2021

Metagenomics & Metatranscriptomics

EECS 730 MINI REVIEW

PEDURI MADHU

Contents

1. 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.

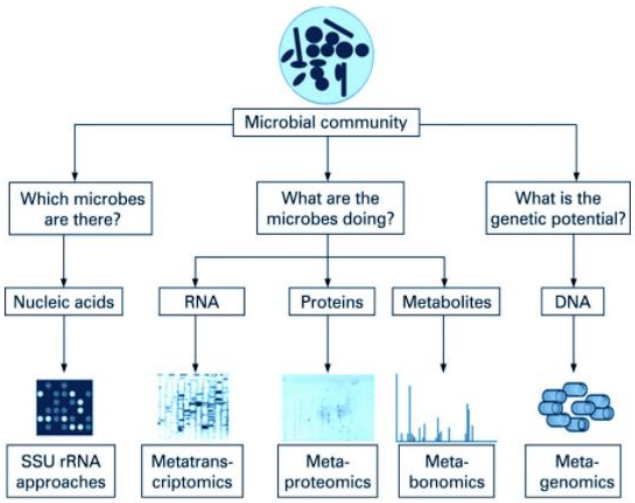

Figure – 1: Branches of study on Microbial community [6]

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.

2. 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.

2.1. 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.

2.1.1. 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.

2.1.2. 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.

Sr. #	DNA extraction method	References
1	MoBio PowerSoil® DNA Isolation Kit	Fierer *et al.* (2010), Gilbert *et al.* (2014), Human Microbiome Project Consortium (2012), and Metcalf *et al.* (2015)
2	Zymo ZR Fecal DNA extraction kit	Gajer *et al.* (2012)
3	Qiagen QIAamp® DNA stool Mini Kit	Mirsepasi *et al.* (2014)
4	MP Biomedicals FastSpin Soil DNA kit	Eren *et al.* (2015)
5	Phenol: chloroform-based DNA isolation	Pechal *et al.* (2013), Singh *et al.* (2014a), and Zheng *et al.* (2013)

Table 2.1 Commonly used DNA extraction methods [8]

2.2. 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,

2.2.1. 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.

2.2.2. 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:

Factors	16S rRNA sequencing	Shotgun Metagenomic Sequencing
Taxonomic resolution: Genus, species, strain?	Bacterial genus (sometimes species); dependent on region(s) targeted	Bacterial species (sometimes strains and single nucleotide variants, if sequencing is deep enough)
Databases	Established, well-curated	Relatively new, still growing
Sensitivity to host DNA contamination	Low (but PCR success depends on the absence of inhibitors and the presence of a detectable microbiome)	High , varies with sample type (but this can be mitigated by calibrating the sequencing depth)
Bias	Medium to high (retrieved taxonomic composition is dependent on selected primers and targeted variable region)	Lower (while metagenomics is "untargeted", experimental and analytical biases can be introduced at various stages)

Table 2.2 Comparison factors for metagenomic sequencing. [9]

2.3. 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,

2.3.1. 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)

2.3.2. 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.

2.4. 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:

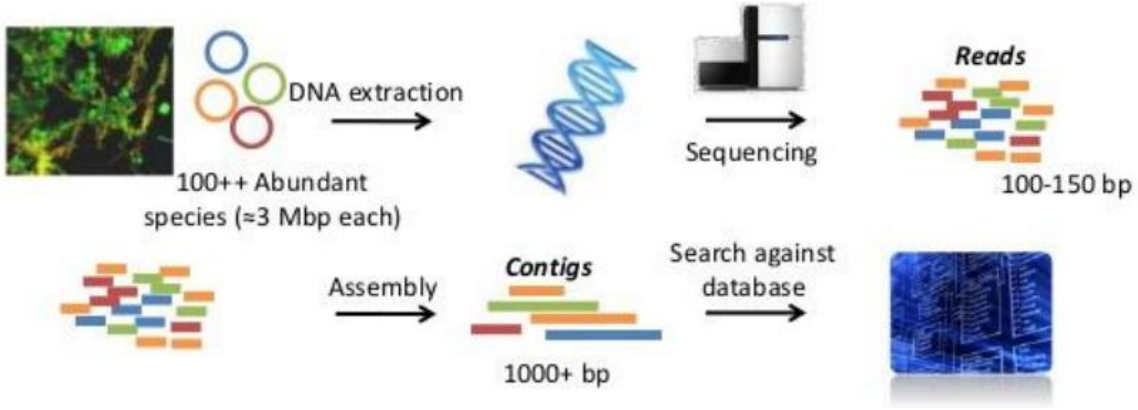

Figure – 2 Life cycle of Metagenomics [6]

