Characterization of Microbial Diversity using Metabarcoding

MIO, Mars 2024

Different names … same meaning!

16S amplicon sequencing

16s rDNA Gene Tag Sequencing

Metabarcoding

barcoding

16S metagenomics

eDNA metabarcoding

Definitions

A new name for an old concept

"**DNA Barcoding**" appears recently in literature (*Floyd et al. 2002*)

but was firstly reported in 1993 (*Arnot et al. 1993*)

- *sensu stricto*: Use of a standardized DNA region as a **tag** for rapid and accurate **species** identification

- *sensu lato*: Identification of **any taxonomical level** using any DNA fragment

Mitochondrial Cytochrome Oxydase 1 → Animal

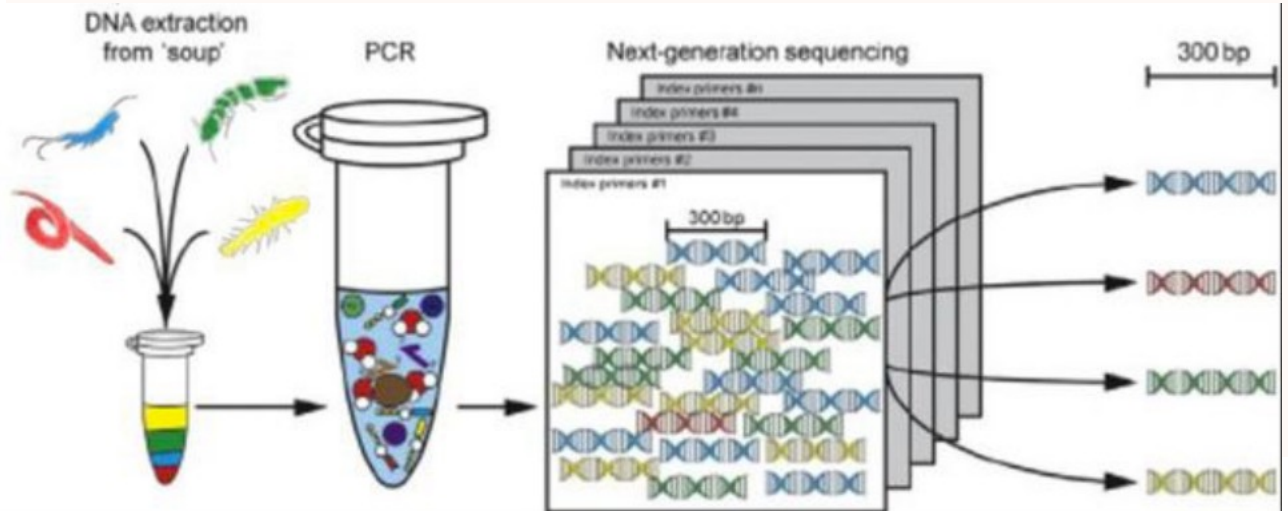
Chloroplast part → Plant

16S rDNA → Bacteria/Archaea

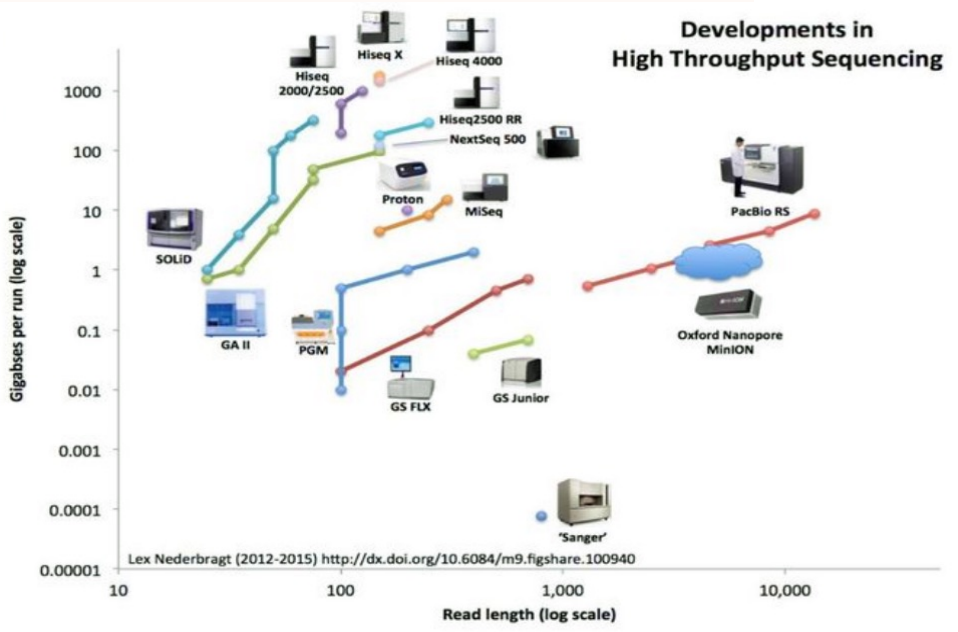
Metabarcoding…

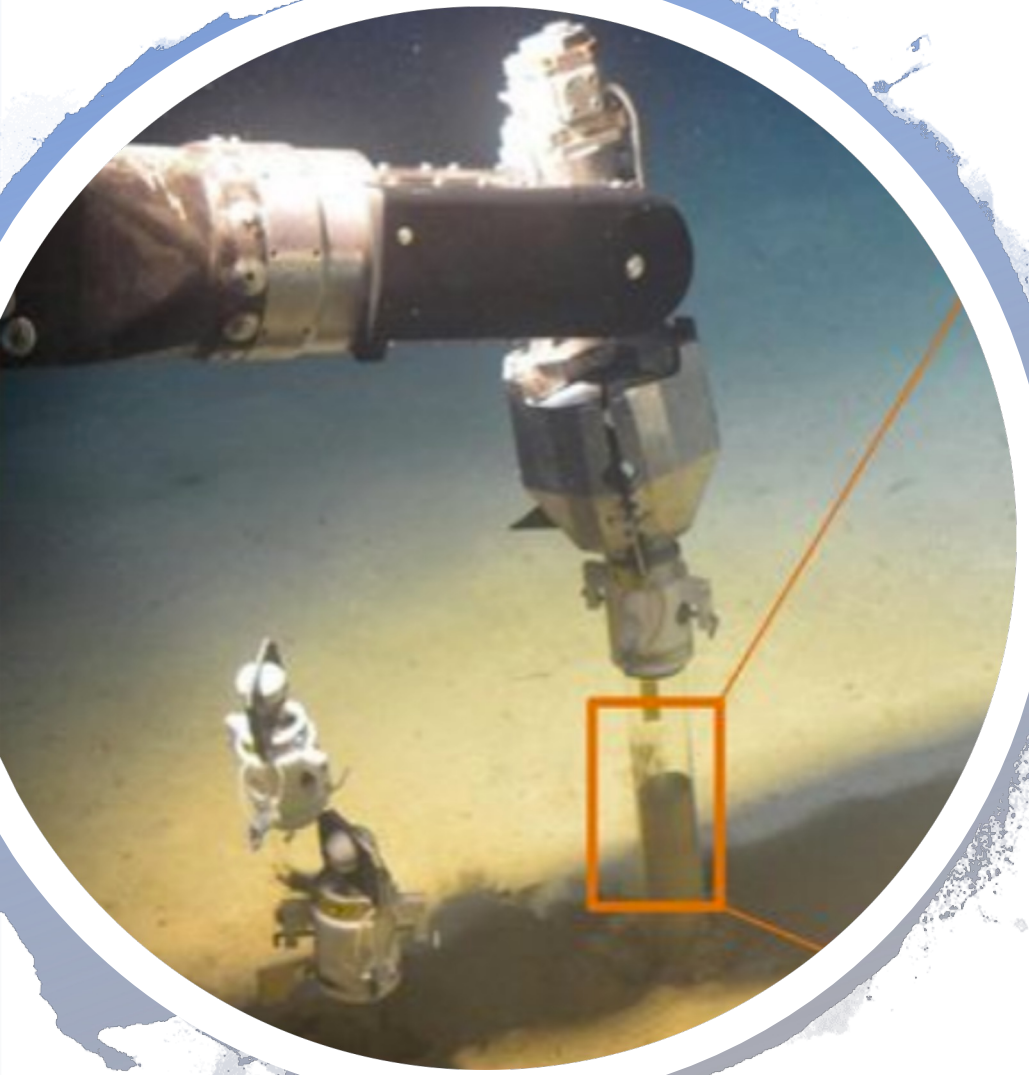
Use to designate **high-throughput multispecies identification** using DNA extracted from an **environmental sample** or from bulk samples of entire organisms (*Taberlet et al. 2009*)

