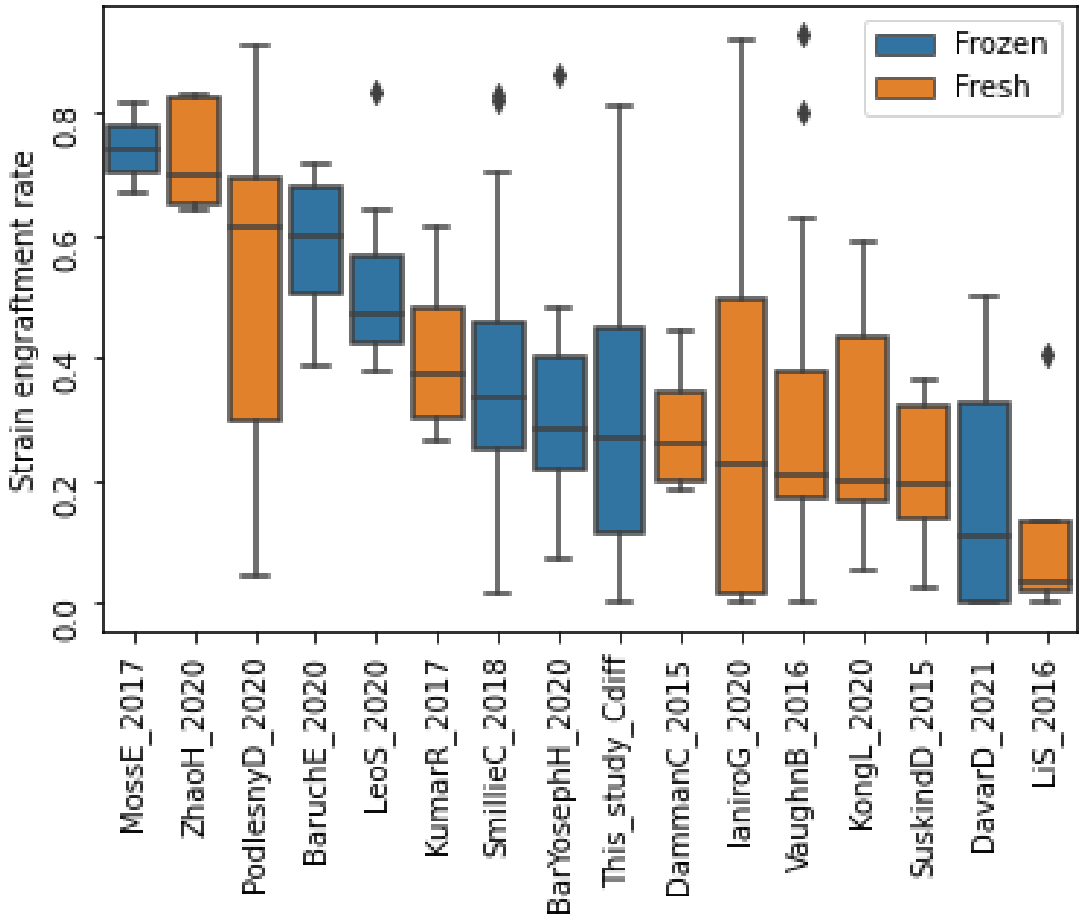
Supplementary Figure 6. Strain sharing rates between donor and post-FMT samples is non-significantly higher in datasets using related or a mixture of related and unrelated donors compared to those using only unrelated donors (Related and Mixed vs Unrelated median difference=0.0919, permutation test p=0.3829).



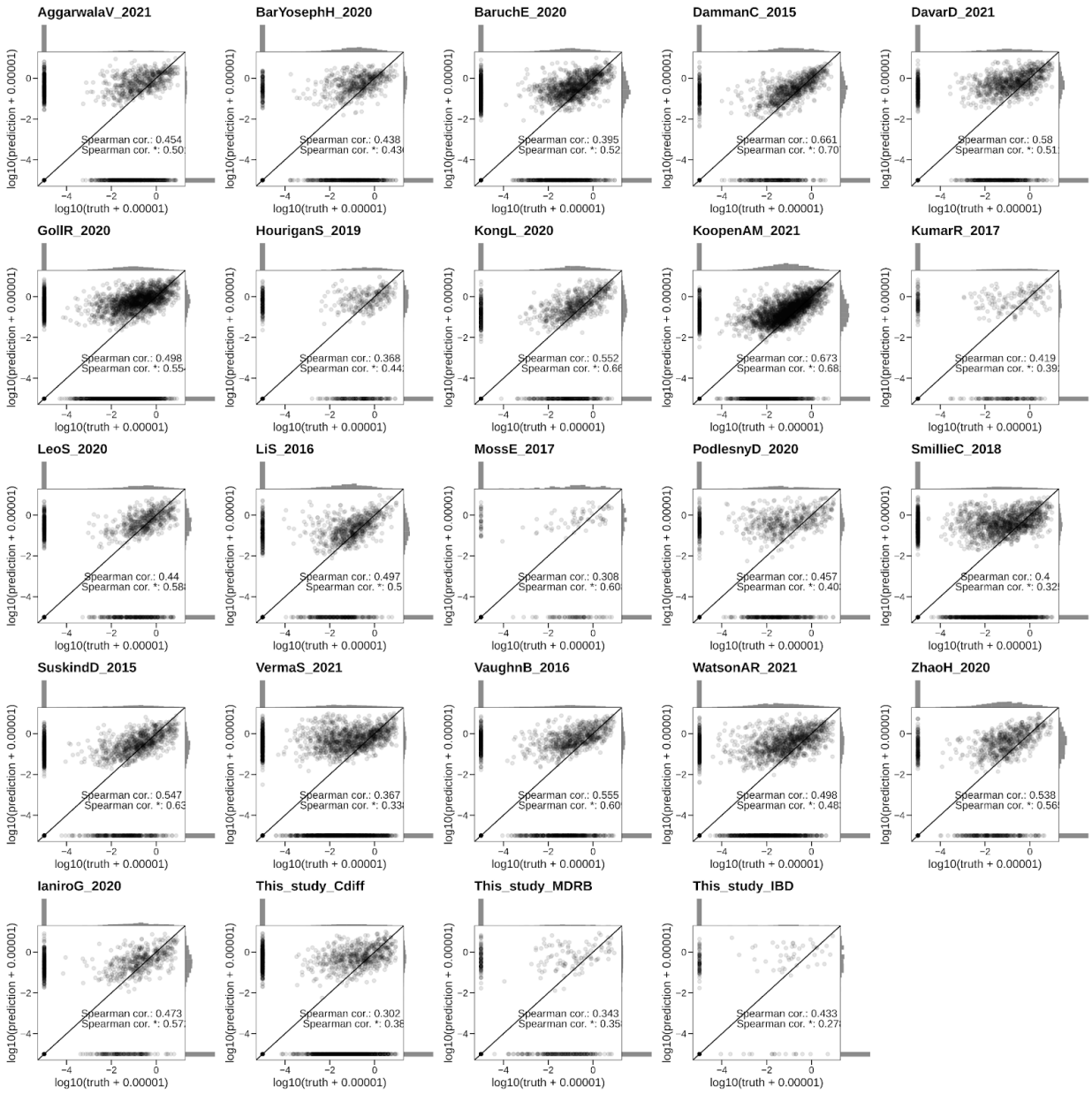
Supplementary Figure 7. Strain engraftment rates would be in many cases overestimated if the amount of strains shared between pre-FMT and donor samples was not subtracted. To avoid noisy values, only the FMTs in which at least 5 strains could engraft (as defined by the strain engraftment rate, Methods) is displayed.



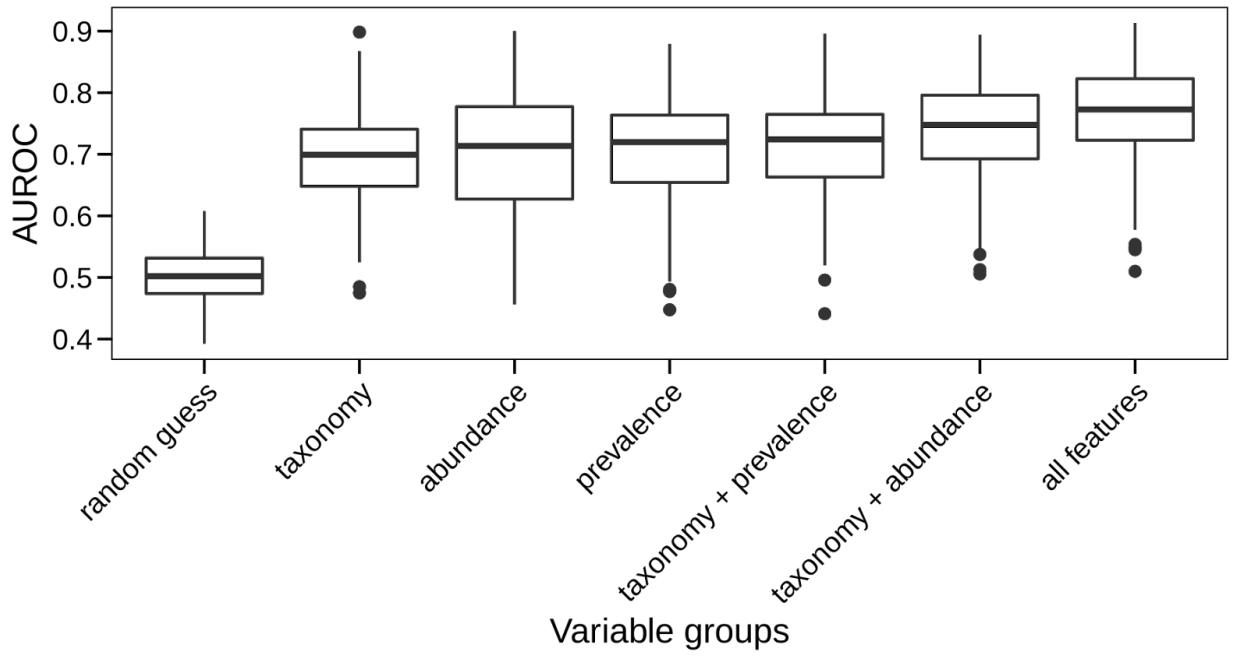
Supplementary Figure 8. PCoA ordination on strain sharing rate distances and variance explained by number of components, suggesting that two dimensions are not sufficient to linearly separate the clusters induced by dataset or donor batch effects. Unique combinations of color and shape correspond to samples associated with one donor subject.



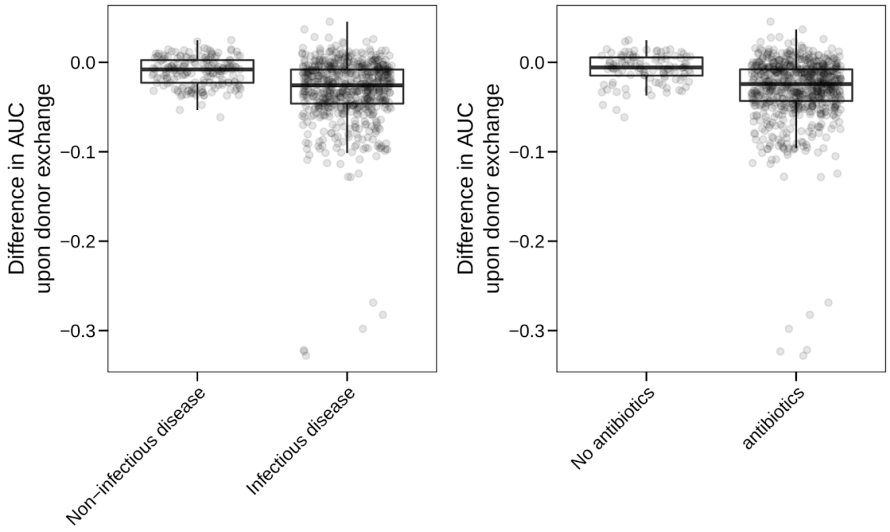
Supplementary Figure 9. Strain engraftment rates in the datasets included in this study do not differ significantly based on whether the FMT was performed on fresh or frozen material (median difference=0.05, permutation test, p=0.60).



