LongTrack: long read metagenomics-based precise tracking of bacterial strains and their genomic changes after fecal microbiota transplantation

Yu Fan[1], Mi Ni[1], Varun Aggarwala[1], Edward A. Mead[1], Magdalena Ksiezarek[1], Lei Cao[1], Michael A. Kamm[2], Thomas J. Borody[3], Sudarshan Paramsothy[4], Nadeem O. Kaakoush[5], Ari Grinspan[6], Jeremiah J. Faith[1] and Gang Fang[1#]

[1] *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA*

[2] *Department of Gastroenterology and Medicine, St. Vincent's Hospital, Melbourne, VIC, Australia*

[3] *Centre for Digestive Diseases, Sydney, NSW, Australia*

[4] *University of Sydney, Sydney, Australia*

[5] *School of Biomedical Sciences, University of New South Wales, Sydney, Australia*

[6] *Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.*

[#] Address correspondence to: *gang.fang@mssm.edu*

Abstract

Fecal microbiota transplantation (FMT) has revolutionized the treatment of recurrent *Clostridioides difficile* infection (rCDI) and is being evaluated across other diseases. Accurate tracking of bacterial strains, the functional units of the microbiota, that stably engraft in recipients is critical for understanding the determinants of strain engraftment, evaluating their correlation with clinical outcomes, and guiding the development of therapeutic bacterial consortia. While short-read sequencing has advanced our understanding of FMT, it faces challenges in the strain-level *de novo* metagenomic assembly of microbiome samples. In this study, we described a novel framework, *LongTrack*, that uses long-read metagenomic assemblies and reliable informatics tailored for FMT strain tracking. We first highlighted LongTrack's advantage over short-read approaches especially when multiple strains co-exist in the same sample. We then showed LongTrack uncovered hundreds of engrafted strains across six FMT cases of rCDI and inflammatory bowel disease patients, validated by complementary bacterial culturing. Furthermore, beyond conventional strain tracking, long reads also allowed us

to assess the genomic and epigenomic stability of successfully engrafted strains during the 5-year follow-ups, revealing structural variations in their genomes, which may be associated with strain adaptation in a new host environment. Combined, our study advocates the use of long read metagenomics and LongTrack to enhance strain tracking in future FMT studies, paving the way for the development of more effective defined biotherapeutic as an alternative to FMT.

Introduction

Fecal microbiota transplantation (FMT) has emerged as a new therapeutic strategy, transforming the treatment paradigms for patients with recurrent *Clostridioides difficile* infection (rCDI)[1–3]. By transferring gut microbiota from a healthy donor into a patient, FMT effectively restores the patient's gut microbial ecology[4,5]. Potential applications of FMT extend beyond rCDI onto treating a broader array of diseases[6,7]. Notably, the use of FMT in the treatment of inflammatory bowel disease (IBD) is under investigation with multiple clinical trials[8–10]. Additionally, there is growing interest in the potential of FMT as an adjunctive tool in cancer immunotherapy[11–13] .

The FMT procedure operates on the principle of transferring gut microbiota, which constitutes a diverse array of bacterial strains, the functional impact units of the gut microbiota[14–16], from a healthy donor into a recipient[4,17]. Therefore, the precise tracking of these strains, from the point of transplantation to their eventual engraftment within the recipient's gut, is of critical importance[18,19]. Specifically, reliable strain tracking can help researchers understand factors that facilitate or hinder engraftment[5,17]. Moreover, comprehensive strain tracking can provide vital information on how the engraftment of specific strains correlates with clinical outcomes[18,20]. Lastly, a detailed understanding of bacterial strain engraftment can guide the development of defined bacterial consortiums, composed of select bacterial strains with beneficial effects, for safer therapeutic applications[19–21].

Several previous studies, powered by high throughput metagenomic shotgun sequencing, have provided valuable insights into strain engraftment in FMT in various diseases[18,20–25]. These studies relied on short-read sequencing and metagenomic analyses, which can detect single nucleotide polymorphisms (SNPs) and/or gene clusters that help determine the source of a strain observed in the recipient after FMT[22–24,26–28]. However, as demonstrated in a recent study that did comprehensive bacteria isolation combined with strain tracking using high-purity and high-completeness genome sequences of bacterial isolates[20], existing methods that use short-read sequencing alone face challenges when multiple strains co-exist in a single microbiome sample[29–34]. Specifically, the complexity of recovering the number of strains in a metagenome poses a critical challenge in subsequent strain tracking[29,35].

Bacterial isolation effectively addressed this challenge by converting a complex microbiome sample into a finite set of discrete bacteria isolates[20]. Once the genomes of individual bacterial isolates from donor and pre-FMT recipient samples are assembled (high-purity and high-completeness), downstream tracking (post-FMT) can be reliably achieved based on the unique k-mers of each strain in short-read metagenomic sequencing samples even with modest-to-low sequencing depth[20]. However, standard bacterial culturing methods tend to miss certain bacterial taxa due to the selective nature of culture media[36]. Additionally, it is not feasible (labor intensive and time consuming) to perform comprehensive bacterial culture across a large number of fecal samples, making it challenging to scale up to large FMT cohorts. Strain-resolved *de novo* meta-genome assembly, if achievable, can greatly complement the bias and scalability of culture-based approaches.

In this context, long-read sequencing could empower bacterial strain tracking in FMT. Specifically, the ability of long-read sequencing to sequence DNA molecules of thousands of kilobases in a single read is much more effective at *de novo* genome assembly, constructing metagenome-assembled genomes (MAGs) at strain resolution[32,37,38]. With this rationale, we

aimed to address two questions in this study. (Q1) How to reliably use long-read MAGs for strain tracking? To answer this question, we developed a new framework, *LongTrack*, that uses tailored informatics to reliably use long-read MAGs for strain tracking (**Fig. 1**). We built upon a recent study[20] that conducted a systematic culture of bacterial strains from FMT donors and recipients for rCDI with 5-year follow-ups (**Fig. 1A**), which serve as a reliable, complex and realistic ground truth for rigorous evaluation of LongTrack (**Fig. 1B**). We systematically evaluated LongTrack in comparison with short-read MAGs and alternative methods and observed superior performance of LongTrack over short-read based approaches, especially in distinguishing co-existing bacterial strains within a microbiome sample for reliable FMT strain tracking (**Fig. 1C**). (Q2) How stable are the genomes of the engrafted strains (before and after FMT)? We asked this question to expand beyond the standard scope of strain tracking building on the FMT cohort with longitudinal sample collections across five years. Quantitative analysis building on the unique advantage of long reads for reliable strain-level mapping uncovered complex structural genomic variations within the same strains that differ between donor and post-FMT samples (**Fig 1D**). In addition, we also built on the unique ability of long-read sequencing technology for the direct detection of bacterial DNA methylation[39–43] (which contributes to epigenetic heterogeneity and plasticity[41,44]), and highlight epigenetic variations within the same strains during FMT and follow-ups. These in-depth genomic and epigenomic analyses suggest the versatile capacity of individual strains during FMT and long-term adaptation. Combined, this study introduces *LongTrack* (**Fig. 1C and 1D**), a novel framework that advocates the incorporation of long-read sequencing to improve the tracking of bacterial strains and their genomic dynamics in future FMT studies, paving the way towards better understanding FMT and the development of more effective defined biotherapeutics.

Results

High purity, strain-level Long-read MAGs serve as a solid foundation for LongTrack.

The success of strain tracking in FMT depends on obtaining strain-resolved and high purity genomes from both the donor and the recipient before FMT. Strain resolution is crucial as it allows differentiation among co-existing strains of the same species, while the high purity of MAGs avoids chimeric contigs or bins (comprising mixed strains) and supports accurate strain tracking.

Bacterial isolation is an effective method for achieving this, as sequenced bacteria isolates derived from individual colonies yield discrete genomes with high purity and completeness. With this rationale, we built on the comprehensive collection of 1000+ bacteria isolates with individually assembled genomes from a recent study[20] as a ground truth that is reliable, complex, and realistic to evaluate the long-read MAGs, and their accuracy in FMT strain tracking. This study involved a cohort of seven healthy donors and thirteen recipients using FMT for the treatment of rCDI. For evaluation purposes, we selected four FMT cases from this cohort (3 donors and 4 recipients) from which, 184 unique strains (105 species) were isolated, individually sequenced, and assembled[20]. In total, 25 fecal samples were taken from the donors and recipients at multiple time points before and after FMT for up to 5 years (**Supp Fig. S1**). We built on the previously generated short-read sequencing data[20], and generated new long-read sequencing data for 16 samples (**Methods**. Additionally, we generated deeper short-read data with three times the depth of the long-read data for two samples to enhance detection power and enable a more comprehensive evaluation (**Methods**).

To construct strain-level long-read MAGs, we employed metaFlye v2.9[45] for metagenomic assembly from long-read Nanopore sequencing and MaxBin2[46] for species-level binning. Furthermore, we utilized Strainberry[32] for haplotype phasing each species-level bin and

employed metaWRAP[47] for final strain-level binning. (**Supp Fig. S2; Methods**). From the three donors and four recipients, we obtained a total of 460 long-read MAGs (**Fig. 2A**), which collectively represent 460 unique strains (across 312 species), including 60 strains that were assembled from isolate genomes (**Fig. 2A**). Additional short-read Illumina sequencing data was generated for *de novo* assembly of Donor 1 (D1) and Recipient 1 pre (R1), resulting in 155 short-read MAGs, to facilitate methods evaluation. Contamination levels for all long-read MAGs and SR MAGs were assessed using CheckM, revealing significantly lower contamination in long-read MAGs (**Fig. 2B**). Furthermore, long-read MAGs exhibited significantly higher contig N50 values compared to short-read MAGs at similar relative abundances, indicating higher assembly quality (**Fig. 2C**).

In our in-depth evaluation, we focused on sample D1, which had the most bacterial isolates (136 unique strains; **Methods**), out of which 31 and 23 were shared with the 124 long-read MAGs and 135 short-read MAGs, respectively (**Fig. 2D**). Notably, long read MAGs include all 23 MAGs recovered by short-read MAGs along with 8 additional strains (**Fig. 2D**). Both long-read and short-read MAGs spanned the same 17 taxonomic orders. When compared to isolate genomes, long-read MAGs contained 8 unique orders, while isolate genomes featured 5 unique orders (**Fig. 2E**). These overlapping strains provide an opportunity to evaluate the reliability of long-read MAGs in comparison with short-read MAGs generated from the same sample, using genomes assembled from bacterial isolates as ground truths (**Methods**). The properties of long-read MAGs and short-read MAGs were thoroughly evaluated regarding their strain resolution and purity. First, we assessed their contamination levels by mapping long-read and short-read MAGs to their corresponding isolate genomes (**Fig. 2F; Supp Fig. S3A; Methods**). Counting the percentage of unaligned bases (MAGs to isolate genomes) as contamination, this method builds on the valuable isolate genomes to provide more reliable insights than estimates by CheckM (**Supp Fig. S3B)**. The contamination levels were lower for long-read MAGs (1.34%)

compared to short-read MAGs (4.77%), both for species with a single strain and those with co-existing strains (**Fig. 2E and Supp Fig. S3C and S3D**). We further evaluated the similarity between long-read and short-read MAGs and their corresponding isolate genomes using k-mer similarity, which consistently indicated long-read MAGs exhibit a higher similarity (**Supp Fig. S4A**). Importantly, long-read sequencing effectively recovered the same number of strains as bacterial culture from the same donor sample (**Fig. 2F)**. In contrast, these co-existing strains were not completely recovered in the short-read MAGs (**Fig. 2F**). It is worth noting that, despite the comparable completeness of long-read and short-read MAGs (**Supp Fig. S4B**), the higher purity of long read MAGs (**Fig. 2E**) and strain resolution is expected to be more suitable for FMT strain tracking, because even a partial MAG that reliably represents individual strains can provide unique k-mers valuable information for strain tracking[20,48], as we will also evaluate in the next section.

While the shared strains demonstrate the reliability of long-read MAGs, it's important to note that long-read MAGs and bacterial isolates capture different additional strains and species due to their inherent methodological differences. Specifically, long-read MAGs, based on metagenomic sequencing, favor strains with modest-to-high relative abundance and tend to miss low-abundance ones, while bacterial culture favors species and strains selected by culture media rather than abundance. For instances, *Akkermansia muciniphila* and *Faecalibacterium prausnitzii* (**Fig. 2A**), two species that have been associated with several diseases and treatment responses[49,50], have modest-to-high relative abundances (0.63% and 3.94% respectively) in metagenomic sequencing but were missed by conventional culture methods used in the previous study[20]. Conversely, *Escherichia coli* and *Enterococcus faecalis* were cultured from D1 but not recovered by long-read sequencing because their relative abundances were both below 0.01%. 37.1% and 65.9% of the total metagenomic reads of D1 can be mapped to the isolate genomes and long-read MAGs (**Methods**), respectively. When we

combined the isolate genomes and long-read MAGs, >76% of the total metagenomic reads can be mapped to the merged reference (**Fig. 4C**). Interestingly, the 31 common strains (13.5% of the 229 strains from long-read MAGs and cultured isolates combined) represented a substantial proportion (30.7%) of the total metagenomic reads, because of the relatively high abundance of these strains. Isolate genomes and long-read MAGs uniquely explain 6.5% and 38.77% of the total reads, respectively. These highlight the complementarity between long-read MAGs and bacterial culture, each providing valuable insights into the microbial community, as we will further investigate in the context of FMT strain tracking.