→ Take advantage of Next Generation Sequencing (NGS)



Gill et al. 2016





Environmental DNA (eDNA)

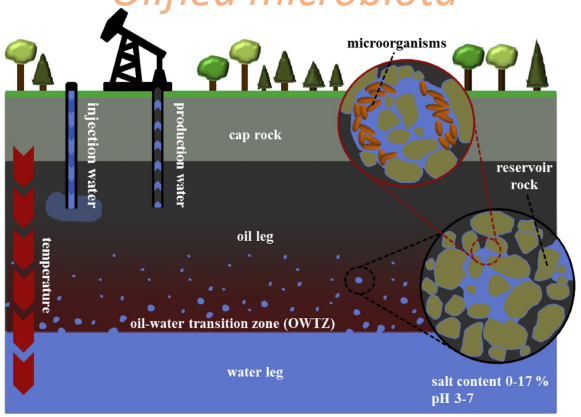
DNA extracted from air, water, soil without isolating any microorganism

Goal of Metabarcoding: Who is there?

Characterization of the **taxonomic diversity** inhabiting various ecosystems using direct eDNA

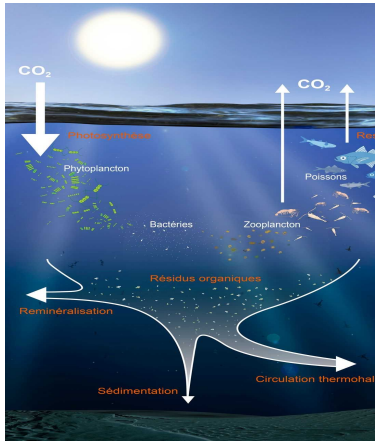
→ Multispecies sample
→ Target 16S rDNA gene marker for Bacteria/Archaea identification
→ Use high-throughput sequencing (= massive data→ complex ecosystem)

Oilfied microbiota



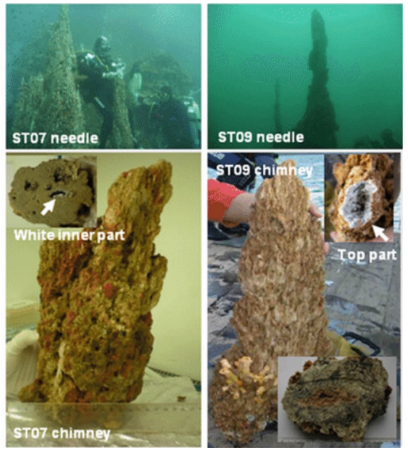
Pannenkens et al. 2019

Biological pump of CO2



Tamburini et al.

Submarine hydrothermal sources



Erauso et al.

More than "Who is there?"....

- **Relationships Taxonomic Composition – Environmental Factors (EF)**

EF: ph, temperature, salinity, trace element, hydrostatic pressure, nutrient availability etc

- **How environmental factors impact community composition/structure**?

→Abundance shifts? Absence/Presence? Gradient? Diversity level?

Pollution Response

- **How similar/different are your samples**

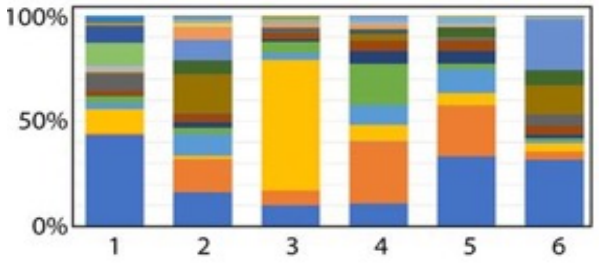
→ What are the major actors driving these similarities?

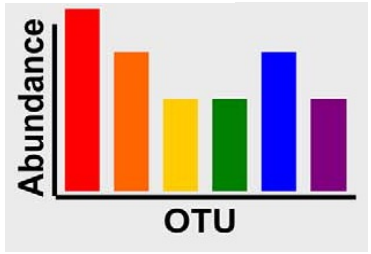
- **Dynamics of communities distribution at time and/or spatial scale**

i.e. Seasons, geographic location, ocean depth, perturbations etc

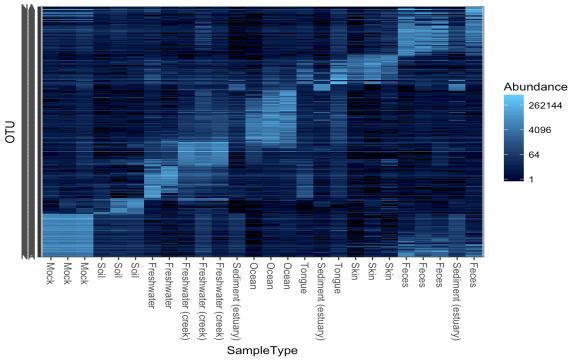
What kind of results you can expect?

