



Check for updates

METHOD ARTICLE

REVISED Do you cov me? Effect of coverage reduction on metagenome shotgun sequencing studies [version 4; peer review: 2 approved, 2 not approved]

Previously titled: Do you cov me? Effect of coverage reduction on species identification and genome reconstruction in complex biological matrices metagenome shotgun high throughput sequencing

Federica Cattonaro ¹, Alessandro Spadotto ¹, Slobodanka Radovic ¹,
Fabio Marroni ^{1,2}

¹IGA Technology Services Srl, Udine, Udine, 33100, Italy

²Department of Agricultural, Food, Environmental and Animal Sciences (DI4A), University of Udine, Udine, 33100, Italy

v4 First published: 08 Nov 2018, 7:1767 (
<https://doi.org/10.12688/f1000research.16804.1>)

Second version: 22 Mar 2019, 7:1767 (
<https://doi.org/10.12688/f1000research.16804.2>)

Third version: 29 Jul 2019, 7:1767 (
<https://doi.org/10.12688/f1000research.16804.3>)

Latest published: 22 Jan 2020, 7:1767 (
<https://doi.org/10.12688/f1000research.16804.4>)

Abstract

Shotgun metagenomics sequencing is a powerful tool for the characterization of complex biological matrices, enabling analysis of prokaryotic and eukaryotic organisms and viruses in a single experiment, with the possibility of reconstructing *de novo* the whole metagenome or a set of genes of interest. One of the main factors limiting the use of shotgun metagenomics on wide scale projects is the high cost associated with the approach. We set out to determine if it is possible to use shallow shotgun metagenomics to characterize complex biological matrices while reducing costs. We used a staggered mock community to estimate the optimal threshold for species detection. We measured the variation of several summary statistics simulating a decrease in sequencing depth by randomly subsampling a number of reads. The main statistics that were compared are diversity estimates, species abundance, and ability of reconstructing *de novo* the metagenome in terms of length and completeness. Our results show that diversity indices of complex prokaryotic, eukaryotic and viral communities can be accurately estimated with 500,000 reads or less, although particularly complex samples may require 1,000,000 reads. On the contrary, any task involving the reconstruction of the metagenome performed poorly, even with the largest simulated subsample (1,000,000 reads). The length of the reconstructed assembly was smaller than the length obtained with the full dataset, and the proportion of conserved genes that were identified in the meta-genome was drastically reduced compared to the full sample. Shallow shotgun metagenomics can be a useful tool to describe the structure of complex matrices, but it is not adequate to reconstruct—even partially—the metagenome.

Open Peer Review


Reviewer Status

	Invited Reviewers			
	1	2	3	4
version 4 (revision) 22 Jan 2020			 report	 report
version 3 (revision) 29 Jul 2019			 report	 report
version 2 (revision) 22 Mar 2019		 report	 report	
version 1 08 Nov 2018	 report	 report		

- Alejandro Sanchez-Flores** , National Autonomous University of Mexico (UNAM)), Cuernavaca, Mexico
- José F. Cobo Diaz** , Université de Brest, Plouzané, France

Keywords

high-throughput sequencing, metagenome, metagenomics, next generation sequencing, alpha diversity, complex matrices

3 **Francesco Dal Grande** , Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

LOEWE Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

4 **Marcus Claesson** , University College Cork, Cork, Ireland

Shriram Patel, University College Cork, Cork, Ireland

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Federica Cattonaro (fcattanaro@igatechnology.com), Fabio Marroni (marroni@appliedgenomics.org)

Author roles: **Cattonaro F:** Conceptualization, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Spadotto A:** Investigation; **Radovic S:** Conceptualization, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Marroni F:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Metagenome sequencing of B1 and B2 was financed by Corvelva (non-profit association, Veneto, Italy), in the frame of a service contract with IGA Technology Services. No other grants were involved in supporting the work.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Cattonaro F *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Cattonaro F, Spadotto A, Radovic S and Marroni F. **Do you cov me? Effect of coverage reduction on metagenome shotgun sequencing studies [version 4; peer review: 2 approved, 2 not approved]** F1000Research 2020, 7:1767 (<https://doi.org/10.12688/f1000research.16804.4>)

First published: 08 Nov 2018, 7:1767 (<https://doi.org/10.12688/f1000research.16804.1>)

REVISED Amendments from Version 3*Reviewers Francesco Dal Grande*

We shared the issue of the reviewer regarding the detection of a false positive in the mock community and we further investigated the issue. Besides the full not database (which has been used throughout the study for its general applicability), we repeated the analysis on the mock community using three additional databases and we noticed that basically no database is free from the problem of misclassification. The fact the taxonomic classification of false positives differs across databases suggests that the problem is in presence of (few) incorrectly classified sequences in the databases itself. We thus added a warning in the discussion, regarding the need of carefully interpreting the results, especially when unexpected species are identified.

Reviewer Marcus Claesson (in collaboration with Shriram Patel)

Following the reviewers' suggestion, we performed a Procrustes analysis to assess if decrease of coverage affected beta-diversity estimates, and show our results in [Figure 6](#). We observed that the reduction of coverage only moderately altered the beta-diversity estimates. As expected, the greater the reduction of coverage, the greater the effect. When sequencing 10,000 reads the data were too sparse to obtain reliable beta-diversity estimates.

Any further responses from the reviewers can be found at the end of the article

Introduction

Shotgun metagenomics offers the possibility to assess the complete taxonomic composition of biological matrices and to estimate the relative abundances of each species in an unbiased way^{1,2}. It allows to agnostically characterize complex communities containing eukaryotes, bacteria and also viruses.

Metagenome shotgun high-throughput sequencing has progressively gained popularity in parallel with the advancing of next-generation sequencing (NGS) technologies^{3,4}, which provide more data in less time at a lower cost than previous sequencing techniques. This allows the extensive application to study the most various biological mixtures such as environmental samples^{5,6}, gut samples⁷⁻⁹, skin samples¹⁰, clinical samples for diagnostics and surveillance purposes¹¹⁻¹⁴ and food ecosystems^{15,16}. Another, more traditional approach currently used to assign taxonomy to DNA sequences is based on the sequencing of target conserved regions. Metabarcoding method relies on conserved sequences to characterize communities of complex matrices. These include the highly variable region of 16S rRNA gene in bacteria¹⁷, the nuclear ribosomal internal transcribed spacer (ITS) region for fungi¹⁸, 18S rRNA gene in eukaryotes¹⁹, cytochrome c oxidase sub-unit I (*COI* or *cox1*) for taxonomical identification of animals²⁰, *rbcL*, *matK* and *ITS2* as the plant barcode²¹. Metabarcoding has the advantage of reducing sequencing needs, since it does not require sequencing of the full genome, but just a marker region. On the other hand, given the commonly used approaches, characterization of microbial and eukaryotic communities requires different primers and library preparations²². In addition, bias in the amplification of the targeted sequence is a major issue in targeted metagenomics studies and constitutes an important limitation of metabarcoding²³. Several studies suggested that whole shotgun metagenome sequencing is more effective in the characterization of metagenomics samples compared to target

amplicon approaches, with the additional capability of providing functional information regarding the studied approaches^{24,25}.

Current whole shotgun metagenome experiments are performed obtaining several million reads^{5,7}. However, obtaining a broad characterization of the relative abundance of different species might be achieved with lower number of reads.

To test this hypothesis, we analyzed ten samples (eight sequenced in the framework of this study and two retrieved from the literature) derived from different complex matrices using whole metagenomics approach and tested accuracy of several summary statistics as a function of the reduction of the number of reads used for analysis. The selection of samples belonging to different matrices with distinct characteristics enabled to understand if the results are generally applicable and, if this is not the case, which are the features with the greatest impact on results.

In summary, the aim of the present work is to test the effect of the reduction of sequencing depth on 1) estimates of diversity and species richness in complex matrices; 2) estimates of abundance of the species present in the complex matrix, and 3) completeness of *de novo* reconstruction of the genome of the species present in the samples. To assess the consistency of our approach, we selected samples characterized by different levels of species richness and by different relative abundance of prokaryotic and eukaryotic organisms and viruses.

Methods**Samples description and DNA extraction**

The following samples were used in the present work: the mock community DNA sample "20 Strain Staggered Mix Genomic Material" ATCC® MSA-1003™ (short name: A1), two biological medicines (B1 and B2), two horse fecal samples (F1 and F2), three food samples (M1, M2, and M3), and two human fecal samples (V1 and V2).

Biological medicines were two different lots of live attenuated MPRV vaccine, widely used for immunization against measles, mumps, rubella and chickenpox in infants. Lyophilised vaccines were resuspended in 500 µl sterile water for injection and DNA extracted from 250 µl using Maxwell® 16 Instrument and the Maxwell® 16 Tissue DNA Purification Kit (Promega, Madison, WI, USA) according to the manufacturer's instructions. The vaccine composition declared by the producer is the following: live attenuated viruses: 1) Measles (ssRNA) Swartz strain, cultured in embryo chicken cell cultures; Mumps (ssRNA) strain RIT 4385, derived from the Jeryl Linn strain, cultured in embryo chicken cell cultures; Rubella (ssRNA) Wistar RA 27/3 strain, grown in human diploid cells (MRC-5); Varicella (dsDNA) OKA strain grown in human diploid cells (MRC-5).

Horse feces from two individuals were processed as follows: 100 mg of starting material stored in 70% ethanol were used for DNA extraction using the QIAamp PowerFecal DNA Kit (QIAGEN GmbH, Hilden, Germany), according to the manufacturer's instructions.

Food samples were raw materials of animal and plant origin, used to industrially prepare bouillon cubes. DNA extractions from those three samples were performed starting from 2 grams of material each, using the DNeasy mericon Food Kit (QIAGEN GmbH, Hilden, Germany), according to the manufacturer's instructions. The declared sample composition was *Agaricus bisporus* for M1, spice (*Piper nigrum*) for M2 and mix of animal extracts for M3.

The mock community declared components are: 0.18% *Acinetobacter baumannii* (ATCC 17978), 0.02% *Actinomyces odontolyticus* (ATCC 17982), 1.80% *Bacillus cereus* (ATCC 10987), 0.02% *Bacteroides vulgatus* (ATCC 8482), 0.02% *Bifidobacterium adolescentis* (ATCC 15703), 1.80% *Clostridium beijerinckii* (ATCC 35702), 0.18% *Cutibacterium acnes* (ATCC 11828), 0.02% *Deinococcus radiodurans* (ATCC BAA-816), 0.02% *Enterococcus faecalis* (ATCC 47077), 18.0% *Escherichia coli* (ATCC 700926), 0.18% *Helicobacter pylori* (ATCC 700392), 0.18% *Lactobacillus gasseri* (ATCC 33323), 0.18% *Neisseria meningitidis* (ATCC BAA-335), 18.0% *Porphyromonas gingivalis* (ATCC 33277), 1.80% *Pseudomonas aeruginosa* (ATCC 9027), 18.0% *Rhodobacter sphaeroides* (ATCC 17029), 1.80% *Staphylococcus aureus* (ATCC BAA-1556), 18.0% *Staphylococcus epidermidis* (ATCC 12228), 1.80% *Streptococcus agalactiae* (ATCC BAA-611), 18.0% *Streptococcus mutans* (ATCC 700610).

DNA purity and concentration were estimated using a NanoDrop Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA) and Qubit 2.0 fluorimeter (Invitrogen, Carlsbad, CA, USA).

Human fecal samples V1 and V2 derive from a study investigating the virome composition of feces of uncontacted Amerindians²⁶. Data are publicly available on Sequence Read Archive (SRA, <https://www.ncbi.nlm.nih.gov/sra/>). The two samples with the highest sequencing depth were chosen; accession numbers are SRR6287060 and SRR6287079, respectively.

Whole metagenome DNA library construction and sequencing

DNA library preparations were performed according to manufacturer's protocol, using the kit Ovation® Ultralow System V4 1–96 (Nugen, San Carlos, CA). Library prep monitoring and validation were performed both by Qubit 2.0 fluorimeter (Invitrogen, Carlsbad, CA, USA) and Agilent 2100 Bioanalyzer DNA High Sensitivity Analysis kit (Agilent Technologies, Santa Clara, CA). Obtained DNA concentrations were as follows: A1 8 ng/μl (total amount = 640 ng), B1 10.7 ng/μl (total amount = 535 ng), B2 9.41 ng/μl (total amount = 470.5 ng), F1 42.3 ng/μl (total amount = 4,230 ng), F2 22.6 ng/μl (total amount = 2,260 ng), M1 16.6 ng/μl (total amount = 1,494 ng), M2 1.87 ng/μl (total amount = 168.3 ng), M3 16 ng/μl (total amount = 640 ng).

Cluster generation was then performed on Illumina cBot and flowcell HiSeq SBS V4 (250 cycle), and sequenced on HiSeq2500 Illumina sequencer producing 125bp paired-end reads.

Samples F1 and F2 were loaded on flowcell HiSeq Rapid SBS Kit v2 (500 cycles) producing 250bp paired-end reads. The estimated library insert sizes were: 539 bp (A1), 531 bp (B1), 536 bp (B2), 620 bp (F1), 620 bp (F2), 342 bp (M1), 178 bp (M2), 496 bp (M3). Samples were sequenced in different runs and pooled with other libraries of similar insert sizes.

The CASAVA Illumina Pipeline version 1.8.2 was used for base-calling and de-multiplexing. Adapters were masked using cutadapt²⁷. Masked and low quality bases were filtered using *erne-filter* version 1.4.6.²⁸.

