End-to-end assessment of fecal bacteriome analysis: from sample processing to DNA sequencing and bioinformatics results

Ana Paula Christoff[1†], Giuliano Netto Flores Cruz[1†], Aline Fernanda Rodrigues Sereia[1], Laís EikoYamanaka[1], Paola Paz Silveira[1] Luiz Felipe Valter Oliveira[1*]

[1]BiomeHub, Florianopolis, Brazil

[†]These authors have contributed equally to this work

*** Correspondence:**

Luiz Felipe Valter de Oliveira

felipe@biome-hub.com

Keywords: gut microbiome, intestinal, Brazil, DNA sequencing, NGS, bacteria.

Abstract

Intestinal microbiome, comprising the whole microbiota, their genes and genomes living in the human gut have significant roles in promoting health or disease status. As many studies showed so far, identifying the bacterial components of the microbiome can reveal important biomarkers to help in the disease comprehension to a further adequate treatment. However, the human nature is quite variable considering the genetic components associated with life styles, directly reflecting on the gut microbiome. Thus, it is extremely important to know the populational microbiome background in order to draw conclusions regarding the health and disease conditions. Also, methodological best practices and knowledge about the methods being used are essential for the results quality and applicability with clinical relevance. In this way, we standardized the sample collection and processing methods used for the Probiome assay, a test developed to identify the Brazilian bacteriome from stool samples. EncodeTools Metabarcode pipeline of analysis was developed to obtain the best result from the samples. This pipeline uses the information of amplicon single variants (ASVs) in 100% identical oligotype clusters, and performs a *de novo* taxonomical assignment based on similarity for unknown sequences. To better comprehend the results obtained in Probiome assays, is essential to know the intestinal bacteriome diversity of Brazilians. Thus, we applied the standardized methods herein developed and began characterizing our populational data to allow a better understanding of the Brazilian bacteriome profiles and how they can be related to other microbiome studies.

Introduction

The human body hosts a diverse microbial community composed by bacteria, fungi, virus, and small eukaryotes that along with their genes and genomes comprise the human microbiome. All this microbial living in our bodies, mainly in our intestine, serves as a source of genetic and metabolic diversity. Most of our gut microbiota is composed of bacteria [1,2] and their diversity influence the human health by playing a role in the digestive, neurological, or immunological systems disorders [3–5]. Two larger projects made significant contributions in the understanding of the healthy microbiota and their host, the Metagenomics of the Human Intestinal Tract (MetaHIT) [1] and the Human Microbiome Project (HMP) [6,7] . More recently, the American Gut project also contributed to the knowledge of intestinal microbiome profiles from populations in the United States, United Kingdom and Australia [8]. These microbiome projects, along with several others conducted around the world, have the primary goal of understanding the dynamics and variations in the human intestinal microbiome to characterize it regarding health and disease conditions.

The intestinal microbiome varies widely among individuals, also fluctuating over human development and time. These variations increase the complexity of the human microbiome comprehension, becoming more challenging to define what is a healthy status for a population and an individual [9]. Additionally, each population has its particularities regarding their genetic background, physiology, lifestyle, nutrition, and habits that can influence the microbiota [10,11]. A recent study published with Chinese populations revealed that geography has a substantial interference with microbiome profiles, hampering the universal application of microbiota-associated disease models that were developed based on specific populations [12]. Thus, it is extremely relevant to have microbiome information about the specific target population to allow conclusions regarding their health and disease conditions.

All these research studies were fundamental to improve the knowledge regarding microbiome characterization along with the technical and biological challenges that must be addressed and controlled in the best possible ways [13]. The experimental reproducibility is critical, giving the potential of clinical application for the obtained results. Moreover, adequate sample collection and storage is a requirement for maintaining the original microbial composition, since the improper storage can allow selective microorganisms to overgrow leading to microbial profile biases and consequently misleading the interpretation of the results [14,15]. Several efforts have also been made to address variations and standardize DNA extraction, amplicon 16S rRNA gene sequencing, and bioinformatics analysis, as done by the Microbiome Quality Control (MBQC) project consortium [16]. Moreover, usage of amplicon sequence variants (ASVs), the exact DNA sequence read, instead of the

OTU picking (generally clustering sequences at 97% similarity) improves the resolution for microbiome results [13,17,18].

In this paper, we present an end-to-end assessment of a human intestinal bacteriome analysis for Brazilian populations, covering all the process from sample storage, amplicon library preparation, high-throughput DNA sequencing, and bioinformatics analysis. We introduced a new pipeline of analysis: EncodeTools Metabarcode, and generated 16S rRNA amplicon data for fecal samples of the Brazilian subjects to begin an understanding of the bacteriome compositional patterns in such a diverse population whose gut microbiome profiles are yet to be characterized.

Material and Methods

Sample collection and processing

Stool samples were collected using the Probiome kit (BiomeHub, Brazil) which includes a sanitary seat cover capable of retaining the stool and allows the proper sample collection with a sterile flocked swab - 520CS01 (Copan, USA) or 25-3606-H BT (Puritan, USA). The swabs have a breakpoint that allows the swab tip containing the collected sample to be inserted into a provided microtube with 1ml of fecal stabilization solution - ZSample (BiomeHub, Brazil). Each subject can take the entire kit home and perform the fecal sample collection individually. The samples were homogenized by microtube inversion and then forwarded to BiomeHub laboratory (Florianopolis, Brazil) for sample processing within 30 days after collection. In the laboratory, DNA was extracted from the preserved stool using the DNeasy PowerSoil Kit (QIAGEN, Germany) according to the manufacturer instructions. At each batch of DNA extraction, a negative control was included (CNE). A set of 206 stool samples that used the above collection and processing methods were randomly selected from the mischaracterized BiomeHub database. No possible correlations or associations with the fecal donors can be made from this bacterial sequences or any data included in this study. These samples, collected and anonymously processed as described above, along 2018, represent a Brazilian populational diverse subset comprising 65.4% female and 34.5% male from various geographical locations.

Experimental subsets for sample storage, ZSample stability and DNA extraction tests

ZSample stability solution and stool sample preservation at room temperature were evaluated along 30 days. A single stool specimen was self-collected by an anonymous donor in seventeen replicates and stored in ZSample Probiome tubes to be analyzed at T0 (maximum of two hours after

collection), T15 (15 days after sample collection) and T30 (30 days after collection). Five of the replicates were analyzed in T0, and six replicates were analyzed at each T15 and T30.

Additionally, batch effects for the ZSample lot production in stool sample preservation was evaluated along four batches of the solution produced at 0, 2, 9 and 18 months before the stool sample collection. Twenty-four replicates of a fecal sample from an anonymous donor were collected using the four solution lots listed above. For each lot, six replicates were obtained, three of them were processed in T0 (maximum of two hours after collection) and the other three in T30 (30 days after collection). All samples remained at room temperature in ZSample solution during the 30-day storage. Furthermore, these fecal samples collected and stored in ZSample were inoculated in a general culture media (PCA - plate count agar) and incubated at 35ºC for three days, to evaluate cellular bacterial viability.

DNA extraction of fecal samples stored in ZSample was further tested in four different methods: DNeasy PowerSoil kit, DNeasy PowerSoil Pro kit, DNeasy PowerSoil Pro modified and QIAamp PowerFecal DNA kit, all from QIAGEN, Germany. In DNeasy PowerSoil Pro modified its original bead beating tubes with zirconium beads were replaced for the traditional PowerSoil silica bead tubes. Fecal samples were donated by five anonymous subjects, and processed with four experimental replicates for each extraction kit, in a total of 80 samples.

DNA library preparation and sequencing

The *16S rRNA* amplicon sequencing libraries were prepared using the V3/V4 primers (341F CCTACGGGRSGCAGCAG and 806R GGACTACHVGGGTWTCTAAT) [19,20] in a two-step PCR protocol. The first PCR was performed with V3/V4 universal primers containing a partial Illumina adaptor, based on TruSeq structure adapter (Illumina, USA) that allows a second PCR with the indexing sequences similar to procedures described previously [21]. Here, we add unique dual-indexes per sample in the second PCR. Two microliters of individual stool sample DNA were used as input in each first PCR reaction. The PCR reactions were carried out using Platinum Taq (Invitrogen, USA) with the conditions: 95°C for 5 min, 25 cycles of 95°C for 45s, 55°C for 30s and 72°C for 45s and a final extension of 72°C for 2 min for PCR 1. In PCR 2 the conditions were 95°C for 5 min, 10 cycles of 95°C for 45s, 66°C for 30s and 72°C for 45s and a final extension of 72°C for 2 min. All PCR reactions were performed in triplicates. The final PCR reactions were cleaned up using AMPureXP beads (Beckman Coulter, USA) and samples were pooled in the sequencing libraries for quantification. At each batch of PCR, a negative reaction control was included (CNR). The DNA concentration of the libraries was estimated with Picogreen dsDNA assays (Invitrogen, USA), and then the pooled libraries were diluted for accurate qPCR quantification using KAPA Library

