Article

Exploring quantitative metagenomics studies using Oxford Nanopore sequencing: A computational and experimental protocol

Rohia Alili[1,2,5*], Eugeni Belda[3*#], Phuong Le[1], Thierry Wirth[4,5,] Jean-Daniel Zucker[1,6], Edi Prifti[1,6], Karine Clément[,1,2]

*These authors contributed equally

Institutional addresses

[1] Sorbonne Université, INSERM, Nutrition and obesities; systemic approaches (NutriOmics), Paris, France : rohia.alili@aphp.fr, phuongleee@gmail.com.

[2] Assistance Publique Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Nutrition department, CRNH Ile de France, Paris France : karine.clement@inserm.fr.

[3] Integrative Phenomics, Paris, France. e.belda@integrative-phenomics.com.

[4] Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, Université des Antilles, EPHE, Paris 75005, France. thierry.wirth@mnhn.fr.

[5] PSL University, EPHE, Paris 75014, France .

[6] IRD, Sorbonne Université, UMMISCO, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes, F-93143, Bondy, France : jdzucker@gmail.com, edi.prifti@ird.fr.

# Corresponding authors : Eugeni Belda, Integrative Phenomics, 8 Rue des Pirogues de Bercy, 75012 Paris, France.

Abstract. Background : The gut microbiome plays a major role in chronic diseases, of which several are characterized by an altered composition and diversity of bacterial communities. Large-scale sequencing projects allowed characterizing the perturbations of these communities. However, translating these discoveries into clinical applications remains a challenges. To facilitate routine implementation of microbiome profiling in clinical settings, portable, real-time, and low-cost sequencing technologies are needed. Results : Here, we propose a computational and experimental protocol for whole genome quantitative metagenomics studies of human gut microbiome with Oxford Nanopore sequencing technology (ONT) that could be applied to other microbial ecosystems. We developed a bioinformatic protocol to analyse ONT sequences taxonomically and functionally and optimized pre-analytic protocols including stool collection and DNA extraction methods to maximize read length. This is a critical parameter for the sequence alignment and classification. Our protocol was evaluated using simulations of metagenomic communities which reflect naturally occuring compositional variations. Next, we validated both protocols using stool samples from a bariatric surgery cohort, sequenced with ONT, Illumina and SOLiD technologies. Results revealed similar diversity and microbial composition profiles. Conclusion : This protocol can be implemented in the clinical or research setting, bringing rapid personalized whole genome profiling of target microbiome species.

Keywords: quantitative metagenomics, microbiome, obesity, gut microbiota, microbial DNA extraction, sequencing, Simulation, Oxford Nanopore Technologies, MinION.

1. Introduction

In recent years, there has been a burst in knowledge related to gut microbiota screening in chronic diseases. The increasing access to high throughput sequencing has led to the discovery of alterations in the composition of intestinal microbiota in many human disorders, including metabolic diseases. Currently, there is a real challenge to discover reproducible gut microbial signatures for diseases in order to develop generalizable diagnostic and prognostic tools, which makes their clinical use difficult.

In the field of metabolic diseases, microbial diversity is generally representative of microbiome and host health, as exemplified in previous studies such as MetaHit [1], HMP [2], Metacardis [2] and others covering severe obesity, bariatric surgery [3], diabetes, NAFLD/NASH [4], and cirrhosis [5] [6]. In mild [7] and severe obesity [8] for instance, we previously showed that reduced microbial richness linked to altered composition was found in 40% to 75% of the subjects and was associated with a more deleterious host phenotype. Even with these established signatures in metabolic diseases, the gut microbiome varies greatly in composition and abundance from one individual to another.

Presently, most microbiome research studies are carried out using 16S ribosomal RNA genes or whole genome shotgun sequencing (WGS), the latter requiring extensive computational ressources and pipelines. Moreover, not all medical and research centers are able to set up high-end shotgun sequencing platforms due to multiple constraints. As opposed to previously existing technologies, ONT proposes real-time sequence data generation with fewer resources and a small benchtop footprint. In the context of metagenomics, the long reads generated by ONT have lead to major improvements in the *de-novo* assembly of microbial genomes from metagenomic samples [19], [20]. It has been applied to target pathogen and viral profiling [21], [22], [23], [24] as well as the characterization of microbial communities in diverse environments from 16S data [25]. However, there is a need to define standardized wet-lab and bioinformatics protocols for the use of ONT in large-scale quantitative metagenomic studies given that most of the quantitative metagenomic bioinformatics pipelines are adapted to short reads [25].

In addition to biological variation, gut microbiome quantification is subject to technical variation along the pre-analytical process, including sample collection and extending to DNA extraction, library preparation and sequencing but also along the bioinformatics analytical protocols [9]. This is observed in the literature with frequently non reproducible results [10],[11],[12],[13], highlighting the need for technical standardization [14]. Even though progress has been made with the work of different international consortia [15]-[16] to standardize protocols, there is still a need for fast-track and affordable microbiome screening protocols in clinical settings. For example, among critical steps prior to sequencing is DNA extraction. Costea et al. [16] reported variability in microbial composition and diversity with different DNA extraction protocols. Extraction protocols, with or without bead-beating, increase the representation of gram-positive bacteria as is also the case for different DNA extraction kits - the richness is higher and reads are longer with the Qiagen compared to Magnapure kits [17]. Library preparation has also an impact on the relative abundances of taxonomic and functional microbial objects [10]. Finally, the bioinformatic pipelines can yield consequent variability in microbial ecosystem description [18].

Here, we have explored protocols with the quest to optimize ONT for microbiome analyses and have proposed a complete protocol, including wet-lab preparation (i.e. sample collection, DNA extraction, and library preparation) as well as data processing and analysis. In particular, we have set up a customized analytical pipeline to estimate microbial composition and diversity as well as to classify ONT reads using latest bacterial gene catalogs along with functional profiling. This protocol is open-access, allowing for replication, and implemention within world's medical or research centers (https://git.ummisco.fr/ebelda/nanopore).

2. Materials and methods

2.1. Study design : To determine the optimal parameters for quantitative estimation of microbiome ecosystems, we first optimized our bioinformatics pipeline based on controlled experiments of simulated data [26]. We simulated sequence data based on a set of known bacterial genomes as well as abundance distribution profiles similar to real metagenomics varied in terms of composition, richness and sequencing depth. The simulator took into account the particularity and biases of ONT sequences. Next, we built and adapted a bioinformatics pipeline, and we searched for the best hyperparameters to minimize the difference between the estimated quantified features (abundance, richness) and the real abundance used to parameterize the simulation **(Figure 1a)**.

In addition, we conducted multiple wet lab experiments to establish an optimized pre-analytical protocol, from stool collection and DNA extraction, fragmentation, to end-repair steps (**Figure 1b**). Finally, we validated our protocol and pipeline using human stool samples sequenced with different technologies (ONT, Illumina and SOLiD) (**Figure 1c**).

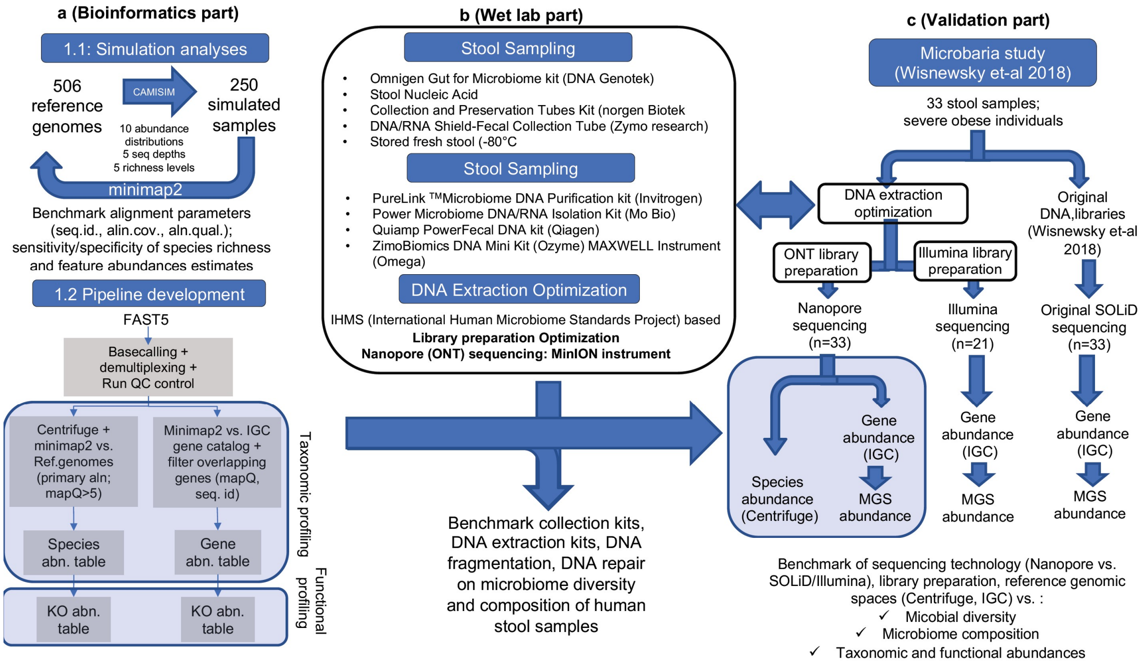

a (Bioinformatics part)
1.1: Simulation analyses
506 reference genomes
CAMISIM
250 simulated samples
10 abundance distributions
5 seq depths
5 richness levels
minimap2
Benchmark alignment parameters (seq.id., aln.cov., aln.qual.); sensitivity/specificity of species richness and feature abundances estimates
1.2 Pipeline development
FAST5
Basecalling + demultiplexing + Run QC control
Centrifuge + minimap2 vs. Ref.genomes (primary aln; mapQ>5)
Minimap2 vs. IGC gene catalog + filter overlapping genes (mapQ, seq. id)
Species abn. table
Gene abn. table
KO abn. table
KO abn. table
Taxonomic profiling
Functional profiling
b (Wet lab part)
Stool Sampling
Omnigen Gut for Microbiome kit (DNA Genotek)
Stool Nucleic Acid
Collection and Preservation Tubes Kit (norgen Biotek
DNA/RNA Shield-Fecal Collection Tube (Zymo research)
Stored fresh stool (-80°C
Stool Sampling
PureLink ™Microbiome DNA Purification kit (Invitrogen)
Power Microbiome DNA/RNA Isolation Kit (Mo Bio)
Quiamp PowerFecal DNA kit (Qiagen)
ZimoBiomics DNA Mini Kit (Ozyme) MAXWELL Instrument (Omega)
DNA Extraction Optimization
IHMS (International Human Microbiome Standards Project) based
Library preparation Optimization
Nanopore (ONT) sequencing: MinION instrument
Benchmark collection kits, DNA extraction kits, DNA fragmentation, DNA repair on microbiome diversity and composition of human stool samples
c (Validation part)
Microbaria study (Wisnewsky et-al 2018)
33 stool samples; severe obese individuals
DNA extraction optimization
Original DNA,libraries (Wisnewsky et-al 2018)
ONT library preparation
Illumina library preparation
Nanopore sequencing (n=33)
Illumina sequencing (n=21)
Original SOLiD sequencing (n=33)
Gene abundance (IGC)
Species abundance (Centrifuge)
MGS abundance
Gene abundance (IGC)
MGS abundance
Gene abundance (IGC)
MGS abundance
Benchmark of sequencing technology (Nanopore vs. SOLiD/Illumina), library preparation, reference genomic spaces (Centrifuge, IGC) vs. :
✓ Micobial diversity
✓ Microbiome composition
✓ Taxonomic and functional abundances

Figure 1: Summary of the workflow : (a) Simulated data processing. (b) Wet-lab optimization. (c) Summary of ONT sequencing comparison with Illumina and SOLiD technologies.

2.2. ONT microbiome-like simulated data : We set up a data simulation framework to estimate the performance of the quantification pipeline, while maximizing representation to real human gut microbial ecosystems. We used 506 reference genomes included in the construction of the IGC human gut gene catalog [28]. We simulated ten samples (M1:M10) whose abundances followed a Pareto distribution estimated using real metagenomic profiles of metagenomic species [8][27] computed on the same IGC catalog [28]. We included two important variables for the quantification of microbial ecosystems into the simulation: richness (number of present species) and the sequencing depth (i.e. the number of reads generated by the sequencing). We simulated the variation in richness from 50 to 450 species (R50:R450), as well as the sequencing depth ranging from 1x to 5x the complete coverage of the genomes present. In total, 250 samples were simulated using the CAMISIM software (option: Nanosim tool with default error parameters for the E.coli example) [29] (**Figure 1a**).

2.3. Bioinformatics workflow for taxonomic binning of ONT sequencing : The proposed bioinformatic workflow for quantitative metagenomic (QM) analyses from ONT shotgun sequencing starts with fast5 files generated by the MiniKNOW^{tm} software. The first step of the workflow consists of base calling and demultiplexing the fast5 files into fastq files. Here we used Albacore (v2.1.10) and Guppy (v2.1.3) ONT base callers from the community site available to ONT customers [30] together with custom R scripts that parses the the *sequence_summary.txt* files, generated during the base calling step, to generate different visualizations of the quality of the sequencing (active channels in the flow cell, distribution of active channels through time, yield in terms of reads of the run and the read length distribution).

The taxonomic binning of ONT reads is carried out using two different reference resources.

- *Centrifuge-based taxonomic binning :* Centrifuge [32] was used for the taxonomic binning of individual ONT reads using their comprehensive reference database of more than 8000 reference genomes from prokaryotes and viruses (including human reference genome). This step allows to exclude human sequence reads. To remove spurious taxonomic assignments, we additionally mapped read bins product of the initial Centrifuge classification against the corresponding reference genome from centrifuge database using Minimap2 with *map-ont* option optimized for ONT reads [33]. Based on simulation experiments results, only sequences with a minimum mapQ score of 5 were retained for subsequent analyses (see results). A species relative abundance table was generated by summing the counts of each taxonomic bin (NCBI taxonomy identifiers) from the filtered Centrifuge results. This relative abundance table was combined with the experiment metadata information and a reference taxonomic table reconstructed from Centrifuge NCBI taxonomy identifiers using the R package *taxize* v0.9.95 [34] using phyloseq v1.30.0[35], generating a phyloseq-class R object. This object can be used for microbial ecology analyses (rarefaction, alpha-diversity, beta-diversity, and differential abundance analysis).

- *IGC-based taxonomic binning* : A complementary approach consisted of quantifying the abundance of microbial genes. Here, ONT reads were aligned against the Integrated Gene Catalog of reference genes of the human gut microbiome (IGC) [28] using Minimap2 with *map-ont* option [33]. The alignment of long ONT reads over short or fragmented IGC genes provided two different types of multiple mappings (an ONT read mapped over several genes). First, a long ONT read could cover a genomic region harboring more than one gene, so different genes can be mapped over non-overlapping regions of an ONT read, providing a structural annotation of the corresponding DNA region. Second, multiple genes can be also mapped in overlapping regions of a read. These second multiple mappings were filtered out using *GenomicRanges* and *plyrRanges*R packages [36]·[37] allowing to retain the genes with the highest mapQ score and sequence identity across each alignment region. The raw gene abundance table was reconstructed by counting the number of times each gene was mapped by ONT reads. From this gene count table, the abundance of Metagenomic Species (MGS; co-abundant gene groups clustered from 1267 human gut metagenomes used to construct the IGC catalog[27]) were estimated as the mean value of the 50 most connected genes in each MGS as proposed in the original study[27].

2.4. Bioinformatics workflow for functional profiling of ONT sequencing : The final step of the bioinformatics workflow consisted in the quantification of KEGG orthology groups (KO groups) [68]. KO abundances were quantified from the results of both taxonomic binning approaches. From IGC abundance tables, KO abundances were quantified using available reference annotation from the IGC gene catalog as the sum of the individual abundances of genes annotated with different KO groups [38]. For taxonomic results produced from Centrifuge quantification, we retrieved the KO content of KEGG genomes from the KEGG API [39] for which species-level pan-genomes were reconstructed for all species-level bins based on NCBI taxonomy and matched with genomic sequences in the Centrifuge database. Based on this matching, the abundance of KO groups from Centrifuge results were computed as the sum of the abundances of the species containing these KO groups. The pan-genome strategy fits with the compressed nature of Centrifuge genomes at the species level, followed to reduce the size of the indexes and improve the overall performance of the classification process [32].

