# Gut microbiota survey at species- and strain-level using long-amplicon sequencing, and its application to trace species engraftment

Alfonso Benítez-Páez[1,2*], Annick V. Hartstra[3], Max Nieuwdorp[3], Yolanda Sanz[1*]

1 Microbial Ecology, Nutrition & Health Research Unit. Institute of Agrochemistry and Food Technology, Spanish National Research Council (IATA-CSIC). 46980 Paterna-Valencia, Spain.

2 Host-Microbe Interactions in Metabolic Health Laboratory. Príncipe Felipe Research Center (CIPF). 46012 Valencia, Spain.

3 Department of Internal and Vascular Medicine, Amsterdam University Medical Centres. 1105 AZ Amsterdam, The Netherlands.

Running title: *rrn* operon sequencing for assessing species engraftment

* To whom any correspondence should be addressed: ABP E-mail abenitez@cipf.es, YS E-mail yolsanz@iata.csic.es.

Word count: 6473

# Abstract

The precise identification of bacterial species inhabiting the human gut is essential to define the role of certain gut microbiota in human health and disease unambiguously. At this regard, the sequencing of bacterial *rrn* regions (16S-ITS-23S) has provided reliable identification of microbiota at the species level previously. On the other hand, fecal microbiota transplantation (FMT) is a clinical approach with recognized performance for treating infection, immune and metabolic disease. However, the demonstration of its underlying species engraftment is challenging because there are no cost-effective strategies for accurately identification and tracing the transfer of donor microbiota to recipients. We used Nanopore-based device to sequence long amplicons from bacterial *rrn* operons to reveal strain-level variation in predominant members of the human gut microbiota. By assessing the donor and recipient microbiota before and after FMT, we evaluated whether this molecular approach helps to reveal strain-level genetic variation and demonstrate microbe engraftment. We reconstructed several *rrn* regions derived from predominant gut microbiota, revealing novel genetic variation. Strict control over sequencing data quality and major microbiota covariates was critical for accurately estimating the changes in gut microbial species abundance in the recipients after FMT. Despite the modest per-base accuracy of Nanopore data, we could detect strain-level variation by predicting single-nucleotide variants (SNVs) at *rrn* regions in a species-specific manner. We showed that the FMT trial explored successfully induced donor-strain (e.g., *Parabacteroides merdae)* engraftment in recipients according to the mapping and assessment of nucleotide frequencies at selected SNVs. Despite the modest per-base accuracy of Nanopore reads, the data showed sufficient quality to allow the accurate estimation of SNVs and the tracing of microbe engraftment. This methodology and data analysis represent a cost-effective approach for tracing genetic variability, which is necessary to better understand the health effects of the human microbiome under the application of microbiota replacement strategies.

***Trial registration.*** The study was prospectively registered at the Dutch Trial registry - NTR4488 (https://www.trialregister.nl/trial/4488).

**Keywords**: Nanopore sequencing, MinION, gut microbiota, species engraftment, single-nucleotide variation, species-level resolution, *rrn* operon.

## Introduction

The microbiome research advances rapidly and the study of microbial communities demands the implementation of sequencing strategies permitting the identification of microbial species and strains present in samples and allowing the potential tracing of microbes engraftment in different clinical trials. Among them, fecal microbiota transplantation (FMT) is widely recognized as an efficient therapy for *Clostridioides difficile* infection that outperforms antibiotic treatment (Cammarota et al., 2014). Despite the potential risk for the transmission of infectious diseases by FMT, this approach has shown efficacy in ameliorating inflammatory gut and metabolic disorders. The mode of FMT action is thought to depend on the engraftment of live microbes from donors to recipients, although some clinical trials leave the door open for hypotheses about alternative mechanisms (Ott et al., 2017). Consequently, the demonstration of disease amelioration linked to donor microbes thriving in recipient guts is challenging and remains rare but is necessary to ascertain the roles of particular microbes in human health. Metagenomics based on DNA shotgun sequencing is currently the only strategy that enables us to reach such a resolution level, but it is still very expensive (ranging between 450-900USD depending on the coverage), and few strain profiling results have been provided by FMT studies adopting such protocols (Aggarwala et al., 2021;Verma et al., 2021;Wilson et al., 2021). Thus, this approach is too costly to employ in most studies aiming to define microbiota features associated with health or disease based on a large number of samples to minimize the noise introduced by other covariates that contribute to the high interindividual variability of microbiota. As a consequence, most of FMT microbiome surveys have been performed through the targeted amplification of universal gene markers, mainly including a few hypervariable regions (e.g., V3 and/or V4) of the bacterial 16S rRNA gene (Yu et al., 2020;Rinott et al., 2021). This protocol shows limitations since it only allows changes affecting the global microbiome structure to be captured and does not permit precise taxonomic identifications at the species level, mostly because of the limited resolution of these gene markers in discriminating

closely related microbes. This issue is of particular relevance to accurately trace species engraftment in FMT or demonstrate microbe colonization in probiotic-based clinical trials. We previously developed a Nanopore-based amplicon sequencing method to improve the resolution of taxonomic identifications by studying the microbial genetic variability of nearly complete 16S rRNA genes (Benitez-Paez et al., 2016). This approach has been demonstrated to perform well in a wide variety of microbiota studies (Shin et al., 2016;Cusco et al., 2018;Shin et al., 2018;Sakai et al., 2019). More recently, we have also pioneered a new methodology combining the sequencing of an extremely variable multilocus region with sample multiplexing (Benitez-Paez and Sanz, 2017) to improve the species-level characterization of complex microbial communities (Kerkhof et al., 2017;Cusco et al., 2018). The implementations mentioned above of nanopore-based technology and their multiple applications to assess microbial communities allow us to gain knowledge in the configuration of different analytical pipelines improving the processing of such data as updated protocols are delivered (Ciuffreda et al., 2021;Kerkhof, 2021;Pavlovic et al., 2021).

Despite the promising performance of the *rrn* region for microbial species identification, one of the main pitfalls of this methodology is the lack of a reference database for comparison with the long reads generated from Nanopore-based devices. In this context, we have made a great effort to compile a great deal of genetic information on *rrn* sequences from thousands of bacterial species (*rrn_db*, freely available at https://github.com/alfbenpa/rrn_db) (Benitez-Paez and Sanz, 2017). Nevertheless, the limited number of properly annotated DNA sequences available in public repositories of bacteria from multiple ecosystems will restrict the utilization of this approach in less explored habitats (Cusco et al., 2018;Peker et al., 2019). Moreover, recent studies have indicated that the architecture of the *rrn* region is not equally conserved in all bacteria; thus, microbial communities enriched in the Deinococcus-Thermus, Chloroflexi, and Planctomycetes phyla might be difficult to study by using such pipelines (Brewer et al., 2020). Regardless, the utility of *rrn*-based analysis for surveying biological diversity has been explored across kingdoms, showing

promising results in the identification of metazoans by metabarcoding approaches (Krehenwinkel et al., 2019).

Despite the issues concerning unlinked 16S rRNA and 23S rRNA markers that can be addressed in certain bacterial groups and environmental samples via the study of individual regions, we think the long-read amplicon sequencing of bacterial *rrn* regions could help fill the existing gap to demonstrate the transfer and stable colonization of microbes under FMT strategies in a cost-effective manner. For this reason, we aimed to perform the amplicon sequencing of the nearly complete *rrn* regions of the human intestinal microbiota from donors and recipients of FMT by using Nanopore-based sequencing. This clinical trial aimed to alleviate recipients' metabolic syndrome via FMT from post-Roux-en-Y gastric bypass (RYGB) donors. We obtained *de novo* information on the *rrn* region from predominant bacterial species inhabiting the human gut, one of the most extensively investigated environments, enriching the *rrn* existing database with the annotations of dozens of microbial species and strains. We estimated the changes in the human gut microbiota at the species level resulting from an FMT intervention. More importantly, we explored the potential of this approach to reveal strain-level variation to track microbiota engraftment, and the results demonstrated that our FMT trial achieved microbe transfer between donors and recipients, which likely mediated health improvements.

## Materials and Methods

*Subjects, samples, and clinical data*

Samples were obtained upon receiving informed consent from participants in a previous FMT clinical trial carried out under the framework of the MyNewGut project (http://www.mynewgut.eu/) between September 2014 and March 2017. The details of the study design are publicly available elsewhere (Hartstra et al., 2020). Briefly, a total of twenty-four fecal samples were analyzed in the present study. Twenty samples were obtained from 10 recipients undergoing allogenic FMT, who

provided one sample before (PRE-FMT fecal samples) and one sample 4 weeks after (POST-FMT fecal samples) the intervention. The samples of 4 donors were also analyzed, and the effect of this variable on the recipients was considered in the data analysis. This subset of samples was mainly assessed to explore differential species engraftment in multiple FMT recipients with common donors.

*DNA extraction, multilocus amplification, and sequencing*

Microbial DNA was recovered from 100 mg feces samples by using the QIAamp® Fast DNA Stol Mini kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions and omitting cell disruption by mechanical methods (bead-beating) to preserve DNA with the high molecular weight needed to amplify amplicons of 5 kbp in length. The *rrn* region comprising the nearly full-length bacterial RNA ribosomal operon (16S, ITS, and 23S) was amplified as previously reported (Benitez-Paez and Sanz, 2017). Dual-barcoded purified PCR products were mixed in equimolar proportions before sequencing library preparation. In total, three different libraries were prepared from ~1 µg of mixed amplicon DNA (containing 7, 8, or 10 barcoded samples) using the SQK-LSK108 sequencing kit (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's instructions to produce 1D reads. Each library was individually loaded into FLO-MIN107 (R9.5) flow cells (Oxford Nanopore Technologies, Oxford, UK), and sequencing was carried out in a portable MinION™ MkIb sequencer (Oxford Nanopore Technologies, Oxford, UK), operated with *MINKNOW* v1.10.23 software (Oxford Nanopore Technologies, Oxford, UK). Flow cells were primed according to the manufacturer’s instructions, and an ~18 h run of 1D sequencing (outperforming the 2D approach for this multiplex and amplicon setup, and improving de-multiplexing by using the barcode+primer tags) was then executed for each library and flow cell.

