Gut microbiota strain richness is species specific and affects engraftment

Chen-Liaw, A.; Aggarwala, V.; Mogno, I.; Haifer, C.; Li, Z.H.; Eggers, J.; ... ; Faith, J.J.

Article

Gut microbiota strain richness is species specific and affects engraftment

https://doi.org/10.1038/s41586-024-08242-x

Received: 14 November 2022

Accepted: 17 October 2024

Published online: 27 November 2024

Check for updates

Alice Chen-Liaw[1], Varun Aggarwala[1,2,3], Ilaria Mogno[1,2], Craig Haifer[4,5], Zhihua Li[1,2], Joseph Eggers[2], Drew Helmus[6], Amy Hart[7], Jan Wehkamp[7], Esi S. N. Lamousé-Smith[7], Robert L. Kerby[8], Federico E. Rey[8], Jean Frédéric Colombel[6], Michael A. Kamm[9], Bernat Olle[10], Jason M. Norman[10], Rajita Menon[10], Andrea R. Watson[10], Emily Crossette[10], Elisabeth M. Terveer[11,12], Josbert J. Keller[11,13,14], Thomas J. Borody[15], Ari Grinspan[6], Sudarshan Paramsothy[4,16], Nadeem O. Kaakoush[17], Marla C. Dubinsky[6] & Jeremiah J. Faith[1,2]✉

Despite the fundamental role of bacterial strain variation in gut microbiota function[1–6], the number of unique strains of a species that can stably colonize the human intestine is still unknown for almost all species. Here we determine the strain richness (SR) of common gut species using thousands of sequenced bacterial isolates with paired metagenomes. We show that SR varies across species, is transferable by faecal microbiota transplantation, and is uniquely low in the gut compared with soil and lake environments. Active therapeutic administration of supraphysiologic numbers of strains per species increases recipient SR, which then converges back to the population average after dosing is ceased. Stratifying engraftment outcomes by high or low SR shows that SR predicts microbial addition or replacement in faecal transplants. Together, these results indicate that properties of the gut ecosystem govern the number of strains of each species colonizing the gut and thereby influence strain addition and replacement in faecal microbiota transplantation and defined live biotherapeutic products.

Strain-level variation in gut microbiome composition shapes the influence of the microbiota on the host and host health[1–6]. A range of tools have enabled initial explorations of strain-level gut microbiota structure, transmission and sharing[7–22]. Metagenomic analyses have demonstrated that human gut species within an individual can vary in their degree of polymorphism, indicating that different species may have variable numbers of strains in the gut[12,15,23]. Despite these advances, we still lack insight into the fundamental organization of strains in the human gut microbiota, including how many unique strains of a species can stably colonize the gut (that is, the strain richness (SR) of a species).

Compared with more open ecosystems, like the ocean or soil, the gut ecosystem represents a unique semi-closed ecological niche that maintains constant temperature, diverse nutrients and a semicontinuous unidirectional flow of nutrients. These physical parameters can influence the niche availability for different strains. Other salient factors include the frequent use of antibiotics in clinical care that could reduce strain diversity in susceptible species[24], the variation in pangenome size across species leading to variable accessory genomes sizes, host health status and cultural factors that might limit the consumption of sufficiently diverse sources of gut microbes to maximally fill gut niches[25–28]. Considering these limitations, defining the SR of human gut species will provide a new framework for exploring, understanding and predicting the interactions of donor and recipient microbial strains in faecal microbiota transplants (FMT). Among the bacterial species commonly found in the human colon, *Bacteroides fragilis* is the only species whose strain-level structure has been well studied. It is documented that *B. fragilis* almost always maintains a single unique strain per gut microbiota[7,28–30]. In the stomach, *Helicobacter pylori* is known to maintain a single unique strain per individual[31]. For two common skin bacterial species, *Staphylococcus epidermidis* has also been shown to maintain a higher SR than *Staphylococcus aureus*[32].

To quantify the SR of common species in the gut microbiota, we performed high-throughput culture and metagenomics sequencing. Using more than 7,000 sequenced gut bacterial isolates from more than 90 common gut species, we quantified the average SR across species in the human gut, as well as in soil and lake environments. We explored ecological and functional properties underpinning high and low SR. We studied the transmission of SR in recurrent *Clostridioides difficile* infection (rCDI) FMT recipients[8]. We then explored the impact of supraphysiologic administration of SR using pooled FMT comprising three to seven donor stools in a second cohort of FMT

[1]Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Icahn Genomics Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Reliance Foundation Institution of Education and Research, Jio Institute, Navi Mumbai, India. [4]Concord Clinical School, University of Sydney, Sydney, New South Wales, Australia. [5]School of Clinical Medicine, University of New South Wales, Sydney, New South Wales, Australia. [6]Dr. Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Janssen R&D, Spring House, PA, USA. [8]Department of Bacteriology, University of Wisconsin—Madison, Madison, WI, USA. [9]Department of Gastroenterology and Medicine, St. Vincent's Hospital, Melbourne, Victoria, Australia. [10]Vedanta Biosciences, Cambridge, MA, USA. [11]Netherlands Donor Feces Bank, Leiden University Medical Center, Leiden, The Netherlands. [12]Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands. [13]Department of Gastroenterology, Haaglanden Medical Center, The Hague, The Netherlands. [14]Department of Gastroenterology and Hepatology, Leiden University Medical Center, Leiden, The Netherlands. [15]Centre for Digestive Diseases, Sydney, New South Wales, Australia. [16]Department of Gastroenterology and Hepatology, Macquarie University Hospital, Sydney, New South Wales, Australia. [17]School of Medical Sciences, University of New South Wales, Sydney, New South Wales, Australia. ✉e-mail: jeremiah.faith@mssm.edu

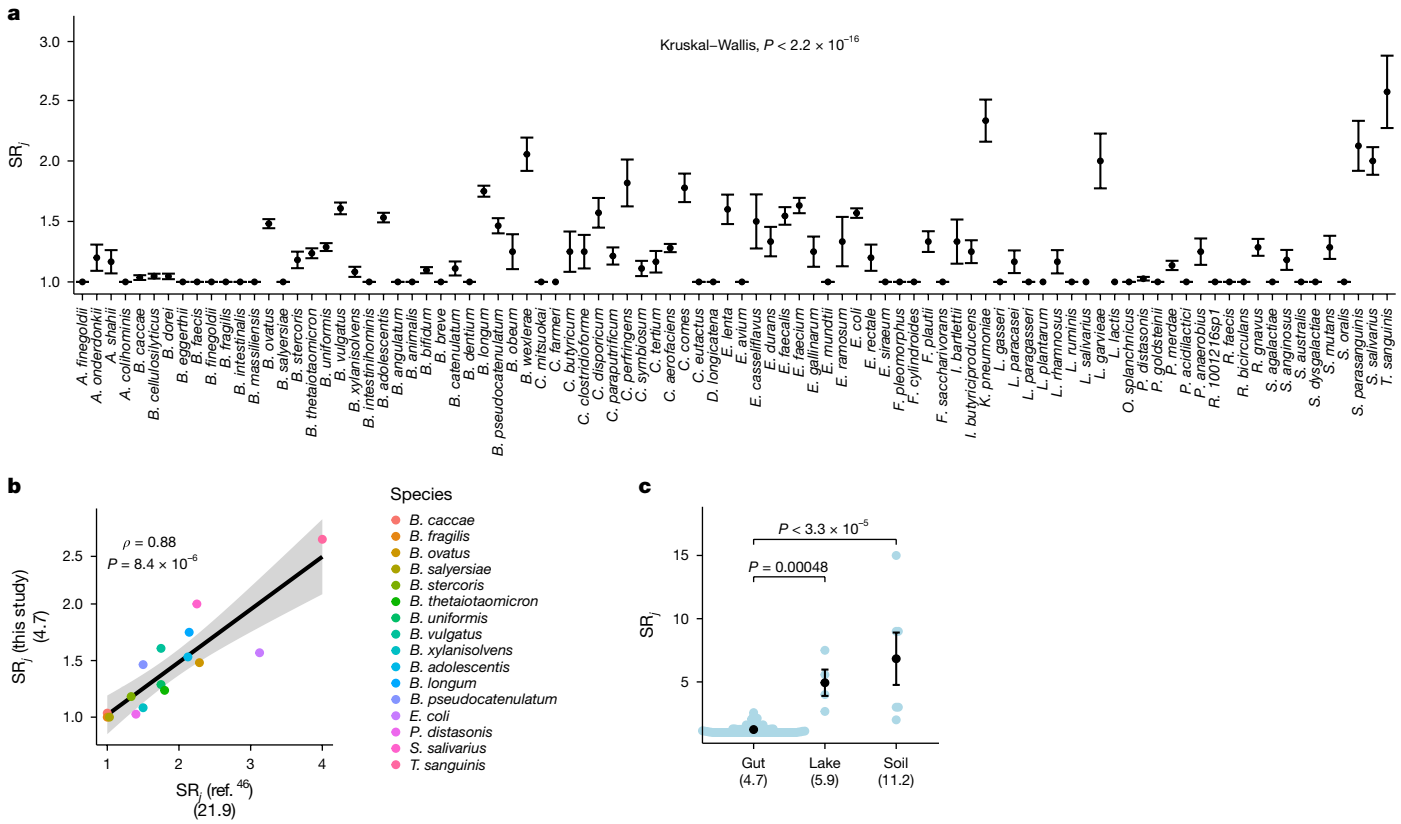


Fig. 1 | SR_j of 92 human gut species. a, SR_j varies by species (Kruskal–Wallis, $P < 2.2 \times 10^{-16}$). Data are represented as mean value ± s.e.m. **b**, SR_j using the isolate genome set in this study ($n = 4{,}773$) is highly correlated with SR_j estimated from an independent set of bacterial genomes ($n = 1{,}947$) isolated from 11 humans[46]. Spearman rank correlation was applied. Grey area, 95% confidence interval. **c**, The SR_j of human gut species is lower than SR_j of species isolated from lake and soil microbiomes (Kruskal–Wallis test). Blue points, average SR_j for a species found in each of the environments; black points, mean of the environment; error bars, s.e.m.

recipients with ulcerative colitis (UC)[33]. We explored how SR capacity predicts whether a donor strain engrafts and replaces a recipient strain, coexists with a recipient strain or fails to engraft while the recipient strain remains in the microbiota. Our insights highlight the potential requirement for maintenance doses to maintain a high SR and provide insights for understanding engraftment outcomes on the basis of the ecological constraints imposed by the average SR of each species.

Average SR of gut species

To estimate the SR of gut bacterial species in the human gut, we sequenced 5,113 gut bacteria cultured as pure isolates from the stool of 100 people (54 healthy, 37 with inflammatory bowel disease (IBD), and 9 with rCDI; Supplementary Tables 1 and 2). We focused on 92 commensal gut bacterial species that were isolated from at least three different people. Pairwise genome distances from isolates of the same species (conspecific) cultured from two unrelated people typically have a fractional *k*-mer distance of at least 0.04, whereas the *k*-mer distances between conspecific isolates from the same person are almost always far less than 0.04 (Extended Data Fig. 1a). As previously described[34–36], these distances are the result of most isolates of a given species in an individual person being replicates of the same strain that is reisolated many times, while the same strain is very rarely shared between unrelated people[34]. Therefore, we count the number of unique strains of a species (that is, SR) within a person *i* for a given species *j* (SR_{ij}) as the number of genomes with a *k*-mer distance of at least 0.04, corresponding to an average nucleotide identity (ANI)[37] of approximately 99.5%, which was identified independently as a genomic-based strain marker[38] (Extended Data Fig. 1b). To calculate the average SR for a species *j* across our study population (SR_j), we average SR_{ij} across all people in our cohort that harbour species *j* within their microbiota. Isolates from the rCDI participants were excluded from these initial analyses given the established observation that the rCDI microbiota are iatrogenically depleted in terms of diversity and microbiota density[39–41].

Across 92 species in 91 people, we observed that SR_j in the human gut microbiome varies significantly by species ($P = 2.2 \times 10^{-16}$, Kruskal–Wallis; Dunn's Test; Extended Data Table 2), is less than 2.0 strains per species on average (1.23 ± 0.33; mean ± s.d.), and below 3.0 for all species (Fig. 1a and Extended Data Table 1). SR_j of the commensal species *B. fragilis* was 1.0, which aligned with expectations from previous findings[7,29,30]. As an initial validation, we estimated *B. fragilis* SR_{ij} and SR_j using the same methodology on a deeply sampled *B. fragilis* dataset[7] with 524 sequenced bacterial isolates from 12 people. We find identical results with $SR_j = 1.0$ in the validation cohort (Extended Data Fig. 1c). For our initial dataset, we sequenced, on average, 3.55 genomes per species per person. To determine whether this limited sampling depth results in a large underestimation of SR_j, we sampled deeply across a subset of ten bacterial species from four main gut phyla whose estimated SR_j ranged between 1.00 and 2.00 (Extended Data Fig. 1d–f and Extended Data Table 3). In addition to performing deeper culturing of the original donor sample, we also transferred some of the original human donor stools to ex-germ-free mice consuming various diets (high-casein, high psyllium or mixed carbohydrate), since previous studies have demonstrated that different dietary conditions and the mouse gut environment can enrich for different strains[42–45]. After a 149% increase in sampling for these organisms, we observed only 41 new strains across a further 778 isolates (7.2% increase in unique strains), indicating that we

were near saturation of the SR_j for these species. As expected, the new strains belonged to species with higher SR_j (Extended Data Fig. 1d–h). In addition to this deeper sampling for a subset of species, we also generated rarefaction curves and found our average sampling depth per species per person was typically beyond the point of steepest increase in the rarefaction (Extended Data Fig. 1i–l). To test the generalizability of these results in a second validation cohort beyond *B. fragilis*, we calculated SR_j for other gut species from a previous study[46], where 1,947 gut bacterial isolates were cultured from 11 healthy people[46] (21.9 mean genomes per species per person). We observed a highly significant correlation between SR_j across the two datasets ($P = 8.4 \times 10^{-6}$, Spearman correlation; Fig. 1b) providing an independent validation of the SR_j estimates across a broad taxonomic range.