Taxonomic profile

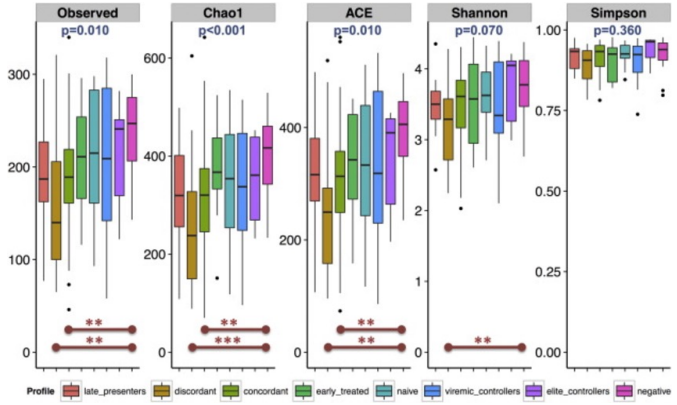




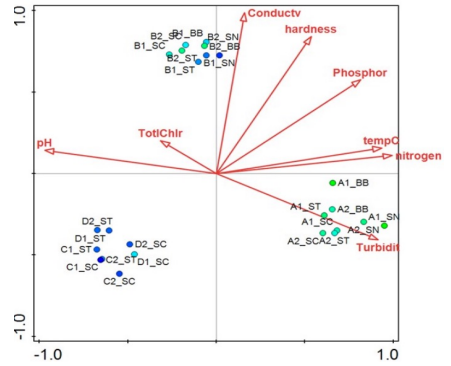
Time or space Gradients



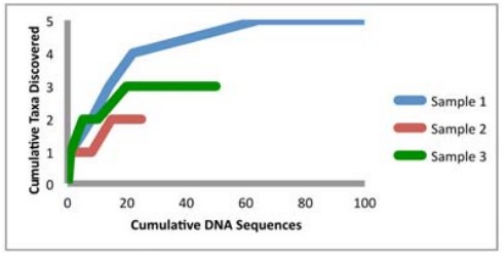
Within-sample alpha diversity



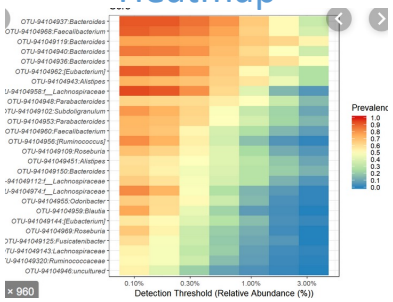
Multivariate analysis



species richness



Heatmap



Who is there? Easy?

- **You can not use the full length of 16S gene**

→ Have to work with a short 16S region

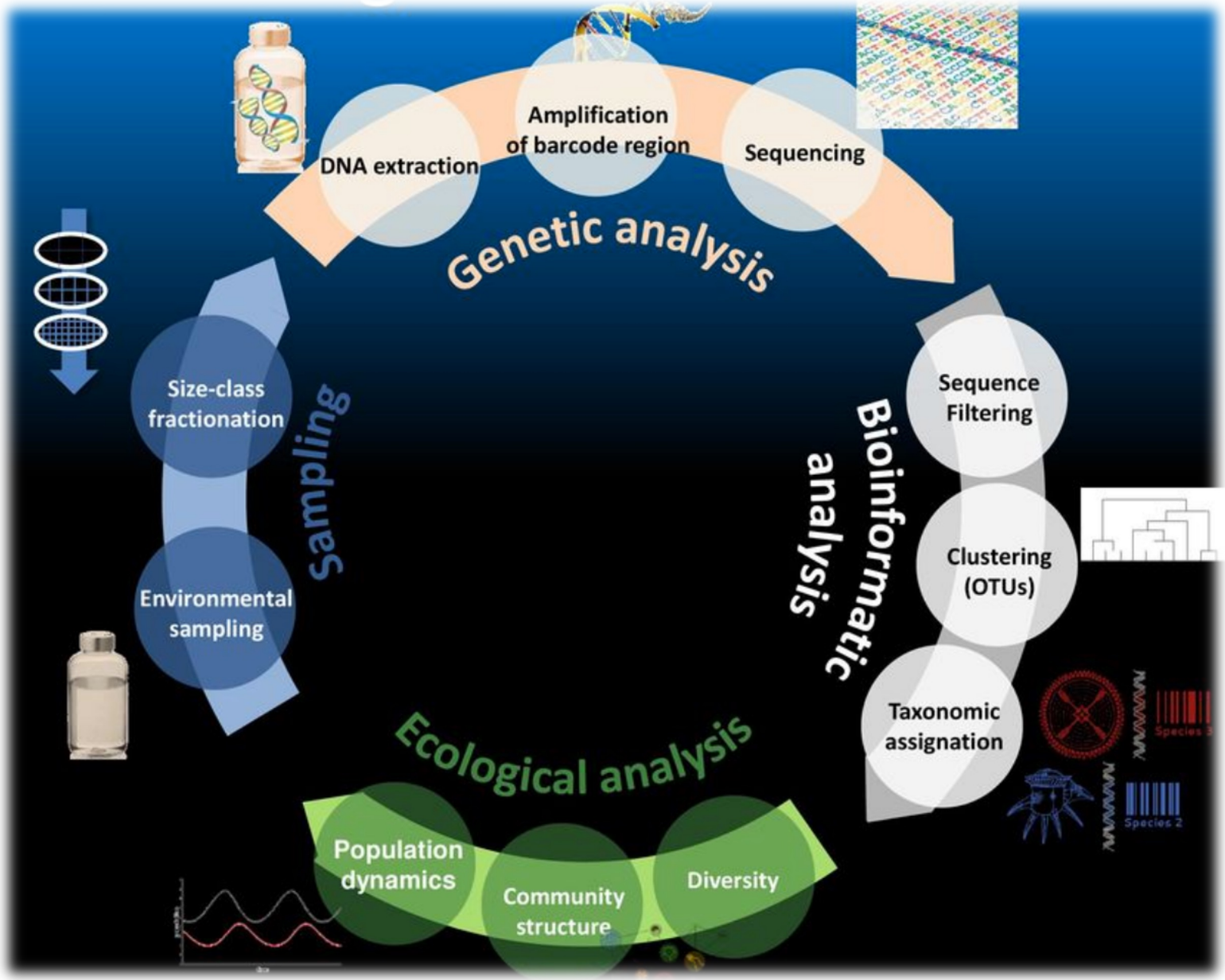
→ Loss of resolutive power for taxonomic identification