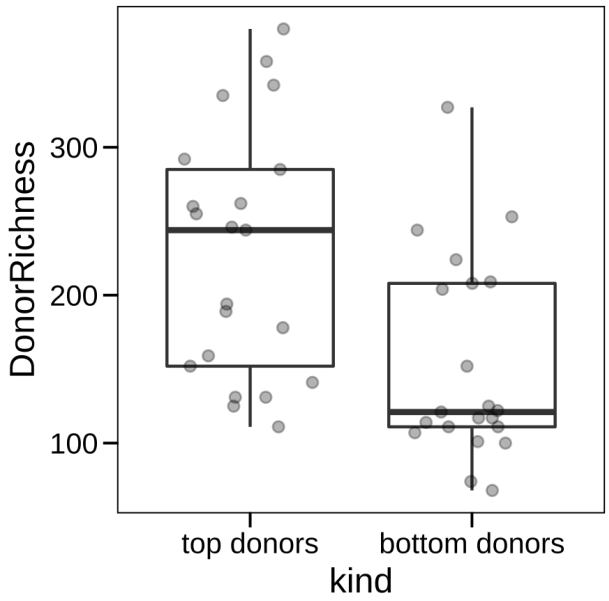
Supplementary Figure 10. Scatterplots by dataset of real vs. predicted bacterial abundances in post-FMT samples. Abundances were predicted using a RF regression model (see Methods). The asterisk designates the Spearman correlation when omitting truly absent species predicted to be absent. Corresponds to Figure 4G.



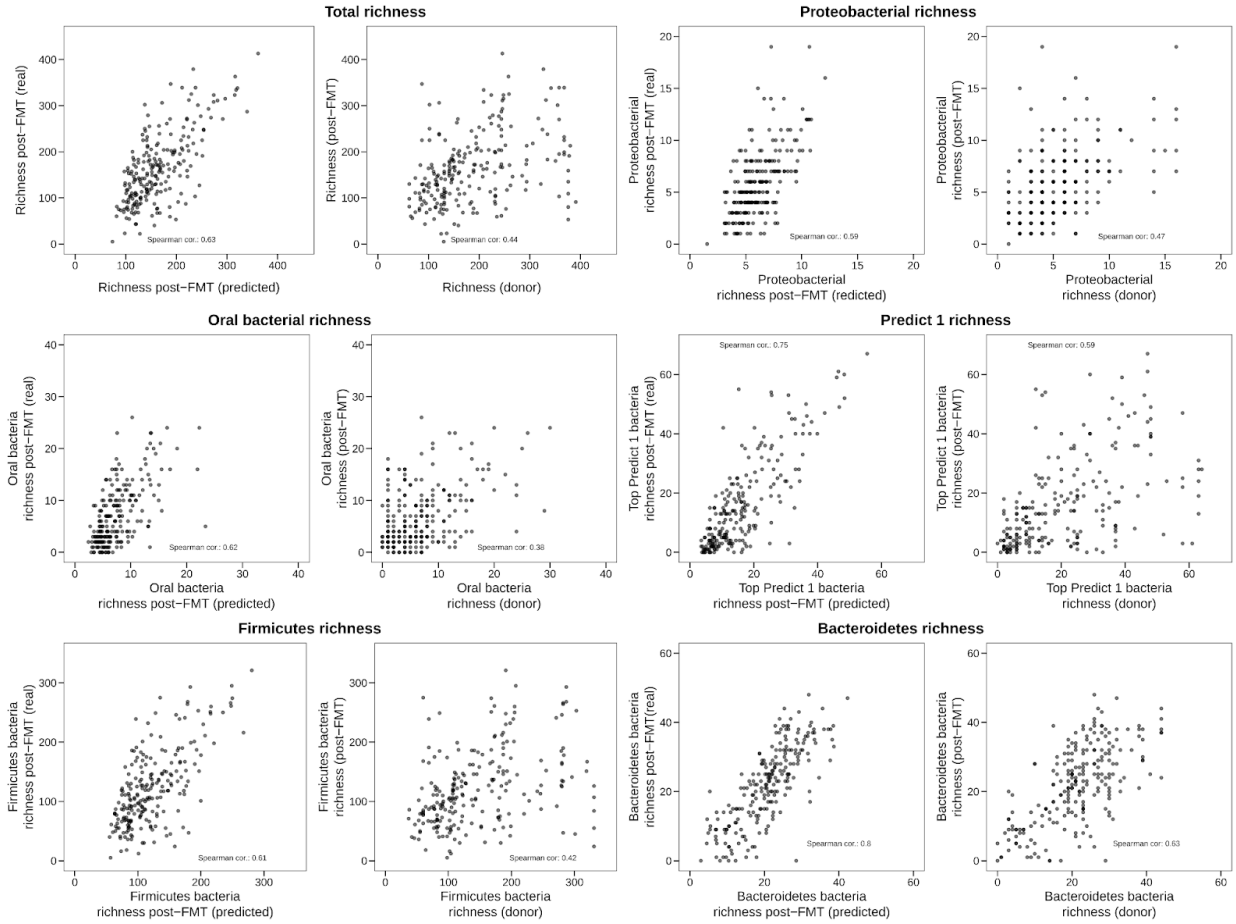
Supplementary Figure 11. LODO AUROC values of post-FMT presence/absence predictions using RF classifiers with varying sets of features. The illustrates that the information content in microbial taxonomic labels is roughly equal to that of abundance or prevalence.



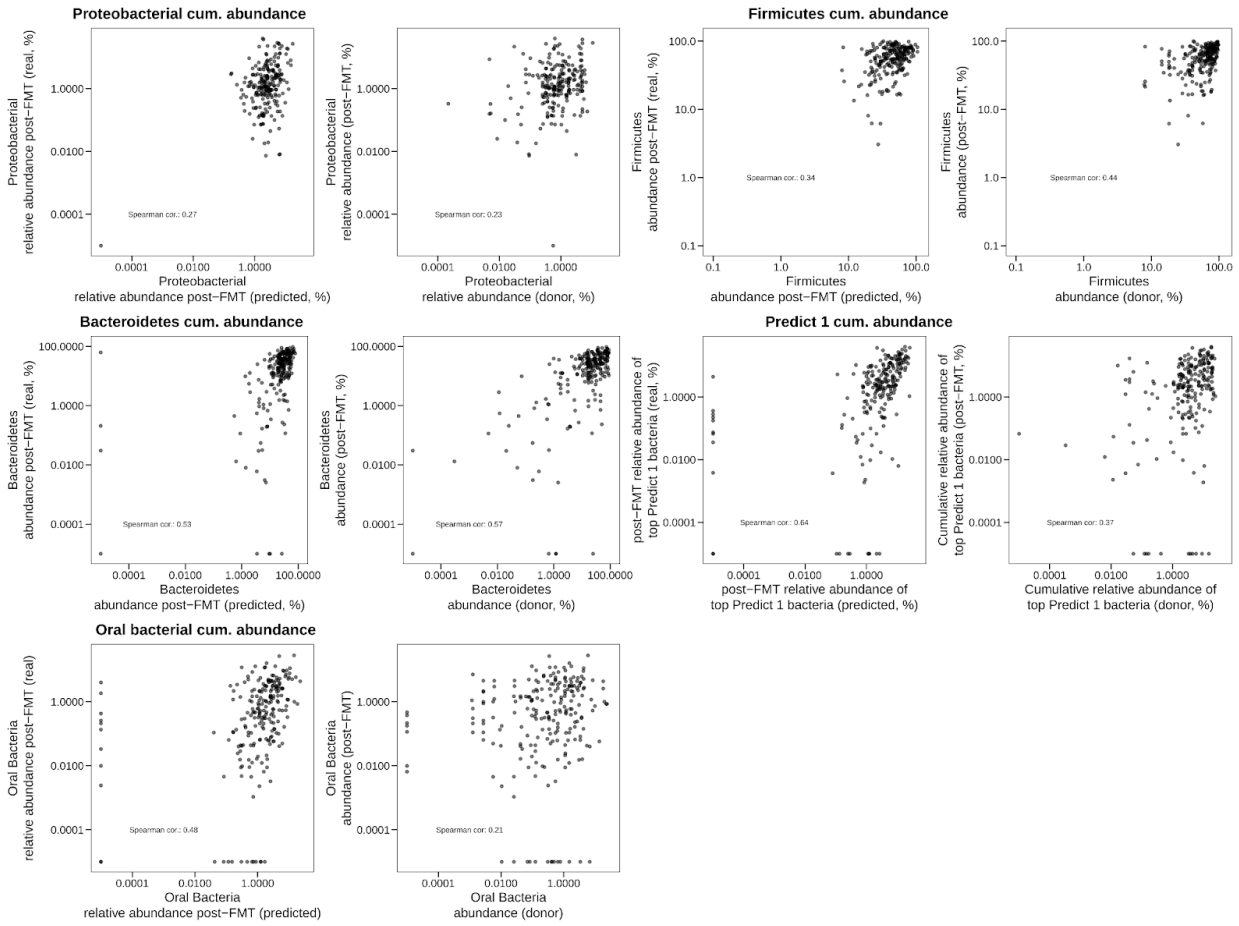
Supplementary Figure 12. Boxplots of the difference in AUC upon simulated donor exchange. Wilcoxon test p-value <0.001 for both infectious vs. non-infectious disease and antibiotics vs. no antibiotics comparison.



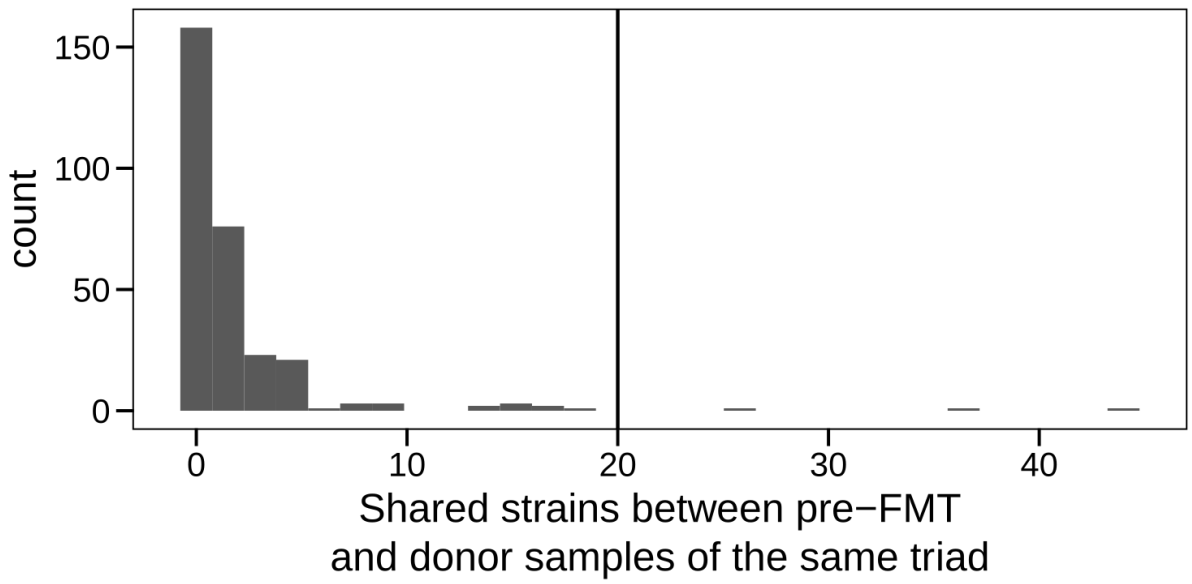
Supplementary Figure 13. Donor species richness of donors that led to a maximum increase in predicted post-FMT richness upon exchange into an FMT triad in each dataset ("top donors") and donors that led to a maximum decrease in predicted post-FMT richness in each dataset ("bottom donors").



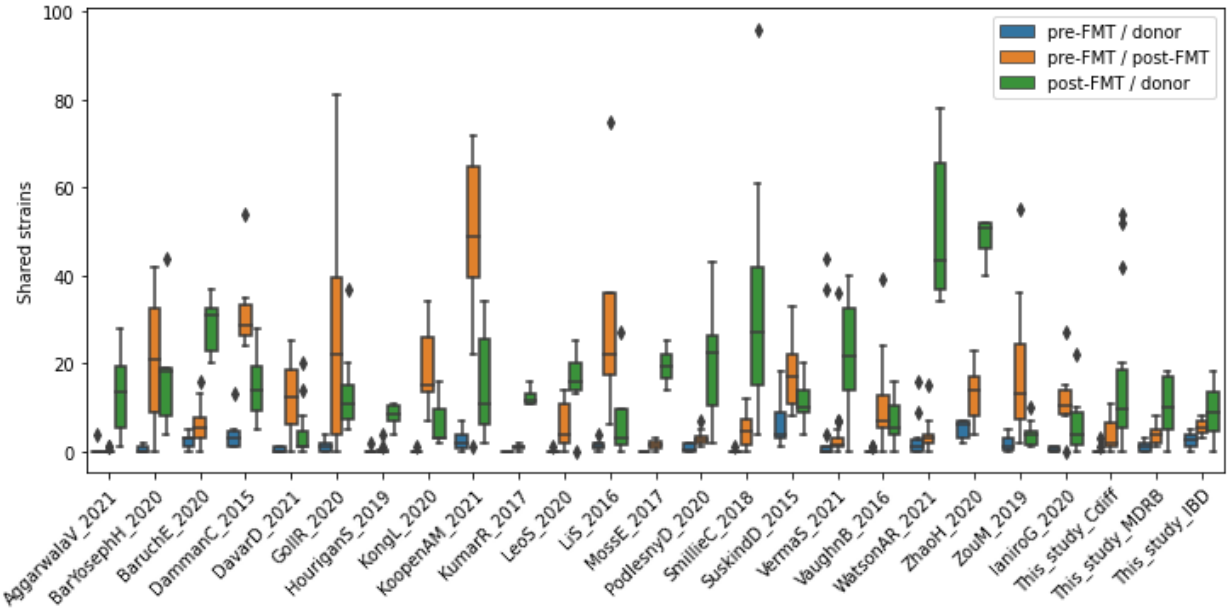
Supplementary Figure 14. Comparisons of the predicted total species richness of bacterial groups in post-FMT samples. Predictions on the y-axis come from the RF classifier, predictions on the x-axis correspond to the cumulative richness in donor samples.



