Chapter

Metagenomics-Based Phylogeny and Phylogenomic

Ayixon Sánchez-Reyes and Jorge Luis Folch-Mallol

Abstract

Phylogenetic relationships among microbial taxa in natural environments provide key insights into the mechanisms that shape community structure and functions. In this chapter, we address the current methodologies to carry out community structure profiling, using single-copy markers and the small subunit of the rRNA gene to measure phylogenetic diversity from next-generation sequencing data. Furthermore, the huge amount of data from metagenomics studies across the world has allowed us to assemble thousands of draft genomes, making necessary the comparison of whole genomes composites through phylogenomic approximations. Several computational tools are available to carry out these analyses with considerable success; we present a compendium of those open source tools, easy to use and with modest hardware requirements, with the aim that they can be applied by biologists non-specialists to study microbial diversity in a phylogenetic context.

Keywords: metagenomics profiling, phylogenetic diversity, phylogenetic metadata representation

1. Introduction

Next-generation sequencing technologies have transformed our perception of diversity and microbial distribution in natural ecosystems and have contributed substantially to the discovery of totally new microbial landscapes in such distinctive environments as the gut of mammals, the vegetal rhizosphere, vascular tissues of higher plants, and even in volcanic lakes [1–3]. There are two general approaches to profile microbial communities through next-generation sequencing techniques: shotgun sequencing of total DNA isolated directly from the environment and sequencing of variable regions coming from SSU-rRNA genes (we know these approaches as metagenomics methods since all involve the culture-independent genomic analysis of microbiomes on a particular environment [4, 5]). Both approaches have been widely used to trace microbial diversity at increasingly fine taxonomic levels, either by capturing a representative fraction of the total gene content or by amplicon sequencing techniques like the popular bacterial 16S rRNA. Each method has advantages and disadvantages, and the selection depends on several factors like taxonomic level resolution, cost, sensitivity, and primer bias, among others. One of the challenges associated with metagenomics methods is the analysis of massively generated data. Both the sequencing of amplicons and environmental DNA produces millions of short DNA sequences (reads), which must undergo preprocessing and quality control, before they can be used to extract

biologically useful information from them. One of the goals of massively sequencing data analysis is to obtain the patterns of phylogenetic diversity in ecological communities, an important trait in order to assess the classic ecological questions "Who is there?" or "What they are doing?" and provide better understandings into the phylogenetic relationships among microbial community taxa. Extracting phylogenetic information from massive sequencing reads is not a trivial task; however, it can be achieved with reasonable success by using several profiling tools adapted both to the analysis of amplicons of ribosomal genes and to the conserved genes between different domains [6, 7]. The microbial community structure has been approached mostly using the 16S SSU-rRNA gene as phylogenetic marker, mainly due to lower sequencing costs and an acceptable relation of specificity-resolution in taxonomic assignments [8], while methods that use single-copy markers obtained from shotgun sequencing reads or assembled samples are gaining relevance because they have demonstrated strain-level resolution [9, 10], a really hard issue when analyzing complex microbiomes.

To date, several computational tools have been developed to carry out community profiling and phylogenetic inferences from next-generation sequencing data with considerable success. In this chapter we present a compendium of open-source tools and easy-to-use with modest hardware requirements, with the aim that they can be applied by biology non-specialists to study microbial diversity in a phylogenetic context. We show several practical examples explained step by step, in order to provide to the reader, the replication using their own data.

We have selected tools for use on a local computer through the Unix command line, and tools are available from dedicated servers, with easy access and intuitive use. The examples described in the chapter were tested on a Dell Optiplex 7010 desktop, 6T6ZYV1 Series, Intel (R) Core (TM) i5-3550 CPU at 3.30 GHz, Memory 12 GiB.

2. Community structure profiling across microbial samples using single-copy markers

With the advent of massive DNA sequencing technologies, several methods have been developed to assign shotgun reads to microbial taxonomic categories. These methods aim to perform a microbial community profiling that infers its relative structure, and they are very important to understand how microbiomes work in nature, their phylogenetic composition, and even their dynamics and evolutionary history. The starting point for these analyzes is a set of reads obtained by massive sequencing whose length is variable (as little as 50–75 bp up to >1000 bp) depending on the platform used (Illumina, Ion Torrent, PacBio RS). We can understand by a read the sequence of bases from a single discrete molecule of DNA, obtained in a massively parallel manner [11]. However, currently most metagenomics studies use a range of a short-read sequencing instruments between 100 and 600 bp in order to maximize counting reads and lower costs. These short-reads contain the genomic, phylogenetic, and functional information of the microbiome into millions of discrete DNA fragments, which are sufficient to make a reliable estimate of the phylogenetic diversity present in a microbial sample (**Figure 1**).

