Scholars Journal of Agriculture and Veterinary Sciences

OPEN ACCESS

Abbreviated Key Title: Sch J Agric Vet Sci
ISSN 2348–8883 (Print) | ISSN 2348–1854 (Online)
Journal homepage: https://saspublishers.com/sjavs/

Soil Metagegomics: Approaches, Bioinformatics Tools and Applications

Tiziana Maria Sirangelo[1*], Grazia Calabrò[2]

[1]Life Science Department, University of Modena and Reggio Emilia, Italy
[2]Computer Engineering Department, University of Calabria Italy

DOI: 10.36347/sjavs.2020.v07i06.003 | **Received:** 09.06.2020 | **Accepted:** 17.06.2020 | **Published:** 20.06.2020

***Corresponding author:** Tiziana Maria Sirangelo

Abstract

Review Article

Most of soil microbes have not yet been individuated and their metabolisms are unknown. Recent studies indicate that classic plate counting techniques were able to determine only a part of the overall soil microbial community and that the metagenomics approach, based on a culture independent method, allows to study it in depth and to overcome the previous limits. The relevance of this approach is growing as it allows exploring factors affecting the soil fertility, to deepen the interactions between microorganisms and plants, and to identify new molecules with pharmacological activities such as antibiotics. This work investigates the metagenomics approach and the most commonly used sequencing methods, focusing on bioinformatics tools for each of them. Some of the most recent metagenomics soil applications, concerning both amplicon and whole genome shotgun sequencing are described. Therefore, metagenomics, used in combination with other omics approaches, based on metabolomics, transcriptomics and/or proteomics, is discussed and focused as it can provide a comprehensive soil microbiome draw.

Keywords: Metagenomics, Soil, Amplicon sequencing, Shotgun sequencing, Bioinformatics.

INTRODUCTION

Soil can be considered a complex biological system characterized by a remarkable microbial diversity. The complexity of this diversity results from many factors, influencing each other, including pH, water, soil structure, climatic changes and biotic activities.

Prokaryotes and fungi represent the main components of the microbial biodiversity of the soil. Prokaryotes are present with a large number of different taxa and are the most abundant group in many soils. Fungi are also abundant and although they have been studied for centuries, the modern molecular biology techniques show that strategies based on in vitro cultivation have strongly underestimated the total diversity and relevance of soil fungal communities [1]. Animal organisms live in the soil and nematodes are considered among the most relevant, both in terms of species richness and abundance, and are expected to exceed one million individuals per square meter [2]. Nematodes are considered key species in the soil ecosystem being involved in processes such as the decomposition of organic matter and the recycling of nutrients, and many soil nematode taxons represent one of the largest sources of biotic stress for agricultural plants.

The classic plate counting techniques determined only a low percentage of the overall soil microorganisms. Current investigations indicate that more than 99% of the microorganisms living in natural environments are not culturable and therefore not available research activities [3]. Instead, today it is possible to investigate the microbial species living in soil through culture independent methods. The metagenomics approach, based on DNA extraction, and on its purification and sequencing, with or without cloning, allows to determine in depth the microbiological diversity of the soil. This is a significant methodological challenge, for comparison human sequencing covered 3 Gbp, while soil can concern 1000 Gbp of microbial sequences per gram of soil [4]. However, recent developments in sequencing techniques have made this goal possible.

The metagenomics analysis of soil allows to achieve many aims, like to deepen the study of the microbiological factors affecting the soil fertility, to better understand the interactions between microorganisms and plants, to optimally exploit beneficial microorganisms in agriculture, to discover new genes that can allow the bioremediation of polluted soils, and to identify new molecules having pharmacological activities such as antibiotics.

Phases of a metagenomics project

Metagenomics is becoming an increasingly used studying approach and soil metagenomics can now be the subject of many research projects. Therefore, here it may be important to briefly illustrate what the procedures necessary for properly carrying out a metagenomic project are.

The first phase is pre-sequencing, in which the project aims are defined, taking into account the available sequencing power and computational analysis. It consists of the *Experimental Design* of the project and must also include an assessment of the complexity of the microbial community being studied.

