BioInfo4EEG

ECOSYSTEM ECOLOGY

Bioinformatics Course

Session 1

Theory: Understanding the **metabarcoding** technology

Practice: Initial **processing** of **.fastq** reads

- DNA sequencing and analysis is **just another tool** we use to improve our understanding of ecosystem processes.
- Just as there are multiple ways of measuring enzymatic activity, elements, etc, there are **multiple ways to obtain and analyse DNA seq data**.
- DNA sequencing and analysis is still evolving and changing, with **no clear consensus** on many of the steps of the technique.
- I will show you how I have analysed DNA seq data, but it is just **one version** of how to do it. There is always room for improvement. In fact, I will explain where I think I've made mistakes, and how to improve them.

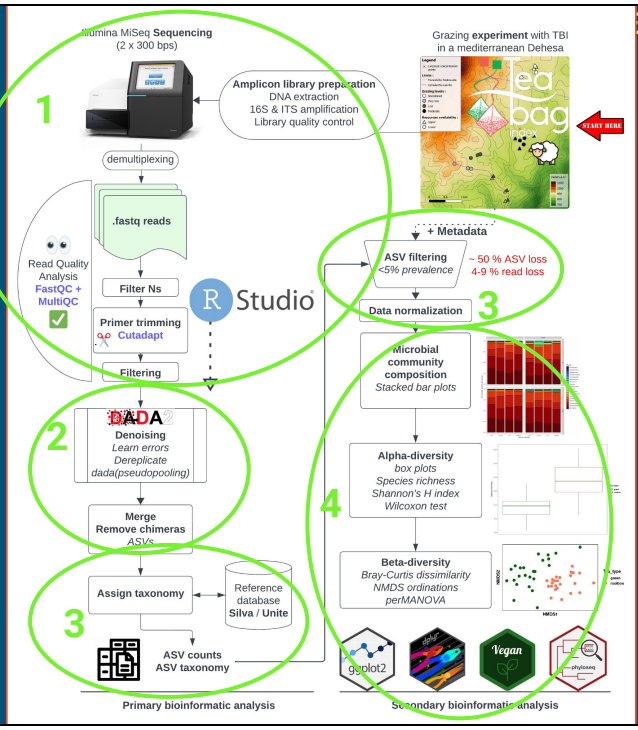
Course **plan**

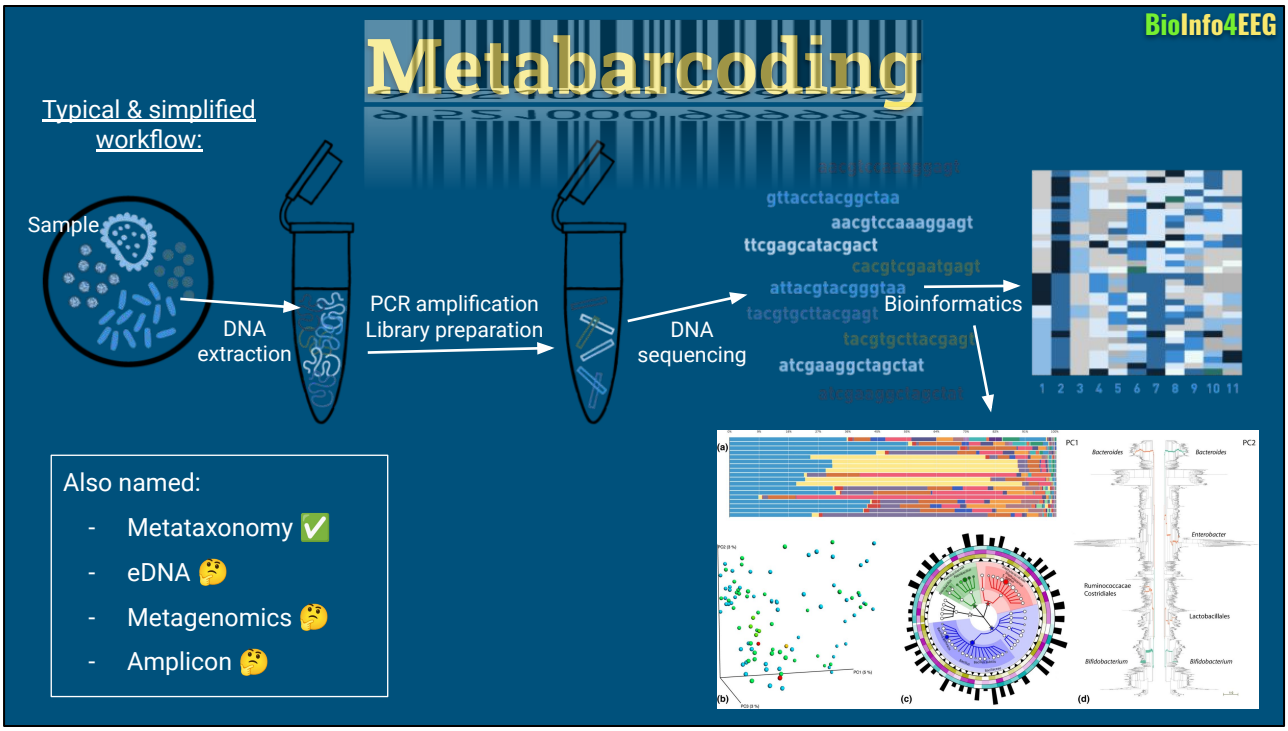
Session 1
Theory: Understanding the **metabarcoding** technology
Practice: Initial processing of **.fastq** reads