The taxonomic composition of a microbial community can be estimated from a set of short-reads by assigning each read to the most likely microbial lineage [12]. Historically, a single gene target approach has been the gold standard for assigning taxonomy in the Prokaryote domain, through the 16S ribosomal RNA gene. However, this presents important biases related to copy-number variations and significant intraspecific differences ~6%. In this sense, both clade-specific and universal single-copy phylogenetic markers genes have gained popularity among the

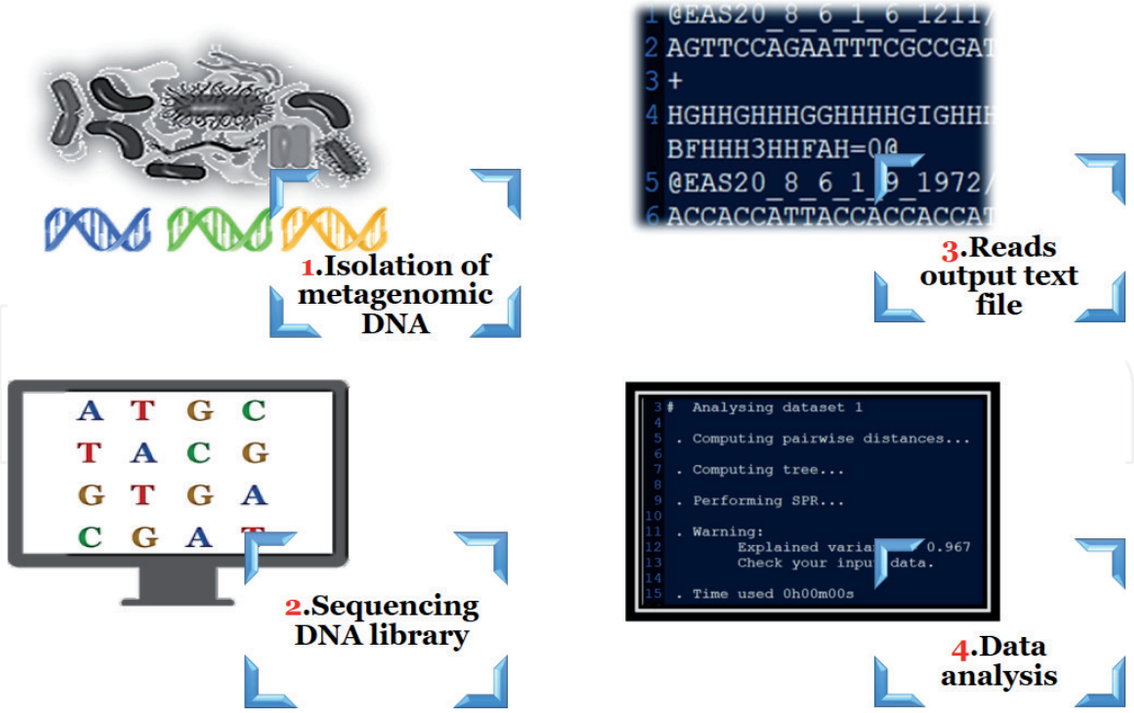


Figure 1.
General overview of metagenomics analysis in a microbial sample by next-generation sequencing: (1) isolation of metagenomic DNA, (2) sequencing DNA library, (3) reads output text file, and (4) data analysis.

scientific community since they are not subject to intragenomic diversity, are rarely subjects of horizontal transfer, and have proven robustness to delineate species and prokaryotic strains in multiple studies, because several genes can be combined to reconstruct phylogenies [13, 14]. Although each method selects its own set of clade-specific or universal markers, most of these genes encode proteins with functional relevance in housekeeping metabolism (**Table 1**). To make the analysis, the coding nucleotide sequences are generally used as they offer better resolution than amino acid sequences in closely related organisms [16]. This simplifies the computational analysis as the short-reads could be compared unambiguously without the need to translate them into proteins, which could generate artifacts given the small size of the reads.

One of the most popular tools for microbial profiling based on clade-specific marker genes is the MetaPhlAn classifier [12, 17]. MetaPhlAn maps the experimental reads against a collection of 231 markers for species-level comparisons and >115,000 markers for higher taxonomic levels. Among the advantages of this classifier is that no preprocessing is required, so raw data can be uploaded and analyzed. The main disadvantage for non-specialists is that MetaPhlAn works through the command line in a Unix architecture.

2.1 Profiling a textile dye degrader microbiome with MetaPhlAn2

Next we described the steps performed for profiling a microbial community capable of degrading the textile dye HC Blue no. 2. Also we show a graphical representation of the profiling phylogenetic metadata. This general strategy can be applied to profile any microbial community from short-reads obtained by massive sequencing. Symbol convention: Comments (#); executable commands ($). The raw data are available on [18].

You can find a complete MetaPhlAn guide on the author's site: https://bitbucket.org/biobakery/biobakery/wiki/metaphlan2.

