METHOD ARTICLE

REVISED Do you cov me? Effect of coverage reduction on species identification and genome reconstruction in complex biological matrices by metagenome shotgun high-throughput sequencing [version 2; peer review: 3 not approved]

Federica Cattonaro, Alessandro Spadotto, Slobodanka Radovic, Fabio Marroni

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

v2 **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)
Latest published: 29 Jul 2019, **7**:1767 (https://doi.org/10.12688/f1000research.16804.3)

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. However, we demonstrate that—for some applications—it is possible to use shallow shotgun metagenomics to characterize complex biological matrices while reducing costs. 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 alpha diversity estimates, species abundance, detection threshold, and ability of reconstructing the metagenome in terms of length and completeness. Our results show that a classification of prokaryotic, eukaryotic and viral communities can be accurately performed even using very low number of reads, both in mock communities and in real complex matrices. With samples of 100,000 reads, the alpha diversity estimates were in most cases comparable to those obtained with the full sample, and the estimation of the abundance of all the present species was in excellent agreement with those obtained with the full sample. On the contrary, any task involving the reconstruction of the metagenome performed poorly, even with the largest simulated subsample (1M 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.

Keywords

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

Open Peer Review

Reviewer Status ✗ ✗ ?

	Invited Reviewers		
	1	**2**	**3**
REVISED **version 3** published 29 Jul 2019			? report
REVISED **version 2** published 22 Mar 2019		✗ report	✗ report
version 1 published 08 Nov 2018	✗ report	✗ report	

1 **Alejandro Sanchez-Flores**, National Autonomous University of Mexico (UNAM)), Cuernavaca, Mexico

2 **José F. Cobo Diaz**, Université de Brest, Plouzané, France

3 **Francesco Dal Grande**, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany
LOEWE Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

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

Corresponding authors: Federica Cattonaro (fcattonaro@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 (MPRV vaccines, Prorix Tetra, GlaxoSmithKline) was financed by Corvelva (non-profit association, Veneto, Italy), in the frame of a contract work 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.

How to cite this article: Cattonaro F, Spadotto A, Radovic S and Marroni F. **Do you cov me? Effect of coverage reduction on species identification and genome reconstruction in complex biological matrices by metagenome shotgun high-throughput sequencing [version 2; peer review: 3 not approved]** F1000Research 2019, **7**:1767 (https://doi.org/10.12688/f1000research.16804.2)

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

REVISED **Amendments from Version 1**

In this version we incorporated all the suggestions of the reviewers.

Reviewer Alejandro Sanchez-Flores:

- We added a mock community (sample A1)
- We added several details on sample type and input DNA quality.
- We clarified that the aim of the work is to estimate the effect of the sequencing depth on metagenomics studies; thus, we chose heterogeneous samples to obtain results of general applicability.
- We added estimate of the detection threshold of rare species at varying sequencing depths.
- We provide more detail regarding the use of megahit for *de novo* assembly. We share online a pipeline to reproduce the main manuscript analyses.
- We are now performing BUSCO separately on each species and then reporting average statistics.
- We replaced Krona charts with a barplot. Krona charts are available online as html interactive graphs.

Reviewer José F. Cobo Diaz:

- We rewrote the introduction to clarify that the aim of the work is to analyze the effect of varying sequencing depth in the characterization of complex matrices sequenced via whole genome shotgun. The first version erroneously convinced the readers that our focus was on analysis of functional data and/or on pathogens detection.
- We added estimate of the detection threshold of rare species at varying sequencing depths.
- We describe parameters used for bioinformatics analysis. We share online a pipeline to reproduce the main manuscript analyses.
- We provide several measures related to species detection by our approach. Our approach for species classification is accurate. However, a small number of reads (possibly due to sequencing errors) is responsible for the inflation of the number of observed species.
- We incorporated Good's coverage and Pielou's evenness index in the analysis.
- We modified the discussion to clarify the aim of the work and the conclusions that can be drawn based on our observations.

See referee reports

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, fungi, 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 samples[7–9], skin samples[10], clinical samples for diagnostics and surveillance purposes[11–14] 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 bacteria[17], the nuclear ribosomal internal transcribed spacer (ITS) region for fungi[18], 18S rRNA gene in eukaryotes[19], cytochrome c oxidase sub-unit I (*COI* or *cox1*) for taxonomical identification of animals[20], *rbcL*, *matK* and *ITS2* as the plant barcode[21]. 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 preparations[22]. In addition, 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[23].

Current whole shotgun metagenome experiments are performed obtaining several million reads[5,8]. However, obtaining a broad characterization of the relative abundance of different species, might easily be achieved with lower number of reads.

To test this hypothesis, we analzyed 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. In addition, publicly available viral particle enriched sequencing data was used to extend our analysis to viruses. Finally, we included in the study a mock community sample with known species composition.

Some of the samples were predominantly composed by eukaryotic organisms, while others were composed by prokaryotes or viruses; some were represented by very few dominant species while others had greater diversity. Results that were observed across such dissimilar samples are likely to be of general validity.

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 faecal samples (V1 and V2).

Biological medicines were two different lots of live attenuated MPRV vaccine, widely used for immunisation 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 fluorometer (Invitrogen, Carlsbad, CA, USA).

Human fecal samples V1 and V2 derive from a study investigating the virome composition of feces of Amerindians[24]. The two samples with the highest sequencing depth were choosen. Sequences were retrieved from SRA (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 fluorometer (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 cutadapt[25]. Masked and low quality bases were filtered using erne-filter version 1.4.6.[26]. 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[27].

