UNIVERSITY OF TRENTO

PhD Program in Biomolecular Sciences

Department of Cellular, Computational and Integrative Biology – CIBIO

37th Cycle

Transmission of human microbiome: Computational metagenomic tools and biomedical applications

Tutor
Prof. Nicola SEGATA
University of Trento

Advisor
Prof. Francesco ASNICAR
University of Trento

Ph.D. Thesis of

Michal PUNČOCHÁŘ

University of Trento

Academic Year 2023-2024

Declaration

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

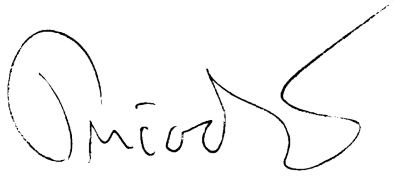


Table of contents

Abstract

The human gut microbiome is important for health and its disruption can lead to disease. Transmission from other people is the main way we obtain microbiomes in our gut. It has been established that the first microbes are acquired from mother upon birth and more recent findings show that people co-housing share a substantial portion of their microbiome. However, the impact of social contact on the developing gut microbiome after the initial seeding from the family remains unexplored. Fecal microbiota transplantation (FMT) is a clinical procedure to modulate a recipient's gut microbiome in order to improve health or response to treatment and can be viewed as a forced transmission of the whole microbial community. Despite many single cohort studies published, its success in treating *Clostridioides difficile* infection and promising preliminary results in cancer immunotherapy, the underlying mechanisms of microbial engraftment and the links to clinical outcomes remain to be understood, preventing informed clinical decisions. In my doctoral thesis I explore the development and application of computational methods of strain sharing to study microbiome transmission and engraftment in FMT.

During my PhD work I was continuously collecting metagenomic assembled genomes (MAGs) and clustering them to define species-level genome bins (SGBs) that allow for detecting transmission of otherwise uncharacterized species in our cohorts. I developed and maintained a computational pipeline for strain sharing detection among metagenomic samples. In Chapter 2 I applied the strain sharing pipeline to over 1,000 samples from 43 babies attending nursery for the first time with weekly microbiome sampling along with their educators and family. My pipeline supported the data analysis, which highlighted the importance of social contact in the development of the infant’s microbiome by observing substantial strain transmission among babies after only one month of attending the nursery. It enabled the tracking of individual strains spreading among the nursery shedding light to the dynamics of microbial transmission.

In Chapter 3 we pooled together all available FMT cohorts with shotgun metagenomic sequencing for the total of 226 donor-recipient pairs. I participated in the data and metadata collection, applied the strain sharing detection pipeline and analyzed the strain engraftment from donors to recipients. We showed that recipients using antibiotics or with an infectious disease both showed higher engraftment rates and discovered that the combined route of delivery (for example both capsules and colonoscopy) yields higher engraftment rates than a single route. Although the mechanism of action of FMT is disease specific, we found an overall link between engraftment rate and clinical success of the treatment. By looking at the engrafted species we observed a pattern of Bacteroidetes and Actinobacteria species engrafting at a higher rate than Firmicutes. My application of the strain sharing methodology and data analysis played an important role in better understanding the patterns of engraftment in FMT and by helping clinicians to make more informed decisions.

Overall my PhD work contributed to successful applications of computational methods of strain sharing detection to gain new insights into microbial transmission and clinical applications.

Chapter 1: Introduction

The human gut microbiome is important

Microorganisms, that is the bacteria, archaea, viruses, micro-eukaryotes such as protozoa, algae, and fungi, are found almost everywhere on planet Earth leaving only few places truly sterile. When they form an interacting community, they are collectively referred to as a microbiome[1]. Microbiomes can be found in many different environments, including oceans, lakes, and soils, as well as in association with hosts with which they form symbiotic relationships. In humans, we recognize several important microbiomes classified by the different body sites they occupy, among which the most important are the gut, skin, oral, and vaginal. The human gut microbiome, located in the gastrointestinal tract, is arguably the most diverse, complex, and impactful on the function of the host.

Gut microbes are fed mainly by undigested food components, with some species being able to also degrade mucin in the mucus layer[2]. The produced metabolites can in turn be taken up by other microbes to form complex interaction networks or be directly taken up by the host. These microbiome interactions with the host range from providing nutrients[3], regulating the immune system and inflammation[4] to influencing brain chemistry via so-called gut-brain axis[5]. Conversely, several diseases have been linked to gut microbiome through the same interaction links, such as inflammatory bowel disease[6], cardiovascular disease[7] or cancer[8].

Gut microbes are acquired by transmission

What shapes an existing microbiome has been studied extensively. Factors like diet[9], age[10], smoking[11], or medical drugs[12], have been associated with different species compositions. On the other hand, the full picture about how we obtain microbes in the first place is coming to light only recently. The intestinal environment is very different from that of our surroundings like soil, water and object surfaces or even our skin and mouth, mostly due to different oxygen levels, and thus the species of our gut don’t thrive outside of it. This makes the colonization of the gut by microbes from our surrounding environment unlikely. The leading hypothesis is that we acquire new microbes through transmission from other hosts, mainly other humans. The spread of viral-borne diseases or bacterial pathogens through person-to-person transmission is a common knowledge, but this paradigm has only recently been applied to the non-pathogen class of microbes.

The most studied angle has been the maternal seeding at birth, when babies born sterile obtain their first microbes upon the immediate contact with their mother[13–16]. More recently a large study by our group (with my minor contribution, see Chapter 4) has shown the diminishing extent of the mother-derived microbes with a person's age and that a substantial portion can be shared with people in close contact, for example in co-housing[17], shown also by others[18], putting forward the hypothesis of transmission due to social contact. Transmission of microbes

from other sources such as food[19] or farm animals[20] have been observed but seems to be minor in contribution.

The route of transmission is vaginal, fecal and skin to oral at birth and possibly through milk during breastfeeding, while it is believed that in adult life the fecal-oral route predominates as observed in pathogen transmission[21]. The microbes can be transmitted through a direct skin contact, object surfaces[22] or as air-borne particles[23]. Certain bacterial traits like sporulation, aerotolerance and dormancy facilitate the temporary survival in the outside environment and passage through the digestive tract and thus could create advantage for some species[21].

The understanding of microbiome transmission is necessary to understand microbiome acquisition. It is not only a basic question in human biology and microbiology, but it can expand our understanding of human health and disease. For example, some diseases traditionally considered non-communicable such as type-II diabetes, inflammatory bowel disease or cardiovascular disease, can have a microbial component and thus could become communicable via microbiome transmission[24]. Moreover, personal decisions like hygiene and home cleaning or social connectedness like urban vs. rural environment, co-housing, attending nursery, homeschooling etc. impact the microbiome transmission and could indirectly impact our health[25].

Fecal microbiota transplantation

As the microbiome is linked to several diseases, naturally there has been an interest to induce changes in the microbiome in order to improve the host condition. Antibiotics aim to eradicate microbial members while probiotics aim to add specific members. Possibly the most impactful method is the fecal microbiota transplantation (FMT) where the aim is to transplant a whole microbial community from a donor to a recipient using stool as a microbe-rich medium.

FMT has been approved for clinical practice to treat *Clostridioides difficile* infection and exhibits a success rate around 90%[26]. By restoring a functional microbial community, the pathogenic strain is out-competed or inhibited via several possible mechanisms[27,28]. Consequently, there have been attempts to apply FMT to other intestine related disorders like metabolic syndrome, ulcerative colitis or irritable bowel syndrome although with less relative success[29]. Importantly, maybe surprisingly, FMT has been found to improve response to immunotherapy in patients with melanoma, non-small lung cancer or renal carcinoma by several phase I and II trials[30–34]. Here patients responding well to the therapy are used as donors, instead of healthy individuals as would be common in other scenarios. We can hypothesize that the positive response to immunotherapy after FMT is due to acquisition of certain beneficial strains or loss of unbeneficial strains, but the exact mechanism is not yet discovered. Even before the disease-specific mechanisms leading to successful treatment, the microbiological basis of donor community mixing into recipient's existing community resolving into a final state has not been understood.

Metagenomics to study microbiomes

There are several ways to obtain a direct sample from an intestinal microbiome including colonoscopy or biopsy, all of which are invasive procedures limiting their uses to only clinically justified scenarios. Fecal samples, on the other hand, are non-invasive and relatively easy to collect, and provide a sort of aggregate of the microbial community along the full length of the gastrointestinal tract. Despite certain discrepancies reported between the microbiome communities from stool, colon biopsy and colon effluent samples[35,36], fecal samples became widely adopted by the community to provide a proxy of the gut microbiome composition and allow the collection of large numbers of samples.

Isolation and cultivation followed by genome sequencing has historically been the main way to study the microbiome's members and their genomes. This process allows for a detailed study of a given species in laboratory conditions, but it is slow and difficult as it requires to determine the right conditions to grow each species, especially since the intestine is an anaerobic environment. This effort leaves a very large proportion of species diversity undiscovered, sometimes called the microbial dark matter analogously to the invisible portion of matter in physics. Previously, the sequencing of 16S ribosomal RNA (rRNA) genes has been employed to identify the microbial members via sequence similarity comparison or phylogenetic reconstruction, as the 16S rRNA gene is present in all prokaryotes and is highly conserved[37]. The same conservation property also presents a downside, that due to its low variability it is limited in taxonomic resolution. Although from sequencing of the full-length 16S gene it is possible to distinguish between species, studies typically employ sequencing of only 1-2 variable regions, and in any case the fine strain-level differences needed to study transmission are not captured [38]. With the lowering cost of DNA sequencing, the sequencing of whole genomes of all microbiome members (metagenomics) overcame the limitations of isolation and cultivation, as well as the taxonomic resolution of 16S rRNA amplicon sequencing[39].

Metagenomic assembly and binning are a set of algorithms applied to disentangle the genomic content of individual species genomes within a metagenome, which are called metagenome-assembled genomes (MAGs). MAGs enable the study of the genetic content of all bacterial members, including those that make up the uncultured part of the microbiome[40]. The main limitation is that MAGs can be reconstructed only from organisms above certain abundance thresholds, which can be partially mitigated by increasing sequencing depth. Moreover, once a MAG of a species is obtained, detection of such species at a lower coverage in other metagenomic samples can be achieved using short read mapping approaches. Whole genome metagenomic sequencing is particularly suited to study bacteria and archaea and even though it is able to capture genomes of certain microeukaryotes[41] or DNA viruses[42], application to these other domains is limited unless adapting specific techniques[43,44]. In this thesis I focus mostly on the prokaryotic (that is bacterial and archaeal) part of the microbiome.

Prokaryotic species definition and profiling

A fundamental entity of a microbiome is a single cell with its genome. Prokaryotic cells continuously clone themselves replicating their genome, introducing mutations due to errors of the DNA polymerase[45]. These mutations are inherited upon further cloning thus creating descendant lineages. Another mechanism that introduces genetic variability is horizontal gene transfer, where cells from different lineages can exchange genetic material and incorporate it into their genomes[46]. On the scale of cell populations, the prevalence of individual lineages will change by random genetic drift or by selection if the mutations carry fitness advantage or disadvantage[47].

In classical biology, species is defined as a group of organisms that can sexually reproduce within but not across the species boundary, but this definition is not directly applicable to bacteria and archaea[48]. Nonetheless, species-like natural clustering is observed at the genomic level possibly owing to mechanisms like selective sweeps[49] and recombination being more likely between more similar lineages leading to within-species cohesion[47]. Genomic similarity assessed in the laboratory through the DNA-DNA hybridization technique became the gold standard to define bacterial species[50]. With the rise of genome sequencing, large amounts of genomes can be compared and it has been observed that species correspond to clusters of genomes with around 95% similarity[51–53]. Thanks to the widespread adoption of metagenomic sequencing, several large collections of MAGs have been published, which over the years grew from hundreds of thousands[54,55] to millions[56–58]. This has been complemented by the development of efficient tools for genome comparison like Mash[59] and FastANI[52] and even more recently those specific for MAGs avoiding biases due to possible incompleteness and contamination like skani[60]. MAGs clustered by genome sequence similarity can define novel species-like clusters, most of which were never cultured and isolated[54,56–58]. In this work I utilize the system of species-like taxonomic units called species-level genome bins (SGBs) defined as clusters of MAGs and genomes from isolate sequencing with 95% average ANI as introduced in ref. [54].

Several tools have been developed for detection and estimation of relative abundance of species in a metagenomic sample. These tools work directly with sequenced reads without the need for metagenomic assembly and thus can detect taxa with lower coverage. Among the most used tools are Kraken suite using fast k-mer matching[61], mOTUs using universal phylogenetic marker genes[62] and MetaPhlAn using species-specific marker genes[63]. In this work I utilize MetaPhlAn 4, which employs the SGB system and SGB-specific marker gene database[63].

Computational methods to study microbiome transmission

Characterized prokaryotic species have assigned taxonomy and for those uncharacterized we can adopt the SGB system described in the previous section to define new species. However, there's still much diversity even within the species where the individual strains with different single nucleotide variants (SNVs) or larger genomic structural changes can exhibit different

phenotypes with regards to metabolism, pathogenicity or demographic distribution[64–66]. And regardless of phenotype, we can utilize these strain-level differences to detect transmission and spread of microbes within groups of people.

The definition of strain is rather loose and means a group of cells descending from a common ancestor with relative genetic similarity. For the purpose of studying transmission, we define strains in an operative way. Mutations within a lineage arise spontaneously by chance and those favorable in terms of fitness sweep to high frequencies[49]. Still, those *de novo* mutations accumulating within a person's lifetime will be comparatively few to the difference between two strains that have a common ancestor much further back in time[67]. Exploiting this fact, by detecting the same mutations, in particular SNVs, across different samples, we can infer recent transmission events as the same mutations would unlikely co-arise by chance.

The SNVs can be tracked from metagenomic sequencing and can be compared directly using metrics based on matches and mismatches like average nucleotide identity (ANI)[68,69] or can be used to build phylogenetic trees using algorithms based on evolutionary models[70]. Phylogenetic trees add one more step to the computational analysis, but allow us to study the whole picture of evolutionary relationships in addition to the pair-wise genomic similarity[66]. Common challenges of strain tracking methods include the higher coverage requirement to reliably detect SNVs, the noise of sequencing error and the co-existence of genetically different strains of the same species[71] that are difficult to disentangle from short-read sequencing. In this work I utilize StrainPhlAn that uses SNVs in MetaPhlAn's marker genes to reconstruct strain-level phylogenetic trees for each SGB[70]. StrainPhlAn uses majority voting to determine the dominant alleles at positions covered with multiple reads and drops positions where no clear dominant allele exists. This way the dominant strain for each SGB in each sample is represented.

In the StrainPhlAn framework phylogenetic distances are calculated as distances along branches between the tree leaves. By comparing them to a predefined threshold it is determined whether two strains are similar enough to call a transmission event. The threshold has to allow for a certain number of SNPs between the strains to accommodate possible *de novo* mutations occurring since the transmission event and the noise due to sequencing errors. In this thesis I adopt a data driven approach of determining such thresholds developed in ref. [17]. I consider a distribution of distances between strains from unrelated individuals, which represents the variability of unrelated strains and can be used as a null distribution for the null hypothesis of no transmission event. By setting a threshold as a percentile of this null distribution we effectively set the false positive rate (FPR). To improve these thresholds when possible, I consider another distribution of distances between strains from longitudinal samples taken less than 6 months apart, which is typically a sharp peak around 0 and represents the variability due to *de novo* mutations and sequencing error as most strains persist in individuals in this time-frame[67]. This can be used as the alternative distribution for the hypothesis of strain sharing event. The optimal threshold separating those two distributions is then determined with Youden's index. These distributions and thus the strain sharing event thresholds are determined

for each SGB separately, which takes into account the possible differences between species in mutation rates, growth rates and biases due to marker gene selection.

Aims of the thesis

This thesis has two main aims:

1. Understand the role of social contact in development of an early life human gut microbiome through the study of microbiome transmission

The microbiome in the first months and years undergoes development that is influenced by factors like breastfeeding, birth mode and exposure to pets or siblings and stabilizes to become adult-like at circa 3 years of age[72]. The early stages of microbial transmission from mother and other family members have been extensively described[13,16,73,74], the impact of social contact in the later infant stages, however, remains unexplored. Determining the role of social contact in early life gut microbiome development is the Aim 1 of my thesis.

2. Understand the dynamics of microbial engraftment and its links to clinical variables and outcomes in FMT

Despite many single-cohort studies conducted, the usually small sample size and variability of diseases and clinical parameters prevents the discovery of generalizable patterns. Understanding the dynamics of microbial engraftment and its link to clinical variables is the Aim 2 of my thesis.

Structure of the thesis

Following this Introduction are two result Chapters, each targeting one of the two Aims. Chapter 4 will summarise other works to which I contributed to and the thesis will end with a Discussion, where I summarize the key findings, current state of research and discuss future research directions.

Chapter 2: Baby-to-baby strain transmission shapes the developing gut microbiome

Context and contribution

The aim of this work was to apply computational methods to detect strain sharing among individuals, with the goal to investigate the development of the early-life microbiome in infants driven by social contact in the nursery context. I contributed to this work by expanding the SGB database by incorporating MAGs from this and other cohorts in order to accurately detect and track strains of as many species as possible. I implemented and ran the strain sharing pipeline implementing checks and validations and manual curation of the strain phylogenies to ensure correct strain matching, which is a crucial step for all the downstream analysis and findings. I participated in data analysis in collaboration with the other two co-first authors, specifically that shown in panels Figure 1b, e, f, Figure 3d, e and Figure 4a, b. I contributed to the manuscript by writing the method sections pertaining to the SGB database, strain profiling and strain sharing pipeline, and by providing feedback on the draft manuscript.

Reference

Liviana Ricci*, Vitor Heidrich*, Michal Punčochář* et. al. Baby-to-baby strain transmission shapes the developing gut microbiome. Accepted for publication in *Nature*

Inserted manuscript

Baby-to-baby strain transmission shapes the developing gut microbiome

Liviana Ricci[1,*], Vitor Heidrich[1,*], Michal Punčochář[1,*], Federica Armanini[1], Matteo Ciciani[1], Amir Nabinejad[2], Farnaz Fazaeli[2], Elisa Piperni[1,2], Charlotte Servais[1], Federica Pinto[1], Mireia Valles-Colomer[3], Francesco Asnicar[1], Nicola Segata[1,2,4,^]

1 Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy
2 European Institute of Oncology, Scientific Institute for Research, Hospitalization and Healthcare, Milan, Italy
3 Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain
4 Department of Nutritional Sciences, King's College London, London, UK
^ correspondence to nicola.segata@unitn.it

SUMMARY

The early infant microbiome is largely primed by microbial transmission from the mother between birth and the first few weeks of life [1–3], but how interpersonal transmission further shapes the developing microbiome in the first year remains unexplored. Here we report a metagenomic survey to model microbiome transmission in the nursery setting among babies

attending the first year, their educators, and their families (*n* = 134 individuals). We performed dense longitudinal microbiome sampling (*n* = 1,013 fecal samples) during the first year of nursery and tracked microbial strain transmission within and between nursery groups across three different facilities. We detected extensive baby-to-baby microbiome transmission within nursery groups even just after only one month of nursery attendance, with nursery-acquired strains accounting for a proportion of the infant gut microbiome comparable with that from family by the end of the first term. Baby-to-baby transmission continued to grow over the nursery year, in an increasingly intricate transmission network with single strains spreading in some classes, and with multiple baby-acquisition and species-transmissibility patterns. While having siblings was associated with higher microbiome diversity and reduced strain acquisition from nursery peers, antibiotic treatment is the condition most accounting for increased influx of strains. This study shows that microbiome transmission between babies is extensive during the first year of nursery, and points to social interactions in infancy as crucial drivers of infant microbiome development.

INTRODUCTION

The early infant microbiome assembles via intricate and partially stochastic microbial acquisitions which have the mother as the primary source and other family members as additional ones [1–5]. The infant microbiome then evolves during the following few years with complex dynamics that later result in a more stable adult-like microbiome [6]. While early family-to-infant microbiome strain transmission has been quite extensively investigated [1–4,7], later infant developmental stages including those involving interaction with other peers in social contexts have received very little attention.

Because the person-to-person intra-generational microbiome transmission has been recently revealed to be extensive and impacting the personal microbiome make-up [8], we hypothesized that early social contexts such as nurseries might exert a large impact on infant microbiomes via baby-to-baby transmission. Beyond work on pathogen spreading [9,10] and linked immune competence development [11–13], microbiome investigations in nurseries were limited in observing increased microbial diversity among attendants [14]. This leaves a major gap in the understanding of the dynamics of human microbiome maturation during the key first 1,000 days of life [15].

Here we present microTOUCH-baby, a strain-resolved longitudinally-dense metagenomic study modeling interpersonal gut microbiome transmission between babies attending the nursery for the first time and their close contacts, including family members and nursery educators.

RESULTS

The microTOUCH-baby study

We set up the microTOUCH-baby cohort to study the dynamics of microbiome development and transmission among babies of about one year of age and their close social interactions network (**Methods**). Participants included 43 babies attending the first year of nursery (median age at nursery admission = 10 months), 7 co-living siblings, 39 mothers, and 30 fathers of the babies, and 5 pets from the participants' houses, as well as 10 nursery educators (134 volunteers in total, **Fig. 1A**, **Supplementary Table 1**). Infant participants were enrolled from three public nurseries in Trento (Italy). Babies spent on average 8 hours per weekday (after the "settling-in period", **Methods**) in one of two classes with limited shared activities and spaces, followed by different educational staff.

Sampling started before the beginning of the first term (T01), hence before participants from different families had any nursery-related contact among them, and ended after the Christmas nursery closure (**Fig. 1A**). During nursery attendance, we collected stool samples of the babies on a weekly basis, while educators and parents were less densely sampled (**Methods**). For all participants in group 1 of nursery A, sample collection continued through the second term. Two additional follow-up samples were collected for all participants at nursery year's conclusion (TA) and at the end of the summer break (TB) (**Fig. 1A**).

Overall, we collected and metagenomically sequenced 1,013 microbiome samples (avg. sequencing depth = 15.61 Gbp, **Methods**). Host metadata information included exact age, past and current host-health data, antibiotic exposures, maternal delivery information (**Methods**, **Fig.**

1A, **Supplementary Table 2–3**), as well as diet questionnaires (**Methods**). Metagenomes were processed via MetaPhlAn 4 [16] to generate taxonomic profiles at species-level genome bin (SGB) [17] resolution (**Supplementary Table 4, Ext. Data Fig. 1A**), including yet-to-characterize species (i.e. unknown SGBs - uSGBs - accounting for 46.37% of total SGBs). We then used StrainPhlAn 4 [16,18] to generate strain-level phylogenies for 311 known SGBs (kSGBs) and 201 uSGBs that were used to infer microbiome strain transmission (**Methods**) [8].

Compositional baby microbiome landscape

We first observed expected microbiome structures [1,19,20], with large compositional divergence between adults and babies (**Fig. 1B**, **Ext. Data Fig. 1B, 2, 3**, **Supplementary Table 5**), age-dependent differences in babies (**Ext. Data Fig. 4A-E**), and diet-dependent microbial stratification in adults (**Ext. Data Fig. 4F**), but not in babies after accounting for age (**Supplementary Table 6**). Interestingly, at T01 (median age = 10 months) the impact of maternal intrapartum antibiotic prophylaxis against *Streptococcus* B and of mode of delivery on alpha diversity was already not detected as statistically significant (Mann-Whitney U test, $n = 37$, $U = 137$, $P = 0.68$ and $n = 37$, $U = 109$, $P = 0.89$ respectively, **Ext. Data Fig. 5A-D**).

Some compositional patterns were suggestive for a role of microbiome transmission. Babies having a sibling had, for example, an overall higher SGB richness compared to babies without brothers and sisters ($n = 40$, $U = 271$, $P = 0.012$; **Fig. 1C**, **Supplementary Table 6**), further supporting previous observations [21,22] and suggesting that siblings may provide important sources for infant microbiome enrichment. In contrast, babies with pets exhibited lower overall SGB richness ($n = 40$, $U = 61$, $P = 0.012$; **Fig. 1D**), but significance was lost after adjusting for age (**Supplementary Table 6**). Babies' alpha diversity increased during the 3 months of nursery attendance (**Fig. 1E**) and while the total pool of microbial species detected among babies in the nursery did not change noticeably throughout the study (**Ext. Data Fig. 5E**), the inter baby beta-diversity decreased significantly (7% avg. decrease; $n = 116$ baby pairs, $T = 1{,}026$, $P = 7.0e\text{-}11$; **Fig. 1F**). As overall this might be indicative of baby microbiome convergence influenced by interindividual transmission, we performed strain-level transmission analysis to investigate this hypothesis.

Mapping strain sharing in the nursery

Extending our StrainPhlAn-based validated pipeline [8] (**Methods**), we defined a strain-sharing event as the identification of the same strain (i.e. differing by a genetic distance lower than the pre-computed optimal species-specific threshold distinguishing between inter- and intra-individual genetic distance distributions) in different microbiome samples. Strain-sharing rates (SSR) are computed as the number of strains shared between a pair of microbiome samples over the number of species with profiled strains present in both samples (**Methods**). Applied on the task of inferring mother-baby transmission, the pipeline estimated a 50% median SSR for babies at the beginning of the study which is highly consistent with previous results irrespective of population (**Ext. Data Fig. 5F**).

Overall, we captured over 9.47M instances of the same SGB typed at strain-level in different samples (including those from the same participant and from different participants), with a total

of 5.97% of cases in which the same strain of the SGB was present, resulting in 565,258 detected strain-sharing events (**Supplementary Table 7–8**). Within-subject strain-sharing accounted for 27.9% of the total (157,599 events, with 99% likelihood of samples from the same individual sharing at least one strain, and 87% at least 5) but also strain-sharing between different subjects in the same family was very high (51,483, 9.1% of the total, with 86% likelihood of sharing at least one strain, 47% at least 5), with rarer between-family strain sharing instances at T01 (46% likelihood of sharing at least one strain and 3% at least 5; **Ext. Data Fig. 5G**). While most strain-sharing over the first term was observed among individuals from different families (356,176, 63%), this reflected the >75x greater number of between-family comparison pairs; after normalizing for the number of comparisons, one order of magnitude fewer strains were shared between families versus within family (0.7 vs 7.9 strains shared per sample pair; **Supplementary Table 7**). The 0.7 average strains shared by unrelated individuals represent the cohort's microbiome sharing background, including untraced social interaction before T01, clonal strains spreading into the nursery-associated local community, and possible false positive instances among other factors.

Tracking multi-host strain transmission

As a representative example of the combined capabilities of our study design and metagenomic pipeline to trace complex strain transmission chains, we illustrate the interpersonal transfer of a nursery-acquired strain of *Akkermansia muciniphila* (SGB9226) in group 1 of nursery B. A strain from this species was first introduced in the nursery group by a baby (B05) who likely obtained it from their mother, passed to another baby (B06) to then be found in their mother (M06) and father (F06), in the latter replacing another *A. muciniphila* strain (**Fig. 2A**). *A. muciniphila* strains contain CRISPR arrays that can be used as unique genetic tags for strains [23,24] that further confirmed *A. muciniphila* strain identity across volunteers (**Methods**). Metagenomic assembly also validated such transmission patterns for the limited number of strains (8 out of 19 StrainPhlAn-positive samples) that could be reconstructed into draft genomes of sufficient quality, with high genomic similarity between assemblies from samples with the same strain according to StrainPhlAn (pairwise avg. ANI: 99.97% which aligned with same-strain boundaries independently estimated elsewhere [25,26]). We note that the missing detection of *A. muciniphila* strains (grey circles, **Fig. 2A**) was overall consistent with the absence of the species as shown in a high-sensitivity, SGB-specific PCR (**Methods, Ext. Data Fig. 6A**). Within this example, we found only one sample in which we missed the metagenomic strain profiling to be PCR-positive at SGB-level, concordantly with a non-zero relative abundance (0.04%) in its MetaPhlAn profile (B05_T08, **Ext. Data Fig. 6B**), being thus the single case in **Fig. 2A** of SGB9226 falling below the limit of detection for strain profiling. Another transmission chain example involved *Alistipes finegoldii* (SGB2301) and included an educator (**Ext. Data Fig. 6C**), further contributing to show the potential of our approach to recapitulate microbial transmission in nurseries.

We also explored potential gut microbiome transmission between household pets and their families. Anecdotally (given the only 5 pets considered), we overall identified a low total number of pet-human strain-sharing events, with intra-family pet-baby strain-sharing significantly higher than inter-family (Fisher's exact test, $n = 211$, $P = 0.005$, **Ext. Data Fig. 6D-E**). Strains found to be transmitted between babies and pets belonged to human-associated species that had also

been previously detected in pet gut microbiomes (*Faecalimonas umbilicata*, *Ruminococcus gnavus*, *Clostridium* sp AT4 and *Phocaeicola vulgatus* [27–30]), indicating they may be ecologically fit to overcome host species boundaries.

Strain spreading patterns in the nursery

We then examined the changes in the collective composition of the human microbiome in nurseries. First, we found the overall pool of distinct strains to decrease over time (i.e. avg. nursery strain heterogeneity decreasing from 0.91 at T01 to 0.77 at T15, Mann-Whitney U test, n = 454, U = 34,312, P = 1.3e-11). Considering that the total reservoir of microbial species did not increase (**Ext. Data Fig. 5E**), this indicates that some strains within the same species may have spread among babies and prevailed over other strains initially present (**Ext. Data Fig. 7A**).

We then focused on strains that showed efficient spreading within a nursery. We found 8 cases of strains initially detected in no more than one baby before nursery start (T01) reaching ≥ 50% prevalence afterwards (**Fig. 2B**). Among these, a *Streptococcus gallolyticus* (nursery A) and a *Bifidobacterium pseudocatenulatum* (nursery B) strain were introduced in the nursery after approximately the first month of attendance and progressively spread to seven and eight babies, respectively (**Fig. 2B**). While *S. gallolyticus* spread appeared to dwindle after reaching the maximum diffusion, *B. pseudocatenulatum* presence was steadily detected, consistently with the high prevalence of the *Bifidobacterium* genus in the infant population [2]. Other cases of bacterial strain diffusion involved *Escherichia coli* and *Veillonella dispar* in nursery B, and *Clostridium innocuum* in nursery C which was possibly limited in its spread by other conspecific strains and niche preemption dynamics [31].

Baby microbiomes built via transmission

Quantification of strains shared between babies attending the same nursery over time revealed they had, on average, more shared strains at the end of the first term than before nursery admission (avg. number of strains shared with any other baby at T01 = 2.5 and at T15 = 7.2 or 8.8 when disregarding strains already present at T01 and only for babies with samples available at both time points; **Fig. 3A**). Accordingly, while at T01 baby strain-sharing relations were not recapitulating nursery attendance, at T15 they clustered consistently with it (**Fig. 3B**, **Supplementary Table 9**, **Methods**). We thus found strong evidence of quantitatively relevant acquisition of nursery-specific microbial profiles by babies, occurring via interindividual strain transmission even in the relatively short time frame of the first nursery term.

