Analyzing the differences between reads and contigs when performing a taxonomic assignment comparison in metagenomics

Pablo Rodríguez-Brazzarola, Esteban Pérez-Wohlfeil, Sergio Díaz-del-Pino, Ricardo Holthausen, and Oswaldo Trelles

University of Malaga, Campus de Teatinos, 29071, Malaga, Spain
{pabrod,estebanpw,sergiodiazdp,
ricardoholthausen,ortrelles}@uma.es
http://www.bitlab.es

Abstract. Metagenomics is an inherently complex field in which one of the primary goals is to determine the compositional organisms present in an environmental sample. Thereby, diverse tools have been developed that are based on the similarity search results obtained from comparing a set of sequences against a database. However, to achieve this goal there still are affairs to solve such as dealing with genomic variants and detecting repeated sequences that could belong to different species in a mixture of uneven and unknown representation of organisms in a sample. Hence, the question of whether analyzing a sample with reads provides further understanding of the metagenome than with contigs arises. The assembly yields larger genomic fragments but bears the risk of producing chimeric contigs. On the other hand, reads are shorter and therefore their statistical significance is harder to asses, but there is a larger number of them. Consequently, we have developed a workflow to assess and compare the quality of each of these alternatives. Synthetic read datasets beloging to previously identified organisms are generated in order to validate the results. Afterwards, we assemble these into a set of contigs and perform a taxonomic analysis on both datasets. The tools we have developed demonstrate that analyzing with reads provide a more trustworthy representation of the species in a sample than contigs especially in cases that present a high genomic variability.

Keywords: taxonomic assignment, sequencing analysis, metagenomic comparison

1 Introduction

A drastical reduction of time and cost per sequencing experiment has taken place, dropping from a 10,000$ per megabase down to a few cents, due to the major breakthroughs in sequencing technologies that have occurred in the last decades [1]. These techniques produce a huge amount of data overcoming the

data generation problem, which was the main barrier during the early Genomic Era. Biologists are now facing a torrent of data which have paved the way towards the analysis of numerous unknown biological communities and the research of pioneering scientific areas such as metagenomics (beyond genomes).

The goal of metagenomics is to study microbial communities, also known as microbiotas, in their natural environment, without requiring to aisle and cultivate the species that make up such community. This field brings a profound transformation in multiple fields, such as: biology, medicine, ecology, agriculture, and biotechnology [2]. Despite these benefits, metagenomic sequence data presents several challenges. For instance, most communities are so diverse that most genomes are not utterly represented by reads. The difficulty of performing direct comparisons through sequence alignment is even greater due to distinct reads from the same gene that may not overlap. However, when they do overlap it is not always noticeable whether they are from the same or different genomes, challenging the sequence assembly. Additionally, its informatic analysis is more complicated when dealing with poor quality reads, detecting repeated sequences from similar organisms, and genomic variants or species that have not yet been sequenced within a sample in which the representation of organisms is uneven and unidentified [3].

A primary objective in metagenomics is portray the organisms present in an environmental sample. A correct classification of the species within a sample with enable a further insight about several issues such as: the microbial ecosystems models used to describe and predict community-based microbial processes, changes, and sustainability; the global scale descriptions of the role of the human microbiome in different health states in individuals and populations; and the exploitation of the remarkably versatile and diverse biosynthetic capacities of microbial communities to generate beneficial industrial, health, and food products.

Tools such as MEGAN [4], FANTOM [5] or RAST [6] perform a taxonomic analysis with reads and are also prepared to work with contigs, since each approach has advantages and disadvantages. Analyzing contigs provide larger genomic fragments, nevertheless this entails a risk of generating chimeric contigs due to the heterogeneity of the sample. On the other hand, with reads this risk is non-existent, however the analysis is affected by several factors such as the quality and length of the sequences, thus may generate matches with low statistical significance.

The main contributions of this paper are a set of tools that analyze the quality of the taxa assigned to the metagenomic sample and establishes statistical differences between reads and contigs in order to provide a better judgement to properly identify the correct taxa distribution in a metagenomic sample. It also provides a workflow that employs the previous tools to propose suggestions on how to perform an optimal taxonomic analysis of a metagenomic sample, either with reads or with contigs.