Session 2
Theory: Denoising, **ASVs** vs **OTUs**
Practice: **DADA2** denoising algorithm

Session 3
Theory: metabarcoding **databases**
Practice: assign **taxonomy** & ASV **filtering**

Session 4
Theory: **downstream** analysis, microbial ecology
Practice:





Metabarcoding (or, less popular, **metataxonomy**) is the technique of identifying the **composition of species** that make up the **community** of a sample. This is achieved by:

1. extracting **environmental DNA (eDNA)**
2. amplifying certain sequences (**amplicons**)
3. **sequencing** that DNA
4. **bioinformatic processing** of DNA reads
5. **assign taxonomy** to unique DNA sequences

This technique is widely **used**, although not exclusively, to obtain information on the **microbial community** of a certain environment.

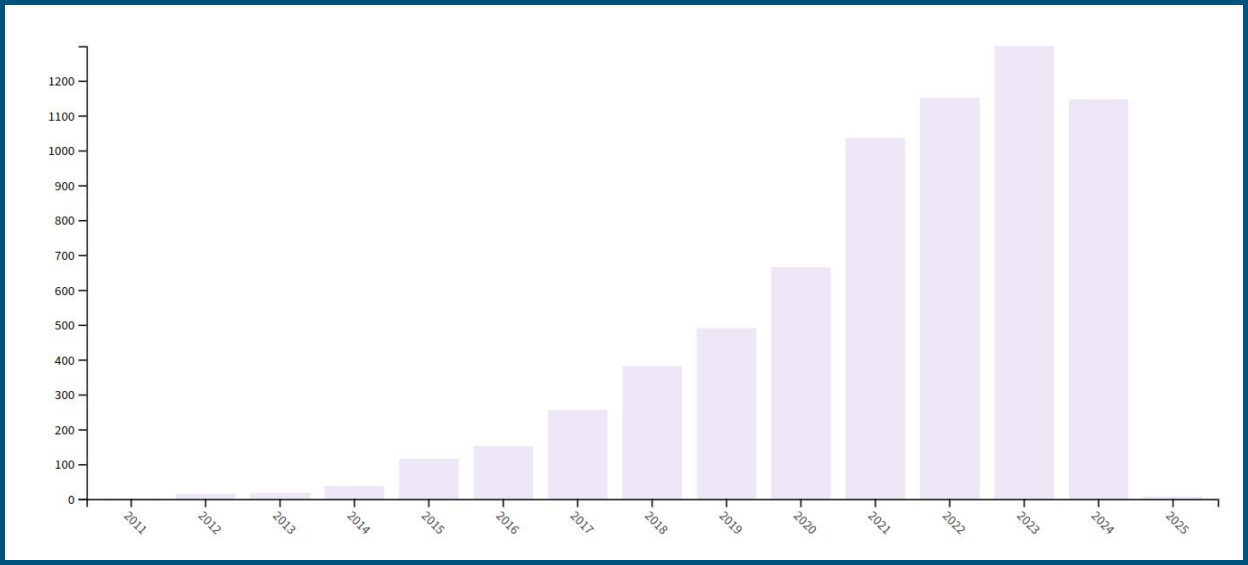
Metabarcoding is also known as **Metataxonomy.**
eDNA is environmental DNA
Metagenomics encompasses metabarcoding, but also includes many other techniques. It is used to describe the sequencing and analysis of multiple genomes

(not just a barcode region)

Amplicon refers to the amplified region of DNA, product of PCR.