Investigating longitudinal gut microbiome changes, babies showed the lowest rate of SGB retention (defined as the Jaccard similarity between samples from initial and final timepoints of the same individual) (**Ext. Data Fig. 7B**) and the highest rate of strain replacement (defined as 1-SSR) among the retained SGBs (**Fig. 3C**) compared to adults. A median 44.4% of the retained SGBs in babies showed baseline strain replacement during the 5 months of the study. In contrast, all other participants replaced a much lower fraction of strains in their gut (medians below 11.1%), with strain replacement rates correlated although non-significantly with age among non-baby participants (Spearman's test, n = 68, ρ = 0.22, P = 0.071; **Ext. Data Fig. 7C**).

This reflects the expected high plasticity of the infant gut microbiome with its rapidly evolving ecosystem and limited colonization resistance [6,32].

To assess the extent to which nursery attendance affects infant microbiome assembly via microbiome transmission, we quantified and compared the strain-sharing rate between pairs of babies within the same group or nursery, and across different nurseries at each timepoint (**Fig. 3D**). Strain-sharing among babies in the same nursery group was significantly higher after approximately only one month of nursery attendance in comparison with babies from different nurseries (median SSR 8.3% vs 0% at T04; permutation test for medians, n = 249, P = 0.001). This is all the more noteworthy in view of the first two weeks of the nursery's "settling-in period" during which babies attend discontinuously and for shorter periods. In addition, at the end of the first term (T15), SSR in the same nursery group reached an avg. of 20.2%, significantly higher than the SSR between babies attending different nurseries (4.6%; permutation test for medians, n = 312, P < 0.001) and higher than the SSR among babies attending the same nursery but in different groups (16.1%; permutation test for medians, n = 122, P = 0.079, significant at T08 P = 0.026, T10 P < 0.001, and T13 P = 0.001).

By extending the investigation to the second term of nursery, we found baby-baby SSR within the same nursery (regardless of group) to reach a median 33.3% at the end of school year (TA) (vs median 17.9% at T15; Wilcoxon signed-rank test, n = 58, T = 86, P = 6.2e-9, **Fig. 3E**), with a progressive increase occurring during the whole second term, as observed for the class that was densely sampled over such period (group 1 of nursery A, **Ext. Data Fig. 7D**). Although baby-baby SSR decreased during the summer break (TB), it remained significantly higher compared to post-Christmas break levels (T15) (median 23.7% at TB vs 17.9% at T15; Wilcoxon signed-rank test, n = 31, T = 68, P = 2.0e-4). These results highlight that social relations outside of the household and continued spatial proximity are key determinants of infant microbiome transmission and development at levels that are substantially higher than what was recently observed for adults [8].

Nursery strains match family contribution

Parent-baby strain-sharing rate at T01 averaged 37.3% for mothers and 19.6% for fathers, consistently with available reports [1,4,8,33–35]. Such patterns persisted throughout the first term (**Fig. 4A**). Contributions of sibling strains to the baby was even higher (avg. SSR = 56.2%; **Fig. 4B**). As expected, strain transmission between babies and individuals from different families remained negligible throughout the first term (**Fig. 4A-B**), a testament for the reliability of the strain transmission inference approach.

To establish the relative contribution of strain-transmission from the nursery with respect to strain transmission from the family, we computed, for each baby, the proportion of strains in the infant microbiome that were exclusively shared with, and hence putatively acquired from, either family members or other babies in the nursery group (**Methods**) and we refer to it as "proportion of strains acquired". We found that the proportion of strains acquired from the nursery group – but not of strains acquired from the family – significantly changed over time. The proportion of strains acquired from family members fluctuated from an average of 24.0% per baby at T01 to 20.0% at the end of the first term of nursery (Wilcoxon signed-rank test, n = 25, T = 112, P =

0.18; **Ext. Data Fig. 7E**), while those putatively acquired from the nursery group increased from an average of 6.5% to 28.4% at the end of the first term (Wilcoxon signed-rank test, n = 25, T = 0, P = 6.0e-8; **Ext. Data Fig. 7E**), significantly surpassing the proportion of strains acquired from the family (Mann-Whitney U test, n = 52, U = 463, P = 0.023; **Fig. 4C**). This indicates that after only 3 months of nursery attendance, babies had proportionally more strains acquired from nursery peers than from their family.

A similar trend was observed when quantifying the relative abundance of strains acquired either from the family or the nursery group (**Fig. 4C**). Family contribution slightly diminished over time (from an avg. 33.2% at T01 to 20.6% at T15; Wilcoxon signed-rank test, n = 25, T = 72, P = 0.014; **Ext. Data Fig. 7F**) while the contribution from the nursery group greatly expanded (reaching an avg. of 39.6% at T15 from a starting 10.2%; Wilcoxon signed-rank test, n = 25, T = 18, P = 1.5e-5; **Ext. Data Fig. 7F**). Strains shared with both family and group also increased significantly (from avg. 0.9 to 8.5%; Wilcoxon signed-rank test, n = 25, T = 0, P = 4.4e-4; **Ext. Data Fig. 7F**), likely reflecting reciprocal transmission between family and nursery (**Fig. 2A**). Overall, this suggests that the nursery collectively contributes to a larger extent to infant strain composition than the family by the end of the first term (39.6 vs 20.6% at T15; Mann-Whitney U test, n = 52, U = 479, P = 0.01; **Ext. Data Fig. 7F**).

Long-term nursery effect on transmission

The extended longitudinal analysis of group 1 of nursery A revealed that the proportion of strains acquired from nursery peers continued to gradually increase during the second term (**Ext. Data Fig. 8A**). Samples from all babies across nurseries at year-end (TA) confirmed comparable contributions of family and nursery to the baby (17.6% median proportion of strains acquired from nursery vs 15% from family; Mann-Whitney U test, n = 19, U = 218, P = 0.29; **Ext. Data Fig. 8B**), that non-significantly tended toward a greater family contribution after summer nursery closure, (8.7% median proportion of strains acquired from nursery vs 16.7% from family; Mann-Whitney U test, n = 17, U = 122, P = 0.43; **Ext. Data Fig. 8B**).

Babies exhibited lower strain retention and higher strain replacement across the summer break (i.e. between TA and TB) compared to adults, despite no differences in the carriage of SGBs typed at the strain level (**Ext. Data Fig. 8C-F**). Interestingly, family-acquired strains were significantly more retained and less replaced in babies over the summer break than nursery-acquired strains (Wilcoxon signed-rank test, n = 11, T = 5, P = 0.019 and P = 0.022 respectively; **Ext. Data Fig. 8G-H**), suggesting that continuous seeding linked to continued contact is a factor behind long-term colonization.

Siblings affect baby strain acquisition

Hypothesizing a potential role of siblings in the transmission patterns, we found that at T01 babies showed higher SSR with their siblings (avg. 52.3%) than with their fathers (24.9%; Mann-Whitney U test, n = 36, U = 147, P = 0.026) as well as with their mothers, although non-significantly (46.1%; Mann-Whitney U test, n = 36, U = 120, P = 0.47; **Ext. Data Fig. 8I**). Of note, an average of 10.4 strains were shared exclusively with siblings at T01, while only 2.0 and 2.4 were shared exclusively with the mother or the father (**Ext. Data Fig. 8J**), possibly reflecting

closer intestinal ecology, physical interaction, and development stage, which are likely some of the same factors leading to the higher nursery-strain acquisition observed in our cohort.

We further observed that having a sibling was associated with babies acquiring significantly fewer strains from their nursery group compared to babies without a sibling at T15 (Mann-Whitney U test, n = 28, U = 117, P = 0.004; **Fig. 4D**). While causality cannot be inferred, this might be linked to early acquisition from siblings "saturating" the overall strain acquisition potential which would be in line with babies with a sibling having higher alpha diversity (**Fig. 1C**) and acquiring less new SGBs than only-children (**Fig. 4E**). Notably though, while all babies both spread and acquired strains in the nursery, the ratio between acquired and donated strains varied widely between babies (**Fig. 4F**).

The most transmissible species

We next assessed species-level transmissibility by counting the number of strain-sharing events for each SGB in our cohort over the total potential number of strain-sharing events (**Methods**). Microeukaryotic taxa were not found abundant enough in infants to try to infer transmission, with *Blastocystis,* the most common human gut microeukaryote [36], identified in 9.18% of the samples but never in babies (**Supplementary Table 12**). Focusing thus on prokaryotic taxa, out of the 64 SGBs with highest transmissibility (henceforward "T") over all participant categories (**Ext. Data Fig. 9A, Supplementary Table 13**), many kSGBs encompassed aerotolerant (*S. gallolyticus, Rothia mucilaginosa, B. pseudocatenulatum*) and spore-forming species (e.g. *Tyzzerella nexilis* and *Clostridium fessum*). We also identified the spore-forming *Clostridioides difficile* among the most transmissible SGBs between baby-baby pairs only (T = 0.38, prevalence in babies = 24% and in adults = 0%), in line with widespread carriage in asymptomatic infants [37,38]. Exceptions to this trend were prevalent non-sporulating human gut anaerobes (such as *Blautia wexlerae* and *Faecalibacterium prausnitzii*).

SGB transmissibility correlated with SGB prevalence in both adults (Spearman's test, n = 461, ρ = 0.35, P_{adj} = 9.8e-14) and babies (Spearman's test, n = 461, ρ = 0.40, P_{adj} = 1.2e-17; **Ext. Data Fig. 9B-C**). Highest transmissibility scores were highlighted for SGBs shared in baby-siblings pairs, namely *A. finegoldii*, *Bacteroides ovatus* and *Bacteroides caccae*, the butyrate-producing *Roseburia intestinalis* and *Agathobaculum butyriciproducens* [39,40], *Bifidobacterium bifidum*, and *Bifidobacterium breve* (all with T = 1; **Ext. Data Fig. 9A**). *B. caccae* strains were also commonly transmitted between mothers and babies, alongside strains of two undescribed *Clostridium* spp, *Phocaeicola vulgatus*, and the typically maternally-derived *B. bifidum* and *B. pseudocatenulatum* [32,41]. Highly transmitted SGBs between fathers and babies included *Clostridium* sp AM333, *Lachnospira* spp, and the aerotolerant and bile-resistant *Sutterella wadsworthensis.* Finally, with the exception of the microaerophilic *Streptococcus salivarius* and *S. wadsworthensis* (T = 0.83 and 0.82, respectively), highly transmitted SGBs between mother-father pairs included multiple bifidobacteria and *Blautia* spp. Interestingly, many of the species are fiber-degrading specialists in the gut [42–44], with known beneficial effects on the host [45], indicating that within-family microbial transmission may hold a favorable potential for health-associated microbiome development.

We looked further into our dataset to identify species differentially more transmitted

baby-to-baby in the nursery setting compared to baby-mother and baby-father pairs (**Supplementary Table 14**). *B. breve*, a highly prevalent and health-promoting species in (breast-fed) babies [6,46], was differentially more transmissible among baby pairs, compared to mother-baby pairs, as it was the case also for *Dorea formicigenerans,* an age progression biomarker in infants [47] (**Ext. Data Fig. 10A-B**). *Bifidobacterium longum* subsp. *infantis,* a specialized gut colonizer of breast-fed infants [46] with anti-inflammatory effects [48], was detected exclusively in babies in our cohort (**Methods**, **Ext. Data Fig. 10C-D**), with prevalence peaking at ~50% mid-term (T08), before declining (**Ext. Data Fig. 10E**); its transmission was significantly higher than *B. longum* subsp. *longum* among baby-baby pairs (T = 85.3% vs 19.4%, respectively; Fisher's exact test, n = 142, P = 5e-12), showing that the acquisition of subsp. *infantis* may specifically occur via interpersonal transmission among babies.

Host factors and microbiome transmission

In addition to the effect of having a sibling, age also significantly affected strain donation (increasing frequency in older babies, Spearman's test, n = 39, ρ = 0.43, P = 0.007), but not strain acquisition (n = 39, ρ = 0.24, P = 0.14; **Ext. Data Fig. 11A-C**). Interestingly, potentially delayed microbial colonization at birth (due to Cesarean delivery or intrapartum antimicrobial prophylaxis) did not influence infant microbial strain acquisition in the nursery (**Ext. Data Fig. 11D-E**), in line with no T01 alpha diversity differences (**Ext. Data Fig. 5A-B**). Analysis of the influence of infant diet on strain-sharing revealed that infants consuming milk at T01, particularly maternal milk, exhibited elevated albeit not statistically significant strain-sharing rates with their mothers at T01 (**Ext. Data Fig. 12A-B**). Further exploration of dietary impacts on strain acquisition and donation patterns failed to identify significant associations (**Ext. Data Fig. 12C-K**), suggesting an overall negligible impact of diet on interpersonal microbiome transmission; however, putative dietary effects on the establishment of specific strains in a recipient microbiome cannot be definitely excluded, given the limited granularity of our dietary data.

Antibiotics effect on strain acquisition

Finally, we assessed the impact of antibiotic interventions on adult and infant interpersonal transmission exploiting the recorded antibiotic administration events that included amoxicillin alone (n = 7 events) and in combination with clavulanic acid (n = 13), betamethasone dipropionate (n = 6) and the macrolide azithromycin (n = 4), routine treatments for bronchitis, inflammatory skin conditions, upper respiratory, ear and intestinal infections [49,50].

Antibiotic treatment significantly reduced the absolute number of retained strains between consecutive timepoints in both adults (avg. 86.4 Ctrl pre-post vs 60.1 ATB pre-post; Mann-Whitney U test, n = 74, U = 729, P = 5.1e-4) and babies (avg. 24.3 Ctrl pre-post vs 14.1 ATB pre-post; Mann-Whitney U test, n = 77, U = 1,169, P = 1.3e-5, **Ext. Data Fig. 12L**). Even for SGBs typed at the strain level that were present in both timepoints, the strain retention rate was also significantly diminished after treatments in adults (avg. 93.8% Ctrl pre-post vs 88.4% ATB pre-post; Mann-Whitney U test, n = 74, U = 631, P = 0.028; **Fig. 5A**) and in babies (avg. 90.6% Ctrl pre-post vs 70.2% ATB pre-post; Mann-Whitney U test, n = 76, U = 1,215, P =

2.9e-7, **Fig. 5A**), but to a greater extent in the latter (ATB pre-post avg. strain retention rate adults vs babies: 88.4 vs 70.2%; Mann-Whitney U test, n = 53, U = 466, P = 0.001; **Fig. 5A**).

After antibiotic use, the gut microbiomes of babies were replenished with new strains (**Fig. 5B-C**). This was driven by both the acquisition of new SGBs (avg. SGB acquisition rate 30.4% Ctrl pre-post vs 49.2% ATB pre-post; Mann-Whitney U test, n = 59, U = 164, P = 2.9e-4; **Fig. 5B, Ext. Data Fig. 12M**), and by the strain replacement within SGBs (avg. 2.1 replaced strains and 7.1% replacement rate Ctrl pre-post vs 3.9 and 13.6% ATB pre-post; Mann-Whitney U test, n = 59, U = 209, P = 0.003 and P = 0.004; **Fig. 5C, Ext. Data Fig. 12N**). In contrast, adult microbiomes appeared to be less prone to new colonizations after antibiotic treatment via either means of strain acquisition (ATB pre-post avg. SGB acquisition rate adults vs babies: 34.2 vs 49.2%; Mann-Whitney U test, n = 33, U = 74, P = 0.041, **Fig. 5B**; ATB pre-post avg. strain replacement rate adults vs babies: 7.5 vs 13.6%; Mann-Whitney U test, n = 33, U = 69, P = 0.026; **Fig. 5C**), suggesting that while infant microbiomes tend to be more impacted by antibiotic therapy, their richness is also more easily recovered.

Conclusions

Our longitudinal strain-resolved metagenomic framework revealed that the infant gut microbiome largely assembles, expands, and modifies in the nursery via extensive baby-baby strain transmission extending earlier work on family-to-baby transmission [1–4,34,35] and overviews of the infant microbiome in nurseries [14,51,52]. After a few months of nursery attendance, the microbial strains acquired from peers in the same nursery group accounted for a larger proportion of the infant microbiome than those from the mother and – more generally – family members (**Fig. 4C**), who are known to exert the greatest influence on babies' microbiome in the first months of life. Contributions to the infant microbiome of family and nursery were not influenced by birth practices or feeding regimes (**Ext. Data Fig. 11D-E, 12A-K**), and became comparable by the end of the second term, possibly indicative of strains being shared with both family and nursery. In addition to their already established effects in emotional and cognitive development [53,54], social relations among peers in the nursery are thus a hub for microbial enrichment during infancy, particularly of key early-life gut colonizers such as *Bifidobacterium longum* subsp. *infantis* and *Bifidobacterium breve* [6,46] (**Ext. Data Fig. 10A-E**).

Horizontal infant microbiome transmission as opposed to vertical transmission, does not occur only in nursery settings as we found that baby-sibling strain-sharing surpasses transmission between parent-baby pairs (**Ext. Data Fig. 5G**) and correlates with a later decrease in infant microbial acquisition in the nursery (**Fig. 4D**). Even pets might contribute with strains to the infants but not to the adults (**Ext. Data Fig. 6D-E**), and while limited and somewhat conflicting evidence on the effect of having a pet on human microbiomes has been produced [55–57], larger studies specifically focused on strain transmission and medium-term retention should be promoted. Overall, our data further reinforce the role of horizontal intragenerational (and possibly inter host species) over vertical intergenerational transmission not only in adults [8] and nurseries (the main point of the present work), but also within a family context.

In several cases, we observed very effective spread of a single strain within nurseries (**Fig. 2B**). Such diffusion patterns are akin to typical pathogenic outbreaks within closed communities [58,59].

However, while pathogenic spread typically elicits an acute immunological reaction and/or requires treatment, leading to somewhat rapid clearance after transmission, for gut microbiome members colonization may be long-lasting as we reported in several cases in our cohort (**Fig. 2B**), even though it remains unsettled whether colonization persisted for many years after the end of nursery school. Moreover, further elucidation of the phenotypes linked to the propagation of fast-spreading strains may be highly relevant toward a better comprehension of the factors favoring the development of a healthy host-microbiome mutualism.

Among the factors that may influence microbial transmission in babies, we found antibiotic usage as the strongest one. Despite the infant microbiome being highly perturbed by antibiotic treatment during the first year of life, as previously reported [60–62], it is also fast-recovering via extensive strain acquisition (**Fig. 5**, **Ext. Data Fig. 12M-N**), consistently with antibiotic treatment before fecal microbiota transplantation increasing donor strain engraftment in adults [63]. However, we found strong evidence that the extent and the rate of post-antibiotic strain acquisition was substantially higher in babies compared to adults (**Fig. 5**), and this clearly reinforces the risks – but potentially also the opportunities – of infant antibiotic intake connected with a deep reprogramming of the structure of the infant microbiome induced by post-antibiotic strain acquisition. Whether the rapid acquisition of microbial diversity after antibiotic courses in babies is driven specifically by the high level of peer-to-peer interaction in the nursery environment should be investigated further, but it is reasonable to hypothesize that prolonged isolation within the family of antibiotic-treated babies would result in a slower microbiome recovery and acquisition of fewer infant-specific microbial species.

Methodologically, our strain-sharing pipeline models the genetics of the dominant strains of each species (SGBs) present in any given microbiome sample [64] to enable identification of strain transmission events. Although recent surveys have pointed out the usual presence in the gut of a single strain of each species [25], further advances in metagenomic strain-profiling tools could reveal the complexity of multiple coexisting conspecific strains and shed light on their role in influencing strain(s) transmission dynamics and long-term colonization in the gut microbiome.

Overall, our results reveal the centrality of social factors in shaping the infant microbiome via interindividual microbial transmission, thus rebalancing social interactions as key to building a healthy microbiome, beyond their epidemiological role in the spread of (opportunistic) pathogens. Continued efforts on this topic should be focused on investigating the transmission of further microbiome components such as phages, plasmids, and operons, as well as on applying experimental tools to profile the microbial features favoring diverse modes of transmission.

MAIN TEXT REFERENCES

1. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133–145.e5 (2018).
2. Yang, B. *et al.* Development of gut microbiota and bifidobacterial communities of neonates in the first 6 weeks and their inheritance from mother. *Gut Microbes* **13**, 1–13 (2021).
3. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**, 146–154.e4 (2018).
4. Dubois, L. *et al.* Paternal and induced gut microbiota seeding complement mother-to-infant transmission. *Cell Host Microbe* **32**, 1011–1024.e4 (2024).
5. Heidrich, V., Valles-Colomer, M. & Segata, N. Human microbiome acquisition and transmission. *Nat. Rev. Microbiol.* (2025) doi:10.1038/s41579-025-01166-x.
6. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
7. Brito, I. L. *et al.* Transmission of human-associated microbiota along family and social networks. *Nat Microbiol* **4**, 964–971 (2019).
8. Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).
9. Song, Z., Chen, L., Sun, S., Yang, G. & Yu, G. Unveiling the airborne microbial menace: Novel insights into pathogenic bacteria and fungi in bioaerosols from nursery schools to universities. *Sci. Total Environ.* **929**, 172694 (2024).
10. Andrup, L. *et al.* Reduction of acute respiratory infections in day-care by non-pharmaceutical interventions: a narrative review. *Front Public Health* **12**, 1332078 (2024).
11. Ball, T. M. *et al.* Siblings, day-care attendance, and the risk of asthma and wheezing during childhood. *N. Engl. J. Med.* **343**, 538–543 (2000).
12. Côté, S. M. *et al.* Short- and long-term risk of infections as a function of group child care attendance: an 8-year population-based study. *Arch. Pediatr. Adolesc. Med.* **164**, 1132–1137 (2010).
13. Kamper-Jørgensen, M., Wohlfahrt, J., Simonsen, J., Grønbaek, M. & Benn, C. S. Population-based study of the impact of childcare attendance on hospitalizations for acute respiratory infections. *Pediatrics* **118**, 1439–1446 (2006).
14. Amir, A. *et al.* Gut microbiome development in early childhood is affected by day care attendance. *NPJ Biofilms Microbiomes* **8**, 2 (2022).
15. Enav, H., Bäckhed, F. & Ley, R. E. The developing infant gut microbiome: A strain-level view. *Cell Host Microbe* **30**, 627–638 (2022).
16. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
17. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
18. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
19. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
20. Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: A cross-sectional study. *BMC Microbiol.* **16**, 90 (2016).
21. Christensen, E. D. *et al.* The developing airway and gut microbiota in early life is influenced by age of older siblings. *Microbiome* **10**, 106 (2022).
22. Laursen, M. F. *et al.* Having older siblings is associated with gut microbiota development during early childhood. *BMC Microbiol.* **15**, 154 (2015).
23. Karcher, N. *et al.* Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol.* **22**, 209 (2021).
24. Mavromatis, K. *et al.* Complete genome sequence of the bile-resistant pigment-producing anaerobe Alistipes finegoldii type strain (AHN2437T). *Stand. Genomic Sci.* **8**, 26–36 (2013).
25. Chen-Liaw, A. *et al.* Gut microbiota strain richness is species specific and affects engraftment. *Nature* **637**, 422–429 (2025).
26. Rodriguez-R, L. M. *et al.* An ANI gap within bacterial species that advances the definitions of intra-species units. *MBio* **15**, e0269623 (2024).
27. Abdugheni, R. *et al.* Comparative genomics reveals extensive intra-species genetic divergence of the prevalent gut commensal Ruminococcus gnavus. *Microb Genom* **9**, (2023).
28. Ganz, H. H. *et al.* The Kitty Microbiome Project: Defining the Healthy Fecal 'Core Microbiome' in Pet Domestic Cats. *Veterinary Sciences* **9**, 635 (2022).
29. Chen, L. *et al.* Gut microbiome of captive wolves is more similar to domestic dogs than wild wolves indicated by metagenomics study. *Front. Microbiol.* **13**, 1027188 (2022).
30. Huang, Z., Pan, Z., Yang, R., Bi, Y. & Xiong, X. The canine gastrointestinal microbiota: early studies and

research frontiers. *Gut Microbes* **11**, 635–654 (2020).
31. Woelfel, S., Silva, M. S. & Stecher, B. Intestinal colonization resistance in the context of environmental, host, and microbial determinants. *Cell Host Microbe* **32**, 820–836 (2024).
32. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
33. Hildebrand, F. *et al.* Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host Microbe* **29**, 1167–1176.e9 (2021).
34. Bogaert, D. *et al.* Mother-to-infant microbiota transmission and infant microbiota development across multiple body sites. *Cell Host Microbe* **31**, 447–460.e6 (2023).
35. Lou, Y. C. *et al.* Infant gut strain persistence is associated with maternal origin, phylogeny, and traits including surface adhesion and iron acquisition. *Cell Rep Med* **2**, 100393 (2021).
36. Piperni, E. *et al.* Intestinal Blastocystis is linked to healthier diets and more favorable cardiometabolic outcomes in 56,989 individuals from 32 countries. *Cell* (2024) doi:10.1016/j.cell.2024.06.018.
37. Kubota, H. *et al.* Longitudinal investigation of carriage rates, counts, and genotypes of toxigenic Clostridium difficile in early infancy. *Appl. Environ. Microbiol.* **82**, 5806–5814 (2016).
38. Li, Z. *et al.* Clostridioides difficile infection in infants: a case report and literature review. *Gut Pathog.* **15**, 31 (2023).
39. Duncan, S. H., Louis, P. & Flint, H. J. Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. *Appl. Environ. Microbiol.* **70**, 5810–5817 (2004).
40. Ahn, S. *et al.* Agathobaculum butyriciproducens gen. nov. sp. nov., a strict anaerobic, butyrate-producing gut bacterium isolated from human faeces and reclassification of Eubacterium desmolans as Agathobaculum desmolans comb. nov. *Int. J. Syst. Evol. Microbiol.* **66**, 3656–3661 (2016).
41. Feehily, C. *et al.* Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat. Commun.* **14**, 3015 (2023).
42. Solvang, M. *et al.* Beyond purified dietary fibre supplements: Compositional variation between cell wall fibre from different plants influences human faecal microbiota activity and growth in vitro. *Environ. Microbiol.* **25**, 1484–1504 (2023).
43. Chung, W. S. F. *et al.* Modulation of the human gut microbiota by dietary fibres occurs at the species level. *BMC Biol.* **14**, 3 (2016).
44. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME J.* **6**, 1535–1543 (2012).
45. Flint, H. J., Duncan, S. H., Scott, K. P. & Louis, P. Links between diet, gut microbiota composition and gut metabolism. *Proc. Nutr. Soc.* **74**, 13–22 (2015).
46. Sela, D. A. *et al.* The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18964–18969 (2008).
47. Fahur Bottino, G. *et al.* Early life microbial succession in the gut follows common patterns in humans across the globe. *Nat. Commun.* **16**, 660 (2025).
48. Henrick, B. M. *et al.* Colonization by B. infantis EVC001 modulates enteric inflammation in exclusively breastfed infants. *Pediatr. Res.* **86**, 749–757 (2019).
49. Tsalik, E. L. *et al.* Efficacy and safety of azithromycin versus placebo to treat lower respiratory tract infections associated with low procalcitonin: a randomised, placebo-controlled, double-blind, non-inferiority trial. *Lancet Infect. Dis.* **23**, 484–495 (2023).
50. Schouwenburg, S. *et al.* A Pooled Population Pharmacokinetic Study of Oral and Intravenous Administration of Clavulanic Acid in Neonates and Infants: Targeting Effective Beta-Lactamase Inhibition. *Clin. Pharmacol. Ther.* (2024) doi:10.1002/cpt.3423.
51. Hermes, G. D. A., Eckermann, H. A., de Vos, W. M. & de Weerth, C. Does entry to center-based childcare affect gut microbial colonization in young infants? *Sci. Rep.* **10**, 10235 (2020).
52. Thompson, A. L., Monteagudo-Mera, A., Cadenas, M. B., Lampl, M. L. & Azcarate-Peril, M. A. Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *Front. Cell. Infect. Microbiol.* **5**, 3 (2015).
53. Bronfenbrenner, U. *The Ecology of Human Development: Experiments by Nature and Design*. (Harvard University Press, 1979).
54. Ilyka, D., Johnson, M. H. & Lloyd-Fox, S. Infant social interactions and brain development: A systematic review. *Neurosci. Biobehav. Rev.* **130**, 448–469 (2021).
55. Kates, A. E. *et al.* Household Pet Ownership and the Microbial Diversity of the Human Gut Microbiota. *Front. Cell. Infect. Microbiol.* **10**, 73 (2020).
56. Du, G., Huang, H., Zhu, Q. & Ying, L. Effects of cat ownership on the gut microbiota of owners. *PLoS One* **16**, e0253133 (2021).
57. Tun, H. M. *et al.* Exposure to household furry pets influences the gut microbiota of infants at 3–4 months following various birth scenarios. *Microbiome* **5**, 40 (2017).
58. Badovinac, V. P. & Harty, J. T. Adaptive immunity and enhanced CD8+ T cell response to Listeria

monocytogenes in the absence of perforin and IFN-gamma. *J. Immunol.* **164**, 6444–6452 (2000).
59. Wu, Y. *et al.* Epidemiological study of post-pandemic pediatric common respiratory pathogens using multiplex detection. *Virol. J.* **21**, 168 (2024).
60. Xu, Y. *et al.* Antibiotic exposure prevents acquisition of beneficial metabolic functions in the preterm infant gut microbiome. *Microbiome* **10**, 103 (2022).
61. Reyman, M. *et al.* Effects of early-life antibiotics on the developing infant gut microbiome and resistome: a randomized trial. *Nat. Commun.* **13**, 893 (2022).
62. Uzan-Yulzari, A. *et al.* Neonatal antibiotic exposure impairs child growth during the first six years of life by perturbing intestinal microbial colonization. *Nat. Commun.* **12**, 443 (2021).
63. Ianiro, G. *et al.* Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. *Nat. Med.* **28**, 1913–1923 (2022).
64. Andreu-Sánchez, S. *et al.* Global genetic diversity of human gut microbiome species is related to geographic location and host health. *Cell* (2025) doi:10.1016/j.cell.2025.04.014.

