METHOD ARTICLE

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 1; referees: awaiting peer review]

Federica Cattonaro, Alessandro Spadotto, Slobodanka Radovic, Fabio Marroni

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

v1 **First published:** 08 Nov 2018, **7**:1767 (https://doi.org/10.12688/f1000research.16804.1)
Latest published: 08 Nov 2018, **7**:1767 (https://doi.org/10.12688/f1000research.16804.1)

Abstract

Shotgun metagenomics sequencing is a powerful tool for the characterization of complex biological matrices, enabling analysis of prokaryotic and eukaryotic organisms in a single experiment, with the possibility of *de novo* reconstruction of 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. Here we compared the results obtained on full size, real datasets with results obtained by randomly extracting a fixed number of reads. The main statistics that were compared are alpha diversity estimates, species abundance, and ability of reconstructing the metagenome in terms of length and completeness. Our results show that a classification of the communities present in a complex matrix can be accurately performed even using very low number of reads. 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 sensibly 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 *de novo*—even partially—the metagenome.

Keywords

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

Open Peer Review

Referee Status: *AWAITING PEER REVIEW*

Discuss this article

Comments (0)

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 1; referees: awaiting peer review]** *F1000Research* 2018, **7**:1767 (https://doi.org/10.12688/f1000research.16804.1)

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

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 for agnostic characterization of complex communities containing eukaryotes, fungi, bacteria and also viruses, using both DNA and RNA as a starting material. In addition, the whole metagenome approach can be used not only to simply identify DNA and RNA virus in a complex matrix, but also to study the genetic diversity in virus populations[3–5], and to identify potential adventitious agents in biopharmaceutical manufacturing[6,7].

Metagenome shotgun high-throughput sequencing has progressively gained popularity in parallel with the advancing of next-generation sequencing technologies[8,9], 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[10,11], gut samples[12–14], skin samples[15], clinical samples for diagnostics and surveillance purposes[16–19], food ecosystems[20,21] and drugs manufactured using biological sources as vaccines[22].

The aim of whole metagenome approaches is not only to study the taxonomic composition of biological substrates but also to identify which genes and metabolic pathways are present with the aim to understand functional capacities in the studied microbiota[13,23]. Recently the approach has been also used to analyze the ensemble of genes that may encode antibiotic resistance in various microbial ecosystems (i.e. soil), which are defined as the resistome[24].

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[27], the nuclear ribosomal internal transcribed spacer (ITS) region for fungi[28], 18S rRNA gene in eukaryotes[29], cytochrome c oxidase sub-unit I (*COI* or *cox1*) for taxonomical identification of animals[30], *rbcL*, *matK* and *ITS2* as the plant barcode[31]. Considering the large amount of genetic diversity within and between virus families, a universal metabarcoding approach is not applicable to detect virus nucleic acids in complex biological samples.

The selection of conserved regions has the advantage of reducing sequencing needs, since it does not require sequencing of the full genome, just a small region. On the other hand, given the currently used approaches, characterization of microbial and eukaryotic communities requires different primers and library preparations[32]. 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 sample[33].

Current whole shotgun metagenome experiments are performed obtaining several million reads[10,13]. 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 performed sequencing using whole metagenomics approach of seven samples derived from different complex matrices to characterize their composition, and subsequently tested the accuracy of several measures when downsampling the number of reads used for analysis including the performance of *de novo* assembly in the ability to reconstruct both entire genomes and genes.

Methods

Samples description and DNA extraction

The following samples were used in the present work: two samples collected from a live attenuated virus vaccine (B1 and B2), two horse fecal samples (F1 and F2), and three food samples (M1, M2, and M3).

Biological medicines were two different lots of live attenuated MPRV vaccine (Prorix Tetra, Glaxo SmithKline) 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 using Maxwell® 16 Instrument and the Maxwell® 16 Tissue DNA Purification Kit (Promega, Madison, WI, USA) according to the manufacturer's instructions.

Horse feces from two individuals were collected as follows: 100 mg of starting material stored in 70% ethanol were processed 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.

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).

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).

Cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturization and hybridization of the sequencing primers was then performed on Illumina cBot and flowcell HiSeq SBS V4 250 cycle kit, loaded on

HiSeq2500 Illumina sequencer producing 125bp paired-end reads (for samples B1, B2, M1, M2 and M3) and 250bp paired-end reads (for samples F1 and F2).

The CASAVA Illumina Pipeline version 1.8.2 was used for base-calling and de-multiplexing. Adapters were masked using cutadapt[34]. Masked and low quality bases were filtered using erne-filter version 1.4.6[35].

Bioinformatics analysis

The bioinformatics analysis performed in the present work are summarized in Figure 1.

