Exploring Quantitative Metagenomics Studies using Oxford Nanopore Sequencing: A Computational and Experimental Protocol

Rohia ALILI
Unité de recherche mixte 1269 - NutriOmics https://orcid.org/0000-0002-0158-4250
Eugeni BELDA (e.belda@integrative-phenomics.com)
Integrative Phenomics https://orcid.org/0000-0003-4307-5072
Karine CLEMENT
UMRS 1269 - NutriOmics
Phuong Le
Sorbonne Université,
Edi PRIFTI
IRD : UMMISCO
Jean-Daniel ZUCKER
IRD : UMMISCO
Thierry WIRTH
Institut de Systématique Evolution Biodiversité: Institut de Systematique Evolution Biodiversite

Methodology

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

Posted Date: December 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-131495/v1

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.

[2] Assistance Publique Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Nutrition department, CRNH Ile de France, Paris France.

[3] Integrative Phenomics, Paris, France.

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

[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.

Email addresses :

Rohia ALILI : rohia.alili@aphp.fr

Eugeni BELDA : e.belda@integrative-phenomics.com

Karine CLEMENT : karine.clement@inserm.fr

Phuong LE : phuongleee@gmail.com

Edi Prifti : edi.prifti@gmail.com

Thierry WIRTH : thierry.wirth@mnhn.fr

Jean-Daniel ZUCKER : jdzucker@gmail.com

Corresponding authors

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

Abstract

Background : The gut microbiome plays a major role in chronic diseases, several of which are characterized by an altered diversity and composition of bacterial communities. Large-scale sequencing projects allowed the characterization of these microbial community perturbations. However, a gap remains in how these discoveries can be translated into clinical applications. 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 the human gut microbiome with Oxford Nanopore sequencing technology (ONT). We developed a bioinformatic pipeline to process ONT sequences based on the evaluation of different alignment parameters in the estimation of microbial diversity and composition. We also optimized stool collection and DNA extraction methods to maximize read length, a critical parameter for the sequence alignment and classification. Our analytical pipeline was evaluated using simulations of metagenomic communities to reflect naturally occuring compositional variations. We then validated our experimental and analytical pipeline with stool samples from a bariatric surgery cohort sequenced with ONT and Illumina, revealing comparable diversity and microbial composition profiles. These results were compared to those previously obtained with SOLiD sequencing, where differences were observed, possibly explained by variations in library preparation steps. Finally, we found that sequences obtained with ONT allowed assembly of complete genomes for disease-related species.

Conclusion : This protocol can be implemented in the clinical or individual 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.

Background

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 challenge to discover reproducible gut microbial signatures for diseases in order to develop diagnostic and prognostic tools.

In the field of metabolic diseases, microbial diversity and richness 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.

In addition to biological variation, gut microbiome quantification is subject to technical variation along the experimental process starting from sample collection and extending to DNA extraction, library preparation and sequencing, and bioinformatics analytical protocols[9]. This is observed in the literature where results are frequently irreproducible[10],[11],[12],[13]. Reports highlight 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 to develop methods for fast-track and affordable microbiome screening in clinical settings. Among critical steps

is DNA extraction prior to sequencing. Costea et al.[16] for example reported biases in microbial composition and diversity varying 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. For example, the richness is higher and reads are longer with the Qiagen compared to Magnapure kits[17]. The library preparation has also an impact on relative abundances of bacterial features and functional annotation[10]. Finally, the bioinformatic pipelines can yield consequent variability in microbial ecosystem description[18].

Presently, most microbiome research studies are carried out using 16S ribosomal RNA genes or whole genome shotgun sequencing (WGS), the latter requiring quite a long time for data processing and result producing. 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 technology, Oxford Nanopore Technology (ONT) proposes real-time sequence data generation with fewer resources and a small benchtop footprint. Due to the recent development of this technology, there is a need to define standardized wet-lab and bioinformatics protocols for ONT in metagenomics. Importantly, annotation and quantification of ONT-generated long reads is properly needed as current bioinformatics and metagenomics pipeline are adapted to short reads. Here, we have explored methods to optimize ONT for microbiome analyses and propose a fine-tuned protocol, including wet-lab preparation (i.e. sample collection, DNA extraction, and library preparation) and 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 current bacterial gene catalogs along with functional profiling. This protocol is open-access, allowing it to be replicated, and implemented by medical or research centers.

Materials and methods

Study design

To determine the optimal parameters to maximize sequence mapping and prepare for wet-lab experiments, we first optimized our bioinformatics pipeline using a simulation framework, a classical approach in computer science[19]. We simulated sequence data based on a set of known bacterial genomes as well as their abundance distribution based on existing data, which we varied in terms of composition, richness and sequencing depth. The simulator we used took into account the particularity of ONT sequences and biases. Next, we built and tuned our bioinformatics pipeline considering the best hyperparameters to minimize the difference between the estimated quantified features (abundance, richness) with those used to parameterize the simulation **(Figure 1A)**.

In parallel to the bioinformatic pipeline optimization, we conducted wet lab experiments to establish an optimized protocol for stool collection and DNA extraction, DNA fragmentation, and DNA end-repair on microbiome composition. Using the results of these experiments in coordination with our bioinformatics pipeline, we proposed an optimized protocol for DNA quality and increased ONT sequencing read length (**Figure 1B**). Finally, we validated our protocol and pipeline using human stool samples sequenced in parallel with ONT and Illumina technology and previously sequenced with SOLiD technology (**Figure 1C**).

Simulated ONT microbiome data

We designed the data simulation process to maximize representation of real human gut microbial ecosystems. We used 506 reference genomes that were used with

metagenomic assemblies to build the IGC[1]-[20]-[21] human gut gene catalog. We simulated ten samples (M1:M10) whose abundances were calculated using a Pareto distribution, which was estimated using real metagenomic profiles[8] based on the abundance of metagenomic species[20] computed on the same IGC catalog[21]. These empirical distributions were used to define generation probabilities of the 506 reference genomes. We included two important elements 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 in microbial communities that could be observed ranging 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)[22] (**Figure 1A**).

The bioinformatics workflow

The proposed bioinformatic workflow for quantitative metagenomic (QM) analyses from ONT shotgun sequencing starts with fast5 files generated by the MiniKNOW^{tm} software. The Fast5 file format is derived from the HDF5 standard and has a hierarchical structure in which the metadata associated with individual reads (i.e. position in the flow cell, start time of the sequencing of the DNA fragment, time elapsed to pass through the flow cell pore and when it ends) as well as the events (i.e. aggregated bulk current measurements for each DNA fragment) are stored and then decoded by base calling algorithms into fastq format. The first step of the workflow consists of base calling and demultiplexing the fast5 files into fastq files. Here we used Albacore and Guppy ONT base callers from the community site available to ONT

customers[23]. Since the start of our work, ONT had discontinued development on Albacore in favor of the more performant Guppy[24].

Secondly, the *sequence_summary.txt* files, generated during the base calling step, are used to generate different QC visualizations along with information embedded in the fast5 files. This allows evaluating the quality of the sequencing run in terms of number of active channels in the flow cell, the distribution of active channels through time, the yield in terms of reads of the run and the read length distribution. The third step of the workflow consists in the taxonomic binning of ONT reads using two different reference resources. The first one uses Centrifuge [25] for the taxonomic binning of individual reads using a comprehensive reference database of more than 8000 reference genomes from prokaryotes and viruses (including the reference human genome). Thus, this step allows excluding human sequence reads. To remove spurious taxonomic assignments, we additionally mapped against the corresponding reference genome from centrifuge database using Minimap2 with *map-ont* option optimized for ONT reads [26]. Based on simulation experiments results, only sequences with a minimum mapQ score of 5 were retained for subsequent analyses. A species abundance table was generated by summing the counts of each NCBI taxid from the filtered Centrifuge results. This abundance table was combined with the experiment metadata information and a reference taxonomic table reconstructed from Centrifuge NCBI taxids using the R package *taxize*[27] into a S4 *phyloseq* object. This object was used to carry out standard microbial ecology analyses (rarefaction, alpha-diversity, beta-diversity, and differential abundance analysis) with *Phyloseq* or *Vegan*R packages [28].

A complementary approach consisted of quantifying the abundance of microbial genes. For this, ONT reads were aligned against the IGC [21] gene catalog using

Minimap2 with *map-ont* option [26]. The alignment of long ONT reads over short or fragmented IGC genes provided a preliminary structural annotation of the reads in the sense that a long read could cover a genomic region harboring more than one gene, but also provided situations of multiple genes mapping in overlapping regions of a read. Overlaps were filtered out using *GenomicRanges* and *plyrRanges*R packages[29]-[30] allowing to quantify 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 read.

The final step consisted in quantifying functional features (i.e. taxonomic binning). Functional features were KEGG orthology groups (KO groups), quantified from the gene abundance table, using available reference annotation from the IGC catalog [31]. For taxonomic results produced from Centrifuge quantification, we retrieved the KO content of KEGG genomes from the KEGG API [32] 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 [25].

