RecyclesEU

recycleseu

Recycles EU

Introduction to metagenomics and metabarcoding approaches: Potentialities and pitfalls

Leandro Gammuto

Department of Biology, University of Pisa

RECYCLES WORKSHOP
Metagenomics and metabarcoding approaches to describe ecological systems and infer their development

5th, 6th & 7th of July 2022

GA: 872053 — H2020 - MSCA - RISE-2019

European Commission

TALK STRUCTURE

1 - Barcoding and Metabarcoding

2 - General flow of a Metabarcoding experiment
    1 Sampling
    2 DNA extraction
    3 Primers selection
    4 PCR

3 - Metagenomic: General principles and considerations

BARCODING AND METABARCODING

The characterization of the microbial community is commonly carried out via PCR amplification of taxonomic marker genes (called “**DNA barcodes**”).

These markers are typically **100 to 600** bp long.

They need to be sufficiently variable to provide deep taxonomic resolution and are simultaneously flanked by **conserved regions** to cover a broad range of taxa.

BARCODING AND METABARCODING

The combination of **High Throughput Sequencing (HTS)** with barcoding has been named “**metabarcoding**”

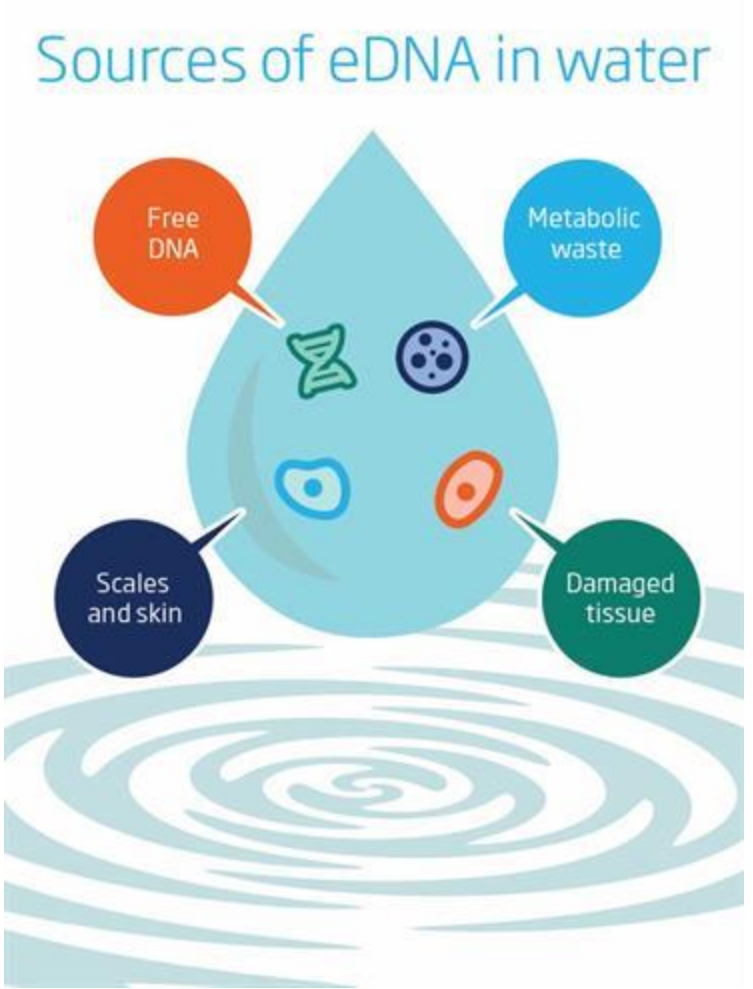
No need for species isolation and cultivation

The relative short length of these markers does not always allow a resolution to species level

