[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France ; [2] Laboratoire Alphabio, Marseille, France ; [3] CRCM, Inserm, CNRS, Institut Paoli-Calmettes, Aix-Marseille Université, 13009, Marseille, France ; [4] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France

Shallow sequencing: a cost-effective and accurate alternative to WGS for taxonomic profiling ?

Benoit Goutorbe[1,2,3], Anne-Laure Abraham[1], Mahendra Mariadassou[1], Valentin Loux[1,4], Olivier Rué[1,4], Anne Plauzolles[2], Ghislain Bidaut[3], Philippe Halfon[2] and Sophie Schbath[1]

Context

In the rising research area **microbiota-associated health outcomes**, clinical researchers have to deal with the critical choice of the analytic technique used to characterize patients' microbiota. This choice is usually binary, with **metabarcoding**, a low-cost and an efficient way to identify and quantify organisms present in an ecosystem (taxonomic profiles) which has a limited resolution and suffers from well known biases (amplification biases, variation in copy numbers, *etc*), **and whole genome sequencing** (WGS), which offers deep insights about both taxonomic and functional profiles but is much more expensive (about 10 times the cost of metabarcoding) and produces data massively more complex to analyze. Due to a huge inter-patient variability**, large cohorts are needed** to extract reliable information. Thus, a large majority of studies are carried out using metabarcoding techniques. Shallow WGS (WGS at very low sequencing depth, down to 500 K reads/sample) is one of the techniques that could fit into this technological gap : a recently published paper [1] demonstrated the huge potential of this approach but left **many important questions unanswered**. Our aims is then to evaluate reliability and limitations of shallow shotgun metagenomic data.

Materials and Methods

Grinder [2] was used to produce synthetic metagenomic dataset of 10 millions reads/sample with realistic complexity, based on 19 gut microbiota samples [3] (richness = 126 ± 27 species/sample). Taxonomic profiles were collected from *curatedMetagenomicData* [4] and converted to UHGG's taxonomy [5]. Representative genomes for each species was used, and no sequencing errors were introduced so far. We also used publicly available raw sequencing data from clinical studies about hepatic diseases [6] and immunotherapy's response in cancer [7].

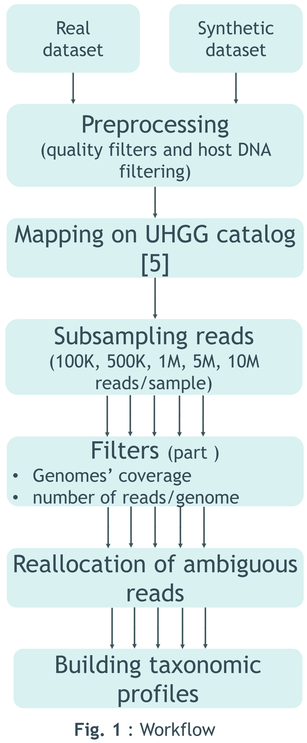

Fig. 1 : Workflow

We used BWA-MEM [8], BWA-ALN [9] and Bowtie2 [10] with end-to-end and local presets, to map reads to UHGG (Fig. 1).

Unambiguous reads (only 1 hit in the catalog) were used to perform filters on the profiles. Other reads were only used to fine-tune the estimation of genomes' relative abundance (GRA).

1 - Taxonomic profiles at low sequencing depth: reads mapping and filters

BWA-MEM and BWA-ALN mapped more reads than Bowtie2, without increasing FP* rates, which incited us to focus on these 2 algorithms. BWA-MEM tended to produce more ambiguous mapping than BWA-ALN but fewer FPs. Based on AUCs under ROC curves (Fig. 2) to discriminate TPs* and FP*s, we chose to use BWA-MEM mapping and a threshold on genomes' coverage of 0.6%, giving a precision* of 95%.

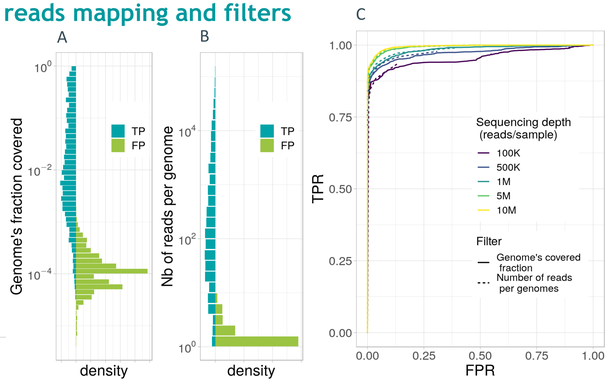

Fig. 2 : TPs* and FPs* distributions relatively to their fraction covered (A) and number of reads (B) at **500K reads/sample**. ROC curves (C) to discriminate both groups.