*Data preprocessing*

Fast5 files were processed with the *Albacore* v2.1.3 basecaller (minimum q-score filter = 7), and *fasta* files were retrieved for downstream analyses. The barcode and primer (forward or reverse)

sequence information was used for demultiplexing into the DNA reads generated from forward and reverse primers according to previous procedures (Benitez-Paez and Sanz, 2017). A size filtering step was configured to retain those reads of least 1,500 nt in length. Barcode and primer sequences were then removed by trimming 50 nucleotides at the 5' end of forward and reverse reads. The sequencing chemistry and Nanopore data-processing algorithms were the latest available by January 2018, when data was generated.

*Assembly of rrn regions*

The study design and the assembly of *rrn* sequences are graphically explained in Figure 1. This pipeline carries out what we refer to as the strand-wise and reference-assisted method for *rrn* region reconstruction via the following steps:

- Demultiplexing of Nanopore data using barcode and primer sequence information generated two datasets per sample comprising DNA reads led from the 5´ end of *rrn* amplicons (initiated from forward primer mapping at the 16S rRNA gene start site = forward reads), and DNA reads beginning at the 3' end of *rrn* amplicons (initiated from reverse primer mapping at the 23S rRNA gene end site = reverse reads).
- Reads from all 24 samples were merged into forward and reverse read subsets.
- Forward read binning into precise microbial species via competitive alignment against the nonredundant 16S NCBI database (release January 2018). The *LAST* aligner (Kielbasa et al., 2011) with a -s 2 -q 1 -b 1 -Q 0 -a 1 -r 1 configuration was used for this purpose. The alignment score was the main criterion for selecting top hits. In the case of multiple hits with the same top score, the alignment was discarded for downstream processing. High-quality alignments were retained according to sequence identity (sequence identity above the 33rd percentile ≥ 85%) and length information (alignments ≥ 1500 nt).
- A maximum of 500 forward reads per species (randomly shuffled), producing high-quality alignments, were selected, and aligned through iterative refinement methods implemented

in *MAFFT* v7.310 with the default parameters (Katoh and Standley, 2013), and the consensus sequence was then obtained by using the *hmmbuild* and *hmmemit* algorithms implemented in *HMMER3* (Eddy, 2011). No significant improvements in consensus *rrn* sequences were obtained using more than 500 sequences per species.

- Consensus sequences were annotated according to the NCBI taxonomy and used as a reference to bin reverse reads in a similar manner as that used for forward reads. The thresholds for the selection of high-quality alignments based on reverse reads and the preliminary assembly of *rrn* regions were 85% sequence identity (upper 45th percentile) and $\geq$ 3500 nt in length.
- Reverse reads were compared against the consensus sequences obtained from the forward read assemblies and binned into the respective species according to NCBI taxonomic annotation. A maximum of 500 reverse reads per species (randomly shuffled), producing high-quality alignments, were selected, merged with a maximum of 500 forward reads assigned reciprocally to the same species (randomly shuffled), and aligned through iterative refinement methods implemented in *MAFFT*, after which a new consensus sequence was obtained by using the *hmmbuild* and *hmmemit* algorithms implemented in *HMMER3*.
- Final consensus sequences were annotated according to the concordant NCBI taxonomy, obtained from forward and reverse reads, and used as a reference to study the abundance and prevalence of gut microbiota members at the species level.
- A BLAST-based search of the *rrn* assemblies against the nonredundant nucleotide and reference 16S NCBI databases was performed to evaluate the representation of the reconstructed and identified operons in public databases. The identification of *rrn* assemblies was compared against the SILVA database (Quast et al., 2013) with SINA aligner (Pruesse et al., 2012). Top hit selection was based on the taxonomy score, TS = $Log_{10}$[alignment score*sequence identity*alignment length].

*Long-read mapping and variant calling*

The *rrn* database (Benitez-Paez and Sanz, 2017) was updated with more than two hundred new *rrn* operon sequences assembled in this study (labeled operONT) and re-annotated to different taxonomic levels according to the NCBI taxonomy database. The mapping of sample reads was performed against *rrn_DBv2* (https://github.com/alfbenpa/rrn_DBv2) by competitive alignment using the *LAST* aligner, as stated above. Taxonomic assignment was based on the best hit retrieved by the calculation of TS (see above), and reliable taxonomic assignments were filtered to those based on alignments with at least 85% sequence identity that were longer than 1500 nt. Species read counts were normalized using the *compositions::clr* R v3.6 function, and a covariate-controlling linear mixed model-based method was used to explore the differential abundance of taxonomic features between conditions as stated in the paragraphs below. Species engraftment was evaluated by the selection of all sample reads mapped to the *Parabacteroides merdae* and *Faecalibacterium prausnitzii* species. To detect informative sites, based on single-nucleotide variants that were probably linked to strain variation, we first proceeded to map all sample reads matching these species to the respective reference *rrn* sequences using *LAST* aligner and the parameters set across this study (see *rrn* assembly). The *maf-convert* algorithm (*LAST* aligner toolkit) was used to generate the respective *SAM* files. The algorithms compiled within *SAMtools* v1.3.1 were used to index, order, and pileup reads as well as to retrieve information regarding the nucleotide frequency per site and coverage (*vcf* files). Site selection was based on *rrn* positions showing no nucleotide dominance (<70% frequency for a single nucleotide), without reductions in the coverage (>75% of the relative coverage), thus obtaining sites with a balanced representation of at least the two predominant alleles likely representing strain variation across all samples. Second, the sample reads that were mapped to the reference species were individually assessed to detect meaningful changes in intestinal bacterial genotypes between the recipients PRE-FMT and POST-FMT samples. The changes in nucleotide frequencies at all selected SNV sites (three for *P. merdae* and six for *F. prausnitzii*) were assessed by calculating the Bray-Curtis dissimilarity index between pairs of

POST-FMT and PRE-FMT samples or POST-FMT and corresponding donor samples following statistical evaluation by using the Wilcoxon signed rank test for paired samples.

*Sanger sequencing and validation of Parabacteroides merdae SNVs*

The reference *rrn* sequence from *P. merdae* was submitted to the Primer-BLAST web server (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) to retrieve species-specific primer pairs to selectively amplify this *rrn* sequence and primers flanking SNV-1254, SNV-2293, and SNV-3527 for Sanger sequencing. Comparisons against the nonredundant NCBI database and *Parabacteroides merdae* [taxid:46503] as a reference organism were fixed as checking parameters for primer prediction. The *rrn* region from *P. merdae* was amplified via 28 PCR cycles including the following stages: 95°C for 20 s, 61°C for 30 s, and 72°C for 150 s. Phusion High-Fidelity Taq Polymerase (Thermo Scientific) and the pm623 (TGCCGTTGAAACTGGGTTACTTGA) and pm3785 (TTTCGCACAGCCATGTGTTTTGTT) primer pairs were used in the amplification reaction. The PCR products were cleaned with an Illustra GFX PCR DNA and gel band purification kit (GE Healthcare, Chicago, IL, USA) and sequenced by using Sanger technology in an ABI 3730XL sequencer (STAB-VIDA. Caparica, Portugal) using the primers described above (pm623 and pm3785) in addition to the pm1855 (ACCCCTTACGGAGTTTATCGTGGA) and pm2686 (TTCGCGTCTACTCACTCCGACTAT) primers. The sequencing electropherograms from the *ab1* files were visualized with FinchTV v1.4.0 (Geospiza Inc.).

*Diversity and taxonomic analyses at the species level*

The prevalence and abundance of a total of 2,519 species contained in the database were evaluated, and diversity analyses were completed taking into account the species with a > 0.01% relative abundance on average (~250 species in total). Alpha diversity descriptors such as Chao's index, Shannon's entropy, Simpson's reciprocal index, and dominance were obtained by using *QIIME* v1.9.1 (Caporaso et al., 2010). Similarly, *QIIME* was used to calculate the Bray-Curtis dissimilarity indexes among samples and to perform multivariate exploratory (principal coordinate analysis -

PCoA) and statistical (PERMANOVA) analyses. A linear mixed model (LMM – *nlme::lme* R v3.6 function) analysis was also conducted on clr-transformed data to detect differential features in the microbiota before and after the intervention. The inherent variation due to individual features was set as a random effect for each variable analyzed (fixed effect). The possible covariates of the clinical and fecal microbiota that showed differences between the study groups (p-value $\leq$ 0.05) were selected. Recognized covariates of microbiota, such as age, sex, baseline BMI, and sequencing batch, were identified as significant in this study and included as random effects in the LMM. To identify microbial species potentially linked to clinical variables that were altered as a consequence of FMT, Kendall's $\tau$ (tau) parameter was estimated between variable pairs and corrected for multiple testing using the false discovery rate (FDR) approach. Associations were selected according to an FDR p-value $\leq$ 0.1. Graphics were generated in R v3.6 using the *ggplot2* and *ggridges* packages.

# Results

*Database enrichment by assembly and taxonomic identification of human gut microbiota rrn sequences*

The evaluation of the gut microbiota in this FMT clinical trial including subjects with metabolic syndrome was aimed at revealing microbiota engraftment; it required the comprehensive analysis of Nanopore sequencing data to generate reliable genetic information supporting species- and strain-level variation additionally to that compiled in our previous *rrn* repository (Benitez-Paez and Sanz, 2017). We retrieved a total of ~516 k reads after base calling from three MinION™/R9.5 sequencing runs, ~430 k reads after size trimming (83.3% retained), and ~427 k reads that were successfully demultiplexed (82.7%). The numbers of forward (from 5´ *rrn* amplicons) and reverse (from 3´ *rrn* amplicons) reads obtained after demultiplexing were 210,736 and 216,617 (0.49 and 0.51 proportions), respectively, showing no bias during the sequencing mode (1D library prep). During the first step of *rrn* assembly (Figure 1), based on mapping forward reads against the nonredundant NCBI 16S database and the selection of high-quality alignments, we detected the

presence of 381 different species. However, preliminary *rrn* assembly was initiated only for those species with at least 10X coverage (250 in total), supporting high-quality consensus *rrn* sequences in downstream analyses. The mapping of reverse reads against the preliminary *rrn* assemblies permitted us to confirm the detection of 229 microbial species.