2.5. Study participants for wet-lab experiments : Stool samples used for wet-lab protocol optimization were collected from healthy French volunteers (n=15; men=8, BMI 18-25 kg/m^2) from the European "Metacardis" cohort [2]. For the comparison between sequencing technologies, we used 33 baseline samples from the Microbaria study [8], where the gut microbiome of subjects with severe obesity was characterized before and after bariatric surgery [8].

2.6. Sample collection and bacterial DNA extraction for pre-analytic protocol experiments : Fresh stools were collected with two different methods: 1) a dry spoon tube (SARSTED), which requires storage at -80°C and 2) a tube containing DNA/RNA stabilizing solution, which can be kept at room temperature, -20°C, or -80°C depending on storage duration. For the latter collection method, we tested three available commercial kits including 1)"DNA/RNA Shield-Fecal Collection Tube" (Zymo marketed by Ozyme), 2) "Stool Nucleic Acid Collection and Preservation Tubes" (Norgen Biotek) and 3) "Omnigen Gut for Microbiome" (DNA Genotek).

To extract bacterial DNA, we tested four different commercial kits using manual extraction protocols: 1) “PureLink ™ Microbiome DNA Purification Kit” (Invitrogen), 2) “Qiamp PowerFecal

DNA Kit" (Qiagen), 3) "ZimoBiomics DNA Mini Kit" (Ozyme) and 4) "Power Microbiome RNA/DNA isolation kit" (Mo Bio). We used the "MAXWELL Instrument" a robotic station from Promega that extracts DNA from 16 samples simultaneously. We, also tested automated extraction with two different kits: "Maxwell RSC Buffy Coat DNA Kit" (Promega 1) and "Maxwell RSC PureFood GMO and Authentication Kit" (Promega 2). Extracted stool DNA yield and quality was evaluated with a fluorometer (Qubit, Life Technologies) and Nanodrop (Thermo Scientific), respectively.

2.7. Optimization of DNA extraction, DNA fragmentation and End Repair : DNA extraction tests were performed from stool samples collected in dry tubes from three healthy subjects from the MetaCardis cohort (BMI<20kg/m^2) at three sampling times for subject 01. After collection, stool samples were aliquoted and immediately stored at -80°C. Each sample was extracted according to the protocols proposed by the manufacturer. After extraction, the samples were evaluated using the "Qubit" fluorometer to estimate the DNA yield obtained in ng/µl and using Nanodrop to evaluate DNA quality.

2.8. Library preparation and sequencing : We used 1.5 µg of DNA to perform the library construction. Extracted DNA was fragmented in g-tubes from Covaris, and DNA end repair was performed using the NEBNext FFPE Repair Mix from New England Biolabs (NEB). We used NEBext's NEBNext Ultra II End Repair / dA-Tailing Module (NEB) for the "end prep" step, 1D Native barcoding genomic DNA kit (ONT) and "NEB Blunt / TA Ligase Master Mix kit (NEB) for DNA multiplexing and adapters ligation. We used Agentcourt AMPure XP (Beckman Coulter) beads for DNA purification.

Whole genome metagenomic sequencing was performed with a ONT's MinION tool using flow cells on which 12 samples were simultaneously loaded per run. 33 samples from the Microbaria study were sequenced in parallel with ONT and Illumina Novaseq (2x150bp PE reads). Illumina sequences were processed following the same procedure as described in the original Microbaria study [8] in order to estimate microbial gene richness and the abundances of metagenomic species based on the 9.9-million-gene integrated gene catalog (IGC catalog)[28].

2.9. Statistical-ecological analyses : All statistical analyses were performed on R v.3.6. Wilcoxon rank-sum tests (for 2-level categorical variables) and Kruskal-Wallis tests (for categorical variables with more than two levels) were used to compare differences in microbial diversity between experimental conditions in different experiments. P-values<0.05 (alpha-level) were considered as significant. Spearman correlation tests were used to compare the abundance of taxonomic and functional features between sequencing technologies (SOLiD, Illumina, Nanopore) in Microbaria samples followed by correction for multiple comparison with Benjamini-Hochberg method. Adjusted P-values <0.05 were considered as significant.

Raw abundance tables product of ONT sequencing were rarefied to the minimum sequencing depth in each experiment before ecological analyses. Permutational analyses of variance (PERMANOVA) with the *adonis* function of vegan R package [44] were used to evaluate the impact of different covariates on microbiome composition in different experiments using Bray-Curtis beta-diversity dissimilarity matrix computed from genus-level abundance data. Alpha-diversity were estimated with phyloseq v1.30.0[35]

3. Results

3.1. Metagenome simulations identified key pipeline parameters for ONT microbiome quantification

The metagenome simulation approach allowed evaluating the impact of different steps and parameters in the bioinformatic pipeline. The evaluation accuracy consisted of comparing the estimated abundance of microbial features (i.e. species abundance) with the original values used to generate the sequences of over 100 million long reads for 250 simulated metagenomes (see methods; additional files Supplementary Table S1 and S2). These reads were aligned against the 506 reference genome catalog using minimap2 aligner with the *map-o*nt configuration, designed for optimal performance and accuracy with ONT sequencing data [33]. On average, 381.000 reads per sample (94%) were aligned against the reference genomes. The reads that could not be aligned were on average 2.5 times shorter in size (average read length=3168 bp, sd=24) compared to those that could (average read length =8064 bp, sd=8) (Figure 2a, Supplementary Table S3), suggesting that read length is a key parameter.

We evaluated the accuracy of the estimated species abundance and richness to the reference values used for the simulation based on filtering using primary alignment parameter (PA), defined as the best alignment of a single reads among all possible multiple alignments. We compared species abundance and richness of PA-filtered minimap2 results with those obtained without applying the PA-filtering step (raw abundances/richness)

For species richness, we observed that raw quantifications detected all reference species in simulated samples (100% samples with recall values equal to 1; Figure 2B). Quantifications based on PA-filtered reads detected all reference species (recall values equal to 1) in 95.8% of the simulated samples, with missing species observed in 21 simulations across different community compositions with low simulated sequencing depth (2 samples with 2x simulated depth of reference genomes; 19 with 1x sequencing depth; **Figure 2b**). However, both raw and PA-filtered reads overestimated the number of species especially in the low-richness community compositions. For all community compositions, taxonomic profiles from PA-filtered data reached higher precision values in species richness estimates compared to raw alignments (**Figure 2c**). Importantly, PCoA analyses using a Bray-Curtis beta-diversity dissimilarity matrix showed that PA-filtered samples were more similar to the corresponding reference distributions (R2 effect size dissimilarities=0.04, P-value=0.001; Permanova test) compared with raw alignment samples (R2 effect size dissimilarities=0.51, P-value=0.001; Permanova test) (**Figure 2d**). This suggests that the noise introduced by secondary alignments significantlty decreased the precision of the pipeline.

Figure 2

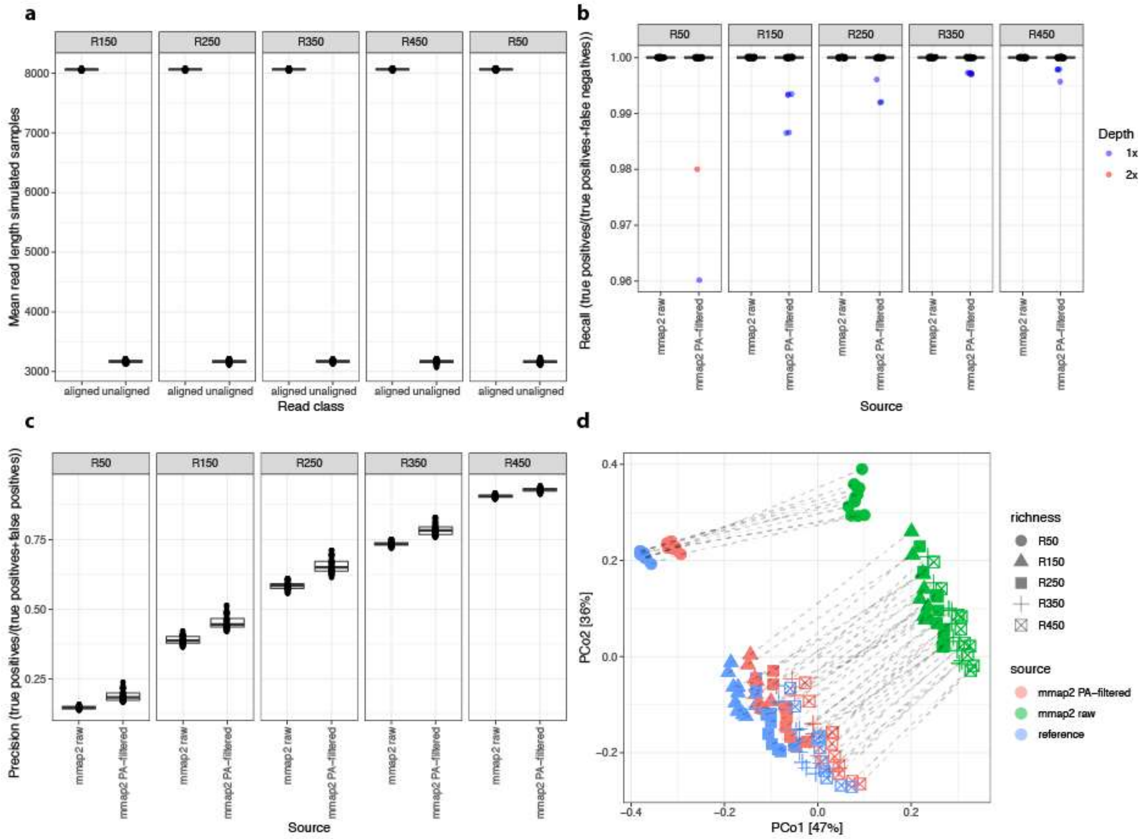


Figure 2: Metagenomic profiles from simulated samples between minimap2 results and minimap2 results filtered from secondary alignments. (a) Boxplots of mean lengths of ONT reads of 250 simulated samples (y axis) between those aligned and unaligned over the 506 reference genomes from minimap2 results (x-axis). (b) Boxplots of recall values of species richness estimates in 250 simulated samples (y-axis) between metagenomic profiles inferred from all minimap2 alignments (mmap2 raw) and from minimap2 primary alignments only (mmap2APfilt,

doi:10.20944/preprints202108.0104.v1

x-axis). For simulated samples not reaching the recall of 1 the sequencing depth is highlighted in different colours (c) Boxplots of precision values of species richness estimates in 250 simulated samples (y-axis) between metagenomic profiles inferred from all minimap2 alignments (mmap2 raw) and from minimap2 primary alignments only (mmap2APfilt, x-axis). (d) Principal Coordinates Analysis (PCoA) of metagenomic profiles from the reference and 250 simulated samples inferred from all minimap2 alignments (mmap2raw) and from minimap2 primary alignments only (mmap2APfilt, x-axis). Dashed lines connect points coming from the same sample (reference, simulated ones; 3 points per sample).

3.2. Alignment identity and alignment quality affects workflow precision

We next evaluated the impact of filtering read alignments at different thresholds of sequence identity on the accuracy of the estimated microbiome profiles. The recall values were close to 1 for species richness when filtering by identity levels up to 40%. This means that all reference species in each simulated sample were detected by the workflow. When progressively increasing identity levels from 50% to 90%, the fraction of reference species not detected notably decreased (Figure 3a). This resulted however in the increasing of the precision of the estimated richness as the filtering lowered the number of false positives (Figure 3b). At the level of similarities between relative abundance estimates, the Spearman Rho's of the correlation between the estimated species abundance and the reference values decreased as the alignment identity threshold increased across all different community compositions in similar way as recall values, showing that the loss of species with high stringent identity thresholds leads to decrease in the overall similarities of simulated relative abundance profiles with the reference abundances (Figure 3c). This has an impact on the overall microbiome composition similarities based on ordination framework. When considering overall microbial composition, the higher the stringency of the alignment identity the more dissimilar the metagenomic profiles were from the reference composition of simulated samples (Figure 3d), despite the presence of false positives. Overall, these results showed that common approaches to filter read alignments used in the context of second-generation NGS technologies (e.g. identity thresholds above 80%-90% sequence identity) were not directly applicable to high error-prone ONT sequencing data. Additional parameters were needed to be explored in order to improve the accuracy of the resulting metagenomic profiles.

Figure 3

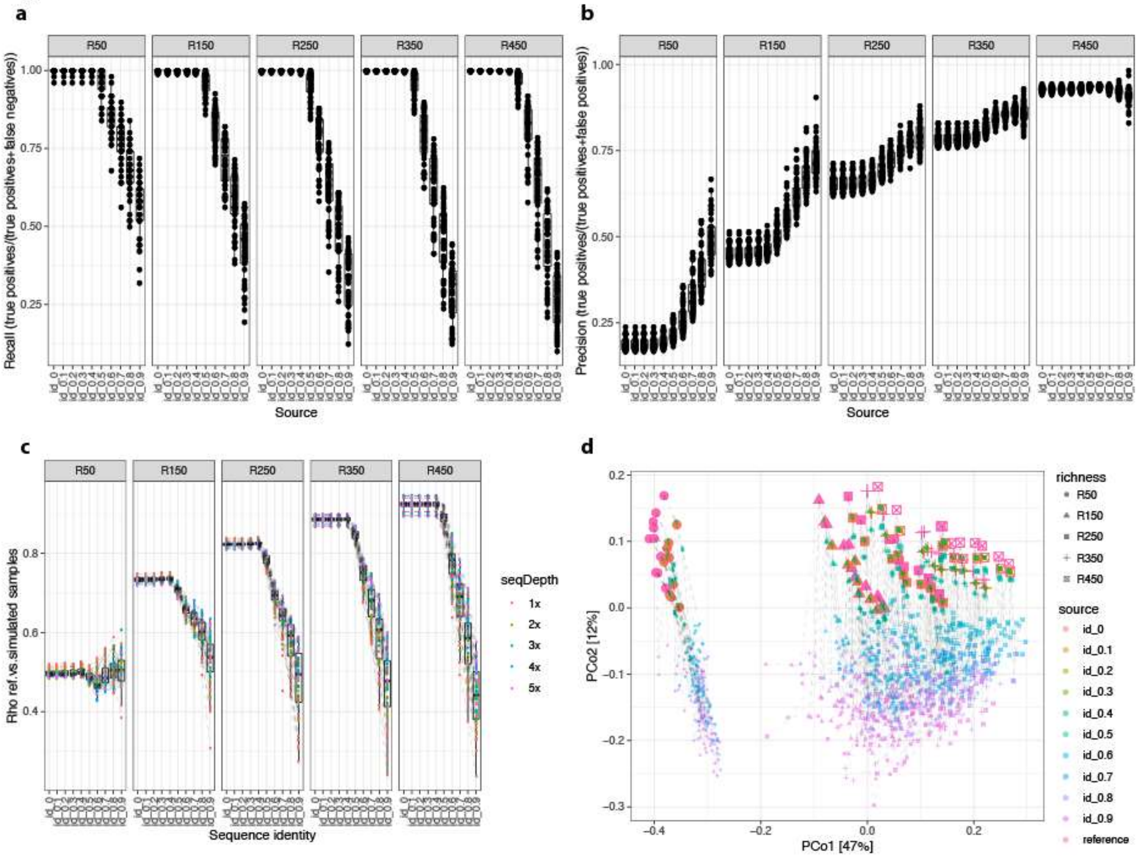


Figure 3: Impact of filtering minimap2 primary alignments of ONT reads at different thresholds of sequence identity. Boxplots of recall (a) and precision (b) values of species richness estimates in 250 simulated samples (y-axis) between metagenomic profiles inferred from primary alignments of ONT reads filtered by different thresholds of sequence identity (from 0 to 90%; x-axis) stratified by the number of species in reference metagenomic profiles. (c) Boxplots of Spearman's Rho coefficients in correlation analyses between taxonomic profiles of reference and simulated samples (y-axis) at different thresholds of sequence identity (from 0 to 90%; x-axis) stratified by the number of species in reference metagenomic profiles. Points are colored according with the sequencing depth of simulated samples. (d) PCoA of reference and simulated samples inferred from minimap2 primary alignments filtered by different thresholds of sequence identity (from 0 to 90%; x-axis). Dashed lines connect points coming from the same sample (reference, simulated ones; 11 points per sample) with different shapes assigned to samples from different reference species richness. Points corresponding to reference samples and simulated samples with no filtering by sequence identity (id_0) are highlighted with larger point sizes.