The second phase is *Sampling*, during which the utmost effort must be made to preserve DNA quality. The so-called "metadata" and additional soil samples for further possible analyses are also collected at this stage. When we carried out the *DNA Extraction*, we must consider that it should be representative of all cells of the considered sample and that the resulting nucleic acids must be sufficient in quantity for the subsequent project activities. Furthermore, DNA extraction protocols should be chosen coherently with the kind of investigation performed: the large use of commercial kits makes the data obtained from different studies more consistent and comparable.

The *Sequencing* phase is the one in which DNA sequences are produced, according to the technological platform used. The sequencer machine "reads" portions of the DNA called "fragments" which were previously predisposed. The details of the method and the length of the fragments vary according to the technique adopted.

The next step involves the *Assembling* of the sequences obtained in continuous sets which have to be the longest possible called *contig*. The aim is to establish the original order in which the bases were present in the DNA strand. The quality of the results is very relevant in many metagenomics projects. Before processing the sequences, the following activities are necessary: removing adapters from reads, filtering reads, removing any contaminants, identifying and removing any chimera sequences that may have been generated during the previous step, and making data ready to be analysed.

The *Annotation* phase is the one in which we try to give a name to the assembled sequences, by indicating their supposed function. In addition to annotation the soil metagenomics requires a phase in which the sequences are attributed to specific taxonomic groups (*Binning*). The taxonomic assignment is based on the presence of phylogenetic markers and provides an overview of the species that play a primary role in the microbial community.

In the *Downstream analyses* of metagenomic data, the diversity of a given microbiome is typically described in terms of alpha diversity, that is the diversity of a given sample, while the one between different samples as beta diversity. During this step, data diversity has to be properly modelled and visualized.

The flow diagram described above is illustrated in Figure 1.

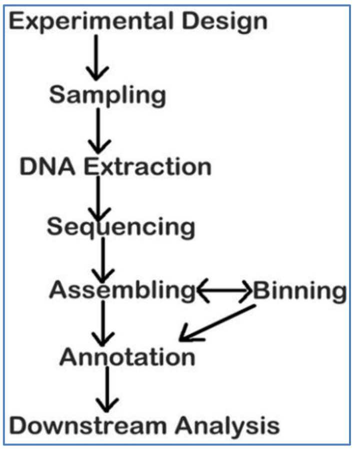

Fig-1: Flow Diagram of a metagenomics project

Sequencing approaches and bioinformatic tools

Multiple approaches to sequencing can be followed. The *Amplicon Sequencing*, the *Whole Genome Shoutgun Sequencing* and the *Cloning Library Sequencing* are described below.

Amplicon Sequencing

Studying a microbiota/microbiome community by using a metagenomic approach based on 16S and 18S rRNA profiling is a technique that has become common practice in the latest years [5, 6].

In the data generation phase 16S rRNA gene amplicon sequences are obtained. DNA extraction and quantification, library preparation and sequencing are the activities performed in this step.

The choice of the primers depends on several factors, such as the compatibility with previous research works and primer specificity. The choice of the region is also critical: in fact, phylogenetic information depends on the length of the 16S rRNA gene [7]. The selected regions are amplified during the library preparation activities.

For the sequencing activity, one of the most popular tools is the MiSeq system, based on Illumina technology [8].

Chimeras, which are generated by incomplete template extension and recombination among sequences, can result into an overestimation of diversity and can be deceiving. Several software options for chimera filtering, including UCHIME [9], and DECIPHER [10] are available. Noise introduced by sequencing errors generally impacts alpha diversity but has little influence on beta diversity analysis.

Quality-filtering parameters can be implemented by using the QIIME system [11], and its scripts are applied on files generated from the previous step. Generally, each sequence is assigned to its sample of origin by using a barcode. Reads that do not match any barcode are discarded.

Processing the reads consists of classifying each read on the basis of the taxa which have the highest probability of being related to this read.

After the quality filtering step, sequences are clustered into OTUs (Operational Taxonomic Units), which provide a name for grouped bacteria. The term 'OTU' refers to clusters of organisms, grouped by DNA sequence similarity of a specific taxonomic marker gene (for instance, 16S rRNA). For several years, OTUs have been the most commonly used units of microbial diversity, at different taxonomic levels. It is used in the context of numerical taxonomy as a pragmatic definition to describe the group of organisms being studied.