Those 229 *rrn* sequences were then subjected to cross-identification to evaluate their representation in major nucleotide repositories using conventional methods and databases. We encountered limitations in this analysis given the scarce taxonomic information for this multilocus region, despite the fact it contains the 16S rRNA gene, the classical marker for species bacterial identification. When we explored the SILVA "ssu" and "lsu" databases (for the analysis of 16S and 23S rRNA genes, respectively) through the SINA aligner, we obtained classifications mainly at the genus level given that this is the generally deepest taxonomic level found in this database. Accordingly, we retrieved genus matches for 155 (68%) *rrn* sequences using the 16S marker and 133 (58%) genus matches using the 23S marker. Most of the remaining assemblies were correctly identified at the family and order levels. Similar performances were observed when using the 16S taxonomic classifications of the RDP and Greengenes databases. Additionally, we submitted the set of 229 *rrn* sequences to the BLAST server of NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) for comparison with the nonredundant nucleotide collection and the reference 16S database. Top hits from the nonredundant nucleotide database (based on the TS score - see methods) indicated that only 124 *rrn* sequences (54%) showed alignments covering ≥ 95% of the query length. The global assessment of *rrn* sequences against this database also produced hits predominantly annotated as "uncultured bacteria" (54%). Among the subset of alignments covering ≥ 95% of the query length, we observed that those supporting a species match had an average sequence identity of 98.61 ± 1.69 (mean ± sd), whereas those supporting genus and "uncultured bacteria" matches had average sequence identities of 94.69 ± 3.43 and 95.46 ± 3.67, respectively. Similar distributions were observed for alignments covering less than 95% of the query sequences, among which the 23S

region was preferentially explored (alignment length ~2,630 nt on average). Examples of full *rrn* sequences for which matched annotations were retrieved, including those with relatively low, intermediate, and high coverage, are disclosed in Figure S1.

In summary, we obtained the best results when comparing the *rrn* sequences against the reference 16S NCBI database (release May 2019). This enabled us to identify 183 of the sequences (80%) at the species level and 23 (10%) at the genus level and match them with the original annotation obtained during our reference-based *rrn* assembly. The remaining 23 *rrn* sequences (10%) matched species belonging to different genera, but in all cases, they were related to species of the Enterobacteriaceae family (e.g., *Shigella*, *Escherichia*, *Citrobacter*, and *Salmonella*) that were difficult to distinguish by inspecting only the 16S rRNA gene sequence (Naum et al., 2008). The global assessment of our assembled set of *rrn* sequences against the 16S NCBI database is shown in Figure 2A, where five levels of information are compiled, including sequence identity, alignment length, coverage, coverage ratio (forward vs reverse reads employed in the assembly), and match level. We observed species matches even when coverage was very low (e.g., assemblies based on ten reads) and species matches with modest sequence identity (~89%), possibly indicating the accumulation of high genetic variability at the strain level in certain species. The comparison of indels and mismatch proportions retrieved from the alignments supporting the best hits against the NCBI databases of the 16S and nonredundant nucleotide collections (the latter of which discriminates 23S alignments from those of whole *rrn* sequences) showed an expected progressive increase in either mismatch or indel proportions from queries identified at the species, genus, and other taxonomic levels (unclassified bacteria, with no further identification beyond kingdom level) (Figure 2B). However, this was only observed for 16S and 23S alignments separately given that for alignments covering the entire *rrn* sequence, the highest mismatch and indel rates were detected for identifications at the genus level (Figure 2B).

*Species abundance shifts associated with FMT*

Sample-derived forward and reverse reads were mapped to assess shifts in diversity and taxonomic features. Among all reads, only 43% supported alignments with the highest quality (according to identity and length) and were used in downstream analyses. After taxonomic assignment, we found no large changes in any of the alpha diversity descriptors analyzed. Nevertheless, we found that FMT increased the richness (Chao's index) of the microbiota in six out of the ten recipients (Figure S2), which showed a gain of 31 species on average. In the remaining four recipients, there was a reduction in richness resulting in an average loss of 16 species as a result of FMT. The beta diversity evaluation based on the Bray-Curtis dissimilarity index indicated no major shifts in the microbial structure of recipients as a consequence of FMT (PERMANOVA = 1.02, $p$ = 0.403). Nevertheless, the gut microbiota composition of the recipients was strongly influenced from greater to lesser extents by the sequencing run (PERMANOVA = 3.69, $p$ = 0.001), sex (PERMANOVA = 1.70, $p$ = 0.032), and donor (PERMANOVA = 1.69, $p$ = 0.004). The results of this multivariate analysis are shown in Figure 3A. Globally, we observed that POST-FMT samples tended to be mapped closer to those from donors. To further assess this hypothesis, we compared the distances (Bray-Curtis weighted metrics) between the respective donors and the PRE-FMT microbiota or the POST-FMT microbiota. As a result, we noted that donor microbiota-PRE-FMT pairs had a trend to be more dissimilar than the donor microbiota-POST-FMT pairs, as indicated by a decrease in the Bray-Curtis distance (PRE-FMT vs Donor median = 0.79, POST-FMT vs Donor median = 0.71, average delta = 0.09, $p$ = 0.075) (Figure 3B).

We performed an LMM analysis, which provided similar results to the beta diversity evaluation. We found that sequencing batch, sex, and donor were the main covariates influencing the microbiota data. Furthermore, we found that age and baseline BMI also explained the gut microbiota variation among the subjects of this study to some extent. After including the variables mentioned above as random effects in the model, a list of microbial species showing alterations as a

consequence of FMT was retrieved (Table 1), and some of them exhibiting a positive abundance changes as a consequence of the FMT are plotted in Figure 3C. Ten different microbial species were found to be differentially abundant when paired samples obtained before and after FMT were compared, and only three of them (e.g., *Bifidobacterium adolescentis*) seemed to show a decline because of transplantation (Table 1).

*Transferred and engrafted gut microbial species*

To assess species engraftment, we next designed a proof of concept analysis to discriminate strain-level variation and thus trace microbe engraftment more accurately. For this purpose, we selected *Parabacteroides merdae*, a predominant species in the samples assessed that showed the greatest increase as a result of FMT (Figure 4A), and *Faecalibacterium prausnitzii*, a highly abundant species in the analyzed samples for which evident transfer between the donor-recipient pairs was not observed (Figure 4B), to detect single-nucleotide variations (SNVs) associated with the strains. After a massive analysis of the total dataset merging the species-specific *rrn* reads derived from all subjects (see methods), we identified three potential informative SNV sites in the *P. merdae rrn*, whereas six were found in the *F. prausnitzii* counterpart (Figure 4C-D). In an attempt to explore the natural existence of such genotypic variation inferred from our analysis in a predictive manner, we looked for the presence of similar variants in the NCBI non-redundant nucleotide database using the 100 nt surrounding region as a query for all *P. merdae* SNVs. The inferred genetic variability was replicated at least for SNV-3527, where multiple matches presented A and G nucleotides at this position (Figure S3). When we studied the nucleotide frequencies of these particular SNVs across samples, a clear transfer pattern of the *P. merdae* genotype from donor to recipient was detected, as indicated by the decreased genetic distance (Bray-Curtis index for assessing the abundance of the four nucleotides in all SNVs at once, one-sided Wilcoxon Signed-Rank test = 11, $p$ = 0.051) between POST-FMT and donor samples relative to that of POST-FMT and PRE-FMT pairs (Figure 4E). By contrast, the genetic distances retrieved after the comparison of *F. prausnitzii* genotypes

between POST-FMT or PRE-FMT samples and their donors did not indicate the transfer of strains belonging to this species from donors to recipients (Figure 4F). Comparative delta Bray-Curtis distance (POST vs DONOR minus PRE vs DONOR samples) between both species also shed light on the likely transference of *P. merdae* only, thus obtaining a distinctive tendency towards decreasing the distance between POST-FMT and donor samples (Figure 4G, $p = 0.056$). The direct Sanger sequencing of informative *P. merdae* SNV-1254, SNV-2293, and SNV-3527 in two recipients and their common donor supported the results of our long-read-based assessment of FMT (Figure 4H). Globally, we observed the presence of a mixture of strains in some of the samples analyzed, given the base calling profile visualized in the electropherograms (e.g., SNV-3527 of Donor1 and 06-Pre samples). This pattern of strain coexistence was more evident in the POST-FMT samples of both analyzed recipients (see SNV-2293 and SNV-3527 in Figure 4H). The predominant haplotype observed in Donor1 (A1254-C2293-A3527) was transferred to the two recipients explored and became dominant after FMT. Similar patterns of transfer were observed in other donor-recipient pairs. Globally, our results demonstrated that the increased abundance of *P. merdae* in POST-FMT samples was a direct consequence of donor strain transmission. This is likely the case for other bacterial species that increased as a result of FMT.

*Correlation between clinical variables and species abundance after FMT*

Among the multiple clinical variables evaluated in these subjects with metabolic syndrome, FMT induced a trend of positive changes in the markers of glucose metabolism and blood pressure in the recipients (Table 2). Fasting insulin ($p = 0.030$), fasting glucose ($p = 0.074$), and consequently HOMA-IR ($p = 0.005$) tended to improved in the recipients after FMT. In line with these findings, there was an apparent decrease in the glycosylated hemoglobin concentration after FMT ($p = 0.060$). On the other hand, systolic and diastolic blood pressures were lowered as a consequence of donor FMT treatment, being reduced by 10% ($p = 0.028$) and 19% ($p = 0.0008$), respectively (Table 2). The concentrations of fecal SCFAs, such as fecal butyrate ($p = 0.018$) and acetate ($p = 0.033$),

were decreased after FMT, whereas that of propionate was not ($p$ = 0.280). Regarding markers of lipid metabolism, HDL cholesterol levels in plasma were reduced ($p$ = 0.032). By computing rank-based Kendall's $\tau$ (tau) parameter, we established associations between these clinical variables and the alterations in the abundance of microbial species as a consequence of FMT (see Table 1). Species such as *P. merdae*, which was increased after FMT, were negatively correlated with diastolic blood pressure ($\tau$ = -0.44 and FDR = 0.062), whereas *A. shahii* was negatively correlated with the abundance of fecal acetate and butyrate ($\tau$ = -0.50, -0.42 and FDR = 0.017, 0.042, respectively). The above correlations were not detected for other closely related and abundant species (e.g., *Parabacteroides johnsonii*, *Parabacteroides distasonis*, *Parabacteroides goldsteinii*, *Alistipes finegoldii*, *Alistipes onderdonkii*, *Alistipes putredinis* or *Alisti*pes *indistinctus*).

