UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

# WORKSHOP FRIBOURG 2025

# INTRODUCTION TO METAGENOMICS DATA ANALYSIS OF MICROBIAL COMMUNITIES

**Laurent Falquet**
**Jeferyd Yepes-Garcia**

SIB Swiss Institute of Bioinformatics

## Modalities of the workshop

**Each topic is organised in 1h theory + 2h exercises (see moodle page)**
**https://moodle.unifr.ch/course/view.php?id=288715**

**or https://moodle.unifr.ch and search for**
**« SIBMetagenomicsWorkshop »**

**Key: SIB_Metagenomics25**

**Coffee breaks (near the teaching room)**
**Lunch breaks not included, but two affordable cafeteria nearby.**
**Social Apero this evening at the small cafeteria (5:30pm).**

## Program of the workshop

| Day | Topic |
|---|---|
| DAY1am | **General Introduction**, goals & opportunities<br>**Intro to NGS and QC** of the reads |
| DAY1pm | **Taxonomic classification**, contaminant removal |
| DAY2am | **Whole Genome Shotgun sequencing part 1**<br>Comparing Pipelines & workflow managers (i.e SnakeMake, Nextflow) |
| DAY2pm | **Whole Genome Shotgun sequencing part 2**<br>Downstream analyses, gene prediction prokka, annotation, function, pathways, eggNOG, KEGG, AntiSmash secondary pathways in bacteria |
| DAY3am | **Targeted sequencing** ASV vs OTU |
| DAY3pm | **Targeted sequencing** Downstream analysis |



## Who am I ?

Laurent Falquet
Swiss, born in Geneva

**PhD in Biochemistry at UniGe**
**Post-doc at the SIB Swiss Institute of Bioinformatics**
**Responsible of the Swiss EMBnet node**
**Project manager at Vital-IT High Performance Computing center**

Since 2013
**Senior Lecturer manager of the Bioinformatics core facility at UniFr, Switzerland**
**Group leader of the Swiss Institute of Bioinformatics**

## Microbiome amplicon projects: past

Exploring the phyllosphere of plants resistant to pathogens
**(Arabidopsis thaliana, leaf surface, 16S)**

Evaluating items of scientific findings in cases where the Defense says: 'It is my twin brother'
**(human saliva, 16S, rpoB)**

Involved in data analysis of the PromESSinG:
Promoting EcoSystem Services in Grapes https://www.promessing.eu/
**(soil, 16S, ITS)**

Targeted vaccination of the plant microbiome for sustainable crop protection: SOLANUM (SOiL ANtifungal vaccination Using native Microbiome) **with Prof. Laure Weisskopf, UniFr**
**(soil & leaves, 16S)**

Metagenomics analysis of mother & infant gut microbiome **with Prof. Petra Zimmermann, HFR**
**(gut, milk & saliva, WGS)**



## Metagenomics projects: current

Started in September 2022:

**Developing high-throughput strategies for efficient rice straw degradation through metagenomics data analysis and network reconstruction of metabolic pathways by deep learning approaches. (WGS, rice straw & soil)**

A collaboration with Prof. D.Uribe and Prof. E.Barreto, Colombia

Started in September 2023:

**(soil 16S + WGS, tomato and wheat rhizosphere)**

A collaboration with Dr. Bilal Rahmoune, Algeria

Other: MC member COST Action Machine Learning for Microbiome (finished) and WG member COST Action MiCropBiome

## Your turn!

## Please introduce yourself and your interest in metagenomics



UNI FR

UNI FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

# WORKSHOP FRIBOURG 2025

# WHAT IS METAGENOMICS?

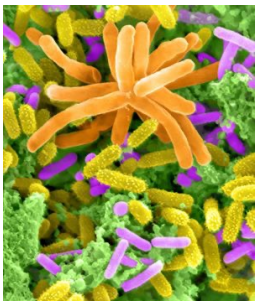
**Laurent Falquet**
**Jeferyd Yepes-Garcia**

SIB
Swiss Institute of Bioinformatics

## What is metagenomics?

**"The study of the DNA of uncultured organisms"**

> 99% of all microbes do not grow *in vitro*



*Methylamine-enriched community of Lake Washington sediment featuring Methylotenera cells Photo © Dennis Kunkel Microscopy, Inc. (Color by Ekaterina Latypova)*

### Some definitions…

### A genome:

Entire genetic information of a single organism

### A metagenome:

Entire genetic information of an assemblage/mixture of organisms

### A metatranscriptome:

Entire collection of transcripts of an assemblage/mixture at a given time



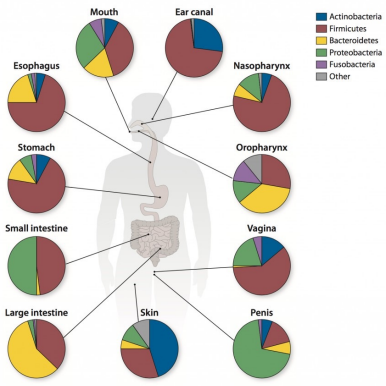
https://en.wikipedia.org/wiki/Metagenomics

