Molecular and cellular biology

6/10

Meta-omics approaches to microbial communities' analysis

3 main omics approaches will be held by elio rossi

DNA-seq processes utilizes biochemical methods in order to determine the correct order of nucleotide bases in a DNA macromolecule using sequencing machines

- Sanger Seq Tech (chain termination method):

- Mixture of target DNA with ssDNA primers
- DNA polymerase kicks in and start its work
- PCR with fluorescent chain terminating ddNTPs
- Size separation through gel electrophoresis
- Laser excitation and detection by seq machines

It has a very low error rate, may take long sequences and has a low-throughput but is expensive, requires a correct primer and is sequence-specific

New Generation Techs:

- Illumina is a solid-phase based (bridge amplification) technique, which exploits sequencing by synthesis to generate multiple copies and detect light emitted after each incorporation of nucleotide
- Roche 454 uses emulsion PCR and detection of light emission occurs after each incorporation of nucleotide
- Ion torrent emulsion PCR, it detects hydrogen ions released during the sequencing process (pH charges)
- ABI/SOLiD emulsion PCR, sequencing by ligation and detection of light emitted by incorporated probes

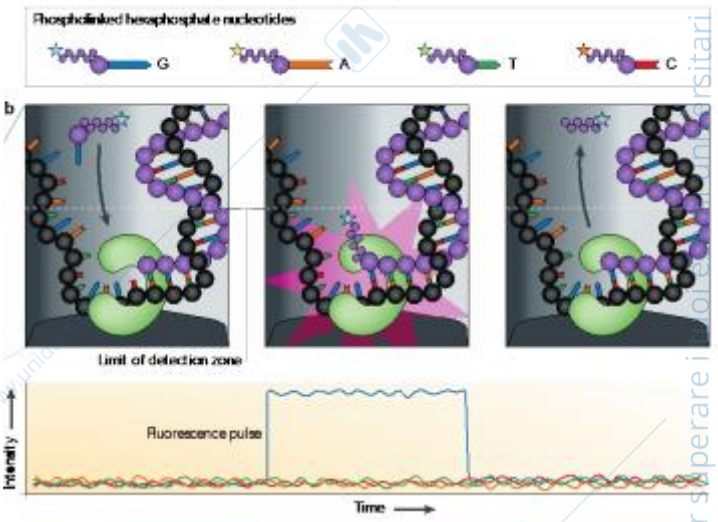
Even though these technologies are very advantageous, they have some issues to be taken into consideration:

- If short reads range is a feature:
    - The original sequence should be reconstructed
    - Due to the fact that genomes contains numerous repeated sequences longer than the NGS reads, sometimes leading to misassembles, gapped/fragmented/incomplete genomes
    - Can detect SNPs and small INDELs, but not large genomic rearrangements
    - Can't distinguish alternative splicing transcript isoforms
- It is PCR based:
    - Amplification errors
    - Underrepresentation of repeated regions
    - Biased in high GC% regions
    - No direct RNA sequencing (requires cDNA synthesis)

- Detect signal as an average between multiple copies of the same DNA fragments (signal increased)
    - Don’t have single DNA molecule resolution

In 2011, we have a new long-read sequencing technique, PacBio, which exploits Single molecule real-time sequencing (SMRT).

SMRT relies on a sequencing-by-synthesis approach: a DNA polymerase is bound within a zeptoliter-scale (10 ^-21 L) observation chamber, called zero-mode waveguide (ZMW), and uses a strand from the native sequencing library as a template for the read, incorporating fluorescently labelled nucleotides as they diffuse into the ZMW. Each incorporated nucleotide is briefly immobilized at the polymerase active site, emitting a fluorescent pulse in the corresponding colour channel



More in general, LR sequencing has few key characteristics:

• PCR-free, single molecule real time sequencing

• Massive parallel reactions, high yield (up to 4Tb/run on ONT)

• Lower cost (as low as 0.12 $ / Mb), but high running costs (ONT is cheaper)

• Long (> 1000 bp / 10 Kb up to 2 Mb on ONT)

• Higher error rate (12 - 13% on both platforms but steadily improving)

• Single-end reads (PacBio), 1D/2D presented as single consensus sequence

(ONT, 2D higher accuracy)

• Simpler sample preparation, but complex data analysis

• Direct RNA sequencing

• Direct identification of DNA modification, i.e. methylation

• Work on a mixture of DNA species

Raw sequences must be processed in order to check the quality, fix errors and remove low quality reads; in order to check data quality we can use FastQC (or others) for short-reads platform, as well as specialized tools for long reads (not well defined yet, but NanoPack, SequelTools are good candidates)

8/10

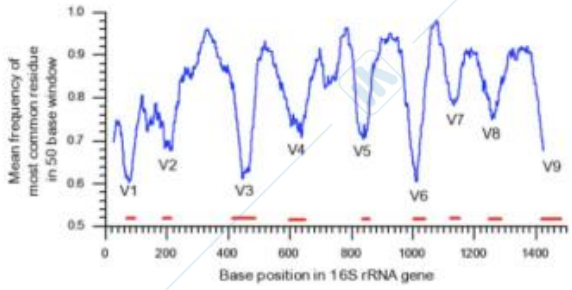
Molecular Taxonomy

Used for classification based on genetic similarities

- Whole Genome
- Single representative genes
    - Sequence gene of interest
    - Compare seq against specific databases
    - Define a cut-off for similarity
    - Assign taxonomic category based on the best match

It is not possible to use every gene of a bacteria for these studies, we need a strong evolutionary conserved gene, that can be used as evolutionary chronometers. **16S rRNA provided a phylogenetic framework that still underpins contemporary microbial ecology and systematic**:

- It belongs to the 30S subunit
- Used for shine-Dalgarno recognition seq
- Many interactions intra and inter molecule -> high evolutionary conservation
- Variable regions (V1-V9) interspersed with conserved regions, hypervariable segments which provide high resolution at short phylogenetic distances. These features are really important
- No exchange- limited examples of rRNA gene-sharing between organisms