To determine how the SR_j structure of the human gut compares with that of other ecosystems, we compared the SR_j of the human gut microbiome with that of soil and lake microbiomes. Species in the human gut demonstrated much lower SR_j (1.23 ± 0.33) than species sampled from soil (6.83 ± 5.07; 112 sequenced isolates) or lake (4.93 ± 2.08; 94 sequenced isolates; Fig. 1c). In addition, in contrast to the rarefaction curves for gut species, we were unable to approach SR_j saturation for lake or soil species even when we sampled 20–30 isolates of the same species (Extended Data Fig. 1m,n). This startling difference indicates that unique features of the gut, such as its semi-closed ecosystem with unidirectional flow, stable temperature and fast microbial growth rates, may limit SR_j.

Factors influencing SR

Although the broad characteristics of the gut microbial ecosystem seem to limit species SR compared with more open environments such as lakes and soil (Fig. 1c), the highly significant differences in SR_j across different species in the gut indicate factors that might influence a given species' ability to maintain more than one strain in a host. For instance, colonization with several strains that are frequent in humans could increase the chance that at least one of these strains is transmitted to a new person. Therefore, we compared SR_j with the prevalence of each species *j* in the cohort (Fig. 2a and Supplementary Table 3). We observe a highly significant correlation between species frequency and average SR_j ($P = 1.2 \times 10^{-5}$, Spearman correlation). However, there are numerous exceptions with low frequency and high SR_j, as well as high frequency and low SR_j, including the frequent and known example of *B. fragilis* that has anti-microbial mechanisms to limit SR_j to 1.0. The low frequency (less than 5%), high SR_j (greater than 1.2) species were largely from Proteobacteria ($N = 1$) and Firmicutes ($N = 9$), while the high frequency (more than 20%), low SR_j species were from Bacteroidetes ($N = 5$) and Actinobacteria ($N = 1$) with no significantly enriched phyla. The large number of Bacteroidetes among the high frequency (more than 20%), low SR_j species indicates that other members of this phyla also maintain mechanisms to limit SR_j, similar to *B. fragilis*, although *B. fragilis* is notably the only species in this group with $SR_j = 1.0$. Many of these outliers probably have unique mechanisms to maintain high or low SR_j.

From a comparative genomics standpoint, species that have, on average, more accessory genome *k*-mers and a smaller core genome would be expected to have more genetic flexibility to find suitable niches that do not overlap with conspecific strains in the same host. When comparing the mean pairwise distances between species with $SR_j > 1.2$ versus for species with $SR_j < 1.1$, we found significantly greater pairwise distances in species with $SR_j > 1.2$ compared with species with $SR_j < 1.1$ (0.35 ± 0.001 versus 0.30 ± 0.007; $P = 2.2 \times 10^{-16}$, Kruskal–Wallis), indicating that larger accessory genomes in species with $SR_j > 1.2$ might enable greater SR_j within individual microbiotas (Fig. 2b). Differences in genome size and the enormous evolutionary distances across bacteria from different phyla in the gut make a direct gene-based comparison across all taxa impractical. However, for five genera for

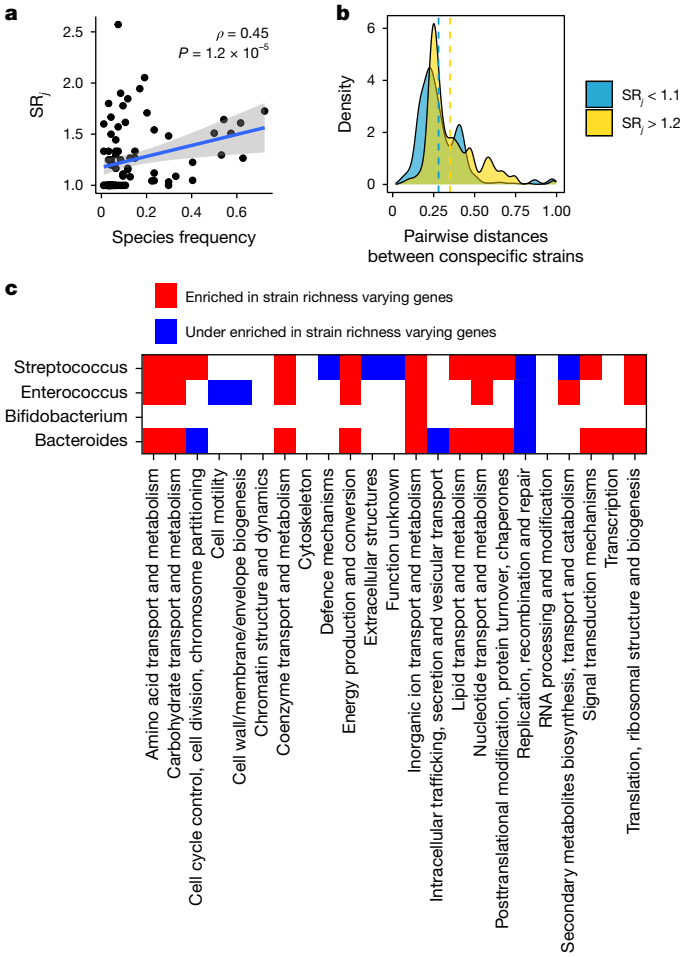


Fig. 2 | Factors influencing SR_j. a, Spearman rank correlation of SR_j versus species frequency in our cohort. No corrections were used. Grey area, 95% confidence interval. **b**, Pairwise *k*-mer distances between unique conspecific strains for species with $SR_j < 1.1$ and $SR_j > 1.2$. **c**, Functional categories over- and under-represented in genes that vary between species with $SR_j < 1.1$ versus $SR_j > 1.2$.

which we had at least six species with quantified SR_j, we annotated the genomes of 1,015 strains from 45 species with prokka[47] and used roary[48] to quantify the fraction of the genome represented by core genes. Four out of the five genera (*Bacteroides*, *Bifidobacterium*, *Clostridium* and *Enterococcus*) had the same trend of lower SR_j being associated with a larger core genome (Extended Data Fig. 2a), although only the association between SR_j and Bifidobacterium was significant ($P = 0.03$). Analogous to our results in the intestine, in the stomach, *H. pylori* has an SR_j of 1.0 and has a very large core genome[31]. We used the scoary algorithm[49] to identify genes enriched in species with $SR_j > 1.2$ versus for species with $SR_j < 1.1$. Genes were functionally annotated with eggNOG[50] and functional enrichment was determined by Fisher's exact test (Supplementary Table 4). Replication, recombination and repair was the only functional category universally under-enriched—probably representing an artefact from the representation of these genes in the core genome relative to the accessory genome. In contrast, numerous functions related to metabolism were over-enriched (Fig. 2c), which provides further evidence that the increased accessory gene content associated with increased SR_j is driving the metabolic diversity that potentially enables niche differentiation between several strains for SR_j high species.

Finally, the health status of the host might influence SR_j. For example in IBD, UC and Crohn's disease (CD), the gut microbiome has reduced genus and species diversity compared with healthy controls[51–54],

indicating that host health status might influence the average SR for a species, SR_j. To determine whether SR_j is also altered in IBD, we compared the average SR_j of species that were isolated from both healthy and IBD microbiotas. We found that UC ($N = 16$) and CD ($N = 21$) human gut microbiomes do not differ in SR_j when compared with healthy ($N = 54$) microbiomes, at least for the number of participants assessed in this study (Extended Data Fig. 2b,c paired Wilcox test for species comparisons). These results indicate that the reduction in species and genus diversity in IBD does not have a corresponding drop in strain diversity.

Overall, these results indicate that several factors, including species prevalence in the human population, core/accessory genome size and metabolic use diversity, influence species richness. Studying the outlier species to these trends in future in vitro and in vivo studies could yield new insights into mechanisms driving SR_j.

Transmissibility and stability

Previously, we demonstrated that most donor microbiota strains stably engraft post-FMT in patients rCDI[8], indicating that donor SR_j might be transmitted by means of FMT. We isolated and sequenced 1,008 unique strains from 7 FMT donors and 13 rCDI patients. We used the *Strainer* strain detection algorithm to quantify the presence of donor strains in the recipients using the donor strain genomes and recipient metagenomes[8]. The SR_j estimated from FMT donor metagenomics data was correlated significantly with the SR_j that we measured from cultured isolates in our complete cohort in Fig. 1 (Extended Data Fig. 3a). This allowed us to track the transmission of donor strain-level community structure.

In this US-based FMT study, single donors were used for each recipient, with six donors giving stool to separate individual rCDI recipients and one donor (D283) providing stool to seven different rCDI recipients for a total of 13 1:1 donor–recipient pairs (Fig. 3a)[55]. We quantified the number of donor strains per species found in the donors and the number of donor strains that were subsequently detected in recipients 8 weeks post-transplant (Fig. 3b and Extended Data Fig. 3c). Donor SR_j and the recipient SR_j at week 8 post-FMT were significantly correlated ($P = 2.2 \times 10^{-5}$; Spearman's rank correlation; Fig. 3c). Average SR_j in recipients at week 8 post-FMT showed very little loss of donor strains and resembled the cultured SR_j we initially measured in our cohort (Extended Data Fig. 3b). Together, these results demonstrate that FMT can transmit a healthy donor strain-level community structure. To determine whether transferability of SR_j is consistent across cohorts, we tracked donor strains in six more rCDI recipients from two donors from the Leiden University Medical Center, the Netherlands[56]. In this validation cohort, we also find donor SR_j is correlated with recipient SR_j ($P = 0.014$; Spearman's rank correlation; Extended Data Fig. 3d).

A subset of donors ($n = 3$) and recipients ($n = 3$) from the US-based FMT cohort had samples collected 5-years post-transplant, which allowed us to measure the stability of SR in each species over time. In the donors, we found a significant correlation between donor SR_j at time point 0 and 5 years later (Fig. 3d), indicating long-term stability of SR_j in healthy, untransplanted people up to 5 years. In the recipients, we also find significant correlations between recipient SR_j at 8 weeks post-FMT and 5-years post-FMT (Fig. 3e), indicating that FMT can durably restore a healthy strain-level structure to recipients. However, these 5-year SR_j stability results are based on a limited number of people and would benefit from validation in larger cohorts.

Supraphysiologic SR in FMT

Both the analyses of unperturbed gut microbiota and of the transmission of unperturbed gut microbiota indicate that SR_j ranges from approximately one to three strains per species and can be transmitted durably through FMT. Whereas recent studies have examined the transmission of strains by means of FMT, none have quantified SR_j

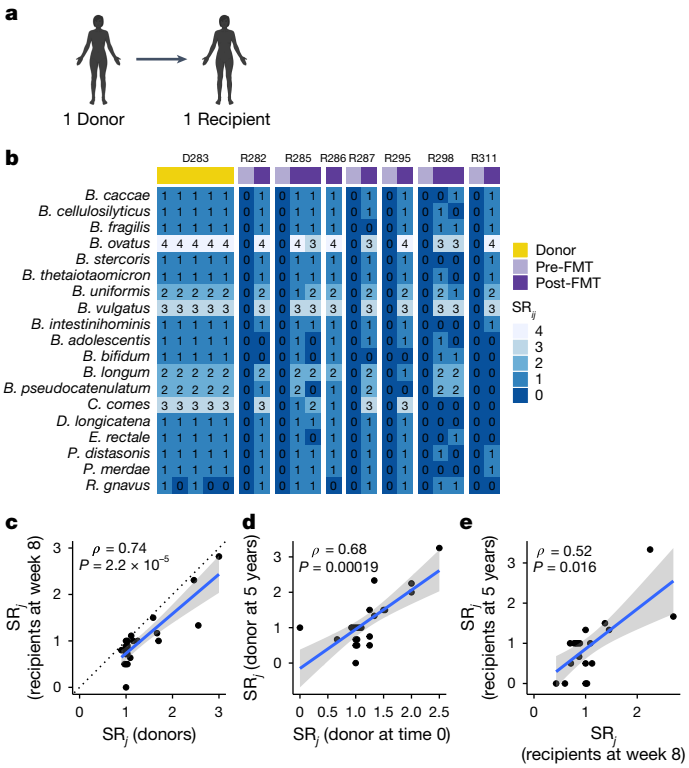


Fig. 3 | FMT durably transmits healthy donor SR_j to rCDI patients. **a**, Schematic of FMT experimental design with 1:1 donor–recipient pairs. **b**, Representative heatmap showing the transmission of SR_{ij} from donor D283 to seven different recipients (R282, R285 and so on). The numbers in the cells of the heatmap indicate the number of strains of the given species detected in the donor or recipient across five time points for donor D283 and up to three time points for the seven recipients. **c**, Spearman rank correlation between recipient SR_j at week 8 post-FMT and donor SR_j. **d**, Spearman rank correlation between donor SR_j at 5 years and time 0 (pre-FMT stool sample). **e**, Spearman rank correlation between recipient SR_j at 5 years post-FMT and 8 weeks post-FMT. **c**–**e**, Each datapoint represents the average SR_j for a tracked species in **b**; grey area, 95% confidence interval, and Spearman rank correlations are two-tailed. Credit: **a**, © Jeremiah Faith/123rf.com.