We further investigated the FNs* and their relative abundances in the expected profiles, to characterize the loss of information due to low sequencing depth. Our simulations showed that 500 K reads/sample was enough to identify all populations down to a proportion of 10^{-4} (Fig. 3).

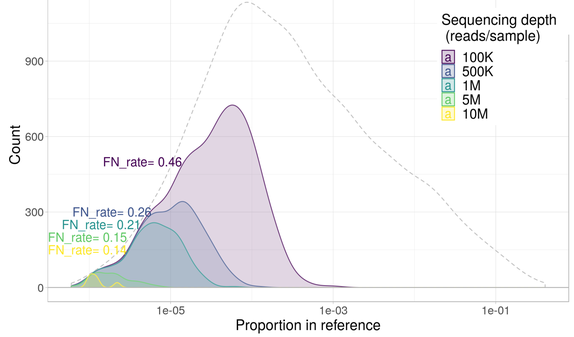

Fig. 3 : False negative rates, and their distribution relatively to their abundances in the reference profiles and sequencing depth

2 - Reallocating ambiguous reads allows a better estimation of GRAs

Finally, we built the taxonomic profiles by estimating GRAs: we divided the mean coverage for each genome by its length, and normalized across all genomes detected to sum to 1. We showed that reallocating ambiguous reads, according to probabilities proportional to the amount of unambiguous reads mapped to each of the tied hits, enhances correlation between expected and estimated GRA, regardless of sequencing depth (Fig. 4).

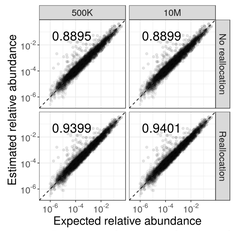

Fig. 4 : Pearson's correlation between expected and estimated GRA

3 - Downstream statistical analysis is robust towards low sequencing depth

The analysis of complete and subsampled datasets from clinical studies revealed that structures in the distance matrix between samples (Bray-Curtis distances, PERMANOVA's *p-values* regarding clinical outcomes, AGE, BMI, SEX) were **not affected by sequencing depth**. We also reproduced a random forest-based classification to discriminate patients as performed in [6] and noticed no loss of discriminative power between full depth (29 ± 19 M reads/sample, AUC = 0.904) and shallow (500 K reads/sample, AUC = 0.902) datasets.

Conclusions and perspectives

Strengths of shallow sequencing:

- Useful for routine analysis in well known ecosystems
- Accurate down to 500K reads/sample
- Little loss of statistical power for diagnosis-like classification

Limitations:

- Requires good reference databases for accurate taxonomic profiling.
- Adapted only for model organisms

Shallow sequencing is very promising for clinical use and diagnostic tools based on the human gut microbiome.

*TP : True Positive (detected and expected) ; FP : False Positive (detected but not exepected) ; FN : False Negeative (expected but not detected) ; precision = $\frac{TP}{TP+FP}$

This work was financially supported by ANRT and Laboratoire Alphabio thanks to a CIFRE scholarship

Contact
benoit.goutorbe@inrae.fr

Acknowledgment
We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources.

References

[1] B. Hillmann et al., Evaluating the Information Content of Shallow Shotgun Metagenomics, 2018.
[2] F. Angly et al., Grinder: a versatile amplicon and shotgun sequence simulator, 2012.
[3] N. Qin et al., Alterations of the human gut microbiome in liver cirrhosis, 2014.
[4] E. Pasolliet al., Accessible, curated metagenomic data through ExperimentHub, 2018.
[5] A. Almeida et al., A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome, 2019.
[6] R. Loomba et al., Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease, 2017.
[7] V. Matson et al., The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients, 2018.
[8] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, 2013.
[9] H. Li and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, 2009.
[10] B. Langmead et al., Fast gapped-read alignment with Bowtie 2, 2012.