Early studies of microbial diversity from environmental samples were technically challenging and required the isolation and DNA sequencing of cloned rRNA gene inserts from recombinant libraries via Sanger sequencing. Each sequence served as a proxy for the occurrence of a microbial genome in an environmental sample. Comparisons of environmental rRNA gene sequences with the rapidly growing rRNA database from cultured bacteria revealed the complexity microbial diversity. Polymerase chain reactions (PCRs) for amplifying and cloning rRNA gene sequences and automated DNA sequencing machines, accelerated the rate of discovery

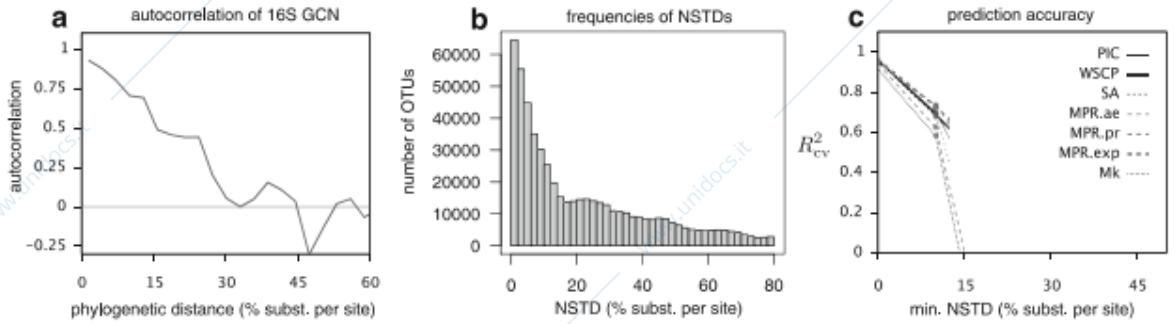
Meta-taxonomic is a high-throughput evolution of 16S sequencing applied to environmental microbial communities (common amplicon sequencing method) and includes 16S rRNA sequencing as well as various techniques that profile selected organisms or single marker genes (it differs by metagenomics on the scale of the target that is sequenced).

It is important to have in mind the right properties a gene should have in order to be considered as universally useful for bacteria classification, and this properties could be the following:

- Universal or almost universal to allow classification of most taxa
- Highly conserved sequence to allow designing universal primers, yet variable enough to resolve closely related organisms
- Low redundance in the genome
- Low horizontal gene transfer rates

For example, if 16S has many useful properties for sequencing, it also has many drawbacks, like the counts of 16S sequence variants, often in PCR amplicons, are used to estimate proportions of bacterial and archaeal taxa in microbial communities. Multiple copies can lead to overestimation of the species (biased towards clades with greater 16S gene copy numbers)

The copy number of 16s rRNA gene has some phylogenetic inertia: the copy number of 16s rRNA gene in a certain genome can be inferred by knowing the multiplicity of the gene in nearby genomes. Different tools try to correct based on this, but the number of copies is relatively volatile in evolution and therefore when there are no genomes very close to the one of interest the prediction can have large error.



16S and other amplicon-based technologies are limited in their phylogenetic ranges; for example, 16S rRNA gene studies primarily target bacteria, with some crossover, whereas 18S or internal transcribed spacer (ITS) studies typically target fungi. Although highly sensitive, these technologies also suffer from contamination, amplification, and extraction biases.

Now, the general procedure for data robustness and reliability requires multiple steps:

1. Quality control (removal of sequencing adaptors, quality trimming, removal of sequencing duplicates, removal of chimeras, i.e. fusion of different sequences due to abortion of PCR). For example, Fastqc program can be used to perform trimming and assessment of the data considered.
2. Taxonomic assignment or classification (OTUs picking, taxonomic assignment).
3. Abundance estimation
4. Diversity measurement and comparison between conditions and samples (alpha and beta diversity)
5. Data visualization and exploration (inc. clustering, heatmaps, meta-data stratification)

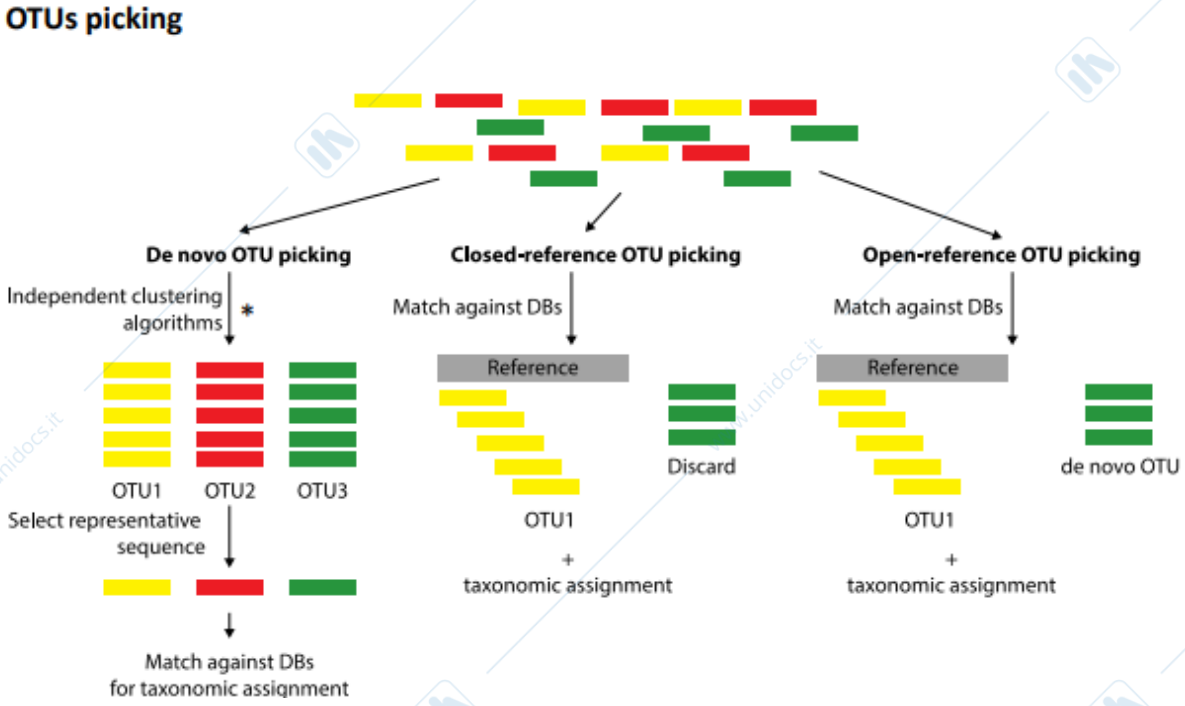
In 16S metataxonomic approaches is preferred the use of **Operational Taxonomic Units** (**OTU**) which are operational definitions used to classify groups of closely related individuals. OTUs are cluster of similar sequence variants of the 16S rDNA marker gene sequence. Each of these cluster is intended to represent a taxonomic unit of a bacteria species or genus depending on the sequence similarity threshold: 97% identity threshold of the 16S gene sequences to distinguish bacteria at the species level; 95% sequence identity for genus.