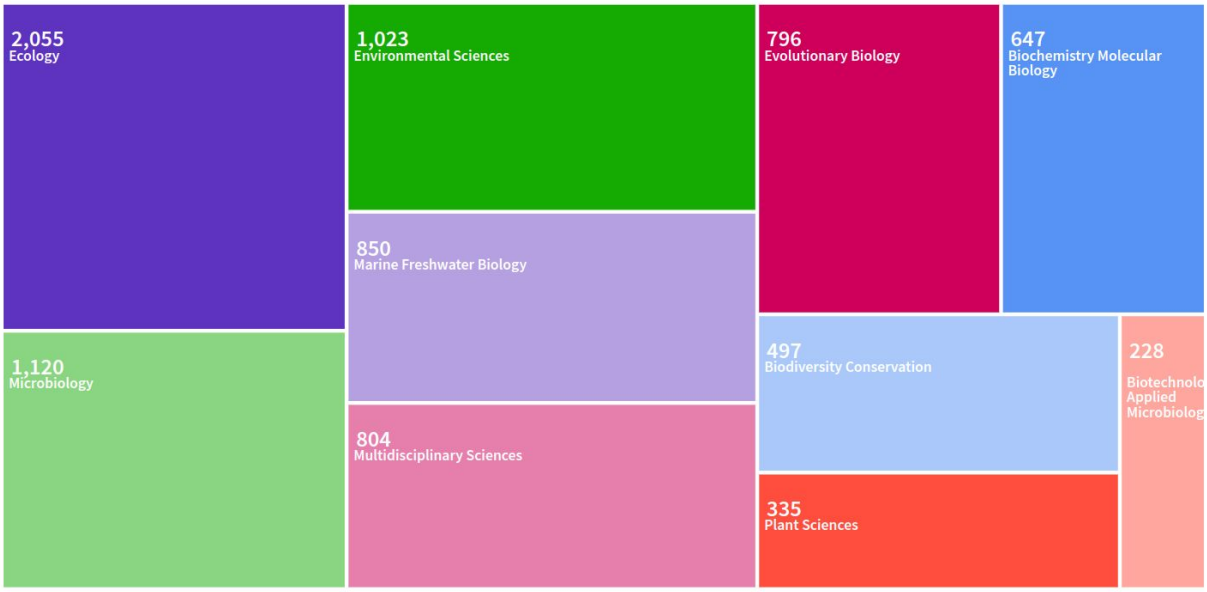
BioInfo4EEG

Searching for “**metabarcoding**” in Web of Science…



BioInfo4EEG

Searching for “**metabarcoding**” in Web of Science…



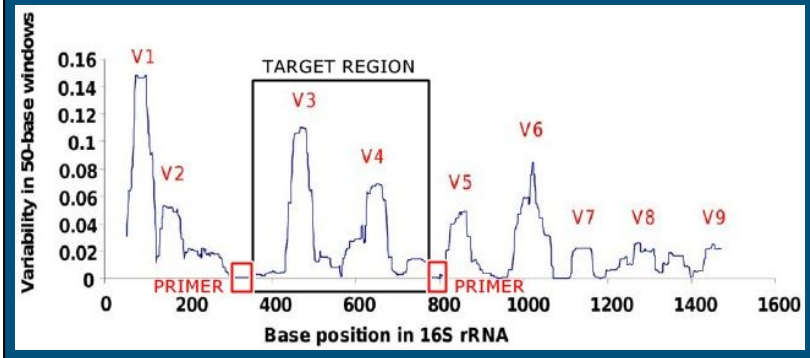
Metabarcoding

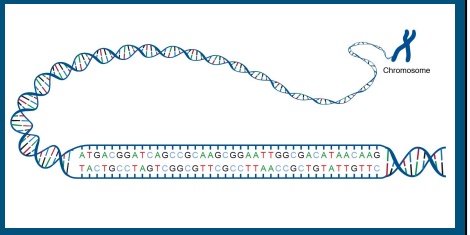
BioInfo4EEG

1. Barcodes

What would the **perfect** molecular **barcode** achieve?

"A good DNA barcode should be **ubiquitous**, have both **low intra-specific** and **high inter-specific variability** (high **taxonomic resolution**) and possess **conserved flanking sites** for developing universal PCR primers for wide taxonomic application. It should also, for the practical user, be reflected in large **reference databases**."





DNA barcode properties:

Ubiquity: within a group of living organisms, the DNA barcode is present in all of them. Example of ubiquitous genes: RuBisCO (present in chloroplasts), COI (mitochondria). The more ubiquitous, the more species we can "see".

Low intra-specific variability: within a species, all individuals have the "same" sequence.

High inter-specific variability: between species, there are remarkable differences in the sequence that make them distinguishable from one another.

Conserved flanking sites: upstream and downstream of the sequence, conserved regions of DNA where primers can bind to and amplify that region. In line with **ubiquity**.

Database presence / existing reference: a barcode is not useful for the practical user if that DNA sequence does not exist and is indexed in a database. Of what use is a DNA sequence that does not tell use where it comes from? However, in this case we know whether it is the egg or the chicken that comes first: databases typically are build and maintained around used, proof-checked and functioning barcodes.

What would a "perfect" barcode achieve?

Amplify a DNA sequence that is **ubiquitous across the whole tree of life** (from

viruses to all prokarya and eukarya), is **unique within each species** (low intra-specific and high inter-specific variability) and all species possess conserved flanking sequences for primer annealing.