Clusters of orthologous groups of proteins (COG)	Protein name
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S7
COG0052	Ribosomal protein S2
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1
COG0085	DNA-directed RNA polymerase, beta subunit
COG0087	Ribosomal protein L3
COG0088	Ribosomal protein L4
COG0090	Ribosomal protein L2
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8
COG0097	Ribosomal protein L6P/L9E
COG0098	Ribosomal protein S5
COG0099	Ribosomal protein S13
COG0100	Ribosomal protein S11
COG0102	Ribosomal protein L13
COG0103	Ribosomal protein S9
COG0124	Histidyl-tRNA synthetase
COG0172	Seryl-tRNA synthetase
COG0184	Ribosomal protein S15P/S13E
COG0185	Ribosomal protein S19
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0200	Ribosomal protein L15
COG0201	Preprotein translocase subunit SecY
COG0202	DNA-directed RNA polymerase, alpha subunit/40 kD subunit
COG0215	Cysteinyl-tRNA synthetase

Table 1.
Universal single-copy phylogenetic marker genes employed in metagenomics-based phylogenies for delineation of prokaryotic species (modified from [15]).

Installing MetaPhlAn2
with an activated Bioconda channel in Linux, type the following command:
$ *conda install metaphlan2*
# this will install the software with all its dependencies
 # Generate a taxonomic profile
Type the following command:
$ python /path/to/metaphlan2.py /path/to/textile_microbiome.fastq.gz --input_type fastq > textile_microbiome _profile.txt

#Sample ID	Abundance MetaPHlAn2_analysis (%)
k__Bacteria	56.02708
k__Archaea	43.94783
k__Viruses	0.02509
k__Bacteria|p__Firmicutes	45.48396
k__Archaea|p__Euryarchaeota	43.94783
k__Bacteria|p__Proteobacteria	8.46518
k__Bacteria|p__Actinobacteria	2.07794
k__Viruses|p__Viruses_noname	0.02509

The taxonomic levels are: Kingdom, k; and Phylum, p. The table was trimmed to show only up to the phylum level; to read complete report, see [7].

Table 2.
Features of the abundance table for a textile dye degrader microbiome profile.

The output profile (called: textile_microbiome _profile.txt) contains the computed clade's abundances (**Table 2**).
 # Capture phylogenetic relatedness with GraPhlAn
In order to visualize microbial abundances on a phylogeny we'll use GraPhlAn tool [19].
Installing GraPhlAn
Type the following two commands:
$ brew tap biobakery/biobakery
$ brew install graphlan
# In order to know the installation directory type the following command:
$ which graphlan
 # Input files
Type the following commands sequentially:
*$ python path/to/merge_metaphlan_tables.py *_profile.txt > merged_abundance_table.txt*
$ python path/to/export2graphlan.py --skip_rows 1,2 -i merged_abundance_table.txt --tree merged_abundance.tree.txt --annotation merged_abundance.annot.txt --most_abundant 100 --abundance_threshold 1 --least_biomarkers 10 --annotations 5,6 --external_annotations 7 --min_clade_size 1
 # Create the phylogeny
Type the following commands sequentially:
$ python path/to/graphlan_annotate.py --annot merged_abundance.annot.txt merged_abundance.tree.txt merged_abundance.xml
$ python path/to/graphlan.py --dpi 300 merged_abundance.xml merged_abundance.png --external_legends

Finally, you will obtain:

- a cladogram called: merged_abundance.png
- an annotation file called: merged_abundance_annot.png
- a legend file called: merged_abundance_legend.png

You can change the format of the final results to pdf, just modifying the name: *merged_abundance.png* to *merged_abundance.pdf* in the last command. A representation of the annotated cladogram is shown in **Figure 2**. The size of the nodes correlates with microbial community relative abundances.