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

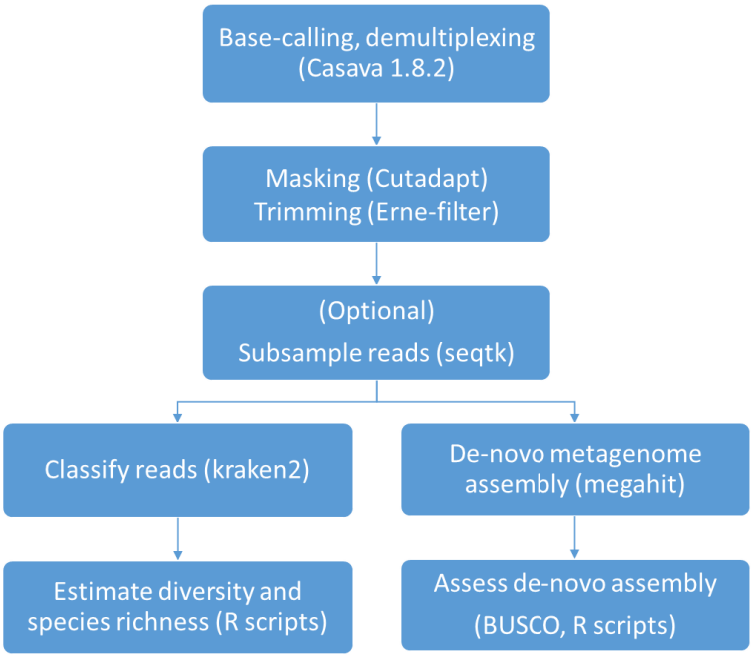

Figure 1. Workflow of the main bioinformatics analysis performed in the present work.

To classify the largest possible number of prokaryotes, eukaryotes and viruses, reads were classified against the complete NCBI nt database using kraken2, version 2.0.6[28]. 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[29].

Observed number of taxa, Chao1 species richness[30], Good's coverage[31], Shannon's diversity index[32] and Pielou's index[33] were estimated using the R package vegan version 2.4.2[34] or base R functions. The number of observed taxa was computed as the number of species to which at least one read was assigned. The number of singletons is defined as the number of species identified by only one read. The number of core species is the number of species with frequency equal or greater than 1‰. We then define the measure S90, obtained as follow: a) sort species in decreasing abundance, b) perform cumulative sum of the species abundance, and c) report how many of the ordered species are needed to reach an abundance equal or greater to 90% of the total number of reads.

Assembly of the metagenome was performed using megahit version 1.1.2[35] with default parameters, with kmer sizes varying as follows: 21, 29, 39, 59, 79, 99, 119, 141. Reconstructed contigs were classified at the species level using kraken2. Completeness of the assemblies of each species was assessed using BUSCO[36]. 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 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 (predominanty composed by fungi) for which the fungal database was used.

Unless otherwise specified, all the analysis were performed using R 3.3.3[37].

Results

Sample composition and downsampling

Summary statistics for the full samples included in the study are shown in Table 1.

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 millions (sample F1). The number of species identified in each sample was very high, ranging from 2,508 in sample B1 to 29,661 in sample F1. However, the 20 most abundant species accounted for a large proportion of the reads in each sample, from 74.62% in M2 to 99.75% in B1, and the 100 most abundant species accounted for 84.7% in M2 and 99.8% in B1. In sample A1 98.8% of the reads were assigned to the 20 declared species, and only 1.2% of reads were either unassigned or uncorrectly attributed to other species. 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 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. Reads that were either unclassified or assigned to rare phyla were aggregated under the name "Unknown/Other". Samples B1, B2 and M3 where mainly composed of Chordata, sample M1 was mostly

Table 1. Summary statistics for the full samples included in the study.

Sample	N reads	Core	N species	Singletons	% Top 20	% Top 100	S90
A1	4,969,245	16	2,571	1,191	98.91	99.66	7
B1	11,031,061	4	2,507	1,299	99.75	99.81	1
B2	3,830,083	9	4,597	1,795	98.83	99.27	2
F1	12,472,553	99	29,660	14,750	21.16	38.45	2,795
F2	10,780,450	106	25,607	12,374	19.94	36.67	2,947
M1	1,898,011	2	3,206	1,469	99.03	99.35	1
M2	1,558,975	132	9,637	3,377	36.68	61.68	1,218
M3	1,867,879	19	5,566	1,999	95	97.38	7
V1	1,300,221	76	6,372	2,114	73.91	86.05	186
V2	2,001,984	9	3,177	1,605	98.96	99.38	2

Core: number of species with frequency greater than 1‰. **N species**: number of species identified in the sample; include species identified by one or more reads. **Singletons**: number of species identified in the sample by only one read. % **Top 20**: percentage of reads assigned to the 20 most abundant species. % **Top 100**: percentage of reads assigned to the 100 most abundant species. **S90**: Number of species accounting for 90% of the reads.