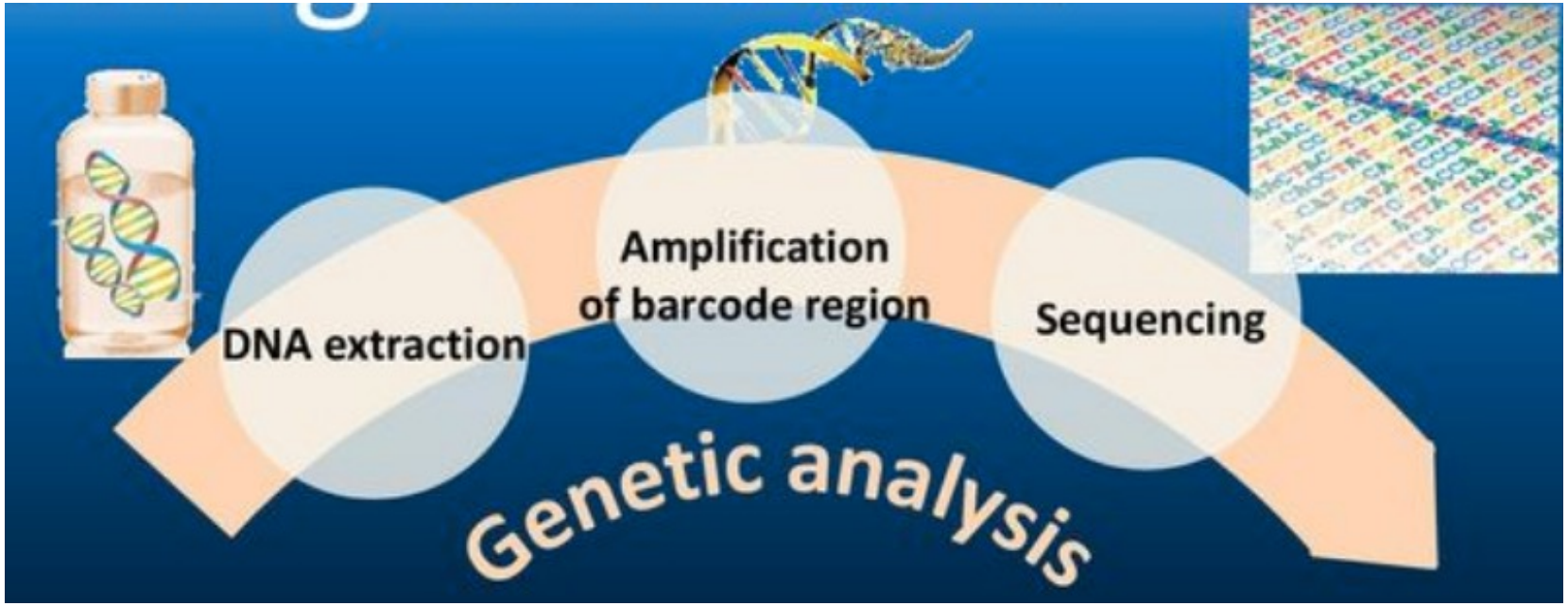
- **PCR & Sequencing introduce errors in sequences/reads**

How to deal with that??

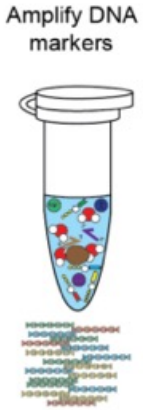
→ Increase taxonomic misidentification

Metabarcoding Global Overview

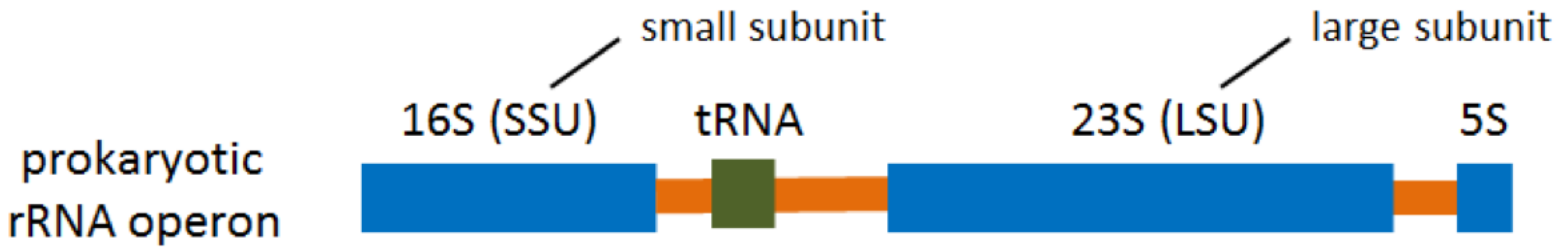




- **Choice of gene Marker**
- **Choice of the Region to amplify**
- **Choice of Sequencing Technology (NGS)**



16S marker for bacterial/Archaeal identification

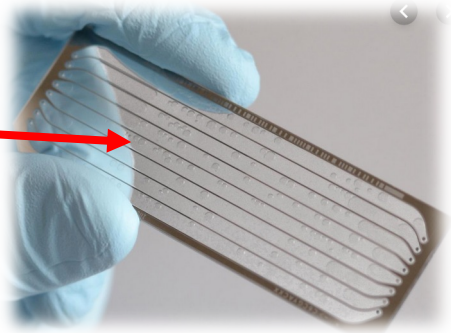


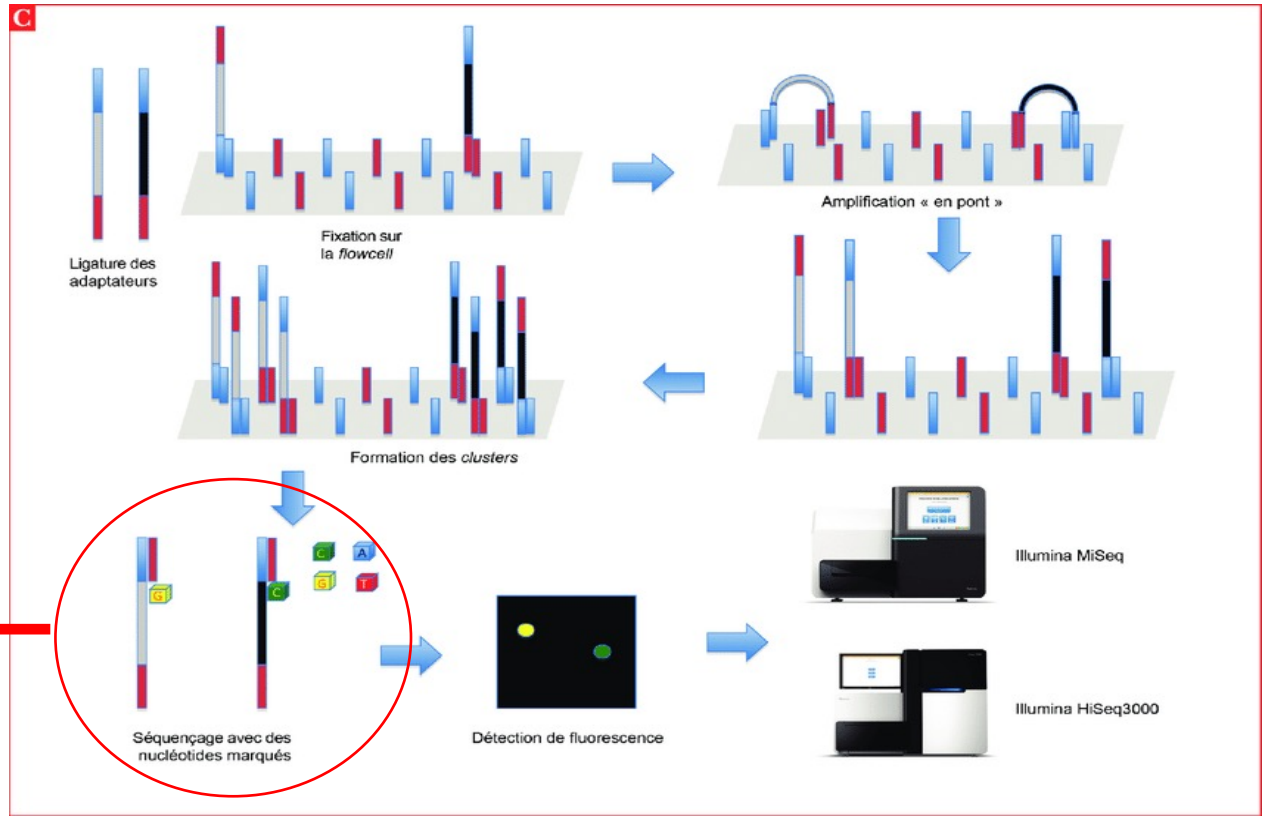
Type	LSU	SSU
prokaryotic	5S - 120 bp 23S - 2906 bp	16S - 1542 bp