Study participants for wet-lab experiments

Stool samples used for wet-lab protocol optimization were collected from healthy 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 samples from the

Microbaria study, where the gut microbiome of subjects with severe obesity was characterized before and after bariatric surgery[8].

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 the duration of storage. For the latter collection method, we tested four available commercial kits including "DNA/RNA Shield-Fecal Collection Tube" (Zymo marketed by Ozyme), "Stool Nucleic Acid Collection and Preservation Tubes" (Norgen Biotek) and "Omnigen Gut for Microbiome" (DNA Genotek).

To extract bacterial DNA, we tested five different commercial kits using manual extraction protocols: “PureLink ™ Microbiome DNA Purification Kit” (Invitrogen), “Qiamp PowerFecal DNA Kit” (Qiagen), “ZimoBiomics DNA Mini Kit” (Ozyme) and “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 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.

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 with flow cells on which 12 samples were loaded per run. Upon receipt of the flow cells and, prior to sequencing, pore counts were measured using the Platform QC script (MinKNOW from Version 1.4 and further). Flow cells were replaced into their packaging, sealed with tape, and stored at 4°C until further use.
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)[21].

Statistical 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 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. Permutational analyses of variance (PERMANOVA) with the *adonis* function of vegan R package [33] 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.

Results

Metagenome simulations identified key parameters for ONT microbiome quantification

The metagenome simulation approach allowed the evaluation of the impact of different bioinformatic parameters. The evaluation accuracy consisted of comparing the estimated abundance of the microbial features (i.e. species abundance) with the original values used before sequence generation. Over 100 million long reads were generated out of the 250 simulated metagenomic samples from 10 different microbiome compositions at 5 different levels of community complexity for species richness and 5 different levels of sequencing depth (see methods; additional files **Tables S1** and **S2**). These reads were aligned against the 506 reference genome catalog using minimap2 aligner with the map-ont configuration, designed for optimal performance and accuracy with ONT sequencing data [26]. On average, 381000 reads per sample (94% of sample reads) were aligned against the reference genomes. We observed that the reads that could be aligned were 2.5 longer in size (average read length across 250 samples=8064 bp, sd=8) compared to those that could not (average read length=3168 bp, sd=24) (**Figure 2A, Additional file Table S3**), suggesting that read length was a key parameter to optimize ONT sequencing on reference genomes.

We evaluated the accuracy of the estimated species abundance and richness to the reference values used for the simulation. Taxonomic profiles were obtained from all minimap2 aligned reads as well as from all primary alignments (PA), defined as the best alignment of a single reads among all possible multiple alignments.

For species richness, we observed that both raw and PA-filtered reads detected all reference species in the simulated samples, reaching recall values close to 1 under all simulated community compositions (**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 with reference and simulated samples showed that PA-filtered samples were more similar to the corresponding reference distributions (R2 effect size dissimilarities=0.04, pvalue<0.05; Permanova test) than raw alignment samples (R2 effect size dissimilarities=0.51, p<0.05; Permanova test) (**Figure 2D**). This suggests that the noise introduced by secondary alignments decreased the precision of the workflow.

Alignment identity and read quality affected workflow precision

We evaluated the impact of filtering at different thresholds of identity alignment on the accuracy of the estimated microbiome profiles. Alignment sequence identity (i.e. the ratio of the number of matching bases over the number of bases in a read alignment, including gaps) is a commonly used parameter to filter read alignments, with values around 0.8-0.9 to filter small reads from 2nd generation sequencing technologies against reference genome sequences. This is particularly important for potential error-prone reads generated with ONT sequencing. For species richness, the recall values were close to 1 for identity levels up to 40%, meaning that all reference species in each simulated sample were detected by the workflow. When progressively increasing identity levels from 50% to 90%, a significant fraction of reference species was not detected (**Figure 3A**), although increasing the precision of the estimated richness as

the filtering lowered the number of false positives (**Figure 3B**). However, 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 3C**), despite the presence of false positives. Finally, as for the recall, we observed that the Spearman Rho correlation between the estimated species abundance and the reference values decreased as the alignment identity threshold increased. This was in agreement with ordination results (**Figure 3D**). Overall, these results showed that common approaches to filter read alignments used in the context of second-generation NGS technologies were not directly applicable to ONT sequencing data, probably a consequence of its high error rates. Additional parameters were needed to be explored in order to improve the accuracy of the resulting metagenomic profiles.

Relevance of mapping quality scores in workflow precision

The mapping quality score (mapQ) as computed by minimap2 provided relevant information. High values corresponded to long reads and for which the scores assigned to secondary alignments were weak when compared with primary alignments[26]. MapQ score distributions, based on primary alignments of simulated datasets (**Additional file S1**) were evaluated with 11 different thresholds from [0:50], 0 being no filtering. For species richness, we observed an inverse relationship between the stringency of mapQ filtering scores and recall and positive with precision of the workflow (**Figure 4A-B**). Importantly, we observed that mapQ score was more sensitive than the percentage of identity in estimating of species richness. For simulated samples with 50 species each at the lowest filtering threshold (mapQ=5), the achieved precision values were higher (mean=0.55, sd=0.1; Figure 4B) than those reached when filtering with the most stringent sequence identity threshold (90%

sequence identity; mean=0.49, sd=0.06; **Figure 3B**). Regarding community composition, our results displayed an inverse relationship between the filtering stringency and the similarity of the estimated communities with the reference simulated samples based on PCoA ordination (**Figure 4C**).

Finally, we explored the similarity between estimated species abundance and the reference values using Spearman correlations. We observed different results for different species richness of simulated samples. Whereas in simulated samples with low richness (R50, R150) the similarities between the estimated and the reference species abundance vectors increased with the mapQ threshold from 5 to 30, this was not the case for more complex samples, where the similarities did not improve as the mapQ threshold was increased further than mapQ=5. On the contrary, the similarity significatively decreased in simulated samples with 450 species (**Additional file S2**).

Based on these results and considering the trade-off between the accuracy of species richness estimates and the accuracy of microbiome composition for complex microbial communities, we proposed the threshold of mapq5 as the main criteria to process ONT sequence alignments for quantitative metagenomic studies. This filtering was used to process the subsequent experiments aiming to optimize the wet-lab protocols. Moreover, the sequencing of the ZymoBIOMICS mock community and its quantification 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 (**Additional file S3**). 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).

Steps to optimize 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.

DNA extraction kits influenced read length distribution

When testing all DNA extraction kits on dry spoon stool samples from healthy volunteers, all kits except the “ZimoBiomics DNA Mini Kit” (Ozyme) provided sufficient DNA quanity and quality for sequencing. Thus, we examined library preparation and sequencing all kits except the Ozyme kit. The Invitrogen and Mo Bio kits showed a higher portion of long reads (> 1.1 kb, **Figure 5A**). However, 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 examination of alpha diversity showed a significantly higher number of present species for the Invitrogen kit compared to the other kits tested (**Figure 5E**). Based on these observations, the Invitrogen kit was selected as the preferred extraction kit for sequencing.

DNA fragmentation and end-repair steps did not influence alpha diversity

The first step in ONT's library preparation protocol DNA fragmentation to generate 8kb fragments [34]. Using 3 different samples from one subject, DNA fragmentation had no effect on read length distribution (**Figure 5B**), alpha diversity (Wilcoxon rank-sum test P-value>0,05; **Figure 5F**), or microbiome composition based on PCoA ordination (**Figure 5I**). Therefore, we decided to exclude this step from our experimentation framework. The ONT DNA preparation protocol also 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. We found that sequence length (**Figure 5C**) and microbial diversity (**Figure 5G**) were not significantly affected by DNA-end repair. This step was then omitted from our proposed protocol.

DNA extraction kits affected workflow outcomes

According to a PCoA and Permanova analyses, we observed that the differences between microbiome composition of the different replicates were mainly explained by the sampling time (R2=0.45, p<0.001), followed by donor (R2=0.03, P-value<0.05) and DNA extraction kit (R2=0.02; P-value<0.01) (**Figure 5H**). Importantly, we observed no effect of sequencing runs nor fragmentation as well as end repair tests **(Figure 5I & 5J)**, suggesting no obvious batch effect.

We observed a bimodal distribution of read lengths across the Invitrogen and Mo Bio DNA library preparation kits, whereas with the Promega and Qiagen kits the distribution was skewed towards smaller reads **(Figure 5A)**. Based on simulation results, read length had a major impact on the fraction of reads that could be aligned against reference sequences. This ultimately impacted the yield of sequencing runs and the amount of information obtained for microbiome composition.

We observed that this was also the case with real data when we classified reads as short or long based on the median of log2-transformed read lengths and we compared the fraction of classified reads between both groups (**Figure 5A**). We found that the fraction of classified sequences was significantly higher for long reads (log2-length > 9.96; P-value<0.05; Paired Wilcoxon rank-sum 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))**. Based on the simulated and real data, we hypothesized that longer ONT reads would increase the accuracy of taxonomic binning. Consequently, in the next step, we aimed at increasing the read length by optimizing DNA extraction and library preparation using the Invitrogen kit.