## What is a microbiome?


### The microbiome (or microbiota):

is an ecological community of commensal, symbiotic and pathogenic microorganisms found in and on all multicellular organisms studied to date from plants to animals. A microbiota includes bacteria, archaea, protists, fungi and viruses.

https://en.wikipedia.org/wiki/Microbiota

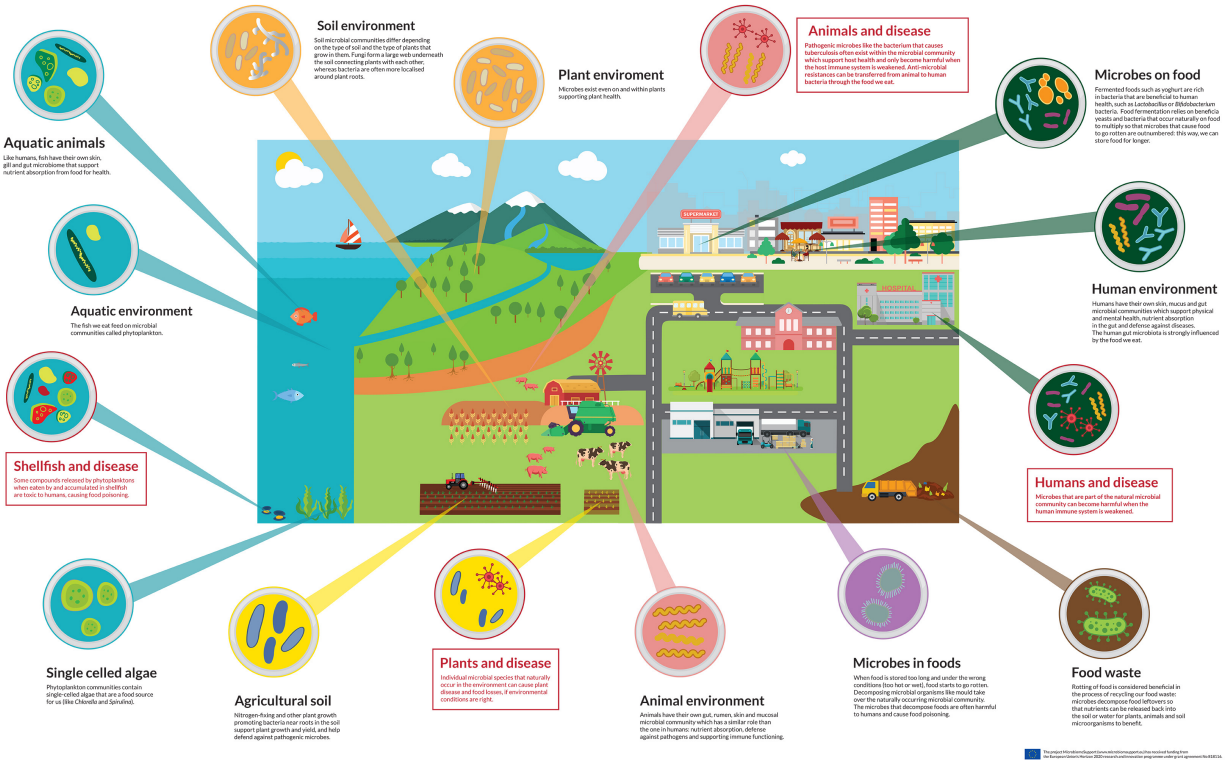



⚠ **Flora or microflora** synonym should not be used in scientific literature. It can be found in popular literature or in a yogurt/probiotic advertisement destined to the general public.