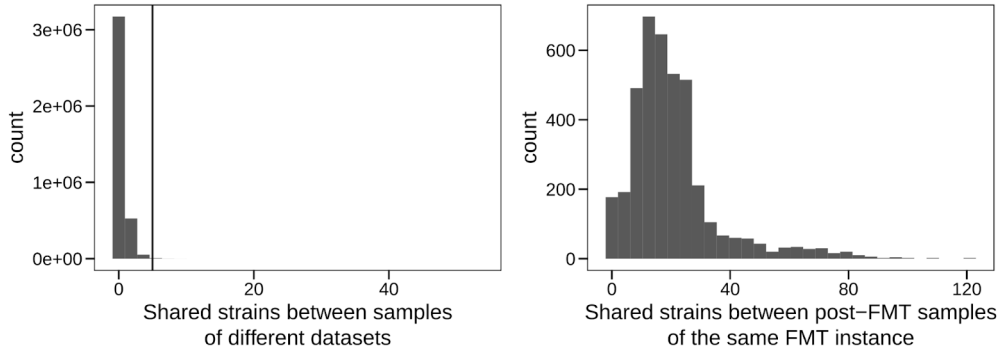
Supplementary Figure 15. Comparisons of the predicted cumulative abundance of bacterial groups in post-FMT samples. Predictions on the y-axis come from the RF regressor, predictions on the x-axis correspond to the cumulative abundance in donor samples.



Supplementary Figure 16. Three recipients share an exceptionally high number of strains (>20) with their donor before the FMT, strongly suggesting an error in the metadata.



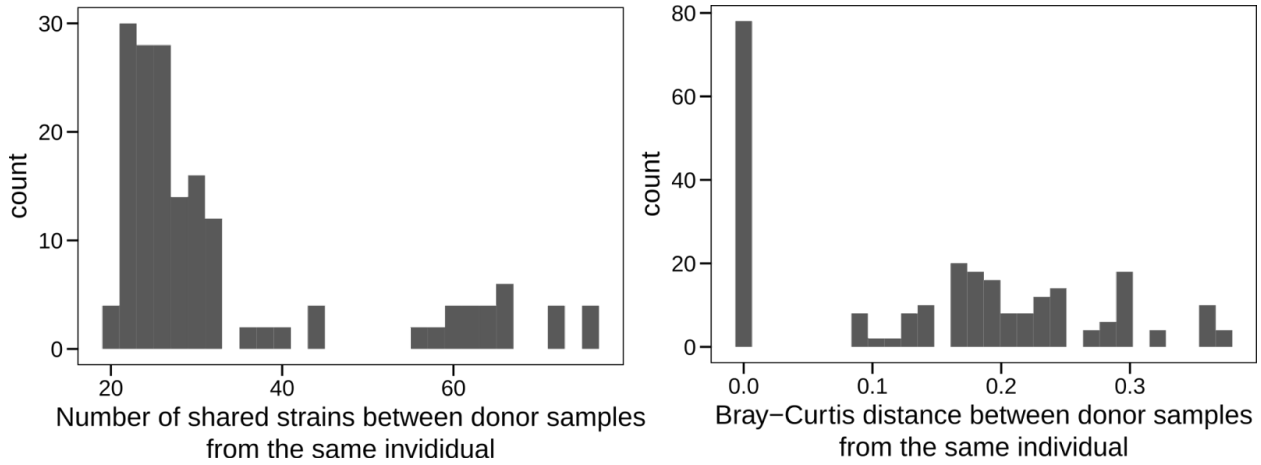
Supplementary Figure 17. Distributions of the number of shared strains within FMT triads by dataset and sample group. Note that The ZouM_2019 dataset is the only dataset in our meta cohort that displays a median of only one shared strain between corresponding post-FMT and donor samples. For this and other reasons, it was excluded from further analyses (see Methods).



Supplementary Figure 18. Distribution of the number of shared strains between samples coming from different datasets (left) and between post-FMT samples of the same FMT triad (right), showing that unrelated samples rarely share more than 5 strains.

sample name		UWIBD01 D049B0	UWIBD01 D049B1	UWIBD01 D202B0	UWIBD01 D202B1	UWIBD01 D341B0	UWIBD01 D554B0	UWIBD01 D554B1	UWIBD01 D693B0	UWIBD01 D693B1	UWIBD01 D862B0	UWIBD01 D862B1
	subject name	inferred_ donor_5	inferred_ donor_5	inferred_ donor_1	inferred_ donor_1	inferred_ donor_2	inferred_ donor_2	inferred_ donor_2	inferred_ donor_4	inferred_ donor_4	inferred_ donor_3	inferred_ donor_3
UWIBD01 P016T1	P016	3	0	0	0	41	47	44	1	2	2	3
UWIBD01 P016T2	P016	3	1	0	1	28	27	25	0	2	4	2
UWIBD01 P016T3	P016	0	0	0	0	20	22	20	0	2	1	1
UWIBD01 P081T1	P081	1	1	0	0	0	1	1	39	41	2	1
UWIBD01 P081T2	P081	1	1	0	0	2	1	1	15	14	3	1
UWIBD01 P081T3	P081	1	1	0	0	0	0	0	7	8	2	1
UWIBD01 P224T1	P224	1	0	3	5	0	0	0	1	1	1	0
UWIBD01 P224T2	P224	0	0	5	5	0	0	0	1	1	0	0
UWIBD01 P224T3	P224	1	0	5	5	0	0	0	1	1	1	0
UWIBD01 P271T1	P271	0	0	0	0	2	4	3	2	2	13	8
UWIBD01 P271T2	P271	0	0	0	0	0	0	0	0	0	0	0
UWIBD01 P271T3	P271	0	0	0	0	0	0	0	0	0	0	0
UWIBD01 P279T1	P279	2	1	0	2	30	33	35	4	3	2	0
UWIBD01 P279T2	P279	2	2	0	2	21	29	29	6	6	5	1
UWIBD01 P279T3	P279	1	1	0	1	20	22	19	3	4	3	2
UWIBD01 P788T1	P788	2	2	0	0	0	0	0	2	2	0	0
UWIBD01 P788T2	P788	8	5	0	0	1	1	1	2	2	0	0
UWIBD01 P788T3	P788	5	5	0	0	0	0	0	0	0	0	0

Supplementary Figure 19. Matrix of shared strains between all donor samples and all post-FMT samples in the DammanC_2015 cohort, which was used to infer donor-recipient mapping in this dataset.



Supplementary Figure 20. Donor samples from the same individual are reasonably stable in terms of species-level composition and strain identity.

REFERENCES

Aggarwala, Varun, Ilaria Mogno, Zhihua Li, Chao Yang, Graham J. Britton, Alice Chen-Liaw, Josephine Mitcham, et al. n.d. “Quantification of Discrete Gut Bacterial Strains Following Fecal Transplantation for Recurrent Clostridioides Difficile Infection Demonstrates Long-Term Stable Engraftment in Non-Relapsing Recipients.” https://doi.org/10.1101/2020.09.10.292136.

Albanese, Davide, and Claudio Donati. 2017. “Strain Profiling and Epidemiology of Bacterial Species from Metagenomic Sequencing.” *Nature Communications* 8 (1): 2260.

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.

Asnicar, Francesco, Sarah E. Berry, Ana M. Valdes, Long H. Nguyen, Gianmarco Piccinno, David A. Drew, Emily Leeming, et al. 2021. “Microbiome Connections with Host Metabolism and Habitual Diet from 1,098 Deeply Phenotyped Individuals.” *Nature Medicine* 27 (2): 321–32.

Asnicar, Francesco, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, et al. 2020. “Precise Phylogenetic Analysis of Microbial Isolates and Genomes from Metagenomes Using PhyloPhlAn 3.0.” *Nature Communications* 11 (1): 2500.

Baruch, Erez N., Ilan Youngster, Guy Ben-Betzalel, Rona Ortenberg, Adi Lahat, Lior Katz, Katerina Adler, et al. 2021. “Fecal Microbiota Transplant Promotes Response in Immunotherapy-Refractory Melanoma Patients.” *Science* 371 (6529): 602–9.

Bar-Yoseph, Haggai, Shaqed Carasso, Shlomit Shklar, Alexander Korytny, Razi Even Dar, Haneen Daoud, Roni Nassar, et al. 2021. “Oral Capsulized Fecal Microbiota Transplantation for Eradication of Carbapenemase-Producing Enterobacteriaceae Colonization With a Metagenomic Perspective.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 73 (1): e166–75.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Usinglme4.” *Journal of Statistical Software*. https://doi.org/10.18637/jss.v067.i01.

Baunwall, Simon Mark Dahl, Mads Ming Lee, Marcel Kjærsgaard Eriksen, Benjamin H. Mullish, Julian R. Marchesi, Jens Frederik Dahlerup, and Christian Lodberg Hvas. 2020. “Faecal Microbiota Transplantation for Recurrent Clostridioides Difficile Infection: An Updated Systematic Review and Meta-Analysis.” *EClinicalMedicine* 29-30 (December): 100642.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2020. “Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.11.19.388223.

———. 2021. “Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3.” *eLife* 10 (May). https://doi.org/10.7554/eLife.65088.

Beghini, Francesco, Edoardo Pasolli, Tin Duy Truong, Lorenza Putignani, Simone M. Cacciò, and Nicola Segata. 2017. “Large-Scale Comparative Metagenomics of Blastocystis, a Common Member of the Human Gut Microbiome.” *The ISME Journal* 11 (12): 2848–63.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2012. “GenBank.” *Nucleic Acids Research* 41 (D1): D36–42.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Brito, Ilana L., Thomas Gurry, Shijie Zhao, Katherine Huang, Sarah K. Young, Terrence P. Shea, Waisea Naisilisili, et al. 2019. “Transmission of Human-Associated Microbiota along Family and Social Networks.” *Nature Microbiology* 4 (6): 964–71.