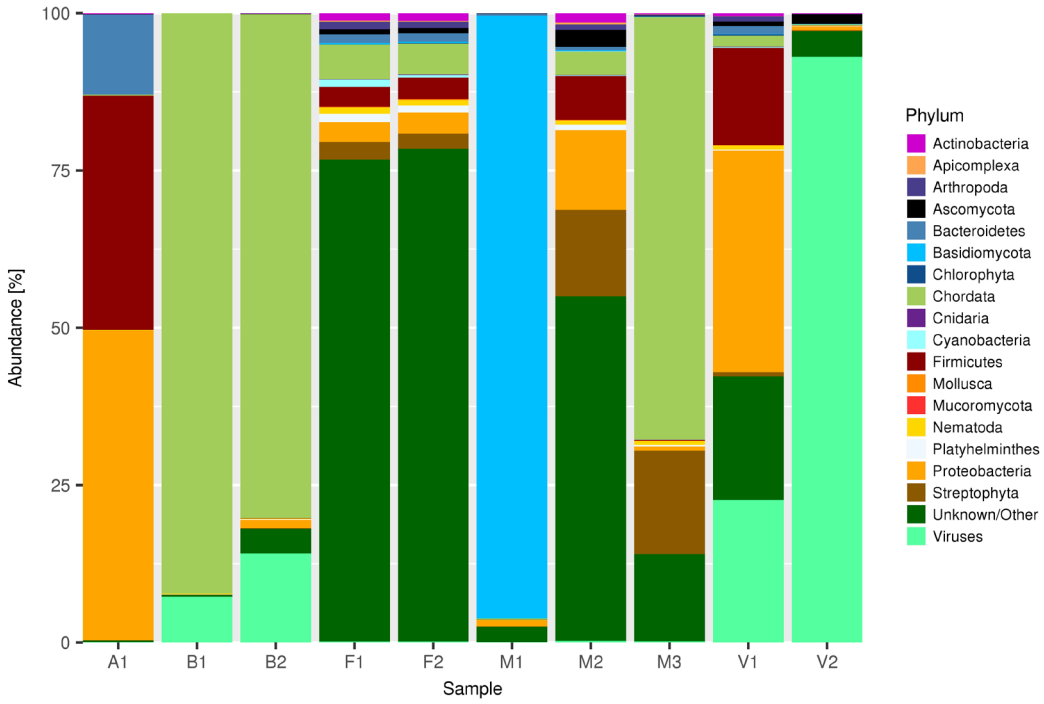

Figure 2. Phylum composition of the samples. Only phyla represented by at least 1% of the reads are shown. Viruses are presented at division level. Unclassified reads and reads assigned to rare phyla are aggregated under the name "Unknown/Other".

composed by 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.

Mock community analysis

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%.

Figure 3 shows the scatterplot (in logarithm scale) of the observed and expected abundance of organisms of the mock community at different taxonomic levels, from Sepcies to Phylum when using the full-dataset (4.9 M of reads). The correlation between the two measures is high even at the species level (r=0.87), and increases for higher taxonomic levels reaching 0.95 at Class and Phylum level.

The correlation between expected and observed abundancies of the 20 mock species remained high when decreasing sequencing depth, and Pearson's correlation coefficient remains stable at 0.87 at all the investigated sequencing depths. Results for the hwole depth, 1,000,000 reads, 25,000 reads and 10,000 reads, together with 95% intervals are shown in Figure 4.

Diversity analysis

Figure 5 shows the variation of several summary statistics as a function of the number of reads used for the analysis, from the smallest (10,000 reads) on the left, to the full dataset on the right. Panels A and B show the observed number of taxa and the value of Chao1 (expected number of taxa) respectively. The two measures have very similar trend, with a swift decrease in horse feces (F1 and F2) when going from full set to 1,000,000 reads, and a relatively slow decrease in all other samples and subsets.

Downsampling has different effects on the observed and estimated number of species in different samples. For most samples, even a robust downsampling led to only a slight reduction in the estimated species richness. However, for samples F1 and F2, characterized by a high number of species including rare ones, the downsampling led to a significant reduction (panels A and B). Good's coverage (panel C) remained nearly constant when more than 100K reads were sequenced. Lower sequencing depth determined a decrease in Good's coverage, especially for samples F1, F2, M2 and V1.

Shannon's diversity index (panel D) 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 all samples.

Pielou's index (panel E) is a measure of the species' distribution evenness. Values close to 1 denote equifrequent species, and lower values denote uneven distribution of species relative abundance. The effect of the number of reads on Pielou's index is moderate.

Species abundance and detection threshold

Figure 6 shows the correlation in species abundance estimation between the full dataset and a reduced dataset of 100,000 reads. The linear correlation coefficient between the two datasets was >0.99 in all five subsampling replicates. The plot