136 Quantification Kit for Illumina platforms (KAPA Biosystems, MA). The libraries pools were adjusted
137 to a final concentration of 11.5 pM (for V2 kits) or 18 pM (for V3 kits) and sequenced in a MiSeq
138 system (Illumina, USA), using the standard Illumina primers provided in the manufacturer kit. Single-
139 end 300 cycle runs were performed using V2x300, V2x300 Micro, V2x500 or V3x600 sequencing
140 kits (Illumina, USA), always generating 283bp size amplicons suitable for analysis. Coverage of
141 50,000 reads was set to each sample sequenced.
142
143 *Bioinformatics analysis - EncodeTools Metabarcode pipeline*
144 The sequenced reads obtained were processed using EncodeTools Metabarcode pipeline
145 (BiomeHub, Brazil) a bioinformatics pipeline developed *in-house* and described below. Illumina
146 FASTQ files were quality filtered and the primers were trimmed to yield a resulting read of 283bp.
147 Only one mismatch is allowed in the primer sequences and the whole read is discarded if this criterion
148 is not met. Sequenced reads smaller than expected or with remaining Illumina sequence adapter were
149 discarded. After this initial quality assessment, identical read sequences (100% identity) were grouped
150 into oligotypes and analyzed with Deblur package [22] to remove possible erroneous reads. After,
151 VSEARCH [23] was used to remove chimeric amplicon reads. The oligotype clusterization with 100%
152 identity provides a higher resolution for the amplicon sequencing variants (ASVs), also called sub-
153 OTUs (sOTUs) [13] - herein denoted as oligotypes. An additional filter was implemented to remove
154 oligotypes below the frequency cutoff of 0.2% in the final sample counts, *i.e.*, given a library size of
155 1,000 reads, oligotypes with less than two reads were filtered out. We also implemented a negative
156 control filter, as in each processing batch we have negative controls for the DNA extraction and PCR.
157 If any oligotypes were observed in the negative controls, they are checked against the samples and
158 automatically removed from the sample results if present. The remaining oligotypes in the samples
159 were used for taxonomic assignment with the BLAST tool [24] against a reference genome database.
160 This database was constructed with complete and draft bacterial genomes, focused on clinically
161 relevant bacteria, obtained from NCBI and *in-house* genome sequencings. It is composed of 11,750
162 sequences comprising 1,843 different bacterial taxonomies. Taxonomy was assigned to each oligotype
163 using a lowest common ancestor (LCA) algorithm. If more than one reference can be assigned to the
164 same oligotype with equivalent similarity and coverage metrics (*e.g.* two distinct species mapped to
165 oligotype "A" with 100% identity and 100% coverage), the EncodeTools Metabarcode Taxonomy
166 Assignment algorithm leads the taxonomy to the lowest level of possible unambiguous resolution
167 (genus, family, order, class, phylum or kingdom), according to the similarity thresholds previously
168 established previously [25]. The bacterial profile obtained at the end of the pipeline is shown in
169 taxonomy proportions for the analyzed sample.

170

171 *Experimental subsets for robustness, sensibility and specificity of the EncodeTools Metabarcode*
172 *pipeline*

173 EncodeTools Metabarcode pipeline was tested and calibrated using internal data generated on
174 diverse hospital microbiome DNA samples obtained and processed as previously described [26].
175 Eight different microbiome samples were evaluated (A-H). Seven of them (A-G) were diverse
176 environmental swab samples and one was an artificial microbial community - mock (sample-H) -
177 composed of: *Acinetobacter baumanii, Bacillus subtilis, Enterococcus faecalis, Escherichia coli,*
178 *Klebsiella pneumoniae, Listeria monocytogenes, Pseudomonas aeruginosa, Salmonella enterica* and
179 *Staphylococcus aureus.* The 16S rRNA amplicon library preparation for these eight different samples
180 (A-H) was processed as described above in a total of 28 replicates per sample. These libraries
181 replicates were prepared by three different operators in three separated MiSeq runs, totalizing 224
182 sample assays along with 22 negative controls. Eleven amplicon library replicates were prepared for
183 each of the eight samples by a single operator for an intra-run technical reproducibility test and
184 sequenced in a single V2x300 Illumina MiSeq run. Inter-run technical reproducibility test was done
185 re-sequencing these eleven replicates amplicon libraries in a V3x600 Illumina MiSeq run. All
186 sequencing runs were a single-end of 300 cycles. Then, two additional operators prepared the same
187 amplicon libraries for the eight samples, in triplicates, for inter-run repeatability and robustness. These
188 libraries were sequenced in two separated V2x300 Illumina MiSeq runs, one for each operator's
189 library. All data generated were compared and used to evaluate the reproducibility, repeatability,
190 sensibility and specificity for our amplicon library preparation along with DNA sequencing, and the
191 EncodeTools Metabarcode pipeline of analysis.

192

193 *Data comparison and diversity analysis*

194 The results from all samples were integrated into an oligotype table (analogous to OTU table),
195 whose rows are samples and columns are oligotypes. For each oligotype, taxonomic lineage was
196 computed. A typical data analysis input was comprised of oligotype, taxonomy, and metadata tables.
197 The raw sequences were used to construct phylogenetic trees using FastTree 2.1 [27] and these were
198 used to calculate weighted UniFrac [28] distances when suitable. Further analyzes were conducted
199 inside the R statistical software environment (R version 3.6.0), using the Phyloseq package [29].
200 DESeq2, EdgeR, and metagenomeSeq packages were used for differential abundance analyses [30–
201 32]. Nonparametric comparisons included *Kruskal-Wallis* and *Wilcoxon* tests as implemented in base
202 R and in coin R package, respectively [33]. Other R packages used in this study are listed in
203 Supplementary Table 1.

Alpha-diversity was computed using the plot_richness function from the Phyloseq R package with default parameters. Note that Phyloseq by default calculates the Simpson Diversity Index as 1 - D. Here, we transform the value back to $D = \sum_i^n p_i^2$ (p_i is the proportional abundance for the ith taxonomy). Beta-diversity used proportion-normalized abundances as noted by [34] and [35]. Bray-Curtis Dissimilarity and weighted UniFrac were both calculated using Phyloseq's distance function. Correlation coefficients between sample groups used mean taxonomy proportions within each group.

Differential abundance analysis was performed using four distinct methods, all of which using the above cited packages with default options unless stated otherwise: DESeq2 and EdgeR were used to fit Negative Binomial models with relative log expression scaling [30,31,35]; metagenomeSeq applied a zero-inflated log-normal model with cumulative-sum scaling [32]; finally, rarefaction (with Phyloseq) was also applied followed by exact *Wilcoxon-Mann-Whitney* test, as implemented in the Coin R package, as this is a very traditional method, even though it has been characterized by its lack of power [34,35]. Rather than accepting the significance calls from all methods or arbitrarily choosing one of them, here we considered as significantly differentiated those taxa that were detected by at least two distinct methods simultaneously. Effect sizes were reported as fold-changes in the log_2 scale (log_2 *FC*) for all but the *Wilcoxon-Mann-Whitney* method, whose effect size estimates were computed as $Z_{score}/\sqrt{N}$ for sample size *N*. P-value correction for multiple comparisons was performed using the Benjamini-Hochberg procedure.

Results

Stool sample storage for bacteriome analysis

To validate our ZSample storage solution concerning the bacterial composition maintenance in fecal samples, we analyzed replicated samples stored at T0, T15 and T30 days. After DNA extraction and amplicon sequencing, we evaluated the bacterial profile from these samples through diversity and correlation analysis. Alpha and beta diversities showed no significant differences for the bacterial genera detected across sample storage times (Figures 1A and 1B). Additionally, high correlations were observed among bacterial profiles from all time points (Pearson and spearman's > 0.92) (Figure 1C). Figures 1D and 1E show the bacterial relative abundance profiles across the replicated samples analyzed in T0, T15 and T30. Some variations could be observed; however, they were no more related to the storage time than with inter-replicates variation. The overall diversity and relative abundance of each bacterial genus detected remained equivalent in all the samples across the storage time. Data for correlations and bacterial abundance for other taxonomy levels (phylum, family

and species) can be seen in Supplementary Figure 1. Taken together, these results indicate that ZSample properly maintains the original bacterial profile in samples stored at room temperature for at least 30 days. Moreover, no bacterial cellular viability was detected in the sample cultivation tests that were performed aerobically to resemble more closely how the samples are stored and manipulated along with the processes.