Thus, we next explored the mapping quality score (mapQ) as parameter for filtering ONT sequence alignments. MAPQ, as computed by minimap2, assigns high values to long reads and for which the scores assigned to secondary alignments were weak when compared with primary alignments[33]. 11 different mapQ thresholds (from 0 to 50 at steps of 5) were evaluated based on primary alignment of simulated datasets (Supplemental Fig 1). In terms of recall in species richness estimates, we observed similar decrease with the stringency of mapQ threshold as observed with alignment identity, although reaching overall higher recall values for the ensemble of simulated data (Figure 4a; mean recall ± standard deviation=0.826±0.053 vs 0.816±0.237 for mapQ filtering vs. alignment identity filtering respectively). Similar results were observed for precision, which increases with the stringency of the mapQ threshod reaching overall high values for the ensemble of simulated data (Figure 4b; mean precision ± standard deviation=0.95±0.089 vs 0.65±0.24 for mapQ filtering vs. alignment identity filtering respectively). When both filtering strategies were compared in terms of F1 scores, defined as the harmonic mean of precision and recall (high F1 scores being good trade-off between both metrics [66]) we observed that the filtering by mapQ produces significantly higher F1 scores than filtering by alignment identity under all community compositions regarding species richness (Figure 4c; P-value <0.05; Wilcoxon rank-sum test).

Finally, in terms of similarity between estimated species abundance and the reference values, we observed different results for different complexity of simulated communities. In low richness samples (R50, R150), the similarities increased with the mapQ threshold from 5 to 30, but this was not the case for more complex samples (R250-R450), where the similarities did not improve as the mapQ threshold was increased further than mapQ=5 (Figure 4d). On the contrary, the similarity of simulated abundances with the reference abundances significatively decreased with the strigency of mapQ filtering in simulated samples with 450 species (Supplementary Fig 2). Pairwise comparison of F1 scores in species richness estimates and Spearman's rhos of pairwise similarities with referece abundances across simulations with different mapQ filter threshold shows that both metrics were strongly associated (Supplemental Figure 3), which suggest that the overall accuracy of species richness estimates determines how well the simulated abundance profiles resembles the reference abundance data. These results show that for complex microbiome communities, like those of the human gut, the mapQ threshold of 5 gave the best results in terms of species richness estimations (based on F1 scores) and similarity regarding the estimated relative abundance profiles with the reference data (Supplemental Figure 3). This parameter however should be adapted depending on the estimated complexity of the target microbial community.

Figure 4

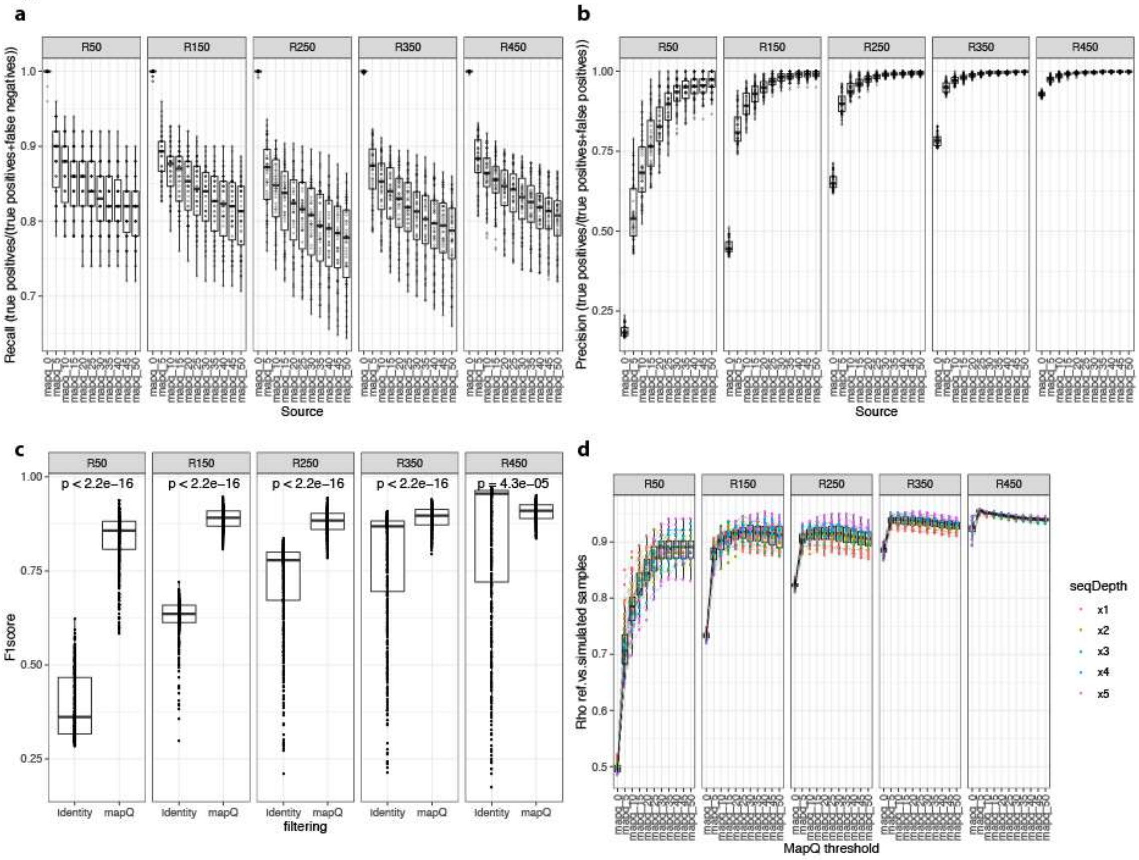


Figure 4: Impact of filtering minimap2 primary alignments of Nanopore reads at different thresholds of mapQ score. Boxplots of recall (a) and precision (b) values of species richness estimates in 250 simulated samples (y-axis) between metagenomic profiles inferred from primary alignments of ONT reads filtered by different thresholds of mapQ score (from 0 to 50; x-axis) stratified by the number of species in reference metagenomic profiles. (c) Boxplots of F1scores (harmonic mean of precision and recall) in species richness estimates between simulated datasets filtered by alignment identity and mapQ scores stratified by the number of species in reference metagenomic profiles. P-values product of Wilcoxon rank-sum tests are showed for each pairwise comparison. (d) Boxplots of Spearman's Rho coefficients in correlation analyses between taxonomic profiles of reference and simulated samples (y-axis) at different thresholds of sequence identity (from 0 to 90%; x-axis) stratified by the number of species in reference metagenomic profiles. Points are colored according with the sequencing depth of simulated samples.

3.3. Validation of the bioinformatic pipeline with ZymoBIOMICS mock community

The quantification of the ZymoBIOMICS mock community combining Centrifuge taxonomic binning and filtering by minimap2 alignment of read bins vs. the corresponding reference genomes with parameters derived from simulation experiments (primary alignments only, min. mapQ=5) reproduced the composition of the mock community with high accuracy and reducing the number of miss-assignments in comparison with classification based on Centrifuge only (Supplemental Fig 4). This also led to a higher overall similarity of microbiome composition (estimated as 1- Bray-Curtis beta diversity) with the reference mock community with the combination of Centrifuge and minimap2 filtering (0.91) than with raw Centrifuge results (0.88).

3.4. DNA extraction kits influenced read length distribution

When testing DNA extraction kits on dry spoon stool samples from healthy volunteers, all kits except the "ZimoBiomics DNA Mini Kit" (Ozyme) provided sufficient DNA quantity and quality for sequencing. Thus, we examined library preparation and sequencing all kits except the Ozyme kit. We observed a bimodal distribution of read lengths across the Invitrogen and Mo Bio DNA library preparation kits, with high proportion of long reads (>1.1Kb), whereas with the Promega and Qiagen kits the distribution was skewed towards smaller reads (Figure 5a). The Mo Bio kit led to the production of sequences with a mean of 8.5kb in size while the Invitrogen kit produced sequences up to 24 kb. The Qiagen kit and the two Promega (Promega 1 and Promega 2) kits yielded sequences up to 17kb but with a higher portion of short reads. The fraction of classified sequences was significantly higher for long reads (log2-length > 9.96; P-value=7.5e-^{09}; Wilcoxon signed-rank test), with on average 39% of long reads successfully classified after the two-step's procedure based on Centrifuge compared with the 24% for shorter reads (log2-length<9.96 (Figure 5d)), confirming initial observations with simulated data about the importance of this parameter in the taxonomic classification of ONT reads. The examination of alpha diversity showed also significant differences by DNA extraction kit, with high diversity levels in Invitrogen samples (Figure 5e, P-value=0.0061, Kruskal-Wallis test; P-value=9.4x10-4 Invitrogen vs. Promega 1 and Promega 2 kits; Post-hoc pairwise Dunn's test). Finally, we observed that the differences between microbiome composition of the different replicates were mainly explained by the collection day (R2=0.45, P-value=0.001), followed by sample donor (R2=0.03, P-value=0.001) and DNA extraction kit (R2=0.02; P-value=0.001) (Figure 5h; Permanova test, marginal effects on multivariate model with collection day, DNA extraction kit, sample donor, DNA fragmentation and DNA end repair). Based on these observations, the Invitrogen kit was selected as the preferred extraction kit for sequencing.

3.5. DNA fragmentation and end-repair

The first step in ONT's library preparation protocol DNA fragmentation to generate 8kb fragments [45]. Using 3 different samples from one subject, DNA fragmentation had no effect on read length distribution (Figure 5b), species richness (Figure 5f; P-value>0.05, Wilcoxon rank-sum test), or microbiome composition based on PCoA ordination (Figure 5i; R2=0.001, P-value=0.949, Permanova test). Therefore, we decided to exclude this step from our experimentation framework.

The ONT DNA preparation protocol recommends *DNA-end repair*. We evaluated the effect of DNA-end repair on read length and microbial diversity by extracting DNA from stools of the same three subjects using the Invitrogen kit and excluding the DNA fragmentation step. We found similar profiles of read length distributions between samples with end-repair and no end-repair steps (Figure 5c), no significant impact on species richness (Figure 5g; P-value >0.05, Wilcoxon rank-sum test) and no significant impact on microbiome composition (Figure 5J; R2=0.001, P-value=0.877, Permanova test). This step was then omitted from our proposed protocol.

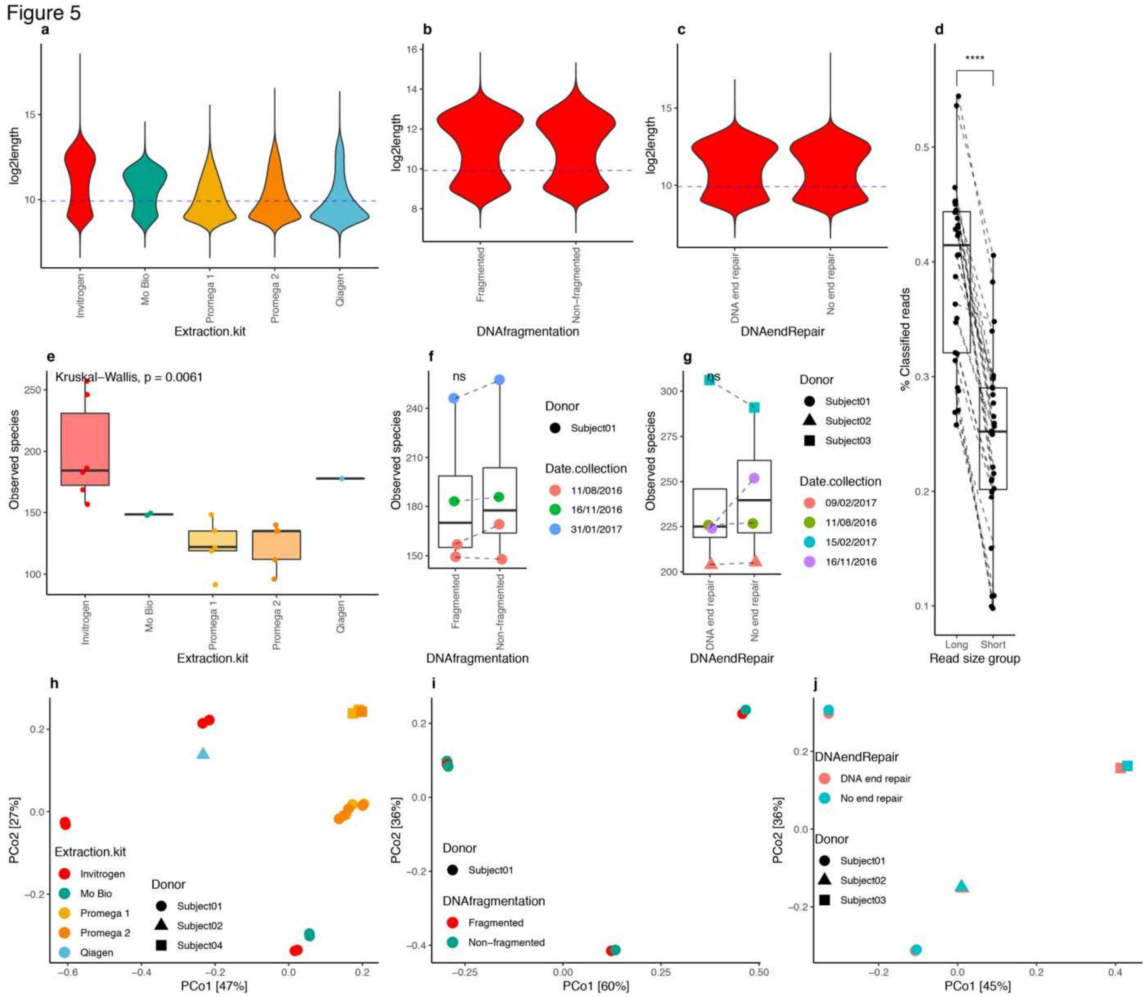


Figure 5: DNA extraction kits, fragmentation and end repair impact over human stool metagenomic composition from ONT sequencing data. Read length distributions of ONT reads across different DNA extraction kits (a, n=29) and between DNA fragmentation (b, n=6 paired samples Fragmented/non-fragmented) and DNA end repair (c, n=6 paired samples end vs. no end repair) steps for Invitrogen samples. Blue dashed lines correspond to the median value of log2-transformed read lengths used to stratify reads as long or short. (d) Differences between the fraction of classified reads by Centrifuge approach between long and short reads for 29 samples in panel a. (e) Differences in microbial diversity (observed species) between extraction kits (n=29). (f) Differences in microbial diversity (Observed species) by DNA fragmentation (n=4 paired samples). (g) Differences in microbial diversity (Observed Species) by DNA end-repair step (n=4 paired samples). (h) PCoA ordination of 29 samples in panel A colored by extraction kit. (i) PCoA ordination of 8 samples in panel f. (j) PCoA of 8 samples of panel g colored by DNA end-repair step. ns (panel f, g) =Non-significant differences in paired Wilcoxon rank-sum tests. ***=P-value<0.0001 in paired Wilcoxon rank-sum test.

3.6. Optimized DNA extraction protocol improved ONT read length and microbial diversity estimation

Based on the simulated and real data, read length was an important parameter to maximize the efficiency of taxonomic classification of ONT reads. Consequently, we aimed at increasing the read length by optimizing DNA extraction and library preparation using the Invitrogen kit. We followed recommendations from the IHMS consortium [16](Figure 6a**)** but to increase the proportion of long reads, we improved the sequencing library preparation protocol by modifying two main steps. The first one was the "End-prep" step which prepares the binding of the adapter to the DNA after two incubation periods. We used the "NEBNext Ultra II End Repair /dA-Tailing Module" from New England Biolabs (NEB) company. In the ONT protocol, “End-prep” reaction incubating is recommended for 5 minutes at 20°C followed by 5 minutes at 65°C. However, the NEB kit recommends a first incubation at 20°C for 30 minutes followed by a second incubation at 65°C for 30 minutes. Given the lack of effects of the ONT end prep protocol, we attempted end repair using NEB kit and recommended protocol (Figure 6b**).**

In the library preparation step, DNA was purified by using "Agencourt AMPure XP" beads (Beckman Coulter), which use Solid-phase reversible immobilization (SPRI) paramagnetic bead technology that selectively binds nucleic acids according to type and size. Agencourt AMPure XP utilizes an optimized buffer, Polyethylene glycol (PEG), to selectively bind DNA fragments. The size of the fragments eluted from the beads is determined by PEG concentration. For example, if 50µl of beads are added to a 50µl DNA sample, a SPRI/DNA ratio of 1 is obtained. When this ratio was changed, the length of the fragments binding and/or remaining in the solution also changed. The SPRI/DNA ratio was disproportionately associated with the DNA fragment size, which is due to fragment size affecting the total charge carried by the molecule. Thus, long DNA fragments would have a greater proportion of negative charges, which promotes their electrostatic interaction with the beads and allows a priority link to the carboxyl molecules. The ONT protocol was developed based on DNA fragmentation of 8Kb sequence length, and the SPRI/DNA ratio must be equal to 1. In order to promote the selection of larger DNA fragments by paramagnetic beads, we reduced the SPRI/DNA ratio to 0.4 **(**Figure 6b**)**. The chosen ratio was based on the SPRI technology documentation [46] and ONT users' recommendations from the "Community" forum [30].