FIGURE LEGENDS

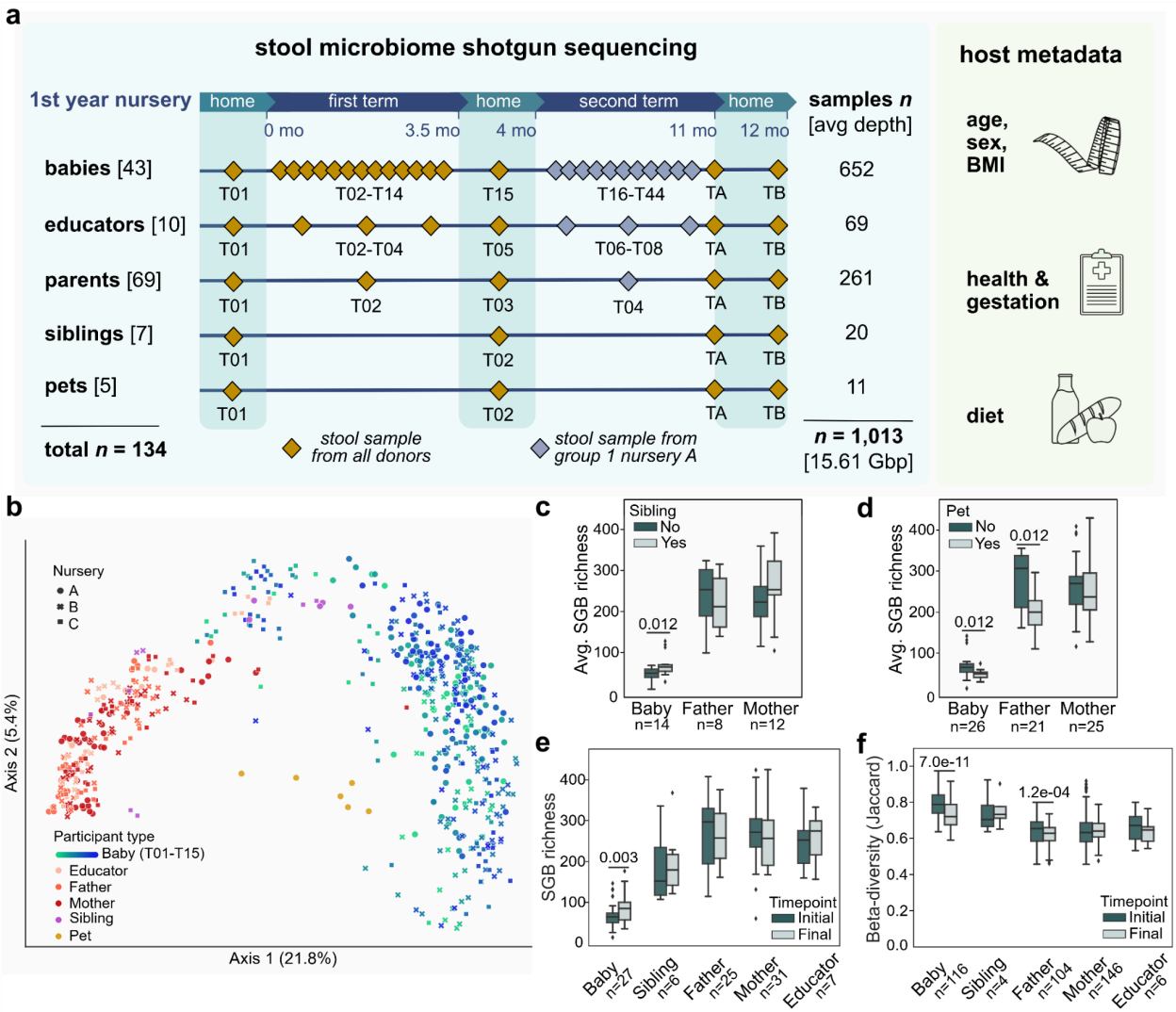


Figure 1. The microTOUCH-baby study and species-level microbiome configurations before and after nursery attendance. (A) microTOUCH-baby study design and overview. **(B)** Species-level microbiome composition overview of the microTOUCH-baby cohort during first term (principal coordinate analysis on Jaccard dissimilarity, $n = 646$). Samples are coloured by host categories, and shapes indicate nursery. Baby samples' color intensity is according to timepoint (from initial-T01 to final-T15). **(C-D)** Average species-level genome bin (SGB) richness across all timepoints and for all individuals in each family member category **C,** having vs not-having a sibling and **D,** having vs non-having a pet. **(E-F) E,** Change in alpha (SGB richness) and **F,** beta diversity (Jaccard dissimilarity) across participant types between beginning and the end of the first term, with *n* indicating the number of subject-subject pairs. Beta diversity refers to the all-vs-all within-nursery dissimilarities. In box plots (C-F), box edges show the lower and upper quartiles, the center line indicates the median, and whiskers expand the interquartile range (IQR). *P*-values are reported when statistically significant (two-sided Mann-Whitney U test in C-D and two-sided Wilcoxon signed-rank tests for E-F), all other comparisons are non significant.

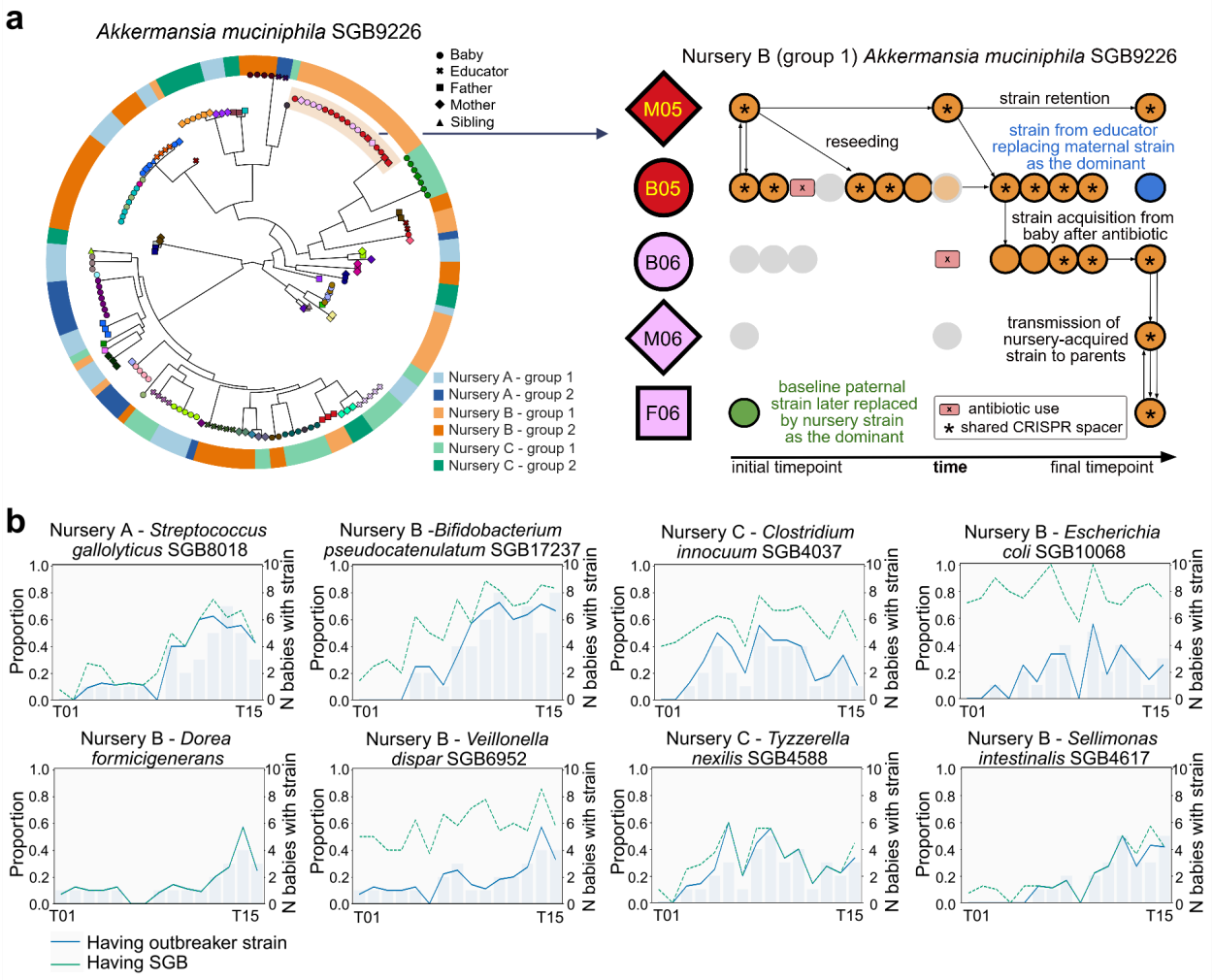


Figure 2. **Interindividual strain transmission and nursery spreading during the first term. (A)** Strain-level profiling for *Akkermansia muciniphila* SGB9226 (left) uncovers the chain of transmission events of one strain of this species in group 1 of nursery B (right). Participant types are identified by shape (mother, diamond; baby, circle; father, square) containing participant identifiers composed of the first letter indicating participant type (M, mother; B, baby; F, father) and the family number; familiar relations are also highlighted by same-color filling. On the right, each circle represents a sample collected from the participants depicted, with color filling indicating the identity of the strain of *A. muciniphila* detected in the sample (except gray, used to indicate the species-level genome bin (SGB) was not detected/typable at the strain level) and arrows indicating the most likely transmission event. Light orange and grey circle identifies SGB9226-positive sample (B05_T08) in which a strain could not be profiled by StrainPhlAn. The identification of shared CRISPR spacers of the target strain of *A. muciniphila* (orange circles) across different samples is indicated by an asterisk. **(B)** Strains present at most in one baby before nursery admission (T01) and spreading to other participants in the same nursery, reaching ≥50% prevalence in following timepoints, until T15. Left and right y-axes show the proportion and number of babies in which the outbreaker strain was detected, respectively. The left y-axis also refers to the proportion of babies in which the SGB was detected (i.e. their prevalence in the nursery group).

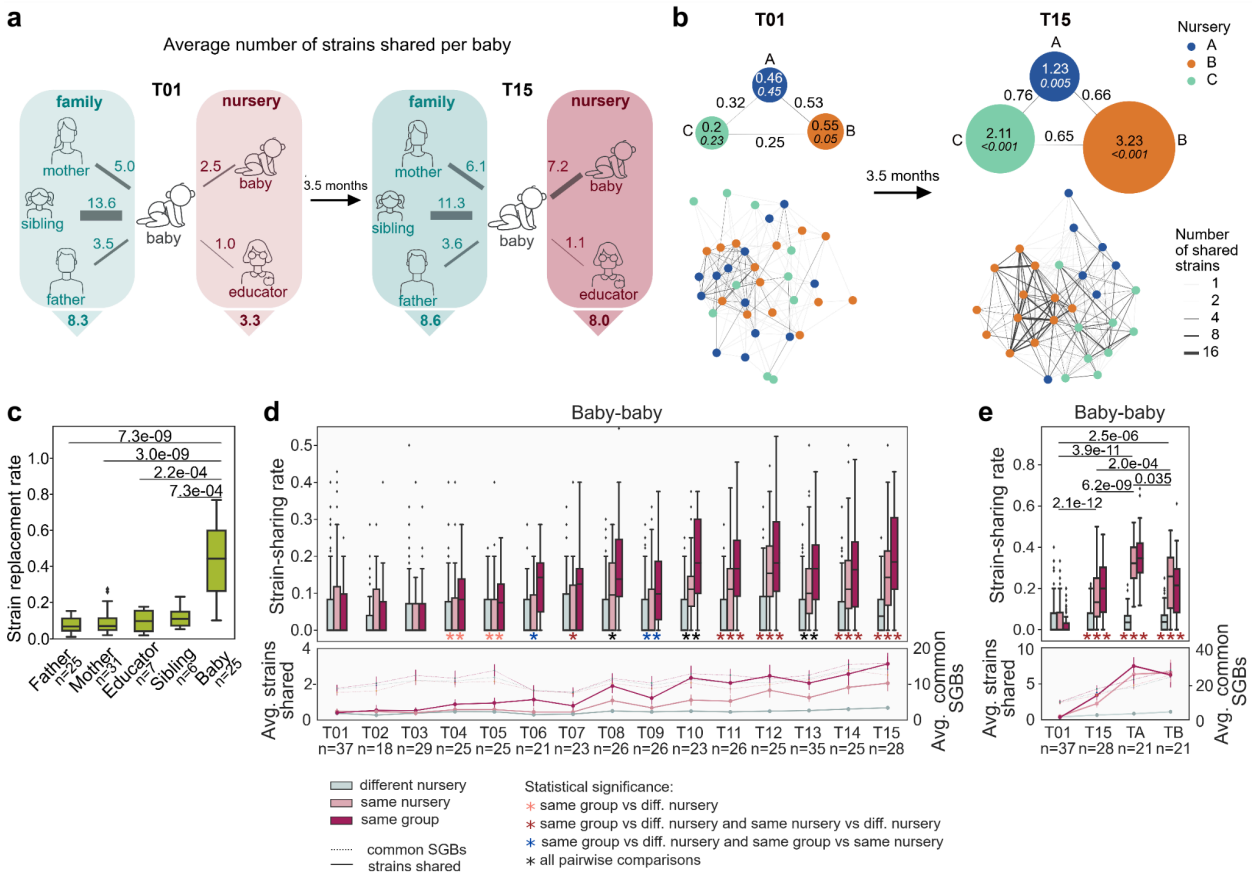


Figure 3. Strain-sharing across hosts before, during and after the first term of nursery attendance. **(A)** Average number of strains shared between each baby and other participants at T01 and T15. The triangles under the boxes report the average number of strains shared between the baby and any participant in "family" (mother, father, sibling) or "nursery" (other babies, educator). **(B)** Average number of shared strains between baby pairs in the same vs different nursery; *P*-values (two-sided permutation test for means; **Methods**) for intra- vs inter-nursery comparisons are shown in italics in the circle. Statistics for A and E are in **Supplementary Tables 9** and **10**. Networks are built on strain-sharing matrices among all babies at T01 and T15. **(C)** Strain replacement rate (one minus the strain-sharing rate) between initial and final timepoints. *P*-values are reported when statistically significant (two-sided Mann-Whitney U tests), all other pairwise comparisons are non significant. In (C-E) box plots, box edges show the lower and upper quartiles, the center line indicates the median, and whiskers expand the IQR. **(D)** Baby-baby strain-sharing rate and average number of strains shared throughout the first term. In D and E statistical significance asterisks refer to the highest significant *P*-value adjusted for multiple comparisons (two-sided permutation test for medians; **Methods**) for the set of comparisons indicated in the legend, with $P < 0.01$, **, $P < 0.001$, ***. Left and right y-axes indicate avg. strains shared and avg. common SGBs. **(E)** Baby-baby strain-sharing rate and average number of strains shared at T01, at the beginning and the end of the second term (T15 and TA), and after the summer break (TB), across all babies in all nurseries. At the top, *P*-values are reported when statistically significant (two-sided Wilcoxon signed-rank test evaluating longitudinal SSR for paired baby-baby pairs attending the same nursery).

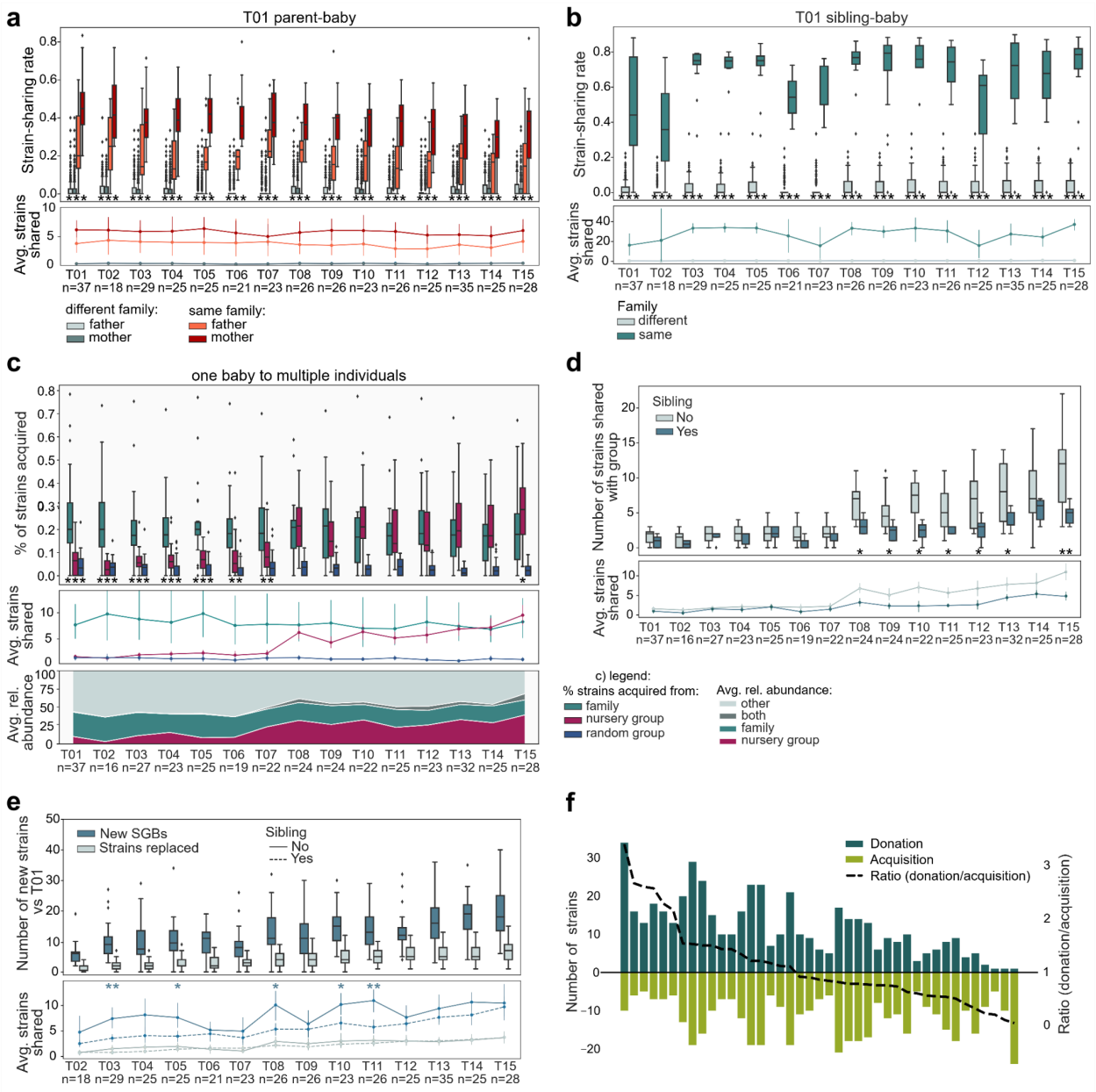


Figure 4. **Dynamics of vertical and horizontal strain transmission during the first term of nursery. (A)** Strain-sharing rate and average number of strains shared between pairs of babies and parents (at T01) from the same vs different families at each timepoint. In (A-B), statistical significance asterisks refer to two-sided permutation tests for medians (**Methods**) adjusted for multiple comparisons for same family vs different family across each family member type, with $P < 0.001$, ***. Exact P-values for Fig. 4A to 4E are provided in **Supplementary Table 11**. In all box plots, box edges indicate the lower and upper quartiles, the center line represents the median, and whiskers expand the IQR. **(B)** Strain-sharing between pairs of babies and siblings (T01) from the same vs different families at each baby timepoint. **(C)** Proportion of strains acquired from group vs family, and corresponding cumulative relative abundance (bottom panel). For each baby timepoint, comparisons were performed against past/contemporaneous samples of the family and the nursery group (**Methods**). Statistical significance asterisks refer only to the proportion of strains acquired from the same group vs the family. In (C-E) the two-sided Mann-Whitney U test was used, with $P < 0.05$, *; $P < 0.01$, **; $P < 0.001$, ***. **(D)** Association between having a sibling and the number of strains acquired from the nursery group. **(E)** Breakdown between acquisition of new SGBs typed at the strain level and strain replacement for the strains acquired from the nursery (top panel) and

association between either means of strain acquisition and having a sibling (bottom panel). Statistical significance asterisks refer to the comparison between SGB acquisition from nursery for babies with vs. without siblings. **(F)** Number of strains either donated (dark green) or acquired (light green) by each baby over the first term.

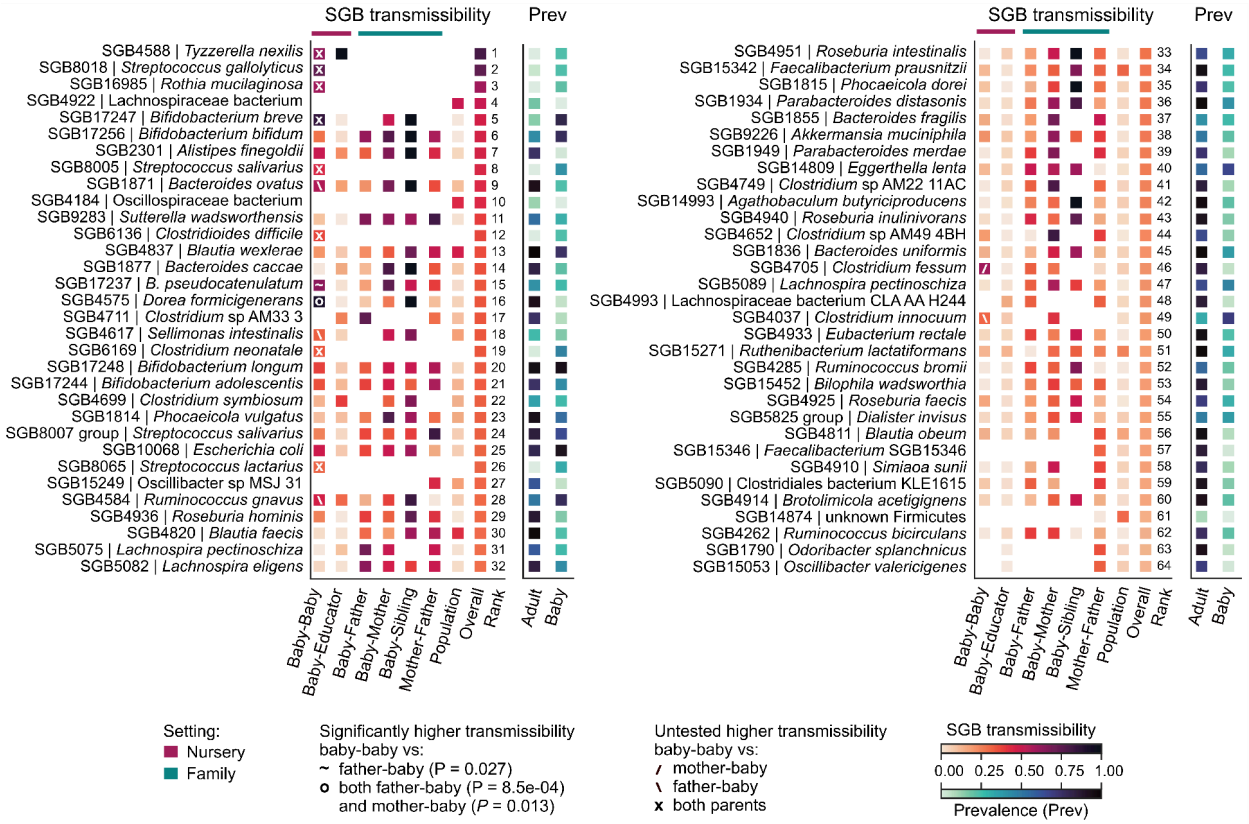


Figure 5. Antibiotic use is associated with lower strain retention and, in babies only, higher strain acquisition. (A) Strain retention rate in adult and infant participants (*n* = 69 and 41, respectively, in panels A-C) that underwent antibiotic treatment (ATB pre-post) vs untreated controls (Ctrl pre-post). In all box plots, box edges indicate the lower and upper quartiles, the center line represents the median, and whiskers expand the IQR. Statistical significant *P*-values in all panels refer to two-sided Mann-Whitney U tests. All other comparisons are non significant. **(B)** Acquisition rate of species-level genome bins (SGBs) typed at the strain level in babies and adults following antibiotic use. **(C)** Strain replacement rate in babies and adults following antibiotic use. Siblings and pets are excluded. Comparisons were performed between consecutive Ctrl pre-post and ATB pre-post timepoints (one per volunteer).

METHODS

Cohort description and recruitment

A total of 134 volunteers comprising babies (4-15 months old at nursery start, median = 10, M = 18, F = 25) about to attend the first year of nursery school, their parents (29-50 years old, median = 36, M = 30, F = 39), siblings (2-21 years old, median = 2, M = 3, F = 4) and house pets (N = 5, two cats and three dogs), and educators (34-56 years old, median = 38.5, F = 10) were recruited and enrolled across three nursery schools (here identified as A, B, C), each with two distinct classes, in the municipality of Trento (Italy) in June 2022. The classes within the same nursery shared few activities (i.e. baby drop-off and pick-up) and spaces throughout the day, and were followed by different educators. The protocol of this study was approved by the Ethics Committee of the University of Trento (protocol nr. 2022-040) and by the Ethics Panel of the European Research Council Executive Agency upon evaluation of the project (microTOUCH Grant agreement ID 101045015). Upon enrollment, volunteers were asked to provide informed consent and complete metadata questionnaires. Consent for participation of infants was obtained directly from parents.

Metadata collection and organization

Date of birth, sex, anthropometric data (weight, height), antibiotic treatment in the three months preceding the start of the study or supplemented during its course, in addition to information regarding putative contacts with other volunteers preceding the beginning of nursery were collected for volunteers of all ages. Metadata specifically collected for babies included gestation length, mode of delivery and general diet at nursery admission (breast or formula milk feeding and weaning date of start). Adult participants were also required to provide information regarding past or ongoing chronic conditions and relative treatments, and putative maternal anti-*Streptococcus* B prophylaxis during birth. Diet metadata for infants and adults are detailed in the next section.

Dietary information collection and analysis

Briefly, most babies had begun weaning at T01 (*n* weaned = 38, *n* non-weaned = 2, NA = 3) and received identical solid meals while in the nursery. The majority followed a mixed feeding approach during weaning, combining solid foods with any type of milk (*n* mixed diet = 24, *n* exclusively solid food = 14, NA = 5). Among those babies receiving milk supplementation, feeding types were relatively balanced (*n* breastfed = 9, *n* formula-fed = 10, *n* receiving both = 5). Finally, adults detailed their long term dietary habits via the compilation of the EPIC Food Frequency Questionnaire (FFQ). FFQs were used to calculate the healthy Plant-based Diet Index (hPDI) [65]. Quality and quantity of plant-based foods were derived from FFQs for a total of eighteen food groups, and divided into quintiles and assigned positive or negative scores. Participants whose intake exceeded the highest quintile received a score of 5, while those below the lowest quintile received a score of 1. Healthy plant-based foods received positive scores, whereas less healthy or unhealthy plant-based and animal-based foods received a negative score. A final score was derived by summarizing the scores of each participant. Metadata were collected and utilized after pseudonymization of volunteers IDs.

Sample collection

Sample collection began a week prior to the start of the first term of nursery (August 2022) and ended after the Christmas holidays (January 2023) for all volunteers. During the first two weeks the nursery organized a “settling-in phase", in which infants were gradually introduced to the nursery and attended it for ~3 hours/weekday. In the following weeks infants attended the nursery for ~8 hours/weekday. Throughout the term length (~14 weeks), stool samples of infant participants were collected weekly (from before nursery admission-T01 to at the end of Christmas holidays-T15) by the nursery staff or the researcher in the nursery from nappies stored at RT on the same day of use, using collection tubes for specimen collection containing 9 ml of DNA/RNA Shield buffer (Zymo). Sample collection was extended until the end of the second term of the year (~30 weeks, ending July 2023) for all donors in group 1 of nursery A, including babies, parents, educators and pets, maintaining sampling timepoint frequencies and modalities. Two follow-up timepoints were collected for all participants enrolled, at the end of the year of nursery (July 2023, “TA”) and at the end of the summer break (August/September 2023, “TB”). The samples collected were moved to the lab and DNA-extracted within 2 weeks of delivery. Samples collection of babies during summer or winter breaks timepoints together with those of siblings and pets were performed directly at home by the parents and stored at RT until the beginning of nursery (maximum 2 weeks later). All adult participants' samples were self collected following detailed instructions, delivered to the lab and processed as previously. Educators donated monthly, while parents collected one additional sample halfway the study period, in addition to initial and final sample timepoints.

DNA extraction and sequencing

After vortex homogenization, DNA was extracted using the DNeasy PowerSoil Pro Kit (Qiagen), following the directions of the HMP protocol (Human Microbiome Project Consortium, 2012b). Additional homogenized aliquots were stored at -20ºC. DNA was quantified using Qubit 2.0 fluorometer (Thermo Fisher Scientific). Sequencing libraries were prepared using the Nextera DNA Library Preparation Kit (Illumina), as described by the manufacturer’s guidelines. The sequencing was performed on the Illumina NovaSeq 6000 platform following manufacturer’s protocols. Sequencing depth was set at 15 Gbp.

Metagenome quality control and preprocessing

Stool samples sequences were pre-processed using the pipeline described at https://github.com/SegataLab/preprocessing. Briefly, metagenomic reads were quality-controlled and reads of low quality (quality score <Q20), fragmented short reads (<75 bp), and reads with >2 ambiguous nucleotides were removed with Trim Galore (v0.6.6). Contaminant and host DNA was identified with Bowtie2 (v2.3.4.3) [66] using the -sensitive-local parameter, allowing confident removal of the phiX 174 Illumina spike-in and human-associated reads (hg19/GRCh37 human genome release). Remaining high-quality reads were sorted and split to create standard forward, reverse, and unpaired reads output files for each metagenome. Metagenomes with at least 1 Gbp were included in the analysis (n = 1,021), while metagenomes with insufficient sequencing depth were excluded (n = 5).

Species-level profiling

Profiling at the resolution of species-level genome bins (SGBs) was performed with MetaPhlAn version 4.1 [16,67] using the vJun23_202307 markers database and using the –unclassified_estimation parameter (**Supplementary Table 4**). SGBs with <0.1% relative abundance in all stool samples were removed from taxonomic profiles for calculation of diversity indices.

Building strain-level phylogenetic trees

To reliably detect strain-sharing events we augmented our dataset with oral samples from the same cohort not analyzed in this study (*n* = 342) and additional samples from 16 public longitudinal cohorts. To do so, we queried the curatedMetagenomicData v3.18 ([68]) for stool samples sequenced at least at 1 Gbp depth from healthy westernized human subjects with at least two time-points per subject and three such subjects per dataset. We went through the corresponding manuscripts and excluded studies involving an intervention between the sampled time-points. Thus we have included samples satisfying the above criteria from the following datasets: ShaoY_2019 [32], MehtaRS_2018 [69], VatanenT_2016 [70], HMP_2019_ibdmdb [71,72], BackhedF_2015 [73], CosteaPI_2017 [74], YassourM_2018 [3], KosticAD_2015 [75], LouisS_2016 [76], FerrettiP_2018 [1], HallAB_2017 [77], WampachL_2018 [78], NielsenHB_2014 [79], Heitz-BuschartA_2016 [80], ChuDM_2017 [81], AsnicarF_2017 [82]. In total, phylogenetic trees were built using data from 1,405 samples from this cohort and 4,322 samples from additional cohorts.

For all the SGBs detected in our cohort we queried MetaRefSGB, a microbial genomic database containing >156k isolate genomes and >952k metagenome-assembled genomes (MAGs) as of vJun23 [16], for isolate genomes or MAGs from food sources, and included them to the trees as references [83].

To build each tree we included all the samples for which the SGB was detected by MetaPhlAn. SGBs detected in less than 20 samples from our cohort were discarded, remaining with 1,363 SGBs. The strain-level phylogenetic trees were built for each SGB with StrainPhlAn v4.1 [16,67]. For 1,107 SGBs we were able to build the phylogenetic tree, the remaining 256 didn't have a sufficient number of samples with enough marker genes with minimal coverage as reported by StrainPhlAn.

Assessment of strain-sharing rates

We discarded trees built from alignments shorter than 1000 nt (*n* = 111 SGBs discarded). In the phylogenetic trees with genomes from a food source we considered the ANI on the marker genes (mutation rates in StrainPhlAn) and discarded samples closer than 99.85% ANI as likely coming from food. When more than 20% of the samples would be discarded, we dropped the SGB altogether (*n* = 7 SGBs discarded).

We calculated phylogenetic distances between all samples as the length of the shortest path between the samples along the tree branches. We normalized distance distributions within each tree by dividing by the median distance.