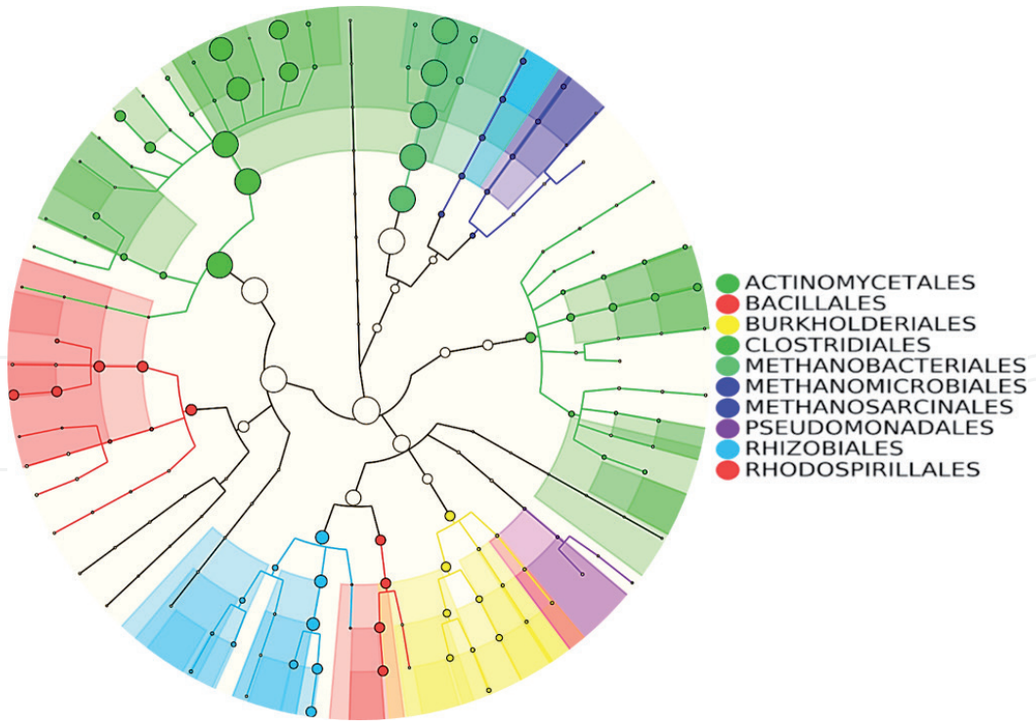


Figure 2.
Cladogram produced by GraPhlAn software with metadata from MetaPhlan community profiling to order level. Taxon abundance is proportional to the circle diameter.

3. Phylogenetic diversity of microbial communities based on 16S rDNA gen

Estimating the taxonomic and phylogenetic diversity of a microbial community is also possible through sequencing and analysis of small ribosomal RNA subunit (16S rRNA) gene, whenever this sequence has been considered for a long time a stable marker, crucial in the microbial systematics of the last 30 years. 16S ribosomal ribonucleic acid is a key component of the small subunit of prokaryotic ribosomes, central player in the cellular biology of microorganism; it serves as a linker for the process of translating genetic information to proteins [20]. Because DNA is much easier to sequence than RNA, DNA segment coding for 16 rRNA is obtained for the purposes of sequencing (**Figure 3**). This gene fragment meets several features that have made it a "quasi-gold standard" for bacterial taxonomy:

- It is a ubiquitous gene in the Bacteria and Archaea domains.
- Within its ~1500 bp, it has discrete regions with enough variability to establish a phylogenetic signal among phyla and even genus.
- It has conserved regions that allow the design of "universal primers," a very useful feature in massive sequencing.
- It has several databases enriched with sequences from almost all international projects where 16S sequences are obtained (**Table 3**). For example, the 16S ribosomal RNA (Bacteria and Archaea) database from the National Center for Biotechnology Information (NCBI) contains near to 20,831 curated records and more than 17 million of total records (consulted date: 2019/08/05: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome).

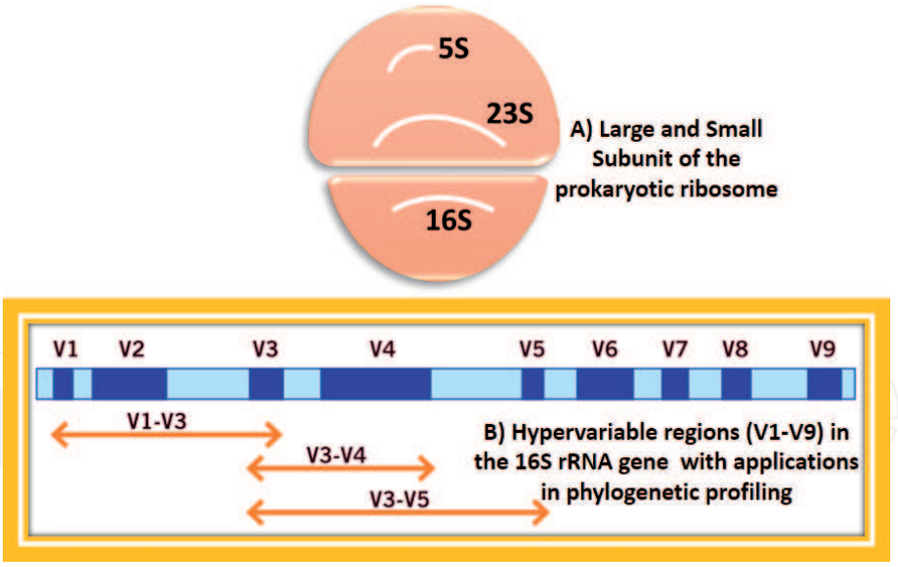


Figure 3.
Prokaryotic ribosome general representation and variable sequence regions used in microbial phylogenetic diversity estimations.

Database	Available SSU sequences	Current release	Citation/link
16S NCBI database	20831[a]	2019	[21]/https://blast.ncbi.nlm.nih.gov/
SILVA rRNA database	23629[b]	2018	[22]/https://www.arb-silva.de
Ribosomal Database Project (RDP)	16277	September 2016	[23]/https://rdp.cme.msu.edu
EzBioCloud 16S database	13132[c]	2019	[24]/https://www.ezbiocloud.net/
Genomic-based 16S ribosomal RNA database (GRD)	13202	2015	https://metasystems.riken.jp/grd/download.html

[a]*Not redundant manually curated small (16S, SSU) subunit ribosomal RNA sequences.*
[b]*The dataset contains 23,629 SSU sequences representing a single bacterial type strain up to June 2017.*
[c]*Phylotypes with validly published names.*

Table 3.
Most popular public databases for depositing and analyzing sequences of the 16S ribosomal gene.