(refs. 17,18,57). Thus, we still do not know SR_j elasticity, SR_j upper limits, or what role ecologic and environmental pressures play in SR_j in the gut, as factors such as our antibiotic and hygiene practices may artificially reduce contemporary SR_j below a physiologic or ecologic limit. Pooled donor FMT offers a unique opportunity to test the ecologic limit of SR_j and determine whether SR_j can be stably increased.

In the FOCUS clinical trial of FMT in participants with UC, 14 donor stools were combined into 21 different donor stool batches comprising 3–7 unique donor stools per batch (Extended Data Fig. 4a). Each multi-donor batch was given to one or more recipients initially by means of colonoscopy followed by 40 enema doses[33]. For species that are common across healthy people, the pooled FMT approach provides a unique situation in which a supraphysiologic number of strains for many species—much higher than what we found in our calculation of SR_j in untransplanted people (Fig. 1a and Extended Data Table 1)—was administered to each recipient. Thus, this is an excellent occasion to query the upper limits of SR_j and its stability in recipients over time.

Using Strainer, we quantified the number of donor strains present in the individual donors, in the donor batches, and in the recipient pre- and post-FMT time points (Fig. 4b). We find variable engraftment efficiency across the strains in each species, indicating that some strains may be more fit than others and that multi-donor FMT may lead to a more resilient and transmissible recipient microbiota (Extended Data Fig. 4b). As with FMT participants in the rCDI trial, we found that

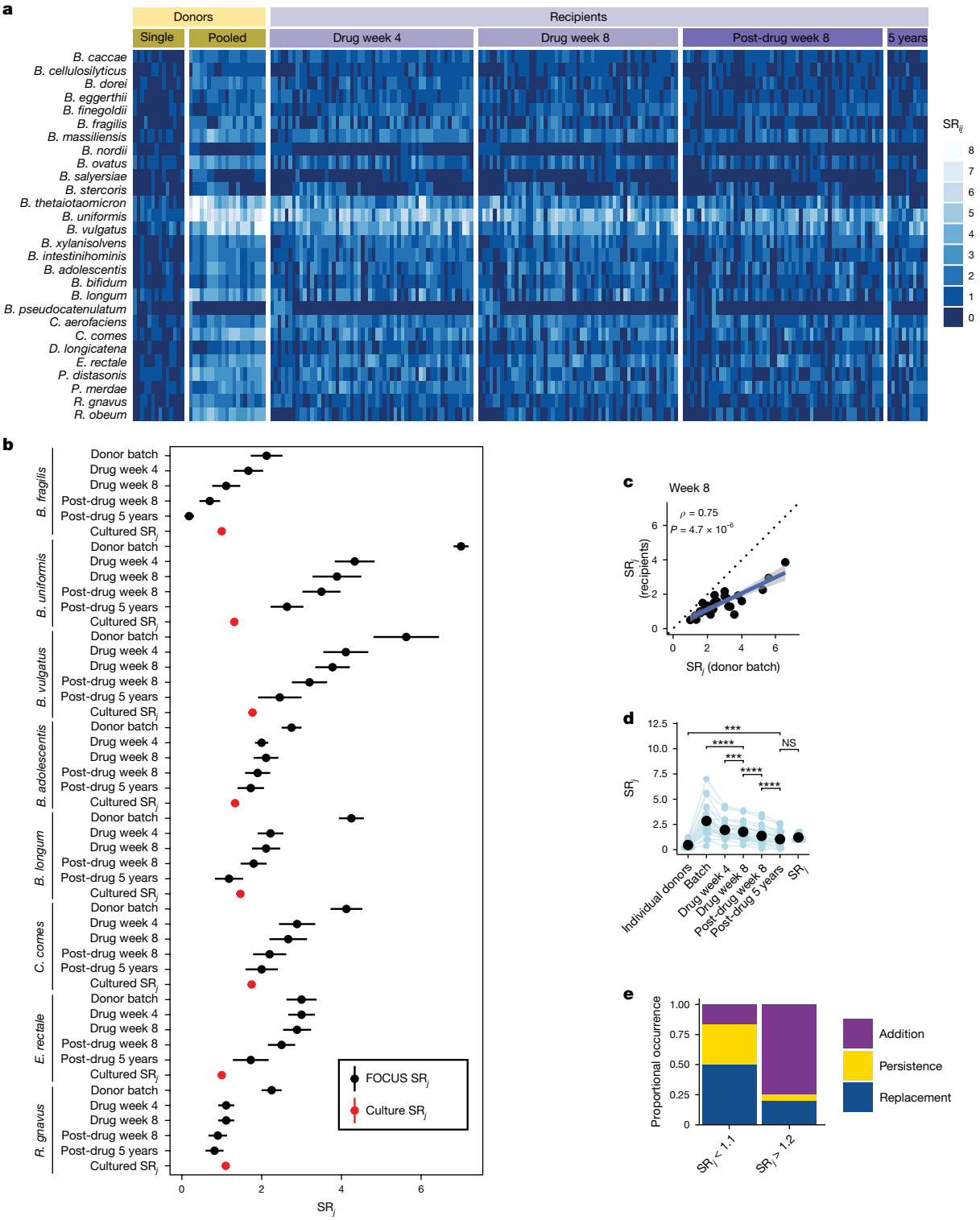


Fig. 4 | Supraphysiologic manipulation of SR_j in FMT recipients converges to the population baseline observed in untransplanted people. a, Heatmap showing the SR_{ij} of single donors, donor batches and recipients at post-FMT time points both during and after FMT drug administration. **b**, SR_j for representative species across the donor batch, recipient post-FMT time points and cultured cohort from Fig. 1a (individual donors). Data are presented as the mean value ± s.e.m. with each point representing the SR_j for a species at each time point or for the cultured cohort. **c**, Spearman rank correlation (two-tailed) between recipient SR_j at drug week 8 and donor batch SR_j. Each point represents the SR_j of a species as measured in the pooled donor batch versus as measured in the recipient post-FMT week 8. **d**, SR_j across donors, batches, recipient time points and culture (previously measured in Fig. 1a and Extended Data Table 1). Blue points, tracked SR_j of a species as measured in each donor group (individual donors versus pooled batch) or in each recipient post-FMT time point; black points, mean SR_j across the overall time point or group. Two-sided Wilcoxon tests were used to compare groups. **e**, Proportional occurrence of addition, persistence and replacement events across species with low SR_j and high SR_j. $^{***}P < 10^{-3}$, Wilcoxon test; $^{****}P < 10^{-4}$, Wilcoxon test; NS, not significant by Wilcoxon test. Exact P values: individual donors versus recipients 5 years ($P = 5.1 \times 10^{-5}$), donor batch versus recipients week 4 ($P = 6 \times 10^{-8}$), recipients week 4 versus week 8 ($P = 8.5 \times 10^{-4}$), recipients week 8 versus week 16 ($P = 1.9 \times 10^{-5}$), recipients week 16 versus recipients 5 years ($P = 1.2 \times 10^{-5}$), recipients 5 years versus cultured SR_j ($P = 0.071$).

Strainer quantification of SR_j across the individual donors (Extended Data Fig. 4c) was correlated significantly with the SR_j measured across our cultured cohort in Fig. 1a. As expected, the multi-donor pools harboured a higher SR_j than in single individual people (2.65 ± 1.27 versus 1.23 ± 0.33; Fig. 4a,b).

After 8 weeks of transplantation, we find a strong correlation between the SR_j of the donor batch and the SR_j of the recipients (Fig. 4c). However, even after an intensive regimen of 40 faecal transplants over 8 weeks, only a proportion of donor SR_j engrafted in recipients resulting in recipient SR_j that was significantly less than the theoretical maximum based on the number of detected strains in the donor batches (1.33 ± 0.75 versus 2.65 ± 1.27; $P = 4.00 \times 10^{-6}$, pairwise Wilcoxon signed-rank). These results demonstrate that even an intensive course of supraphysiologic SR_j through FMT for 40 times over the course of 8 weeks has a limited effect on the maximum SR_j in the recipients.

When we track recipient SR_j longitudinally, we observe a significant, progressive decrease in SR_j in the recipients after the last FMT dose at week 8 (Fig. 4b,d). By 5 years after the final transplant, SR_j in the recipients converged near the population-wide SR_j measured in people not receiving faecal transplants (Fig. 4b,d and Extended Data Table 4). We found this trend to be consistent across all species, irrespective of species prevalence or SR_j. Together, these results demonstrate that it is possible to therapeutically increase SR_{ij} slightly with maintenance dosing (Fig. 4a,d). However, this increase in recipient SR_j is temporary as recipient SR_j returns to the baseline SR_j that we quantified in unperturbed healthy and disease microbiotas (Fig. 4b,d). These results indicate that SR_j in an individual person is a species-specific capacity that is limited by the gut microbiota ecosystem.

The species-specific capacities in SR have implications for the expectations of strain engraftment by FMT or defined live biotherapeutic product. For species with low SR_j that are harboured in the recipient, we would expect that therapeutic strains of the same species would either engraft and replace the recipient strain (replacement) or would fail to engraft while the recipient strains persist in the recipient (persistence) (Extended Data Fig. 4d). For species with higher capacities for SR_j, we would expect that therapeutic strains would more easily engraft without replacing recipient strains (addition)—at least until the capacity for the species is reached. To test this hypothesis, we evaluated the frequency of persistence, replacement or addition in the context of the pooled donor FMT. As expected, for species with low average SR_j ($SR_j < 1.1$) therapeutic outcomes are dominated by persistence and replacement whereas outcomes for species with higher average SR_j ($SR_j > 1.2$) are significantly enriched for strain additions ($P = 0.0128$; Fisher's exact test; Fig. 4e). Our results show that SR_j is a key characteristic of gut species that underpins engraftment outcomes.

Discussion

Using high-throughput bacterial culturing and sequencing coupled with metagenomics and strain tracking, we find that the average SR of species in the healthy human gut microbiome varies by species and ranges between 1.00 and 2.57 (Fig. 1a). Previous investigations of microbiota composition in people with IBD at genus and species levels[51–53] have shown reduced diversity compared with healthy people. We find no difference in SR_j between healthy and IBD microbiotas (Extended Data Fig. 2b,c). Comparing the SR_j of the human gut microbiome with environmental microbiomes demonstrated that this low level of bacterial strain diversity is unique to the semi-closed human gut ecosystem (Fig. 1c). In both macroecology and microecology, species diversity often follows a hump-shaped, unimodal distribution where very low productivity ecosystems have low species diversity, moderate productivity ecosystems have the highest species diversity and high productivity ecosystems have lower species diversity[58,59]. Here we find a parallel trend at the strain level where moderate productivity ecosystems such as the soil and lake ecosystems have higher strain diversity than the far more productive microbial ecosystem of the gut, where the microbial biomass per gram is several orders of magnitude higher than soil or lake despite having a relatively high turnover rate from flushing of the intestinal contents. A second factor probably enabling a higher SR_j in these non-gut environments is their vastly increased potential for spatial segregation[59].

Besides ecological parameters, the microbial factors that influence how conspecific strains co-exist within the gut microbiota are still being discovered. It is known that the several species of the gut microbiota mediate colonization resistance to enteric pathogens through interspecies inhibition of growth, direct killing or production of bacteriocins[60–66]. Within a species, competition for nutrient resources becomes a main factor in colonization resistance as conspecific strains often have overlapping ecological niches[67]. For example, some less virulent *C. difficile* strains can decrease germination of enterotoxigenic *C. difficile* strains by competing for limited amino acid resources[68]. Another example is *B. fragilis*, which expresses colonization factors that inhibit colonization by new conspecific strains[29]. It is likely that both interspecies and intraspecies interactions along with other host-microbial interactions involving host immunity and codiversification apply selective pressures to certain bacterial species resulting in low richness near 1.0.

Although most of the bacterial species we measured had low richness, there were also species that had SR_j ranging from 2.0 to 3.0 (Fig. 1a). Many open questions remain about the origins and mechanisms underlying the co-existence of several strains within a species. Our investigations offer some hints to underlying mechanisms including an association between average SR in a species and the prevalence of the species (Fig. 2a). We also observe that conspecific strains from species with higher SR ($SR_j > 1.2$) have significantly greater mean pairwise genomic distances compared with strains from species with lower SR ($SR_j < 1.1$), indicating that larger accessory genomes confer a greater ability to use diverse niches and may facilitate greater SR_j within the human gut (Fig. 2b). Indeed, we find that the gene functions that vary across species with low versus high SR_j include numerous functions related to metabolism (Fig. 2c), indicating that the increased accessory gene content that is associated with increased SR_j drives greater metabolic versatility.

One further parameter that may play a role in facilitating higher SR_j is gut anatomy. One recent study on the human skin microbiome found that skin pores impose random bottlenecks as *Cutibacterium acnes* migrates into pores, reducing intraspecies competition and enabling co-existence of several *C. acnes* lineages[9]. In the gut, though the bulk of the niche volume is in the well-mixed lumen (allowing mixing and competition), intestinal crypts may promote co-colonization by several strains by partitioning a species population and reducing competition for shared resources[29,69].