Browne, Hilary P., B. Anne Neville, Samuel C. Forster, and Trevor D. Lawley. 2017. “Transmission of the Gut Microbiota: Spreading of Health.” *Nature Reviews. Microbiology* 15 (9): 531–43.
Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods*. https://doi.org/10.1038/nmeth.3176.
Burns, Alyssa M., Michelle A. Zitt, Cassie C. Rowe, Bobbi Langkamp-Henken, Volker Mai, Carmelo Nieves Jr, Maria Ukhanova, Mary C. Christman, and Wendy J. Dahl. 2016. “Diet Quality Improves for Parents and Children When Almonds Are Incorporated into Their Daily Diet: A Randomized, Crossover Study.” *Nutrition Research* 36 (1): 80–89.
Cammarota, Giovanni, Gianluca Ianiro, Colleen R. Kelly, Benjamin H. Mullish, Jessica R. Allegretti, Zain Kassam, Lorenza Putignani, et al. 2019. “International Consensus Conference on Stool Banking for Faecal Microbiota Transplantation in Clinical Practice.” *Gut* 68 (12): 2111–21.
Chang, Ju Young, Dionysios A. Antonopoulos, Apoorv Kalra, Adriano Tonelli, Walid T. Khalife, Thomas M. Schmidt, and Vincent B. Young. 2008. “Decreased Diversity of the Fecal Microbiome in Recurrent Clostridium Difficile-Associated Diarrhea.” *The Journal of Infectious Diseases* 197 (3): 435–38.
Costello, S. P., W. Soo, R. V. Bryant, V. Jairath, A. L. Hart, and J. M. Andrews. 2017. “Systematic Review with Meta-Analysis: Faecal Microbiota Transplantation for the Induction of Remission for Active Ulcerative Colitis.” *Alimentary Pharmacology & Therapeutics* 46 (3): 213–24.
Csardi, Gabor, Tamas Nepusz, and Others. 2006. “The Igraph Software Package for Complex Network Research.” *InterJournal, Complex Systems* 1695 (5): 1–9.
Damman, Christopher J., Mitchell J. Brittnacher, Maria Westerhoff, Hillary S. Hayden, Matthew Radey, Kyle R. Hager, Sara R. Marquis, Samuel I. Miller, and Timothy L. Zisman. 2015. “Low Level Engraftment and Improvement Following a Single Colonoscopic Administration of Fecal Microbiota to Patients with Ulcerative Colitis.” *PloS One* 10 (8): e0133925.
Dancer, Stephanie J. 2004. “How Antibiotics Can Make Us Sick: The Less Obvious Adverse Effects of Antimicrobial Chemotherapy.” *The Lancet Infectious Diseases* 4 (10): 611–19.
Davar, Diwakar, Amiran K. Dzutsev, John A. McCulloch, Richard R. Rodrigues, Joe-Marc Chauvin, Robert M. Morrison, Richelle N. Deblasio, et al. 2021. “Fecal Microbiota Transplant Overcomes Resistance to Anti-PD-1 Therapy in Melanoma Patients.” *Science* 371 (6529): 595–602.
De Groot, Pieter F., M. N. Frissen, N. C. De Clercq, and M. Nieuwdorp. 2017. “Fecal Microbiota Transplantation in Metabolic Syndrome: History, Present and Future.” *Gut Microbes* 8 (3): 253–67.
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.
Ferri, Maurizio, Elena Ranucci, Paola Romagnoli, and Valerio Giaccone. 2017. “Antimicrobial Resistance: A Global Emerging Threat to Public Health Systems.” *Critical Reviews in Food Science and Nutrition* 57 (13): 2857–76.
Gardiner, B. J., A. Y. Tai, D. Kotsanas, M. J. Francis, S. A. Roberts, S. A. Ballard, R. K. Junckerstorff, and T. M. Korman. 2015. “Clinical and Microbiological Characteristics of Eggerthella Lenta Bacteremia.” *Journal of Clinical Microbiology* 53 (2): 626–35.
Goll, Rasmus, Peter Holger Johnsen, Erik Hjerde, Joseph Diab, Per Christian Valle, Frank Hilpusch, and Jorunn Pauline Cavanagh. 2020. “Effects of Fecal Microbiota Transplantation in Subjects with Irritable Bowel Syndrome Are Mirrored by Changes in Gut Microbiome.” *Gut Microbes* 12 (1): 1794263.
Green, Jessica Emily, Jessica A. Davis, Michael Berk, Christopher Hair, Amy Loughman, David

Castle, Eugene Athan, et al. 2020. “Efficacy and Safety of Fecal Microbiota Transplantation for the Treatment of Diseases Other than Clostridium Difficile Infection: A Systematic Review and Meta-Analysis.” *Gut Microbes* 12 (1): 1–25.
Hourigan, Suchitra K., Michelle Ahn, Keylie M. Gibson, Marcos Pérez-Losada, Grace Felix, Melissa Weidner, Ian Leibowitz, et al. 2019. “Fecal Transplant in Children With Clostridioides Difficile Gives Sustained Reduction in Antimicrobial Resistance and Potential Pathogen Burden.” *Open Forum Infectious Diseases* 6 (10): ofz379.
Hu, Youjia, Ping Jin, Jian Peng, Xiaojun Zhang, F. Susan Wong, and Li Wen. 2016. “Different Immunological Responses to Early-Life Antibiotic Exposure Affecting Autoimmune Diabetes Development in NOD Mice.” *Journal of Autoimmunity* 72 (August): 47–56.
Ianiro, Gianluca, Leonardo H. Eusebi, Christopher J. Black, Antonio Gasbarrini, Giovanni Cammarota, and Alexander C. Ford. 2019. “Systematic Review with Meta-Analysis: Efficacy of Faecal Microbiota Transplantation for the Treatment of Irritable Bowel Syndrome.” *Alimentary Pharmacology & Therapeutics* 50 (3): 240–48.
Ianiro, Gianluca, Marcello Maida, Johan Burisch, Claudia Simonelli, Georgina Hold, Marco Ventimiglia, Antonio Gasbarrini, and Giovanni Cammarota. 2018. “Efficacy of Different Faecal Microbiota Transplantation Protocols for Clostridium Difficile Infection: A Systematic Review and Meta-Analysis.” *United European Gastroenterology Journal* 6 (8): 1232–44.
Ianiro, Gianluca, Rita Murri, Giusi Desirè Sciumè, Michele Impagnatiello, Luca Masucci, Alexander C. Ford, Graham R. Law, et al. 2019. “Incidence of Bloodstream Infections, Length of Hospital Stay, and Survival in Patients With Recurrent Clostridioides Difficile Infection Treated With Fecal Microbiota Transplantation or Antibiotics: A Prospective Cohort Study.” *Annals of Internal Medicine* 171 (10): 695–702.
Ianiro, Gianluca, Ernesto Rossi, Andrew M. Thomas, Giovanni Schinzari, Luca Masucci, Gianluca Quaranta, Carlo Romano Settanni, et al. 2020. “Faecal Microbiota Transplantation for the Treatment of Diarrhoea Induced by Tyrosine-Kinase Inhibitors in Patients with Metastatic Renal Cell Carcinoma.” *Nature Communications* 11 (1): 1–6.
Ianiro, G., M. Sanguinetti, A. Gasbarrini, and G. Cammarota. 2017. “Predictors of Failure after Single Faecal Microbiota Transplantation in Patients with Recurrent Clostridium Difficile Infection: Results from a 3-Year Cohort Study: Authors’ Reply.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 23 (11): 891.
Jakobsson, Hedvig E., Cecilia Jernberg, Anders F. Andersson, Maria Sjölund-Karlsson, Janet K. Jansson, and Lars Engstrand. 2010. “Short-Term Antibiotic Treatment Has Differing Long-Term Impacts on the Human Throat and Gut Microbiome.” *PloS One* 5 (3): e9836.
Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. “MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies.” *PeerJ* 7 (July): e7359.
Kong, Lingjia, Jason Lloyd-Price, Tommi Vatanen, Philippe Seksik, Laurent Beaugerie, Tabassome Simon, Hera Vlamakis, Harry Sokol, and Ramnik J. Xavier. 2020. “Linking Strain Engraftment in Fecal Microbiota Transplantation With Maintenance of Remission in Crohn’s Disease.” *Gastroenterology* 159 (6): 2193–2202.e5.
Kootte, Ruud S., Evgeni Levin, Jarkko Salojärvi, Loek P. Smits, Annick V. Hartstra, Shanti D. Udayappan, Gerben Hermes, et al. 2017. “Improvement of Insulin Sensitivity after Lean Donor Feces in Metabolic Syndrome Is Driven by Baseline Intestinal Microbiota Composition.” *Cell Metabolism* 26 (4): 611–19.e6.
Korpela, Katri, Paul Costea, Luis Pedro Coelho, Stefanie Kandels-Lewis, Gonneke Willemsen, Dorret I. Boomsma, Nicola Segata, and Peer Bork. 2018. “Selective Maternal Seeding and Environment Shape the Human Gut Microbiome.” *Genome Research* 28 (4): 561–68.
Kumar, Ranjit, Nengjun Yi, Degui Zhi, Peter Eipers, Kelly T. Goldsmith, Paula Dixon, David K. Crossman, et al. 2017. “Identification of Donor Microbe Species That Colonize and Persist

Long Term in the Recipient after Fecal Transplant for Recurrent Clostridium Difficile." *NPJ Biofilms and Microbiomes* 3 (June): 12.
Lahti, Leo, Anne Salonen, Riina A. Kekkonen, Jarkko Salojärvi, Jonna Jalanka-Tuovinen, Airi Palva, Matej Orešič, and Willem M. de Vos. 2013. "Associations between the Human Intestinal Microbiota, Lactobacillus Rhamnosus GG and Serum Lipids Indicated by Integrated Analysis of High-Throughput Profiling Data." *PeerJ* 1 (February): e32.
Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.
Lang, Michel, Martin Binder, Jakob Richter, Patrick Schratz, Florian Pfisterer, Stefan Coors, Quay Au, Giuseppe Casalicchio, Lars Kotthoff, and Bernd Bischl. 2019. "mlr3: A Modern Object-Oriented Machine Learning Framework in R." *Journal of Open Source Software* 4 (44): 1903.
Lee, Sonny T. M., Stacy A. Kahn, Tom O. Delmont, Alon Shaiber, Özcan C. Esen, Nathaniel A. Hubert, Hilary G. Morrison, Dionysios A. Antonopoulos, David T. Rubin, and A. Murat Eren. 2017. "Tracking Microbial Colonization in Fecal Microbiota Transplantation Experiments via Genome-Resolved Metagenomics." *Microbiome* 5 (1): 50.
Leo, Stefano, Vladimir Lazarevic, Myriam Girard, Nadia Gaïa, Jacques Schrenzel, Victoire de Lastours, Bruno Fantin, et al. 2020. "Metagenomic Characterization of Gut Microbiota of Carriers of Extended-Spectrum Beta-Lactamase or Carbapenemase-Producing Enterobacteriaceae Following Treatment with Oral Antibiotics and Fecal Microbiota Transplantation: Results from a Multicenter Randomized Trial." *Microorganisms* 8 (6). https://doi.org/10.3390/microorganisms8060941.
Li, Ming, Pin Liang, Zhenzhen Li, Ying Wang, Guobin Zhang, Hongwei Gao, Shu Wen, and Li Tang. 2015. "Fecal Microbiota Transplantation and Bacterial Consortium Transplantation Have Comparable Effects on the Re-Establishment of Mucosal Barrier Function in Mice with Intestinal Dysbiosis." *Frontiers in Microbiology* 6 (July): 692.
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.
Li, Youxian, and Kenya Honda. 2021. "Towards the Development of Defined Microbial Therapeutics." *International Immunology*, July. https://doi.org/10.1093/intimm/dxab038.
Lloyd-Price, Jason, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A. Brantley Hall, Arthur Brady, et al. 2017. "Strains, Functions and Dynamics in the Expanded Human Microbiome Project." *Nature* 550 (7674): 61–66.
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.
Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J. Xavier, and Dirk Gevers. 2015. "ConStrains Identifies Microbial Strains in Metagenomic Datasets." *Nature Biotechnology* 33 (10): 1045–52.
Mahnert, Alexander, Christine Moissl-Eichinger, Markus Zojer, David Bogumil, Itzhak Mizrahi, Thomas Rattei, José Luis Martinez, and Gabriele Berg. 2019. "Man-Made Microbial Resistances in Built Environments." *Nature Communications* 10 (1): 968.
Mirdita, Milot, Lars von den Driesch, Clovis Galiez, Maria J. Martin, Johannes Söding, and Martin Steinegger. 2017. "Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments." *Nucleic Acids Research* 45 (D1): D170–76.
Moayyedi, Paul, Michael G. Surette, Peter T. Kim, Josie Libertucci, Melanie Wolfe, Catherine Onischi, David Armstrong, et al. 2015. "Fecal Microbiota Transplantation Induces Remission in Patients With Active Ulcerative Colitis in a Randomized Controlled Trial." *Gastroenterology* 149 (1): 102–9.e6.
Moss, Eli L., Shannon B. Falconer, Ekaterina Tkachenko, Mingjie Wang, Hannah Systrom,

Jasmin Mahabamunuge, David A. Relman, Elizabeth L. Hohmann, and Ami S. Bhatt. 2017. “Long-Term Taxonomic and Functional Divergence from Donor Bacterial Strains Following Fecal Microbiota Transplantation in Immunocompromised Patients.” *PloS One* 12 (8): e0182585.

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.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2020. “Vegan: Community Ecology Package.” https://CRAN.R-project.org/package=vegan.

Olm, Matthew R., Alexander Crits-Christoph, Keith Bouma-Gregson, Brian A. Firek, Michael J. Morowitz, and Jillian F. Banfield. 2021. “inStrain Profiles Population Microdiversity from Metagenomic Data and Sensitively Detects Shared Microbial Strains.” *Nature Biotechnology*, January. https://doi.org/10.1038/s41587-020-00797-0.

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.

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.

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

Podlesny, Daniel, and W. Florian Fricke. 2020. “Microbial Strain Engraftment, Persistence and Replacement after Fecal Microbiota Transplantation.” *bioRxiv*. medRxiv. https://doi.org/10.1101/2020.09.29.20203638.

———. 2021. “Strain Inheritance and Neonatal Gut Microbiota Development: A Meta-Analysis.” *International Journal of Medical Microbiology: IJMM* 311 (3): 151483.