3.1 16S community profiling by analysis of ribosomal amplicons

Microbial diversity is measured as a function that depends on the richness and abundance of distinct taxons among any community [25]. Obtaining representative DNA sequences from the entire community is essential to make valid inferences. Profiling a microbial community through 16S gene analysis generally consists of four steps (**Figure 4**). To date, several computational tools have been developed to analyze microbial communities through the 16S gene marker; however, estimating the total microbial diversity in any environment is a still a major challenge [6, 26–28], influenced by several factors, among them we want to mention two: (I) processing huge amounts of data moves within the limits of modern computing and (II) the need for some expertise that can cost years of training. Fortunately, many tools have been developed in recent years, aiming to make bioinformatics platforms dedicated to this type of analysis more human-friendly, and there are dedicated sites exclusively to deposit computational alternatives for almost all needs, for example, https://github.com/.

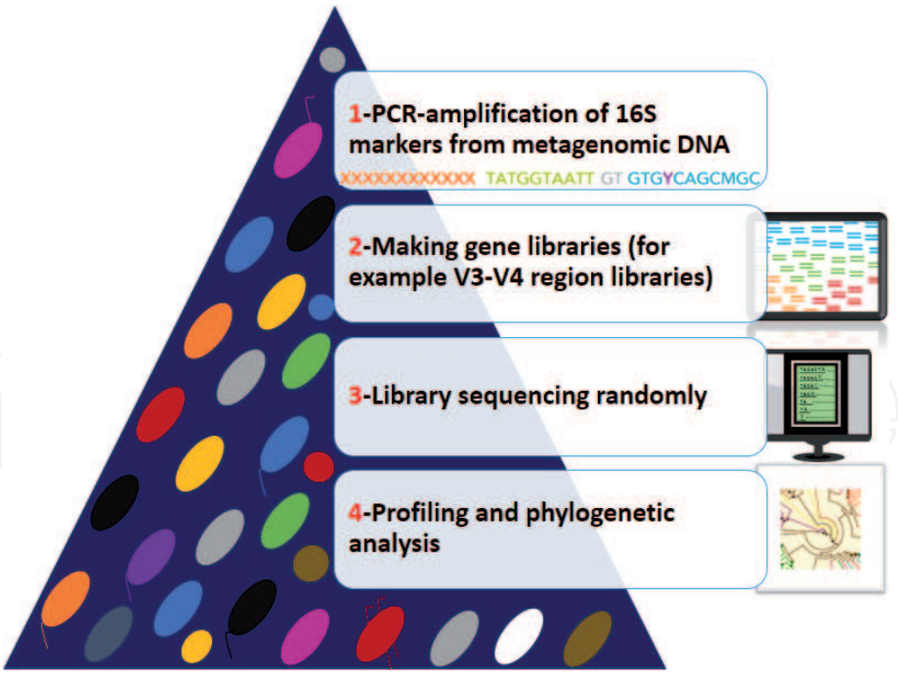


Figure 4.
General steps for profiling a microbial community through 16S gene analysis.

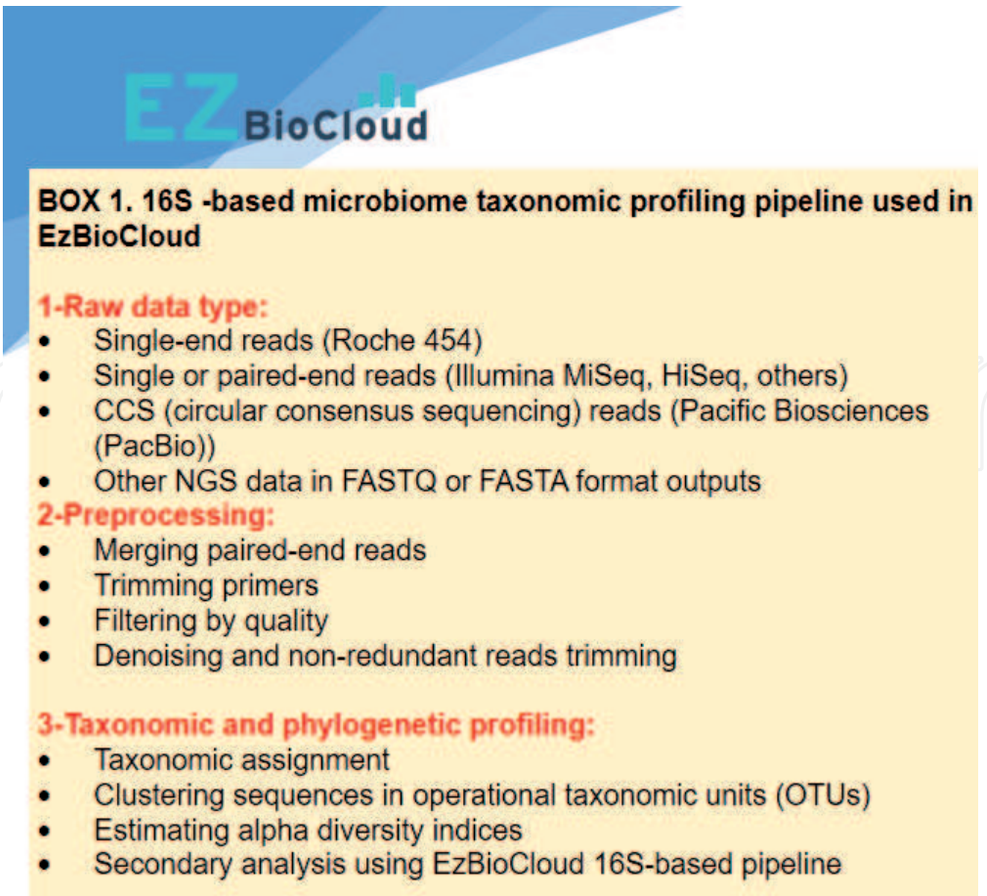
A good example of these multiplatforms to profile microbial communities is the Microbiome Taxonomic Profiling (MTP) pipeline from EzBioCloud site (https://www.ezbiocloud.net/contents/16smtp) [24]. Among its fundamental advantages are: it is free, knowledge of Linux environment is not needed to carry out the analyses, and several types of outputs such as functional profiles, taxonomic and phylogenetic structure, as well as on-demand comparison with other published microbiome data are fully available. New users of EzBioCloud will be required to open a local server account (https://www.ezbiocloud.net/signup?from=addMTP); after that you can upload up to 100,000 reads for sample and begin the analysis. We list general steps to perform a profiling on the platform (**Box 1**).