Limitations to this study are that, although the gut microbiota is relatively stable, there is still a small amount of strain acquisition and loss that probably continually occurs in an individual person. The duration of these transient colonizations will vary, but our pooled donor FMT results indicate that it can take weeks to years for therapeutically inflated SR_j of a species to converge. However, in untransplanted people, a transient increase in strain count is probably limited to a few species and varies between people. Therefore, the impact of these dynamics on average SR for a species is probably minimal when averaged across more than 90 people. Although we studied SR_j in 123 people between the cohort in this study and the validation studies[7,46], SR_j measures will improve and be available for more species as more people are studied. Another limitation is that no current tool can evaluate every bacterial cell in a person's gut microbiome at the strain level. Therefore, we cannot know whether there is a large amount of strain diversity at very low abundance that is undetectable with current methods. Although our deeper sampling of a few species indicated that we had not found all strains of a species in every person, the overall impact of

deeper sampling was minimal, and our rarefaction analyses indicate that our sampling depth was close to, or beyond, the maximal rate of increase on the collectors curve. In addition, our estimates of SR_j were confirmed with two independent datasets[7,46]. Finally, there is no widely accepted definition of a bacterial strain. However, we used a genomics-based threshold that was demonstrated empirically to identify bacterial strains shared over time[34], in between family members[34] and across faecal transplants[8]. Our analyses in the set of bacterial genomes in this analysis further support a strain threshold of around 0.96 *k*-mer overlap (Extended Data Fig. 1a,b) and align with a similar empirical observation using ANI[38].

Another limitation to this study is that SR_j was tested only in stool, with no testing of SR_j for bacterial populations residing in the small bowel, those adherent to mucus or those residing in epithelial crypts. However, due to the low number of viable mucosal adherent bacterial cells, as well as the difficulty in investigating the small bowel microbiome, we used faecal samples as a proxy for the gut microbiome, as do most microbiome studies[3,7,12,27,70]. In addition, our culturing approach has inherent biases towards culturable species[71,72] as well as more common species. For example, some rarer and difficult-to-culture gut species, such as *Prevotella copri*[73–75] and *Akkermansia muciniphila*[76], have been implicated in human health. Although we do have isolates of these species in our culture collection, we were unable to collect sufficient isolates from enough people to include these species in our analyses.

Despite many potential facilitators for high strain diversity, our results demonstrate that the human gut ecosystem does not permit limitless SR_j. When pooled donor FMT is used to administer supraphysiologic SR_j to recipients, we find only a temporary increase in recipient SR_j that eventually returns to the population baseline (Fig. 4). Nevertheless, this result also indicates that FMT maintenance dosing could allow sustained higher SR_j. In one-to-one donor–recipient pairs, donor SR_j is consistently transmittable and stable in rCDI recipients (Fig. 3). Together, these findings emphasize the importance of understanding the influence of human gut anatomy and physiology on SR_j and its future translational applications. In FMT and defined live biotherapeutic products, where success depends on the successful engraftment of bacterial strains[8,57,77], careful consideration of which strains to include, how many strains of a species to include and whether resident strains should be removed is essential to future development. In line with our findings, a recent meta-analysis found that recipient resident species play an outsized role in inhibiting donor strain engraftment[18]. Although there are many forces at the individual host level that can influence the engraftment of strains, there do seem to be strains within species that seem more likely to be engrafted even when transmitted in multi-strain pools (Extended Data Fig. 4b). Together with the limited SR_j capacity of the gut, these data indicate a theoretical and practical basis for removing risk- or disease-associated strains from the gut by administration of other milder strains from the same species that are more fit and can occupy the same species niche. Likewise, these results also indicate that attempts to dose supraphysiologic SR_j to recipients would require continuous administration or will result in only a temporary increase, with a lower SR_j remaining in the long term.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08242-x.

1. Yang, C. et al. Immunoglobulin A antibody composition is sculpted to bind the self gut microbiome. *Sci. Immunol.* **7**, eabg3208 (2022).
2. Parida, S. et al. A procarcinogenic colon microbe promotes breast tumorigenesis and metastatic progression and concomitantly activates notch and β-catenin axes. *Cancer Discov.* **11**, 1138–1157 (2021).
3. Arthur, J. C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
4. Britton, G. J. et al. Defined microbiota transplant restores Th17/RORγt$^+$ regulatory T cell balance in mice colonized with inflammatory bowel disease microbiotas. *Proc Natl Acad. Sci. USA* **117**, 21536–21545 (2020).
5. Yang, C. et al. Fecal IgA levels are determined by strain-level differences in bacteroides ovatus and are modifiable by gut microbiota manipulation. *Cell Host Microbe* **27**, 467–475.e6 (2020).
6. Spindler, M. P. et al. Human gut microbiota stimulate defined innate immune responses that vary from phylum to strain. *Cell Host Microbe* **30**, 1481–1498.e5 (2022).
7. Zhao, S. et al. Adaptive evolution within gut microbiomes of healthy people. *Cell Host Microbe* **25**, 656–667.e8 (2019).
8. Aggarwala, V. et al. Precise quantification of bacterial strains after fecal microbiota transplantation delineates long-term engraftment and explains outcomes. *Nat. Microbiol.* **6**, 1309–1318 (2021).
9. Conwill, A. et al. Anatomy promotes neutral coexistence of strains in the human skin microbiome. *Cell Host Microbe* **30**, 171–182.e7 (2022).
10. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
11. Drewes, J. L. et al. Transmission and clearance of potential procarcinogenic bacteria during fecal microbiota transplantation for recurrent *Clostridioides difficile*. *JCI Insight* **4**, 130848 (2019).
12. Vatanen, T. et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat. Microbiol.* **4**, 470–479 (2019).
13. Zheng, W. et al. High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* **376**, eabm1483 (2022).
14. Gupta, S. et al. LB973 Cutaneous surgical wounds have distinct microbiomes from intact skin. *J. Invest. Dermatol.* **142**, B24 (2022).
15. 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).
16. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
17. Valles-Colomer, M. et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).
18. Schmidt, T. S. B. et al. Drivers and determinants of strain dynamics following fecal microbiota transplantation. *Nat. Med.* **28**, 1902–1912 (2022).
19. Schloss, P. D., Iverson, K. D., Petrosino, J. F. & Schloss, S. J. The dynamics of a family's gut microbiota reveal variations on a theme. *Microbiome* **2**, 25 (2014).
20. Tamburini, F. B. et al. Precision identification of diverse bloodstream pathogens in the gut microbiome. *Nat. Med.* **24**, 1809–1814 (2018).
21. Dsouza, M. et al. Colonization of the live biotherapeutic product VE303 and modulation of the microbiota and metabolites in healthy volunteers. *Cell Host Microbe* **30**, 583–598. e8 (2022).
22. Siranosian, B. A. et al. Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults. *Nat. Commun.* **13**, 586 (2022).
23. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).
24. Blaser, M. J. *Missing Microbes: How the Overuse of Antibiotics Is Fueling Our Modern Plagues* (Henry Holt and Co., 2014).
25. Ni, J., Wu, G. D., Albenberg, L. & Tomov, V. T. Gut microbiota and IBD: causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 573–584 (2017).
26. Raffals, L. E. et al. The development and initial findings of a study of a prospective adult research cohort with inflammatory bowel disease (SPARC IBD). *Inflamm. Bowel Dis.* **28**, 192–199 (2022).
27. Schnorr, S. L. et al. Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 3654 (2014).
28. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).
29. Lee, S. M. et al. Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501**, 426–431 (2013).
30. Verster, A. J. et al. The landscape of type VI secretion across human gut microbiomes reveals its role in community composition. *Cell Host Microbe* **22**, 411–419.e4 (2017).
31. Han, S. R. et al. *Helicobacter pylori*: clonal population structure and restricted transmission within families revealed by molecular typing. *J. Clin. Microbiol.* **38**, 3646–3651 (2000).
32. Saheb Kashaf, S. et al. Staphylococcal diversity in atopic dermatitis from an individual to a global scale. *Cell Host Microbe* **31**, 578–592.e6 (2023).
33. Paramsothy, S. et al. Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. *Lancet* **389**, 1218–1228 (2017).
34. Faith, J. J. et al. The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).
35. Faith, J. J., Colombel, J.-F. & Gordon, J. I. Identifying strains that contribute to complex diseases through the study of microbial inheritance. *Proc. Natl Acad. Sci. USA* **112**, 633–640 (2015).
36. Faith, J. J. et al. Strain population structure varies widely across bacterial species and predicts strain colonization in unrelated individuals. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.17.343640 (2020).
37. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
38. Rodriguez-R, L. M. et al. An ANI gap within bacterial species that advances the definitions of intra-species units. *mBio* **15**, e02696-23 (2024).
39. Seekatz, A. M., Rao, K., Santhosh, K. & Young, V. B. Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent *Clostridium difficile* infection. *Genome Med.* **8**, 47 (2016).

40. Contijoch, E. J. et al. Gut microbiota density influences host physiology and is shaped by host and microbial factors. *eLife* **8**, e40553 (2019).
41. Weingarden, A. et al. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent *Clostridium difficile* infection. *Microbiome* **3**, 10 (2015).
42. Faith, J. J., McNulty, N. P., Rey, F. E. & Gordon, J. I. Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* **333**, 101–104 (2011).
43. Bahtiyar Yilmaz, A. et al. Long-term evolution and short-term adaptation of microbiota strains and sub-strains in mice. *Cell Host Microbe* **29**, 650–663.e9 (2021).
44. McNulty, N. P. et al. Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome. *PLoS Biol.* **11**, e1001637 (2013).
45. Walker, A. W. et al. Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J* **5**, 220–230 (2011).
46. Poyet, M. et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019).
47. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
48. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
49. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* **17**, 238 (2016).
50. Hernández-Plaza, A. et al. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **51**, D389–D394 (2023).
51. Forbes, J. D. et al. A comparative study of the gut microbiota in immune-mediated inflammatory diseases—does a common dysbiosis exist? *Microbiome* **6**, 221 (2018).
52. Pascal, V. et al. A microbial signature for Crohn's disease. *Gut* **66**, 813–822 (2017).
53. Michail, S. et al. Alterations in the gut microbiome of children with severe ulcerative colitis. *Inflamm. Bowel Dis.* **18**, 1799–1808 (2012).
54. Antharam, V. C. et al. Intestinal dysbiosis and depletion of butyrogenic bacteria in *Clostridium difficile* infection and nosocomial diarrhea. *J. Clin. Microbiol.* **51**, 2884–2892 (2013).
55. Hirten, R. P. et al. Microbial engraftment and efficacy of fecal microbiota transplant for *Clostridium difficile* in patients with and without inflammatory bowel disease. *Inflamm. Bowel Dis.* **25**, 969–979 (2019).
56. Terveer, E. M. et al. Faecal microbiota transplantation for *Clostridioides difficile* infection: four years' experience of the Netherlands Donor Feces Bank. *United Eur. Gastroenterol. J.* **8**, 1236–1247 (2020).
57. 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).
58. Smith, V. H. Microbial diversity-productivity relationships in aquatic ecosystems. *FEMS Microbiol. Ecol.* **62**, 181–186 (2007).
59. Graham, J. H. & Duda, J. J. The humpbacked species richness-curve: a contingent rule for community ecology. *Int. J. Ecol.* **2011**, e868426 (2011).
60. Sassone-Corsi, M. et al. Microcins mediate competition among Enterobacteriaceae in the inflamed gut. *Nature* **540**, 280–283 (2016).
61. Becattini, S. et al. Commensal microbes provide first line defense against *Listeria monocytogenes* infection. *J. Exp. Med.* **214**, 1973–1989 (2017).
62. Caballero, S. et al. Cooperating commensals restore colonization resistance to vancomycin-resistant *Enterococcus faecium*. *Cell Host Microbe* **21**, 592–602.e4 (2017).
63. Ducarmon, Q. R. et al. Gut microbiota and colonization resistance against bacterial enteric infection. *Microbiol Mol Biol Rev* **83**, e00007–e00019 (2019).
64. Donia, M. S. et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414 (2014).
65. Chatzidaki-Livanis, M., Coyne, M. J. & Comstock, L. E. An antimicrobial protein of the gut symbiont *Bacteroides fragilis* with a MACPF domain of host immune proteins. *Mol. Microbiol.* **94**, 1361–1374 (2014).
66. Roelofs, K. G., Coyne, M. J., Gentyala, R. R., Chatzidaki-Livanis, M. & Comstock, L. E. Bacteroidales secreted antimicrobial proteins target surface molecules necessary for gut colonization and mediate competition in vivo. *mBio* **7**, e01055–16 (2016).
67. Cohan, F. M. Transmission in the origins of bacterial diversity, from ecotypes to phyla. *Microbiol. Spectr.* https://doi.org/10.1128/microbiolspec.mtbp-0014-2016 (2017).
68. Leslie, J. L. et al. Protection from lethal *Clostridioides difficile* infection via intraspecies competition for cogerminant. *mBio* **12**, e00522–21 (2021).
69. Fung, C. et al. High-resolution mapping reveals that microniches in the gastric glands control *Helicobacter pylori* colonization of the stomach. *PLoS Biol.* **17**, e3000231 (2019).
70. Balato, A. et al. Human microbiome: composition and role in inflammatory skin diseases. *Arch. Immunol. Ther. Exp. (Warsz.).* **67**, 1–18 (2019).
71. Browne, H. P. et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
72. Lagier, J. C. et al. Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16**, 540–550 (2018).
73. De Filippis, F. et al. Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. *Cell Host Microbe* **25**, 444–453.e3 (2019).
74. Scher, J. U. et al. Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* **2**, e01202 (2013).
75. Fehlner-Peach, H. et al. Distinct polysaccharide utilization profiles of human intestinal *Prevotella copri* isolates. *Cell Host Microbe* **26**, 680–690.e5 (2019).
76. Lopez-Siles, M. et al. Alterations in the abundance and co-occurrence of *Akkermansia muciniphila* and *Faecalibacterium prausnitzii* in the colonic mucosa of inflammatory bowel disease subjects. *Front. Cell. Infect. Microbiol.* **8**, 281 (2018).
77. Louie, T. et al. VE303, a defined bacterial consortium, for prevention of recurrent clostridioides difficile infection: a randomized clinical trial. *JAMA* **329**, 1356–1366 (2023).