Optimized DNA extraction protocol improved microbial diversity estimation

To optimize DNA extraction, we followed recommendations from the IHMS consortium [16]**(Figure 6A).** 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 [35] and ONT users' recommendations from the "Community" forum [23].

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 1.38 vs 2.11, respectively (P-value = 0.0007) and the absorbance ratio at 260/280 increased significantly 1.73 vs 1.89, respectively (P-value = 0.0046) (**Table 1**).

Table 1: Improvement of DNA yield and quality with the Optimized Invitrogen protocol.

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

Nanodrop data Estimation of the amount of DNA, protein contamination by given by the 260/280 ratio and by impurities and solvents contamination estimated by the 260/230 ratio (n = 6, Wilcoxon test).

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. With this new dataset, we confirmed that the fraction of classified reads was higher for long ONT reads (**Figure 7B**; P-value= $4.88e^{-04}$, paired Wilcoxon rank-sum test). We observed that 29.72% of the reads with the optimized Invitrogen protocol were successfully classified after the processing with the Centrifuge strategy in comparison with 23.92% of the reads with the standard protocol (P-value=0.031,

paired Wilcoxon rank-sum test). We observed an increased microbial diversity in the samples with the optimized protocol although these differences were not statistically significant (P-value=0.31; Paired Wilcoxon rank-sum test) (**Figure 7C**). The PCoA (**Figure 7D**) confirmed that differences in microbiome composition across samples are explained mainly by subjects. 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 and to a more accurate microbial diversity assessment.

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 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 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). For each sampling kit, we tested the effect of sample storage temperature on DNA extraction and did not observe any significant difference in the distribution of classified reads or microbial diversity in any of the three different collection kits at -80°C (P-value=0.73; Kruskal-Wallis test) or at room temperature (P-value=0.44; Kruskal-Wallis test) **(Figure 8B)**, 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=0.0018, Kruskal-Wallis test). Similarly, donor was 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.

Optimized ONT protocol compared to main stream sequencers

We compared ONT obtained QM profiles with those generated with other sequencing technologies from the Microbaria study[8]. We selected 33 baseline 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 above, based on Centrifuge and mapping over the IGC catalog. DNA from 21 of the 33 samples were extracted with the optimized Invitrogen protocol and sequenced using Illumina technology. Quantitative metagenomic profiles were generated by mapping reads against the IGC gene catalog [8].

First, we compared the estimates of microbial diversity from ONT (GeneRichness 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 diversity estimates based on ONT sequencing with the reference gene richness estimates from SOLiD sequencing based on IGC gene catalog (Spearman Rho=0.59, P-value=3e-04 for Observed Species based on Centrifuge results; Spearman Rho=0.74, P-value=2e-06 for Gene Richness based on ONT read mapping over IGC catalog; **Figure 9A & 9B**). These similarities increased with the use of reference databases maximizing the genomic information of the gut microbiome ecosystem (gut microbiome gene catalogs vs. generic Centrifuge database). However, the similarity was higher with gene richness estimates based on Illumina sequencing despite differences in library preparation (Spearman Rho=0.86, p<2.2e-16; **Figure 9C**). When we integrated the scaled diversity profiles of different sources available for these 33 samples ordering them based on the reference gene richness[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. This included a switch in the sample showing the highest diversity (i.e. MB12 with Illumina and ONT sequencing; MB21 with SOLiD sequencing; **Figure 9D**).

This 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 (**Additional file S4A**) and in hierarchical clustering analyses (**Additional file S4B**).

Potential of the ONT workflow for the detection of 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 in 91 of the 95 common taxonomic features with SOLiD quantifications (96%), of which 72 (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 quantifications (93%), of which 78 (77 %) were significantly associated (FDR<0.05, Spearman correlations) (**Figure 9E**). Using IGC results at MGS level, 137 common taxonomic features with SOLiD quantifications were positively associated, 128 of which (93%) were significantly associated (FDR<0.05, Spearman correlations), whereas 133 of the 137 common taxonomic features with Illumina quantifications were positively associated (98 %), 122 of which (90%) were significantly associated (FDR<0.05, Spearman correlations) (**Additional file, Table 4**). Importantly, we observed that the similarities between taxonomic features between ONT and Illumina quantifications were higher than between ONT and SOLiD sequencing (**Additional file S5;** $p<0.05$ for comparisons at species and genus level with ONT Centrifuge results; $p<0.005$ 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 (**Additional file S6A**). This difference may be related to the different content of both genomic reference spaces (Centrifuge genomes;

IGC gene catalog), which can have a major impact on the quantification of functional modules if differences in composition also result in differences in gene content. Importantly, we observed that DNA extraction also impacted the functional profiles, with ONT functional profiles being more similar to Illumina functional profiles based both on Centrifuge and IGC quantifications (**Additional file 6B**; P-value=0.0046 for Centrifuge-based quantifications; P-value=6.1e-14 for IGC-based quantifications; Wilcoxon rank-sum test of Spearman's Rho distributions between ONT-SOLiD comparisons and ONT-Illumina comparisons).

We reproduced these associations at the functional level. We found significant positive associations between the sporulation module md:M00485 (KinABCDE-Spo0FA (sporulation control) two-component regulatory system) and microbial diversity (**Additional file S7**), which was in agreement with estimations of 50%-60% of bacteria from gut microbiome of healthy individuals producing resilient spores. This sporulation phenotype is an unappreciated and basic feature of the human microbiome with a key impact in bacterial persistence and the spread of microbes between individuals [36]. This was also the case for the negative association between modules involved in the biosynthesis of bacterial Lipopolysaccharide (LPS) and microbial diversity (**Additional file S8**), in line with the association of obesity and other metabolic disorders with an increase of blood LPS concentration[37].

The ONT workflow enabled de novo genome assembly

We finally explored the potential of ONT sequencing for de-novo assembly of bacterial genomes from shotgun metagenomics data. We carried out de-novo assembly of the genomes of *Bacteroides vulgatus* (a bacterial species with the highest mean abundance based on Centrifuge quantifications, **Additional file** S9A) and *Akkermansia muciniphila*, a bacterium whose abundance in this cohort was lower than

in other overweight or obese cohorts [38]. We observed that de-novo assemblies from the corresponding read bins product of Centrifuge workflow gave genomic scaffolds that covered 88% of the *Bacteroides vulgatus ATCC 8482* genome at 99% Average Nucleotidic Identity (ANI) and 78% of the *Akkermansia muciniphila isolate Urmite* at 95% ANI (**Additional file S9B & S9C**). These reference genomes were the closest relatives to the assembled ONT genomes based on mash genomic distances vs. RefSeq genomes[39]. The structural annotation of these genome assemblies also revealed a highly fragmented proteome, which is reflected in the smaller gene lengths observed in ONT assemblies in comparison with the corresponding reference genomes (**Additional file S9D & S9E**) and in the large number of genes in ONT assemblies in comparison with the corresponding reference genomes (15532 CDS in *Bacteroides vulgatus* assembly vs. 4104 in *Bacteroides vulgatus ATCC 8482* genome; 2900 CDS in *Akkermansia muciniphila* assembly vs. 2101 CDS in *Akkermansia muciniphila isolate Urmite* genome). Thus, we were able to recover near complete genomes for individual species bins directly from ONT metagenome sequencing of stool samples but the error-prone character of this sequencing technology prevented accurate structural annotation of these genomes uniquely with ONT sequencing data.

Discussion

We present here a novel protocol and analytical pipeline enabling the quantification of the gut microbiome features using Oxford Nanopore Technologies. This technology potentially 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 established protocol enables the study of the gut microbiome in the context of clinical applications or group studies. To generate optimal results, we optimized protocols for the wet-lab (from DNA extraction to sequencing) and data analysis. We also compared the final results to state-of-the-art sequencing methods (Illumina and SOLiD) in an already 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 that 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. 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. Also, the alignment quality based on Nanopore-adapted sequence aligners like minimap2 allowed for improved accuracy in the estimation of microbial diversity and composition in complex ecosystems such as the human gut.

Regarding sample processing, a first step required the elaboration of a DNA extraction protocol from human stools that provide high DNA quality. Several 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],[40],[41]. Authors in

these areas have used different DNA extraction protocols and different sequencing techniques (Illumina, SOLiD, Ion Proton). Multiple studies have also noted “batch”[42] effects and differences in data analyses [43]-[44], which introduce biases in study analytical comparisons. Thus, the need for procedure standardization has been highlighted by several reports, as illustrated by the IHMS consortium[45]. In this study, the authors compared 21 DNA extraction methods using whole genome metagenomic shotgun sequencing with Illumina HiSeq2000 technology. They 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 "Standard Operating Procedure "SOPs" with recommendations that would improve DNA extraction in terms of yield and quality.