Bioinformatics analysis

The bioinformatics analysis performed in the present work are summarized in **Figure 1**; a standard pipeline for reproducing the main steps of analysis is available on GitHub (<http://www.doi.org/10.5281/zenodo.2593798>).

Since different read lengths among samples may constitute an additional confounder in analysis, 250 bp long reads belonging

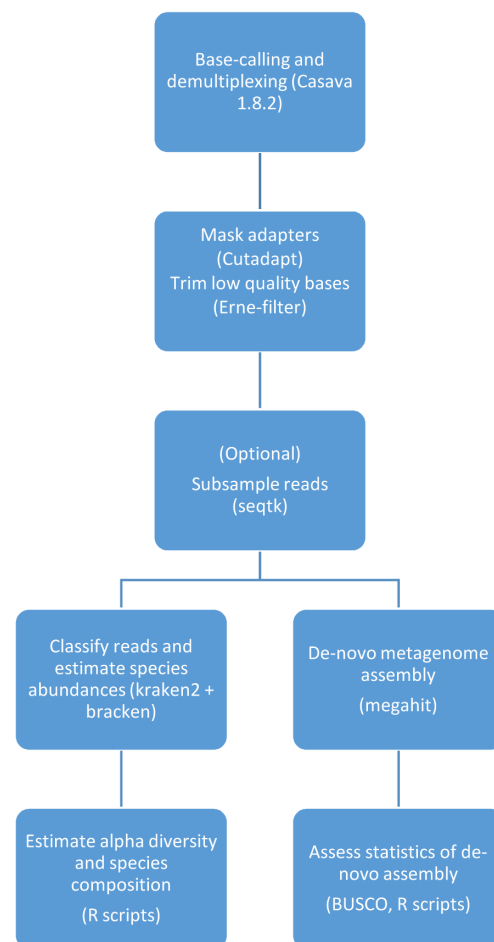


Figure 1. Workflow of the main bioinformatics analysis performed in the present work.

to F1, F2, V1 and V2 were trimmed to a length of 125bp using fastx-toolkit version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/) before analysis.

Reduction in coverage was simulated by randomly sampling a fixed number of reads from the full set of reads. Subsamples of 10,000, 25,000, 50,000, 100,000, 250,000, 500,000 and 1,000,000 reads were extracted from the raw reads using *seqtk* version 1.3. To estimate the variability due to random effects, subsampling was replicated five times for each simulated depth and 99% confidence limits were estimated and plotted.

To classify the largest possible number of prokaryotes, eukaryotes and viruses, reads were classified against the complete NCBI nt database using *kraken2*, version 2.0.6²⁹. The nt database was converted to *kraken2* format using the built-in *kraken2-build* script with default parameters. Among the most significant parameters, *kmer* size for the database is by default set to 35 and the minimizer length to 31. A simplified representation of species composition was obtained using *Krona*³⁰. To obtain accurate species abundances *Bracken*, version 2.2³¹ was used on species supported by at least 10 reads; since the reads used in all the experiments were 125bp, the *bracken* database was built using 125bp kmers.

The threshold for declaring a species as present was set according to results of a performance analysis on the mock community (A1) for which species presence and abundance was known. Performance was assessed using F1-score, calculated as $2*TP/(2*TP+FP+FN)$, as previously reported³². F1-score is a measure used in performance analysis when the number of true negatives is extremely high or unknown.

The effect of the selected database on reads classification was assessed on the mock community full sequencing experiment, by observing the variation in present and absent species when using different databases. In addition to the nt database built explicitly for this study, the standard *kraken2* database, the *minikraken2* v1 database and the *minikraken2*_v2 were used. The standard *kraken2* database contains complete genomes in RefSeq for the bacterial, archaeal, and viral domains, along with the human genome and a collection of known vectors (*UniVec_Core*), and the *minikraken2* v1 database contains RefSeq bacteria, archaea, and viral libraries, and the *minikraken2*_v2 database contains RefSeq bacteria, archaea, and viral libraries and the GRCh38 human genome.

Bracken database was built for 125bp kmers for the standard database. *Minikraken2* instead is distributed as a prebuilt database, from which it is not possible to build the *bracken* database, but for which *bracken* databases with kmers distribution of 100bp, 150bp and 200 kmers are available. kmers 100bp and 150bp were tested, since they are the closest to the read length used in this study.

Observed number of taxa, Shannon's diversity index³³ and Pielou's index³⁴ were estimated using the R package *vegan* version 2.4.2³⁵ or base R, version 3.3.3³⁶ functions. The number of

observed taxa was computed as the number of species passing the detection threshold.

Shannon diversity index is estimated as

$$H = - \sum_{i=1}^N p_i * \ln(p_i)$$

Where N is the total number of species and p_i is the frequency of the species i .

Pielou's evenness index is estimated as

$$J = \frac{H}{\ln S}$$

Where H is Shannon's diversity index and S is the total number of observed species. The value $\ln S$ corresponds to the maximum possible value of H , observed when all species have the same frequency.

The effect of sequencing depth on beta-diversity was assessed using the *procrustes* and *protest* functions, implemented in *vegan*.

Assembly of the metagenome was performed using *megahit* version 1.1.2³⁷ with default parameters, with *kmer* sizes varying as follows: 21, 29, 39, 59, 79, 99, 119, 141. Reconstructed contigs were binned at the species level using *kraken2*, and only contigs assigned to species above the detection threshold were used for further analysis. Completeness of the assemblies of each species was assessed using *BUSCO*³⁸. For each species, the proportion of the reconstructed genes was measured as the proportion of genes that were fully reconstructed, plus the proportion of genes that were partially reconstructed. For each sample, results were then averaged over detected species to provide the average proportion of reconstructed genes. *BUSCO* analysis was performed on prokaryotic database for all the samples with the exception of M1 (predominantly composed by fungi) for which the fungal database was used.

Unless otherwise specified, all the analysis were performed using R 3.3.3³⁶.

Results

Determination of detection threshold

The mock community sample "20 Strain Staggered Mix Genomic Material" (ATCC® MSA-1003™) was used as a reference to control performance of sequencing and classification procedures at various depth. The community includes a total of 20 bacterial species, of which 5 have a frequency of 0.02%, 5 a frequency of 0.18%, 5 a frequency of 1.8% and 5 a frequency of 18%.

Results of the performance analysis on the mock dataset are shown in Table 1. The highest F1 score (0.8) was obtained when applying a 0.1% threshold. Using this threshold, 14 species were correctly identified while 6 of them were not detected. Five out of the 6 undetected species had a nominal frequency of 0.02%; the sixth undetected species was *Helicobacter pylori*, with a nominal frequency of 0.18%, for which we recorded

a frequency of 0.096%, below the 0.1% threshold. The only false positive was *Shigella flexneri*, a species highly related to *Escherichia coli*³⁹, that was observed at a frequency of 0.128%. Based on these results we used a threshold of 0.1% for declaring a species as present in a sample.

Sample composition

Summary statistics for the samples included in the study are shown in Table 2.

The number of reads obtained in the samples selected for the present study ranged from slightly more than 1 million (sample

V1) to more than 12 million (sample F1). The number of species identified in each sample ranged from 4 in sample B1 to 138 in sample M2. Figure 2 summarizes the composition of each sample at the Phylum level. Viruses are aggregated at the division level. Only phyla more abundant than 1% were plotted.

Table 2. Summary statistics for the full samples included in the study.

Sample	N reads	N species
A1	4,969,245	15
B1	11,031,061	4
B2	3,830,083	9
F1	12,472,553	127
F2	10,780,450	126
M1	1,898,011	5
M2	1,558,975	138
M3	1,867,879	21
V1	1,300,221	84
V2	2,001,984	12

Sample: Short name of the sample.
N reads: Number of reads obtained for the sample in the full sequencing experiment.
N species: number of species identified in the sample

Table 1. Results of performance analysis. Threshold (%): detection threshold, expressed as percentage of assigned reads. **TP:** true positives. **FP:** false positive. **FN:** false negatives. **F1:** F1 score.

Threshold (%)	TP	FP	FN	F1
0.001	19	188	1	0.17
0.005	19	49	1	0.43
0.01	19	32	1	0.54
0.05	15	6	5	0.73
0.1	14	1	6	0.8
0.5	10	0	10	0.67

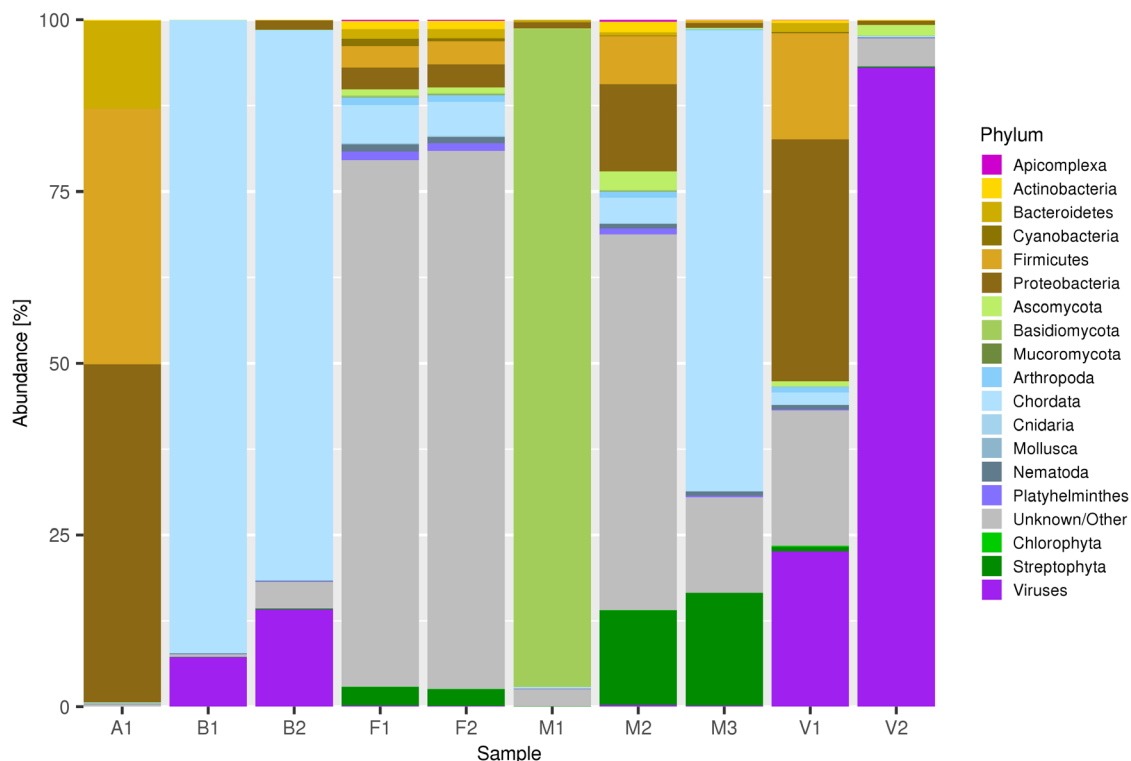


Figure 2. Phylum composition of the samples. Only phyla represented by at least 1% of the reads are shown. Viruses are presented at division level. Unclassified reads and reads assigned to rare phyla are aggregated under the name "Unknown/Other".

Reads that were either unclassified or assigned to rare phyla were aggregated under the name “Unknown/Other”. Samples B1, B2 and M3 were mainly composed of Chordata, sample M1 was mostly composed of Basidiomycota, and sample V2 was mainly composed of Viruses. Samples F1, F2 and, to a lesser extent, M2 were characterized by a large proportion of reads unclassified or assigned to rare phyla. For a more detailed view of raw taxonomy composition, interactive html Chrona are available for download on Open Science Framework

(<https://osf.io/y7c39/>), under the project “Do you cov me”, DOI: [10.17605/OSF.IO/Y7C39](https://doi.org/10.17605/OSF.IO/Y7C39).

Effect of the choice of database

The effect of the selected database on reads classification was assessed only on the mock community. Results are shown in [Table 3](#). All the Spearman correlation coefficients were >0.9 (not shown). Estimated abundances according to the minikraken2 v1 database were very similar to those obtained

Table 3. Effect of the database choice on the assignment of species in the mock community.

