Characterizing the limits of shallow shotgun metagenomics for taxonomic profiling of human gut microbiota in clinical studies

Benoit Goutorbe, Anne-Laure Abraham, Mahendra Mariadassou, Anne Plauzolles, Ghislain Bidaut, Philippe Halfon, Sophie Schbath

Characterizing the limits of shallow shotgun metagenomics for taxonomic profiling of human gut microbiota in clinical studies

Benoit Goutorbe (✉ benoit.goutorbe@inrae.fr)
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas
Anne-Laure Abraham
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas
Mahendra Mariadassou
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas
Anne Plauzolles
Clinical Research and R&D Department, Laboratoire Européen Alphabio, 13003, Marseille
Ghislain Bidaut
Centre de Recherche en Cancérologie de Marseille (CRCM), Université Aix-Marseille U105, Inserm UMR1068, CNRS UMR7258, Institut Paoli Calmettes, 13009, Marseille
Philippe Halfon
Clinical Research and R&D Department, Laboratoire Européen Alphabio, 13003, Marseille
Sophie Schbath
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas

Research Article

Keywords: microbiota, metagenomics, shotgun metagenomics, shallow shotgun metagenomics

Posted Date: July 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1306026/v1

RESEARCH

Characterizing the limits of shallow shotgun metagenomics for taxonomic profiling of human gut microbiota in clinical studies

Benoit Goutorbe[1,2,3*]
, Anne-Laure Abraham[1]
, Mahendra Mariadassou[1]
, Anne Plauzolles[2]
, Ghislain Bidaut[3]
, Philippe Halfon[2]
and Sophie Schbath[1]

*Correspondence:
benoit.goutorbe@inrae.fr
[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
[2]Clinical Research and R&D Department, Laboratoire Européen Alphabio, 13003 Marseille, France
[3] Cibi plateform, Centre de Recherche en Cancérologie de Marseille (CRCM), Institut Paoli-Calmettes, Aix-Marseille Université U105, Inserm U1068, CNRS UMR7258, 13009, Marseille, France
Full list of author information is available at the end of the article

Abstract

Background: Shallow shotgun metagenomics (SSM) has been recently suggested as a promising strategy to study human microbiota, providing nearly identical taxonomic profiles to deep shotgun metagenomics but at a sequencing cost as low as that of metabarcoding. To help clinical researchers determine whether shallow sequencing is appropriate for their projects, it is crucial to ascertain the accuracy of the information it provides, compared to deep sequencing. Here, we design a mapping-based workflow to build taxonomic profiles from SSM data and assess its accuracy at varying sequencing depths at both sample and cohort levels using extensive simulations and several public data sets.

Results: To identify genuinely present species and spuriously identified ones, we propose a novel data-driven filtering method based on machine learning techniques that largely outperforms basic filtering strategies based on predefined thresholds, resulting in reliable taxonomic profiles at different sequencing depths, ranging from 50K to 10M reads/samples. Up to 90% of species with relative abundances higher than 4.10^{-4} were recovered correctly at 500K reads/sample showing that only information about rare taxa is lost at shallow depths. Furthermore, our results clearly show that SSM is able to correctly recover relevant biological signal from the confidently identified taxa, such as differences between groups of patients and diagnosis-like classification.

Conclusions: This study confirms that SSM is suitable for clinical research on human gut microbiota. We recommend that researchers should consider moving from 16S to SSM to limit biases in taxonomic profiles, or moving from deep to shallow sequencing, when functional analyses are not the main focus, to reduce costs and be able to include more patients in research projects.

Keywords: microbiota; metagenomics; shotgun metagenomics; shallow shotgun metagenomics

Background

High throughput sequencing (HTS) unraveled the quantity, diversity and complexity of host-microbiota interactions in health and human diseases by allowing culture-free analyses of microbial ecosystems [1] [2] [3]. Metabarcoding, which consists in targeted sequencing of a phylogenetic marker (often the 16S rRNA gene for bacteria), has been widely used, as it allows to assess the biodiversity within and between samples, and to obtain an approximate taxonomic identification (most frequently down to the genus level) of the microorganisms present in a sample [4]. Besides, whole genome sequencing (WGS, also referred as shotgun sequencing) allows deeper taxonomic resolution (down to species, or even strain level) [5] [6], functional profiling (identification and quantification of genes, metabolic pathways) [7], and *de novo* assembly of uncultured organism genomes as Metagenome-Assembled Genomes (MAGs) [8]. Even though it is often claimed that human microbiota research would need to head toward WGS to deeply understand host-microbiota interaction [5] [9] [10] [11], metabarcoding is still widely used, notably because of its substantially lower sequencing and data processing costs. Indeed, researchers often favour the number of samples to include many patients and/or have multiple samples per patient (longitudinal approach), rather than the amount of information per sample.

Shallow shotgun metagenomics (SSM) have been recently suggested as an alternative middle route [12], which is both cost-competitive with metabarcoding and as informative as deep WGS for taxonomic profiling. Whereas tens of millions of reads per sample are typically used to characterize human gut microbiota samples with standard (deep) shotgun sequencing [13] [14], SSM typically deals with fewer than 1M reads/sample, thus drastically reducing sequencing costs. Previous works suggest that species level taxonomic profiles obtained by mapping reads on reference genomes were highly similar to deep WGS, and that the resulting α-diversity metrics were barely impacted by limiting the sequencing depth to $\sim 500K - 1M$ reads/sample [12] [15] [16] [17]. Furthermore, mapping on reference genomes was more efficient than on a marker genes catalog in the context of SSM [15]. By contrast, the depth required for functional analysis depends on the level of granularity of the analysis: 500K reads/sample may be sufficient to identify KEGG orthology groups [12], but 3M to 5M reads/sample are necessary to accurately detect genes and pathways [15] [18], and very deep sequencing up to $\sim 60 - 80M$ reads/sample is needed to study antimicrobial resistance genes [16] [19]. However, complementary investigations on the reliability of taxonomic profiles constructed by mapping SSM reads on a catalog of representative genomes are needed. In particular, as we deal with critically low sequencing depths, it is essential to retrieve as much information as possible from each read, including ambiguous reads that map to several genomes, and to identify genomes present in low abundances. To the best of our knowledge, previous works used catalogs that did not include MAGs (Metagenome-Assembled Genomes), limiting resulting profiles to cultured organisms. Huge metagenomic efforts have been made to study the human gut microbiota, allowing to build exhaustive catalogs, that gather both cultured and uncultured organisms, expanding taxonomic resolution for following metagenomic studies [20]. In addition, there is still limited knowledge about the potential loss of biological signal recovery in clinical metagenomics data sets when switching from deep WGS to SSM.