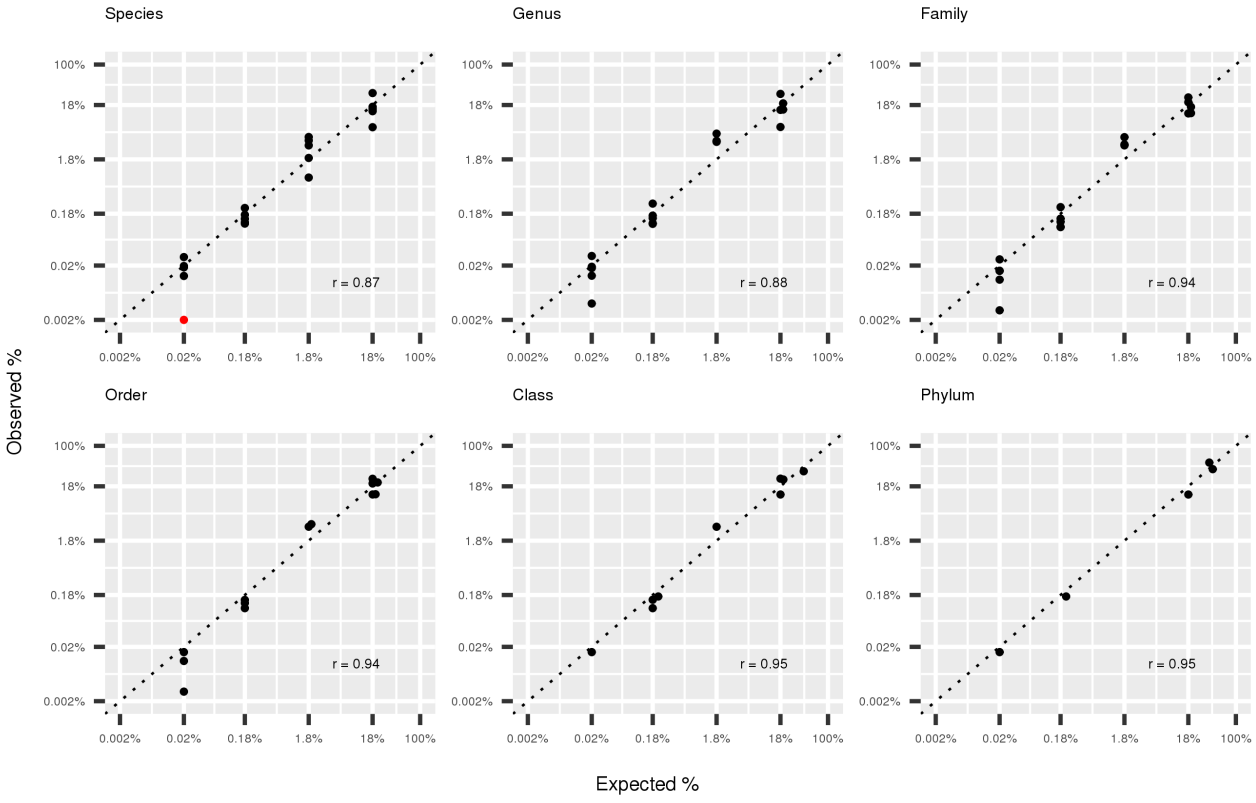

Figure 3. Log-log scatterplot of observed and expected abundance of bacterial organisms present in the mock community "20 Strain Staggered Mix Genomic Material" (ATCC® MSA-1003™). In red *Actinomyces odontolyticus* identified at frequency <0.002%, arbitrarily plotted at 0.002%.

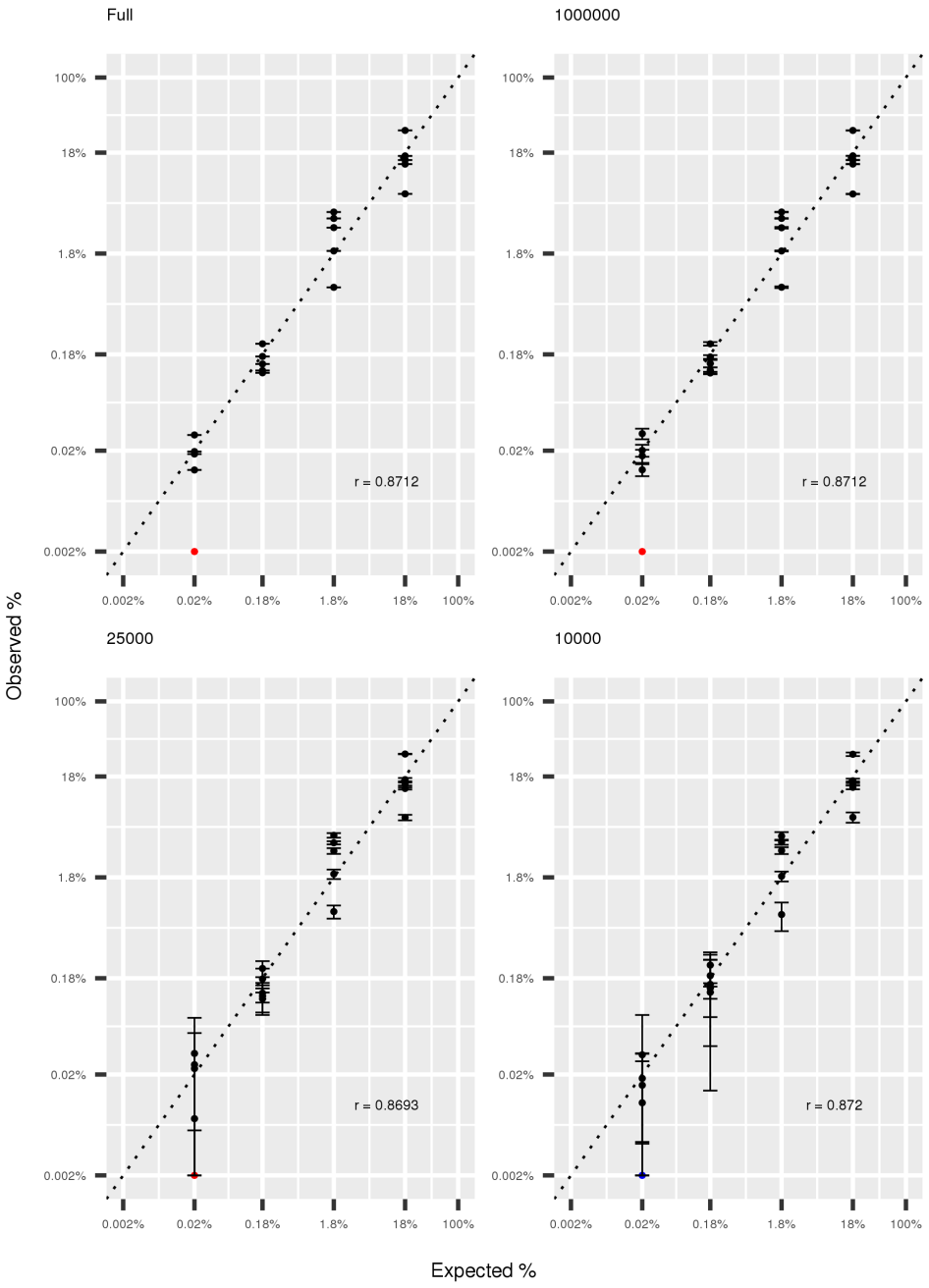

Figure 4. Log-log scatterplot of observed and expected abundance of bacterial species present in the mock community "20 Strain Staggered Mix Genomic Material" (ATCC® MSA-1003™) at varying sequencing depths. In red *Actinomyces odontolyticus* identified at frequency <0.002%, arbitrarily plotted at 0.002%. Error bars represent 95% confidence intervals obtained from five resampling experiments.

is in log-log scale to emphasize differences in low abundance species. Only the relative abundance estimation of species with frequencies lower than 0.01% (i.e . species represented by 1 read out of 10,000) was affected by subsampling. The same pattern was observed in all examined samples.