OTUs are based on sequence identity (%ID), generally 97%. The degree of sequence variability depends on several factors, such as the region of the 16S rRNA, the length of the amplicon, and the particular taxa.

The adoption of a specific OTU-picking method may have a relevant impact on data interpretation. OTU clustering algorithms are divided into the three following categories: de novo, closed reference, and open reference. In de novo OTU picking, sequences are clustered into OTUs, without any external reference sequences. Conversely, closed-OTU picking approach uses a reference database, and sequences that fail to match the reference ones are discarded. Open-reference OTU picking includes two steps. First, a closed-reference OTU picking is performed, which is followed by de novo clustering for sequences that do not match reference sequences.

Open-reference OTU picking is generally preferred, since it retains all sequence data. In reference-based OTU picking, sequences are clustered against a reference database such as Greengenes [12], Ribosomal Database Project (RDP) [13], or SILVA [14]. The addition of more sequences to the databases over time improves matching efficiency.

The most widely used software packages for the analysis of 16S rRNA amplicon data are QIIME [11] and MOTHUR [15]. Both packages are open source and have open access online tutorials and forums. QIIME and MOTHUR are primarily accessed through an interface based on the command-line model. QIIME works on the basis of a set of scripts which are able to turn sequencing data from raw sequences to interpretable and easily readable data stored in databases, to generate graphics and statistics starting from sample metadata. Some of these scripts include one or more other software packages, such as UCLUST [16] and RDP classifier [13]. QIIME scripts implement statistical tests, alpha and beta diversity indices, and data visualization tools. QIIME can be run on a large variety of platforms, from personal computers to online clouds.

QIIME (QIIME1) may be defined as a collection of custom tools and wrappers around other software packages that makes it easy to customize metagenomic analysis. However, its excessive flexibility often makes it hard to track the provenance of data. In order to overcome these aspects, the QIIME2 software is now available online [17]. It has a very different model for data analysis, by wrapping information into one object (artifact), which contains data and metadata.

QIIME2 also uses "Sequence Variants" (SV) rather than "Operational Taxonomic Units" (OTU). Since new methods control errors sufficiently, SVs can be resolved with greater accuracy, down to the level of single-nucleotide differences over the sequenced gene region. Therefore, it is possible to state that SVs combine the advantages of closed-reference OTUs with the benefits of a finer taxonomic resolution [18].

Alpha diversity can be elaborated by using QIIME, often by integrating it with other software packages, such as Phyloseq. It is an open access bioinformatic package which is able to import, analyse and graphically display complex phylogenetic sequencing data which have already been clustered into OTUs.

Whole Genome Shotgun sequencing

With new advances in DNA technology, the cost of sequencing has decreased and Whole Genome Shotgun (WGS) metagenomic sequencing was applied in many research projects to study all microorganisms genes present in uncultured communities. This method is based on the coverage of the genome outside the small 16S rRNA region: thus, a strain level discrimination is made possible. Through shotgun metagenome sequencing is possible to investigate deeper layers of the soil microbial communities, providing an unbiased view on the phylogenetic and functional composition of its microbial communities [19].

The extraction of total DNA from the soil community is followed by a fragmentation phase in order to break DNA strands into pieces, which are then purified, amplificated and sequenced by using the available platform.

The obtained data can be analysed with MEGAN (MEtaGenome Analyzer) [20], a software package that allows optimized analysis of large metagenomic datasets. MEGAN analysis starts with collecting reads and, then, compares them to sequence databases using BLAST or similar algorithmics. Among reference database RefSeq [21], GenBank [22], or Pathosystems Resource Integration Center (PATRIC) [23] are the most relevant.

Furthermore, MEGAN assigns a tax on ID to processed read results based on NCBI taxonomy. Data can be analysed also by using MG-RAST (MetaGenomic Rapid Annotations using Subsystems Technology) [24]. MG-RAST is open-source web-based software that supports automatic phylogenetic and functional analysis of metagenomes. The related server is also one of the biggest repositories about metagenomic data. The pipeline is able to automatically produce functional assignments to the shotgun sequences by making comparisons to databases both at nucleotide and amino-acid level. The applications also make available tools for comparing different metagenomes.