2 Methods

The definitions, procedures and algorithms employed to compare reads and contigs when analyzing a metagenomic sample are described in this section. First, we defined a set of conditions that describe the taxonomic concordance between the contig (handles as once sequence) and the reads that assemble it on a specific taxonomic rank in order to achieve a reasonable comparison.

- **General Definitions:**
  Let S be the set composed by the Reads and Contigs nucleotide sequences
  Let T be the set composed by the Taxa in a specific taxonomic rank and None

$$s \in S \ \wedge \ t \in T$$
$$Taxon(s) \rightarrow t$$

- **Consistency (C):** Both, the read and the contig, have the same taxon assigned or were not assigned at all.

$$Taxon(Read) = Taxon(Contig)$$

- **Weak Inconsistency (WI):** One of the sequences has been assigned to a taxon, but the other one was not assigned to any. These relationships are classified based on which sequence was unassigned. Granted that the read does not match a taxon in the specified taxonomy rank, it will be defined as a Weak Inconsistency by Read (WIR). But if the unassigned sequence was the contig then it will be classified as a Weak Inconsistency by Contig (WIC).

$$Taxon(Read) = None \ \wedge \ Taxon(Contig) = x \ \wedge \ x \neq None$$
or
$$Taxon(Read) = X \ \wedge \ Taxon(Contig) = None \ \wedge \ x \neq None$$

- **Strong Inconsistency (SI):** Both, the read and the contig, are assigned to a taxon in the selected taxonomic rank, but to different taxa. Note that if either the read or the contig is not assigned, it will be classified as a WI.

$$Taxon(Read) \neq Taxon(Contig)$$

Having settled the previous definitions, a workflow has been designed with the intent of analyzing the levels of concordance between reads and the contigs they assemble and retrieve reliable comparison results (see Supplementary Material - Figure 1). Such workflow begins with metagenomic reads and a reference database as input. Firstly, the reads are assembled into contigs using MEGAHIT [7], an assembler developed for large and complex metagenomic NGS (Next Generation Sequencing) reads. Afterwards, both sets of sequences are mapped against a reference database to acquire the possible species that each sequence

came from. These results are fed to MEGAN, a microbiome analysis tool that uses the last common ancestor (LCA) algorithm to assign each sequence to a taxa. Finally, the retrieved information from previous steps is processed by our developed software in order to generate a set of results that assesses the quality of the taxa assigned to the reads and contigs and provides statistical insight about these results. The subsequent section provides a detailed description of the internal functioning of the workflow.

2.1 Detecting Differences between the Reads and Contigs

The developed toolkit has been designed for comparing the results obtained after performing a primary sequence comparison and a biological taxonomic analysis between reads and contigs. The output information provided by this tool is composed by:

1. The associations for each contigs and the reads that assembled it.
2. Concordance of the taxa assigned between the reads and contig assembled.
3. Coverage of the reference database.
4. Ratios of highest scoring matching species per sequence in a metagenomic dataset.
5. Correct taxonomic classification percentage.

The internal procedure implemented to obtain these results are described in the following section.

- **The associations for each contigs and the reads which assembled it:** The associations between the reads and contigs are extracted from the BLASTN [8] output obtained by performing a DNA primary sequence alignment between them. Other comparison tools can be used by adding an specific parser. This result is processed to obtain two collections of the relationships between the reads and the contigs: one in which the reads are assigned to the contig that it assembled; and the other where the contigs are partnered with the group of reads used to assemble it.
- **Concordance of the taxa assigned between the contigs and the reads that assembled it:** Firstly, the identifier of all the sequences that have been assigned to a taxon in the selected biological classification rank are extracted from the MEGAN results. Afterwards, this information is used to classify the previously obtained associations between reads and contigs, based on the concordance level of the taxon assigned to a contig and the reads that assembles it.
- **Coverage of the reference database:** The amount of base pairs that were aligned to the database obtained from the results after executing the BLASTN with each set of sequences is compared to the number of base pairs in such database to obtain the following results:
  - Total coverage of the database for each set of sequences
  - Total coverage of the database that the reads and contigs match together

- **Ratios of highest scoring matching species per sequence in a metagenomic dataset:** The average of top scoring matches resulting from the sequence alignment against the reference database is calculated for each of the datasets. Afterwards, the measurement obtained from each dataset is compared to decide which one provides less variable matches.
- **Correct taxonomic classification:** This measurement can only be calculated when the original taxon for each read is previously known. The percentage of sequences assigned to the taxon to which they belong is calculated for both datasets. An assignment is correct for a read if such is matched to the correct taxon. However, it is impossible to know the proper specie for a contig because they can be assembled from reads that belong to different organisms. Therefore, we define the assigned taxon to a contig as the one to which the majority of the reads that assemble it belong to.