2nd generation sequencers

BioInfo4EEG

Platform	Use	Sequencing Technology	Amplification Type	Principle	Read Length (bp)	Limitations
454 pyrosequencing	Short read sequencing	Seq by synthesis	Emulsion PCR	Detection of pyrophosphate released during nucleotide incorporation.	400–1000	May contain deletion and insertion sequencing errors due to inefficient determination of homopolymer length.
Ion Torrent	Short read sequencing	Seq by synthesis	Emulsion PCR	Ion semiconductor sequencing principle detecting H^+ ion generated during nucleotide incorporation.	200–400	When homopolymer sequences are sequenced, it may lead to loss in signal strength.
Illumina	Short read sequencing	Seq by synthesis	Bridge PCR	Solid-phase sequencing on immobilized surface leveraging clonal array formation using proprietary reversible terminator technology for rapid and accurate large-scale sequencing using single labeled dNTPs, which is added to the nucleic acid chain.	36–300	In case of sample overloading, the sequencing may result in overcrowding or overlapping signals, thus spiking the error rate up to 1%.
SOLiD	Short read sequencing	Seq by ligation	Emulsion PCR	An enzymatic method of sequencing using DNA ligase. 8-Mer probes with a hydroxyl group at 3′ end and a fluorescent tag (unique to each base A, T, G, C) at 5′ end are used in ligation reaction.	75	This platform displays substitution errors and may also under-represent GC-rich regions. Their short reads also limit their wider applications.

3rd generation sequencers

PacBio Single-molecule real-time sequencing (SMRT) technology	Long-read sequencing	Seq by synthesis	Without PCR	The SMRT sequencing employs SMRT Cell, housing numerous small wells known as zero-mode waveguides (ZMWs). Individual DNA molecules are immobilized within these wells, emitting light as the polymerase incorporates each nucleotide, allowing real-time measurement of nucleotide incorporation	average 10,000–25,000	The higher cost compared to other sequencing platforms.
Nanopore DNA sequencing	Long-read sequencing	Sequence detection through electrical impedance	Without PCR	The method relies on the linearization of DNA or RNA molecules and their capability to move through a biological pore called “nanopores”, which are eight nanometers wide. Electrophoretic mobility allows the passage of linear nucleic acid strand, which in turn is capable of generating a current signal.	average 10,000–30,000	The error rate can spike up to 15%, especially with low-complexity sequences. Compared to short-read sequencers, it has a lower read accuracy.

Limitations in barcode **length** come from the length of reads that the sequencing machines produce. The most widely used and cheaper technology in metabarcoding is **Illumina**, which produces **~300bp** long reads. Combining forwards and reverse reads, this can enable the sequencing of **~550 bp long** DNA barcodes.