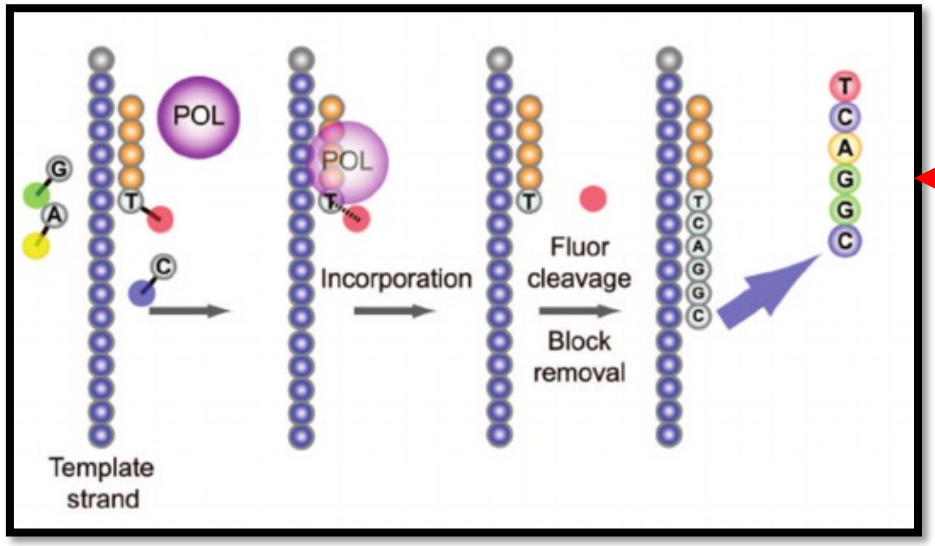
1. **Use of 5S : low phylogenetic power**
2. **Use of 23S: too long**
3. **Use of 16S : ideal length 1500 pb (adapted for Sanger sequencing)**

Next Generation Sequencing (NGS): Illumina Miseq

Flowcell

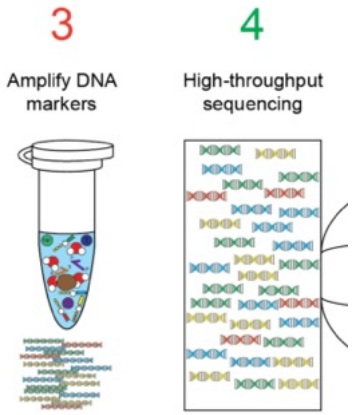






Only remember → Sequencing by reversible dye terminators

Illumina Miseq Sequencing Performance



Run Time	4–55 hours
Maximum Output	15 Gb
Maximum Reads Per Run	25 million †
Maximum Read Length	2 × 300 bp

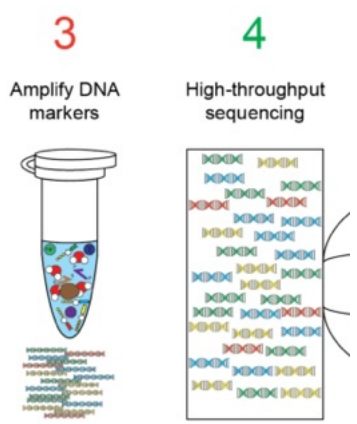
Massive data for complex ecosystem! good!