This section provides a comprehensive evaluation of long-read MAGs with a large collection of isolate genomes from matched fecal samples. We showed that long-read MAGs are more reliable than short-read MAGs for resolving strain-level MAGs *de novo* and at better purity. Our analyses also highlighted the complementarity of long-read MAGs and bacterial culture in capturing different strains and species. This section lays the groundwork for the subsequent section, which will further investigate how to reliably use long-read MAGs for precise FMT strain tracking (**Fig. 2A)**.

LongTrack enables reliable use of Long-read MAGs for FMT strain tracking

Once individual isolate genomes are assembled, strain tracking can be done reliably based on unique 31-mers selected from isolate genomes and short-read sequencing of post-FMT samples, even with modest sequencing depth, serving as a gold-standard for evaluation[20]. This k-mer based strategy (called Strainer) for strain tracking demonstrated superior performance compared to several SNP-based approaches[20] including inStrain[27], StrainFinder[23] and ConStrains[22], particularly in resolving strains with modest-to-low abundance and multiple co-existing strains. Our independent evaluation of LongTrack for strain tracking confirmed these

findings, showing consistent results with the previous study[20] in that the k-mer based approach is more sensitive than the SNP-based method: while the SNP-based method works well for strains with relatively modest-to-high abundance, it tends to miss strains with modest-to-low abundance (**Supp Fig. S5**).

However, the existing k-mer based method for strain tracking is not compatible with long-read MAGs, because while some of the long-read MAGs are nearly complete, others are not (taxa with modest-to-low abundance). K-mers selected from incomplete long read MAGs are not necessarily strain-specific and thus can lead to false strain tracking results (**Methods**). Therefore, to reliably use long-read MAGs for FMT strain tracking, we made a few important methodology adaptations to ensure a more thoughtful selection of unique k-mers to avoid false positives (**Methods**). Specifically, for similar coexisting strains in one fecal sample, or two similar strains each in the donor sample and recipient pre-FMT samples, we selected unique k-mers only from their common genomic regions (**Fig. 1C; Methods**). The rationale of this adaptation is to avoid the selection of unreliable k-mers that appear to be unique to a strain but are not, due to the incompleteness of some long-read MAGs, which is the key difference between high purity long read MAGs and the complete genomes of bacterial isolates[20]. We used negative controls to estimate the confidence of strain tracking, based on which a strain is determined to be present (confidence score >95%) or absent in a post-FMT sample (**Methods**). To evaluate the performance of LongTrack, we conducted precision and sensitivity analyses for strain tracking using long-read MAGs versus short-read MAGs across three donors and four recipients, with multiple fecal sampling time points (326 unique strain presence/absence events). LongTrack achieved perfect precision (1.00) and sensitivity (1.00) (**Fig. 3A**) when compared to isolate genomes as the gold standard [20]. In contrast, strain tracking using short-read MAGs showed lower sensitivity (0.77) (**Fig. 3A**). Additionally, we compared the number of unique k-

mers used for strain tracking in cases involving multiple co-existing strains, using the original Strainer method (with high completeness isolate genomes) versus LongTrack with high-purity long-read MAGs (**Fig. 3B**). While some k-mers from shared genomic regions were filtered out, *LongTrack* retained a sufficient number of unique k-mers, ensuring reliable strain tracking in post-FMT samples (**Fig. 3A and 3B**).Specifically, we first built upon the two comprehensively characterized FMT cases, D1->R1, and D1-> R2, for which strain tracking results are well validated against the gold standard[20]. For the 31 common strains as described in the previous section, we evaluated long-read MAGs vs. isolate genomes for strain tracking and found that they had 100% consistency (**Fig. 3C**). In contrast, short-read MAGs allowed us to track 23 of the 31 strains (74%), but failed to track the other 8 strains in post-FMT samples: this was largely because it was not able to resolve the multiple strains coexisting within the same donor sample as described in the previous section. In addition, we also observed that short-read MAGs led to six false positives tracking, which was due to the >5% contamination levels when compared to the isolate genomes. For instance, the short-read MAGs *Bifidobacterium adolescentis* and *Bacteroides uniformis* have contamination levels of 8.76% and 5.8%, respectively. Specifically, 28% of the contamination of the *B. adolescentis* is from a *Bacteroides vulgatus* strain within the sample (**Supp Fig. 6A)** which led to the subsequent unreliable selection of unique k-mers and the ultimate false positive tracking results in the Year 5 post-FMT sample from R1 (**Fig. 3C**). If short-read MAGs with >5% contamination were marked as unreliable, the sensitivity reduced from 0.77 to 0.39 (**Fig. 3A**), and accuracy of short-read MAGs for strain tracking decreased from 74% to 45% (**Supp Fig. 6B)**, while LongTrack maintained 100% consistency with the isolate genome-based strain tracking (**Supp Fig. 6B**). We also applied LongTrack to two additional rCDI FMT cases: D2->R3 and D3->R4, and also found 100% consistency with well-validated results based on isolate genomes for the strains shared between long-read MAGs and gold standard i.e. isolate genomes (**Fig. 3D and 3E**). A noteworthy observation is that completeness does not significantly affect tracking accuracy, provided the long-read MAGs are of high purity.

For instance, the *Bacteroides caccae* long-read MAG, despite having a low completeness of 17.24%, contained sufficient unique k-mers to accurately track strains in the post-FMT samples (**Fig 3C**).

To further understand the performance of LongTrack compared with Strainer with short-read MAGs, we designed additional evaluations with two mock FMT scenarios with *in silico* spike-in of specific strains into D1 and R1 (**Methods**). In the first evaluation, we spiked-in three strains of *Phocaeicola massiliensis* (Str 1-3) to D1, and a fourth strain of *P. massiliensis* (Str 4) to R1 pre-FMT (**Fig. 3F**). All four strains were added to R1 post-FMT. While LongTrack were able to recover all four strains and support accurate tracking of the four strains in R1 post-FMT, short-read MAGs only recovered two of the four strains (**Fig. 3F**). In the second evaluation, we spiked-in ten pairs of strains, each with high similarity (ANI>99.5%), from ten different species, one strain to D1 and the other to R1 pre-FMT, similarly for all the ten pairs. All twenty strains were added to R1 post-FMT (**Fig. 3H**). LongTrack correctly tracked all the ten pairs of strains in post-FMT, while Strainer with short-read MAGs missed two of the strains (**Fig. 3H**).

In summary, the above evaluations support the reliability of LongTrack for FMT strain tracking. Specifically, long-read MAGs can reliably (better than short-read MAGs) resolve co-existing strains with high purity (**Supp Fig. S2C**), and LongTrack can identify strain-specific k-mers from long-read MAGs to differentiate highly similar strains between the donor sample and the recipient pre-FMT sample in FMT strain tracking. This evaluation served as the foundation for the next section, where we used long-read MAGs to track uncultured strains to better understand strain engraftment in FMT for rCDI and IBD beyond culture-based approaches.

LongTrack helps better understand strain engraftment in rCDI and IBD patients

Building on the reliability of LongTrack-based FMT strain tracking as evaluated in the previous section, we next aimed to track the uncultured strains missed by the previous culture-based study[20] to more comprehensively understand strain engraftment in FMT. Across all the four cases of rCDI FMT, long-read sequencing recovered 396 MAGs with diverse taxonomy distribution (**Fig. 4A; Methods**) that were not captured by the previous culture-based study.

We performed the same type of analysis to measure the engraftment of donor strains in the recipients, focusing on uncultured strains from long-read. To track the strains for up to five years after FMT, we used a metric named *proportional engraftment of donor strains* (PEDS), which is calculated as the number of donor strains detected in a recipient post-FMT divided by the total number of strains identified from the donor in the long-read MAGs. We found consistent and stable engraftment of donor strains (**Fig. 4B),** and individual trajectories of strain engraftment for donor-recipient pairs (**Figs. S7A-C)**, respectively. In these individuals, we report an average PEDS of 98% of uncultured strains at the first week, which stabilized ~96% at 8 weeks (a commonly noted clinical endpoint for measuring efficacy) and remained high at 63% even 5 years later. This demonstrates that gut microbiota manipulation by FMT can lead to a fairly stable long-term engraftment of donor‘s uncultured strains in patients with rCDI.

In parallel with tracking donor strains, we also monitored the persistence of recipient strains post-FMT using a metric called *proportional persistence of recipient strains (PPRS),* defined similarly to PEDS but for recipient strains. We observed a major reduction (47%) in PPRS of recipient strains post-FMT, indicating a major shift in the gut microbiota following FMT (**Fig. 4C**). The absence of *Clostridioides difficile* in post-FMT samples from recipient R1 is noteworthy, suggesting that the transplantation led to a more balanced and health-promoting gut microbiota (**Fig. S7A**). However, some recipient strains demonstrated resilience and were still detectable years post-FMT (**Fig. S7A-C**), which highlights the dynamic nature of the gut

microbiota and underscores the potential for certain recipient strains to withstand the microbial reshuffling induced by FMT.

The above analyses of PEDS and PPRS at the final time are consistent with the previous study[20] that focused on strains resolved by bacterial cultures, but highlight the hundreds of uncultured strains identified through long-read MAGs, providing a more complete view of donor strain engraftment in FMTs for rCDI. When we examine the individual trajectories of long-read MAGs recovered strains for donor-recipient pairs (**Figs. S7A-C**), we found several strains that stably engrafted in the recipient after five years include many species linked to human health and diseases. One notable example, *A. muciniphila* (**Fig. S7AD, S7B**), initially present in donor D1 and donor D2, is a mucin-degrading bacterium that has been linked to health benefits in the context of several human diseases and treatment outcomes[50–55]. It has been shown to play a role in preventing the overgrowth of harmful pathogens such as *C. difficile* through the modulation of gut microbiota and metabolites[32]. Moreover, its presence is thought to bolster gut barrier integrity and modulate the immune response, thereby creating an environment less favorable for *C. difficile* colonization and infection[33]. *A. muciniphila* strains from donors (D1 and D2) demonstrated stable engraftment in post-FMT recipients, with persistence observed for up to five years (**Fig. S7A, S7B**). In contrast, an *A. muciniphila* strain from pre-FMT recipient R3 was not detected in the corresponding post-FMT recipient R3 four weeks after the FMT (**Fig. S7B**). *Faecalibacterium prausnitzii* (**Fig. S7A)** is another species well resolved by long-read MAGs but not the large-scale culture in the previous study[20]. Specifically, it's known for its anti-inflammatory effects and its ability to produce short-chain fatty acids like butyrate, which are beneficial for the health of the colon and the entire gastrointestinal tract[49,57–61] . This bacterium is often considered a marker of a healthy gut and has been the focus of research for its potential role in maintaining gut health and preventing inflammatory bowel diseases (IBD)[57,58,62]. *F. prausnitzii* strain from donor D1 was detected in post-FMT recipient R1 eight

weeks after FMT, but not observed five years after the FMT (**Fig. S7A**). *Ruminococcus bromii* (**Figs. S7A, S7C)** is another species tracked by long-read MAGs, known for its ability to break down resistant starches into short-chain fatty acids like butyrate[63]. These short-chain fatty acids contribute to gut health, serve as an energy source for colon cells, and have anti-inflammatory properties[63,64]. *R. bromii* strains from donors D1 and D3 showed stable engraftment in post-FMT recipients, with persistence observed up to the last collection time point (**Figs. S7A, S7C**).

Next, we expanded the tracking analysis to strains represented by long-read MAGs from recipient post-FMT samples and reversely tracked them in earlier time points (**Methods**). The rationale of this design was that some strains were indeed present in donor or recipient pre-FMT but at too low abundance to be recovered by long-read MAGs, yet their abundances can be higher at recipient post-FMT samples, hence assembled by long reads and can be reversely tracked in earlier time points to have a more complete view of strain engraftment in FMT. Indeed, this design helped us identify 17, 8 and 10 additional strains from the three FMT cases D1->R1, D2->R3, and D3-R4, respectively (**Supp Fig. S8**). All these cases have clear evidence that they were from either the donor sample or recipient pre-FMT samples but just at low abundance so were not recovered by long-read MAGs from pre-FMT samples. This strategy helps make the study of strain engraftment in FMT more comprehensive. Interestingly, in this analysis, we also observed a few cases where a novel strain was recovered from recipient post-FMT samples, but no evidence supported the presence of these strains in any of the earlier times, neither donor nor recipient pre-FMT (**Supp Fig. S9**), which suggested that these strains may have been acquired from the sourced unrelated to the donor or recipient. For example, a strain of *Acidaminococcus intestini* from the year 5 sample of R1 post-FMT was absent at every earlier time or in the donor (**Supp Fig. S9A**), was likely acquired from the environment sometime between week 8 and year 5, unless its abundance was below detections limit in all the earlier time points.