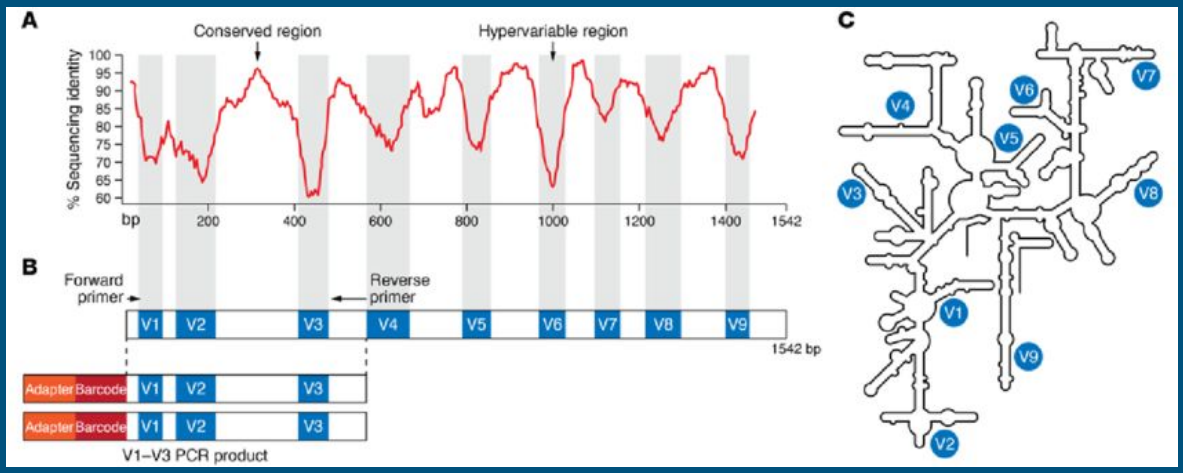
Metabarcoding

1. Barcodes

BioInfo4EEG

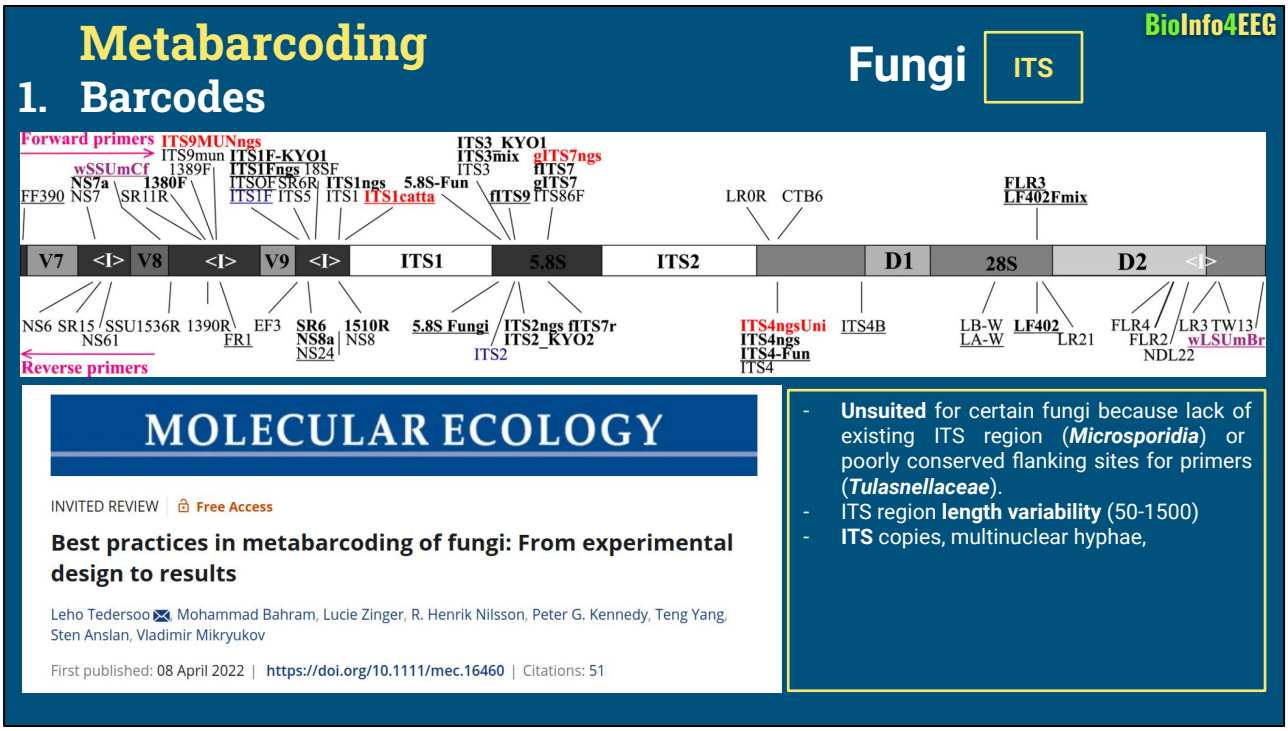
Prokaryota (Bacteria & Archaea)

16S rRNA



DNA sequence coding for the **16S** subunit of the **ribosomal RNA** (rRNA)

The rRNA is indispensable in forming the ribosome, the essential riboprotein for translation of mRNA to protein. Limitations in 2nd generation sequencing technology make it impossible to sequence the entire 16S sequence, so a subset of hypervariable region(s) are selected as barcodes.



ITS stands for **Internal Transcribed Spacer**, a DNA region that separates the subunits genes.

From Tedersoo et. al fungal metabarcoding best practices:

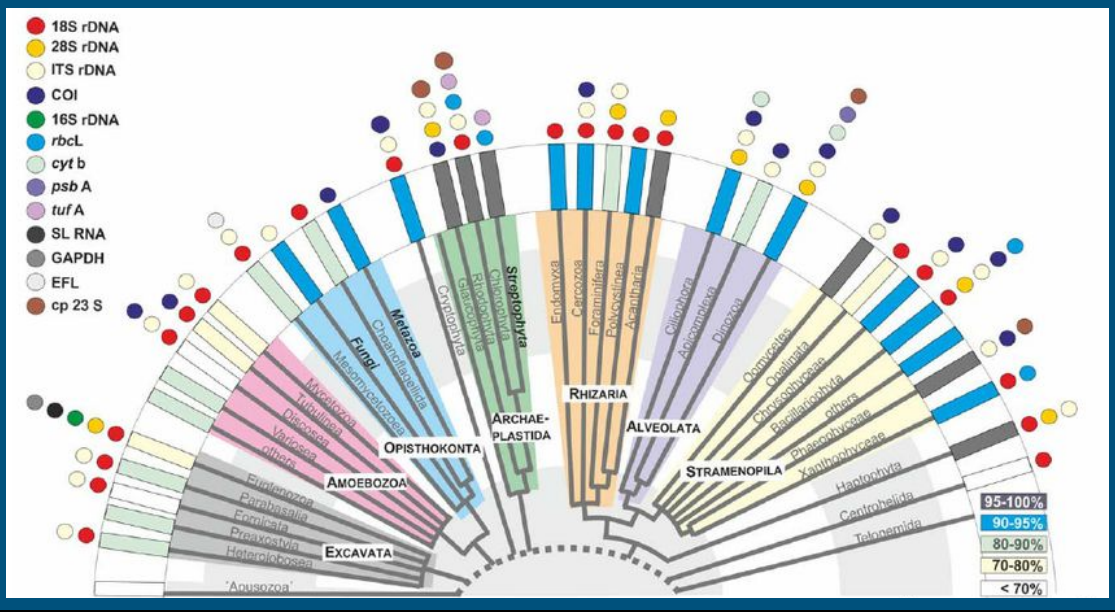
- "The ITS region is **unsuited** to target certain fungi such as Microsporidia (intracellular animal parasites) that may lack this region and certain Tulasnellaceae (orchid root symbionts) that have **mutations in primer sites**"
- The **size** of full-length ITS ranges from 250 (some *Saccharomycetales*) to around 1500 bases (e.g., some *Cantharellales* and various unicellular groups), but *Microsporidea* may have only a few bases of rudimentary ITS sequences. The ITS1 and ITS2 subregions taken separately **vary from 50 to around 1000 bases**. There is also **great length variation** in 18S and 28S rRNA genes, which is mostly ascribed to introns.
- The arbuscular mycorrhizal *Glomeromycota* have multinucleate hyphae with highly variable ITS copies, which has rendered the rRNA 28S and 18S gene fragments of broad use as well (Kolaříková et al., 2021).