3 Results and Discussion

In order to apply the previously described workflow and to obtain valid comparison results, the metagenomic reads dataset has to be properly designed, meaning that the real taxon for each read must be known beforehand for the purpose of enabling us to assess and to establish whether it is better to perform a taxonomic analysis with reads or with contigs.

To achieve this goal, two use cases have been developed to achieve authentic results that fulfill the requirement of knowing the original specie for each read. One fully synthetic where the reads are originated from each genome in the database. The other one is semi-synthetic, where the reads come from a selection of genomes that are representative of the classes in a real metagenomic sample (See Figure 1).

The differences between the cases are the initial dataset of reads and the reference database. These metagenomic reads and the databases employed are obtained through the following approaches:

- **Fully synthetic use-case/dataset (FSD):** The database selected are the gastrointestinal tract genomes provided by Human Microbiome Project (HMP)[10] and the reads are obtained by executing Better Emulation for Artificial Reads (BEAR)[11] with the HMP dataset. An equitably number of reads are generated from each genome in the reference database to obtain a very mixed sample of reads from all the different species to which they will be compared. The total number of reads is 521,334.
- **Semi-synthetic use-case/dataset (SSD):** Following the class taxonomic distribution from the study Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients[9], a set of genomes were selected that belonged to each of the classes. These samples were used with BEAR to generate a set of reads proportional to the class distribution obtained by analyzing the article. The remaining percentage of the metagenomic sample (9%) was obtained by generating a

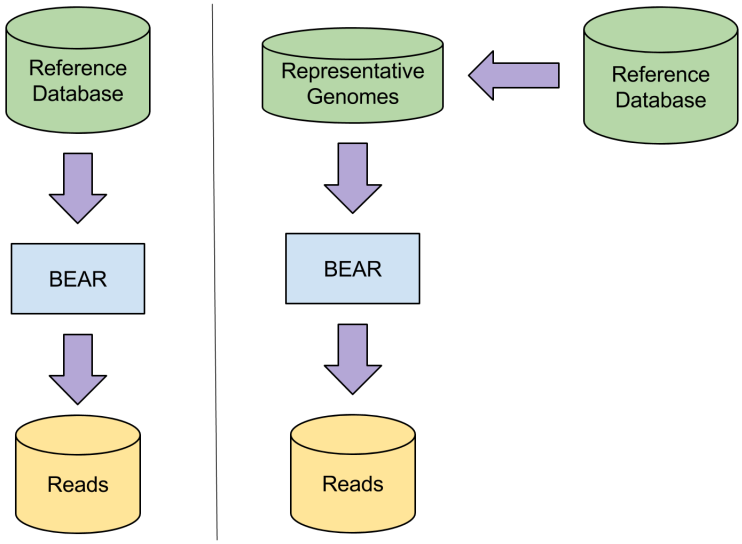

Fig. 1. On the left: Generation of fully synthetic reads. On the right: Generation of semi synthetic reads. BEAR (referenced in the next paragraph)

set of random reads that followed the nucleotide distribution from the rest of the dataset. In order to provide a soil sequencing framework, the reference database for the soil microbial genomes selected is RefSoil[12]. The total number of reads is 499,991.

The species distribution for both metagenomic datasets is represented in the Figure 2.

The workflow depicted in the Methods section has been applied to each of the use cases with the parameters described in the Table 1 of the Supplementary Material. Afterwards, the output from the developed tools for each use case is interpreted to obtain the following results:

- **Comparison with the Original Distribution:** The species distribution obtained by performing a taxonomical analysis with the reads and contigs is compared with the original dataset in Figure 3.
  For the FSD both reads and contigs seem to have differences against the original dataset, however it is not noticeable which one is more similar to the authentic dataset. This is not the case for the SSD since the reads present an almost identical distribution of species in comparison to the original, while the contigs clearly have noticeable differences. This is further verified in the next section.
- **Root Mean Square Error (RMSE) after the Taxonomical Analysis:** The RMSE is calculated for both reads and contigs using the original

