Strand-wise and bait-assisted assembly of nearly-full *rrn* operons applied to assess species engraftment after faecal microbiota transplantation

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 Centre (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.

Abstract

Background. Effective methodologies to accurately identify members of the gut microbiota at the species and strain levels are necessary to unveiling more specific and detailed host-microbe interactions and associations with health and disease.

Methods. MinION™ MkIb nanopore-based device and the R9.5 flowcell chemistry were used to sequence and assemble dozens of *rrn* regions (16S-ITS-23S) derived from the most prevalent bacterial species in the human gut microbiota. As a method proof-of-concept to disclose further strain-level variation, we performed a complementary analysis in a subset of samples derived from an faecal microbiota transplantation (FMT) trial aiming amelioration of glucose and lipid metabolism in overweight subjects with metabolic syndrome.

Results. The resulting updated *rrn* database, the data processing pipeline, and the precise control of covariates (sequencing run, sex, age, BMI, donor) were pivotal to accurately estimate the changes in gut microbial species abundance in the recipients after FMT. Furthermore, the *rrn* methodology described here demonstrated the ability to detect strain-level variation, critical to evaluate the transference of bacteria from donors to recipients as a consequence of the FMT. At this regard, we showed that our FMT trial successfully induced donors' strain engraftment of e.g. *Parabacteroides merdae* species in recipients by mapping and assessing their associated single nucleotide variants (SNV).

Conclusions. We developed a methodology that enables the identification of microbiota at species- and strain-level in a cost-effective manner. Despite its error-prone nature and its modest per-base accuracy, the nanopore data showed to have enough quality to estimate single-nucleotide variation. This methodology and data analysis represents a cost-effective manner to trace genetic variability needed for better understanding the health effects of the human microbiome.

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, faecal microbiota transplantation, single nucleotide variation, species-level resolution, *rrn* operon.

Background

Studying complex human-associated microbial communities demands the development of cost-effective sequencing strategies providing informative DNA pieces and high coverage outputs, thus permitting to discern the specific species or strains inhabiting an ecosystem. Metagenomics based on DNA shotgun sequencing is, up to date, the only strategy that enable us to reach such a resolution level, but it is still highly expensive. This makes it unaffordable for most of the studies aiming to define microbiota features associated with health or disease including large number of samples to minimize the noise introduced by other covariates that contribute to the high inter-individual variability of the microbiota. As a consequence, most of the microbiome analyses have been performed through targeted amplification of universal gene markers, mainly few hypervariable regions (e.g. V3 and/or V4) of the bacterial 16S rRNA gene. This protocol shows limitations since it only allows to capture changes affecting the global microbiome structure but does not permit precise taxonomic identifications at the species levels, mostly because of the limited resolution of this gene marker among closely related species. We have developed a nanopore-based sequencing method to improve the resolution of taxonomy identifications by studying the microbial genetic variability of the nearly-full 16S rRNA gene previously [1]. This approach has also demonstrated to perform well in a wide variety of microbiota inventories [2-5]. More recently, we have also pioneered a new methodology combining the sequencing of an extremely variable multi-locus region with sample multiplexing [6] for improving the species-level characterisation of complex microbial communities [2, 7]. 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 to compare the long-reads generated from nanopore-based devices. In this regard, we have made a great effort to compile a vast amount of genetic information of *rrn* sequences from thousands of bacterial species [6]. Nonetheless, the limited amount of properly annotated DNA sequences available in public repositories of bacteria from multiple ecosystems will restrict the utilisation of this approach for less explored habitats [2, 8]. Moreover, recent studies indicate that the architecture of the *rrn* region is not equally conserved in all bacteria; thus, microbial communities enriched in Deinococcus-Thermus, Chloroflexi, Planctomycetes phyla might be difficult to be studied by such pipeline [9]. Anyhow, the utility of the *rrn* to survey the biological diversity has been explored across-kingdoms, showing promising results also for the identification of metazoans by metabarcoding approaches [10].

Although the issues concerning unlinked 16S rRNA and 23S rRNA markers can be addressed, in certain bacteria groups and environmental samples, through the study of individual regions, the database completeness and proper annotation are pivotal for the complete implementation of this molecular appraisal. For that reason in this study, we aimed to perform a bait-assisted assembly of the nearly-full *rrn* regions of the human intestinal microbiota by using nanopore-based sequencing. Therefore, we provided *de novo* information of the *rrn* region of microorganisms from the human gut, one of the environments most extendedly investigated, enriching the *rrn* database with annotation of dozens of microbial species and strains. As a proof of concept of the utility of this methodology we have estimated the changes in the gut microbiota, at the species-level, as a consequence of a faecal microbiota transplantation (FMT) intervention in humans, using the updated *rrn* database. Furthermore, we have explored the potential of this approach to unveil the strain-level variation to track species engraftment after FMT.

Methods

Subjects, samples, and clinical data

Samples were obtained upon informed consent from a previous FMT clinical trial carried out in the frame of the MyNewGut project. The details of the study design are publicly available elsewhere [11]. Briefly, a total of twenty-four faecal samples were analysed in the present study. Twenty samples were obtained from 10 recipients involved in the allogenic FMT, who provided one sample before (PRE-FMT faecal samples) and 4 weeks after (POST-FMT faecal samples) the intervention. The samples of 4 donors were also analysed and the effect of this variable in the recipients was considered in the data analysis. This subset of samples was mainly assessed to explore the differential species engraftment in multiple FMT recipients with common donors.

DNA extraction, multi-locus amplification, and sequencing

Microbial DNA was recovered from 100 mg faeces by using the QIAamp® Fast DNA Stol Mini kit (Qiagen, Hilden, Germany) according to manufacturer's instructions and omitting cell disruption by mechanical methods (bead-beating) to preserve DNA with high molecular weight. The *rrn* region comprising the nearly-full bacterial RNA ribosomal operon (16S, ITS, and 23S) was amplified as previously published [6]. Dual-barcoded purified PCR products were mixed in equimolar proportions before sequencing library preparation. In total, three different libraries were prepared from ~1 ug mixed amplicon DNA (containing 7, 8, and 10 barcoded samples, respectively) 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 respective FLO-MIN107 (R9.5) flowcells (Oxford Nanopore Technologies, Oxford, UK) and sequencing was carried out in the portable sequencer MinION™ MkIb (Oxford Nanopore Technologies, Oxford, UK) operated with *MINKNOW* v1.10.23 software (Oxford Nanopore Technologies, Oxford, UK). Flowcells were primed according to manufacturer instructions, and then a ~18h run of 1D sequencing was executed for each library and flowcell, respectively.

Data pre-processing

Fast5 files were processed with the *albacore* v2.1.3 basecaller 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 [6]. A size filtering step was configured to retain those reads with at least 1,500 nt in length. Barcode and primer sequences were then removed by trimming 50 nucleotides at 5' end of forward and reverse reads.

Bait-assisted rrn assembly

The study design and the assembly of *rrn* is graphically explained in Figure 1 and performed in different steps as follows:

- reads from all the 24 samples were merged respectively into forward and reverse subsets.
- forward read binning into precise microbial species by using competitive alignment against the non-redundant 16S NCBI database (release January 2018). The *LAST* aligner [12] with -s 2 -q 1 -b 1 -Q 0 -a 1 -r 1 configuration was used for such aim. Alignment score was the main criterion to select top-hits. In case of multiple hits with same top score, the alignment was discarded for downstream processing. The sequence identity (sequence identity above the 33th percentile ≥ 85%) and length information (alignments ≥ 1500 nt) were used to retaining high-quality alignments.
- a maximum of 500 forward reads per species (randomly shuffled), producing high-quality alignments, were selected, aligned through iterative refinement methods implemented in *MAFFT* v7.310 and default parameters [13], and then consensus sequenced obtained by using *hmmbuild* and *hmmemit* algorithms implemented in *HMMER3* [14]. No significant improvements on consensus *rrn* were obtained using more than 500 sequences per species.
- consensus sequences were annotated according to NCBI taxonomy and used as reference to binning reverse reads in similar manner as made for forward reads. The thresholds for selection of high-quality alignments based on reverse reads and preliminary assemble of *rrn* regions were 85% sequence identity (upper 45th percentile) and ≥ 3500 nt in length.
- reverse reads were compared against the consensus sequences obtained from forward reads assemblies and binned into respective species according to the NCBI taxonomy annotation. A maximum of 500 reverse reads per species (randomly shuffled), producing high-quality

alignments, were selected, merged with maximum 500 forward reads assigned reciprocally to same species (randomly shuffled), aligned through iterative refinement methods implemented in *MAFFT*, and then a new consensus sequenced was obtained by using *hmmbuild* and *hmmemit* algorithms implemented in *HMMER3*.