In addition to the four cases of FMT for the treatment of rCDI, we further applied the LongTrack to two FMT cases for the treatment of IBD[8,65] (**Supp Fig. 10**), which represents pre-FMT recipients with more complex microbiome than rCDI patients. Different from the FMT cases for rCDI patients, the two IBD patients (R5 and R6) each received fecal samples from six donors (D4 - D9), as multiple donors are often used in the FMT treatment of IBD considering the more diverse and resilient pre-FMT gut microbiome in IBD patients compared to rCDI patients[8,65]. This greater complexity of FMT provided a great opportunity to demonstrate the power of long-read MAGs for strain tracking. Here, we conducted long-read sequencing on two of the donors (D4 and D5) and the two pre-FMT, recovered 188 MAGs (R5: 45; R6: 52; D4: 52; D5: 39). Out of these strains, 58 (64%) and 32 (35%) strains from the two donors successfully engrafted the recipient 16 weeks post-FMT (8 weeks of FMT dosage then 8 weeks after cessation) for the two recipients, respectively (**Figs. 4D, 4E**). We also estimated PPRS post-FMT and observed that 80% and 96% of strains from recipient pre-FMT were present after 16 weeks post-FMT (**Fig. 4D, 4E**). Compared to the FMT cases for rCDI at a similar time point (8 weeks), we observed lower engraftment of donor strains and higher persistence of recipient strains post-FMT in the two FMT cases for IBD, which is consistent with the expectation that the gut microbiome of IBD patients is more diverse than those of rCDI patients, thus exhibiting a stronger resilience upon FMT introduction of donor strains[8]. Because fecal samples of IBD patients exhibited greater diversity than those from patients with rCDI, we observed several coexisting strains in the recipient post-FMT samples, where strains from donors and recipients were detected. Specifically, co-existing strains of the same species were observed for 19 and 20 species in the two IBD FMT cases (**Fig. 4F, 4H**). Notably, while nearly all strains from the pre-FMT recipient samples remained stable 16 weeks post-FMT, some donor strains were not detectable at 16 weeks post-FMT, suggesting endogenous strains in IBD patients pre-FMT are more resilient (than donor strains) in the same host after FMT.

In summary, LongTrack empowered the tracking of hundreds of additional strains missed by conventional culture from the same FMT cases in a previous study, and provided a more comprehensive view of strain engraftment in six FMT cases across rCDI and IBD patients. The newly resolved strains that stably engrafted the recipient post-FMT include species that have been implicated with healthy benefits across several human diseases, such as *A. muciniphila* and *F. prausnitzii.* The complex co-existing strains post-FMT in the two IBD FMT cases also demonstrated the power of long-read MAGs based strain tracking for more complex diseases beyond rCDI.

Long-read sequencing of post-FMT samples uncovered genomic structural variations within engrafted strains.

In addition to determining strain presence/absence during FMT, it's crucial to understand how donor strains adapt to the new microbial and host environments in the recipient. The pressure to adapt can drive the selection for genomic variations in these donor strains, a topic of interest but not yet extensively studied in FMT research[66–68]. While short-read sequencing is powerful in detecting single nucleotide variations (SNVs) and small indels, long-read sequencing has great advantages in the detection of more complex structural variations (SVs, such as insertions, deletions, and inversions). In addition, another advantage of third-generation long-read sequencing technology, represented by the Nanopore and PacBio platforms, is that they allow the direct detection of bacterial DNA methylation[39–43,69]. Therefore, we aimed to use long-read sequencing of donor and post-FMT samples to assess the stability and dynamics of SVs and DNA methylation of individual strains across the longitudinal samples collected over a five-year post-FMT.

For SV analyses, we focused on the 255 long-read MAGs and 41 isolated genomes without co-existing strains from the rCDI donors (D1-D3), because genomic analysis of single strain simplifies the calling and interpretation of strain-level SVs in this exploration. A systematic analysis of long-read sequencing data of the post-FMT recipients (R1-R4) across various time points led to the detection of 38 SVs, which included 6 deletions, 20 insertions, and 12 inversions, as determined by stringent filtering criteria to maximize reliability of genome-specific read mapping (**Methods**). Of these, 32 SVs were detected in eight strains that were *de novo* assembled into near-complete long-read MAGs in both donor and post-FMT samples. Importantly, comparison of paired donor and post-FMT samples confirmed that the SVs originated from the same strain, with genome-wide ANI >99.70% (mean ANI = 99.91%). The remaining 6 SVs were detected in a strain that could not be *de novo* assembled into a high-completeness MAG. Nevertheless, rigorous local genomic examination confirmed that these SVs also originated from the same strain, with local ANI >99.77% (mean ANI = 99.81%). Further comparison of read alignments of SV regions— specifically, a 1323 bp insertion in *Alistipes onderdonkii* and a 1270 bp deletion in *Alistipes putredinis*—showed that both two long-read sequencing methods provided clear and consistent flanking sequence alignments (**Fig. 5A**), confirming they were from the same strain. In contrast, short-read sequencing plots displayed multiple mismatches, making it challenging to reliably attribute the sequences to the same strain (**Fig. 5A**),

Of the 38 SVs (**Fig. 5B**), 15 were located within genic regions, while 23 were found in intergenic regions. Notably, 5 SVs were identified in TonB-dependent receptor genes or in their upstream regions (**Fig. 5B**). These receptors, predominantly located in the outer membrane of Gram-negative bacteria, are integral for nutrient transport, particularly iron, into bacterial cells[70]. Previous research has suggested that deletions in these TonB-dependent receptor genes might be associated with immune-mediated selective pressures[71]. For example, an inversion in a

strain of *Bacteroides stercoris* from D1 had high frequency exclusively in late-stage post-FMT sample, across TonB-dependent receptor SusC genes, which may be associated with immune adaptation to the new host environment **(Fig. 5C**). In *A. putredinis* D1, the deletion shown in **Fig 5C** encompasses an IS110 family transposase gene designated ISCaa7, was observed with a high prevalence post-FMT **(Fig. 5C**). The ISCaa7 gene encodes a transposase, an enzyme that mediates the movement of transposons (or jumping genes) within the genome of a single cell[72]. This alteration may affect genetic diversity, evolution, and antibiotic resistance propagation, highlighting the impact on microbial adaptability[73]. Moreover, in *A. onderdonkii* D1, an insertion proximal to the RhaS gene **(Fig. 5C**), encoding an HTH-type transcriptional activator, was observed with a high prevalence exclusively in donor D1 and in R1 early-stage post-FMT, suggesting possible involvement in the modulation of gene expression **(Fig. 5C**). RhaS is pivotal in the catabolism of rhamnose through the activation of related genes[74], and its involvement suggests a possible role in environmental response and metabolic adaptation.

We also observed some more complex SVs that involve bacterial restriction-modification systems (RMS) that can influence global changes in DNA methylation, which can impact bacterial gene expression and adaptation to different environments[41,44,75–78]. For example, we detected a complex SV spanning a Type I RMS S subunits of *Parabacteroides merdae.* The nested inverted repeats create a complex genomic shuffling among six different genotypes (**Fig. 5D**). Notably, genotype 5, predominantly observed in the donor, had a drastic reduction post-FMT, while genotype 6 was absent in the donor but significantly increased in post-FMT recipient. Because the specificity unit, S, of the Type I RMS system determines the sequence recognition motif of the methyltransferase, the genotype switches of this RMS locus can create genome-wide methylation changes[41]. To test this hypothesis, we used PacBio sequencing data, which offers advantages in detecting 6-methyladenine (6mA) at single molecule resolution[42,69,76], and indeed observed a DNA methylation motif CA6mAYNNNNNNRTC exclusively in post-FMT

samples (**Fig. 5E**). This observation suggests some SVs can influence genome-wide changes in methylation, which might empower the *P. merdae* strain to better adapt to the new gut environment of the recipient post-FMT. Further investigation revealed that the other DNA methylation motifs in this *P. merdae* strain were stable between donor and recipient across different time points (**Fig. 5E**). In addition, we examined 11 unique methylation motifs across 10 single strains with high-coverage ONT long-read data in Donor D1 (**Fig. 5F**). Among these, 1 motif was identified as 5-methylcytosine (5mC), 2 as 4-methylcytosine (4mC) motifs, and 8 as 6mA motifs, all of which conform to Type II methylation patterns. These motifs demonstrated consistency and stability in both donors and recipients over various time points, which correlates with the absence of SVs in their flanking genomic regions.

In summary, to assess the genomic stability of engrafted strains, we built on the strengths of long-read sequencing data from post-FMT samples to detect both complex genomic SVs and epigenomic variations. Some of these variations are observed in genes with functions relevant to bacterial adaption to the new host or nutrition factors, which may play a crucial role in bacterial adaptability, and gene expression, and may influence pathogenicity and antibiotic resistance.

Discussion

In this study, we present a novel framework, LongTrack, that uses long-read sequencing and tailored informatics for FMT strain tracking, addressing a critical gap in the comprehensive tracking of microbial strains.

Technological and methodological advancements: We have employed long-read MAGs with higher confidence in taxonomic assignments[79,80] to overcome the limitations of short-read sequencing in the *de novo* assembly of co-existing, highly related bacterial strains. Long-read

MAGs demonstrated superior reliability in strain differentiation, crucial for subsequent strain tracking using short-read sequencing. A major methodological innovation of LongTrack is the tailored approach for the selection of strain-specific k-mers from long-read MAGs that effectively leverage the higher purity of long-read MAGs and reliably handle the incompleteness of long-read MAGs (for low-abundance strains) compared to separately assembled bacterial isolate genomes[20]. Based on strain-specific k-mers selected by LongTrack from long-read MAGs, subsequent strain tracking can be reliably performed using short-read sequencing data. This LongTrack framework was rigorously validated against gold standard methods based on isolate genomes from large-scale bacterial culture. While previous isolate-based methods[20] have inherent biases—for instance, strains like as *A. muciniphila* and *F. prausnitzii* can theoretically be cultured from stool samples under more stringent conditions—the inclusion of isolate genomes in our study was primarily intended to facilitate the evaluation of long-read MAGs versus short-read MAGs, particularly in resolving co-existing conspecific strains.

Comprehensive application to FMT cases for both rCDI and IBD patients demonstrated the method's applicability across diverse microbiome complexities. Notably, the application of long-read sequencing allowed the identification of a large number of strains across diverse taxonomy, many previously uncultured strains, such as as *A. muciniphila* and *F. prausnitzii* [20], and revealed a more complete insights into their long-term engraftment in rCDI patients up to five years post-FMT. In IBD FMT trios, this approach demonstrated its effectiveness in more complex microbiomes with greater strain diversity, and the FMT from multiple donors to IBD patient.

Beyond identifying strains that are engrafted post-FMT, it is also imperative to explore the genomic dynamics of donor strains that might reflect their adaptation to their new surroundings within the recipient's microbial and host environments. This adaptation process, which may select for certain genomic changes in the engrafted strains, remained a relatively

unexplored area in FMT research. A number of recent studies highlighted the importance of SVs (insertions, deletions, and inversions), in addition to SNVs, in microbiome adaptation[66,81,82]. However, reliable detection of strain-level SVs, especially large SVs, across the longitudinal samples poses a challenge to short-read sequencing but can be better approached by long-read sequencing[83]. Consequently, our study has utilized long-read sequencing to examine the stability and variations in SVs in individual strains across longitudinal samples collected over five years following FMT, unveiling dynamic changes over time may reflect strain adaptation to the new environment. For some of the SVs that involve bacterial DNA methylation systems, we further investigate their impact on genome-wide DNA methylation building on the unique capability of third-generation long-read sequencing for bacterial epigenomic analysis. Our findings suggest a new dimension to characterize how donor strains adapt in post-FMT recipients over time, which may lead to enhanced efficacy of defined live biotherapeutics.

While LongTrack offers novel insights, it does inherently face limitations in recovering super low abundance strains. To mitigate this, we propose two complementary strategies. First, we would like to emphasize the complementarity and integration between long-read sequencing and culture-based methods that can provide a more comprehensive view of the FMT strain engraftment. This synergistic approach is especially valuable for tracking strains that are either challenging to culture or present in low abundance, thus ensuring a more complete microbial representation. Second, the challenges posed by low-abundance strains to metagenomic approaches can be mitigated by targeted enrichment methods. Specifically, recent studies have demonstrated efficient enrichment of specific taxa of interests based on flow cytometry[84] or bacterial methylome patterns[85]. These enriched subsets can then be sequenced and analyzed separately, enhancing the recovery and tracking of low abundance strains.

Figure legend

Figure 1. Description of FMT samples, matched bacteria isolates and the core design of LongTrack for reliable FMT strain tracking and detecting intra-strain structural variations in engrafted strains.

A, Stool samples were collected from three FMT donors and four FMT recipients at multiple time points over five years. Comprehensive bacterial culture, isolate genome assemblies, and short read-metagenomic sequencing have been performed in the previous study by Aggarwala et al, and shown along with FMT samples. Long read metagenomic sequencing data were newly generated for this study. **B**, Long-read data from donors' and recipients' samples before FMT were used to construct *de novo* metagenome-assembled genomes (long-read MAGs). The overlapping isolate genomes enable a comprehensive evaluation of the reliability of long-read MAGs, which serves for the foundation of LongTrack. **C**, Schematics of the core design of LongTrack for FMT strain tracking based on strain-level long-read (LR) MAGs. Strain-unique k-mers (k=31) were identified for each strain-level LR MAG from donor or pre-FMT recipient as the markers by eliminating common k-mers present in public database, or the other MAGs in the same FMT donor and recipient (**Methods**). To ensure specificity, strain-unique k-mers are only selected from shared regions across multiple strain-level LR MAGs within the same species. The presence of these final unique k-mers of a specific MAG (above a confidence score; **Methods**) in a post-FMT recipient's short-read metagenomic data indicates that the corresponding MAG is present (green color) in the post-FMT recipient sample. **D**, Long-read metagenomic sequencing data were also utilized for strain-specific detection of structure variations (SVs) that differ between donor and post-FMT samples. The references for putative genotypes were created and used to map and quantify specific SVs in each sample, allowing tracking of SV changes between hosts and/or along time. The long-read approach has unique advantage in strain-level read mapping that allows reliably detection of intra-strain variations,

which are challenging for short reads due to their limitations in accurately mapping reads to specific strains.

