

SCIENTIFIC REPORTS



OPEN

Methodology challenges in studying human gut microbiota – effects of collection, storage, DNA extraction and next generation sequencing technologies

Marina Panek¹, Hana Čipčić Paljetak¹, Anja Barešić², Mihaela Perić¹, Mario Matijašić¹, Ivana Lojkić³, Darija Vranešić Bender⁴, Željko Krznarić⁵ & Donatella Verbanac¹

The information on microbiota composition in the human gastrointestinal tract predominantly originates from the analyses of human faeces by application of next generation sequencing (NGS). However, the detected composition of the faecal bacterial community can be affected by various factors including experimental design and procedures. This study evaluated the performance of different protocols for collection and storage of faecal samples (native and OMNIgene.GUT system) and bacterial DNA extraction (MP Biomedicals, QIAGEN and MO BIO kits), using two NGS platforms for 16S rRNA gene sequencing (Illumina MiSeq and Ion Torrent PGM). OMNIgene.GUT proved as a reliable and convenient system for collection and storage of faecal samples although favouring *Sutterella* genus. MP provided superior DNA yield and quality, MO BIO depleted Gram positive organisms while using QIAGEN with OMNIgene.GUT resulted in greatest variability compared to other two kits. MiSeq and IT platforms in their supplier recommended setups provided comparable reproducibility of donor faecal microbiota. The differences included higher diversity observed with MiSeq and increased capacity of MiSeq to detect *Akkermansia muciniphila*, [*Odoribacteraceae*], *Erysipelotrichaceae* and *Ruminococcaceae* (primarily *Faecalibacterium prausnitzii*). The results of our study could assist the investigators using NGS technologies to make informed decisions on appropriate tools for their experimental pipelines.

The scientific interest and mounting knowledge on human gut microbiota have led to many important findings associating the composition of bacterial taxa in the human gastrointestinal tract with disorders from the neurologic, psychiatric, respiratory, cardiovascular, gastrointestinal, hepatic, autoimmune, metabolic and oncologic spectra^{1,2}. The main source of the information regarding human gut microbiota originated from the analyses of the faeces. The importance of this easily available metabolic waste comes from its microbial content found to hold the potential in diagnosis, disease prediction and therapeutic intervention. This justifies it being considered a “virtual organ”³ or a human tissue^{4,5}. The analyses of human intestinal microorganisms were, until recently, performed by culture-dependent methodologies, limiting the outputs to the cultivable species. The availability of novel tools, primarily next-generation sequencing (NGS), enabled the assessment of genes and genomes contained within complex microbial communities. The most widely used NGS method for the taxonomic and phylogenetic evaluation of bacterial community composition relies on 16S rRNA gene amplicon analysis^{4,6–9}.

¹University of Zagreb School of Medicine, Center for Translational and Clinical Research, Šalata 2, 10000, Zagreb, Croatia. ²MRC London Institute of Medical Sciences, Du Cane Road, London, W12 0NN, UK. ³Croatian Veterinary Institute, Department for Virology, Savska cesta 143, 10000, Zagreb, Croatia. ⁴University Hospital Zagreb, Department of Internal Medicine, Unit of Clinical Nutrition, Kišpatičeva 12, 10000, Zagreb, Croatia. ⁵University of Zagreb School of Medicine, Department of Internal Medicine, Kišpatičeva 12, 10000, Zagreb, Croatia. Marina Panek, Hana Čipčić Paljetak and Anja Barešić contributed equally to this work. Correspondence and requests for materials should be addressed to M.P. (email: mihaela.peric@mef.hr)

However, detected composition of the faecal bacterial community can be affected by experimental design and procedures, including sampling and storage protocol as well as DNA extraction method.

Sample collection procedures represent one of the first crucial steps that ensure integrity and stability of the collected material. Although an early study claimed no difference in community structure upon collection under field conditions¹⁰, subsequent reports showed importance of the collection procedure^{11–14} with accuracy increased by prompt sample downstream processing (within 2–3 hours^{11–14}), immediate stabilization^{12,13,15} and appropriate storage conditions^{11–14,16,17}. Faeces represents one of the most complex biological materials for bacterial DNA isolation as it also contains remnants of human DNA, food DNA and many inhibitors hampering subsequent PCR amplification and NGS procedures¹⁸. DNA extraction from faeces represents the essential step in obtaining good quality DNA and ensuring accurate identification of microbial composition and relative abundance¹⁹. It is therefore critical to evaluate the available methodologies for bacterial DNA extraction from faeces and optimize protocols and procedures that would provide sufficient amount, purity and integrity of the DNA, thus obtaining high quality sample for further studies.

During the last decade sequencing tools have gradually shifted from conventional Sanger sequencing technology to NGS^{9,20} and currently available technologies with sequencers from Pacific Bioscience, Roche, Thermo Fischer Scientific and Illumina, have all been successfully applied for the analyses of complex biological samples^{7,9,19,21,22}. NGS is used for determining bacterial composition in a sample based on DNA fragment detection from the total DNA isolate. Although shotgun sequencing provides the most complete information on the entire gene pool within the sample, the high amount of generated data requires substantial bioinformatic efforts in sequence assembly, mapping and analyses. In many studies, both clinical and environmental, sequencing of 16S rRNA gene amplicons, covering variable regions of the gene, is the method of choice for the analyses of bacterial community composition, providing cost-effectiveness, sufficient resolution and sequencing depth^{1,7,8,20}. The 16S rRNA gene is approximately 1500 bp in size and consists of nine variable regions separated by conserved regions^{19,23}. Illumina and Thermo platforms with their MiSeq and Ion Torrent (IT) Personal Genome Machine (PGM) benchtop sequencers are increasingly being used for 16S rRNA-based analyses of diverse bacterial populations^{7,9,19,21,22}. Although both Illumina and IT platforms sequence DNA by detecting the nucleotide addition during DNA synthesis, the two platforms function on different principles, which could affect their performance and ultimately result in variation of the obtained data^{7,9,24–27}.

As each step in the experimental pipeline introduces variation influencing the final output, there is an unmet need for standardization of methodology which would enable reliable and reproducible analysis of valuable human biological samples for studying gut microbiota. This study evaluated the performance of different protocols for collection and storage of faeces (native versus OMNIgene.GUT system) as well as bacterial DNA extraction (with MP Biomedicals, Qiagen and MO BIO commercial kits) using the manufacturer recommended methodology for 16S rRNA gene sequencing on two commercially available platforms (Illumina MiSeq and IT PGM).

Results

Stool samples from 4 individuals were collected and processed as described in Fig. 1, yielding a total of 96 donor and two mock sample sequences.

DNA yield and quality. Summary of DNA isolation methods and obtained DNA concentrations per sampling parameter are given in Supplementary Table 2. The yield and quality of DNA differed between kits, with the average yield and the estimated purity of DNA, both of which varied between kits in the MP > QIA > MO BIO order (Fig. 2). Extracted DNA concentrations were normalized by faecal weight (Fig. 2a). MP yielded approximately three times more DNA than MO BIO and QIA kits (0.34 ± 0.018 , 0.09 ± 0.03 and 0.12 ± 0.02 DNA (ng/ μ l) per mg faeces for MP, MO BIO and QIA, respectively). The best A_{260}/A_{280} absorbance ratio was obtained with MP and QIA, with mean values of 2.00 and 1.91, respectively, while mean ratio for MO BIO was significantly lower at 1.55 (Fig. 2b).

A slightly higher DNA yield was obtained when using OMNIgene.GUT collection system compared to the conventional faeces collection at both sampling time points (Supplementary Fig. 1).

Next generation sequencing – platform comparison. A total of 48 human faecal microbiota DNA samples were obtained, amplified and profiled by 16S RNA amplicon sequencing using Illumina MiSeq (V3–V4 region) and Ion Torrent PGM (V2, V4, V8 and V3, V6–V7, V9 regions). Different primer sets were used to enable comparison of supplier recommended protocols.

