Short Communication

Oral microbiome research – A Beginner's glossary

Priya Nimish Deo, Revati Shailesh Deshmukh

Department of Oral Pathology and Microbiology, Bharati Vidyapeeth Deemed to be University, Dental College and Hospital, Pune, Maharashtra, India

Abstract

Oral microbiome plays a key role in the etiology of oral diseases and is linked to many diseases in other parts of the body as well. This makes the oral microbiome an area of interest for researchers globally. A meticulous planning of the research project is the first and most crucial step while conducting an oral microbiome study. For beginners in this field, it is essential to be familiar with the terminologies used in oral microbiome research for a better understanding. The purpose of this article is to familiarize new researchers to the frequently used terms for the field of oral microbiome research.

Keywords: Microbiome, metagenomics, sequencing

Address for correspondence: Dr. Priya Nimish Deo, Department of Oral and Pathology and Microbiology, Bharati Vidyapeeth Deemed to be University, Dental College and Hospital, Katraj-Dhankawadi, Pune Satara Road, Pune - 411 043, Maharashtra, India.
E-mail: priyanimishdeo@gmail.com
Submitted: 31-Dec-2021, **Accepted:** 02-Feb-2022, **Published:** 31-Mar-2022

MICROBIOME

The term microbiome was coined by scientist Joshua Lederberg, a Nobel Prize laureate, to describe the ecological community of symbiotic, commensal and pathogenic micro-organisms.[1]

MICROBIOTA

Microbiota refers to the assembly of micro-organisms present in a defined environment. This term "microbiota" was first defined by Lederberg and McCray who pointed out the significance of micro-organisms inhabiting in the human body in health and disease states.[2]

METAGENOMICS

Metagenomics is the direct analysis of genomes which are obtained from different environments. The term metagenomics is used interchangeably with 16S ribosomal RNA (rRNA) sequencing. 16S rRNA sequencing is a marker gene approach and does not target the whole genome, while metagenomics is a shotgun sequencing approach for the genomic analysis of the microbes from a particular environment. It catalogs all micro-organisms both culturable and nonculturable from complex environmental samples.[3]

META-TRANSCRIPTOMICS

Meta-transcriptomics refers to the genes that are expressed as a whole by a community.[4] It is an approach to reveal information about transcriptionally active populations rather than just the genetic content of bacterial populations, as shown by metagenomic analysis.[5]

METAPROTEOMICS

Meta-proteomics is an upcoming complementary approach for metagenomics and meta-transcriptomics.

Access this article online

Quick Response Code:

Website: www.jomfp.in

DOI: 10.4103/jomfp.jomfp_455_21

How to cite this article: Deo PN, Deshmukh RS. Oral microbiome research – A Beginner's glossary. J Oral Maxillofac Pathol 2022;26:87-92.

It is used to analyze the function of microbial communities. The term "metaproteomics" was defined in 2004 as "the large-scale characterization of the entire protein complement of environmental microbiota, at a given point in time." It is a dynamic tool to study the presence and abundance of proteins in oral microbiome samples.[6]

METATAXONOMICS

Amplicon metataxonomics generally target 16S rRNA genes because its sequence is similar enough across the microbiome taxa to be amplified by universal polymerase chain reaction (PCR) primers and also distinct enough to be used for taxonomic classification of species.[7]

16S RIBOSOMAL RNA GENES

16S rRNA gene is a gene which is invariably present in all prokaryotic organisms.[8]

It is around 1600 base pairs in length and contains nine hypervariable regions (V1-V9) that can be used for bacterial identification.[9]

ALPHA DIVERSITY

Alpha diversity is the diversity within a sample. For example, saliva sample. The three alpha diversity indices used commonly used in research are Chao 1 index, Shannon-Wiener index and Simpson index.[10]

BETA DIVERSITY

Beta diversity describes differences in the microbiota in between samples or groups. It is basically used to study whether the differences between in the microbiota compositions in between the groups are significant. The two common indices to measure beta diversity are Bray-Curtis dissimilarity and UniFrac distance.[10]

RICHNESS AND EVENNESS OF SPECIES

Richness is defined as a measure of various kinds of micro-organisms in a particular community. Evenness compares the similarity (homogeneity) of the population size of each species.[11]

PIPELINE

It is a defined sequence of processing steps that is used to the conversion of raw data into meaningful data.[12]

DNA SEQUENCING

DNA sequencing is a process of determining or identifying the exact order of nucleotides sequence (adenine, guanine, cytosine and thymine) in a DNA.