To define strain-sharing events we calculated a threshold best separating the within-subject distribution from the across-subjects distribution of phylogenetic distances. For the within-subject distribution, pairs of samples from the same subject sampled maximum 6 months apart were considered using maximum one pair per subject. Among the possible pairs, we chose the one maximizing the chances to type the strains of the SGB of interest in both samples, and we did so by choosing the pair for which the sample with the lowest coverage for the SGB between the two samples was the highest among all pairs. In case of ties we maximized the higher estimated coverage of the two. The sample coverage was estimated as the sequencing depth times the SGB's relative abundance. For the across-subjects distribution we pick pairs of samples coming from different datasets, one sample per subject maximizing the coverage. When there were less than 50 pairs in the across-subject distribution we discarded the SGB (*n* = 365 SGBs discarded). When there were less than 20 pairs in the within-subject distribution (*n* = 276 SGBs), we calculated the threshold as the 3rd percentile of the across-subject distribution, i.e. setting the expected false discovery rate (FDR) to 3%. When the within-subject distribution had at least 20 pairs (*n* = 309 SGBs), the threshold separating the distributions was calculated as maximizing the Youden's index, unless the expected FDR exceeded 5%, in which case we set the threshold as the 5th percentile of the across-subject distribution (*n* = 61 SGBs). After manual curation of the trees, including outlier branches removal, further 112 SGBs were discarded. In total, we assessed strain-sharing for 512 SGBs.

For each pair of samples, we called a strain-sharing event of an SGB when their phylogenetic distance in the corresponding tree was lower than the corresponding calculated threshold.

For each sample we considered the SGBs in which corresponding trees it was placed, i.e. profiled by StrainPhlAn. And finally, we define the strain-sharing rate as the number of strain-sharing events divided by the number of SGBs profiled at the strain level in both of them by StrainPhlAn. For pairs with less than 5 SGBs profiled in common we set the strain-sharing rate as undefined and such pairs were excluded from strain-sharing rate analysis. Strain retention rate is defined as the within-individual strain-sharing rate. Within-individual strain replacement rate is defined as the number of strains not retained longitudinally among retained SGBs over the number of retained SGBs.

Samples found to be contaminated or mislabeled according to strain-sharing analysis (and validated with CrocoDeEL v1.0.6 [84]) were removed post-hoc (*n* = 8), reducing the dataset to 1,013 metagenomes. Samples collected during antibiotic treatment (*n* = 26) were excluded from all following analyses.

Within-SGB strain heterogeneity, strain-sharing networks, SGB transmissibility

Within-SGB strain heterogeneity (**Ext. Data Fig. 7A**) was computed on a per-nursery basis as the number of strains of a given SGB present among babies at a given timepoint over the number of babies having the SGB at the same timepoint. Within-SGB strain heterogeneity = 1 indicates there is no strain-sharing among babies having the SGB (i.e. they all have a different strain of the SGB).

Strain-sharing matrices were used to build unsupervised strain-sharing networks (**Fig. 3B**) with the R packages ggraph (v2.2.1) and tidygraph (v1.3.1), where only nodes with degree >0 are shown.

SGB transmissibility (**Ext. Data Fig. 9A**) was computed as the number of individual pairs sharing the strain over the total number of potentially callable strain-sharing events involving the SGB (i.e. the number of pairs of individuals in which the SGB was present and typable according to StrainPhlAn v4.1) [8]. When the SGB was not present among at least 3 pairs within a category, transmissibility of the SGB within the category was set as undefined. Differential transmissibility between baby-baby and baby-mother or baby-father pairs was assessed for all SGBs having at least 10 baby-baby pairs sharing it, with application of a Fisher's exact test (including FDR control) in case there were at least 10 baby-mother or 10 baby-father pairs sharing the SGB, otherwise the number of SGB- and strain-sharing events were reported for each group without application of the test.

Comparison of the contribution of familial and nursery strains to the infant microbiome

To compare the contribution of familial and nursery strains to infants' microbiome composition, for each baby, we computed at each baby timepoint (T01 to T15) the number of strains shared either with any member of the family or with any other infant of the nursery group, disregarding strains shared with both (unless otherwise stated). This allowed us to compute the proportion of strains for a given baby microbiome that was putatively acquired from either the nursery or family (referred to in the text and figures as the "proportion of strains acquired"), as the strains exclusively shared with either one of the two groups of individuals. Considering that family members were less densely sampled than infants, we considered a maximum of three samples for each of the other babies in the nursery group (baseline-T01, halftime-T08, and final-T15, when available, emulating sampling timeline of parents), adding other babies and family samples to the longitudinal analysis considering the time in which they were sampled (i.e. looking for strain-sharing only with samples collected in the past or contemporaneous to the target timepoint of the target baby). This likely explains the non-linear increases of proportion of strains acquired from the nursery group at baby T08 and T15 observed (e.g. **Fig. 4C**). Moreover, as a negative control for the considerably larger number of individuals in the nursery group (avg. n = 7) in comparison with the family (avg. n = 2), we also analyzed the strain-sharing dynamics with a random nursery group from another nursery.

Metagenomic assembly and CRISPR analysis

Metagenome-assembled genomes (MAGs) were generated through a previously validated metagenomic assembly pipeline [17], including assembly of contigs with MEGAHIT (v1.1.1) [85], calculation of contigs coverage with Bowtie2 (v2.2.9) [66], binning of contigs with MetaBAT2 (v2.12.1) [86], and quality-checking of bins with CheckM (v1.1.3) [87]. Medium- and high-quality MAGs were identified following the criteria previously proposed [88] and low-quality MAGs were discarded. Average nucleotide identity (ANI) between MAGs was computed using skani (v.0.2.1) [89].

To validate the chain of transmission events of the strain of *Akkermansia muciniphila* SGB9226 shown in **Fig. 2A**, CRISPR arrays were identified from MAGs using MinCED (v0.4.2, default parameters) [90]. CRISPR spacers and repeats were extracted from raw sequencing reads using Crass (v1.0.1, parameters -d 20 -D 55 -s 20 -S 55 --longDescription) [91]. Following the identification of 5 CRISPR arrays and 39 CRISPR spacers in this set of MAGs, we looked for the CRISPR spacers of this strain in the metagenomic reads of the whole dataset. While single CRISPR spacers were found in the metagenomic reads of up to 16/19 samples containing the strain (**Fig. 2A**), none of the 39 CRISPR spacers could ever be found in the remaining 627 metagenomes, providing independent validation for the trajectory of the strain in the nursery group.

***Bifidobacterium longum* strains assignment to subspecies**

To evaluate the transmissibility of distinct subspecies of *Bifidobacterium longum* SGB17248 in our dataset, we constructed a StrainPhlAn 4 phylogenetic tree of all 591 *B. longum* strains identified in our cohort, which revealed two distinct clusters (cluster_1 and cluster_2, **Ext. Data Fig. 10C-E**). Cluster_1 consisted exclusively of strains present in infants and hence was hypothesized to belong to *B. longum* subsp. *infantis*, while cluster_2 contained strains from both infants and adults, possibly representing other subspecies of *B. longum*. To definitively assign these strains to subspecies, we succeeded in generating MAGs (as described above) for 262 of the 591 strains and then calculated the ANI between these MAGs and 15 reference genomes representing the three known *B. longum* subspecies [92] (five reference genomes each for subsp. *longum*, subsp. *infantis*, and subsp. *suis*). The 24 MAGs from cluster_1 showed highest similarity to *B. longum* subsp. *infantis* reference genomes (median ANI = 98.04%), compared to lower ANI values with subsp. *longum* (95.60%) and subsp. *suis* (96.13%). Conversely, the 238 MAGs from cluster_2 displayed the highest genomic similarity to *B. longum* subsp. *longum* reference genomes (median ANI = 98.87%), with substantially lower similarity to subsp. *suis* (96.82%) and subsp. *infantis* (95.39%). The 591 *B. longum* strains as profiled with StrainPhlAn 4 were then assigned according to their genome-level assignment.

***Blastocystis* detection and ST profiling**

The presence of *Blastocystis* in metagenomic samples was assessed using a previously validated computational workflow [93]. In brief, nine reference genomes for eight distinct *Blastocystis* STs (i.e., ST1 [ST1_LXWW01], ST2 [ST2_JZRJ01], ST3 [ST3_JZRK01], ST4 [GCF_000743755 & ST4_BT1_JZRL01], ST6 [ST6_JZRM01], ST8 [ST8_JZRN01], and ST9 [ST9_JZRO01]) were mapped against metagenomic reads with Bowtie2 (v2.5). Then, SAMtools (v1.19) and bedtools (v2.30) were used to compute the breadth of coverage of each genome. We reported a sample to be positive for a *Blastocystis* ST if the respective genome had a breadth of coverage of at least 10%.

PCR validation of transmission of an *Akkermansia muciniphila* SGB9226 strain

To further test how much potential problems with limit of detection in the metagenomic approach could influence strain transmission inference, we implemented a SGB-specific PCR assay for *A. muciniphila* (SGB9226) and applied it to the representative example depicted in **Fig. 2A**. We designed the primers (F-5'-TGACTGGACTCTATTGCCTGAAG-3' and

R-5'-GCCTTTCAATATGCCCTTCGTAC-3'; amplicon length = 101 bp) to recognize the SGB9226-specific core gene (UniRef90_A0A2N8IRV1; identified by MetaPhlAn 4 [16]), using ConsensusPrime (v.1.0, with set consensus similarity = 0.8 and consensus threshold = 0.95) [94] and Primer3 (v. 2.6.1, with set primer size = 18-28 bp, optimal Tm = 57-63°C, GC content = 40-60%, PRIMER_MAX_HAIRPIN_TH = 24.0, PRIMER_INTERNAL_MAX_HAIRPIN_TH = 24.0, PRIMER_MAX_END_STABILITA = 9.0) [95]. Assay sensitivity was independently evaluated using a spike-in approach with DNA from the *A. muciniphila* type strain ATCC BAA-835 (from 1M down to 1 single genome copy) into an *A. muciniphila*-negative fecal test sample. The assay achieved a limit of detection equivalent to a single genome copy of *A. muciniphila* (**Ext. Data Fig. 6A**). PCRs were performed using GoTaq® G2 Hot Start Green Master Mix (Promega) with 500 nM of each primer. The thermal cycling program included an initial denaturation at 95°C for 10 minutes, followed by 40 cycles of annealing at 62°C. Positive bands were observed by electrophoresis on a 2% agarose gel.

Statistical analysis

Statistical analyses were performed in Python (v3.10.12) using libraries scikit-bio (v0.5.9), scipy (v1.10.1), and statsmodels (v0.14.0). Cross-sectional comparisons between groups were performed using the Mann-Whitney U test (two groups) or the Kruskal-Wallis test with post-hoc Dunn tests (multiple groups) for independent observations. Cross-sectional dependent observations were compared with a permutation test for medians (or means, when comparing the number of shared strains), with *P*-value calculated as the proportion of times (out of a 1,000 permutations) that the observed difference in the median between groups with shuffled labels is equal or more extreme than the observed in the correctly labeled data. Longitudinal comparisons between timepoints were performed using the Wilcoxon signed-rank test. Jaccard dissimilarity matrices were computed from taxonomic profiles and compositional differences between groups were evaluated using PERMANOVA. When appropriate, correction for multiple testing was applied using the Benjamini-Hochberg procedure (P_{adj}), with significance defined as $P_{adj} < 0.05$.

ADDITIONAL REFERENCES

65. Satija, A. *et al.* Healthful and unhealthful plant-based diets and the risk of coronary heart disease in U.s. adults. *J. Am. Coll. Cardiol.* **70**, 411–422 (2017).
66. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
67. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, (2021).
68. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature Methods* vol. 14 1023–1024 Preprint at https://doi.org/10.1038/nmeth.4468 (2017).
69. Mehta, R. S. *et al.* Stability of the human faecal microbiome in a cohort of adult men. *Nat Microbiol* **3**, 347–355 (2018).
70. Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**, 842–853 (2016).
71. Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol* **3**, 337–346 (2018).
72. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
73. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).
74. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).
75. Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in development and in progression toward

type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).
76. Louis, S., Tappu, R.-M., Damms-Machado, A., Huson, D. H. & Bischoff, S. C. Characterization of the Gut Microbial Community of Obese Patients Following a Weight-Loss Intervention Using Whole Metagenome Shotgun Sequencing. *PLoS One* **11**, e0149564 (2016).
77. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).
78. Wampach, L. *et al.* Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat. Commun.* **9**, 5091 (2018).
79. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
80. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* **2**, 16180 (2016).
81. Chu, D. M. *et al.* Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* **23**, 314–326 (2017).
82. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, (2017).
83. Carlino, N. *et al.* Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. *Cell* 1–21 (2024).
84. Goulet, L. *et al.* CroCoDeEL: accurate control-free detection of cross-sample contamination in metagenomic data. *bioRxiv* (2025) doi:10.1101/2025.01.15.633153.
85. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
86. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
87. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
88. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
89. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
90. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
91. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* **41**, e105 (2013).
92. Mattarelli, P., Bonaparte, C., Pot, B. & Biavati, B. Proposal to reclassify the three biotypes of Bifidobacterium longum as three subspecies: Bifidobacterium longum subsp. longum subsp. nov., Bifidobacterium longum subsp. infantis comb. nov. and Bifidobacterium longum subsp. suis comb. nov. *Int. J. Syst. Evol. Microbiol.* **58**, 767–772 (2008).
93. Beghini, F. *et al.* Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. *ISME J.* **11**, 2848–2863 (2017).
94. Collatz, M., Braun, S. D., Monecke, S. & Ehricht, R. ConsensusPrime - a bioinformatic pipeline for alignment filtering and ideal consensus primer design. *Research Square* (2022) doi:10.21203/rs.3.rs-1030641/v2.
95. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).

Data availability

Shotgun metagenomic data for microTOUCH-baby are available at the NCBI-SRA under accession number PRJNA1140720, with respective sample-wise metadata information available in **Supplementary Table 2** and at https://doi.org/10.5281/zenodo.17663257. Accession numbers for sequencing data from additional longitudinal datasets can be found in the original publications. The human genome release used for host decontamination is available at NCBI RefSeq (GCF_000001405.13).

Code availability

All the software used in this study are available in the MetaPhlAn4 package (which includes StrainPhlAn 4 and the script for strain transmission inference), available at http://segatalab.cibio.unitn.it/tools/metaphlan with the open-source code at https://github.com/biobakery/MetaPhlAn.

Acknowledgments
We thank all study participants for their commitment and A. Zanetti, L. Zambaldi and R. Vit of the Childcare Service Office of the Trento Council (Trento, Italy), all nursery coordinators and educators for their collaboration; the CIBIO Next Generation Sequencing Facility of the University of Trento (V. de Sanctis, R. Bertorelli, P. Cavallerio and C. Valentini) for sequencing the metagenomic libraries, B. Servaes for support in designing figure icons. This work was supported by the European Research Council (ERC-CoG microTOUCH-101045015) to N.S.

Author contributions N.S. conceived and supervised all steps of the study, N.S. and L.R. performed the study design, L.R. managed cohort recruitment, data, and sample acquisition, F.P. participated in cohort recruitment, C.S., F. Armanini and L.R. performed DNA extraction and shotgun sequencing libraries preparation, V.H., M.P., M.C., E.P., M.V.C., and F. Asnicar performed and supported the computational analyses, A.N. and F.F. performed SGB-specific PCRs, L.R., V.H. and N.S. interpreted the data and wrote the manuscript. All authors revised the manuscript and approved the final version for submission.

Competing interests The authors declare no competing interests.

EXTENDED DATA LEGENDS

Extended Data Figure 1. Quality metrics and baseline SGB richness per participant type. (A) Distribution of unclassified reads percentage across all samples analyzed in the study (*n* = 1,013), as estimated using MetaPhlAn v4.1. **(B)** T01 alpha diversity represented as species-level genome bin (SGB) richness divided by participant type. In box plots, box edges indicate the lower and upper quartiles, the center line represents the median, and whiskers expand the interquartile range (IQR). Statistically significant *P*-values refer to two-sided Mann-Whitney U tests. All other pairwise comparisons are non significant.

Extended Data Figure 2. Most abundant species-level genome bins (SGBs) in babies and adults at baseline. Heatmap shows SGBs with prevalence ≥30% at 0.1% relative abundance in either babies or adults. The SGBs are ordered based on the difference between the median relative abundance among babies and adults.

Extended Data Figure 3. Microbiome compositional structure and association with species-level genome bin (SGB) richness. (A) Ordination overview of the species-level microbiome composition of the microTOUCH-baby cohort (principal coordinate analysis on Jaccard dissimilarity, *n* = 646), with samples coloured by SGB richness and shapes indicating participant types. **(B)** Association between SGB richness and first four PCoA ordination components. Statistical significance refers to the two-sided Spearman's test (*n* = 646).

Extended Data Figure 4. Correlation between microbiome diversity and baby age, or adult diet quality. (A–B) A, Association between baby microbiome diversity at T01 (*n* = 36) and **B,** T15 (*n* = 27) and baby age before nursery start (T01). **(C–D) C,** Association between Jaccard dissimilarity at T01 (*n* = 491

baby-baby pairs) and **D,** T15 (*n* = 312 baby-baby pairs) and age difference (months), in babies. **(E)** Association between Jaccard dissimilarity during the first term of nursery attendance (T01-T15) and baby age (months) at T01 (*n* = 26). **(F)** Association between species-level genome bin (SGB) richness before nursery start (T01) and healthy Plant-based Diet Index (hPDI), as an indicator of diet quality, in adult volunteers (*n* = 69). In (A-F) statistical significance refers to the two-sided Spearman's test.

Extended Data Figure 5. Perinatal associations with baseline baby microbiome diversity, longitudinal baby species-level genome bin (SGBs) richness, or strain-sharing rates between babies and family members. (A-B) A, Association between alpha diversity in babies at T01 and maternal intrapartum antibiotic prophylaxis against *Streptococcus* B and **B,** mode of delivery. In box plots (A-B, F-G), box edges represent the lower and upper quartiles, the center line indicates the median, and whiskers expand the IQR. Statistical significance refers to two-sided Mann-Whitney U tests, with $P \geq 0.05$, ns for comparisons in A-B. **(C-D)** Ordination overview of the species-level microbiome composition of babies at T01 (principal coordinate analysis on Jaccard dissimilarity, *n* = 37) in relation with birth practices. Statistical significance refers to the two-sided PERMANOVA test. Samples are coloured according to **C**, administration of maternal intrapartum antibiotic prophylaxis and **D,** mode of delivery. **(E)** Change in total species-level genome bin (SGB) richness of all babies in the three nurseries throughout the first term of nursery. Average SGB richness is shown by the black dotted line. **(F)** Mother-baby strain-sharing rates in Valles-Colomer *et al*. (V-C), 2023, stratified in the [1-15] and [4-15] months age ranges (median SSR= 48% and 45%), and in this study (median SSR = 50%). Only pairs of samples with at least 10 SGBs in common were included. Statistical significance refers to two-sided Mann-Whitney U tests, V-C 2023 [1-15] months vs this study (U = 474.0), and V-C 2023 [4-15] months vs this study (U = 198.0), with $P \geq 0.05$, ns, for both comparisons. **(G)** Strain-sharing rate between family members pairs within the same vs different families before nursery start (T01). Statistically significant *P*-values refer to two-sided permutation test for medians (**Methods**) adjusted for multiple comparisons, with *n* indicating the number of subject-subject pairs.

Extended Data Figure 6. Validation and dynamics of multi-host strain transmission. (A-B) Validation of the chain of transmission of a strain of *Akkermansia muciniphila* SGB9226 (Fig. 2A) using a SGB-specific PCR assay; **A,** Sensitivity of the SGB9226-specific PCR assay, assessed by testing a SGB9226-negative human fecal DNA spiked with seven ten-fold dilutions of genomic DNA from *A. muciniphila* ATCC BAA-835, corresponding to an estimated 10^6 to 1 genome copies ("D1" to "D7"). Controls included the human fecal matrix alone ("Spike-in Matrix") and a no-template control (NTC). **B,** Application of the SGB9226-specific PCR to fecal samples in Fig. 2A. Sample IDs include the volunteer type and family number, while per volunteer longitudinal samples are identified by timepoint. Relative abundances according to MetaPhlAn 4 ("Rel. Abun., %") and sample IDs are colored dark blue (positive) or grey (negative) based on strain identification according to StrainPhlAn (Fig. 2A). Genomic DNA from *A. muciniphila* ATCC BAA-835 is included as a positive control, alongside a NTC. **(C)** Phylogenetic tree of *Alistipes finegoldii* SGB2301 (left) and chain of transmission events of one strain in group 1 of nursery B (right). Participant types are identified by shape (educator, cross; baby, circle; mother, diamond) containing participant identifiers (with the initial identifying the participant type followed by specific family number). Familial relations are highlighted by same-color filling. Each circle represents a sample collected from participants depicted, with color filling indicating the identity of the strain of *Alistipes finegoldii* SGB2301 detected in the sample (except gray, used to indicate the SGB was not detected/typable) and arrows indicating the most likely transmission event. **(D-E)** Strain-sharing between pets (*n* = 5) and human hosts within the same family (turquoise) and across different ones (gray). **D,** The average total number of strains shared between participant pairs across contemporaneous samples are reported above the connecting lines; in brackets are indicated the number of pet-human pairs with at least one strain shared over the total number of pet-human pairs. Statistical significance according to a two-sided Fisher's exact test is depicted for the same family vs different family pet-baby comparison. All other comparisons are non significant. **E,** Average number shared strains between pet and human across contemporaneous samples, in different vs same family. In the box plots, box edges indicate the lower and upper quartiles, the center line represents the median, and whiskers expand the IQR. Statistical significance *P*-values refer to two-sided Mann-Whitney U tests, with *n* indicating the number of pet-human pairs. All other comparisons are non significant.

Extended Data Figure 7. Trends of strain transmission, retention, replacement and strain contributions of family and nursery to the infant microbiome. (A) Change in average number of strains per SGB typed at the strain level present in babies separated by nursery at each timepoint. The dotted line shows the trend in the average among the three nurseries. Species-level genome bin (SGB) retention rate during the first term in different participant types. **(B)** The metric is defined as the Jaccard similarity between samples from initial and final timepoints of the same subject. Statistically significant *P*-values refer to two-sided Mann-Whitney U tests. All other pairwise comparisons are non significant. In box plots (B, E-F), box edges indicate the lower and upper quartiles, the center line represents the median, and whiskers expand the IQR. **(C)** Association between overall strain replacement rates and group age among non-baby participants. Statistical significance refers to the two-sided Spearman's test (n = 68). **(D)** Longitudinal average baby-baby strain-sharing rate (with each baby-baby observation in the background) and average number of strains shared during the entire first year of nursery, for group 1 of nursery A. Only pairs with at least 5 species-level genome bins (SGBs) in common were included. In the avg. strains shared (left *y*-axis)/avg. common SGBs (right *y*-axis) subplot, the vertical lines report the standard deviation, with n indicating the number of baby-baby pairs. **(E)** Percentage of strains acquired from either their family, nursery group and a random different group from another nursery (as a control) at T01 and T15. **(F)** Relative abundance (ra) of strains acquired from either family, nursery group and both, at T01 and T15. Statistically significant *P*-values refer to two-sided Mann-Whitney U tests, except for the ra family vs ra group comparison, for which the two-sided Wilcoxon signed-rank test was used. All other comparisons are non significant.

Extended Data Figure 8. Longitudinal strain acquisition and strain retention during and after the second term. (A-B) A, Proportion of strains acquired from group vs family, corresponding average number of strains acquired and total relative abundance, during the entire first year of nursery, for group 1 of nursery A only and **B,** at the end of the second trimester (TA) and at the end of summer break (TB), for all participants in the study. In (A-B), for each baby timepoint, comparisons were performed against past/contemporaneous samples of the family and the nursery group (**Methods**). Statistically significant *P*-values refer to a two-sided Mann-Whitney U test for the proportion of strains acquired from the same group vs the family. Comparisons at other timepoints are non significant. In all box plots (A-F, I-J), box edges indicate the lower and upper quartiles, the center line represents the median, and whiskers expand the IQR. **(C)** Proportion of SGBs typed at strain-level that were lost after the summer break (TB) over those present at year-end (TA), in adult and infant participants. In C–F, significant *P*-values refer to a two-sided Mann-Whitney U test. All other comparisons are non significant. **(D)** Proportion of SGBs typed at strain-level that were newly acquired during the summer break (TB) over the total present at TB, in adult and infant participants. **(E)** Proportion of strains retained after summer break (TB) over the total found at year-end (TA), in adult and infant participants. **(F)** Proportion of strains replaced at TB over the total strains present at TA, in adult and infant participants. **(G–H) G,** Proportion of baby strains acquired from either family or nursery that were retained or **H,** replaced by different conspecific strains after summer break (TB) over the total strains found at year-end (TA), in infant participants. In G–H, significant *P*-values refer to two-sided Wilcoxon signed-rank tests. **(I)** Strain-sharing rate between babies and other family members at T01. **(J)** Average number of strains shared between pairs of babies and siblings or one parent or more than one family member at the time. In I-J, statistically significant *P*-values refer to two-sided Mann-Whitney U tests. All other pairwise comparisons are non significant.

Extended Data Figure 9. Species transmissibility in the nursery and family settings. (**A**) The 64 species-level genome bins (SGBs) typed at the strain level with highest transmissibility scores (**Methods**) in the nursery and among family members, divided per participant type pairs and ordered by their overall transmissibility score (avg. across all participant type pairs). The prevalence among adults (parents and educators) and babies at T01 are shown. SGBs included had a transmissibility score of ≥ 0.3 (computed over at least 10 pairs containing the SGB) for at least one participant type pair. The transmissibility score was not computed for participant type pairs having less than 3 pairs containing the SGB. Symbols in squares highlight significantly higher (**Methods**) baby-baby transmissibility over baby-adult pairs (with exact *P*-values provided in the legend and in **Supplementary Table 14**). "Untested higher transmissibility" indicates a transmissibility difference (transmissibility baby-baby - transmissibility baby-adult) ≥ 0.3 for cases in which not enough SGB-sharing pairs were identified among baby-adult pairs for reliable statistical comparison (**Methods**), and was also highlighted using symbols (legend). **(B)** Association

between SGB transmissibility between pairs of different participant types and SGB prevalence (in adults, babies, or the ratio of prevalence baby/adults). “Population” includes all unrelated adults in the microTOUCH-baby study cohort, while “Overall” reports the average of all category pairs in which the SGB transmissibility could be computed, including unrelated volunteers. Statistical significance asterisks refer to two-sided Spearman’s tests, with $P_{adj} < 0.05$, *; $P_{adj} < 0.01$, **; $P_{adj} < 0.001$, ***. For example, higher transmissibility of prevalent SGBs was captured among unrelated volunteers, such as adults (Spearman’s test, $n = 427$ SGBs, $\rho = 0.24$, $P_{adj} = 4.6e\text{-}6$) and babies in the nursery (Spearman’s test, $n = 129$, $\rho = 0.21$, $P_{adj} = 0.031$). **(C)** Association between SGB transmissibility between pairs of different participant types (Baby-Baby, $n = 129$ SGBs; Baby-Father, $n = 80$; Baby-Mother, $n = 97$) and SGB prevalence ratio (Baby/Adult). *P*-values and correlation coefficients (ρ) according to a two-sided Spearman’s test are shown at the bottom of the figure.

Extended Data Figure 10. Highly transmissible species baby-to-baby. (A–B) A, Phylogenetic tree of *Bifidobacterium breve* SGB17247 and **B,** *Dorea formicigenerans* SGB4575 that were found to be differentially more transmitted baby-to-baby in the nursery setting, compared to parent-baby pairs. In phylogenetic trees (A-C), participants are identified by shape and colour in the inner ring, while nursery group affiliation by colour in the outer ring, as indicated in the legend. **(C)** Phylogenetic tree of *Bifidobacterium longum* SGB17248. Colour coding and symbols as in A-B**. (D)** Intra-subspecies average nucleotide identity (%) used to define subspecies-level clusters within *B. longum***.** In box plots, box edges indicate the lower and upper quartiles, the center line represents the median, and whiskers expand the IQR. **(E)** Prevalence of *B. longum* subsp. *longum* and subsp. *infantis* in babies during the first term of nursery.

Extended Data Figure 11. Correlation of baby strain donation and acquisition with age or birth practices. (A–C) A, Association between number of strain donations **B**, acquisitions and **C,** ratio of donations over acquisitions and baby age (months) at T01. Statistical significance refers to the two-sided Spearman’s test ($n = 39$). **(D-E) D,** Association between the number of strains shared with the nursery group and mode of delivery (C-section or vaginal) and **E,** maternal intrapartum anti-*Streptococcus* B prophylaxis. In both box plots, box edges show the lower and upper quartiles, the center line indicates the median, and whiskers expand the IQR. Statistical significance asterisks refer to two-sided Mann-Whitney U tests, with $P < 0.05$, *. All other comparisons are non significant. In both the avg. strains shared subplots, the vertical lines report the standard deviation.

Extended Data Figure 12. Effect of infant feeding regime and antibiotic use on strain transmission. (A–B) A, Mother-baby strain-sharing rate in babies drinking vs not drinking any type of milk and **B,** in babies receiving maternal milk vs formula milk at T01. In box plots (A-N), box edges represent the lower and upper quartiles, the center line indicates the median, and whiskers expand the IQR. In (A-N), statistically significant *P*-values refer to two-sided Mann-Whitney U tests. All other comparisons are non significant. **(C-E)** Association between number of strain donations and **C**, weaning status **D,** supplementation of milk and **E,** milk type at T01. **(F-H)** Association between number of strain acquisitions and **F,** weaning status **G**, supplementation of milk and **H,** milk type at T01. **(I–K) I,** Association between ratio of strain donations over acquisitions and weaning status, **J,** supplementation of milk and **K,** milk type at T01. **(L-N) L**, Absolute number of retained strains, **M**, absolute number of newly acquired SGBs (i.e. acquisition of strains assigned to previously undetected SGBs) and **N**, absolute number of strains replaced in adult and infant participants ($n = 69$ and 41) that underwent antibiotic treatment (ATB pre-post) vs untreated controls (Ctrl pre-post). Siblings and pets are excluded. Comparisons were performed between consecutive Ctrl pre-post and ATB pre-post timepoints (one per volunteer).

Chapter 3: Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases

Context and contribution

In this work, we assembled a meta-cohort of 226 donor-recipient pairs across 24 studies and 8 different disease types to infer patterns of microbial engraftment during fecal microbiota transplantation (FMT) and its relations to disease phenotype, clinical procedure and outcomes. I participated in the data collection by downloading some of the datasets and searching literature for additional cohorts beyond the initial search during revision of the manuscript. I performed the metadata collection and harmonization. I implemented and ran the strain sharing pipeline and performed checks for mislabeling of certain samples. I performed the downstream analysis, and created panels in Figures 1b-e, Figure 2 and Figure 3 and the layout of Figures 1, 2, 3. I participated in results interpretation and manuscript writing together with the other co-first and senior authors.