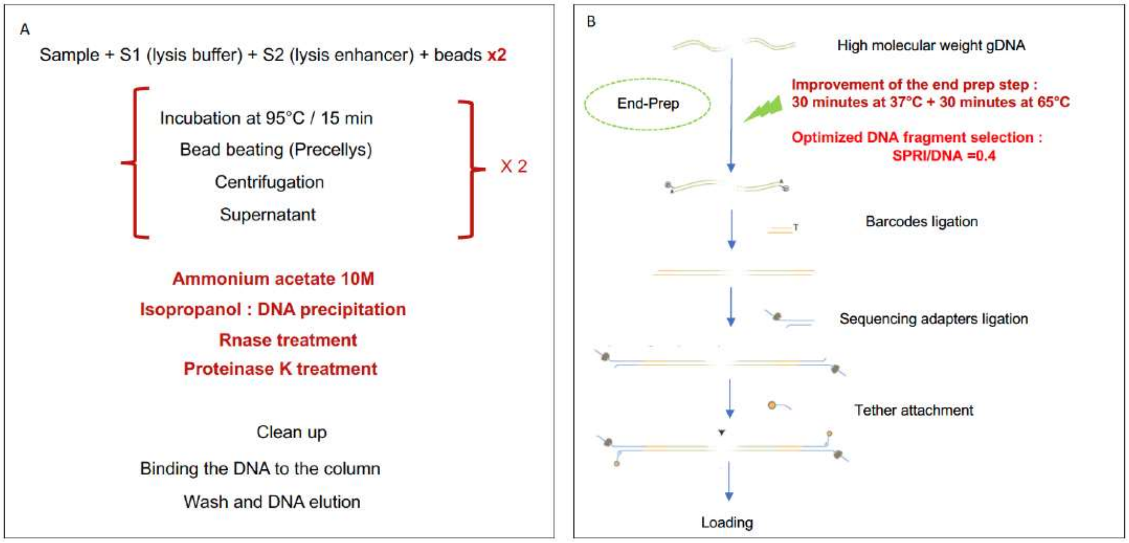

A
Sample + S1 (lysis buffer) + S2 (lysis enhancer) + beads x2
Incubation at 95°C / 15 min
Bead beating (Precellys)
Centrifugation
Supernatant
X 2
Ammonium acetate 10M
Isopropanol : DNA precipitation
Rnase treatment
Proteinase K treatment
Clean up
Binding the DNA to the column
Wash and DNA elution
B
High molecular weight gDNA
End-Prep
Improvement of the end prep step :
30 minutes at 37°C + 30 minutes at 65°C
Optimized DNA fragment selection :
SPRI/DNA =0.4
Barcodes ligation
Sequencing adapters ligation
Tether attachment
Loading

Figure 6: Optimization of DNA extraction and library preparation protocols. (a) Steps of the bacterial DNA extraction protocol. In black, the steps include in the protocol of the Invitrogen kit, in red the improvement steps recommended by the IHMS consortium. (b) Improvement of library preparation by application of NEB recommendation and decrease the SPRI/DNA ratio.

Thus, we performed two modifications (End-prep and DNA purification) with the Invitrogen extraction protocol, referred to as "Optimized Invitrogen". This optimization step was performed for six samples from one healthy subject (from the MetaCardis study), collected at six time points. Each sample was extracted using the standard "Invitrogen" protocol and with the "optimized" protocol. DNA yields extracted from this optimized protocol were five-time greater than the ones obtained with the standard kit (55 ng vs. 300 ng, P-value < 0.0001). The ratio of the absorbance at 260/230 was higher with the optimized protocol 2.11 vs. 1.38 respectively (P-value = 0.0007) and the absorbance ratio at 260/280 increased significantly 1.89 vs 1.73 in the non optimized one respectively (P-value = 0.0046) (Supplementary Table S5).

The read length was also improved (Figure 7a). The standard Invitrogen protocol produced two populations of reads with average read lengths of 500 and 6,000 bp while the optimized protocol produced a single read population with an average read length of 6,000 bp. According with this, we observed a significant increase in the fraction of classified reads in comparison with the fraction obtained with the initial protocol recommended by Invitrogen (Figure 7b; 29.72% with optimized protocol vs. 23.92% with original protocol; P-value=0.031, Wilcoxon signed-rank test). We observed an increased microbial diversity (observed species) in four of the 6 samples with the optimized protocol even if this variation remains insignificant (P-value=0.31; Wilcoxon signed-rank test) (Figure 7c). Finally, PCoA (Figure 7d) showed that differences in microbiome composition across samples are explained mainly by the collection date of the samples (R2=0.89, P-value=0.001, Permanova test), with no significant effect of extraction kit on overall microbiome composition (R2=0.026, P-value=0.906, Permanova test). Altogether, the optimized DNA extraction protocol exhibited a better DNA yield and purity, longer sequences than the usual protocol, leading to a significant yield improvement (fraction of classified reads) of the taxonomic binning with no impact on the overall microbiome composition of the samples.

Figure 7

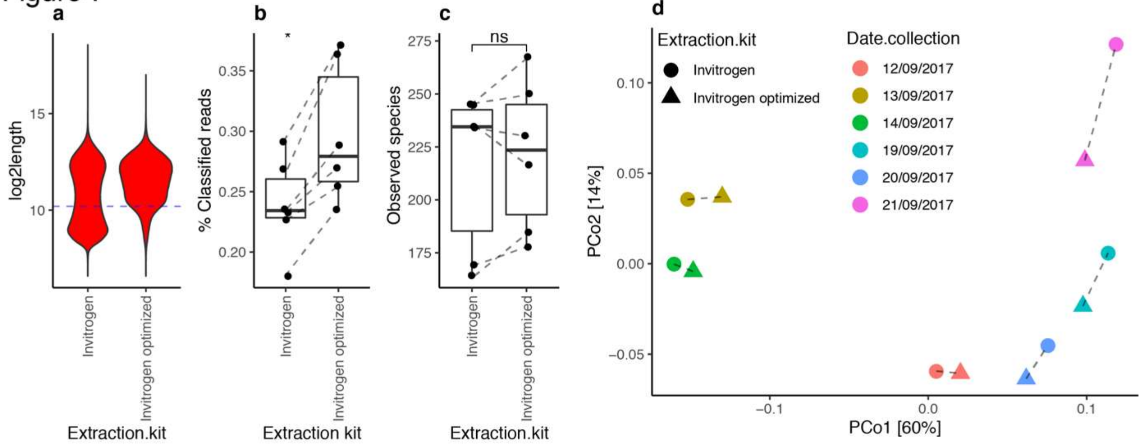


Figure 7: Impact of Invitrogen optimized protocol over human stool metagenomic composition from Nanopore sequencing data. (a) Read length distributions of ONT reads in n=6 samples extracted with original (Invitrogen) and optimized (Invitrogen optimized) protocols. Blue dashed lines correspond to the median value of log2-transformed read lengths used to stratify reads as long or short. (b) Differences in the fraction of classified reads between protocols (n=6 paired samples; *=P-value<0.05, Paired Wilcoxon rank-sum test). (c) Differences in microbial diversity between protocols (n=6 paired samples; ns=P-value>0.05, Paired Wilcoxon rank-sum test). (d) PCoA ordination from genus-level beta-diversity matrix (Bray-Curtis) of 12 samples extracted with Invitrogen optimized kit and original Invitrogen kit. Dashed lines connect samples coming from the same fecal stool sample collected at different dates.

3.7. Impact of stool sampling and storage on sequence length and diversity

Subjects' stool samples were initially collected in a dry spoon tube and rapidly frozen at -80°C to ensure the stability of the bacterial DNA. However, an increasing number of sampling systems contain a solution that can stabilize bacterial DNA at room temperature for periods ranging from 60 days (DNA Genotek) up to 2 years (NORGEN Biotek and OZYME). We evaluated the effects of room temperature (RT) stabilized samples on bacterial DNA extraction, library preparation and sequencing. We prepared six DNA libraries from stools of 12 healthy subjects collected by different protocols in three different stabilizing kits : "Omnigen Gut for Microbiome" (DNA Genotek), "Stool Nucleic Acid Collection and Preservation Tubes"(Norgen Biotek) and "DNA/RNA Shield-Fecal Collection Tube" (Ozyme). Regarding read lengths, we observed similar unimodal distribution towards long reads across all experiments (Figure 8a). We did not observe significant differences between the three collection kits in the fraction of classified reads (Figure 8b; P-value=0.38 in -80° group; P-value=0.28 in RT group). Also, we observed no significant differences in species richness (Figure 8c; P-value=0.71 in -80° group, P-value=0.41 in RT group; Kruskal-Wallis test), although we could notice a tendency with Norgen and Omnigen kits to decrease microbial diversity at room temperature in comparison with -80°C storage (Figure 8c). In contrast, we observed significant variations in microbial diversity by donor (Figure 8d; P-value=0.0017, Kruskal-Wallis test), being the variable with the highest impact on microbiome composition by Permanova analyses (R2=0.82; P-value=0.001) in comparison with collection kit (R2=0.11, P-value=0.009) and storage conditions (R2=0.07; P-value=0.029) (Figure 8e & 8f). Thus, sampling methods at room temperature with DNA stabilization performed similarly to snap frozen samples and the choice of the sampling kit might depend on practical feasibility sampling for the subject as well as kit price. We chose the Ozyme kit for its practicality for users to collect stool samples and its relative costs.

Figure 8

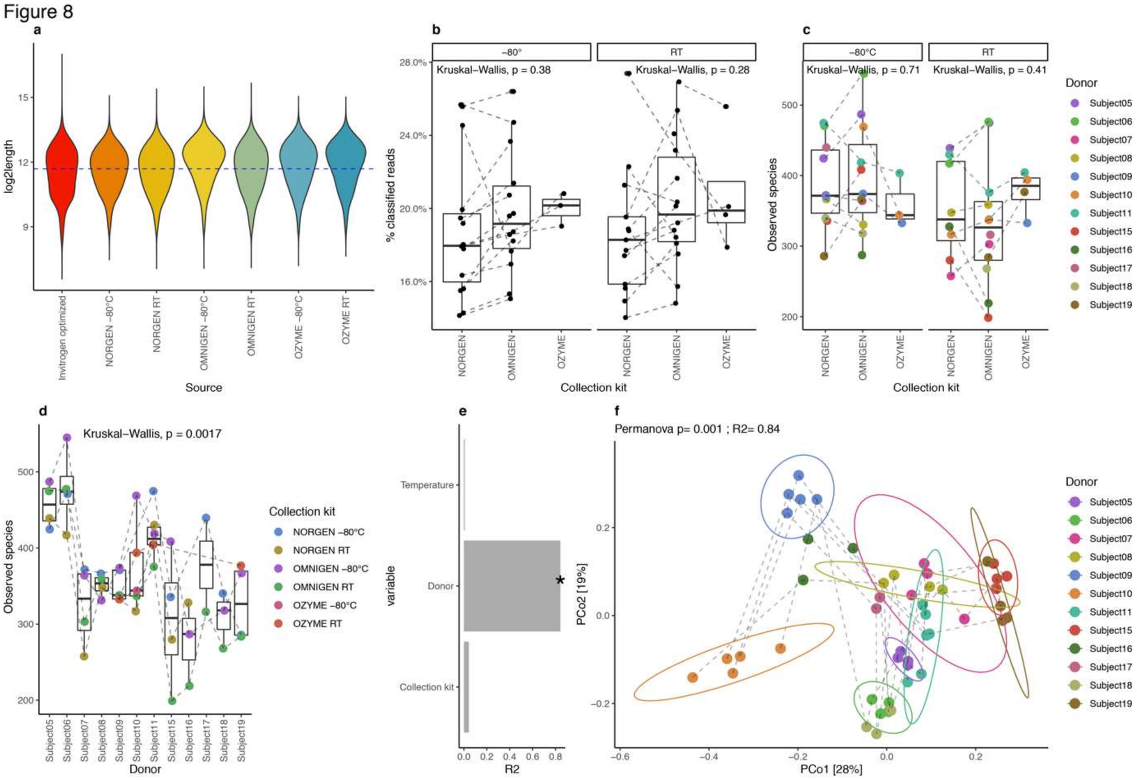


Figure 8: Impact of collection kits and storage conditions on metagenomic human stool composition from Nanopore sequencing data. (a) log2-read length distribution of ONT reads across collection kits and temperature storage conditions. For comparison, the log2-read length distribution of initial Invitrogen optimized reads is included. Dashed blue line represents the median log2-read length from the entire dataset. (b) Difference in the fraction of classified reads by Centrifuge strategy between collection kits stratified by storage condition. (c) Differences in microbial diversity between collection kits stratified by storage condition. (c) Differences in microbial diversity (Observed species) between donors of fecal samples in this experiment. (e) Impact of difference experimental variables (donor, temperature, collection kit) over microbiome composition of studied samples. The barplot represents the effect sizes (R2) from Permanova tests of variables in Y-axis over a beta-diversity distance matrix computed from Centrifuge-based genus abundance data (*=P-value<0.05, Permanova test). (F) PCoA ordination of samples from collection kits experiments coloured by donor. Dashed lines connect samples collected with same collection kits (Omnigen, Ozyme, Norgen).

3.8. Optimized ONT protocol compared to IlluminaSOLiD sequencing

We compared ONT obtained QM profiles with those generated with other sequencing technologies from the Microbaria study[8]. We selected 33 pre-surgery samples covering the extremes of microbiome diversity defined as microbial gene richness (13 samples from individuals with High Gene Count (HGC) and 20 samples from individuals with Low Gene Count (LGC)). ONT abundance profiles were generated using the two bioinformatics workflows described in the methods section, based on Centrifuge and mapping over the IGC catalog. DNA from 21 of the 33 samples were also extracted with the optimized Invitrogen protocol and sequenced using Illumina technology. Quantitative metagenomic profiles from Illumina samples were generated by mapping reads against the IGC gene catalog as described in [8].

First, we compared the estimates of microbial diversity from ONT (gene richness from IGC mapping and Observed Species from Centrifuge classification) and Illumina sequencing (Gene richness from IGC mapping) with the gene richness inferred from the original SOLiD sequencing of these samples. SOLiD sequencing generated 4.38e+07 single reads of 35 bases (sd=1.86e+07) per sample on average, representing 1.53e+09 base pairs overall (sd=6.52e+08). With the ONT, we generated an average of 1.53e+05 reads per sample between 200bp and 24 Kb (sd=6.07e+04), representing 4.19e+08 bp overall (sd=1.607e+08).

We observed significant positive associations between Centrifuge-based diversity estimates from ONT sequencing with the reference gene richness from SOLiD sequencing based on IGC gene catalog (Spearman Rho=0.59, P-value=3e-04 for Observed Species based on Centrifuge results; Figure 9a). These similarities increased with the use of IGC as reference database for diversity estimations (Spearman Rho=0.74, P-value=2e-06 for Gene Richness based on ONT read mapping over IGC catalog; Figure 9b). However, the similarity was higher with gene richness estimates based on Illumina sequencing despite differences in library preparation (Spearman Rho=0.86, P-value<2.2e-16; Figure 9c). When we integrated the scaled diversity profiles (dividing each diversity estimate by the maximum value in each sequencing source for these 33 samples (ranges from 0 to 1)) and order them based on the reference gene richness in the original publication[8], we observed that DNA extraction had an impact on diversity. Both ONT and Illumina sequencing using the same DNA extraction method showed similar variations in microbial diversity estimates compared to the reference SOLiD data (Figure 9d). This included a switch in the sample showing the highest diversity (i.e. MB12 sample with Illumina and ONT sequencing; MB21 sample with SOLiD sequencing; Figure 9d).