Figure 2. High-purity, strain-level long-read MAGs serve as a solid foundation for LongTrack.

A, The Venn diagram displays the number of strains that were recovered by long-read (LR) MAGs (yellow color) or isolate genomes from bacterial culture (red color) from 3 donors and 4 recipients, with the left bar plot illustrating examples of species with MAGs with modest-to-high relative abundance but were missed by culture, and with right bar plot illustrating examples of isolates that were cultured but not recovered by long-read sequencing because of their very low abundances. **B**, Contamination levels of 460 LR MAGs (from three donors and four recipients) and 135 SR MAGs (from Donor 1, D1, for which we generated ultra-deep data) assessed using CheckM, showing significantly lower contamination in LR MAGs ($p < 0.01$, **). **C**, Distributions of relative abundance (log scale) and contig N50 (kb) for LR MAGs and SR MAGs. **D**, The Venn diagram displays the number of strains that can be recovered by LR MAGs (yellow color), short-read (SR) MAGs (blue) or isolate genomes (red) in D1. LR MAGs include all the 23 strains recovered by SR MAGs, along with 8 additional strains not assembled by short reads. **E,** Order-level taxonomic comparison of LR MAGs, SR MAGs, and isolate genomes from D1. **F**, Contamination level compared to corresponding isolate genomes for 60 LR MAGs (from Donor 1-3 and Recipient 1-4) and 23 SR MAGs (from D1), evaluated across various species with a single strain or multiple strains. LR MAGs consistently show lower contamination. **H**, Long reads (LR) sequencing have an unique advantage to resolve multiple co-existing strains from the same species than short reads (SR) sequencing.

Figure 3. Comprehensive evaluation of LongTrack for FMT strain tracking.

A, Precision and sensitivity of strain tracking using LongTrack (methods tailored for long read MAGs) versus Strainer with short-read MAGs (SR MAGs). Across 3 donors and 4 recipients with multiple fecal sampling time points (326 unique strain presence/absence events), LongTrack achieved 100% consistency with isolate-based strain tracking, even under stringent quality filtering (**Methods**). In contrast, Strainer with SR MAG-based strain tracking (n = 186) showed lower sensitivity (0.77 overall, reduced to 0.39 under stringent filtering). **B**, Numbers of unique k-mers used for strain tracking in cases of multiple co-existing strains by the original Strainer method (with complete genomes assembled for isolates) and the LongTrack method. Although some k-mers from shared regions are filtered out by LongTrack, enough unique k-mers remain for reliable strain tracking in post-FMT samples. **C**, Presence (green) or absence (gray) of strains in post-FMT recipients determined by strain-specific unique k-mers selected from the isolate genomes (considered the *gold standard*), LR MAGs, or SR MAGs. Discrepancies with the gold standard are indicated by an "X" marker. **(C-E).** The evaluation of strain tracking was performed for four FMT cases of rCDI: D1->R1 (**C**), D1->R2 (**C**), D2->R3 (**D**), and D3->R4 (**E**). **F**, Further evaluation of FMT strain tracking using *in silico* spike-in data that represent the co-existence of multiple strains of the same species in the donor sample. Specifically, we spiked-in three strains of *Phocaeicola massiliensis* (Str 1-3) to D1, and a fourth strain of P. massiliensis (Str 4) to R1 pre-FMT. All four strains were added to R1 post-FMT. **H**, Further evaluation of FMT strain tracking using *in silico* spike-in data that represent cases when there are high-similarity strains present in between the donor and pre-FMT samples using *in silico* spike-in data. Specifically, we spiked-in ten pairs of strains, each with high similarity (ANI>99.5%), from ten different species, one strain to D1 and the other to R1 pre-FMT, similarly for all the ten pairs. All twenty strains were added to R1 post-FMT.

Figure 4. LongTrack uncovered hundreds of engrafted strains across six FMT cases of rCDI and IBD patients.

A, Diverse taxonomy distribution of long read MAGs that are not cultured in the previous study by Aggarwala *et al*. **B**, Proportional engraftment of donor strains (PEDS) in post-FMT recipients at various time points of the 5-year follow up. **C**, Proportional persistence of recipient strains (PPRS) in post-FMT recipients at various time points of the 5-year follow up. **D** & **E**, PEDS and PPRS analysis for the two FMT cases for IBD patients R5 (**D**) and R6 (**E**), where three colors are used to represent the strains from the two Donors (D4 and D5) or the recipient. (**F** & **H**) The presence and absence of co-existing strains in the two IBD patients (R5, **F**; R6, **H**) after FMT that originated from the two Donors (D4 and D5) or the recipient.

Figure 5. Long reads uncovered complex structural variations in engrafted strains during the 5-year follow-up after rCDI FMT.

A, IGV plots illustrating the advantage of long reads in the detection of intra-strain structural variation (SV) detection, including a 1323 bp insertion in *Alistipes onderdonkii* and a 1270 bp deletion in *Alistipes putredinis*, over short-read sequencing. Flanking sequences of SVs show clear and consistent alignments in all long-read IGV plots, indicating they are from the same strain. In contrast, short-read plots display multiple mismatches, making it difficult to reliably determine whether the sequences belong to the same strain. **B**, The heatmap plot summarizing all the SVs detected in the donors and post-FMT recipients at early (E) and late (L) time points after FMT including inversions (INV), deletions (DEL), and insertions (INS). Samples with insufficient coverage over a specific SV region are marked by an asterisk (*). SVs are also categorized by their location: intergenic regions (IR) and genic regions (GR). **C**, Three examples of SVs with changes in the frequency of the two genotypes across donor and post-FMT recipient at different time points. The first one illustrates an inversion (detected in *Bacteroides*

stercoris), the second one illustrates deletion (detected in *Alistipes putredinis*) changes in genotype frequency both in the early and late stage post-FMT, while the third one shows insertion (detected in *Alistipes onderdonkii*) and change in genotype frequency only at late stage post-FMT. Pie charts demonstrate the frequency of two genotypes shown across the donor (D1) and post-FMT recipient (R1) at early (E) and late (L) time points after FMT. **D**, A complex inversion involving six genotypes is located on restriction endonuclease subunit S of the RMS system in *Parabacteroides merdae*. The frequency of the genotypes varies across donor and post-FMT recipient at different time points, with Genotype 6 highlighted (green) for its increases after FMT. **E**, DNA adenine methylation (6mA) motifs in the *P. merdae* strain detected across donor and post-FMT recipient with different time points; CA6mAYNNNNNNRTC corresponds to Genotype 6 of this RMS phase variant, which is only detectable after FMT. **F**, DNA methylation motifs detected across several single strains assessing their epigenomic stability in the donor (D1) and post-FMT recipient (R1) over time.

Acknowledgments

This work was supported in part by the staff and resources of the Microbiome Translational Center and Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. This work was supported by grant no. R35 GM139655 (G.F.) from the National Institutes of Health. G.F. is a Hirschl Research Scholar by Irma T. Hirschl/Monique Weill-Caulier Trust and a Nash Family Research Scholar.

Competing interests

S.P. has served as a consultant / steering committee member for Vedanta Biosciences and has received speaker / advisory board fees from AbbVie, Dr Falk Pharma, Ferring, Janssen and Takeda. J.F. is a consultant for Vedanta Biosciences and Genfit. The remaining authors declare no competing interests.

Author contributions

Initiated study: G.F. and J.F. Developed methodology: Y.F. and G.F. Data analysis: Y.F., J.F. and G.F. Data interpretation: Y.F., M.N., V.A., E.A.M., M.K., L.C., J.F. and G.F. Collection and interpretation of clinical data: M.A.K., T.J.B., S.P., N.O.K. and A.G. Supervised research: G.F. Wrote first draft of paper: Y.F. and G.F. Approved paper: all authors.

Reference

1. Feuerstadt, P. *et al.* SER-109, an Oral Microbiome Therapy for Recurrent *Clostridioides difficile* Infection. *N Engl J Med* **386**, 220–229 (2022).
2. Khanna, S. *et al.* Efficacy and Safety of RBX2660 in PUNCH CD3, a Phase III, Randomized, Double-Blind, Placebo-Controlled Trial with a Bayesian Primary Analysis for the Prevention of Recurrent *Clostridioides difficile* Infection. *Drugs* **82**, 1527–1538 (2022).
3. Kelly, C. R. *et al.* Effect of Fecal Microbiota Transplantation on Recurrence in Multiply Recurrent *Clostridium difficile* Infection: A Randomized Trial. *Ann Intern Med* **165**, 609–616 (2016).
4. Ooijevaar, R. E., Terveer, E. M., Verspaget, H. W., Kuijper, E. J. & Keller, J. J. Clinical Application and Potential of Fecal Microbiota Transplantation. *Annu Rev Med* **70**, 335–351 (2019).
5. Li, S. S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586–589 (2016).
6. Pamer, E. G. Fecal microbiota transplantation: effectiveness, complexities, and lingering concerns. *Mucosal Immunol* **7**, 210–214 (2014).
7. Bethlehem, L. *et al.* Microbiota therapeutics for inflammatory bowel disease: the way forward. *Lancet Gastroenterol Hepatol* **9**, 476–486 (2024).
8. Paramsothy, S. *et al.* Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. *The Lancet* **389**, 1218–1228 (2017).
9. Rossen, N. G. *et al.* Findings From a Randomized Controlled Trial of Fecal Transplantation for Patients With Ulcerative Colitis. *Gastroenterology* **149**, 110-118.e4 (2015).
10. Chu, N. D. *et al.* Dynamic Colonization of Microbes and Their Functions after Fecal Microbiota Transplantation for Inflammatory Bowel Disease. *mBio* **12**, e00975-21.

11. Sevcikova, A. *et al.* The Impact of the Microbiome on Resistance to Cancer Treatment with Chemotherapeutic Agents and Immunotherapy. *Int. J. Mol. Sci.* **23**, 488 (2022).

12. Davar, D. *et al.* Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. *Science* **371**, 595–602 (2021).

13. Routy, B. *et al.* Fecal microbiota transplantation plus anti-PD-1 immunotherapy in advanced melanoma: a phase I trial. *Nat Med* **29**, 2121–2132 (2023).

14. Zhao, S. *et al.* Adaptive evolution within the gut microbiomes of healthy people. *Cell Host Microbe* **25**, 656-667.e8 (2019).

15. Ghensi, P. *et al.* Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics. *npj Biofilms Microbiomes* **6**, 1–12 (2020).

16. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).

17. Porcari, S. *et al.* Key determinants of success in fecal microbiota transplantation: From microbiome to clinic. *Cell Host & Microbe* **31**, 712–733 (2023).

18. 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).

19. Duvallet, C. *et al.* Framework for rational donor selection in fecal microbiota transplant clinical trials. *PLOS ONE* **14**, e0222881 (2019).

20. 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).

21. Podlesny, D. *et al.* Identification of clinical and ecological determinants of strain engraftment after fecal microbiota transplantation using metagenomics. *Cell Rep Med* **3**, (2022).

22. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* **33**, 1045–1052 (2015).

23. Smillie, C. S. *et al.* Strain-tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* **23**, 229-240.e5 (2018).

24. Tamburini, F. B. *et al.* Precision Identification of Diverse Bloodstream Pathogens in the Gut Microbiome. *Nat Med* **24**, 1809–1814 (2018).

25. Schmidt, T. S. B. *et al.* Drivers and determinants of strain dynamics following fecal microbiota transplantation. *Nat Med* **28**, 1902–1912 (2022).

26. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, e65088 (2021).

27. Olm, M. R. *et al.* InStrain enables population genomic analysis from metagenomic data and sensitive detection of shared microbial strains. *Nat Biotechnol* **39**, 727–736 (2021).

28. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, 1633–1644 (2023).

29. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat Rev Microbiol* **18**, 491–506 (2020).

30. Gounot, J.-S. *et al.* Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. *Nat Commun* **13**, 6044 (2022).

31. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation – a benchmark of computational metagenomics software. *Nat Methods* **14**, 1063–1071 (2017).

32. Vicedomini, R., Quince, C., Darling, A. E. & Chikhi, R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat Commun* **12**, 4485 (2021).

33. Tamburini, F. B. *et al.* Short- and long-read metagenomics of urban and rural South African gut microbiomes reveal a transitional composition and undescribed taxa. *Nat Commun* **13**, 926 (2022).

34. Ghazi, A. R., Münch, P. C., Chen, D., Jensen, J. & Huttenhower, C. Strain Identification and Quantitative Analysis in Microbial Communities. *J Mol Biol* **434**, 167582 (2022).