We applied these IHMS recommendations to the PureLink™ Microbiome DNA Purification kit from Invitrogen, and we further optimized the microbial DNA extraction protocol. The DNA yield from the optimized protocol was significantly improved, compared to the conventional kit in our hands. We worked on two critical steps, bacterial wall lysis and protein/RNA elimination. Two rounds of lysis process gave two supernatants with higher DNA yield. DNA recovery by isopropanol precipitation followed by elimination of protein and RNA contamination allowed further DNA purity. We also explored other methods of DNA stabilization. Sampling conditions, storage and harmonization have previously been shown to be critical in affecting microbiome results. Although storing fecal samples at 4°C appeared to protect bacterial DNA from degradation to some extent, a reduction in microbial diversity was observed [46]. 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[47]. However, these studies were conducted before the development

and widespread use of commercially available fecal collection kits with stabilizing solution. In our current study, our results suggest that sample storage temperature is not a significant factor as long as guidelines from manufacturers are followed. Furthermore, 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 selected the DNA extraction protocol and improved the library preparation protocol to increase the proportion of long reads. For this purpose, we optimized the end-prep and DNA purification steps in the library. Applying this approach, we obtained a unimodal read length distribution curve and observed an increase of the proportion of long reads. 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. We observed that 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 developed in this study, which emphasizes the importance of DNA library preparation protocols in quantitative metagenomic profiles in our experimental design.

The long reads obtained with ONT sequencing allowed us to perform "de novo" assemblies for the most abundant species in the Microbaria cohort (*B. vulgatus*) as well as potential metabolic disease-relevant bacteria such as *Akkermansia muciniphila,* whose abundance in the sequenced samples is lower in comparison with other less obese or overweight individuals, and is associated with metabolic health[38]-[48]. These assemblies showed a high level of completion when compared with closest relatives, highlighting the strong potential of ONT sequencing in the assembly of bacterial genomes directly from shotgun metagenomics data. In comparison, the assembly of the *Akkermansia muciniphila* genome in a previous study required "SOLiD" and shotgun sequencing to cover the genome with a scaffold and 56 gaps[49]. However, the structural annotation of the assembled genomes revealed a caveat associated with this sequencing technology, which is the poor quality of the assemblies, that we observed, with highly fragmented genes as a consequence of the presence of frameshifts in the assembled sequences (as shown in **Additional file S8**). This is explained by the low read accuracy of ONT reads in comparison with short-read technologies, which makes hybrid strategies combining long reads product of ONT sequencing with small reads product of short-read sequencing the most prevalent approach for accurate genome assemblies [50]-[51].

It is also important to mention that longer reads generated from ONT did not compensate for the overall sequence coverage in comparison with other sequencing technologies. The low throughput of ONT sequencing is one of its major drawbacks for quantitative metagenomic studies of complex microbial ecosystems like the human gut microbiome. However, the results obtained showed a high similarity in bacterial diversity estimation between the two sequencing methods in our test samples. Despite improvements in experimental protocol such as DNA extraction and library

preparation, 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[52].

Nanopore-based technology such as that used by ONT, 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%)[53] compared to Illumina (0,1%)[54]. However, in our current work, ONT consistently replicated results for intestinal microbiome diversity and the composition of main phyla in patients with severe obesity.

Conclusion

ONT should be considered as a suitable sequencing method in a clinical setting, and we propose to pair this technology with AI-based approaches in the future, which may allow the identification of interpretable and accurate signatures for disease diagnosis and prognosis in an effort to adapt future personalized microbiome based therapies [55].The present work indeed provides a useable and complete experimental and analytical workflow for the use of ONT sequencing in quantitative metagenomic studies. Both computational and wet-lab steps has been benchmarked to improve the precision of quantitative metagenomic profiles and microbial diversity estimates. The developed bioinformatic pipeline allows the quantification of metagenomic features in different reference spaces (non-redundant gene catalogs, reference genomes with Centrifuge) from raw ONT FAST5 files, including functional characteristics of metagenomic species. This is a key dimension of metagenomic analyses to understand the impact of variations in the composition of the microbiome from a functional point of view. Wet lab experiments revealed that it is DNA extraction technique rather than the sequencing technology what predominantly drives the observed variation in microbial diversity and microbiome composition. Moreover, despite variations in library preparation metagenomic profiles inferred from ONT sequencing, data appears close to those obtained with Illumina sequencing in terms of microbial diversity and taxonomic and functional profiles. Through simulations and wet lab experiments read length emerged as a critical parameter to optimize in order to maximize the classification of ONT reads. The experimental protocol developed allowed this optimization providing not only a correct coverage of the taxonomic profiles with the reduction of unclassified reads, but also in the assembly of near complete microbial genomes even from limited number of reads. This could contribute

to improve the knowledge and completeness of microbial genome assemblies of complex ecosystems like the human gut.

Overall, despite caveats and limitations of ONT sequencing regarding throughput and sequencing errors, 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.

Ethics approval and consent to participate

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.

Availability of data and material

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

Competing interests

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.

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.

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.

Acknowledgements

We thank Tim Swartz (Integrative Phenomics) for critical reading of the manuscript and English correction.

Authors' information

Eugeni Belda is employee of Integrative phenomics but this activity is not related to the topic of this publication

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. Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. Proc Natl Acad Sci U S A. 2020.
20. 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.
21. 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.
22. Fritz A, Hofmann P, Majda S, Dahms E, Dröge J, Fiedler J, et al. CAMISIM: simulating metagenomes and microbial communities. Microbiome. 2019;7:17.
23. https://nanoporetech.com/community.
24. 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.
25. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26:1721–9.
26. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
27. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Res. 2013;2:191. doi:10.12688/f1000research.2-191.v2.
28. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE. 2013;8:e61217.
29. 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.
30. 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.
31. http://meta.genomics.cn/meta/home.
32. 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.
33. https://CRAN.R-project.org/package=vegan.
34. https://community.nanoporetech.com/protocols/native-barcoding-genomic-dna/.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.

40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. http://www.microbiome- standards.org/.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. Next Gener Seq Appl. 2014;1.
55. 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.

Figures :

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.

Figure 2: Metagenomic profiles from simulated samples between minimap2 results and minimap2 results filtered from secondary alignments. (A) Boxplots of mean lengths of Nanopore 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, x-axis). (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).

Figure 3: Impact of filtering minimap2 primary alignments of Nanopore 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 Nanopore 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) PCoA of reference metagenomic profiles and metagenomic profiles of the 250 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. (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.

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 Nanopore 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) PCoA of reference metagenomic profiles and metagenomic profiles of the 250 simulated samples inferred from minimap2 primary alignments filtered by different thresholds of mapQ score (from 0 to 50; x-axis). Dashed lines connect points coming from the same sample (reference, simulated ones) with different shapes assigned to samples from different reference species richness. Points corresponding to reference samples and simulated samples with no filtering by mapQ (0) and filtered by mapQ>5 (5) are highlighted with larger point sizes. (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.

Figure 5: DNA extraction kits, fragmentation and end repair impact over human stool metagenomic composition from Nanopore sequencing data. Read length distributions of nanopore 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 by DNA fragmentation (n=4 paired samples). (G) Differences in microbial diversity 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 E. (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. ***=Pvalue<0.0001 in paired Wilcoxon rank-sum test

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.

Figure 7: Impact of Invitrogen optimized protocol over human stool metagenomic composition from Nanopore sequencing data. (A) Read length distributions of ONT reads across different DNA extraction kits including the optimized

Invitrogen protocol. 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 Invitrogen optimized kit and original Invitrogen kit (n=6 paired samples; *=P-value<0.05, Paired Wilcoxon rank-sum test). (C) Differences in microbial diversity between Invitrogen optimized kit and original Invitrogen kit (n=6 paired samples; ns=P-value>0.05, Paired Wilcoxon rank-sum test). (D) PCoA ordination 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.

Figure 8: Collection kits and storage conditions Impact 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. (D) Differences in microbial diversity 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).

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.

Additional files

Additional file 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).

Additional file 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).

Additional file S3: 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)

Additional file S4: 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=34 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.

Additional file S5: 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.

Additional file S6: 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.

Additional file S7: 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.

Additional file S8: 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.

Additional file S9: Summary of *Bacteroides vulgatus* and *Akkermansia muciniphila* genomes the assemblies from Microbaria Nanopore sequencing data. (A) Mean ± standard error of the top 26 bacterial species with the highest abundance in 33 Microbaria samples based on Centrifuge workflow. (B) MUMMER alignment dot plot between *Bacteroides vulgatus* assembly from Nanopore (ONT) reads (y axis) and the reference genome of *Bacteroides vulgatus ATCC 8482* (x axis). Each point represents a contig in y-axis matching the reference genome on x-axis, with red representing the same orientation and blue representing inverse orientation of contig in y axis vs. reference genome. Points across the diagonal represents assembled contigs collinear with the reference genome on x-axis. (C) MUMMER

alignment dot plot between *Akkermansia muciniphila* assembly from Nanopore (ONT) reads (y axis) and the reference genome of *Akkermansia muciniphila isolate Urmite* (x axis). (D) Comparison of log2-gene length distributions between genes from *Bacteroides vulgatus* Nanopore (ONT) assembly and genes from *Bacteroides vulgatus ATCC 8482* reference genome. (E) Comparison of log2-gene length distributions between genes from *Akkermansia muciniphila* Nanopore (ONT) assembly and genes from *Akkermansia muciniphila isolate Urmite* reference genome.