Figure 9

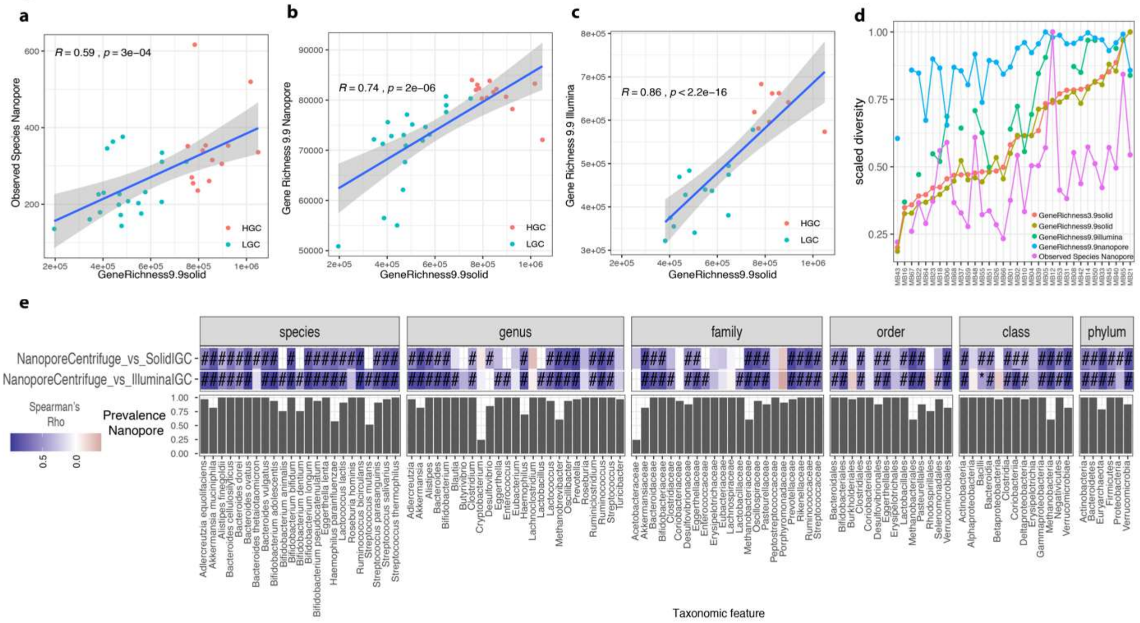


Figure 9: Comparison of quantitative metagenomic profiles of Microbaria samples between sequencing technologies. Correlation between gene richness from SOLiD sequencing (x-axis) and Observed Species inferred from Nanopore(ONT) sequencing data using Centrifuge approach (a, n=33), gene richness inferred from ONT sequencing data (b, n=33) and gene richness inferred from Illumina sequencing data (c, n=21). The strength of the similarities was evaluated with Spearman correlation test (Spearman's Rho and P-value included in the scatter plots). (d) Lineplots representing the scaled diversity (from zero to 1) of Microbaria samples from different diversity metrics based on SOLiD, ONT and Illumina sequencing data. Samples in x-axis are ordered based on the scaled diversity of the gene richness from the original Microbaria study (GeneRichness3.9SOLiD). (e) Heatmap of Spearman's Rho representing similarities in abundance vectors of taxonomic features in x-axis between ONT quantifications based on Centrifuge data and Illumina and SOLiD quantifications based on metagenomic species of the IGC gene catalog (y-axis; #=P-valueadj<0.05, BH method; *=P-value<0.05). On the bottom of the heatmap is represented the prevalence of taxonomic features in x-axis based on ONT sequencing data.

The high similarity between ONT and Illumina datasets was confirmed in an ordination framework where we integrated the genus-level abundance profiles from IGC quantification with the three sequencing technologies (ONT, Illumina, SOLiD), where we observed that samples product of the same DNA extraction method (Illumina and ONT) are closer in a PCoA ordination (Supplemental Figure S5A) and in hierarchical clustering analyses (Supplemental Figure S5B).

3.9. ONT pipeline detects target species and functional profiles

Regarding taxonomic feature quantification, we found a good agreement between ONT sequencing and both Illumina and SOLiD sequencing data. Based on Centrifuge, we observed a positive correlation of relative abundances in 91 of the 95 common taxonomic features with SOLiD quantifications (96%; mean Spearman Rho ± standard deviation=0.68±0.15 (species), 0.58±0.32 (genus), 0.53±0.3 (family), 0.5±0.26 (order), 0.51±0.24 (class), 0.62±0.18 (phylum)). 72 of these features (76%) were significantly associated (FDR<0.05, Spearman correlations). Similarly, we observed a positive correlation in 94 of the 101 common taxonomic features with Illumina quantification (93%; mean Spearman Rho ± standard deviation=0.74±0.22 (species), 0.62±0.26 (genus), 0.63±0.27 (family), 0.63±0.24 (order), 0.66±0.2 (class), 0.68±0.22(phylum)). 78 of these features (77 %) were significantly associated (FDR<0.05, Spearman correlations) (Figure 9e).

Using quantification of Metagenomic Species (MGS) based on ONT mapping over IGC gene catalog, the relative abundances of 137 common taxonomic features with SOLiD quantifications were positively associated (mean Spearman rho ± standard deviation=0.68±0.12 (species), 0.64±0.16 (genus), 0.63±0.17 (family), 0.58±0.2 (order), 0.59±0.2 (class), 0.58±0.16 (phylum)), 128 of which (93%) were significantly associated (FDR<0.05, Spearman correlations). Similar comparison with MGS relative abundances product of Illumina sequencing gave 133 of the 137 common taxonomic features positively associated (98 %, mean Spearman Rho ± standard deviation=0.77±0.14 (species), 0.75±0.19 (genus), 0.72±0.21 (family), 0.65±0.22 (order), 0.67±0.2 (class), 0.64±0.26 (phylum)), 122 of which (90%) were significantly associated (FDR<0.05, Spearman correlations) (Supplementary Table S4).

Importantly, these results also showed that the similarities in the relative abundances of taxonomic features between ONT and Illumina quantifications were significatively higher than between ONT and SOLiD sequencing (Supplementary Fig S6**;** P-value<0.05 for comparisons at species and genus level with ONT Centrifuge results; P-value<0.005 for comparisons at species, genus, and family level with ONT IGC results).

We made similar observations with functional profiles based on KEGG modules. Using Centrifuge, 76% and 72% of the functional modules were positively associated with the equivalent modules quantified with Illumina and SOLiD sequencing respectively, whereas this fraction substantially increased to 98% and 98% with ONT abundance data based on IGC quantifications (Supplementary Fig S6A). This difference may be related to the different content of both genomic reference spaces (Centrifuge genomes and IGC gene catalog), which can have a major impact on the quantification of functional modules if differences in database composition also result in differences in gene content. Importantly, we observed that DNA extraction also impacted the similarity between functional profiles, with ONT functional profiles being more similar to Illumina functional profiles based both on Centrifuge (P-value=0.0046 Wilcoxon rank-sum tests of Spearman's Rho distributions between ONT-SOLiD comparisons and ONT-Illumina comparisons; mean Spearman rho ± standard deviation=0.24±0.32 (ONT Centrifuge vs. Illumina IGC functional module abundances), 0.20±0.28 (ONT Centrifuge vs. SOLiD IGC functional module abundances) and IGC quantifications (P-value=0.0046 Wilcoxon rank-sum tests of Spearman's Rho distributions between ONT-SOLiD comparisons and ONT-Illumina comparisons; mean Spearman rho ± standard deviation=0.53±0.22 (ONT IGC vs. Illumina IGC functional module abundances), 0.44±0.18 (ONT IGC vs. SOLiD IGC functional module abundances).

Finally, we reproduced with ONT data previously reported associations between functional modules and microbiome diversity at similar strength as with Illumina and SOLiD data. We found significant positive associations between the sporulation module md:M00485 (KinABCDE-Spo0FA (sporulation control) two-component regulatory system) and microbial diversity (Supplemental Figure S8), which was in agreement with estimations of 50%-60% of bacteria from gut microbiome of healthy individuals producing resilient spores, being a basic feature of the human microbiome with a key impact in bacterial persistence and the spread of microbes between individuals [47]. This was also the case for the negative association between modules involved in the biosynthesis of bacterial Lipopolysaccharide (LPS) and microbial diversity (Supplemental Fig S9), in line with the association of obesity and other metabolic disorders with an increase of blood LPS concentration[48].

4. Discussion

Here, we presented a novel protocol and analytical pipeline enabling the quantification of the gut microbiome features using Oxford Nanopore Technologies. This technology supports easy access and use of high throughput sequencing at competitive costs as well as fast data production and analyses of the results. We believe this protocol enables the study of gut microbiome samples in the context of clinical applications or group studies. We improved protocols both for the wet-lab (from DNA extraction to sequencing) and data analysis. We also compared results to second generation sequencing methods (Illumina and SOLiD) in previously described patient cohort. This was driven by 1) an initial assessment of the best parameters in terms of alignment of ONT sequencing reads from simulated metagenomic datasets with different levels of complexity and, 2) the development of a bioinformatic pipeline which combines rapid k-mer based classification of ONT reads with read alignments vs. reference genomes to improve the quantification of microbiome species diversity and composition. This also included the taxonomic and functional profiling of ONT metagenomics data from reference genomes (Centrifuge approach) and gut microbiome non-redundant gene catalogs. Previous studies have proposed similar approaches based on Centrifuge for the real-time metagenomic profiling from ONT data [24] and more specifically for the metagenomic profiling of fecal and oral swabs [67], but not allowing functional profiling nor the metagenomic profiling using non-redundant gut microbiome gene catalogs that maximizes the genomic knowledge of the gut microbiome ecosystem. The simulation experiments revealed that filtering strategies commonly used with second generation sequencing technologies, such as high sequence identity thresholds, could not be extrapolated to highly error-prone reads such as those produced by ONT. In contrast, the mapping quality based on nanopore-adapted sequence aligners like minimap2 showed significantly better performance in terms of precision and recall of species richness composition estimates at different complexities of simulated communities. MapQ scores of 5 gave the best results regarding estimates of species richness and relative abundances of microbial species, being particulary suitable for complex ecosystems like the human gut.

Regarding sample processing, we elaborated a DNA extraction protocol from human stools that provide high DNA quality. Studies over the past years have used bacterial DNA or RNA to explore microbial communities in diverse ecosystems including stool samples from large cohorts [2],[51],[52]. Authors have used different DNA extraction protocols and different sequencing techniques (Illumina, SOLiD, Ion Proton). Multiple studies have also noted "batch"[53] effects and differences in data analyses [54]-[55], which introduce biases in analytical comparisons. Thus, the need for procedure standardization has been highlighted by several reports, as illustrated by the IHMS consortium[56]. They compared 21 DNA extraction methods using whole genome metagenomic shotgun sequencing with Illumina HiSeq2000 technology and assessed the taxonomic profile and functional variability while standardizing the stages of stool collection, bacterial DNA stabilization, library preparation and sequencing. This resulted in the generation of recommendations that would improve DNA extraction in terms of yield and quality.

Taking into account IHMS recommendations, we further optimized the microbial DNA extraction protocol which showed DNA yield improvement. We worked on two critical steps, bacterial wall lysis and protein/RNA elimination.

Sampling conditions, storage and harmonization have also been shown to be critical in affecting microbiome results. Although storing fecal samples at 4°C appeared to protect bacterial DNA from degradation, a reduction in microbial diversity was observed [57]. A previous study showed that prior storage of stool samples at 4°C (one hour) before placing them at -20°C, had a large impact on the taxonomic composition at the genus and species level[58]. However, these studies were conducted before the development and widespread use of commercially available fecal collection kits with stabilizing solution. Here, our results suggest that sample storage temperature is not a significant factor as long as guidelines from manufacturers are followed. The effect of sample storage kit type or temperature on sequencing and microbiome results are largely outweighed by inter-donor variation.

Since we identified read length as a critical criterion for subsequent bioinformatics analyses, we also improved the library preparation protocol to increase the proportion of long reads by optimizing the end-prep and DNA purification steps in the library. Finally PCoA of different wet-lab experiments showed that individual microbiome composition drove most of the variation observed in microbial diversity and quantitative metagenomic profiles obtained from ONT sequencing data, with no apparent batch effects associated to different wet-lab steps (DNA fragmentation, DNA end-repair, Collection kits, DNA library preparation or sequencing run).

To examine the relevance of our pipeline in human cohorts, we performed comparison of the results obtained with ONT sequencing with those obtained with SOLiD and Illumina sequencing on human stool samples collected in the "Microbaria" study [8]. For gene richness, microbiome composition and functional modules, the similarity was higher between ONT and Illumina sequencing compared to SOLID. ONT and Illumina sequences were generated from the same DNA extracted with the optimized protocol, which emphasizes the importance of DNA extraction protocols in quantitative metagenomic profiles.

The low throughput of ONT sequencing is however one of its major drawback for quantitative metagenomic studies of complex microbial ecosystems like the human gut microbiome. Nevertheless the results showed a high similarity in bacterial diversity estimation between the two sequencing methods in our test samples. Despite improvements in experimental protocol, the demultiplexing step needs to be improved. For instance, the ratio of unclassified reads was about 25%, knowing that the sequencing depth of ONT was low, and the error rate elevated. However, the low classification rates (25%) did not seem to impact the bacterial diversity and the bacterial compositions estimates. In this context, we could consider ONT in the context of quantitative metagenomic studies as a “shallow-sequencing” method in the line of proposed low-sequencing depth approaches to characterize microbial ecosystems more accurately than 16S barcoding approaches and with lower costs than deep shotgun sequencing[62].

5. Conclusion

Nanopore-based technology is proposed as easily accessible due to relatively low costs and a small benchtop footprint, providing an avenue to perform NGS in clinical settings. Admittedly, this technology has some drawbacks such as relatively modest sequencing depth, and error rates that remain high (2 - 5%) [63] compared to Illumina (0,1%)[64]. Through accurate assessment of experimental and bioinformatic steps our current work demonstrate that this technology is suitable to carry out quantitative metagenomic studies in the human gut microbiome. The maximization of ONT read lengths by our experimental protocol, in addition of being key in maximizing the efficiency of the taxonomic binning (fraction of classified reads), could be of major importance in additional aspects of metagenomic analyses like de-novo assembly of microbial genomes and the improvement of genomic completion of Metagenomic Species (MGS) derived from human gut microbiome gene catalogs. Our bioinformatic pipeline also extends for the first time the scope of metagenomic profiling of ONT reads to these gene catalogs, which has been of pivotal importance as reference genomic spaces used in multiple quantitative metagenomic studies, opening the way for the validation of disease biomarkers derived from these studies in clinical practice. In this context, we show that ONT consistently replicated results obtained with other sequencing technologies for intestinal microbiome diversity and the composition of main phyla in patients with severe obesity. This proposed workflow paves the way to taxonomic and functional profiling of microbial communities with this sequencing technology at competitive costs and fast data, which corresponds to a great need in the microbiome community.

Supplemental Figures and tables

Supplemental table S1 : Number of reads generated for the 250 simulated samples with CAMISIM.

Supplemental table S2 : Compositional profile of the 250 simulated samples based on 506 reference genomes used for CAMISIM simulations. Columns 1 to 5 corresponds to reference genomic information of the 506 genomes used in the simulations. Columns 6 to 255 corresponds to the relative abundances of the 506 reference genomes in the

doi:10.20944/preprints202108.0104.v1

250 simulated samples based on 10 reference pareto distributions (M1-M10) and 5 different species richness compositions (R50-R450)

Supplemental table S3 : Summary of the results of nanopore read alignments against 506 reference genomes in 250 simulated samples. For each sample is provided the mean, median standard deviation and standard error of read lengths of aligned and unaligned reads (class column).

Supplemental table S4 : Comparison of taxonomic features based on MGS quantification from IGC gene catalog between Nanopore quantifications and Illumina and Solid quantifications. For each taxonomic features is reported the Spearman correlation between pairwise abundance estimates (comparison column), the pvalue and the adjusted qvalue for multiple comparisons in each level of comparison column, and the prevalence of the feature in the 33 samples based on Nanopore sequencing data.

Supplemental table S5 : Improvement of DNA yield and quality with the Optimized Invitrogen protocol. Nanodrop data Estimation of the amount of DNA, protein contamination given by the 260/280 ratio and by impurities and solvents contamination estimated by the 260/230 ratio (n = 6, Wilcoxon test).

Supplemental Fig S1 : Density distributions of mapQ scores in primary alignments of 250 simulated samples stratified by the number of species in reference samples (50 samples per reference species richness).

Supplemental Fig S2 : Statistical comparison of differences between reference and simulated samples at different thresholds of mapQ scores. At each level of species richness (from 50 to 450 species, 50 samples per level), we compare the distributions of the similarities in species abundances between reference and simulated samples (Spearman's Rho coefficients of correlations between reference and simulated species abundance vectors) for all possible pairs of mapQ thresholds evaluated with Trukey's post-hoc pairwise tests. The 95% family-wise confidence level in the difference between pairs of mapQ threshold is represented colored by the significance of the difference according to adjusted P-values in Tukey's tests. If we focus on the mapQ=5, we observe that higher mapQ values leads to higher similarities between reference and simulated species abundance vectors (positive values in the confidence levels of the differences) in R50 and R150 simulated samples, whereas this is not the case for more compex/rich simulated samples (R250-R450), where we observe that the similarities with the reference decrease as we increase the stringency of the mapQ filtering (negative values in the confidence levels of the differences, being significant for R450 samples).