Clustering and annotation of OTUs was performed by the same pipeline for both sequencing platforms. Average number of counts per sample was 257.806 ± 23.076 for MiSeq and 110.512 ± 7.529 for IT. A comparison of OTU assignment on each taxonomic level (from species to phylum) revealed that, all OTUs were assigned from phylum to order level for both platforms. At lower taxonomic levels, however, higher proportion of the assigned OTUs was obtained for samples sequenced with MiSeq than those sequenced with IT (Fig. 3).

Microbiota content complexity. Further insight into the effects introduced by the variations of the experimental pipeline was gained through detailed taxa level analyses, focusing mainly on phylum and family levels since former provides maximal and latter lowest taxonomic level where >85% of annotation was achieved. The annotation at genus level, although lower and variable, was considered sufficient for further analyses but the species level was not further considered due to its inferior and varying annotation efficiency (Fig. 3).

The first platforms comparison was performed on the mock community sample with even DNA quantities from a defined bacterial community. The level of annotated OTUs for the mock community was comparable to that achieved in faeces samples (Fig. 3). The richness of detected taxa on both platforms was in good agreement

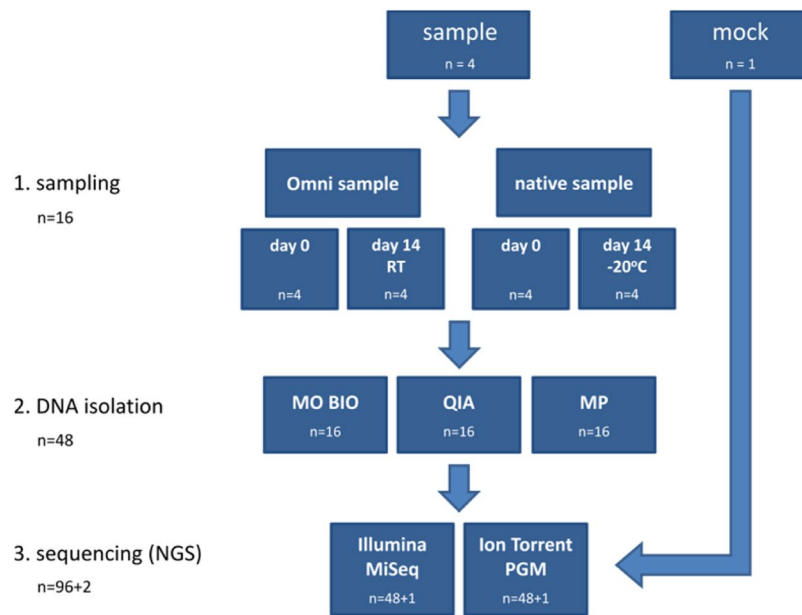


Figure 1. Experimental design. Schematic representation of procedures and conditions for isolation and analysis of microbial DNA from faeces. The figure presents two different sample collection methods, two different storage conditions/sampling time points, three different DNA extraction kits, as well as two different sequencing platforms. Omni = OMNIgene.GUT, RT = room temperature. +1/+2 refer to non-faecal mock samples.

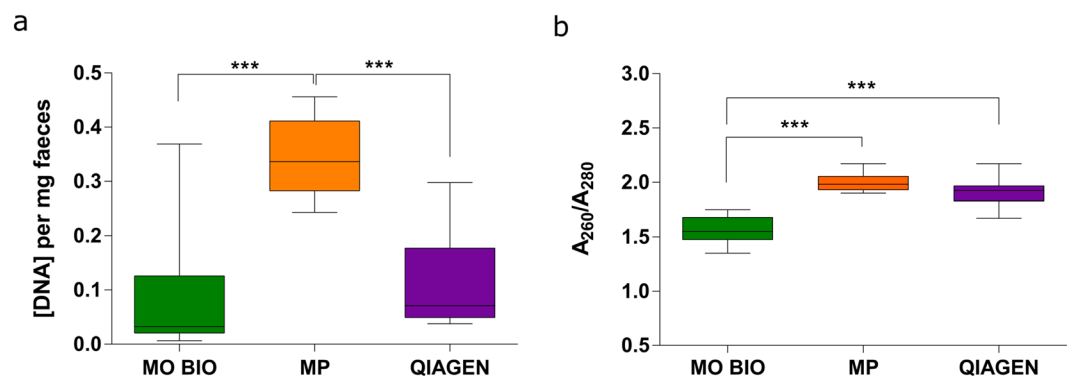


Figure 2. Yield and purity of extracted DNA varies between isolation kits. (a) DNA yield and (b) quality obtained using different DNA extraction kits (n = 16 samples per kit). DNA yield is expressed as DNA concentration (ng/μl) normalized by quantity of faeces used for extraction. Bonferroni-adjusted unpaired, two-sided t-test: ***p < 0.001.

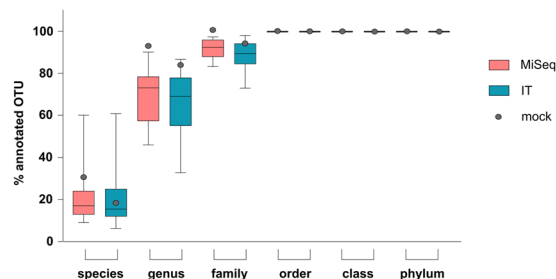


Figure 3. Annotated OTUs per platform at all taxonomic levels. Percentage (%) of annotated OTUs on each taxonomic level for both Illumina MiSeq (MiSeq) and Ion Torrent PGM (IT) sequencing platforms.

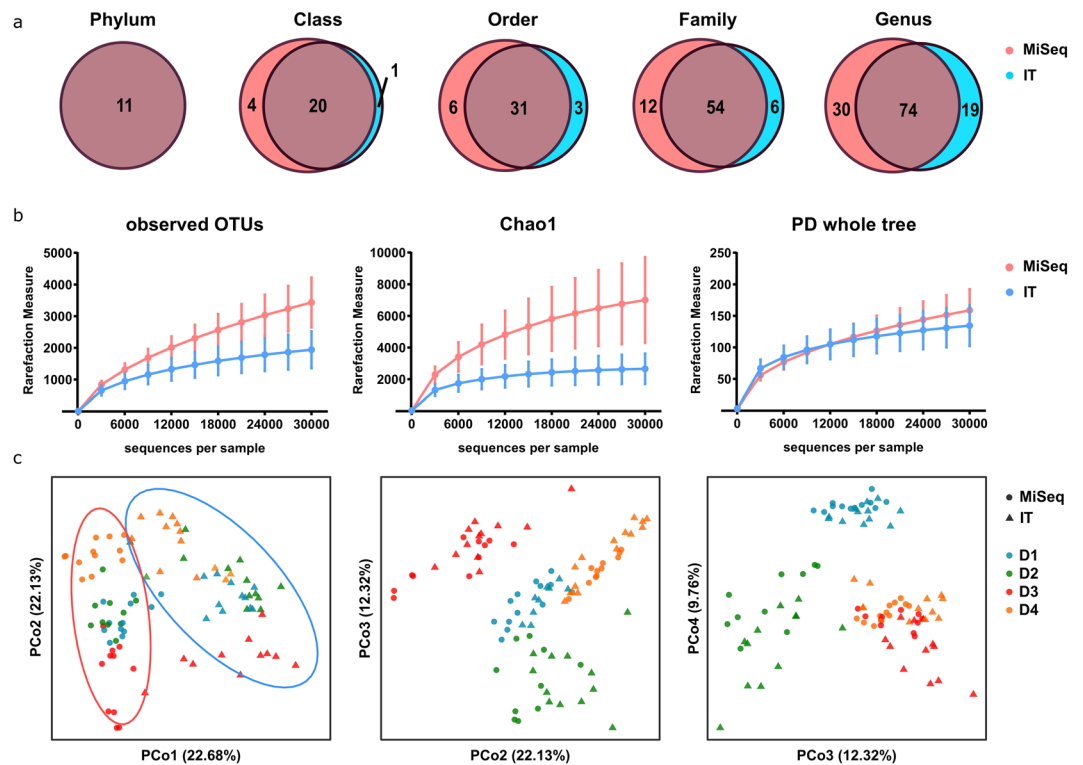


Figure 4. Comparison of diversity between MiSeq and IT platforms. **(a)** Venn diagrams of common and unique number of taxa for individual platform at each taxonomic level, **(b)** alpha diversity indices after rarefaction: observed OTUs, Chao1 index and PD_whole_tree index, **(c)** PCoA plots of beta diversity for first four coordinates marked per platform (MiSeq circled in magenta, IT in turquoise) and per donor.

with the theoretical distribution of all taxa present. However, the evenness diverged to some extent from the theoretically expected values (Supplementary Fig. 2). At genus level all 17 genera were properly assigned (annotation 92% and 83% for MiSeq and IT, respectively) (Fig. 3). This was not the case for species level where only 7/20 were identified (annotation 30% for MiSeq and 17% for IT).