Different read lengths among samples may constitute an additional confounder in analysis. To obtain homogeneous read length across samples, reads sequenced belonging to F1 and F2 were trimmed to a length of 125 bp using fastx-toolkit version 0.0.13 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 assess variability, subsampling was performed 5 times for each sample and forr each read abundance.

Classification of reads against the NCBI nt database downloaded on May 2018 was performed using Kraken 2 version 2.0.6-beta[36] to estimate species abundance and Shannon diversity index. A simplified representation of species composition was obtained using Krona version 2.6[37].

Chao1[38] species richness and Shannon's diversity[39] were estimated using the R package vegan version 2.4.2[40].

Assembly of the metagenome was performed using megahit version 1.1.2[41]. Completeness of the assembly was assessed using BUSCO version 3.0.2[42]. 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. BUSCO analysis was performed on prokaryotic database for all the samples with the exception of M1 (mostly composed by fungi) for which the fungal database was used. Samples B1 and B2 were also compared against the eukaryotic BUSCO database; results for the prokaryotic database are reported.

Unless otherwise specified, all the analysis were performed using R 3.3.3[43].

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

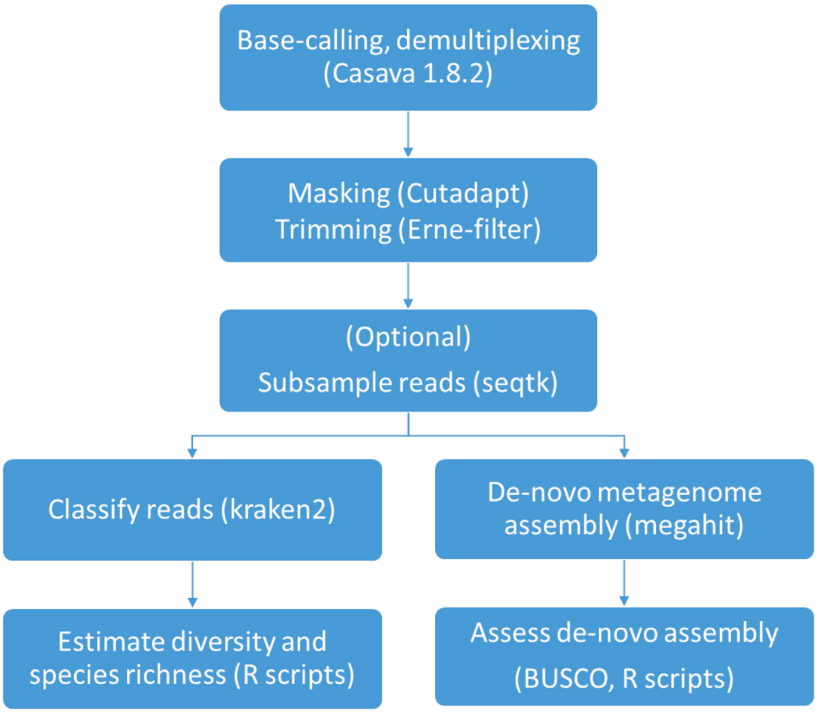

Figure 1. Workflow of the main bioinformatics analysis performed in the present work.

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

Sample	N reads	N species	Singletons
B1	11,031,061	2508	1299
B2	3,830,083	4598	1795
F1	12,472,553	29661	14750
F2	10,780,450	25608	12374
M1	1,898,011	3207	1469
M2	1,558,975	9638	3377
M3	1,867,879	5567	1999

N species, number of species identified in the sample including species identified by one or more reads; Singletons, number of species identified in the sample by only one read.

(sample M2) to more than 12 million (sample F1). Our subsampling, ranging from 10,000 to 1,000,000 reads, led to a reduction in size of 36% (1,000,000 out of 1,558,975) in M2 to 0.08% of the original size (10,000 reads out of 12,472,553) in F1.

Samples used in this study had different levels of species composition (Figure 2). Some samples, such as M1, B1 and B2 were dominated by a single species, while others, in particular fecal samples, showed high heterogeneity in species composition.

Diversity and species richness

Figure 3 shows the variation of the value of Chao1 estimator, representing the estimated number of species in each sample when varying the number of reads used for the estimation, from the smallest number on the left, to the full dataset on the right. The value of Chao1 estimator for the full dataset is plotted on the right side of the plot, at the rightmost fecal samples F1 and F2 had an estimated number of species greater than 40,000, much higher than all the other samples, for which less than 20,000 species were estimated (less than 10,000 B1, B2, M1 and M3).

The effect of downsampling on the estimated number of species has different effects 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, which were characterized by a high number of overall species and rare species, the downsampling led to a significant reduction in the estimated species richness.

Shannon's diversity index is a widely used method to assess the biological diversity of ecological and microbiological communities. Figure 4 depicts the effect of subsampling on the Shannon's diversity index. The effect of subsampling on Shannon's diversity index is smaller than the effect on the estimated number of species. The variation in Shannon diversity index with subsampling is negligible for all samples, even reducing the number of reads from the full size to 100,000 or less.