3. Metatranscriptomics

Metatranscriptomics analysis helps us comprehend how the microbiome responds to the environment by studying the functional analysis of genes expressed by the microbiome. It can also estimate the taxonomic composition of the microbial species. It furnishes scientists with the confirmation of predicted open-reading frames (ORFs) and potential identification of novel sites of transcription and/or translation from microbial genomes. Since this branch focuses on what genes are expressed, it allows to understand the active functional profile of the entire microbial community. The overview of the gene expression in each sample is obtained by capturing the total mRNA of the microbiome. Metatranscriptomics can enable more complete generation of protein sequences databases for Metaproteomics. A generic pipeline of this process consists of steps – RNA Purification, Sequencing, Assembly and Statistical analysis.

3.1. RNA Purification

Conventional DNA isolation methods cannot be directly applied to RNA due to differences in its structure, RNA is a single stranded while DNA is double stranded. Presence of RNases, a group of enzymes that degrade RNA molecules, are abundant in the environment, including on hands and on surfaces and it is difficult to remove/destroy RNases completely. Therefore, assiduous aseptic approaches should be used to isolate RNA. Below are two sample procedures for RNA extraction / purification,

Sampling: The samples should be obtained quickly and aseptically and should be processed immediately or snap frozen. Generally, samples are frozen directly on the field using either liquid nitrogen, or dry ice/acetone to stop metabolism without damaging cell structures, however when samples are thawed RNases will be active. When planning sampling we should anticipate how to stabilize RNA because this is done before freezing.

Stabilization: All cells have intracellular RNase, the mRNA in bacteria generally have a few minutes life span so RNA can be degraded while purified. Moreover, transport and purification can induce the synthesis of new mRNA changing expression profiles. Several reagents may serve to inactivate endogenous RNase. The simplest is to start with the isolation process before freezing adding guanidinium thiocyanate–phenol–chloroform solution, commercially known as TRIzol or Qiazol. One of the most popular stabilizers is RNAlater containing EDTA, sodium citrate, and ammonium sulfate, it is used for all cell types and has been tested in bacteria. RNAprotect is another stabilizer designed for bacteria.

RNA Enrichment: Below are two sample methods for enrichment.

- One careful approach is to use probes that identify specific mRNA regions that are attached to magnetic beads. This process involves annealing of probes to target sequences (rRNA) followed by their removal with the use of a magnet.
- Use the RNeasy preparation method (Qiagen), the phenol-chloroform based TRIzol procedure, and separate the nuclear and the cytoplasmic fraction using the non-organic extraction kit PARIS (Life Technologies). All RNA extractions should be performed in duplicates. One can detect nuclear un spliced RNAs (bioanalyzer peak >4000 nt) in TRIzol RNA, Qiagen RNA and nuclear RNA using this process.

3.2. Library preparation

Purpose of library preparation is to produce cDNA of certain size that is flanked by adapters. So, library preparation requires fragmenting the RNA, first strand synthesis, second strand synthesis, coupling adapters, and validating the library. Sequence service providers can perform the library preparation. cDNA should be of a certain size to optimize sequencing, depending on the platform is the size fragments must be. Fragmentation can be done with enzymes, metals, heat, or sonication. Incubation times for fragmentation must be optimized for each case, as the integrity of each sample is usually different.

We have different tools that perform library preparation like TruSeq Stranded, Total RNA Library Preparation from Illumina (TS), SMARTer Stranded RNA-Seq Kit from Clontech (SMART), the Ovation RNA-Seq System V2 (OV) and the Encore Complete Prokaryotic RNA-Seq System.