Next, donor samples were evaluated with respect to total number of commonly annotated taxa at different phylogenetic levels and their diversities (Fig. 4). Raw OTU counts analysis identified many commonly detected, but also a number of platform-specific taxa at each level (Fig. 4a). Diversity indices were calculated to describe the complexity of samples (alpha) and distinguish differences between samples (beta). Alpha diversity was assessed through three indices (Fig. 4b). The more rapid increase in observed OTUs and higher Chao1 (a measure highly favouring singletons and doubletons) with MiSeq than IT was noted. However the difference is lost when phylogenetic distance of detected taxa is taken into account and expressed using PD_whole_tree measure. Additionally, the same indices point to higher diversity in samples where MP was used for DNA extraction (Supplementary Fig. 3). Principle coordinate analysis of beta diversity expressed as weighted UniFrac (Fig. 4c), revealed a major contribution to the differences in diversity first by platform (PCo1) followed by donors (PCo2–4).

Analysis of microbiota composition. Next, relative abundance of the bacterial phyla detected within the entire sample pool was compared, aggregating all the sequencing data according to platform, thus excluding variations due to other experimental parameters (Supplementary Fig. 4). *Bacteroidetes* and *Firmicutes* phyla accounted for most of the taxon-assigned OTUs identified in all samples ($\geq 90\%$ of all OTUs) (Supplementary Fig. 4a). At the family level, the most abundant taxa identified were *Bacteroidaceae*, *Lachnospiraceae*, *Ruminococcaceae* and *Prevotellaceae* (accounting for $>80\%$ of assigned OTUs), displaying comparable abundance ratios between platforms (Supplementary Fig. 4b).

The relative abundance patterns among sample collection procedures and DNA extraction kits were compared on family level (Supplementary Fig. 5). Although OMNIGene.GUT system notably favoured the isolation of *Prevotellaceae* in all tested samples, a good agreement in terms of relative abundance of detected families was found between native and OMNIGene.GUT samples (Omni) (Supplementary Fig. 5a). Variation in *Bacteroidaceae*/*Lachnospiraceae* abundance ratio was observed for samples extracted with different DNA extraction kits (Supplementary Fig. 5b).

Further analysis on phylum (Supplementary Fig. 6) and family (Fig. 5) level included all combinations of experimental parameters tested. Setting the fresh native sample (native 0) as a reference point and comparing the kits (Fig. 5a), it was observed that MO BIO kit tends to provide higher proportion of *Bacteroidaceae* (48.6%) and less *Lachnospiraceae* (17.3%), compared to MP (33.4% and 25.2%) and QIA extraction kit (24.8% and 28.4%, respectively). Comparing the frozen native samples (native 14) with their corresponding reference points revealed

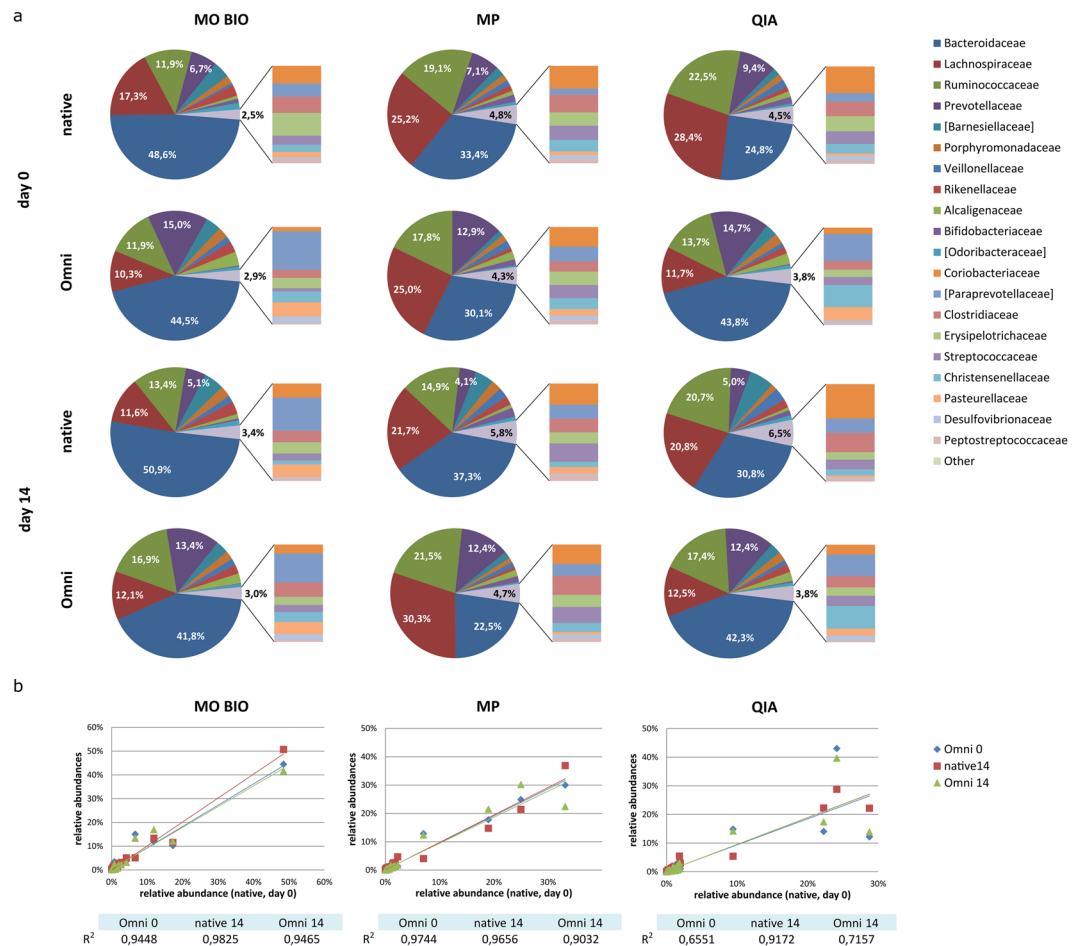


Figure 5. Relative abundance of families depends on sample collection/storage and DNA extraction procedure. Analysis of (a) DNA extraction procedures (MO BIO, MP, QIA) with respect to sampling time point (day 0, day 14) and sample collection procedure (native, Omni). Twenty most abundant families are presented, the remainder have been summed as “Other”. 10 major families are displayed on a pie chart, while the remaining ones are shown on a side bar. (b) Native 0 sample relative family abundances plotted against corresponding native 14, as well as Omni 0 and Omni 14 relative family abundances for each kit. R² values for linear regression on each time point are displayed in the table below.

the increase in *Bacteroidaceae*/*Lachnospiraceae* ratio and identification of slightly less *Prevotellaceae* across all extraction kits. Collecting the samples with OMNIGene.GUT stabilizing system (Omni 0, Omni 14) tended to increase the *Bacteroidaceae*/*Lachnospiraceae* ratio for QIA and MO BIO extracted samples. Altogether, application of MP and MO BIO resulted in fewer discrepancies across all tested parameters (Fig. 5b). QIA displayed variations between native and Omni samples most notably in the ratio of two dominant families, *Bacteroidaceae* and *Lachnospiraceae*.