Figure 5 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 is >0.99 in all the replicates. The plot is in log-log scale to emphasize differences in low abundance species. Only species with frequencies lower than 0.01% (i.e. species represented in 1 read out of 10,000) show some effect of subsampling on the relative abundance estimation. All the seven samples share a similar behavior.

In Figure 6 we show the results obtained by reducing the number of sampled reads to 10,000 reads per sample. Similar to what we observed for larger subsamples, the linear correlation coefficient between species abundance estimate in the full and the reduced dataset was high in all the samples ($r>0.95$) and in all the replicated subsampling. The abundance of species with frequency greater than 1/1000 (0.1%) is correctly estimated in the subsamples, while for rare species the estimate is not precise. Species with frequencies <0.01% are by definition absent in the subsample obtained with 10,000 reads, and were arbitrarily set to a frequency of 0.001% to provide the reader with an idea on their abundance and distribution in the original sample.

Metagenome reconstruction

While characterizing and measuring species present in a complex matrix is an important task, some studies aim at reconstructing (partially or entirely) the metagenome via a *de novo* approach. We thus investigated the effect of coverage reduction on this task. We reconstructed *de novo* the metagenome of the full and reduced datasets, and compared the reconstructed genomes. Results are summarized in Figure 7. As expected, the size of the assembly is strongly influenced by the read number. Assemblies obtained using the full set of reads had a size ranging from slightly more than 1 Mb (sample B1) to nearly 100 Mb (F1 and F2). A decrease in the number of reads used for the assembly lead to a steady decrease in assembly size in all samples, although with different slopes. Assembly sizes obtained using 1,000,000 reads ranged from less than 1 Mb (F1 and F2) to slightly more than 10 Mb (M1), and those obtained using 100,000 reads ranged from less than 100 Kb (F1 and F2) to less than 1 Mb (all the remaining samples).

However, the total assembly length is not necessarily a sufficient measure to describe assembly goodness and completeness[42,44]. Since we are interested in assessing the completeness of the reconstructed metagenome, we used BUSCO to report the proportion of genes covered by any given assembly[42]. Figure 8 reports the proportion of metagenome completeness estimated by BUSCO from full and from the reduced dataset obtained by randomly sampling 1,000,000 reads. The prokaryotic BUSCO dataset was used for all samples with the exception of sample M1, composed prevalently by a mushroom, for which the fungal BUSCO database was used. The full samples F1 and F2 reconstructed a fairly complete proportion of the BUSCO genes (>90%), while the reduced dataset reconstructed less than 20%.

Similar trends can be observed with other datasets. Given the lower number of reads sequenced in other samples, the performance in reconstructing the BUSCO genes was generally poor, but reducing to 1 million reads led to a further decrease in performance, suggesting that this is a clearly suboptimal number of reads. Samples B1 and B2 show a very poor performance