In Figure 7 we show the results obtained by reducing the number of sampled reads to 10,000 reads per sample. Similarly to what we observed for 100,000 reads depth, the linear correlation coefficient between species abundance estimate in the full and the reduced dataset was high (r>0.95) for all the samples and in all five subsampling replicates. Only rare species with frequencies lower than 1/1000 (0.1%) in full dataset showed some deviation.

Since the reduction of the sequencing depth inevitably affects the ability of detecting rare species, we determined the minimum frequency required for a species to be identified

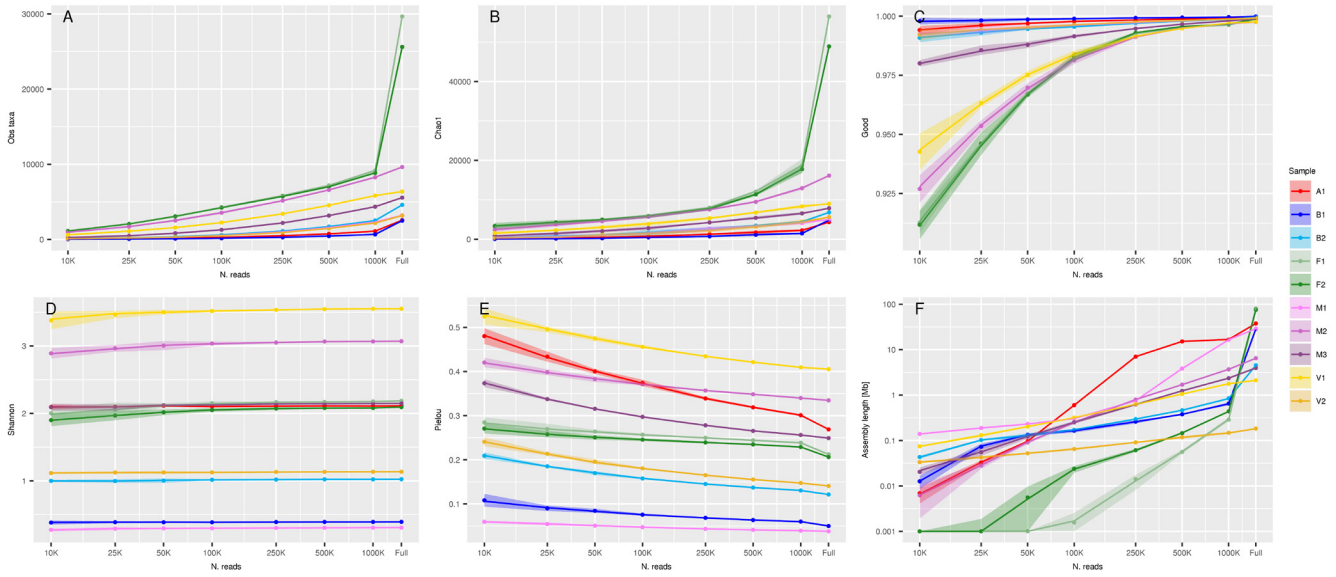

Figure 5. Effect of reduction of sequencing depth on: **A**) Observed number of taxa, **B**) Chao1 estimated number of taxa, **C**) Good's Coverage, **D**) Shannon's diversity index, **E**) Pielou's diversity index, and **F**) Total length of *de novo* assembly. In all panels X axis is in log scale and 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. "Full" represents the values obtained with the full set of reads.

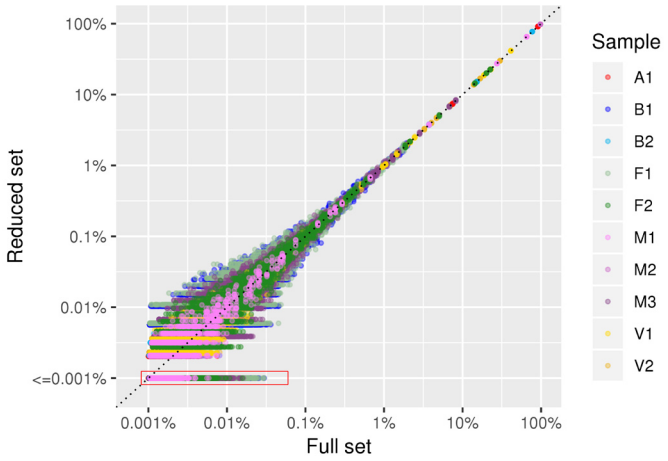

Figure 6. Correlation of species abundance estimated using the full dataset and a set composed of 100,000 reads. Data for all the five subsampled replicates are plotted. Each point (colored by sample of origin) represents a given species. Both axis are plotted in log scale to facilitate visualization of low abundance species. A red box encompasses datapoints of species that were present in the full set and absent in the reduced set, for which the frequency in the reduced set was set at "<=0.001%".

at each sequencing depth. This detection threshold at any given sequencing depth was defined as follows: a) for each sample, identify all the species that are present in all five subsampling replicates; b) among the species identified, for each sample select the one with the lowest frequency in the full dataset; c) average the lowest frequencies across all samples. Table 2 shows the average and standard deviation of the detection threshold across the ten samples at any sequencing depth. At 10,000 reads depth, the detection threshold was 0.0124%. This means that species with frequencies higher than 0.0124% in full dataset were consistently identified also in the reduced datasets, while species with lower frequencies may be lost. At 1,000,000 reads depth the detection threshold was 0.00006% (*i.e.* 60 reads per million).