Reference

Ianiro, G*., Punčochář, M*., Karcher, N*. *et al.* Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. *Nat Med* **28**, 1913–1923 (2022). Open access https://doi.org/10.1038/s41591-022-01964-3 [75]

Inserted manuscript

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

Gianluca Ianiro*, Michal Punčochář*, Nicolai Karcher*, Serena Porcari, Federica Armanini, Francesco Asnicar, Francesco Beghini, Aitor Blanco-Míguez, Fabio Cumbo, Paolo Manghi, Federica Pinto, Luca Masucci, Gianluca Quaranta, Silvia De Giorgi, Giusi Desirè Sciumè, Stefano Bibbò, Federica Del Chierico, Lorenza Putignani, Maurizio Sanguinetti, Antonio Gasbarrini, Mireia Valles-Colomer, Giovanni Cammarota, Nicola Segata

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, which prevents informed applications of this therapeutic approach. Here, we performed an integrated shotgun metagenomic systematic meta-analysis of new and publicly available stool microbiomes

collected from 226 triads of donors, pre-FMT recipients and post-FMT recipients across eight different disease types. By leveraging improved metagenomic strain-profiling to infer strain sharing, we found that recipients with higher donor strain engraftment were more likely to experience clinical success after FMT ($P = 0.017$) when evaluated across studies. Considering all cohorts, increased engraftment was noted in individuals receiving FMT from multiple routes (for example, both via capsules and colonoscopy during the same treatment) as well as in antibiotic-treated recipients with infectious diseases compared with antibiotic-naïve patients with noncommunicable diseases. Bacteroidetes and Actinobacteria species (including *Bifidobacteria*) displayed higher engraftment than Firmicutes except for six under-characterized Firmicutes species. Cross-dataset machine learning predicted the presence or absence of species in the post-FMT recipient at 0.77 average AUROC in leave-one-dataset-out evaluation, and highlighted the relevance of microbial abundance, prevalence and taxonomy to infer post-FMT species presence. By exploring the dynamics of microbiome engraftment after FMT and their association with clinical variables, our study uncovered species-specific engraftment patterns and presented machine learning models able to predict donors that might optimize post-FMT specific microbiome characteristics for disease-targeted FMT protocols.

MAIN

Fecal microbiota transplantation (FMT) is the medical procedure of transferring human fecal matter from a healthy donor to a recipient to treat a disease related to microbiome imbalance. FMT has shown nearly 90% success rate for the treatment of recurrent *Clostridioides difficile* infection (rCDI)[1,2], for which it is approved in clinical practice[3]. FMT has been explored more recently for other diseases associated with microbiome alterations[4,5,6] or to support other therapies[7,8,9], but its efficacy is usually lower and less consistent over cohorts than for rCDI[10,11]. Some factors that may explain this variability include the presence or abundance of single bacteria and the diversity of the patient microbiome at baseline[5,6], clinical characteristics of the disease[12], the composition of the donor's gut microbiome[13], specific aspects of the FMT working protocols (for example route of delivery, amount of infused feces)[14] and differential engraftment among species[5,6]. Yet, it is generally unknown how strain engraftment might be linked with clinical remission after FMT.

The mechanisms and dynamics dictating which donor microbes can engraft in the recipient are poorly understood. Initial studies able to track the transmission of donor strains to the recipient have been performed on very few donor–recipient pairs[15]. Availability of larger FMT trials and the advances in strain-resolved metagenomics enabled deeper analyses that started unraveling the engraftment efficiency of FMT across diseases and led to the development of statistical models to predict the post-FMT microbiome composition[16]. Such investigations remained confined to single cohorts[16,17,18,19,20,21], with unanswered questions about cross-cohort and cross-condition generalizability. As deeper strain-level metagenomics is possible[22,23,24] and not limited to well characterized microbial taxa[25], and as more metagenomic datasets are becoming available[7,8,9,15,16,17,18,26,27,28,29,30,31,32,33,34,35,] an integrative metagenomic analysis may allow uncovering general patterns of microbial engraftment and connected clinical outcomes.

Here, we present a systematic meta-analysis of 24 studies that investigated FMT in different clinical settings for which we employed new strain-resolved metagenomic approaches to unravel the dynamics of FMT engraftment and its links with clinical outcomes.

Results

A meta-analysis of public and new FMT metagenomic datasets

We retrieved all FMT studies with publicly available data that assessed microbiome composition of donors and recipients (pre- and post-FMT) through shotgun metagenomics (Methods). This search yielded a total of 21 datasets (Table 1 and Supplementary Table 1)[7,8,9,15,16,17,18,26,27,28,29,30,31,32,33,34,35]. In each study, we removed samples that were not sequenced at sufficient depth (<1 Gbp) or with evidence of mislabeling (Methods). The retained metagenomes belong to 203 FMT procedures for which at least one sample is available from each member of the 'FMT triad': the pre-FMT recipient, the post-FMT recipient and the corresponding donor. When multiple post-FMT samples were available, we selected the post-FMT sample collected closest to 1 month after FMT, as 30 days was the value that minimizes the overall time deviation (Supplementary Fig. 1; Methods).

Table 1: Summary and main characteristics of the FMT datasets included in this meta-analysis

Disease	No. of datasets (new datasets)	No. of recipients (new recipients)	No. of samples (new samples)	Median no. of post-FMT samples [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
Multidrug-resistant bacteria colonization	3 (1)	21 (5)	109 (13)	1.0 [2.0]	Infectious	Italy, Israel, France, the 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	the 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)**	**1,371 (116)**	**2 [2]**		

Numbers in parenthesis refer to data collected specifically for the present study.

We additionally sequenced 116 stool samples (23 FMT triads) from three cohorts of patients with rCDI, inflammatory bowel disease (IBD) and clinically relevant colonization by multidrug-resistant bacteria (MDRB) (Table 1 and Supplementary Table 2; Methods) enrolled in prospective case series from Italy (Fondazione Policlinico Gemelli IRCCS and Bambino Gesù Children's Hospital; Methods) and sequenced at a higher read depth than most existing FMT datasets (Supplementary Figs. 2 and 3).

In total, 1,371 samples and 226 FMT triads from 24 different cohorts (Table 1) were included in the analysis covering nine clinical conditions, including rCDI (n = 9), IBD (n = 5), MDRB (n = 3), melanoma (n = 2) and metabolic syndrome (n = 2), and single cohorts of irritable bowel syndrome (IBS), Tourette syndrome and diarrhea induced by tyrosine kinase inhibitors[7,8,9,15,16,17,18,26,27,28,29,30,31,32,33,34,35]. Studies enrolled adult participants with the exception of HouriganS_2019 (ref. [29]), ZhaoH_2020 (ref. [35]), This_study_MDRB and This_study_IBD and originated from countries with Mediterranean (France, Italy, Israel) and Northern European lifestyles (Germany, the Netherlands, Norway), in North America (USA) and China. All samples were processed following the same computational pipeline, from quality-control to analysis by strain-level profiling including yet-to-be-characterized species based on the species-level genome bins (SGBs; Methods) framework[25]. While we used all 1,371 samples (together with 4,443 samples from unrelated longitudinal datasets) to optimally delineate strain identity, we limited the analyses to one post-FMT sample per FMT triad (559 samples).

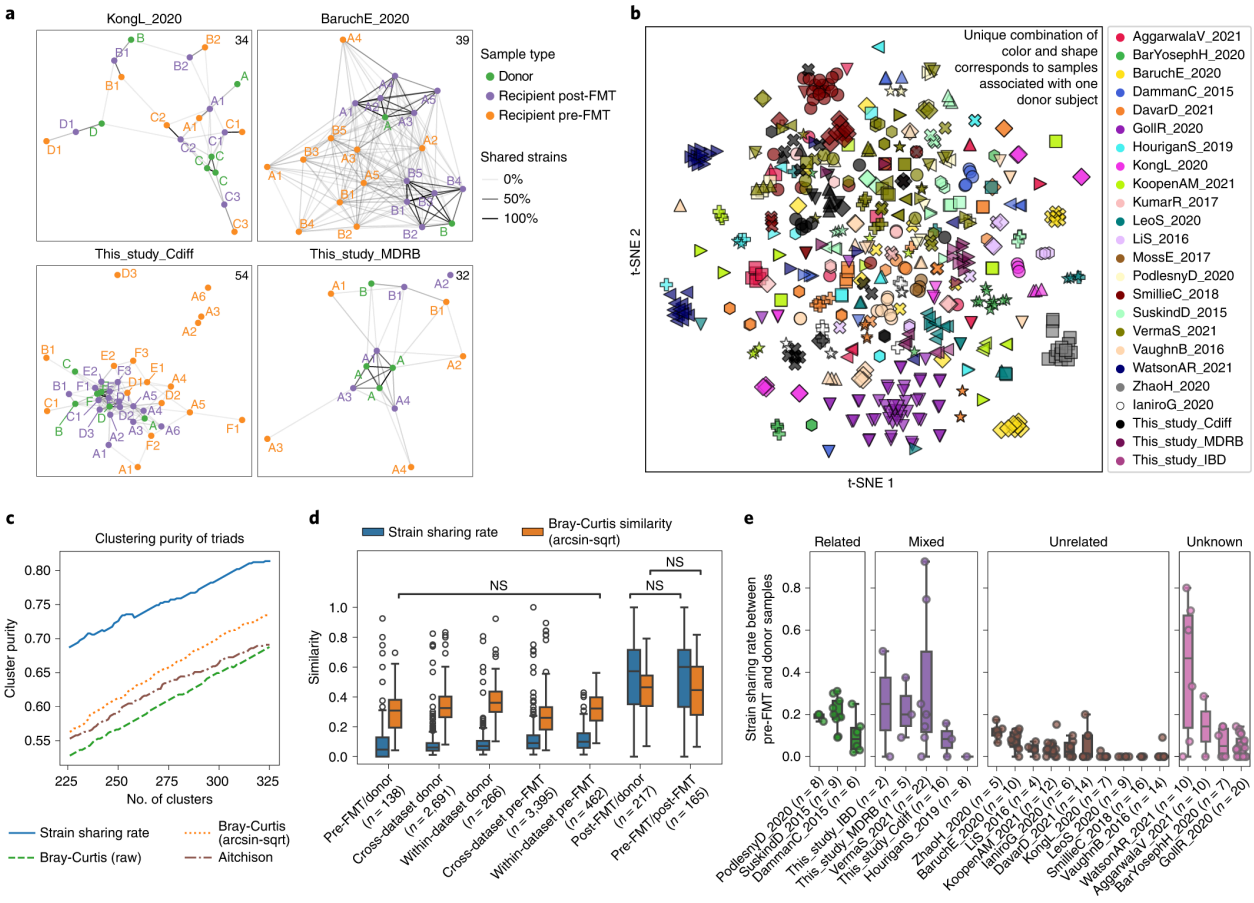
Strain-level metagenomics can assess microbial engraftment

To identify the transfer and engraftment of the donor microbiome in the recipient, we exploited the observation that microbial strains are generally specific to individuals and rarely found shared between unrelated individuals[22,23,36]. We adopted an operational species-specific definition of 'strain'[37,38] by comparing phylogenetic distance distributions of microbial genetic profiles of a given species sampled from the same individual over multiple timepoints with those distributions obtained comparing profiles from unrelated individuals (Supplementary Table 3; Methods). We implemented the approach and the species-specific cut-offs that define strain identity within StrainPhlAn 4 (ref. [39]), which we empowered with a custom database of marker

gene sequences from around 729,000 microbial genomes and metagenome-assembled genomes (MAGs). With such references, StrainPhlAn is able to detect and model strains belonging to a total of 4,992 yet-to-be-characterized species[25]; that is, unknown SGBs (uSGBs; Methods).

The StrainPhlAn-based pipeline allowed generating a map of between-samples strain sharing events that we recapitulated in undirected networks based on the number of common strains between samples (Fig. 1a and Extended Data Fig. 1; Methods). These networks confirmed that samples from the same FMT triad tend to share many more strains than unrelated individuals, whereas they are connected only weakly to samples of other FMT triads in the same cohort (Fig. 1a, PERMANOVA by FMT triad and dataset on strain sharing-based dissimilarity metric, R^2 = [0.05–0.61] and Q < 0.1 in 14 of the 24 datasets; Supplementary Table 4).

Fig. 1: Overview of microbial strain sharing in FMT studies.



a, Strain-sharing networks of the two new FMT cohorts with *C. difficile* and MDRB colonization and of two published FMT cohorts[9,30]. Nodes represent samples and are colored by role in FMT triads. The letters correspond to the donor subject and letter/number combinations indicate both associated donors (the letter) and FMT instance membership (the number) of pre-/post-FMT samples. Edges report strain sharing (minimum 2) and their opacity is scaled to the maximum number of shared strains in each dataset (indicated in the top right corner). Extended Data Fig. 1 reports the networks of all 24 datasets. **b**, Ordination of samples from all cohorts based on strain sharing rates (t-SNE with perplexity = 20). See Extended Data Fig. 3 for a PCoA ordination. **c**, Strain-sharing enabled more precise reconstructions of

the true FMT triads compared with species-level β-diversities (Extended Data Fig. 2 and Supplementary Table 5). We compare the K-medoids clustering purity of FMT triads 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 with Bray–Curtis similarity. Significance was assessed by Mann–Whitney U-tests and the two-tailed P values were FDR-adjusted using the BH method. All pairwise tests are significant except for those labeled NS. All P and Q values are reported in Supplementary Table 6. **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; Methods). Boxplots report the median and upper/lower quartiles, whiskers are at 1.5 times higher/lower of the upper/lower quartiles.

To account for different numbers of strains that can be analyzed over samples, we defined the strain-sharing rate metric as the number of strains found identical in two samples divided by the number of species with available strain profiles that are present in both samples (Methods). K-medoids clustering on strain-sharing rates yielded clusters of higher purity with respect to FMT triad membership than β-diversity measures (Fig. 1c, Extended Data Fig. 2 and Supplementary Table 5; Methods) and a t-distributed stochastic neighbor embedding (t-SNE) projection also separates samples by FMT triad membership (Fig. 1b and Extended Data Fig. 3). Strain-level metagenomics can thus accurately describe strain sharing events within FMT triads.

Donor–recipient relationship influences post-FMT engraftment

Strain-sharing rates were much higher between post-FMT and donor samples (median 57%), and between pre-FMT and post-FMT samples (60%) than between donors and pre-FMT recipients (4.8%). The substantial increase in donor–recipient strain sharing after FMT is also significantly stronger than the decrease in β-diversity (Wilcoxon signed-rank test, $P = 7 \times 10^{-23}$; Fig. 1d and Supplementary Table 6; Methods), confirming that the strain identity-based profiling approach better captures the microbiome remodeling induced by FMT compared with species-level β-diversity measures.

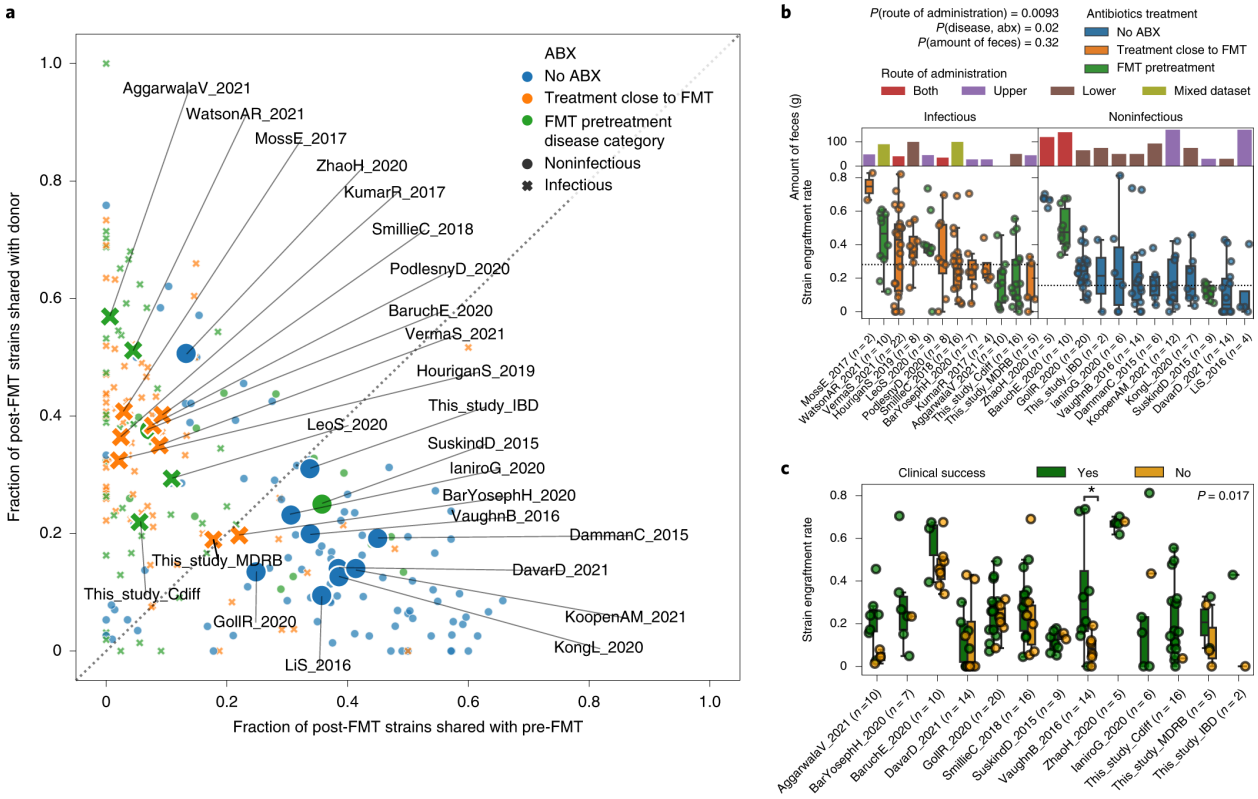
Overall, 58.4% of post-FMT samples shared more strains with corresponding donor samples than with their pre-FMT. However, the difference in shared strains between donor/post-FMT samples and pre-FMT/post-FMT samples differed substantially across FMT triads (median = −3; range = −96–75; Extended Data Fig. 1 and Supplementary Fig. 4). We also found that pre-FMT recipients shared more strains with related (usually cohabitating) donors than with unrelated donors (that is, donors in the original studies that were specified as unrelated, or recruited through public advertisement or hospital cohorts (Fig. 1e), related versus unrelated, permutation test $P < 1 \times 10^{-4}$, median strain sharing rate difference = 0.18). This also holds in datasets in which only a subset of the donors and recipients were related (Fig. 1e; $P < 1 \times 10^{-4}$). We accounted for these potential baseline strain sharing biases by subtracting them from post-FMT engraftment rates, resulting in significantly lower estimates (Wilcoxon signed-rank test, $p = 1 \times 10^{-9}$; Supplementary Fig. 5 and Extended Data Fig. 4; Methods). Together, these data

confirm that the extent of donor microbiome engraftment is variable and influenced by pre-FMT donor–recipient relatedness.

Combined FMT administration associates with strain engraftment

To assess the main determinants of post-FMT strain engraftment, we first performed a multivariate analysis including clinical variables that could potentially influence engraftment (infectious/noninfectious disease, antibiotics treatment), recipient and donor microbiome characteristics (α-diversity, species-level dissimilarity and strain sharing rate at baseline, recipient age and geographical region) and other procedural features that were consistently available across datasets (administration of fresh/frozen stool, amount of feces administered, route of administration and bead-beating steps in the DNA extraction protocol; Methods). By fitting a partial least squares (PLS) regression model (Methods), we found that only the first two components were significantly associated with engraftment, explaining 18.7% ($Q = 6 \times 10^{-10}$) and 4.6% ($Q = 3.8 \times 10^{-3}$) of the variation (Extended Data Fig. 5). Only FMT administration through a mixed route combining upper gastrointestinal tract administration (by capsules, enteroscopy, nasogastric tube, nasoduodenal tube or upper endoscopy) and lower gastrointestinal tract administration (by colonoscopy) was significantly associated with the first PLS component ($P = 0.016$). Indeed, route of delivery emerged as the variable most significantly associated with strain engraftment also in univariate testing ($P = 0.0093$; Fig. 2b). So far, no consensus exists as to a recommended route of administration in FMT protocols[40] and, whereas our results suggest that combined routes increase the engraftment likelihood, the observation is based on only four studies adopting this approach. Importantly, intake of antibiotics (14 studies with antibiotic intake before FMT, 10 without) and disease category (12 studies on infectious diseases, 12 on noninfectious) were significantly associated with strain engraftment in cohorts that employed a single administration route (n = 19 datasets, permutation test antibiotics treatment and infectious disease versus no antibiotics and noninfectious disease, $P = 0.027$), and both were associated with the first two PLS components while being highly correlated with each other (Supplementary Fig. 6; Methods).

Fig. 2: Variability of strain engraftment and retention across disease, antibiotics 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 antibiotics administration and disease category. Small points represent individual FMT triads and large labeled marks represent per dataset averages. **b**, Variability of 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 (two-tailed permutation test route of administration mixed versus lower or upper $P = 0.0093$, antibiotics and infectious disease versus no antibiotics and noninfectious disease in datasets employing single route of administration $P = 0.02$, amount of feces $P = 0.32$). **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 one 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 (two-tailed $P = 0.017$), that was significant in only 1 of the 13 datasets when considered individually (VaughnB_2016 Mann–Whitney U-test with two-tailed $P = 0.039$). Boxplots plots report the median and upper/lower quartiles, whiskers are at 1.5 times higher/lower of the upper/lower quartiles.

FMT engraftment is linked to antibiotics and infectious diseases

We examined the extent of donor strain engraftment over strain retention in FMT recipients by comparing the fraction of donor strains detectable in the post-FMT sample (fraction of donor strains) with 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 pretreatment before FMT—had a significantly

higher fraction of donor strains compared with the fraction of retained strains, as was previously reported in the context of ulcerative colitis[41] (Wilcoxon signed-rank test, $P = 2 \times 10^{-16)}$, while the opposite was true for recipients who did not receive antibiotics (Wilcoxon signed-rank test, $P = 1 \times 10^{-5}$, Fig. 2a). Antibiotic treatment thus seems to lead to enhanced donor strain engraftment and decreased strain retention in the FMT recipient, possibly by reducing colonization resistance in the recipient[42]. Recipients with infectious diseases also had comparatively higher fractions of donor strains compared with the fraction of retained strains (Wilcoxon signed-rank test, $P = 8 \times 10^{-16}$), while the opposite was true in patients with noninfectious diseases (Wilcoxon signed-rank test, $P = 6 \times 10^{-4}$).

Patients with recurrent or resistant infectious diseases often have a long history of repeated antibiotic courses and are pretreated with specific antibiotics before FMT, while only two of the noninfectious disease cohorts (SuskindD_2015 and BaruchE_2020) underwent treatment with antibiotics before FMT. The SuskindD_2015 cohort of patients with Crohn's disease received rifaximin before FMT and exhibits strain sharing patterns similar to datasets with noninfectious disorders, consistent with previous results showing that rifaximin does not lead to substantial shifts in microbiome composition[43]. On the contrary, the BaruchE_2020 melanoma cohort, in which patients were pretreated 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[44] and neomycin[45] treatment. Antibiotic use may also explain the successful engraftment observed in patients with infectious diseases treated with artificial microbiome consortia[46].

Administration of stool samples from multiple donors could also maximize the diversity of engrafted bacteria in the recipient[47]. For the only study available adopting mixed donor feces (GollR_2020), we found the second highest median strain engraftment rate among the cohorts without antibiotics treatment. We also observed an exceptionally high microbial strain sharing between donors and post-FMT recipients, comparable with datasets of infectious diseases and pre-FMT antibiotics, in the ZhaoH_2020 cohort, where FMT was given for Tourette syndrome (a noninfectious disorder) without antibiotic preconditioning. Besides using a mixed administration route, this cohort included children whose microbiome is less resistant to colonization from incoming strains.

Overall, these results show that the fractions of donor-derived and retained strains after FMT are influenced by antibiotic administration and by the presence of an infectious disease, which are both hypothesized to reduce microbiome colonization resistance. However, since antibiotic treatment and infectious diseases were closely entangled variables in our meta-cohort (Supplementary Fig. 6), it was not possible to unravel their relative contribution to strain engraftment and retention. Nonetheless, as both variables are known to lead to a decreased microbial diversity[48,49] and, given that the substantially lower microbial α-diversity is probably 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.

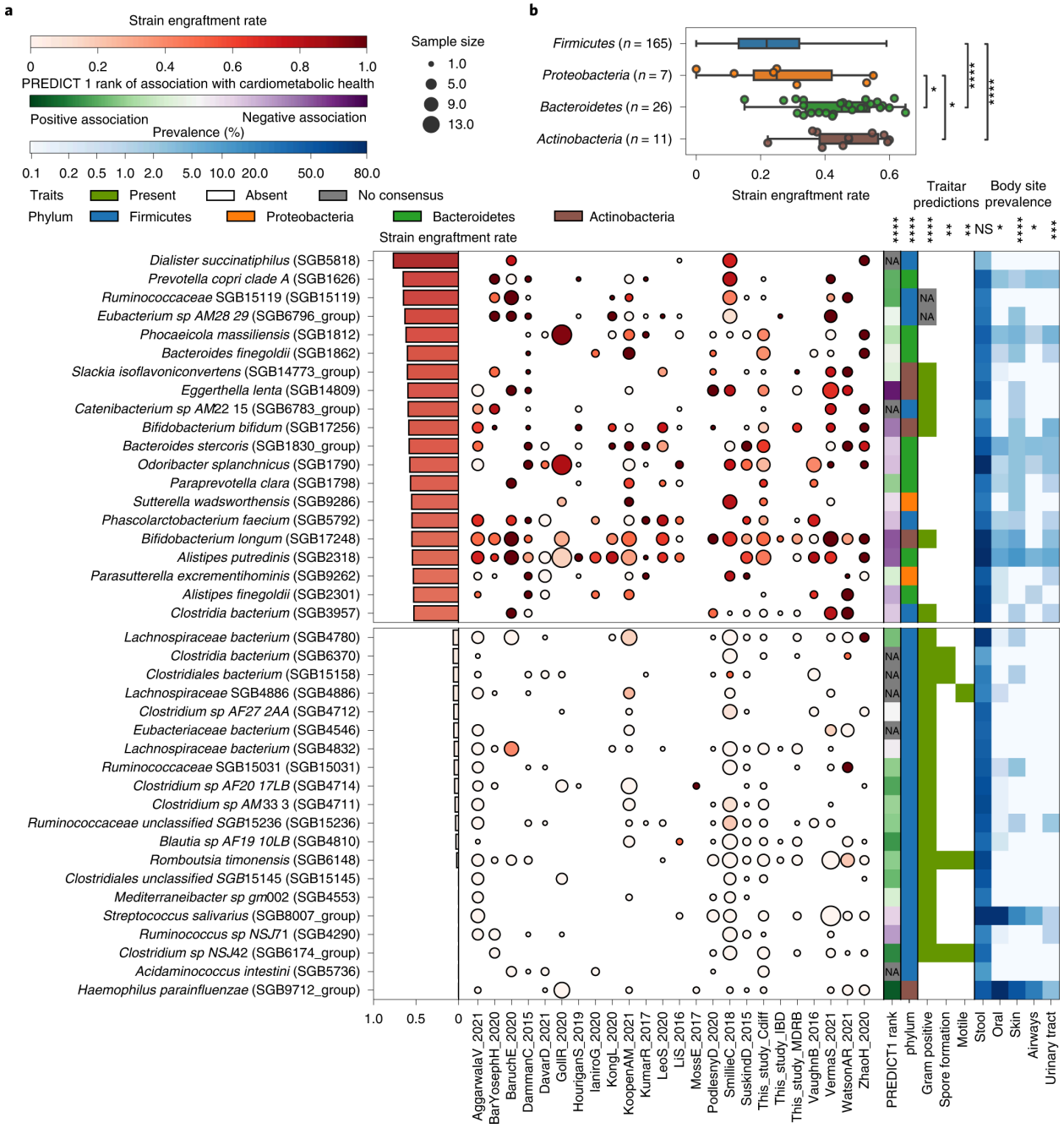
Links between strain engraftment and clinical success of FMT

Previous studies suggest that strain engraftment might be associated with clinical success of FMT, but consolidated evidence is still lacking[5,6]. We thus compared the strain engraftment rates with the clinical success of each FMT triad for the datasets with appropriate clinical data available (Supplementary Table 1; Methods). When considering single studies, we found that recipients experiencing clinical success showed significantly higher engraftment only in the VaughnB_2016 cohort (Mann–Whitney U-test, $P = 0.039$). When analyzing all cohorts together, we found an overall positive association between strain engraftment rate and clinical response to FMT (Fig. 2c) that proved significant according to a blocked permutation test and a Wilcoxon signed-rank test on medians ($P = 0.017$ and $P = 0.040$, respectively) and borderline significant using a random effects model meta-analysis ($P = 0.051$; Methods). We similarly tested for an association between the species-level similarity between post-FMT and donor samples and clinical success, which yielded a significantly positive association when evaluating species-level microbial abundances with a blocked permutation test (Bray–Curtis similarity between post-FMT and donor samples $P = 0.018$) but not with the other tests (random effects model meta-analysis $P = 0.072$, Wilcoxon signed-rank test on medians $P = 0.414$; Supplementary Fig. 7), and no significant association was found when considering overlap in species presence (Jaccard similarity between post-FMT and donor samples; Supplementary Fig. 8). The limited total sample size, the binary categorization of success of clinical treatments, and the heterogeneity of conditions tested represent limitations in our analyses, but the results overall suggest that both higher microbial engraftment and, partially, the overall convergence of microbial species abundances between recipient and donor might improve clinical success of FMT.

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

We then computed species-specific strain engraftment rates over all FMT triads for the 211 microbial species for which the strain engraftment rate could be estimated with sufficient confidence (that is, that could engraft in at least 15 FMT triads and four different datasets; Fig. 3a and Supplementary Table 7; Methods). Overall, we found significant differences in engraftment rates across bacterial phyla (Kruskal–Wallis test, $P = 3 \times 10^{-11}$), as Bacteroidetes and Actinobacteria spp. (26 and 11 species, respectively) displayed higher average strain engraftment rates ($45 \pm 12\%$ and $46 \pm 12\%$, respectively; Fig. 3b and Supplementary Table 8) compared with Firmicutes and Proteobacteria ($23 \pm 14\%$ and $29 \pm 20\%$, respectively; post hoc Dunn tests, $Q < 0.1$; Fig. 3b).

Fig. 3: Bacterial strain engraftment rates across datasets and their associations with phenotypic properties, cardiometabolic health and prevalence.



a, Overall and within-dataset strain engraftment rates and associations of species with predicted phenotypic properties[62], cardiometabolic health[53] 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 with continuous variables were tested with Spearman's rank correlation tests, while those with binary categorical variables were tested with the 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 10). Significance levels (NS, nonsignificant, *$Q < 0.05$, **$Q < 0.01$, ***$Q < 0.001$, ****$Q < 1 \times 10^{-4}$) 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 measure (Methods). **b**, Strain engraftment rates are significantly associated with bacterial phyla (Kruskal–Wallis test, $P = 3 \times 10^{-11}$; post hoc Dunn tests FDR corrected using the BH method, Firmicutes versus Bacteroides $Q = 8.0 \times 10^{-9}$, Firmicutes versus Actinobacteria $Q = 3 \times 10^{-5}$, Proteobacteria versus Bacteroidetes $Q = 0.037$, Proteobacteria versus Actinobacteria $Q = 0.037$, the remaining pairs are NS, that is $Q > 0.1$). The Euryarchaeota and Verrucomicrobia phyla were omitted from the analysis as only one species in each of those phyla was assessed in our analysis. Boxplots report the median and upper/lower quartiles, whiskers are at 1.5 times higher/lower of the upper/lower quartiles. NA, not applicable.