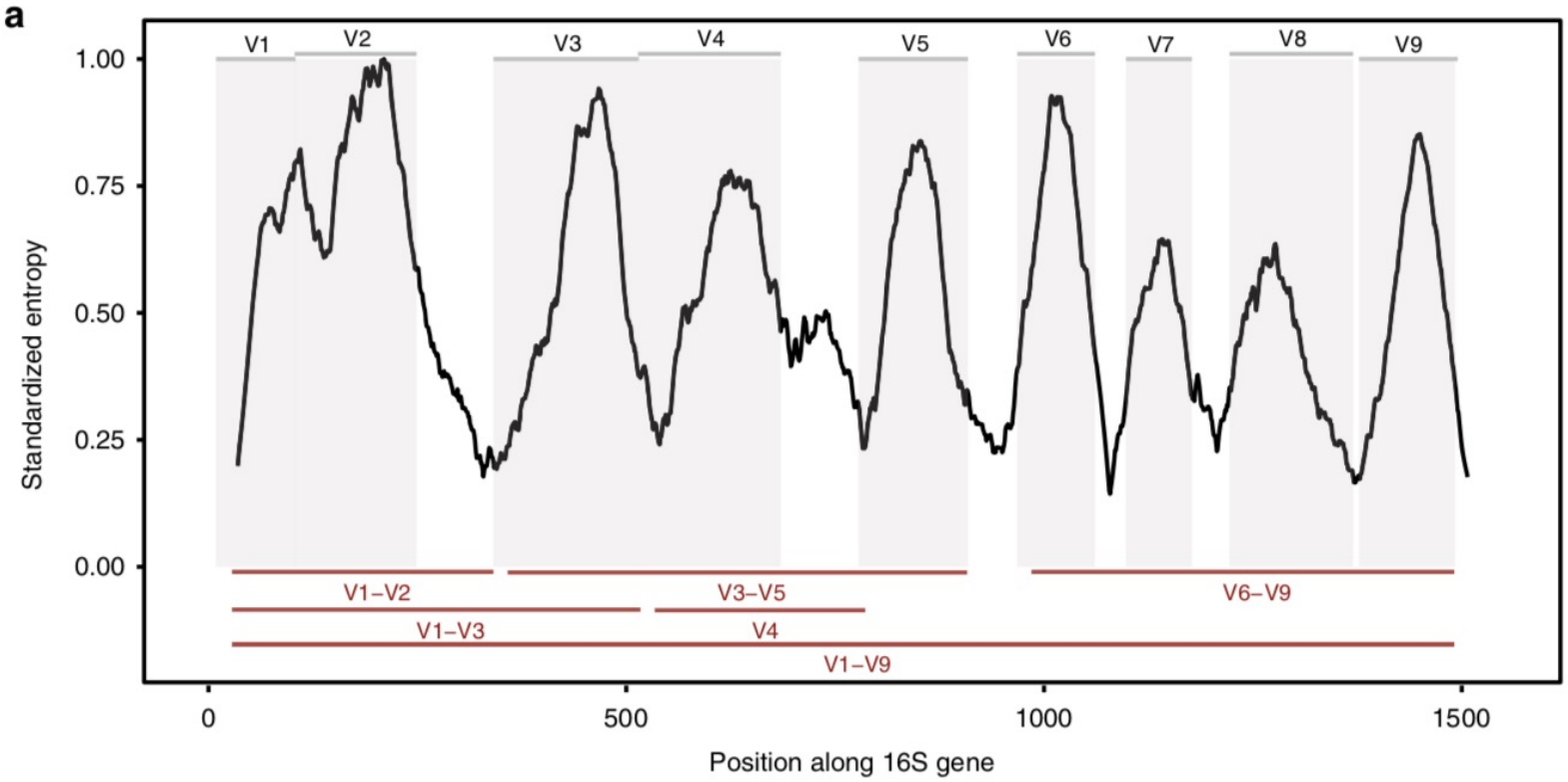
V3 Kit

Technology limit! oops!! Loss of resolutive power

	MiSeq Reagent Kit v2				MiSeq Reagent Kit v3	
Read Length	1 × 36 bp	2 × 25 bp	2 × 150 bp	2 × 250 bp	2 × 75 bp	2 × 300 bp
Total Time*	~4 hrs	~5.5 hrs	~24 hrs	~39 hrs	~21 hrs	~56 hrs
Output	540–610 Mb	750–850 Mb	4.5–5.1 Gb	7.5–8.5 Gb	3.3–3.8 Gb	13.2–15 Gb

Phylogenetic resolutive power & variable regions of 16S

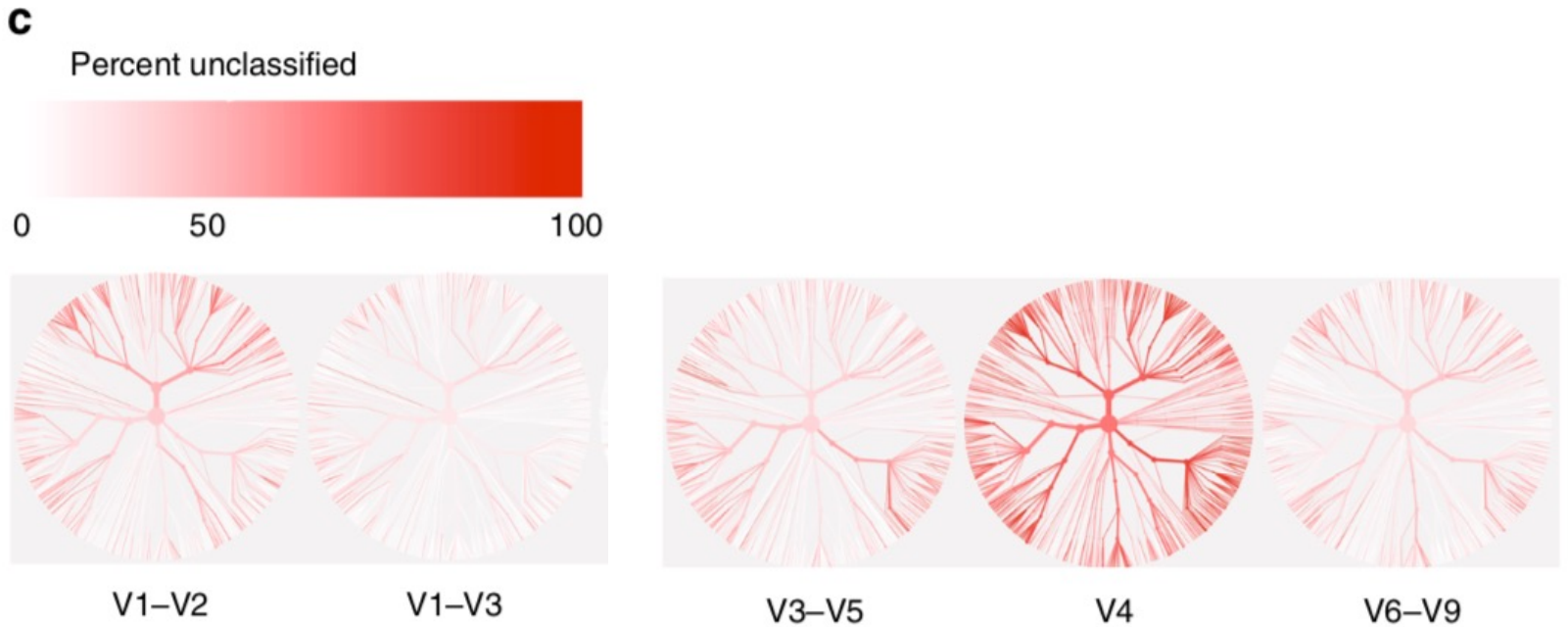




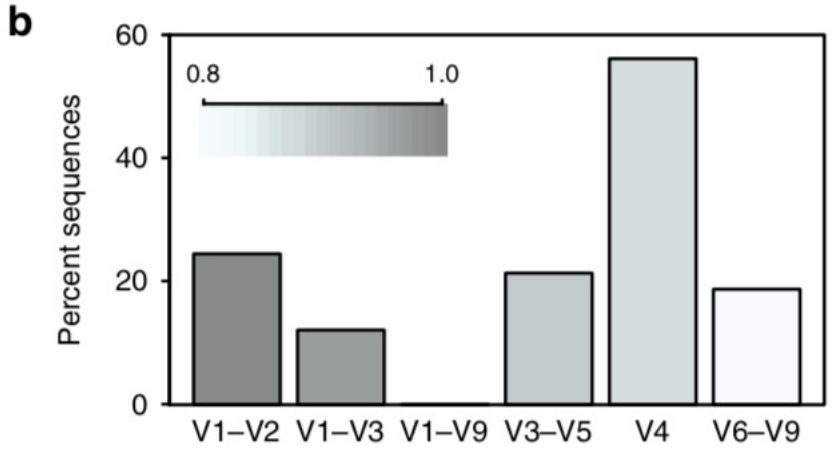
Multiple Sequence Alignment (MSA)
Of ALL 16S