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 F). As expected, the size of the assembly was strongly influenced by the sequencing depth.

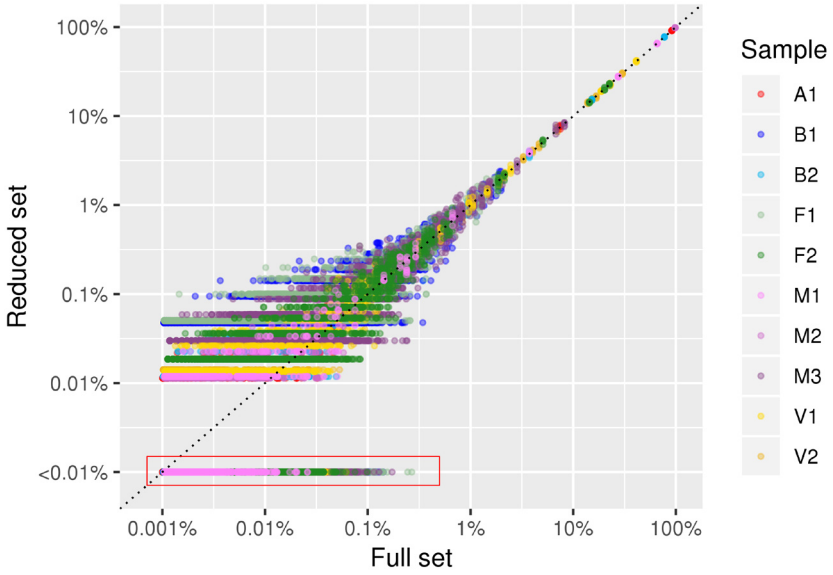

Figure 7. Correlation of species abundance estimated using the full dataset and a set composed of 10,000 reads. Data for all the five subsampled replicates are plotted. Each point (colored by sample of origin) represents a given species. Both axis are plotted in log scale to facilitate visualization of low abundance species. A red box encompasses datapoints of species that were present in the full set and absent in the reduced set, for which the frequency in the reduced set was set at "<0.01%".

Table 2. Detection threshold as a function of the sequencing depth. N. reads: Number of reads. **Detection threshold (%):** Detection threshold averaged across the ten samples, **SD(%):** Standard deviation of the detection threshold.

N. reads	Detection threshold (%)	SD (%)
10,000	0.01242	0.006312
25,000	0.00348	0.001719
50,000	0.00189	0.001078
100,000	0.00069	0.000536
250,000	0.0001	6.53E-05
500,000	0.00007	4.77E-05
1,000,000	0.00006	4.9E-05

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 analysis[36] was used as an additional measure to assess the completeness of the reconstructed metagenome. The proportion of reconstructed genes in full (X axis) and reduced (Y axis) datasets obtained by randomly sampling 1,000,000 reads is shown in Figure 8. In samples A1 and M1, on average 80% of the BUSCO genes were reconstructed in the full dataset. Reducing sequencing depth to 1,000,000 reads lowered the porportion of reconstructed genes in the two samples to 50% or less. In the remaining samples the proportion of reconstructed genes was very low even in the full dataset and the reduction of sequencing depth did not significantly alter the proportion.

Discussion

We set out to test the effect of the reduction of sequencing depth on 1) estimates of diversity and species richness; 2) estimates of abundance of the species present, 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 started by determining the general characteristics of our samples. All the samples resulted as a mixture of a large number of species, nearly half of which were singletons (*i.e.* represented by one read). A control sample A1 comprised 2,572 species, while it should contain only 20 of them. However, A1 core set (species with a frequency of at least 0.1%) was made up by 16 species. Based on product specifications, 15 species in the mock community had a frequency greater than 0.1% and we observed all of them. In addition, we erroneously identified *Staphylococcus lugdunensis* (with a frequency of 0.11%), probably due to misclassification of other *Staphylococcus* reads. We devised the S90 measure which reports the number of the species (sorted by decreasing abundance) accounting for 90% of the reads. For several samples the S90 is less than 10, while for highly complex matrices as F1 and F2, is 2,795 and 2,947 respectively. The abundance of rare species might be

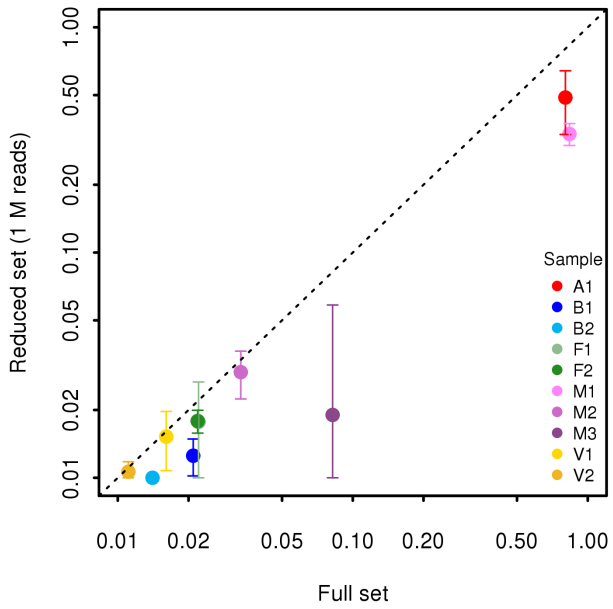

Figure 8. Completeness of the BUSCO genes in the full dataset (X axis) and in the largest of the reduced datasets (consisting of 1,000,000 reads, Y axis); error bars are based on the five replicate experiments performed for each sample. The plot is in log-log scale.