https://owlcation.com/stem/The-Ecology-of-the-Human-Microbiome

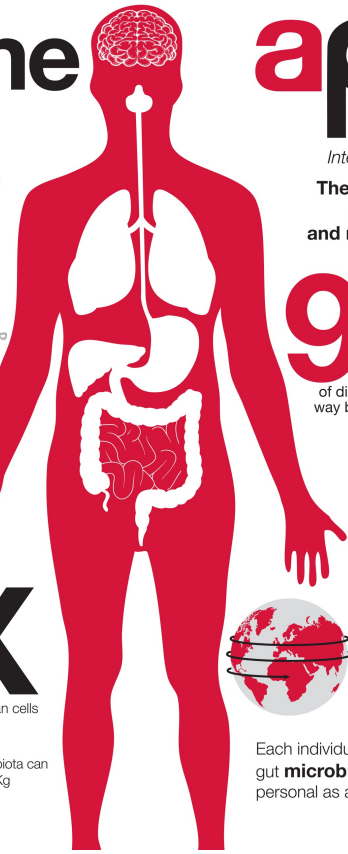
# Microbes are everywhere




**UNIVERSITÉ DE FRIBOURG / UNIVERSITÄT FREIBURG** | Biochemistry/Bioinformatics Unit | Falquet Laurent

https://worldmicrobiomeday.com/resources/

# Microbiome
IN NUMBERS



Microbiome Ireland
*Interfacing Food & Medicine*

**100 Trillion**
symbiotic microbes live in and on every person and make up the human microbiota

The human body has more microbes than there are stars in the milky way

**95%**
of our microbiota is located in the GI tract

**150:1**
The genes in your microbiome outnumber the genes in our genome by about 150 to one

The surface area of the **GI tract** is the same size as 2 tennis courts

**>10,000**
Number of different microbial species that researchers have identified living in and on the human body

You have **1.3X** more microbes than human cells

**2kg** The gut microbiota can weigh up to 2Kg

**The microbiome is more medically accessible and manipulable than the human genome**

It is thought that **90%** of disease can be linked in some way back to the gut and health of the microbiome

**5:1**
**Viruses:Bacteria** in the gut microbiota

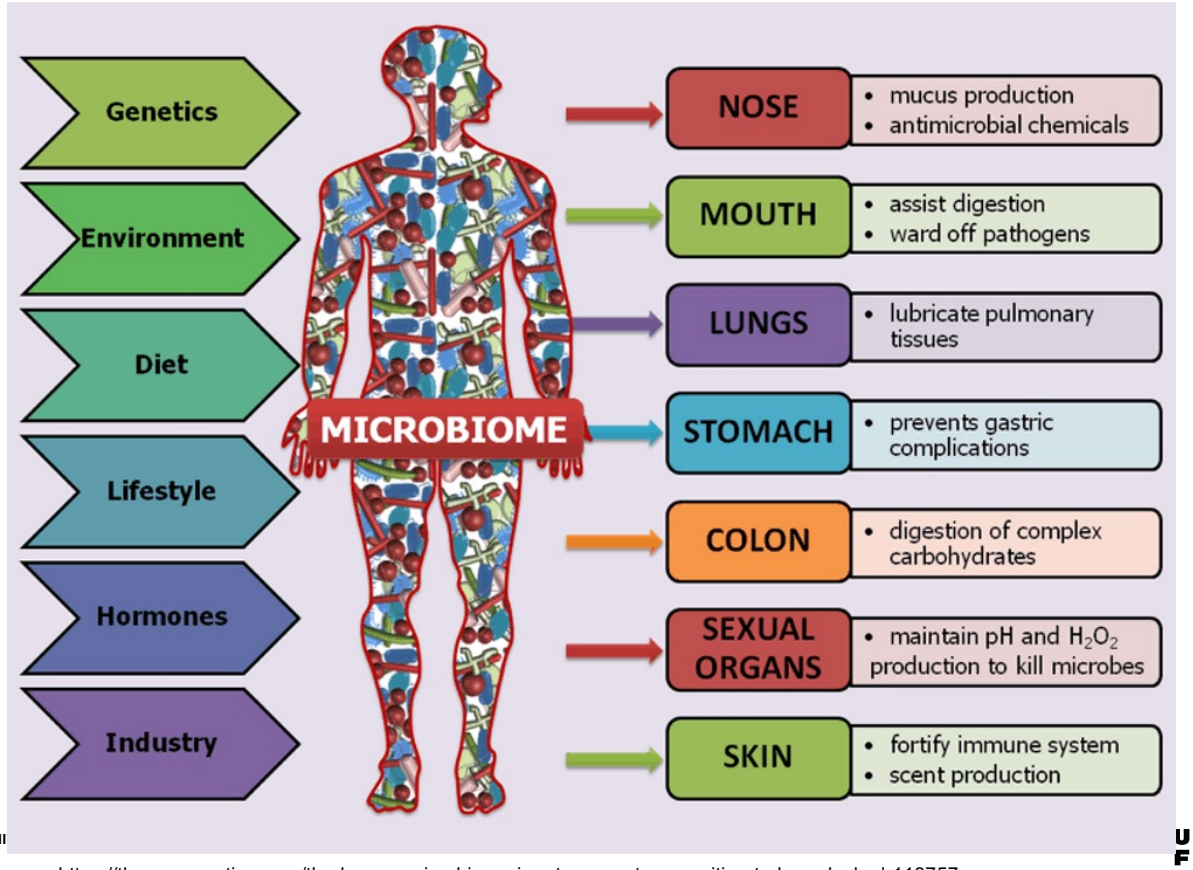
**2.5** The number of times your body's microbes would circle the earth if positioned end to end