WGS method was integrated with metatranscriptomic or metaproteomic approaches to investigate microbial community function [25]. At the same time, specific databases for genome annotation sequences, such as Subsystems ontology [26] and protein, such as SwissProt [27], have been created.

To study the microbiome of complex environments amplicon sequencing and whole genome shotgun sequencing are the most used approaches. Both can be applied in microbiome studies, depending on the investigation.

Some studies concerning water samples extracted in remote location found that amplicon sequencing method can detect a larger number of phyla than WGS [28]. In this study, however, the considered areas were not well investigated and were constituted only by a limited number of sequenced genomes.

Instead, in many studies concerning human microbiome, WGS were considered preferable to the amplicon sequence approach [29]. Similarly, in a recent investigation about the soil metagenome, the results show how WGS metagenomics offers finer resolution for microbial community structure and dynamics compared to 16S method, able to detect only more dominant organisms in samples [30].

Generally, WGS can be more expensive than 16S, requiring more complex data analysis [31].

Clone library sequencing

Clone library sequencing is based on the extraction of soil DNA followed by amplification of partial or full length of 16S rRNA and 1492R. The obtained sequences are then ligated and copied into plasmids (BACs, bacterial artificial chromosome, that can be 150 kb long) to divide the sequencing job into sections**.**

Clones are then purified and sequenced. Sequences are assembled and checked. Phylum, class, order, family, or OTU placement is individuated when a clone matches the similarity thresholds. When similarity to a database sequence is considered acceptable, the related clone may represent a novel subfamily [32].

Clone library sequencing and shotgun sequencing are not exclusive terms; clone library sequencing projects usually use large clone libraries to section the project into a set of shotgun sequencing projects. The basic principle is the same in both methods; the difference is that in the cloning-based sequencing method libraries of the pieces of DNA clones are firstly made. The assembling of DNA contigs will then be a lot easier to manage from the computational point of view. In shotgun sequencing the same is done but without the cloning phase.

For all metagenomic studies and for all sequencing approaches, the submission of sequence files and the metadata associated with each sample to public databases is a useful step which may further improve reliability of reference databases.

Several database initiatives exist for this purpose, including: QIIME, MG-RAST, and NCBI’s and EMBL-EBI’s respective short-read archives (SRA) data repository such as INSDC, the International Nucleotide Sequence Database Collaboration, which encompasses NCBI, EMBL-EBI and DDBJ (Annotated/Assembled Sequences database).

INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, along with contextual information relating to samples. Moreover, the European Nucleotide Archive (ENA) is produced and maintained by the European Bioinformatics Institute and is a member of the INSDC.

In addition, to ensure reproducibility, it is important to standardize protocols for sample processing and sequencing. The Genomic Standard Consortium created a standard for reporting marker gene sequences and established the minimum information required about a marker gene sequence (MIMARKS) [33] that is just a part of the MIxS family of standards, making it possible to describe a wide range of ‘omics’ data sets.

Applications of metagenomics in soil microbial community

Amplicon Sequencing applications

16S rRNA gene amplicon sequencing was used in a study investigating how bacterial chitinolytic communities respond to chitin and pH alteration in soil [34]. In other studies [35] strengths and limitations of this approach were discussed, while 18S rRNA soil genes were analysed in [36] and [37].

Other analyses displayed the diverse microbial community of vermicomposting systems [38]. Particularly, by using 16S rRNA gene pyrosequencing, Proteobacteria, Actinobacteria, Tenericutes, Bacteroidetes, Chloroflexi, Firmicutes and Planctomycetes phyla were detected in wormbed leachate.

Studies investigated how novel oligonucleotide primers reveal a high diversity of microbes which drive phosphorous turnover in soil [39]. Subsequent studies underlined how this high-throughput methods offer novel possibilities compared to cultivation-based approaches, and discussed several key points relevant in order to minimize potential biases occurring during library preparation and the subsequent bioinformatic activity [40]. Issues including soil sampling strategies, DNA extraction, and metadata collection were treated in comprehensive way in [41].