35. Yan, Y., Nguyen, L. H., Franzosa, E. A. & Huttenhower, C. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Medicine* **12**, 71 (2020).

36. Tidjani Alou, M. *et al.* State of the Art in the Culture of the Human Microbiota: New Interests and Strategies. *Clin Microbiol Rev* **34**, 10.1128/cmr.00129-19 (2020).

37. Kim, C. Y., Ma, J. & Lee, I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat Commun* **13**, 6367 (2022).

38. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* **38**, 701–707 (2020).

39. Tourancheau, A., Mead, E. A., Zhang, X.-S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat Methods* **18**, 491–498 (2021).

40. Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* **36**, 61–69 (2018).

41. Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet* **20**, 157–172 (2019).

42. Kong, Y. *et al.* Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science* **375**, 515–522 (2022).

43. Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* **30**, 1232–1239 (2012).

44. Sánchez-Romero, M. A. & Casadesús, J. The bacterial epigenome. *Nat Rev Microbiol* **18**, 7–20 (2020).

45. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**, 1103–1110 (2020).

46. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

47. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

48. Faith, J. J. *et al.* Strain population structure varies widely across bacterial species and predicts strain colonization in unrelated individuals. 2020.10.17.343640 Preprint at https://doi.org/10.1101/2020.10.17.343640 (2020).

49. Leylabadlo, H. E. *et al.* The critical role of *Faecalibacterium prausnitzii* in human health: An overview. *Microbial Pathogenesis* **149**, 104344 (2020).

50. Cani, P. D., Depommier, C., Derrien, M., Everard, A. & de Vos, W. M. *Akkermansia muciniphila*: paradigm for next-generation beneficial microorganisms. *Nat Rev Gastroenterol Hepatol* **19**, 625–637 (2022).

51. Derrien, M., Belzer, C. & de Vos, W. M. *Akkermansia muciniphila* and its role in regulating host functions. *Microbial Pathogenesis* **106**, 171–181 (2017).

52. Bae, M. *et al. Akkermansia muciniphila* phospholipid induces homeostatic immune responses. *Nature* **608**, 168–173 (2022).

53. Wu, Z. *et al. Akkermansia muciniphila* Ameliorates *Clostridioides difficile* Infection in Mice by Modulating the Intestinal Microbiome and Metabolites. *Front Microbiol* **13**, 841920 (2022).

54. Zheng, M. *et al.* The role of *Akkermansia muciniphila* in inflammatory bowel disease: Current knowledge and perspectives. *Front. Immunol.* **13**, (2023).

55. Derosa, L. *et al.* Intestinal *Akkermansia muciniphila* predicts clinical response to PD-1 blockade in patients with advanced non-small-cell lung cancer. *Nat Med* **28**, 315–324 (2022).

56. Everard, A. *et al.* Cross-talk between *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. *Proc Natl Acad Sci U S A* **110**, 9066–9071 (2013).

57. Cao, Y., Shen, J. & Ran, Z. H. Association between *Faecalibacterium prausnitzii* Reduction and Inflammatory Bowel Disease: A Meta-Analysis and Systematic Review of the Literature. *Gastroent Res Pract* **2014**, e872725 (2014).

58. Miquel, S. *et al. Faecalibacterium prausnitzii* and human intestinal health. *Curr Opin Microbiol* **16**, 255–261 (2013).

59. Zhao, H. *et al.* Systematic review and meta-analysis of the role of *Faecalibacterium prausnitzii* alteration in inflammatory bowel disease. *J Gastroenterol Hepatol* **36**, 320–328 (2021).

60. Sokol, H. *et al. Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A* **105**, 16731–16736 (2008).

61. Zhang, J. *et al.* An immune-competent human gut microphysiological system enables inflammation-modulation by *Faecalibacterium prausnitzii. npj Biofilms Microbiomes* **10**, 1–11 (2024).

62. Quévrain, E. *et al.* Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease. *Gut* **65**, 415–425 (2016).

63. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. *Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *ISME J* **6**, 1535–1543 (2012).

64. Abell, G. C. J., Cooke, C. M., Bennett, C. N., Conlon, M. A. & McOrist, A. L. Phylotypes related to *Ruminococcus bromii* are abundant in the large bowel of humans and increase in response to a diet high in resistant starch. *FEMS Microbiol Ecol* **66**, 505–515 (2008).

65. Chen-Liaw, A. *et al.* Gut microbiota strain richness is species specific and affects engraftment. *Nature,* 10.1038/s41586-024-08242-x (2024).

66. Durrant, M. G. & Bhatt, A. S. Microbiome genome structure drives function. *Nat Microbiol* **4**, 912–913 (2019).

67. Behling, A. H. *et al.* Horizontal gene transfer after faecal microbiota transplantation in adolescents with obesity. *Microbiome* **12**, 26 (2024).

68. Moss, E. L. *et al.* Long-term taxonomic and functional divergence from donor bacterial strains following fecal microbiota transplantation in immunocompromised patients. *PLOS ONE* **12**, e0182585 (2017).

69. Kong, Y., Mead, E. A. & Fang, G. Navigating the pitfalls of mapping DNA and RNA modifications. *Nat Rev Genet* **24**, 363–381 (2023).

70. Noinaj, N., Guillier, M., Travis J. Barnard & Buchanan, S. K. TonB-Dependent Transporters: Regulation, Structure, and Function. *Annual Review of Microbiology* **64**, 43–60 (2010).

71. Dingemans, J. *et al.* The deletion of TonB-dependent receptor genes is part of the genome reduction process that occurs during adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis lung. *Pathog Dis* **71**, 26–38 (2014).

72. Babakhani, S. & Oloomi, M. Transposons: the agents of antibiotic resistance in bacteria. *J Appl Microbiol* **58**, 905–917 (2018).

73. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev* **31**, 10.1128/cmr.00088-17 (2018).

74. Kolin, A., Jevtic, V., Swint-Kruse, L. & Egan, S. M. Linker Regions of the RhaS and RhaR Proteins. *J Bacteriol* **189**, 269–271 (2007).

75. Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiol Mol Biol Rev* **77**, 53–72 (2013).

76. Beaulaurier, J. *et al.* Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat Commun* **6**, 7438 (2015).

77. Srikhanta, Y. N., Fox, K. L. & Jennings, M. P. The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat Rev Microbiol* **8**, 196–206 (2010).

78. Oliveira, P. H. & Fang, G. Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria. *Trends Microbiol* **29**, 28–40 (2021).

79. Gehrig, J. L. et al. evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microbial Genomics* **8**, 000794 (2022).

80. Kuleshov, V. et al. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat Biotechnol* **34**, 64–69 (2016).

81. Zhernakova, D. V. *et al.* Host genetic regulation of human gut microbial structural variation. *Nature* **625**, 813–821 (2024).

82. West, P. T., Chanin, R. B. & Bhatt, A. S. From genome structure to function: insights into structural variation in microbiology. *Curr Opin Microbiol* **69**, 102192 (2022).

83. Zhou, X. *et al.* De Novo Structural Variations of *Escherichia coli* Detected by Nanopore Long-Read Sequencing. *Genome Biol Evol* **15**, evad106 (2023).

84. Ji, P., Zhang, Y., Wang, J. & Zhao, F. MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat Commun* **8**, 14306 (2017).

85. Cao, L. *et al.* mEnrich-seq: methylation-guided enrichment sequencing of bacterial taxa of interest from microbiome. *Nat Methods* **21**, 236–246 (2024).

Materials and Methods

Fecal sample collection and DNA extraction

The adult fecal samples were from a previous study[1] with Institutional Review Board approval (rCDI: HS# 11-01669, IBD: HREC/13/SVH/69). Fecal DNA was extracted with a QIAamp PowerFecal Pro DNA Kit (Qiagen, 51804) according to the manufacturer's instructions. Briefly, 64 - 436 (average is 181) mg of fecal sample, and 800 µl of Solution CD1 were added to the Bead Tube (provided). After gentle vortexing, the sample was homogenized with the TissueLyser II machine (Qiagen, 85300) for 10 min at 25 Hz speed (5 min each side). The homogenized samples were centrifuged to collect the supernatant at 15,000 x g for 1 min. The subsequent column-based extraction steps were performed as described in the QIAamp PowerFecal Pro DNA Kit protocol.

Library preparation and sequencing

Nanopore long-read sequencing

Native libraries were prepared using the Oxford Nanopore Technology Inc. (New York, NY, USA) Native barcoding kit (EXP-NBD104 and 114 for R9 flow cells, and SQK-NBD114.24 for R10 flow cells), Ligation Sequencing Kit (SQK-LSK109 for R9, and SQK-LSK114 for R10). Prior to library preparation, genomic DNA (gDNA) was sheared using Covaris g-tubes (Covaris, Woburn, MA, USA 520079) at 6000 RPM in an Eppendorf 5424R centrifuge (Eppendorf, Enfield, CT, USA, 5406000640) and cleaned up with AMPure XP beads (Beckman Coulter, Indianapolis, IN, USA, A63882) at 0.8 x concentration. Whole genome amplification (WGA) samples were prepared using REPLI-g Mini Kits (QIAGEN, Hilden, Germany, 150023) according to the protocol with 12.5 ng of input DNA per sample and 16 h incubation at 30 °C, followed by denaturation at 65 °C for 3 minutes. Next, WGA samples were treated with T7 endonuclease I (NEB M0302S)

to maximize nanopore sequencing yield according to ONT documentation, then cleaned up with AMPure XP beads as above. Subsequently, samples were processed the same as above native libraries. Samples were sequenced on R9.4.1 and R10.4.1 MinION flow cells and post-run basecalled using Guppy (v5.0.7 for R9.4.1 data and 6.3.4 for R10.4.1 data)[2] with the SUP model dna_r9.4.1_450bps_sup.cfg and dna_r10.4.1_e8.2_400bps_sup.cfg, respectively.

Illumina short-read sequencing

Library construction for Illumina sequencing was executed utilizing the seqWell Plexwell 384 kit. DNA was systematically barcoded, and ligation products underwent purification. Size selection and PCR amplification of the DNA fragments followed. The prepared library was then sequenced on an Illumina HiSeq platform, generating paired-end reads of 150 base pairs (bp) each.

PacBio long-read sequencing

DNA SMRTbell libraries were prepared according to the manufacturer's instructions for the SMRTbell Express Template Prep Kit 2.0, then each library was annealed to sequencing primer together with the Sequel 2.1 polymerase, followed by loading to the a 8M Sequel II sequencing chip (for Sequel II) and 25M Revio sequencing chip.

Metagenomic assembly, strain-level separation and taxonomic annotation

Long-read metagenomic assembly

Initially, the raw long reads were subjected to a metagenomic assembly for long-read MAGs process using the metaFlye tool (v2.9)[3] with "---nano-raw --meta --keep-haplotypes --iterations 2". The resulting metagenome-assembled genomes (MAGs) were partitioned into individual bins corresponding to the species level using the MaxBin (v2.0) software[4]. To annotate the species-level bins, we applied the Genome Taxonomy Database Toolkit (GTDB-tk v1.6.0)[5]. For each

species-level bin, we utilized Strainberry (v1.1)[6] to separate the contigs into the strain level based on haplotype phasing. Subsequently, these separated contigs were binned to strain level using the "binning" module of metaWRAP (v1.3.2)[7], which could combine the results from three metagenomic binning software - MaxBin2[4], metaBAT2[8], and CONCOCT[9]. The quality of the reconstructed long-read MAGs was evaluated using the "bin_refinement" module in metaWRAP (v1.3.2)[7], which combines the results of multiple binning algorithms and selects high-quality bins based on completeness and contamination estimates, as well as estimated by CheckM (v.1.1.3)[10]. The strain-level long-read MAGs with completeness of more than 10% were selected for further analysis.

Short-read metagenomic assembly

We conducted a metagenomic assembly of the short reads using metaSPAdes (v3.15.2)[11] to construct short-read MAGs. Subsequently, we utilized metaWRAP (v1.3.2)[7] to bin and refine the reconstructed short-read MAGs. The completeness and contamination level for the short-read MAGs were estimated using CheckM (v.1.1.3)[10], and only those with completeness of more than 10% were selected for further analysis. Taxonomic assignment of the reconstructed short-read MAGs was performed using GTDB-tk (v1.6.0)[5].

Evaluation of long-read MAGs

The previous study identified a total of 156 unique strains, which were from 156 different bacterial isolates previously isolated from three distinct donors[1]. Specifically, 136 bacterial isolates were from Donor 1 (D1), 13 were from Donor 2 (D2), and 7 were from Donor 3 (D3). The 156 bacterial isolates were either directly isolated from the donors or from post-FMT recipients, which could be traced back to their original donors. The resulting assemblies for each isolate serve as reference individual genomes. To detect shared strains between long-read MAGs/short-read MAGs and genomes of bacterial isolates, we utilized fastANI (1.32)[12] for

average nucleotide identity (ANI) calculation, applying a 99.4 % cut-off[3]. We assessed the contamination level by mapping long-read and short-read MAG to its corresponding isolate genome, calculated as the percentage of unaligned bases. Mummer (v4.0.0b)[14] was employed to identify the mapping regions between long-read MAG/short-read MAG and its corresponding genome of isolate. We also assessed the genomic similarity of shared strains between the long-read MAGs/short-read MAGs and their corresponding isolate genomes determined by k-mer similarity as in the previous study[1], with the following formula:

$$\text{Kmer Similarity} = \frac{Number\ of\ shared\ kmer\ between\ Genome\ A\ and\ Genome\ B}{Lesser\ total\ kmer\ count\ of\ either\ Genome\ A\ and\ Genome\ B}$$

Where:

K-mer size is set to 31.