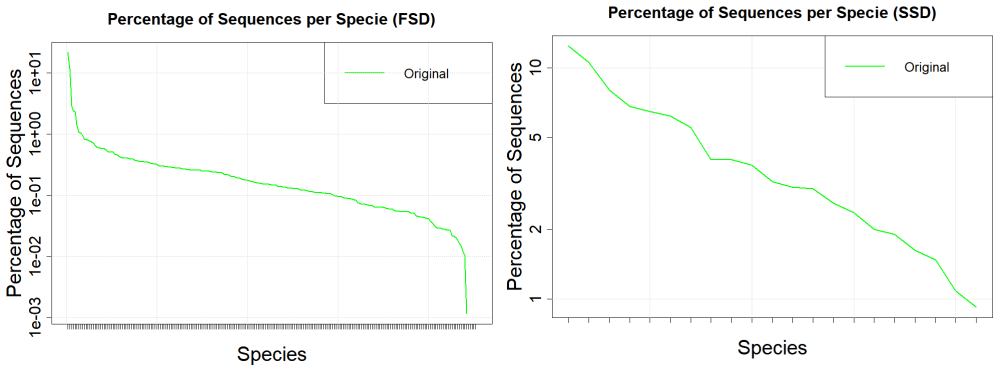

Fig. 2. Percentage distribution of the species in a metagenomic original dataset. On the left: Distribution for the fully synthetic dataset. On the right: Distribution for the semi synthetic dataset.

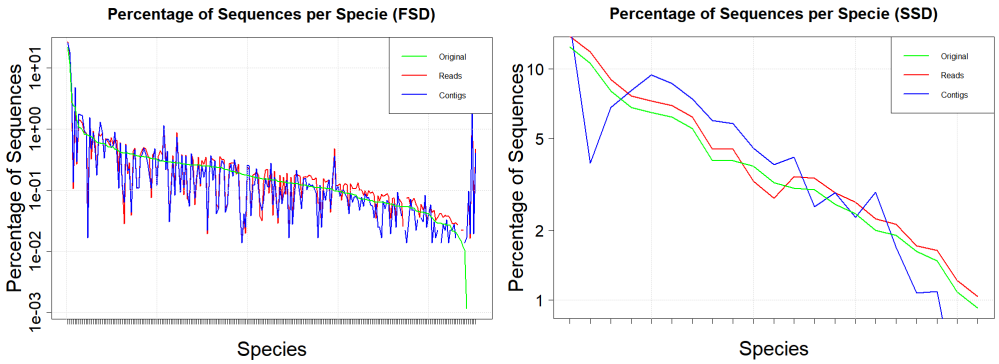

Fig. 3. Percentage distribution of the species in a metagenomic original dataset (green), reads (red) and contigs (blue). On the left: Distribution for the fully synthetic dataset. On the right: Distribution for the semi synthetic dataset.

dataset as reference. This implies that if the RMSE is lower for a set of sequences (reads or contigs), the mapping of this dataset is more appropriate to describe the ideal distribution of species in the metagenome (Table 1)

Table 1. Root of the Mean Squared Error of the assignment of species for reads and contigs compared to the original dataset for both datasets.

Dataset	RMSE for FSD	RMSE for SSD
Reads	0.3187	0.4031
Contigs	0.3858	4.2534

The RMSE describes the average difference for each dataset in comparison to the original, and provides further insight about how correct are the assignments of species per dataset. Reaffirming the results from the Figure 4, it is observed that in both use cases the reads have a lower RMSE than the contigs.

- **Inconsistencies Found:** A concordance level is established to each of the associations between a contig and the read that it assembles. Identifying the types of inconsistencies aids us at the moment of determining the reason behind the RMSD. If there are more weak inconsistencies at the species taxonomic rank, then most of the reads or contigs involved were assigned to a taxon in a higher and less specific taxonomic rank. The detected inconsistencies and the percentage of relationships they represent are shown in the Table 2.

Table 2. Number of inconsistencies assigned for each use case.

