Contents

[1. Introduction 2](#_Toc71764641)

[2. Metagenomics 3](#_Toc71764642)

[2.1. DNA Extraction 3](#_Toc71764643)

[2.1.1. Sampling 3](#_Toc71764644)

[2.1.2. Extraction methods 3](#_Toc71764645)

[2.2. Sequencing 4](#_Toc71764646)

[2.2.1. Shotgun sequencing 4](#_Toc71764647)

[2.2.2. Amplicon (16S RNA) sequencing 5](#_Toc71764648)

[2.3. Assembly 6](#_Toc71764649)

[2.3.1. Reference-based assembly 6](#_Toc71764650)

[2.3.2. De Novo assembly 7](#_Toc71764651)

[2.4. Binning 8](#_Toc71764652)

[3. Metatranscriptomics 9](#_Toc71764653)

# Introduction

Microbiomes are omnipresent and they are found in and on different places like in the soil, ocean and on many living organisms as well. Changes in the habitat of a microbiome cause a disrupting effect on the well-being of the environment in which they reside. For example, Human gut is a habitat for different microbiomes and changes in these can affect the health of the gut. Below figure-1 [6] explains different branches of study on Microbial community. Metagenomics is the study of the genetic material of these microbiomes recovered directly from environmental samples. Metatranscriptomics is the study of the diversity of the active genes within such community.

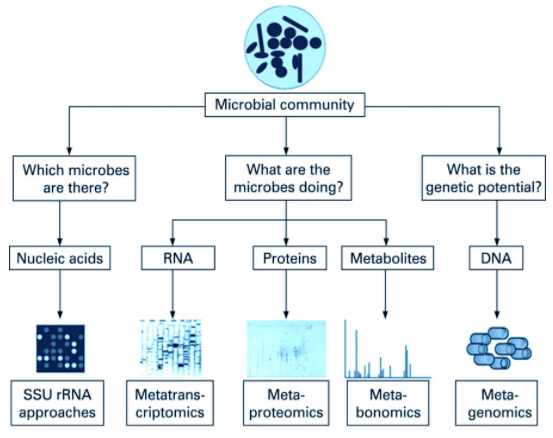


Figure – 1: Branches of study on Microbial community [6]

Early studies focused on 16S ribosomal RNA (rRNA) sequences which are relatively short, often conserved within a species, and generally different between species. Multitude of 16S rRNA sequences have been found that do not belong to any discovered cultured species and there are many such homogenous organisms in a sample. Conventional methods of genetic sequencing require culturing of identical cells as a source of DNA. Large groups of micro-organisms, such as microbiomes, cannot be cultured and sequenced and setbacks such as these made cultivation-based methods elude vast majority of microbial diversity and detect less than 1% of the bacterial and archaeal species in a sample.

Metagenomics and Metatranscriptomics focus on microbial communities to analyze the microbial DNA that is extracted directly from communities in environmental samples. These processes alleviate the need for isolation and cultivation of individual species. These techniques determine whole gene expression profiling of complex microbial communities and provide information about differences in the active functions of microbial communities which appear to be the same in terms of their composition.

# Metagenomics

Metagenomics perform Taxonomic and functional analysis of the microbiome. A typical metagenomic process can be broadly segregated in to four steps, Extraction, Sequencing, Assembly and Binning.

## DNA Extraction

Extraction of microbial DNA from an environment sample is vital step that can affect all further steps in metagenomics. This is different to conventional extraction methods that are used for animal or plant tissues due to the uncertainty in quantity and diversity of the microbial species present in the sample. An explicit sample from environment is a black box due to this challenge and make us rely on prior experience and publicly available sequence data.

### Sampling

Method of identifying an environment and collection of a microbial sample from it is called as sampling. There are many factors to be considered that affect the final sequence outcome from a sample. These include,

* From where the sample was collected – e.g., Human, water, air, or surface.
* How the sample was collected – e.g., Swabbing, scrapping, biopsy, or bulk substrate collection.
* How much quantity was collected and how the sample will be stored after collection.
* How long it was stored prior to extraction – e.g., temperature and environmental conditions.
* Method that was used for extraction.

### Extraction methods

For and unbiased estimate of microbial composition in a genomic analysis, choosing an extraction method that can recover both unicellular (prokaryotic) and multi-cellular (eukaryotic) DNA efficiently is instrumental, especially if our sequencing method is shot gun. Studies of DNA extracted from environmental samples are focusing on 16S rRNA sequence data analysis, it has been shown that due to differences in the cell wall and membrane structure of bacteria, effectiveness of DNA extraction can depend on the extraction protocol used. Morgan and coworkers (2010) suggested to use multiple DNA extraction procedures for a single sample to increase the likelihood of including every organism in the tested sample. If the target community is associated with a host, e.g., human or plant, then physical fractionation or selective lysis can be employed to ensure host DNA is kept to a minimum. Host material can also be removed during bioinformatics filtering and mapping. Regardless of the approach used, it is important to remember that extraction and isolation methods can introduce bias in terms of microbial diversity, yield, and fragment lengths. It is highly recommended that the exact same extraction method be used when comparing samples. Following is the broad classification of extraction methods,

**Direct methods**: This method starts with disrupting the cell wall of the microbiome and muster nucleic acids from bacteria to the extraction buffer. We will separate the extraction buffer from the soil particles and nucleic acids are isolated from the extraction buffer avoiding contaminants such as humus acids, heavy metal ions and proteins. Using of extraction buffer has a downside of being a choice between expected DNA quantity and purity.