## Discussion

The emergence of single-molecule and synthesis-free sequencing methods and portable devices for conducting such analyses has democratized genomics, making it a disruptive technology with applications across multiple life science and clinical disciplines. In general, the central advantage of third-generation sequencing platforms, despite their higher error rates, is considered to be the ability to produce very long DNA reads that can easily be applied to resolve eukaryotic genomes and their repetitive structures (Belser et al., 2018;Deschamps et al., 2018;Jain et al., 2018;Tyson et al., 2018;Li et al., 2019;Masonbrink et al., 2019). This strength has also been advantageous for better assessing the composition of complex microbial communities, making it possible to expand the genetic information classically used in microbiota surveys and retrieve reliable taxonomic identifications at the species level (Benitez-Paez et al., 2016;Schloss et al., 2016).

Here, we explored the *rrn* amplicon sequencing of the human gut microbiota through Nanopore sequencing and the *de novo* assembly of sequences of this multilocus hypervariable region to gain insights into the corresponding genetic markers and to demonstrate their potential use for deeply

characterizing complex communities at the species and strain levels, permitting the assessment of microbe engraftment between donors and recipients in an FMT clinical trial. The utility of *rrn* operon sequencing via the MinION™ platform has already demonstrated a notable performance in resolving microbial communities at the species level, with the potential to define strain variability for particular species (Dowden et al., 2020;Ibironke et al., 2020).

In our study, the identification of the *rrn* assemblies was based mostly on sequence identities higher than 99%, achieving a similar accuracy as in previous studies based on *de novo* genome assembly (Deschamps et al., 2018;Tyson et al., 2018). Furthermore, we were presumably able to detect variability at the strain level because the retrieved *rrn* sequences allowed the correct identification of bacterial species showing 89% and 91% sequence identity against references when using the 16S rRNA gene and the entire *rrn* region as queries, respectively. Nevertheless, the possibility that some of the identified species could be novel species cannot be disregarded despite the great amount of genetic and taxonomic information on microbes inhabiting the gut environment compiled during recent years (Li et al., 2014;Pasolli et al., 2019;Almeida et al., 2020). Interestingly, we also achieved correct species identification through *rrn* sequencing, supported by very-low-coverage assemblies (at least ten reads). Since R7.3 was initially released by 2015, the constantly improving chemistry of Nanopore devices has enabled better assemblies to be obtained with lower coverage (Jain et al., 2018;Solares et al., 2018). Therefore, releases with improved chemistry (e.g., R10 and later) are also expected to influence the quality of the obtained assemblies and to drastically increase the sensitivity of this approach to reliably and rapidly detect and measure the presence of more microbial species/strains in the gut microbiota.

We provided *de novo* reliable assemblies for more than two hundred *rrn* region sequences of human gut microbes. The taxonomic identification of such assemblies indicated that they were better represented in the 16S rRNA NCBI database (different releases used during assembly and annotation), probably because this is the most studied and most commonly used bacterial taxonomic

marker. Furthermore, the large number of high sequence identity matches (>95%) retrieved for "unculturable bacteria" when the *rrn* assemblies were compared against the NCBI non-redundant nucleotide collection highlights the high level of unannotated DNA sequences (metagenome-assembled genomes, MAGs) in major public repositories. The assembled *rrn* regions could represent the basis for the taxonomic identification of these unclassified entries given that the distribution of mismatches was close to that observed for positively identified *rrn* assemblies. Additionally, future comparative assessments of *rrn* sequences in a large collection of MAGs (Almeida et al., 2020) could help to solve taxonomic issues in gene catalogs of the human microbiome. The annotation of the final *rrn* assemblies obtained against different databases and algorithms indicated that we were able to reconstruct a large proportion of *rrn* regions from approximately 40 new microbial species, and more than 150 new strains were absent in the first database release (Benitez-Paez and Sanz, 2017). These numbers, limited mainly by sequencing coverage, could also be reduced by the DNA extraction method used, in which a physical disruption step was omitted to maintain DNA integrity and ensure the successful amplification of the 5 kbp amplicons. However, the homogeneous processing of all samples resulted in comparable microbial profiles before and after FMT and did not affect the species-specific evaluation conducted in this study.

By using a dataset consisting of approximately 400 K Nanopore reads derived from 24 samples, employing the updated *rrn* database, and controlling the covariates influencing the microbiota, we revealed changes in the human gut microbiota at the species and strain levels as a consequence of an FMT intervention. This long amplicon-based approach enabled us to detect increases in species such as *Parabacteroides merdae*, *Butyricimonas virosa*, *Prevotella copri*, and *Sutterella massiliensis* as a result of FMT. In addition to age, sex, sequencing runs (recognized stochastic run-to-run variation confounding the interpretation of microbiota data (Schloss et al., 2011;Gibbons et al., 2018)), and baseline BMI, we found that the donor was a critical covariate contributing to the

impact of FMT on the recipient microbiota when using a unique donor for multiple recipients. These results are in agreement with those previously published on the basis of short reads from the V4 hypervariable region of the 16S rRNA bacterial gene explored in a larger cohort of subjects (Hartstra et al., 2020), for instance, the significant increase of *P. copri* after FMT; however, our results outperform the taxonomic resolution achieved previously, thus providing a more accurate gut microbiota survey. Interestingly, a recent FMT study on the treatment of ulcerative colitis (UC) in which the microbiota was analyzed by sequencing short reads also showed increases in *Sutterella* species as a consequence of the intervention (Paramsothy et al., 2019). Similarly, the abundance of *Butyricimonas* species was shown to be increased as a consequence of an FMT intervention aimed at eradicating antibiotic-resistant bacteria (Bilinski et al., 2017). These findings suggest that some species may frequently show changes as a result of FMT in humans, regardless of the condition of the recipient. Nevertheless, the presumable transfer of species between donors and recipients and their engraftment could not be confirmed in those studies based on short-read amplicon technology due to the lack of sufficient resolution. Additionally, potential replacement with closely related species in the recipient could be impossibly monitored in conventional microbiota surveys based on short reads. To shed light on the ability of the *rrn* sequencing approach to assess SNVs that are likely associated with strain diversity, we designed a proof of concept analysis selecting *P. merdae*, a species that exhibited a remarkable increase in the recipients' gut after FMT, and *F. prausnitzii*, which seemed to not be affected by FMT intervention. The combined information on different informative SNVs from the *P. merdae rrn* sequence clearly suggests that the increase in this species observed in recipients after undergoing FMT was likely because of strain transfer from the donors. This inference is further supported by the existence of nucleotide variants inferred from Nanopore data processing in either the samples evaluated in the present study or the genomic data present in public repositories and derived from different sequencing methodologies.

The taxonomic resolution achieved with our sequencing approach will help to more firmly support a causal relationship between the changes in the gut microbiota and the improvements in a recipient's metabolic markers and blood pressure, since robust evidence of a correlation of the transfer of bacterial strains from donor to recipient with improved clinical variables can be provided. Nevertheless, the direct roles of particular species, such as *P. merdae*, and strains thereof in the improvement of cardiometabolic health markers need to be further explored in future (single-strain) intervention studies. The performance of Nanopore-generated data in the identification of single-nucleotide polymorphisms (SNPs) in eukaryotic and prokaryotic organisms has been previously reported (Bainomugisa et al., 2018;Malmberg et al., 2019). Altogether, these results suggest that despite the error-prone nature of these data, well-processed Nanopore reads show sufficient quality to estimate SNVs, in addition to the recognized utility of these reads for chromosome assembly.

The protocols described here for identifying SNVs will be pivotal in the near future for performing more accurate microbiota assessments in FMT clinical trials with the aim of demonstrating microbe engraftment and establishing reliable genotype-to-phenotype associations between human diseases and the microbiome at the strain level. In addition to its cost effectiveness, the data derived from our method could be analyzed in a reference-based (read mapping against the *rrn* database) or reference-independent manner (read assembly into discrete similarity clusters), making this approach versatile for strain surveillance and discovery. Overall, the features mentioned above should be highly useful for defining subtle genetic variation in the human microbiome and profiling the identified variants as harmful or beneficial to human health, which is part of an envisioned field of research in the framework of the epidemiology of microbial communities and the human microbiome (Yan et al., 2020).

## Conclusions

The molecular methods and data analyses presented here enabled reliable microbial surveys at the species level and the inference of strain variations, taking into account the most abundant human intestinal microbiota members. To the best of our knowledge, this is the first microbiota assessment of fecal samples derived from donors and recipients involved in an FMT clinical trial to demonstrate microbe engraftment by exploring genetic diversity at the strain level using amplicon sequencing. This long-read approach allows the estimation of species- and strain-level changes in the microbiota at a lower cost (here estimated in one-fourth to one-sixth of the shotgun sequencing cost and including reagents for DNA extraction, PCR, and library preparation) than the expensive shotgun-DNA-sequencing-based metagenomics approach, which has thus far been the only method with the ability to provide this level of information. Thus, we expect our methodology's affordability will help improve microbiome surveys to better define the human gut microbial species associated with health and disease. Despite our promising results, our methodology needs further development to expand the knowledge of *rrn* diversity across the human gut microbiome and other environments. Moreover, we firmly think the utilization of improved releases of Nanopore chemistry (R10 and onwards) and the adoption of updated methods for basecalling, polishing, and alignment will provide a more robust tool for microbiome research.

## List of abbreviations

BMI, body mass index; FDR, false discovery rate; FMT, fecal microbiota transplantation; LMM, linear mixed model; NCBI, National Center for Biotechnology Information; MAG, metagenome-assembled genome; PCoA, principal coordinate analysis; PCR, polymerase chain reaction; RDP, Ribosomal Database Project; rrn, bacterial ribosome RNA operon (16S-ITS-23S); SNP, single-nucleotide polymorphism, SNV, single-nucleotide variation; UC, ulcerative colitis.

# Declarations

*Ethics approval and consent to participate*

The study was prospectively registered at the Dutch Trial registry (https://www.trialregister.nl/trial/4488) and was conducted according to the guidelines laid down in the Declaration of Helsinki and the ethical standards of the responsible local committee on human experimentation of the Amsterdam UMC (location AMC) (Hartstra et al., 2020). The study was registered on August 1st, 2014. The first participant was enrolled on September 1st, 2014.