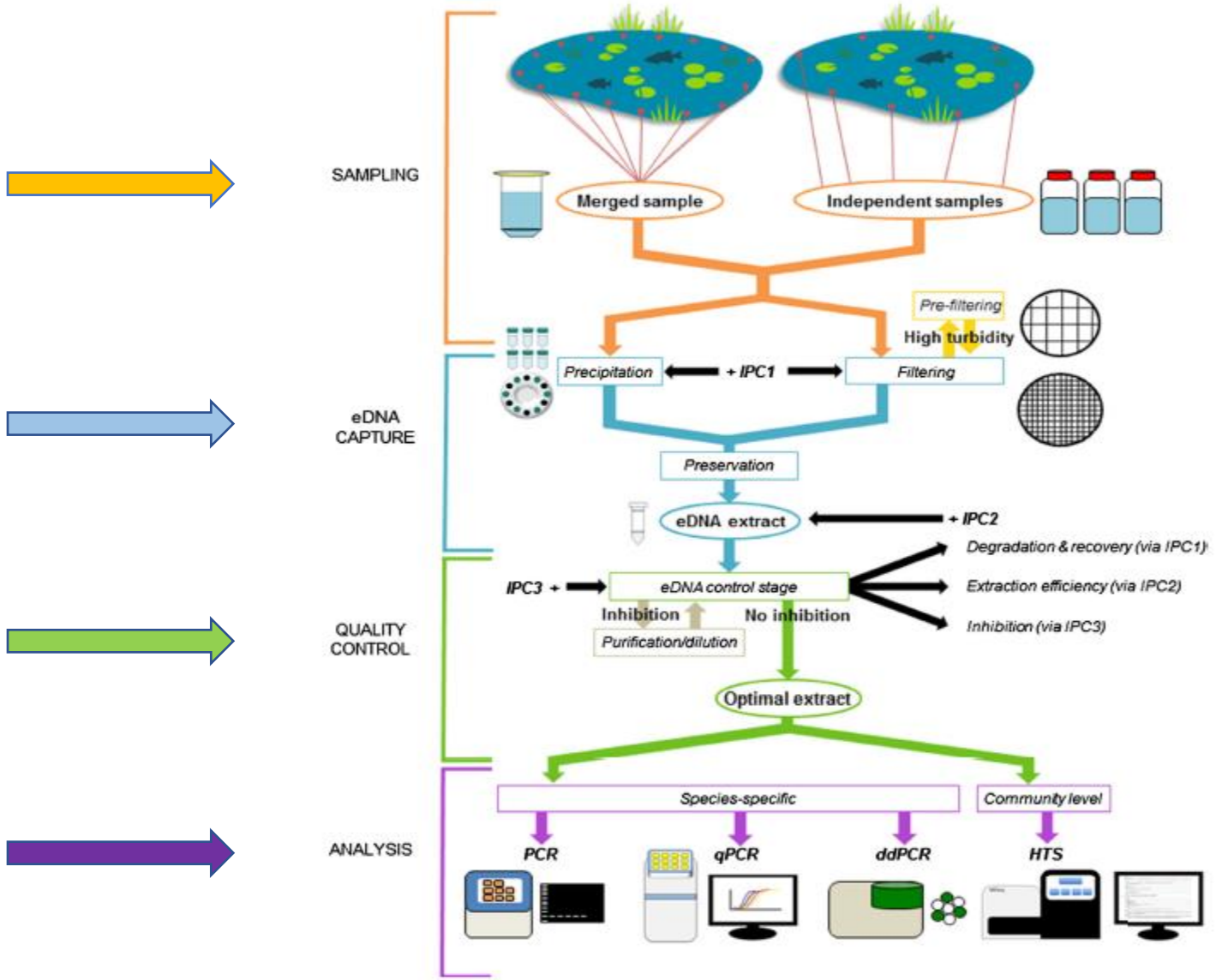
ENVIRONMENTAL DNA - eDNA

Environmental DNA or **eDNA** describes the genetic material present in environmental samples such as sediment, water, and air, including whole cells, extracellular DNA and potentially whole organisms

eDNA can be captured from environmental samples and preserved, extracted, amplified, sequenced, and categorized based on its sequence



GENEARL WORKFLOW OF A METAARCODING EXPERIMENT



A LOT OF STUFF CAN GO WRONG

This approach consists of multiple wet laboratory procedures and requires bioinformatics and computational statistics.

Therefore, sufficient technical knowledge and informed choice **at each step** are essential for successful microbial detection and taxonomic identification.

Significant biases can occur from the cumulative effect of both **systematic** and **random** errors throughout the whole workflow, including sampling, DNA extraction, amplicon library preparation, sequencing and bioinformatics.