Species	Abundance	v1_100	v1_150	v2_100	v2_150	standard	nt
Acinetobacter baumannii	0.18	0.219	0.219	0.219	0.219	0.177	0.326
Actinomyces odontolyticus	0.02	NP	NP	NP	NP	NP	NP
Bacillus cereus	1.8	4.002	4.039	3.962	4	3.567	4.259
Bacteroides vulgatus	0.02	NP	NP	NP	NP	NP	NP
Bifidobacterium adolescentis	0.02	NP	NP	NP	NP	NP	NP
Clostridium beijerinckii	1.8	3.027	3.023	2.975	2.971	2.524	4.69
Cutibacterium acnes	0.18	0.136	0.136	0.137	0.137	0.134	0.134
Deinococcus radiodurans	0.02	NP	NP	NP	NP	NP	NP
Enterococcus faecalis	0.02	NP	NP	NP	NP	NP	NP
Escherichia coli	18	23.793	23.986	23.619	23.95	25.945	20.043
Helicobacter pylori	0.18	NP	NP	NP	NP	NP	NP
Lactobacillus gasseri	0.18	NP	NP	NP	NP	NP	0.103
Neisseria meningitidis	0.18	0.118	0.118	0.118	0.119	0.121	0.12
Porphyromonas gingivalis	18	11.834	11.829	11.844	11.84	11.811	11.706
Pseudomonas aeruginosa	1.8	3.067	3.078	3.055	3.065	3.145	3.233
Rhodobacter sphaeroides	18	22.588	22.568	22.448	22.426	23.22	23.047
Staphylococcus aureus	1.8	2.098	1.994	1.781	1.734	2.494	1.537
Staphylococcus epidermidis	18	13.76	13.942	13.821	13.938	13.112	16.701
Streptococcus agalactiae	1.8	1.067	1.075	1.079	1.089	0.645	1.026
Streptococcus mutans	18	10.898	10.889	10.877	10.864	10.851	11.543
Bacillus anthracis	ND	NP	NP	0.125	0.114	NP	NP
Bacillus thuringiensis	ND	0.359	0.343	0.322	0.311	0.773	NP
Escherichia albertii	ND	0.194	0.194	0.179	0.179	NP	NP
Escherichia marmotae	ND	0.113	0.113	0.111	0.111	NP	NP
Homo sapiens	ND	NP	NP	0.109	0.11	0.116	NP
Salmonella enterica	ND	0.97	0.697	1.573	1.165	NP	NP
Shigella dysenteriae	ND	0.239	0.242	0.209	0.212	NP	NP
Shigella flexneri	ND	NP	NP	NP	NP	NP	0.128
Staphylococcus lugdunensis	ND	NP	NP	NP	NP	0.108	NP
Streptococcus pyogenes	ND	NP	NP	NP	NP	0.495	NP

Species: Binomial nomenclature of the species. **Abundance:** declared abundance. **v1_100:** estimated abundance using minikraken2 v1 database and bracken database kmer length of 100. **v1_150:** estimated abundance using minikraken2 v1 database and bracken database kmer length of 150. **v2_100:** estimated abundance using minikraken2 v2 database and bracken database kmer length of 100. **v2_150:** estimated abundance using minikraken2 v2 database and bracken database kmer length of 150. **standard:** estimated abundance of species using standard database. **nt:** estimated abundance of species using nt database. **ND:** Not declared in the mock community. **NP:** Not present according to the detection threshold of 0.1%.

according to the minikraken2 v2 database (irrespective of the kmer length which slightly altered the results). The main differences between the two were observed for species not declared in the mock community (*i.e.* false positives, indicated by ND in Abundance column), such as *Homo sapiens*, *Bacillus anthracis* and *Salmonella enterica*.

Homo sapiens contamination was detected with the Minikraken2 v2 (the only Minikraken2 containing human sequences) and using the standard database. No contamination was detected using the nt database, for which *Homo sapiens* was recorded with a frequency of 0.086% and was therefore below the detection threshold of 0.1%. *Bacillus anthracis* was only detected when using the Minikraken2 v2 database. *Salmonella enterica* was detected in all the experiments involving the Minikraken2 databases, but it was the only species for which substantial variation in the estimated frequencies was observed, ranging from 0.697% to 1.573% according to Minikraken2 v1 with 150 kmers and Minikraken2 V2 with 100 kmers, respectively.

False positives were generally low: the total contribution of false positives ranged from 0.168% for nt to 2.628% for Minikraken2 v2 100 kmers. However, none of the databases was immune to false positives. The nt database showed only one false positive, *Shigella flexneri*, while minikraken and standard databases showed more than one false positive each.

The estimated abundances of the declared species were in excellent agreement across databases, with some exceptions. For *Escherichia coli*, with declared abundance of 18%, estimated abundances ranged from 20.043% to 25.954% when using nt and the standard database, respectively. *Staphylococcus epidermidis*, with declared abundance of 18%, was estimated at 13.112% by the standard database and 16.701% by nt. Finally, *Clostridium beijerinckii*, declared at 1.8% was estimated at 2.524% by the standard Kraken database and at 4.69% by nt database.

Species abundance

The effect of reducing sequencing depth on the accuracy of taxonomical classification was assessed by using the mock community, given its known composition. Expected and observed abundances of the 20 mock species maintained a high correlation ($r=0.94$) even when decreasing sequencing to 10,000 reads (Figure 3). However, decreasing sequencing depth caused an increase in uncertainty, as shown by the broader confidence intervals for lower depths, particularly for rare species.

We also measured the correlation in species abundance between the full and reduced datasets in all the samples (Figure 4). The correlation between the two quantities was in general high and improved at increasing sequencing depths. The average Spearman's correlation coefficient between full and reduced samples ranged from 0.71 in the 25,000 reads subsample to 0.94 in the 1,000,000 reads subsample.

Diversity analysis

We further evaluated the impact of reducing sequencing depth on several diversity measures, such as the observed number of

taxa, Shannon's diversity index and Pielou's diversity index (Figure 5).

Samples F1, F2, M2 and V1 had more than 50 taxa, and all the remaining samples had less than 25 (Table 2 and Figure 5A). Downsampling only produced significant differences in the four samples with high number of observed taxa (panel A). In samples F1 and F2, intermediate levels of downsampling (*e.g.* 100,000 reads) caused an increase in the number of species exceeding the 0.1% abundance threshold.

Shannon's diversity index (panel B) is a widely used method to assess biological diversity of ecological and microbiological communities. The effect of sequencing depth on Shannon's diversity index is negligible for most samples, with the exception of the samples with the richest species composition (F1, F2, and M2) for which downsampling led to a significant variation in the estimate.

Pielou's index (panel C) is a measure of the species' distribution evenness. Values close to 1 denote equipotent species, and lower values denote uneven distribution of species relative abundance. The effect of the number of reads on Pielou's index is negligible for all samples.

Effect of sequencing depth on beta-diversity using procrustes analysis is shown in Figure 6. Procrustes analysis is shown between the full dataset (starting point of the arrows) and each replicate of each of the reduced dataset (arrival points of the arrowheads). The smallest dataset was excluded for the analysis because it did not have sufficient information for procrustes analysis. The relative positions between the full dataset and the reduced datasets tended to remain similar across different subsampling, with some notable exception, such as V2 in the 250,000 reads sample and A1 in the 500,000 sample. F1 and F2, always clustered together. Correlation was good between all the matrices, but it was excellent ($r>0.9$) only for the comparison of the full dataset with the 1,000,000 reads subsample.

Completeness of *de novo* assembly

We investigated the effect of coverage reduction on the completeness of *de novo* assembly. We reconstructed the metagenome of the full and reduced datasets and compared the completeness of the reconstructed genomes. Results are summarized in Figure 5 (panel D). As expected, the size of the assembly was strongly influenced by the sequencing depth. Assembly size for the full dataset ranged from less than 1 Mb (V2) to nearly 100 Mb (F1 and F2). A decrease in the sequencing depth led to a steady decrease in assembly size in all samples. At 1,000,000 reads the size ranged from slightly more than 100 kb (V2) to slightly more than 10Mb (A1 and M1).

BUSCO analysis³⁸ was used as an additional measure to assess the completeness of the reconstructed metagenome.

First, we assessed the performance on the A1 mock community for the full set of reads (4,969,245 reads) and for the largest subset, (1,000,000 reads). Genomes of species present at 0.02% and 0.18% were not reconstructed, while genomes of species present at 1.8% or 18% were reconstructed. The proportion

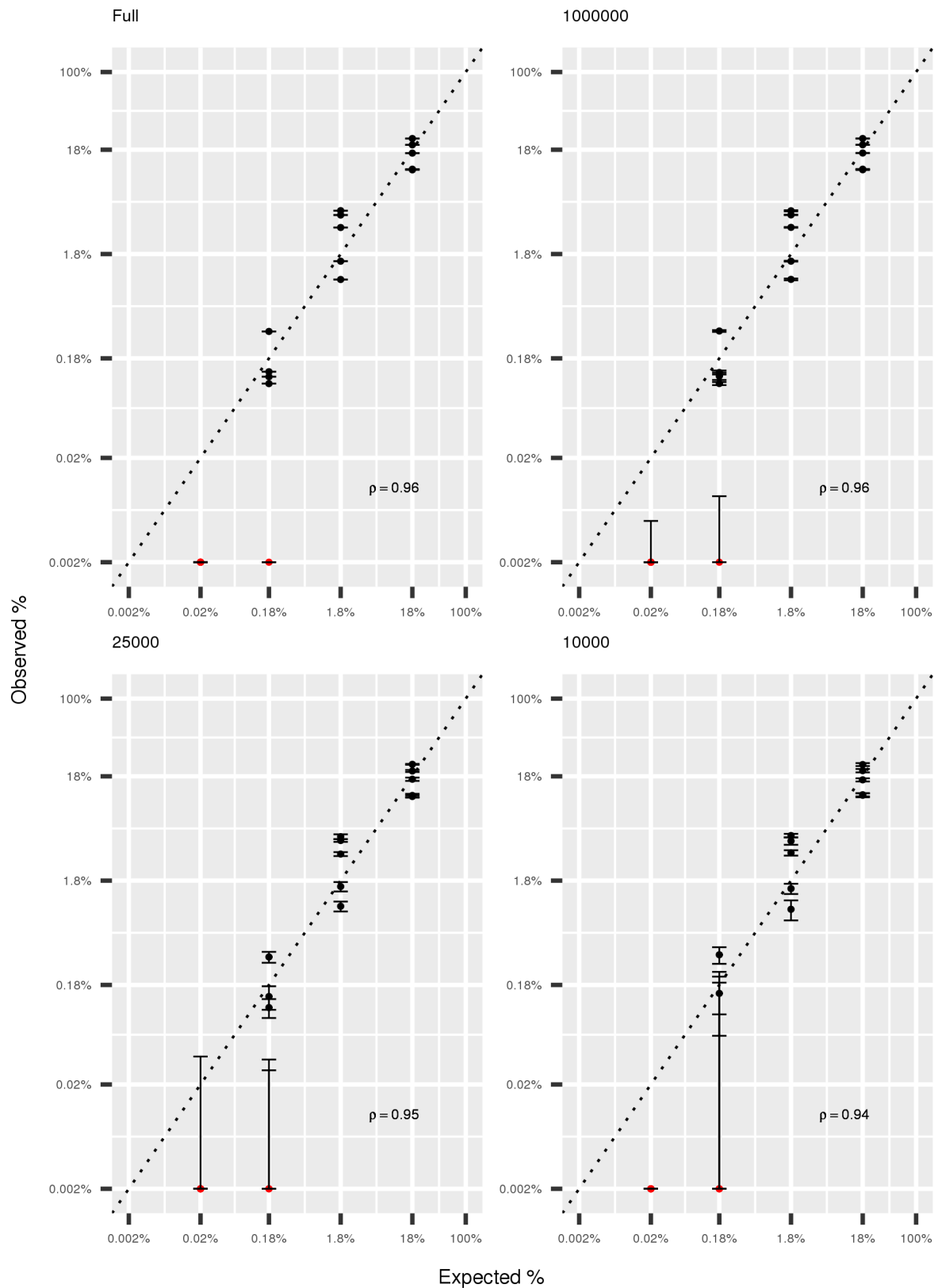


Figure 3. Observed and expected abundance of bacterial species present in the mock community "20 Strain Staggered Mix Genomic Material" (ATCC® MSA-1003™) at varying sequencing depths. In red, species identified at frequency lower than the selected threshold of 0.1% and arbitrarily plotted at 0.002%. Error bars represent 95% confidence intervals obtained from five resampling experiments. Both axes are plotted in log scale to facilitate visualization of rare species.

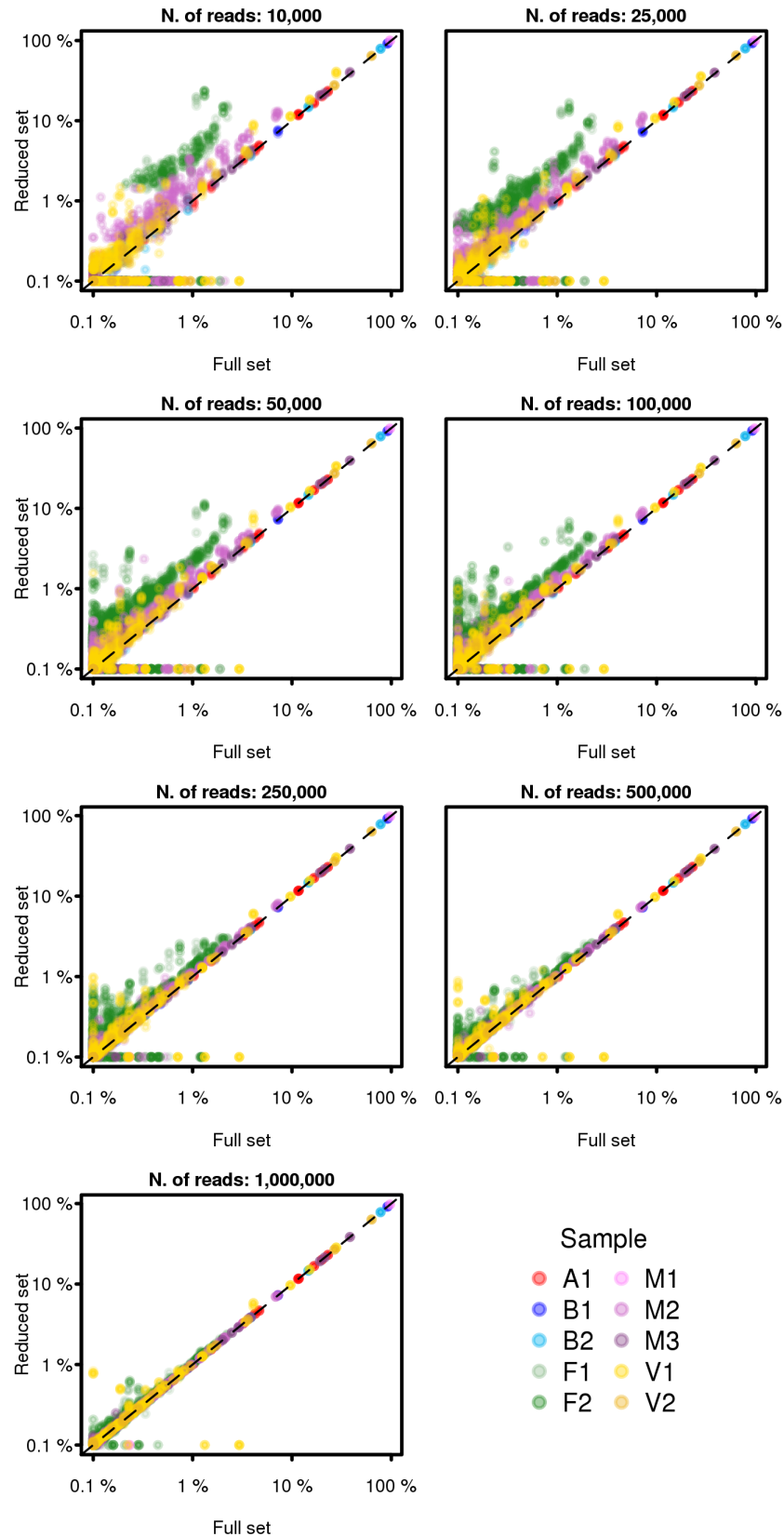


Figure 4. Correlation of species abundance estimated using full and reduced datasets. Data for all the five subsampled replicates are plotted. Each point (colored by sample of origin) represents a given species. Both axes are plotted in log scale to facilitate visualization of rare species.