Quagliariello, Andrea, Federica Del Chierico, Sofia Reddel, Alessandra Russo, Andrea Onetti Muda, Patrizia D’Argenio, Giulia Angelino, et al. 2020. “Fecal Microbiota Transplant in Two Ulcerative Colitis Pediatric Cases: Gut Microbiota and Clinical Course Correlations.” *Microorganisms* 8 (10). https://doi.org/10.3390/microorganisms8101486.

Rands, Chris M., Harald Brüssow, and Evgeny M. Zdobnov. 2019. “Comparative Genomics Groups Phages of Negativicutes and Classical Firmicutes despite Different Gram-Staining Properties.” *Environmental Microbiology* 21 (11): 3989–4001.

Rossen, Noortje G., Susana Fuentes, Mirjam J. van der Spek, Jan G. Tijssen, Jorn H. A. Hartman, Ann Duflou, Mark Löwenberg, et al. 2015. “Findings From a Randomized Controlled Trial of Fecal Transplantation for Patients With Ulcerative Colitis.” *Gastroenterology* 149 (1): 110–18.e4.

Ruggles, Kelly V., Jincheng Wang, Angelina Volkova, Monica Contreras, Oscar Noya-Alarcon, Orlana Lander, Hortensia Caballero, and Maria G. Dominguez-Bello. 2018. “Changes in the Gut Microbiota of Urban Subjects during an Immersion in the Traditional Diet and Lifestyle of a Rainforest Village.” *mSphere* 3 (4). https://doi.org/10.1128/mSphere.00193-18.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30

(14): 2068–69.
Segata, Nicola. 2018. “On the Road to Strain-Resolved Comparative Metagenomics.” *mSystems* 3 (2). https://doi.org/10.1128/mSystems.00190-17.
Smillie, Christopher S., Jenny Sauk, Dirk Gevers, Jonathan Friedman, Jaeyun Sung, Ilan Youngster, Elizabeth L. Hohmann, et al. 2018. “Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation.” *Cell Host & Microbe* 23 (2): 229–40.e5.
Soldi, Sara, Sotirios Vasileiadis, Francesca Uggeri, Mariachiara Campanale, Lorenzo Morelli, Maria Vittoria Fogli, Fiorella Calanni, Maria Grimaldi, and Antonio Gasbarrini. 2015. “Modulation of the Gut Microbiota Composition by Rifaximin in Non-Constipated Irritable Bowel Syndrome Patients: A Molecular Approach.” *Clinical and Experimental Gastroenterology* 8 (December): 309–25.
Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology* 35 (11): 1026–28.
Suskind, David L., Mitchell J. Brittnacher, Ghassan Wahbeh, Michele L. Shaffer, Hillary S. Hayden, Xuan Qin, Namita Singh, et al. 2015. “Fecal Microbial Transplant Effect on Clinical Outcomes and Fecal Microbiome in Active Crohn’s Disease.” *Inflammatory Bowel Diseases* 21 (3): 556–63.
Suzek, Baris E., Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and UniProt Consortium. 2015. “UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches.” *Bioinformatics* 31 (6): 926–32.
Terveer, Elisabeth M., Tom van Gool, Rogier E. Ooijevaar, Ingrid M. J. G. Sanders, Eline Boeije-Koppenol, Josbert J. Keller, Aldert Bart, Ed J. Kuijper, and Netherlands Donor Feces Bank (NDFB) Study Group. 2020. “Human Transmission of Blastocystis by Fecal Microbiota Transplantation Without Development of Gastrointestinal Symptoms in Recipients.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 71 (10): 2630–36.
Tett, Adrian, Edoardo Pasolli, Giulia Masetti, Danilo Ercolini, and Nicola Segata. 2021. “Prevotella Diversity, Niches and Interactions with the Human Host.” *Nature Reviews. Microbiology* 19 (9): 585–99.
Tito, Raul Y., Samuel Chaffron, Clara Caenepeel, Gipsi Lima-Mendez, Jun Wang, Sara Vieira-Silva, Gwen Falony, et al. 2019. “Population-Level Analysis of Blastocystis Subtype Prevalence and Variation in the Human Gut Microbiota.” *Gut* 68 (7): 1180–89.
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.
Valles-Colomer, Mireia, Aitor Blanco-Míguez, Paolo Manghi, Francesco Asnicar, Federica Armanini, Fabio Cumbo, Kun D. Huang, et al. n.d. “The Person-to-Person Transmission Landscape of the Gut and Oral Microbiomes.” *Under Submission*.
Van Rossum, Thea, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. 2020. “Diversity within Species: Interpreting Strains in Microbiomes.” *Nature Reviews. Microbiology* 18 (9): 491–506.
Vaughn, Byron P., Tommi Vatanen, Jessica R. Allegretti, Aiping Bai, Ramnik J. Xavier, Joshua Korzenik, Dirk Gevers, Amanda Ting, Simon C. Robson, and Alan C. Moss. 2016. “Increased Intestinal Microbial Diversity Following Fecal Microbiota Transplant for Active Crohn’s Disease.” *Inflammatory Bowel Diseases* 22 (9): 2182–90.
Watson, A. R., J. Fuessel, I. Veseli, and J. Z. DeLongchamp. 2021. “Adaptive Ecological Processes and Metabolic Independence Drive Microbial Colonization and Resilience in the Human Gut.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.03.02.433653v2.abstract.

Weimann, Aaron, Kyra Mooren, Jeremy Frank, Phillip B. Pope, Andreas Bremges, and Alice C. McHardy. 2016. “From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer.” *mSystems* 1 (6). https://doi.org/10.1128/mSystems.00101-16.

Willmann, Matthias, Maria J. G. T. Vehreschild, Lena M. Biehl, Wichard Vogel, Daniela Dörfel, Axel Hamprecht, Harald Seifert, Ingo B. Autenrieth, and Silke Peter. 2019. “Distinct Impact of Antibiotics on the Gut Microbiome and Resistome: A Longitudinal Multicenter Cohort Study.” *BMC Biology* 17 (1): 76.

Wilson, Brooke C., Tommi Vatanen, Thilini N. Jayasinghe, Karen S. W. Leong, José G. B. Derraik, Benjamin B. Albert, Valentina Chiavaroli, et al. 2021. “Strain Engraftment Competition and Functional Augmentation in a Multi-Donor Fecal Microbiota Transplantation Trial for Obesity.” *Microbiome* 9 (1): 107.

Xiao, Yandong, Marco Tulio Angulo, Songyang Lao, Scott T. Weiss, and Yang-Yu Liu. 2020. “An Ecological Framework to Understand the Efficacy of Fecal Microbiota Transplantation.” *Nature Communications* 11 (1): 3329.

Zellmer, Caroline, Mohamad R. A. Sater, Miriam H. Huntley, Majdi Osman, Scott W. Olesen, and Bharat Ramakrishna. 2020. “Shiga Toxin–Producing Escherichia Coli Transmission via Fecal Microbiota Transplant.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 72 (11): e876–80.

Zeng, M. Y., N. Inohara, and G. Nuñez. 2017. “Mechanisms of Inflammation-Driven Bacterial Dysbiosis in the Gut.” *Mucosal Immunology* 10 (1): 18–26.

Zhao, Hui-Jun, Xi Luo, Yi-Chao Shi, Jian-Feng Li, Fei Pan, Rong-Rong Ren, Li-Hua Peng, et al. 2020. “The Efficacy of Fecal Microbiota Transplantation for Children With Tourette Syndrome: A Preliminary Study.” *Frontiers in Psychiatry / Frontiers Research Foundation* 11 (December): 554441

Chapter 5 | Other contributions

5.1 | A novel computational tool for profiling of carbohydrate-active enzymes in the human gut and its application in colorectal cancer cohorts

Carbohydrate metabolism of gut microbes is of great importance to understand nutrition and host-microbe interactions, yet annotation of carbohydrate-active enzymes (CAZy) is based on ill-adjusted statistical cutoffs and profiling of these gene families in gut microbes is cumbersome. Quinten Ducarmon, a visiting PhD student from the University of Leiden, and me have been developing a CAZy profiling tool with better-adjusted statistical cutoffs for more precise annotation as well as a hierarchical substrate classification scheme for CAZy families to facilitate downstream interpretation of results.

In this project, I have been responsible for the statistical evaluation of HMMs (cross-validation of HMMs to optimize E-values for each CAZy class specifically). I have also been assisting Quinten in the implementation of his scripts and have guided the project conceptually.

This manuscript is currently in preparation for submission.

Development of a novel computational tool for profiling of carbohydrate-active enzymes in the human gut and its application in colorectal cancer cohorts

Ducarmon, Q. R.*, Karcher, N.*, H.L.P. Tygat, C. Schudoma, G. Zeller
* Equal contributions

In preparation for submission

Abstract. Carbohydrate-active enzymes (CAZymes) are essential for the synthesis and breakdown of (complex) glycans and glycoconjugates. They are present in all living species, but are especially diverse in bacteria. In the gut microbiome, CAZymes are crucial for metabolizing complex carbohydrates of dietary and host origin, such as fiber and mucins, respectively. Currently, dbCAN2 is the most widely used computational tool for annotation of CAZymes in genomic data, but it cannot be directly applied to metagenomic data. dbCAN2 can identify protein sequences that are similar to those present in the CAZy database (the most comprehensive data and knowledge base on CAZymes) using Hidden Markov models (HMMs) specifically built for each CAZyme (sub-)family. However, detection accuracy (E-values cutoffs) has not been calibrated for these HMMs. A second challenge for wide application of this tool for metagenome analysis is the lack of systematic substrate annotations for CAZyme families, which are currently primarily grouped based on amino acid

sequence similarity. A hierarchical annotation of CAZyme substrates would however be needed for functional interpretation of CAZyme profiles. To close this gap, the main aim of this study was to build the first tool for computing CAZyme profiles from shotgun metagenomic data, which can be interpreted in terms of substrate specificities. This entailed optimization of HMM E-values for precise detection of CAZymes and construction of a novel hierarchical substrate scheme to facilitate functional interpretation. Application of this tool using data from eight different colorectal cancer (CRC) cohorts revealed that CRC metagenomes were enriched in microbial CAZymes involved in glycosaminoglycan metabolism (p-value 7.44e^-04) and in peptidoglycan metabolism (3.52e^-02), and depleted in CAZymes involved in dietary fiber metabolism (p-value 3.68e^-04) as compared to control metagenomes, suggesting that known dietary risk factors, such as increased meat consumption/decreased fiber consumption in CRC, are reflected in the gut microbial CAZy repertoire.

5.2 | Sulfoquinovose is a select nutrient of prominent bacteria and a source of hydrogen sulfide in the human gut

Hanson *et al.* conducted a series of experiments that lead to the discovery that sulfoquinovose (SQ) - a sulfonic acid derivative of glucose contained in leafy greens - can be metabolized by certain gut microbes, leading to the generation of hydrogen sulfide. Based on their initial *in vitro* data, *E. rectale* was one of the main candidates with SQ metabolism potential.

To expand the understanding of SQ metabolism in the human gut, I constructed Hidden Markov Models for SQ metabolism genes and conducted a comprehensive search for these genes in large MAG collections., confirming *E. rectale* and revealing several other human gut bacteria as putative SQ metabolizers in the human gut.

Sulfoquinovose is a select nutrient of prominent bacteria and a source of hydrogen sulfide in the human gut

Hanson, B.T., Dimitri Kits, K., Löffler, J., Burrichter, A.G., Fiedler, A., Denger, K., Frommeyer, B., Herbold, C.W., Rattei, T., Karcher, N., Segata, N., Schleheck, D., Loy, A., 2021
ISME J. 15, 2779–2791

Abstract. Responses of the microbiota to diet are highly personalized but mechanistically not well understood because many metabolic capabilities and

interactions of human gut microorganisms are unknown. Here we show that sulfoquinovose (SQ), a sulfonated monosaccharide omnipresent in green vegetables, is a selective yet relevant substrate for few but ubiquitous bacteria in the human gut. In human feces and in defined co-culture, *Eubacterium rectale* and *Bilophila wadsworthia* used recently identified pathways to cooperatively catabolize SQ with 2,3-dihydroxypropane-1-sulfonate as a transient intermediate to hydrogen sulfide (H2S), a key intestinal metabolite with disparate effects on host health. SQ-degradation capability is encoded in almost half of *E. rectale* genomes but otherwise sparsely distributed among microbial species in the human intestine. However, re-analysis of fecal metatranscriptome datasets of four human cohorts showed that SQ degradation (mostly from *E. rectale* and *Faecalibacterium prausnitzii*) and H2S production (mostly from *B. wadsworthia*) pathways were expressed abundantly across various health states, demonstrating that these microbial functions are core attributes of the human gut. The discovery of green-diet-derived SQ as an exclusive microbial nutrient and an additional source of H2S in the human gut highlights the role of individual dietary compounds and organosulfur metabolism on microbial activity and has implications for precision editing of the gut microbiota by dietary and prebiotic interventions.