A LOT OF STUFF CAN GO WRONG

In addition, the use of DNA metabarcoding for microbial identification has some important limitations:

- Variable number of copies of the selected gene markers
- Biases in the taxonomic annotation of sequences depending on the variable region chosen for the analyses
- Low taxonomic resolution at the species level for some microbial groups

ETEROGENITY OF METHODOLOGIES

Within published eDNA metabarcoding studies, the methodologies often differ substantially, making cross-study comparisons impossible

As eDNA can be applied to numerous ecosystems, collection methods must correspond with these different sample types.

Microbes and pollen can be easily collected from the air, biofilms can be swabbed or scraped, water can be precipitated or filtered, and sediments can be processed

STEP 1 - SAMPLING

For water samples, large volumes of water and multiple field replicates should be used; however, the choice between precipitation and filtration (and subsequently filter size) is reliant on the target taxa, as different taxa can be isolated more efficiently in different ways.

For soil and sediment samples, larger volumes of sample over larger spatial scales are required from larger size classes of organisms, and samples should be extracted from multiple locations to avoid heterogeneity of samples and to better describe the biodiversity of the area.

STEP 1 - SAMPLING

STEP 1 - SAMPLING

Negative field controls are also essential to ensuring valid sampling and to identify contamination.

Periodically filtering clean water as field blanks and processing the filters with the same protocol as the samples.

TIP: Bring the water **on the field** and process it at the same time of regular samples, not separately in lab

STEP 2 – DNA EXTRACTION

The analytical success of molecular techniques is significantly affected by a successful **DNA extraction**.

Main steps in DNA extraction:

- Disruption of cells
- Denaturation of proteins and nucleoprotein complexes
- Inactivation of nucleases
- Removal of PCR inhibitors
- Recovery of the DNA

STEP 2 – DNA EXTRACTION

There are two main kinds of DNA extraction methods:

Kit-based extraction procedures are probably the most common and are generally effective.

- Different kits may be more effective depending on the target taxa
- Usually are less costly
- Harder to customize to meet the experimenter needs

Phenol chloroform extractions are also effective, and may isolate more genetic material than other methods in certain scenarios

- More expensive
- Greater yield of DNA material
- More flexible to customizations

STEP 2.1 – CHECK DNA QUALITY

The DNA template that will be used for the subsequent PCRs should be checked for its **quality** and **quantity**.

The easiest way to assess DNA quality is by a spectrophotometer or with instruments like qubit or nanodrop.

It’s always a good idea to make a test PCR on the extracted DNA to be sure that it is possible to amplify it. The presence of **PCR inhibitors** (e.g. humic acids) may be a serious problem. In case of negative results trying PCR of serial **dilutions** can be a good starting point.

STEP 3 – CHOOSING THE BARCODING GENE

Perhaps one of the most important considerations in eDNA metabarcoding studies is **PCR primer design**.

Different primers and regions **differ** in coverage, resolution, and bias between taxa.

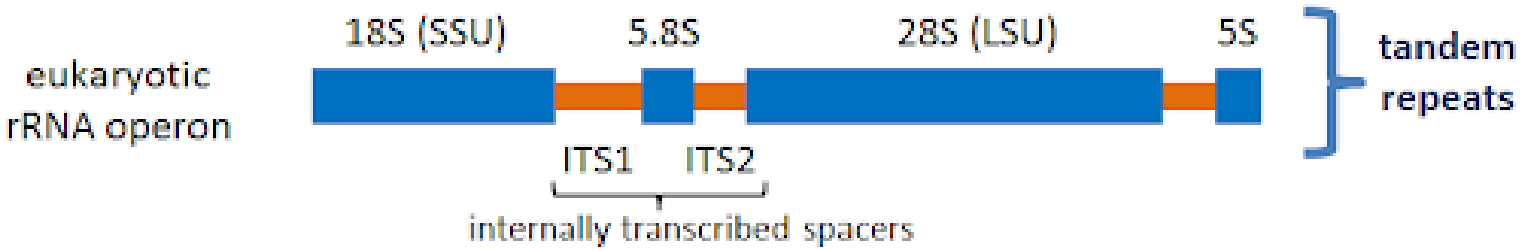
It’s not unusual to use different markers in the same experiment to exploit the features of each of them in order to cover the various downsides.