OTUs picking is an important step and there are three main approaches to do it:

1. De novo picking
    a. Pros: Does not rely on reference databases, especially the samples under study contain fewer known species, such as extreme environments.
    b. Cons: Affected by sequencing errors. Some sequences are not real but produced by the experimental procedure. When clustering using this method, they will be mistaken for an independent OTUs. Filtering and quality control are essential.

2. Closed-reference clustering
    a. Pros: The obtained OTU has high credibility and is faster
    b. Cons: Only the sequence of known species can be obtained, and the information of unknown species is lost.

3. Open-reference clustering, that has the characteristics of the above two clustering methods, that is, the sequence is aligned with the reference sequence, and the unaligned sequence is then subjected to de novo clustering. It has the advantages of the above two methods, but it cannot be used for the combined analysis of different 16S regions



There are obviously more accurate methods overcoming OUT, which are the **Exact amplicon sequence variants (ASVs)**, a variable with the exact purpose of replacing OTUs as atomic unit of analysis. ASVs have demonstrated sensitivity and specificity as good or better than OTU methods and better discriminate ecological patterns, also having an increased resolution and recyclability across studies, reproducibility in future data sets. Plus, they are not limited by incomplete reference databases.

Quantification

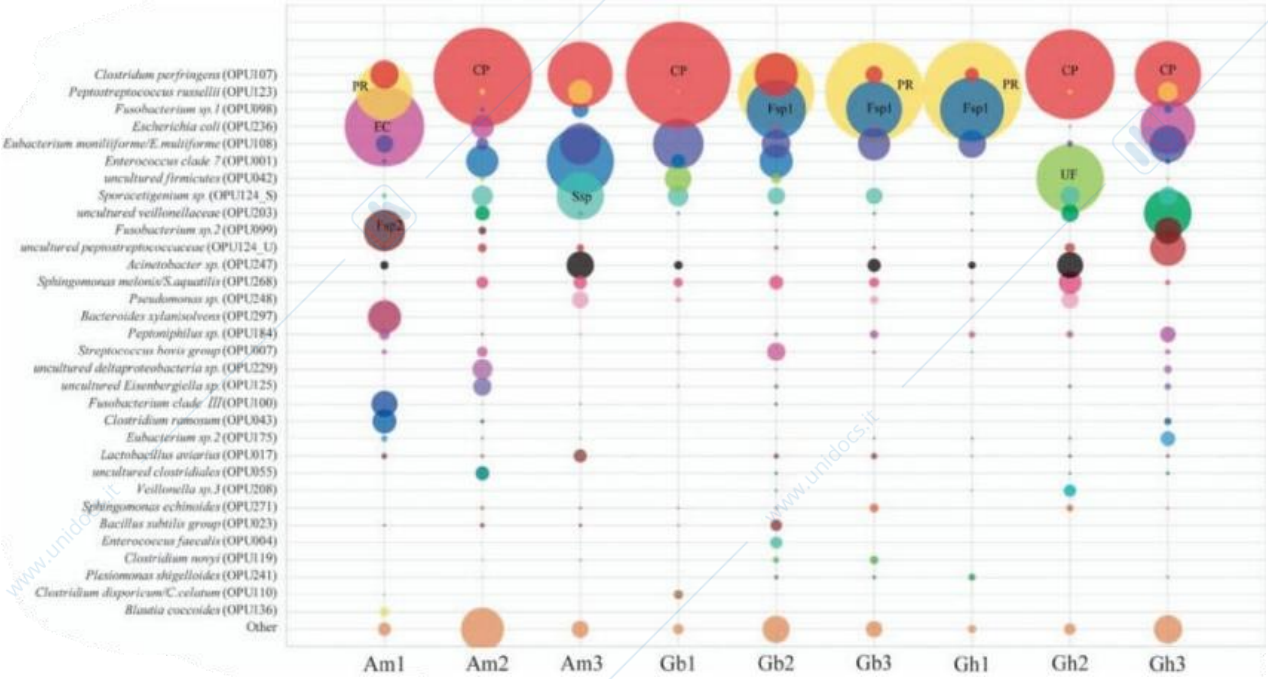
The output of the classification also provides a quantification for each observed taxonomical category. Quantification is proportional to the number of sequences assigned to a given OUT/taxon (that's why all the issues we discussed before can affect our results) after normalization.

Normalization: Accounting for sequencing depth differences and library size. Commonly normalized their OTU matrices by rarefying or drawing without replacement from each sample such that all samples have the same number of total counts

Visualization

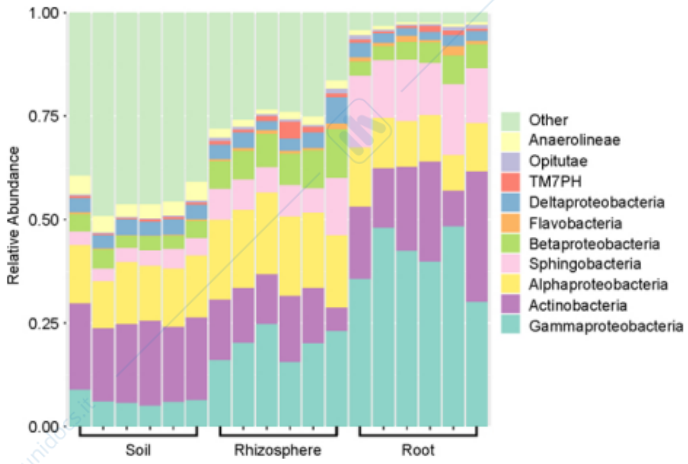
Depending on the kind of data and experiment we are considering, we can use multiple and diverse graphs to represent research findings:

Example 1

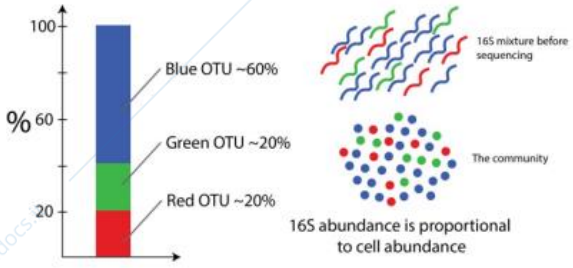


Abundances of the top ten most abundant OPUs. The abundances are shown as percentages of reads in each vulture. A total of 32 OPUs were collected as being the top 10 within the set of all vultures. The others indicated the rest of the OPUs for a given vulture. Near-complete 16S sequence using PacBIO (long reads). Clostridium perfringens causes food poisoning (toxins)

Example 2 → Stacked bar



Relative abundance of bacterial classes in root, rhizosphere, and soil communities (compare across environments)



Stacked bar chart showing relative abundance of classes present in a representative 16S dataset containing 6 replicates for each sample type

13/10/2021

Microbial Diversity data analysis

It is possible to define indices of the total diversity present in certain sample/environment. There are two main diversity typologies:

- Alpha diversity, which regards within samples. Based on total number of species (species richness), the relative abundances of the species (species evenness) or indices that combine these two dimensions. **Shannon's diversity index** most common, combines richness and evenness. A large value is given by the presence of many species with well-balanced abundances. The Shannon index increases as both the richness and the evenness of the community increase

$$H = -\sum_{i=1}^{s} p_i \ln(p_i)$$

where

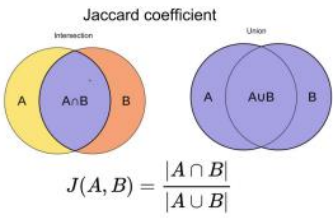
H = the Shannon index value

p_i = the proportion of individuals found in the ith species

ln = the natural logarithm

s = the number of species in the community

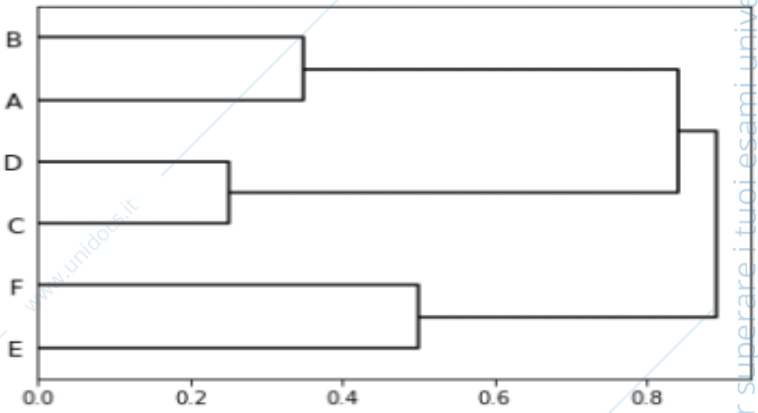
- Beta diversity, which regards between samples. Most biological questions requires the comparison of different microbial communities, which can be made in either a Qualitative or Quantitative manner:
  - Qualitative measures are most informative when communities differ primarily by what can live in them (e.g., at high temperatures). It is divided into:
    - Phylogenetic-lacking, comprehending the Jaccard similarity index → $d_{jaccard}$ = 1-J



- Phylogeny-based, which can be either weighted (Jaccard Bray-Curtis) or unweighted (UniFrac)

- Quantitative measures are ideally suited to revealing community differences that are due to changes in relative taxon abundance (e.g., when a particular set of taxa flourish because a limiting nutrient source becomes abundant

Hierarchical clustering is used to represent graphically how different objects are related. On this example, we can see a typical hierarchical cluster, and as we can notice, samples A and B are diverse in an Alpha configuration, while A and C are diverse in a Beta sense. The displayed distance measures define the degree of Alpha/Beta diversity. Heatmaps are also used to represent frequency matrices.

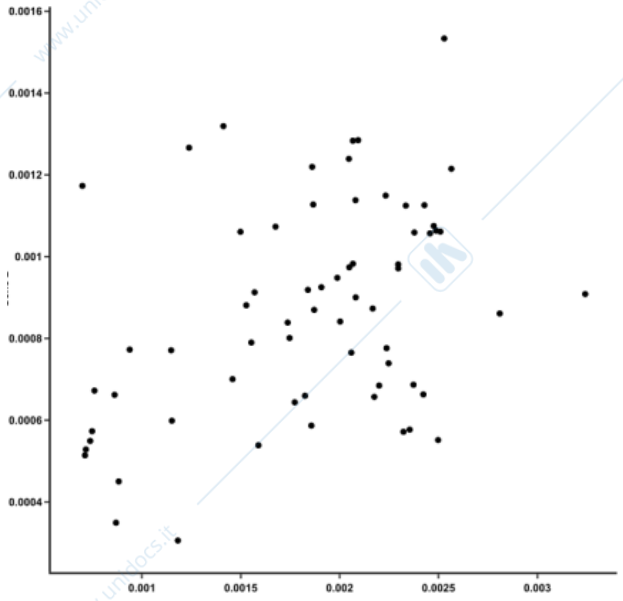


Consider that the specific methods used in clustering, both distance metrics (UniFrac, Jaccard, Bray-Curtis, Euclidean) and clustering (UPGMA, WPGMA) can affect the outcome and the interpretation of clustering analyses.