5.3 | Understanding the functional repertoire and within-community genetic polymorphism of uncharacterized species in the human gut from MAGs

In 2019, Pasolli *et al.* published a large collection of MAGs assembled from human gut microbiomes. Among other analyses, this also included the characterization and contrasting of the functional potential of MAGs.

In this context I annotated the ORFs of all 150,000 MAGs with uniprot information and visualized the high-level functional differences between MAGs via ordination techniques. Furthermore I assessed the within-community genetic heterogeneity for MAGs by mapping reads against species core gene sequences (per MAG) followed by quantification of polymorphic sites.

Similarly, for the comparative genomics study of Tett *et al.*, I was involved in functional annotation and comparative functional analysis of different *Prevotella* clades.

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

Pasolli, E., Asnicar, F. *, Manara, S. * , Zolfo, M. *, Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M.C., Rice, B.L., DuLong, C., Morgan, X.C., Golden, C.D., Quince, C., Huttenhower, C., Segata, N., 2019
* Equal Contribution | Cell 176, Issue 3, pages 649-662

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

The *Prevotella copri* complex comprises four distinct clades underrepresented in westernized populations

Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., De Filippis, F., Magnabosco, C., Bonneau, R., Lusingu, J., Amuasi, J., Reinhard, K., Rattei, T., Boulund, F., Engstrand, L., Zink, A., Collado, M.C., Littman, D.R., Eibach, D., Ercolini, D., Rota-Stabelli, O., Huttenhower, C., Maixner, F., Segata, N., 2019. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host Microbe 26, 666–679.e7.

Abstract. *Prevotella copri* is a common human gut microbe that has been both positively and negatively associated with host health. In a cross-continent meta-analysis exploiting >6,500 metagenomes, we obtained >1,000 genomes and explored the genetic and population structure of *P. copri*. *P. copri* encompasses four

distinct clades (>10% inter-clade genetic divergence) that we propose constitute the *P. copri* complex, and all clades were confirmed by isolate sequencing. These clades are nearly ubiquitous and co-present in non-Westernized populations. Genomic analysis showed substantial functional diversity in the complex with notable differences in carbohydrate metabolism, suggesting that multi-generational dietary modifications may be driving reduced prevalence in Westernized populations. Analysis of ancient metagenomes highlighted patterns of *P. copri* presence consistent with modern non-Westernized populations and a clade delineation time predating human migratory waves out of Africa. These findings reveal that *P. copri* exhibits a high diversity that is underrepresented in Western-lifestyle populations.

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

Thomas, Manghi *et al.* performed a meta-analysis of colorectal cancer metagenomics cohorts, showing a consistent gut microbial signature of people affected by colorectal cancer over several shotgun metagenomic cohorts.

For this work, I contributed dataset collection and data exchange with a competing research group with which this paper was published back-to-back.

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

Thomas, A.M.*, Manghi, P.*, Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Wirbel, J., Schrotz-King, P., Ulrich, C.M., Brenner, H., Arumugam, M., Bork, P., Zeller, G., Cordero, F., Dias-Neto, E., Setubal, J.C., Tett, A., Pardini, B., Rescigno, M., Waldron, L., Naccarati, A., Segata, N., 2019
* Authors contributed equally. Nat. Med. 25, 667–678.

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.

5.5 | Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species

Manara *et al.* performed a MAG-based study to explore uncharted microbial complexity in the non-human primate gut microbiome, revealing many previously uncharacterized bacterial species in the gut of non-human primates, most of which are not shared with humans.

In this context, I supervised a Master's student which assisted the analysis. I furthermore assisted in understanding if we can find signals of co-diversifying clades between humans and non-human primates akin to what Moeller *et al.* have done.

Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species

Manara, S., Asnicar, F., Beghini, F., Bazzani, D., Cumbo, F., Zolfo, M., Nigro, E., Karcher, N., Manghi, P., Metzger, M.I., Pasolli, E., Segata, N., 2019

Genome Biol. 20, 299.

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

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

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

Chapter 6 | Conclusions and Outlook

In this thesis, I set out to better understand the genetic and functional diversity as well as the transmission dynamics of human gut commensal strains. I first performed two proof-of-concept large-scale comparative genomics studies of *Eubacterium rectale* and *Akkermansia muciniphila* (Chapter 2 and Chapter 3), showing that newly generated MAGs from human gut shotgun metagenomes can be leveraged to conduct such studies in a similar way to those previously done on isolate genomes, but at a substantially larger scale. These studies are among the first relying mostly on MAGs and extended our understanding of the strain-level genetic variation of the two selected commensals both within species as well as between species. They further yielded insights into biogeographic stratification, association patterns with human and non-human hosts, and revealed entirely novel clades that would have been erroneously conflated into single species based on their 16S rRNA gene sequence. The discovery of discrete structural polysaccharide synthesis loci variants within strains of the same species of both bacteria further illustrate the resolution and fidelity offered by MAGs. Overall, these studies call for similar analyses on the hundreds of other human gut bacterial species that have not yet been examined at large scale and high resolution.

Next, I aimed to verify whether strain-resolved metagenomics can be employed to track the transmission of microbiome members between individuals and whether specific transmission patterns can be identified. To this end, I focused on microbiome transmission with fecal microbiota transplantation (FMT) studies in order to assess the mechanisms and extent to which incoming strains engraft in a microbial community (Chapter 4). I conducted a strain-level meta-analysis of 24 FMT cohorts, in which I could show that antibiotic intake and infectious diseases (recurrent *C. difficile* or multi-drug resistant bacterial overgrowth) are associated with a much higher propensity of the recipient to take up strains from the FMT donor. Importantly, strain engraftment was significantly associated with clinical success, which was previously proposed in the context of ulcerative colitis, but had not been directly shown before. I furthermore designed a machine learning framework to predict the post-FMT microbiome that can also be used to select the optimal donor to maximize/minimize a specific post-FMT microbiome feature of interest such as bacterial richness or proteobacterial species count. Overall, this work revealed differences in strain engraftment characteristics over study populations and taxonomic groups and presents a framework that represents a first step towards facilitating better clinical FMT outcomes.

The expertise in genome-centric, strain-level analysis I gained through working on these projects allowed me to also contribute to other studies about carbohydrate-active enzyme identification in microbial genomes, novel sulfur-carbohydrate metabolism routes in the human gut as well as functional comparisons of a variety of MAGs from human- and non-human primate guts (Chapter 5).

Together, my work has leveraged state-of-the-art computational strain-level analysis tools to advance our genomic understanding of human gut microbes, opening up new avenues to conduct targeted comparative genomics studies to better understand specific human gut microbes of interest. It has also sparked the development of a novel approach to select donor individuals for FMT, which might provide practical clinical benefits in the future.

In the following sections I would like to discuss and provide an outlook for some specific findings I made during my thesis.

Human gut MAGs are a suitable and valuable resource to conduct comparative genomics studies

In Chapters 2 and 3, I relied predominantly on MAGs as a resource to set up and conduct comparative genomics studies. A substantial fraction of MAGs of recent high-throughput metagenomic assembly efforts are designated to be of high quality, which means they are estimated to be more than 90% complete and less than 5% contaminated according to state-of-the-art tools (Parks et al. 2015). In line with this, a study showcased the fidelity of MAGs by concomitant isolate genome reconstruction from the same community and subsequent comparison of MAGs against the isolate genome ‘ground truth’ (Alneberg et al. 2018). Conversely, other studies show that, while a large fraction of genetic material is faithfully recovered by MAGs, genetically divergent sequences are not correctly binned (Nelson, Tully, and Mobberley 2020; Meziti et al. 2021), and metagenomic assembly and binning has been put into question from a conceptual point of view (Shoemaker, Chen, and Garud 2021; Garud et al. 2019; Van Rossum et al. 2020; Meziti et al. 2021; Orakov et al. 2021). Here, I would like to argue why some human gut MAGs can be expected to be truly of high quality and also briefly discuss future research perspectives for MAG quality control.

In a highly-cited community-driven, thus largely impartial, metagenomics methods benchmarking study, assembly and binning software was tested on in-silico datasets of varying levels of taxonomic complexity, coverage distributions and strain mixtures (Sczyrba et al. 2017). Assemblers consistently produced contigs corresponding to the dominant strain in cases where community complexity was low, a single strain was much more highly abundant than other conspecific strains, and that strain was sufficiently abundant (Sczyrba et al. 2017). Importantly, in cases in which these conditions were not met (particularly when closely related strains were co-present at similar abundances), assemblers yielded far smaller fractions of ground truth genomes, i.e. they produced more incomplete assemblies of the genomes of individual species. As such, assemblers appear to work conservatively: they have a propensity to yield incomplete assemblies, which are very unlikely to be binned into complete MAGs.

Similarly, metagenomic binners performed much better in cases of single-strain dominance than when multiple strains were co-present in a sample (Sczyrba et al. 2017). The human gut is an environment where most species are dominated strongly by one strain (Truong et al. 2017). Furthermore, even though thousands of bacterial species can be found in the human gut, owing to extremely large inter-individual differences in gut community composition and unequal abundance distributions of bacterial species, each individual human gut harbours a select few species at comparatively high abundances (Pasolli et al. 2017). This, together with the continuing rise in numbers of publicly available shotgun metagenomes, provides good conditions for a growing number of high-quality MAGs to be produced from human gut shotgun metagenomes.

Despite these considerations, there are drawbacks of MAGs that need to be discussed. First of all, even under ideal conditions, metagenomic assemblers and binners will never yield perfect genomes (Sczyrba et al. 2017): owing to the lack of long-range information sharing of Illumina short-read sequences, assemblies will by definition be discontinuous if repeat structures are present, although this is also true for isolate genome assembly (Peona et al. 2021). Binning tools - essentially clustering contigs based on coverage and tetranucleotide frequencies - will fail to correctly assign contigs to bins when genomic elements are duplicated or when sequence composition diverges too much from the bulk of the genome. Furthermore, current state-of-the-art MAG quality assessment tools (checkM (Parks et al. 2015)) work by determining the fraction of present and duplicated clade-specific marker genes to assess MAG completeness and contamination, respectively. The drawback of this is that the quality of the accessory genome cannot be assessed. Recently, a method was presented that looks at the intra- and inter-contig ‘taxonomic congruence’ to judge mis-assemblies and binnings (Orakov et al. 2021), which is a step into the right direction, although it is not clear how this tool can handle bins containing frequently horizontally-transferred genetic elements.

Recently, Illumina short-read sequencing was combined with Nanopore sequencing to yield complete and closed microbial genomes directly from microbial communities (Moss, Maghini, and Bhatt 2020), illustrating the potential value of novel sequencing technologies. Nanopore sequencing - which determines the sequence of DNA by measuring voltage changes induced by individual nucleotides passing through a protein pore (Wang, Yang, and Wang 2014) - is less accurate compared to Illumina sequencing, but offers invaluable long-range mapping information which reduces the computational complexity and allows a higher contiguity of genome assemblies when combined with more accurate Illumina short read sequences. An alternative new long read sequencing technology, PacBio single molecule sequencing, which performs sequencing-by-synthesis on individual DNA molecules (Eid et al. 2009), was initially

plagued by high error rates. Yet, recently further technological advances permit repeated sequencing of the same molecule, which dramatically reduces error rates (Wenger et al. 2019). Regardless of which long read sequencing technology will establish itself as dominant in the coming years, the long range information contained in them will provide invaluable data for metagenomic assembly.

Taken together, high-quality gut MAGs are well suited for comparative genomics studies and represent a valuable resource to genomically explore human gut bacteria. I expect that future studies will leverage MAGs extensively to explore and generate testable hypotheses for species of interest, such as disease-associated bacteria (see below). New sequencing technologies and improvements in assembly and binning technology as well as MAG quality assessment will likely contribute to further improvements of MAGs.