BioInfo4EEG

Metabarcoding

1. Barcodes

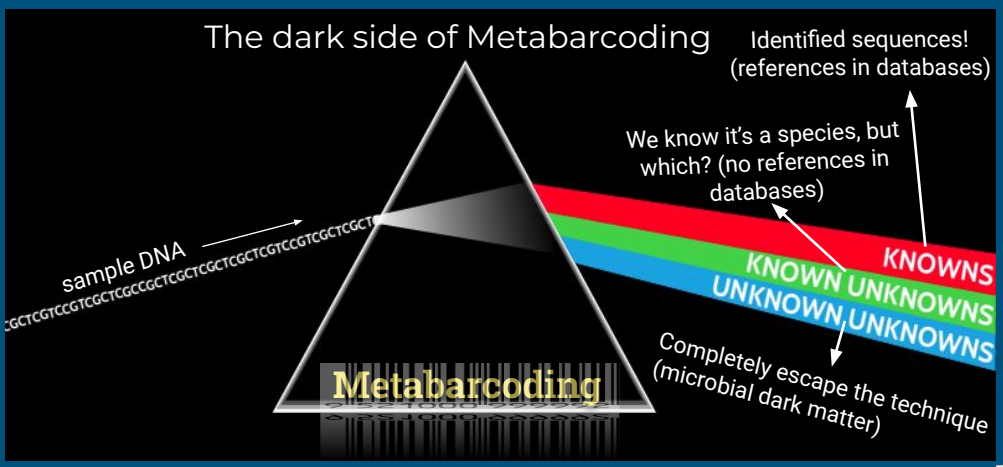
Protists 18S



Metabarcoding
1. Barcodes

BioInfo4EEG

Does a **perfect barcode** exist?



Here comes... The **BIAS** police!

(You don't want this guy to be your reviewer 😱)

Limitations:

Microbial Dark Matter.

There is no barcode that works for everything. Some taxa will always be left out.

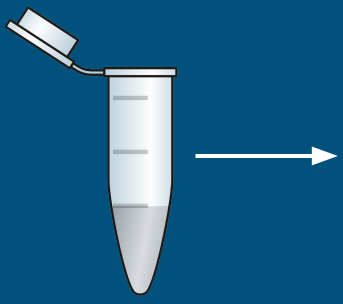
From Microbial dark matter:
Analyses of the 16S rRNA gene from environmental samples revealed that **fewer than half of the known microbial phyla are represented by at least one cultivated representative**. Moreover, among all microbial isolates, **more than 88% belong to only four bacterial phyla** (from among the more than 1,500 estimated phyla): *Proteobacteria*, *Firmicutes*, *Actinobacteria* and *Bacteroidetes*.

INSERT FACT: X of Y phyla are unidentified through metabarcoding technology
A considerable limitation of the technique