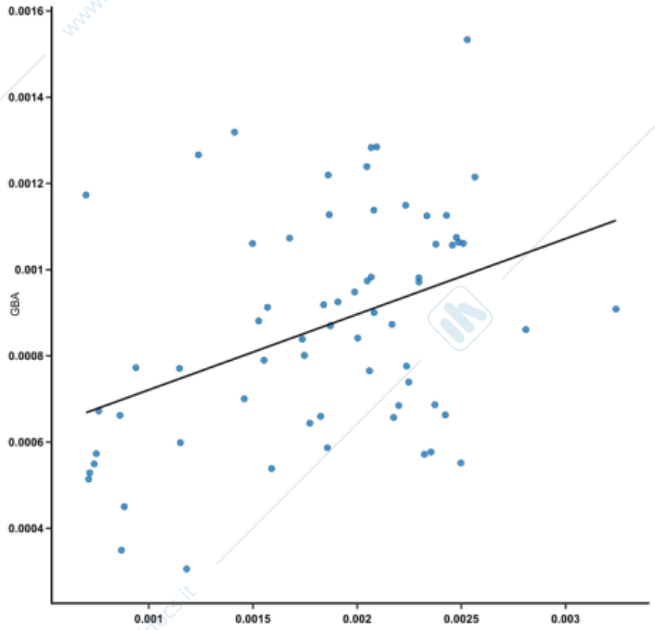
Dimensionality reduction

Since most of the times results over an experiment can yield loads of data to work on, it can be useful to lighten these data somehow. One common approach is to perform some dimensionality reduction technique to be able to extract the most important data and visually represent it in 2 or maximum 3 dimensions to identify patterns. Dimensionality reduction can be made by PCA (Principal Component Analysis), an ordination with sample by observation count data, such as the OUT tables, of by PCoA (Principal Coordinate Analysis), that performs an ordination with a distance matrix as input.

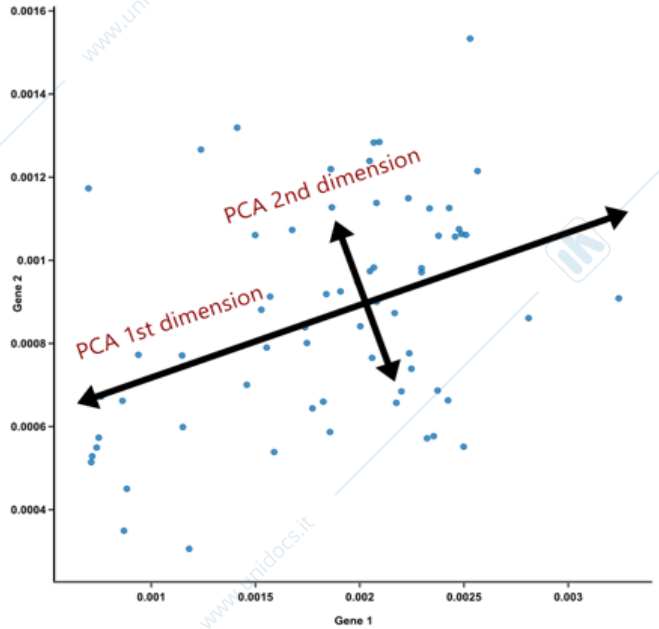
PCA achieve ordination with sample by observation count data by Euclidean distance computation between the samples, which is then followed by a PCoA (Principal Coordinate Analysis) that performs ordination with a distance matrix as input.



← Refined data without reduced dimensionality. Species 1 compared to Species 2 in all 60 samples



← Principal component 1 (PC1) is a line that goes through the centre of that cloud and describes it best. Must convey the maximum variation among data points and contain minimum error. Consider a case in which we have 15 species, i.e. 15-D. We can't graph it.



<--So we use a PCA that anchors a line at the centre of the 15-D cloud and rotate in 15 directions, all the while acting as a "mirror," on which the original 60 dots are projected. This rotation continues until the total distance among projected points is maximum. The rotating line now describes the most variation among 60 samples and is fit to be PC1. PC2 is the second line that meets PC1, perpendicularly, at the centre of the cloud, and describes the second most variation in the data. We continue until we have 15 PCs representing the entire variance of the dataset. Each species has a different contribution to defining each PC.

Lecture 3

Metataxonomics differs from Metagenomics for many reasons:

- Focuses on a single marker gene, while metagen focus on precise functions (Enzymes of biotechnological interest, Virulence genes, antibiotic resistance genes)
- Provides a taxonomic characterization of the community. Instead, metagen provides a thorough description of an environment by characterizing the functional potential of the whole community.
- Provides relative abundance of OTUs
- Indirect and approximate information on metabolic potential (if we have available data about the identified bacterium). On the contrary, Metagenomics is extremely versatile and can be used for numerous applications yet is more expensive since for having high-quality and detailed results we need do generate a large amount of sequencing data.

In general Metagen is more focused on those species' abundances in a certain sample, so that it is possible to characterize the functional potential of that environment.

Metagenomics - Shotgun sequencing related

Sample collection and processing

While the sequencing data may contain errors or limitation that can be fixed in silico (to some extent), wrong processing cannot. Sample collection and preservation protocols can affect both quality and accuracy of metagenomics data. The effect size of these steps, in some circumstances, can be greater than the effect size of the biological variables of interest. It is pivotal to collect sufficient microbial biomass (i.e. within human body). However, enrichment methods can be used for environments in which microbes are scarce, although these procedures can introduce bias into sequencing data.

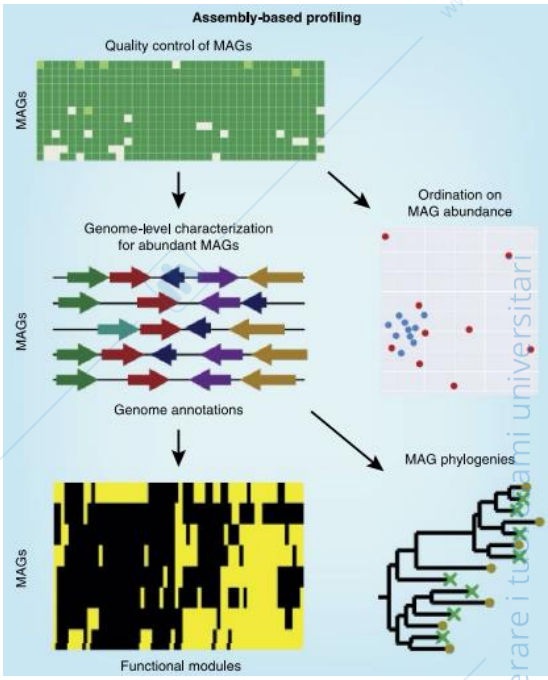
DNA extraction methodology can affect the composition of downstream sequence data too. The extraction method must be effective for diverse microbial taxa. Sequencing results may be dominated by DNA derived only from easy-to-lyse microbes. Moreover, contamination can occur during sample processing stages since kit or laboratory reagents may contain variable amounts of microbial contaminants. Metagenomics data sets from low-biomass samples are particularly vulnerable to this problem, because there is less 'real' signal to compete with low levels of contamination. It could be a good strategy to Include a blank sample.