Supplemental Fig S3 : Similarity in species richness and species abundances estimations across different mapQ thresholds. (A) Distribution of F1 scores in species richness estimates (harmonic mean of precision and recall; y-axis) across simulated data filtered by different mapQ thresholds (x-axis) and stratified by the complexity of simulated microbial communities. (B) Distribution of similarities between simulated and reference abundance profiles (Spearman Rho's) across simulated data filtered by different mapQ thresholds (x-axis) and stratified by the complexity of simulated microbial communities. (c) Correlogram of Spearman Rho's between F1 scores in panel A and similarities between simulated and reference abundance profiles in panel B. *=P-value<0.05 Spearnan Rank test.

Supplemental Fig S4 : Taxonomic profile of ZymoBIOMICS mock community inferred from Nanopore sequencing. The reference composition of ZymoBIOMICS mock community is compared with the taxonomic profile obtained from Nanopore sequencing data with Centrifuge only and with Centrifuge combined with filtering of read bins by minimap2 mapping against the corresponding reference genomes with parameters derived from simulation experiments (primary alignments only, min. mapQ=5)

Supplemental Fig S5 : Comparison of microbial composition of Microbaria samples between Nanopore, Illumina and SOLiD sequencing data. (a) PCoA of samples from Microbaria study based on genus-level MGS abundance data from three different sequencing technologies (n=33 for Nanopore (ONT) and SOLiD; n=21 for Illumina). Significant effect of sequencing technology in microbiome composition is observed in PERMANOVA test (P-value=0.001; R2=0.11), with sample points from Illumina and ONT sequencing data (both generated with Invitrogen optimized protocol) closer than sample points from SOLiD sequencing (different DNA extraction method) (b) Hierarchical clustering of Microbaria samples product of different sequencing methods based on same genus-level MGS abundance data as PCoA in panel A. Sample points from Illumina and ONT sequencing over same biological sample tends to cluster together in the dendrogram.

Supplemental Fig S6 : Comparison of similarities in the abundance of taxonomic features between Nanopore and SOLiD-Illumina sequencing data. (a) Correlations of taxonomic feature abundances at different levels of taxonomic hierarchy between Nanopore(ONT) abundance data based on Centrifuge approach and Illumina and SOLiD abundance data (based on MGS from IGC catalog). (b) Correlations of taxonomic feature abundances at different levels of taxonomic hierarchy between ONT abundance data and Illumina and SOLiD abundance data based on MGS from IGC catalog. Dashed lines connect the same taxonomic feature across comparisons. ** P-value<0.01, Paired Wilcoxon rank-sum test.

Supplemental Fig S7: Comparison of similarities in KEGG functional modules abundance between Nanopore (ONT) and SOLiD-Illumina sequencing data. (a) Vulcano plots comparing the results of Spearman correlations of individual KEGG functional modules between ONT and Illumina-SOLiD sequencing data. (b) Comparison of similarities in module abundance data (Spearman's Rho) between ONT abundance data (from Centrifuge and from MGS abundance data) and Illumina and SOLiD abundance data (based on MGS abundance data). P-values from pairwise Wilcoxon rank-sum tests of Spearman's rho distributions between comparisons in x-axis are shown above the violin plots.

Supplemental Fig S8: Scatterplots of KEGG Sporulation module M00485 abundance and microbial diversity across different quantifications of diversity and module abundance based on Nanopore (ONT), SOLiD and Illumina sequencing data. Results of Spearman correlation tests are shown for each comparison.

Supplemental Fig S9: Scatterplots of abundances of KEGG LPS biosynthesis modules (M00060, M00063) and microbial diversity across different quantifications of diversity and module abundance based on Nanopore (ONT), SOLiD and Illumina sequencing data. Results of Spearman correlation tests are shown for each comparison.

Author contribution : K.C., E.P, E.B and J-D.Z conceived and designed the project. K.C. is the primary investigator of the clinical study, from which, samples were used in this work. R.A. performed all the wet-lab experiments and established the protocol (from stool sample processing to sequencing). E.B developed the databases, analytical pipelines and performed metagenomics and functional analysis. E.P. conceived the simulation and validation protocol and supervised the development of the bioinformatics workflows. K.C, J-D.Z., E.P., R.A., P.L. T.W. and E.B. contributed to results' discussion. R.A., E.B., E.P. and K.C wrote the paper and all authors commented and edited the manuscript.

Funding : This study was supported by the European Union FP7 Metacardis (grant agreement HEALTH- F4-2012-305312), H2020 EPoS (H2020- PHC-2014-634413), the Innovative Medicines Initiative 2 (IMI-2) Joint Undertaking under grant agreement No. 777377. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This study was also supported by Transatlantic Networks of Excellence Award from the Leducq Foundation (17CVD01) and JPI (A healthy diet for a healthy life; 2017-01996_3). The clinical study (Microbaria) was sponsored by Assistance Publique Hopitaux de Paris.

Informed Consent Statement : All subjects from the MetaCardis study or the Microbaria study provided written informed consent and the study was conducted in accordance with the Helsinki Declaration. MetaCardis study is registered in clinical trial https://clinicaltrials.gov/show/NCT02059538 and Microbaria study was registered in clinicaltrial.gov (NCT01454232).

Consent for publication : All authors gave consent for publication.

Data Availability Statement : Sequences have been deposited in the European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA) (Private access until paper acceptance). The computational pipeline is freely available in https://git.ummisco.fr/ebelda/nanopore. Other data are available on request.

Acknowledgements : We thanks Agnieszka Sekowska for his contribution to the reflection of the implementation of this project. We thank Tim Swartz (Integrative Phenomics) for critical reading of the manuscript and English correction.

Conflicts of Interest KC is a consultant for Danone Research, LNC therapeutics and CONFO therapeutics for work unassociated with the present study. J-D.Z. is consultant for Quinten for work unassociated with the present study.

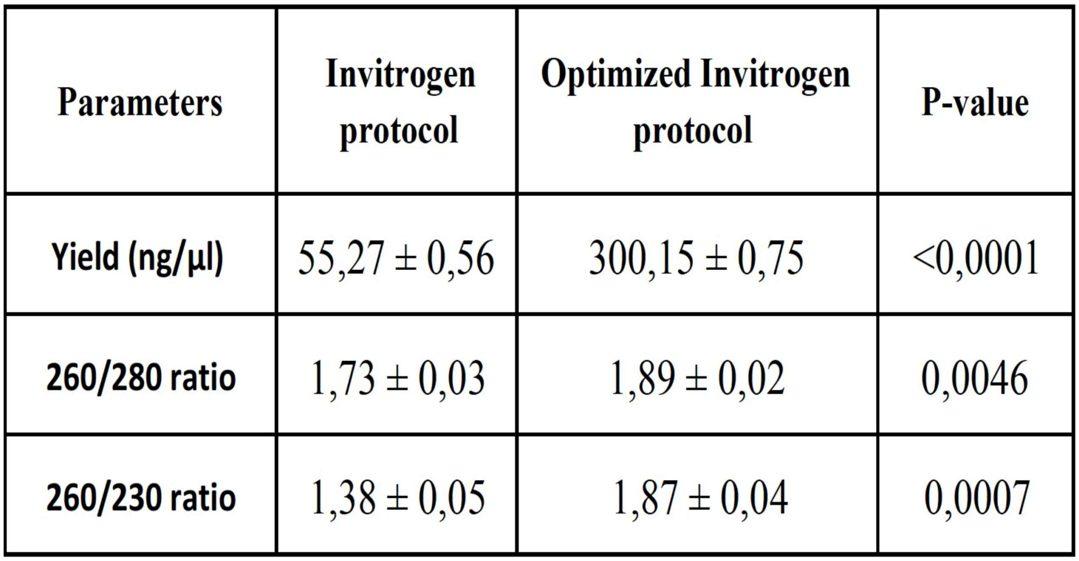
References

1. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
2. Vieira-Silva S, Falony G, Belda E, Nielsen T, Aron-Wisnewsky J, Chakaroun R, et al. Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. Nature. 2020;581:310–5.
3. Aron-Wisnewsky J, Gaborit B, Dutour A, Clement K. Gut microbiota and non-alcoholic fatty liver disease: new insights. Clin Microbiol Infect. 2013;19:338–48.
4. Aron-Wisnewsky J, Vigliotti C, Witjes J, Le P, Holleboom AG, Verheij J, et al. Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic disorders. Nat Rev Gastroenterol Hepatol. 2020;17:279–97.
5. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006;444:1027–31.
6. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513:59–64.
7. Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, et al. Dietary intervention impact on gut microbial gene richness. Nature. 2013;500:585–8.
8. Aron-Wisnewsky J, Prifti E, Belda E, Ichou F, Kayser BD, Dao MC, et al. Major microbiota dysbiosis in severe obesity: fate after bariatric surgery. Gut. 2019;68:70–82.
9. Raes J, Bork P. Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol. 2008;6:693–9.
10. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci USA. 2015;112:14024–9.
11. Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, et al. Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities. PLOS ONE. 2013;8:e74787. doi:10.1371/journal.pone.0074787.
12. Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, et al. Processing faecal samples: a step forward for standards in microbial community analysis. BMC Microbiology. 2014;14:112. doi:10.1186/1471-2180-14-112.
13. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, 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. 2014;9:e88982. doi:10.1371/journal.pone.0088982.
14. Voigt AY, Costea PI, Kultima JR, Li SS, Zeller G, Sunagawa S, et al. Temporal and technical variability of human gut metagenomes. Genome Biol. 2015;16:73.
15. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449:804–10.
16. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35:1069–76.
17. Harstad H, Ahmad R, Bredberg A. Nanopore-based DNA sequencing in clinical microbiology: preliminary assessment of basic requirements. bioRxiv. 2018;:382580. doi:10.1101/382580.
18. Nayfach S, Pollard KS. Toward Accurate and Quantitative Comparative Metagenomics. Cell. 2016;166:1103–16. doi:10.1016/j.cell.2016.08.007.
19. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nature Biotechnology. 2020;38:701–7. doi:10.1038/s41587-020-0422-6.
20. Maghini DG, Moss EL, Vance SE, Bhatt AS. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. Nat Protoc. 2021;16:458–71.
21. Deng X, Achari A, Federman S, Yu G, Somasekar S, Bártolo I, et al. Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. Nat Microbiol. 2020;5:443–54.
22. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, et al. Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. Nat Microbiol. 2020;5:430–42.
23. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37:783–92.
24. Sanderson ND, Street TL, Foster D, Swann J, Atkins BL, Brent AJ, et al. Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. BMC Genomics. 2018;19:714.
25. Urban L, Holzer A, Baronas JJ, Hall MB, Braeuninger-Weimer P, Scherm MJ, et al. Freshwater monitoring by nanopore sequencing. Elife. 2021;10.
26. Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. Proc Natl Acad Sci U S A. 2020.
27. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat

Biotechnol. 2014;32:822–8.
28. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–41.
29. Fritz A, Hofmann P, Majda S, Dahms E, Dröge J, Fiedler J, et al. CAMISIM: simulating metagenomes and microbial communities. Microbiome. 2019;7:17.
30. https://nanoporetech.com/community.
31. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology. 2019;20:129. doi:10.1186/s13059-019-1727-y.
32. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26:1721–9.
33. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
34. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Res. 2013;2:191. doi:10.12688/f1000research.2-191.v2.
35. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE. 2013;8:e61217.
36. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology. 2013;9:e1003118. doi:10.1371/journal.pcbi.1003118.
37. Lee S, Cook D, Lawrence M. plyranges: a grammar of genomic data transformation. Genome Biology. 2019;20:4. doi:10.1186/s13059-018-1597-8.
38. http://meta.genomics.cn/meta/home.
39. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40 Database issue:D109-114.
40. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
41. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.
42. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
43. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
44. https://CRAN.R-project.org/package=vegan.
45. https://community.nanoporetech.com/protocols/native-barcoding-genomic-dna/.
46. James@cancer. CoreGenomics: How do SPRI beads work? CoreGenomics. 2012. http://core-genomics.blogspot.com/2012/04/how-do-spri-beads-work.html. Accessed 18 Nov 2020.
47. Hp B, Sc F, Bo A, N K, Ba N, Md S, et al. Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation. Nature. 2016;533:543–6. doi:10.1038/nature17645.
48. Krajmalnik-Brown R, Ilhan Z-E, Kang D-W, DiBaise JK. Effects of gut microbes on nutrient absorption and energy regulation. Nutr Clin Pract. 2012;27:201–14.
49. Dao MC, Belda E, Prifti E, Everard A, Kayser BD, Bouillot J-L, et al. Akkermansia muciniphila abundance is lower in severe obesity, but its increased level after bariatric surgery is not associated with metabolic health improvement. Am J Physiol Endocrinol Metab. 2019;317:E446–59.
50. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016;17:132. doi:10.1186/s13059-016-0997-x.
51. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe. 2015;17:690–703.
52. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25:968–76.
53. Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, et al. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome. 2014;2:19.
54. McOrist AL, Jackson M, Bird AR. A comparison of five methods for extraction of bacterial DNA from human faecal samples. J Microbiol Methods. 2002;50:131–9.
55. Ariefdjohan MW, Savaiano DA, Nakatsu CH. Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. Nutr J. 2010;9:23.
56. http://www.microbiome- standards.org/.
57. Ott SJ, Musfeldt M, Timmis KN, Hampe J, Wenderoth DF, Schreiber S. In vitro alterations of intestinal bacterial

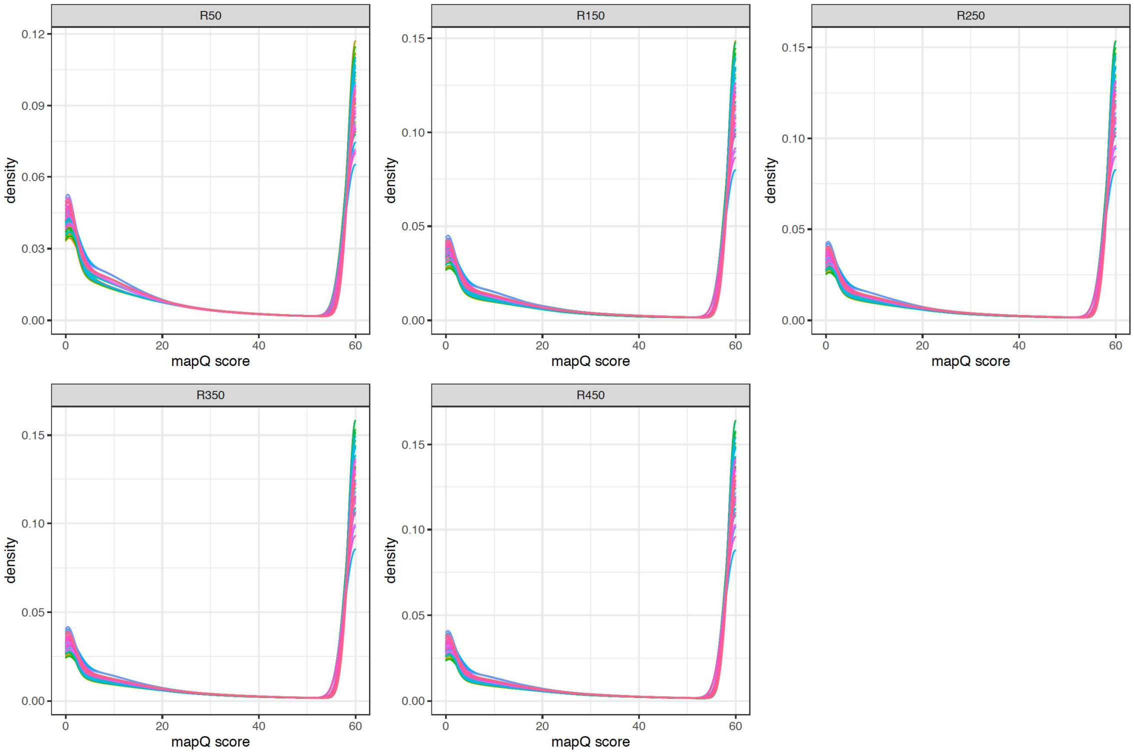
microbiota in fecal samples during storage. Diagn Microbiol Infect Dis. 2004;50:237–45.
58. Cardona S, Eck A, Cassellas M, Gallart M, Alastrue C, Dore J, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. BMC Microbiol. 2012;12:158.
59. Dao MC, Everard A, Aron-Wisnewsky J, Sokolovska N, Prifti E, Verger EO, et al. Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. Gut. 2016;65:426–36.
60. Caputo A, Dubourg G, Croce O, Gupta S, Robert C, Papazian L, et al. Whole-genome assembly of Akkermansia muciniphila sequenced directly from human stool. Biol Direct. 2015;10:5.
61. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biology. 2018;19:90. doi:10.1186/s13059-018-1462-9.
62. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, et al. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems. 2018;3. doi:10.1128/mSystems.00069-18.
63. Tedersoo L, Drenkhan R, Anslan S, Morales-Rodriguez C, Cleary M. High-throughput identification and diagnostics of pathogens and pests: Overview and practical recommendations. Mol Ecol Resour. 2019;19:47–76.
64. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. Next Gener Seq Appl. 2014;1.
65. Prifti E, Chevaleyre Y, Hanczar B, Belda E, Danchin A, Clément K, et al. Interpretable and accurate prediction models for metagenomics data. Gigascience. 2020;9.
66. Vanessa R. Marcelino, Philip T. L. C. Clausen, Edward C. Holmes & al. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data
67. Christoph Ammer-Herrmenau, R. Alili, E. Belda, K. Clement, and A. Neesse "Comprehensive wet-bench and bioinformatics workflow for complex microbiota using Oxford Nanopore Technologies". Msystemsjournal. 2021.
68. M Kanehisa & S Goto. "KEGG: kyoto encyclopedia of genes and genomes". Nucleic Acids Res. 2000 Jan 1.