Additionally, evaluation of Gram+ and Gram− bacteria extraction efficiency revealed high dependence on collection and storage procedure when using QIA (Fig. 6) as it isolated more G− than G+ bacteria compared to native samples after OMNIGene.GUT stabilisation. MO BIO was less efficient in extracting G+ organisms, while MP kit provided more G+ and an even G+/G− ratio. Unlike QIA, both MO BIO and MP displayed balanced extraction patterns across collection and stabilization procedures.

The relative abundances of all samples are presented in Fig. 7b. Although each parameter investigated contributed to sample variability, PCA analyses of transformed data (Fig. 7a) confirmed that the most prominent driver of sample clustering was subject specificity, irrespective of any other experimental or platform-related parameter, with each donor maintaining a unique and distinct faecal microbiota profile. Finally, MiSeq relative abundances follow a log-linear trend more closely than IT, with higher correlation and narrower error margins (Fig. 7c).

Statistical analyses of covariates. The analysis of all covariates presented in Fig. 1 on (clr) transformed data²⁸ was performed on each level down to genus.

Overall statistics. Wilcoxon rank test was performed for each covariate pair. Statistically significant differences in transformed taxon counts were observed between platforms and kits on family (MiSeq:IT p = 0.026; MOBIO:MP p = 0.049) and genus levels (MiSeq:IT p < 0.001; MOBIO:MP, MOBIO:QIA, MP:QIA all p < 0.001).

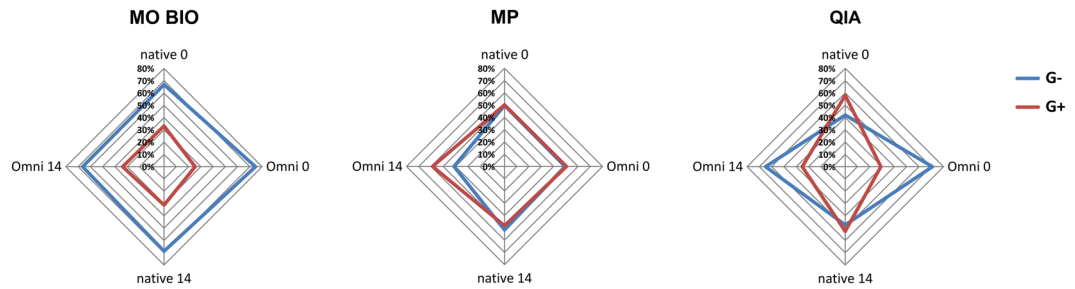


Figure 6. The influence of sample collection/storage protocol and DNA extraction kits on the detected level of Gram+ and Gram- bacteria. The graphs show DNA extraction procedures (MO BIO, MP, QIA) in different sample collection (native, Omni) and storage conditions (day 0, day 14).

Omni 0 was found different from both native 0 and native 14 groups of samples only at the genus level ($p = 0.003$ and 0.004 , respectively).

Taxon significance. In order to test the effect of each covariate on taxon composition, two tests were performed on clr-transformed data: Kruskal Wallis rank test testing each pair of covariates, and Wilcoxon rank test on Monte-Carlo simulated data testing a covariate vs. all others, e.g. one kit vs. other two. Table 1 lists the main contributors driving differences between platforms and DNA extraction kits at phylum and family levels. Other taxonomic levels and combinations of covariates with statistical significance are available in Supplementary Information (Supplementary Tables 3 and 4).

Platform comparison revealed significant differences among majority of detected phyla. *Actinobacteria*, *Firmicutes*, *Proteobacteria* and *Tenericutes* were found more abundant on IT while *Lentisphaerae* and *Verrucomicrobia* were found more abundant on MiSeq. These differences were statistically significant but the effect sizes were not substantial to support genuine relevance, except *Verrucomicrobia*, which was both highly significant and had a high effect size (a measure quantifying the magnitude of difference between two groups). Even though *Verrucomicrobia* was not a highly abundant phylum, MiSeq detected 10 fold more *Verrucomicrobia* than IT (clr values). It is worth mentioning that the only contribution to *Verrucomicrobia* phylum stems from *Akkermansia muciniphila* species. At the family level, there were ten families that contributed to platform differences, all more abundant on MiSeq, but only four [*Odoribacteraceae*], *Erysipelotrichaceae*, *Ruminococcaceae* and *Verrucomicrobiaceae* have demonstrated high significance and effect sizes. Covariance biplots for organisms with statistical significance ($p < 0.05$) on family and genus levels (showing highest platform discrepancy) are available in Supplementary Information (Supplementary Fig. 7).

For the DNA extraction kits comparison, MP tends to extract *Actinobacteria* and *Firmicutes* more efficiently than other two kits, while MOBIO is the least efficient in extracting representatives from *Actinobacteria* phylum, confirmed with both the Kruskal Wallis and the Wilcoxon tests. The majority of significantly differing families (Wilcoxon rank test) between kits were found in MP versus other kits (Table 1), with all families more efficiently extracted with MP (1.9–3.2 fold). Kruskal Wallis test revealed that this difference can be accounted for by disparities between MP and MOBIO kits, with QIA coming out as less specific (Table 1). The families with sufficiently high effect size are [*Mogibacteriaceae*], *Clostridiaceae*, *Coriobacteriaceae*, *Lachnospiraceae*, *Peptostreptococcaceae* and *Streptococcaceae*. When MOBIO is compared to other two kits, only three of the above mentioned taxa (*Mogibacteriaceae*], *Coriobacteriaceae*, and *Streptococcaceae*) show significance which stems from lower abundance (by a factor of 2–3.8) these taxa achieve with MOBIO kit (Supplementary Tables 3 and 4).

For the collection-storage combination, Omni 0 was found to enable better extraction of *Betaproteobacteria* class, *Burkholderiales* order, *Alcalligenaceae* family and *Suterella* genus, which was substantiated by Wilcoxon rank test (Supplementary Tables 3 and 4).

Discussion

With the recent advances in sequencing technologies cost-effective methodologies for microbial composition determination from complex environmental samples became widely available to broad research community. However, each methodological step comprises of multiple technical variables and potentially introduces variation in the final results. These steps include methods of sample collection, stabilisation and storage; DNA extraction and sequencing technology which play an important role in preserving the bacterial content of the sample. Consequently, thorough selection of experimental design and subsequent interpretation of the results needs to be implemented. All this wealth of advanced technological and methodological options exposes the researchers entering the field to plethora of choices only a subset of which is tailored to their particular needs. In addition, the data analysis of the experimental outputs create further level of complexity related to interpretation and comparison across studies in the same research field^{27–30}.

Current collection protocols propose the immediate use of fresh faeces or rapid freezing at -80 or -20 °C as optimal^{11,12,16,31}. Our results suggest that freezing at -20 °C preserves the specific microbial community composition in faecal samples (Fig. 6 and Supplementary Fig. 5), confirming these literature data. Although it was reported that freezing significantly increased *Firmicutes* to *Bacteroidetes* relative abundance ratio using quantitative PCR analysis³², our study identified a decrease in the *Firmicutes*/*Bacteroidetes* ratio, as well as a decrease in the ratio of *Lachnospiraceae* to *Bacteroidaceae*, two major families within *Firmicutes* and *Bacteroidetes* phyla.