From raw reads try to reconstruct the genomes (or long fragment of them: contigs).

Similar to single genome assembly, but metagenomes have unique challenges:

- Different coverage (n° sequences / genome) depending on species abundance: low-abundance bacteria have very fragmented genomes. Increase sequence depth can cope with this issue (more expensive)

- Samples can contain different strains of the same bacterial species that differs for a single nucleotide variant or by the presence or absence of an entire gene or operon. The assembler produces a very fragmented genome
- There are specialized tools such as Meta-IDBA, MEGAHIT, MetaSPAdes (different results for different datasets. Try different approaches and see what is working better)



Results include potentially millions of contigs (fragments of genomes) with no specific assignment. We need to run **"binners"** to link the contigs back to the genomes they derived from.

Contig **'binning'** (MetaBAT, CONCOT, GroopM): group contigs into species (Always based on sequence identity. >5% difference, different species). Supervised binning methods use databases of already sequenced genomes to label contigs into taxonomic classes. Unsupervised (clustering) methods look for natural groups in the data and don't require a reference (discovery of new previously unannotated species)

Taxonomic profiling a microbial community (assembly-free)

This technique mitigates assembly problems related to metagen. Plus, is faster than assembly-based approaches (less computation demand but profiling of low-abundance organisms that cannot be assemble). Profiling requires the existence of "complete" reference genomes: previously uncharacterized microbes are difficult to profile. However, the total collection of available genomes are always expanding. Some environments have more genomes available, i.e. human gut microbiome can use assembly-free, while others lacks data availability, i.e. ocean and soil that require the use assembly-based.

There are two assembly-free profiling strategies:

1. Map reads to a database obtained from sequenced genomes (KRAKEN, KRAKEN2, CLARK) which requires a large amount of computational resources (RAM, CPU). It can generate many false positives, but can be corrected using lowest common ancestor (LCA) strategies or coupled with compositional interpolated Markov models
2. Map reads against (multiple) representative or discriminatory markers in genomes (MetaHIT, MetaPhlan2), which is faster and less computationally intensive (best choice for large datasets). It can also be used for strain-level (i.e. sub-species variations) comparative microbial genomics in metagenomes (StrainPhlAn)

Metagenomics- Functional profiling

Metagenomes contains information about the functional potential of each microorganism in each community. **Functional profiling** means assigning a molecular function to genes and or sequences present in your metagenomic data,

Can be used to:

- Predict the gene content of our metagenomics dataset
    - From MAGs (of good quality!): we can ORF finding tools originally developed for analysis of single genomes (with some metagenomic parameter set)
    - From sequences: use specialized tools such as FragGeneScan: predict protein-coding regions from short read sequences
- Assign a molecular function using specialized databases and mapping tools (attention: information is limited to the available molecular functions, so the lack of annotations for accessory genes in most microbial species results in lack of information about metagenomes functional potency).
    - EggNOG, evolutionary genealogy of genes, Non-supervised Orthologous Group. It is a public resource in which thousands of genomes are analysed at once to establish orthology relationships between all their genes. It is focused on providing comprehensive functional annotations for the inferred orthologs; predictions across thousands of genomes covering the three domains of life and viruses; hierarchical resolution of orthology assignments and fine-grained relationships (i.e. in-paralogies) based on phylogenetic analysis.
    - The KEGG database resource, which has been developed for the purpose of uncovering cellular and organism-level functions from large-scale molecular-level datasets, especially gene sets in the complete genomes.
    - UniProt

Functional profiling specialized version

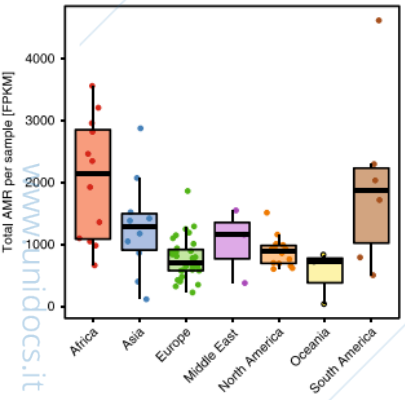
General purpose databases focus on providing good results in annotating the most biological functions. Many times you are interested in very specific biological aspects, which requires specialized resources that are more accurate in identification of genes with a specific biological role.

Example 1 – Carbohydrate-active enzymes are produced by all organisms to accomplish enzymatic modification of carbohydrate-containing compound both intra- and extracellularly. This enzyme group is relevant for understanding central biological processes such as sugar metabolism, protein glycosylation and, on an ecological level, for global biomass synthesis and degradation. Carbohydrate-active enzymes are used in medical and industrial biotechnology.