Figures

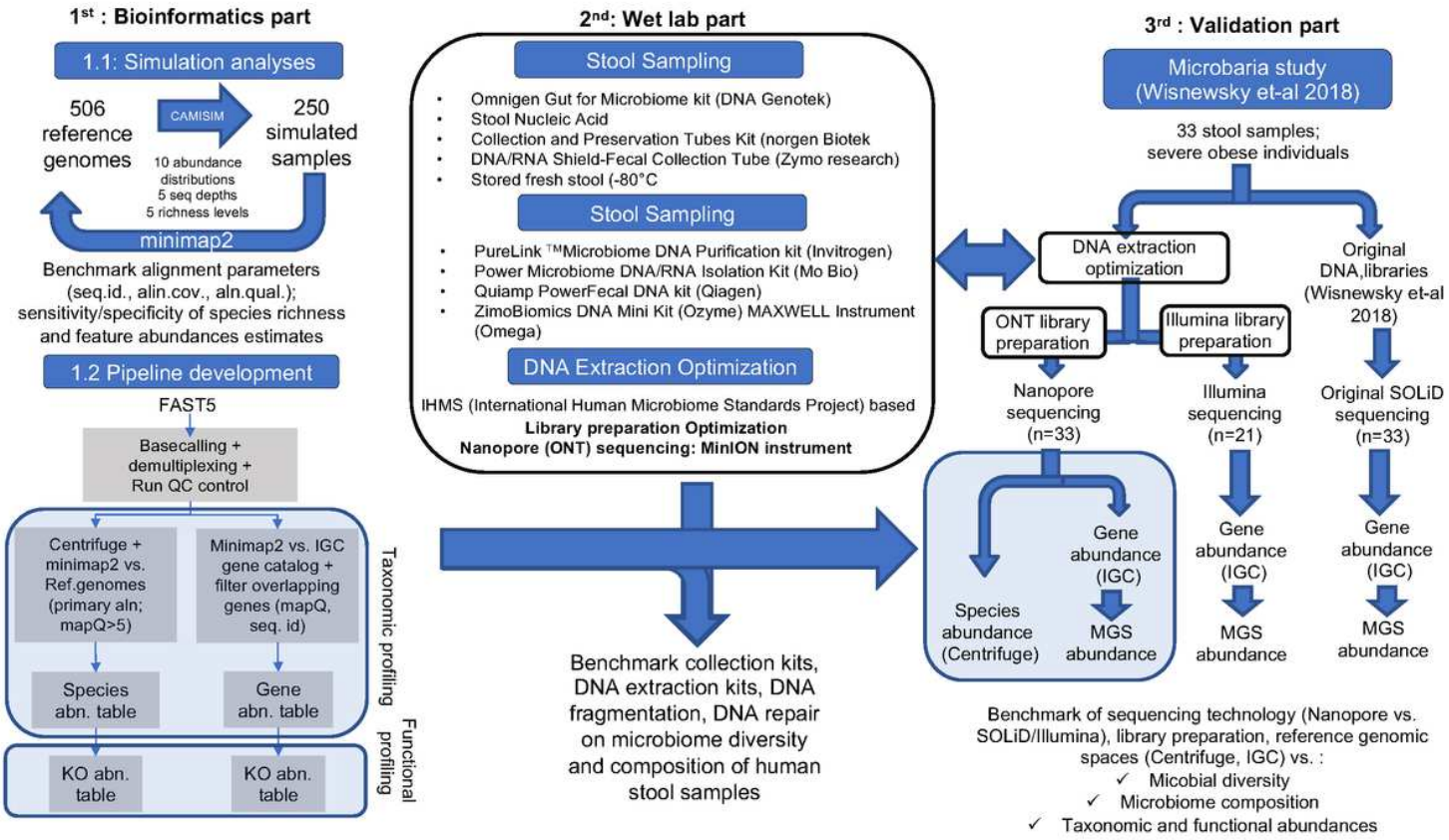

Figure 1

Summary of the workflow : (A) Simulated data processing. (B) Wet-lab optimization. (C) Summary of ONT sequencing comparison with Illumina and SOLiD technologies.

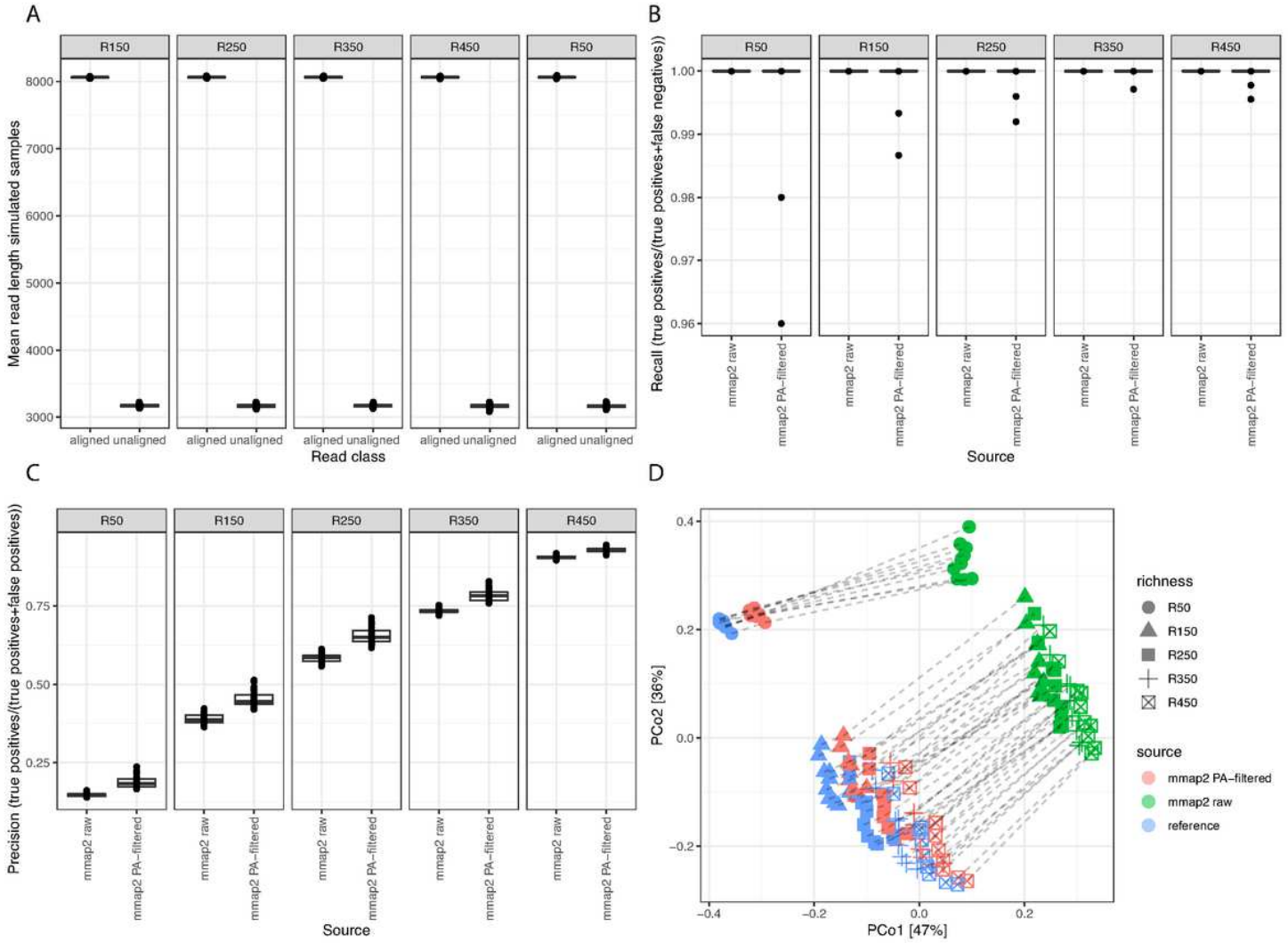

Figure 2

Metagenomic profiles from simulated samples between minimap2 results and minimap2 results filtered from secondary alignments. (A) Boxplots of mean lengths of Nanopore 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, x-axis). (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).