Amplicon sequencing metagenomic analysis, DNA extraction from Krossfjorden sediment, Illumina MiSeq platform and MG-RAST analysis of NGS data were also applied in [42].

In the same year, by using high-throughput ITS-amplicon sequencing, other studies investigated fungal community profiles in agricultural soils treated with different tillage, fertilization and crop rotation conditions [43].

In a recent study the Illumina technology and the amplicon sequencing approach were applied to analyse the microbiota and its metabolic capabilities in polluted soils [44]. The results detected several bacterial pathogens belonging to *Salmonella enterica*, *Pseudomonas aeruginosa*, *Escherichia coli*, and *Staphylococcus aureus* and provided useful clues for limiting the spread of dangerous microorganisms in the soil.

Genome shotgun sequencing applications

A quite recent study adopted the genome shotgun approach for investigate the changes of the microbial composition of grassland soils submitted to 2°C infrared heating for 10 years [45]. The results showed that some metabolic pathways such as cellulose degradation, CO_2 production and nitrogen cycling were improved under these particular experimental conditions.

A relevant study analysed soybean and corn cultivated soil and the effect of N fertilization on its microbiome, by using the Illumina sequencing platform [46].

Based on WGS metagenomic analysis, an investigation examined Cd-contaminated soil samples for exploring their microbial diversity [47]. The study also explored the associated metabolic pathway network in cluster of orthologous groups and Kyoto Encyclopedia of Genes and Genomes (KEGG).

In a study, WGS was applied to analyze the impact of different farming practices involving tillage techniques and N-fertilization on the microbiome of cultivated soils. Taxonomic contig binning approaches resulted to the individuation of Metagenomically Assembled Genomes (MAGs) considered dominant member of the soil microbial environment [48].

A metagenomic survey of soil microbial communities along a rehabilitation chronosequence after iron ore mining was carried out in [49]. In this investigation a paired-end library sequencing technology (NextSeq 500 Illumina) was used.

In a recent research activity [30] available National Ecological Observatory Network (NEON) soil metagenomic sequencing data in combination with open access Metagenomics Rapid Annotation (in MG-RAST server) were used to illustrate advantages of WGS compared to amplicon sequencing approach. The results showed as WGS leads to a more detailed microbial resolution and allowed to detect a larger number of bacteria, archaea, viruses, and eukaryote genera.

Clone library sequencing applications

Several research activities used this sequencing method, among them, a study performed to characterize the soil acidobacterial diversity, by considering several types of soil and the pyrosequencing method [50]. In another investigation soil and sediment samples were studied in response to oil leak, and clone library sequencing method was used for detecting their microbiota [51].

CONCLUSIONS AND DISCUSSION

Soil microorganisms are involved in important processes, such as plant growth and the cycling of carbon and other nutrients. However, most of soil microbes have not yet been detected and their mechanisms are unknown.

Metagenomics approach supports the prediction of the soil microbial community and can be successfully applied in addressing researches related to the agricultural field. As the amount of soil metagenomic data is more and more growing, it is necessary to share information, coordinating sequencing and bioinformatics activities. The Terragenome initiatives [4] were born to support this aim, promoting, at the same time, cooperative works involving scientists in metagenomic analysis.

At the same time, alongside the metagenomics approach, metabolomic methods can be used to assess the genetic variation among different agricultural species. These data in combination with other profiles generated from other omics activities, based on transcriptomics and/or proteomics can be used to draw a complete and exhaustive overview of the soil microbiota. This combined omics method is coherent with the metaphenome concept, that encompasses all omics fields, that is, metagenomics, metatranscriptomics, metaproteomics and metabolomics. In fact, "metaphenome approach" may be considered as the product of the combined genetic potential of the microbiome (metagenomics) and the environment, such as available resources, biotic and abiotic factors [52].

The integrated omics approach was applied with different aims, for instance, for supporting sustainable agriculture and for exploring the rhizosphere microbial community [3] and also for investigating the role of soil microbiota and of specific metabolites in the Anaerobic soil disinfestation (ASD), for plants disease control [53].