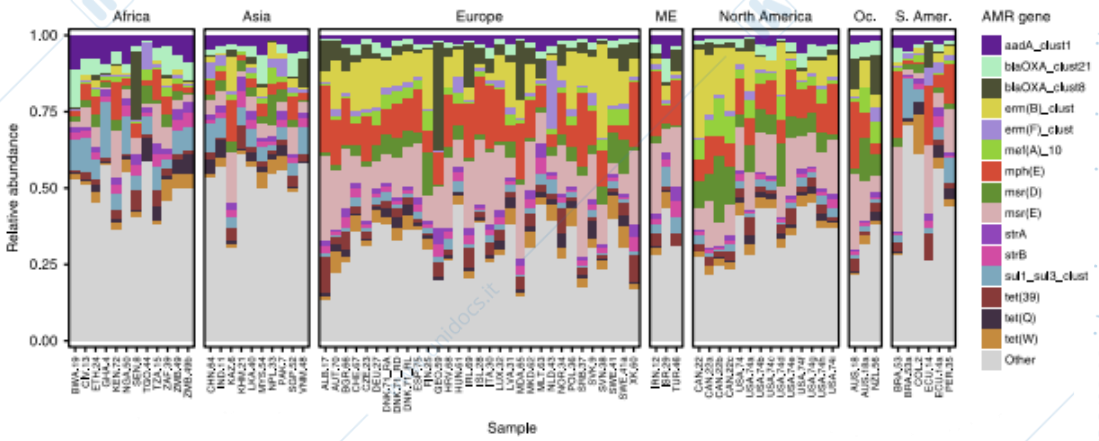
The Carbohydrate-Active enZYmes Database, or CAZy, is a sequence-based family classification of enzymes that assemble, modify and breakdown oligo- and polysaccharides, meant for the unique classification of carbohydrate-active enzymes including carefully curated information about enzyme sequence, structure, and function. It is Regularly updated with newly sequenced genomes from NCBI and Metagenome annotation through collaboration.

Example 2 - Antibiotic resistance genes analysis involves ResFinder, which is a web-based database that uses BLAST for identification of acquired antimicrobial resistance genes in whole genome data,

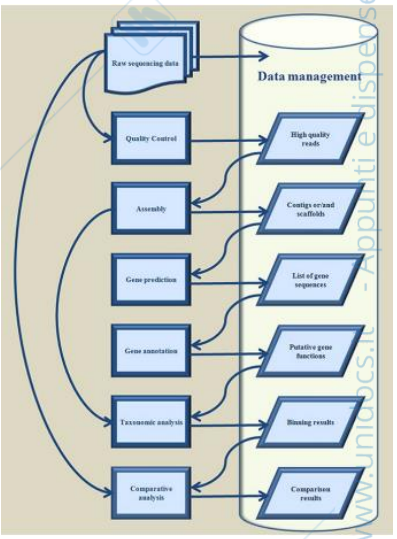
and MEGARes 2.0, a database for classification of antimicrobial drug, biocide, and metal resistance determinants in metagenomic sequence data.



Used metagenomes from global sewage sampling, identified antimicrobial resistance genes using ResFinder (after strict QC), mapped the distribution in the world.



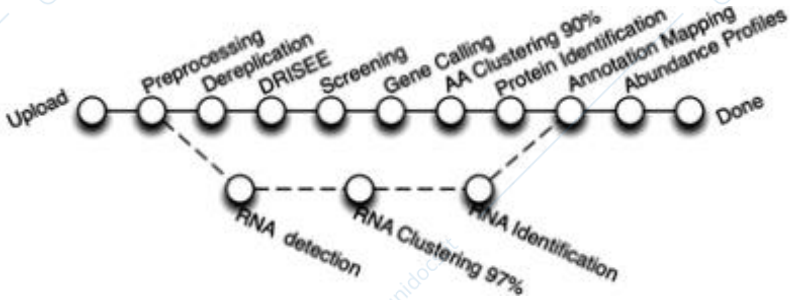
Pipeline



As the standard protocol for analysing metagenomes comprises steps that can be made in a mostly unsupervised way, a lot of pipelines performing the whole job (and including the best practices) have been developed.

An example is given by MG-RAST (Metagenomic Rapid Annotations using Subsystems Technology):

- Data hygiene: includes steps for quality control and artifacts removal: trimming, dereplication, DRISEE (Duplicate Read Inferred Sequencing Error Estimation), screening the reads using Bowtie aligner and removing the reads showing matches close to model organisms' genomes.
- Gene identification: Identifies gene sequences (FragGeneScan) and rRNA genes (search SILVA db)
- Feature annotation: builds clusters of proteins at 90% identity level using the UCLUST, chose the longest sequence of each cluster, search using sBLAT against the M5nr database (non-redundant integration of GenBank, SEED, IMG, UniProt, KEGG and eggNOGs databases). rRNA reads are searched against SILVA, Greengenes and RDP.
- Profile generation: abundance profiles (taxonomy, function)
- Data visualization

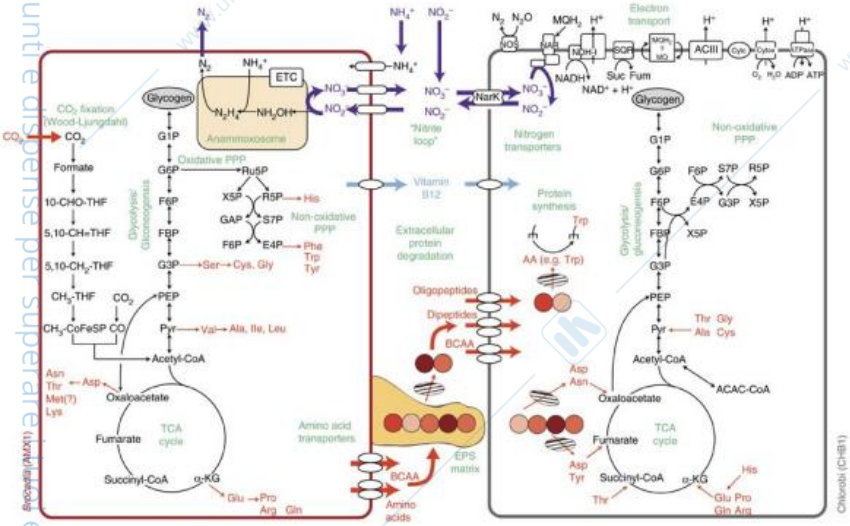


Function of an entire community

From shotgun metagenomics we can reconstruct the functional activity of a certain species in the community or biological system. Through the integration of the functional information for all the species in the community we can therefore be able to somehow assemble all these functions into what could be called a meta-function, comprising all the activities of all the cells in the community.

Keep in mind: species living in a community are not isolated. They can compete, collaborate, depend one from the others...they can release/use different compounds coming from the diet, or the other species. Their metabolic activity is influenced by the surrounding species. By studying the genomes or the genes present in a certain community, we can sometimes infer these kind of interactions.

An example is given by the metabolic network analysis reveals microbial community interactions in anammox granules.