Inspect each Nucleotide Position in MSA :
Mainly Conserved? -> low Entropy
Variable nucleotide? -> High entropy

Variable 16S region: Fail to classify at the species taxonomic level



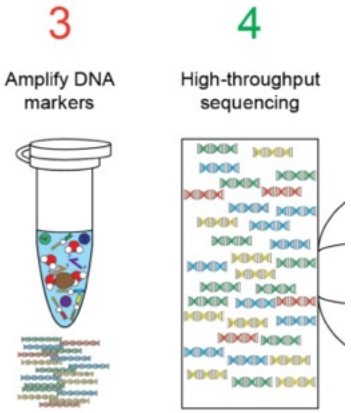
Johnson et al. Nature 2019



56% not assigned to species level with targeted V4 region

Greengene database
RDP classifier (Threshold 80)

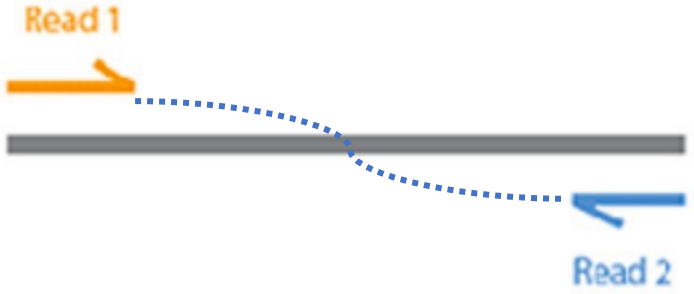
16S paired-end sequencing



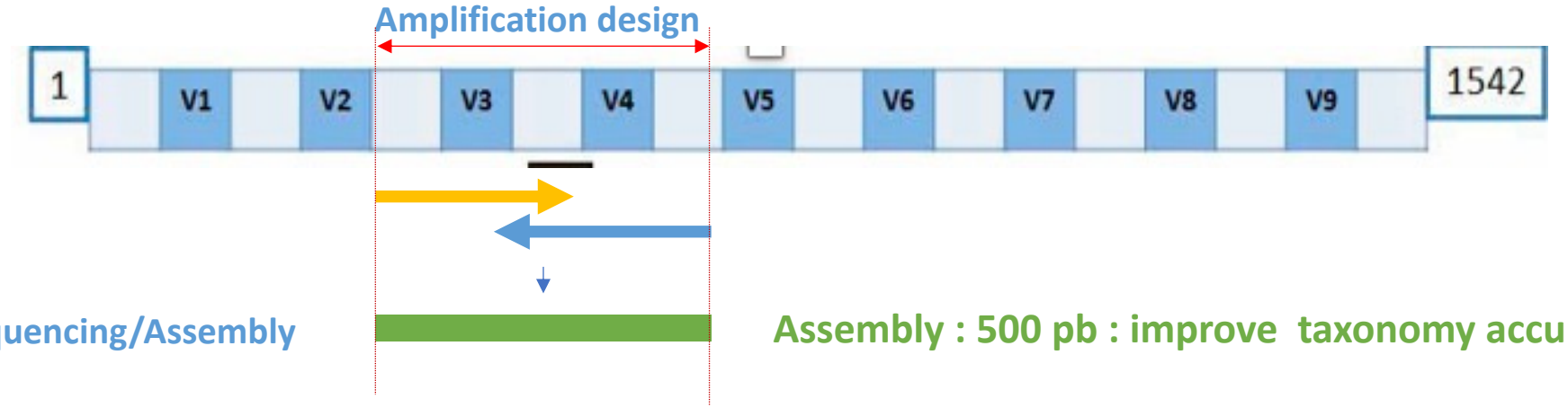
Run Time	4–55 hours
Maximum Output	15 Gb
Maximum Reads Per Run	25 million †
Maximum Read Length	2 × 300 bp

- **Miseq Illumina: Max read length is 300 bp**
- **Paired-end strategy increase the size of your target 16S region (Amplicon)**

Paired end : 2 X 300bp (Forward, Reverse)