Shared k-mers are those that appear in both Genome A and B.

Total k-mer count is the count of all distinct 31-mers in a genome.

The k-mer similarity between two genomes A and B was determined by generating a hash for genomes with k-mer size 31 and quantifying the proportion of k-mers shared in both genomes A and B divided by the lesser total k-mer count of either

The proportion of reads associated with each strain/MAG

We analyzed long-read and short-read metagenomic sequencing data from D1 by mapping to a merged set of genomes, which included genomes from bacterial isolates and uncultured long-read MAGs or short-read MAGs. For the long-read data, we used 136 genomes from bacterial isolates and 105 uncultured long-read MAGs. The proportion of reads associated with each strain/MAG was determined by calculating the percentage of metagenomic reads that were successfully mapped to each genome/MAG in this merged set. Similarly, the short-read data were aligned to the same 136 genomes of bacterial isolates, supplemented with 120 uncultured short-read MAGs, to quantify the proportion of reads associated with each strain/MAG.

LongTrack strain tracking based on strain-specific k-mers selected from long-read MAGs

We built on the methodology of the previous study[1], but tailored it with important modifications to adapt it to long reads MAGs, given their lower completeness relative to genomes of bacterial isolates. This modified approach is designated as the "LongTrack" tool. In summary, we identified unique k-mers (k=31) for each strain-level long-read MAG from donor or pre-FMT recipient as the markers by eliminating common k-mers present in >100K NCBI bacterial sequenced strains (ncbi2018-11-20), 116 unrelated metagenomic samples as used in a previous study[1], and other long-read MAGs in the same FMT cases. We then applied an additional filter to only retain unique k-mers that originate from common regions shared among multiple strain-level long-read MAGs within the same species in the same FMT case. Mummer (v4.0.0b)[14] was employed to identify the common genomic regions. Subsequently, we assigned short reads from post-FMT recipient's microbiome data, which contained the unique k-mers, to corresponding long-read MAG. These assigned reads were mapped back to the corresponding long-read MAG genome using Bowtie 2.4.1[15] with parameters optimized for accuracy ('very-sensitive', 'no-mixed', and 'no-discordant'). If assigned reads can align to regions containing unique k-mers of corresponding long-read MAG, it indicates that the MAG is present in the post-FMT recipient (or can be tracked). Finally, we determined the confidence score by comparing read enrichment in five unrelated samples (negative controls) as done in the previous study[1]. The results with low confidence were discarded.

In the IBD cases, the recipients had received fecal microbiota transplantation from six donors (D4 - D9), and strain tracking was performed for long-read MAGs from two donors (D4 and D5). To avoid the influence of common k-mers from the remaining four donors, we further excluded the k-mers that were present in the short-read metagenomic seqeuncing data of the other four donors. Subsequent procedures were consistent with those applied to the cases.

Evaluation of LongTrack for FMT strain tracking

Evaluation of LongTrack for FMT strain tracking through comparison with gold standard

For a comprehensive assessment of FMT strain tracking, we utilized the shared strains between long-read MAGs/short-read MAGs and genomes of bacterial isolates in four FMT cases of rCDI. This evaluation was conducted by comparing tracking results in the post-FMT recipients at different time points based on genomes of bacterial isolates (considered the gold standard) and long-read MAGs/short-read MAGs.

The accuracy of strain tracking based on long-read MAGs /short-read MAGs was evaluated using the formula: the number of consistent tracking results between gold standard and long-read MAGs/short-read MAGs divided by the total number of tracking results, expressed as a percentage. The tracking results were presented as a heatmap using the R package "pheatmap v1.0.12".

Evaluation of FMT strain tracking between LongTrack and StrainFinder, which serves as a representative methods of SNP-based tracking

To evaluate FMT strain tracking between LongTrack and StrainFinder[16], we utilized *in silico* spike-in datasets. For donor D1, we spiked in 3% *Clostridium tyrobutyricum* and 2% *Turicibacter sanguinis* into long-read microbiome dataset. The simulated Nanopore long reads with an average length of 5 kb, were generated using nanosim (v 3.0.0)[17] for metagenomic assembly. The error profile was trained using real sequencing data from a large number of nanopore sequencing reads, and the read length distribution was modeled to match that of the actual Donor 1 sequencing data. Similarly, we produced simulated short-read data with identical abundances of these strains for donor D1 in the short-read microbiome datasets. For the post-FMT recipient R1, we spiked in short reads to the short-read metagenomic sequencing data, which was used for strain tracking. We designed two distinct spike-in datasets: A) The first dataset A (low abundance) featured a composition of 0.25% *Clostridium tyrobutyricum*, along with paired short reads, each having a length of 150 bp. Furthermore, this dataset included 0.1%

Turicibacter sanguinis, with paired short reads, also 150 bp in length. B) The second dataset B (high abundance) presented a composition of 2.5% *Clostridium tyrobutyricum*, accompanied by paired short reads, each measuring 150 bp in length. Additionally, this dataset contained 1% *Turicibacter sanguinis*, once again with paired short reads, each 150 bp in length. We simulated paired short reads with a length of 150 bp using randomreads.sh from the bbmap package (v38.86)[18].

Evaluation of LongTrack for FMT strain tracking when there are high-similarity strains present in between the donor and pre-FMT samples.

In this evaluation, we spiked-in ten pairs of strains, each with high similarity (ANI>99.5%), from ten different species, one strain to D1 and the other to R1 pre-FMT, similarly for all the ten pairs. All twenty strains were added to R1 post-FMT. To facilitate metagenomic assembly, we generated simulated long-read data/short-read data for D1 and R1 pre-FMT into their long-read/short-read microbiome datasets. Furthermore, we simulated short-read paired reads, each with a length of 150 bp, to R1 post-FMT short-read microbiome datasets. The simulated Nanopore long reads with an average length of 5 kb, were generated using the nanosim (v 3.0.0)[17]. The short reads were generated using the randomreads.sh from the bbmap package (v38.86)[18]. LongTrack was performed for strain tracking analysis.

Evaluation of LongTrack for FMT strain tracking when there are co-existing strains of the same species in one donor sample.

In this evaluation, we spiked-in three strains of *Phocaeicola massiliensis* (Strains 1-3) to D1 and a fourth strain of *P. massiliensis* (Strain 4) to R1 pre-FMT. For metagenomic assembly, long-read data/short-read data were generated for these strains in D1 and R1 pre-FMT microbiome datasets. Furthermore, we simulated short-read paired reads of 4 strains (Strains 1-4), each with a length of 150 bp, to R1 post-FMT short-read microbiome datasets. The simulated Nanopore long reads with an average length of 5 kb, were generated using the

nanosim (v 3.0.0)[17]. The short reads were generated using the randomreads.sh from the bbmap package (v38.86)[18]. LongTrack was performed for strain tracking analysis.

Reverse tracking and novel strain identification

We initially retrieved 402 uncultured long-read MAGs from post-FMT recipients (R1-R4). We then identified long-read MAGs shared between the post-FMT recipient and the respective donor or pre-FMT recipient using an ANI exceeding 99.4% in the same FMT cases. We confirmed that 39 long-read MAGs are unique to the post-FMT recipients and investigated whether they were present in the metagenomic sequencing data of the corresponding donor and pre-FMT recipient using LongTrack. If the long-read MAGs were not tracked in microbiome data of donor, pre-FMT or any earlier time points of post-FMT, we considered them to be novel strains.

Detection of structural variation in engrafted strains using long-read sequencing

We utilized NGMLR (v 0.2.7)[19] to perform the alignment of long-read microbiome data of post-FMT recipients (R1-R4) across various time points to the uncultured long-read MAGs and isolated genomes from the respective rCDI donors (D1-D3). Following the alignment process, we employed Sniffles (v1.0.12)[20] for the detection of structural variations (SVs) within the long-read microbiome data. To ensure the reliability of our SV detection, we applied several filtering criteria: 1) We excluded SVs involving multiple strains within the same species, as these are more prone to mismatches and may not accurately represent structural variations. 2) For mapping quality control, we retained aligned reads with identity greater than 98%, length more than 1000 bp, and MAPQ (Mapping Quality) score greater than 60. 3) We kept SVs that had a coverage depth of greater than 5 in long-read metagenomic sequencing data of donor and at least one post-FMT, ensuring that the detected structural variations were well-supported by sequencing data.

Methylation analysis using long read sequencing

De novo methylation motif discovery using Nanopore long-read sequencing

We utilized Nanodisco (v1.0.2)[21] to identify *de novo* methylation motifs in long-read MAGs from the D1 . First, raw fast5 reads from both native and WGA metagenomic sequencing data of D1 were aligned to the long-read MAGs with “nanodisco preprocess”. The alignments for each long-read MAG were then segregated. Only long-read MAGs exhibiting a coverage above 20 x were retained to the following analysis. We then computed current differences (PA) between the native (retaining methylation) and WGA (methylation-free) data for each genomic position using “nanodisco difference”. To conclude, “nanodisco motif and refine” was employed for *de novo* methylation motif discovery based on these current differences. For D1 5-year, R1 post-FMT 8-week, and R1 post-FMT 5-year, the raw fast5 reads from native microbiome data were aligned to the retained long-read MAGs from D1 with “nanodisco preprocess” and separated the alignments into each MAG. Subsequently, we computed the current differences by comparing the native data with the WGA data from D1 for each MAG using “nanodisco difference”. “Nanodisco refine” was employed to verify the presence of D1 detected methylation motifs in the samples of D1 5-year, R1 post-FMT 8-week, and R1 post-FMT 5-year based on the current differences.

De novo methylation motif discovery using PacBio long-read sequencing

For PacBio long-read data of D1 (generated by the Sequel II platform), the raw reads pre-processing was performed with SMRTlink (v8.0). Raw multiplexed files were first de-multiplexed with lima. We mapped the subreads to long-read MAGs using pbalign (v0.4.1) with default parameters and separated the alignments into each MAG. The IPD ratio was calculated with ipdSummary (v2.4.1) from the alignments for each MAG. The *de novo* methylation motifs were

detected by using motifMaker (in SMRTlink v8.0 package) with default parameters. For PacBio long-read data of D1 5-year, R1 8-week and R1 5-year (generated by the Revio platform), the raw reads pre-processing was performed with SMRTlink (v13.0), and methylation motif detection was performed by our in-house scripts.

Method reference

1. 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).
2. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
3. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
4. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
5. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
6. Vicedomini, R., Quince, C., Darling, A. E. & Chikhi, R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat. Commun.* **12**, 4485 (2021).
7. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
8. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
9. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

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

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

12. 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).

13. Rodriguez-R, L. M. *et al.* An ANI gap within bacterial species that advances the definitions of intra-species units. 2022.06.27.497766 Preprint at https://doi.org/10.1101/2022.06.27.497766 (2023).

14. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).

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

16. Smillie, C. S. *et al.* Strain-tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* **23**, 229-240.e5 (2018).

17. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience* **6**, gix010 (2017).

18. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).

19. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

20. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* 1–10 (2024) doi:10.1038/s41587-023-02024-y.

21. Tourancheau, A., Mead, E. A., Zhang, X.-S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods* **18**, 491–498 (2021).