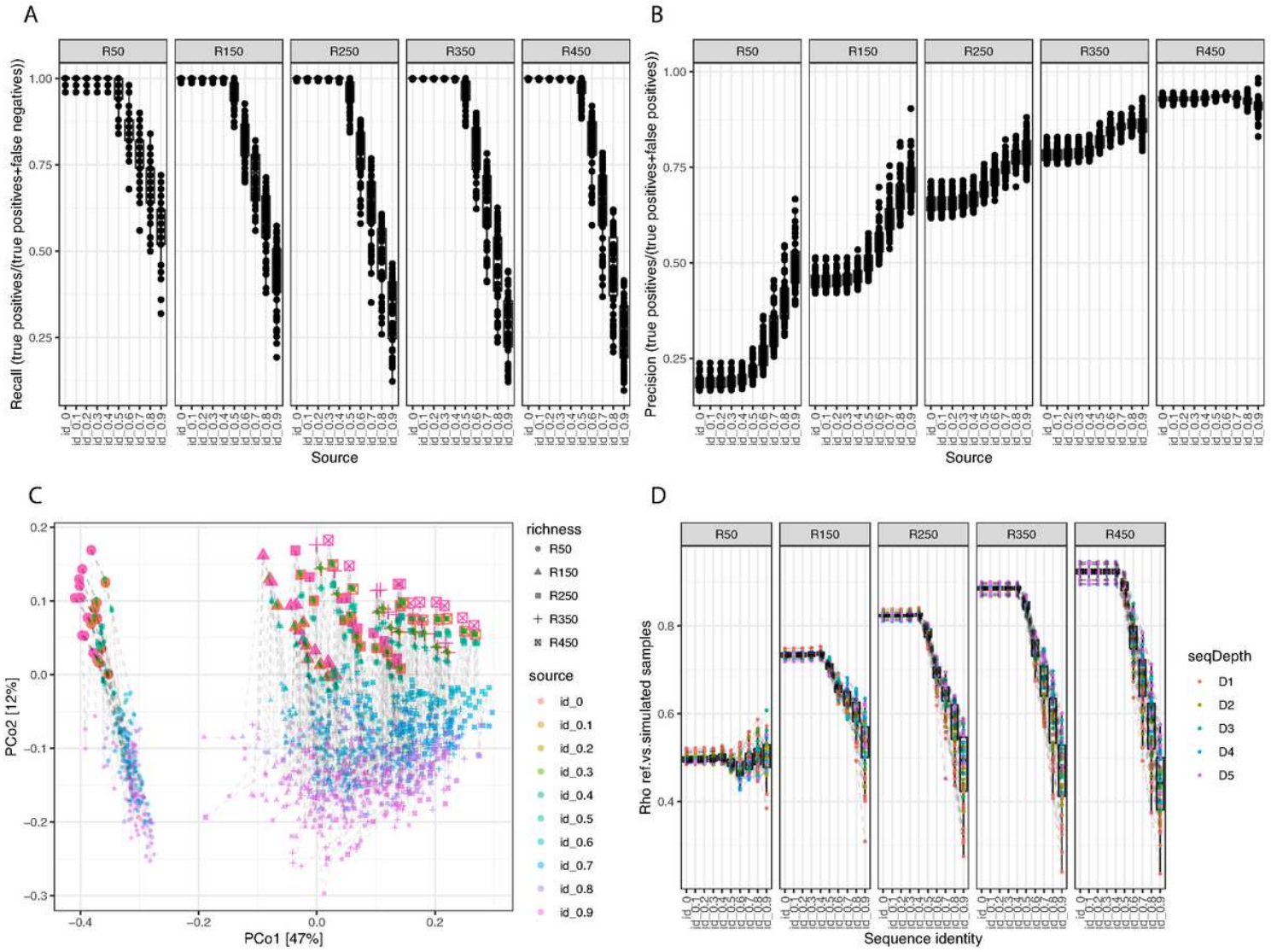

Figure 3

Impact of filtering minimap2 primary alignments of Nanopore 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 Nanopore 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) PCoA of reference metagenomic profiles and metagenomic profiles of the 250 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. (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.

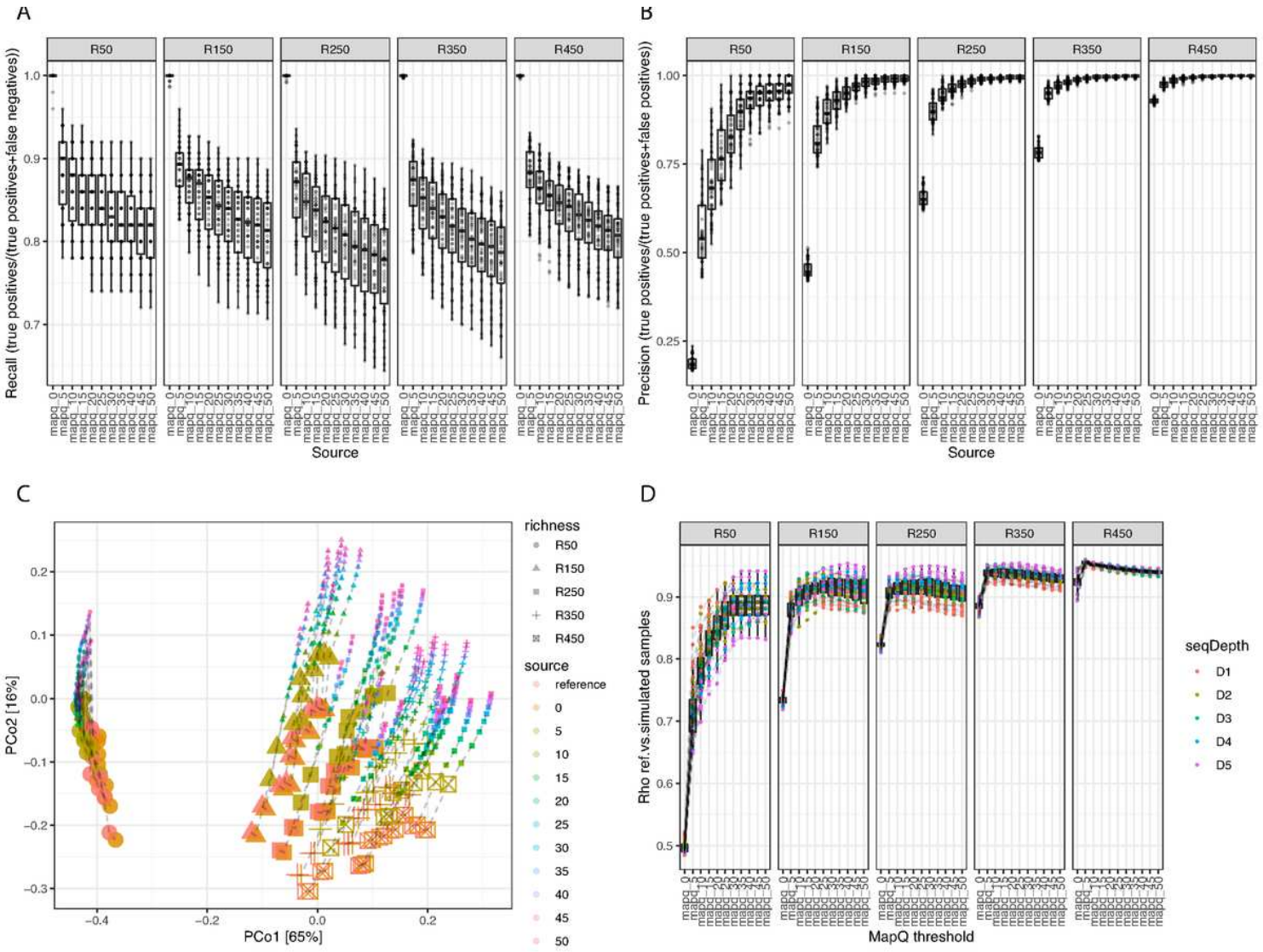

Figure 4

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 Nanopore 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) PCoA of reference metagenomic profiles and metagenomic profiles of the 250 simulated samples inferred from minimap2 primary alignments filtered by different thresholds of mapQ score (from 0 to 50; x-axis). Dashed lines connect points coming from the same sample (reference, simulated ones) with different shapes assigned to samples from different reference species richness. Points corresponding to reference samples and simulated samples with no filtering by mapQ (0) and filtered by mapQ>5 (5) are highlighted with larger point sizes. (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.