Methods

Human participants

For the US rCDI study with colonoscopic delivery, written consent was obtained from all participants recruited in the study using a protocol approved by the Mount Sinai Institutional Review Board (HS no. 11-01669). Donors and patients who received FMT for rCDI or rCDI and IBD were described in a previous study analysed with 16S ribosomal RNA amplicon sequencing[55]. For the Fecal Microbiota Transplantation for Chronic Active Ulcerative Colitis (FOCUS) study, written informed consent was obtained from all patients before screening. Donors and patients who received FMT for UC were described in a previous study[33]. More patients were recruited at The Mount Sinai Hospital under IRB 16-00021. Human participant health status is available in Supplementary Table 1. As a second rCDI FMT validation cohort, stools were collected from two donors and six recipients from a published nasoduodenal FMT study in Leiden (Extended Data Table 5)[21,78,79]. Patients provided informed consent for collection of stool samples and outcome data of FMT for research purposes, which was approved by the Leiden University Medical Center Medical Ethics Committee (P15.145).

Faecal sample preparation and high-throughput anaerobic bacterial isolation

We followed a previously described protocol[5,40,80]. Briefly, faecal samples were aliquoted on dry ice or liquid nitrogen and stored at −80 °C. Under strict anaerobic conditions, stool from each donor was blended into culture medium[80] and stored at −80 °C. We used a well-established, robotized platform that enables isolation and culturing of a high proportion of bacteria found in the human gut[4–6,34,80]. Briefly, clarified and diluted donor stool was plated onto a variety of solid selective and non-selective media under anaerobic, microaerophilic and aerobic conditions selected to promote the greatest growth of a diverse array of all stool microbes. Plates were incubated for 48–72 h at 37 °C. A total of 384 single colonies from each donor microbiota were picked individually and regrown in medium for 48 h under anaerobic conditions. Regrown isolates were identified at the species level using a combination of matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (Bruker Biotyper) with optional 16S rDNA amplicon sequencing. If either mass spectrometry or 16S rDNA amplicon sequencing identified the presence of two or more bacterial species present, these isolates were not selected for archiving. The 384 isolates were next de-replicated by first maximizing species diversity and second by filling the remainder of the 96-well archive plate with several isolates of the same species. Culture libraries were archived in multi-well plates. DNA was then extracted by bead beating and extraction in phenol chloroform and stored at −20 °C. Across these culture collections, we have found a person's cultured strains represent most bacteria-assigned reads in the metagenome with approximately 70% of the bacterial metagenome mapping to the cultured strain genomes[8].

Construction of whole-genome libraries and Illumina sequencing

We constructed the Illumina library using the seqWell Plexwell 384 kit or an Illumina Nextera kit. DNA was barcoded, ligation products were purified and finally we performed an enrichment PCR. Samples were pooled in equal proportions and size-selected before sequencing with an Illumina HiSeq (paired-end 150 base pairs (bp)). The sequence data files (FASTQ) for whole-genome assemblies are available from the National Center for Biotechnology Information (NCBI) (BioProject IDs: PRJNA880610, PRJNA1093465 and PRJNA637878). Genome quality metrics were computed with QUAST[81]. To ensure genome sequences were of sufficient quality for downstream analyses, we required each genome to have N50 > 10 K (average N50 > 150 K). Although MALDI-TOF and 16S rRNA sequencing provided some initial criteria to eliminate gross contamination, it is critical to also evaluate genome purity to avoid overinflating strain counts through chimeric genome assemblies. All genomes isolated from a single person were compared with each other to check for genomic contaminations from different strains within the same human stool, as we find by far the most common genome error in high-throughput culture collections is a chimeric assembly of two strains from the same person. Genomes with perfect matches to several strains in the same person or containing a large proportion of genomic content from two or more species are removed. Finally, each genome is mapped to a database of 156,403 sequenced bacterial genomes from NCBI to quantify the *k*-mer overlap for each genome with other conspecific strains. This mapping is used to confirm the MALDI-TOF species name, provide a name if the species is not in the Bruker Biotyper MALDI-TOF database and determine whether the length of the genome is within the expected length for that species. The genome length distribution of strains in a species is very consistent and a large deviation from the expected length distribution (more than 3 s.d.) is investigated as potential contamination. Strain genomes were annotated with prokka v.1.12 (ref. 47), and pan-genomic analyses of genome annotations were performed with roary[48]. All strain genomes are available on NCBI. Accessions for each strain are in Supplementary Table 2.

Strain enrichment in gnotobiotic mice

To potentially enrich for strains at lower abundance in the human donor, we colonized 6- to 8-week-old ex-germ free C57BL/6J mice with the same human stool samples previously used for direct culturing and administered two different diets to the colonized mice. Mice of both sexes were assigned randomly for these experiments. We selected the two diets (41% high casein, Harlan TD.09054 and 5% psyllium, Harlan TD.150229) from our published screen of more than 40 custom mouse diets[82], as well as a standard carbohydrate rich mouse chow (LabDiet 5K67). We collected faecal pellets after 2 weeks of diet administration and stored them at −80 °C. Mouse faecal pellets were then used for depth-focused high-throughput culture to sample the SR of our selected species (Extended Data Table 3) more deeply. Mice were housed according to standard guidelines with 12 h dark/light cycles, 18–23 °C and 40–60% humidity. All animal experiments in this study were approved by Institutional Animal Care and Use Committee of the Icahn School of Medicine (protocol: IACUC-2013-1385) and were performed in accordance with the approved guidelines for animal experimentation at the Icahn School of Medicine at Mount Sinai.

Selection of species for SR_j validation and anaerobic bacterial culture

We selected ten gut bacterial species for their membership across the four main phyla of the gut and for their range of SR_j on the basis of preliminary calculations. To enable greater sampling depth for these species, we adjusted our previous breadth-focused culturing approach by plating clarified stool samples on a selected range of environmental conditions designed to cultivate our target species. Clarified stool samples were diluted to grow single colonies. Next, 384 colonies were picked for each donor sample and regrown in liquid medium in multi-well plates. Each isolate was then identified by a combination of MALDI-TOF mass spectrometry and whole-genome sequencing. Using the MALDI identification, the original 384 isolates were de-replicated and about ten isolates of each of the target species were archived in multi-well plates. To eliminate potentially mixed wells, optical density $(OD)_{600}$ of the cultures were collected after each culture and, if the OD_{600} varied significantly from the previous ODs, these wells were not selected. In addition, MALDI-TOF was used to check the bacterial identities after each growth step (after 384-well growth and again after picking and archiving). Only those wells that demonstrated a consistent identification of a single bacterial species were used. DNA was then extracted by bead beating and stored at −20 °C.

Calculation of SR_j with bacterial isolates and metagenomics

As in our previous analyses[4,5,34], bacterial isolates with less than 96% whole-genome similarity were defined as unique strains, otherwise they were considered as several isolates for the representative strain. To calculate an average SR for a given species from cultured isolates, SR_j, we calculated the SR of a species *j* within an individual *i*: SR_{ij}. We then average SR_{ij} across the people in our cohort to arrive at an average SR_j. We only measure SR_{ij} where a person harboured species j within their microbiota. Thus, SR_j is a measure of the number of strains a species stably maintains if it is present within a microbiota. We only quantified those species that had at least two isolates sampled from at least three people to allow for a broad sampling of SR_j across our cultured cohort but with several people sampled per species for a more accurate measurement. For our metagenomics analyses, we used our previously published metagenomics algorithm Strainer[8]; we tracked the presence or absence of a strain in a sample for quantification of SR_j.

Detection of sequenced strains using Strainer

We previously described Strainer[8] for tracking discrete bacterial strain genomes in metagenomes. In brief, we identify a set of informative sequence features, or *k*-mers, from a bacterial genome that can uniquely identify a given strain. We first initialize this informative *k*-mer set by removing those shared extensively with bacterial genomes and faecal metagenomes from unrelated, non-cohabitating people, where the probability of the occurrence of the same strain is very low. Next, we update this informative *k*-mer set by removing those that co-occur on metagenomics reads with uninformative *k*-mers. Finally, we assign each sequencing read in a metagenomics sample of interest to a unique strain by comparing the distribution of *k*-mers on a read with the informative *k*-mers identified earlier, and with controls to find statistical significance. In the rCDI trial, we applied Strainer to metagenomics samples from seven donors and 13 recipients over several time points (pre-FMT to 5 years post-FMT). We sequenced an average of around 5.2 million reads from a total of 85 metagenomics samples across donors and recipients. For the FOCUS UC trial, we also applied *Strainer* on metagenomics samples from 14 donors, 21 pooled batch samples and 63 recipients over several time points (pre-FMT to 5 years post-FMT). The average sequencing depth of metagenomics was 2.4 million reads and we tracked 1,421 unique strains from the donors in the recipients. In the FOCUS UC trial, treated participants were either given an endoscopic FMT followed by 40 enemas over 8 weeks (active arm of placebo-controlled trial) or 40 enemas over 8 weeks without endoscopically administered FMT (optional open-label arm). Both groups were combined for our analyses. For understanding the impact of pooled donors on SR limits, we focused on strains that engrafted in at least 30% of recipients.

Pan-genomic functional analysis of SR_j high and low strains

To determine whether specific genes, genomic features or gene functions might influence SR_j, we analysed the five genera (Bacteroides, Bifidobacterium, Clostridium, Enterococcus and Streptococcus) for which at least six strain isolates were available for at least six different species. All strains for all species in these genera were annotated with prokka[47]. For species in these genera with at least six distinct strains, we quantified the core genome fraction using roary[48], where the core genome fraction was estimated as the number of core genes at the sixth strain relative to the number of genes in the species' genome. To identify genes differentially present between $SR_j > 1.2$ (high) and $SR_j < 1.1$ (low) strains, we used the scoary algorithm to calculate enrichment[49]. The genus *Clostridium* was not used for this analysis as all species have high SR_j. To prevent the difference in the number of strains available per species from disproportionately influencing the comparisons, we randomly subsampled the median number of strains observed for all species in the genus. All species in the genus (including those with fewer than six strains) were used for this comparison to maximize genetic diversity. Clusters of orthologous groups (COG) functions, used to quantify functional enrichment, were assigned with eggNOG v.6.0 (ref. 50). To calculate functional enrichment, all genes significantly associated with SR_j group were aggregated into contingency tables by COG functional category using the Benjamini–Hochberg corrected *P* values from scoary with a threshold of less than 0.05 (ref. 83). Functional enrichment for each COG in each genus was calculated using Fisher's exact test with Benjamini–Hochberg corrected *P* values with a threshold of less than 0.05.

Strain addition, replacement and persistence

In the context of a microbial therapeutic intervention where one or more strains of given species are introduced into a recipient microbiota harbouring a different strain of the same species, the newly introduced strains could be added to the recipient microbiota, replace the original recipient strain or the recipient strains could persist without colonization by the new strains. To quantify these outcomes in the context of the multi-donor FMT trial for recipients with UC, we tracked the presence of donor strains and recipient strains in the recipient metagenomes for six recipients where we cultured and sequenced bacterial strains 5 years after completion of FMT. We observed 27 qualifying episodes across the six donors where one or more strains from the same species were found in the donor and the recipient. Addition was defined as when a recipient strain was retained and one or more donor strains were engrafted. Persistence was defined as when a recipient strain was retained and no donor strains were engrafted. Replacement was defined as when one or more donor strain colonized and the recipient strain was decolonized. Enrichment of strain addition in the context of species with $SR_j > 1.2$ determined by Fischer's exact test.

Statistical analysis and plotting

We analysed data in RStudio (v.2022.02.3+492, v.2023.06.0+421). Heatmaps were created using the R packages ComplexHeatmap v.2.12.0 and circlize v.0.4.15. Kruskal–Wallis was used for one-way data with multiple groups. Spearman rank tests were used to assess the significance of potential correlations between variables. Wilcoxon signed-rank tests were used to determine the potential significance of paired observations. Rarefaction curves were generated using the R package vegan v.2.6.4 using a minimum sampling size of seven (only those people with at least seven isolates of a species were used). Pangenomes were analysed using scoary v.1.6.16, prokka v.1.12, and roary v.3.13.0.

Reporting summary

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

Data availability

Sequence data files (FASTQ) for all cultured and whole-genome assembled sequences are stored in the SRA under project number PRJNA880610. Previously published whole-genome assembled sequenced can be found under project number PRJNA637878. Sequence data files (FASTQ) for all metagenomic sequencing samples from the US FMT study for rCDI patients can be found under project number PRJNA637878. Sequence data files (FASTQ) for all metagenomic sequencing samples from the Leiden FMT validation cohort for rCDI patients can be found under project number PRJEB44737. Sequence data files for metagenomic sequencing samples from the pooled donor FMT trial for UC patients can be found at PRJEB26357. *B. fragilis* isolate whole genomes[7] that were used for validation can be accessed at project number PRJNA524913. Isolate whole genomes from ref. 46 can be accessed at project number PRJNA544527. Source data for Extended Data Fig. 1g,h are based on source data from Extended Data Fig. 1d–f. Source data for Fig. 2b and Extended Data Fig. 1a,b are available at Zenodo (https://doi.org/10.5281/zenodo.13942097)[84]. Source data are provided with this paper.

Code availability

The strain-tracking algorithm, Strainer, was published previously[8]. Code for Strainer can be accessed at https://bitbucket.org/faithj02/strainer-metagenomics/.