Fig.1

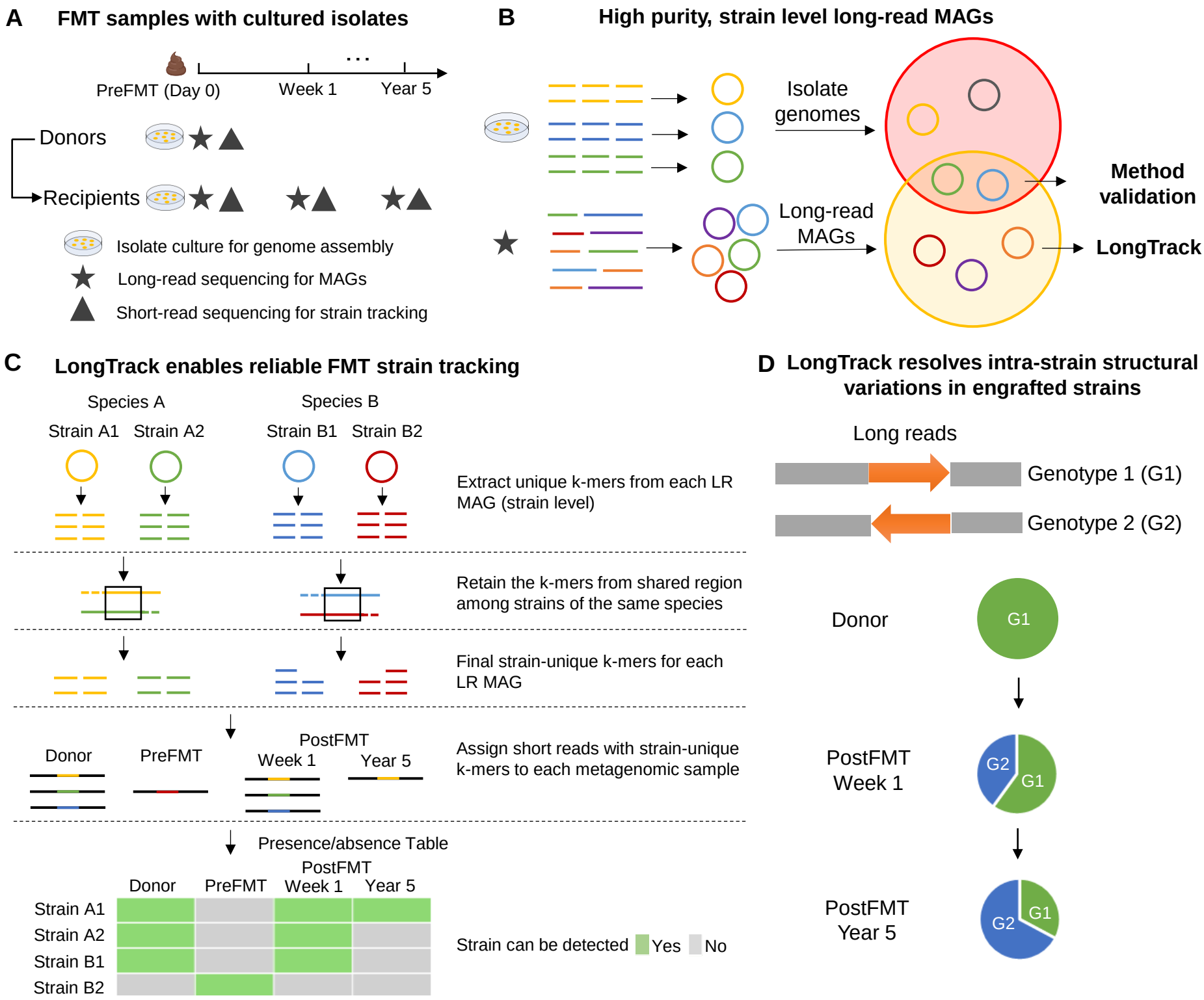
A
FMT samples with cultured isolates
PreFMT (Day 0)
Week 1
Year 5
...
Donors
Recipients
Isolate culture for genome assembly
Long-read sequencing for MAGs
Short-read sequencing for strain tracking
B
High purity, strain level long-read MAGs
Isolate genomes
Long-read MAGs
Method validation
LongTrack
C
LongTrack enables reliable FMT strain tracking
Species A
Species B
Strain A1
Strain A2
Strain B1
Strain B2
Extract unique k-mers from each LR MAG (strain level)
Retain the k-mers from shared region among strains of the same species
Final strain-unique k-mers for each LR MAG
Donor
PreFMT
PostFMT
Week 1
Year 5
Assign short reads with strain-unique k-mers to each metagenomic sample
Presence/absence Table
Strain can be detected
Yes
No
D
LongTrack resolves intra-strain structural variations in engrafted strains
Long reads
Genotype 1 (G1)
Genotype 2 (G2)
Donor
PostFMT Week 1
PostFMT Year 5
G1
G2

Fig.2

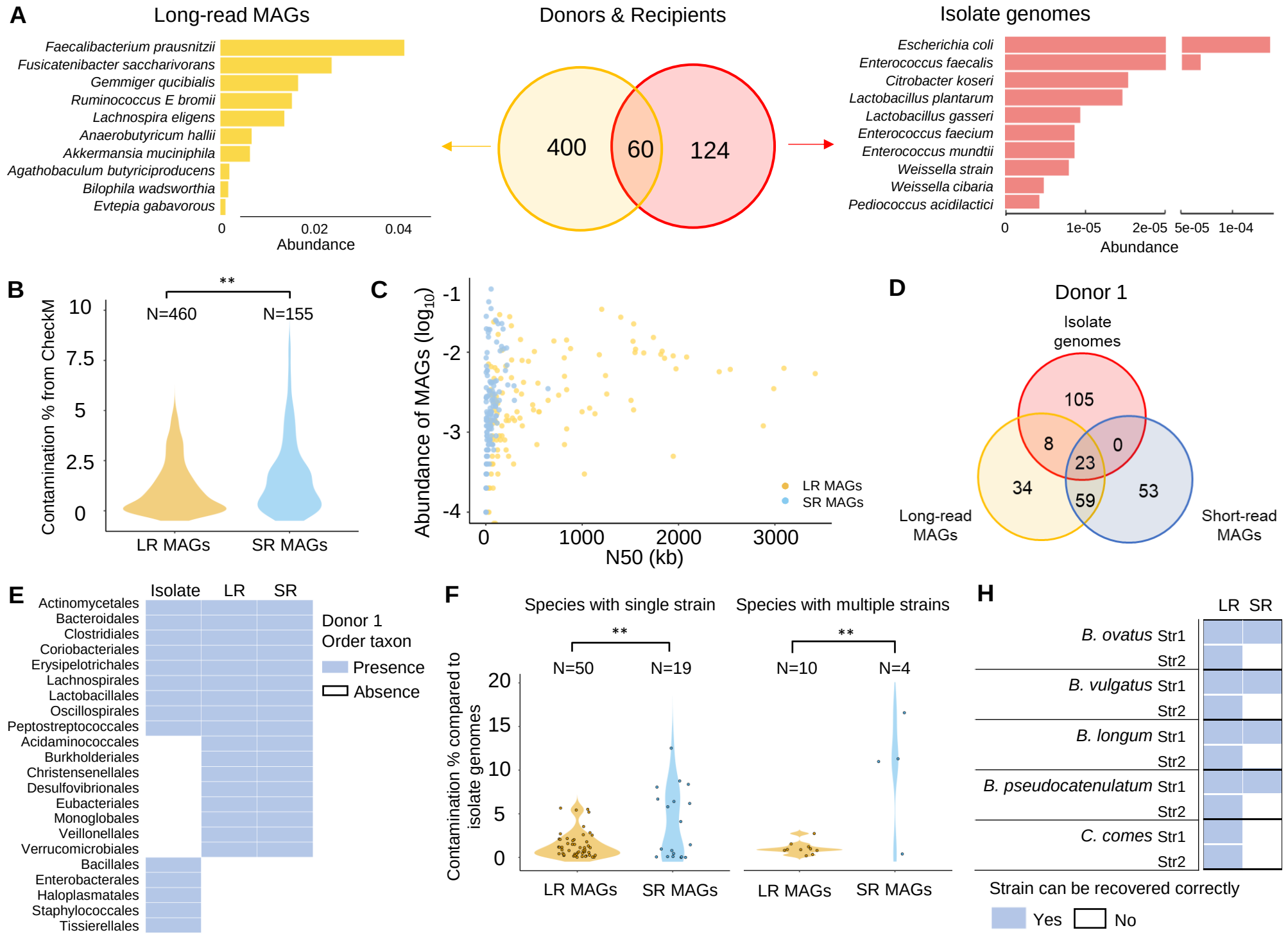
A
Long-read MAGs
Faecalibacterium prausnitzii
Fusicatenibacter saccharivorans
Gemmiger qucibialis
Ruminococcus E bromii
Lachnospira eligens
Anaerobutyricum hallii
Akkermansia muciniphila
Agathobaculum butyriciproducens
Bilophila wadsworthia
Evtepia gabavorous
0
0.02
0.04
Abundance
Donors & Recipients
400
60
124
Isolate genomes
Escherichia coli
Enterococcus faecalis
Citrobacter koseri
Lactobacillus plantarum
Lactobacillus gasseri
Enterococcus faecium
Enterococcus mundtii
Weissella strain
Weissella cibaria
Pediococcus acidilactici
0
1e-05
2e-05
5e-05
1e-04
Abundance
B
**
N=460
N=155
Contamination % from CheckM
10
7.5
5
2.5
0
LR MAGs
SR MAGs
C
Abundance of MAGs (log10)
-1
-2
-3
-4
0
1000
2000
3000
N50 (kb)
LR MAGs
SR MAGs
D
Donor 1
Isolate genomes
105
8
0
23
34
59
53
Long-read MAGs
Short-read MAGs
E
Isolate
LR
SR
Actinomycetales
Bacteroidales
Clostridiales
Coriobacteriales
Erysipelotrichales
Lachnospirales
Lactobacillales
Oscillospirales
Peptostreptococcales
Acidaminococcales
Burkholderiales
Christensenellales
Desulfovibrionales
Eubacteriales
Monoglobales
Veillonellales
Verrucomicrobiales
Bacillales
Enterobacterales
Haloplasmatales
Staphylococcales
Tissierellales
Donor 1
Order taxon
Presence
Absence
F
Species with single strain
Species with multiple strains
**
N=50
N=19
**
N=10
N=4
Contamination % compared to isolate genomes
20
15
10
5
0
LR MAGs
SR MAGs
LR MAGs
SR MAGs
H
LR
SR
B. ovatus Str1
Str2
B. vulgatus Str1
Str2
B. longum Str1
Str2
B. pseudocatenulatum Str1
Str2
C. comes Str1
Str2
Strain can be recovered correctly
Yes
No

Fig.3

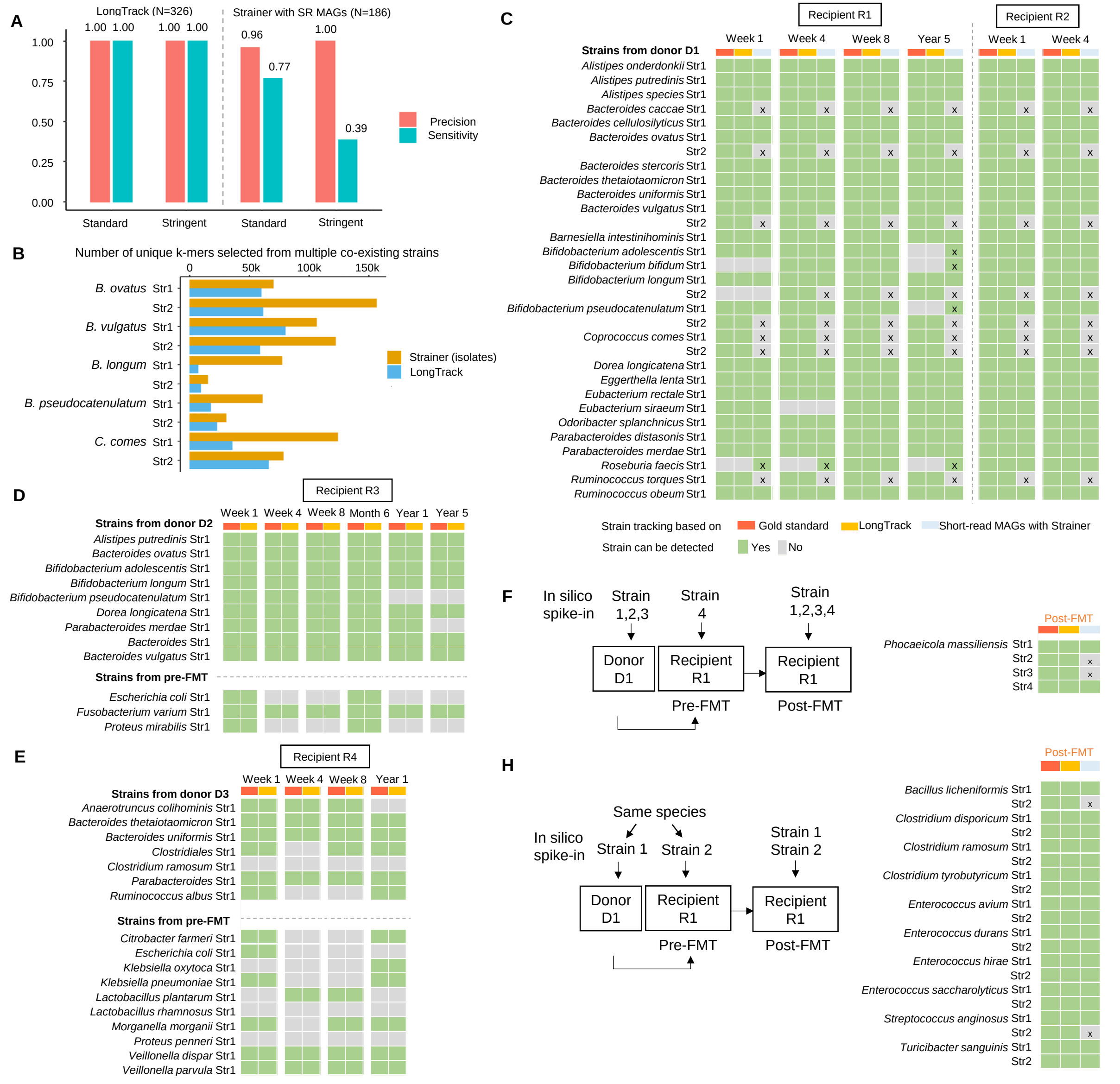
A
LongTrack (N=326)
Strainer with SR MAGs (N=186)
1.00 1.00 1.00 1.00 0.96 0.77 1.00 0.39
Precision
Sensitivity
Standard
Stringent
Standard
Stringent
B
Number of unique k-mers selected from multiple co-existing strains
Strainer (isolates)
LongTrack
C
Recipient R1
Recipient R2
Week 1
Week 4
Week 8
Year 5
Strains from donor D1
Strain tracking based on
Gold standard
LongTrack
Short-read MAGs with Strainer
Strain can be detected
Yes
No
D
Recipient R3
Week 1
Week 4
Week 8
Month 6
Year 1
Year 5
Strains from donor D2
Strains from pre-FMT
E
Recipient R4
Week 1
Week 4
Week 8
Year 1
Strains from donor D3
Strains from pre-FMT
F
In silico spike-in
Strain 1,2,3
Strain 4
Strain 1,2,3,4
Donor D1
Recipient R1
Recipient R1
Pre-FMT
Post-FMT
Post-FMT
Phocaeicola massiliensis
H
Same species
In silico spike-in
Strain 1
Strain 2
Strain 1
Strain 2
Donor D1
Recipient R1
Recipient R1
Pre-FMT
Post-FMT
Post-FMT
Bacillus licheniformis
Clostridium disporicum
Clostridium ramosum
Clostridium tyrobutyricum
Enterococcus avium
Enterococcus durans
Enterococcus hirae
Enterococcus saccharolyticus
Streptococcus anginosus
Turicibacter sanguinis