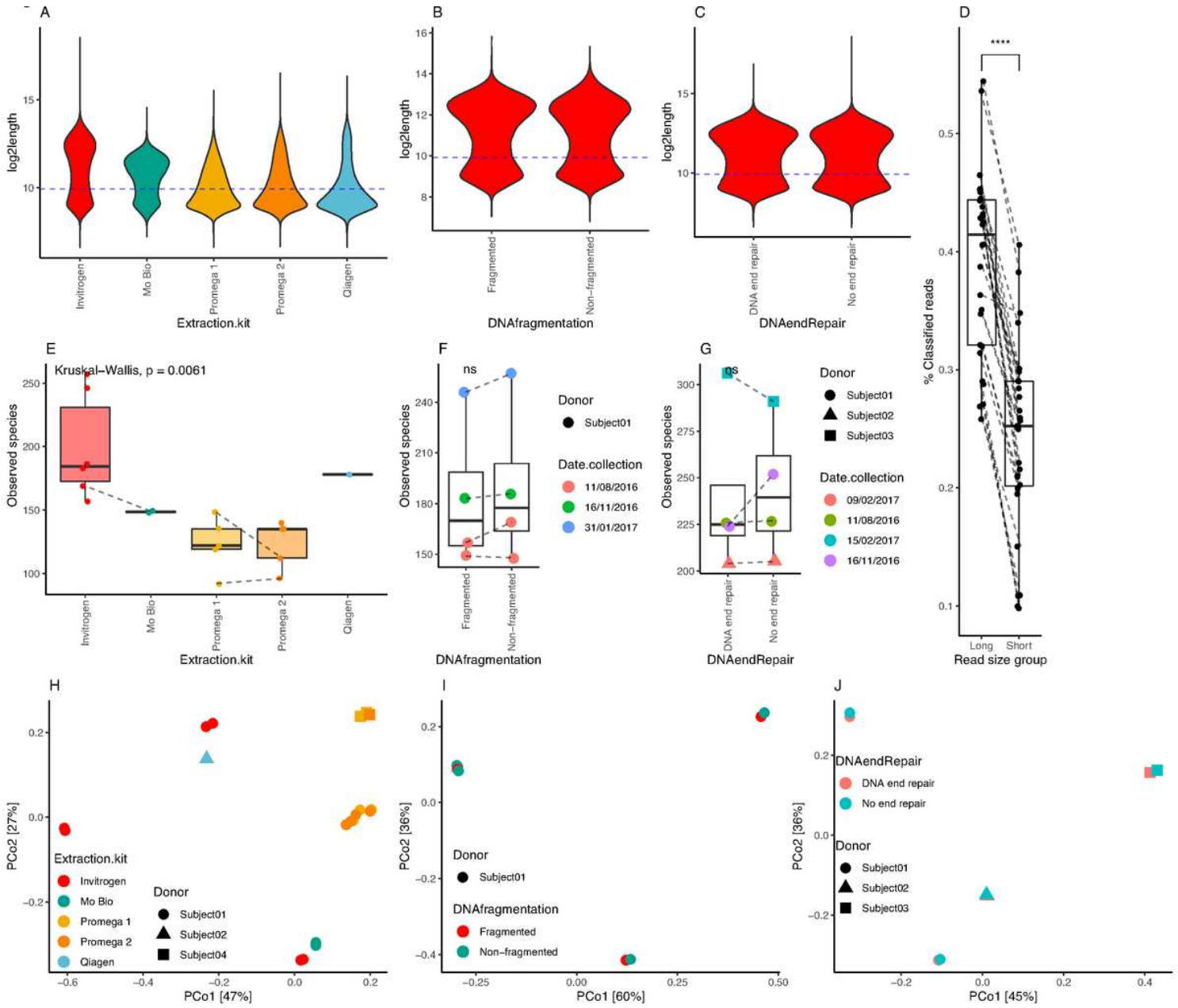

Figure 5

DNA extraction kits, fragmentation and end repair impact over human stool metagenomic composition from Nanopore sequencing data. Read length distributions of nanopore 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 by DNA fragmentation (n=4 paired samples). (G) Differences in microbial diversity 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 E. (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. ***=Pvalue<0.0001 in paired Wilcoxon rank-sum test

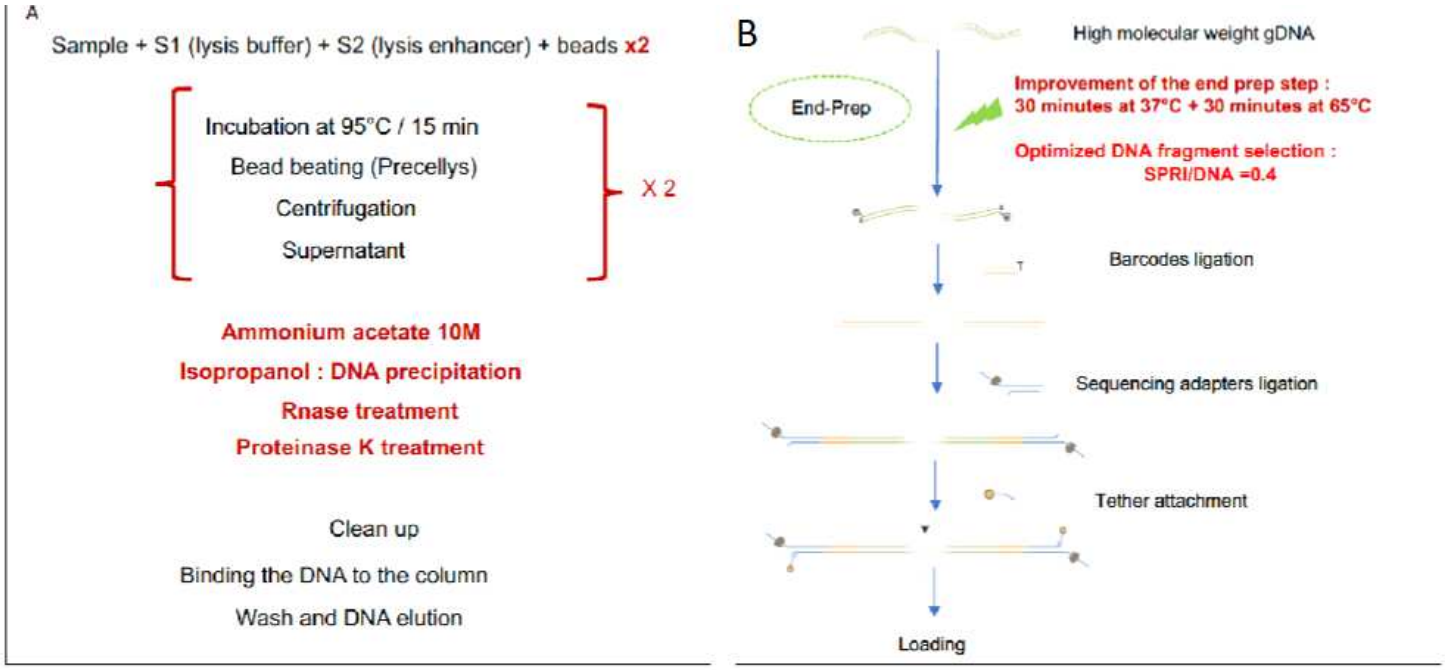

Figure 6

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.

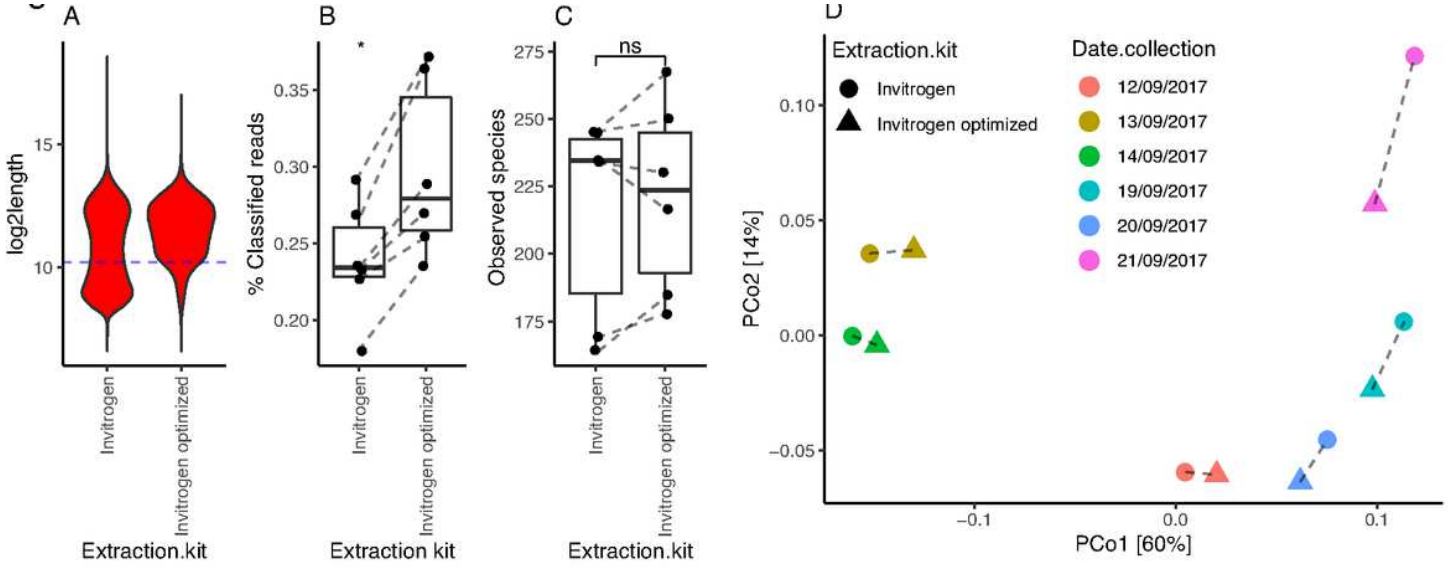

Figure 7

Impact of Invitrogen optimized protocol over human stool metagenomic composition from Nanopore sequencing data. (A) Read length distributions of ONT reads across different DNA extraction kits

including the optimized Invitrogen protocol. 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 Invitrogen optimized kit and original Invitrogen kit (n=6 paired samples; *=P-value<0.05, Paired Wilcoxon rank-sum test). (C) Differences in microbial diversity between Invitrogen optimized kit and original Invitrogen kit (n=6 paired samples; ns=P-value>0.05, Paired Wilcoxon rank-sum test). (D) PCoA ordination 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.