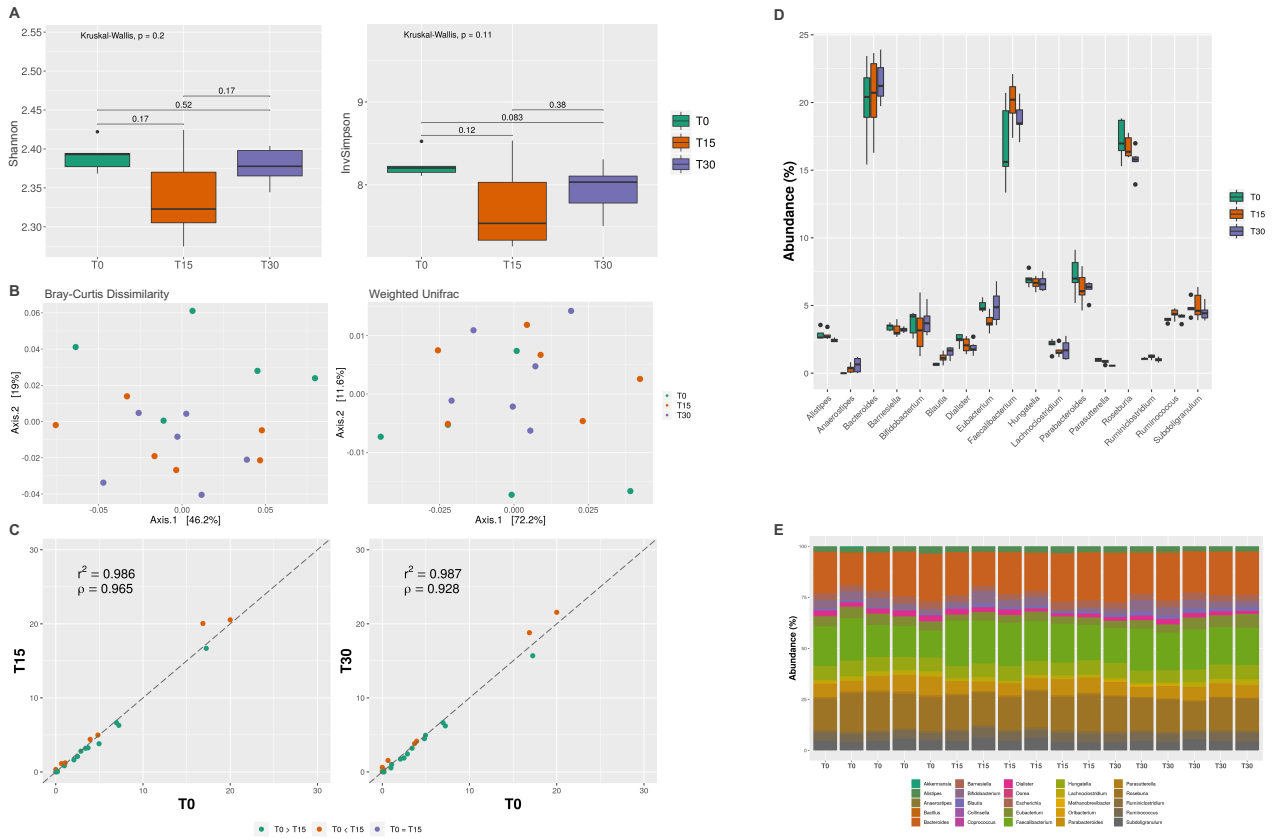


Figure 1. Fecal sample storage and bacterial profile along 30 days. Data presented in this figure is for bacterial genera analyzed after T0, T15 and T30 storage days in ZSample. **(A)** Shannon and InvSimpson alpha diversity analysis were performed with no significant differences among T0, T15 and T30 (*Kruskal-Wallis* p>0.05). Wilcoxon lacks of significance (p>0.05) is also showed above boxplots for pairwise comparisons between T0xT15, T15xT30 and T0xT30. **(B)** Beta diversities (Bray-curtis and Weighted UniFrac) didn't show any specific sample grouping or deviation related to the storage time. **(C)** A correlation analysis was performed between T0-T15 and T0-30 showing values > 0.92 for Pearson (r^2) and Spearman (ρ) coefficients. **(D)** Genera abundances along the sample storage demonstrate some inter-replicate variations higher than the storage time variation itself. **(E)** The proportional abundances for genera detected along the storage time in each replicate are shown. This also demonstrates the process reproducibility along different replicates and time.

As an additional validation step, we evaluated the batch effects of different ZSample lot productions in the bacterial profiles obtained in T0 or after 30 days (T30) of room temperature storage. High correlations (Pearson and Spearman's > 0.94) were obtained for bacterial genera comparisons between storage in T0 and T30 (Figure 2A), and also for lots produced with differences in fabrication date of up to 18 months (Pearson and Spearman's > 0.89) (Figure 2B). More detailed correlations considering other taxonomy levels as phylum, family and species are shown in Supplementary Figure 2. No significant of bacterial gain or loss due to the storage was observed in the data analyzed. Although relative abundances for bacterial phylum, family, genus or species demonstrate that the

bacterial profile in the samples have some replicate variations, these were not correlated with the ZSampe production batch (Figure 2C).

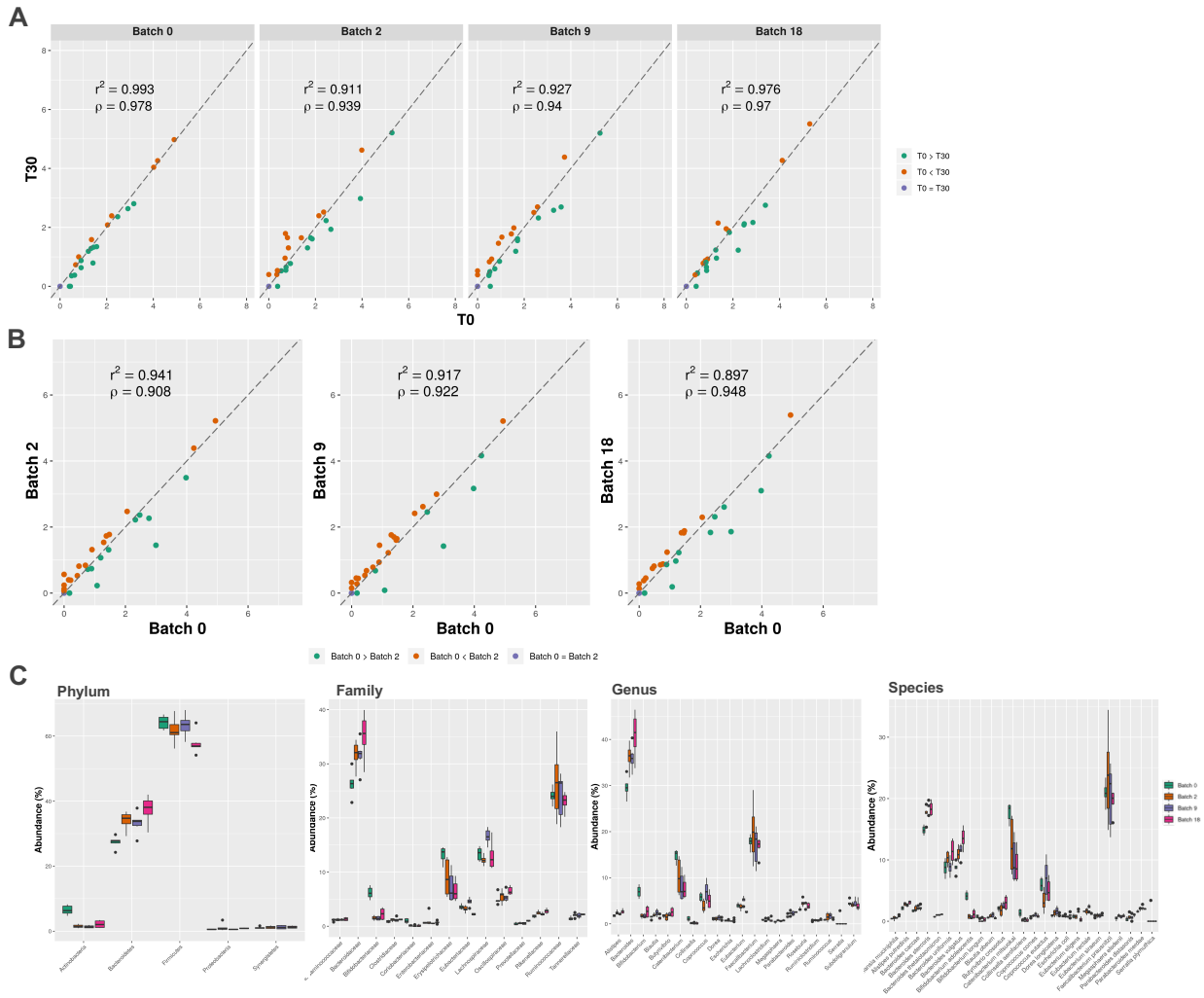


Figure 2. ZSample batch effects in fecal sample storage. Four ZSample lot production (0 - produced in the processing day -, 2, 9 and 18 months of difference from the manufacturing time) were evaluated. **(A)** A correlation analysis showed equivalent results (Pearson - r^2 and Spearman - ρ coefficients) for the lot solution batches along the time of storage T0-T30, and **(B)** among different lot production batches. **(C)** Bacterial abundances analyzed for phylum, family, genus and species in each production lot. The relative abundance levels are maintained regardless of the solution batch. Lot variations are in the same scale as the intra-replicates variations.

In addition to sample storage, DNA extraction from fecal samples in ZSample was evaluated for four different methods (Supplementary Figure 3). We observed higher correlations and similar diversities for the bacterial profiles obtained with DNeasy PowerSoil, DNeasy PowerSoil Pro modified and QIAamp PowerFecal DNA kit. The recently launched DNeasy PowerSoil Pro kit recovers a higher amount of DNA on average, showing an increased abundance of Firmicutes with reduced Bacteroidetes, Proteobacteria and Verrucomicrobia (Supplementary figure 3). Moreover, no differences related to ZSample solution in the different methods of DNA extraction were observed.