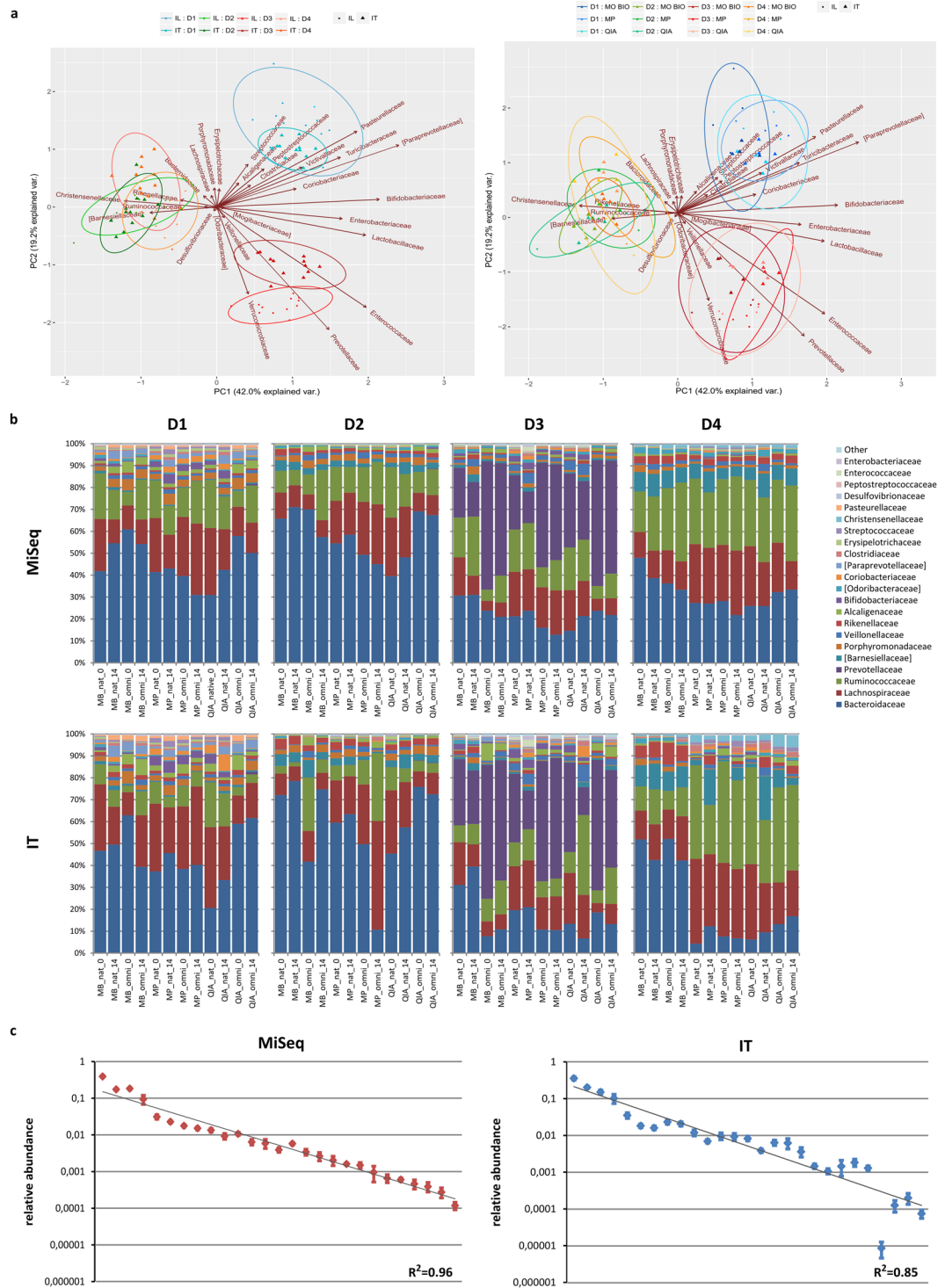


Figure 7. Donor specificity is maintained independent of experimental pipeline. **(a)** PCA bi-plots of clr transformed data on Illumina MiSeq and Ion Torrent PGM sequencing platforms. **(b)** Relative abundance profiles of individual subjects on each platform for top 20 families. **(c)** Relative abundances per platform represented as mean \pm SD, with R^2 values displayed from linear model fitted to log₁₀ values.

Also, we found a decline in *Proteobacteria* phylum in frozen faecal samples. These changes, however, were not found statistically relevant. The use of preservative buffers and nucleic acid stabilizing reagents, such as TE-buffer or RNAlater was found to negatively influence DNA yield and quality, alter microbial diversity and proportions of certain taxa^{12,33,34}. However, studies evaluating OMNIgene.GUT, a commercially available faeces collection/stabilization system, indicated this system was equivalent to, or superior in maintaining sample microbiota composition to freezing, resulting in very few alterations to fresh stool samples^{12,15}. We found that stabilizing the faecal

	Kruskal-Wallis				Wilcoxon					
	IL:IT	MOBIO:MP	MOBIO:QIA	MP:QIA	IL:IT	effect	MOBIO:other	effect	MP:other	effect
phylum										
<i>Actinobacteria</i>		< 0.001			0.04	0.336	0.004	0.602	0.001	0.621
<i>Firmicutes</i>	0.039	< 0.001		0.035					0.018	0.426
<i>Lentisphaerae</i>	0.009				0.037	-0.297				
<i>Proteobacteria</i>					0.011	0.434				
<i>Tenericutes</i>	0.044				0.044	0.299				
<i>Verrucomicrobia</i>	< 0.001				< 0.001	-0.616				
family										
[<i>Mogibacteriaceae</i>]		< 0.001		0.023			0.045	0.445	0.002	0.641
[<i>Odoribacteraceae</i>]	< 0.001				< 0.001	-0.61				
<i>Bacteroidaceae</i>	0.001									
<i>Bifidobacteriaceae</i>		0.003							0.016	0.348
<i>Clostridiaceae</i>		< 0.001							0.01	0.526
<i>Coriobacteriaceae</i>		< 0.001					< 0.001	0.733	0.001	0.666
<i>Desulfovibrionaceae</i>	< 0.001				0.015	-0.395				
<i>Erysipelotrichaceae</i>	< 0.001				0.002	-0.54				
<i>Lachnospiraceae</i>	0.019	< 0.001		0.023					0.001	0.635
<i>Peptostreptococcaceae</i>		< 0.001							0.001	0.559
<i>Porphyromonadaceae</i>	< 0.001				0.028	-0.424				
<i>Ruminococcaceae</i>	< 0.001				0.001	-0.557				
<i>Streptococcaceae</i>		< 0.001					< 0.001	0.755	0.001	0.674
<i>Veillonellaceae</i>	< 0.001				0.014	-0.413				
<i>Verrucomicrobiaceae</i>	< 0.001				< 0.001	-0.769				
<i>Victivallaceae</i>	0.001				0.007	-0.451				

Table 1. Taxon significance between platforms and kits. Only tests yielding Benjamini-Hochberg adjusted p-values less than 0.05 obtained using Kruskal Wallis test or Wilcoxon rank test, and the matching effect size are shown. Tests performed on centred log ratio normalized data per OTU. Effect size is calculated as the ratio of median difference between groups and largest median variation within groups, values $> \pm 0.5$ indicate relevance.

sample in OMNIgene.GUT tended to preserve *Proteobacteria*, a phylum which was depleted in frozen samples (Supplementary Fig. 6), as already observed¹². OMNIgene.GUT favours isolation of *Prevotellaceae* family, but with no statistical significance (Fig. 6). Although we did find a significant increase in *Sutterella* genus in samples collected using OMNIgene.GUT (Supplementary Table 5), also documented by Choo *et al.*¹², our findings suggest that microbiota composition in OMNIgene.GUT collected samples was comparable to freshly processed faeces.