In the present study, we aim (i) to build a mapping-based workflow for SSM data to build accurate taxonomic profiles, that uses state of the art reference genome database, (ii) to evaluate the effect of sequencing depth on taxonomic profiles and, and (iii) underlying signal recovery (*e.g.* stratification of patients into groups) for a clinical study. To do so, we used extensive simulations and analyzed several publicly available data sets.

Methods

Data

Simulated data sets.

We retrieved taxonomic profiles of 100 human gut microbiotas from Qin 2014 [21] through *curatedMetagenomicData* [22]. The samples have a richness of 98 ± 15 species per sample, and species' relative abundances range from 5.10^{-1} down to 10^{-6} (geometric mean 5.10^{-4}). These profiles were given using the NCBI's taxonomy and were converted into UHGG's taxonomy [20] v1.0 by choosing the UHGG species with the taxonomic assignation closest to the NCBI one (if several NCBI species were tied, one was chosen randomly), resulting in profiles with the exact same complexity, approximately the same phylogenetic composition but including some uncultured organisms (MAGs). UHGG genomes are clustered into "species clusters" (thereafter referred to as *species*), that share at least 95% of identity on 30% of the genomes. One representative genome is chosen in each cluster for inclusion in the mapping catalog. We generated profiles using a randomly selected genome of each targeted species, in order to mimic the case where the strain we observe is not the one present in the catalog. For each of the 100 samples, we simulated 10M paired end reads, using *Grinder* [23] v0.5.3 (length of 2*125bp, insert size normally distributed with an average of 500bp and standard deviation of 50 bp without sequencing error) and subsampled the read sets at 5 M, 1 M, 500 K, 100 K, 50 K and 10 K reads/sample.

Real data sets.

We used data from 3 clinical studies, for a total of $N = 439$ samples covering patients from several continents and clinical conditions (healthy patients, hepatic diseases at different stages, cancer patients). Loomba *et al.* (2017) [24] compared the gut microbiota of 86 patients suffering from hepatic diseases at different stages ($N = 14$ fibrosis vs $N = 72$ NAFLD). Matson *et al.* (2018) [25] compared, among 39 patients having metastatic melanomas, those who responded to anti-PD-1 immunotherapy ($N = 15$) and those who did not ($N = 24$). Qin *et al.* (2014) [21] compared patients having liver cirrhosis ($N = 169$) and a group of healthy controls ($N = 145$), with a discovery and a validation cohort for both groups. We analyzed these data sets at full depth and subsampled them to mimic shallow sequencing in the remainder, as described above.

Bioinformatics pipeline

Reads were pre-processed using *trimmomatic* [26] v0.39 to remove low quality reads (with average quality below 30) and reads shorter than 80 nucleotides. For real data sets, reads were also mapped to the human genome (hg38) to filter out host contamination. Remaining reads were then mapped to the UHGG catalog [20] v1.0 using *bwa mem* [27] v0.7.17 local aligner, with option $-h50$ to allow up to 50 reported hits. Mapping files were then processed using *samtools* [28] v1.9 and custom scripts developed under Python v2.7.13.

Ambiguous reads (*i.e.* reads that map to several genomes) occur frequently when mapping reads to a catalog of reference genomes (26% and 42% of the mapped reads in simulated and real data sets respectively), due to highly conserved genes and mobile elements notably. Thus, we split mapped reads into unambiguous reads (*i.e.* mapping to one genome only), and ambiguous reads. For each genome, we computed the reads count (RC) and the fraction of the genome covered (FC) by at least one read, using either all reads or unambiguous reads only (uRC and uFC), as well as a specificity ratio (SR) defined by the number of unambiguous reads divided by the total number of reads mapped to this genome (uRC/RC).

In order to estimate the species' s relative abundances, we first compute the representative genomes' average coverage $C_s = \frac{1}{\ell_s}\sum_i r_{i,s}$, with ℓ_s being the length of the representative genome of species s and $r_{i,s}$ the length of read i that is unambiguously mapped to s. We then obtain the relative abundance by normalizing across species to sum up to 1: $A_s = \frac{C_s}{\sum_j C_j}$. We refine this estimation by reallocating the ambiguous reads, randomly assigning them to one of their hits with a probability proportional to A_s.

Simulations analysis

Direct mapping of short reads on reference genomes produces many false positives (genomes covered by reads but not present in the sample) that need to be filtered out. We used simulated profiles, with known composition, to determine the most efficient way to classify the genomes into true positives (TP) and false positives (FP). In order to assess methods and compare them to each other, we computed the area under the receiver operating characteristic (ROC) curve (AUC) for this classification task, using *evabic* R package (https://github.com/abichat/evabic). We also implemented an automated threshold search, that limits the false discovery rate ($FDR = \frac{FP}{TP+FP}$) to a maximum of 0.1, and compared false negative (FN) rates at this threshold across methods and sequencing depths.