- final consensus sequences were annotated according to concordant NCBI taxonomy, obtained from forward and reverse reads, and used as a reference to study abundance and prevalence of gut microbiota at the species level.
- a blast-based search of *rrn* assemblies against then non-redundant nucleotide and reference 16S NCBI databases was accomplished to evaluate the identity of the operons reconstructed. Similarly, annotation of *rrn* assemblies was evaluated against the SILVA database [15] through SINA aligner [16]. 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 [6] was updated with more than two-hundred new *rrn* operon sequences assembled in this study (labeled as operONT) and re-annotated to different taxonomy levels according to the NCBI taxonomy database. The mapping of sample reads was performed against the *rrn_DBv2* by competitive alignment using *LAST* aligner similarly as above stated. Taxonomy assignment was based on best hit retrieved by calculation of TS (see above) and reliable taxonomy assignments were filtered to those based on alignments with at least 85% sequence identity and longer than 1500 nt. Species read counts were normalised, and a covariate-controlling linear mixed models based method was used to explore the differential abundance of taxonomy features between conditions as stated in the next paragraph. The species engraftment was evaluated by a selection of all sample reads mapped to the *Parabacteroides merdae* and *Faecalibacterium prausnitzii* species. To detect informative sites, based on single nucleotide variants probably linked to strain variation, we firstly proceed to map all selected reads to the respective reference *rrn* using *LAST* aligner and the parameters set across this study (see *rrn* assembly). The *maf-convert* algorithm (*LAST* aligner toolkit) was used to create respective *sam* files. The algorithms compiled into *samtools* v1.3.1 were used to index, order, and pileup reads as well as retrieving the information regarding to the nucleotide frequency per site and coverage (*vcf* files). The site selection was based on positions of *rrn* with the lowest frequency of the dominant allele, without reductions in the coverage (>75% of the relative coverage), thus obtaining sites with a balanced representation of at least the two most predominant alleles. Secondly, the samples reads mapped to the reference species were individually assessed to detect meaningful changes between recipients' PRE-FMT and POST-FMT in intestinal bacterial genotypes. Changes in nucleotide frequencies per SNV site were assessed by calculating the Bray-Curtis dissimilarity index between pairs of POST-FMT samples and PRE-FMT or respective donor samples following statistical evaluation by using Wilcoxon signed rank test for paired samples.

Sanger sequencing and validation of Parabacteroides merdae SNVs

The reference *rrn* from *P. merdae* was submitted to the Primer-Blast web server (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) to retrieve specific primer pairs to amplify selectively this *rrn* and primers flanking the SNV-743, SNV-1975, and SNV-3016 for Sanger sequencing. The comparison against the non-redundant NCBI database and *Parabacteroides* [taxid:375288] as reference organism were fixed as checking parameters for primer prediction. The *rrn* region from *P. merdae* was amplified by 28-PCR cycles, including the following stages: 95ºC for 20 s, 61ºC for 30 s, and 72ºC for 150 s. Phusion High-Fidelity Taq Polymerase (Thermo Scientific) and the pm485 (TTTCGCACAGCCATGTGTTTTGTT) and pm3647 (TGCCGTTGAAACTGGGTTACTTGA) primer pair, were used in the amplification reaction. PCR products were cleaned with Illustra GFX PCR DNA and gel band purification kit (GE Healthcare, Chicago, IL, USA) and sequenced in by Sanger technology in an ABI 3730XL sequencer (STAB-VIDA. Caparica, Portugal) using the described primers above (pm485 and pm3647), additionally to the primers pm1584 (TTCGCGTCTACTCACTCCGACTAT) and pm2415 (ACCCCTTACGGAGTTTATCGTGGA). The sequencing electropherograms from *ab1* files were visualised with FinchTV v1.4.0 (Geospiza Inc.).

Diversity and taxonomic analyses at the species level

Prevalence and abundance of a total of 2,519 species contained in the database was evaluated, and diversity analyses were completed taking into account the species with > 0.01% of 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 obtaining by using *qiime* v1.9.1 [17]. Similarly, *qiime* was used to calculate Bray-Curtis dissimilarity index among samples and to perform multivariate exploratory (Principal Coordinate Analysis - PCoA) and statistical (PERMANOVA) analyses. A linear mixed model (LMM - *nlme* R package) analysis was also conducted on log-transformed and normalised data to detect differential features in the microbiota before and after the intervention. Inherent variation due to individual features was set as a random effect for each variable analysed (fixed effect). The possible covariates of the clinical and faecal microbiota that showed differences between the study groups (p-value $\leq$ 0.05) were selected. Recognised covariates of microbiota such as age, sex, baseline BMI, and sequencing batch were identified as significant in this study as well and, also, included as random effects in the LMM. To identify microbial species potentially linked to clinical variables altered as a consequence of the FMT the Kendall's τ (tau) between variable-pairs was estimated and corrected for multiple testing, using false discovery rate (FDR) approach. Associations were selected when FDR p-value $\leq$ 0.1. Graphics were performed on R v3.6 using *ggplot2* and *ggridges* packages.

Results

Assembly and taxonomic identification of human gut microbiota-derived rrn sequences

We retrieved a total of ~516k reads after base calling from three MinION™/R9.5 sequencing runs, ~430k reads after size trimming (83.3% retained), and ~427k reads successfully demultiplexed (82.7%). The number of the forward and reverse reads obtained after demultiplexing was 210,736 and 216,617 (0.49 and 0.51 proportions), respectively. During the first step of the bait-based *rrn* assembly (Figure 1), based on mapping of forward reads against the non-redundant NCBI 16S database and selection of high-quality alignments, we detected the presence of 381 different species. However, the preliminary *rrn* assembly was initiated only for those species with at least 5X coverage, 250 in total. The mapping of reverse reads against the preliminary *rrn* assemblies permitted to confirm the detection of 229 microbial species. Figure S1 shows the initial assessment of alignments produced from forward and reverse reads according to different sequencing batches of our study. Additionally, the high-quality alignments (see methods) resulting from mapping forward and reverse reads against respective reference databases were evaluated to determine mismatch and indels proportions in a microbial species-wise manner. The above analysis for the top-20 most abundant species detected in the strand-based mapping of reads is depicted in Figure 2. After measuring the proportion of mismatches and indels (opened gaps in queries and targets under alignment scoring configuration - see methods), we observed that forward reads produced alignments on 16S rRNA gene sequences with a homogeneous distribution of indels across the species (observation extended to less abundant species), and that variation in mismatch rates were peculiarly more pronounced in species such as *Oscillibacter valericigenes*, *Phascolarctobacterium faecium*, and *Roseburia hominis*, thus suggesting a probable detection of strain-associated genetic variation for the microbial communities evaluated. Similar patterns were observed for the appraisal of the alignments produced from reverse reads indicating there were no drastic changes in the quality of alignments originated by both subsets of reads. The absence of several and notable shifts in the distribution of mismatches in reverse reads was expected since the reference database used for such mapping is thought to contain already the potential genetic variation uncovered from the previous forward reads binning (Figure 2).

After merging forward and reverse reads to obtain final *rrn* assemblies, we retrieved a total of 229 *rrn* sequences that were subject of cross-identification to evaluate the taxonomy annotation using different methods and databases used by dozens of taxonomy classifiers as reference. We found limitations to do so given the scarce taxonomic information of this multi-locus region despite the fact that it contains the classical marker for species bacterial identification, the 16S rRNA gene. When we explored the SILVA "ssu" and "lsu" databases (for analysis of 16S and 23S rRNA genes, respectively) through the SINA aligner, we obtained classifications towards genus level given this is the deeper taxonomy level predominantly found in this database. Accordingly, we retrieved a genus match 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 family and order levels. Similar performances were observed when using the 16S taxonomic classification of the RDP and Greengenes databases. Additionally, we submitted the set of 229 *rrn* sequences to the Blast server at NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to be compared to the non-redundant nucleotide collection and the reference 16S database. Top hits from the non-redundant nucleotide database (based on the TS score - see methods) indicated that only 124 *rrn* sequences (54%) produced 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 query length, we observed that those supporting the species match had an averaged sequence identity of 98.61 ± 1.69 (mean ± sd), whereas those supporting genus and "uncultured bacteria" matches had an averaged sequence identity of 94.69 ± 3.43 and 95.46 ± 3.67, respectively. Similar distributions were observed for alignments covering less than 95% of query sequences, where predominantly the 23S region was preferentially explored (alignment length ~2,630 nt on average). Examples of full *rrn* sequences retrieving matched annotations and including those with lower, mid, and higher coverage are disclosed in Figure S2.