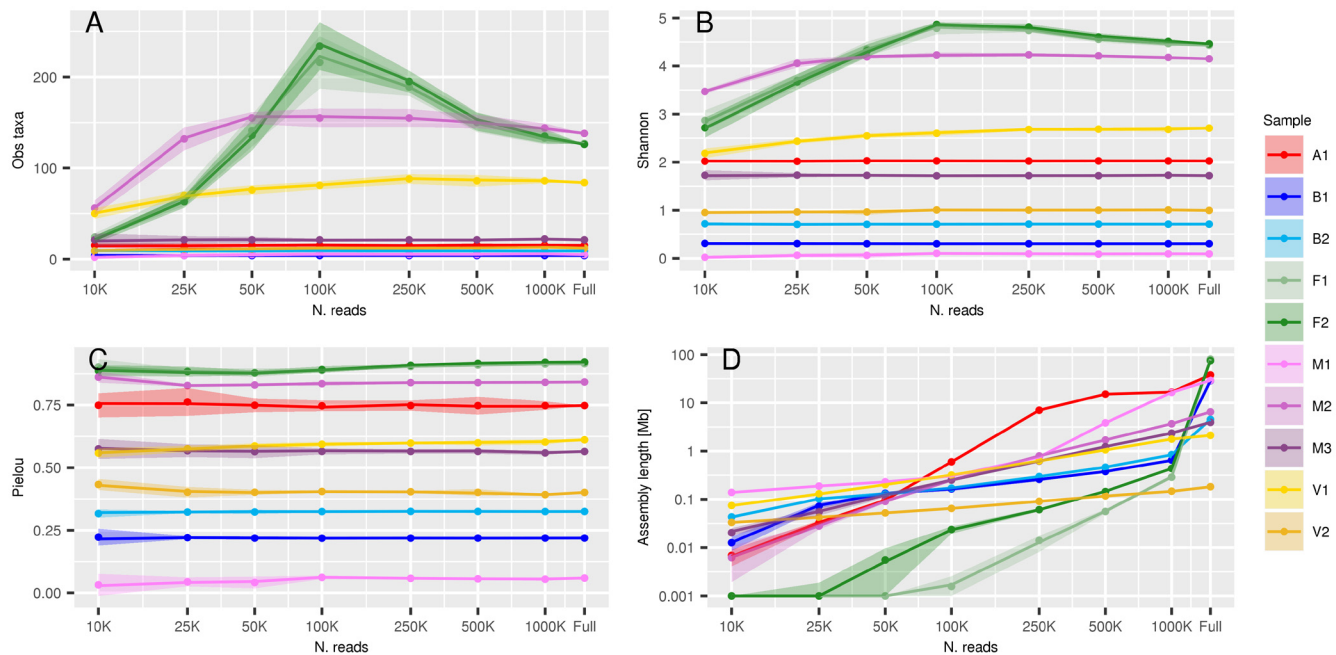


Figure 5. Effect of reduction of sequencing depth on: **A)** Observed number of taxa, **B)** Shannon's diversity index, **C)** Pielou's diversity index, and **D)** Total length of *de novo* assembly. In all panels X axis is in log scale and Y axis is in linear scale with the exception of panel F, in which both axes are in log scale. Shaded areas represent the confidence limits of resampling experiments. The number of reads used for the analysis is shown from the smallest (10,000) on the left, to the full dataset on the right.

of BUSCO genes reconstructed using the whole set of reads ranged 59%–99% for species present at 1.8% and 93%–99% in the most abundant species (Figure 7). At 1,000,000 reads the proportion of reconstructed BUSCO genes dropped to 0.7%–3% for species present at 1.8% while it ranged 93%–99% for the species present at 18%.

In addition, we plot in Figure 8 the proportion of reconstructed genes in full (X axis) and reduced (Y axis) datasets obtained by randomly sampling 1,000,000 reads. The proportion of reconstructed BUSCO genes is very low even in the full samples, indicating that in general the sequencing depth is still too low to obtain an accurate reconstruction of the metagenome. Only in samples A1 and M1, the average proportion of BUSCO genes reconstructed in the full sample was greater than 10%. Reducing sequencing depth to 1,000,000 reads significantly lowered the proportion of reconstructed genes in all the samples, as testified by the fact that all the points and their confidence limits lie below the diagonal.

Discussion

We set out to test the effect of the reduction of sequencing depth in metagenome shotgun sequencing experiments on 1) estimates of diversity and species richness; 2) estimates of species abundance, and 3) completeness of *de novo* reconstruction of the genome of the species present in complex matrices. We selected ten heterogeneous samples that underwent whole genome DNA-sequencing. This was also true for vaccine samples B1 and B2, several components of which are ssRNA viruses, and could not be detected using this approach. Indeed, the determination of

the ssRNA components in vaccines was not the aim of the present study.

We used the mock community to determine the optimal detection threshold and then performed all the analysis enforcing the selected threshold. Five of the species composing the mock community had a declared abundance of 0.02% and could not by definition be detected using the threshold. However, the threshold caused the appearance of only one false positive, and resulted in a F1 score of 0.8. The false positive species is *Shigella flexneri* a sister species of *Escherichia coli*, and is likely a result of misclassification of a proportion of reads. A possible explanation is that in the used database, genomic sequence of *Escherichia coli* is classified as *Shigella flexneri*. We thus further investigated if the use of different databases could change this behavior, by classifying the mock community reads against the standard database and the two Minikraken2 databases distributed with kraken2. We noticed that while the performance was overall in excellent agreement, there were some differences in the results obtained with each database. In particular, each database recorded at least one false positive species. Thus, we suggest researchers to cautiously interpret results, especially when unexpected species are identified.

To the best of our knowledge, this is the first published work reporting the observed frequencies of a mock community using shotgun high-throughput sequencing. However, previous studies performed very extensive analysis on target 16s sequencing of mock communities, and reported large deviations from the expected values, dependent on sequencing

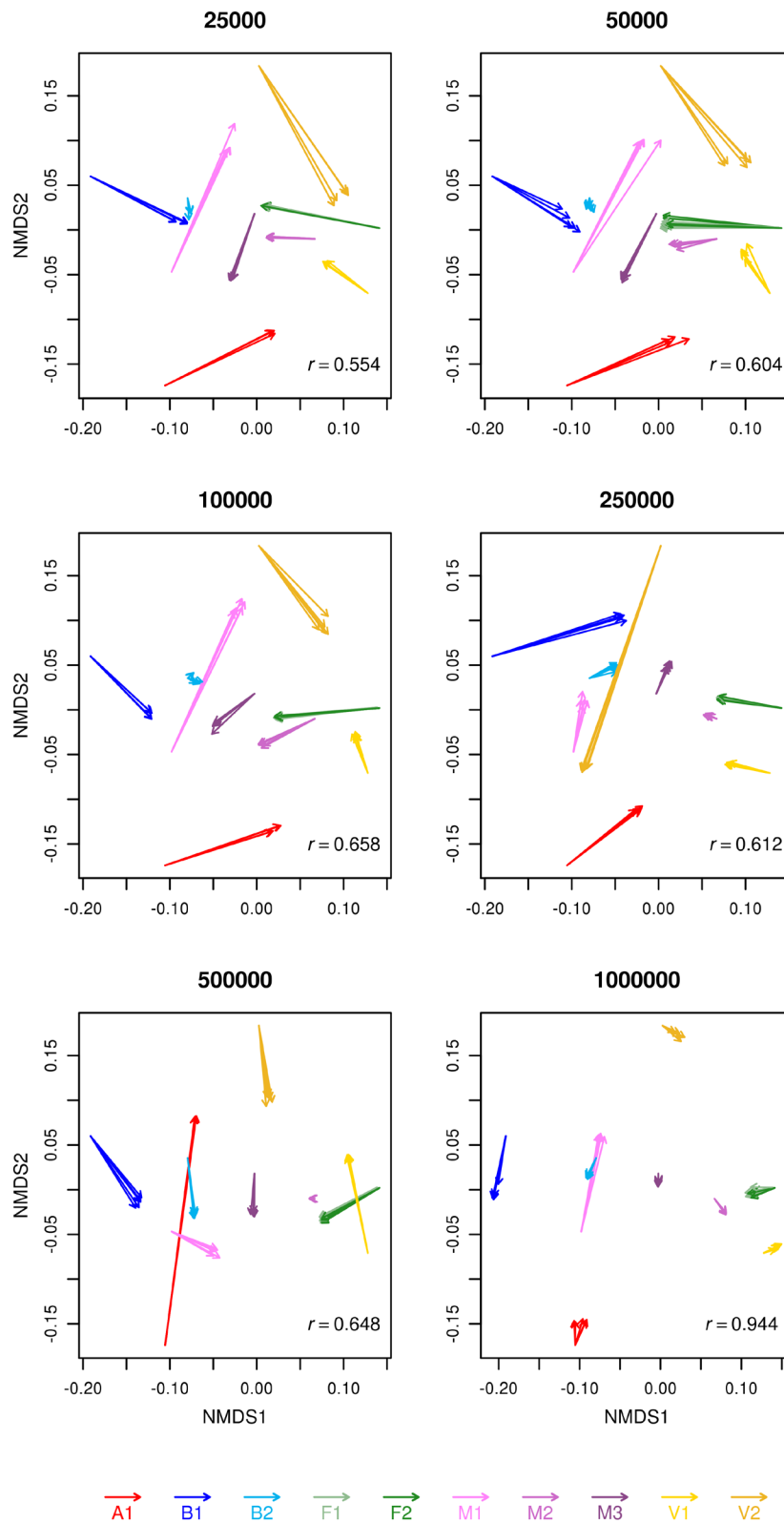


Figure 6. Procrustes analysis. Arrows start from the NMDS matrix of the full sample and the arrowhead ends at the NMDS matrices of the five subsampled replicates for each sample. The correlation between the matrices are shown at the bottom right of each plot. All correlations were significant ($p \leq 0.001$).

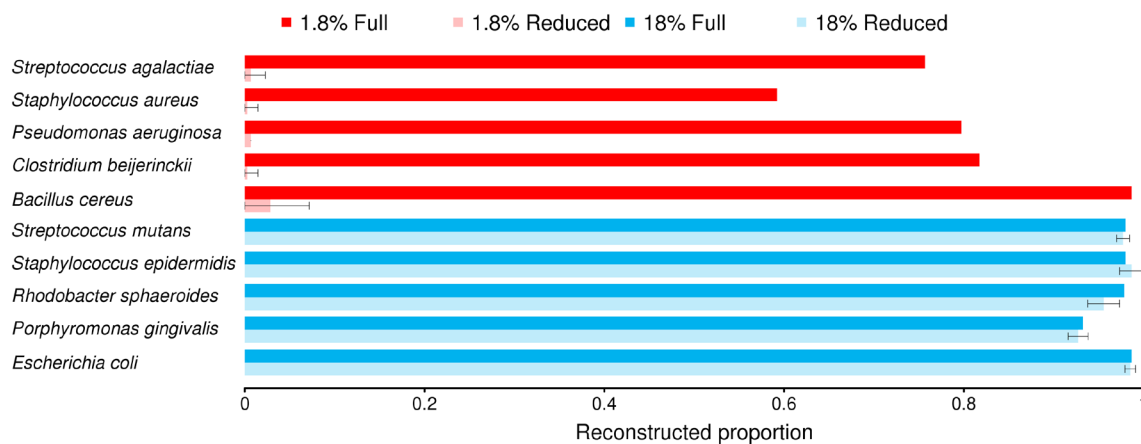


Figure 7. Proportion of BUSCO genes reconstructed in the full set of reads, and in a subset of 1,000,000 reads. Error bars represent 95% confidence intervals based on five subsampling experiments. **1.8%:** Species with nominal abundance of 1.8%; **18%:** Species with nominal abundance of 18%; **Full:** results using full set of reads; **Reduced:** results using a subsample of 1,000,000 reads.