**Indirect methods**: In this method we first disperse the soil matrix to separate bacterial cells, with high DNA quality, encompassing full diversity of microbial life in the sample. To disperse the soil matrix, one can use physical or chemical methods. We then isolate and purify the extracted DNA sample.

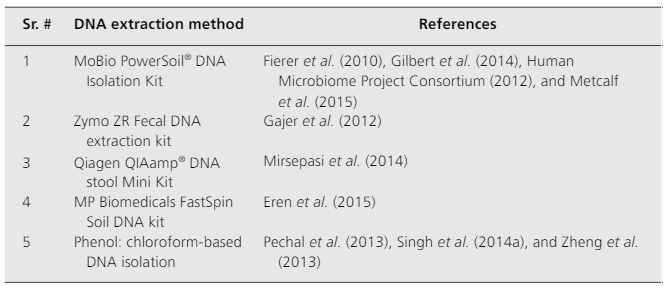


Table 2.1 Commonly used DNA extraction methods [8]

## Sequencing

Sequencing DNA is the process of determining the order of the chemical building blocks, called bases, that make up the DNA molecule. Sequence information can be used to determine the stretches that contain genes, and the ones that contain regulatory instructions. Due to the length of bacterial genomes, 0.5 – 10 Mega base pairs (Mbp), it is impossible to sequence the entire genome in one reaction. Instead, small pieces called ‘reads’ are sequenced first. We have below two main types of sequencing techniques for metagenomics,

### Shotgun sequencing

This is performed for taxonomic profiling (diversity and abundance), as well as functional analysis. This complex technique allows for parallel sequencing of DNA from all organisms within the community, with high coverage for species-level detection. This method fragments the DNA into many small random pieces and can read all genomic DNA in a sample, rather than just one specific region of DNA. For microbiome studies, this means that shotgun sequencing can identify and profile bacteria, fungi, viruses, and many other types of microorganisms at the same time. As genomes are sequenced, it is also possible to identify and profile microbial genes that are present in the sample (the metagenome), which provide additional information about microbiome functional potential. Below are the steps that briefly explain the process,

* Extract DNA from your sample.
* Tagmentation, a process which cleaves and tags DNA with adapter sequences, priming the fragmented DNA for ligation of molecular barcodes.
* Clean up the fragmented DNA sample to remove reagent impurities from Tagmentation.
* Perform PCR to amplify DNA samples, as well as adding molecular barcodes to each sample.
* Size selection and remove the DNA to remove impurities after the PCR steps.
* Muster samples together in equal proportions and perform Library quantification of the pooled samples.
* Sequence mustered samples.

### Amplicon (16S RNA) sequencing

Amplification and sequencing of targeted marker loci (e.g., 16S rDNA for bacteria/archaea, 18S rDNA for eukaryotes, fungi) is currently the most common culture‐independent molecular method for the detection and characterization of the microbial community structure from a particular environment. With this method, a marker locus of a target community is amplified directly from the extracted DNA using specific universal PCR primers, and the amplified product is then sequenced in parallel (no traditional cloning step is needed before sequencing) on a next‐generation sequencing platform of choice. Below are the steps that briefly explain the process,

* Extract DNA from your sample and perform PCR on your DNA sample to amplify one or more selected hypervariable regions (V1-V9) of the 16S rRNA gene, as well as adding molecular ‘barcodes’ to each cleaned DNA sample (to multiplex multiple samples)
* Clean up and remove impurities from the amplified DNA.
* Muster samples together in equal proportions for Library quantification.
* Sequence mustered samples.

**Few important factors to compare:**

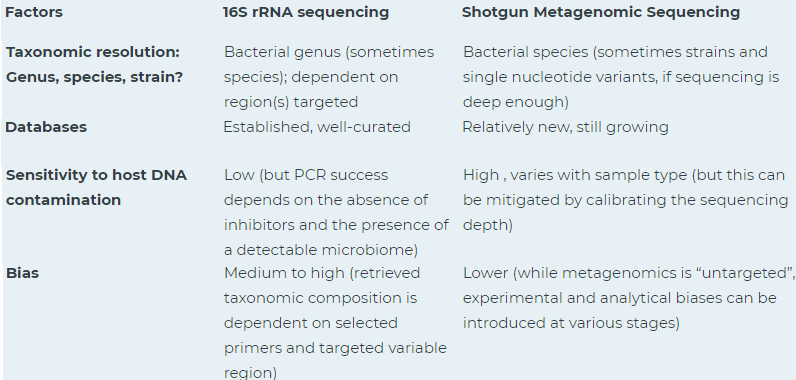


Table 2.2 Comparison factors for metagenomic sequencing. [9]

## Assembly

Assembly is the process of combining sequence reads into contiguous stretches of DNA called contigs, based on sequence similarity between reads. The consensus sequence for a contig is either based on the highest-quality nucleotide in any given read at each position or based on majority rule. The main challenge in metagenomic assembly arises from the heterogeneous nature of metagenomic data. Most environments contain an uneven representation of the member species, and furthermore, the organisms in the environment frequently belong to clusters of closely related strains whose genomes are largely similar but differ due to mobile genetic elements and point mutations.