In summary, we obtained the best results when comparing the *rrn* sequences against the reference 16S NCBI database (release May 2019). This enable us to identify correctly 183 of the sequences (80%) at the species level and 23 (10%) at genus level. The remained 23 *rrn* sequences (10%) matched with species belonging to different genus, but in all cases were related to species of the Enterobacteriaceae family (e.g. *Shigella*, *Escherichia*, *Citrobacter*, *Salmonella*) difficult to distinguish by inspecting only the 16S rRNA gene sequence [18]. The global assessment of our assembled set of *rrn* sequences against the 16S NCBI database is shown in Figure 3A, where five levels of information are compiled including sequence identity, alignment length, coverage, coverage ratio (forward vs reverse reads employed in assembly), and level of match. We observed species matches even when coverage was very low (e.g. assemblies based on ten reads), and species matching with modest sequence identity (~89%), possibly indicating accumulation of high genetic variability at the strain level for certain species. The comparison of indels and mismatch proportions retrieved from alignments supporting best hits against NCBI databases, the 16S and non-redundant nucleotide collection (this last discriminating 23S alignments from those of whole *rrn*), showed an expected progressive increase in either mismatch and indels proportions from queries identified at the species, genus, and other taxonomy levels (unclassified bacteria included) (Figure 3B). However, this was only observed for 16S and 23S alignments separately since for those covering the entire *rrn* the highest mismatch and indels rates were detected for identifications at genus level (Figure 3B).

FMT associated microbiota shifts

All forward and reverse reads were mapped to assess shifts in diversity and taxonomic features. Of all reads, only 43% supported alignments with the top quality (identity and length), and used in downstream analyses. After the taxonomy assignment, we found not drastic changes in any of the alpha diversity descriptors analysed. Notwithstanding, we found that the FMT increased the richness (Chao's index) of the microbiota in six out of the ten recipients (Figure S3), who showed a gain-of 31 species on average. In the remaining four recipients there was a reduction of richness resulting in an averaged loss of 16 species as a result of the FMT. The beta diversity evaluation, based on the Bray-Curtis dissimilarity index, indicated minor shifts in the microbial structure of recipients as a consequence of the FMT (PERMANOVA = 1.02, p = 0.403). Nevertheless, the gut microbiota composition of the recipients was strongly influenced, from larger to a lesser extent, by the sequencing run (PERMANOVA = 3.69, p = 0.001), the sex (PERMANOVA = 1.70, p = 0.032), and the donor (PERMANOVA = 1.69, p = 0.004). The result of this multivariate analysis is shown in Figure 4A. Globally, we observed that POST-FMT samples tended to map closer to those from donors. To further assess this hypothesis, we compared the distances (Bray-Curtis metrics) between the respective donors and the PRE-FMT microbiota, and the donors and the POST-FMT microbiota. As a result, we noted that donors' microbiota and PRE-FMT pairs were more dissimilar when compared to POST-FMT pairs, as indicated by the decreased Bray-Curtis distance. (p = 0.075) (Figure 4B).

Additionally, we performed a LMM analysis disclosing similar results than beta diversity evaluation. We found that sequencing batch, sex, and donors were the main covariates influencing the microbiota data. Furthermore, we found that age and baseline BMI also explain, to some extent, the gut microbiota variation between the subjects involved in this study. After including the variables mentioned above as random effects in the model, a list of microbial species altered as a consequence of the FMT was retrieved (Table 1). Seventeen different microbial species were found to be differentially abundant when comparing paired samples obtained before and after FMT, and only three of them seemed to decline because of the transplantation, e.g. *Bifidobacterium adolescentis*. Moreover, a kind of species replacement effect between *Ruminococcus bicirculans* and *Ruminococcus callidus* was observed. According to abundance and the occurrence pattern, *R.*

callidus seemed to occupy the niche of *R. bicirculans*. On the other hand, we detected several potential bacterial consortia consisting of closely related species, such as those included in the *Parabacteroides*, *Butyricimonas*, and *Sutterella* genera which all tended to raise as a consequence of the FMT (Table 1).

Gut microbial species transferred and engrafted

In order to assess species engraftment, we next deeply analysed the donor and recipient microbiota pairs at species and strain level. For that purpose, we selected *Parabacteroides merdae*, a predominant species in the samples assessed and showing a remarkable change as a result of the FMT (Figure 5A), and *Faecalibacterium prausnitzii*, a highly abundant species in the samples analysed as well, for which an evident transference between donor-recipient pairs was not observed (Figure 5B), to detect single nucleotide variation (SNV) associated with strains. After massive analysis of the total dataset, we found three potential informative SNV sites in the *P. merdae rrn*, whereas six were found in the *F. prausnitzii* counterpart (Figure 5C-D). When we studied the nucleotide frequencies of these particular SNVs across samples, a clear transference pattern of *P. merdae* genotype from donors-to-recipients was detected as indicated by the decreased genetic distance (Bray-Curtis) between POST-FMT and donor samples when compared to that of POST-FMT and PRE-FMT pairs (Figure 5E). By contrast, the genetic distances retrieved after comparison of *F. prausnitzii* genotypes between POST-FMT or PRE-FMT samples and their donors did not indicate transference of strains belonging to this species from donors to recipients (Figure 5F). Direct Sanger sequencing of *P. merdae* SNV-743, SNV-1975, and SNV-3016 in two recipients and their common donor, supported the accuracy of our long-read based assessment of FMT (Figure 5G). Globally, we observed the presence of a mix of strains in some of the samples analysed, given the basecalling profile visualised in the electropherograms (e.g. SNV-743 of Donor1 and 06-Pre samples). This pattern of strain co-existence was more evident in POST-FMT samples of both recipients analysed (see SNV-743 and SNV-1975 in Figure 5G). The predominant haplotype observed in the Donor1 (A743-C1975-G3016) was transferred to the two recipients explored, and became dominant after the FMT. Similar patterns of transference 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's strain transmission. This is likely the case for other bacterial species increased as a result of the FMT.

Correlation between clinical variables and the species abundance after FMT

Among multiple clinical variables evaluated in this cohort of subjects, FMT induced remarkable changes in markers of glucose metabolism and blood pressure of the recipients (Table 2). Fasting insulin (p = 0.030), fasting glucose (p = 0.074), and consequently, the HOMA-IR (p = 0.005) were improved in recipients after the FMT. In line with these findings, there was a decrease in the glycosylated haemoglobin concentration after the FMT as well (p = 0.060). On the other hand, the systolic and diastolic blood pressures were also lower as a consequence of the donor FMT treatment and reduced by 10% (p = 0.028) and 19% (p = 0.0008), respectively (Table 2). The concentration of faecal SCFAs such as faecal butyrate (p = 0.018) and acetate (p = 0.033), were decreased after the FMT but not that of propionate (p = 0.280). Regarding markers of lipid metabolism, the HDL cholesterol levels in plasma were reduced (p = 0.032). By computing Kendall's τ (tau) parameter, we established associations between these clinical variables and the abundance of microbial species altered as a consequence of the FMT. Species such as *R. bicirculans*, which were reduced after FMT, correlated positively with the reduction of Hba1c, HOMA-IR, and plasma insulin (τ = 0.56, 0.49, 0.39 and FDR = 0.017, 0.028, 0.081, respectively), whereas blood pressure parameters correlated negatively with abundance of *B. coccoides* (τ = -0.55, -0.43 and FDR = 0.009, 0.049, respectively for Diastolic and Systolic blood pressure), and to a lesser extent with *P. merdae* abundance (τ = -0.42, FDR = 0.091 for Diastolic blood pressure). The above correlations were not detected for other closely related species (e.g. *Ruminococcus albus*, *Ruminococcus gnavus*, *Blautia* *luti*, *Blautia wexlerae*, *Parabacteroides johnsonii*, *Parabacteroides distasonis*, or *Parabacteroides goldsteinii*).