Six Firmicutes SGBs were among the set of the 20 most-engrafting species, including two 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, *D. succinatiphilus*—the SGB with the highest likelihood to engraft (76%)—and *Phascolarctobacterium faecium* are both members of the Negativicutes class, characterized by a cell-wall composition containing lipopolysaccharides, which results in a negative Gram stain[50]. As such, these Firmicutes species may have characteristics not completely in line with those of the typical members of this phylum, possibly explaining their comparatively high engraftment rates. Among the top-engrafting non-Firmicutes species, we found several *Bacteroidales*: *Prevotella copri clade* A[51] (strain engraftment rate = 65%), *Bacteroides finegoldii* (60%), *Bacteroides stercoris* (58%), *Alistipes putredinis* (54.2%), *Alistipes finegoldii* (53%) and *Phocaeicola massiliensis* (62%). Among Actinobacteria, the dysbiosis-associated species *Eggerthella lenta*[52] (strain engraftment rate = 60%) and two *Bifidobacteria* (*B. bifidum* 58%, *B. longum* 55%) also exhibit high engraftment likelihood. In contrast, 19 out of the 20 least-engrafting species (strain engraftment rate < 6.5%) belonged to Firmicutes, of which 16 were members of the Clostridiales order. *Acidaminoccus intestini*, *Streptococcus salivarius* and four other unnamed and uncharacterized Firmicutes species were never found to detectably engraft in the FMT recipient despite being fairly prevalent in the donor (Fig. 3a and Supplementary Table 8). These data suggest that the engraftment potential of microbes differs among phyla and species, and that such engraftment likelihoods could be considered in future therapeutic protocols when selecting fecal donors or designing artificial microbial consortia to use instead of FMT.

We also assessed the potential transmission of eukaryotic microbes, and found that only *Blastocystis* was detectable at enough coverage to infer transmission (Methods). Most FMT screening procedures exclude donors with *Blastocystis*[3], so its prevalence in donors is lower than in most 'Westernized' populations[53,54,55]: we detected only five donors positive for *Blastocystis* in two cohorts (BarYosephH_2020, SmillieC_2018). No transmission could be inferred (Supplementary Table 9; Methods) while two retention events in recipients were detected based on *Blastocystis* subtyping. While *Blastocystis* is increasingly reported to be linked with favorable health conditions[53,55,56], it does not seem to play a role in FMT, possibly due

to donor screening procedures, and transmission via FMT was reported as asymptomatic elsewhere[57].

Engraftment is linked with predicted bacterial phenotypes

We assessed whether the taxonomic differences in strain engraftment (Supplementary Table 8) we detected were associated with predicted microbial phenotypic properties. The more resistant Gram-negative species had a higher engraftment likelihood (Mann–Whitney U-test $Q = 3 \times 10^{-6}$; Supplementary Table 10), and only a few Gram-positive bacteria were among the most-engrafting species (Fig. 3a). Since most Firmicutes are Gram positive, this association may be driven by characteristics of the Firmicutes phylum unrelated to cell-wall structure. Spore-forming and motile species also tended to display reduced engraftment (Mann–Whitney U-test $Q = 0.007$ and $Q = 0.008$, respectively; Fig. 3a and Supplementary Table 8). All of the above suggests that species engraftment may be facilitated by specific microbial features although more refined knowledge of phenotypic traits is needed to infer mechanistic hypotheses underlying these associations.

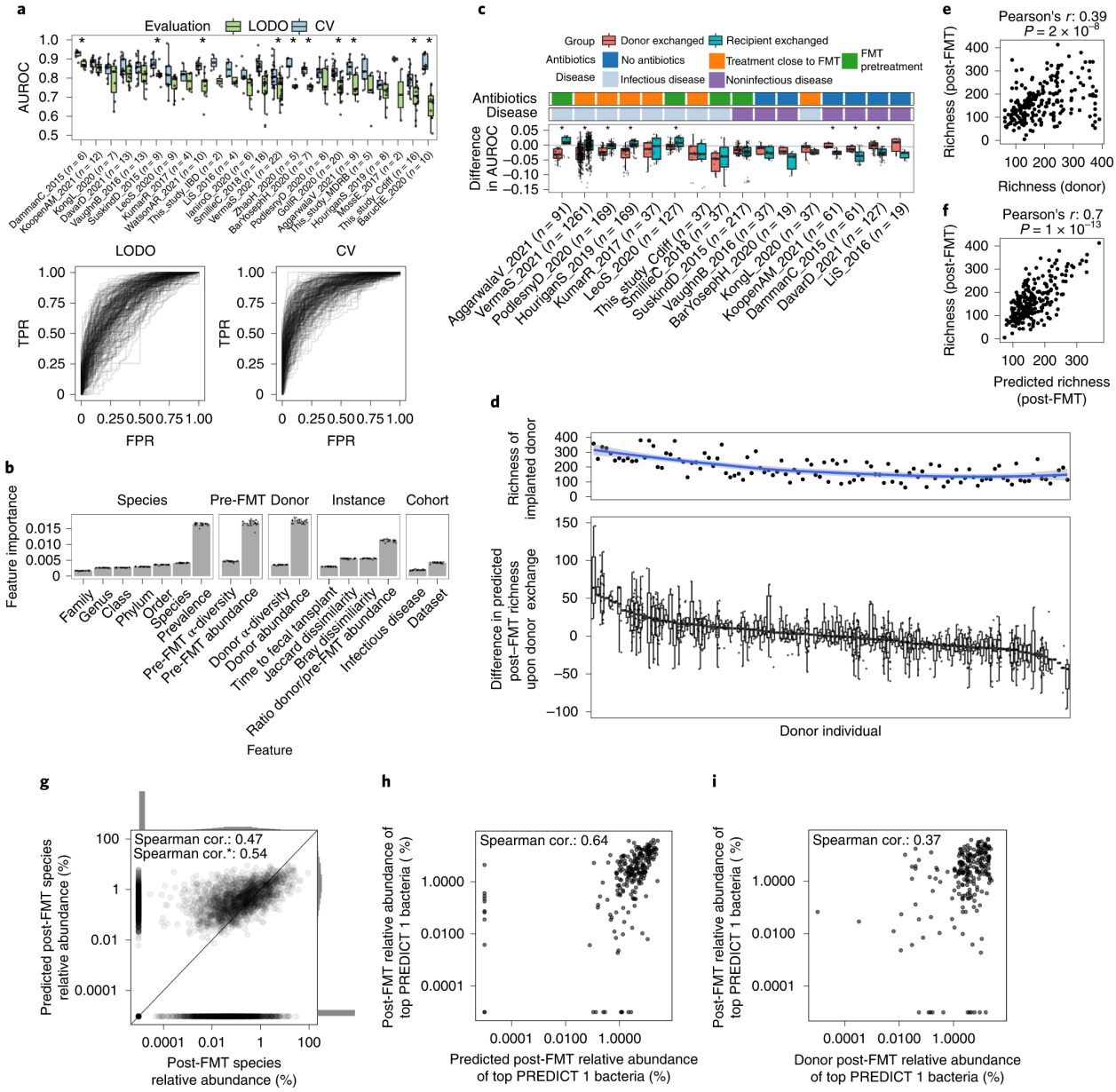
While screening for pathogens is routinely performed as part of FMT protocols, the ability of noninfectious but disease-associated microbes to engraft remains unknown. Interestingly, microbes negatively associated with cardiometabolic health in the PREDICT 1 study[53] tended to engraft more frequently (Spearman's $\rho = 0.36$, $P = 4 \times 10^{-7}$), possibly due to more aggressive host colonization strategies or higher adaptive potential to dysbiotic or inflamed gut environments such as those found in FMT recipients. Although species prevalence in the gut of healthy individuals did not significantly correlate with engraftment across 9,120 gut metagenomic samples from 56 public studies (Supplementary Table 11; Spearman correlation, $P > 0.05$), the prevalence of bacteria in nonintestinal human body sites was associated with higher engraftment (Mann–Whitney U-test, $P = 8 \times 10^{-4}$). This suggests that ability to engraft is linked to the microbes' capability of surviving in diverse environments. Finally, we found no association between the engraftment of individual species and clinical success (Fisher's exact test, $Q > 0.1$; Supplementary Table 8). Together, these results show a remarkable variability in the 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 while promoting those with positive health associations.

Machine learning can predict post-FMT microbial composition

Understanding what are the donor and pre-FMT microbiome factors dictating the post-FMT microbiome configuration could facilitate precision-medicine approaches for targeted microbiome modulations. Since donor strain engraftment accounts only partially for the post-FMT microbiome composition, as strains can also persist or be acquired from the environment, we developed machine learning (ML) 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, microbial abundances and α-diversity in pre-FMT and donor samples and microbial prevalence in unrelated cohorts (Methods). We found that these models predict post-FMT species composition with an area under the receiver

operating characteristic curve (AUROC) ranging from 0.77 to 0.91 (average = 0.85, s.d. = 0.03; Extended Data Fig. 6 and Supplementary Table 12) in a fivefold cross-validation (CV) setting (Fig. 4a; Methods).

Fig. 4: RF models predict post-FMT microbiome composition and the effect of different donors on the post-FMT microbiome.



a, RF predictions of the presence or absence of species post-FMT. LODO and CV AUROC are reported and represented as true positive rates (TPR) versus false positive rates (FPR). **b**, The relative importance of microbial features in the LODO model (n = 24 for each bar). Data are presented as mean, error bars correspond to s.d. **c**, Distribution of the changes in AUROC values for the LODO models of a upon donor exchange (Methods). **d**, Top panel, species richness of FMT donors. The blue line is a locally estimated scatterplot smoothing fit, the

shaded area corresponds to the 95% confidence interval. Bottom panel, difference in post-FMT species richness upon donor exchange with respect to the predicted post-FMT species richness of the real triad n(total) = 1,317. **e**, Donor species richness is positively correlated with recipient's post-FMT species richness (Pearson's correlation test, $r = 0.39$, $P = 2 \times 10^{-8}$). **f**, Predicted post-FMT species richness is strongly correlated with the actual post-FMT richness (Pearson's correlation test, $r = 0.7$, $P = 1 \times 10^{-13}$). **g**, An RF regression model is able to predict bacterial abundances in the post-FMT microbiome. The asterisk designates the Spearman correlation (cor.) when omitting truly absent species predicted to be absent. Individual datasets are reported in Supplementary Fig. 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, Donor abundance is a worse predictor of the cumulative abundance of the top 20% PREDICT 1 bacteria than the RF regression model. Boxplots report the median and upper/lower quartiles, whiskers are at 1.5 times higher/lower of the upper/lower quartiles.

We next 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 and accuracy values were expectedly lower than in the CV setting, AUROC values were above 0.7 in all but 3 of the 24 cohorts (average = 0.77, s.d. = 0.05); Extended Data Fig. 7, Supplementary Fig. 9 and Supplementary Table 12). Finally, we evaluated RF regression models to predict the post-FMT abundance of bacterial species (Methods). These models provided estimates of the abundance of species in the post-FMT microbiome that were significantly correlated with those assessed by microbiome sequencing of the post-FMT samples (Spearman correlation 0.47, $P < 1 \times 10^{-16}$; Fig. 4g and Supplementary Fig. 10). We thus conclude that, whereas the prediction potential of the post-FMT microbiome composition is partially dependent on the cohort, substantial prediction ability is maintained across datasets.

Analysis of the importance of each feature 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 α-diversity of donor and recipient microbiomes, the β-diversity between donor/recipient pairs, or disease context (Fig. 4b). Single taxonomic features (that is, the species or genus labels) proved not particularly important despite differences in strain engraftment rates over different clades (Fig. 3a). This observation was, however, likely due to the effect of information redundancy and hierarchy on the importance estimates as, when we considered all taxonomic levels together, the importance of the taxonomy was comparable with that of bacterial prevalence or abundance (Supplementary Fig. 11). Further evaluation of species-wise strain engraftment rates as well as predicted microbial phenotypes (Fig. 3a and Supplementary Table 8) showed no relevant additional increase in prediction ability (mean change in CV AUROC upon addition of strain engraftment rate = –0.007, s.d. = 0.015; Supplementary Fig. 12; mean change in CV AUROC upon addition of predicted phenotypes = 0.005, s.d. = 0.019; Supplementary Fig. 13). 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.

ML models can pinpoint suitable FMT donors

To better understand to what 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 selected 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 (Fig. 4c). The performance decrease upon donor exchange was particularly pronounced in cohorts of infectious diseases and in patients pretreated with antibiotics (Mann–Whitney U-tests, $P < 0.001$; Extended Data Fig. 8), consistent with a higher fraction of donor strains engrafting in the recipient in these conditions (Fig. 2a). The choice of donor thus has a higher influence on the post-FMT microbiome in patients with infectious disease and/or those that were treated with antibiotics.

Finally, we investigated whether ML models can pinpoint particularly suitable donor individuals for improving microbiome features in recipients based on their individual microbiomes. We first evaluated the donor effect in modulating post-FMT species richness—a microbiome feature linked with community stability and resilience[58] and with clinical success in the context of ulcerative colitis[5]. Upon exchange of donors in triads, we found that some donors led to a consistent increase in predicted post-FMT richness compared with the original donor, whereas others led to a decreased predicted post-FMT richness (Fig. 4d). We also found that the donors with higher richness were predicted to induce higher richness in the recipient post-FMT (Fig. 4d and Supplementary Fig. 14), and such predictions of post-FMT richness using the real donor were much more accurate than the donor's richness alone (Pearson's $r = 0.7$ versus $r = 0.39$, $P = 1 \times 10^{-13}$ versus $P = 2 \times 10^{-8}$; Fig. 4e,f).

We then exploited this framework to pinpoint donors that are predicted to maximize the probability of the presence of other predefined groups of microbes in the post-FMT samples, such as Firmicutes, species found in the oral cavity (Supplementary Table 13), or the set of species found positively linked with cardiometabolic health in the PREDICT 1 study[53] (Extended Data Fig. 9 and Supplementary Table 14). In all these situations, our models proved more accurate in predicting a given trait than using the quantitative microbial features of the donor as a direct estimator. We finally evaluated a regression model to predict the cumulative relative abundance of the same microbial groups, finding that the model can predict the cumulative abundance of microbes positively linked with cardiometabolic health better than the donor abundances alone (Fig. 4h,i), although the results are variable across different clades (Extended Data Fig. 10). Taken together, these results illustrate that our ML framework provides predictive models of the composition of the post-FMT microbial communities that might be useful for choosing a suitable donor given a specific post-FMT microbiome feature of clinical relevance, such as post-FMT microbiome richness.

DISCUSSION

In our meta-analysis of 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 such variability was explained best by mixed FMT administration routes (combining upper and lower gastrointestinal (GI) tract), by the administration in the recipient of antibiotics before FMT (therapeutically or as preconditioning), and by the recipient being affected by infectious diseases. These findings could explain the discrepancies in the effectiveness of FMT between rCDI and chronic or noninfectious disorders[4,5,6]. Our results provide further support for administering FMT by combined routes and including antibiotic preconditioning in FMT working protocols to increase donor microbiome engraftment, even though the potential side effects of antibiotic treatments for noninfectious diseases[59] should be considered.

We found differential strain engraftment likelihoods associated with microbial taxonomy and phenotypic properties. Some species with immune modulation potential (for example, *Bifidobacteria* spp.), Gram-negative bacteria and some species with proinflammatory potential (for example, *Eggerthella lenta*) were more likely to engraft than most Firmicutes, including putative butyrate-producing bacteria. As FMT is performed in patients and not on healthy volunteers, it remains to be elucidated whether the general higher engraftment rates of proinflammatory microbes reflect intrinsic phenotypic traits that favor transmission and colonization in a new environment or rather a better fitness for an inflamed and dysbiotic environment. Additionally, the implementation of targeted, fine-tuned bacterial consortia as an alternative to traditional FMT would avoid the transfer of potentially detrimental bacteria (including pathogens that could remain undetected upon screening[60]), but it is still unclear whether such consortia can represent a suitable alternative to the complexity of FMT[61].

Finally, we developed an ML model to 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 with a previous, single-cohort study[16]. 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 favorable 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 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.

In our analysis, we integrated all available metagenomic datasets of FMT in clinical settings, but the small sample size of single studies as well as the heterogeneity of diseases and clinical protocols still prevent more clear-cut identification of predictors of post-FMT microbiome engraftment. Moreover, the link we observed between engraftment and clinical success of the FMT treatment needs to be substantiated in appropriately sized studies with higher number of patients in both outcome arms (for example, clinical failures for rCDI are relatively rare) and with more fine-grained evaluation of clinical success. Dedicated studies and randomized controlled

trials are also needed to clarify the influence of protocol-related variables, such as antibiotic preconditioning or combined routes of delivery, on strain engraftment. Estimates of engraftment rates can also be refined both by sequencing samples at higher depth and by developing computational methods able to profile multiple strains from the same species co-colonizing an individual and by better accounting for nonbacterial members of the microbiome. These further improvements and investigations are needed to effectively translate the metagenomic support to FMT protocols into clinical practice.

METHODS

Metagenomic dataset search strategy and selection

We systematically searched PubMed, Scopus and ISI Web of Knowledge as of 8 February 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 nonautologous FMT; (2) shotgun metagenomic analysis of donor feces and of recipient feces (before and after treatment) was performed. We excluded studies in which the only therapeutic treatment for the disease was based on antibiotics. We further excluded those studies using microbial consortium-based transplantation approaches (instead of donor stool-based transplantations), those in which fewer than three 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 patients treated with nonautologous FMT. If studies used stool 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 nonoriginal studies (reviews, meta-analyses, editorials, and so on). The eligibility of each study was assessed independently by two reviewers (N.K. and S.P.), and any disagreements were resolved by the opinion of a third reviewer (G.I.).

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 them through private correspondence.

Metadata extraction and curation

Metadata extraction was carried out independently by two reviewers (N.K. and S.P.), using a data collection form. Discrepancies between the two reviewers were resolved by the opinion of a third investigator (G.I.). The following data were extracted from each study if available: author names, publication year, Bioproject Accession code, sequencing depth, study location, number of total samples, study disease, number of recipients and donors, donor type (that is, 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. Data were not analyzed by sex or gender due to lack of this information in most of the published datasets.

Newly collected metagenomic datasets

Three Italian cohorts were newly collected as case series and sequenced in the context of this study. A first cohort (This_study_Cdiff) was collected between February 2021 and August 2021 at the Fondazione Policlinico Gemelli IRCCS in Rome, Italy, and included 16 adult subjects with recurrent *C. difficile* infection and no history of other GI disorders or GI surgery. Patients were treated with a single fecal transplant from six different donors, and their stool was collected just before FMT and at different timepoints (7, 15, 30, 60, 180 and 240 days) after FMT. FMT was performed with frozen fecal material. Donor selection and manipulation of fecal material were performed following international guidelines3. All patients underwent FMT by colonoscopy, after bowel lavage and a 3-day vancomycin regimen, as previously described1. A total of 94 stool samples were sequenced. A second cohort (This_study_IBD) was collected from May 2017 to October 2017 at the Ospedale Bambino Gesù IRCCS in Rome, Italy, and included two pediatric patients with mild-to-moderately active IBD despite traditional treatments, without any active GI infection, placed central venous catheter or critical illness or comorbidity. They received a single FMT (one patient from a related donor, the other from an unrelated donor). Stool samples were collected and sequenced at follow-up visits up to 30 days after treatment, yielding eight metagenomic samples. A third cohort (This_study_MDRB), from the Ospedale Pediatrico Bambino Gesù IRCCS in Rome, Italy, included, between October 2018 and March 2019, five pediatric patients with large bowel colonization with 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[63]. Written informed consent was obtained from all participants (or the parents of pediatric participants). No compensation was provided to the participants. Consistent metadata of all 115 samples newly collected in this study can be found in Supplementary Table 2.

Samples were collected using a stool collector with a DNA stabilization buffer, brought directly by patients to the FMT centers in a refrigerated box within 6 h from collection, and then stored at –80 °C for up to 36 months before being shipped in dry ice to the CIBIO Department (Trento, Italy) for DNA extraction and sequencing. DNA extraction was performed using the DNeasy PowerSoil Pro Kit (Qiagen) according to the manufacturer's procedures. No human DNA sequence depletion or enrichment of microbial or viral DNA was performed. DNA concentration was measured with Qubit (Thermo Fisher Scientific) and DNA was then stored at –20 °C. Sequencing libraries were prepared using the Illumina DNA Prep (M) Tagmentation kit (Illumina) following the manufacturer's guidelines. Sequencing was performed on the Illumina NovaSeq 6000 platform at a target sequencing depth of 7.5 Gbp following the manufacturer's protocols.

Newly generated shotgun metagenomic sequences were preprocessed and quality controlled using the pipeline available at https://github.com/SegataLab/preprocessing and KneadData within bioBakery v.3 (ref. [23]). Shortly, reads were quality controlled and those of low quality (average quality score <Q20), fragmented (<75 bp) and with more than two ambiguous nucleotides were removed with Trim Galore (v.0.6.6). Contaminant and host DNA was identified with Bowtie2 (v.2.3.4.3)[64] 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.

Definition of clinical response across studies

To evaluate the association between microbial engraftment and clinical success, we identified all studies that expressed clinical outcomes as binary variables, for which single individual metadata were available or could be retrieved from the publication via manual curation, and for which both the clinically successful and the unsuccessful groups had at least one FMT triad. Ten published studies (AggarwalaV_2021, BarYoseph_2020, BaruchE_2020, DavarD_2021, GollR_2020, SmillieC_2018, SuskindD_2015, VaughnB_2016, ZhaoH_2020, IaniroG_2020) and the three new cohorts (This_Study_Cdiff, This_Study_IBD, This_Study_MDRB) were included. Clinical success was defined as C. difficile infection cure in three studies (AggarwalaV_2021, SmillieC_2018, This_Study_Cdiff), as eradication of MDRB in two studies (BarYoseph_2020, This_Study_MDRB), as objective tumor regression by imaging according to iRECIST criteria[65] in two studies (BaruchE_2020, DavarD_2021), as reduction by more than 75 points in the IBS-Severity Scoring System (IBS-SSS) in GollR_2020, as resolution of diarrhea in IaniroG_2020, as reduction by >25% in the Yale Global Tic Severity Scale (YGTSS-TTS) and reduction by more than three in the Harvey-Bradshaw Index (HBI) change without an increase in IBD-related medications in VaughnB_2016, as clinical remission expressed as Pediatric Crohn's Disease Activity Index (PCDAI) of less than ten in SuskindD_2015, and as clinical remission expressed as Pediatric Ulcerative Colitis Activity Index (PUCAI) of less than ten in This_Study_IBD.

Building the expanded SGB database

SGBs are clusters of microbial genomes and MAGs defined to have no more than 5% pairwise genetic divergence[25]. SGBs can contain taxonomically labeled microbial genomes from isolate sequencing (kSGBs) or can lack taxonomic contextualization from isolate sequencing (uSGBs; that is, SGBs with no cultured isolate). 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 reference isolate genomes from Pasolli et al.[25] and further expanded using the same approach with 616,805 MAGs from different human body sites, animal hosts and other environments, together with 155,767 reference genomes in the National Center for Biotechnology Information GenBank database[66] available as of November 2020. MAGs were assembled from metagenomes by applying metaSPAdes[67] (v.3.10.1) or MEGAHIT[68] (v.1.1.1) to each sample separately as reported in Pasolli et al.[25]. Obtained assembled contigs longer than 1,500 nucleotides were binned into MAGs with MetaBAT2 (ref. [69]) (v.2.12.1). We executed CheckM (v.1.1.4)[70] 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[71] on the quality-controlled sequences, and dereplicating sequences at 99.99% genetic identity. A total of 729,195 genomes (560,076 MAGs (Supplementary Table 15) and 169,119 reference genomes) were kept in the extended database used for species- and strain-level profiling, thus leveraging reference-based profiling with information provided by metagenome assembly. Reference isolate genomes and MAGs were then clustered into SGBs spanning at least 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.[25]. 'phylophlan_metagenomic'—a subroutine of PhyloPhlAn 3[72] that applies Mash[71] to estimate the whole-genome average nucleotide identity among genomes—was used to assign MAGs to SGBs. Reference genomes and MAGs for which no SGB with at least 5% average genetic distance was present in the database were assigned to new SGBs based on the average linkage hierarchical clustering (with the dendrogram cut at 5% genetic distance). Similarly, when no GGBs or FGBs below the genetic distance threshold existed, SGBs were assigned to new GGBs and FGBs following the same procedure.

Prokka (v.1.12 and v.1.13)[73] was used to annotate the open reading frames of all reference genomes and MAGs. Coding sequences were assigned to a UniRef90 cluster[74] by performing a Diamond search (v.0.9.24)[75] of the coding sequences on the UniRef90 database (v.201906) and assigning a UniRef90 identifier when the mean sequence identity to the centroid sequence was greater than 90% and covered more than 80% of the centroid sequence. Sequences that could not be assigned to any UniRef90 cluster following this procedure were de novo clustered with MMseqs2 (ref. [76]) to SGBs following the Uniclust90 criteria[77].

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 the

corresponding GGB (up to the genus level) if this included reference genomes, and of the FGB (up to the family level) if that included reference genomes. Alternatively, if no reference genomes were contained in the FGB, a phylum-level taxonomic label 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 = 1,419; 100 samples being excluded from downstream analyses) using MetaPhlAn 4 (ref. [23,39]) with default parameters and the custom extended SGB database. uSGBs with fewer than five MAGs were discarded, as there is a higher risk of them being 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[77]) 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. The obtained core genes were then split into fragments of 150 nt, and such fragments were then aligned against the genomes of all SGBs using Bowtie2 (v.2.3.5.1; –sensitive option)[64]. Marker genes of a SGB were defined as core genes whose fragments were found in less than 1% of the genomes of any other SGB. When fewer than ten marker genes were found for a SGB, conflicts were defined as occurrences of more than 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 fewer than ten markers were discarded at this point. Merged SGBs (SGB_group) profiled in this study can be found in Supplementary Table 3. The resulting 5.1 M marker genes (average: 189 ± 34.25 s.d. marker genes/SGB) were used as a new reference database for MetaPhlAn 4 (species-level profiling) and StrainPhlAn 4 (strain-level profiling). The presence of *Blastocystis* and the identification of its different subtypes was inferred with a mapping-based computational pipeline described elsewhere[55].

Strain-level profiling of metagenomic samples

Strain profiling was performed with a modified version of StrainPhlAn 3 (ref. [23]) using the custom SGB marker database described above that has been released as StrainPhlAn 4[39]. We modified the StrainPhlAn code to change the sample and marker filtering behavior to allow for profiling more samples and SGBs. A sample was kept as long as it had at least 20 markers (parameter–sample_with_n_markers) and a marker was kept as long as it was present in 50% of the samples (parameter–marker_in_n_samples). After this first filtering, we retained samples with at least ten 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 accuracy of strain sharing detection and to more confidently define strain identity, we additionally considered samples from curatedMetagenomicData (cMD) R package[78] (v.3.15). We

included 4,443 human gut metagenomic samples from 962 individuals older than 6 years from 'Westernized' populations (as defined in cMD) that were sampled longitudinally, obtained from 18 datasets (Supplementary Table 11). For each subject and each SGB, two samples being at most 6 months apart were selected. When more than two timepoints close in time were available, we selected the pair that maximized the lower estimated coverage of the SGB among the two samples, that is, maximized their chance to pass the filtering steps in StrainPhlAn. In case of ties, we took those with higher coverage. Coverage of an SGB in a sample was estimated as [sample sequencing depth] × [relative abundance of the SGB] / [estimated genome length], with estimated genome length being extracted from the MetaPhlAn enlarged database described above. For kSGBs this is determined using only the genome lengths of the reference genomes in the kSGB, whereas 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 (that is, those that are used to select markers, parameter–samples) if they had an estimated coverage of at least 2X that of a given SGB genome, otherwise they were added as secondary samples (that is, those that are added only after the markers are selected with the primary samples, parameter–secondary_samples). In total, 1,033 SGBs that were detected in at least 20 primary samples were profiled at the strain level. To exclude strains likely coming from food sources, we included 216 MAGs in 19 SGBs (Supplementary Table 16) coming from food samples[79] and used them in the StrainPhlAn profiling with the –secondary_references parameters. Samples that had StrainPhlAn mutation rates less than 0.0015 to any food MAG were discarded following the same procedure as in (Valles-Colomer et al., manuscript in preparation). SGBs in which more than 20% of the samples would be discarded using this criterion—constituting in large part of strains regularly found in food—were fully excluded (n = 3 SGBs: *Bifidobacterium animalis* SGB17278, *Lactobacillus acidophilus* SGB7044, *Streptococcus thermophilus* SGB8002). Additionally, we excluded 7 SGBs for which the marker genes alignment length was shorter than 1,000 nucleotides, and another 11 SGBs for which StrainPhlAn was not successful in building a phylogenetic tree.

Inference of strain transmission events

We obtained phylogenetic distances between strains as their leaf-to-leaf branch lengths along the trees (that is, patristic distances) produced by StrainPhlAn (built on marker genes alignments, retaining positions with at least 1% variability), normalized by dividing them by the median phylogenetic distance. As no consensus definition of strain is currently available, to infer strain identity and supported by the clear bimodal distribution of patristic distances of strains from the same individual with the highest peak in 0 (ref. [22]), 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) whenever enough data were available. For all strain-level profiled SGBs, we determined the phylogenetic distance threshold that best separates strains from the same subject (different post-FMT timepoints 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 resulting calculated threshold was greater than the fifth 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 (FDR). For SGBs for which fewer than 50 same-individual comparisons but at least 50 unrelated comparisons were available (in which optimal thresholds cannot reliably be estimated), we used the third percentile of the interindividual phylogenetic distances of subjects in different datasets, which corresponded to the median of all the calculated percentiles in (Valles-Colomer et al., manuscript in preparation). SGBs for which fewer than 50 unrelated comparisons were available (n = 17) were discarded. The SGB-specific phylogenetic distance thresholds for all 995 strain-level analyzed SGBs 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 mislabeled samples[80]. We identified and excluded post-FMT samples (n = 21 out of 1,419) that did not share any strain with neither their corresponding pre-FMT sample nor the donor's sample—something highly unexpected due to the high temporal stability of the gut microbiome[22,23,36,81] and thus potential cases of sample mislabeling. We also identified outliers with more than 20 shared strains between pre-FMT and donor samples while being from two supposedly unrelated individuals (n = 2 cases; Supplementary Fig. 15), most probably not representing true recipient–donor pairs. The third outlier with more than 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 (Bray–Curtis = 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 (Extended Data Fig. 1) and ZouM_2019 was the only dataset with a median of only one strain shared between post-FMT and donor samples (Supplementary Fig. 16), 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, either from the metadata or through private correspondence with the authors. Therefore, 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 very rarely share more than five strains (0.08% of pairs of samples), whereas longitudinal post-FMT samples frequently share more than 15 (56.8% of pairs of samples; Supplementary Fig. 17) as also reported elsewhere[22]. 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 eight or more strains with one donor subject, while no post-FMT samples of the same recipient shared eight or more strains with any other donor subject (Supplementary Fig. 18), we used the criterion of sharing eight or more strains to infer donor–recipient matching in the dataset.