261 *High throughput amplicon sequencing robustness and analysis using EncodeTools Metabarcode*
262 *pipeline*

263 Even with the possible variations intrinsic of the method and process, the 16S rRNA amplicon
264 approach must be highly reproducible. Based on this, we performed repeatability and reproducibility
265 (robustness) tests to evaluate our method bias and variations in amplicon library preparation along
266 with the bioinformatics analysis. For the intra-run technical reproducibility test (Supplementary
267 Figures 4A-C) compared to the inter-run reproducibility tests (Supplementary Figures 4D-F) high
268 correlations were obtained, with lower variations in samples alpha and beta-diversities. The overall
269 within-sample correlations for the results obtained with the three different operators can be seen in
270 Figure 3A. Considering all the library and sequencing process variation, different operators, reagents'
271 lots, plastics and laboratory equipment (*e.g.* thermocyclers and pipettes) the Pearson and Spearman
272 correlation indices showed considerably high values, mainly above 0.9 for all samples. Alpha diversity
273 for the three independent batches of amplicon library preparations and sequencing, performed by three
274 different operators, showed equivalent indexes (Figure 3B). Beta diversity analysis also demonstrated
275 sample-related grouping patterns, which indicates within-sample distances were consistently smaller
276 than between-sample distances (Figure 3C). Negative controls showed a small number of sequenced
277 reads (from 10 to 45), with different and random profiles, while the samples themselves presented
278 from 1,882 to 47,528 reads with consistent bacterial pattern among the replicates.

279 All the analyses presented in this paper were performed using the EncodeTools Metabarcode
280 pipeline, as described in methods. Besides providing more reliable taxonomic classification due to the
281 LCA feature, this pipeline allows us to access the oligotypes present in a given sample that
282 corresponds to the real amplicon sequence variants (ASVs), and are independent of taxonomic
283 assignment. Oligotype information provides a higher-resolution view of the sample diversity and its
284 DNA sequence composition, so we used that approach to evaluate both our pipeline (EncodeTools
285 Metabarcode) and the robustness of our amplicon library preparation method. As observed in Figure
286 3 and Supplementary Figure 4, satisfactory correlations and small within-sample variability were
287 observed for the conjunction of experimental methods and the bioinformatics pipeline. To further
288 characterize the latter in terms of sensitivity and specificity, we also extended the analysis to a
289 bacterial mock as described below.

290 Specificity and sensibility of the EncodeTools Metabarcode pipeline were measured using the
291 bacterial mock results (sample - H) along with the robustness assays. 88.2 ± 2.7% sensitivity and
292 100% specificity for species level was achieved, given the possible resolution of taxonomical
293 assignment for some 16S rRNA sequences (Supplementary Figure 5). Meanwhile, 99.3 ± 2.7%

294 sensibility and 100% specificity was achieved for the genus level. At family level, the sensibility and
295 specificity reached 100%.

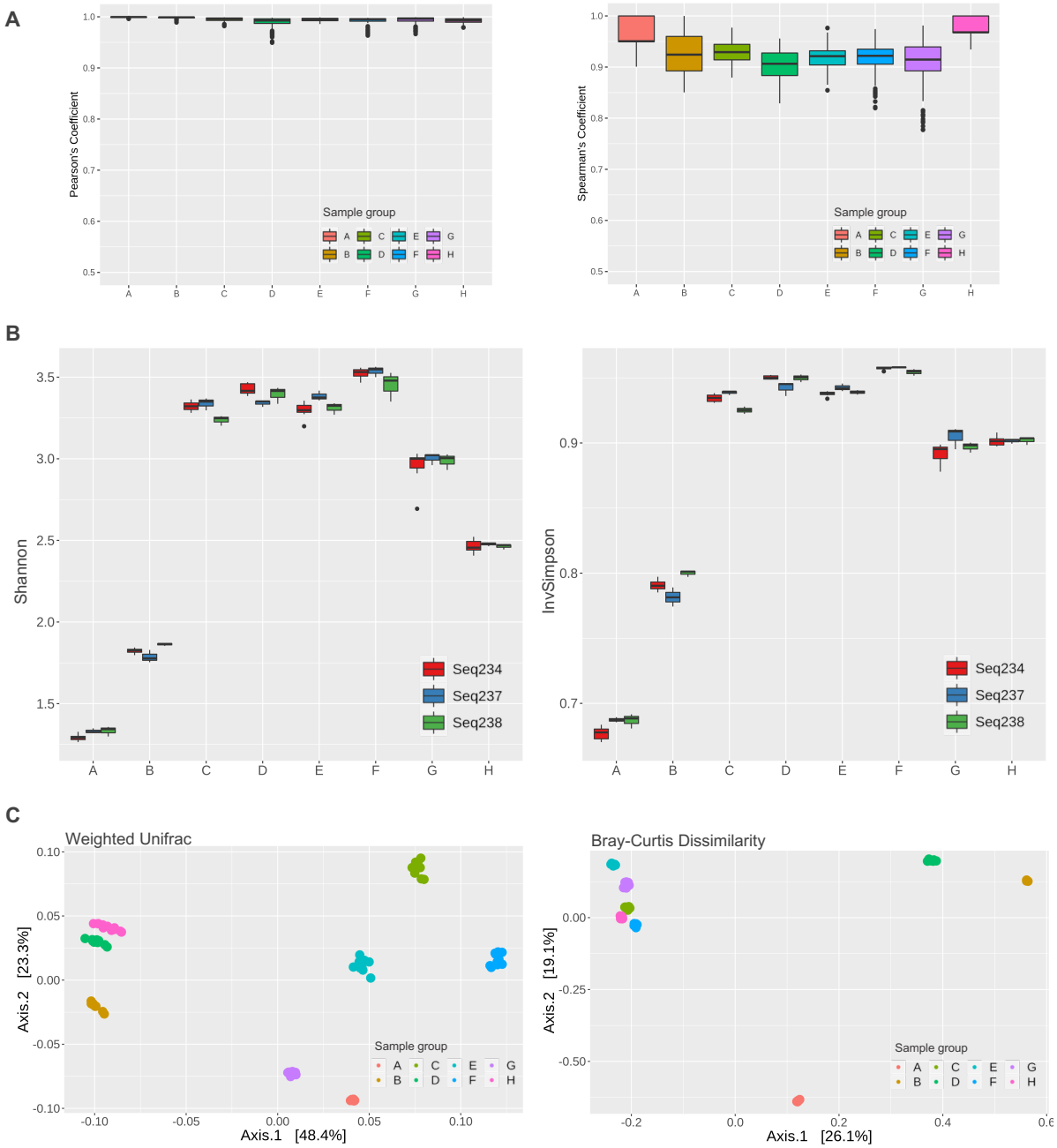
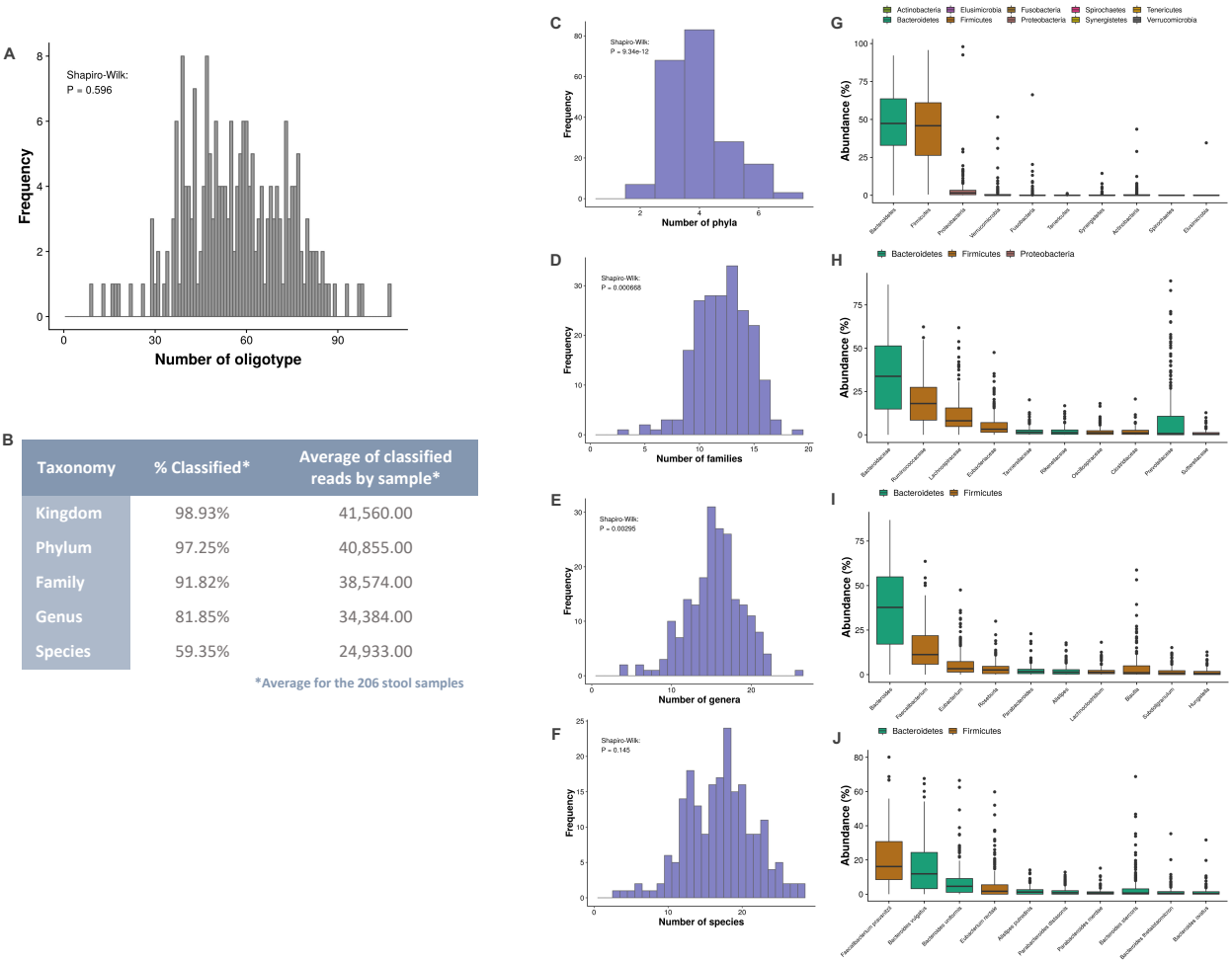