Fig.4

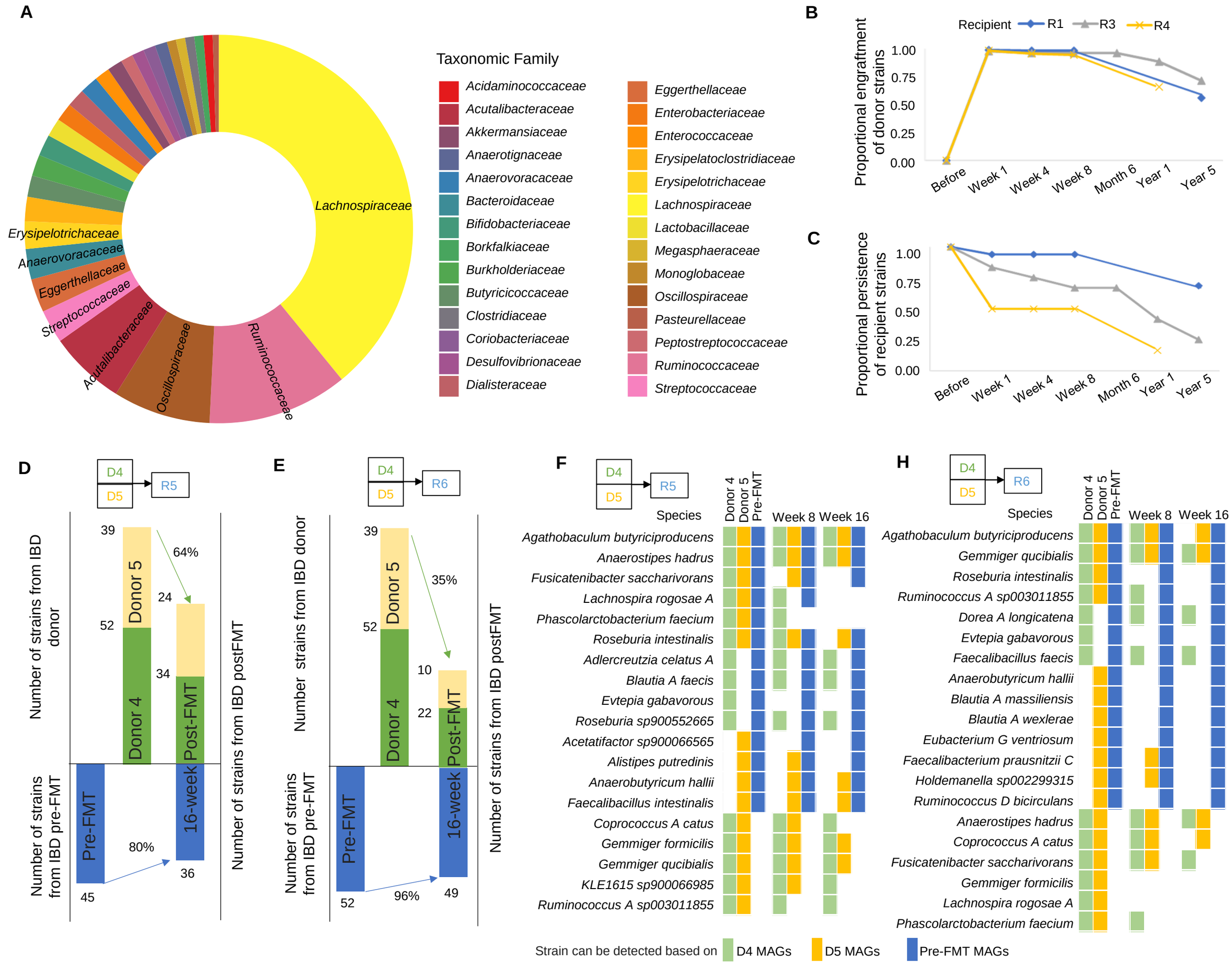
A
Lachnospiraceae
Ruminococcaceae
Oscillospiraceae
Acutalibacteraceae
Streptococcaceae
Eggerthellaceae
Anaerovoracaceae
Erysipelotrichaceae
Taxonomic Family
Acidaminococcaceae
Acutalibacteraceae
Akkermansiaceae
Anaerotignaceae
Anaerovoracaceae
Bacteroidaceae
Bifidobacteriaceae
Borkfalkiaceae
Burkholderiaceae
Butyricicoccaceae
Clostridiaceae
Coriobacteriaceae
Desulfovibrionaceae
Dialisteraceae
Eggerthellaceae
Enterobacteriaceae
Enterococcaceae
Erysipelatoclostridiaceae
Erysipelotrichaceae
Lachnospiraceae
Lactobacillaceae
Megasphaeraceae
Monoglobaceae
Oscillospiraceae
Pasteurellaceae
Peptostreptococcaceae
Ruminococcaceae
Streptococcaceae
B
Recipient R1 R3 R4
Proportional engraftment of donor strains
1.00
0.75
0.50
0.25
0.00
Before
Week 1
Week 4
Week 8
Month 6
Year 1
Year 5
C
Proportional persistence of recipient strains
1.00
0.75
0.50
0.25
0.00
Before
Week 1
Week 4
Week 8
Month 6
Year 1
Year 5
D
D4
D5
R5
Number of strains from IBD donor
39
52
Donor 5
Donor 4
64%
24
34
Post-FMT
Pre-FMT
16-week
80%
45
36
Number of strains from IBD pre-FMT
Number of strains from IBD postFMT
E
D4
D5
R6
Number strains from IBD donor
39
52
Donor 5
Donor 4
35%
10
22
Post-FMT
Pre-FMT
16-week
96%
52
49
Number of strains from IBD pre-FMT
Number of strains from IBD postFMT
F
D4
D5
R5
Species
Donor 4
Donor 5
Pre-FMT
Week 8
Week 16
Agathobaculum butyriciproducens
Anaerostipes hadrus
Fusicatenibacter saccharivorans
Lachnospira rogosae A
Phascolarctobacterium faecium
Roseburia intestinalis
Adlercreutzia celatus A
Blautia A faecis
Evtepia gabavorous
Roseburia sp900552665
Acetatifactor sp900066565
Alistipes putredinis
Anaerobutyricum hallii
Faecalibacillus intestinalis
Coprococcus A catus
Gemmiger formicilis
Gemmiger qucibialis
KLE1615 sp900066985
Ruminococcus A sp003011855
H
D4
D5
R6
Species
Donor 4
Donor 5
Pre-FMT
Week 8
Week 16
Agathobaculum butyriciproducens
Gemmiger qucibialis
Roseburia intestinalis
Ruminococcus A sp003011855
Dorea A longicatena
Evtepia gabavorous
Faecalibacillus faecis
Anaerobutyricum hallii
Blautia A massiliensis
Blautia A wexlerae
Eubacterium G ventriosum
Faecalibacterium prausnitzii C
Holdemanella sp002299315
Ruminococcus D bicirculans
Anaerostipes hadrus
Coprococcus A catus
Fusicatenibacter saccharivorans
Gemmiger formicilis
Lachnospira rogosae A
Phascolarctobacterium faecium
Strain can be detected based on
D4 MAGs
D5 MAGs
Pre-FMT MAGs

Fig.5

A

Alistipes onderdonkii

Short-read Alignment

Nanopore Long-read Alignment

PacBio Long-read Alignment

1323bp Insertion

Alistipes putredinis

Short-read Alignment

Nanopore Long-read Alignment

PacBio Long-read Alignment

1270bp Deletion

B

Strain	Type	Donor	Early	Late	Gene/downstream gene
Agathobacter rectalis D1	INV			*	IR: DNA/RNA helicase
	INS			*	IR: Type II toxin-antitoxin system prevent-host-death family
	DEL			*	IR: Nucleotidyltransferase domain-containing protein
	INS			*	IR: Transcriptional regulator
Alistipes onderdonkii D1	INS				IR: HTH-type transcriptional activator RhaS
	INV				GR: Restriction endonuclease subunit S
	INV				IR: Transcription antitermination protein RfaH
	INS				IR: Hypothetical protein
	INV				IR: Hypothetical protein
Alistipes putredinis D1	DEL				GR: IS110 family transposase ISCaa7
	INS				IR: Hypothetical protein
	INV				GR: Type-1 restriction enzyme EcoKI specificity protein
	INS				GR: OmpA family protein
	DEL				IR: DUF4836 family protein
	INS				IR: Hypothetical protein
	INV				IR: Site-specific integrase
	INS				GR: Hypothetical protein
	INS				IR: Lysine 2,3-aminomutase
	INS				IR: Peptide chain release factor RF2
Bacteroides stercoris D1	INV				GR: TonB-dependent receptor SusC
Bacteroides uniformis D1	INV				IR: UpxY family transcription antiterminator
	INS				GR: Alkaline phosphatase
Bifidobacterium adolescentis D1	INS			*	GR: H(+)/Cl(-) exchange transporter ClcA
Parabacteroides merdae D1	INV				GR: Restriction endonuclease subunit S
	INS			*	IR: TonB-dependent receptor
	INV			*	GR: Hypothetical protein
	INS			*	GR: TonB-dependent receptor SusC
	INV			*	IR: Hypothetical protein
	DEL				GR: TonB-dependent receptor SusC
Gemmiger formicilis D2	DEL				IR: Hypothetical protein
	INV				GR: Restriction endonuclease subunit S
	INS				IR: ABC transporter substrate-binding protein
	INS				IR: Type II secretion system protein
	INS		*		IR: tRNA-Ala(tgc)
	DEL		*		IR: Protein VirD4
	INS		*		IR: Hypothetical protein
	INS		*		GR: DUF87 domain-containing protein
Phocaeicola plebeius D2	INS				GR: TonB-dependent receptor P26

SV frequency: 0 – 1

* Low/No coverage; IR: Intergenic region; GR: Genic region

C

Bacteroides stercoris

SusC SusC

Genotype 1

Genotype 2

D1 R1 E R1 L

Alistipes putredinis

Soj ISCaa7

Genotype 1

Genotype 2

D1 R1 E R1 L

Alistipes onderdonkii

rhaS

Genotype 1

Genotype 2

D1 R1 E R1 L

Genotype 1

Genotype 2

E: Early Stage

L: Late Stage

D

Parabacteroides merdae Type I MTases

R M S1 S2 S3

Genotype 1

Genotype 2

Genotype 3

Genotype 4

Genotype 5

Genotype 6

D1 R1 E R1 L

Genotype 1 Genotype 3 Genotype 5 E: Early Stage

Genotype 2 Genotype 4 Genotype 6 L: Late Stage

E

Species	Motifs	Donor 1 Day 0	Donor 1 Year 5	Post-FMT R1 Week 8	Post-FMT R1 Year 5
Parabacteroides merdae	TCCGG6mA	Yes	Yes	Yes	Yes
	CA6mAGNNNNNNNNTTC	Yes	Yes	Yes	Yes
	GRNT6mAYNNNNNNNRTTA	Yes	Yes	Yes	Yes
	CA6mAYNNNNNNNRTC			Yes	Yes

Methylation detection: Yes / No

Only in Genotype 6

F

Species	Motifs	Day 0	Year 5	Week 8	Year 5
Agathobacter faecis	GGW4mCC	Yes	Yes	Yes	Yes
Akkermansia muciniphila	RA6mATTY	Yes	Yes	Yes	Yes
Anaerobutyricum hallii	GGNC5mC	Yes	Yes	Yes	Yes
Anaerostipes hadrus	CTS6mAG	Yes	Yes	Yes	Yes
Bacteroides uniformis	AC6mAGG	Yes	Yes	Yes	Yes
	AGT6mACT	Yes	Yes	Yes	Yes
Mediterraneibacter faecis	GGW4mCC	Yes	Yes	Yes	Yes
Ruminococcus D bicirculans	CTGC6mAG	Yes	Yes	Yes	Yes
Bifidobacterium bifidum	CTGC6mAG	Yes	Yes	Yes	Yes
Dorea longicatena	GC6mATC	Yes	Yes	Yes	Yes
Eubacterium R sp000434995	A4mCGGC	Yes	Yes	Yes	Yes
	GCC6mAT	Yes	Yes	Yes	Yes
	C6mATG	Yes	Yes	Yes	Yes