Alongside the opportunity to apply multi-omics approaches for soil microbiome investigation, standardized procedures that allow to share and compare results across projects are becoming very relevant. Many initiatives were started to support these aims, among them, the National Ecological Observatory Network (NEON) that provide high-quality, integrated, and standardized data about soil metagenomics analysis [30].

REFERENCES

1. Lumini E, Bianciotto V, Bonfante P. La biodiversità fungina nel suolo: un approccio di metagenomica - Giornata di studio su: Il Metagenoma del suolo: problematiche di ricerca e prospettive applicative – Firenze. 2010.
2. Jeffery S, Gardi C, Jones A, Montanarella L, Marmo L, Muco L, Ritz K, Peres G, Römbke J, Van Der Putten WH. European atlas of soil biodi-versity, Edito da European Commission, Publications Office of the European Union, Luxembourg. Jones K.L., Tod. 2010.
3. Gupta N, Vats S, Bhargava P. Sustainable Agriculture: Role of Metagenomics and Metabolomics in Exploring the Soil Microbiota. In book: In Silico Approach for Sustainable Agriculture, Springer. 2018: 183-199.
4. Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD, Bailey MJ, Nalin R, Philippot, L. TerraGenome: a consortium for the sequencing of a soil metagenome. Nature Reviews Microbiology. 2009; 7(4): 252.
5. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE. Conducting a Microbiome Study. Cell. 2014;158: 250-262.
6. Aguiar-Pulido V, Huang W, Suarez-Ulloa V, Cickovski T, Mathee K, Narasimhan G. Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis. Supplement: Bioinformatics Methods and Applications for Big Metagenomics Data. 2016; 12: 5-16.
7. Soergel DA, Dey N, Knight R, Brenner SE. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences, ISME J. 2012; 6:1440-1444.
8. Illumina. An introduction to Next-Generation Sequencing Technology.2013.
9. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27:2194-2200.
10. Wright ES, Yilmaz LS, Noguera DR. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Appl. Environ. Microbiol. 2012; 78:717-725.
11. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods. 2011; 7: 335-6.
12. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012; 6: 610-618.
13. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009; 37: 141-145.
14. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013; 41:590-596.
15. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing

microbial communities. Appl. Environ. Microbiol. 2009; 75:7537-7541.

16. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26: 2460-2461.
17. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Caporaso JG. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018; 6: 90.
18. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker gene data analysis. The ISME Journal. 2017;119.
19. Rastogi G, Sani RK. Molecular techniques to assess microbial community structure, function, and dynamics in the environment. In: Ahmad. editors. Microbes and Microbial Technology: Agricultural and Environmental Applications. New York: Springer Science+Business Media, LLC; 2011.
20. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016; 12(6): e1004957.
21. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44: 733–745.
22. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD. GenBank. Nucleic Acids Res. 2018;46: 41–47.
23. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42: 581–591.
24. Keegan KP, Glass EM, Meyer F. MG-RAST. A Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods in Molecular Biology. 2016; 1399: 207–233.
25. Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One. 2008; 3: e3042.
26. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014; 42: 206–214.
27. Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank: current status. Nucleic Acids Res. 1994; 22: 3578–3580.
28. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci Rep. 2017;7: 6589.
29. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016;469: 967–977.
30. Brumfield KD, Huq A, Colwell RR, Olds JL, Leddy MB. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. PLoS ONE. 2020; 15(2): e0228899.
31. van Nimwegen KJM, van Soest RA, Veltman JA, Nelen MR, van der Wilt GJ, Vissers LELM. Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. Clin Chem. 2016; 62: 1458–1464.
32. DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microbial Ecology. 2007; 53(3):371-383.
33. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 2011; 29, 415-420.
34. Kielak AM, Cretoiu MS, Semenov AV, Sorensen SJ, and van Elsas JD. Bacterial chitinolytic communities respond to chitin and pH alteration in soil. Appl Environ Microbiol. 2013; 79:263–272.
35. Poretsky R, Rodriguez RL, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One. 2014; 9:e93827.
36. Lentendu G, Wubet T, Chatzinotas A, Wilhelm C, Buscot F, Schlegel M. Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a multiple barcoding approach. Mol Ecol. 2014; 23:3341–3355.
37. Wang Y, Tian RM, Gao ZM, Bougouffa S, Qian PY. Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. PLoS One. 2014; 9:e90053.
38. Romero-Tepal EM, Contreras-Blancas E, Navarro-Noya YE, Ruiz-Valdiviezo VM, Luna-Guido M, Gutierrez-Miceli FA. Changes in the bacterial community structure in stored wormbed leachate. J Mol Microbiol Biotechnol. 2014; 24: 105–13.
39. Bergkemper F, Kublik S, Lang F, Kruger J, Vestergaard G, Schloter M, Schulz S. Novel oligonucleotide primers reveal a high diversity of microbes which drive phosphorous turnover in soil. J Microbiol Methods. 2016; 125:91–97.
40. Schöler A, Jacquiod S, Vestergaard G. Analysis of soil microbial communities based on amplicon sequencing of marker genes. Biol Fertil Soils. 2017; 53: 485–489.