Proposed metabolic interactions between Brocadia (AMX1) and Chlorobi (CHB1) in anammox granules.

Microbial communities mediating anaerobic ammonium (NH_4^+) oxidation (anammox) represent one of the most energy-efficient environmental biotechnologies for nitrogen removal from wastewater. Little is known about the functional role heterotrophic bacteria play in anammox granules

Limitations of Metagenomics

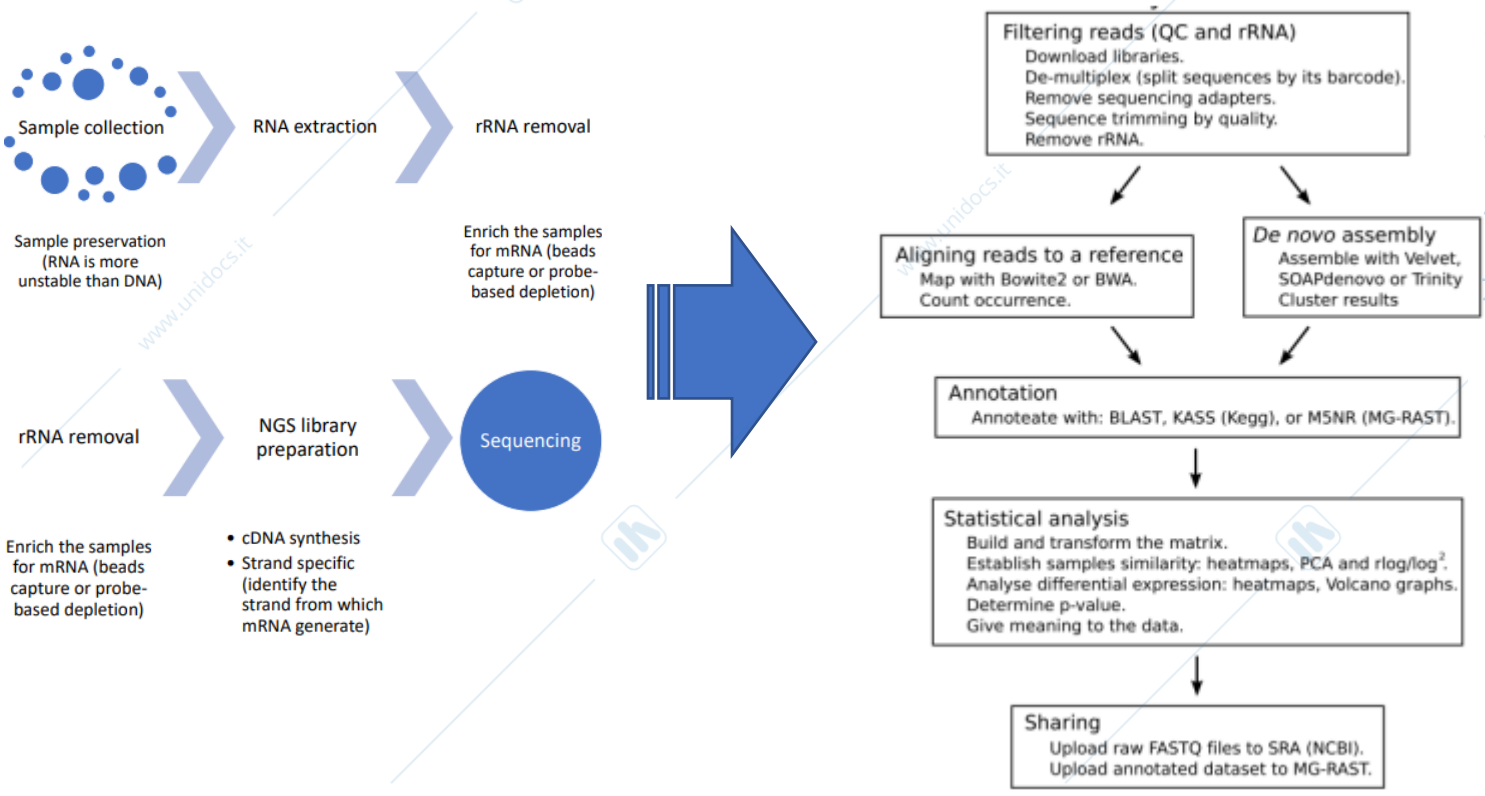
- Entry-level access: expensive to sequence and analyse large numbers of metagenomes without access to sequencing and computational facilities.
- Comprehensiveness of genome catalogues: genomes available are biased toward model organisms, pathogens, and easily cultivable bacteria. All metagenomic computational tools rely to some extent on available genomes.
- Biases in functional profiling: lack of validated annotations for most genes, an issue that can be mitigated only by expensive and low-throughput gene-specific functional studies.
- Microbial dark matter: several members of a microbiome might have not been characterized before with culture-based methods or with metagenomics. Assembly-based approaches can recover part of this 'microbial dark matter'. A fraction of reads may still remain unused after assembly, and the size of this fraction is highly dependent on community structure and complexity. Metagenomics - Limitations
- Live or dead dilemma: DNA persists in the environment after the death of the host cell, so sequencing results may not be representative of the active microbial population.

- Humans-associated microbiome sequencing: sequencing host-associated microbiome with shotgun metagenomics is very challenging due the extremely high fraction of human DNA and the low microbial biomass.
- Disentangling cause from effect: hypotheses from metagenomic studies should be followed up with experimental work to validate correlations and associations.

Gene expression defines the functional role of a species, meaning what a bacterium is doing at a certain moment in time and space. What are the most expressed activities? Since these are related to the environmental condition in which organisms live, they provide a precise picture of the role of each bacterium and of the community.

By analysing the RNA content of a microbial community, we can discover many important information: like the gene activity relationship (the number of different genes expressed in a microbial community across all species), or the gene expression abundance (the highest expressed genes in a specific environmental condition, as well their most active pathway), and a differential gene expression analysis (genes that change the most in different conditions).

Now, compared to metagenomics, Metatranscriptomics analysis require a much higher depth of sequencing (number of reads) than metagenomics studies, in particular if the communities are extremely complex. Below is shown a typical workflow



An example of this