The evaluated three DNA extraction kits successfully extracted bacterial DNA from faecal samples. MP kit was the most efficient, extracting approximately three times more DNA than MO BIO and QIA kits per mg faeces. The use of MP and QIA kits resulted in high quality of the extracted DNA, while MO BIO provided DNA of inferior quality, consistent with literature data^{18,35–38}. Furthermore, DNA extraction from faecal samples collected using OMNIgene.GUT system resulted in higher DNA yield irrespective of the extraction kit utilized, thus decreasing the amount of faecal material required, as previously reported¹⁵. MP kit was also associated with higher richness of isolated taxa as observed in alpha diversity analysis (observed OTUs, Chao1 and PD_whole_tree) (Supplementary Fig. 3). This is probably due to its effective G+ cell wall disruption, resulting in extracting significantly more G+ bacteria including *Bifidobacteriaceae* (and *Actinobacteria* phylum) than the other two kits (Table 1, Supplementary Table 4). Statistical analysis indicated differences between MP and MO BIO kits, with MO BIO depleting *Lachnospiraceae*, the most abundant G+ family, as well as other G+ organisms from *Firmicutes* and *Actinobacteria* phyla. The largest effect sizes were observed for *Coriobacteriaceae*, *Streptococcaceae* and *Lachnospiraceae*. The discrepancy in extracting efficiency of G+ organisms (illustrated in Supplementary Fig. 6) was previously suggested to relate to suboptimal lysing conditions in MO BIO extraction method³⁶. The additional bead beating step proposed for MO BIO protocol³⁸ was included in this study, but it did not improve the G+ recovery as was expected. When QIA was compared to other kits, Wilcoxon rank test on simulated data did not detect significant differences and only two significant families were identified using the Kruskal Wallis test (Table 1). This could be a consequence of intrinsic variability among QIA extracted samples with OMNIgene.GUT collection system (Fig. 6) having an influence on the G+/G- relative abundance ratio in these samples (Supplemental Fig. 6). Although the recently published study aiming to standardise extraction procedures³⁸ recommended QIA based protocol for faecal DNA extraction (protocol Q), the results of the MP protocol from the same study (protocol #14) demonstrated high DNA quantity and quality, as well as community diversity and G+ recovery, positioning this extraction method high on the performance scale.

The influence of using different NGS platforms, MiSeq and IT, on microbial community composition analyses was already compared using different experimental setups and conditions^{7,9,26,27}. While most of the published

work assessed the performance of NGS approaches using the same primer set(s) across platforms, this study focused on investigating the comparability of taxonomic composition obtained with supplier recommended sequencing protocols, including the 16S rRNA gene regions and suggested primers.

The performance of platforms varied with respect to annotation levels. MiSeq annotation slightly outperformed IT at lower taxonomy levels since higher proportions of the assigned OTUs were obtained at family, genus and species levels (Fig. 3). This could be attributed to several platform set-up related factors, including differences in sequencing depth, targeted 16S gene region, amplicon sequencing lengths and/or single vs. pair-end sequencing technology^{7,9,27}. Additionally, the mock community revealed that both sequencing technologies have the ability to detect all bacteria present in the mock sample with >99% of detected OTUs annotated correctly down to family level. In contrast to previous platform comparison using even and uneven mock community designed for that study⁹, we identified all 17 genera present in our mock community sample (HM-782D). The detected false positive genera comprised <1% of total OTU counts, resulting in overall substantial correspondence to theoretical composition unlike the abovementioned study. The different capacity of platforms to capture mock community diversity in these two studies could be related to the differences in community composition, tested 16S gene region and primer sets used. The relative abundances of taxa in our study were not fully reproduced and MiSeq in general had closer proportions and improved detection of *Actinobacteria* (*Actinomycetaceae* (*Actinomyces odontolyticus*) and *Propionibacteriaceae* (*Propionibacterium acnes*)), when compared to the expected values. This is in accordance with previous work on the same mock sample⁷ that also found marked differences with these organisms and attributed them to premature read truncation observed on the IT platform. In general, we found more deviations from theoretical abundance with IT than with MiSeq sequencing of mock community (Supplementary Fig. 2).

Since defined mock community lacks the complexity of the human faecal microbiota, further evaluation included comparison of human samples with respect to total number of commonly annotated taxa and detected diversity on two platforms. Alpha diversity indices demonstrated tendency of MiSeq technology to detect more organisms in the samples than IT but the phylogenetic distance of these taxa was similar (Fig. 4b). Additionally, the major driver of diversity between samples, presented as the beta diversity, was the platform used for the sequencing step (PCoA1: 22.7% variance), followed by the donor (PCoAs2–4 explaining 22.1%, 12.3% and 9.8% of variance), totalling 66.9% of total variance between samples. No major trends related to kit, storage or extraction method were observed.

The assessment of relative abundances did not reveal marked differences between platforms (Supplementary Fig. 4). However, the statistical analyses taking into consideration the compositional nature of the NGS data and using Kruskal Wallis test on clr transformed dataset demonstrated statistical differences at all taxonomic levels, showed numerous taxa in this study to be exposed to technology driven bias (Supplementary Tables 3 and 4). However, as the level of significance does not by itself predict the magnitude of the differences further evaluations were performed. Primarily, to ensure that false positives (type I error) are avoided and to account for the technical variation and imprecision intrinsic to NGS datasets, Monte-Carlo simulations from Dirichlet distribution were generated on the original data followed by the evaluation of the effect sizes and additional parametric and non-parametric statistics. Statistical significance (Kruskal Wallis) revealed a substantial number of taxa (ten on family level) apparently guiding the platform difference (Supplementary Tables 3 and 4). After accounting for technical variation (Wilcoxon) the number declined (eight families remained). Scrutinizing for the effect size and filtering out only the taxa with medium to large effect sizes³⁹ the number was further reduced (four families with effect size >0.5). We considered these taxa as the most relevant drivers of platform differences. The notable finding of this study, and not reported so far, was the increased capacity of MiSeq to detect *Verrucomicrobia* (*Akkermansia muciniphila* species was the only representative of this phylum) as well as *Faecalibacterium* genus with abundances being 10 and 5 fold higher than with IT, respectively. This finding may be of special interest to researchers of health promoting bacteria since both *Faecalibacterium prausnitzii*, a butyrate-producing species in the *Ruminococcaceae* family, and *Akkermansia muciniphila*, a mucus-colonizing and degrading bacteria, showed sufficient efficacies to be considered next generation probiotics⁴⁰. Other taxa of interest that were statistically significant and 2 fold more abundant on MiSeq were [*Odoribacteraceae*], *Erysipelotrichaceae* and *Ruminococcaceae* (Supplementary Table 3). Additionally, the property that also separated MiSeq from IT was its improved consistency in detecting organisms across samples (fitness to linear model) and lower error margins (Fig. 7c) which could potentially be attributed to the fact that MiSeq pair-end sequencing produces longer reads versus single-ended IT (mean length \pm SD - amplicon 451 ± 14 bp and read 213 ± 17 bp, respectively). Although previous reports suggested that the sequencing approach of using more than one primer pair covering more 16S rRNA gene regions should better represent the content of a complex bacterial sample^{9,41}, we failed to see this improvement with experimental setup here described. The higher overall richness in MiSeq in this study could be attributed to the higher average sequencing depth (2.3 fold in MiSeq) or to the lower required matching sequence for MiSeq (V3-V4 region) to assign taxonomy to an OTU, than for IT (V2-V4, V6-V9).

In conclusion, OMNIgene.GUT was found to enable reliable faecal sample collection, transport and storage (comparable to freshly collected material) with respect to DNA amount and quality, except when QIA kit was subsequently used for DNA extraction. Out of three kits used, MP provided most efficient DNA extraction resulting in significantly higher quantity and quality of the DNA, as well as higher richness of faecal microbiota content. Despite the fact that this study investigated optimized NGS platform technologies which differ in multiple parameters (16S rRNA gene region(s), primers, amplicon lengths, sequencing technology (single vs. pair-end)), when analysed using the same processing pipeline, both MiSeq and IT adequately reproduce donor faecal microbiota specificity. However, all the technology-related bias described here suggests that particular care should be given to appropriate methodology choices suitable for the microbial profiling in a specific study.