Metabarcoding

BioInfo4EEG

2. DNA extraction

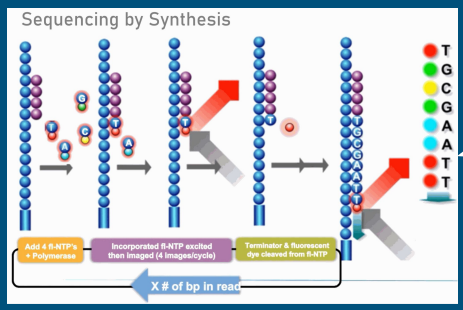


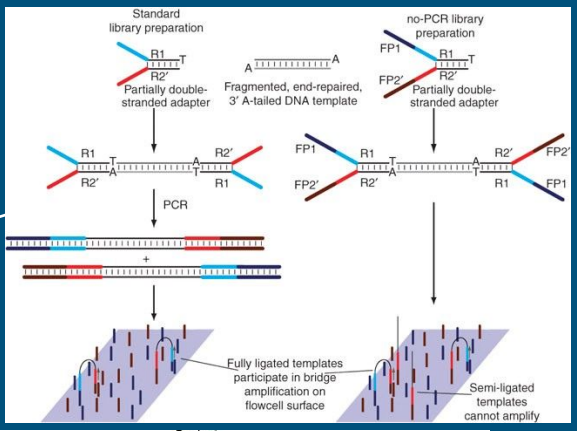
Metabarcoding

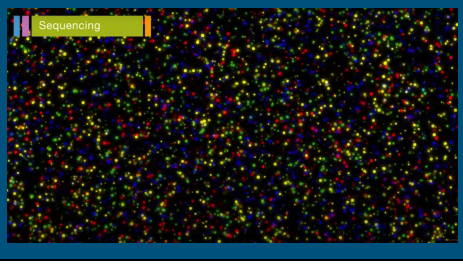
BioInfo4EEG

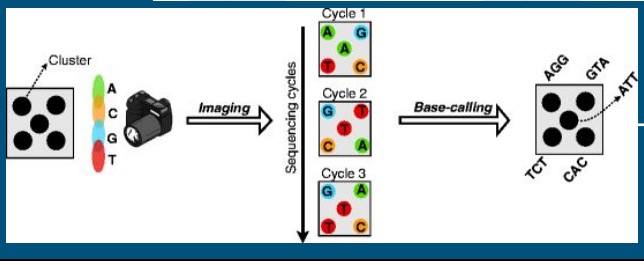
3. DNA sequencing

Illumina Sequencing









.fastq

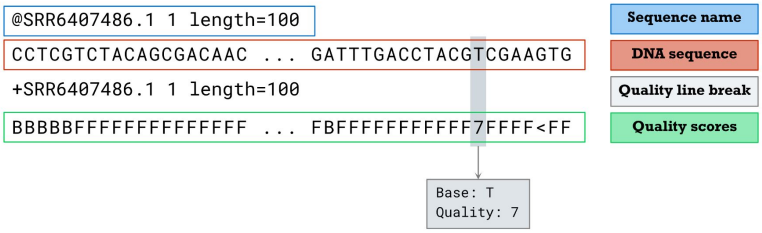
BioInfo4EEG

Metabarcoding

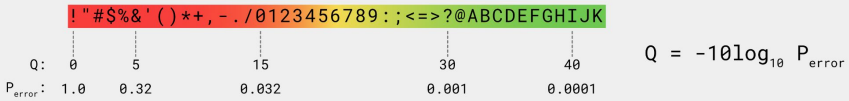
4. The DNA sequence read - .fastq file format

FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFBFFFFFFFFFF7FFFF<FF
```

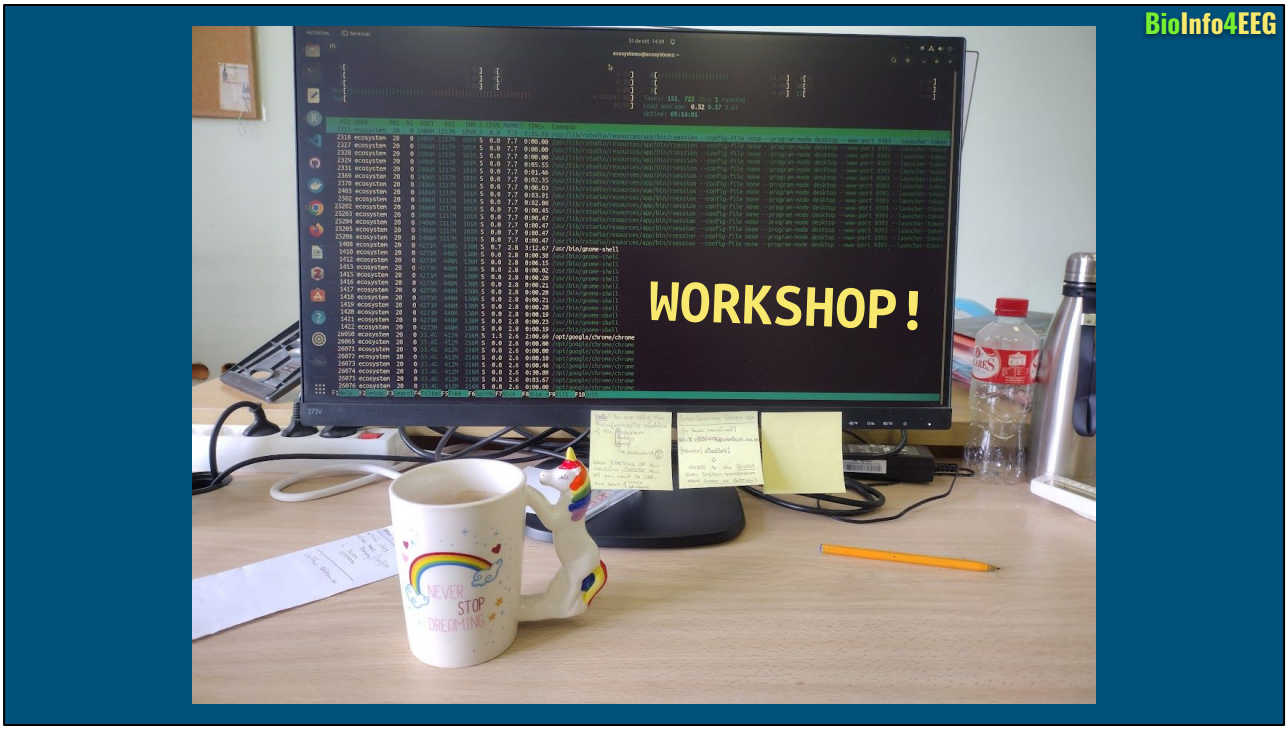


Quality scores as ASCII characters:



$$Q = -10\log_{10} P_{error}$$

Symbol	Nucleotide Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
N	A or C or G or T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	Not T
H	Not G
D	Not C
B	Not A



This course, especially the **workshop** part, is designed around the bioinformatic processing of reads **on the linux machine of the Ecosystem Ecology Group**. So, that when new raw metabarcoding data arrives, programs and tools will be already installed and commands should work **on that machine**.

BioInfo4EEG

List of tools & commands used:

bash

`conda activate metatax`	Activates the conda environment (metatax contains fastqc, multiqc and cutadapt)
`fastqc -o [OUTPUT PATH] *fastq.gz`	Executes fastqc program on any file ended in fastq.gz (in the working directory)
`multiqc --interactive *`	Multiqc takes in .fastqc files and compiles multiple samples QC into one .html (working directory must contain .fastqc files)

Key steps in the Rscript: code/01_Trimming_and_Filtering.R