Type of Inconsistency	Found on FSD (%)	Found on SSD (%)
Weak Inconsistency by Read	21,393 (4.10)	4,003 (0.80)
Weak Inconsistency by Contig	24,183 (4.64)	1,622 (0.32)
Hard Inconsistency	4,464 (0.84)	2,231 (0.45)

- **Inconsistency Resolution:** Inconsistencies can be solved by selecting a less specific taxonomic group to cover a broader range of taxa that a sequence could be assigned. In both use cases the sequences belong to bacterias, therefore the discrepancy between the assignment a contig and the read that assembles it will be sorted out in the taxonomic group Domain. This can be appreciated in Figure 4.

  The heterogeneity of the samples provoke that a noticeable amount of contigs are assembled from reads from different species. This confirms that the inconsistencies arise during the assembly process.
- **Coverage and Mapping Comparison against the Reference Database:**

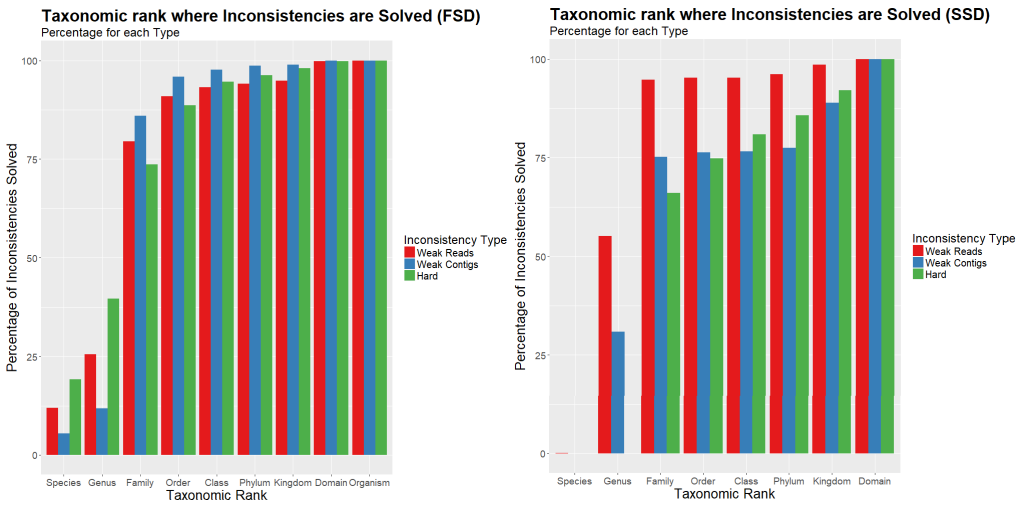

Fig. 4. Percentage of inconsistencies solved at different taxonomic ranks. In both use cases, over 50% of the inconsistencies are resolved if the desired taxonomic group to analyze is the family. On the left: inconsistency resolution for the fully synthetic dataset. On the right: inconsistency resolution for the semi synthetic dataset.

For each use case, the ratio of top scoring matches after performing the sequence alignment to the reference database and the percentage of nucleotides covered by the full set of sequences is depicted in the Table 3.

Table 3. Mapping and coverage comparison between reads and contigs for each use case

Measurement	**FSD**		**SSD**	
	Reads	Contigs	Reads	Contigs
Ratio of Matches per Sequence	7.05	7.50	4.52	8.38
% Coverage of Database	21.21	7.16	5.59	3.37
Common % within Use Case		6.42		3.03
Common Coverage % against Contigs	89.66	Not Applicable	89.91	Not Applicable

For both use cases, reads have a lower average of top scoring matches since the contigs tend to have more matches due to assembly noise generated by forming contigs from reads belonging to different species. It is also noteworthy that over 85% of the nucleotides covered by contigs are also covered by reads, yet reads cover a wider range of the database meaning that they provide more information that may be of interest depending on the interests of the metagenome experiment.

– **Correct Assignment of a Taxon for each Sequence Comparison:** Both use cases fulfil the prerequisite to calculate this measurement, to know

beforehand the original taxon of a read. The resulting assessment is described in the Table 4.