Discussion

The emergence of single-molecule and synthesis-free based sequencing methods and its portable devices has democratised the genomics, making itself a disruptive technology with application across multiple life-science and clinical disciplines. In general, the central claim of the third-generation sequencing platforms, despite their higher error-rates, is the ability to produce very long DNA reads with a handy application to resolve eukaryote genomes and their repetitive structures [19-24]. This strength has also been advantageous to better assess the composition of complex microbial communities, making it possible to expand the genetic information classically used in microbiota surveys and retrieving reliable taxonomy identifications at the species-level [1, 25].

Here we explored the *rrn* of the human gut microbiota through the nanopore sequencing and *de novo* assembly of this multi-locus hypervariable region to gain insights into these genetic markers and their potential use to deeply characterize complex communities at species and strain level. Through the pipeline for the *rrn* assembly, we realised that the indels rate during the strand-wise alignments was always higher than mismatches, likely an effect of the alignment parameters (see methods). Notwithstanding, close inspections of the alignments suggest that such indels are produced in homopolymeric regions (containing single or di-nucleotide repeats), a common failure during basecalling of nanopore data [21, 26]. On the other hand, the final assemblies were recovered with less than 1% indels proportions when compared to database references, thus indicating they were drastically attenuated when compared to the strand-wise evaluation, and suggesting that assemblies have largely been improved to correct typical errors of nanopore data, and likely retaining the nucleotide changes (mismatches) linked to species/strains genetic variability.

In our study, identifications of assembled *rrn* were based mostly on sequence identities higher than 99%, achieving a similar accuracy as in previous studies based on *de novo* genome assembly [20, 22]. Furthermore, we have been able to detect the variability at strain level because the *rrn* sequences retrieved 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 query, respectively. Nevertheless, the possibility that some of the species identified could be novel ones cannot be disregarded despite the vast amount of genetic and taxonomy information of microbes inhabiting the gut environment compiled during last years [27-29]. Interestingly, we have also obtained correct species identification through *rrn* sequencing supported by very low coverage assemblies, at least ten reads. Since the R7.3 was initially released, the constantly improved chemistry on nanopore devices enables better assemblies with lower coverages [21, 30]. Therefore, the improved chemistry releases (e.g. R10 and later) are also expected to influence the quality of assemblies and to increase drastically the sensitivity of this approach to detect and measure reliably and rapidly the presence of more microbial species/strains in the gut microbiota.

We provided *de novo* reliable assemblies for more than two-hundred *rrn* regions of human gut microbes. The taxonomic identification of such assemblies indicated that the best results were those obtained with the 16S rRNA NCBI database (different releases used during assembly and annotation) probably because this is the most studied and used marker for bacteria taxonomy. Furthermore, the high amount of high sequence identity matches (>95%) retrieved with "unculturable bacteria" when *rrn* assemblies were compared against the NCBI non-redundant nucleotide collection, highlights the high level of uncertainty of taxonomic assignations done based on metagenome assembled genomes (MAGs) released so far in major public repositories. The assembled *rrn* regions could represent the basis for taxonomy identification of these unclassified entries given the distribution of mismatches was close to that observed for positively identified *rrn* assemblies. Additionally, future *rrn* comparative assessments on large collection of MAGs [27] could help to solve taxonomy issues on gene catalogues 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 absent in the first database release [6].

By using approximately a dataset consisting of 400K nanopore reads derived from 24 samples, the updated *rrn* database, and controlling the covariates influencing the microbiota, we proved the validity of the methodology to assess changes in the human gut microbiota at species and strain level as a consequence of a FMT intervention. This long amplicon-based approach, enable us to detect the increase of several species from the *Parabacteroides*, *Butyricimonas*, *Ruminococcus*, and *Sutterella* species as a result of FMT. Additionally to age, sex, sequencing run, and baseline BMI, we found that the donor is a critical covariate of the impact of the FMT in the recipient microbiota when using a unique donor for multiple recipients. These results are partly in agreement with those previously published using short-reads from V4 hypervariable regions of the 16S rRNA bacterial gene [11], however, our results outperform the taxonomy resolution reached previously, thus providing a more accurate gut microbiota survey. Interestingly, a recent FMT study to treat ulcerative colitis (UC) where the microbiota was analysed sequencing short-reads also showed increases of *Sutterella* species as a consequence of the intervention [31]. Similarly, the abundance of *Butyricimonas* species was increased as a consequence of an FMT intervention to eradicate antibiotic-resistant bacteria [32]. These findings suggest that some species could be often shifted as a result of the FMT in humans, regardless the condition of the recipient.

Nonetheless, the presumable transference of species between donor and recipients and their engraftment could not be confirmed based on the short-read amplicon technology due to the lack of sufficient resolution. Also the potential replacement between closely related species in the recipient could be overseen in conventional microbiota surveys based on short-reads. To shed light on the ability of the *rrn* sequencing approach to assess SNVs likely associated with the strain diversity, we selected *P. merdae*, a species that exhibited a remarkable increase in the recipients' gut after FMT, and *F. prausnitzii*, which seemed to be not affected by the FMT intervention. The combined information of three different informative SNVs from the *P. merdae rrn* demonstrated that the increase of this species in recipients after FMT was more than likely because of the strain transference from donors.

The taxonomic resolution achieved with our sequencing approach also would help to support more firmly a causal relationship between the changes in the gut microbiota and the improvements in the recipients' metabolic markers and blood pressure since robust evidence of the transference of bacterial strains from the donor to the recipient that correlated to the improved clinical variables could be provided. Notwithstanding, the direct implication of particular species such as *B. coccoides* and *P. merdae* and their strains in the improvement of cardio-metabolic health markers would need to be further explored in future (single strain) intervention studies. The performance of nanopore-generated data for identification of single-nucleotide polymorphisms (SNP) on eukaryote and prokaryote organisms has been previously reported [33, 34]. Altogether, those results suggest that despite the error-prone nature of these data, well-processed nanopore reads have enough quality to estimate SNVs, additionally to its recognised utility for chromosome assemblies.

The capacity of the methodology described in here to unveil SNVs will be pivotal in the near future to establish reliable genotype-to-phenotype associations between human diseases and microbiome at the strain-level. Additionally to its cost-effectivity, the data derived from our method could be analysed in a reference-based (read mapping against the *rrn* database) or reference-independent manner (read assembly into discrete similarity clusters), making this approach versatile either for strain surveillance and discovery. All in all, the features mentioned above should be central to define subtle genetic variation in the human microbiome and profiling such variants as harmful or beneficial for human health, what is part of an envisioned field of research in the frame of the epidemiology of microbial communities and the human microbiome [35].

Conclusions

The updated version of the *rrn* database will be useful to do reliable microbial surveys at the species level and, potentially, to infer strain variations taking into account the most abundant members of the human intestinal microbiota. This long-read approach allows the detection of species- and strain-level changes in the microbiota at lower cost compared to the expensive shotgun-DNA-sequencing-based metagenomics approach that up to date is the only one with the ability to provide such level of information. Thus, the affordability of this methodology will help to improve microbiota surveys aiming to discriminate the human gut microbial species associated with health and disease. This methodology has been proven to perform well for the identification of species and strains transferred and engrafted in the gut microbiota of the new recipients receiving FMT, as a proof of concept. Considering our promising results, future studies should be conducted to expand the knowledge of *rrn* diversity across this and other environments, using improved releases of the nanopore chemistry, and to provide more robust tools for microbiome research progressing towards their standardization.

List of abbreviations

BMI, body mass index; FDR, flase discovery rate; FMT, faecal 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), 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) [11]. Registered on August 1st, 2014. First participant was enrolled on September 1st, 2014.

Availability of data and material

The albacore-basecalled *fast5* files obtained from respective runs are publicly available in the European Nucleotide Archive upon 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 conflict of interest to declare.

Funding

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

Authors'contributions

ABP conceived and designed the study. ABP performed sequencing experimental research and data analysis, AVH and MN performed 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.

References

1. Benitez-Paez A, Portune KJ, Sanz Y: **Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer.** *Gigascience* 2016, **5:**4.

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

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

4. Shin H, Lee E, Shin J, Ko SR, Oh HS, Ahn CY, Oh HM, Cho BK, Cho S: **Elucidation of the bacterial communities associated with the harmful microalgae Alexandrium tamarense and Cochlodinium polykrikoides using nanopore sequencing.** *Sci Rep* 2018, **8:**5323.