factual for samples with very high complexity such as feces. Still, species represented by only one read are unlikely to be real. A proportion of singleton species is probably originated from sequencing errors and/or from errors in the classification against the database. In addition, especially for low input samples, it is possible that contaminants in laboratory reagents artificially increase the number of observed species[38,39]. Nevertheless, determining the relative and absolute contribution of these biases to metagenomics studies is out of the scope of the present paper.

The choice of the database against which sequences are matched can affect results. In the present study, we matched our sequences against the full NCBI nt database, because this allows to classify reads belonging to any given organism. However, this might cause some drawback in accuracy. As an example, in the vaccine sample B1, we identified 61 reads attributed to *Elaeophora elaphi*, a nematode, only found as a parasite of the liver of deers[40]. It is therefore highly unlikely that such organism might really be present in the vaccine sample. Repeating the analysis on the standard database, only consisting of *Homo sapiens*, bacteria and viruses, 57 out of 61 reads were assigned to *Homo sapiens* and the remaining 4 were unassigned (data not shown). Possible explanations are that a) some contamination from *Homo sapiens* is present in the deposited sequence of *Elaeophora elaphi*, or b) some reads belonging to *Homo sapiens* are attributed by mistake to genuine *Elaeophora elaphi* sequences.

Such marginal missclassification problems do not affect the results of our study, but clearly indicates that researchers should be very cautious when reporting contaminants or unexpected results from metagenomics studies.

In our study we kept the read length constant at 125 bp across experiments. Previous studies (although limited to targeted approaches) showed the effect of read length on the evaluation of the composition of complex matrices[41]. Even though an extensive assessment of the effect of read length on the ability to characterize complex matrices was beyond the scope of the present work, we compared the results obtained for horse fecal samples (F1 and F2) when using 250 bp long reads. The use of shorter sequences led to a strong increase in the proportion of unclassified reads, from 56% to 74% in F1 and from 58% to 75% in F2.

We performed a benchmark of the entire workflow with the help of a mock community with known composition. By comparing the expected and observed relative abundance of the 20 bacterial species included in the mock community we concluded that the workflow is accurate at all taxonomic levels (Figure 3). One species, *Actynomices odontolyticus*, with expected frequency of 0.02%, was observed with a much lower frequency (<0.002%, represented as a red dot in Figure 3 and Figure 4). Other species showed only slight deviations from expected frequencies in our experiment. To the best of our knowledge, this is the first published work reporting the observed frequencies of a mock community using WGS. However, previous works performed very extensive studies on target 16s sequencing of mock communities, and reported large deviations from expectation, depending on sequencing primers, extraction method and sequencing platform[42]. We tested the effect of decrease in sequencing depth on deviations from expected frequency (Figure 4) and observed that even when sampling 10,000 reads the average correlation between expected and observed abundances remained high (r=0.87), although the variance among resampling experiments was high.

To assess the requirements in sequencing depth for characterizing complex matrices, we measured the variation of several diversity indexes while reducing the sequencing depths. We measured the number of observed taxa, Chao1 (or number of expected taxa), Good's coverage, Shannon's diversity index and Pielou's evenness index.

Chao1 estimator is obtained as

$$S_{Chao1} = S_{Obs} + \frac{f_1(f_1-1)}{2(f_2+1)}$$

Where S_{obs} is the number of observed species in the sample, f_1 is the number of species observed once, and f_2 is the number of species observed twice.

Under our experimental conditions, the number of observed and estimated taxa followed similar trends. Both of them were heavily affected by the small proportion of reads attributed to unique or rare taxa.

Good's coverage (G) is defined as

$$G = 1 - \frac{f_1}{N}$$

where f_1 is the number of singletons and N is the total number of reads. G is heavily affected by the sequencing depth. Significant variation in G is observed when using 100,000 reads or less.

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. Thus Shannon diversity index is affected more by variation in the frequencies of highly abundant species than by the loss of rare species. In our study, Shannon's index was very stable across sample sizes.

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, thus Pielou's index approaches 1 when all the species are evenly distributed. In our study, Pielou's index showed a slight increase as the number of sampled reads decreased.

Horse fecal samples F1 and F2 are characterized by a very large number of observed species (29,660 and 25,607, respectively), while all the other samples have lower number of species, ranging from 2507 in B1 to 9637 in M2. Chao1 captures this differences, showing that F1 and F2 have greater diversity estimates Measures such as the number of observed taxa and Chao1 capture this differences, showing that F1 and F2 have greater diversity estimates. Shannon's and Pielou's indices, on the contrary, rely on the frequency distribution of the species. Therefore, samples that have a relatively high number of common species with comparable frequencies tend to have high Shannon's diversity indices. Samples (such as M1) dominated by a single species, have very low Shannon and Pielou indices. The effect of sequencing depth on nearly all indices is moderate; we thus conclude that biological matrices with different levels of complexities, composed by different admixture of prokaryotes, eukaroytes and viruses can be satisfactorily characterized via WGS even at sequencing depth lower than 1,000,000 reads.

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 even in case of substantial reduction of the number of sequenced reads, species abundances as low as 0.1% can be reliably estimated (Figure 6 and Figure 7).