Supplemental Figures

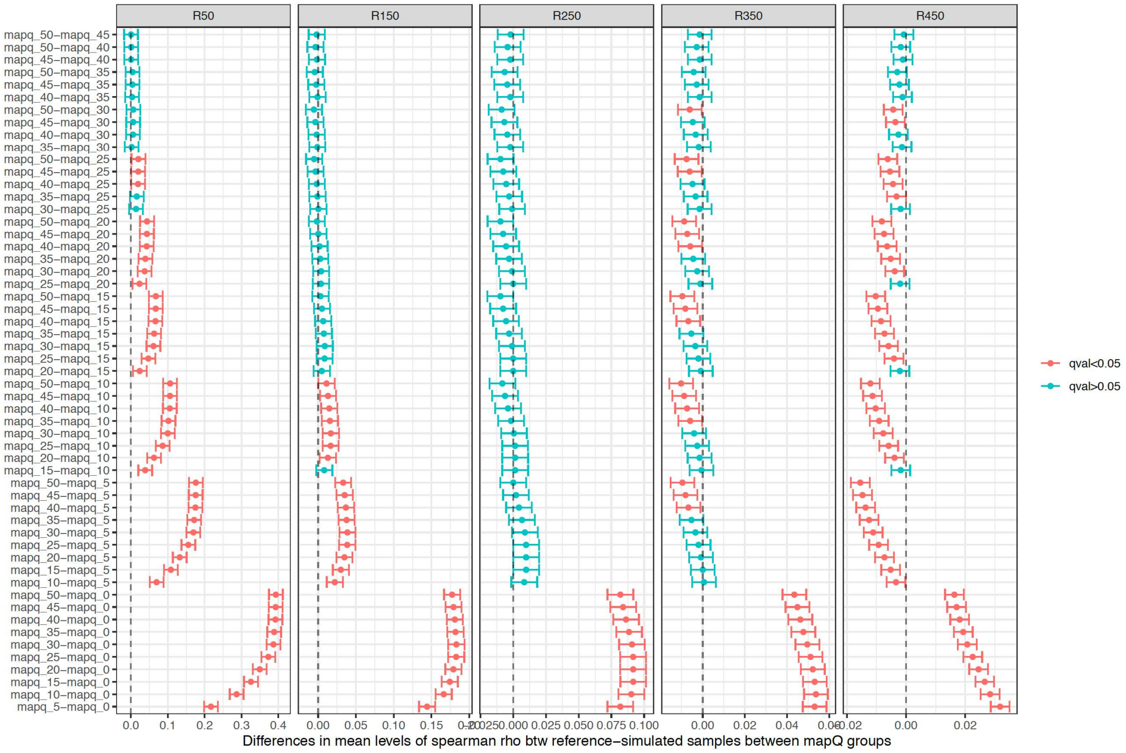


Parameters	Invitrogen protocol	Optimized Invitrogen protocol	P-value
Yield (ng/µl)	55,27 ± 0,56	300,15 ± 0,75	<0,0001
260/280 ratio	1,73 ± 0,03	1,89 ± 0,02	0,0046
260/230 ratio	1,38 ± 0,05	1,87 ± 0,04	0,0007

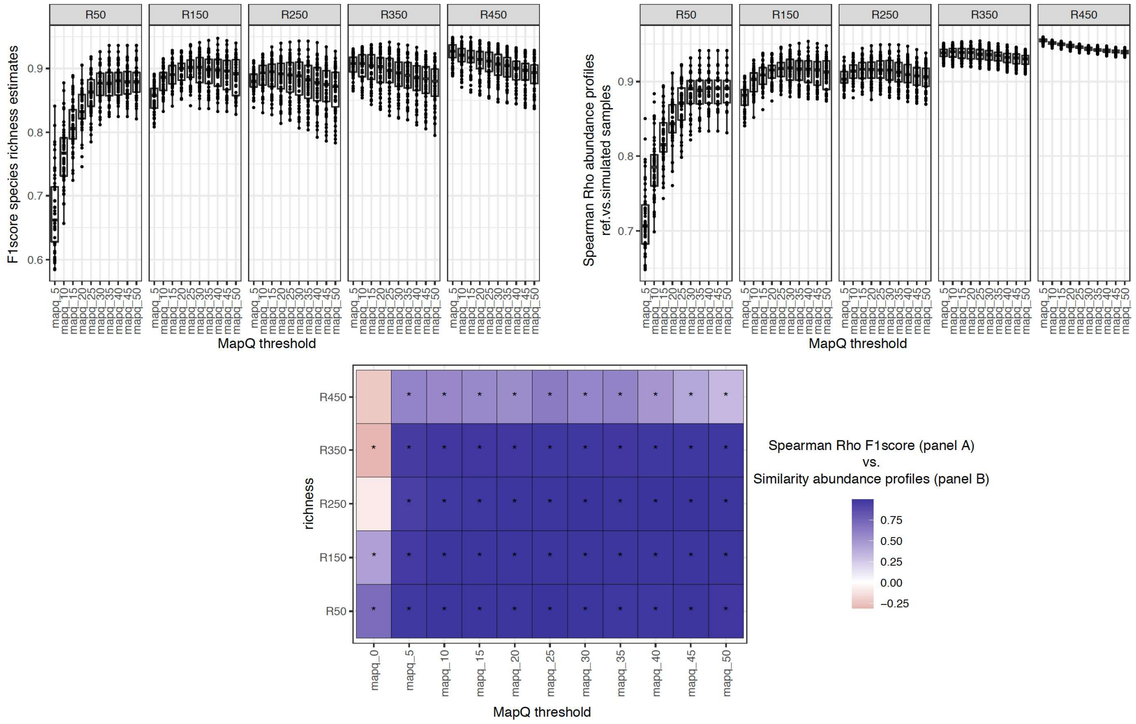
Supplemental table S1: Improvement of DNA yield and quality with the Optimized Invitrogen protocol. Nanodrop data Estimation of the amount of DNA, protein contamination given by the 260/280 ratio and by impurities and solvents contamination estimated by the 260/230 ratio (n = 6, Wilcoxon test).



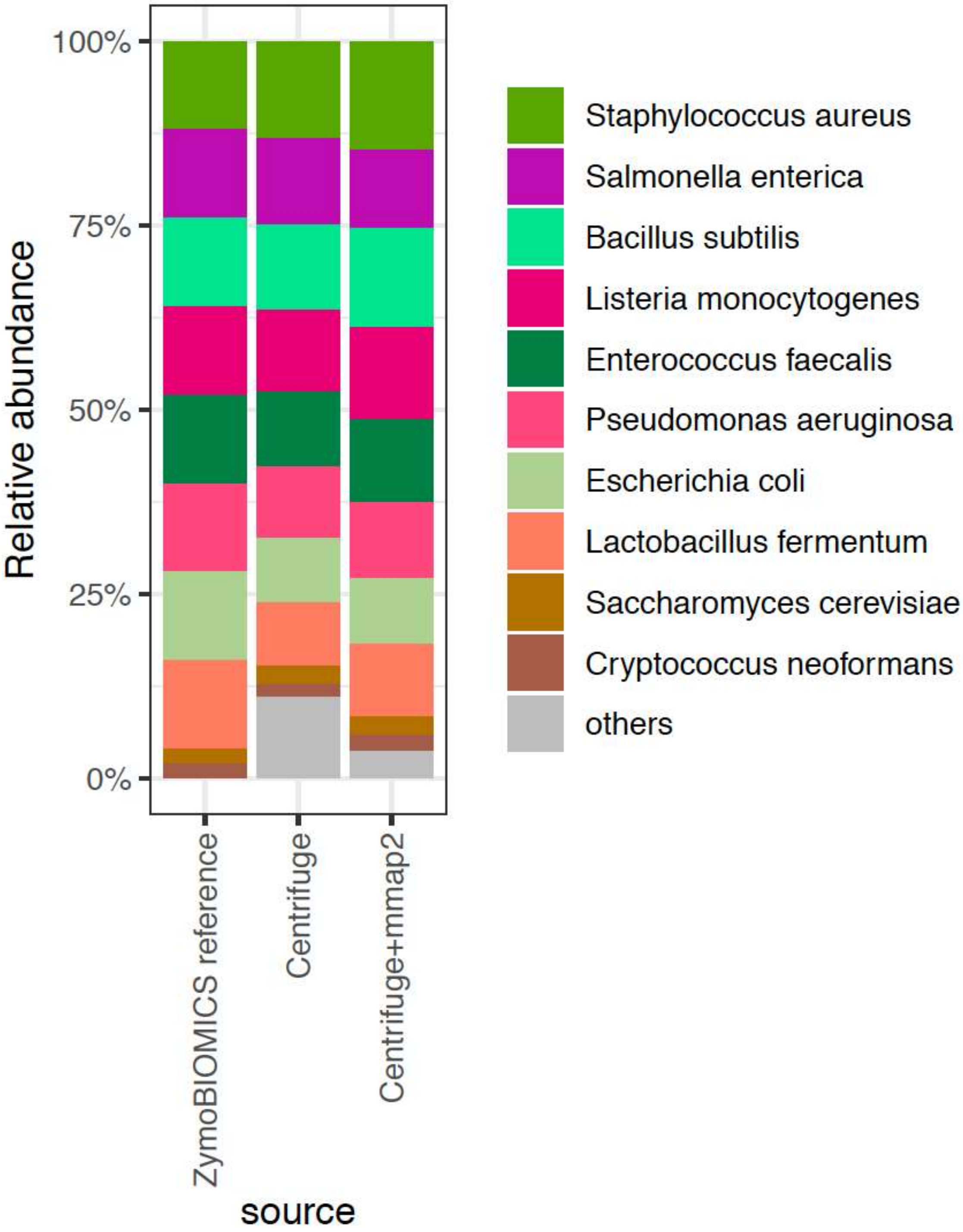
Supplemental Fig S1: Density distributions of mapQ scores in primary alignments of 250 simulated samples stratified by the number of species in reference samples (50 samples per reference species richness).



Supplemental Fig S2: Statistical comparison of differences between reference and simulated samples at different thresholds of mapQ scores. At each level of reference species richness (from 50 to 450 species, 50 samples per level), we compare the distributions of the similarities in species abundances between reference and simulated samples (Spearman's Rho coefficients of correlations between reference and simulated species abundance vectors) for all possible pairs of mapQ thresholds evaluated with Trukey's post-hoc pairwise tests. The 95% family-wise confidence level in the difference between pairs of mapQ threshold is represented colored by the significance of the difference according to adjusted P-values in Tukey's tests. If we focus on the mapQ=5, we observe that higher mapQ values leads to higher similarities between reference and simulated species abundance vectors (positive values in the confidence levels of the differences) in R50 and R150 simulated samples, whereas this is not the case for more compex/rich simulated samples (R250-R450), where we observe that the similarities with the reference decrease as we increase the stringency of the mapQ filtering (negative values in the confidence levels of the differences, being significant for R450 samples).

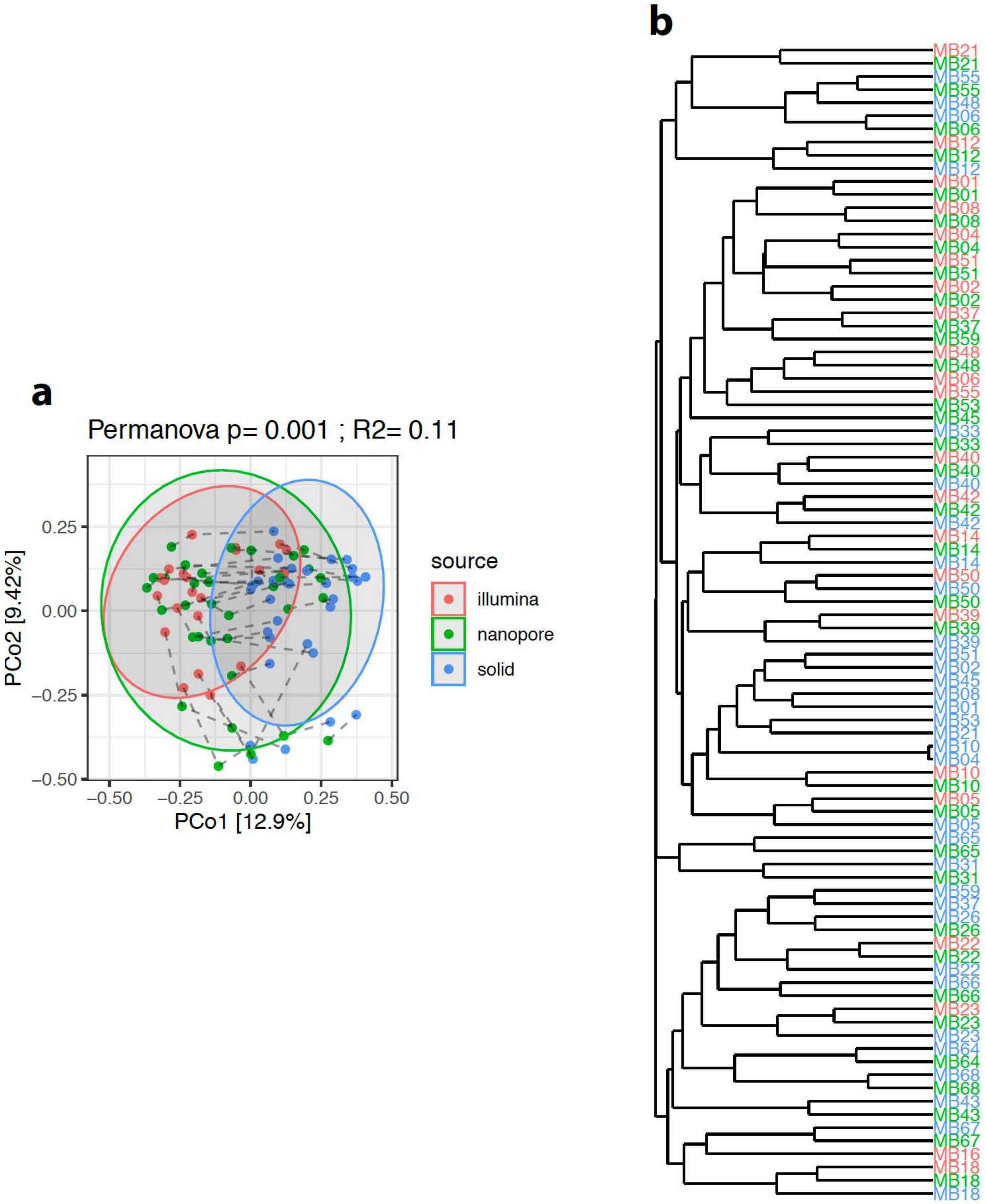


Supplemental Fig S3: Similarity in species richness and species abundances estimations across different mapQ thresholds. (A) Distribution of F1 scores in species richness estimates (harmonic mean of precision and recall; y-axis) across simulated data filtered by different mapQ thresholds (x-axis) and stratified by the complexity of simulated microbial communities. (B) Distribution of similarities between simulated and reference abundance profiles (Spearman Rho's) across simulated data filtered by different mapQ thresholds (x-axis) and stratified by the complexity of simulated microbial communities. (c) Correlogram of Spearman Rho's between F1 scores in panel A and similarities between simulated and reference abundance profiles in panel B. *=P-value<0.05 Spearnan Rank test.