5. Shin J, Lee S, Go MJ, Lee SY, Kim SC, Lee CH, Cho BK: **Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing.** *Sci Rep* 2016, **6:**29681.

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

7. Kerkhof LJ, Dillon KP, Haggblom MM, McGuinness LR: **Profiling bacterial communities by MinION sequencing of ribosomal operons.** *Microbiome* 2017, **5:**116.

8. Peker N, Garcia-Croes S, Dijkhuizen B, Wiersma HH, van Zanten E, Wisselink G, Friedrich AW, Kooistra-Smid M, Sinha B, Rossen JWA, Couto N: **A Comparison of Three Different Bioinformatics Analyses of the 16S-23S rRNA Encoding Region for Bacterial Identification.** *Front Microbiol* 2019, **10:**620.

9. Brewer TE, Albertsen M, Edwards A, Kirkegaard RH, Rocha EPC: **Unlinked rRNA genes are widespread among Bacteria and Archaea.** *bioRxiv* 2019**:**705046.

10. Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, Shoobridge JD, Graham N, Patel NH, Gillespie RG, Prost S: **Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale.** *Gigascience* 2019, **8**.

11. Hartstra AV, Schüppel V, Imangaliyev S, Schrantee A, Prodan A, Collard D, Levin E, Dallinga-Thie G, Ackermans MT, Winkelmeijer M, et al: **Infusion of donor feces affects the gut-brain axis in humans with metabolic syndrome.** *Mol Metab* 2020**:**in press.

12. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC: **Adaptive seeds tame genomic sequence comparison.** *Genome Res* 2011, **21:**487-493.

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

14. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS Comput Biol* 2011, **7:**e1002195.

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

16. Pruesse E, Peplies J, Glockner FO: **SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes.** *Bioinformatics* 2012, **28:**1823-1829.

17. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, et al: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods* 2010, **7:**335-336.

18. Naum M, Brown EW, Mason-Gamer RJ: **Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the enterobacteriaceae?** *J Mol Evol* 2008, **66:**630-642.

19. Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, Genete M, Berrabah W, Chevre AM, Delourme R, et al: **Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps.** *Nat Plants* 2018, **4:**879-887.

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

21. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al: **Nanopore sequencing and assembly of a human genome with ultra-long reads.** *Nat Biotechnol* 2018, **36:**338-345.

22. Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP: **MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome.** *Genome Res* 2018, **28:**266-274.

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

24. Masonbrink R, Maier TR, Muppirala U, Seetharam AS, Lord E, Juvale PS, Schmutz J, Johnson NT, Korkin D, Mitchum MG, et al: **The genome of the soybean cyst nematode (Heterodera glycines) reveals complex patterns of duplications involved in the evolution of parasitism genes.** *BMC Genomics* 2019, **20:**119.

25. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK: **Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system.** *PeerJ* 2016, **4:**e1869.

26. Zascavage RR, Thorson K, Planz JV: **Nanopore sequencing: An enrichment-free alternative to mitochondrial DNA sequencing.** *Electrophoresis* 2019, **40:**272-280.

27. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, et al: **A unified catalog of 204,938 reference genomes from the human gut microbiome.** *Nat Biotechnol* 2020.

28. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, et al: **An integrated catalog of reference genes in the human gut microbiome.** *Nat Biotechnol* 2014, **32:**834-841.

29. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, et al: **Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000**

Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 2019, **176:**649-662 e620.

30. Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS: **Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing.** *G3 (Bethesda)* 2018, **8:**3143-3154.

31. Paramsothy S, Nielsen S, Kamm MA, Deshpande NP, Faith JJ, Clemente JC, Paramsothy R, Walsh AJ, van den Bogaerde J, Samuel D, et al: **Specific Bacteria and Metabolites Associated With Response to Fecal Microbiota Transplantation in Patients With Ulcerative Colitis.** *Gastroenterology* 2019, **156:**1440-1454 e1442.

32. Bilinski J, Grzesiowski P, Sorensen N, Madry K, Muszynski J, Robak K, Wroblewska M, Dzieciatkowski T, Dulny G, Dwilewicz-Trojaczek J, et al: **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* 2017, **65:**364-370.

33. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ, Coin LM: **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* 2018, **4**.

34. Malmberg MM, Spangenberg GC, Daetwyler HD, Cogan NOI: **Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (Brassica napus L.).** *Sci Rep* 2019, **9:**8688.

35. Yan Y, Nguyen LH, Franzosa EA, Huttenhower C: **Strain-level epidemiology of microbial communities and the human microbiome.** *Genome Med* 2020, **12:**71.

Tables

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

Genus	Species	Abundance	Abundance	Prevalence	Prevalence	*p*-value
		PRE-FMT	POST-FMT	PRE-FMT	POST-FMT	
		(N = 10)[1]	(N = 10)[1]	(N = 10)	(N = 10)	
Parabacteroides	*P. merdae*	3.44 ± 0.61	4.36 ± 0.26	100%	100%	0.002
	P. johnsonii	1.46 ± 1.58	2.97 ± 0.48	50%	100%	0.005
	P. goldsteinii	1.60 ± 1.75	2.65 ± 1.08	50%	90%	0.032
Butyricimonas	*B. paravirosa*	1.18 ± 1.27	2.44 ± 0.89	50%	90%	0.031
	B. virosa	2.44 ± 1.39	3.33 ± 0.50	80%	100%	0.033
Ruminococcus	***R. bicirculans***[2]	**1.84 ± 1.66**	**0.58 ± 1.34**	**60%**	**20%**	0.017
	R. callidus	0.41 ± 0.86	1.64 ± 1.51	20%	60%	0.010
Sutterella	*S. massiliensis*	3.05 ± 1.39	4.17 ± 0.69	90%	100%	0.021
	S. wadsworthensis	2.27 ± 1.73	3.44 ± 0.86	70%	100%	0.019
Bacteroides	*B. finegoldii*	0.83 ± 1.35	2.17 ± 1.23	30%	80%	0.046
Bifidobacterium	***B. adolescentis***[2]	**2.32 ± 1.60**	**1.95 ± 1.40**	**70%**	**70%**	0.024
Blautia	*B. coccoides*	1.33 ± 1.45	2.44 ± 0.89	50%	90%	0.034
Coprococcus	*C. eutactus*	2.08 ± 1.58	2.85 ± 1.62	70%	80%	0.045
Desulfovibrio	*D. piger*	1.76 ± 1.95	2.90 ± 1.29	50%	90%	0.035
Paraprevotella	*P. clara*	2.06 ± 1.52	3.66 ± 0.46	70%	100%	0.011
Prevotella	*P. bivia*	0.63 ± 1.03	1.99 ± 1.44	30%	70%	0.038
Terrisporobacter	***T. mayombei***[2]	**1.96 ± 1.72**	**1.24 ± 1.36**	**60%**	**50%**	0.029

[1] Data expressed as the mean of the number of normalised reads in log10 scale ± standard deviation (sd).

[2] Results underlined are those of bacterial species decreasing after FMT.

Table 2. Clinical variables altered after the 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
Faecal butyrate (µmol/g)	88.6 ± 44.2	60.3 ± 40.9	*v* = -31.6, *p* = 0.018
Faecal 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 analysed applying a LMM (PRE-FMT group as reference), HDL = high density lipoprotein, Hba1c = glycosylated haemoglobin.

Figure legends

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

Figure 2. Comparative analysis of the stranded-based alignment of nanopore reads to respective baits before assembly. In each case, the alignments for the top 20 most abundant species were evaluated in terms of the indels and mismatch content across the full set of reads mapped. The species occurrence of the individual forward and reverse read assessments is linked by dashed lines. These density ridgeline plots were designed with the *ggridges* R package.

Figure 3. Taxonomic identification of final *rrn* assemblies. A - Scatter plot showing information for the sequence identity supporting the identification of assembled *rrn* against the NCBI 16S database, and the coverage (number of reads) accounted for the respective assemblies. Additional levels of information are included in the plot such as taxonomy level match, bias coverage between forward and reverse reads, and alignment size (see graph symbols and their colour and size scale). B - The indel and mismatch content evaluation in the alignments resulting from cross-identification of *rrn* assemblies against the NCBI 16S database and the non-redundant nucleotide collection (GenBank). Those values were discriminated by the taxonomic level of identification according to the original species-level annotation of respective *rrn*s (see the colour legend). UB, uncultured bacteria.