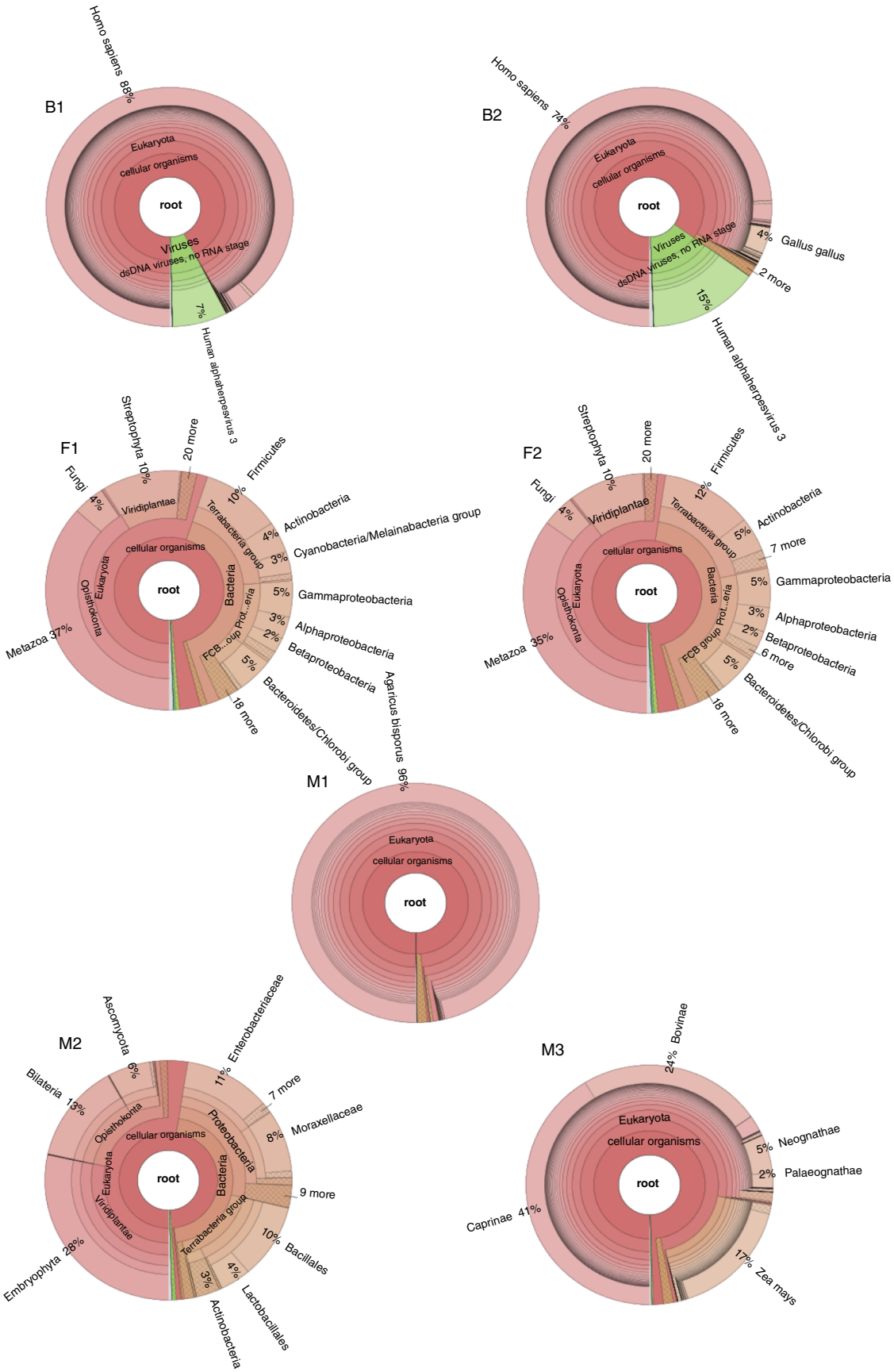

Figure 2. Graphical representation of the composition of the seven studied samples.

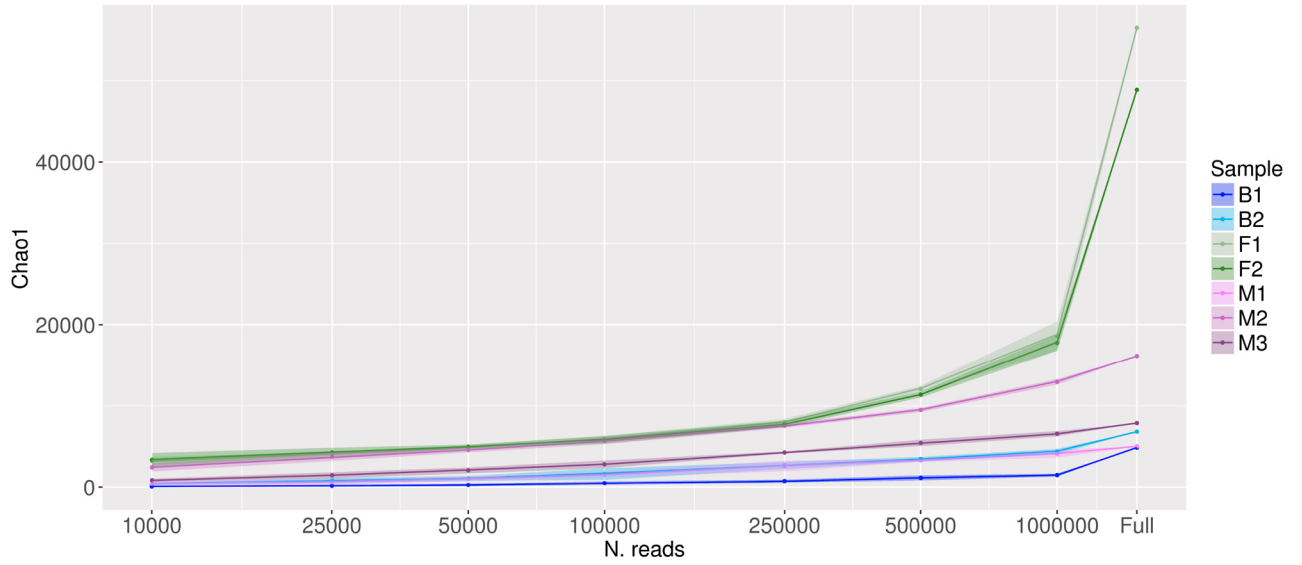

Figure 3. Effect of decreasing the number of reads on Chao1 diversity estimate. X axis is in log scale, Y axis is in linear scale. Shaded areas represent the confidence limits of resampling experiments. "Full" represents the values obtained with the full set of reads (number of reads per sample listed in column 2 of Table 1).