The platform consists of a very intuitive and user-friendly presentation that guides the beginner user at every stage of the analysis. The first step is the uploading of the next-generation sequencing data (16S amplicon reads). After that, you can request for the MTP pipeline, and the analysis starts. In a relatively short time, you can access the result portal with the preprocessing results resumed in pre-filtered reads (by removing low-quality and chimeric amplicons), statistics about read lengths, and taxonomic read assignments at species level.

Other outputs in results portal are related with several diversity indices, taxonomic composition and hierarchy, and graphical implementations like Krona [29]. MTP implements seven different diversity indices; among them is the phylogenetic diversity index, a measure of biodiversity that considers phylogenetic difference between taxons and ponders several variables like taxonomic diversity and species abundances or distributions.

3.2 Extracting 16S sequences from assembled data

In occasions, we do not have a set of DNA short-reads, but assembled composites in contiguous regions of variable size. Such is the case of genomes assembled from metagenomes or contigs from complex metagenomes. Inferring taxonomic diversity from this type of data usually requires other strategies. One of the most useful is to predict all the rRNA sequences contained in the assembly and cluster them



Box 1.
16S-based microbiome taxonomic profiling pipeline used in EzBioCloud.

according to their identity (this implies making a list of nonredundant sequences) to define operative taxonomic units. A simple way to address this problem is through the use of Barrnap software [27]; it works through the Unix command line and has the advantage of consuming few computational resources, so that several complex microbiomes can be analyzed in a personal computer for extraction of rRNA sequences. Barrnap gives us an output with all predicted sequences; this includes 5S, 16S, and 23S rRNA in the case of bacteria. The sequences can be saved on-demand in a text file and subsequently analyzed by a third-party phylogenetic processing software to establish evolutionary relationships between taxa. A suitable platform for this objective is SeaView [28], which contains sequence alignment and curing utilities, as well as a set of phylogenetic reconstruction methods, like PhyML, which uses maximum likelihood algorithms and seven different evolutionary models. It is also possible to use distance methods such as Neighbor Joining and BioNeighbor Joining, both with seven different methods to calculate distances between sequences. The platform is open access and has the advantage of being a graphical application that works on Unix and Windows, as well as being very intuitive.

4. Open-source software for phylogenetic and phylogenomic surveys

Genome-based comparisons play an essential role in the current taxonomy and phylogenetic of Bacteria and Archaea domains and eventually will replace the single gene target approach ruled by 16S rRNA gene phylogeny. The exponential growth of complete genomes and genome drafts with significant completeness values and low contamination (<5%) in international databases has resulted in an approach to phylogenetic analysis where the whole information has become in a more conservative