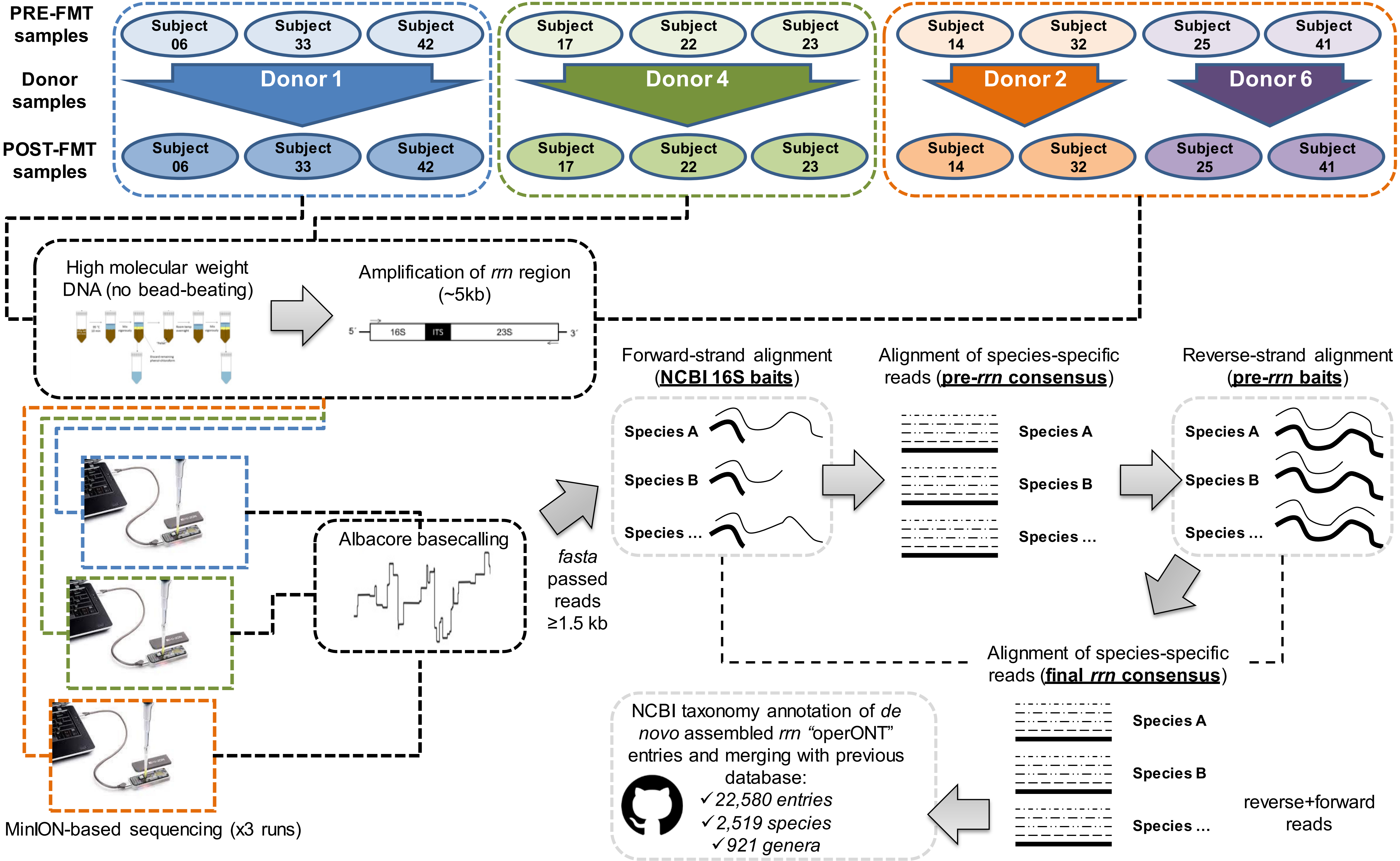
Figure 4. Beta diversity of the microbial communities assessed by *rrn* sequencing. A - Scatter plot compiling data from the multivariate analysis (principal coordinate analysis - PCoA) of the microbiota from recipients and donors involved in the FMT. The donor and recipient samples and the sampling time points, are defined according to the legend on top. PCo; principal coordinate (the two most informative are shown). B - A genetic distance-based approach to evaluate microbiota transference between donors and recipients pairs. The microbial community structures of PRE-FMT and POST-FMT samples were compared with donors, respectively, through the calculation of the Bray-Curtis dissimilarity index and represented as boxplots. The Wilcoxon signed-rank test for paired samples was used to compare differences in genetic distances.

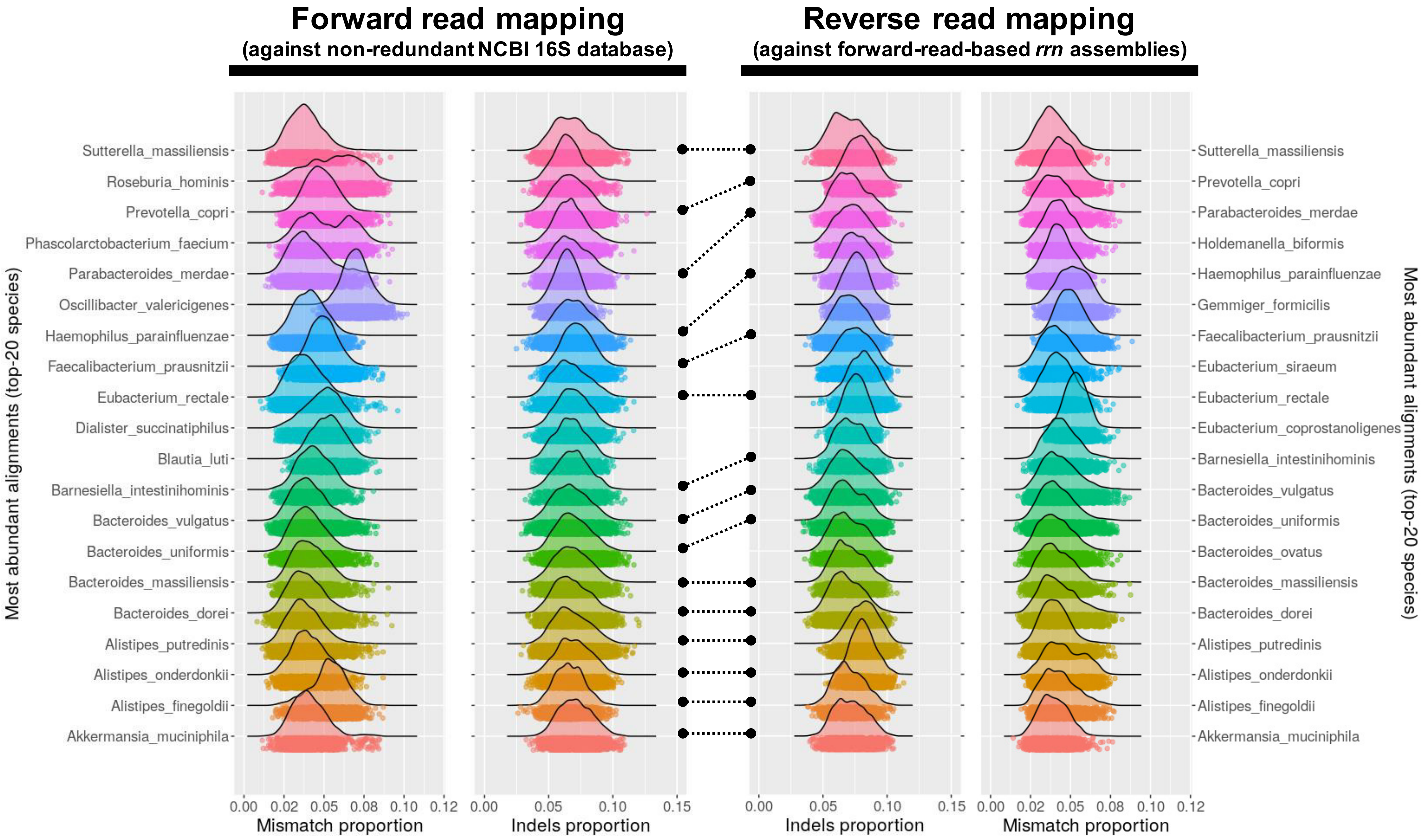
Figure 5. Single nucleotide variation (SNV) analysis to detect species transference and engraftment. The species *Parabacteroides merdae* and *Faecalibacterium prausnitzii rrn*s were used to unveil SNVs within each species and their abundance as a consequence of the FMT (A and B panels, respectively). C and D - The mapping, indexing, and pileup of thousands of reads on *P. merdae* and *F. prausnitzii rrn* enable us to uncover polymorphic sites exhibiting an even frequency for at least two nucleotides. Those potential SNVs are highlighted inside open circles, and the position number in the *rrn* is also indicated. The structural arrangement of the *rrn* is drawn on the x-axis. The black lines indicate the relative frequency of the dominant allele, whereas the grey 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 - Electropherograms obtained from Sanger sequencing for the strains of *P. merdae* SNV-743, SNV-1975, and SNV-3016. The predominant alleles inferred for every sample (based on the Q-score of basecalling) supported the hypothesis of that the strains were transferred between donors and recipients pairs, as anticipated during the nanopore-based assessment.