3.3. RNA Sequence

We follow the below steps to perform sequencing for libraries that were prepared.

- We first split the libraries into individual files, this is also known as de-multiplexing. If we are using barcodes to mix several samples in a single run, then samples are split based on its barcode sequence.
- We then remove sequencing adapters. Removing these templates for the sequencing is important and could help with sequence assembly.
- **Sequence trimming:** Each sequenced base has its own quality value, which is known as Phred score that serves as a proxy probability calculator. A Phred value of 30 accounts for 1 error in every 1000 bases or a 99.9 % of accuracy. This is a good standard to make a cut-off. Visualize the overall quality of your sequences via boxplots.
- **Filter rRNA:** A quick way to do this step is by using a rRNA DB and MegaBLAST. There are other strategies using Interpolated Markov Models like Infernal and SSU-align and will help at this stage.

3.4. Sequence Assembly

For RNA sequencing, we have two types,

Reference-based assembly: In this method, we should provide reference sequences. Both reference files and Metatranscriptomics FASTQ files should be indexed. After the alignment, we need to take the SAM/BAM resulting file and count the occurrence of each gene model (if available). The counting of each gene could be accomplished with R. R is a computer language intended for statistical computing and graphics.

De Novo Assembly: This method is applicable, when we do not have reference RNA data to proceed with the assembly. In this method, preprocessed, high-quality reads can now be assembled into putative transcripts using de novo assemblers. Given that most microbiomes are not adequately characterized with reference genomes, de novo assemblers provide a reference scaffold representing longer, expressed genome segments that can provide a reference set of genes. This provides users the ability to find homologs in a more straightforward fashion, establish taxonomic origin, and serve as a reference for expression analysis. The current state of de novo assembly for Metatranscriptomics datasets is still in its very early stages. Only a handful of tools have been specifically developed for Metatranscriptomics, but their efficacy on multiple datasets has not been tested and their hardware, or memory requirements across an array of community complexities and data volume, have also not been established.

Annotation: Genome annotation in the process of identifying functional elements along the sequence of a genome and thus categorizing it. In this we annotate each transcript with a hierarchical schema. If we know the species that is being compared and its annotated genome sequences, we could perform BLAST searches directly to them. If we have any sequences without homologs, we need to go up to the next hierarchy which is a bacterial DB.

3.5. Statistical Analysis

Statistical analysis is performed to understand the features, of microbiome, like how many different genes are expressed in that microbial community, the highest expressed genes in a specific environment and highest change in expression levels between different conditions. Below are few steps that can be performed,

- Build a count matrix. This could be done by counting the mapped reads or to segregate the sequences obtained from different experimental conditions according to their identity and count the number of occurrences in each sample/experiment. This step is required for parsing the annotation data to the Data Analysis pipeline.
- Transform your matrix to a common factor to accomplish differential expression. When measuring distances and sample similarities, it is better to apply normalizations like regularized-logarithm transformation (rlog) or DESeq, which uses a negative binomial distribution.
- Compute the distance on the r log/log 2 transformed data, Principal Component Analysis to assess sample/treatment similarity and visualize using heat maps. With the transformed matrix, we can describe the dissimilarity between samples/replicates/experiments by means of clustering analysis.
- Perform the differential expression analysis. In this point, we need to calculate the log 2-fold changes between our treatments (control vs. experiment).
- Perform multiple testing correction, to calculate the amount of false discovery rate (FDR) and then assess the significance of the adjusted p-value. This is to derive how much of false positives could be accepted.
- Visualize the amount of significant differentially expressed genes. We can do this by means of Volcano plots, and heatmaps.
- Using Ontology tools like SEED, COG, GO, KO, connect the most abundant features with its annotation.
- Make sense of the known and annotated genes to derive new working hypothesis about their gene expression under the experimental circumstances. The whole dataset of significant genes could be divided into two main groups: genes with known functions, and genes with unknown functions.
- For the genes with a known function, a process of data mining will be necessary to comprehend the functions and processes involving their participation. Upload your RNA-seq experiments to appropriate databases and repositories.
- For the genes with unknown function, design further experiments to discover their function (mutants, heterologous expression, etc.)