NEXT-GENERATION SEQUENCING

Next-generation sequencing (NGS) is a comprehensive method used to describe:[13]

1. Template preparation for the genomic DNA for downstream analysis
2. Generation of millions or billions of short DNA sequences called reads in a massively parallel manner
3. Alignment of the reads to sequences from known database
4. Assembling of the aligned sequences and discovery of new genetic variants.

Different NGS platforms are available for performing the sequencing of millions of DNA fragments. It is a high throughput method. Individual fragments of DNA are mapped to the reference databases and analyzed by bioinformatics.[14]

AMPLICON SEQUENCING

It is the ultra-deep sequencing of PCR amplification products for analyzing of the genetic variations.[15]

DE NOVO SEQUENCING

De novo sequencing is the generating of the first genetic sequence for a micro-organism which does not have any prior sequence data.[12]

WHOLE GENOME SEQUENCING

It is an alternative approach to 16S rRNA sequencing. It uses random primers to sequence overlapping regions of a genome. The taxa are more accurately defined at the species level using whole-genome sequencing (WGS). WGS requires extensive data analysis.[16]

SHOTGUN SEQUENCING

Shotgun sequencing is a process in which a long DNA molecule is randomly broken into fragments which are sequenced. Each DNA fragment is from a different source in a long DNA molecule.[17]

DNA AMPLICONS

DNA amplicons are sections/fragments of DNA which are the products of amplification. PCR is the most important

method for amplicon generation. These amplification products are then sequenced and compared with known microbiome databases.[18]

PCR amplification produces around thousands to millions of amplicons of the target DNA. These amplicons are then sequenced using high-throughput sequencing and nucleotide sequences called as reads are obtained.[19]

READS

Shotgun and NGS procedure involves shredding of the genomic DNA into smaller pieces/fragments which are then sequenced. The raw sequenced fragments are known as reads.[20]

FRAGMENT READ

A read which is produced from a fragment library. They are generated from single end of a small fragment of DNA in the order of 100–500 base pairs based on the sequencing platform. Fragment paired-end reads– These are two reads which are produced from each end of DNA fragment from a fragment library. Mate-paired read– They are two reads formed from each end of a large fragment of DNA with a predefined size range.[21]

COVERAGE

The number of times the sequenced nucleotide bases are covered by the target genome. E.g.,– ×30 coverage means that every base pair from the reference genome was covered by approximately 30 reads.[21]

DNA BARCODE

DNA barcode is a DNA sequence which is used for the identification of a target molecule during DNA sequencing. DNA barcode libraries are classified into two groups– randomly generated libraries and rationally designed libraries. Randomly generated libraries are produced by physically assembling oligonucleotides in the pool. Rationally generated libraries are designed using computer modeling (*in silico*) and then manufactured.[22]

The fragments of DNA sequences which enable to identify unknown species are called as DNA barcodes and the process is described as DNA barcoding.[23]

ADAPTORS/ADAPTOR SEQUENCES

They are short oligonucleotide sequences which are ligated at the ends of DNA fragments of interest. This is done to combine with primers for amplification. This is a part of library preparation.[24]

ADAPTOR TRIMMING

Adaptor trimming is an essential step for analyzing NGS data when reads are more in length than the target DNA/RNA fragments. Short oligonucleotides called adapter sequences are ligated to the ends of DNA fragments of interest so that primers can be used to amplify them. The adapter sequence is read out, sometimes partially, next to the unknown target DNA sequence when the sequencing read length is greater than that of the target DNA. It is critical to identify and trim the adapter sequence to recover the target DNA sequence.[24]

LIBRARY PREPARATION

The conventional NGS preparation protocol consists of three basic steps:[25]

1. Fragmentation– It is the first step in library preparation. The DNA molecules are mechanically or enzymatically fragmented into small uniform fragments around 200–400 base pairs
2. Adaptor ligation– The sequencing adaptors are ligated (tied) to the fragments
3. Amplification– After PCR amplification, the DNA library is set to go through many quality control steps to be loaded into the NGS machine.