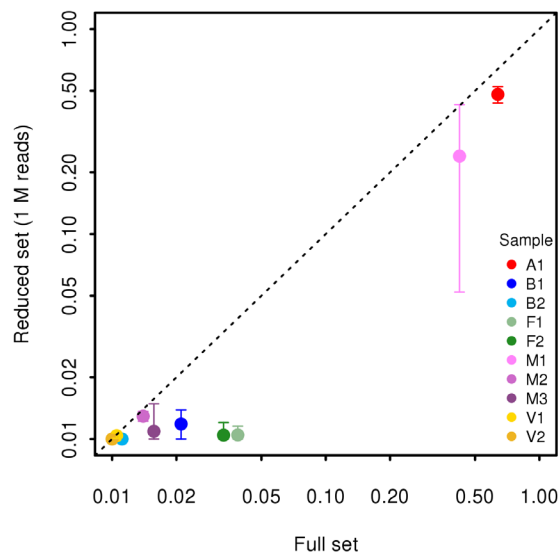


Figure 8. Completeness of the BUSCO genes in the full dataset (X axis) and in the largest of the reduced datasets (consisting of 1,000,000 reads, Y axis); error bars are based on the five replicate experiments performed for each sample. The plot is in log-log scale.

primers, extraction method and sequencing platform⁴⁰. We tested the effect of decrease in sequencing depth on deviations from expected frequency (Figure 3) and observed that even when sampling 10,000 reads the average correlation between expected and observed abundances remained high ($r=0.94$), although the variance among resampling experiments was high.

Horse fecal samples F1 and F2 and the food sample M2 are characterized by a large number of observed species (127, 126 and 138, respectively), while all the other samples have lower number of species, ranging from 4 in B1 to 84 in V1. The greater diversity of F1, F2 and M2 compared to others is confirmed by Shannon's and Pielou's indices, although Pielou's index also

assigns high variability to the mock community A1. The effect of sequencing depth on nearly all indices is moderate, although researchers should be aware that very complex samples (such as F1, F2 and M2 in our study) require high sequencing depth (1 million reads) to ensure that all indices are correctly estimated, since we observed the reduction in coverage could result in under- or over-estimation of the number of taxa and of Shannon's diversity index (Figures 5A and 5B).

We then set out to assess the changes in the estimated relative frequency of each individual species when reducing the number of sequenced reads. Accurate estimate of the relative abundance of each species is an important task when the aim is a) to detect species with a relative abundance above any given threshold, b) to differentiate two samples based on different abundance of any given species composition, or c) to cluster samples based on their species composition.

Our results show that species abundances can be reliably estimated for most samples even in case of substantial reduction of sequencing depth. However, researchers should be aware that for complex samples (horse fecal samples F1 and F2, in our study), extreme reduction in coverage might result in biases in the estimation of species abundances (Figure 4).

Finally, we assessed the effect of a reduction in the sequencing coverage on the ability of reconstructing *de novo* the metagenome. Our results suggest that 1 million reads are clearly suboptimal for *de novo* assembly for all the tested samples. Assembly size obtained subsampling 1 million reads are significantly smaller than those obtained with the full depth in all samples, included M1, M2 and M3, for which the full sequencing depth was less than 2 million reads (Figure 5D).

Additional analysis were performed to assess the effect of down-sampling on the completeness of the *de novo* assembly. First, we used BUSCO to assess the completeness of assemblies of the species used in the mock community A1 sample, and to compare

the performance in the full set and in the larger reduced set (1 million reads). No BUSCO genes were reconstructed for species with frequencies of 0.02% and 0.18%, and we show results only for the 10 species with frequency of 1.8% or greater. The full sequencing depth (~5 million reads) enabled the reconstruction of the majority of BUSCO genes in all the species, ranging 59% (*Staphylococcus aureus*) to 99% (*Bacillus cereus* and most of the species with 18% frequency). The ability of reconstructing BUSCO genes in assemblies obtained with 1 million reads was unchanged for the species with 18% abundance, while it dramatically decreased for species at frequency 1.8% (Figure 7).

We then performed a similar analysis on all the samples. Our results show that downsampling had a strongly negative effect on the proportion of reconstructed genes in all the study samples (Figure 8).

Our results clearly indicate that the proportion of genes reconstructed with BUSCO in the full dataset is very low for all samples, with the exception of the two samples M1, predominantly composed by one fungal species, and A1, composed by a limited number of small genomes, some of which with uniform and high abundance. In addition, detailed analysis of BUSCO performance in sample A1 revealed that only the genomes of the most frequent species could be reconstructed (Figure 7), even at full sequencing depth, amounting to nearly 5 million reads. Reduction of sequencing depth resulted in significant reduction of performance in all samples, as shown by the fact that the point estimates of the proportion of reconstructed genes and their confidence limits are below the diagonal in Figure 8. These results indicate that a complete reconstruction of the metagenome of a complex matrix requires at least several million reads. Our conclusions are also important for research aimed at the reconstruction of an interesting part of the meta-genome, such as genes involved in antibiotic resistance⁴¹. The decrease in performance observed in the genes' reconstruction will be likely observed for any gene category. Researchers aiming at a *de novo* reconstruction of the metagenome (although partial) must keep in mind that several millions of reads are needed to attain reliable results.

Researchers should be cautious when the fraction of reads that can be used to classify the microbial community is low. This might happen if the sample includes a substantial proportion of poorly characterized organisms, i.e. organisms not present in current databases, or if the samples come from biopsy or blood, thus containing a large proportion of the host tissue. In both cases, the amount of reads that can be used for the classification is already much lower than the number of produced reads, and further reduction is discouraged.

In the present work we tested the feasibility of using metagenome shotgun shallow high-throughput sequencing to analyze complex samples for the presence of eukaryotes, prokaryotes and virus nucleic acids for monitoring, surveillance, quality

control and traceability purposes. We show that, if the aim of the experiment is a taxonomical characterization of the sample or the identification and quantification of species, a low-coverage shotgun high-throughput sequencing is a good choice, provided that at least 500,000 reads are sequenced. On the other hand, if one of the aims of the study relies on *de novo* assembly, substantial sequencing efforts are required. The number of reads required for the reconstruction of the meta-genome, depends on several factors such as the number of species in the sample, their genome size and abundance and length of the sequencing reads. An estimation needs to be performed for each experiment based on specific goals and sample characteristics.

Data availability

Underlying data

Raw reads generated in the present study are available at NCBI Sequence Read Archive.

Sample A1 is available under accession number [SRP174028](https://identifiers.org/insdc.sra/SRP174028): <https://identifiers.org/insdc.sra/SRP174028>.

Samples F1 and F2 are available under accession number [SRP163102](https://identifiers.org/insdc.sra/SRP163102): <https://identifiers.org/insdc.sra/SRP163102>.

Samples B1 and B2 are available under accession number [SRP163096](https://identifiers.org/insdc.sra/SRP163096): <https://identifiers.org/insdc.sra/SRP163096>;

and samples M1, M2 and M3 are available under accession number [SRP163007](https://identifiers.org/insdc.sra/SRP163007): <https://identifiers.org/insdc.sra/SRP163007>.

Extended data

Open Science Framework: Do you cov me. <https://doi.org/10.17605/OSF.IO/Y7C39>⁴².

This project contains the raw html graphs, produced using Krona.

Software availability

Pipeline for performing the standard analysis included in this work available from: <https://github.com/fabiomarroni/doyoucovme>.

Archived code at time of publication: <https://doi.org/10.5281/zenodo.2593798>²⁷.

License: [GNU GPL-3.0](https://www.gnu.org/licenses/gpl-3.0.html).

Acknowledgments

The authors would like to thank Dr Loretta Bolgan for fruitful scientific discussions and Corvelva (non-profit association, Veneto, Italy) to give us the permission to use their own metagenome sequencing data (samples B1 and B2) for the paper purposes; Dr Federica Cattapan (Mérieux NutriSciences Italia and Chelab S.r.l., Italia) to provide the DNAs of M1, M2, M3 samples and Dr Carol Hughes (Phytorigins Ltd., United Kingdom) to give us the biological samples F1, F2 and to both of them to give us the permission to use their samples for whole metagenome sequencing and analysis.

References

- Quince C, Walker AW, Simpson JT, *et al.*: **Shotgun metagenomics, from sampling to analysis.** *Nat Biotechnol.* 2017; **35**(9): 833–44.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Forbes JD, Knox NC, Ronholm J, *et al.*: **Metagenomics: The Next Culture-Independent Game Changer.** *Front Microbiol.* 2017; **8**: 1069.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bragg L, Tyson GW: **Metagenomics using next-generation sequencing.** *Methods Mol Biol.* 2014; **1096**: 183–201.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Desai N, Antonopoulos D, Gilbert JA, *et al.*: **From genomics to metagenomics.** *Curr Opin Biotechnol.* 2012; **23**(1): 72–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sunagawa S, Coelho LP, Chaffron S, *et al.*: **Ocean plankton. Structure and function of the global ocean microbiome.** *Science.* American Association for the Advancement of Science; 2015; **348**(6237): 1261359.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wilhelm RC, Cardenas E, Leung H, *et al.*: **A metagenomic survey of forest soil microbial communities more than a decade after timber harvesting.** *Sci data.* Nature Publishing Group; 2017; **4**: 170092.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Qin J, Li R, Raes J, *et al.*: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature.* Nature Publishing Group; 2010; **464**(7285): 59–65.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hamady M, Knight R: **Microbial community profiling for human microbiome projects: Tools, techniques, and challenges.** *Genome Res.* 2009; **19**(7): 1141–52.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Human Microbiome Project Consortium: **Structure, function and diversity of the healthy human microbiome.** *Nature.* Nature Publishing Group; 2012; **486**(7402): 207–14.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oh J, Byrd AL, Deming C, *et al.*: **Biogeography and individuality shape function in the human skin metagenome.** *Nature.* Nature Publishing Group; 2014; **514**(7520): 59–64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wilson MR, Naccache SN, Samayoa E, *et al.*: **Actionable diagnosis of neuroleptospirosis by next-generation sequencing.** *N Engl J Med.* Massachusetts Medical Society; 2014; **370**(25): 2408–17.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wilson MR, Suan D, Duggins A, *et al.*: **A novel cause of chronic viral meningoencephalitis: Cache Valley virus.** *Ann Neurol.* 2017; **82**(1): 105–14.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Greninger AL, Messacar K, Dunnebacke T, *et al.*: **Clinical metagenomic identification of Balamuthia mandrillaris encephalitis and assembly of the draft genome: the continuing case for reference genome sequencing.** *Genome Med.* 2015; **7**(1): 113.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Forbes JD, Knox NC, Peterson CL, *et al.*: **Highlighting Clinical Metagenomics for Enhanced Diagnostic Decision-making: A Step Towards Wider Implementation.** *Comput Struct Biotechnol J.* Elsevier; 2018; **16**: 108–20.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mayo B, Rachid CT, Alegria A, *et al.*: **Impact of next generation sequencing techniques in food microbiology.** *Curr Genomics.* 2014; **15**(4): 293–309.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oniciuc EA, Likotrafiti E, Alvarez-Molina A, *et al.*: **The Present and Future of Whole Genome Sequencing (WGS) and Whole Metagenome Sequencing (WMS) for Surveillance of Antimicrobial Resistant Microorganisms and Antimicrobial Resistance Genes across the Food Chain.** *Genes (Basel).* 2018; **9**(5): pii: E268.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Caporaso JG, Lauber CL, Walters WA, *et al.*: **Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.** *Proc Natl Acad Sci U S A.* 2011; **108** Suppl 1: 4516–22.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Schoch CL, Seifert KA, Huhndorf S, *et al.*: **Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.** *Proc Natl Acad Sci U S A.* National Academy of Sciences; 2012; **109**(16): 6241–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hugert LW, Muller EE, Hu YO, *et al.*: **Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia.** Voolstra CR, editor. *PLoS One.* Public Library of Science; 2014; **9**(4): e95567.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hebert PD, Cywinska A, Ball SL, *et al.*: **Biological identifications through DNA barcodes.** *Proc Biol Sci.* 2003; **270**(1512): 313–21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Fazekas AJ, Kuzmina ML, Newmaster SG, *et al.*: **DNA barcoding methods for land plants.** *Methods Mol Biol.* 2012; **858**: 223–52.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Uyaguari-Diaz MI, Chan M, Chaban BL, *et al.*: **A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples.** *Microbiome.* BioMed Central; 2016; **4**(1): 20.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Brooks JP, Edwards DJ, Harwich MD Jr, *et al.*: **The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies.** *BMC Microbiol.* BioMed Central; 2015; **15**(1): 66.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ranjan R, Rani A, Metwally A, *et al.*: **Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing.** *Biochem Biophys Res Commun.* NIH Public Access; 2016; **469**(4): 967–77.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Eloe-Fadrosh EA, Ivanova NN, Woyke T, *et al.*: **Metagenomics uncovers gaps in amplicon-based detection of microbial diversity.** *Nat Microbiol.* Nature Publishing Group; 2016; **1**(4): 15032.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Siqueira JD, Dominguez-Bello MG, Contreras M, *et al.*: **Complex virome in feces from Amerindian children in isolated Amazonian villages.** *Nat Commun.* Nature Publishing Group; 2018; **9**(1): 4270.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet J.* 2011; **17**(1): 10–2.
[Publisher Full Text](#)
- Del Fabbro C, Scalabrini S, Morgante M, *et al.*: **An extensive evaluation of read trimming effects on Illumina NGS data analysis.** Seo JS, editor. *PLoS One.* Public Library of Science; 2013; **8**(12): e85024.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biol.* BioMed Central; 2014; **15**(3): R46.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ondov BD, Bergman NH, Phillippy AM: **Interactive metagenomic visualization in a Web browser.** *BMC Bioinformatics.* 2011; **12**(1): 385.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lu J, Breitwieser FP, Thielen P, *et al.*: **Bracken: estimating species abundance in metagenomics data.** *PeerJ Comput Sci.* PeerJ Inc.; 2017; **3**: e104.
[Publisher Full Text](#)
- Marroni F, Scaglione D, Pinosio S, *et al.*: **Reduction of heterozygosity (ROH) as a method to detect mosaic structural variation.** *Plant Biotechnol J.* 2017; **15**(7): 791–793.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shannon CE: **A Mathematical Theory of Communication.** *Bell Syst Tech J.* 1948; **27**(3): 379–423.
[Publisher Full Text](#)
- Pielou EC: **The measurement of diversity in different types of biological collections.** *J Theor Biol.* Academic Press; 1966; **13**: 131–44.
[Publisher Full Text](#)
- Oksanen J, Blanchet G, Friendly M, *et al.*: **vegan: Community Ecology Package.** 2017.
[Reference Source](#)
- R Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2018.
[Reference Source](#)
- Li D, Liu CM, Luo R, *et al.*: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.** *Bioinformatics.* 2015; **31**(10): 1674–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* Oxford University Press; 2015; **31**(19): 3210–2.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Zuo G, Xu Z, Hao B: **Shigella strains are not clones of Escherichia coli but sister species in the genus Escherichia.** *Genomics Proteomics Bioinformatics.* Elsevier; 2013; **11**(1): 61–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Fouhy F, Clooney AG, Stanton C, *et al.*: **16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform.** *BMC Microbiol.* BioMed Central; 2016; **16**(1): 123.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Adu-Oppong B, Gasparrini AJ, Dantas G: **Genomic and functional techniques to mine the microbiome for novel antimicrobials and antimicrobial resistance genes.** *Ann N Y Acad Sci.* 2017; **1388**(1): 42–58.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Marroni F: **Do you cov me.** 2019.
<http://www.doi.org/10.17605/OSF.IO/Y7C39>