78. Nooij, S. et al. Fecal microbiota transplantation influences procarcinogenic *Escherichia coli* in recipient recurrent *Clostridioides difficile* patients. *Gastroenterology* **161**, 1218–1228.e5 (2021).
79. Nooij, S. et al. Long-term beneficial effect of faecal microbiota transplantation on colonisation of multidrug-resistant bacteria and resistome abundance in patients with recurrent *Clostridioides difficile* infection. *Genome Med.* **16**, 37 (2024).
80. Britton, G. J. et al. Microbiotas from humans with inflammatory bowel disease alter the balance of gut Th17 and RORγt$^+$ regulatory T cells and exacerbate colitis in mice. *Immunity* **50**, 212–224 (2019).
81. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
82. Llewellyn, S. R. et al. Interactions between diet and the intestinal microbiota alter intestinal permeability and colitis severity in mice. *Gastroenterology* **154**, 1037–1046.e2 (2018).
83. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
84. Chen-Liaw, A. Source Data. *Zenodo* https://doi.org/10.5281/zenodo.13942097 (2024).

Acknowledgements We thank C. Fermin, E. Vazquez and G. N. Escano for gnotobiotic husbandry. This work was supported in part by the staff and resources of the Mount Sinai Gnotobiotic Facility and the Scientific Computing Division at the Icahn School of Medicine at Mount Sinai. This work was supported by the National Institutes of Health grants (nos. NIDDK DK112978, NIDDK DK124133, NIDDK DK123749), an NIH F30 to A.C.-L. (NIDDK DK131862), Crohn's and Colitis Foundation awards (no. 650451 to V.A.; no. 651867 to J.J.F.; no. 988415 to N.O.K.) and Janssen Research & Development.

Author contributions A.C.-L. and J.J.F. wrote the manuscript. A.C.-L., I.M., Z.L. and J.E. performed the microbiology, mouse model and molecular biology experiments on human bacterial isolates. R.L.K., F.E.R., I.M., Z.L., A.C.-L. and J.J.F performed the microbiology and molecular biology experiments on environmental microbes. A.G., A.C.-L., J.J.F. and V.A. performed the US rCDI clinical study and analysis. E.M.T., J.J.K., B.O., J.M.N., R.M., A.R.W., E.C., A.C.-L. and J.J.F. performed the Netherlands rCDI clinical study, data generation and analysis. S.P., N.O.K., C.H., M.A.K., T.J.B., V.A., A.C.-L. and J.J.F. performed the UC FMT clinical study, data generation and analysis. M.C.D., D.H., J.F.C., A.H., J.W., E.S.N.L-S., J.J.F. and A.C.-L., performed the non-FMT human clinical collections, data generation and analysis. All authors read and provided critical feedback and approved the final manuscript.

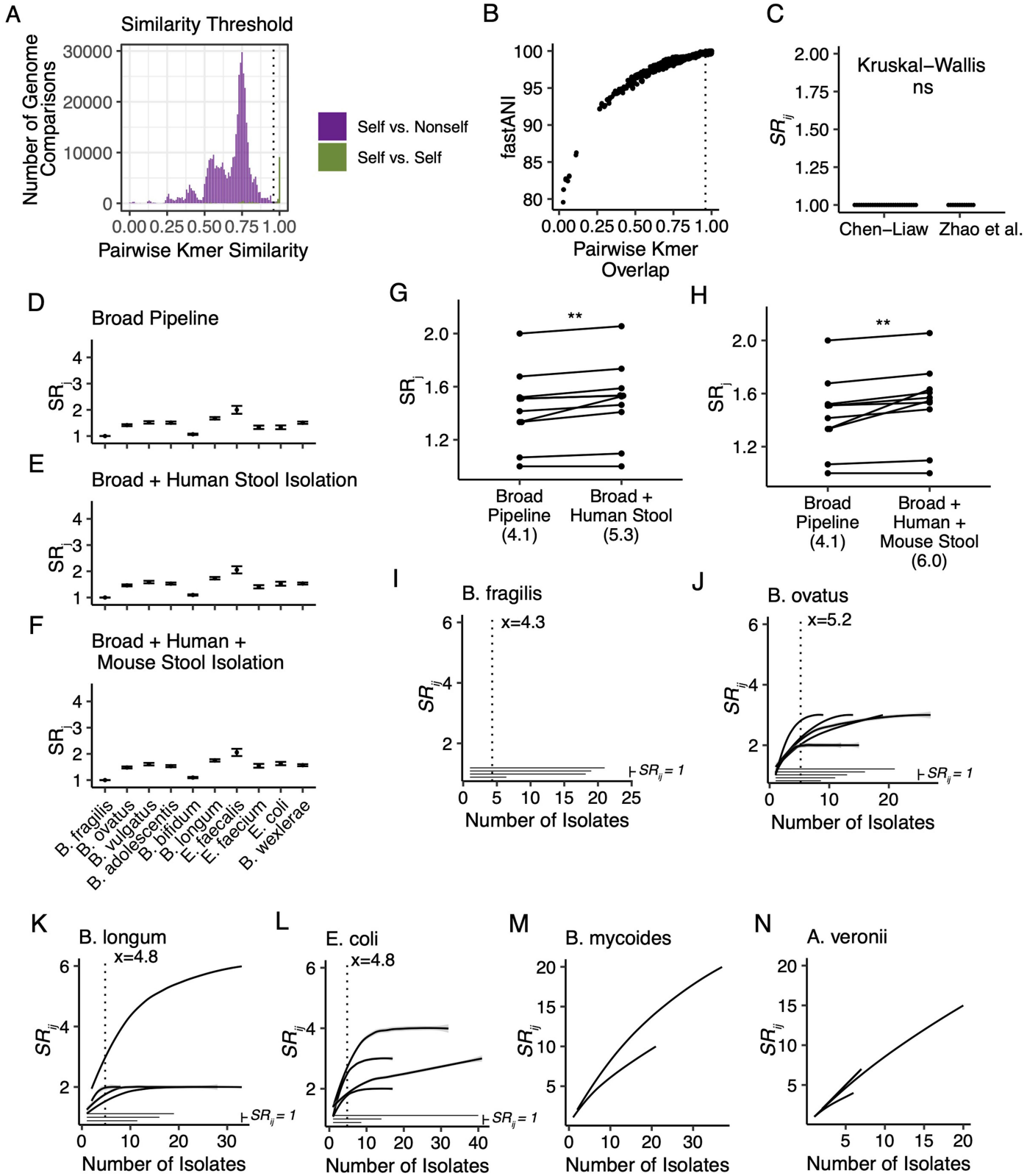
Competing interests J.J.F. is a scientific advisory board member and consultant to Vedanta Biosciences, Inc. A.H., J.W., E.S.N.L.-S. are employees of Janssen Research & Development. B.O., J.M.N., R.M., A.R.W. and E.C. are employees of Vedanta Biosciences. J.K. and E.T. received research grants from Vedanta Biosciences. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08242-x.

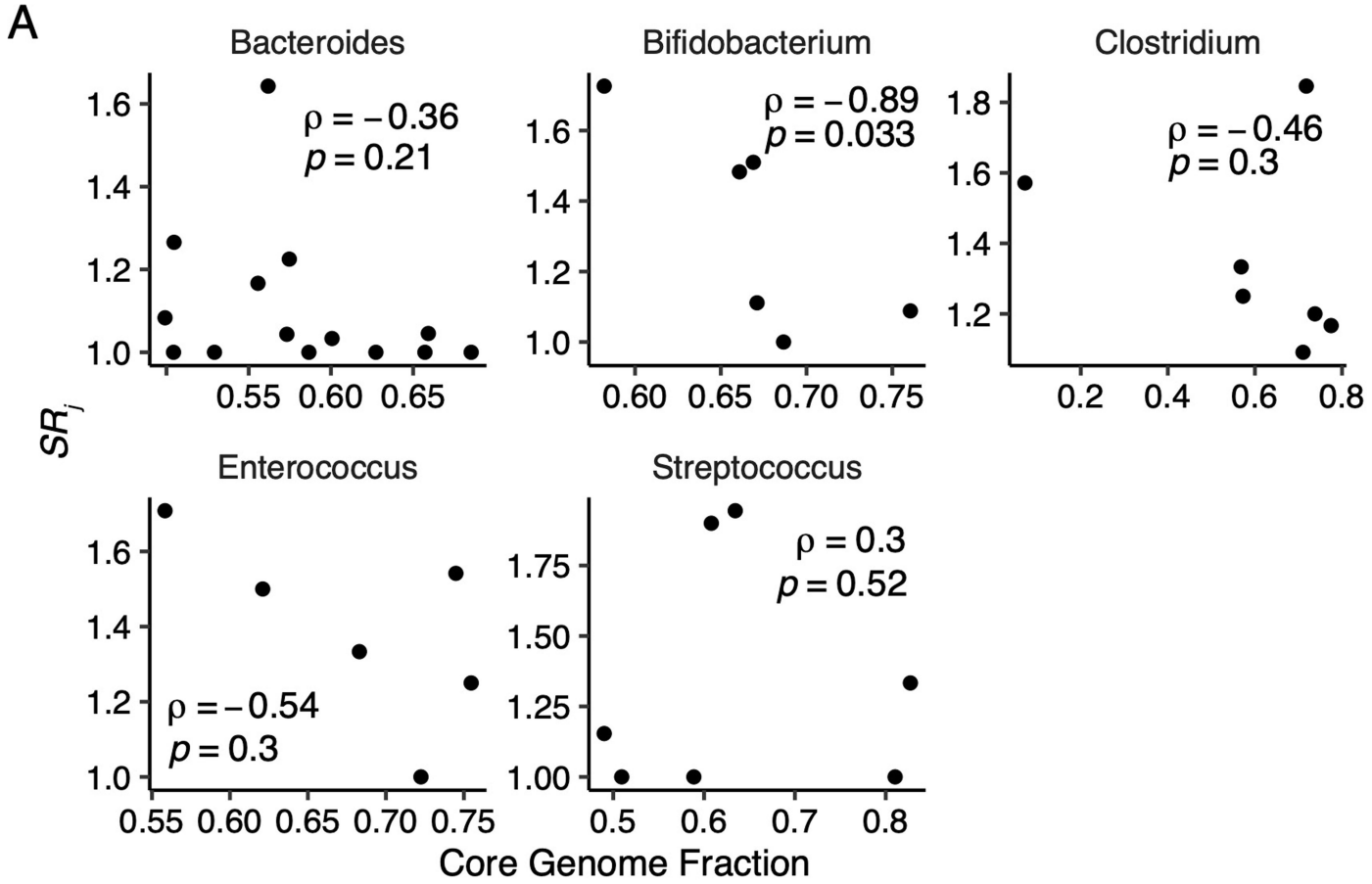
Correspondence and requests for materials should be addressed to Jeremiah J. Faith.

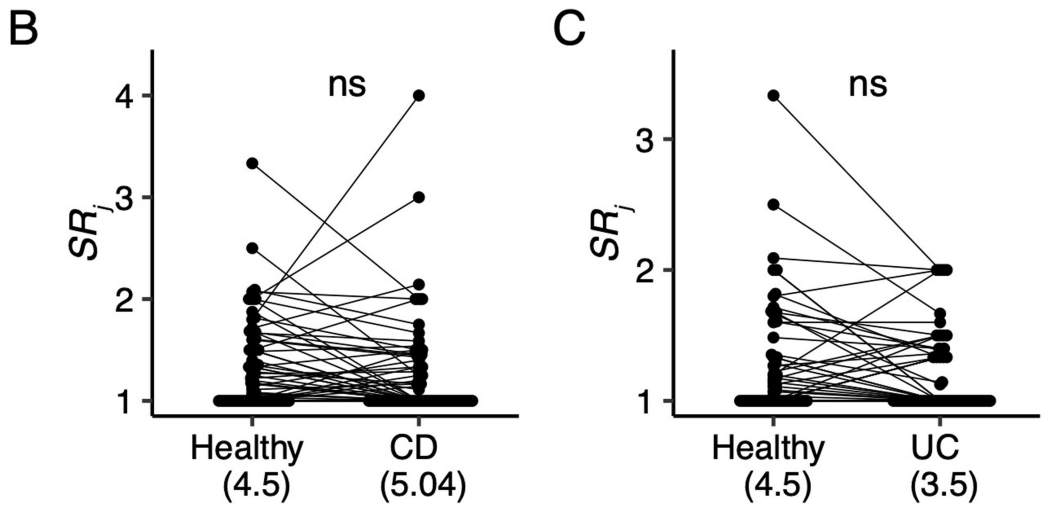
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