*Availability of data and material*

The Albacore-basecalled *fast5* files obtained from specific runs are publicly available in the European Nucleotide Archive under accession number PRJEB33947. The updated *rrn* database (*rrn_DBv2*) is publicly accessible at the GitHub repository https://github.com/alfbenpa/rrn_DBv2.

*Competing interests*

The authors have no conflicts of interest to declare.

*Funding*

This study was supported by the EU Project MyNewGut (No. 613979) of the European Commission 7th Framework Programme and grant AGL2017-88801-P from the Ministry of Science, Innovation and Universities (MICIU; Spain), which funded the extension of the contract of ABP. The Miguel Servet CP19/00132 grant from the Spanish “Carlos III” Health Institute (ISCIII) to ABP is fully acknowledged.

*Authors’ contributions*

ABP conceived and designed the study. ABP performed the experimental sequencing research and data analysis, and AVH and MN performed the clinical research. ABP and YS directed the study. ABP and YS wrote the manuscript. All authors reviewed and approved the final version of the manuscript.

*Acknowledgements*

The authors thank the Bioinformatics and Biostatistics Unit from Principe Felipe Research Center (CIPF) for providing access to the cluster, co-funded by European Regional Development Funds (ERDF/FEDER) in Valencian Community 2014-2020.

# References


Aggarwala, V., Mogno, I., Li, Z., Yang, C., Britton, G.J., Chen-Liaw, A., Mitcham, J., Bongers, G., Gevers, D., Clemente, J.C., Colombel, J.F., Grinspan, A., and Faith, J. (2021). Precise quantification of bacterial strains after fecal microbiota transplantation delineates long-term engraftment and explains outcomes. *Nat Microbiol* 6**,** 1309-1318.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., Segata, N., Kyrpides, N.C., and Finn, R.D. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*.

Bainomugisa, A., Duarte, T., Lavu, E., Pandey, S., Coulter, C., Marais, B.J., and Coin, L.M. (2018). A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb Genom* 4.

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.C., Falentin, C., Genete, M., Berrabah, W., Chevre, A.M., Delourme, R., Deniot, G., Denoeud, F., Duffe, P., Engelen, S., Lemainque, A., Manzanares-Dauleux, M., Martin, G., Morice, J., Noel, B., Vekemans, X., D'hont, A., Rousseau-Gueutin, M., Barbe, V., Cruaud, C., Wincker, P., and Aury, J.M. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants* 4**,** 879-887.

Benitez-Paez, A., Portune, K.J., and Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer. *Gigascience* 5**,** 4.

Benitez-Paez, A., and Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION portable nanopore sequencer. *Gigascience* 6**,** 1-12.

Bilinski, J., Grzesiowski, P., Sorensen, N., Madry, K., Muszynski, J., Robak, K., Wroblewska, M., Dzieciatkowski, T., Dulny, G., Dwilewicz-Trojaczek, J., Wiktor-Jedrzejczak, W., and Basak, G.W. (2017). Fecal Microbiota Transplantation in Patients With Blood Disorders Inhibits Gut Colonization With Antibiotic-Resistant Bacteria: Results of a Prospective, Single-Center Study. *Clin Infect Dis* 65**,** 364-370.

Brewer, T.E., Albertsen, M., Edwards, A., Kirkegaard, R.H., Rocha, E.P.C., and Fierer, N. (2020). Unlinked rRNA genes are widespread among bacteria and archaea. *ISME J* 14**,** 597-608.

Cammarota, G., Ianiro, G., and Gasbarrini, A. (2014). Fecal microbiota transplantation for the treatment of Clostridium difficile infection: a systematic review. *J Clin Gastroenterol* 48**,** 693-702.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., Mcdonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7**,** 335-336.

Ciuffreda, L., Rodriguez-Perez, H., and Flores, C. (2021). Nanopore sequencing and its application to the study of microbial communities. *Comput Struct Biotechnol J* 19**,** 1497-1511.

Cusco, A., Catozzi, C., Vines, J., Sanchez, A., and Francino, O. (2018). Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. *F1000Res* 7**,** 1755.

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G., and Lin, H. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun* 9**,** 4844.

Dowden, R.A., Mcguinness, L.R., Wisniewski, P.J., Campbell, S.C., Guers, J.J., Oydanich, M., Vatner, S.F., Haggblom, M.M., and Kerkhof, L.J. (2020). Host genotype and exercise exhibit species-level selection for members of the gut bacterial communities in the mouse digestive system. *Sci Rep* 10**,** 8984.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* 7**,** e1002195.

Gibbons, S.M., Duvallet, C., and Alm, E.J. (2018). Correcting for batch effects in case-control microbiome studies. *PLoS Comput Biol* 14**,** e1006102.

Hartstra, A.V., Schuppel, V., Imangaliyev, S., Schrantee, A., Prodan, A., Collard, D., Levin, E., Dallinga-Thie, G., Ackermans, M.T., Winkelmeijer, M., Havik, S.R., Metwaly, A., Lagkouvardos, I., Nier, A., Bergheim, I., Heikenwalder, M., Dunkel, A., Nederveen, A.J., Liebisch, G., Mancano, G., Claus, S.P., Benitez-Paez, A., La Fleur, S.E., Bergman, J.J., Gerdes, V., Sanz, Y., Booij, J., Kemper, E., Groen, A.K., Serlie, M.J., Haller, D., and Nieuwdorp, M. (2020). Infusion of donor feces affects the gut-brain axis in humans with metabolic syndrome. *Mol Metab* 42**,** 101076.

Ibironke, O., Mcguinness, L.R., Lu, S.E., Wang, Y., Hussain, S., Weisel, C.P., and Kerkhof, L.J. (2020). Species-level evaluation of the human respiratory microbiome. *Gigascience* 9.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Nieto, T., O'grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J., and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 36**,** 338-345.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30**,** 772-780.

Kerkhof, L.J. (2021). Is Oxford Nanopore sequencing ready for analyzing complex microbiomes? *FEMS Microbiol Ecol* 97.

Kerkhof, L.J., Dillon, K.P., Haggblom, M.M., and Mcguinness, L.R. (2017). Profiling bacterial communities by MinION sequencing of ribosomal operons. *Microbiome* 5**,** 116.

Kielbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res* 21**,** 487-493.

Krehenwinkel, H., Pomerantz, A., Henderson, J.B., Kennedy, S.R., Lim, J.Y., Swamy, V., Shoobridge, J.D., Graham, N., Patel, N.H., Gillespie, R.G., and Prost, S. (2019). Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience* 8.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., Juncker, A.S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., Liang, S., Zhang, D., Zhang, Z., Chen, W., Zhao, H., Al-Aama, J.Y., Edris, S., Yang, H., Wang, J., Hansen, T., Nielsen, H.B., Brunak, S., Kristiansen, K., Guarner, F., Pedersen, O., Dore, J., Ehrlich, S.D., Meta, H.I.T.C., Bork, P., Wang, J., and Meta, H.I.T.C. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 32**,** 834-841.

Li, Q., Li, H., Huang, W., Xu, Y., Zhou, Q., Wang, S., Ruan, J., Huang, S., and Zhang, Z. (2019). A chromosome-scale genome assembly of cucumber (Cucumis sativus L.). *Gigascience* 8.

Malmberg, M.M., Spangenberg, G.C., Daetwyler, H.D., and Cogan, N.O.I. (2019). Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (Brassica napus L.). *Sci Rep* 9**,** 8688.

Masonbrink, R., Maier, T.R., Muppirala, U., Seetharam, A.S., Lord, E., Juvale, P.S., Schmutz, J., Johnson, N.T., Korkin, D., Mitchum, M.G., Mimee, B., Den Akker, S.E., Hudson, M., Severin, A.J., and Baum, T.J. (2019). The genome of the soybean cyst nematode (Heterodera glycines) reveals complex patterns of duplications involved in the evolution of parasitism genes. *BMC Genomics* 20**,** 119.

Naum, M., Brown, E.W., and Mason-Gamer, R.J. (2008). Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the enterobacteriaceae? *J Mol Evol* 66**,** 630-642.

Ott, S.J., Waetzig, G.H., Rehman, A., Moltzau-Anderson, J., Bharti, R., Grasis, J.A., Cassidy, L., Tholey, A., Fickenscher, H., Seegert, D., Rosenstiel, P., and Schreiber, S. (2017). Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection. *Gastroenterology* 152**,** 799-811 e797.

Paramsothy, S., Nielsen, S., Kamm, M.A., Deshpande, N.P., Faith, J.J., Clemente, J.C., Paramsothy, R., Walsh, A.J., Van Den Bogaerde, J., Samuel, D., Leong, R.W.L., Connor, S., Ng, W., Lin, E., Borody, T.J., Wilkins, M.R., Colombel, J.F., Mitchell, H.M., and Kaakoush, N.O. (2019). Specific Bacteria and Metabolites Associated With Response to Fecal Microbiota Transplantation in Patients With Ulcerative Colitis. *Gastroenterology* 156**,** 1440-1454 e1442.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M.C., Rice, B.L., Dulong, C., Morgan, X.C., Golden, C.D., Quince, C., Huttenhower, C., and Segata, N. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176**,** 649-662 e620.

Pavlovic, J., Cavalieri, D., Mastromei, G., Pangallo, D., Perito, B., and Marvasi, M. (2021). MinION technology for microbiome sequencing applications for the conservation of cultural heritage. *Microbiol Res* 247**,** 126727.

Peker, N., Garcia-Croes, S., Dijkhuizen, B., Wiersma, H.H., Van Zanten, E., Wisselink, G., Friedrich, A.W., Kooistra-Smid, M., Sinha, B., Rossen, J.W.A., and Couto, N. (2019). A Comparison of Three Different Bioinformatics Analyses of the 16S-23S rRNA Encoding Region for Bacterial Identification. *Front Microbiol* 10**,** 620.

Pruesse, E., Peplies, J., and Glockner, F.O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28**,** 1823-1829.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glockner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41**,** D590-596.