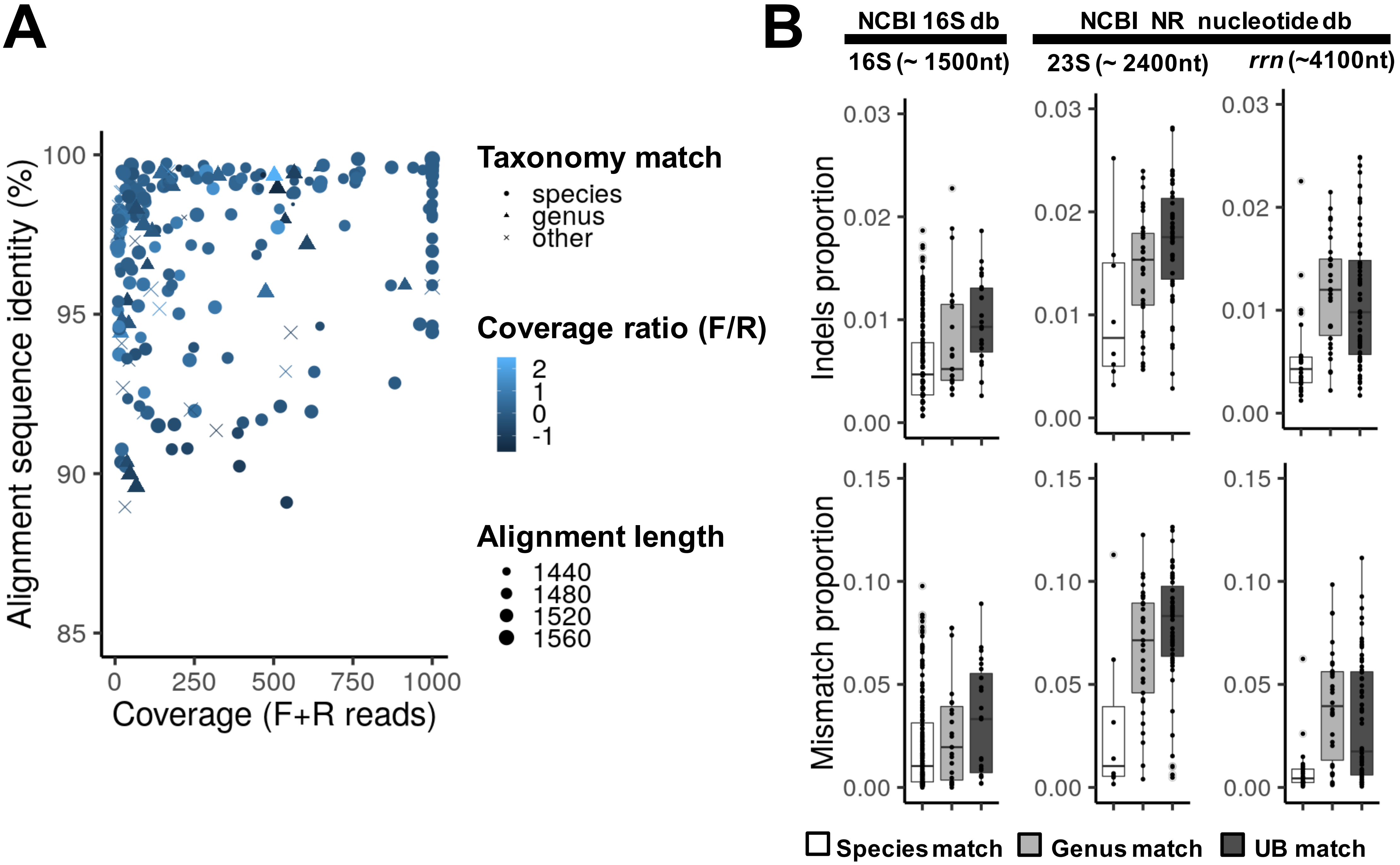
Figure S1. Comparative analysis of the alignments based on forward and reverse reads. Sequence identity, mismatch and indels proportions, are represented as histograms. Sequencing runs discriminate the distributions (see colour legend). Vertical dotted lines of the sequence identity plots indicate the threshold for selection of high-quality reads for downstream analyses.

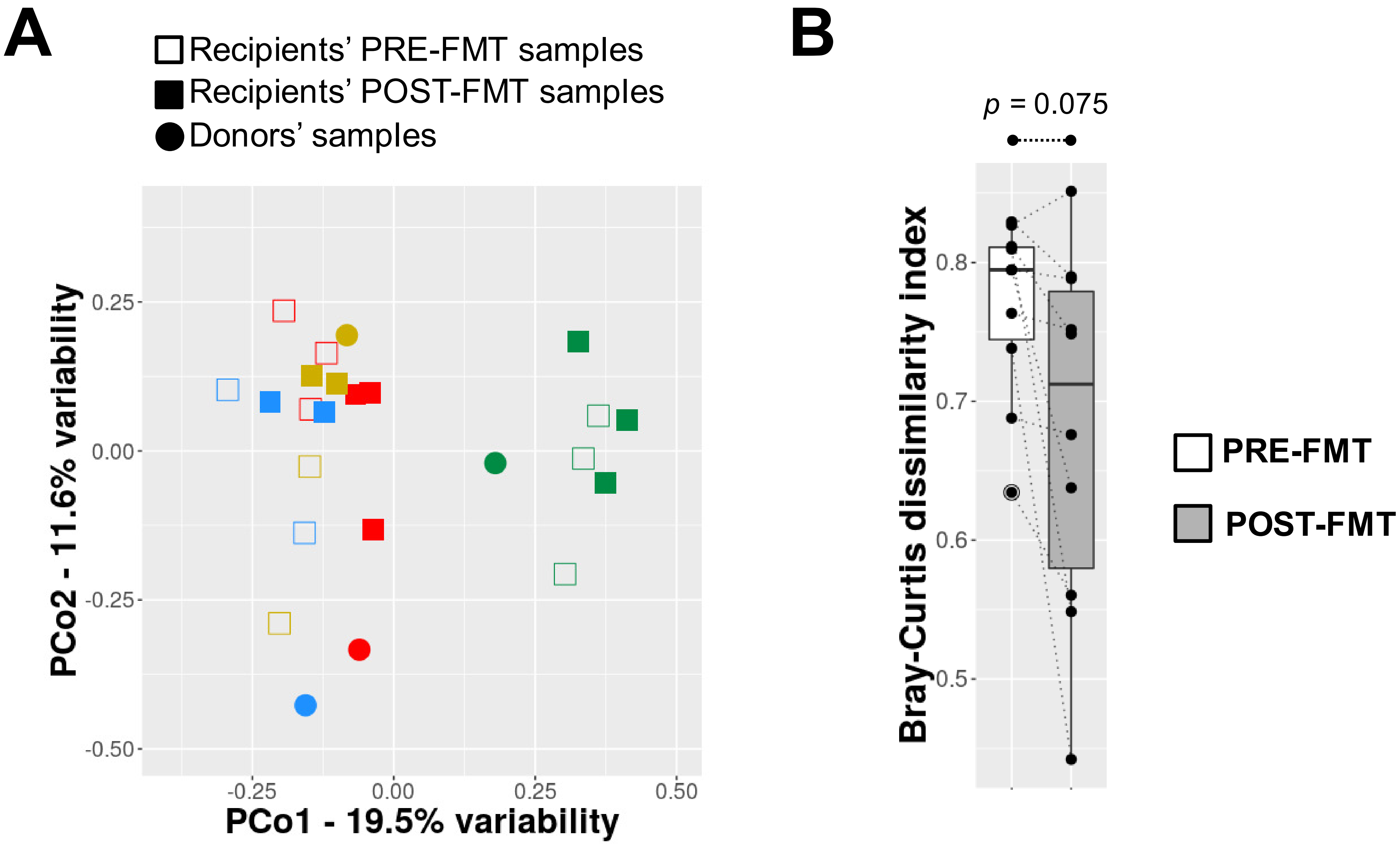
Figure S2. Blast-based results showing correct identification of assembled *rrn*s. The top hits supporting the identification are shown for six different *rrn*s assembled with low, mid or high coverage (number of reads).