We first evaluated how genomes' features (RC, uRC, FC, uFC and SR) can be used independently to classify the genomes, and then combined them to train classifiers. We used logistic regression (LR), linear discriminant analysis (LDA) and random forests (RF), to perform classification, with uRC, uFC, SR and total sequencing depth as input features. Finally, we used a 4-fold cross validation procedure to evaluate the performance of these methods and determine suitable thresholds for each method and sequencing depths.

Real data sets analysis

We analyzed real data sets using (1) RF-based filters fitted on the simulations data and thresholds that control FDR at each sequencing depth, and (2) a basic filtering that discards all species with a relative abundance beyond 10^{-4}, FC beyond 10^{-2} or uFC beyond 10^{-4}. This filtering is inspired by the one used by Santiago-Rodriguez *et al.* [15]. Note that it is a quite permissive threshold due to the low sequencing depths to which it is applied.

We evaluated α-diversity using species richness and Shannon diversity, and β-diversity using Jaccard distance and Bray-Curtis dissimilarity index, computed with the R package *phyloseq* [29]. In order to assess the impact of sequencing depth on taxonomic profiles, we evaluated the correlation between subsampled and deep α-diversity measures using Spearman and Pearson correlation as well as the correlation between species' relatives abundances at full and shallower depths. Additionally, we measured the distance between low depth samples and their full depth counterparts. Finally, for each data set, we evaluated whether differences between groups of interest were preserved at lower sequencing depths using the following criteria :

- difference in α-diversity between groups, through a Wilcoxon test on the aforementioned metrics,
- structure in the β-diversity matrix, through a PERMANOVA analysis of the aftermentioned dissimilarity indices,
- biomarker discovery, using a Wilcoxon test with Benjamini-Hochberg correction to recover differentially abundant species,
- patients' classification in their groups of interest, using random forests trained on taxonomic profiles together with species richness and Shannon's diversity, and performing an iterative feature selection step as described by Loomba *et al.* [24], with 10 repetitions at each step to take into account the variability of the classification method. For Loomba-2017, we asked authors for patients age and BMI, and imputed missing values with the mean value of the cohort. For Qin-2014, we performed one classification with all patients, and another classification where the discovery cohort was used for training and the validation cohort for testing.

In order to perform unbiased comparisons of p-values and AUCs across sequencing depths, we used only samples having at least 10M high quality reads per sample for Loomba-2017 ($N = 77$) and Matson-2018 ($N = 39$) and only 5M reads for Qin-2014 ($N = 236$) as using the same 10M depth would have reduced the cohort analyzed down to 172 patients.

Results

A tailored filtering strategy greatly improves species recovery

We first used the 100 simulated metagenomes to design an appropriate pipeline to build taxonomic profiles from SSM reads mapped on a reference genomes catalog. Our raw mapping data from simulations shows that a small number of reads (8% of unambiguously mapped reads) are mapped to a large number of unexpected genomes, not part of the original profile used for read simulation. Overall, those numerous unexpected genomes have a low coverage but result in a great number of false positives if no filtering is applied (FDR = 0.92). The basic filtering approach, which discards the rarest and least covered species (see above for details), yields an overall FNR = 0.44 and FDR = 0.46. The different features available have contrasted discriminatory powers as measured by AUC: 0.76 for read counts (RC) and 0.87 for fraction covered (FC). The AUC increases when using only unambiguous reads: 0.84 for uRC and 0.90 for uFC (see Table 1 for details). We therefore used the latter two features to build optimized filters. As seen on Figure 1A, a thresholding strategy based on uFC and/or uRC values (corresponding to horizontal and/or vertical lines) to discriminate TPs and FPs, would be suboptimal as it would miss the long tail of genomes with low uRC but comparatively high uFC values. These observations motivated the development of a data-driven classifier that could leverage this pattern. We trained several classifiers (LDA, Logistic regression, Random forest) using uRC, uFC, SR, as well as sequencing depth as predictors to predict genomes' status (present or absent).

We can see in Table 1 that data-driven classifiers largely outperform basic filtering. LDA and LR perform similarly, and yield nearly identical results in training and testing samples in the cross validation process, highlighting very good generalization capabilities. RF appears to be the best method, yielding a nearly perfect classification in training sets, and still outperforming others in the testing sets. RF was thus used in the rest of the work. Table 1 also shows that when choosing a threshold that limits the FDR at 0.1, data-driven classifiers have a drastically lower FN rates ($\sim 0.3 - 0.4$) than the ones obtained with independent thresholding on each genomes features ($\sim 0.6 - 0.9$).

The FN values are still quite high and may correspond to a subset of species that are intrinsically difficult to identify. To further investigate this issue and characterize the information loss induced by SSM, we plotted the distribution of simulated relative abundances of species that were absent in profiles with respect to the sequencing depth, as seen on Figure 1B, using RF-based filtering. We can clearly see the inflation of FN when lowering sequencing depth, but we can also notice that, as expected, the populations that are lost are relatively rare. For instance, at 500K reads/sample, 90% of species with an abundance greater than 4.10^{-4} were detected. In comparison, 90% of the species with an abundance greater than 2.10^{-4} were detected at 1M reads/sample, and this value was down to 3.10^{-5} at 5M reads/sample.

While focusing on TPs, we observed that Pearson correlation between simulated and estimated species relative abundances increases from $\rho = 0.81$ to $\rho = 0.91$ if we add a step of reallocation of ambiguous reads. Therefore, we used profiles after reallocation of the ambiguous reads for the rest of the work.