```r
# PRIMERS [CHANGE ME!]
cat("Forward and Reverse PRIMERS: \n")
FWD <- "CCTACGGGNGGCWGCAG"; print(FWD)
REV <- "GACTACHVGGGTATCTAATCC"; print(REV)
cat("\n")
```

Specify primer sequence

	Forward	Complement	Reverse	RevComp
FWD.ForwardReads	73803	54	54	54
FWD.ReverseReads	56	53	53	58
REV.ForwardReads	58	54	54	194
REV.ReverseReads	76624	53	53	53

Check for primers in the reads

Key steps in the Rscript: code/01_Trimming_and_Filtering.R

```
for(i in seq_along(fnFs)) {
  cmd_output <- system2(cutadapt,
          args = c(R1.flags, # fwd primer & reverse-complement of rev primer
                   R2.flags, # rev primer & reverse-complement of fwd primer
                   "-n", 4, # -n number of primers to search for and trim in a read,
                            # before going on with next read
                   "-o", fnFs.cut[i],
                   "-p", fnRs.cut[i], # output file's PATH
                   fnFs[i],
                   fnRs[i], #input files
                   "--minimum-length", 10,
                   "--max-n", 0, # [IMPORTANT] Remove indeterminations (Ns)
                                 # Ns accumulate at reads' 5' and 3' ends, right
                                 # where primers are located. This makes primer
                                 # trimming difficult to achieve
                   "--cores", 6), # nº of computing cores
          stdout = TRUE,
          stderr = TRUE)
  # Append command output to summary file
  cat(cmd_output, file = cutadapt_summary, append = TRUE, sep = "\n\n")
}
```

Execute cutadapt

Key steps in the Rscript: code/01_Trimming_and_Filtering.R

BioInfo4EEG

```
                  Forward Complement Reverse RevComp
FWD.ForwardReads        0          0       0       0
FWD.ReverseReads        2          0       0       0
REV.ForwardReads        3          0       0       0
REV.ReverseReads        0          0       0       0
> print("Yay! Data should be clean of primers!"); cat("\n")
[1] "Yay! Data should be clean of primers!"
```

Sanity Check: there should be almost no primers remaining in the "cutadapted" data

```
out <- filterAndTrim(fnFs.cut, filtFs,
                     fnRs.cut, filtRs,
                     maxN = 0, # a MUST for DADA2: NO indeterminations (Ns)
                     maxEE = c(4, 6),
                     truncQ = 2,
                     minLen = 100,
                     trimLeft = 5, # remove 10 first bps (low Quality in QA)
                     truncLen = c(267, 232), # [CHANGE ME according to QA
                                             # & length of amplicon region]
                     rm.phix = TRUE,
                     compress = TRUE,
                     multithread = 6, # bool or int with number of threads
                     matchIDs = TRUE,
                     verbose = TRUE
)
```

filterAndTrim() command

This command executes a filtering process based on read quality. Modify the parameters according to the fastqc + multiqc analysis