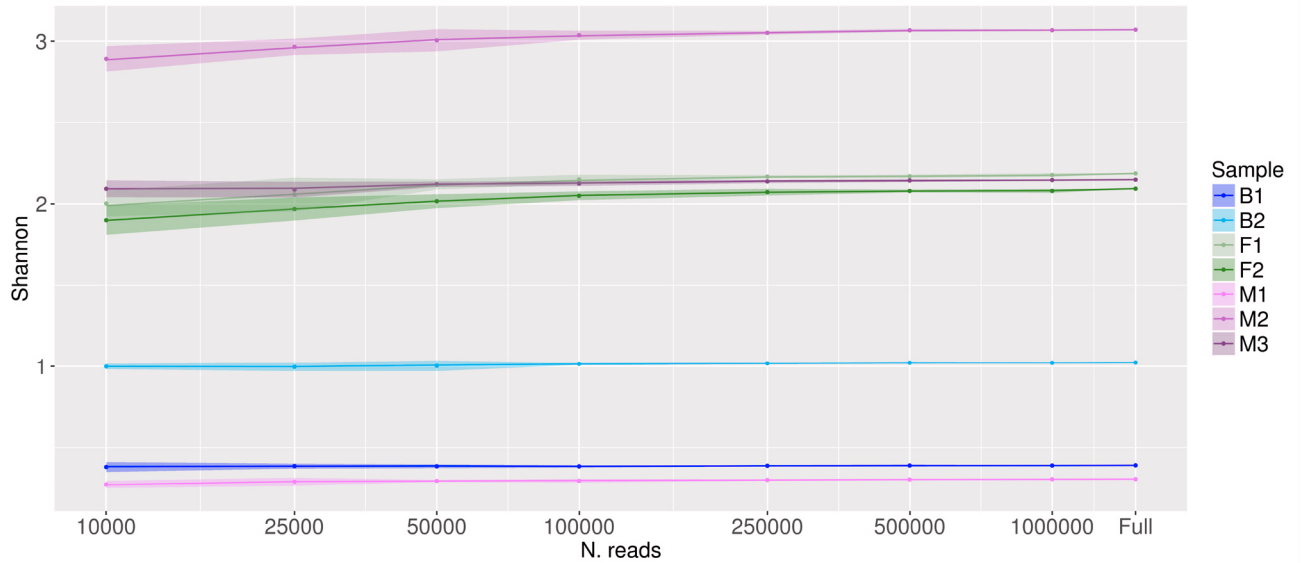

Figure 4. Effect of decreasing the number of reads on Shannon diversity estimate. X axis is in log scale, Y axis is in linear scale. Shaded areas represent the confidence limits of resampling experiments. "Full" represents the values obtained with the full set of reads (number of reads per sample listed in column 2 of Table 1).

because the prokaryotic organisms in the sample are very rare contaminants. Being derived from fetal human cell cultures, a large portion of the metagenome is constituted by human sequences, but given the very small ability in reconstructing *de novo* a genome as large as the human one, the proportion of reconstructed BUSCO genes is very low (<5% both for prokaryotic and eukaryotic BUSCO genes).

Discussion

The aim of the present work was to assess the reliability of low-depth shotgun metagenome sequencing for the characterization of complex matrices, as follows: 1) determining diversity and species richness in complex matrices; 2) estimating abundance of the species present in the complex matrix, and 3) reconstructing *de novo* the genome of the species present in the samples.