In addition, we aimed to determine the threshold of detection for rare species at low sequencing depth. This statistics is of interest when researchers are interested in detecting the presence of a species that might be rare in the sample. Our results show that even very rare species can be accurately detected at low sequencing depth. When subsampling 1,000,000 reads, the frequency threshold for a species to be detected in the reduced sample was measured as 60 reads out of 1,000,000 (0.00006%). Even when the number of reads was unrealistically low (10,000), rare species could still be detected, with a detection threshold estimated to be 0.012%. While the detection threshold can vary according to sample characteristics, we can assume that for most samples rare species can be accurately detected even at low sequencing depth.

Finally, we assessed the effect of a reduction in the sequencing coverage on the accuracy of *de novo* assembly of the metagenome. Our results show that downsampling had a strongly negative effect on the total length of the reconstructed metagenome and on the propoprtion of reconstructed genes (Figure 5F and Figure 8).

BUSCO is widely used for assessing the completeness of genome and transcriptome assemblies for individual organisms, and has benchmark datasets for several lineages. Our results clearly indicate that even 1,000,000 reads is a suboptimal depth in terms of fully sampling the genes present in the complex matrices. This observation needs to be taken into account in the phase of experimental design. Our conclusions are also important for research interested at reconstruction of an interesting part of the meta-genome, such as genes involved in antibiotic resistance[43]. The decrease in performance observed in the genes' recostruction 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. In addition, the proportion of genes reconstructed with BUSCO in the full dataset was very low for all samples, with the exception of the two samples M1, predominanty composed by one fungal species, and A1, composed by a limited number of small genomes. These results indicate that a complete reconstruction of the metagenome of a complex matrix requires at least several millions reads. 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 WGS is a good choice. 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 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.

Samples F1 and F2 are available under accession number SRP163102: https://identifiers.org/insdc.sra/SRP163102.

Samples B1 and B2 are available under accession number 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.

Extended data

Open Science Framework: Do you cov me. https://doi.org/10.17605/OSF.IO/Y7C39[44].

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.

Grant information

Metagenome sequencing of B1 and B2 was financed by Corvelva (non-profit association, Veneto, Italy), in the frame of a contract work 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.

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 Kindom) 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

1. 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

2. 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

3. Bragg L, Tyson GW: **Metagenomics using next-generation sequencing.** *Methods Mol Biol.* 2014; **1096**: 183–201.
PubMed Abstract | Publisher Full Text

4. 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

5. 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

6. 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

7. 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

8. 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

9. 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

10. 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

11. 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

12. 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

13. 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

14. 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

15. 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

16. 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

17. 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

18. 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

19. Hugerth 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

20. 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

21. 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

22. 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

23. 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

24. 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

25. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet J.* 2011; **17**(1): 10–2.
Publisher Full Text

26. Del Fabbro C, Scalabrin 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

27. Marroni F : **fabiomarroni/doyoucovme v1.2 (Version v1.2).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2593798

28. 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

29. 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

30. Chao A: **Non-parametric estimation of the classes in a population.** *Scand J Statist.* Scandinavian Journal of Statistics; 1984; **11**(4): 265–70.
Reference Source

31. Good IJ: **The Population Frequencies of Species and the Estimation of Population Parameters.** *Biometrika.* Oxford University Press Biometrika Trust; 1953; **40**(3/4): 237–264.
Publisher Full Text

32. Shannon CE: **A Mathematical Theory of Communication.** *Bell Syst Tech J.* 1948; **27**(3): 379–423.
Publisher Full Text

33. Pielou EC: **The measurement of diversity in different types of biological collections.** *J Theor Biol.* Academic Press; 1966; **13**: 131–44.
Publisher Full Text

34. Oksanen J, Blanchet G, Friendly M, *et al.*: **vegan: Community Ecology Package**. 2017.
Reference Source

35. 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

36. 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

37. R Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2018.

38. Wally N, Schneider M, Thannesberger J, *et al.*: **Plasmid DNA contaminant in molecular reagents.** *Sci Rep.* 2019; **9**(1): 1652.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Salter SJ, Cox MJ, Turek EM, *et al.*: **Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.** *BMC Biol.* 2014; **12**(1): 87.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Hernández Rodríguez S, Martínez Gómez F, Gutiérrez Palomino P: ***Elaeophora elaphi* n. sp. (Filarioidea: Onchocercidae) parasite of the red deer *(Cervus elaphus)*. With a key of species of the genus *Elaeophora*.** *Ann Parasitol Hum Comp.* EDP Sciences; 1986; **61**(4): 457–63.
PubMed Abstract | Publisher Full Text

41. Wommack KE, Bhavsar J, Ravel J: **Metagenomics: read length matters.** *Appl Environ Microbiol.* American Society for Microbiology; 2008; **74**(5): 1453–63.
PubMed Abstract | Publisher Full Text | Free Full Text

42. 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

43. 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

44. Marroni F: **Do you cov me**. 2019.
http://www.doi.org/10.17605/OSF.IO/Y7C39

Open Peer Review

Current Peer Review Status: ✗ ✗ ✗

Version 2

Reviewer Report 30 May 2019

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

✗ **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 result 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.

Reviewer Report 25 March 2019

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

José F. Cobo Diaz
Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, IBSAM, ESIAB, Université de Brest, Plouzané, France

I appreciate the changes make 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 asume you have done some kind of clustering since you talk about singletons) and taxonomic assignation until you have the number of species expected, 20 in this case. You can also have some less due to problems with species assignation, 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-make. I 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 make 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 asume you have done some kind of clustering since you talk about singletons) and taxonomic assignation until you have the number of species expected, 20 in this case. You can also have some less due to problems with species assignation, 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

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

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