Definition of FMT triads

We considered only complete FMT triads, that is, 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 and there was no indication of which one was used, we picked one randomly since donor samples from the same individual are reasonably stable in terms of species-level composition and strain identity[8,22] (Supplementary Fig. 19). 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 Fig. 1). Where there was more than one round of treatment, we considered only those post-FMT samples that were taken before the second treatment round.

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 defined the fraction of retained strains as 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) and the fraction of donor strains as 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).

Next, we determined the number of engrafted strains as the (absolute) number of shared strains between post-FMT and the donor excluding the strains shared between pre-FMT and the donor samples. In this context we defined four categories that describe the relationship between donor- and recipient individuals (Fig. 1e). ‘Related’: individuals are genetically related or cohabiting/friends; ‘unrelated’: individuals are neither genetically related nor cohabiting/friends as stated in the study manuscript, recruited through public advertisement or hospital’s cohorts; ‘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. 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 four datasets in which the strain could engraft were included in the analyses.

Visualization and ordinations of strain sharing in cohorts

To visualize strain sharing in datasets, we computed networks as well as t-SNE plots based on the number of shared strains between pairs of samples. Unsupervised networks were visualized using the igraph package in R (v.1.2.6)[82] 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 one shared strain are shown. The t-SNE plot was generated using the scikit-learn package[83] in Python (v.1.0.2) with perplexity set to 20 and remaining parameters left default.

Comparing strain- and species-level β-diversities for FMT triad clustering

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 (v.0.2.0) using strain sharing rates dissimilarities (defined as 1 – strain sharing rate) as compared with 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 nonzero abundance and in case of logit transformation the zeros were replaced by the half of the minimal nonzero 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. In 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 9,900 healthy human microbiome samples from 59 datasets spanning different body sites (airways, gastrointestinal tract, oral, skin and urogenital tract; Supplementary Table 11) using MetaPhlAn 4 (ref. [23,39]) with default parameters and the custom SGB database (see above). Only individuals older than 3 years and from cohorts involving industrialized nonrural populations (defined as ‘Westernized’ in cMD[78]) were considered. Age, lifestyle and disease status were considered as reported in cMD[78].

Annotation of SGB phenotypic traits

SGB phenotypes were predicted using Traitar (v.1.1.12)[62] 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 (Fig. 1a) was assessed by PERMANOVA on strain-sharing-based dissimilarities using the adonis function in the vegan package in R (v.2.5–7)[84]. Dissimilarities were computed within each dataset as 1 – (n/M), where n is the number of shared strains and M is the maximum of the number of shared strains.

To compare differences between median strain sharing or engraftment measures (Figs. 1e and 2a,b) in two groups of datasets against the null distribution, permutation tests were applied by randomly permuting the assignments between labels and dataset identifiers 9,999 times.

LOESS fit in Fig. 4d was computed using the geom_smooth function from the ggplot2 (v.3.3.5) in R with standard parameters.

To compare median strain-sharing rates between triads in which the FMT procedure was clinically defined as ‘successful’ and those in which was clinically ‘unsuccessful’ (see above) (Fig. 2c), we applied four statistical tests. First, we used a permutation test applied by randomly permuting the success labels within each dataset 9,999 times. Second, 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 (ref. [85]); the significance was assessed by performing a likelihood-ratio test against a null model without the success indicator variable. Third, we computed median strain sharing rates of successful and unsuccessful groups within each dataset and compared the medians of the successful group with the unsuccessful groups with the Wilcoxon signed-rank test as implemented in the SciPy package[86] (v.1.7.3) in Python. Correction for multiple testing (Benjamini–Hochberg procedure, Q) was applied when appropriate with significance defined at $Q < 0.1$.

Multivariate analysis

A multivariate analysis was performed to assess associations between strain engraftment rates and clinical/nonclinical variables. We included both covariates describing the clinical process, the recipient’s and donor’s microbiomes, and experimental variables consistently available across studies: antibiotics intake (that is, intake close to FMT treatment, intake as a FMT pretreatment or no antibiotic intake); whether the FMT was done to treat an infectious or noninfectious disease; administration of fresh or frozen stool; the amount of feces administered (in grams); the route of FMT administration categorized in ‘upper GI’ routes (capsules, enteroscopy, nasogastric tube, nasoduodenal tube, upper endoscopy, PEG), ‘lower GI’ routes (colonoscopy) and ‘mixed’ routes (FMT protocols utilizing both upper and lower routes for the same recipient); recipient’s age (in years); recipient’s and donor’s α-diversity (Shannon index on species-level abundances); the Bray–Curtis β-diversity and strain-sharing rate between recipient pre-FMT and donor; usage of bead-beating steps for DNA extraction; broad geographic regions based on the recipient’s lifestyle and diet (Mediterranean consisting of Israel, Italy and France[87];

North America consisting of the United States and Canada; Central and Northern Europe consisting of Norway, the Netherlands and Germany; and China). Categorical variables were converted to sets of binary variables, one per each category level (one-hot encoding). All variables were standardized by subtracting the mean and dividing by the s.d.

Since many variables in the analysis are correlated with each other (Supplementary Fig. 6), we performed partial least squares decomposition, which is well-suited for multicollinear data, where the standard linear models are inappropriate. We used the PLSRegression class with parameter scale=False from the scikit-learn[83] Python library (v.1.0.2). The coefficients for each variable composing each component were retrieved through the x_weights_ parameter and the transformed data matrix through the x_scores_ variable returned from the fit_transform method. We regressed each component separately on the strain engraftment rate with ordinary least squares. The first two components were explaining the most the strain engraftment rate and were the only ones significantly associated with it ($R^2 = 0.187$, $Q = 6 \times 10^{-10}$ and $R^2 = 0.046$, $Q = 3.8 \times 10^{-3}$ for the first and second component, respectively; Extended Data Fig. 5). We assessed the association of the variables with the components by hierarchical bootstrap, that is, by resampling the datasets and for each dataset resampling the FMT triads and the associated variables. By resampling the data matrix this way and repeating the PLS decomposition (9,999 iterations) we obtained an estimate of empirical distribution for each weight coefficient.

Machine learning

We used an ML modeling approach to predict the taxonomic composition (presence/absence and relative abundance) of the post-FMT microbiome. To this end, we first organized the data such that each datapoint represented a species in a specific FMT triad. We did not consider species absent in both recipient pre-FMT and donor. As features associated with each datapoint 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 sample collection), species relative abundances for all samples (abundances in the post-FMT were treated as the dependent variables), and Shannon entropy values for pre-FMT and donor samples, information about species (taxonomy, prevalence in an unrelated set of metagenomic samples[23]) and cohort-specific information (dataset, disease infectivity).

We trained RF models[88] both in a LODO as well as in a fivefold CV fashion. In the CV setting, we repeated the entire training/evaluation with five resamplings and averaged the prediction probabilities. To avoid overestimating 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,b). Training and evaluation of RF models was done using the classif.ranger learner (for the presence/absence classifier) and regr.ranger (for the relative abundance regressor) from the mlr3 package (v.0.10) in R[89] with parameter importance = ‘permutation’. We used the unbiased AUROC metric to evaluate the performance of the presence/absence classifier. Feature importance values were obtained directly from the trained RF regression model. Reported AUROC values were calculated per FMT triad and correspond to the AUROC of the predicted post-FMT species against the species actually detected in the post-FMT sample.

The pre-FMT/donor exchange simulations are based on the idea that we can exchange the real pre-FMT/donor individuals with others (from different FMT triads) in silico and then predict and analyze the post-FMT microbiome of these artificial triads. (Fig. 4c,d). Here, 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). In these experiments, we only considered datasets with at least three donors.

To evaluate the ability of the presence/absence classifier to predict continuous post-FMT microbiome traits (Fig. 4e,f,h,i), we computed the predicted species richness of certain groups of bacteria (richness, proteobacterial richness, Firmicutes richness, Bacteroidetes richness, PREDICT 1 species richness (Supplementary Table 14), richness of oral bacterial (Supplementary Table 13). We summed up raw prediction probabilities to estimate richness values. Similarly, for the evaluation of the abundance regressor, we computed the predicted cumulative abundance of the same groups of bacteria described above.

REPORTING SUMMARY

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Newly generated shotgun metagenomics sequencing data are available at the European Nucleotide Archive under accession number PRJEB47909. Metadata are available in Supplementary Table 2 and in curatedMetagenomicData[78].

CODE AVAILABILITY

StrainPhlAn 4 was used for strain-sharing inference, and is available at https://github.com/biobakery/MetaPhlAn. The code to reproduce the ML results can be found under the following link: http://segatalab.cibio.unitn.it/data/FMT_meta.html. All analyses were performed using open-source software.

REFERENCES

1. Ianiro, G. et al. 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. Ann. Intern. Med. 171, 695–702 (2019).
2. Baunwall, S. M. D. et al. Faecal microbiota transplantation for recurrent Clostridioides difficile infection: an updated systematic review and meta-analysis. EClinicalMedicine 29–30, 100642 (2020).
3. Cammarota, G. et al. International consensus conference on stool banking for faecal microbiota transplantation in clinical practice. Gut 68, 2111–2121 (2019).

4. De Groot, P. F., Frissen, M. N., De Clercq, N. C. & Nieuwdorp, M. Fecal microbiota transplantation in metabolic syndrome: history, present and future. Gut Microbes 8, 253–267 (2017).
5. Rossen, N. G. et al. Findings from a randomized controlled trial of fecal transplantation for patients with ulcerative colitis. Gastroenterology 149, 110–118.e4 (2015).
6. Kootte, R. S. et al. Improvement of insulin sensitivity after lean donor feces in metabolic syndrome is driven by baseline intestinal microbiota composition. Cell Metab. 26, 611–619.e6 (2017).
7. Ianiro, G. et al. Faecal microbiota transplantation for the treatment of diarrhoea induced by tyrosine-kinase inhibitors in patients with metastatic renal cell carcinoma. Nat. Commun. 11, 4333 (2020).
8. Davar, D. et al. Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. Science 371, 595–602 (2021).
9. Baruch, E. N. et al. Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. Science 371, 602–609 (2021).
10. Ianiro, G. et al. Systematic review with meta-analysis: efficacy of faecal microbiota transplantation for the treatment of irritable bowel syndrome. Aliment. Pharmacol. Ther. 50, 240–248 (2019).
11. Green, J. E. et al. 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, 1–25 (2020).
12. Ianiro, G., Sanguinetti, M., Gasbarrini, A. & Cammarota, G. Predictors of failure after single faecal microbiota transplantation in patients with recurrent Clostridium difficile infection: results from a 3-year cohort study: authors' reply. Clin. Microbiol. Infect. 23, 891 (2017).
13. Moayyedi, P. et al. Fecal microbiota transplantation induces remission in patients with active ulcerative colitis in a randomized controlled trial. Gastroenterology 149, 102–109.e6 (2015).
14. Ianiro, G. et al. Efficacy of different faecal microbiota transplantation protocols for Clostridium difficile infection: a systematic review and meta-analysis. United European Gastroenterol. J. 6, 1232–1244 (2018).
15. Li, S. S. et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. Science 352, 586–589 (2016).
16. Smillie, C. S. et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. Cell Host Microbe 23, 229–240.e5 (2018).
17. Podlesny, D. et al. Identification of clinical and ecological determinants of strain engraftment after fecal microbiota transplantation using metagenomics. Cell Rep. Med. 3, 100711 (2020).
18. Kumar, R. et al. Identification of donor microbe species that colonize and persist long term in the recipient after fecal transplant for recurrent Clostridium difficile. NPJ Biofilms Microbiomes 3, 12 (2017).

19. Aggarwala, V. et al. Quantification of discrete gut bacterial strains following fecal transplantation for recurrent Clostridioides difficile infection demonstrates long-term stable engraftment in non-relapsing recipients.Nat. Microbiol. 6, 1309–1318 (2021).
20. Wilson, B. C. et al. Strain engraftment competition and functional augmentation in a multi-donor fecal microbiota transplantation trial for obesity. Microbiome 9, 107 (2021).
21. Watson, A. R., Fuessel, J., Veseli, I. & DeLongchamp, J. Z. Adaptive ecological processes and metabolic independence drive microbial colonization and resilience in the human gut. Preprint at bioRxiv https://doi.org/10.1101/2021.03.02.433653 (2021).
22. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 27, 626–638 (2017).
23. Beghini, F. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.eLife 10, e65088 (2021).
24. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol. 39, 727–736 (2021).
25. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176, 649–662.e20 (2019).
26. Bar-Yoseph, H. et al. Oral capsulized fecal microbiota transplantation for eradication of carbapenemase-producing enterobacteriaceae colonization with a metagenomic perspective. Clin. Infect. Dis. 73, e166–e175 (2021).
27. Damman, C. J. et al. Low level engraftment and improvement following a single colonoscopic administration of fecal microbiota to patients with ulcerative colitis. PLoS ONE 10, e0133925 (2015).
28. Goll, R. et al. Effects of fecal microbiota transplantation in subjects with irritable bowel syndrome are mirrored by changes in gut microbiome. Gut Microbes 12, 1794263 (2020).
29. Hourigan, S. K. et al. Fecal transplant in children with Clostridioides difficile gives sustained reduction in antimicrobial resistance and potential pathogen burden. Open Forum Infect. Dis. 6, ofz379 (2019).
30. Kong, L. et al. Linking strain engraftment in fecal microbiota transplantation with maintenance of remission in Crohn's disease. Gastroenterology 159, 2193–2202.e5 (2020).
31. Leo, S. et al. 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, 941 (2020)
32. Moss, E. L. et al. Long-term taxonomic and functional divergence from donor bacterial strains following fecal microbiota transplantation in immunocompromised patients. PLoS ONE 12, e0182585 (2017).
33. Suskind, D. L. et al. Fecal microbial transplant effect on clinical outcomes and fecal microbiome in active Crohn's disease. Inflamm. Bowel Dis. 21, 556–563 (2015).
34. Vaughn, B. P. et al. Increased intestinal microbial diversity following fecal microbiota transplant for active Crohn's disease. Inflamm. Bowel Dis. 22, 2182–2190 (2016).

35. Zhao, H.-J. et al. The efficacy of fecal microbiota transplantation for children with Tourette syndrome: a preliminary study. Front. Psychiatry 11, 554441 (2020).
36. Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550, 61–66 (2017).
37. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. Nat. Rev. Microbiol. 18, 491–506 (2020).
38. Segata, N. On the road to strain-resolved comparative metagenomics. mSystems 3, e00190–e001917 (2018).
39. Blanco-Miguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. Preprint at bioRxiv https://doi.org/10.1101/2022.08.22.504593 (2022)
40. Gulati, M., Singh, S. K., Corrie, L., Kaur, I. P. & Chandwani, L. Delivery routes for faecal microbiota transplants: available, anticipated and aspired. Pharmacol. Res. 159, 104954 (2020).
41. Smith, B. J. et al. Strain-resolved analysis in a randomized trial of antibiotic pretreatment and maintenance dose delivery mode with fecal microbiota transplant for ulcerative colitis. Sci. Rep. 12, 5517 (2022).
42. Kim, S., Covington, A. & Pamer, E. G. The intestinal microbiota: antibiotics, colonization resistance, and enteric pathogens. Immunol. Rev. 279, 90–105 (2017).
43. Soldi, S. et al. Modulation of the gut microbiota composition by rifaximin in non-constipated irritable bowel syndrome patients: a molecular approach. Clin. Exp. Gastroenterol. 8, 309–325 (2015).
44. Jakobsson, H. E. et al. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. PLoS ONE 5, e9836 (2010).
45. Hu, Y. et al. Different immunological responses to early-life antibiotic exposure affecting autoimmune diabetes development in NOD mice. J. Autoimmun. 72, 47–56 (2016).
46. Feuerstadt, P. et al. SER-109, an oral microbiome therapy for recurrent Clostridioides difficile infection. N. Engl. J. Med. 386, 220–229 (2022).
47. Chehri, M. et al. Case series of successful treatment with fecal microbiota transplant (FMT) oral capsules mixed from multiple donors even in patients previously treated with FMT enemas for recurrent Clostridium difficile infection. Medicine 97, e11706 (2018).
48. Willmann, M. et al. Distinct impact of antibiotics on the gut microbiome and resistome: a longitudinal multicenter cohort study. BMC Biol. 17, 76 (2019).
49. Chang, J. Y. et al. Decreased diversity of the fecal microbiome in recurrent Clostridium difficile-associated diarrhea. J. Infect. Dis. 197, 435–438 (2008).
50. Rands, C. M., Brüssow, H. & Zdobnov, E. M. Comparative genomics groups phages of Negativicutes and classical Firmicutes despite different Gram-staining properties. Environ. Microbiol. 21, 3989–4001 (2019).
51. Tett, A., Pasolli, E., Masetti, G., Ercolini, D. & Segata, N. Prevotella diversity, niches and interactions with the human host. Nat. Rev. Microbiol. 19, 585–599 (2021).
52. Gardiner, B. J. et al. Clinical and microbiological characteristics of Eggerthella lenta bacteremia. J. Clin. Microbiol. 53, 626–635 (2015).
53. Asnicar, F. et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat. Med. 27, 321–332 (2021).

54. Tito, R. Y. et al. Population-level analysis of Blastocystis subtype prevalence and variation in the human gut microbiota. Gut 68, 1180–1189 (2019).
55. Beghini, F. et al. Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. ISME J. 11, 2848–2863 (2017).
56. Scanlan, P. D. et al. The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. FEMS Microbiol. Ecol. 90, 326–330 (2014).
57. Terveer, E. M. et al. Human transmission of Blastocystis by fecal microbiota transplantation without development of gastrointestinal symptoms in recipients. Clin. Infect. Dis. 71, 2630–2636 (2020).
58. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. Nature 489, 220–230 (2012).
59. Ferri, M., Ranucci, E., Romagnoli, P. & Giaccone, V. Antimicrobial resistance: a global emerging threat to public health systems. Crit. Rev. Food Sci. Nutr. 57, 2857–2876 (2017).
60. Zellmer, C. et al. Shiga toxin–producing Escherichia coli transmission via fecal microbiota transplant. Clin. Infect. Dis. 72, e876–e880 (2020).
61. Li, Y. & Honda, K. Towards the development of defined microbial therapeutics. Int. Immunol. 33, 761–766 (2021).
62. Weimann, A. et al. From genomes to phenotypes: Traitar, the microbial trait analyzer. mSystems 1, e00101–e00116 (2016).
63. Quagliariello, A. et al. Fecal microbiota transplant in two ulcerative colitis pediatric cases: gut microbiota and clinical course correlations. Microorganisms 8, 1486 (2020).
64. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
65. Seymour, L. et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncol. 18, e143–e152 (2017).
66. Benson, D. A. et al. GenBank. Nucleic Acids Res. 41, D36–D42 (2012).
67. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834 (2017).
68. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 (2015).
69. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359 (2019).
70. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015).
71. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132 (2016).
72. Asnicar, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat. Commun. 11, 2500 (2020).
73. Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069 (2014).

74. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932 (2015).
75. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60 (2015).
76. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028 (2017).
77. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45, D170–D176 (2017).
78. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. Nat. Methods 14, 1023–1024 (2017).
79. & Edoardo, P. et al. Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. Nat. Commun. 11, 2610 (2020).
80. Podlesny, D. & Fricke, W. F. Strain inheritance and neonatal gut microbiota development: a meta-analysis. Int. J. Med. Microbiol. 311, 151483 (2021).
81. Albanese, D. & Donati, C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat. Commun. 8, 2260 (2017).
82. Csardi, G. & Nepusz, T. The igraph software package for complex network research. InterJournal Complex Syst. 1695, https://igraph.org/ (2006).
83. Pedregosa, F. et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
84. Oksanen, J. et al. vegan: Community Ecology Package. https://cran.r-project.org/package=vegan (2020).
85. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4.J. Stat. Softw. 67, 1–48 (2015).
86. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272 (2020).
87. Lăcătușu, C.-M., Grigorescu, E.-D., Floria, M., Onofriescu, A. & Mihai, B.-M. The Mediterranean diet: from an environment-driven food culture to an emerging medical prescription. Int. J. Environ. Res. Public Health 16, 942 (2019).
88. Breiman, L. Random forests. Mach. Learn. 45, 5–32 (2001).
89. Lang, M. et al. mlr3: a modern object-oriented machine learning framework in R. J. Open Source Softw. 4, 1903 (2019).

ACKNOWLEDGEMENTS

We thank all study participants for their commitment and the following authors of included studies for their help in providing data and metadata: H. Bar-Yoseph, R. Goll, H. Koo, S. Leo, C. Morrow, A. Moss, D. Suskind and F. 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 P. Merli, P. De Angelis, G. Angelino, E. Francesco Romeo and L. Gargiullo, S. Pane, S. Martino, L. Romani, P. Bernaschi, A. Finocchi, G. Marucci, F. Rea, S. Faraci, P. D' Argenio, L. Dall'Oglio) and the biobank of the Ospedale Pediatrico Bambino Gesù. Moreover, the staff of the Fondazione Policlinico Gemelli IRCCS thank the Fondazione Roma for the invaluable support to their scientific research. This work was supported by the European

Research Council (ERC-STG project MetaPG-716575) to N.S., by MIUR ‘Futuro in Ricerca’ (grant no. RBFR13EWWI_001) to N.S.; by the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project) to N.S.; by the National Cancer Institute of the National Institutes of Health (1U01CA230551) to N.S.; by the Premio Internazionale Lombardia e Ricerca 2019 to N.S.; by the Italian Ministry of Health with Ricerca Corrente and 5×1000 funds to N.S.; by the EMBO ALTF 593-2020 to M.V.-C.; by the Ricerca Finalizzata Giovani Ricercatori 2018 of the Italian Ministry of Health (project GR-2018-12365734) to G.I.; by the BIOMIS grant of the Italian Ministry of Research to A.G., G.C. and G.I. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. 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 pediatric (2–17 years old) participants. The systematic review was not registered.

AUTHOR INFORMATION

Author notes

These authors contributed equally: Gianluca Ianiro, Michal Punčochář, Nicolai Karcher.

These authors jointly supervised this work: Mireia Valles-Colomer, Giovanni Cammarota, Nicola Segata.

Authors and Affiliations

Digestive Disease Center, Fondazione Policlinico Universitario ‘A. Gemelli’ IRCCS, Rome, Italy
Gianluca Ianiro, Serena Porcari, Silvia De Giorgi, Giusi Desirè Sciumè, Stefano Bibbò, Antonio Gasbarrini & Giovanni Cammarota

Department of Translational Medicine and Surgery, Catholic University of Rome, Rome, Italy
Gianluca Ianiro, Serena Porcari, Silvia De Giorgi, Giusi Desirè Sciumè, Stefano Bibbò, Antonio Gasbarrini & Giovanni Cammarota

Department CIBIO, University of Trento, Trento, Italy
Michal Punčochář, Nicolai Karcher, Federica Armanini, Francesco Asnicar, Francesco Beghini, Aitor Blanco-Míguez, Fabio Cumbo, Paolo Manghi, Federica Pinto, Mireia Valles-Colomer & Nicola Segata

Microbiology Unit, Fondazione Policlinico Universitario ‘A. Gemelli’ IRCCS, Rome, Italy
Luca Masucci, Gianluca Quaranta & Maurizio Sanguinetti

Department of Basic Biotechnological Sciences, Intensivological and Perioperative Clinics, Catholic University of Rome, Rome, Italy
Luca Masucci, Gianluca Quaranta & Maurizio Sanguinetti

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
Federica Del Chierico & Lorenza Putignani

IEO, Istituto Europeo di Oncologia IRCSS, Milan, Italy
Nicola Segata

Contributions

G.I. and N.S. conceived and designed the study. M.P. and N.K. performed the analysis. G.I., N.K. and S.P. performed the literature search. G.I. and G.C. supervised the sample collection and the clinical procedures. G.I., M.V.-C. and N.S. supervised the analysis. M.P., N.K., F. Armanini, F. Asnicar, F.B., A.B.-M., F.C., P.M. and F.P. contributed to data acquisition, data analysis or software development. G.I., S.P., L.M., G.Q., S.D.G., G.D.S., S.B., L.P., F.D.C., M.S., A.G. and G.C. contributed to the clinical procedures and sample collection. G.I., M.P., N.K., M.V.-C. and N.S. interpreted the data and wrote the manuscript. All authors provided critical revision of the manuscript and approved the final version for submission.

Corresponding authors

Correspondence to Gianluca Ianiro or Nicola Segata.

Competing interests

A.G. 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. G.C. has received personal fees for acting as advisor for Ferring Therapeutics. G.I. has received personal fees for acting as speaker for Biocodex, Danone, Sofar, Malesci, Metagenics and Tillotts Pharma, and for acting as consultant/advisor for Ferring Therapeutics, Giuliani, Metagenics and Tillotts Pharma. N.S. reports consultancy and/or SAB contracts with Zoe, Roche, Ysopia, and Freya, Alia Therapeutics, speaker fees by Illumina, and is cofounder of PreBiomics. The other authors have no potential competing interest to disclose.

Peer review information

Nature Medicine thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Alison Farrell, in collaboration with the Nature Medicine team.

Additional information

Chapter 4: Other contributions

Expansion of SGB database

During my PhD I contributed to the expansion of a large collection of MAGs and their clustering into SGBs originally published in ref. [54]. I implemented a pipeline for expansion of the genome database and SGB clusters. I validated new tools skani[60] and CheckM 2[76] to substitute previously utilized MASH[59] and CheckM 1[77], respectively. I expanded the database from circa 700,000 to circa 2,100,000 genomes currently that cluster to circa 240,000 SGBs.

The database has been foundational in the new version 4 of MetaPhlAn[78] profiling tool and its applications[79]. Part of the MAG database coming from food samples has been explored in the curatedFoodMetagenomicData project[19]. I have utilized the MAG database for genomic comparison to help define novel subspecies *Catenibacterium mitsuokai* subsp. *tridentinum* subsp. nov.[80]

- Blanco-Míguez A, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol. 2023;41: 1633–1644.
- Manghi P, Blanco-Míguez A, et al. MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. Cell Rep. 2023;42: 112464.
- Carlino N, et al. Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. Cell. 2024;187: 5775–5795.e15.
- Ricci L, et al. Description of *Catenibacterium mitsuokai* subsp. *tridentinum* subsp. nov., an anaerobic bacterium isolated from human faeces, and emended description of *C. mitsuokai*. Int J Syst Evol Microbiol. 2025;75.

Support in other strain-transmission analyses

By developing, maintaining and applying strain-sharing computational pipeline I contributed to other projects studying mother-to-infant transmission in infants with neuroblastoma[81] or transmission in social networks of lemurs[82]. I also helped in validation of strain sharing results during the revision process of a large microbiome transmission meta-study[17] using our FMT meta-cohort collected in Chapter 3.

- Valles-Colomer M, et al. Neuroblastoma is associated with alterations in gut microbiome composition subsequent to maternal microbial seeding. EBioMedicine. 2024;99: 104917.
- Labisa-Morais F, et al. Bacterial transmission within social groups shapes the underexplored gut microbiome in the lemur Indri indri. ISME J. 2025.
- Valles-Colomer M, Blanco-Míguez A, Manghi P, Asnicar F, Dubois L, Golzato D, et al. The person-to-person transmission landscape of the gut and oral microbiomes. Nature. 2023;614: 125–135.

Support and analysis of engraftment in FMT

The strain sharing computational pipeline has been applied also in other FMT studies following our large meta-study in Chapter 3. For a clinical trial on immunotherapy for melanoma I analyzed the strain engraftment as shown in Fig. 2b-c of the manuscript[32] and I supported the analysis of strain engraftment in other two phase II trials[33,34].

- Routy B, et al. Fecal microbiota transplantation plus anti-PD-1 immunotherapy in advanced melanoma: a phase I trial. Nat Med. 2023;29: 2121–2132.
- Porcari S., Ciccarese Ch., Heidrich V. et. al. Fecal microbiota transplantation plus pembrolizumab and axitinib in metastatic renal cell carcinoma: the randomized phase 2 TACITO trial. *Nat Med. 2026*
- Duttagupta S., Messaoudene M., Hunter, S. et. al. Fecal microbiota transplantation plus immunotherapy in non-small cell lung cancer and melanoma: the phase 2 FMT-LUMINate trial. *Nat Med. 2026*

Other minor contributions

I participated in a few more works, contributions to which were minor or not in line with my PhD topic.

- Karcher N, et al. Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. Genome Biol. 2021;22: 209.
- Blanco-Míguez A, et al. Extension of the Segatella copri complex to 13 species with distinct large extrachromosomal elements and associations with host conditions. Cell Host Microbe. 2023;31: 1804–1819.e9.
- Alves Costa Silva C, et al. Influence of microbiota-associated metabolic reprogramming on clinical outcome in patients with melanoma from the randomized adjuvant dendritic cell-based MIND-DC trial. Nat Commun. 2024;15: 1633.
- Piccinno G, et al. Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. Nat Med. 2025;31: 2416–2429.

Chapter 5: Discussion

Summary

From amplicon to shotgun metagenomic sequencing, researchers were able to detect bacteria on a high throughput basis and classify them into genera and species. It has been shown that natural clustering seems to happen around 95% genomic identity and we exploit the fact to define novel species (SGBs), but the variability within the species is not to be ignored. Strains of the same species can differ functionally by having different metabolizing capabilities or pathogenicity, but also the collection of strains in a person's microbiome provides a sort of fingerprint that is unique to each individual. Looking at genomic variation within bacterial strains we can trace the origin of our commensal microbes similarly to already established methods of tracking pathogenic viruses or bacteria during outbreaks. The development of computational methods to track strains from metagenomic sequencing allows us to answer the basic question of the origins of the microbes in our gut. The possibly most obvious hypothesis of obtaining microbes directly from our environment fails after realizing that gut microbes generally are not found outside our guts, giving rise to the transmission hypothesis that has been put forward recently and explored in Chapter 2. Moreover, fecal microbiota transplantation is a way of "forced transmission" from a donor to a recipient with the objective to improve the recipient's health or treatment outcome. By tracking strains we can determine which engraft and which do not and by understanding the microbiological principles of engraftment we can help guide the clinical application of this method as explored in Chapter 3.

Person-to-person microbiome transmission

Our study of babies attending nursery for the first time (Chapter 2) highlights the importance of social contact in the development of the infant's gut microbiome. The strain sharing between the babies is substantial already after the first month of attendance and by the end of the first term the contribution from the nursery strain pool levels out with the contribution from the family. Moreover, thanks to the longitudinal design, we can make stronger claims about the directionality of the transmission, unlike cross-sectional studies that are usually limited to discuss mere "strain sharing" and not "strain transmission". We were able to track strains at different time-points spreading throughout our studied population gaining novel insights into the transmission dynamics.

Chapter 2 attempts to complete the picture of microbes acquisition throughout our lives, with the overarching theme being transmission due to social contact. From the first and closest contact between a new born baby and its mother, followed by contact with father and siblings to other children in nurseries and throughout adulthood a person will keep exchanging strains with those in close contact. Even though the picture can be considered complete, it is so only by large strokes with important details left to be filled in. The exact process of transmission is not completely understood. We can hypothesise microbes are transmitted through direct skin contact, commonly used object surfaces or as air-borne particles via so-called fecal-oral route

as studied in pathogens. Even though the gut microbes don't thrive on skin or object surfaces, they have to be able to survive at least for a brief time via sporulation or other mechanisms. Different species will have different "transmissibility" as shown in Chapter 2 Figure 5 for the nurseries and also in ref. [17] (analogous is the "engraftability" in Chapter 3 Figure 3), although summarizing transmissibility into a single number is problematic. The path of a strain being transmitted from one person to another can be visualised as a funnel consisting of different rate limiting processes including emission from the originating person to the environment, survival for a given time in the environment, uptake by the other person, passage through digestive tract, and colonization within the existing target community. We can imagine how various species will have different rates at different steps. The more abundant ones will likely spread more into the surroundings while maybe others will be better at surviving in an aerobic environment for prolonged durations or through the highly acidic parts of the digestive tract. In the future breaking down transmission into such processes and studying each individually will be important to completely understand the biological basis of transmission.