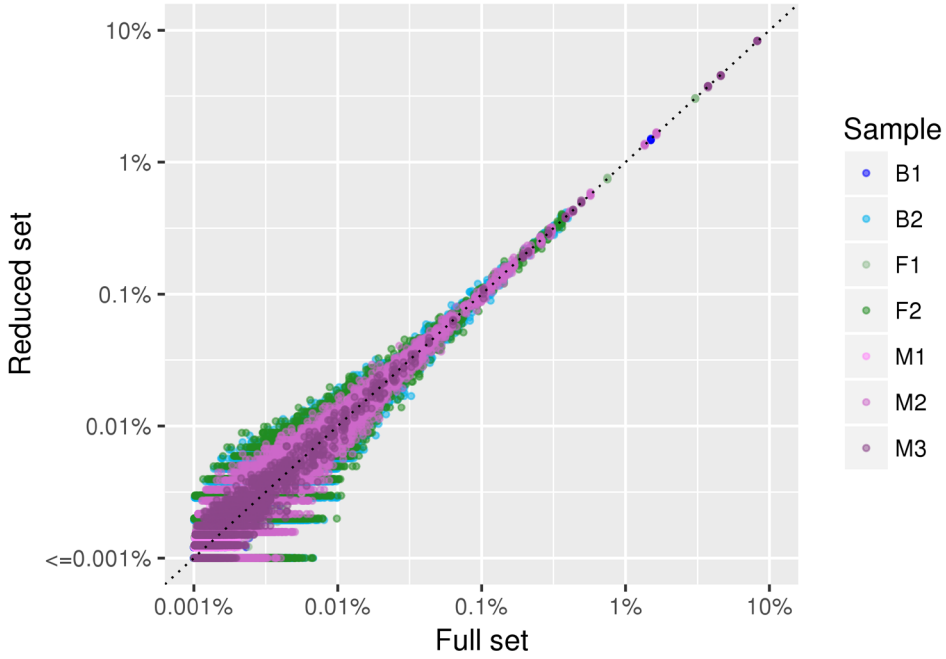

Figure 5. Scatterplot of species abundance estimated using the full set of reads and a set composed of 100,000 reads. Data for all the five replicates of the subsampling are plotted. Each point (colored by sample of origin) represents a given species. The position on the X axis represents the relative abundance of the species in the full dataset, and the position on the Y axis represents the relative abundance of the species in the samples obtained by randomly sampling 100,000 reads. Both axis are plotted in log scale to facilitate visualization of low abundance species.

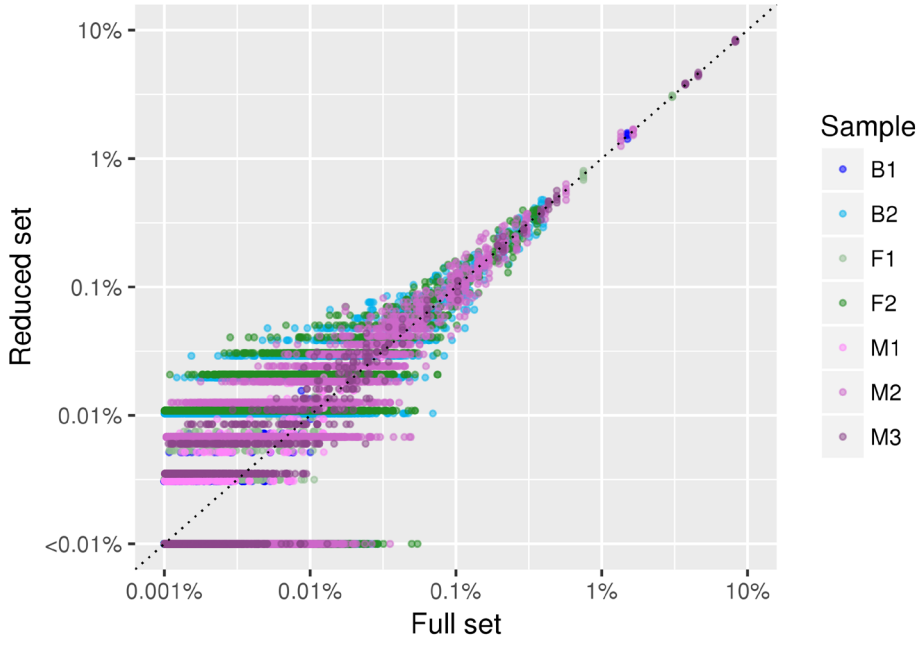

Figure 6. Scatterplot of species abundance estimated using the full dataset of reads and a dataset composed of 10,000 reads. Data for all the five replicates of the subsampling are plotted. Each point (colored by sample of origin) represents a given species. The position on the X axis represents the relative abundance of the species in the full dataset, and the position on the Y axis represents the relative abundance of the species in the samples obtained by randomly sampling 10,000 reads. Both axis are plotted in log scale to facilitate visualization of low abundance species.