Figure 3. Method reproducibility for DNA library preparation, sequencing and analysis. Eight different DNA sample libraries (A-H) processed in replicates by three different operators and sequenced in three different sequencing runs (Seq234, Seq237, Seq238) all of which analyzed with the EncodeTool metabarcode pipeline. (A) Intra-replicates variations were assessed through correlation analysis demonstrating satisfactory results, with Pearson >0.9 and Spearman >0.8. (B) Alpha diversity indexes, Shannon and InvSimpson, obtained for replicates in each sample set were compared in parallel, showing very small differences throughout the results. (C) Beta-diversity analysis using weighted UniFrac and Bray-curtis dissimilarity, showed that each sample bacterial profile remains clustered together, confirming that variations observed in replicates are less relevant than the original bacterial composition from the different samples.

297 The EncodeTools Metabarcode pipeline generates as output an out_metabarcode
298 (Supplementary Table 2). In this table, we can verify all the oligotypes identified in the analysis, the
299 total number of reads for each oligotype, and the taxonomic assignment given to each oligotype -

300 along with their assigned taxonomic lineage (kingdom, phylum, class, order, family, genus and
301 species). This lineage path stops at the last level in which the oligotype could be classified. For
302 example, several Enterobacteria can only be classified at the family level due to the high similarity
303 among their 16S rRNA gene sequences. When the EncodeTools pipeline matches an oligotype with
304 two or more identical reference sequences, belonging to different species, genus or other higher
305 taxonomy level, the oligotype taxonomic assignment is set for the last common level (ancestor) in the
306 taxonomic path. For instance, if an oligotype could not be resolved at the species level, giving its
307 sequence similarity with two or more species, it probably will be classified at the genus or family
308 level. The out_metabarcode table shows us what are these taxonomies, their identities, and their
309 similarities in the analysis. The read sequence for each oligotype can also be visualized in this table,
310 along with a list of samples in which that given exact sequence was found.
311
312 *Brazilian bacteriome profile*
313 The experimental procedures and analyses evaluated in this paper were applied to a subset of
314 over 200 random fecal samples from the Brazilian population. A total of 8,654,114 reads were
315 obtained with an average of 42,010 reads by sample and 2,080 unique oligotypes ranging from 10 to
316 451,065 reads in the global result. The number of bacterial oligotypes for each sample varied mostly
317 between 30 to 90 (Figure 4A) well approximating a Gaussian distribution (Shapiro-Wilk, P = 0.596)
318 in the populational subset evaluated. On average, taxonomic assignment through the EncodeTools
319 Metabarcode pipeline could be obtained for 98.93% of the reads at the bacterial kingdom level,
320 97.25% at phylum, 91.82% at family, 81.85% at genus, and 59.35% at the species level (Figure 4B).
321 In this sample subset, phylum, family and genus distributions did not present a Gaussian pattern
322 (Shapiro-Wilk, P <0.01) (Figures 4C, 4D, 4E and 4F) while species are more normally distributed
323 (Shapiro-Wilk, P= 0.145).
324 Regarding the taxonomic assignment, Bacteroidetes and Firmicutes are the most abundant
325 phyla detected in the Brazilian samples with a median abundance values near to 50% (Figure 4G),
326 followed by phyla Proteobacteria, Verrucomicrobia or Actinobacteria, being the lasts, detected in
327 much lower abundances. In consequence, the most abundant families, genera and species are
328 dominated by taxonomies from Bacteroidetes and Firmicutes phyla.
329 Families Bacteroidaceae, Ruminococcaceae, Lachnospiraceae and Eubacteriaceae are the
330 most abundant families (Figure 4H). Prevotellaceae is also abundant, though its distribution showed
331 relatively lower median and strong positive-skewness, *i.e.*, many high-abundance outliers.
332 *Bacteroides* was the most abundant genera detected, followed by *Faecalibacterium*, *Eubacteria* and
333 *Roseburia* (Figure 4I). At the species level, considering the taxonomies that could be reliably resolved

334 by the EncodeTools pipeline (reflecting ~59.35% of the sequenced reads), *Faecalibacterium*
335 *prausnitzii* is the most abundant species detected in this sample subset (Figure 4J), followed by
336 *Bacteroides vulgatus*, *Bacteroides uniformis*, *Eubacterium rectale* and *Allistipes putrenidis*. Large
337 amounts of *Bacteroides* could not be classified at the species level.



Taxonomy	% Classified*	Average of classified reads by sample*
Kingdom	98.93%	41,560.00
Phylum	97.25%	40,855.00
Family	91.82%	38,574.00
Genus	81.85%	34,384.00
Species	59.35%	24,933.00

*Average for the 206 stool samples

Figure 4. Bacterial profile for the Brazilian fecal microbiome. Over 200 fecal samples from a mischaracterized population were evaluated. **(A)** EncodeTools Metabarcode pipeline showed an oligotype frequency distribution along the samples with more frequent values among 30-90 oligotypes by subject. **(B)** Each sequenced sample yielded approximately 41,500 reads and the EncodeTools Metabarcode Taxonomy Assignment algorithm could attribute phylum taxonomy for an average of 98.93% sample reads. 81.85% of the sequenced reads could be identified at genus level and 59.35% at species level, representing an average of 24,933 reads classified. **(C-F)** Populational distribution of taxonomic assignments in phylum, family, genus and species. Most frequently a subject has between 3 and 4 bacterial phyla, 10 to 15 bacterial families, 10 to 20 bacterial genera, and 12 to 22 bacterial species. Most abundant bacteria for each taxonomy level **(G)** phylum, **(H)**, family, **(I)** genus and **(J)** species are shown accordingly with their populational median distribution.

338

339 Diversity analysis were performed to visualize how these bacterial profiles are distributed in
340 the populational subset evaluated. Alpha diversity indexes (Chao1, Shannon, Simpson and
341 InvSimpson) were calculated for the samples oligotypes (Figure 5A). Chao1 was the only index with
342 a normal distribution (Shapiro-Wilk, P= 0.596). Other indexes did not show a Gaussian distribution;
343 however, they are skewed for some common ranges. The same alpha diversity analysis was performed
344 for phylum, family, genus and species (Supplementary Figures 6A-D). However, all of them presented
345 lower diversity indexes, as expected due to oligotype clustering in higher taxonomic ranks, reducing

the number of taxonomies to account for the analysis. None of these presented Gaussian distribution, except for Chao1 at species level (Supplementary Figure 6D).