Shallow shotgun metagenomics accurately reconstruct taxonomic profiles for abundant taxa

Applying the RF-based filters on real data sets results in profiles with an average diversity of 128 ± 66 species per sample at full depth. The diversity gradually decreases as the sequencing depth decreases, shrinking down to 45 ± 21 at 500K reads/sample for example (fig 2A). In contrast to the observed species count, the Shannon diversity index (fig 2B) is much less impacted by sequencing depth, indicating that the species lost at low sequencing depths are mostly rare ones, as expected. Down to 500K reads per sample, the correlation between full depth and subsampled Shannon indices remains very high. Distances between subsamples and their *reference*, defined as the corresponding sample at full depth, gradually increase when decreasing the sequencing depth (fig 2C), and show high replicability across data sets.

In comparison, basic filtering is more permissive, producing profiles with increased diversities and less affected by reduced sequencing depths (fig 2D, E). Distances to the reference are also smaller than with RF-based filters (fig 2C,F).

These results show that taxonomic profiles are not impacted by the sequencing depth for most abundant species, regardless of the filtering strategy used. For the rarest species, we can see that reducing sequencing depth gradually decreases the number of detected species, especially when using a filtering strategy that ensures few spuriously detected species.

Shallow shotgun metagenomics recover biological signal

Here, we assess the impact of the sequencing depth on biological signal recovery, *i.e.* stratification of patients into groups, in three different data sets covering several diseases and geographical origins. In each study, we track differences in the microbiota composition between two groups of patients according to a clinical condition (see Methods section for details). We aim to see whether the results obtained in deep shotgun sequencing are preserved when switching to SSM.

As expected from previous results, differences in α-diversity between groups are maintained using shallow sequencing: p-values were concordant across sequencing depths (see Fig. 3A), with a strong difference between groups in Qin-2014 data set at all depths, a slight but not significant difference between groups in Loomba-2017 and no difference between groups in Matson-2018.

PERMANOVA analysis leads to similar results (Fig. 3B), showing that the structure of the distance matrix between samples is only marginally impacted by reducing sequencing depths. The p-value between healthy and sick patients in Loomba-2017 slightly increases when reducing the sequencing depth down to 500K but drastically increases when moving to 100K reads/sample and below, revealing that some key species for the stratification of patients may be lost at very low sequencing depths.

Moreover biomarkers discovery (Supplementary File 1) shows that the number of differentially abundant taxa identified in the data sets strongly depends on the sequencing depth, with abundant biomarkers easier to recover at shallow depths than rare ones. Indeed, out of the 6 biomarkers identified in Loomba-2017 with deep sequencing, the 4 markers with a geometrical mean abundance greater than 10^{-2} were recovered at 500K reads/sample. Similarly, in Qin-2014, out of the 56 biomarkers identified with deep sequencing , the 2 markers with a geometrical mean abundance greater than 10^{-2} and 12 of 37 markers with a geometrical mean abundance between 10^{-2} and 10^{-3} were recovered at 500K reads/sample. None of the rares ones, with a geometric mean abundance lesser than 10^{-3} was recovered in both datasets. (see supplementary file 2 for details).

Finally, we performed classification of patients using RF on the taxonomic profiles in Loomba-2017 and Qin-2014 (see Fig. 3C). On Loomba-2017, we observed that, under 5M reads/sample, AUC gradually decreases as sequencing depth becomes smaller, due to the loss of some taxa that are crucial for the classification. On Qin-2014 data set, we could perform a very good classification on both discovery and validation cohorts even at low sequencing depths, with performance gradually decreasing under 5M reads/sample.

Results obtained by basic filtering lead to the same conclusion regarding the impact of the sequencing depth on patients stratification, as shown in supplementary file 3. In comparison with RF-based filtering, p-values tend to be lower and less impacted by sequencing depth on Loomba-2017, and AUCs upper. Basic filters are more permissive, thus more species are identified, especially rare ones at low sequencing depths. As seen in the simulations, they also include many spuriously identified species, but it seems that patients stratification is not widely impacted by the noise created by those spurious species.

Discussion

Our simulations highlighted the need for an efficient filtering strategy while mapping shallow shotgun metagenomic reads against a reference catalog to reconstruct taxonomic profiles. In order to have reliable results for every sequencing depth, depth-dependent thresholds were applied. This step is crucial to prevent misleading interpretations and to provide a trustworthy biological knowledge. Controlling the false discovery rate (FDR) in taxonomic profiles had the direct consequence of decreasing the number of identified species, especially at low sequencing depths. We observed a great benefit of random forest-based filters, and to a lesser extent of other machine learning-based models tested, in comparison with simple filters based on species features independently (read counts and fraction covered, considering all reads and unambiguous reads only), allowing to identify more species and rarer ones for equivalent FDR. On the three real data sets considered, our analysis showed that differences between groups of patients observed at full depth were still recovered at low sequencing depths. Permissive and depth-independent filtering, as performed in previous works on SSM, allowed a little improvement in structure recovery compared to our stringent random forest-based filters. Indeed, these structures were less sensitive to the noise introduced by spuriously identified species in the profiles using basic filtering, than to the removal of key species induced by our stringent random forest-based filter. However, for trustfulness and interpretability of results, downstream analysis should be based on reliably identified taxa even if it induces a moderate loss of biological signal. Regardless of the filtering strategy, our analysis shows that SSM is sufficient to recover most of the differences between groups of patients discovered with deep sequencing, provided that the differences are not based on rare species.

Our simulations led us to develop stringent filters, especially at low sequencing depths, resulting in profiles with limited species richness. Other studies reported a smaller effect on species richness while using similar mapping strategies but less stringent filters [12] [15]. Unlike other published papers, the catalog we used contained metagenome-based assembled genomes, which are often incomplete and have contamination. It enables the identification of uncultured organisms but it can contribute to the noise observed in the mapping data, and thus to the need for stringent filtering.