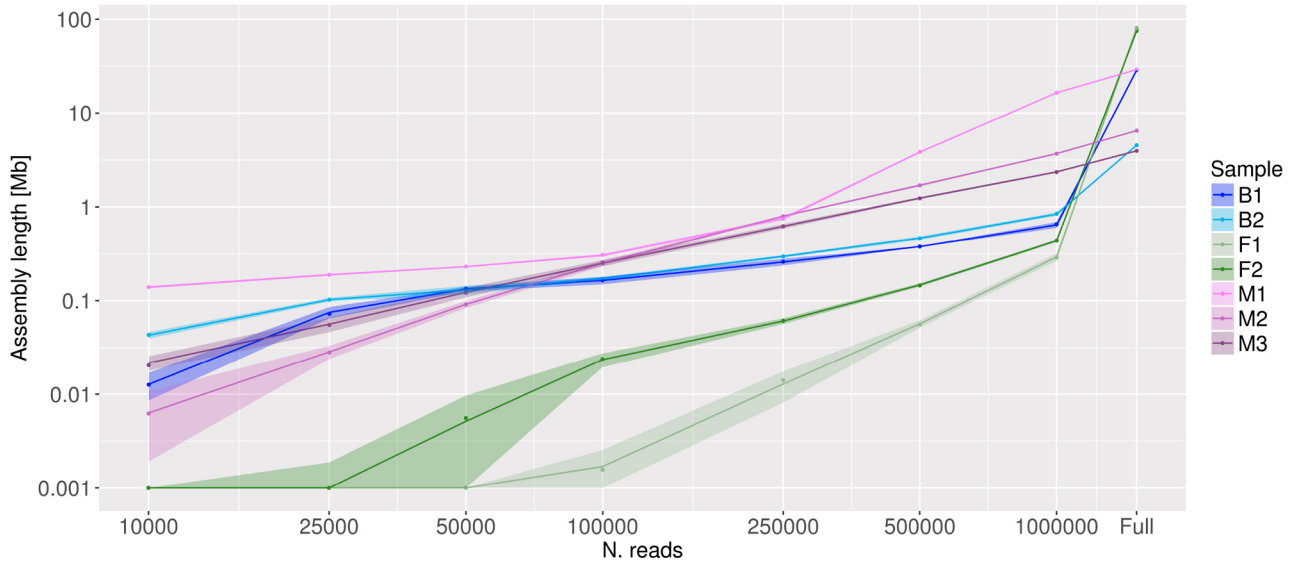

Figure 7. Total length of the *de novo* metagenome assembly in each sample as a function of the number of reads. X and Y axes are in log scale. Shaded areas represent the confidence limits of resampling experiments. "Full" represents the values obtained with the full set of reads (number of reads per sample listed in column 2 of Table 1).

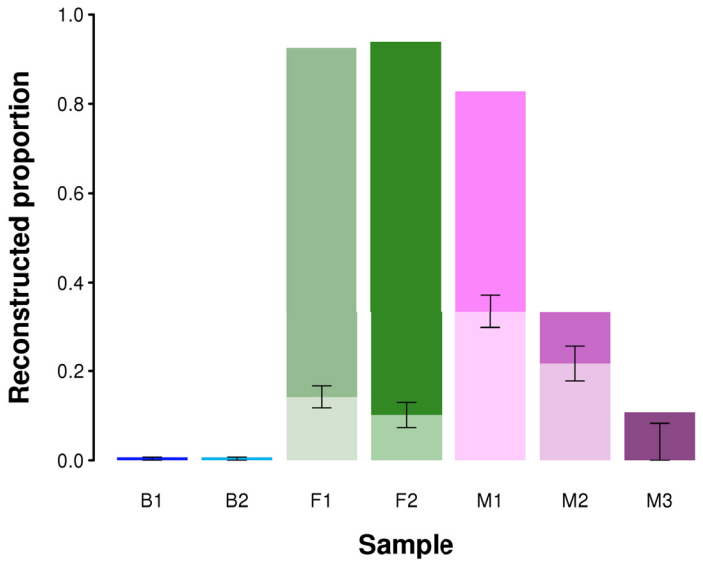

Figure 8. Completeness of the BUSCO genes in the full dataset (darker colors) and in the largest of the reduced datasets (lighter colors). Error bars are based on the five replicate experiments performed for each sample.

We selected seven heterogeneous complex samples, sequenced at varying coverage (ranging 1 to 12 million reads). Shotgun metagenomics experiments—often aiming at reconstructing *de novo* the studied metagenome—have a tendency to generate a very high number of reads per sample[10]. Compared to such studies, all of our samples have relatively shallow coverage of the metagenome, and we tested if even lower coverage could still provide reliable answers to the three main questions listed above.

We used Chao1 as an indicator of species richness and Shannon's diversity index as an estimator of species diversity, and we measured their variation when reducing the number of reads used for the experiment.

An important detail to be considered here is the fact that the two indices behave differently in the full and the reduced samples. We provide an explanation regarding the reasons of this difference.

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.

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, the Chao1 index is heavily affected by the number of rare species that are identified and not from the relative frequencies of the most abundant species, while the Shannon diversity index is affected more by variation in the frequencies of highly abundant species than by the disappearance of rare species.

Samples F1 and F2 are characterized by a very large number of observed species (29,661 and 25,608, respectively), while all the other samples have lower number of species, ranging from 2508 in B1 to 9638 in M2. Chao1 captures this differences, showing that F1 and F2 have greater diversity estimates. The Shannon diversity index, on the contrary, relies not only on the number of observed species, but also on the frequency distribution, and for a given number of species reaches its maximum for equifrequent species. Therefore, samples that have a relatively high number of common species with comparable frequencies tend to have high Shannon's diversity indices.