Rinott, E., Youngster, I., Yaskolka Meir, A., Tsaban, G., Zelicha, H., Kaplan, A., Knights, D., Tuohy, K., Fava, F., Scholz, M.U., Ziv, O., Reuven, E., Tirosh, A., Rudich, A., Bluher, M., Stumvoll, M., Ceglarek, U., Clement, K., Koren, O., Wang, D.D., Hu, F.B., Stampfer, M.J., and Shai, I. (2021). Effects of Diet-Modulated Autologous Fecal Microbiota Transplantation on Weight Regain. *Gastroenterology* 160**,** 158-173 e110.

Sakai, J., Tarumoto, N., Kodana, M., Ashikawa, S., Imai, K., Kawamura, T., Ikebuchi, K., Murakami, T., Mitsutake, K., Maeda, T., and Maesaki, S. (2019). An identification protocol for ESBL-producing Gram-negative bacteria bloodstream infections using a MinION nanopore sequencer. *J Med Microbiol* 68**,** 1219-1226.

Schloss, P.D., Gevers, D., and Westcott, S.L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 6**,** e27310.

Schloss, P.D., Jenior, M.L., Koumpouras, C.C., Westcott, S.L., and Highlander, S.K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* 4**,** e1869.

Shin, H., Lee, E., Shin, J., Ko, S.R., Oh, H.S., Ahn, C.Y., Oh, H.M., Cho, B.K., and Cho, S. (2018). Elucidation of the bacterial communities associated with the harmful microalgae Alexandrium tamarense and Cochlodinium polykrikoides using nanopore sequencing. *Sci Rep* 8**,** 5323.

Shin, J., Lee, S., Go, M.J., Lee, S.Y., Kim, S.C., Lee, C.H., and Cho, B.K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Sci Rep* 6**,** 29681.

Solares, E.A., Chakraborty, M., Miller, D.E., Kalsow, S., Hall, K., Perera, A.G., Emerson, J.J., and Hawley, R.S. (2018). Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3 (Bethesda)* 8**,** 3143-3154.

Tyson, J.R., O'neil, N.J., Jain, M., Olsen, H.E., Hieter, P., and Snutch, T.P. (2018). MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res* 28**,** 266-274.

Verma, S., Dutta, S.K., Firnberg, E., Phillips, L., Vinayek, R., and Nair, P.P. (2021). Identification and engraftment of new bacterial strains by shotgun metagenomic sequence analysis in patients with recurrent Clostridioides difficile infection before and after fecal microbiota transplantation and in healthy human subjects. *PLoS One* 16**,** e0251590.

Wilson, B.C., Vatanen, T., Jayasinghe, T.N., Leong, K.S.W., Derraik, J.G.B., Albert, B.B., Chiavaroli, V., Svirskis, D.M., Beck, K.L., Conlon, C.A., Jiang, Y., Schierding, W., Holland, D.J., Cutfield, W.S., and O'sullivan, J.M. (2021). Strain engraftment competition and functional augmentation in a multi-donor fecal microbiota transplantation trial for obesity. *Microbiome* 9**,** 107.

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

Yu, E.W., Gao, L., Stastka, P., Cheney, M.C., Mahabamunuge, J., Torres Soto, M., Ford, C.B., Bryant, J.A., Henn, M.R., and Hohmann, E.L. (2020). Fecal microbiota transplantation for the improvement of metabolism in obesity: The FMT-TRIM double-blind placebo-controlled pilot trial. *PLoS Med* 17**,** e1003051.

# Tables

Table 1. Changes in abundance and prevalence of gut microbial species after FMT.

| Species | Abundance PRE-FMT (N = 10)[1] | Abundance POST-FMT (N = 10)[1] | LMM estimate[2] | Prevalence PRE-FMT (N = 10) | Prevalence POST-FMT (N = 10) | *p*-value |
|---|---|---|---|---|---|---|
| *Parabacteroides merdae* | 0.58 ± 1.37 | 2.80 ± 0.79 | 2.07 | 100% | 100% | 0.002 |
| *Prevotella copri* | -0.04 ± 0.50 | 0.90 ± 0.57 | 1.35 | 90% | 100% | 0.033 |
| *Sutterella massiliensis* | 0.44 ± 1.61 | 2.36 ± 1.53 | 1.64 | 90% | 100% | 0.003 |
| *Paraprevotella clara* | -0.31 ± 1.14 | 1.18 ± 1.26 | 1.49 | 70% | 100% | 0.016 |
| *Alistipes shahii* | 1.36 ± 0.87 | 2.27 ± 0.66 | 1.05 | 90% | 100% | 0.005 |
| *Butyricimonas virosa* | -0.38 ± 0.86 | 0.43 ± 0.92 | 1.01 | 80% | 100% | 0.013 |
| *Bacteroides eggerthii* | -0.03 ± 0.89 | 0.50 ± 0.57 | 0.49 | 100% | 100% | 0.029 |
| *Clostridium saccharolyticum* | -0.03 ± 0.84 | -0.54 ± 0.76 | -0.57 | 80% | 100% | 0.076 |
| *Bifidobacterium adolescenti* | 0.38 ± 0.54 | -0.50 ± 0.91 | -0.92 | 70% | 70% | 0.027 |
| *Phascolarctobacterium faecium* | 1.64 ± 1.96 | 0.51 ± 1.45 | -1.47 | 70% | 90% | 0.058 |

1 Data expressed as the mean of clr-normalized counts ± standard deviation (sd). N, sample size.

2 LMM estimate resulting from the global pattern of abundance in POST-FMT samples with respect to baseline information (PRE-FMT samples). Species altered as a consequence of FMT ordered by the size effect (LMM estimate).

Table 2. Clinical variables altered after FMT.

| Clinical outcome | PRE-FMT (N = 10)[1] | POST-FMT (N = 10)[1] | Statistics |
|---|---|---|---|
| **Diastolic blood pressure** | 87.4 ± 8.2 | 72.0 ± 9.9 | $v$ = -16.9, $p$ < 0.001 |
| **Systolic blood pressure** | 138.6 ± 15.3 | 124.5 ± 14.2 | $v$ = -14.2, $p$ = 0.028 |
| **Fasting glucose (mmol/L)** | 5.49 ± 0.35 | 5.31 ± 0.50 | $v$ = -0.18, $p$ = 0.074 |
| **Fasting insulin (mg/dL)** | 84.7 ± 30.9 | 63.6 ± 21.9 | $v$ = -19.0, $p$ = 0.030 |
| **HOMA-IR** | 2.96 ± 0.99 | 2.16 ± 0.78 | $v$ = -0.93, $p$ = 0.005 |
| **Hba1c (mmol/L)** | 36.5 ± 4.1 | 35.7 ± 3.6 | $v$ = -0.96, $p$ = 0.060 |
| **Plasma HDL (mmol/L)** | 1.59 ± 0.29 | 1.40 ± 0.23 | $v$ = -0.191 $p$ = 0.032 |
| **Fecal butyrate (µmol/g)** | 88.6 ± 44.2 | 60.3 ± 40.9 | $v$ = -31.6, $p$ = 0.018 |
| **Fecal acetate (µmol/g)** | 434.0 ± 148.3 | 308.2 ± 121.1 | $v$ = -135.7, $p$ = 0.033 |

1 Data expressed as the mean ± standard deviation (sd).

$v$ = variation between groups analyzed by applying an LMM (PRE-FMT group as reference), HDL = high-density lipoprotein, Hba1c = glycosylated hemoglobin.

## Figure legends

**Figure 1**. Graphical description of the study. Data acquisition and processing steps, including sample selection, amplicon sequencing, and the general pipeline for the *de novo* assembly of rrn regions from human gut microbiota, are depicted.

**Figure 2**. Taxonomic identification of final *rrn* assemblies. A - Scatter plot showing information on sequence identity supporting the identification of assembled *rrn* sequences against the NCBI 16S database and the coverage (number of reads) accounting for the respective assemblies. Additional levels of information are included in the plot, such as taxonomic level matches, bias coverage between forward and reverse reads, and alignment length (see graph symbols and their color and size scale, this last feature applicable to other symbols than points). B - Evaluation of indel and mismatch content in the alignments resulting from the cross-identification of *rrn* assemblies against the NCBI 16S database and the nonredundant nucleotide collection (GenBank). The values were discriminated based on the taxonomic level of identification according to the original species-level annotations of the respective *rrn*s (see the color legend). UB, uncultured bacteria.

**Figure 3**. Beta diversity of the microbial communities assessed by *rrn* sequencing. A - Scatter plot compiling data from the multivariate analysis (principal coordinate analysis - PCoA) of the microbiota from recipients and donors participating in FMT. The donor and recipient samples and the sampling time points are defined according to the legend at the top. PCo; principal coordinate (the two most informative coordinates are shown). Color legend corresponds with the distribution of the different PRE-FMT and POST-FMT samples into the groups linked to the four donor samples evaluated. B - A Genetic distance-based approach for evaluating microbiota transfer between donor and recipient pairs. The microbial community structures of the PRE-FMT and POST-FMT samples were compared with those of the donors through the calculation of the Bray-Curtis dissimilarity index and are represented as boxplots. The Wilcoxon signed-rank test for paired samples was used