A good library preparation is of utmost importance for generating good sequence depth and coverage. Different methods are available to achieve this goal.[26]

RAREFACTION

Rarefaction is a method for adjusting the differences in library sizes across samples in order to make alpha diversity comparisons easier. Sanders in 1968 proposed rarefaction, which entails selecting a number of samples equal to or less than the number of samples in the smallest sample, then discarding reads from larger samples at random until the number of samples remaining is equal to this threshold. Diversity metrics can be calculated based on these equal-sized subsamples to compare the ecosystems 'fairly', regardless of sample size differences. [27]

FASTQ

It is the most common output sequence data format from NGS platforms. It is a text-based format. FASTA format– The FASTA format is a format for storing DNA and amino acid sequences. A FASTA file starts

with a single line that describes the sequence. The 'greater' symbol at the beginning of the line distinguishes the description lines from the sequence lines. It is recommended that no more than 80 characters be used for definitions in the standard. The name or a unique identifier for the sequence, as well as other information, is usually included in the description line. Although the structure of this header and the information it contains are not standardized, each database sequence has its own FASTA header.[28]

SEQUENCE ALIGNMENT

It is a process in which a short DNA sequence read generally <250 bp is aligned with a reference genome. This procedure assigns a Phred quality score to each sequence read which indicates the confidence of the alignment process. This step can also be used to calculate the proportion of the mapped reads and the depth of sequencing for one or more loci of interest in the sequenced region. The data are stored in a standard BAM file format (binary alignment map) which is the binary version of MAP format.[29]

DNA ASSEMBLY

DNA assembly is defined as the regeneration of a genome from the large number of short overlapped fragments (reads) obtained by a sequencing machine. The length of every read and the number of reads are determined by the type of sequencer.[30]

PHRED SCORE

A score assigned to each base of a raw sequence in the sequencing platforms is the Phred score. The scores are determined by using predictors of possible errors.[31]

The Phred score is useful for filtering and trimming of sequences.[32]

Illumina reads are typically 25-250 nucleotide long sequences generated in the sequencing machine by a reversible-terminator cyclic reaction linked to base-specific colorimetric signals. Reads can be "single reads" or "paired reads", in which case they represent both ends of the same nucleotide fragment (generally 200-1000 bp long). An internal Illumina software (CASAVA) converts these colorimetric signals into base calls in the FASTQ format. Each nucleotide is associated with an ASCII-encoded quality number corresponding to a PHRED score (Q), which is directly translated into probability P that the corresponding base call is incorrect using the following equation.[33]

CHIMERA

Chimeras are hybrid products of multiple parent sequences that are misinterpreted as new organisms, inflating the appearance of diversity.[34]

Chimeras, which are caused by incomplete template extension and appear to be recombination between dissimilar sequences can lead to inflated diversity.[35]

Some of the amplified sequences can be produced by multiple parent sequences during the PCR amplification process, resulting in chimeras. Chimeric sequences are important for alpha diversity estimates, even though they are technical artifacts rather than actual members of the community.[36]

OPERATIONAL TAXONOMIC UNITS

Operational taxonomic units (OUTs) are common currency of marker gene or 16S rRNA gene studies.[37] OTU Table– Marker gene sequence reads are typically clustered based on sequence similarity, with the assumption that sequences with greater similarity represent more phylogenetically similar organisms, to facilitate taxonomy-independent analyses and to reduce the computational resources required for such analyses. These clusters, also known as OTUs, are a common analytical unit in microbial ecology.[38]

ANNOTATION

Genome annotation entails attaching biologically relevant information to genome sequences by analyzing their structure and composition, as well as taking into account what we know from closely related species that can be used as a reference.[39]