Open Peer Review

Current Peer Review Status:    

Version 4

Reviewer Report 04 March 2020

<https://doi.org/10.5256/f1000research.24347.r58937>

© 2020 Dal Grande F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Francesco Dal Grande 

¹ Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

² LOEWE Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

After reading the authors' response, I agree that the authors have adequately addressed all the referees' comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: metagenomics, metatranscriptomics, community ecology, symbiosis, population genomics, metabarcoding, biotic interactions

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 05 February 2020

<https://doi.org/10.5256/f1000research.24347.r58938>

© 2020 Claesson M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Marcus Claesson 

APC Microbiome Ireland, University College Cork, Cork, Ireland

Shriram Patel

APC Microbiome Ireland, University College Cork, Cork, Ireland

We have now reviewed the authors' response and agree that they have sufficiently addressed our comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbiome in human disease; bioinformatics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 3

Reviewer Report 02 December 2019

<https://doi.org/10.5256/f1000research.22041.r55903>

© 2019 Claesson M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Marcus Claesson

APC Microbiome Ireland, University College Cork, Cork, Ireland

Shriram Patel

APC Microbiome Ireland, University College Cork, Cork, Ireland

This is an interesting piece of work looking to evaluate influence of varying sequencing coverage (depth) on the ability to harness information about taxonomic composition and species diversity and possible use of shallow whole metagenome shotgun sequencing as a potential cost-effective alternative to targeted 16S rRNA gene sequencing in large scale studies.

In general, the manuscript is very well written and make use of staggered mock community along with the actual samples sequenced from diverse environmental origin to optimize required metagenomic sequencing depth to address potential research question.

The authors have used statistics such as alpha diversity, species abundance and completeness of reconstructed genomes to evaluate performance of reduce sequencing efforts. It would be interesting, although not required, to see how overall between sample beta diversity (bray-curtis) changes with varying sequencing depth and in full datasets (considering only actual samples). This could offer insights into whether samples coming from diverse environment clusters together even at varying sequencing depth (as low as 10K)? or does reduce sequencing depth influences overall metagenome composition. Particularly, it would be interesting to see Procrustes analysis between full datasets and reduced datasets (may be at 100K/ 500K reads because estimated alpha diversity reached plateau and most of the species gets covered).

I am confused with statement on page 8. "Intermediate level of down sampling (here 100K reads) caused an increase in observed species, due to increased number of species exceeding the 0.1% abundance cut-off (selected based on mock community)". Does this indicate that with increased sequencing effort

(particularly in horse fecal samples) those species exceeding the cut-off at reduced sequencing depth did not detected?

It would be good if authors can add important limitation of shallow shotgun metagenomic sequencing in discussion. Particularly note on “poorly characterized samples” for which no representative genomes are available in database or “samples coming from biopsy or blood” where host DNA accounts for most of the extracted DNA.

Some General comments:

1. All samples were trimmed to the read length of 125bp. Did authors build bracken database with default read length of 100 or 125? If so, please mention that in the manuscript.
2. Please move formula of alpha diversity indexes in method's section.
3. In the abstract, 'diversity' should be prepended with 'alpha' as it might otherwise include beta-diversity which wasn't analysed.
4. The title is quite long (just a matter of taste)
5. 5th sentence in Intro: technically, fungi are also eukaryotic, so this needs to be reflected.
6. The colour scheme in Figure 2 could be improved. Currently the phyla are ordered alphabetically which is a wasted opportunity for more information. At the least, they should be ordered by kingdom. Unknown could be black/grey/white
7. Correlations for Fig 4 are Pearson, which only should be used if the data follows a normal distribution, otherwise Spearman.
8. Insert “,” for each 1,000 in N. of reads to improve readability
9. All fonts in Figure 5 are too small and unreadable

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbiome in human disease; bioinformatics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 15 Jan 2020

Federica Cattonaro, IGA Technology Services Srl, Udine, Italy

Reviewer 4

Marcus Claesson, APC Microbiome Ireland, University College Cork, Cork, Ireland

Shriram Patel, APC Microbiome Ireland, University College Cork, Cork, Ireland

This is an interesting piece of work looking to evaluate influence of varying sequencing coverage (depth) on the ability to harness information about taxonomic composition and species diversity and possible use of shallow whole metagenome shotgun sequencing as a potential cost-effective alternative to targeted 16S rRNA gene sequencing in large scale studies.

In general, the manuscript is very well written and make use of staggered mock community along with the actual samples sequenced from diverse environmental origin to optimize required metagenomic sequencing depth to address potential research question.

The authors have used statistics such as alpha diversity, species abundance and completeness of reconstructed genomes to evaluate performance of reduce sequencing efforts. It would be interesting, although not required, to see how overall between sample beta diversity (bray-curtis) changes with varying sequencing depth and in full datasets (considering only actual samples). This could offer insights into whether samples coming from diverse environment clusters together even at varying sequencing depth (as low as 10K)? or does reduce sequencing depth influences overall metagenome composition. Particularly, it would be interesting to see Procrustes analysis between full datasets and reduced datasets (may be at 100K/ 500K reads because estimated alpha diversity reached plateau and most of the species gets covered).

We now performed Procrustes analysis between the full dataset and all the reduced sets (we then removed the 10K dataset, because diagnostic measures showed that the MDS on that matrix was not reliable). The analysis is now shown as Figure 6, described, and discussed.

I am confused with statement on page 8. "Intermediate level of down sampling (here 100K reads) caused an increase in observed species, due to increased number of species exceeding the 0.1% abundance cut-off (selected based on mock community)". Does this indicate that with increased sequencing effort (particularly in horse fecal samples) those species exceeding the cut-off at reduced sequencing depth did not detected?

Yes.

The species exceeding the cut-off at reduced sequencing depth were still “detected” at full sequencing depth, but they didn’t exceed the threshold. For example, in the fecal sample 1 (F1), in the full-depth sample, we assigned reads to 6273 species (with an average frequency of 0.02%), but only 124 of them exceeded the threshold; in the 100000 sample we assigned reads to 350 species (with an average frequency of 0.6%), 215 of which exceeded the threshold.

This phenomenon was observed only for the fecal samples, which are the ones with greater complexity and higher number of reads in the full sample. We rewrote part of the results to try to clearly convey our take-home message, i.e.: although reduction in coverage depth usually does not affect estimation of sample diversity, it can in some cases result in an under- or over-estimation of such quantities.

It would be good if authors can add important limitation of shallow shotgun metagenomic sequencing in discussion. Particularly note on “poorly characterized samples” for which no representative genomes are available in database or “samples coming from biopsy or blood” where host DNA accounts for most of the extracted DNA.

We added the following sentence in the discussion: Researchers should be cautious when the fraction of reads that can be used to classify the microbial community is low. This might happen if the sample includes a substantial proportion of poorly characterized organisms, i.e. organisms not present in current databases, or if the samples come from biopsy or blood, thus containing a large proportion of the host tissue. In both cases, the amount of reads that can be used for the classification is already much lower than the number of produced reads, and further reduction is discouraged.

Some General comments:

1. All samples were trimmed to the read length of 125bp. Did authors build bracken database with default read length of 100 or 125? If so, please mention that in the manuscript.

We built a bracken database for 125 kmers. On request of Reviewer 3 we also performed tests on different databases, only for the mock community. One of the additional databases (minikraken) comes as a prebuilt database without possibility of building the bracken index, and we used distributed databases built with 100kmers and 150kmers. We added this information in the methods section.

1. Please move formula of alpha diversity indexes in method’s section.

Done

1. In the abstract, ‘diversity’ should be prepended with ‘alpha’ as it might otherwise include beta-diversity which wasn’t analysed.

We left this unchanged, since we are now also analyzing beta-diversity, and the generic statement of the abstract is true for beta diversity as well.

1. The title is quite long (just a matter of taste)

We changed the title to: Do you cov me? Effect of coverage reduction on metagenome shotgun sequencing studies

1. 5th sentence in Intro: technically, fungi are also eukaryotic, so this needs to be reflected.

We removed the word fungi

1. The colour scheme in Figure 2 could be improved. Currently the phyla are ordered alphabetically which is a wasted opportunity for more information. At the least, they should be ordered by kingdom. Unknown could be black/grey/white

We changed the colour scheme for Figure 2. Protozoan (only apicomplexan detected) are red-violet, bacteria are in shades of brown, fungi are in shades of olive green, vertebrates are in shades of blue, plants in shade of green, unknown are grey, and viruses are violet

1. Correlations for Fig 4 are Pearson, which only should be used if the data follows a normal distribution, otherwise Spearman.

We computed correlations for data of figure 4 as Spearman. We now also present correlation of Figure 3 as spearman's rho, for the same reason (none of the two data followed a normal distribution).

1. Insert “,” for each 1,000 in N. of reads to improve readability

Done

1. All fonts in Figure 5 are too small and unreadable

We increased the font size

Competing Interests: No competing interests were disclosed.

Reviewer Report 08 August 2019

<https://doi.org/10.5256/f1000research.22041.r51765>

© 2019 Dal Grande F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Francesco Dal Grande

¹ Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

² LOEWE Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

I appreciate the changes made in this revision. Specifically, I am glad to see that the authors used the results for the mock community to set the parameters for detecting the presence of species in the other samples. Improved is also the inference of the relative abundances of species using bracken and the presentation of the BUSCO results.

I have only minor suggestions that I hope will help further improving the manuscript.

My only issue is the detection of the false positive (*Shigella flexneri*) for the mock community data set. I agree with the authors that this might likely be the result of misclassification of a small portion of reads. This, however, may also be the result of incorrect taxonomic profiles that may be present in the chosen (full NCBI nt) database. The evaluation of the effects of database taxonomic correctness and composition

on species assignment accuracy is clearly not the scope of the present work. However, since the correct profiling of the mock community is crucial for selecting the best detection threshold for all other data sets, I suggest to strengthen the analysis of the mock community by comparing kracken/bracken results using different databases (only for the mock community): **full NCBI nt** vs. **full bacterial RefSeq** vs. **curated genome database** (i.e. including only the 20 genomes of the species forming the mock community).

Minor points:

- In the abstract, add a line to describe the use of the mock community in your study.
- Figure 1: I would modify the box 'Classify reads (kraken2)' into 'Classify reads and estimate species abundances (kraken2 + bracken)'.
- p. 8: "The effect of the number of reads on Pielou's index is moderate". Please define 'moderate'.
- p. 10: Please move the formulas of the two indices to the Materials and Methods section.

Other corrections:

- p. 6: M1 was mostly composed OF.
- p. 11: "...the performance in the full set and IN".
- p. 12: "..., depends on several factors such as THE number of species ..".

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: metagenomics, metatranscriptomics, community ecology, symbiosis, population genomics, metabarcoding, biotic interactions

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 Jan 2020

Federica Cattonaro, IGA Technology Services Srl, Udine, Italy

Reviewer 3

Francesco Dal Grande, Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany; LOEWE Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

I appreciate the changes made in this revision. Specifically, I am glad to see that the authors used the results for the mock community to set the parameters for detecting the presence of species in the other samples. Improved is also the inference of the relative abundances of species using bracken and the presentation of the BUSCO results.

I have only minor suggestions that I hope will help further improving the manuscript.

My only issue is the detection of the false positive (*Shigella flexneri*) for the mock community data set. I agree with the authors that this might likely be the result of misclassification of a small portion of reads. This, however, may also be the result of incorrect taxonomic profiles that may be present in the chosen (full NCBI nt) database. The evaluation of the effects of database taxonomic correctness and composition on species assignment accuracy is clearly not the scope of the present work. However, since the correct profiling of the mock community is crucial for selecting the best detection threshold for all other data sets, I suggest to strengthen the analysis of the mock community by comparing kracken/bracken results using different databases (only for the mock community): **full NCBI nt** vs. **full bacterial RefSeq** vs. **curated genome database** (i.e. including only the 20 genomes of the species forming the mock community).