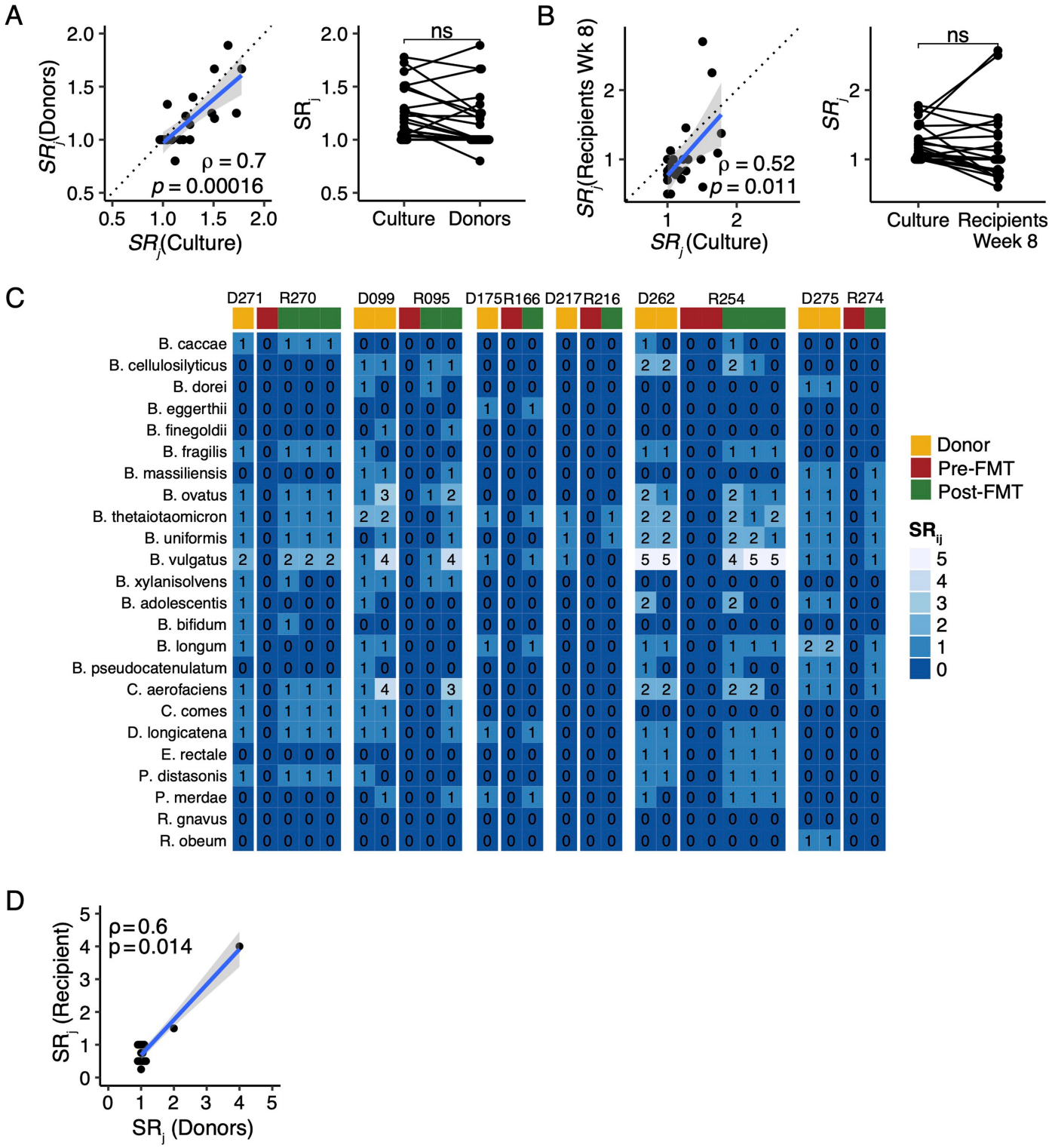
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Determination of SR_j strain threshold and validation of SR_j with deeper sampling of a subset of cultured gut bacterial species. **A**, K-mer overlap was calculated for all pairwise combinations of isolates from a species cultured from two unrelated individuals (purple) with no direct microbial transfer between them or an individual's own microbes from a single timepoint (green). Dotted line shows the threshold of 0.96 k-mer similarity. **B**, fastANI versus pairwise k-mer overlap for isolates of the same species. Dotted line represents k-mer overlap threshold of 0.96 (k-mer distance of <0.04). **C**, Comparison of *B. fragilis* SR_{ij} for the individuals in this study and in the study by Zhao et al. **D**, Preliminary calculation of SR_j using genomes isolated from the Broad pipeline (standard pipeline used to create libraries of cultured gut bacteria). For deeper sampling estimates of SR_j, we isolated additional genomes from the same species from **E**, the original human stool sample and **F**, mouse stool samples from gnotobiotic mice colonized with the same human stool (N = 2-3 mice per microbiota/diet combination) and given unique diets for strain enrichment. Error bars in **D**–**F** represent SEM. Comparison of SR_j calculated with genomes from the broad pipeline or the broad pipeline plus additional genomes from **G**, human stool samples or **H**, human and mouse stool samples. **I**–**L**, Rarefaction curves for validation species. Dotted line shows the mean isolates/species for each species (overall mean isolates/species across the dataset was 4.7 isolates as demonstrated in Fig. 1b). **M**,**N**, Rarefaction curves for a soil species (**M**) and lake species (**N**). **p < 0.01, paired Wilcoxon test, each point represents the average SR_j measured from microbiomes of a healthy or disease state. ns: not significant by paired Wilcoxon test. Grey regions on rarefaction curves indicate 95% confidence intervals.

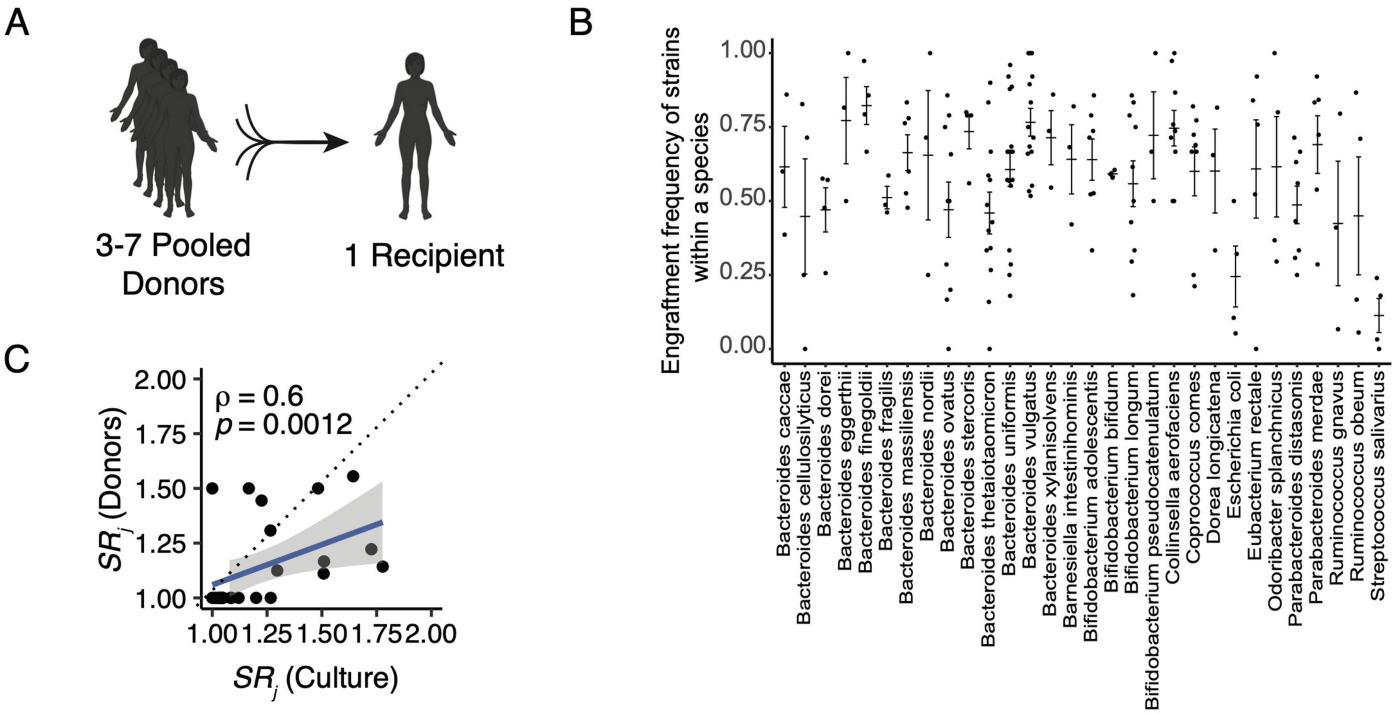




Extended Data Fig. 2 | The influence of core genome size and disease state on strain richness. A, Spearman rank correlation for SR_j versus core genome fraction for several genera. **B**, SR_j was compared for 59 species present in both healthy and CD microbiomes. **C**, SR_j was compared for 51 species present in both healthy and UC microbiomes. Each point represents the average SR_j of a species as calculated using isolates cultured from healthy microbiomes or isolates cultured from disease microbiomes. Lines in **B**,**C** connect the SR_j for each shared species between a healthy subject and a subject with IBD. ns: not significant by paired Wilcoxon test.



Extended Data Fig. 3 | Metagenomics-quantified SR_j correlates with the cultured SR_j measured across our cohort. A, Spearman correlation between donor SR_j as measured by metagenomics and SR_j from our original cultured cohort (first panel) and differences in SR_j between the two groups (second panel). **B**, Spearman correlation between recipient SR_j as measured by metagenomics at week 8 post-FMT and SR_j from our original cohort (first panel) and differences in SR_j between the two groups (second panel). **A**,**B**, Each point represents the average SR_j for species. **C**, Heatmap representing the remaining six donors who donated stool to six different recipients. **D**, Correlation of SR_j across all recipients and all donors in an independent FMT validation cohort (Leiden).



D A strain richness ecological framework for strain persistence, replacement, or addition

Definitions:

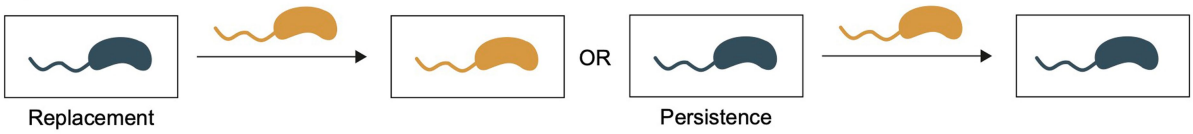
Persistence = donor strain does not engraft, recipient strain remains

Replacement = donor strain engrafts, recipient strain is displaced

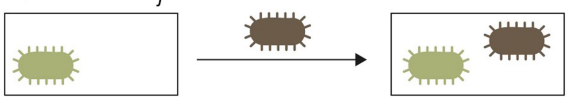
Addition = donor strain engrafts, recipient strain remains

Expected outcomes

Species SR_i < 1.1



Species SR_j > 1.2



Extended Data Fig. 4 | An ecological framework for strain persistence, replacement, or addition based on strain richness. A, Schematic for FMT experimental design with multi-donor stool batches administered to each recipient. **B**, Within a species, bacterial strains vary in their engraftment frequency when administered in the context of a multi-donor FMT product. **C**, Spearman rank correlation between the metagenomics SR_j of the individual FOCUS donors with the overall cultured SR_j measured across our cohort. **D**, Expected outcomes for donor strain engraftment based on species SR_j.

Extended Data Table 1 | SR_j for each species sampled in our cultured cohort

Species	SR_j	Species	SR_j
Alistipes finegoldii	1.00	Enterococcus avium	1.00
Alistipes onderdonkii	1.20	Enterococcus casseliflavus	1.50
Alistipes shahii	1.17	Enterococcus durans	1.33
Anaerotruncus colihominis	1.00	Enterococcus faecalis	1.55
Bacteroides caccae	1.04	Enterococcus faecium	1.63
Bacteroides cellulosilyticus	1.05	Enterococcus gallinarum	1.25
Bacteroides dorei	1.05	Enterococcus mundtii	1.00
Bacteroides eggerthii	1.00	Erysipelatoclostridium ramosum	1.33
Bacteroides faecis	1.00	Escherichia coli	1.57
Bacteroides finegoldii	1.00	Eubacterium rectale	1.20
Bacteroides fragilis	1.00	Eubacterium siraeum	1.00
Bacteroides intestinalis	1.00	Faecalicoccus pleomorphus	1.00
Bacteroides massiliensis	1.00	Faecalitalea cylindroides	1.00
Bacteroides ovatus	1.48	Flavonifractor plautii	1.33
Bacteroides salyersiae	1.00	Fusicatenibacter saccharivorans	1.00
Bacteroides stercoris	1.18	Intestinibacter bartlettii	1.33
Bacteroides thetaiotaomicron	1.24	Intestinimonas butyriciproducens	1.25
Bacteroides uniformis	1.29	Klebsiella pneumoniae	2.33
Bacteroides vulgatus	1.61	Lactobacillus gasseri	1.00
Bacteroides xylanisolvens	1.08	Lactobacillus paracasei	1.17
Barnesiella intestinihominis	1.00	Lactobacillus paragasseri	1.00
Bifidobacterium adolescentis	1.53	Lactobacillus plantarum	1.00
Bifidobacterium angulatum	1.00	Lactobacillus rhamnosus	1.17
Bifidobacterium animalis	1.00	Lactobacillus ruminis	1.00
Bifidobacterium bifidum	1.10	Lactobacillus salivarius	1.00
Bifidobacterium breve	1.00	Lactococcus garvieae	2.00
Bifidobacterium catenulatum	1.11	Lactococcus lactis	1.00
Bifidobacterium dentium	1.00	Odoribacter splanchnicus	1.00
Bifidobacterium longum	1.75	Parabacteroides distasonis	1.03
Bifidobacterium pseudocatenulatum	1.46	Parabacteroides goldsteinii	1.00
Blautia obeum	1.25	Parabacteroides merdae	1.14
Blautia wexlerae	2.06	Pediococcus acidilactici	1.00
Catenibacterium mitsuokai	1.00	Peptostreptococcus anaerobius	1.25
Citrobacter farmeri	1.00	Romboutsia 1001216sp1	1.00
Clostridium butyricum	1.25	Roseburia faecis	1.00
Clostridium clostridioforme	1.25	Ruminococcus bicirculans	1.00
Clostridium disporicum	1.57	Ruminococcus gnavus	1.29
Clostridium paraputrificum	1.21	Streptococcus agalactiae	1.00
Clostridium perfringens	1.82	Streptococcus anginosus	1.18
Clostridium symbiosum	1.11	Streptococcus australis	1.00
Clostridium tertium	1.17	Streptococcus dysgalactiae	1.00
Collinsella aerofaciens	1.28	Streptococcus mutans	1.29
Coprococcus comes	1.78	Streptococcus oralis	1.00
Coprococcus eutactus	1.00	Streptococcus parasanguinis	2.13
Dorea longicatena	1.00	Streptococcus salivarius	2.00
Eggerthella lenta	1.60	Turicibacter sanguinis	2.57