Disease-associated bacteria are appealing candidates for future MAG-based genomics analyses

Many bacterial species have previously been associated with host disease phenotypes. For example, more than 20 bacterial species were consistently enriched in meta-analyses of colorectal cancer (CRC) case-control studies (Wirbel et al. 2019; Thomas et al. 2019). These associations are not perfect and bacterial subspecies genetic variation is a likely factor contributing to differences in virulence and thus differences in disease association. Furthermore, while many species are associated with disease, mechanistic knowledge is typically sparse: In CRC, mechanisms contributing to carcinogenesis or cancer progression have been described for only 3 bacterial species: *Fusobacterium nucleatum (Xu et al. 2007)*, *E. coli (Pleguezuelos-Manzano et al. 2020; Nougayrède et al. 2006)* and *Bacteroides fragilis (Cheng, Kantilal, and Davamani 2020)*. Importantly, only some *E. coli* and *B. fragilis* strains possess these carcinogenic properties. Together, this illustrates that genome-based explorative efforts are especially appealing for disease-associated bacteria, and recent large-scale MAG resources are expected to contribute important genomic information to such studies, especially for the subset of disease-associated bacteria that have no isolate genome available (Wirbel et al. 2019; Nayfach et al. 2019; Pasolli et al. 2019) which is expected to grow even further from these MAG resources.

A valuable next step would be to perform comparative genomics on CRC-associated bacteria with the specific aim to annotate and understand virulence factors such as adhesion proteins, secretion systems and their possible role in CRC. Similarly, one could undertake comparative genomics studies of disease-associated microbial species in inflammatory bowel disease, liver cirrhosis or pancreatic cancer. Importantly, MAGs (or genetic variants of interests identified from them) can be mapped back to host

phenotype. This facilitates guided comparisons of MAGs extracted from a cancer patient to those from non-cancer patients. Indeed, this approach might also facilitate a pan-microbiome comparison where the search for disease associations is expanded to the entire set of microbes in the gut. Another important consideration here is that some of these species might not reach high abundances frequently enough, which might mean that retrieving MAGs for those species is challenging. To circumvent this problem, one might have to view the MAGs one is able to recover as a 'discovery set' to then later quantify certain genomic elements of interest using read-mapping based approaches. This would entail quantifying the presence and/or abundance as well as the genetic makeup of certain genomic elements of interest (that one has discovered through a smaller MAG set) via mapping reads back to contigs or gene catalogues built thereof.

Human-microbe co-diversification in the human gut

In Chapter 2, I show that *Eubacterium rectale* strain-level subspecies cluster extensively by geography with putative concordance of microbial and host phylogeny, suggesting host-microbe co-diversification. The phenomenon of host-microbe co-diversification has been extensively shown and validated for *Helicobacter pylori* living in the human stomach (Falush et al. 2003; Linz et al. 2007). More generally, some but not all human gut bacteria show patterns of biogeographical genetic stratification (Truong et al. 2017; Costea et al. 2017). Previous work has also shown that the phylogeny of certain gut Bacteroidetes families mirror the phylogenies of great apes (gorillas, chimps, bonobos, humans), showing that certain bacterial taxa co-diversified with great apes - including humans - for several million years (Moeller et al. 2016). Furthermore, a more recent study has tested many human gut species for signals of co-diversification and found *E. rectale* as well as a few other species to have phylogenies that are showing signs of co-phylogeny with the human host (Suzuki et al. 2021). While the underlying processes that caused these phenomena might be of different natures and evidently occurred over different time ranges, they all have resulted in an association between host- and host-associated bacterial genotypes.

In theory, geographic clustering of microbial species could be the result not only of co-migration/co-diversification, but the result of selection of certain strains due to differences in host diet, lifestyle or genetics (Garud and Pollard 2020). While it is possible that selection effects have some influence on the biogeographic patterns in human gut bacteria, it seems implausible that they are solely responsible because they cannot explain how bacterial phylogenies mirror those of hosts. Furthermore, this being true would require a ubiquitous distribution of gut bacterial species over the globe akin to the 'everything is everywhere but the environment selects' hypothesis of bacterial ecology (O'Malley 2007). This has recently been disproven to be uniformly true even for

non-host-associated, free-living bacteria (Martiny et al. 2006; Papke and Ward 2004) which can naturally disperse on the globe much more freely than host-associated bacteria.

The fact that only some bacteria show co-diversification might be for a variety of possibly inter-relating reasons, for example because they are not specific to humans, they did not co-migrate with early humans out of Africa, or they exhibit different vertical/horizontal transmission patterns. We might also miss the detection of some co-diversification signal because depending on the time scale of co-diversification as well as mutation rates (Garud et al. 2019), different taxonomic levels have to be considered: as stated above, Moeller et al. find signals of co-diversification of great apes commensals in certain bacterial *families* and not in species (Moeller et al. 2016). Furthermore, horizontal gene transfer could 'dilute' the true phylogenetic signal in genes of species that exhibit extensive horizontal gene transfer, and increasingly cosmopolitan lifestyles in westernized populations are likely to lead to more frequent horizontal strain sharing between individuals of different ethnicities, thus leading to a gradual loss of co-diversification signal in host-associated bacteria over time.

Taken together, it seems likely that the geographical stratification of some gut bacterial species has its roots in co-diversification processes mediated through co-migration with early humans. In the future, increased sampling especially of previously unsampled, non-industrialized societies together with host genotyping will allow researchers to more rigorously test signals of co-diversification. A better understanding of evolutionary processes and transmission patterns in the human gut will aid in interpreting and contextualizing co-phylogeny patterns (or lack thereof) as well. Furthermore, estimation of diversification rates of bacteria would allow researchers to date bacterial phylogenies events to understand if the age of clade coalescence coincides with time information we have about human migration routes.

Investigating the structural polysaccharide landscape in the human gut

In both large scale comparative genomics works that I presented here (Chapters 2 and 3), I discovered large-scale and discrete operon variants that are involved in the buildup of structural carbohydrates, possibly exopolysaccharide (EPS) synthesis (*E. rectale)* or Lipid A synthesis (*A. muciniphila*). While the specific function of these genomic structures remains to be confirmed, differences in EPS and Lipid A/Lipopolysaccharide synthesis operons in human gut bacteria could play a role in host-microbe interactions, a prospect I would like to discuss below.

EPS are formally defined as bacterially-produced polysaccharides secreted into the surroundings (Schmid, Sieber, and Rehm 2015), either to be incorporated into a

bacterial capsule or to be released completely (Angelin and Kavitha 2020), where they play a role in aggregation, adhesion and pathogenicity of microbes. Exopolysaccharides have been experimentally explored in a variety of soil organisms and pathogens and also in the human gut commensals *Bifidobacteria* spp. and *Lactobacillus* spp. (Oerlemans et al. 2021). Yet, the EPS synthesis potential in the human gut has so far not been genomically explored over the entire taxonomic range.

Lipid A is the precursor subunit of Lipopolysaccharide (LPS), which is an important constituent of the cell membrane of gram-negative bacteria (Raetz and Whitfield 2002). Lipid A consists of two glucosamine units covalently bound with several acyl chains (Raetz et al. 2009). Lipid A is a very potent endotoxin detected by Toll Like Receptor 4 (TLR4) (Raetz and Whitfield 2002), but depending on the specific kind of acyl chains attached to the glucosamine base, immune responses via TLR4 can be attenuated greatly (Whitfield and Trent 2014; Kim et al. 2007). In the context of the gut microbiome, these differences in LPS immunogenicity have been linked to differences in autoimmune disorder prevalences between Finish/Estonian and Russian children (Vatanen et al. 2016) and differences in host innate immune reactions to gut commensals have been previously implicated in IBD (Pasternak et al. 2010). Yet, so far a comprehensive understanding of LPS synthesis operon spread and genetic variation as well as the immunogenic potential of synthesized molecules in the human gut is missing.

Using MAG resources, one could rigorously search for structural polysaccharide sugar synthesis operons in human gut bacteria. Both EPS and LPS synthesis machinery is encoded in operons, which means that search strategies leveraging local enrichment of certain genes can be used. Known EPS systems from Bifidobacteria and Lactobacilli are enriched in glycosyltransferases, which means the same is probably true in hitherto undiscovered EPS operon systems. Biosynthetic gene cluster (BGC) prediction tools have shown that the ‘Saccharide’ BGC class is among the most prevalent in human gut bacteria (Cimermancic et al. 2014; Carroll et al. 2021), and BGC prediction tools are expected to facilitate the search for EPS/LPS synthesis operons. For downstream analysis, one could attempt to group EPS/LPS synthesis operon variants at varying degrees of granularity and then associate them with host phenotypes of interest. In the context of LPS, one could furthermore attempt to reconstruct the immunogenicity of encoded LPS variants using known LPS synthesis pathway information (Kanehisa and Goto 2000), which could then be cross-referenced with disease phenotypes of inflammatory diseases like inflammatory bowel disease (IBD).

Machine learning might facilitate engineering of post-FMT parameters through rational donor selection

FMT is a highly effective treatment option against *Clostridiodes difficile* infections, and although it is considered promising for other microbiome-associated conditions, its efficacy varies largely. In Chapter 4, I present the results of an FMT meta-analysis of 20 case-control cohorts over different diseases. As part of the efforts to improve clinical outcomes after FMT, Ott et al. showed that sterile-filtered donor stool successfully treated recurrent *C. difficile* (rCDI) in 5 patients (Ott et al. 2017), which suggests that something other than bacterial strains (metabolites, enzymes, or viruses) could play a mechanistic role in combating *C. difficile* colonization. Indeed, two studies have shown a convergence of recipient and donor viromes after FMT (Zuo et al. 2018; Draper et al. 2018) and recent resources are beginning to comprehensively describe the virome in the human gut (Nayfach et al. 2021). One could extend compositional comparisons of donor and recipient microbiomes with viral taxonomic information. Cross-referencing of phage DNA to *C. difficile* CRISPR spacers could provide tentative evidence of an interaction between specific phages and *C. difficile*, which could be supported by finding those entities associated with clinical success or even decreased prevalence or abundance in *C. difficile* patients post-FMT or in corresponding donors compared to the healthy population. Given the relatively high rate of success in rCDI and concomitant imbalance in retrospective rCDI cohorts, this meta-analysis (and other rCDI-focused studies) would strongly benefit from incorporation of more datasets to increase the number of non-responding cases.

In Chapter 4 I also trained random forest models to predict the post-FMT microbiome (presence/absence) using features from donors, recipients (pre-FMT) as well as microbial and other features, showing that the post-FMT microbiome can generally be predicted. Feature importance analysis showed that bacterial abundances in the pre-FMT and donor samples are the most informative features.

Next, I set out to understand if we can use the above-mentioned machine learning models to pinpoint specific donor individuals that are particularly suitable as FMT donors. To this end, I set up a simulation framework in which I randomly exchanged actual donors with randomly selected ones from a different FMT triad. Following this permutation of donor-related information, I then predicted the post-FMT microbiome of these simulated FMT instances and checked which donor individual leads to a consistent increase in predicted post-FMT richness upon implantation into an FMT instance. This revealed that indeed specific donor individuals lead to consistent increase in predicted post-FMT richness, which was previously tentatively associated with clinical success in ulcerative colitis (Smith et al. 2021). An interesting direction in this context to explore further would be an alternative model architecture: currently, each data point corresponds to a species in an FMT instance (an FMT instance being defined as the pre-FMT recipient sample, donor sample and a single post-FMT sample),

and one model is being trained on all species together. Alternatively, one could train many models, i.e. one model per species. This could benefit the analysis two-fold: first, it could increase predictive performance because it would force the model to pick up differences in engraftment between clades (something that feature importance analysis shows is not being used by models to predict post-FMT microbial presence/absence, which might be because there are too many discrete levels in taxonomic labels); secondly, I could add features of other bacterial species to these models such as the abundance in pre-FMT and donor samples, which is expected to increase predictive performance if there are groups of species that tend to engraft together (or tend to co-exclude).

While microbial richness is an important ecological parameter dictating community stability, I believe the simulation framework I describe in the previous paragraph represents a generalizable way to select donor individuals to modulate not only richness but *any* post-FMT feature of choice: for example, our model could be used to select donors that minimize/maximize the chance of a set of species being detected in the post-FMT sample, which will be particularly useful with a better understanding of the association of particular species with treatment outcome.

Acknowledgements

First of all, I'd like to thank Nicola and Georg for the chance to do my PhD under their supervision. The discussions we had and the feedback you gave me were crucial to my development as a scientist. This PhD wasn't easy for me, but the struggle has given me the opportunity to better myself, and for that I'm very grateful to both of you. I'd also like to thank Mireia for the guidance she has offered towards the end of my PhD, especially during the writing period, which was extremely helpful to me. Thank you all!

Next I'd like to thank my mother for her easy-going nature, support and excellent cooking. Your cooking skills have always made me appreciate food, and especially during difficult times in my PhD cooking has proved a very helpful hobby! I'd also like to thank my brother Martin for the fun biking experiences we've had throughout my PhD, particularly the gnarly all-day tours in Trentino (including the horses that would help us fix our broken tires by licking our salty handlebars) and the fun roller coaster rides in Finale Ligure. It was super fun!