STEP 3 – CHOOSING THE BARCODING GENE

Cytochrome oxidase-I (**COI**) for metazoans and Ribulose bisphosphate carboxylase large chain (**rcbL**) for plants are the standards established by the barcode of life.

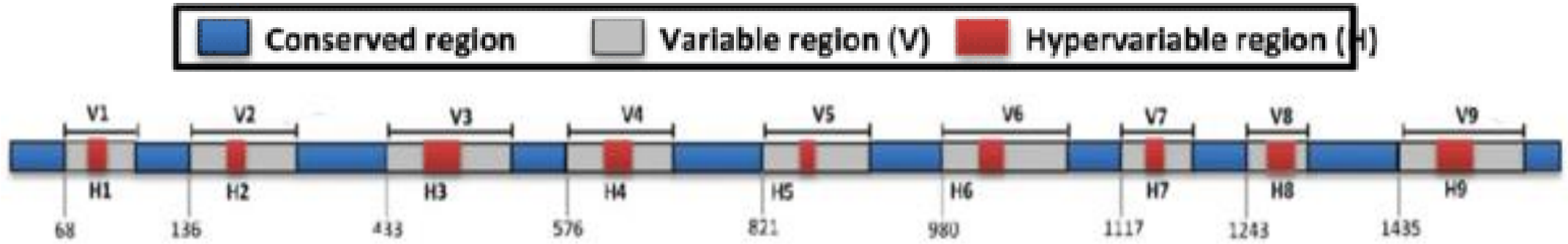
Other regions such as **12S** or **16S** mitochondrial ribosomal RNA may be more appropriate for different taxa.

Internally Transcribed Spacers (**ITS**) is the most used barcoding region for fungi.



STEP 3 – CHOOSING THE BARCODING GENE

In an 16S rRNA gene, there will be region with **different rate of evolution**. Some will be very **conserved** (blue ones), while other will be **highly variables** (grey ones) or even **more variable** (red ones)



STEP 3.1 – CHOOSING THE PRIMERS

Primers for eDNA metabarcoding need to be short enough to amplify degraded samples

Identical within but variable between species, to amplify a variety of species without sacrificing specificity of the target group.

In silico PCR is thus a critical step in the selection of a primer/s in order to:

- Control for appropriate coverage of the target group (i.e., taxonomic coverage and breath).
- The efficient exclusion of outgroups (i.e., taxonomic specificity).
- The ability to discriminate taxa based on nucleotide variability of the amplified marker (i.e. taxonomic resolution).

STEP 3.1 – CHOOSING THE PRIMERS

In addition to these specifications, primer choice has the potential to bias results by **preferentially amplifying** some target sequences more than others

One potential solution to this issue is the use of **multiple primer sets**, particularly evolutionarily independent primer sets coinciding with standardized barcodes for the target taxonomic groups

STEP 3.2 – AMPLICON LENGHT

Another critical aspect linked to the primer choice is the **amplicon length**

Longer sequences will substantially increase annotation accuracy and phylogenetic resolution.

Amplicon libraries created for being sequenced using Illumina paired-end technology will produce amplicon **sizes up to 500 bp.**

STEP 4 – PCR

Once primers have been chosen a PCR reaction is needed to amplify all the potential target in the sample. Those amplicons will constitute the real core of the metabarcoding experiment.

To further reduce primer bias in the amplification process, it is important to determine the **optimal annealing temperature** for the primer pair chosen to avoid the formation of unspecific products.

The use of proofreading DNA polymerases is strongly recommended to reduce **chimera** formation during PCR amplification and to avoid insertion of incorrect nucleotides, which may result in an **overestimation of community richness**.

STEP 4.1 – PCR REPLICAS

Another important factor in PCR and primer design is in **replicates**. Multiple PCR replicates increase species detection and decrease the likelihood of false negatives.