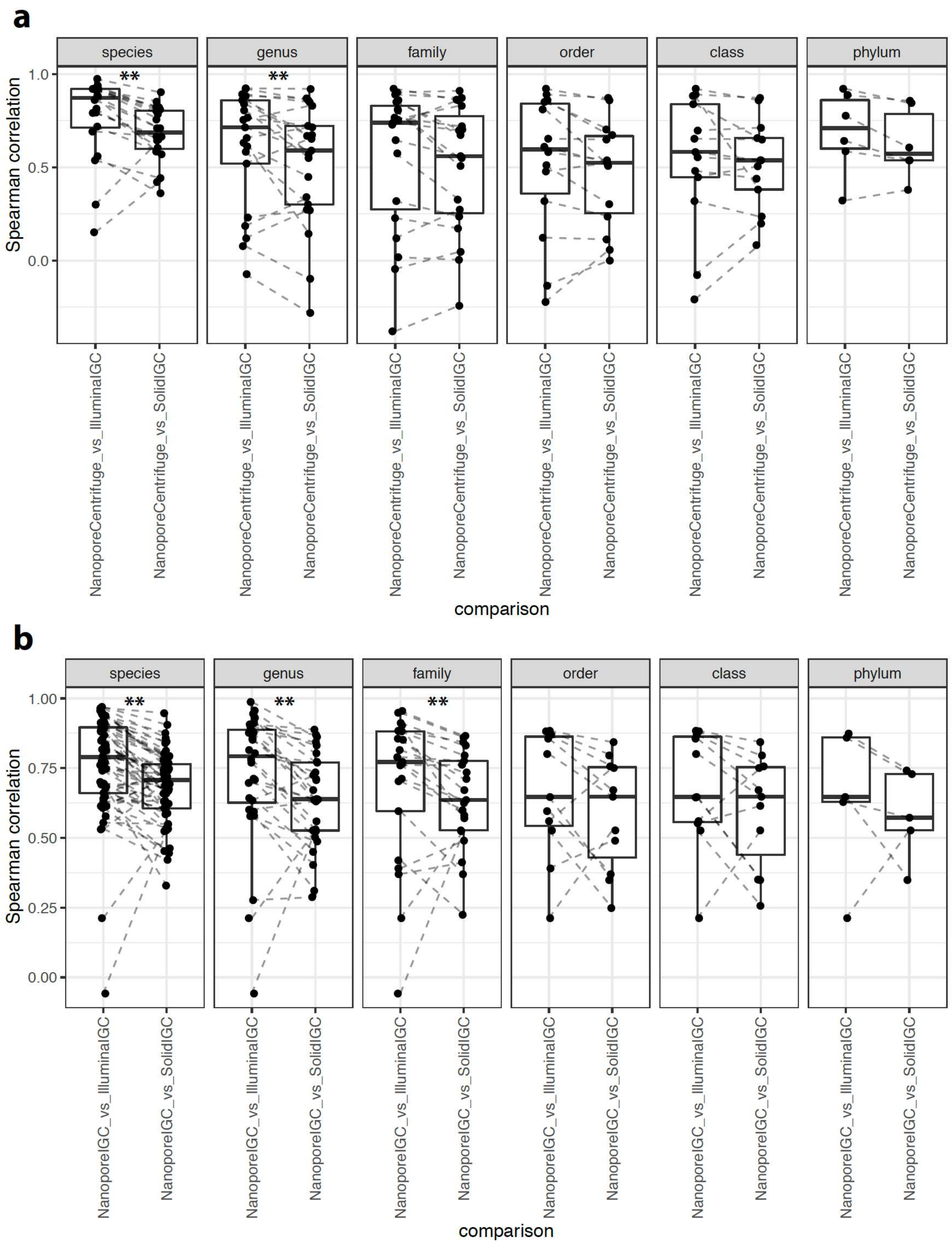
Supplemental Fig S4: Taxonomic profile of ZymoBIOMICS mock community inferred from Nanopore sequencing. The reference composition of ZymoBIOMICS mock community is compared with the taxonomic profile obtained from Nanopore sequencing data with Centrifuge only and with Centrifuge combined with filtering of read bins by minimap2 mapping against the corresponding reference genomes with parameters derived from simulation experiments (primary alignments only, min. mapQ=5)

doi:10.20944/preprints202108.0104.v1

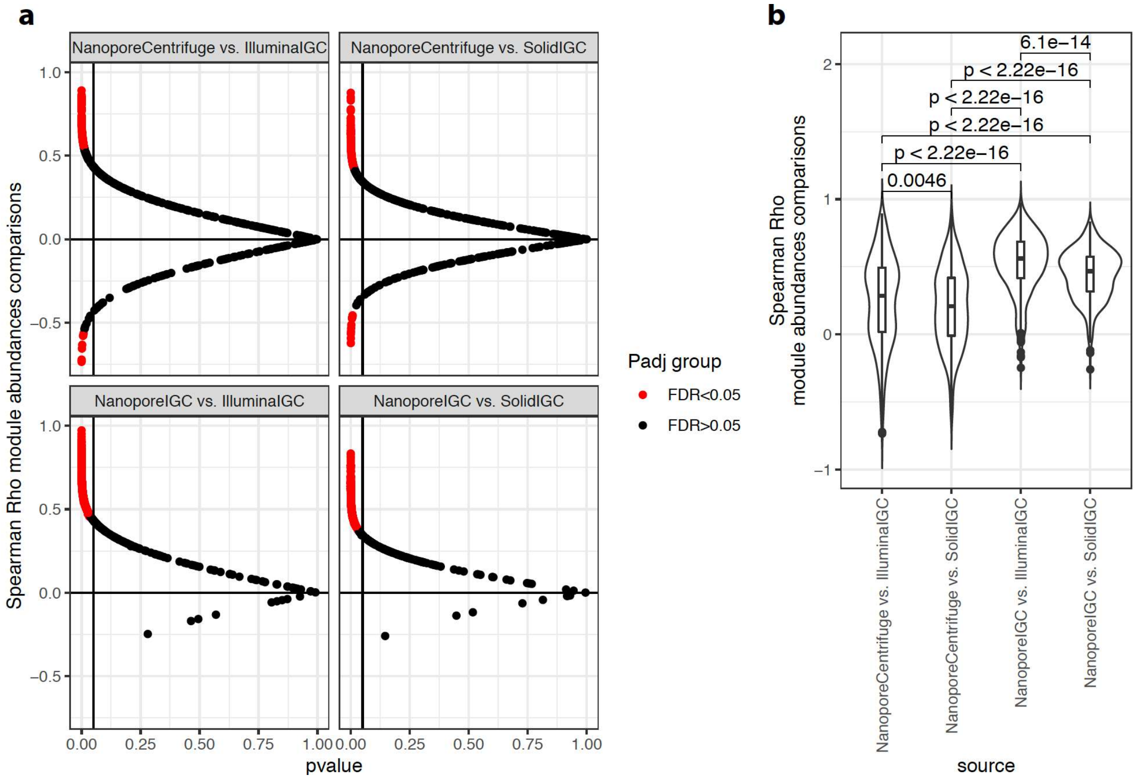


Supplemental Fig S5: Comparison of microbial composition of Microbaria samples between Nanopore, Illumina and SOLiD sequencing data. (a) PCoA of samples from Microbaria study based on genus-level MGS abundance data from three different sequencing technologies (n=33 for Nanopore (ONT) and SOLiD; n=21 for Illumina). Significant effect of sequencing technology in microbiome composition is observed in PERMANOVA test (P-value=0.001; R2=0.11), with sample points from Illumina and ONT sequencing data (both generated with Invitrogen optimized protocol) closer than sample points from SOLiD sequencing (different DNA extraction method) (b) Hierarchical clustering of Microbaria samples product of different sequencing methods based on same genus-level MGS abundance data as PcoA in panel A. Sample points from Illumina and ONT sequencing over same biological sample tends to cluster together in the dendrogram.

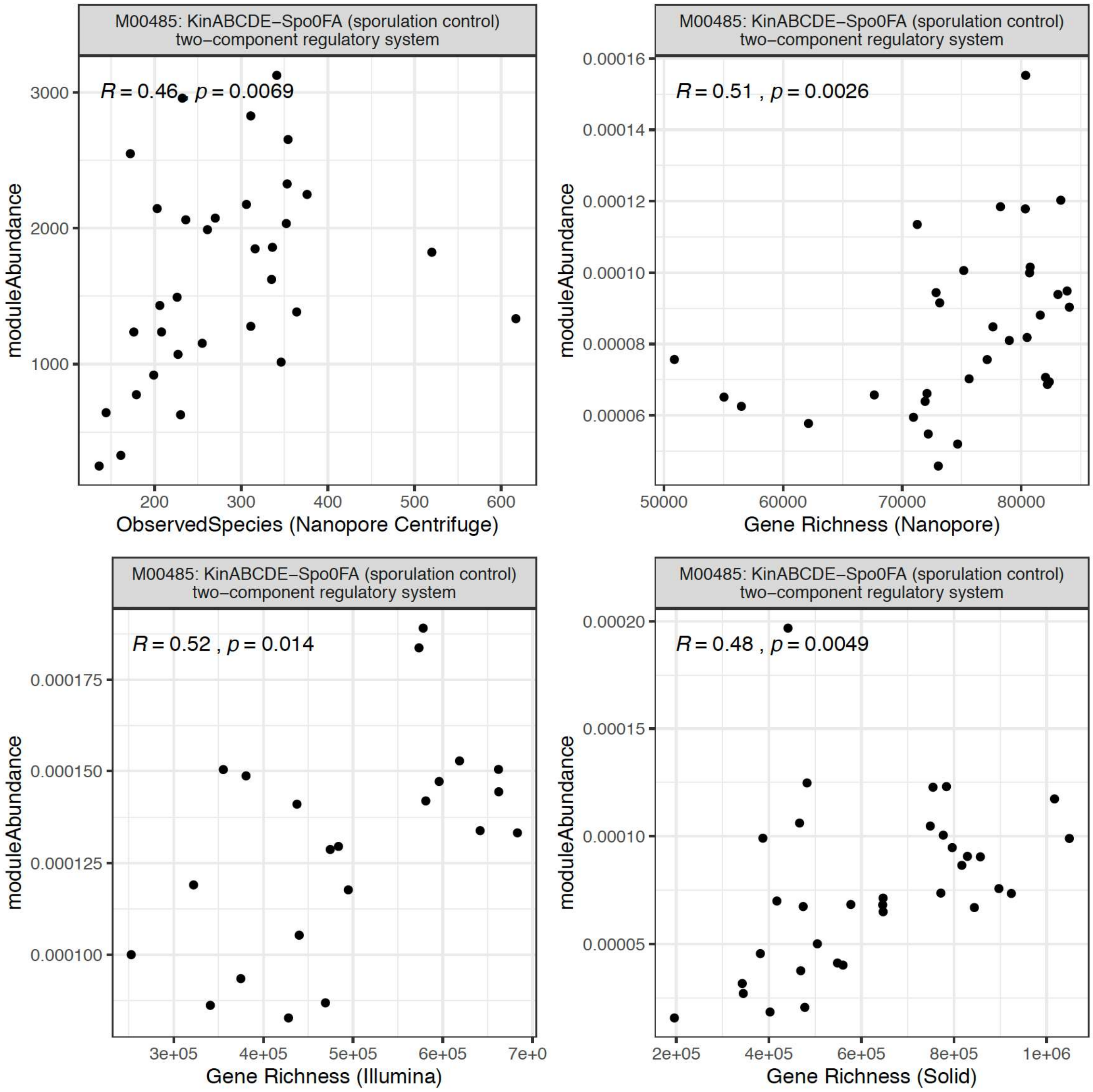
doi:10.20944/preprints202108.0104.v1



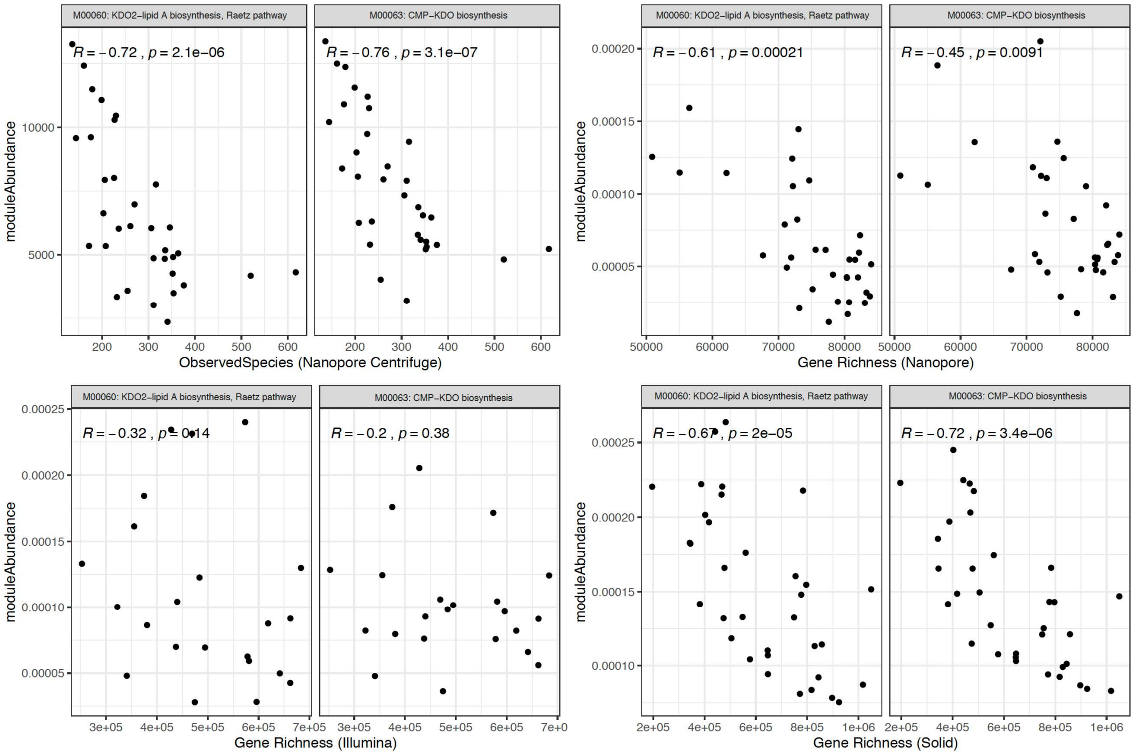
Supplemental Fig S6: Comparison of similarities in the abundance of taxonomic features between Nanopore and SOLiD-Illumina sequencing data. (a) Correlations of taxonomic feature abundances at different levels of taxonomic hierarchy between Nanopore(ONT) abundance data based on Centrifuge approach and Illumina and SOLiD abundance data (based on MGS from IGC catalog). (b) Correlations of taxonomic feature abundances at different levels of taxonomic hierarchy between ONT abundance data and Illumina and SOLiD abundance data based on MGS from IGC catalog. Dashed lines connect the same taxonomic feature across comparisons. ** P-value<0.01, Paired Wilcoxon rank-sum test.



Supplemental Fig S7: Comparison of similarities in KEGG functional modules abundance between Nanopore (ONT) and SOLiD-Illumina sequencing data. (a) Vulcano plots comparing the results of Spearman correlations of individual KEGG functional modules between ONT and Illumina-SOLiD sequencing data. (b) Comparison of similarities in module abundance data (Spearman's Rho) between ONT abundance data (from Centrifuge and from MGS abundance data) and Illumina and SOLiD abundance data (based on MGS abundance data). P-values from pairwise Wilcoxon rank-sum tests of Spearman's rho distributions between comparisons in x-axis are shown above the violin plots.



Supplemental Fig S8: Scatterplots of KEGG Sporulation module M00485 abundance and microbial diversity across different quantifications of diversity and module abundance based on Nanopore (ONT), SOLiD and Illumina sequencing data. Results of Spearman correlation tests are shown for each comparison.



Supplemental Fig S9: Scatterplots of abundances of KEGG LPS biosynthesis modules (M00060, M00063) and microbial diversity across different quantifications of diversity and module abundance based on Nanopore (ONT), SOLiD and Illumina sequencing data**.** Results of Spearman correlation tests are shown for each comparison.