As an example the number of species with a frequency greater than 0.1% was 23 in sample F1 and was 55 in sample M2. Thus, in spite of a much lower number of species in M2 compared to F1, the Shannon diversity is higher in M2 than in F1. Given the differences in behavior between the two indices in certain conditions, we decided to use both of them to have a more complete information on sample diversity when decreasing coverage. Our results show that a substantial reduction of coverage can be safely achieved without compromising the ability of estimating species richness and abundance (Figure 3 and Figure 4), although the estimated number of species is moderately affected by coverage reduction.

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 reducing sequencing to 100,000 reads, species abundances as low as 0.01% can be reliably estimated.

The last questions to which we sought to answer is if a reduction in the sequencing coverage would have a deleterious effect on the ability of *de novo* assembling the metagenome. Our results show that downsampling had a strongly negative effect on the total length of the reconstructed metagenome and on the proportion of BUSCO genes reconstructed with the metagenome assembly.

BUSCO is widely used for assessing the completeness of genome and transcriptome assemblies for individual organisms, and has benchmark datasets for several lineages. It is possible that using BUSCO for assessing completeness of a metagenomics assembly, including both eukaryotic and prokaryotic organisms, results in an underestimation of the completeness of the reconstruction. However, the aim of the present work is not the absolute estimation of the completeness of the metagenomics assembly, but rather the relative variation observed when using a subsample of reads. Our results indicate that even using 1,000,000 reads is clearly suboptimal 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 also affect research aimed at reconstruction of an interesting part of the meta-genome, such as genes involved in antibiotic resistance[24]. The decrease in performance observed in the reconstruction of BUSCO genes will be likely observed for the reconstruction of other gene categories. 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 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 with the aim of monitoring, diagnosis, surveillance, quality control and traceability.

We show that, if the aim of the experiment is a taxonomical characterization of the sample or the identification and quantification of species present in it, then 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, then a higher number of reads is required. We do not provide here a suggestion on the number of reads that are needed when the aim is the (partial) reconstruction of the meta-genome, as it depends on several factors (number of species in the sample, their genome size, and their abundance, length of the sequencing reads, quality of the DNA) and this estimation needs to be performed for each experiment based on detailed understanding of the experiment aims and of sample characteristics.

Data availability

Raw reads are available at NCBI Sequence Read Archive. 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.

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.

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. Edwards RA, Rohwer F: **Viral metagenomics.** *Nat Rev Microbiol.* 2005; **3**(6): 504–10.
PubMed Abstract | Publisher Full Text

4. Sahoo MK, Holubar M, Huang C, *et al.*: **Detection of Emerging Vaccine-Related Polioviruses by Deep Sequencing.** McAdam AJ, editor. *J Clin Microbiol.* 2017; **55**(7): 2162–71.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Martí JM: **Robust Analysis of Time Series in Virome Metagenomics.** *Methods Mol Biol.* 2018; **1838**: 245–60.
PubMed Abstract | Publisher Full Text

6. Richards B, Cao S, Plavsic M, *et al.*: **Detection of adventitious agents using next-generation sequencing.** *PDA J Pharm Sci Technol.* 2014; **68**(6): 651–60.
PubMed Abstract | Publisher Full Text

7. Petricciani J, Sheets R, Griffiths E, *et al.*: **Adventitious agents in viral vaccines: lessons learned from 4 case studies.** *Biologicals.* 2014; **42**(5): 223–36.
PubMed Abstract | Publisher Full Text

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

9. 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

10. 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

11. 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

12. 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

13. 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

14. 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

15. 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

16. 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

17. 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

18. 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

19. 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

20. 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

21. 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

22. Victoria JG, Wang C, Jones MS, *et al.*: **Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus.** *J Virol.* 2010; **84**(12): 6033–40.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Denman SE, Morgavi DP, McSweeney CS: **Review: The application of omics to rumen microbiota function.** *Animal.* 2018; 1–13.
PubMed Abstract | Publisher Full Text

24. 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

25. Staats M, Arulandhu AJ, Gravendeel B, *et al.*: **Advances in DNA metabarcoding for food and wildlife forensic species identification.** *Anal Bioanal Chem.* Springer Berlin Heidelberg; 2016; **408**(17): 4615–30.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Yamamoto S, Masuda R, Sato Y, *et al.*: **Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea.** *Sci Rep.* Nature Publishing Group; 2017; **7**(1): 40368.
PubMed Abstract | Publisher Full Text | Free Full Text

27. 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

28. 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

29. 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

30. 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

31. 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

32. 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

33. 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

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

35. 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

36. 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

37. 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

38. 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

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

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

41. 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

42. 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

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

44. Vezzi F, Narzisi G, Mishra B: **Feature-by-feature--evaluating *de novo* sequence assembly**. Rzhetsky A, editor. *PLoS One.* Public Library of Science; 2012; **7**(2): e31002.
PubMed Abstract | Publisher Full Text | Free Full Text