Materials and Methods

Study design. The outline of the study design is presented in Fig. 1. Subjects participating were healthy adults able to give their consent ($n = 4$, 2 males and 2 females, age range 33–42) who donated a single faeces sample that was collected in its fresh, native state (native) or using OMNIgene.GUT system (Omni). Aliquots of both initial samples were processed further immediately on day 0 using DNA isolation kits from MP Biomedicals (MP), Qiagen (QIA) and MO BIO (MO BIO). The remaining quantities were stored at -20°C and room temperature for native and Omni sample, respectively. DNA isolation was repeated at day 14 using all three kits, thus yielding 12 DNA samples per single faeces sample. All DNA samples were subjected to 16S rRNA gene amplicon sequencing on two NGS platforms – Illumina MiSeq and Ion Torrent PGM. A mock bacterial community sample, was used as a control sample of NGS technologies.

Microbial mock community. Genomic DNA from Microbial Mock Community B (Even, Low Concentration), v5.1 L, for 16S rRNA Gene Sequencing (#HM-782D) was acquired from BEI Resources. It is comprised of DNA from 20 bacterial strains containing equimolar ribosomal RNA operon counts (100,000 copies per organism per μL) and the complete list of organisms can be found in Supplementary Table 1.

Collection and storage. Informed consent was obtained from all participants and all procedures were in accordance with the approved study protocol (University of Zagreb School of Medicine Ethics Committee, case number 380-59-10106-14-55/149). A fraction of faeces sample (1–2 g) was collected fresh in its native state in a screw-cap sterile container. An aliquot was processed within a few hours of collection while the remaining aliquot was frozen at -20°C for 14 days and processed after thawing. The other fraction of the same faeces sample (400–500 mg) was collected in OMNIgene.GUT kit (DNA Genotek Inc, Ottawa, Canada), a system for self-collection and liquid stabilization of microbial DNA from faeces for gut microbiome profiling, according to manufacturer's protocol. Briefly, the faeces sample was thoroughly mixed with 2 mL of stabilization liquid in the provided tube. Again, two aliquots of this faeces sample were processed, one within a few hours of collection and the other at day 14, after two-week storage at room temperature.

DNA extraction. Three commercial faecal DNA extraction kits were evaluated in this study: Power Fecal DNA Isolation Kit (MO BIO, #12830-50), QIAamp Fast DNA Stool Mini Kit (QIAGEN, #51604) and Fast DNA SPIN Kit for Feces (MP Biomedicals, #116570200). The manufacturers' protocols were followed for each kit, and recommendations for application to hard-to-lyse organisms were incorporated (detailed protocols for each DNA extraction procedure are supplied in Supplementary Methods). All extractions were performed by the same experimenter. The sample size for DNA extraction from native faeces ranged from 50–250 mg, while for the OMNIgene.GUT system 250 μL was used. Prior to extraction, samples were homogenized by bead-beating⁴². Homogenization of samples was performed using Minilys homogenizer (Bertin Corp) in tubes prefilled with high quality beads which ensures disruption of the cell walls and release of the DNA molecules in solution. MP kit contains tubes with beads (Lysing matrix E) and for QIA and MO BIO kit homogenization was performed using Soil grinding kit SK38 (Bertin Pharma, #D34016), with 0.1 mm glass beads, 1.4 mm ceramic beads and 1 glass bead of 4.0 mm. After extraction, DNA integrity was assessed by 1% agarose gel electrophoresis.

DNA quantitation. The DNA quantity and quality measurements were performed on two spectrophotometric devices, Nanodrop 2000 and Qubit 3.0 (Thermo Fisher Scientific), based on the absorbance and fluorescence readouts, respectively. The A_{260}/A_{280} absorbance ratio determined with Nanodrop was used to estimate the quality and purity of the extracted DNA while concentration was measured with Qubit 3.0.

DNA concentration and absorbance ratio at A_{260}/A_{280} are reported as mean values \pm SEM (16 samples per group). P-values are reported after applying Bonferroni correction on unpaired two-sided t-test, $p < 0.001$ is reported as ***.

DNA sequencing and post-processing. Faecal bacterial communities were profiled by 16S rRNA amplicon sequencing using two NGS platforms: Illumina MiSeq (MiSeq) and Ion Torrent PGM (IT). Manufacturer recommended primers, reagents and protocols were applied.

For MiSeq sequencing Nextera XT DNA Sample Preparation Kit (Illumina, #FC-131-1096) was used for construction of 16S libraries. Hypervariable regions V3–V4 of 16S rRNA gene were amplified using primers with overhanging adapter sequences for compatibility with Illumina index and sequencing adapters needed for paired-end sequencing, resulting in a single amplicon with mean length of 464 bp⁴³. After PCR amplification and PCR product clean up with Agencourt AMPure XP (Beckman Coulter, #A63881), Illumina sequencing adapters and dual-index barcodes were added to the amplicon target with Nextera XT Index Kit (Illumina, #FC-131-1002). Sequencing was performed on MiSeq platform (Illumina) with MiSeq Reagent Kit v3 (Illumina, #MS-102-3003), according to the manufacturer's instructions.

For IT sequencing 16S libraries from faecal DNA samples were constructed by amplifying seven regions V2, V4, V8 and V3, V6–V7, V9 of bacterial 16S rRNA using primers from Ion 16STM Metagenomics Kit (Thermo Fisher Scientific, #A26216). Amplicons were further processed using manufacturer recommended protocols and ION PGMTM Template OT2 400 Kit (Thermo Fisher Scientific, #4479878). Final DNA library concentration was quantified using Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, #Q32850). Following library quantification the Ion OneTouchTM 2 (Thermo Fisher Scientific, #4474778) system was used to prepare template positive ion sphere particles containing the clonally amplified DNA libraries using the ION PGMTM Template OT2 400 Kit. Template positive sphere density was within the optimal range of 10–30%. Sequencing was performed on Ion Personal Genome Machine (PGM) System (Thermo Fisher Scientific) with Ion PGM Hi-Q Sequencing Kit (Thermo Fisher Scientific, #A25592) and Ion 316 Chip Kit v2 (Thermo Fisher Scientific, #4483188) according to the manufacturer's instructions.

Raw sequencing files from both platforms were processed using QIIME pipeline⁴⁴. For MiSeq, fastq files containing paired end reads were first merged, allowing overlap between mates ('-allow-outies' option), using FLASH⁴⁵. Then for both platforms, fastq files were trimmed, filtered by quality and chimera removed as described in the default QIIME pipeline. Operational Taxonomic Units (OTUs) were assigned using the vsearch⁴⁶ algorithm and PyNast alignment⁴⁷ against the GreenGenes database (version 13_8, May 2013). Cumulative OTU counts for each taxonomy level from phylum to genus were extracted into separate tables from the resulting assignment.

Subsequent processing and analysis was done for each taxonomy level separately according to the Gloor CoDa Microbiome tutorial⁴⁸, using ALDEx2 R package³⁰ with additions of multivariate generalized linear models for testing the significance of multiple factors and their interactions.

Data analysis

Alpha and beta diversity analyses. The compositional diversity within each sample was ascertained using several alpha diversity indices: observed OTUs, followed by the PD whole tree and Chao1, as implemented in the QIIME pipeline. Sequences were rarefied from 0 to 30 000 range. Beta diversity was reported using weighted UniFrac distance measure⁴⁹, and major contributions to the differences in beta diversity were presented as principle coordinate analysis, as implemented in the QIIME pipeline.

Precision analysis. To assess the reproducibility between relative taxa abundances across time points and collection/storage parameters within a DNA extraction kit, abundance values were plotted against reference point (native 0) abundance from the same kit. Linear regression analysis was used to plot the linear regression line and determine R^2 - the goodness of fit of the line to the data.

Aggregate relative abundances at different taxonomic ranks as well as precision analysis were displayed on raw counts.