Next, I'd like to thank my dear friends Dimitris and Sophie who have been pillars of support both inside and outside the lab. Your kind nature saved my butt more than once. I'd also like to thank Hanna und Nils, who became dear friends to me in the past year. You, similarly, are wonderful people who I'm sure are looking into the brightest future possible! I’m extremely glad to have you all as friends. To many more years of climbing together!

Next I'd like to thank all the friends I made during my PhD, particularly Federica and Stefano, Jakob, Toby, Maral, Carlos and Dienty. We've had a great time doing science but also great talks, climbing sessions, dinners and board game nights. Every one of you has left an imprint on me as a scientist and as a person! You’re all awesome!

Finally, I'd like to thank everyone in the Segata and Zeller labs as well as everyone I forgot to mention in the heat of the moment of writing this text. I can assure you it's not due to lack of gratitude, but simply due to my ever-chaotic ways!

References for Chapter 6

Alneberg, Johannes, Christofer M. G. Karlsson, Anna-Maria Divne, Claudia Bergin, Felix Homa, Markus V. Lindh, Luisa W. Hugerth, et al. 2018. “Genomes from Uncultivated Prokaryotes: A Comparison of Metagenome-Assembled and Single-Amplified Genomes.” *Microbiome* 6 (1): 173.

Angelin, J., and M. Kavitha. 2020. “Exopolysaccharides from Probiotic Bacteria and Their Health Potential.” *International Journal of Biological Macromolecules* 162 (November): 853–65.

Carroll, Laura M., Martin Larralde, Jonas Simon Fleck, Ruby Ponnudurai, Alessio Milanese, Elisa Cappio, and Georg Zeller. 2021. “Accurate *de Novo* Identification of Biosynthetic Gene Clusters with GECCO.” *bioRxiv*. https://doi.org/10.1101/2021.05.03.442509.

Cheng, Wai Teng, Haresh Kumar Kantilal, and Fabian Davamani. 2020. “The Mechanism of Bacteroides Fragilis Toxin Contributes to Colon Cancer Formation.” *The Malaysian Journal of Medical Sciences: MJMS* 27 (4): 9–21.

Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, et al. 2014. “Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters.” *Cell* 158 (2): 412–21.

Costea, Paul I., Luis Pedro Coelho, Shinichi Sunagawa, Robin Munch, Jaime Huerta-Cepas, Kristoffer Forslund, Falk Hildebrand, Almagul Kushugulova, Georg Zeller, and Peer Bork. 2017. “Subspecies in the Global Human Gut Microbiome.” *Molecular Systems Biology* 13 (12): 960.

Draper, L. A., F. J. Ryan, M. K. Smith, J. Jalanka, E. Mattila, P. A. Arkkila, R. P. Ross, R. Satokari, and C. Hill. 2018. “Long-Term Colonisation with Donor Bacteriophages Following Successful Faecal Microbial Transplantation.” *Microbiome* 6 (1): 220.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. “Real-Time DNA Sequencing from Single Polymerase Molecules.” *Science* 323 (5910): 133–38.

Falush, Daniel, Thierry Wirth, Bodo Linz, Jonathan K. Pritchard, Matthew Stephens, Mark Kidd, Martin J. Blaser, et al. 2003. “Traces of Human Migrations in Helicobacter Pylori Populations.” *Science* 299 (5612): 1582–85.

Garud, Nandita R., Benjamin H. Good, Oskar Hallatschek, and Katherine S. Pollard. 2019. “Evolutionary Dynamics of Bacteria in the Gut Microbiome within and across Hosts.” *PLoS Biology* 17 (1): e3000102.

Garud, Nandita R., and Katherine S. Pollard. 2020. “Population Genetics in the Human Microbiome.” *Trends in Genetics: TIG* 36 (1): 53–67.

Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.

Kim, Ho Min, Beom Seok Park, Jung-In Kim, Sung Eun Kim, Judong Lee, Se Cheol Oh, Purevjav Enkhbayar, et al. 2007. “Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran.” *Cell* 130 (5): 906–17.

Linz, Bodo, François Balloux, Yoshan Moodley, Andrea Manica, Hua Liu, Philippe Roumagnac, Daniel Falush, et al. 2007. “An African Origin for the Intimate Association between Humans and Helicobacter Pylori.” *Nature* 445 (7130): 915–18.

Martiny, Jennifer B. Hughes, Brendan J. M. Bohannan, James H. Brown, Robert K. Colwell, Jed A. Fuhrman, Jessica L. Green, M. Claire Horner-Devine, et al. 2006. “Microbial Biogeography: Putting Microorganisms on the Map.” *Nature Reviews. Microbiology* 4 (2): 102–12.

Meziti, Alexandra, Luis M. Rodriguez-R, Janet K. Hatt, Angela Peña-Gonzalez, Karen Levy, and Konstantinos T. Konstantinidis. 2021. “The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against

Isolate Genomes Derived from the Same Fecal Sample.” *Applied and Environmental Microbiology* 87 (6). https://doi.org/10.1128/AEM.02593-20.
Moeller, Andrew H., Alejandro Caro-Quintero, Deus Mjungu, Alexander V. Georgiev, Elizabeth V. Lonsdorf, Martin N. Muller, Anne E. Pusey, Martine Peeters, Beatrice H. Hahn, and Howard Ochman. 2016. “Cospeciation of Gut Microbiota with Hominids.” *Science* 353 (6297): 380–82.
Moss, Eli L., Dylan G. Maghini, and Ami S. Bhatt. 2020. “Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing.” *Nature Biotechnology* 38 (6): 701–7.
Nayfach, Stephen, David Páez-Espino, Lee Call, Soo Jen Low, Hila Sberro, Natalia N. Ivanova, Amy D. Proal, et al. 2021. “Metagenomic Compendium of 189,680 DNA Viruses from the Human Gut Microbiome.” *Nature Microbiology* 6 (7): 960–70.
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*, March. https://doi.org/10.1038/s41586-019-1058-x.
Nelson, William C., Benjamin J. Tully, and Jennifer M. Mobberley. 2020. “Biases in Genome Reconstruction from Metagenomic Data.” *PeerJ* 8 (October): e10119.
Nougayrède, Jean-Philippe, Stefan Homburg, Frédéric Taieb, Michèle Boury, Elzbieta Brzuszkiewicz, Gerhard Gottschalk, Carmen Buchrieser, Jörg Hacker, Ulrich Dobrindt, and Eric Oswald. 2006. “Escherichia Coli Induces DNA Double-Strand Breaks in Eukaryotic Cells.” *Science* 313 (5788): 848–51.
Oerlemans, Marjolein M. P., Renate Akkerman, Michela Ferrari, Marthe T. C. Walvoort, and Paul de Vos. 2021. “Benefits of Bacteria-Derived Exopolysaccharides on Gastrointestinal Microbiota, Immunity and Health.” *Journal of Functional Foods* 76 (January): 104289.
O’Malley, Maureen A. 2007. “The Nineteenth Century Roots of ‘Everything Is Everywhere.’” *Nature Reviews. Microbiology* 5 (8): 647–51.
Orakov, Askarbek, Anthony Fullam, Luis Pedro Coelho, Supriya Khedkar, Damian Szklarczyk, Daniel R. Mende, Thomas S. B. Schmidt, and Peer Bork. 2021. “GUNC: Detection of Chimerism and Contamination in Prokaryotic Genomes.” *Genome Biology* 22 (1): 178.
Ott, Stephan J., Georg H. Waetzig, Ateequr Rehman, Jacqueline Moltzau-Anderson, Richa Bharti, Juris A. Grasis, Liam Cassidy, et al. 2017. “Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium Difficile Infection.” *Gastroenterology* 152 (4): 799–811.e7.
Papke, R. Thane, and David M. Ward. 2004. “The Importance of Physical Isolation to Microbial Diversification.” *FEMS Microbiology Ecology* 48 (3): 293–303.
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.
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.
Pasternak, Brad A., Sharon D’Mello, Ingrid I. Jurickova, Xiaonan Han, Tara Willson, Leah Flick, Lisa Petiniot, et al. 2010. “Lipopolysaccharide Exposure Is Linked to Activation of the Acute Phase Response and Growth Failure in Pediatric Crohn’s Disease and Murine Colitis.” *Inflammatory Bowel Diseases* 16 (5): 856–69.
Peona, Valentina, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, et al. 2021. “Identifying the Causes and Consequences of Assembly Gaps Using a Multiplatform Genome Assembly of a Bird-of-Paradise.” *Molecular Ecology*

Resources 21 (1): 263–86.
Pleguezuelos-Manzano, Cayetano, Jens Puschhof, Axel Rosendahl Huber, Arne van Hoeck, Henry M. Wood, Jason Nomburg, Carino Gurjao, et al. 2020. “Mutational Signature in Colorectal Cancer Caused by Genotoxic Pks+ E. Coli.” *Nature* 580 (7802): 269–73.
Raetz, Christian R. H., Ziqiang Guan, Brian O. Ingram, David A. Six, Feng Song, Xiaoyuan Wang, and Jinshi Zhao. 2009. “Discovery of New Biosynthetic Pathways: The Lipid A Story.” *Journal of Lipid Research* 50 Suppl (April): S103–8.
Raetz, Christian R. H., and Chris Whitfield. 2002. “Lipopolysaccharide Endotoxins.” *Annual Review of Biochemistry* 71: 635–700.
Schmid, Jochen, Volker Sieber, and Bernd Rehm. 2015. “Bacterial Exopolysaccharides: Biosynthesis Pathways and Engineering Strategies.” *Frontiers in Microbiology* 6 (May): 496.
Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. “Critical Assessment of Metagenome Interpretation-a Benchmark of Metagenomics Software.” *Nature Methods* 14 (11): 1063–71.
Shoemaker, William R., Daisy Chen, and Nandita R. Garud. 2021. “Comparative Population Genetics in the Human Gut Microbiome.” *Genome Biology and Evolution*, May. https://doi.org/10.1093/gbe/evab116.
Smith, Byron J., Yvette Piceno, Martin Zydek, Bing Zhang, Lara Aboud Syriani, Jonathan P. Terdiman, Zain Kassam, et al. 2021. “Clinical Efficacy and Increased Donor Strain Engraftment after Antibiotic Pretreatment in a Randomized Trial of Ulcerative Colitis Patients Receiving Fecal Microbiota Transplant.” *bioRxiv*. medRxiv. https://doi.org/10.1101/2021.08.07.21261556.
Suzuki, Taichi A., Liam Fitzstevens, Victor T. Schmidt, Hagay Enav, Kelsey Huus, Mirabeau Mbong, Bayode R. Adegbite, et al. 2021. “Codiversification of Gut Microbiota with Humans.” *bioRxiv*. https://doi.org/10.1101/2021.10.12.462973.
Thomas, Andrew Maltez, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, et al. 2019. “Metagenomic Analysis of Colorectal Cancer Datasets Identifies Cross-Cohort Microbial Diagnostic Signatures and a Link with Choline Degradation.” *Nature Medicine* 25 (4): 667–78.
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.
Van Rossum, Thea, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. 2020. “Diversity within Species: Interpreting Strains in Microbiomes.” *Nature Reviews. Microbiology* 18 (9): 491–506.
Vatanen, Tommi, Aleksandar D. Kostic, Eva d’Hennezel, Heli Siljander, Eric A. Franzosa, Moran Yassour, Raivo Kolde, et al. 2016. “Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans.” *Cell* 165 (4): 842–53.
Wang, Yue, Qiuping Yang, and Zhimin Wang. 2014. “The Evolution of Nanopore Sequencing.” *Frontiers in Genetics* 5: 449.
Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. “Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome.” *Nature Biotechnology* 37 (10): 1155–62.
Whitfield, Chris, and M. Stephen Trent. 2014. “Biosynthesis and Export of Bacterial Lipopolysaccharides.” *Annual Review of Biochemistry* 83 (February): 99–128.
Wirbel, Jakob, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, et al. 2019. “Meta-Analysis of Fecal Metagenomes Reveals Global Microbial Signatures That Are Specific for Colorectal Cancer.” *Nature Medicine* 25 (4): 679–89.
Xu, Minghua, Mitsunori Yamada, Mei Li, Hongqi Liu, Shu G. Chen, and Yiping W. Han. 2007.

"FadA from Fusobacterium Nucleatum Utilizes Both Secreted and Nonsecreted Forms for Functional Oligomerization for Attachment and Invasion of Host Cells*." *The Journal of Biological Chemistry* 282 (34): 25000–9.

Zuo, Tao, Sunny H. Wong, Kelvin Lam, Rashid Lui, Kitty Cheung, Whitney Tang, Jessica Y. L. Ching, et al. 2018. "Bacteriophage Transfer during Faecal Microbiota Transplantation in Clostridium Difficile Infection Is Associated with Treatment Outcome." *Gut* 67 (4): 634–43.