41. Vestergaard G, Schulz S, Schöler A, Schloter M. Making big data smart—how to use metagenomics to understand soil quality. Biol Fert Soils. 2017.
42. Kachiprath B, Puthumana J, Gopi J, Solomon S, Krishnan KP, Philip R. Amplicon sequencing based profiling of bacterial diversity from Krossfjorden, Arctic. Data in Brief. 2018; 21: 2522-2525.
43. Sommermann L, Geistlinger J, Wibberg D, Deubel A, Zwanzig J, Babin D, Schlüter A, Schellenberg I. Fungal community profiles in agricultural soils of a long-term field trial under different tillage, fertilization and crop rotation conditions analyzed by high-throughput ITS-amplicon sequencing. PLoS ONE. 2018; 13:e0195345.
44. De Mandal S, Mathipi V, Muthukumaran R. Amplicon sequencing and imputed metagenomic analysis of waste soil and sediment microbiome reveals unique bacterial communities and their functional attributes. Environ Monit Assess. 2019; 191: 778.
45. Luo C, Rodriguez-R LM, Johnston ER, Wu L, Cheng L, Xue K. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. Applied and Environmental Microbiology. 2014; 80(5):1777-1786.
46. Orellana LH, Chee-Sanford JC, Sanford RA, Löffler FE, Konstantinidis KT. Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. Applied and Environmental Microbiology. 2018; 84(2):e01646-e01617.
47. Feng G, Xie T, Wang X. Metagenomic analysis of microbial community and function involved in cd-contaminated soil. BMC Microbiol. 2018; 18: 11.
48. Nelkner J, Henke C, Wentong Lin T, Pätzold W, Hassa J, Jaenicke S, Grosch R, Pühler A, Sczyrba A, Schlüter A. Effect of Long-Term Farming Practices on Agricultural Soil Microbiome Members Represented by Metagenomically Assembled Genomes (MAGs) and Their Predicted Plant-Beneficial Genes. Genes (Basel). 2019; 1(6): 424.
49. Gastauer M, Vera M, de Souza K. A metagenomic survey of soil microbial communities along a rehabilitation chronosequence after iron ore mining. Sci Data. 2019; 6: 190008.
50. Jones RT, Robeson MS, Lauber CL, Hamady M, Knight R, Fierer N. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. The ISME Journal. 2009; 3(4):442-453.
51. Vázquez S, Monien P, Minetti RP, Jürgens J, Curtosi A, Primitz JV. Bacterial communities and chemical parameters in soils and coastal sediments in response to diesel spills at Carlini Station, Antarctica. Science of the Total Environment. 2017; 605:26-37.
52. Jansson JK, Baker ES. A multi-omic future for microbiome studies. Nat Microbiol. 2016; 1:1-3.
53. Hewavitharana SS, Klarer E, Reed AJ, Leisso R, Poirier B, Honaas L, Rudell DR and Mazzola R. Temporal Dynamics of the Soil Metabolome and Microbiome During Simulated Anaerobic Soil Disinfestation. Front. Microbiol. 2019.