As it relies on a learning step, the application of machine learning-based filters we have developed is limited to the training conditions. In our case, usage of random forest-based models to filter genomes should be limited to ecosystems with a similar complexity, sequenced with short reads at a depth included in the range used for the training and mapped to a catalog similar to representative genomes of UHGG in terms of completeness, intra and inter species diversity. More generally, readers should keep in mind that shallow shotgun metagenomics can be considered only for ecosystems for which (nearly) exhaustive reference databases exist, allows to assemble neither genes nor genomes, and can only produce coarse-grained functional analyses.

In addition to our analyses, we provide a pipeline that produces reliable taxonomic profiles at every sequencing depth, as an end-to-end solution for the analysis of shallow shotgun metagenomics data from human gut microbiota samples (see Availability of data and materials).

Conclusions

Our results show that (1) one needs to carefully filter taxonomic profiles retrieved from shallow shotgun metagenomics data to have trustworthy and interpretable results, (2) resulting taxonomic profiles are valid but limited to the most abundant taxa at low sequencing depths, *i.e.* taxa with relative abundance greater than 2.10^{-4} at 1M reads/sample, and greater than 4.10^{-4} at 500K reads/sample, and (3) shallow shotgun metagenomics allow a very good recovery of the structure of a data set, and constitutes a suitable approach to perform diagnosis-like classification of patients. It can be profitable in many gut microbiota-related clinical research projects to use shallow shotgun metagenomics. In comparison with metabarcoding, it allows to identify taxa down to species level and prevent biases related to targeted amplification. In comparison with *deep* shotgun sequencing, the information loss at cohort level is limited, thus it is certainly profitable to reduce the sequencing depth and favour number of samples to be analyzed to produce trustworthy biological knowledge.

Declarations

Acknowledgements

We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing help, computing and storage resources, as well as CRCM's DISC platform.

Funding

This work was financially supported by ANRT and Laboratoire Alphabio through a CIFRE scholarship (2019-0346).

Abbreviations

WGS : whole genome sequencing
SSM : shallow shotgun metagenomics
MAGs : metagenome-based assembled genomes
UHGG : unified human gut genome [20]
RC : read counts (mapped to a given genome)
uRC : unambiguous read counts (mapped to a given genome)
FC : fraction covered by all reads (for a given genome)
uFC : fraction covered by unambiguous reads (for a given genome)
RF : random forest

Availability of data and materials

All data and scripts used in this article are available at https://forgemia.inra.fr/benoit.goutorbe/shallow-shotgun-metagenomics.git.

Ethics approval and consent to participate

Not relevant

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not relevant

Authors' contributions

BG, ALA, MM and SS conceived the ideas and designed the methodology; BG performed the simulations, analysed the data and developed the bioinformatics pipeline; BG led the writing of the manuscript with supervision from ALA, MM, AP, GB and SS. All the authors contributed critically to the drafts and gave final approval for publication.

Author details

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France. [2]Clinical Research and R&D Department, Laboratoire Européen Alphabio, 13003 Marseille, France. [3] Cibi plateform, Centre de Recherche en Cancérologie de Marseille (CRCM), Institut Paoli-Calmettes, Aix-Marseille Université U105, Inserm U1068, CNRS UMR7258, 13009, Marseille, France.