Software	Application	NGS data	License	Environment	Reference
MetaPhlAn	Microbial community profiling	Shotgun sequencing data	Open access	Unix command line	[17]
FOCUS	Taxonomic profiling	Shotgun unannotated sequencing reads	Open access	Unix command line and Web implementation	[30]
Kraken	Assigning taxonomic labels to metagenomic DNA sequences	Shotgun unannotated sequencing reads	Open access	Unix command line	[31]
GraPhlAn	Phylogenetic analysis	Metadata from short-read community profiling	Open access	Unix command line	[19]
PICRUSt	Predictive functional profiling of microbial communities	16S amplicons	Open access	Unix command line	[32]
QIIME	Taxonomic and phylogenetic profiling	16S amplicons	Open access	Unix command line and web implementation	[28]
Mothur	Taxonomic and phylogenetic profiling	16S rRNA gene sequences	Open access	Unix command line	[6]
UBCG	Phylogenomic tree reconstruction	Set of bacterial genomes	Open access	Unix command line	[33]
GToTree	A user-friendly workflow for phylogenomics	Set of bacterial genomes	Open access	Unix command line	[34]
PhylOTU	Identifies OTUs from rRNA sequence by phylogenetic profiles	PCR and shotgun sequenced SSU-rRNA markers	freely available	Unix command line	[35]
PhyloSift	Phylogenetic analysis of genomes and metagenomes	Metagenomic datasets generated by modern sequencing platforms	Freely available	Unix command line	[36]
VITCOMIC2	Phylogenetic representation based on 16S rRNA gene amplicons	16S amplicons	Freely	Web server	[37]
Barrnap	Very fast ribosomal RNA prediction	Assemblies from genomic or metagenomic data	Freely available	Unix command line	[38]
SeaView	Multiplatform for phylogenetic inferences	DNA or protein sequences	Freely available	Unix and Windows environments	[39]

Table 4.
Open-source software for metagenomics-based profiling and phylogenies.

fingerprint of the taxonomic categories. The current challenges for science involve improving existing methods for data acquisition and processing, since comparative analysis, even among modest-sized microbial genomes, can be computationally expensive. Here we present a list of those open-source tools and easy-to-use and modest hardware requirements, with the aim that they can be applied by biologists to study microbial diversity in a phylogenetic context (**Table 4**).

5. Conclusions

Profiling microbial communities from massive sequencing data constitutes a breaking point in the understanding of population structure and dynamics, their ecological functions and the complex relationships established between non-cultivable microorganisms. Through technological developments such as next-generation sequencing and the developing of hundreds of open-access platforms, we have been able to better understand the role of the microbial world in natural ecosystems. This chapter intends to bring the use of computational biology tools to professionals in biological sciences with different expertise, interested in the world of metagenomics analysis. We have started with the basics of microbial community profiling through shotgun sequencing data and its processing using MetaPhlAn software (the reader will notice that there are other tools perhaps more appropriate to their conditions, an interesting option is the FOCUS software that works through a Web server). MetaPhlAn has the advantage of being fully integrated with the GraPhlan phylogenetic reconstruction tools. We dedicate a complete section to the 16S gene-based communities profiling; we illustrate the EzBioCloud platform, a useful tool to obtain ecological and phylogenetic information of microbiomes. An alternative approach to process assembled data is the use of Barrnap software, which is very fast and efficient to extract ribosomal sequences in assembled data, which can be subsequently clustered and processed with phylogenetic construction tools such as SeaView. Finally, we present a list of software that can serve as a guide for the analysis of microbiomes from their taxonomic characterization to the study of phylogenetic relationships between taxa.

Acknowledgements

We thank the supercomputing resources and services of the Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC) from Universidad Nacional Autónoma de México, through the project: LANCAD-UNAM-DGTIC-371.

Conflict of interest

The authors declare no conflict of interest.

Author details

Ayixon Sánchez-Reyes[1*] and Jorge Luis Folch-Mallol[2]

1 Cátedras Conacyt, Institute of Biotechnology-UNAM, Cuernavaca, Morelos, Mexico

2 Biotechnology Research Center, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, Mexico

*Address all correspondence to: ayixon.sanchez@mail.ibt.unam.mx

IntechOpen

References

[1] Press MO, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo C-C, et al. bioRxiv [Internet]. 2017:198713. Cold Spring Harbor Laboratory. Available from: https://www.biorxiv.org/content/10.1101/198713v1 [Accessed: 2019-08-10]

[2] Bakker PAHM, Berendsen RL, Doornbos RF, Wintermans PCA, Pieterse CMJ. The rhizosphere revisited: Root microbiomics. Frontiers in Plant Science. 2013;**4**:1-7. DOI: 10.3389/fpls.2013.00165

[3] Mapelli F, Marasco R, Rolli E, Daffonchio D, Donachie S, Borin S. Microbial Life in Volcanic Lakes. In: Rouwet D, Christenson B, Tassi F, Vandemeulebrouck J, editors. Volcanic Lakes. Advances in Volcanology. Berlin, Heidelberg: Springer; 2015. p. 507-522. DOI: 10.1007/978-3-642-36833-2_23

[4] Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chemistry & Biology. 1998;**5**:R245-R249. DOI: 10.1016/s1074-5521(98)90108-9

[5] Sleator R, Shortall C, Hill C. Metagenomics. Letters in Applied Microbiology. 2008;**47**:361-366. DOI: 10.1111/j.1472-765x.2008.02444.x

[6] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology. 2009;**75**:7537-7541. DOI: 10.1128/aem.01541-09

[7] Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Metagenomic phylogenetic analysis. Nature Methods. 2017;**12**(10):626-638. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3589