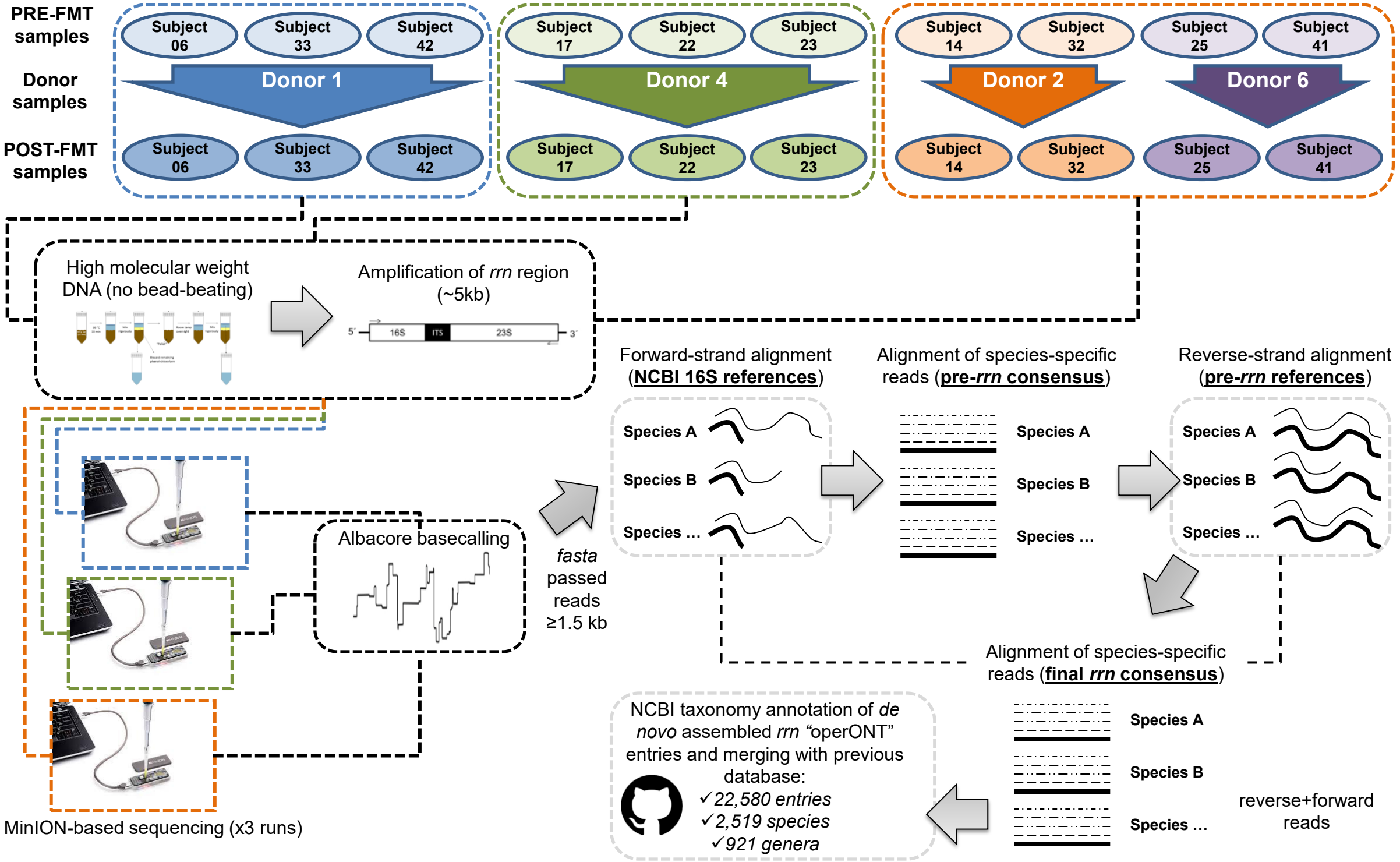
to compare differences in genetic distances. C – Microbiota abundance shifts after FMT. Abundance data distribution for three species selected from those described in Table 1 with meaningful increase after FMT is shown in boxplot fashion.

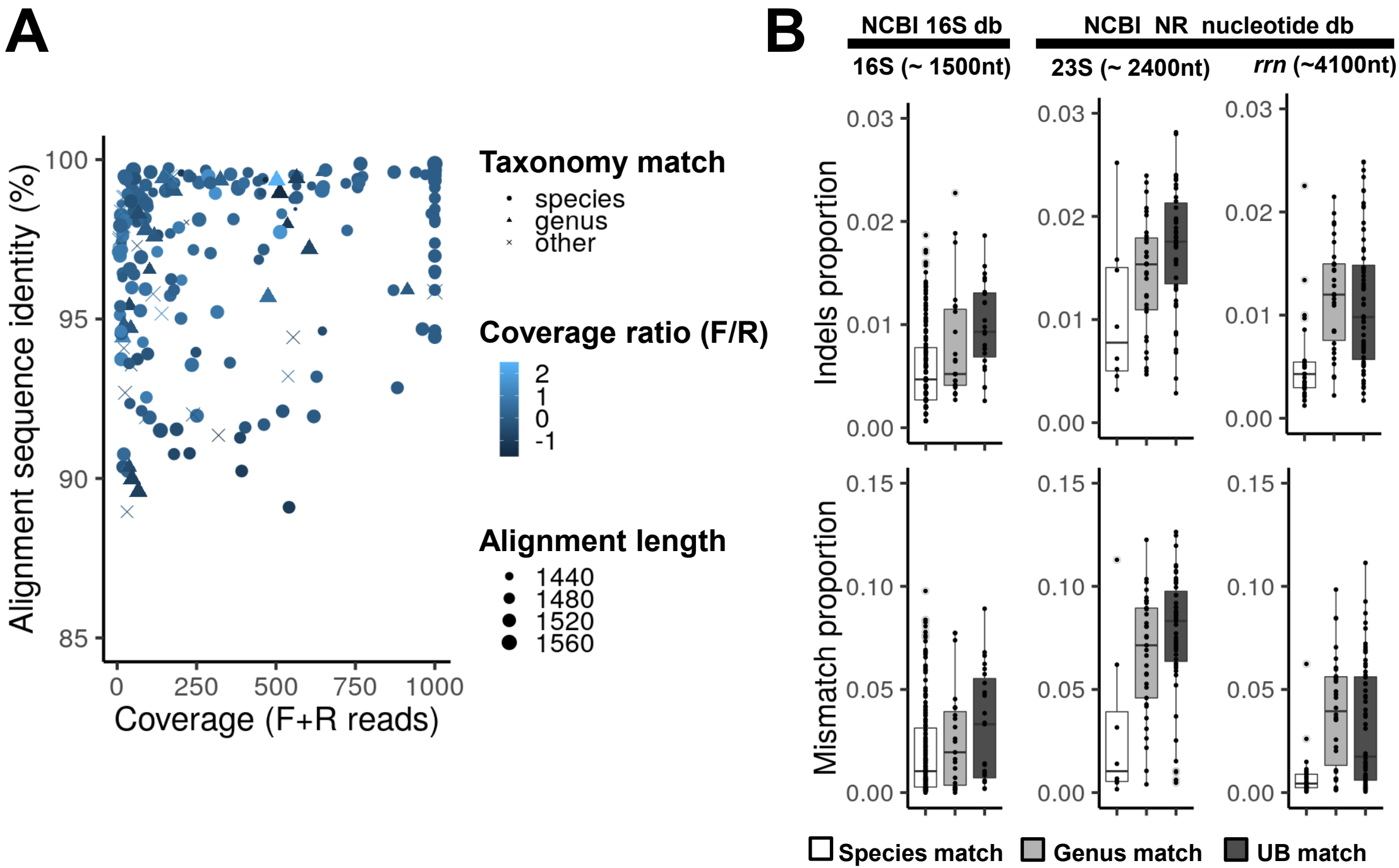
**Figure 4**. Single-nucleotide variation (SNV) analysis to detect species transfer and engraftment. The species *Parabacteroides merdae* and *Faecalibacterium prausnitzii rrn*s were used to reveal SNVs within each species and their abundance as a consequence of FMT (A and B panels, respectively). Variation in one unit of clr-normalized counts is roughly equal to a doubling in raw DNA read counts according to data distribution in both cases. C and D - The mapping, indexing, and pileup of thousands of reads of *P. merdae* and *F. prausnitzii rrns* allowed us to identify polymorphic sites exhibiting an even frequency for at least two nucleotides. These potential SNVs are highlighted inside open circles, and the position number in the *rrn* sequence is also indicated. The structural arrangement of *rrn* is illustrated on the x-axis. The black lines indicate the relative frequency of the dominant allele, whereas the gray lines indicate the relative coverage per site related to the average across the *rrn*. E and F - The distribution of the Bray-Curtis dissimilarity index of the microbiota between POST-FMT and PRE-FMT samples and the corresponding donors, using the combined nucleotide frequencies of SNVs detected in *P. merdae* and *F. prausnitzii*, respectively. The Wilcoxon signed-rank test for paired samples was used to compare the genetic distances of the microbiotas. G – Assessment of delta Bray-Curtis distances computed as POST-FMT vs DONOR samples minus PRE-FMT vs DONOR samples for *Faecalibacterium prasunitzii* and *Parabacteroides merdae* species. H - Electropherograms obtained from Sanger sequencing for the strains of *P. merdae* SNV-3527, SNV-2293, and SNV-1254. The predominant alleles inferred for every sample (based on the Q-score of base calling) supported the hypothesis that the strains were transferred between donor and recipient pairs, as anticipated during the Nanopore-based assessment.

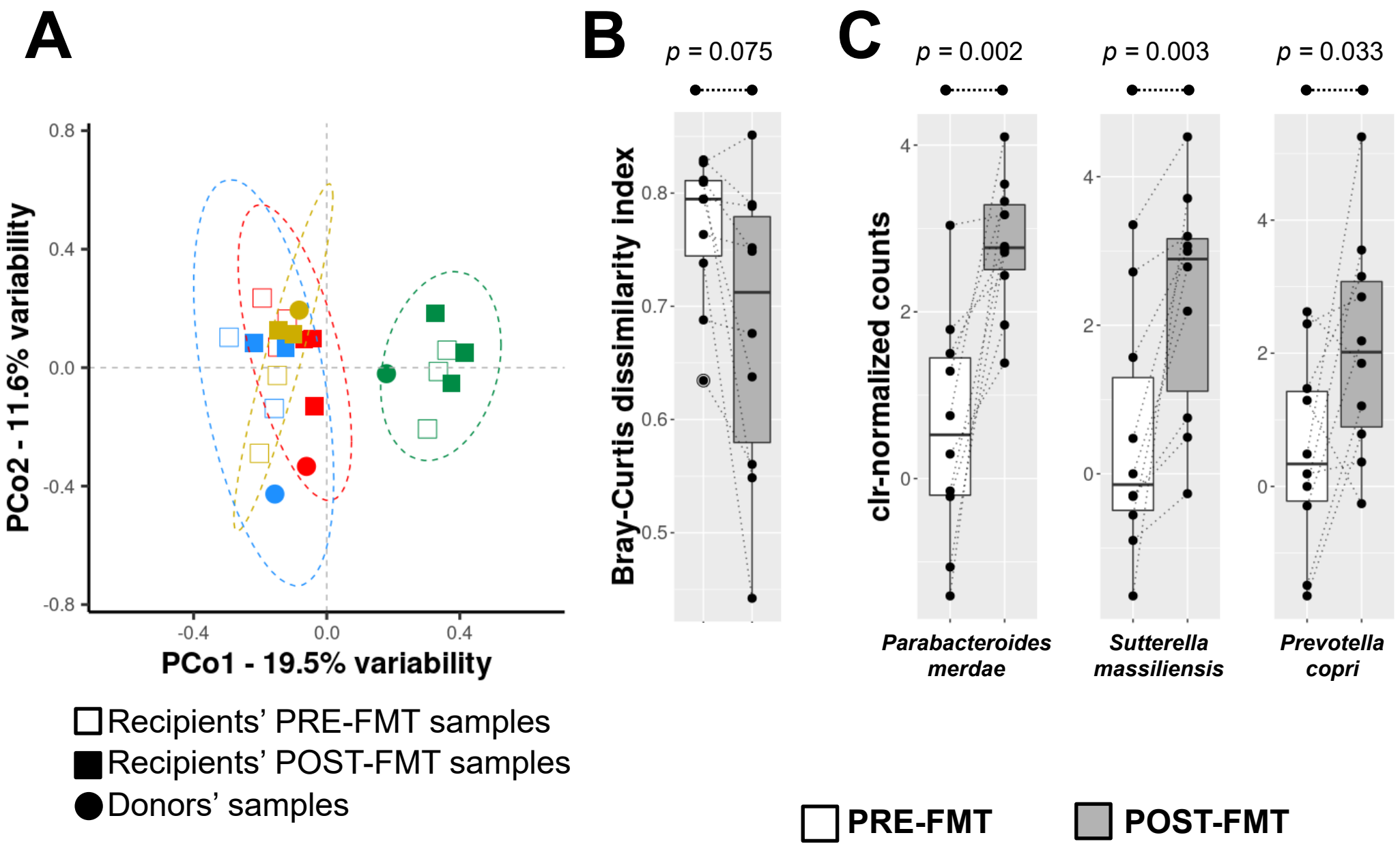
**Figure S1**. BLAST-based results showing the correct identification of assembled *rrn*s. The top hits supporting the identification are shown for six different *rrn*s assembled with low, intermediate or high coverage (number of reads).