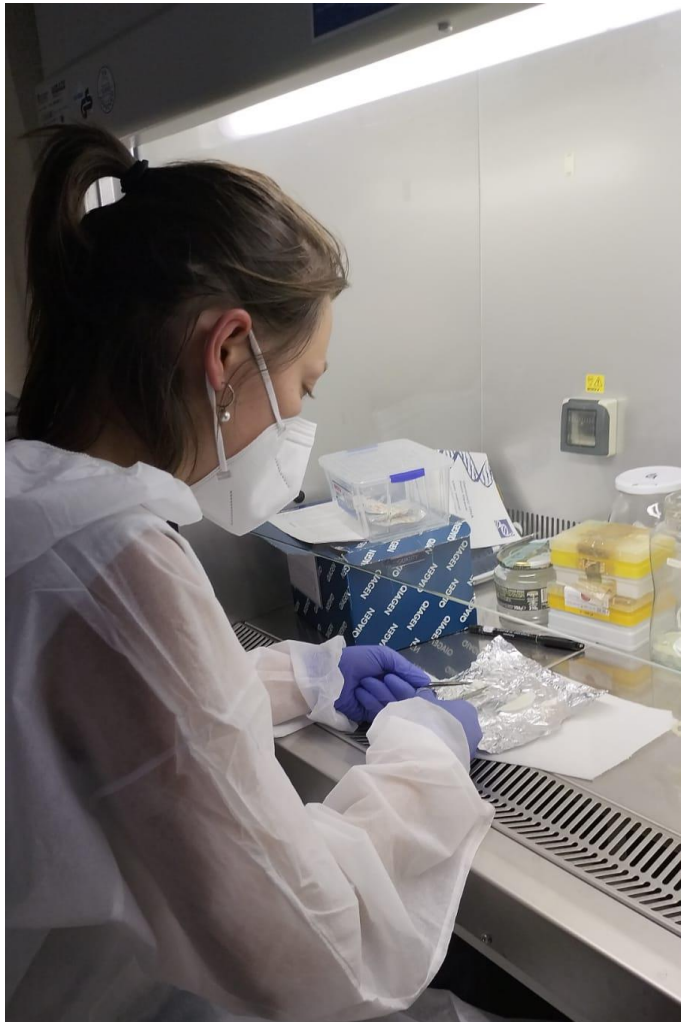
The number of replicates used often differs between studies and depends on several factors like:

- Detection probabilities.
- Research objectives.
- Sequencing depth.
- Primer choice.
- Cost constraints.
- Sequencing platform.

STEP 4.2 – PCR CONTAMINATIONS

In all studies, field equipment, supplies, and personnel **must be separated** from areas where PCR, tissue, and organisms are processed, and decontamination must occur between samples to maintain **independence**

Processing areas in the lab should be cleaned with bleach and UV periodically, and during all procedures, filter pipet tips and clean gloves should be. In addition to the field blanks discussed previously, it is important to have a lab procedure that includes positive controls, negative controls, and replicates at all steps.



THE NEXT STEPS

At this point we are ready for the NGS and subsequent analyses.

These two steps will be covered in detail during the next contributions of this workshop.

DON’T WORRY, YOU’RE NOT GOING TO MISS THEM

BONUS – THE MOK COMMUNITY

The addition of **mock communities** (DNA pools of multiple known species) or positive controls (**single-species DNA**) is also a common practice that can be helpful to:

- Assess the primer bias.
- Asses the error rate of the sequencing.
- Benchmark bioinformatic tools.
- **Determine a relative abundance of sequences in the sample.**
- Correct for compositional bias in case of differential abundance analyses.

Mock or positive controls can also be used to determine a threshold below which an OTU can be considered as an artifact

METAGENOMICS

While partial sequence of microbial DNA (community of microorganisms) is sufficient to assess the information about the diversity of the sampled community, to uncover the genetic potential, we need to analyze the **extended genomic regions**, or even better, **fully restored genomes** from the microbiome (combined genome of the microbiota).

Even assembling a genome is a **difficult task** both due to the complexity of the specific genomes and the many particularities of the sequencing technologies used to accomplish this goal.

METAGENOMICS – GENERAL

In the case of metagenomic data, this task is further complicated by:

- The large volume of data produced.
- The quality of the sequence.
- The unequal representation of members of the microbial community.
- The presence of closely related microorganisms with similar genomes.
- The presence of several strains of the same microorganism.
- An insufficient amount of data for minor community members.

METAGENOMICS - GENERAL

Metagenomic sequencing allows to overcome some of the limitations of the metabarcoding analysis.