Typical workflow of Metatranscriptomics:

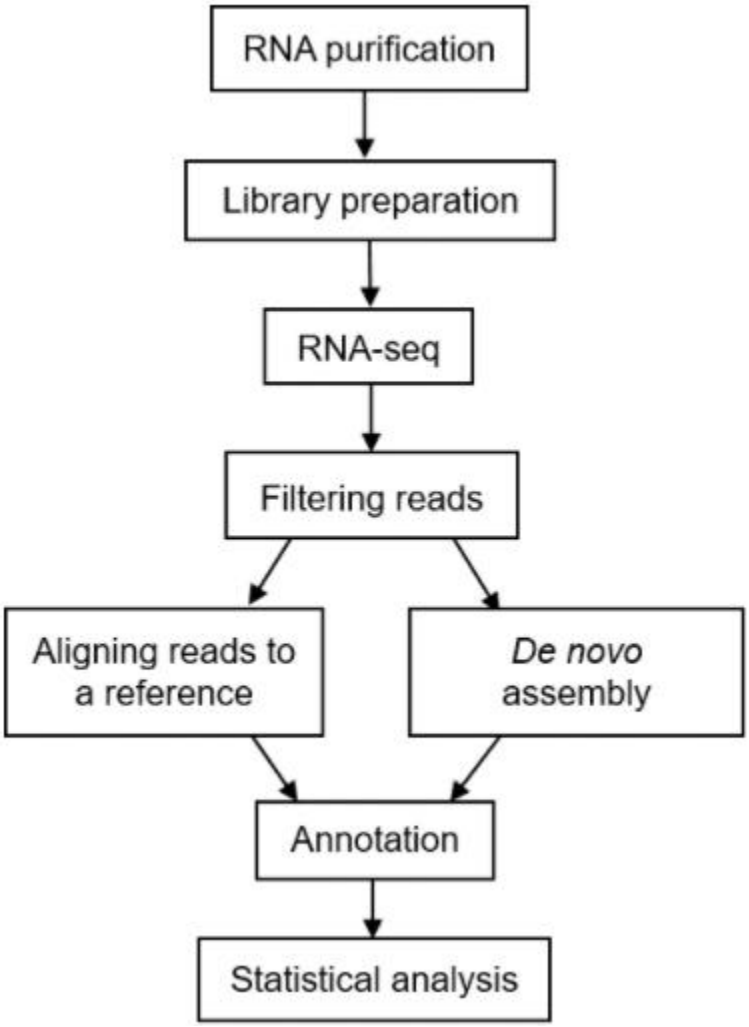

Figure 3.1 A typical workflow for Metatranscriptomics analysis [13]

4. Scope

Below is the short description of Metagenomics and Metatranscriptomics to outline the difference in their scopes.

Metagenomics: It is more about the taxonomical and functional analysis. Analysis that says which microbial species are accounted in the environmental sample. Mostly deals with DNA.

Metatranscriptomics: This is the study of the gene expression and activity of each species that were categorized under metagenomic analysis. Depends on comprehending mRNA of all the species involved in the environmental sample.

Microbial communities:

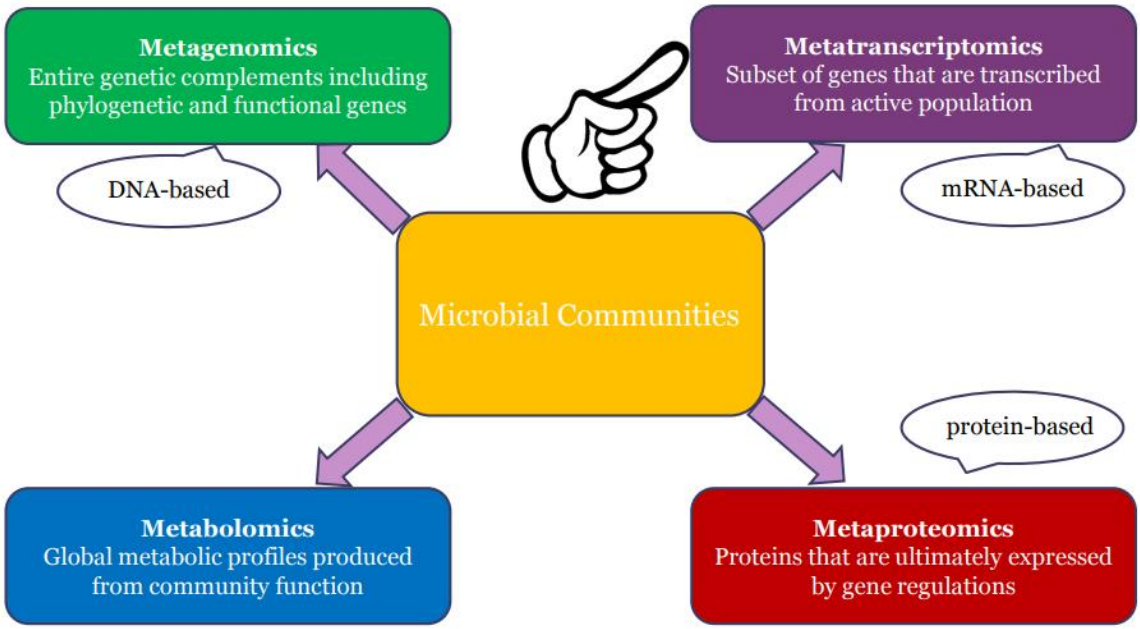

Figure 4.1 Brief scope of each microbial community [4]

4.1. Applications

Following are some of the applications of Metagenomics and Metatranscriptomics,

Human health: Symbiotic bacteria (normal flora) play a key role in protecting us from pathogens, but under certain conditions they can overcome protective host responses and trigger pathological effects. Microbial population analysis can be used as an indicator of an individual's health status and as a powerful tool for the prevention, diagnosis, and treatment of specific diseases.

Immune reactions: The effects of microbiota on the mucosal immune system are thought to be key to affect host physiology. A study of toll-like receptor 5 (TLR5) knockout (KO) mice is an interesting example of the use of Metatranscriptomics to complement metagenomic characterization and 16S rRNA gene profiling of this microbial immune interaction.

Environmental remediation: Metagenomics can improve strategies for monitoring the impact of pollutants on ecosystems and for cleaning up contaminated environments. Increased understanding of how microbial communities cope with pollutants improves assessments of the potential of contaminated sites to recover from pollution and increases the chances of bioaugmentation or bio stimulation trials to succeed.

Food industry: Metagenomics and Metatranscriptomics methods can be used to improve food quality, function, and safety, and provide information related to metabolic activities of microbial communities.

Conclusion

On an average 40 billion bacteria live on and within us which is approximately equal to the total number of cells in our body [8] and a human intestinal microbial ecosystem expresses 100 times more genes than the human host. These points emphasize the importance of taxonomical analysis and gene expression of a microbiome. Metagenomics and Metatranscriptomics together offer a great deal of insight in to the microbial eco system in an environmental sample. We have seen different steps in each process to get a glimpse of these areas and understand the scope of each area.

References

[1]https://en.wikipedia.org/wiki/Metagenomics

[2]https://arxiv.org/ftp/arxiv/papers/1911/1911.11304.pdf

[3]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4869604/

[4]https://www.sdstate.edu/sites/default/files/Metagenomics%20and%20Metatranscriptomics.pdf

[5]https://www.biorxiv.org/content/10.1101/307157v1.full.pdf

[6]http://users.metu.edu.tr/bicgen/research/envt.html

[7]http://www.actabp.pl/pdf/1_2015/151.pdf

[8]https://www.researchgate.net/publication/316740956_An_introduction_to_metagenomic_data_generation_analysis_visualization_and_interpretation

[9]https://blog.microbiomeinsights.com/16s-rrna-sequencing-vs-shotgun-metagenomic-sequencing

[10] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4839964/

[11] https://www.labome.com/method/RNA-Extraction.html

[12] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4148917/

[13] https://www.cd-genomics.com/the-principles-workflow-and-applications-of-metatranscriptomic-sequencing.html