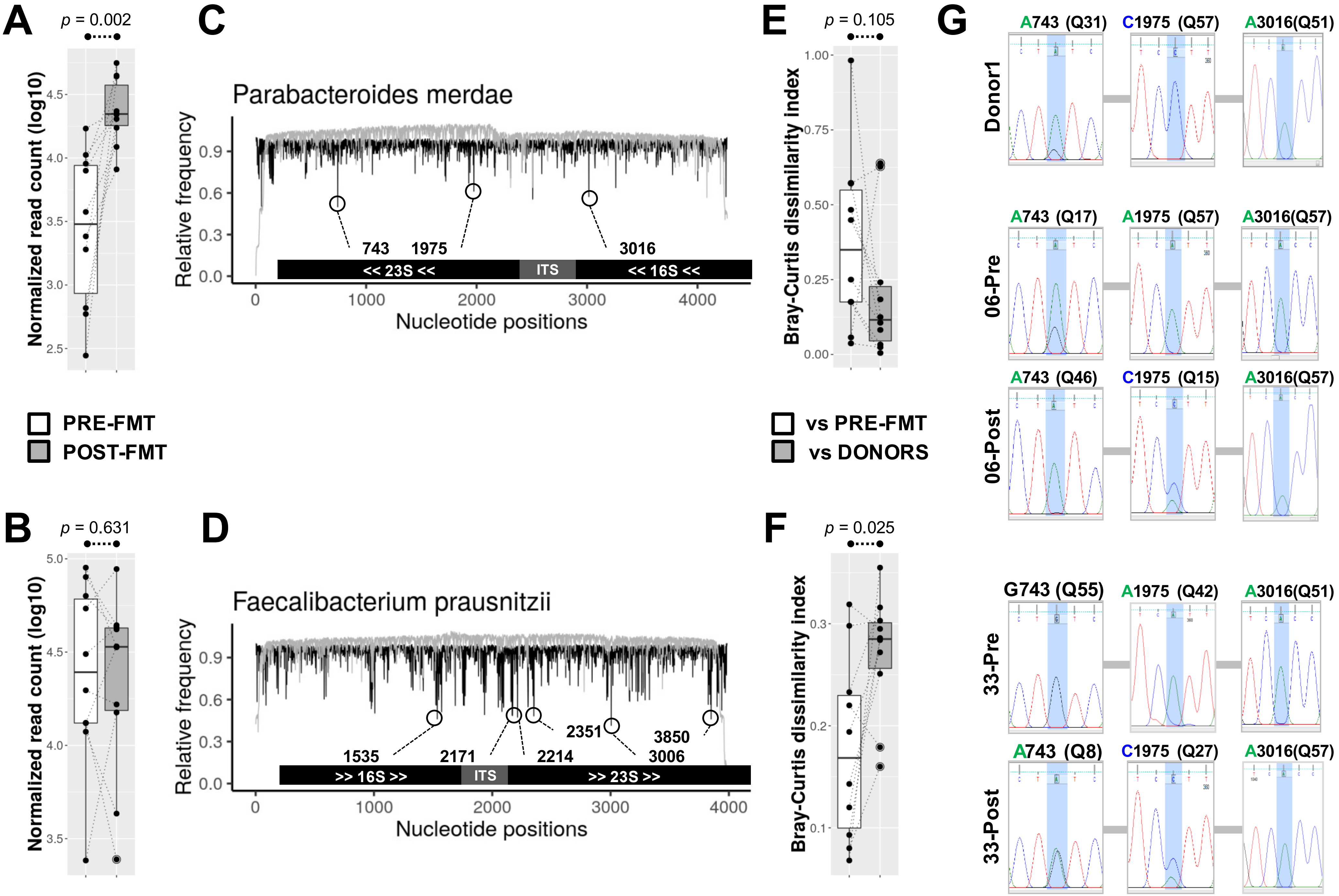
Figure S3. Alpha diversity of the recipients' microbiota s before and after FMT. The distribution of values obtained for four alpha diversity indicators, including the Chao's index, Shannon's index, reciprocal Simpson's index, and dominance index are shown as boxplots. The results of the Wilcoxon signed-rank test applied to establish differences between the two groups of samples (POST-FMT and PRE-FMT) is also shown.


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


Forward read mapping
(against non-redundant NCBI 16S database)
Reverse read mapping
(against forward-read-based *rrn* assemblies)
Most abundant alignments (top-20 species)
Sutterella_massiliensis
Roseburia_hominis
Prevotella_copri
Phascolarctobacterium_faecium
Parabacteroides_merdae
Oscillibacter_valericigenes
Haemophilus_parainfluenzae
Faecalibacterium_prausnitzii
Eubacterium_rectale
Dialister_succinatiphilus
Blautia_luti
Barnesiella_intestinihominis
Bacteroides_vulgatus
Bacteroides_uniformis
Bacteroides_massiliensis
Bacteroides_dorei
Alistipes_putredinis
Alistipes_onderdonkii
Alistipes_finegoldii
Akkermansia_muciniphila
0.00
0.02
0.05
0.08
0.10
0.12
Mismatch proportion
0.00
0.05
0.10
0.15
Indels proportion
0.00
0.05
0.10
0.15
Indels proportion
0.00
0.02
0.05
0.08
0.10
0.12
Mismatch proportion
Sutterella_massiliensis
Prevotella_copri
Parabacteroides_merdae
Holdemanella_biformis
Haemophilus_parainfluenzae
Gemmiger_formicilis
Faecalibacterium_prausnitzii
Eubacterium_siraeum
Eubacterium_rectale
Eubacterium_coprostanoligenes
Barnesiella_intestinihominis
Bacteroides_vulgatus
Bacteroides_uniformis
Bacteroides_ovatus
Bacteroides_massiliensis
Bacteroides_dorei
Alistipes_putredinis
Alistipes_onderdonkii
Alistipes_finegoldii
Akkermansia_muciniphila
Most abundant alignments (top-20 species)


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
□ Recipients' PRE-FMT samples
■ Recipients' POST-FMT samples
● Donors' samples
0.25
0.00
-0.25
-0.50
-0.25
0.00
0.25
0.50
PCo1 - 19.5% variability
PCo2 - 11.6% variability
B
p = 0.075
0.8
0.7
0.6
0.5
Bray-Curtis dissimilarity index
□ PRE-FMT
▨ POST-FMT


A
p = 0.002
Normalized read count (log10)
4.5
4.0
3.5
3.0
2.5
PRE-FMT
POST-FMT
B
p = 0.631
Normalized read count (log10)
5.0
4.5
4.0
3.5
C
Parabacteroides merdae
Relative frequency
0.9
0.6
0.3
0.0
743
1975
3016
<< 23S <<
ITS
<< 16S <<
0
1000
2000
3000
4000
Nucleotide positions
D
Faecalibacterium prausnitzii
Relative frequency
0.9
0.6
0.3
0.0
1535
2171
2214
2351
3850
3006
>> 16S >>
ITS
>> 23S >>
0
1000
2000
3000
4000
Nucleotide positions
E
p = 0.105
Bray-Curtis dissimilarity index
1.00
0.75
0.50
0.25
0.00
vs PRE-FMT
vs DONORS
F
p = 0.025
Bray-Curtis dissimilarity index
0.3
0.2
0.1
G
Donor1
A743 (Q31)
C1975 (Q57)
A3016(Q51)
06-Pre
A743 (Q17)
A1975 (Q57)
A3016(Q57)
06-Post
A743 (Q46)
C1975 (Q15)
A3016(Q57)
33-Pre
G743 (Q55)
A1975 (Q42)
A3016(Q51)
33-Post
A743 (Q8)
C1975 (Q27)
A3016(Q57)