This is a very good point. We were already aware that the choice of the database would affect the accuracy of the results, and the choice to use nt database was motivated by the fact that when studying heterogeneous samples potentially including Eukaryotes the nt would be the database of choice. We avoided by purpose to tackle the aspect of accuracy of databases taxonomic correctness. However, we agree that a simple comparison based on the mock community data would benefit the manuscript and the readers. Thus we tested the following additional databases 1) the “standard” database distributed with kraken2 which is a full bacterial+viral+fungi RefSeq database with the addition of the human genome, and 2) Several “minikraken2” databases that are distributed with kraken2 (the details on the composition of the minikraken2 are provided in the manuscript). We didn’t use the curated database only including the 20 genomes of the species forming the mock community because in that case by definition we will not identify any false positive; even in the case of a real contamination of the mock community all the classified reads would be attributed to one of the 20 genomes, because those are the only genomes present in the database.

Our results show a general good agreement across databases, but some differences were observed. This is especially true for the false positives; each database returns different false positives. It is possible that different databases have – minor – different classification issues. This however should motivate researchers to cautiously interpret results, especially before claiming contaminations from unexpected species in a given

sample. This results are now shown in a Table and discussed.

Minor points:

- In the abstract, add a line to describe the use of the mock community in your study.
- **Done**
- Figure 1: I would modify the box 'Classify reads (kraken2)' into 'Classify reads and estimate species abundances (kraken2 + bracken)'.
- **Done**
- p. 8: "The effect of the number of reads on Pielou's index is moderate". Please define 'moderate'.
- **Very good point. We changed moderate to negligible; indeed Pielou's index is the most stable across sequencing depths.**
- p. 10: Please move the formulas of the two indices to the Materials and Methods section.
- **Done**

Other corrections:

- p. 6: M1 was mostly composed OF.
- **Done**
- p. 11: "...the performance in the full set and IN".
- **Done**
- p. 12: "... depends on several factors such as THE number of species ..".
- **Done**

Competing Interests: No competing interests were disclosed.

Version 2

Reviewer Report 30 May 2019

<https://doi.org/10.5256/f1000research.20298.r48341>

© 2019 Dal Grande F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Francesco Dal Grande 

¹ Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

² LOEWE Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

In this manuscript the authors aimed at evaluating the use of shallow shotgun metagenomic sequencing for the characterisation of species diversity and the reconstruction of genomes in complex Illumina read sets. Overall, the manuscript is well written and contains interesting information that may be useful to others in figuring out a required metagenomic sequencing depth for a given goal.

The manuscript has been vastly improved in the current version, however I feel that it still needs a thorough revision to address a few major issues in order to ensure the general validity of the findings.

The three major issues to address are, in my opinion, the following:

1. **Overestimation of diversity:** Authors decided to base their analyses of diversity on the raw output from kraken2. However, as mentioned by the authors themselves, "species represented by only one read are unlikely to be real". This is quite evident in the report from the 20-species mock community comprising instead >2000 species. I strongly recommend the use of a threshold (e.g., 0.005% of the total amount of reads) to filter out likely false positives. For this purpose, the authors could take advantage of the mock community to evaluate results based on different thresholds and thereby optimise threshold selection.
2. **Inaccuracy of species-level abundances:** in their analysis the authors assumed that read abundances reflect species abundance. However, this is often not the case, especially when closely related taxa are present in the sample; the accuracy of abundance estimation further depends on the database used (Lu *et al* 2017). The authors themselves hint at this when discussing the misclassification of *Staphylococcus lugdunensis*, likely due to the presence of other confounding *Staphylococcus* reads. To address this issue, the authors could use Bracken (from the same developers of kraken, Lu *et al.* 2017). Bracken uses the classification results of kraken to reestimate relative species abundances taking into account how much sequence from each species is identical to other genomes in the database.
3. **Inaccurate assessment of genome reconstruction ability:** considering the classification biases mentioned above and the complexity of the investigated metagenomic data sets, it might be better to base the assessment of the effects of coverage reduction on metagenome reconstruction solely on the mock community data. First, authors would need to bin the metagenomic contigs into individual species (using kraken2 and/or other binning approaches). The individual bins (i.e., species) should then be evaluated for completeness using BUSCO and compared.

In summary, this work (and, by extension, future studies using a similar approach) could greatly benefit from the inclusion of a baseline estimate for species diversity and metagenome reconstruction, even if it is derived from a single mock community. The additional data sets could then be used to validate these estimates against real data.

References

1. Lu J, Breitwieser F, Thielen P, Salzberg S: Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*. 2017; **3**. [Publisher Full Text](#)

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: metagenomics, metatranscriptomics, community ecology, symbiosis, population genomics, metabarcoding, biotic interactions

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 23 Jul 2019

Federica Cattonaro, IGA Technology Services Srl, Udine, Italy

In this manuscript the authors aimed at evaluating the use of shallow shotgun metagenomic sequencing for the characterisation of species diversity and the reconstruction of genomes in complex Illumina read sets. Overall, the manuscript is well written and contains interesting information that may be useful to others in figuring out a required metagenomic sequencing depth for a given goal.

The manuscript has been vastly improved in the current version, however I feel that it still needs a thorough revision to address a few major issues in order to ensure the general validity of the findings.

We thank the reviewer for the suggestions. We implemented them and updated the manuscript accordingly.

The three major issues to address are, in my opinion, the following:

1. **Overestimation of diversity:** Authors decided to base their analyses of diversity on the raw output from kraken2. However, as mentioned by the authors themselves, "species represented by only one read are unlikely to be real". This is quite evident in the report from the 20-species mock community comprising instead >2000 species. I strongly recommend the use of a threshold (e.g., 0.005% of the total amount of reads) to filter out likely false positives. For this purpose, the authors could take advantage of the mock community to evaluate results based on different thresholds and thereby optimise threshold selection.

See answer to point 2.

2. ***Inaccuracy of species-level abundances: in their analysis the authors assumed that read abundances reflect species abundance. However, this is often not the case, especially when closely related taxa are present in the sample; the accuracy of abundance estimation further depends on the database used (Lu et al 2017). The authors themselves hint at this when discussing the misclassification of Staphylococcus lugdunensis, likely due to the presence of other confounding Staphylococcus reads. To address this issue, the authors could use Bracken (from the same developers of kraken, Lu et al. 2017). Bracken uses the classification results of kraken to reestimate relative species abundances taking into account how much sequence from each species is identical to other genomes in the database.***

We took advantage of suggestions 1 and 2 (and from suggestions from reviewer 1) to improve the species abundances estimation. After classifying reads with kraken2, we used bracken to re-estimate species abundance only for species represented by at least 10 reads. Then, using the only gold standard we had (the mock community) we measured performance at difference detection threshold. Our results suggested that a detection threshold of 0.1% was the one resulting in the higher F1 score, minimizing false negatives and false positives while maximizing true positives.

3. ***Inaccurate assessment of genome reconstruction ability: considering the classification biases mentioned above and the complexity of the investigated metagenomic data sets, it might be better to base the assessment of the effects of coverage reduction on metagenome reconstruction solely on the mock community data. First, authors would need to bin the metagenomic contigs into individual species (using kraken2 and/or other binning approaches). The individual bins (i.e., species) should then be evaluated for completeness using BUSCO and compared.***

Results presented in version 2 of our paper are already based on binning approaches, in which we classified contigs using kraken, performed BUSCO for each species and then averaged the proportion of BUSCO genes across species. However, in version 2 we made (in our opinion) a mistake, since we averaged the proportion of BUSCO genes across all species for which at least one BUSCO gene was reconstructed. This led to a slight overestimation of the number of reconstructed BUSCO genes. We thus repeated the analysis by averaging the proportion of BUSCO genes over all the species that were above the detection threshold, including those for which no BUSCO gene was reconstructed. The new approach is now explained in the methods section, and the new plot is now Figure 7. In addition, we liked the idea of using the mock community, and we performed a new analysis, now shown in Figure 6. The results are very interesting and are briefly discussed. Basically, with the full set of reads (around 5M), the majority of BUSCO genes could be reconstructed for species with a nominal abundance of 18% and 1.8%, but not for the rarer species (for which basically no gene could be reconstructed). When only 1M reads are used for the assembly, the proportion of reconstructed BUSCO genes is nearly unchanged in abundant species and drops to less than 10% in species with a nominal frequency of 1.8%. The results and the implications for study designs are briefly discussed in the paper.

Competing Interests: No competing interests were disclosed.

<https://doi.org/10.5256/f1000research.20298.r46099>

© 2019 Cobo Diaz J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



José F. Cobo Diaz 

Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, IBSAM, ESIAB, Université de Brest, Plouzané, France

I appreciate the changes made along the introduction, because the objective of the present study is now more clear. Although the manuscript was improved considerably, there is still a big problem with the data analysis, mainly in reads filtering.

Now that you have included a mock community sample, you need to use this sample to adapt the parameters of reads filtering, clustering step (I assume you have done some kind of clustering since you talk about singletons) and taxonomic assignment until you have the number of species expected, 20 in this case. You can also have some less due to problems with species assignment, but it is crazy to use a 20 species mock community and say that you have 2571 species in this sample. For example, singletons (clustering groups or OTUs (Operational Taxonomical Units) with a unique sequence) are usually removed on metabarcoding pipelines, and in some cases OTUs with less than 0.1% of abundance are removed, assuming that these sequences are sequencing errors (and PCR errors in metabarcoding). Therefore, you have to estimate the minimum percentage of abundance to be considered real (and not due to errors) with the mock sample and apply this cut off value to the rest of samples.

In the same line, to say that 2,507 and 4,597 species were found in vaccines is not correct, where you can expect the DNA from varicella (the other viruses are ssRNA) and the DNA from human and chicken cells used for culture.

Some small changes I suggest:

- Rewrite or suppress last paragraph of introduction, which looks more appropriate to Methodology.
- Add some disadvantages of use metabarcoding approach (being the main one the bias due to primers, with over/under-estimation of some taxa, depending of the primers used).
- At the end of the samples description, you need to put what means SRA (and add the corresponding web-address).
- In samples description, grammatical mistake with human faecal (have to be human fecal).
- Remove this sentence from results: To ensure that our conclusions have a general validity, we selected samples originating from very different sources with different compositions, and sequenced them at different depths.
- Figure 3, with species and genus level is enough.

Thus, the read filtering and hence all the statistical analysis have to be re-made. I do not expect big changes, also at taxonomical level (where only a reduction of "rare species" and unclassified sequences is expected), but it is not convenient to present the results with such great over-estimation of species richness.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: microbial ecology, metabarcoding sequencing, NGS data analysis, bacterial communities, fungal communities

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 23 Jul 2019

Federica Cattonaro, IGA Technology Services Srl, Udine, Italy

I appreciate the changes made along the introduction, because the objective of the present study is now more clear. Although the manuscript was improved considerably, there is still a big problem with the data analysis, mainly in reads filtering.

Now that you have included a mock community sample, you need to use this sample to adapt the parameters of reads filtering, clustering step (I assume you have done some kind of clustering since you talk about singletons) and taxonomic assignment until you have the number of species expected, 20 in this case. You can also have some less due to problems with species assignment, but it is crazy to use a 20 species mock community and say that you have 2571 species in this sample. For example, singletons (clustering groups or OTUs (Operational Taxonomical Units) with a unique sequence) are usually removed on metabarcoding pipelines, and in some cases OTUs with less than 0.1% of abundance are removed, assuming that these sequences are sequencing errors (and PCR errors in metabarcoding). Therefore, you have to estimate the minimum percentage of abundance to be considered real (and not due to errors) with the mock sample and apply this cut off value to the rest of samples.

In the same line, to say that 2,507 and 4,597 species were found in vaccines is not correct, where you can expect the DNA from varicella (the other viruses are ssRNA) and the DNA from human and chicken cells used for culture.

According to your suggestions (and to similar suggestions received from reviewer 3), we now adopted more stringent criteria for determining the presence of a species. Following the suggestion of both reviewers, we leverage the mock community to define a threshold. We use Bracken to refine the species abundance estimation (already providing a very permissive threshold, i.e. ignoring OTUs with less than 10 reads). We then performed a performance analysis to compare Bracken results with the known composition of the mock community, and chose the threshold maximizing the F1 score (harmonic average of precision and recall). The threshold resulting in the best tradeoff was 0.1%.

As a side effect of filtering OTUs with less than 0.1% frequency we do not have any narrow-sense singleton. As a consequence, the number of observed taxa and Chao1 diversity index coincide, and the Good estimator is always 1. We thus removed these two statistics from our panel plot. In addition, we removed the paragraph on the “detection threshold” and the corresponding Table 2, since we are now determining a threshold *a-priori* based on the mock community and this parts are not needed any more.

Some small changes I suggest:

- ***Rewrite or suppress last paragraph of introduction, which looks more appropriate to Methodology.***

We removed the last paragraph.

- ***Add some disadvantages of use metabarcoding approach (being the main one the bias due to primers, with over/under-estimation of some taxa, depending of the primers used).***

We added a sentence and a reference regarding limitation of metabarcoding approaches in the introduction.

- ***At the end of the samples description, you need to put what means SRA (and add the corresponding web-address).***

Done.

- ***In samples description, grammatical mistake with human faecal (have to be human fecal).***

Amended.

- ***Remove this sentence from results: To ensure that our conclusions have a general validity, we selected samples originating from very different sources with different compositions, and sequenced them at different depths.***

Sentence removed.