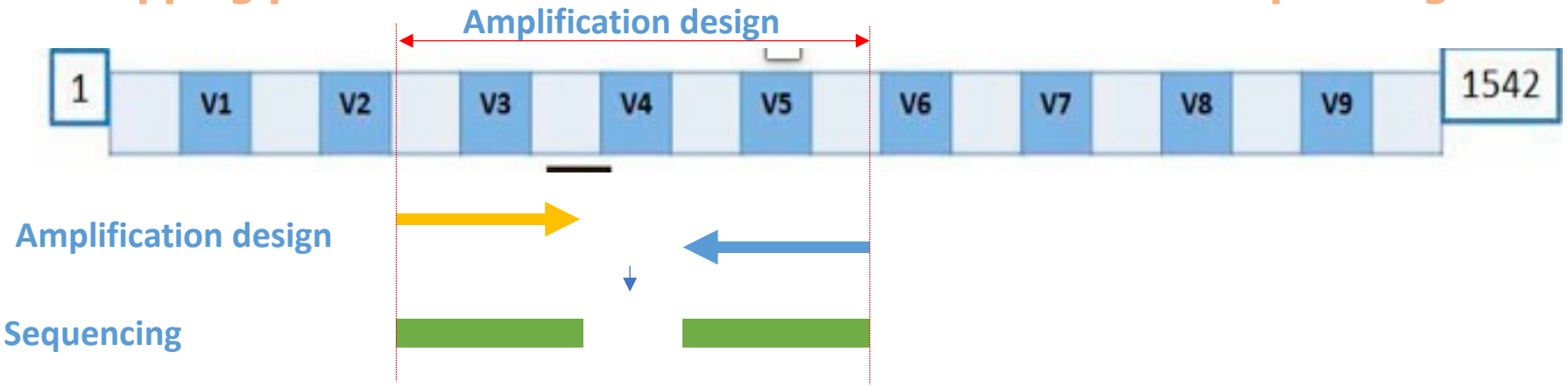


Overlapping paired-end reads : Assembly is possible = increase amplicon size



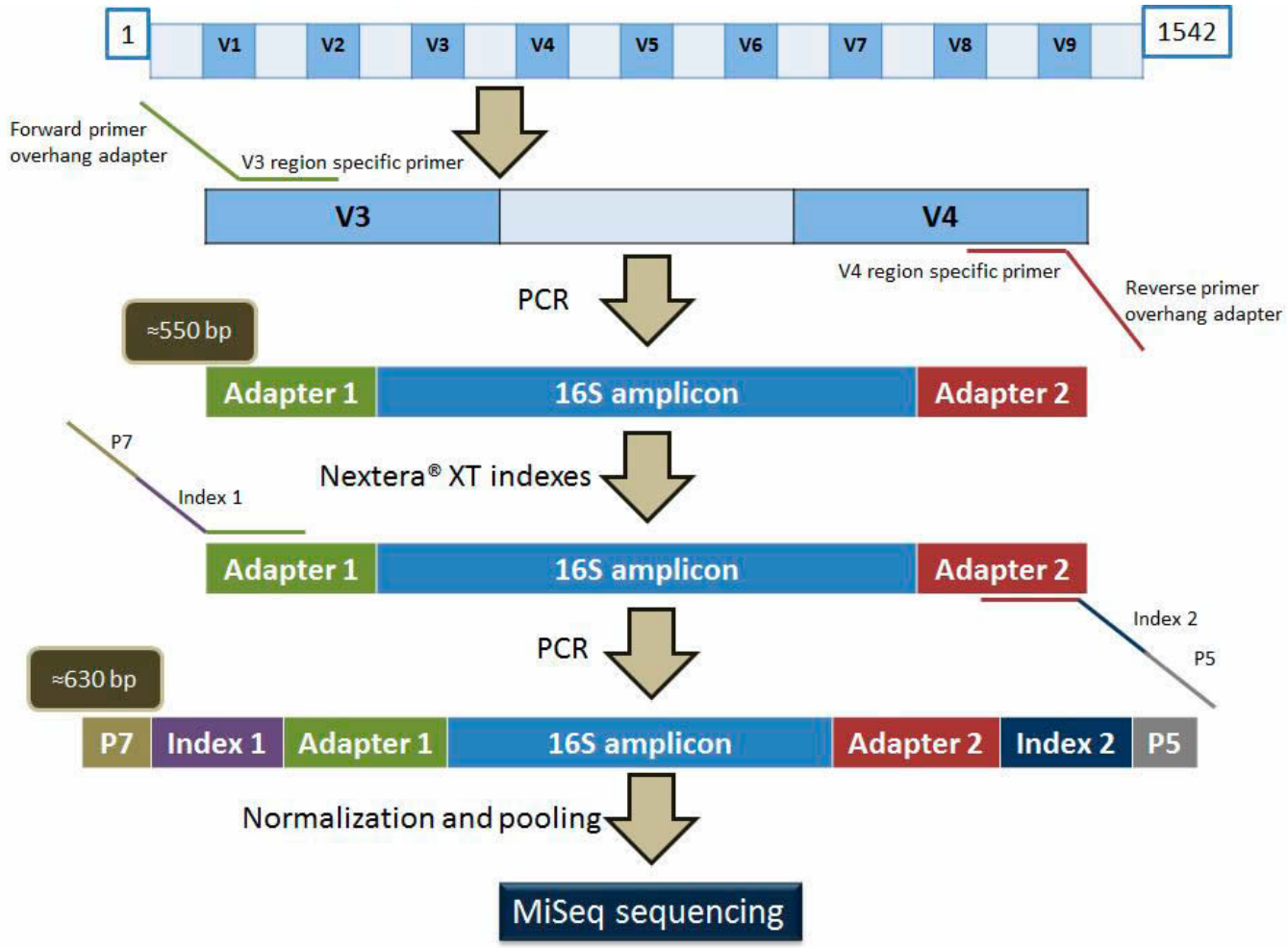
Assembly : 500 pb : improve taxonomy accuracy

Non Overlapping paired-end reads: You can not assemble the sequencing reads!!!

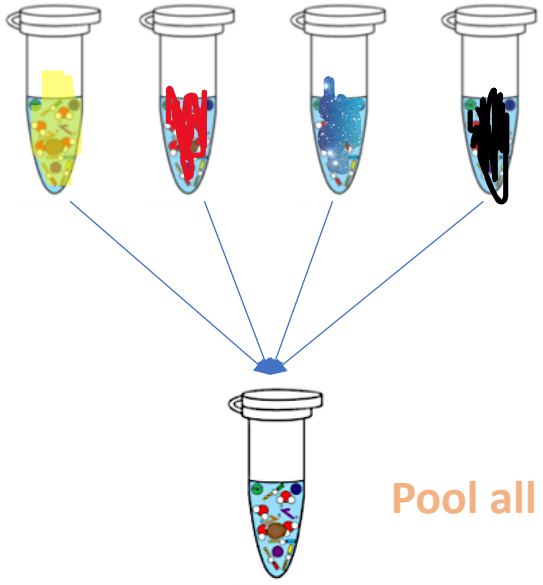


Wrong strategy

16S Miseq paired-end

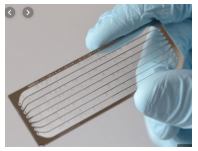


Index concept = Sample multiplexing on the same lane



16S Amplification for each Sample

Pool all samples!!!!!



Sequencing

Analyses : How to discriminate sequences according to sample?
Which sequence belongs to which sample??????
→ Index/barcode strategy

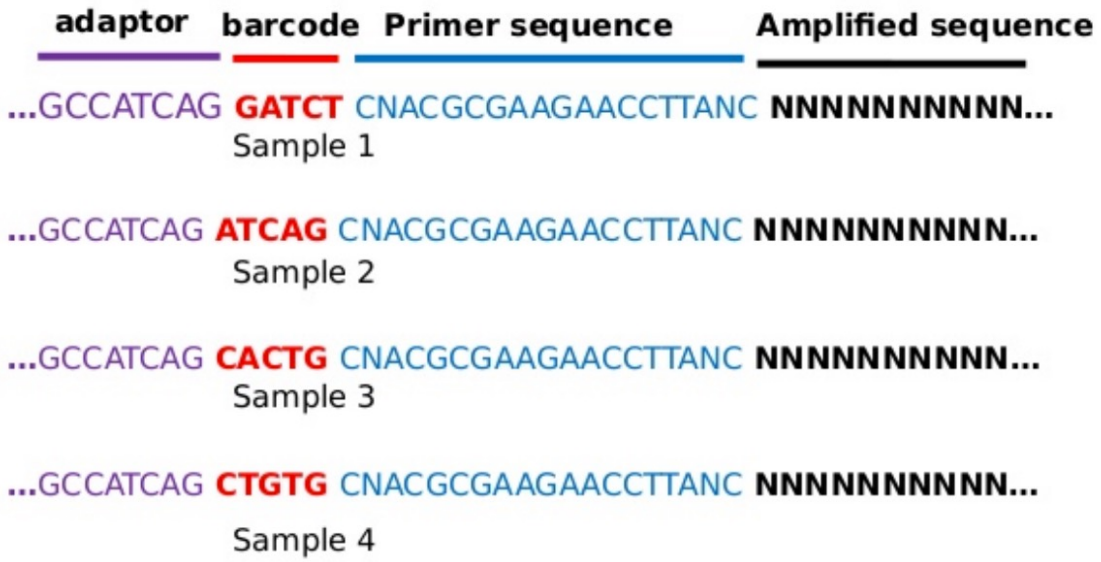
Normal PCR product

NGS-adapted PCR product

Barcoded PCR product for NGS

Add a different Index/barcode for each sample

Barcode



Bioinformatics

Sort sequencing reads of each Sample according to the barcode sequences