Computational prediction of traits like sporulation and aerotolerance is possible by training models on curated datasets[83,84] and could explain why some species are more likely transmitted, but the most difficult step to disentangle is the stable colonization as it depends on the transmitted strain, the target community it has to fit into and the host characteristics. Several factors under the umbrella of colonization resistance like nutrient availability for the arriving strain, cross-feeding from other species or toxin production for increase in competitiveness have to be considered. Here, mechanistic experiments using controlled microbial communities in bioreactors or gut simulators, are invaluable tools to study interactions between microbes, but are not a scalable way to determine the possible behavior of all species and strains. The increasing number of FMT studies with microbiome sampling could be a source of data for computational inference as in FMT all microbes have more or less equal starting point in the engraftment process as the steps of spread and survival outside gut or through digestive tract are bypassed. Possibly this is what our list of most engrafted species in Chapter 3 Figure 3 represents. Being able to accurately model the interaction network of any microbial community is definitely one of the future directions, but at the moment is still out of reach. The ability to retrieve from a database or computationally predict phenotypes like metabolic potential for each strain would be needed. Most microbial protein-coding genes are not characterized and so the development of accurate methods of molecular function prediction [85] are moving us in the right direction.

Since microbiome has been linked to many health and disease states of the host and transmission is the way we acquire microbes, the natural question is how different ways of transmission affect the health outcomes. Firstly it is important to discern whether microbiome differences play a causal role or are just a reflection of the underlying host state or other lifestyle and environmental factors. If the microbiome is determined to be at a cause, the transmission should be taken into account. For example, the effect of attending nursery vs. not, having partners, co-habiting, having pets, hygiene and sanitization habits then all might play a role in human health. Analogously to pathogens, where increased sanitization and isolation lowers the probability of infection, we could hypothesize these practices would lead to lesser acquisition of

new commensal strains, but the long term effects have not been studied so far. The establishment of a causal microbiome-phenotype relationship faces many challenges and therefore many links between microbiome and health and disease remain on the level of associations. Mediation analysis, for example in diet intervention studies, might be an underutilized technique to focus more on in the future. FMT studies are possibly the ultimate causality test as the intervention is limited to the microbiome, but the possible side effects discourage its application in less serious settings. Prebiotics intervention is a safer alternative and can directly test an effect of the given species, although its engraftment (stable colonization) might be limited as it could depend on the presence of other interacting species not administered.

Fecal microbial transplantation

In our FMT meta-analysis (Chapter 3) we showed the most impactful factors on engraftment being antibiotics treatment, infectious disease and combined route of delivery. This provides an important first step towards informed decisions in clinical practice. Despite collecting all available metagenomic studies on FMT, the sample size relative to the vast variability of population demographics, underlying disease and clinical parameters, prevented us from eliciting more finer patterns. Higher engraftment was associated with improved clinical outcome across the studies (Chapter 3 Figure 2) and based on our study it could be suggested that pre-treating recipients with antibiotics and using multiple routes of delivery would maximize the chances. This signal, however, is not strong across all cohorts as the mechanism of action is disease specific and higher engraftment might not necessarily be better. Instead, the addition of certain beneficial strains or the induction of a loss of unbeneficial strains in the recipient might be more precise ways to think about the link of engraftment with the clinical outcomes.

In the ideal future, clinicians would select the best procedure and the best matching donor (or even a defined bacterial consortium) for a given recipient and their medical condition. To continue the path towards this precision medicine ideal, at least two steps are needed. Firstly, understanding the microbiological basis of engraftment on the level of strains and community interaction. This would correspond to the last step of the transmission funnel discussed above, i.e. accurately modelling the interaction between microbes and the host to predict colonization resistance for the transplanted strains. The second step is to uncover the link between the microbiome and target disease or the treatment response. By putting together both steps we can determine the best set of strains to transplant that upon interaction with the recipient's microbiome will produce the desired outcome. Since the analysis performed in Chapter 3, more studies have been published on FMT with metagenomic sequencing. The promising emerging field of application is cancer immunotherapy as FMT due to modulating microbes interacting with the host immunity has the potential to turn treatment non-reponders to responders. One more phase I trial in melanoma (MIMIC) has been published and two other phase II trials in non-small lung cell cancer and renal cancer have been accepted for publication, to all of which I've contributed via computational analysis (Chapter 4: Other contributions). Compared to *C. diff.* infection, the mechanism is not clear and conducting more studies and pooling them

together will be important in the future to find out what about the donor microbiome or the interaction of the donor-recipient microbiomes determines a successful response.

Towards better delineation of species

The computational methods of microbiome analysis rely often on reference genome or gene databases and defined taxonomical units. In this work I utilized mapping approaches within the SGB framework and MetaPhlAn marker gene database allowing for quantification and strain profiling of uncharacterized and low abundant species. The species diversity of the human gut microbiome is well-covered in the genome databases but to extend to other animal hosts or environments like waters or soil a continued effort in sample collection and metagenomic assembly are needed in covering those habitats[86,87]. To define species-like taxonomic units clustering on genomic similarity is usually adopted. Using a universal threshold such as 95% ANI on the common portion of the genome might not be optimal within every lineage and has been disputed[88]. For example, common gut microbes *Phocaeicola dorei* and *Phocaeicola vulgatus* are genomically more similar around 97.5% ANI[89], while *Streptococci* species can display large within-species diversity down to 91% ANI[90] and require more specific phylogenetic analysis to define their evolutionary relationships[91]. In the future, attention should be put towards more careful definitions of species-like taxonomic units to avoid misleading results in downstream analyses. Development of clustering algorithms that are able to adapt the similarity threshold based on local structure of the genomic space might be one direction. Combination of approaches, such as considering both ANI and alignment fraction (the overlap of the genomes) of the whole genome comparison and similarity of universal phylogenetic markers might be another direction. Alignment fraction has been proposed to be used together with ANI for species delineation[92], but its application is limited to genomes from isolate sequencing and potentially complete MAGs as otherwise the observed overlap between MAGs will be driven by their (in)completeness. Developing an algorithm to estimate true alignment fraction between incomplete MAGs could be a challenging future direction of research and development. Using a subset of the genome such as the core genes or phylogenetic marker genes could be useful especially for species with large or very variable accessory genomes.

Present and future of computational methods for transmission

The computational methods to analyze strain sharing for strain transmission and engraftment are based on comparison of genomic content of two strains. The ideal scenario of fully observing a single strain’s genome is not always met in metagenomic analysis. In this work, the presence of other species in the same metagenome is mitigated by using SGB-specific markers defined in the MetaPhlAn database. However, the polymorphisms within the strain population or presence of multiple different strains of the same species remains a challenge as their reads are mixed in metagenomic sequencing and the subsequent alignment. In our analyses with StrainPhlAn we rely on the majority vote on alleles at polymorphic positions in the marker genes in order to reconstruct the dominant strain. The obvious downside is the potential undetected strain transmission or engraftment of non-dominant strains. Additionally, at low coverages due to

random sampling of fragments during sequencing, alleles from only one of the strains might be observed at certain positions leading to incorrect reconstruction of a hybrid strain. Detection of strain sharing of non-dominant strains is especially important in FMT, where the donor and recipient strains may co-exist for a duration of time.

In the future, focus should be put on developing metagenomic tools for strain reconstruction and strain tracking considering polymorphisms and multiple strain co-existing. The task can be computationally defined as either fully reconstructing each strain's genotype from each sample or determining a strain match between samples without regard to number of strains or their relative abundance (that is any strain in one sample matching any strain in the other sample). Genotype reconstruction is inherently more difficult as it requires correctly phasing each strain's alleles across the length of the genome or marker genes, while the direct strain matching can use every allele without the need to assign them to strain genotypes. At present, several tools have been developed in both categories[68,69,93–103], but they often have very high coverage requirements limiting their application to only the most abundant species. So far, current methods of reconstruction utilize either allele frequency to assign alleles to strains, which is prone to misassignment errors due to randomness in sequencing coverage, or assembly-like phasing which requires high coverage of every strain. A combination of allele frequency with phasing through read linkage could be a future direction of reconstruction methods. The long reads technology can be very helpful in phasing strains' alleles across long distances, since long portions of the genome spanning multiple genes can be captured in one read. The error rate is an important factor to consider and technologies with lower error rate like the PacBio HiFi reads would be preferable. Hybrid sequencing combining long reads with short reads sequencing could be an interesting approach where high coverage pileup of short reads could be used to confidently determine the existing alleles in the strain population and long reads would be used for allele linkage. Single cell DNA sequencing similarly can be utilized as the individual cells (that is individual strains) are sequenced separately, but its application to microbial communities is still in the early research stages[104]. Also in short read sequencing using paired-end mode especially with longer fragments can be employed to determine allele linkage.

The tools for direct strain matching so far all utilize the population ANI (pop-ANI) metric, which calculates ANI between two strain populations in two samples considering a match at a position if any allele from one population matches any allele in the other population. Although this approach has the potential to capture a match of a strain that is non-dominant in either of the samples, it can do so only when all polymorphic positions of the non-dominant strain are covered in the sequencing, limiting the matching to only high abundant non-dominant strains. The development of improved methods for strain tracking will enable the further studies of person-to-person transmission and FMT engraftment to be more precise and uncover the fine aspects of multi-strain dynamics.

Determining the genetic or phylogenetic similarity threshold for calling strain sharing events is not a simple task. In an idealized model given a constant mutation rate (*mutations / loci / replication*) and growth rate (*replications / unit of time*) we could by setting a time-frame of the

transmission determine the maximum number of SNPs the two strains can have between them along their genome. However, there are several challenges to this approach. Growth rates and mutation rates are not experimentally known for all species. Moreover they could vary across strains and due to the environment and change through rapid evolution and adaptation, sometimes even giving rise to so-called hypermutators [105,106]. Some of the mutations might be due to recombination between different strains of the same species as suggested previously[67] and explored also more recently[107,108] thus not following the classical evolutionary model. Secondly, in frameworks such as StrainPhlAn only certain marker genes are assessed, which can have different mutation rates compared to the whole genome comparison due to the selection pressure on those genes compared to other parts of the genome – typically the SNP rate observed on marker genes is lower compared to the one between the full genomes. Thirdly, and possibly the most impactfully, along with true SNPs, sequencing errors can introduce further observed differences between strains. Here a tradeoff between reduction of the effect of sequencing errors and sensitivity is encountered as requiring higher coverage for a locus increases the likelihood of observing the true allele, but decreases the limit of detection towards higher abundance strains. When comparing isolate genomes sequenced to a relatively high coverage as done for example in the pathogen tracking field, the sequencing error can be neglected, but in metagenomics working with low abundance species sequencing error can introduce false SNPs. In this thesis a predictive approach is employed where the thresholds are optimally fitted for the task of discriminating short-term persistence from long-term evolution, which does not make assumptions in the evolutionary model and takes into account noise from sequencing error, but presents some other limitations. The thresholds are different for each species and don't have an easy interpretability in terms of SNPs or evolutionary time between strains. Strains genetically closer are more likely to fall below the thresholds and thus within a population an increased strain sharing is expected compared to across populations. For example, even if two people never met, maybe their sources of their strains did and further on. The results of the analyses in this thesis like strain sharing rate between babies in a nursery or engraftment rate between donor and recipient are always presented in comparison to the background: strain sharing between babies from different families before they started nursery or unrelated recipients and donors. The signal-to-noise ratio in these analyses is high enough allowing us to clearly observe patterns and draw conclusions, but in the future more attention should be devoted to understanding the sources of error especially for the study of species transmissibility as different SGBs could have different biases and error rates.

Applications of new computational methods for strain transmission

The strain sharing pipeline that I implemented, optimized and validated presents a powerful framework that can be applied in various settings to answer various research questions. In Chapters 2 and 3 I applied it to study microbiome transmission within a relatively closed community in an early life and engraftment of donor strains after FMT. An example illustrating the power of this approach is a study of microbiome of lemurs to which I contributed (Chapter 4: Other contributions)[82], where virtually all microbial species of this animal are uncharacterized and thus MAG assembly followed by SGB database expansion played a crucial role and

together with the strain sharing pipeline allowed to describe its microbiome composition and the microbiome transmission patterns related to its social behaviour. Application to understudied animal organisms is one direction of application, ranging from endangered species like the *indri indri* lemur, to animals in zoos and potential transmission among them and their human caretakers, to pets and their owners. Conducting longitudinal studies sampling human microbiomes from multiple locations such as skin, oral cavity and gut, will help us to understand the amounts and directions of flow of strains. It has been shown previously that enrichment of oral taxa in the gut is indicative of several possible diseases[109] and thus understanding the strain flow between body sites is important also for human health.

In the human gut, a single strain of each species usually takes dominance[70], but other non-dominant strains can co-exist. As the recent work on strain richness demonstrates[71], the limited strain richness in the gut is not observed in other microbiomes. In the special case of FMT, the donor strains co-exist with the recipient strains at least for a short while until the community resolves to a stable state. In other body-sites like skin or oral plaque the strain richness is usually higher due to possibility of spatial segregation and thus reduced competition. Even higher strain richness is observed in very complex environments like soil. The understanding of strain transmission in those settings will be enabled by the development of methods profiling and tracking multiple non-dominant strains, the technical aspects of which I discussed in the previous section. Even answering the relatively simple questions like whether there are multiple co-existing conspecific strains and possibly the estimation of their number and relative abundance from metagenomic samples would help us understand much more about the strain-level composition of various microbiomes and in the minimum provide an estimate of the validity of tracking only the dominant or high coverage non-dominant strains as done with current methods based on StrainPhlAn or pop-ANI.

To illustrate the varied possibilities of application of the strain sharing methodology, I discuss two more examples. Sample contamination or mislabeling is prevalent in the microbiome sequencing studies and can happen at various stages. By checking the strain sharing between the samples even if the aim of the analysis is not transmission, one can compare the results to the expected outcome, i.e. little to no sharing between unrelated samples, most strains shared among longitudinal samples, etc. and thus uncover and sometimes even correct potential errors not possible to detect from the taxonomic composition alone. Another more futuristic application is in the forensic field. For example, strain sharing between people could provide additional evidence for their close contact. Additionally, matching skin microbiome strains on an object surface could provide evidence for touch, but it is to be determined whether it provides any advantage over the current ways of detection of host DNA.

References

1. Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020;8: 103.

2. Muramatsu MK, Winter SE. Nutrient acquisition strategies by gut microbes. Cell Host

Microbe. 2024;32: 863–874.

3. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, Ventura M. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. Curr Opin Biotechnol. 2013;24: 160–168.

4. Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. Cell Res. 2020;30: 492–506.

5. Ahmed H, Leyrolle Q, Koistinen V, Kärkkäinen O, Layé S, Delzenne N, et al. Microbiota-derived metabolites as drivers of gut-brain communication. Gut Microbes. 2022;14: 2102878.

6. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A. 2007;104: 13780–13785.

7. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun. 2017;8. doi:10.1038/s41467-017-00900-1

8. Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome and human cancer. Science. 2021;371: eabc4552.

9. Retschnig G, Rich J, Crailsheim K, Pfister J, Perreten V, Neumann P. You are what you eat: relative importance of diet, gut microbiota and nestmates for honey bee, Apis mellifera, worker health. Apidologie (Celle). 2021;52: 632–646.

10. Martino C, Dilmore AH, Burcham ZM, Metcalf JL, Jeste D, Knight R. Microbiota succession throughout life from the cradle to the grave. Nat Rev Microbiol. 2022;20: 707–720.

11. Fan J, Zhou Y, Meng R, Tang J, Zhu J, Aldrich MC, et al. Cross-talks between gut microbiota and tobacco smoking: a two-sample Mendelian randomization study. BMC Med. 2023;21: 163.

12. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature. 2018;555: 623–628.

13. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe. 2018;24: 133–145.e5.

14. Korpela K, Costea P, Coelho LP, Kandels-Lewis S, Willemsen G, Boomsma DI, et al. Selective maternal seeding and environment shape the human gut microbiome. Genome Res. 2018;28: 561–568.

15. Wampach L, Heintz-Buschart A, Fritz JV, Ramiro-Garcia J, Habier J, Herold M, et al. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. Nat Commun. 2018;9: 5091.

16. Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, et al. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe. 2018;24: 146–154.e4.

17. Valles-Colomer M, Blanco-Míguez A, Manghi P, Asnicar F, Dubois L, Golzato D, et al. The person-to-person transmission landscape of the gut and oral microbiomes. Nature. 2023;614: 125–135.

18. Brito IL, Gurry T, Zhao S, Huang K, Young SK, Shea TP, et al. Transmission of human-associated microbiota along family and social networks. Nat Microbiol. 2019;4: 964–971.

19. Carlino N, Blanco-Míguez A, Punčochář M, Mengoni C, Pinto F, Tatti A, et al. Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. Cell. 2024;187: 5775–5795.e15.

20. Mahmud B, Vargas RC, Sukhum KV, Patel S, Liao J, Hall LR, et al. Longitudinal dynamics of farmer and livestock nasal and faecal microbiomes and resistomes. Nat Microbiol. 2024;9: 1007–1020.

21. Browne HP, Neville BA, Forster SC, Lawley TD. Transmission of the gut microbiota: spreading of health. Nat Rev Microbiol. 2017;15: 531–543.

22. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science. 2014;345: 1048–1052.

23. Liddicoat C, Sydnor H, Cando-Dumancela C, Dresken R, Liu J, Gellie NJC, et al. Naturally-diverse airborne environmental microbial exposures modulate the gut microbiome and may provide anxiolytic benefits in mice. Sci Total Environ. 2020;701: 134684.

24. Finlay BB, CIFAR Humans, Microbiome. Are noncommunicable diseases communicable? Science. 2020;367: 250–251.

25. Heidrich V, Valles-Colomer M, Segata N. Human microbiome acquisition and transmission. Nat Rev Microbiol. 2025;23: 568–584.

26. Baunwall SMD, Lee MM, Eriksen MK, Mullish BH, Marchesi JR, Dahlerup JF, et al. Faecal microbiota transplantation for recurrent Clostridioides difficile infection: An updated systematic review and meta-analysis. EClinicalMedicine. 2020;29-30: 100642.

27. Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. Nat Rev Gastroenterol Hepatol. 2016;13: 508–516.

28. Andary CM, Al KF, Chmiel JA, Gibbons S, Daisley BA, Parvathy SN, et al. Dissecting mechanisms of fecal microbiota transplantation efficacy in disease. Trends Mol Med. 2024;30: 209–222.

29. Green JE, Davis JA, Berk M, Hair C, Loughman A, Castle D, et al. 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. 2020;12: 1–25.

30. Davar D, Dzutsev AK, McCulloch JA, Rodrigues RR, Chauvin J-M, Morrison RM, et al. Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. Science. 2021;371: 595–602.

31. Baruch EN, Youngster I, Ben-Betzalel G, Ortenberg R, Lahat A, Katz L, et al. Fecal

microbiota transplant promotes response in immunotherapy-refractory melanoma patients. Science. 2021;371: 602–609.

32. Routy B, Lenehan JG, Miller WH Jr, Jamal R, Messaoudene M, Daisley BA, et al. Fecal microbiota transplantation plus anti-PD-1 immunotherapy in advanced melanoma: a phase I trial. Nat Med. 2023;29: 2121–2132.

33. Porcari S, Ciccarese C, Heidrich V, Rondinella D, Quaranta G, Severino A, et al. Fecal microbiota transplantation plus pembrolizumab and axitinib in metastatic renal cell carcinoma: the randomized phase 2 TACITO trial. Nat Med. 2026. doi:10.1038/s41591-025-04189-2

34. Duttagupta S, Messaoudene M, Hunter S, Desilets A, Jamal R, Mihalcioiu C, et al. Fecal microbiota transplantation plus immunotherapy in non-small cell lung cancer and melanoma: the phase 2 FMT-LUMINate trial. Nat Med. 2026. doi:10.1038/s41591-025-04186-5

35. Levitan O, Ma L, Giovannelli D, Burleson DB, McCaffrey P, Vala A, et al. The gut microbiome-Does stool represent right? Heliyon. 2023;9: e13602.

36. Ahn J-S, Lkhagva E, Jung S, Kim H-J, Chung H-J, Hong S-T. Fecal microbiome does not represent whole gut microbiome. Cell Microbiol. 2023;2023: 1–14.

37. Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 2011;166: 99–110.

38. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10: 5029.

39. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35: 833–844.

40. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464: 59–65.

41. Beghini F, Pasolli E, Truong TD, Putignani L, Cacciò SM, Segata N. Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. ISME J. 2017;11: 2848–2863.

42. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6: 960–970.

43. Krinos AI, Mars Brisbin M, Hu SK, Cohen NR, Rynearson TA, Follows MJ, et al. Missing microbial eukaryotes and misleading meta-omic conclusions. Nat Commun. 2024;15: 9873.

44. Hall RJ, Wang J, Todd AK, Bissielo AB, Yen S, Strydom H, et al. Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. J Virol Methods. 2014;195: 194–204.

45. Garushyants SK, Sane M, Selifanova MV, Agashe D, Bazykin GA, Gelfand MS. Mutational signatures in wild type Escherichia coli strains reveal predominance of DNA polymerase

errors. Genome Biol Evol. 2024;16. doi:10.1093/gbe/evae035

46. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. 2005;3: 711–721.

47. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol. 2020;18: 491–506.

48. Doolittle WF. Population genomics: how bacterial species form and why they don't exist. Curr Biol. 2012;22: R451–3.

49. Vasquez KS, Wong DPGH, Pedro MF, Yu FB, Jain S, Meng X, et al. High-resolution lineage tracking of within-host evolution and strain transmission in a human gut symbiont across ecological scales. bioRxiv. 2024. doi:10.1101/2024.02.17.580834

50. Wayne LG, Moore WEC, Stackebrandt E, Kandler O, Colwell RR, Krichevsky MI, et al. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Evol Microbiol. 1987;37: 463–464.

51. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 2005;102: 2567–2572.

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

53. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. 2014;64: 346–351.

54. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell. 2019;176: 649–662.e20.

55. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021;39: 105–114.

56. Schmidt TSB, Fullam A, Ferretti P, Orakov A, Maistrenko OM, Ruscheweyh H-J, et al. SPIRE: A Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Res. 2024;52: D777–D783.

57. Dmitrijeva M, Ruscheweyh H-J, Feer L, Li K, Miravet-Verde S, Sintsova A, et al. The mOTUs online database provides web-accessible genomic context to taxonomic profiling of microbial communities. Nucleic Acids Res. 2025;53: D797–D805.

58. Sun Y, Chen Q, Fan G, Sun Q, Zhou Q, Zhang J, et al. gcMeta 2025: a global repository of metagenome-assembled genomes enabling cross-ecosystem microbial discovery and function research. Nucleic Acids Res. 2025. doi:10.1093/nar/gkaf1115

59. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17: 132.

60. Shaw J, Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat Methods. 2023;20: 1661–1665.

61. Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, et al. Metagenome analysis using the Kraken software suite. Nat Protoc. 2022;17: 2815–2839.

62. Ruscheweyh H-J, Milanese A, Paoli L, Karcher N, Clayssen Q, Keller MI, et al. Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. Microbiome. 2022;10: 212.

63. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol. 2023;41: 1633–1644.

64. Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nat Microbiol. 2019;4: 470–479.

65. Hazen TH, Donnenberg MS, Panchalingam S, Antonio M, Hossain A, Mandomando I, et al. Genomic diversity of EPEC associated with clinical presentations of differing severity. Nat Microbiol. 2016;1: 15014.

66. Andreu-Sánchez S, Blanco-Míguez A, Wang D, Golzato D, Manghi P, Heidrich V, et al. Global genetic diversity of human gut microbiome species is related to geographic location and host health. Cell. 2025;188: 3942–3959.e9.

67. Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLoS Biol. 2019;17: e3000102.

68. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021;39: 727–736.

69. Podlesny D, Arze C, Dörner E, Verma S, Dutta S, Walter J, et al. Metagenomic strain detection with SameStr: identification of a persisting core gut microbiota transferable by fecal transplantation. Microbiome. 2022;10: 53.

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

71. Chen-Liaw A, Aggarwala V, Mogno I, Haifer C, Li Z, Eggers J, et al. Gut microbiota strain richness is species specific and affects engraftment. Nature. 2025;637: 422–429.

72. Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature. 2018;562: 583–588.

73. Dubois L, Valles-Colomer M, Ponsero A, Helve O, Andersson S, Kolho K-L, et al. Paternal and induced gut microbiota seeding complement mother-to-infant transmission. Cell Host Microbe. 2024;32: 1011–1024.e4.

74. Yang B, Ding M, Chen Y, Han F, Yang C, Zhao J, et al. Development of gut microbiota and bifidobacterial communities of neonates in the first 6 weeks and their inheritance from

mother. Gut Microbes. 2021;13: 1–13.

75. Ianiro G, Punčochář M, Karcher N, Porcari S, Armanini F, Asnicar F, et al. Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. Nat Med. 2022;28: 1913–1923.

76. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods. 2023;20: 1203–1212.

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

78. Blanco-Míguez A, Gálvez EJC, Pasolli E, De Filippis F, Amend L, Huang KD, et al. Extension of the Segatella copri complex to 13 species with distinct large extrachromosomal elements and associations with host conditions. Cell Host Microbe. 2023;31: 1804–1819.e9.

79. Manghi P, Blanco-Míguez A, Manara S, NabiNejad A, Cumbo F, Beghini F, et al. MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. Cell Rep. 2023;42: 112464.

80. Ricci L, Selma-Royo M, Golzato D, Servais C, Nabinejad A, Marchi P, et al. Description of Catenibacterium mitsuokai subsp. tridentinum subsp. nov., an anaerobic bacterium isolated from human faeces, and emended description of C. mitsuokai. Int J Syst Evol Microbiol. 2025;75. doi:10.1099/ijsem.0.006798

81. Valles-Colomer M, Manghi P, Cumbo F, Masetti G, Armanini F, Asnicar F, et al. Neuroblastoma is associated with alterations in gut microbiome composition subsequent to maternal microbial seeding. EBioMedicine. 2024;99: 104917.

82. Labisa-Morais F, Valente D, Blanco-Míguez A, Manghi P, Garcia-Valiente A, Andriamaniraka H, et al. Bacterial transmission within social groups shapes the underexplored gut microbiome in the lemur Indri indri. ISME J. 2025. doi:10.1093/ismejo/wraf136

83. Weimann A, Mooren K, Frank J, Pope PB, Bremges A, McHardy AC. From genomes to phenotypes: Traitar, the microbial trait analyzer. mSystems. 2016;1. doi:10.1128/mSystems.00101-16

84. Koblitz J, Reimer LC, Pukall R, Overmann J. Predicting bacterial phenotypic traits through improved machine learning using high-quality, curated datasets. Commun Biol. 2025;8: 897.

85. Zhang Y, Bhosle A, Bae S, Eckenrode K, Huang X, Tang J, et al. Predicting functions of uncharacterized gene products from microbial communities. Nat Biotechnol. 2025. doi:10.1038/s41587-025-02813-7

86. Nishimura Y, Yoshizawa S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. Sci Data. 2022;9: 305.

87. Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, et al. Soil microbial dark matter explored from genome-resolved metagenomics. Research Square. 2023. doi:10.21203/rs.3.rs-2680397/v1

88. Murray CS, Gao Y, Wu M. Re-evaluating the evidence for a universal genetic boundary among microbial species. Nat Commun. 2021;12: 4059.

89. Da Silva Morais E, Grimaud GM, Warda A, Stanton C, Ross P. Genome plasticity shapes the ecology and evolution of Phocaeicola dorei and Phocaeicola vulgatus. Sci Rep. 2024;14: 10109.

90. Kalizang'oma A, Richard D, Kwambana-Adams B, Coelho J, Broughton K, Pichon B, et al. Population genomics of Streptococcus mitis in UK and Ireland bloodstream infection and infective endocarditis cases. Nat Commun. 2024;15: 7812.

91. Jensen A, Scholz CFP, Kilian M. Re-evaluation of the taxonomy of the Mitis group of the genus Streptococcus based on whole genome phylogenetic analyses, and proposed reclassification of Streptococcus dentisani as Streptococcus oralis subsp. dentisani comb. nov., Streptococcus tigurinus as Streptococcus oralis subsp. tigurinus comb. nov., and Streptococcus oligofermentans as a later synonym of Streptococcus cristatus. Int J Syst Evol Microbiol. 2016;66: 4803–4820.

92. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015;43: 6761–6771.

93. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. 2011;12: 119.

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

95. Pulido-Tamayo S, Sánchez-Rodríguez A, Swings T, Van den Bergh B, Dubey A, Steenackers H, et al. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. Nucleic Acids Res. 2015;43: e105.

96. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. 2017;18: 181.

97. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host Microbe. 2018;23: 229–240.e5.

98. Wang S, Jiang Y, Li S. PStrain: an iterative microbial strains profiling algorithm for shotgun metagenomic sequencing data. Bioinformatics. 2021;36: 5499–5506.

99. Li X, Hu H, Li X. mixtureS: a novel tool for bacterial strain genome reconstruction from reads. Bioinformatics. 2021;37: 575–577.

100. Smith BJ, Li X, Shi ZJ, Abate A, Pollard KS. Scalable microbial strain inference in metagenomic data using StrainFacts. Front Bioinform. 2022;2: 867386.

101. Kang X, Luo X, Schönhuth A. StrainXpress: strain aware metagenome assembly from

short reads. Nucleic Acids Res. 2022;50: e101.

102. Vicedomini R, Quince C, Darling AE, Chikhi R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. Nat Commun. 2021;12: 4485.

103. Kazantseva E, Donmez A, Frolova M, Pop M, Kolmogorov M. Strainy: phasing and assembly of strain haplotypes from long-read metagenome sequencing. Nat Methods. 2024;21: 2034–2043.

104. Madhu B, Miller BM, Levy M. Single-cell analysis and spatial resolution of the gut microbiome. Front Cell Infect Microbiol. 2023;13: 1271092.

105. Ramiro RS, Durão P, Bank C, Gordo I. Low mutational load and high mutation rate variation in gut commensal bacteria. PLoS Biol. 2020;18: e3000617.

106. Wei W, Ho W-C, Behringer MG, Miller SF, Bcharah G, Lynch M. Rapid evolution of mutation rate and spectrum in response to environmental and population-genetic challenges. Nat Commun. 2022;13: 4752.

107. Wolff R, Garud NR. Gene-specific selective sweeps are pervasive across human gut microbiomes. Nature. 2026;650: 710–717.

108. Liu Z, Good BH. Dynamics of bacterial recombination in the human gut microbiome. PLoS Biol. 2024;22: e3002472.

109. Manghi P, Antonello G, Schiffer L, Golzato D, Wokaty A, Beghini F, et al. Meta-analysis of 22,710 human microbiome metagenomes defines an oral-to-gut microbial enrichment score and associations with host health and disease. Nat Commun. 2025;17: 196.