These characteristics of the data make it virtually impossible to construct a single assembly of each organisms present in a sample, instead many organisms will be under-sampled and will be assembled in a highly fragmented form, while groups of closely related organisms will end up assembled into a polymorphic structure that can be modeled as a computational graph. Below are the two types of assembly that can be employed for metagenomics,

### Reference-based assembly

This method refers to performing an assembly where the input files would be reads from multiple samples. This contrasts with doing an independent assembly for each sample, where the input for each assembly would be just the reads from that individual sample. This works well if the metagenomic dataset contains sequences where closely related reference genomes are available. However, differences in the true genome of the sample to the reference, such as a large insertion, deletion, or polymorphisms, can mean that the assembly is fragmented or that divergent regions are not covered. Below are some advantages of this method,

* Higher read depth (this can allow to have a more robust assembly that captures more of the diversity in your system, but not always).
* Facilitates the comparison across samples by giving you one reference assembly to use for all.
* Substantially improves the ability to recover genomes from metagenomes due to the awesome power of differential coverage (this concept of using coverage to recover genomes is shown in the figure and slides available above with one sample, but as noted really becomes powerful with multiple)

### De Novo assembly

De Novo refers to assembling a novel genome where there is no reference sequence available for alignment. Sequence reads are assembled as contigs, and the coverage quality of de novo sequence data depends on the size and continuity of the contigs (i.e., the number of gaps in the data). This method requires larger computational resources. Below are the three main paradigms De Novo assembly,

**Greedy:** This is the most simple and intuitive method of assembly. In this method, individual reads are joined together into contigs in an iterative manner starting with the reads that overlap best and ending once no more reads or contigs can be merged.

**Overlap-layout-consensus:** This is a three-step approach begins with a calculation pairwise overlaps between all pairs of reads. The overlaps are computed with a variant of a dynamic programming-based alignment algorithm, making assembly possible even if the reads contain errors. Using this information, an overlap graph is constructed where nodes are reads and edges denote overlaps between them. The layout stage consists of a simplification of the overlap graph to help identify a path that corresponds to the sequence of the genome. More precisely, a path through the overlap graph implies a 'layout' of the reads along the genome.

**De Bruijn Graph:** The de Bruijn graph assembly paradigm focuses on relationship between substrings of fixed length k (k-mers) derived from the reads. The k-mers are organized in a graph structure where the nodes correspond to the k-1 prefixes and suffixes of k-mers, connected by edges that represent the k-mers. In this approach reads are not explicitly aligned to each other, rather their overlaps can be inferred from the fact that they share k-mers. With this graph, assembly problem reduces to finding a Eulerian path – a path through the graph that visits each edge once.

## Binning

This is the process of segregating of reads and contigs, obtained from sequence and assembly process, and assigning them to individual genome of the microbiome sample. Binning assembled sequences into individual groups, which represent microbial genomes, is the key step and a major challenge in metagenomic research. Both supervised and unsupervised machine learning methods have been employed in binning. Binning approach can be divided into taxonomic-dependent binning and taxonomic-independent binning, also called taxonomy binning and genome binning, respectively. Taxonomy binning is a supervised method to compare metagenomic sequences against a database of genomic sequences by making use of aligning algorithms such as blast, bowtie, bwa, minimap or pre-computed databases (k-mers) of previously sequenced microbial genetic sequences.

**Life cycle:**

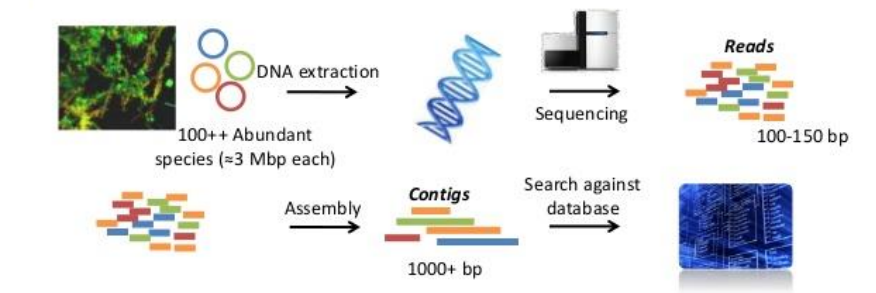


Figure – 2 Life cycle of Metagenomics [6]

# Metatranscriptomics

…