“who is there?” → by performing taxonomic characterization of the community

“what do they do?” → by providing functional annotation using long assembled sequence from metagenomic assemblies

“how do they differ?” → by comparing them via comparative analysis.

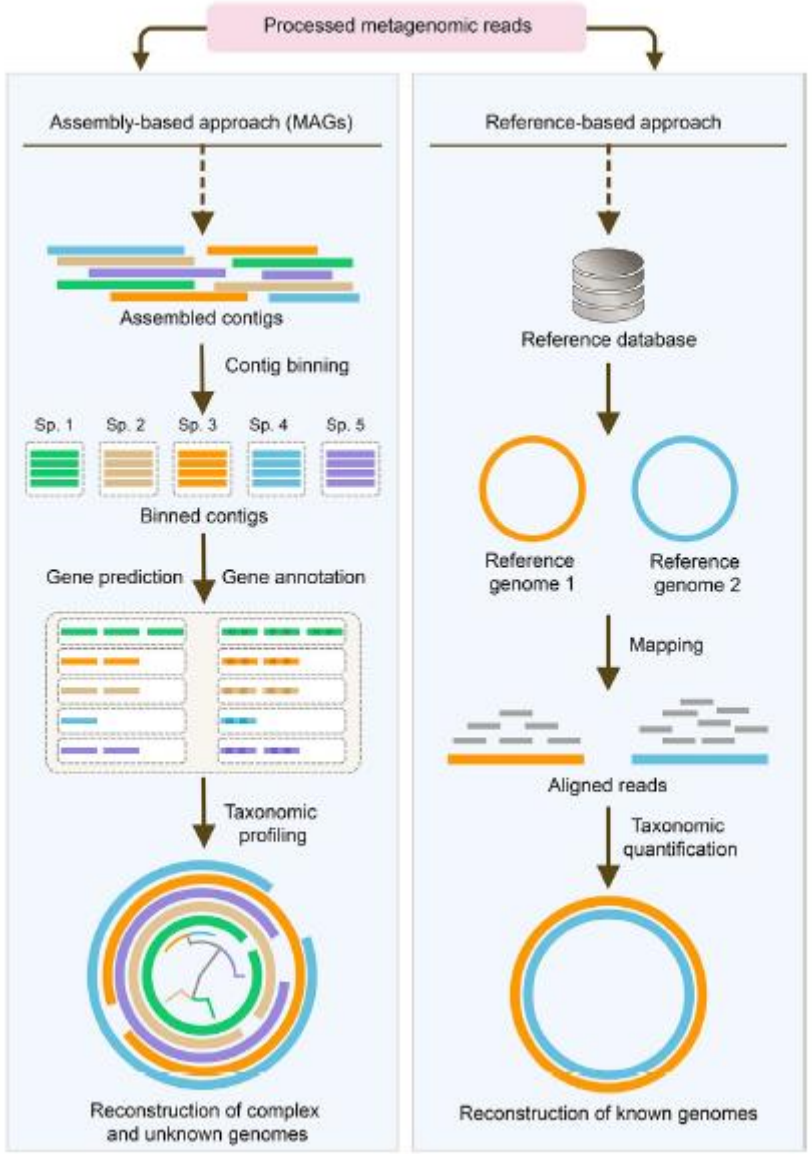
METAGENOMICS - ASSEMBLING

For a thorough analysis of the community, metagenomic reads can be assembled with the help of available reference genomes (**reference-based assembly**) or de novo (***de-novo* assembly**).

Complete bacterial genome, composed by 1 circular chromosome

Metagenomic Assembled Genomes (MAGs)

METAGENOMICS - ASSEMBLING



METAGENOMICS – SOME CONSIDERATIONS

In order to obtain good quality assemblies a good coverage in needed

→ Higher sequencing costs

Higher coverage means higher amount of data (reads) to be processed

→ Higher computational resources needed

Higher sequencing costs and resources needed

→ Less samples can be processed

METAGENOMICS – SOME CONSIDERATIONS

DNA material is directly sequenced

→ No bias linked to primers and PCR

Also incomplete genomes may bring useful information.

→ Possibility to depict functional properties of the whole sample, as well of single organisms.

THAT'S ALL FOLKS!

THANK YOU FOR YOUR ATTENTION