References

1. Fan, Y., Pedersen, O.: Gut microbiota in human metabolic health and disease. Nature Reviews Microbiology **19**(1), 55–71 (2021). doi:10.1038/s41579-020-0433-9. Accessed 2021-04-08
2. Wong, S.H., Yu, J.: Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. Nature Reviews Gastroenterology & Hepatology **16**(11), 690–704 (2019). doi:10.1038/s41575-019-0209-8. Accessed 2021-04-08
3. Clemente, J.C., Manasson, J., Scher, J.U.: The role of the gut microbiome in systemic inflammatory disease. BMJ, 5145 (2018). doi:10.1136/bmj.j5145. Accessed 2021-04-08
4. Golob, J.L., Margolis, E., Hoffman, N.G., Fredricks, D.N.: Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. BMC Bioinformatics **18**(1), 283 (2017). doi:10.1186/s12859-017-1690-0. Accessed 2021-07-26
5. Ellegaard, K.M., Engel, P.: Beyond 16S rRNA Community Profiling: Intra-Species Diversity in the Gut Microbiota. Frontiers in Microbiology **7** (2016). doi:10.3389/fmicb.2016.01475. Accessed 2021-07-26
6. Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., Banfield, J.F.: inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nature Biotechnology **39**(6), 727–736 (2021). doi:10.1038/s41587-020-00797-0. Accessed 2021-07-26
7. Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S.T., Methé, B., Schloss, P.D., Gevers, D., Mitreva, M., Huttenhower, C.: Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. PLoS Computational Biology **8**(6), 1002358 (2012). doi:10.1371/journal.pcbi.1002358. Accessed 2021-07-26
8. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W.: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology **2**(11), 1533–1542 (2017). doi:10.1038/s41564-017-0012-7. Accessed 2021-04-08
9. Yen, S., Johnson, J.S.: Metagenomics: a path to understanding the gut microbiome. Mammalian Genome **32**(4), 282–296 (2021). doi:10.1007/s00335-021-09889-x. Accessed 2021-07-26
10. Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., Carissimi, C.: Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. OMICS: A Journal of Integrative Biology **22**(4), 248–254 (2018). doi:10.1089/omi.2018.0013. Accessed 2021-07-26
11. Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., De Cesare, A.: Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. Scientific Reports **11**(1), 3030 (2021). doi:10.1038/s41598-021-82726-y. Accessed 2021-07-05
12. Hillmann, B., Al-Ghalith, G.A., Shields-Cutler, R.R., Zhu, Q., Gohl, D.M., Beckman, K.B., Knight, R., Knights, D.: Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems **3**(6) (2018). doi:10.1128/mSystems.00069-18. Accessed 2021-06-09
13. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D., Wang, J.: A human gut microbial gene catalog established by metagenomic sequencing. Nature **464**(7285), 59–65 (2010). doi:10.1038/nature08821. Accessed 2021-07-28
14. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N.: Shotgun metagenomics, from sampling to analysis. Nature Biotechnology **35**(9), 833–844 (2017). doi:10.1038/nbt.3935. Accessed 2021-04-08
15. Santiago-Rodriguez, T.M., Garoutte, A., Adams, E., Nasser, W., Ross, M.C., Reau, A.L., Henseler, Z., Ward, T., Knights, D., Petrosino, J.F., Hollister, E.B.: Metagenomic Information Recovery from Human Stool Samples Is Influenced by Sequencing Depth and Profiling Method. Genes, 17 (2020)
16. Gweon, H.S., Shaw, L.P., Swann, J., De Maio, N., AbuOun, M., Niehus, R., Hubbard, A.T.M., Bowes, M.J., Bailey, M.J., Peto, T.E.A., Hoosdally, S.J., Walker, A.S., Sebra, R.P., Crook, D.W., Anjum, M.F., Read, D.S., Stoesser, N.: The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. Environmental Microbiome **14**(1), 7 (2019). doi:10.1186/s40793-019-0347-1. Accessed 2021-07-27
17. Cattonaro, F., Spadotto, A., Radovic, S., Marroni, F.: Do you cov me? Effect of coverage reduction on metagenome shotgun sequencing studies. F1000 Research **7**, 1767 (2020). doi:10.12688/f1000research.16804.4. Accessed 2021-03-02
18. Treiber, M.L., Taft, D.H., Korf, I., Mills, D.A., Lemay, D.G.: Pre- and post-sequencing recommendations for functional annotation of human fecal metagenomes. BMC Bioinformatics **21**(1), 74 (2020). doi:10.1186/s12859-020-3416-y. Accessed 2021-07-23
19. Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S.R., Marinier, E., Van Domselaar, G., Belk, K.E., Morley, P.S., McAllister, T.A.: Impact of sequencing depth on the characterization of the microbiome and resistome. Scientific Reports **8**(1), 5890 (2018). doi:10.1038/s41598-018-24280-8. Accessed 2021-07-27
20. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., Segata, N., Kyrpides, N.C., Finn, R.D.: A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnology **39**(1), 105–114 (2021). doi:10.1038/s41587-020-0603-3. Accessed 2021-02-04
21. Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Chatelier, E.L., Yao, J., Wu, L., Zhou, J., Ni,

S., Liu, L., Pons, N., Batto, J.M., Kennedy, S.P., Leonard, P., Yuan, C., Ding, W., Chen, Y., Hu, X., Zheng, B., Qian, G., Xu, W., Ehrlich, S.D., Zheng, S., Li, L.: Alterations of the human gut microbiome in liver cirrhosis. Nature **513**(7516), 59–64 (2014). doi:10.1038/nature13568. Number: 7516 Publisher: Nature Publishing Group. Accessed 2020-03-04

22. Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D.T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., Huttenhower, C., Morgan, M., Segata, N., Waldron, L.: Accessible, curated metagenomic data through ExperimentHub. Nature methods **14**(11), 1023–1024 (2017). doi:10.1038/nmeth.4468. Accessed 2021-06-09
23. Angly, F.E., Willner, D., Rohwer, F., Hugenholtz, P., Tyson, G.W.: Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Research **40**(12), 94–94 (2012). doi:10.1093/nar/gks251. Accessed 2020-06-19
24. Loomba, R., Seguritan, V., Li, W., Long, T., Klitgord, N., Bhatt, A., Dulai, P.S., Caussy, C., Bettencourt, R., Highlander, S.K., Jones, M.B., Sirlin, C.B., Schnabl, B., Brinkac, L., Schork, N., Chen, C.-H., Brenner, D.A., Biggs, W., Yooseph, S., Venter, J.C., Nelson, K.E.: Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metabolism **25**(5), 1054–10625 (2017). doi:10.1016/j.cmet.2017.04.001. Accessed 2019-10-10
25. Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J.J., Gajewski, T.F.: The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. Science **359**(6371), 104–108 (2018). doi:10.1126/science.aao3290. Accessed 2019-10-10
26. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014). doi:10.1093/bioinformatics/btu170. Accessed 2020-02-25
27. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio] (2013). arXiv: 1303.3997. Accessed 2020-02-25
28. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (2009). doi:10.1093/bioinformatics/btp352. Accessed 2021-10-06
29. McMurdie, P.J., Holmes, S.: phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE **8**(4), 61217 (2013). doi:10.1371/journal.pone.0061217. Accessed 2021-04-08

figures/figure1.png

Figure 1 Simulations results: (A) Unambiguous fraction covered (uFC) and unambiguous read counts (uRC) of genomes present in the simulated profiles (TPs, blue points) or absent (FPs, red points). (B) Distribution of FN species according to their relative abundances in the simulated profiles, using RF-based filters with a 0.1 FDR on the testing set of cross validation. The dotted line represents the distribution of all simulated species.

figures/figure2.png

Figure 2 Comparison across sequencing depths for samples from the 3 data sets considered: richness observed, Shannon diversity and Bray-Curtis distance between subsampled data and reference (full depth data) using RF-based filtering (A, B and C respectively) and basic filtering (D, E and F respectively).

figures/figure3.png

Figure 3 Differences between patients groups in different studies: significance of inter-group difference regarding Shannon diversity index for α-diversity (A), PERMANOVA analysis for β-diversity (B). AUC corresponding to random forest classification (C) was performed in Loomba-2017 and Qin-2014, with a split between discovery and validation cohorts in Qin-2014 as performed on the original paper of this study. Error bars represent standard deviation for 10 repetitions of training process in the classification model.