Extended Data Table 2 | Species with significantly different SR_j based on Dunn test with Benjamini-Hochberg correction

Comparisons	Z	P	P.adjusted	chi2
Bacteroides_massiliensis - Bacteroides_ovatus	-3.83523725	6.27216E-05	0.032819063	173.4122752
Bacteroides_massiliensis - Bifidobacterium_adolescentis	-4.209875492	1.27756E-05	0.013369637	173.4122752
Bacteroides_massiliensis - Bifidobacterium_longum	-4.203917104	1.31168E-05	0.010981345	173.4122752
Bifidobacterium_adolescentis - Flavonifractor_plautii	3.702705977	0.000106656	0.037205176	173.4122752
Bifidobacterium_longum - Flavonifractor_plautii	3.700857602	0.000107436	0.034594398	173.4122752
Bacteroides_ovatus - Odoribacter_splanchnicus	3.580427701	0.000171516	0.047864434	173.4122752
Bifidobacterium_adolescentis - Odoribacter_splanchnicus	3.928317505	4.27711E-05	0.02983998	173.4122752
Bifidobacterium_longum - Odoribacter_splanchnicus	3.907936749	4.65438E-05	0.027833212	173.4122752
Bacteroides_ovatus - Streptococcus_anginosus	4.302491209	8.44442E-06	0.011782779	173.4122752
Bifidobacterium_adolescentis - Streptococcus_anginosus	4.705027737	1.26916E-06	0.002656346	173.4122752
Bifidobacterium_longum - Streptococcus_anginosus	4.728345653	1.13178E-06	0.004737644	173.4122752
Enterococcus_faecalis - Streptococcus_anginosus	3.722161323	9.87624E-05	0.037583573	173.4122752
Enterococcus_faecium - Streptococcus_anginosus	3.823431884	6.58035E-05	0.03060593	173.4122752
Escherichia_coli - Streptococcus_anginosus	3.813792784	6.84251E-05	0.028642752	173.4122752

Extended Data Table 3 | Impact of deeper sampling and ex-germ-free mouse enrichment on SR_j

Species	Broad Pipeline	Broad + Human Stool Isolation	Broad + Mouse Stool Isolation	Broad + Human + Mouse Stool Isolation
Bacteroides fragilis	1.00	1.00	1.00	1.00
Bacteroides ovatus	1.35	1.45	1.38	1.48
Bacteroides vulgatus	1.57	1.67	1.62	1.73
Bifidobacterium adolescentis	1.32	1.34	1.32	1.34
Bifidobacterium bifidum	1.06	1.12	1.06	1.12
Bifidobacterium longum	1.36	1.45	1.38	1.48
Enterococcus faecalis	1.50	1.57	1.69	1.79
Enterococcus faecium	1.63	1.90	1.89	2.10
Escherichia coli	1.45	1.52	1.53	1.58
Blautia wexlerae	2.00	2.14	2.00	2.14

Extended Data Table 4 | SR_j changes over time across the FOCUS FMT trial for UC

Species	Donor Batch	Drug Week 4	Drug Week 8	Post-Drug Week 8	Post-Drug 5 Year	Cultured SRj
Bacteroides caccae	2.00	1.33	1.11	0.70	0.36	1.03
Bacteroides cellulosilyticus	1.00	1.00	0.56	0.50	0.18	1.04
Bacteroides dorei	2.50	1.33	0.89	0.60	0.45	1.05
Bacteroides eggerthii	1.62	1.11	1.00	0.80	0.64	1.00
Bacteroides finegoldii	1.62	1.67	1.44	1.10	0.91	1.00
Bacteroides fragilis	2.12	1.67	1.11	0.70	0.18	1.00
Bacteroides massiliensis	3.12	1.67	1.67	1.30	1.00	1.00
Bacteroides ovatus	3.25	1.44	1.56	0.90	0.64	1.51
Bacteroides salyersiae	0.62	0.44	0.33	0.30	0.36	1.00
Bacteroides stercoris	1.00	0.33	0.44	0.10	0.09	1.17
Bacteroides thetaiotaomicron	5.50	2.56	2.22	1.80	1.18	1.23
Bacteroides uniformis	7.00	4.33	3.89	3.50	2.64	1.27
Bacteroides vulgatus	5.62	4.11	3.78	3.20	2.45	1.64
Bacteroides xylanisolvens	2.38	2.11	1.78	1.50	1.27	1.08
Barnesiella intestinihominis	2.12	1.44	1.33	1.00	0.64	1.00
Bifidobacterium adolescentis	2.75	2.00	2.11	1.90	1.73	1.51
Bifidobacterium bifidum	1.88	1.44	1.11	1.00	0.91	1.09
Bifidobacterium longum	4.25	2.22	2.11	1.80	1.18	1.73
Bifidobacterium pseudocatenulatum	0.38	n/a	n/a	0.30	0.27	1.48
Collinsella aerofaciens	2.50	2.00	2.11	1.70	1.55	1.30
Coprococcus comes	4.12	2.89	2.67	2.20	2.00	1.78
Dorea longicatena	1.38	1.00	1.00	1.00	0.91	1.00
Eubacterium rectale	3.00	3.00	2.89	2.50	1.73	1.20
Parabacteroides distasonis	3.50	1.67	1.33	1.00	0.82	1.05
Parabacteroides merdae	2.75	2.33	2.33	1.70	1.36	1.12
Ruminococcus gnavus	2.25	1.11	1.11	0.90	0.82	1.27

Extended Data Table 5 | Patient characteristics from Leiden FMT validation cohort

Patient ID	Donor or Recipient	Antibiotic duration pretreatment	Days post-FMT	Antibiotic pretreatment prior FMT	FMT indication	Relapse after FMT	FMT Mode	NDFB study alias	Metagenome
D16017	donor	NA	NA	NA	NA	NA	NA	d5	D16017-03Jan2017_S7_R1.fastq.gz
D16017	donor	NA	NA	NA	NA	NA	NA	d5	D16017-1Mar2017_S29_R1.fastq.gz
D16017	recipient	>23	52	vancomycin	rCDI	no	nasoduodenal	p76	P17033-2017-03-16_S4_R1.fastq.gz
D16017	recipient	unknown	53	vancomycin	rCDI	no	nasoduodenal	p15	P17035-2017-04-17_S9_R1.fastq.gz
D16017	recipient	10	48	fidaxomicin	rCDI	no	nasoduodenal	p32	P17057-2017-08-21_S20_R1.fastq.gz
D16017	recipient	10	48	vancomycin	rCDI	no	nasoduodenal	p34	P17061-2017-10-16_S22_R1.fastq.gz
D17001	donor	NA	NA	NA	NA	NA	NA	d6	D17001-2017-02-27_S28_R1.fastq.gz
D17001	donor	NA	NA	NA	NA	NA	NA	d6	D17001-2017-05-08_S27_R1.fastq.gz
D17001	recipient	unknown	48	vancomycin	rCDI	no	nasoduodenal	p17	P17038-2017-05-10_S7_R1.fastq.gz
D17001	recipient	29	32	vancomycin	rCDI	no	nasoduodenal	p38	P17071-2018-01-07_S28_R1.fastq.gz

nature portfolio

Corresponding author(s): Jeremiah Faith

Last updated by author(s): Sep 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	None used
Data analysis	Strainer (published open source algorithm), RStudio (v2022.02.3+492, v2023.06.0+421), scoary v1.6.16, prokka v1.12, roary v3.13.0, eggNOG 6.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequence data files (FASTQ) for all cultured and whole genome assembled sequences are stored in the SRA under project number PRJNA880610. Previously published whole genome assembled sequenced can be found under project number PRJNA637878. Sequence data files (FASTQ) for all metagenomic sequencing samples from the USA FMT study for recurrent C. difficile patients can be found under project number PRJNA637878. Sequence data files (FASTQ) for all

metagenomic sequencing samples from the Leiden FMT validation cohort for recurrent C. difficile patients can be found under project number PRJEB44737. Sequence data files for metagenomic sequencing samples from the pooled-donor FMT trial for UC patients can be found at PRJEB26357. B. fragilis isolate whole genomes from the Zhao et al., Cell Host & Microbe 2019 that were used for validation can be accessed at project number PRJNA524913. Isolate whole genomes from the Poyet et al. Nature Medicine 2019 study can be accessed at project number PRJNA544527.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	For the recurrent CDI study, written consent was obtained from all individuals recruited in the study using a protocol approved by the Mount Sinai Institutional Review Board (HS no. 11-01669). Donors and patients who received FMT for recurrent CDI or recurrent CDI and IBD were described in a previous study analyzed with 16S ribosomal RNA amplicon sequencing32. For the Fecal Microbiota Transplantation for Chronic Active Ulcerative Colitis (FOCUS) study, written informed consent was obtained from all patients prior to screening. Donors and patients were who received FMT for UC were described in a previous study19. Additional patients were recruited at The Mount Sinai Hospital under IRB 16-00021. For analysis, we only considered for analysis the subset of individuals for which donor and recipient stool samples from multiple time points had been collected. As a second rCDI FMT validation cohort, stools were collected from donors and recipients from an FMT study in Leiden under METC approval number P15.145.
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	Demographics of the recurrent CDI study can be found in Aggarwala et al., Nature Micro 2021, for the FOCUS study at Paramsothy et al., The Lancet 2017, and for the rCDI FMT validation cohort at Nooij et al., Gastroenterology 2021. Relevant characteristics such as health status and recent antibiotic use can be found in tables S1 and S5.
Recruitment	Recruitment information is described in the methods of the manuscript, Aggarwala et al., Nature Micro 2021 and Paramsothy et al., The Lancet 2017, and Nooij et al., Gastroenterology 2021.
Ethics oversight	Mount Sinai Institutional Review Board, Leiden University Medical Center Medical Ethics Committee

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the calculations of strain richness, species were considered only if that species had at least 2 isolates each from 3 individuals. We also added some specific text in the Methods section which now reads " We only quantified those species that had at least two isolates sampled from at least three individuals to allow for a broad sampling of SRj across our cultured cohort but with several individuals sampled per species for a more accurate measurement." Given that species may vary across individuals and the enormous labor involved in culturing isolates from stool, we determined three individuals to be a sufficient sampling. We also only calculated strain richness from those individuals that harbored the gut species. For the clinical trial data, we assessed only those individuals that had multiple timepoints available.
Data exclusions	If species did not meet the criteria stated above, we excluded them. Thus, SRj is a measure of the number of strains a species stably maintains if it is present within a microbiota.
Replication	We conducted the gnotobiotic enrichment experiments (Fig S1; described in methods) on a subset of gut species with a range of strain richnesses and across the major phyla to validate our calculations of strain richness from previously isolated strains. Results are detailed in the manuscript.
Randomization	All timepoints from a FMT recipient were grouped and analyzed together. Independent FMT interventions were analyzed separately and not grouped together. All subjects received the same therapy. Our results replicated uniformly across all independent samples.
Blinding	The FMT data came from two previously published clinical trials (Aggarwala et al., Nature Micro 2021 and Paramsothy et al., The Lancet 2017), thus blinding was not necessary. We had access to FMT outcomes and all individuals received the therapy. To detect donor strains in recipient samples, we used a previously published algorithm, Strainer (Aggarwala et al., Nature Micro 2021). rCDI validation data also came from a previously published trial, thus did not need blinding (Nooij et al., Gastroenterology 2021).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Clinical data
☐	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Germ free C57BL/6J mice were bred in-house at the Mount Sinai Immunology Institute Gnotobiotic facility in flexible vinyl isolators. Shortly after weaning (28-42 days old) and under strict aseptic conditions, germ-free mice were transferred to autoclaved filter-top cages outside the breeding isolator and colonized with human microbiotas. Mice were colonized with 200-300 ul of a fecal slurry or pooled cocktail of cultured strains by oral gavage, given only once. Mice were housed according to standard guidelines with 12 hour dark/light cycles, 18-23 degrees C, and 40-60% humidity.
Wild animals	None.
Reporting on sex	Mice of both sexes were randomly assigned for the gnotobiotic enrichment experiments.
Field-collected samples	None.
Ethics oversight	All animal experiments in this study were approved by Institutional Animal Care and Use Committee (IACUC) of the Icahn School of Medicine (protocol: IACUC-2013-1385) and were performed in accordance with the approved guide- lines for animal experimentation at the Icahn School of Medicine at Mount Sinai. Full names of the institutes are provided here and in the manuscript.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Data used in this study came from two previously published clinical trials. They are described in Aggarwala et al., Nature Micro 2021 and Paramsothy et al., The Lancet 2017. Validation rCDI data came from a previously published trial in Nooij et al., Gastroenterology 2021.
Study protocol	Protocols are found in: Aggarwala et al., Nature Micro 2021, Paramsothy et al., The Lancet 2017, and Nooij et al., Gastroenterology 2021.
Data collection	Information on data collection: Aggarwala et al., Nature Micro 2021, Paramsothy et al., The Lancet 2017, Nooij et al., Gastroenterology 2021.
Outcomes	Defined in: Aggarwala et al., Nature Micro 2021, Paramsothy et al., The Lancet 2017, and Nooij et al., Gastroenterology 2021.

Plants

Seed stocks	None
Novel plant genotypes	None
Authentication	None