[8] Schriefer AE, Cliften PF, Hibberd MC, Sawyer C, Brown-Kennerly V, Burcea L, et al. A multi-amplicon 16S rRNA sequencing and analysis method for improved taxonomic profiling of bacterial communities. Journal of Microbiological Methods. 2018;**154**:6-13. DOI: 10.1016/j.mimet.2018.09.019

[9] Silva GGZ, Green KT, Dutilh BE, Edwards RA. SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data. Bioinformatics. 2015;**32**:354-361. DOI: 10.1093/bioinformatics/btv584

[10] Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome research. 2017;**27**:626-638. DOI: 10.1101/gr.216242.116

[11] Goodwin S, Mcpherson JD, Mccombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;**17**:333-351. DOI: 10.1038/nrg.2016.49

[12] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 2012;**9**:811-814. DOI: 10.1038/nmeth.2066

[13] Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nature Methods. 2013;**10**:1196-1199. DOI: 10.1038/nmeth.2693

[14] Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal

delineation of prokaryotic species. Nature Methods. 2013;**10**(9):881-884. DOI: 10.1038/nmeth.2575

[15] Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. Nature Methods. 2013;**10**:881-884. DOI: 10.1038/nmeth.2575

[16] Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. Journal of Bacteriology. 2005;**187**:6258-6264. DOI: 10.1128/jb.187.18.6258-6264.2005

[17] Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 2015;**12**:902-903. DOI: 10.1038/nmeth.3589

[18] Sánchez-Reyes A. Massive sequencing dataset of a textile dye degrader microbiome [Internet]. Mendeley Data, v1. 2019. DOI: 10.17632/t4j8tc3njd.1

[19] Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ. 2015;**3**:1-17. DOI: 10.7717/peerj.1029

[20] Ramazzotti M, Bacci G. 16S rRNA-based taxonomy profiling in the metagenomics era. In: Metagenomics: Perspectives, Methods, and Applications. Cambridge, Massachusetts: Academic Press; 2017. pp. 103-119. DOI: 10.1016/B978-0-08-102268-9.00005-7

[21] Sayers EW, Agarwala R, Bolton EE, et al. Database resources of the national center for biotechnology information. Nucleic Acids Research. 2019;**47**:D23-D28. DOI: 10.1093/nar/gky1069

[22] Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research. 2013;**42**:D643-D648. DOI: 10.1093/nar/gkt1209

[23] Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR. The RDP (Ribosomal Database Project). Nucleic Acids Research. 1997;**25**:109-110. DOI: 10.1093/nar/25.1.109

[24] Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. International Journal of Systematic and Evolutionary Microbiology. 2017;**67**:1613-1617. DOI: 10.1099/ijsem.0.001755

[25] Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. Robust estimation of microbial diversity in theory and in practice. The ISME Journal. 2013;**7**(6):1092-1101

[26] Ramazzotti M, Berná L, Donati C, Cavalieri D. riboFrame: An improved method for microbial taxonomy profiling from non-targeted metagenomics. Frontiers in Genetics. 2015;**6**:1-12. DOI: 10.3389/fgene.2015.00329

[27] Bruno F, Marinella M, Santamaria M. e-DNA meta-barcoding: From NGS raw data to taxonomic profiling. Methods in Molecular Biology. 2015;**1269**:257-278. DOI: 10.1007/978-1-4939-2291-8_16

[28] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 2010;**7**:335-336

[29] Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC

Bioinformatics. 2011;**12**(1):289-294. Available from: http://www.biomedcentral.com/1471-2105/12/385

[30] Silva GGZ, Cuevas DA, Dutilh BE, Edwards RA. FOCUS: An alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ. 2014;**2**:e425. DOI: 10.7717/peerj.425

[31] Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;**15**:R46. DOI: 10.1186/gb-2014-15-3-r46

[32] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology. 2013;**31**:814-821. DOI: 10.1038/nbt.2676

[33] Na SI, Kim YO, Yoon SH, min HS, Baek I, Chun J. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. Journal of Microbiology. 2018;**56**:281-285. DOI: 10.1007/s12275-018-8014-6

[34] Lee MD. GToTree: A user-friendly workflow for phylogenomics. Bioinformatics. 2019. DOI: 10.1093/bioinformatics/btz188

[35] Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, et al. PhyLOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Computational Biology. 2011;**7**:e1001061. DOI: 10.1371/journal.pcbi.1001061

[36] Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. PhyloSift: Phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;**2**:e243. DOI: 10.7717/peerj.243

[37] Mori H, Maruyama T, Yano M, Yamada T, Kurokawa K. VITCOMIC2: Visualization tool for the phylogenetic composition of microbial communities based on 16S rRNA gene amplicons and metagenomic shotgun sequencing. BMC Systems Biology. 2018;**12**:S2. DOI: 10.1186/s12918-018-0545-2

[38] Seemann T. barrnap 0.8: Rapid ribosomal RNA prediction [Internet]. 2013. Available from: https://github.com/tseemann/barrnap [Accessed: 2019-08-10]

[39] Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution. 2010;**27**(2):221-224. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19854763