Tables

Table 1 Classification of mapping hits in present and spuriously identified species: area under ROC curves and false negative rates when threshold is set to tolerate 0.1 FDR. For machine learning-based methods, these measures are split into training and testing sets, using a 4-fold cross validation.

method	AUC		FN rate at threshold	
	training	testing	training	testing
uRC	0.844		0.916	
uFC	0.904		0.655	
LDA	0.947 ± 0.001	0.947 ± 0.002	0.415 ± 0.002	0.416 ± 0.010
Logistic regression	0.958 ± 0.001	0.958 ± 0.002	0.388 ± 0.003	0.389 ± 0.013
Random forest	0.999 ± 0.0001	0.969 ± 0.003	0.037 ± 0.001	0.292 ± 0.008

Figures

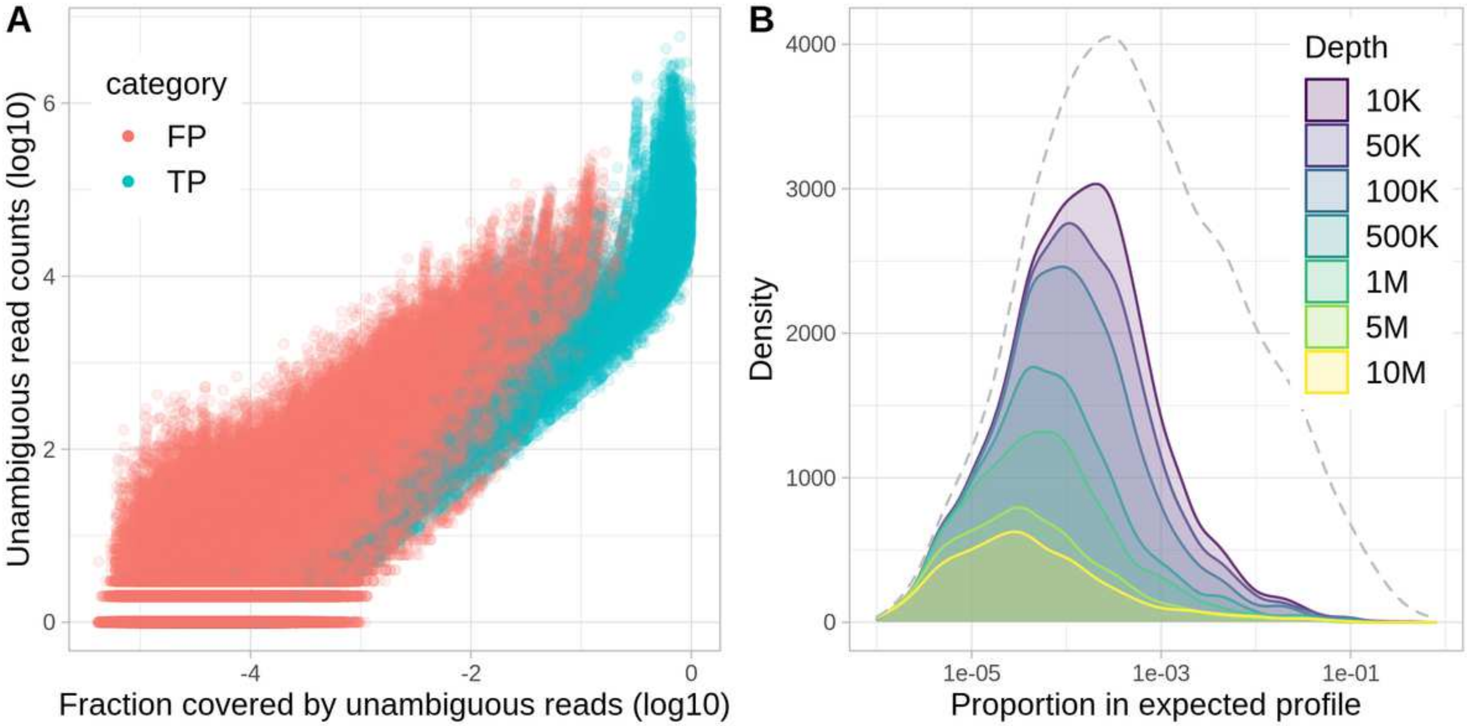

Figure 1

Simulations results: (A) Unambiguous fraction covered (uFC) and unambiguous read counts (uRC) of genomes present in the simulated profiles (TPs, blue points) or absent (FPs, red points). (B) Distribution of FN species according to their relative abundances in the simulated profiles, using RF-based filters with a 0:1 FDR on the testing set of cross validation. The dotted line represents the distribution of all simulated species.