- ***Figure 3, with species and genus level is enough.***

While we were modifying the Figure as per reviewer's request we realized that indeed the results presented at the species level in Figure 3 are also presented in the first panel of Figure 4. Since the results at the genus species did not add much information, we decided to remove Figure 3.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 04 January 2019

<https://doi.org/10.5256/f1000research.18370.r42422>

© 2019 Cobo Diaz J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



José F. Cobo Diaz 

Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, IBSAM, ESIAB, Université de Brest, Plouzané, France

The authors proposed and evaluated the influence of reduce sequencing effort (amount of sequences) for a whole metagenome shotgun analysis, using the Illumina platform, in the species composition and diversity index of the communities studied. Although the idea and hypothesis are good, some problems were found in the experimental design and data analysis.

According to the questions proposed in the peer review form, it is not a new method, only the adaptation of a current methodology to optimize the cost and increase the potential numbers of samples analyzed per run of Illumina platform. Although the introduction is clearly explained, the reasons for use shotgun sequencing, mainly to analyze viruses data and functional data for all the organism, no emphasis on such points was done in the results and discussion. The samples used (vaccines, horse fecal samples and food samples) and the introduction remark the detection of pathogens as the main objective of the approach used, including viruses, which can not be screened by amplicons approaches, like metabarcoding sequencing. I suggest adapting the text and manuscript to focus on pathogens (mainly viruses) found along the sub-samples taken for each sample. At that point, some contaminated samples (or not contaminated samples mixed with known amounts DNA from pathogen viruses) have to be used to determine the lowest pathogen concentration that could be detected for each shotgun sequencing coverage proposed.

Many problems were found with the methodology employed, mainly the parameters used in each step and/or software employed for data filtering and analysis, which are critical for the results, which can have strong variations depending of the parameters used. Hence, the methodology proposed does not allow any replication of the method used. Moreover, there are some mistakes for species designation in the study, with at least 2508 species found in vaccine samples indicating big problems along read filtering and data analysis, because this number of species is often found in more complex systems, such as soils samples from agricultural fields. Moreover, go to species classification using some taxonomical markers, such ITS or 16SrRNA, is risky with sequences lower than 400 bp, and sometimes with bigger sequences. In the current manuscript, the use of non taxonomical marker sequences and 150 bp lengths increase enormously the number of sequences not correctly assigned to species level, and in several cases also for higher taxonomical levels (genus, family...). Therefore, I suggest to clarify how the species assignment was done, because it looks like that each gene-species was considered as one species, and each gene found for a single species was counted as a new species.

Alpha diversity indexes employed are not the best ones, in my opinion, to describe or compare the sub-samples proposed in this manuscript. The chao1 index, an estimator of richness, has a strong influence on the number of singletons obtained in the samples, which due to the complexity of the samples-data tends to be high. Shannon index is influenced by both richness (number of taxa) and evenness (equability, Pielou index), and the reduction of richness due to the loss of rare taxa has a strong influence on this index. I propose to use the number of observed taxa instead of estimated taxa, and any evenness index, like the Pielou index, instead of the Shannon index. Moreover, the use of a coverage index, such Good's coverage index, could be useful to compare the loss of information associated to

sampled size or coverage.

In conclusion, although the raw data can contains some important information, the manuscript has to be improved with new “pathogen contaminated” samples, and be re-written to focus on the detection of pathogens in the samples, which due to the low abundance of the samples could not be detected depending of the shotgun coverage.

Is the rationale for developing the new method (or application) clearly explained?

No

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: microbial ecology, metabarcoding sequencing, NGS data analysis, bacterial communities, fungal communities

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 27 November 2018

<https://doi.org/10.5256/f1000research.18370.r40445>

© 2018 Sanchez-Flores A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Alejandro Sanchez-Flores 

Institute of Biotechnology, National Autonomous University of Mexico (UNAM)), Cuernavaca, Mexico

The authors propose and evaluate a whole metagenome shotgun analysis via a low sequencing yield approach, using the Illumina platform.

In general, the idea and hypothesis are good, but the experimental design itself lacks important controls and there are many variables that are not analyzed and that can potentially bias the results.

My main concern is that the used samples have many variables and despite using a "replicate" for each case, samples within the same type were very different. Also the nature of each sample could have an effect in the DNA isolation, in particular for the vaccine ones. Also, regarding the vaccines, it is not clear to me, if what they are looking for is DNA of potential contaminants, since all viruses in the vaccine are ssRNA. That would be my guess, but is not clear from the text.

The main problem is that to test the influence of the sequencing yield, it would be extremely important to know the initial DNA concentration of each organism in the sample. Therefore, a mock metagenome or controlled sample would be much better as a reference to compare real life cases. In real life cases, the presence of certain organisms detected by the presence of its DNA, is not necessarily an indicator of the availability of alive organisms. Depending on the case, the presence of just the organism DNA could be an indicator of contamination which in the case of vaccines could be really bad. However, in the case of food material, finding DNA of pathogens, has to be associated with microbiology tests. However, with low sequencing yield, is very probable that very DNA in low amounts will be missed, even if this is not changing diversity indexes such as Chao1 and Shannon.

Finally, the main difference where low yield has a significant impact can be observed in the fecal samples. This is expected since among all the tested samples, fecal ones are the most diverse and sub-sampling will really affect them as observed in Figure 3.

Since the composition of each sample is not known *a priori*, then there are some factors that can contribute to biases. As mentioned, the DNA concentration but also its integrity (fragmentation) will affect the library construction; the cited kit requires DNA amplification which will have a bias towards GC rich genomic regions; library size was not described and was not mentioned if the samples were pooled with other libraries with different insert sizes, which affect not only the sequencing quality but the yield.

In terms of bioinformatics analysis, it will be required to put the parameters used for each program, in case someone wants to reproduce this. For Kraken2, it is important to know what is the kmer size to index the database. For MEGAHIT assembly it will be important to know the kmer and step sizes used. For the completeness assessment, the authors used BUSCO, but apparently they are using the whole assembly to assess the completeness. This is not correct, since they must first separate in bins which genomes they have really reconstructed and then they can assess the completeness of them. Probably they can report the an average completeness value for all the reconstructed genomes. By doing the binning they can have a better analysis of what was really reconstructed and how complete it was.

The use of Krona in Figure 2 is not very convenient. The whole point of a Krona graph is that is interactive. If authors want to provide the Krona data to be downloaded it would be possible and recommended. Having said that, I recommend to use bar plots to represent the relative abundance and composition of the samples at a given taxa level.

Again, the idea is very good but the work needs to be improved before indexing.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

No

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, Transcriptomics, Metagenomics, Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 30 Nov 2018

Federica Cattonaro, IGA Technology Services Srl, Udine, Italy

We are grateful for the constructive comments. We agree with all of them and we are planning corrective actions, listed below.

My main concern is that the used samples have many variables and despite using a "replicate" for each case, samples within the same type were very different.

The observation is correct. Actually, the diversity of the samples was sought by purpose in order to be able to generalize the conclusions of our paper. The fact that diversity estimate and species abundance estimation remain reliable even with strong down-sampling for all of the samples is encouraging us to think that this is a general (although not necessarily universal) observation. The same is true for the observation that de-novo assembly quickly loses accuracy when decreasing the number of sequenced reads. Maybe this wasn't made clear enough in the paper, and we will clarify it.

Also the nature of each sample could have an effect in the DNA isolation, in particular for the vaccine ones.

Quantities of DNA isolated from vaccine samples (B1 and B2) were estimated to be ~2 µg using Qbit fluorimeter. However, we will provide a table with all the details about quantity, concentration, quality and size of starting DNA for all samples used in the study.

Also, regarding the vaccines, it is not clear to me, if what they are looking for is DNA of

potential contaminants, since all viruses in the vaccine are ssRNA. That would be my guess, but is not clear from the text.

The vaccine composition declared by the producer is the following:

Live attenuated viruses: Measles (ssRNA) Swartz strain, cultured in embryo chicken cell cultures; Mumps (ssRNA) strain RIT 4385, derived from the Jeryl Linn strain, cultured in embryo chicken cell cultures; Rubella (ssRNA) Wistar RA 27/3 strain, grown in human diploid cells (MRC-5); Varicella (dsDNA) OKA strain grown in human diploid cells (MRC-5).

By DNA-seq we expected to find Varicella (dsDNA) OKA strain DNA (which was found and confirmed by variant analysis with respect to AB097932.1 Human herpesvirus 3 DNA, sub strain vOka). In addition, we found also human and chicken DNA. For human's, we confirmed MRC-5 cell origin by mitochondrial genome variant analysis.

Genotyping analyses gave us confidence on the validity of the obtained results, even though they were beyond the scope of this work.

To identify vaccine's ssRNA viruses we extracted RNA and performed RNA-seq from the same B1 and B2 samples. This aspect also goes beyond the scope of this work.

The main problem is that to test the influence of the sequencing yield, it would be extremely important to know the initial DNA concentration of each organism in the sample. Therefore, a mock metagenome or controlled sample would be much better as a reference to compare real life cases.

A mock community experiment is already on-going by using '10 Strain Staggered Mix Genomic Material (ATCC® MSA-1001™)'. Of course, the data obtained will be integrated in the analysis results.

In real life cases, the presence of certain organisms detected by the presence of its DNA, is not necessarily an indicator of the availability of alive organisms. Depending on the case, the presence of just the organism DNA could be an indicator of contamination which in the case of vaccines could be really bad. However, in the case of food material, finding DNA of pathogens, has to be associated with microbiology tests.

We agree with the observation of the reviewer. However, the aim of this work is to determine if low-pass whole genome sequencing can be an appropriate approach to broadly describe a complex matrix; finding and confirming contaminants in vaccines or DNA pathogens in food samples was beyond of the scope of the paper.

However, with low sequencing yield, is very probable that very DNA in low amounts will be missed, even if this is not changing diversity indexes such as Chao1 and Shannon. Finally, the main difference where low yield has a significant impact can be observed in the fecal samples. This is expected since among all the tested samples, fecal ones are the most diverse and sub-sampling will really affect them as observed in Figure 3.

We agree with the reviewer; we add some thoughts just to clarify. We indeed observed that extremely rare species (with frequencies lower than 1/10000) are lost when subsampling to the most extreme levels. When subsampling to 100K reads we are losing species with a frequency around 1/100,000 (very approximate estimate). However, the effect of losing such species on the global sample diversity as estimated by Shannon diversity index is negligible (see Figure 4, in which we show that reduction in sequencing depth has no dramatic effect on Shannon's diversity

index). The situation is different for the Chao 1 estimator. This is expected and is due to the way Chao1 is computed: this estimator relies heavily on the number of singletons (i.e. species represented by only one read). By subsampling, singletons (i.e. the rarest species) are very likely to be lost. The same phenomenon can be inferred by looking at Figures 5 and 6. Those represent a scatterplot of the relative abundance of species in full sample and reduce samples (100K and 10k reads, respectively). The plots are shown in log log scale to emphasize differences for low-frequency species. Only low-frequency species have some variation in frequency estimation. However, even when sampling only 10K read, species with frequency around 0.1% (i.e. 1/1000) are appropriately quantified. All of these observations led us to conclude that coverage reduction doesn't prevent a satisfactory characterization of complex matrices (with the only exception of Chao 1 estimator).

Since the composition of each sample is not known a priori, then there are some factors that can contribute to biases. As mentioned, the DNA concentration but also its integrity (fragmentation) will affect the library construction; the cited kit requires DNA amplification which will have a bias towards GC rich genomic regions; library size was not described.

The Nugen Ovation® Ultralow System V4 kit used is a standard kit for NGS library preparation (https://www.nugen.com/sites/default/files/DS_v2-Ovation_Ultralow_V2.pdf). It is a standard protocol widely used by the scientific community to perform DNA-seq also from low input DNA quantities (1 ng), even if in our case input DNA was of moderate quantity. Mock community experiment will shed light on eventual biases. DNA concentration and integrity as well as input DNA quantities used in library construction and libraries insert size will be reported in the version 2 of the paper.

It was not mentioned if the samples were pooled with other libraries with different insert sizes, which affect not only the sequencing quality but the yield.

Samples were sequenced in different runs and pooled with other libraries of similar insert sizes. The number of reads obtained per sample reflects and respects their quantities, i.e. nmols that were loaded on the sequencer.

In terms of bioinformatics analysis, it will be required to put the parameters used for each program, in case someone wants to reproduce this. For Kraken2, it is important to know what is the kmer size to index the database. For MEGAHIT assembly it will be important to know the kmer and step sizes used.

All these details will be provided in the version 2 of the paper.

For the completeness assessment, the authors used BUSCO, but apparently they are using the whole assembly to assess the completeness. This is not correct, since they must first separate in bins which genomes they have really reconstructed and then they can assess the completeness of them. Probably they can report the an average completeness value for all the reconstructed genomes. By doing the binning they can have a better analysis of what was really reconstructed and how complete it was.

This is a good point. While our aim was to estimate the total proportion of BUSCO genes that were reconstructed, irrespective of the species of the organism to which they belong, we understand that a practical application is likely to require separating the reconstructed genomes. We will

integrate our analysis by binning the reconstructed genomes.

The use of Krona in Figure 2 is not very convenient. The whole point of a Krona graph is that is interactive. If authors want to provide the Krona data to be downloaded it would be possible and recommended. Having said that, I recommend to use bar plots to represent the relative abundance and composition of the samples at a given taxa level.

We will either provide a link to interactive krona graphs and/or bar plots reporting the relative abundance and composition of the samples.

Competing Interests: No competing interests were disclosed

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research