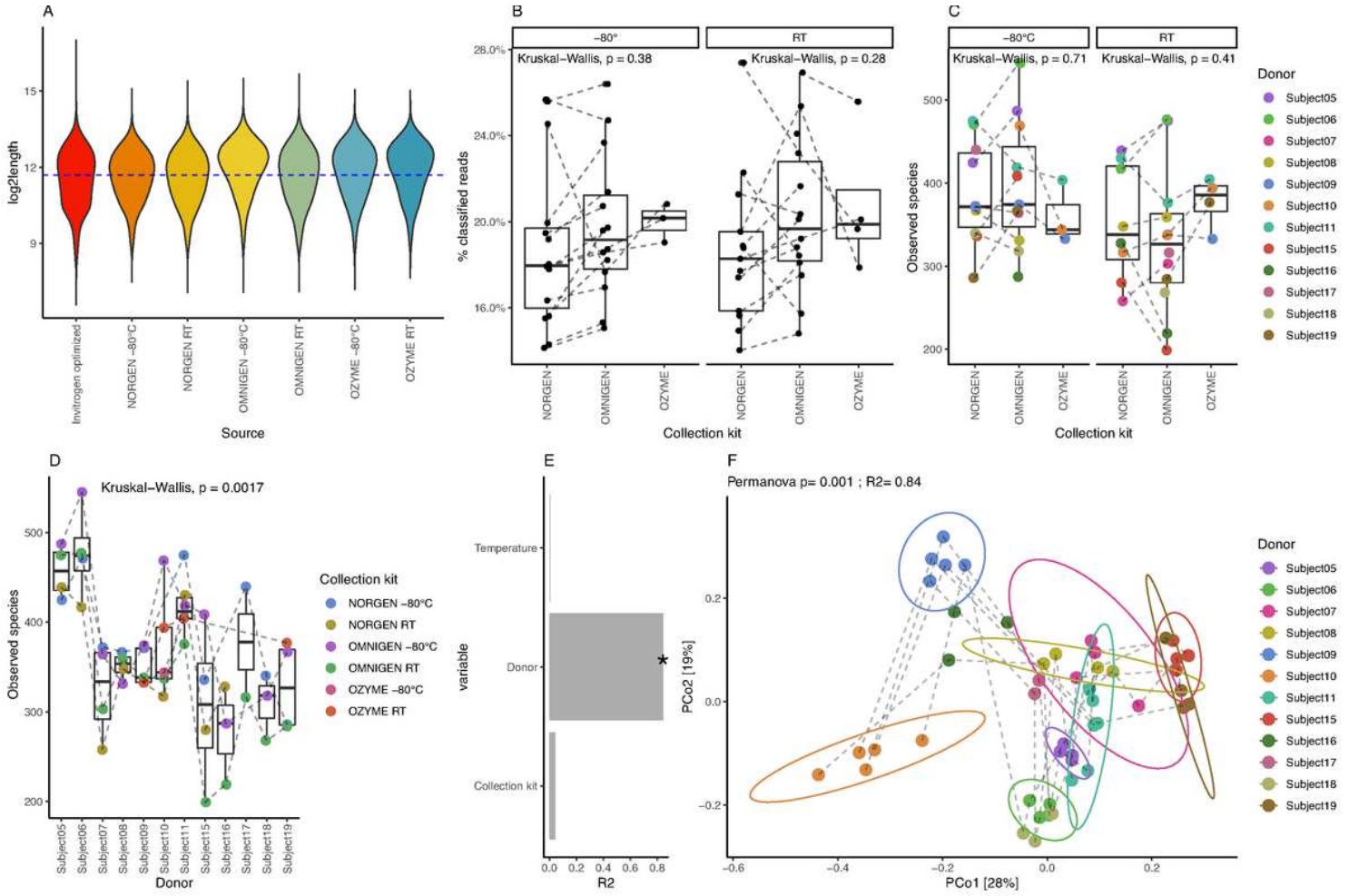

Figure 8

Collection kits and storage conditions Impact 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. (D) Differences in microbial diversity 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).

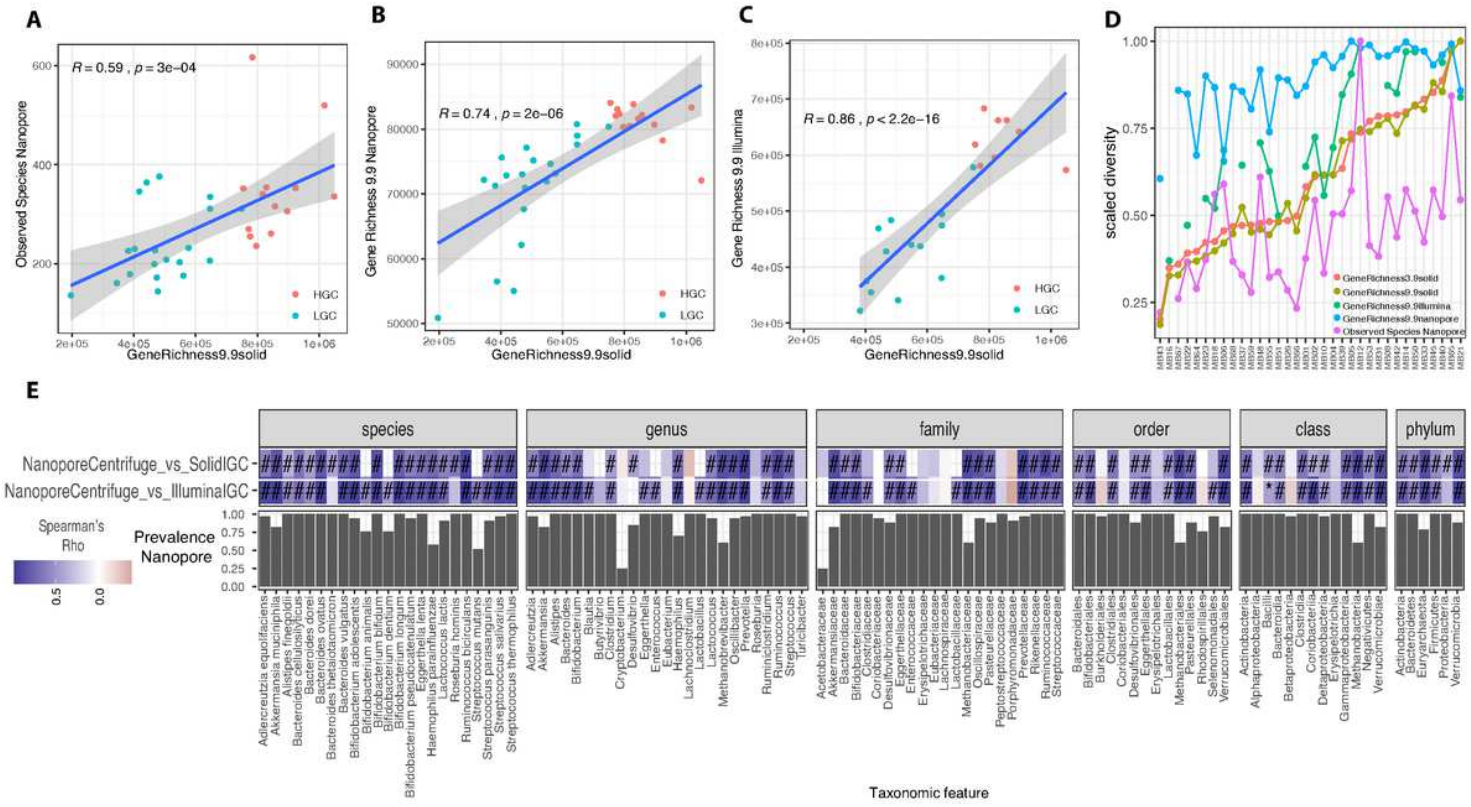

Figure 9

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplFigure1.pdf
- SupplFigure2.pdf
- SupplFigure3.pdf
- SupplFigure4.pdf
- SupplFigure5.pdf
- SupplFigure6.pdf
- SupplFigure7.pdf
- SupplFigure8.pdf
- SupplFigure9.pdf
- SupplementalTables.xlsx