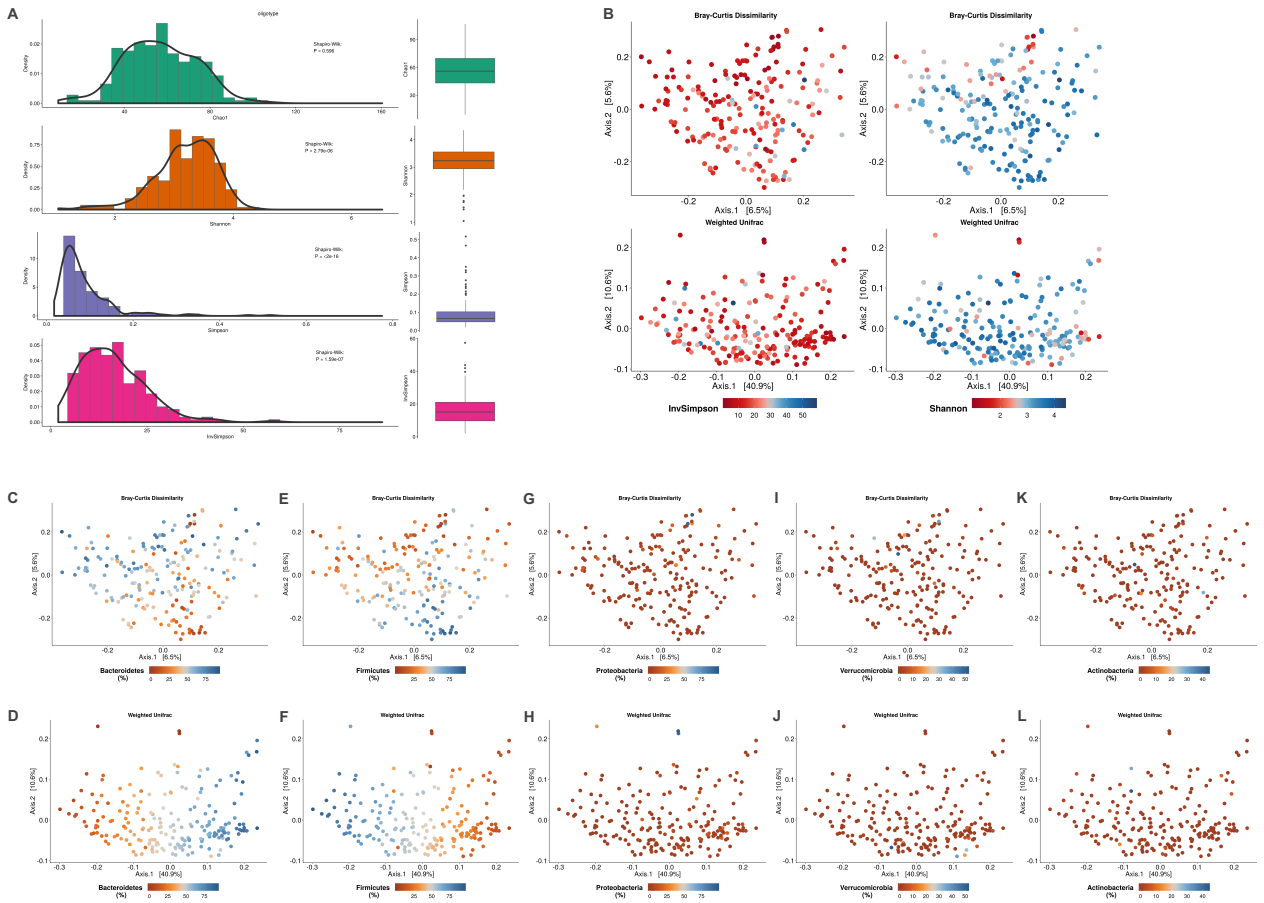


Figure 5. Brazilian bacteriome diversity analysis with oligotype sequences. (A) Alpha-diversity Chao1, Shannon, Simpson and InvSimpson indexes distribution for the data were analyzed. Only Chao1 index approximates to a Gaussian distribution (Shapiro-Wilk 0.59), however other alpha diversity indexes showed pretty narrow skewed data distribution. **(B)** PCoA plots for Bray-Curtis and weighted UniFrac showed a lack of correlation regarding alpha-diversity Shannon and InvSimpson distributions. **(C-L)** Bray-Curtis dissimilarity and weighted UniFrac PCoA plots colored by phylum abundance distribution. Most abundant phyla Bacteroidetes **(C-D)** and Firmicutes **(E-F)** have marked distributions among all the samples analyzed. Three major groups could be seen, samples higher in Firmicutes (that are lower in Bacteroidetes), samples lower in Bacteroidetes (higher in Firmicutes) and samples with equivalent amounts of both phyla. Other less abundant phyla **(G-H)** Proteobacteria, **(I-J)** Verrucomicrobia and **(K-L)** Actinobacteria doesn't seem to contribute to the populational distribution observed in the beta-diversity analysis.

The beta diversity analyses, using both Bray-Curtis dissimilarity and phylogenetic similarity Weighted UniFrac, were performed for the samples' oligotypes. PCoA plots showed that samples were widely dispersed, without specific sample subgroups. In Figure 5B, alpha diversity indexes (Shannon and InvSimpson) didn't seems to explain any significant pattern of sample distribution among the population subset. However, the PCoA arrangement, considering the first two Principal Coordinates (PC's) for both methods, seems to be guided by the two most abundant phyla in the samples: Bacteroidetes and Firmicutes. Samples with higher abundance of Bacteroidetes have smaller amounts of Firmicutes and samples with less Bacteroidetes have more abundant oligotypes attributed to Firmicutes (Figures 5C-F). Less abundant phyla have more homogeneous low abundance distribution among samples (Figures 5G-L). Lower taxonomic levels (family, genus and species) seem to be less

359 correlated with the overall beta-diversity arrangements, at least when considering the first two PCs.
360 Still, some small groupings can be observed for samples with higher amounts of Bacteroidaceae,
361 Ruminococcaceae, Prevotellaceae, *Bacteroides*, *Faecalibacterium*, *Faecalibacterium prausnitzii* and
362 *Bacteroides vulgatus* (Supplementary Figure 6). Other abundance-driven grouping tendencies can be
363 observed in the PCoAs shown in Supplementary Figure 6. However, further analyses are required to
364 establish specific correlations between certain taxonomies and possible enterotypes/bacteriome
365 profiles.

366 Finally, 30 negative controls of DNA extraction (CNE) were analyzed along with 44 negative
367 PCR reaction controls (CNR). This control analysis, performed at each sequencing batch, allows us
368 to detect deviations in the process that could invalidate the sample results. Here, the oligotype numbers
369 as well as the total reads per library obtained for control samples were notably low, which yielded
370 completely different alpha and beta diversity profiles (Supplementary Figure 7). Thus, no significant
371 batch contamination from reagents was detected and the process is capable of reliably representing
372 the original bacteriome samples' compositions.

373

374 **Discussion**

375

376 In this paper, we present an end-to-end assessment of the methodologies that we developed to
377 analyze the bacterial composition of the intestinal microbiome. First, we created a sample collection
378 kit, Probiome, that people can easily take home and use to collect a small amount of fecal sample with
379 a sterile swab and store it at room temperature using a tube containing a stabilizing solution to deliver
380 it to the laboratory within 30 days after sample collection. Then the laboratory performed the
381 following procedures: DNA sample extraction and 16S rRNA amplicon sequencing to access the
382 sample bacterial composition through a bioinformatics - EncodeTools Metabarcode pipeline. All these
383 processes were evaluated to account for processing variabilities and reproducibility of the obtained
384 results.

385 A very high load of microorganisms populates our gut. From the moment of sample collection
386 to the DNA extraction, the bacterial profile can suffer dramatic changes caused by sample degradation
387 or even microorganisms overgrowth. It may favor the detection of some microorganisms over another
388 (*e.g.*, aerobes x anaerobes). Thus, adequate sample storage is necessary until proceeding to the DNA
389 extraction to preserve the real bacterial profile in the samples [14,15,36]. Although immediately
390 freezing seems to be the best choice [36] it is not feasible in large-scale populational studies. Some
391 storage solutions have already been evaluated like RNAlater, OMINIgene-gut, Norgen, Shield, Tris-
392 EDTA, ethanol 70%, 90% or 95% and FTA cards [14,15,36–38]. Generally, these studies evaluated

393 the sample preservation at the short term, from two to seven days, and reported that OMNI, ethanol
394 95%, Norgen and FTA cards were the best preservation alternatives. However, only OMNI and
395 Norgen were shown to impairs bacterial growth in the sample, while RNAlater should be avoided
396 given its poor DNA recovery and alterations in bacterial taxa recovered [14,36,38–40]. Only one of
397 these studies performed a long-term survey of sample preservation at room temperature for eight
398 weeks, showing that OMNIgene-gut, FTA cards and ethanol 95% were the best preservatives with
399 very minimal variations, comparable to technical replicates variations [14]. Another long-term study
400 evaluated 5-year samples stored in RNAlater and frozen at -80 ºC. However, these samples remained
401 6-17 days at room temperature before freezing [41] so this study did not account for the alterations in
402 the microbial profile caused by the room temperature storage during a considerably long period -
403 which is critical given the previous research warnings to avoid RNAlater. Based on all this knowledge,
404 together with the high costs of solutions like OMNI or Norgen and the need for an accessible fecal
405 collection kit in Brazil, we developed our storage solution, ZSample. It was tested regarding bacterial
406 inactivation and profile maintenance for 30 days at room temperature. Variations in the bacterial
407 profile related to different lot productions of the solution were not detected either.

408 After sample collection, the storage lasts until the DNA extraction process, which obtain the
409 microbial genetic information of the sample. This is also an intensive subject of investigation, since
410 different DNA extraction methods can lead to different microbial profiles. Even though subject's
411 differences are known to be one of the greatest sources of variability for human microbiome data,
412 some DNA extraction methods yield more variations than others [16,42–45]. We detected significant
413 variations in the bacterial profile recovered by DNeasy PowerSoil Pro kit. These variations may be
414 attributed to the bead-beating with the zirconium beads during the lysis process. To the best of our
415 knowledge, DNeasy PowerSoil and QIAamp PowerFecal DNA represent the most used kits in
416 microbiome research studies. Hence, aiming to keep consistency with the microbiome profiles
417 reported in the literature, we continued the use of PowerSoil Kit. Nonetheless, it remains to be
418 confirmed which DNA extraction kit yields the most reliable results, *i.e.,* the one which most closely
419 resembles the original samples' bacterial composition.