It is the process of identifying functional elements along with a genome's sequence and thus giving it a meaning. It is required because DNA sequencing generates sequences with unknown functions.[40]

BLAST

Blast stands for-Basic local alignment search tool. It is the most commonly used tool for the calculation of sequence similarity. Different variations of BLAST are available for different sequence comparisons. E.g., -DNA query to a DNA database, a protein query to a protein database.[41]

DENOISING

Denoising aims to carry out filtering of the noisy reads, reduces repetition, remove singletons, chimeric sequences

and correction of errors in marginal sequences. This is a prerequisite step, before clustering. OTU clustering.[42]

CLADOGRAM

Cladogram is defined broadly as – 'any branching diagram, graph or written statement that depicts the relationship between three or more taxa.[43]

INTERACTIVE TREE OF LIFE

It is a web-based application for viewing, manipulating and annotating phylogenetic trees. iTOL was one of the first tools to allow trees to be annotated with various types of extra data.[44]

PHYLOGENY (PHYLOGENETIC TREE)

It is a graphical representation of hypothesized relationships based on genetic differences between sequences.[45] It is a diagram that depicts the relations between taxa (or sequences) and their presumed common ancestors (Nei and Kumar 2000; Felsenstein 2004; Hall 2011). The majority of phylogenetic trees today are based on molecular data, such as DNA or protein sequences. The goals of today's phylogenetic trees include understanding the relationships among the sequences without regard to the host species and inferring the functions of genes that haven't been experimentally studied (Hall *et al*. 2009), There are four steps to constructing a phylogenetic tree: (Step 1) find and acquire a set of homologous DNA or protein sequences, (Step 2) align those sequence data (Step 3) estimate a tree from the aligned sequences and (Step 4) present that tree in such a way that the relevant information is clearly conveyed to others.[46]

CONCLUSION

There is a whole set of new terminologies which a researcher comes across while planning a microbiome study. It is important to use precise terminologies in research work with a clear understanding of its meaning. This article will assist in relating the taxonomy and functionality of the oral microbiome. Hence an attempt of this article for beginners as a guide for oral microbiome research.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

REFERENCES

1. Deo PN, Deshmukh R. Oral microbiome: Unveiling the fundamentals. J Oral Maxillofac Pathol 2019;23:122-8.
2. Marchesi JR, Ravel J. The vocabulary of microbiome research: A proposal. Microbiome 2015;3:31.
3. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform 2021;22:178-93.
4. Aguiar-Pulido V, Huang W, Suarez-Ulloa V, Cickovski T, Mathee K, Narasimhan G. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. Evol Bioinform Online 2016;12 Suppl 1:5-16.
5. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. Bioinform Biol Insights 2016;10:19-25.
6. Bostanci N, Grant M, Bao K, Silbereisen A, Hetrodt F, Manoil D, *et al.* Metaproteome and metabolome of oral microbial communities. Periodontol 2000 2021;85:46-81.
7. Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, *et al.* A robust framework for microbial archaeology. Annu Rev Genomics Hum Genet 2017;18:321-56.
8. Duran-Pinedo AE, Frias-Lopez J. Beyond microbial community composition: Functional activities of the oral microbiome in health and disease. Microbes Infect 2015;17:505-16.
9. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data 2019;6:190007.
10. Qian XB, Chen T, Xu YP, Chen L, Sun FX, Lu MP, *et al.* A guide to human microbiome research: Study design, sample collection, and bioinformatics analysis. Chin Med J (Engl) 2020;133:1844-55.
11. Kim BR, Shin J, Guevarra R, Lee JH, Kim DW, Seol KH, *et al.* Deciphering diversity indices for a better understanding of microbial communities. J Microbiol Biotechnol 2017;27:2089-93.
12. Harison N, Kidner CA. Next generation sequencing and systematics: What can a billion base pairs of DNA sequence data do for you. J Int Assoc Plant Taxonomy 2011;60:1552-66
13. Dewey FE, Pan S, Wheeler MT, Quake SR, Ashley EA. DNA sequencing: Clinical applications of new DNA sequencing technologies. Circulation 2012;125:931-44.
14. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed 2013;98:236-8.
15. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect 2018;24:335-41.
16. 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;469:967-77.
17. Weinstock GM. Genomic approaches to studying the human microbiota. Nature 2012;489:250-6.
18. Washburne AD, Morton JT, Sanders J, McDonald D, Zhu Q, Oliverio AM, *et al.* Methods for phylogenetic analysis of microbiome data. Nat Microbiol 2018;3:652-61.
19. Calle ML. Statistical analysis of metagenomics data. Genomics Inform 2019;17:e6.
20. Li W, Freudenberg J. Mappability and read length. Front Genet 2014;5:381.
21. Yegnasubramanian S. Explanatory chapter: Next generation sequencing. Methods Enzymol 2013;529:201-8.
22. Lyons E, Sheridan P, Tremmel G, Miyano S, Sugano S. Large-scale DNA barcode library generation for biomolecule identification in high-throughput screens. Sci Rep 2017;7:13899.
23. Chaudhary DK, Dahal RH. DNA bar-code for identification of microbial communities: A mini-review. EC Microbiol 2017;7:219-24.
24. Jiang H, Lei R, Ding SW, Zhu S. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC

Bioinformatics 2014;15:182.
25. Feng K, Costa J, Edwards JS. Next-generation sequencing library construction on a surface. BMC Genomics 2018;19:416.
26. Kulski JK. Next Generation Sequencing – An Overview History, Tools and "Omic" Applications. In Next Generation Sequencing - Advances, Applications and Challenges. InTech. 2016. [https://doi.org/10.5772/61964].
27. Willis AD. Rarefaction, alpha diversity, and statistics. Front Microbiol 2019;10:2407.
28. Carpentieri B. Next generation sequencing data and its compression. IOP Conf Ser Earth Environ Sci 2019;362:012059.
29. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, *et al*. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn 2018;20:4-27.
30. Weitschek E, Santoni D, Fiscon G, De Cola MC, Bertolazzi P, Felici G. Next generation sequencing reads comparison with an alignment-free distance. BMC Res Notes 2014;7:869.
31. Liao P, Satten GA, Hu YJ. PhredEM: A phred-score-informed genotype-calling approach for next-generation sequencing studies. Genet Epidemiol 2017;41:375-87.
32. Pereira R, Oliveira J, Sousa M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J Clin Med 2020;9:132.
33. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One 2013;8:e85024.
34. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, *et al*. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 2011;21:494-504.
35. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, *et al*. Conducting a microbiome study. Cell 2014;158:250-62.
36. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, *et al*. Advancing our understanding of the human microbiome using QIIME. Methods Enzymol 2013;531:371-444.
37. Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: Attempting to find consensus "Best Practice" for 16S microbiome studies. Appl Environ Microbiol 2018;84:e02627-17.
38. He Y, Caporaso JG, Jiang XT, Sheng HF, Huse SM, Rideout JR, *et al*. Stability of operational taxonomic units: An important but neglected property for analyzing microbial diversity. Microbiome 2015;3:20.
39. Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, *et al*. Ten steps to get started in Genome assembly and annotation. F1000Res 2018;7:R-148.
40. Abril, JF, Castellano Hereza S. Genome annotation. In: Ranganathan S, Gribskov M, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology. Amsterdam, Netherlands: Elsevier; 2019. p. 195-209S.
41. Madden, Tom. The BLAST sequence analysis tool. The NCBI Handbook; 2002.
42. Kamble A, Sawant S, Singh H. 16S ribosomal RNA gene-based metagenomics: A review. Biomed Res J 2020;7:5-11.
43. Brower AVZ. What is a cladogram and what is not? Cladistics 2016;32:573-6.
44. Letunic I, Bork P. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 2016;44:W242-5.
45. Balloux F, Brønstad Brynildsrud O, van Dorp L, Shaw LP, Chen H, Harris KA, *et al*. From theory to practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. Trends Microbiol 2018;26:1035-48.
46. Barry G. hall, building phylogenetic trees from molecular data with mega. Mol Biol Evol 2013;30:1229-35.