OTU count normalisation and statistical analysis. The obtained sequencing data were evaluated as compositional data⁵⁰ and the statistical analyses were performed as previously described^{28–30}. Briefly, the approach incorporates Bayesian multiplicative replacement of count zeros and centred log-ratio (clr) OTU count data transformation³⁰. The centred log-ratio (clr) transformation²⁸ ensures the data are scalable and sub-compositionally coherent thus making them amenable for both univariate and multivariate analyses. The raw OTU counts on taxonomic levels from phylum to genus were transformed to compositional values using the ALDEx2 (ANOVA-like differential expression analysis) R package.

Principal component analysis (PCA) is shown on zero-imputed and clr transformed counts using 'prcomp' function in R.

Kruskal-Wallis multivariate test was performed to detect overall significance for each variable per OTU. P-values were considered significant if $p < 0.05$ after Benjamini-Hochberg (BH) correction.

The transformed data were further assessed for significant differences among the group of samples collected/processed using Wilcoxon rank test and Welch's t-test, as implemented in ALDEx2 pipeline. Tests were performed on 50 Monte Carlo simulated instances of Dirichlet-distributed original datasets to estimate per-feature technical variation within each sample. All p-values are reported after adjusting for multiple testing with BH correction. Pairwise comparisons of transformed data were performed for pairs of donors, time points, kits and faecal collection type for each taxonomic level (phylum to genus).

References

- Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N. Engl. J. Med.* **375**, 2369–2379 (2016).
- Matijašić, M. *et al.* Modulating Composition and Metabolic Activity of the Gut Microbiota in IBD Patients. *Int. J. Mol. Sci.* **17** (2016).
- Evans, J. M., Morris, L. S. & Marchesi, J. R. The gut microbiome: The role of a virtual organ in the endocrinology of the host. *J. Endocrinol.* **218** (2013).
- Human Microbiome Project Consortium, T. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
- Smith, M. B., Kelly, C. & Alm, E. J. How to regulate faecal transplants. *Nature* **506**, 290–1 (2014).
- Claesson, M. J. & O'Toole, P. W. Evaluating the latest high-throughput molecular techniques for the exploration of microbial gut communities. *Gut Microbes* **1**, 277–278 (2010).
- Salipante, S. J. *et al.* Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl. Environ. Microbiol.* **80**, 7583–7591 (2014).
- Sinclair, L., Osman, O. A., Bertilsson, S. & Eiler, A. Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform. *PLoS One* **10**, e0116955 (2015).
- D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**, 55 (2016).
- Lauber, C. L., Zhou, N., Gordon, J. I. & Knight, R. Effect of storage conditions on the assessment of bacterial community structure in soil and humann-associated samples. *FEMS Microbiol. Rev.* **307**, 80–86 (2010).
- Cardona, S. *et al.* Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol.* **12**, 158 (2012).
- Choo, J. M., Leong, L. E. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* **5** (2015).
- Mathay, C. *et al.* Method optimization for fecal sample collection and fecal DNA extraction. *Biopreserv. Biobank.* **13**, 79–93 (2015).
- Guo, Y. *et al.* Effect of short-term room temperature storage on the microbial community in infant fecal samples. *Sci. Rep.* **6**, 26648 (2016).
- Anderson, E. L. *et al.* A robust ambient temperature collection and stabilization strategy: Enabling worldwide functional studies of the human microbiome. *Sci. Rep.* **6**, 31731 (2016).
- Carroll, I. M., Ringel-Kulka, T., Siddle, J. P., Klaenhammer, T. R. & Ringel, Y. Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. *PLoS One* **7**, e46953 (2012).
- Fouhy, F. *et al.* The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. *PLoS One* **10**, e0119355 (2015).
- Nechvatal, J. M. *et al.* Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. *J. Microbiol. Methods* **72**, 124–132 (2008).

19. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* **2**, 3 (2012).
20. Sanschagrin, S. & Yergeau, E. Next-generation sequencing of 16S ribosomal RNA gene amplicons. *J. Vis. Exp.* 1–7, <https://doi.org/10.3791/51709> (2014).
21. Fadrosch, D. W. *et al.* An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* **2**, 6 (2014).
22. Milani, C. *et al.* Assessing the fecal microbiota: an optimized ion torrent 16S rRNA gene-based analysis protocol. *PLoS One* **8**, e68739 (2013).
23. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
24. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–9 (2012).
25. Lam, H. Y. K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2011).
26. Quail, M. *et al.* A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13**, 1 (2012).
27. Clooney, A. G. *et al.* Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLoS One* **11**, e0148028 (2016).
28. Gloor, G. B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high throughput sequencing data. *Can. J. Microbiol.* **703**, cjm-2015-0821 (2016).
29. Gloor, G. B., Macklaim, J. M., Vu, M. & Fernandes, A. D. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian J. Stat.* **45**, 73 (2016).
30. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).
31. Tedjo, D. I. *et al.* The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *PLoS One* **10**, e0126685 (2015).
32. Bahl, M. I., Bergström, A. & Licht, T. R. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *FEMS Microbiol. Lett.* **329**, 193–197 (2012).
33. Dominianni, C., Wu, J., Hayes, R. B. & Ahn, J. Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiol.* **14**, 103 (2014).
34. Gorzelak, M. A. *et al.* Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. *PLoS One* **10**, e0134802 (2015).
35. Ariefdjohan, M. W., Savaiano, D. A. & Nakatsu, C. H. Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. *Nutr. J.* **9**, 23 (2010).
36. Kennedy, N. A. *et al.* The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9**, e88982 (2014).
37. Wagner Mackenzie, B., Waite, D. W. & Taylor, M. W. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. *Front. Microbiol.* **6**, 130 (2015).
38. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.*, <https://doi.org/10.1038/nbt.3960> (2017).
39. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* (Routledge, 1988).
40. El Hage, R., Hernandez-Sanabria, E. & Van de Wiele, T. Emerging Trends in 'Smart Probiotics': Functional Consideration for the Development of Novel Health and Industrial Applications. *Front. Microbiol.* **8**, 1889 (2017).
41. Rintala, A. *et al.* Gut microbiota analysis results are highly dependent on the 16s rRNA gene target region, whereas the impact of DNA extraction is minor. *J. Biomol. Tech.* **28**, 19–30 (2017).
42. Santiago, A. *et al.* Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* **14**, 112 (2014).
43. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
44. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
45. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
46. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
47. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267 (2010).
48. Gloor G. B. ALDEx2 CoDa microbiome tutorial, https://github.com/ggloor/CoDa_microbiome_tutorial (2016).
49. Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* **73**, 1576–85 (2007).
50. Aitchison, J. *The Statistical Analysis of Compositional Data.* (Chapman and Hall, 1986).

Acknowledgements

Genomic DNA from Microbial Mock Community B (Even, Low Concentration), v5.1 L, for 16S rRNA Gene Sequencing, HM-782D was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project. The authors fully acknowledge the Croatian Science Foundation (Hrvatska zaklada za znanost, HRZZ) for the support of this work under the project MINUTE for IBD (HRZZ grant no IP-11-2013-5467) as well as for supporting Ma.P. under “Young Researchers’ Career Development Project – Training of Doctoral Students” (DR-6-2014). This work was also funded by the Scientific Center of Excellence for Reproductive and Regenerative Medicine (project “Reproductive and regenerative medicine - exploration of new platforms and potentials”, GA KK01.1.1.01.0008 funded by the EU through the ERDF). The authors are grateful to Vladimir Trkulja for critical reading of the manuscript and Kristian Vlahoviček for assistance in bioinformatics.

Author Contributions

H.C.P., Mi.P., A.B. designed research; Ma.P., I.L. performed lab experiments and sequencing; A.B. performed bioinformatics analyses; H.C.P., Mi.P., M.M., A.B. analysed and interpreted data, performed data visualisation; Ma.P., H.C.P., Mi.P., M.M., A.B. wrote manuscript text; D.V.B., Z.K. and D.V. supervised the project. All authors revised and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-23296-4>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018