420 Besides sample storage and DNA extraction, the DNA library preparation for high-throughput
421 sequencing could also have a greater impact on the assessment of the results. In general, there are two
422 main approaches used to assess the gut microbial diversity: a metabarcode analysis, such as 16S rRNA
423 gene for bacterial identification and compositional analysis, and metagenomics approaches which, in
424 addition to bacterial identification, can reveal other microorganisms such as fungi, viruses or
425 eukaryotes, as well as their interaction networks through genes and metabolism inferences. Both
426 methodologies are valid and should be applied in accordance with the expected results. To perform a

427 high-level community profiling, 16S rRNA marker gene is most indicated, whereas to perform
428 functional profiling, metagenomics must be used [13]. Additionally, previous research as the MetaHIT
429 project demonstrated that human intestinal microbiome is composed mainly of bacteria, more than
430 90% of the intestinal DNA recovered was bacterial-related [2]. Also, it was shown that 16S rRNA
431 amplicon sequencing recovers more bacterial diversity than shotgun based metagenomics [46]. Thus,
432 16S rRNA marker gene amplicons are best suited for our analysis and expected results, being the
433 method of choice for this study.

434 We evaluated the reproducibility of our amplicon library preparation performing several
435 replicates and including variables such as different operators, equipment, reagents and dates of
436 processing. These assays were also used to test our pipeline of analysis (EncodeTools Metabarcode)
437 justifying the higher number of replicates performed and a bacterial mock sample with known
438 composition. The EncodeTools Metabarcode pipeline was developed to provide more reliable results,
439 assessing single variations from the sequences with greater confidence and improved taxonomic
440 assignment.

441 The analysis of amplicon sequencing variants (ASVs), grouped into oligotypes composed by
442 sequences with 100% similarity, is the main feature that improves the 16S rRNA gene bacterial
443 profiling [13,17]. We already use this approach of reads clustering since 2014, for hospital
444 microbiome surveillances using 16S rRNA gene high-throughput sequencing [26]. Currently, new
445 bioinformatics tools are available to assist in the accuracy of obtained sequence reads, as the denoising
446 procedures based on software packages like Deblur and DADA2 [18,22,47]. These pipelines help in
447 the detection of sequencing artifacts and erroneous reads, giving a more reliable result regarding
448 oligotypes that may vary by only one nucleotide, as well as being more useful in the detection of real
449 variations among samples [22].

450 In the EncodeTools Metabarcode pipeline, we implemented a *de novo* taxonomic assignment,
451 based on similarity [25], which can classify most of oligotypes at least to the phylum level. Thus,
452 associating our EncodeTools metabarcode pipeline with the 283bp - 16S rRNA V3/V4 oligotypes and
453 a *de novo* taxonomic classification, we can obtain high-quality, highly-reproducible results. Regarding
454 taxonomic assignment, using a read length of 283bp provides a great improvement for taxonomy
455 resolution at several ranks, including at species level. This approach seems to perform even better
456 than some metagenomics approaches in which only 52.8% of the fragments could be assigned to genus
457 and 80% to phylum - while still reporting bacterial dominance within intestinal microbiome [2].

458 The inclusion of negative controls along the process is also important to assess possible
459 contaminations that may occur in the DNA extraction, PCR amplification, sequencing or even in the
460 bioinformatics pipeline, as previously reported [16,48,49]. Contaminations with some bacterial DNA

461 is ubiquitous among DNA extraction kits and laboratory reagents [48], being more relevant for
462 microbial detection in low-biomass samples [50]. In general, we detected very low number of reads
463 in negative controls, with an average of only four oligotypes and highly random bacterial profiles. In
464 each experimental batch, these contaminations must be evaluated to understand the magnitude of their
465 impact on the results, whether they can be filtered from some samples or even if they invalidate the
466 entire result. EncodeTools Metabarcode pipeline has this filtering options embedded in its code to
467 evaluate negative controls from each experimental batch.

468 The procedures described here were applied to a batch of more than 200 fecal samples
469 collected from the Brazilian population. So far, there were no reports for the microbial diversity of the
470 Brazilian gut microbiome, thus we presented a first general overview of this profile for bacterial
471 abundance and distribution. The Brazilian fecal microbiome samples have a consistent distribution of
472 oligotypes, phylum, family, genus and species along the population analyzed, often approximating
473 Gaussian distributions. Alpha and beta diversities have similar distributions to those reported by other
474 studies [51] , in which the main populational dissimilarities are guided by the most abundant phyla.
475 Generally, most of the studies published so far showed that the human intestinal microbiome is mainly
476 composed by Bacteroidetes and Firmicutes phyla [2,3,5–7,9,52], as we observed here. The Brazilian
477 microbiome profile shown here should be further investigated with stratified metadata to better
478 understand patterns and microbial diversities related to populational geography, diet, age, sex, and
479 several other possibly associated/confounding factors. Brazilians compose a very diverse and
480 geographically distributed population. Deep characterization of their microbiome profiles is necessary
481 if we want to better comprehend the applicability of the information derived from other populational
482 studies [8,12,53].

483 In conclusion, we provided an end-to-end assessment of microbiome sample processing and
484 analysis, as well as its applications to the study of the Brazilian fecal bacteriome. Using an effective
485 sample collection method, with a standardized sample processing, DNA sequencing and
486 bioinformatics analysis, we achieved highly reliable results. One of the major gains of the
487 methodology herein presented is the bioinformatics pipeline in which oligotypes represent the pure
488 sample diversity, free of biased taxonomic assignment for generalist groupings as in OTU picking
489 [13,22,54,55]. OTUs are known to underestimate sample diversity. However, sOTUs, ASVs or
490 oligotypes approaches overcome this issue, and even empower 16S rRNA studies to reveal more
491 bacterial diversity than shotgun metagenomics [46,56]. In addition to the oligotype approach, we also
492 gain phylogenetic resolution by sequencing a larger fragment that most studies do, which improves
493 taxonomical assignment in fecal sample characterizations. Using these methodologies, larger sample

494 cohorts should be analyzed for Brazilian population and more detailed comparative studies and meta-
495 analyses must increase the knowledge about the intestinal microbiome of such a diverse population.
496

497 **Conflict of interest**

498 All authors are or were currently full-time employees of BiomeHub (SC, Brazil), a research and
499 consulting company specialized in microbiome technologies.
500

501 **Funding Statement**

502 BiomeHub funded this study.

503

504 **References**

505 1. Qin J, Li R, Raes J, Arumugam M, Burgdorf K, et al. (2010) A human gut microbial gene
506 catalogue established by metagenomic sequencing. Nature 464: 59. doi:10.1038/nature08821.

507 2. Arumugam M, Raes J, Pelletier E, Paslier D, Yamada T, et al. (2011) Enterotypes of the human
508 gut microbiome. Nature 473: 174. doi:10.1038/nature09944.

509 3. Gilbert J, Blaser M, Caporaso J, Jansson J, Lynch S, et al. (2018) Current understanding of the
510 human microbiome. Nat Med 24: 392. doi:10.1038/nm.4517.

511 4. Lloyd-Price J, Abu-Ali G, Huttenhower C (2016) The healthy human microbiome. Genome Med
512 8: 51. doi:10.1186/s13073-016-0307-y.

513 5. Lynch S, Pedersen O (2016) The Human Intestinal Microbiome in Health and Disease. New Engl
514 J Medicine 375: 2369–2379. doi:10.1056/NEJMra1600266.

515 6. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, et al. (2017) Strains, functions and
516 dynamics in the expanded Human Microbiome Project. Nature 550: 61. doi:10.1038/nature23889.

517 7. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger J, et al. (2012) Structure, function and
518 diversity of the healthy human microbiome. Nature 486: 207. doi:10.1038/nature11234.

519 8. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, et al. (2018) American Gut: an
520 Open Platform for Citizen Science Microbiome Research. mSystems 3.
521 doi:10.1128/mSystems.00031-18.

522 9. Lozupone C, Stombaugh J, Gordon J, Jansson J, Knight R (2012) Diversity, stability and
523 resilience of the human gut microbiota. Nature 489: 220. doi:10.1038/nature11550.

524 10. Ley R, Lozupone C, Hamady M, Knight R, Gordon J (2008) Worlds within worlds: evolution of
525 the vertebrate gut microbiota. Nat Rev Microbiol 6: nrmicro1978. doi:10.1038/nrmicro1978.

526 11. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, et al. (2019) Extensive Unexplored Human
527 Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age,
528 Geography, and Lifestyle. Cell. doi:10.1016/j.cell.2019.01.001.

529 12. He Y, Wu W, Zheng H-M, Li P, McDonald D, et al. (2018) Regional variation limits
530 applications of healthy gut microbiome reference ranges and disease models. Nat Med 24: 1532–
531 1535. doi:10.1038/s41591-018-0164-x.

532 13. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, et al. (2018) Best practices for
533 analysing microbiomes. Nat Rev Microbiol 16: 410–422. doi:10.1038/s41579-018-0029-9.

534 14. Song S, Amir A, Metcalf J, Amato K, Xu Z, et al. (2016) Preservation Methods Differ in Fecal
535 Microbiome Stability, Affecting Suitability for Field Studies. Msystems 1: e00021–16.