Table 4. Correct assignment for each sequence comparison

Measurement	**FSD**		**SSD**	
	Reads	Contigs	Reads	Contigs
Properly Assigned	493,226	59,088	454,900	56,242
Wrongly Assigned	3,179,682	611,595	1,933,138	387,162
Total	3,672,908	670,683	2,388,038	443,409
Correct (%)	13	87	19	13
Incorrect (%)	87	91	81	87

The low percentage of properly assigned sequences is caused by the multiple top scoring matches for each sequence previously described. This fact generates a noticeable amount of wrong assignments, but these must be taken into account because in a real metagenomic sample it is impossible to know which is the correct match. This indicates that the reads provide a more accurate assignment in both use cases, whereas contigs provide less sensitive results. However, this also proves that a high percentage of the data obtained from a metagenomic sample is noise originated from various sources. Hence it can be concluded that the assembly is not the only process that needs to be refined in order to obtain more valuable information

4 Conclusions

Even though tools and algorithms in metagenomics have advanced, there are still shortcomings very difficult to solve due to the intrinsic complexity of analyzing a metagenomic sample. Therefore these errors have to be properly addressed to make better tools in the future. Accordingly, the results obtained in this work demonstrate some issues to be resolved in the field of metagenomic assembly.

We have presented in this work several indicators that enable a valid comparison between reads and contigs when performing a taxonomic analysis of a metagenomic sample. We have demonstrated that reads provide a more accurate assignment of taxa, and that the distribution of species resembles in a larger extent the original metagenomic sample distribution, and provide a more specific assignment of taxa than using the contigs.

The measurements established in previous sections suggest that during the assembly process, some reads belonging to different species are put together into a contig as a result of the great heterogeneity of species in a metagenomic sample. In this same stage, another issue arises which is that the distribution of species of contigs assigned less resembles the original since their length will vary depending on how many reads are used to assemble. Yet at the moment of assigning a taxon it will still count as one sequence match even though it

was formed by many of them. Moreover, reads will describe more accurately the proper distribution of species in the metagenomic sample since each belongs to one specie and their length size is uniformly distributed. However, when working with contigs specificity is lost due to the possibility of creating chimeric contigs and the fact that the quality of the assembly will vary strongly on the length and quality of the reads, misrepresenting the original sample.

In terms of future work, the toolkit is being applied to compare the quality of different metagenomic assembly tools and to compare the quality of the assembly using different parameters. Likewise, adjusting the presented workflow to compare the functional analysis between reads and contigs would be very interesting.

5 Acknowledgements

The authors would like to thank Fabiola Carvalho and Ana T. R. Vasconcellos from the LNCC-Brazil for their support. This work has been partially supported by the European project ELIXIR-EXCELERATE (grant no. 676559), the Spanish national projects Plataforma de Recursos Biomoleculares y Bioinformaticos (ISCIII-PT13.0001.0012) and RIRAAF (ISCIIIRD12/0013/0006) and the University of Malaga

References

1. National Human Genome Research Institute, The Cost of Sequencing a Human Genome https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/
2. National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington (DC): National Academies Press (US); 2007. 1, Why Metagenomics? https://www.ncbi.nlm.nih.gov/books/NBK54011/
3. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science, 5, 209. http://doi.org/10.3389/fpls.2014.00209 (2014)
4. Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. MEGAN analysis of metagenomic data. Genome Research, 17(3), 377386. http://doi.org/10.1101/gr.5969107 (2007)
5. Sanli, K., Karlsson, F. H., Nookaew, I. & Nielsen, J. FANTOM: Functional and taxonomic analysis of metagenomes. BMC Bioinformatics 14, 38, doi:10.1186/1471-2105-14-38 (2013).
6. Meyer, F. et al. The metagenomics RAST servera public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386 (2008).
7. Li, D., Liu, C-M., Luo, R., Sadakane, K., & Lam, T-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, doi: 10.1093/bioinformatics/btv033 (2015).
8. Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21097/

9. Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. The ISME Journal, 6(5), 10071017. http://doi.org/10.1038/ismej.2011.159
10. The NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Guyer, M. The NIH Human Microbiome Project. Genome Research, 19(12), 23172323. http://doi.org/10.1101/gr.096651.109 (2009).
11. Johnson, S., Trost, B., Long, J. R., Pittet, V., & Kusalik, A. A better sequence-read simulator program for metagenomics. BMC Bioinformatics, 15(Suppl 9), S14. http://doi.org/10.1186/1471-2105-15-S9-S14 (2014).
12. Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., Howe, A. Strategies to improve reference databases for soil microbiomes. The ISME Journal, 11(4), 829834. http://doi.org/10.1038/ismej.2016.168