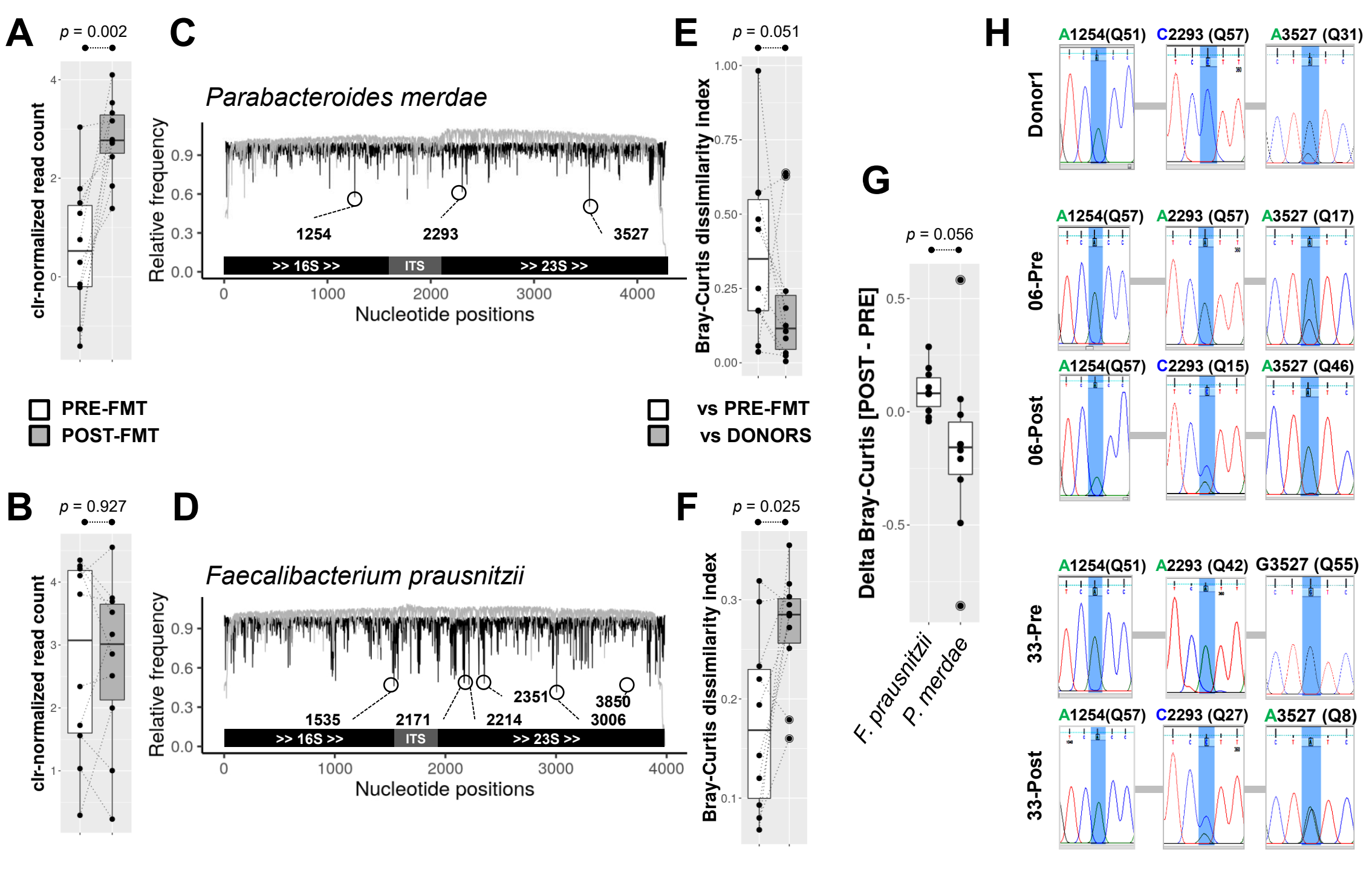
**Figure S2**. Alpha diversity of the recipients' microbiota before and after FMT. The distributions of the values obtained for four alpha diversity indicators, including Chao's index, Shannon's index, the reciprocal Simpson's index, and the dominance index, are shown as boxplots. The results of the Wilcoxon signed-rank test applied to establish differences between the two groups of samples (POST-FMT and PRE-FMT) are also shown.

**Figure S3**. Searching for SNV genotypic variation in the NCBI nonredundant nucleotide database. The nucleotide sequences spanning positions 1180 to 1280, 2250 to 2350, and 3470 to 3570 of the *P. merdae rrn* operon were used as queries in searches against the nonredundant nucleotide database on the BlastN web server (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The results showed the intrinsic variation attributable to SNV-3527 (as an example) present in genomic records from the NCBI nonredundant nucleotide database including strains with genotype A3527 and others with G3527, both of which were variants inferred from our Nanopore-based sequencing. Nucleotides highlighted in yellow show the 5-mer surrounding the region of SNV-3527, revealed in detail by Sanger sequencing, as shown in Figure 4H. The C2293 genotype was retrieved from a unique matching sequence (CL06T03C08 strain), and A1254 was predominantly found in most of the *P. merdae* strains evaluated (data not shown).


PRE-FMT samples
Donor samples
POST-FMT samples
Subject 06
Subject 33
Subject 42
Donor 1
Subject 17
Subject 22
Subject 23
Donor 4
Subject 14
Subject 32
Subject 25
Subject 41
Donor 2
Donor 6
High molecular weight DNA (no bead-beating)
Amplification of rrn region (~5kb)
5'
16S
ITS
23S
3'
Forward-strand alignment (NCBI 16S references)
Alignment of species-specific reads (pre-rrn consensus)
Reverse-strand alignment (pre-rrn references)
Species A
Species B
Species …
Albacore basecalling
fasta passed reads ≥1.5 kb
Alignment of species-specific reads (final rrn consensus)
NCBI taxonomy annotation of de novo assembled rrn "operONT" entries and merging with previous database:
✓22,580 entries
✓2,519 species
✓921 genera
reverse+forward reads
MinION-based sequencing (x3 runs)


A
Alignment sequence identity (%)
100
95
90
85
0
250
500
750
1000
Coverage (F+R reads)
Taxonomy match
species
genus
other
Coverage ratio (F/R)
2
1
0
-1
Alignment length
1440
1480
1520
1560
B
NCBI 16S db
NCBI NR nucleotide db
16S (~ 1500nt)
23S (~ 2400nt)
rrn (~4100nt)
Indels proportion
0.03
0.02
0.01
0.00
Mismatch proportion
0.15
0.10
0.05
0.00
Species match
Genus match
UB match


A
PCo2 - 11.6% variability
PCo1 - 19.5% variability
0.8
0.4
0.0
-0.4
-0.8
-0.4
0.0
0.4
Recipients' PRE-FMT samples
Recipients' POST-FMT samples
Donors' samples
B
p = 0.075
Bray-Curtis dissimilarity index
0.8
0.7
0.6
0.5
C
p = 0.002
p = 0.003
p = 0.033
clr-normalized counts
4
2
0
Parabacteroides merdae
Sutterella massiliensis
Prevotella copri
PRE-FMT
POST-FMT


A
p = 0.002
clr-normalized read count
B
p = 0.927
clr-normalized read count
C
Parabacteroides merdae
Relative frequency
1254
2293
3527
>> 16S >>
ITS
>> 23S >>
Nucleotide positions
D
Faecalibacterium prausnitzii
Relative frequency
1535
2171
2214
2351
3850
3006
>> 16S >>
ITS
>> 23S >>
Nucleotide positions
PRE-FMT
POST-FMT
E
p = 0.051
Bray-Curtis dissimilarity index
F
p = 0.025
Bray-Curtis dissimilarity index
vs PRE-FMT
vs DONORS
G
p = 0.056
Delta Bray-Curtis [POST - PRE]
F. prausnitzii
P. merdae
H
A1254(Q51)
C2293 (Q57)
A3527 (Q31)
Donor1
A1254(Q57)
A2293 (Q57)
A3527 (Q17)
06-Pre
A1254(Q57)
C2293 (Q15)
A3527 (Q46)
06-Post
A1254(Q51)
A2293 (Q42)
G3527 (Q55)
33-Pre
A1254(Q57)
C2293 (Q27)
A3527 (Q8)
33-Post