536 doi:10.1128/mSystems.00021-16.
537 15. Bokulich N, Maldonado J, Kang D-W, Krajmalnik-Brown R, Caporaso J (2019) Rapidly
538 Processed Stool Swabs Approximate Stool Microbiota Profiles. Msphere 4.
539 doi:10.1128/mSphere.00208-19.
540 16. Sinha R, Abu-Ali G, Vogtmann E, Fodor A, Ren B, et al. (2017) Assessment of variation in
541 microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project
542 consortium. Nat Biotechnol 35: nbt.3981. doi:10.1038/nbt.3981.
543 17. Callahan B, McMurdie P, Holmes S (2017) Exact sequence variants should replace operational
544 taxonomic units in marker-gene data analysis. Isme J 11: ismej2017119.
545 doi:10.1038/ismej.2017.119.
546 18. Nearing JT, Douglas GM, Comeau AM, Langille M (2018) Denoising the Denoisers: an
547 independent evaluation of microbiome sequence error-correction approaches. PeerJ 6: e5364.
548 doi:10.7717/peerj.5364.
549 19. Wang Y, Qian P-Y (2009) Conservative fragments in bacterial 16S rRNA genes and primer
550 design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS ONE 4: e7401.
551 doi:10.1371/journal.pone.0007401.
552 20. Caporaso J, Lauber C, Walters W, Berg-Lyons D, Huntley J, et al. (2012) Ultra-high-throughput
553 microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6.
554 doi:10.1038/ismej.2012.8.
555 21. Caporaso J, Lauber C, Walters W, Berg-Lyons D, Lozupone C, et al. (2010) Global patterns of
556 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National
557 Academy of Sciences of the United States of America 108 Suppl: 4516.
558 doi:10.1073/pnas.1000080107/-/DCSupplemental.
559 22. Amir A, McDonald D, Navas-Molina J, Kopylova E, Morton J, et al. (2017) Deblur Rapidly
560 Resolves Single-Nucleotide Community Sequence Patterns. Msystems 2: e00191–16.
561 doi:10.1128/mSystems.00191-16.
562 23. Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source
563 tool for metagenomics. Peerj 4: e2584. doi:10.7717/peerj.2584.
564 24. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J
565 Mol Biol 215: 403–410. doi:10.1016/S0022-2836(05)80360-2.
566 25. Yarza P, Yilmaz P, Pruesse E, Glöckner F, Ludwig W, et al. (2014) Uniting the classification of
567 cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol
568 12: nrmicro3330. doi:10.1038/nrmicro3330.
569 26. Christoff A, Sereia A, Hernandes C, Oliveira L (2019) Uncovering the hidden microbiota in
570 hospital and built environments: New approaches and solutions. Exp Biol Med: 153537021882185.
571 doi:10.1177/1535370218821857.
572 27. Price M, Dehal P, Arkin A (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for
573 Large Alignments. Plos One 5: e9490. doi:10.1371/journal.pone.0009490.
574 28. Lozupone C, Knight R (2005) UniFrac: a New Phylogenetic Method for Comparing Microbial
575 Communities. Applied and Environmental Microbiology 71: 8228. doi:10.1128/AEM.71.12.8228-
576 8235.2005.
577 29. McMurdie P, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis
578 and Graphics of Microbiome Census Data. Plos One 8: e61217. doi:10.1371/journal.pone.0061217.
579 30. Love M, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for
580 RNA-seq data with DESeq2. Genome Biol 15: 550. doi:10.1186/s13059-014-0550-8.

581 31. Robinson M, McCarthy D, Smyth G (2010) edgeR: a Bioconductor package for differential
582 expression analysis of digital gene expression data. Bioinformatics 26: 139–140.
583 doi:10.1093/bioinformatics/btp616.

584 32. Paulson J, Stine O, Bravo H, Pop M (2013) Differential abundance analysis for microbial
585 marker-gene surveys. Nat Methods 10: nmeth.2658. doi:10.1038/nmeth.2658.

586 33. Hothorn T, Hornik K, Wiel MA van de, Zeileis A (2008) Implementing a class of permutation
587 tests: the coin package. Journal of Statistical Software 28: 23.

588 34. Weiss S, Xu Z, Peddada S, Amir A, Bittinger K, et al. (2017) Normalization and microbial
589 differential abundance strategies depend upon data characteristics. Microbiome 5: 27.
590 doi:10.1186/s40168-017-0237-y.

591 35. McMurdie P, Holmes S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is
592 Inadmissible. Plos Comput Biol 10: e1003531. doi:10.1371/journal.pcbi.1003531.

593 36. Choo J, Leong L, Rogers G (2015) Sample storage conditions significantly influence faecal
594 microbiome profiles. Scientific Reports 5: 16350. doi:10.1038/srep16350.

595 37. Vandeputte D, Tito RY, Vanleeuwen R, Falony G, Raes J (2017) Practical considerations for
596 large-scale gut microbiome studies. FEMS Microbiol Rev 41: S154–S167.
597 doi:10.1093/femsre/fux027.

598 38. Chen Z, Hui P, Hui M, Yeoh Y, Wong P, et al. (2019) Impact of Preservation Method and 16S
599 rRNA Hypervariable Region on Gut Microbiota Profiling. mSystems 4.
600 doi:10.1128/mSystems.00271-18.

601 39. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, et al. (2015) Methods for Improving
602 Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of
603 Stool. Plos One 10: e0134802. doi:10.1371/journal.pone.0134802.

604 40. Dominianni C, Wu J, Hayes R, Ahn J (2014) Comparison of methods for fecal microbiome
605 biospecimen collection. Bmc Microbiol 14: 103. doi:10.1186/1471-2180-14-103.

606 41. Tap J, Cools-Portier S, Pavan S, Druesne A, Öhman L, et al. (2019) Effects of the long-term
607 storage of human fecal microbiota samples collected in RNAlater. Sci Rep-uk 9: 601.
608 doi:10.1038/s41598-018-36953-5.

609 42. Mackenzie B, Waite D, Taylor M (2015) Evaluating variation in human gut microbiota profiles
610 due to DNA extraction method and inter-subject differences. Front Microbiol 6: 130.
611 doi:10.3389/fmicb.2015.00130.

612 43. Marotz C, Amir A, Humphrey G, Gaffney J, Gogul G, et al. (2017) DNA extraction for
613 streamlined metagenomics of diverse environmental samples. Biotechniques 62: 290–293.
614 doi:10.2144/000114559.

615 44. Wesolowska-Andersen A, Bahl M, Carvalho V, Kristiansen K, Sicheritz-Pontén T, et al. (2014)
616 Choice of bacterial DNA extraction method from fecal material influences community structure as
617 evaluated by metagenomic analysis. Microbiome 2: 19. doi:10.1186/2049-2618-2-19.

618 45. Costea P, Zeller G, Sunagawa S, Pelletier E, Alberti A, et al. (2017) Towards standards for
619 human fecal sample processing in metagenomic studies. Nat Biotechnol 35: nbt.3960.
620 doi:10.1038/nbt.3960.

621 46. Tessler M, Neumann J, Afshinnekoo E, Pineda M, Hersch R, et al. (2017) Large-scale
622 differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci
623 Rep-uk 7: 6589. doi:10.1038/s41598-017-06665-3.

624 47. Callahan B, McMurdie P, Rosen M, Han A, Johnson A, et al. (2016) DADA2: High-resolution
625 sample inference from Illumina amplicon data. Nat Methods 13: nmeth.3869.

626 doi:10.1038/nmeth.3869.
627 48. Salter S, Cox M, Turek E, Calus S, Cookson W, et al. (2014) Reagent and laboratory
628 contamination can critically impact sequence-based microbiome analyses. Bmc Biol 12: 87.
629 doi:10.1186/s12915-014-0087-z.
630 49. Eisenhofer R, Minich J, Marotz C, Cooper A, Knight R, et al. (2018) Contamination in Low
631 Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol 27: 105–
632 117. doi:10.1016/j.tim.2018.11.003.
633 50. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, et al. (2018) KatharoSeq Enables High-
634 Throughput Microbiome Analysis from Low-Biomass Samples. mSystems 3.
635 doi:10.1128/mSystems.00218-17.
636 51. Zhernakova A, Kurilshikov A, Bonder M, Tigchelaar E, Schirmer M, et al. (2016) Population-
637 based metagenomics analysis reveals markers for gut microbiome composition and diversity.
638 Science 352: 565–569. doi:10.1126/science.aad3369.
639 52. Li J, Jia H, Cai X, Zhong H, Feng Q, et al. (2014) An integrated catalog of reference genes in the
640 human gut microbiome. Nat Biotechnol 32: 834–841. doi:10.1038/nbt.2942.
641 53. Martínez I, Stegen J, Maldonado-Gómez M, Eren M, Siba P, et al. (2015) The Gut Microbiota of
642 Rural Papua New Guineans: Composition, Diversity Patterns, and Ecological Processes. Cell
643 Reports 11: 527–538. doi:10.1016/j.celrep.2015.03.049.
644 54. Janssen S, McDonald D, Gonzalez A, Navas-Molina J, Jiang L, et al. (2018) Phylogenetic
645 Placement of Exact Amplicon Sequences Improves Associations with Clinical Information.
646 Msystems 3: e00021–18. doi:10.1128/mSystems.00021-18.
647 55. Allaband C, McDonald D, Vázquez-Baeza Y, Minich J, Tripathi A, et al. (2018) Microbiome
648 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. Clin Gastroenterol
649 H. doi:10.1016/j.cgh.2018.09.017.
650 56. Rodriguez-R L, Castro J, Kyrpides N, Cole J, Tiedje J, et al. (2018) How Much Do rRNA Gene
651 Surveys Underestimate Extant Bacterial Diversity? Appl Environ Microb 84: e00014–18.
652 doi:10.1128/AEM.00014-18.
653
654