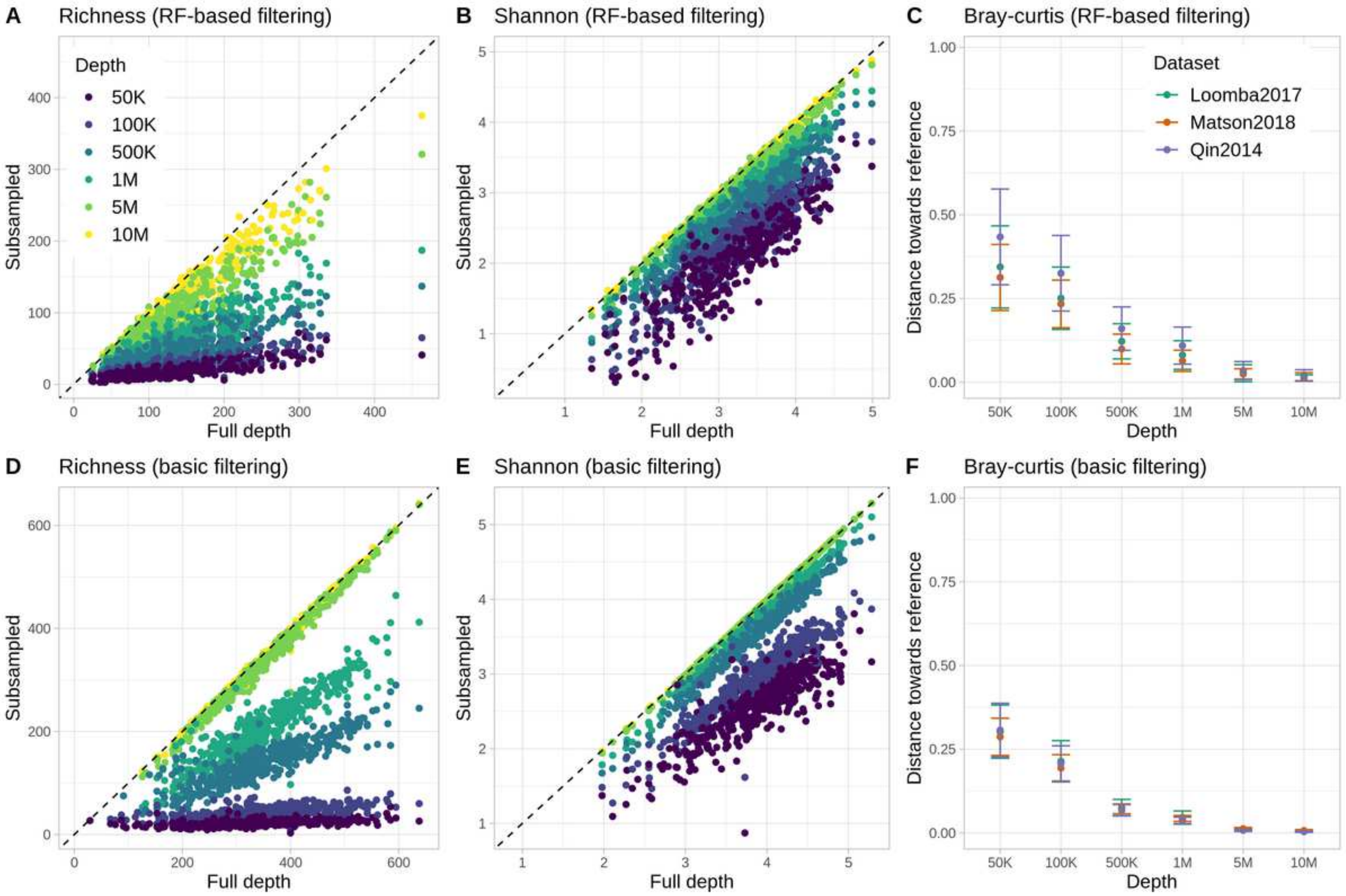

Figure 2

Comparison across sequencing depths for samples from the 3 data sets considered:

richness observed, Shannon diversity and Bray-Curtis distance between subsampled data and reference (full depth data) using RF-based filtering (A, B and C respectively) and basic filtering (D, E and F respectively).

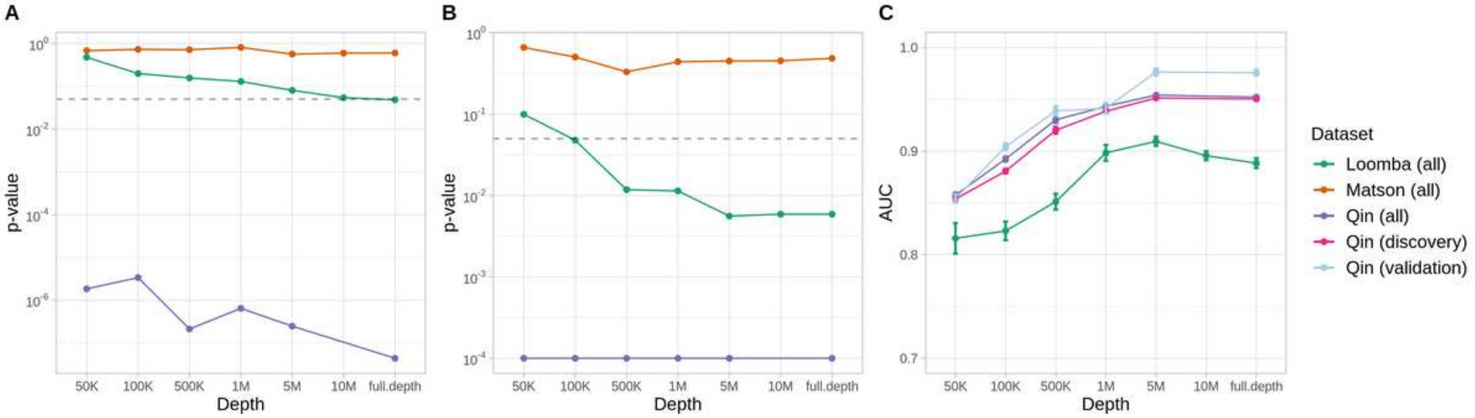

Figure 3

Differences between patients groups in different studies: significance of inter-group difference regarding Shannon diversity index for -diversity (A), PERMANOVA analysis for -diversity (B). AUC corresponding to random forest classification (C) was performed in Loomba-2017 and Qin-2014, with a split between discovery and validation cohorts in Qin-2014 as performed on the original paper of this study. Error bars represent standard deviation for 10 repetitions of training process in the classification model.

Supplementary Files

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

- additionalfile1.xlsx
- additionalfile2.png
- additionalfile3.png
- listofadditionalfiles.pdf