Each individual has a unique gut **microbiota**, as personal as a fingerprint

**UNIVERSITÉ DE FRIBOURG / UNIVERSITÄT FREIBURG** | Biochemistry/Bioinformatics Unit | Falquet Laurent

https://worldmicrobiomeday.com/resources/

## Microbes are everywhere on us




https://theconversation.com/the-human-microbiome-is-a-treasure-trove-waiting-to-be-unlocked-118757

## Examples of roles and impact of the microbiome in the human gut

### Benefits

- digestion, production of vitamins & antioxidants, iron absorption, fighting pathogens

### Diseases

- obesity, diabetes, some cancers, heart disease

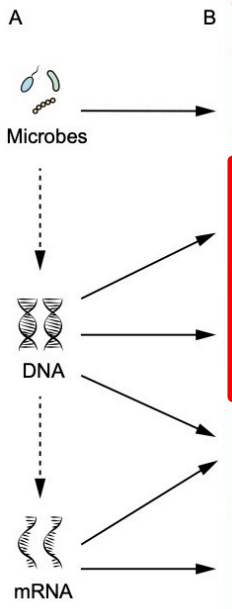
### Immunity

- eczema & allergy, asthma, autoimmune disorders

### Behavior

- psychiatric disorders (depression), autism

## How to study a microbiome/metagenome?




| Method | Advantages | Limitations |
|---|---|---|
| Culturome | • High-throughput<br>• Targeted selection<br>• Provides microbial isolates | • Expensive<br>• Laborious<br>• Influenced by media and the environment |
| Amplicon (16S/18S/ITS) | • Quick analysis<br>• Low-biomass requirement<br>• Applicable to samples contaminated by host DNA | • PCR and primer biases<br>• Resolution limited to genus level<br>• False positive in low-biomass samples |
| Metagenome | • Taxonomic resolution to species or strain level<br>• Functional potential<br>• Uncultured microbial genome | • Expensive<br>• Time-consuming in analysis<br>• Host-derived contamination |
| Virome | • Can identify RNA and DNA viruses<br>• Quick diagnosis | • Most expensive<br>• Difficult to analysis<br>• Severe host-derived contamination |
| Metatranscriptome | • Can identify live microbes<br>• Can evaluate microbial activity<br>• Transcript-level responses | • Complex sample collection and analysis<br>• Expensive and complex in sequencing<br>• Host mRNA and rRNA contamination |

Protein Cell 2021, 12(5):315–330
https://doi.org/10.1007/s13238-020-00724-8



## The choice of the method depends on 3 main questions

**Who is there? Taxonomic profiling**

catalogue of species, diversity, genus, etc.
distribution (how many of each)

**What are they doing? Functional analysis**

genes/proteins, GO terms, metabolic pathways
functional annotation, bioprospection

**How do they compare? Differential analysis**

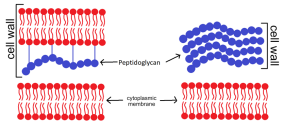
pairwise or multiple comparisons
correlation with environmental factors


**In addition: Meta-analysis**

combining multiple studies

## What is challenging?

- easy + hard




### Lab challenges:

**Different DNA extraction steps can give different sequence sets**

### Data analysis challenges:

**Huge amount of sequences to process**

**Most sequences are absent from the existing reference databases**

**Assembly of communities can be hard with short reads**

**No perfect tool or workflow for analysis exists**

**Different analysis tools can give different results**

Even worse, the same version of a tool can give different results depending on the reference database used



## Technical options for metagenome/microbiome sequencing

### Long reads vs short reads?

| Read type | Advantage | Disadvantage |
|---|---|---|
| Short (Illumina) | Cheaper to produce<br>More sensitive | Many biases |
| Long (Nanopore) | Longer sequences | More expensive<br>High error rate |
| Long (PacBio) | Longer sequences<br>HiFi reads | Less counts |

## Targeted vs WGS

| Amplicon (targeted) | WGS |
|---|---|
| Sequence alignment to a reference database is sufficient | Requires *de novo* assembly |
| | |
| PCR selection of a gene of interest | No selection |
| 16S or ITS or COI or other… | all DNA is sequenced (contaminants!) |
| Error correction (denoising) | QC/cleaning + *de novo* assembly |
| Clustering to ASV or OTU | Binning to MAG |
| Taxonomic assignment | Taxonomic assignment |
| No function available | Functional assignment |
| | |
| Downstream analysis | Downstream analysis |
| E.g., alpha and beta diversity | E.g., comparative genomics |

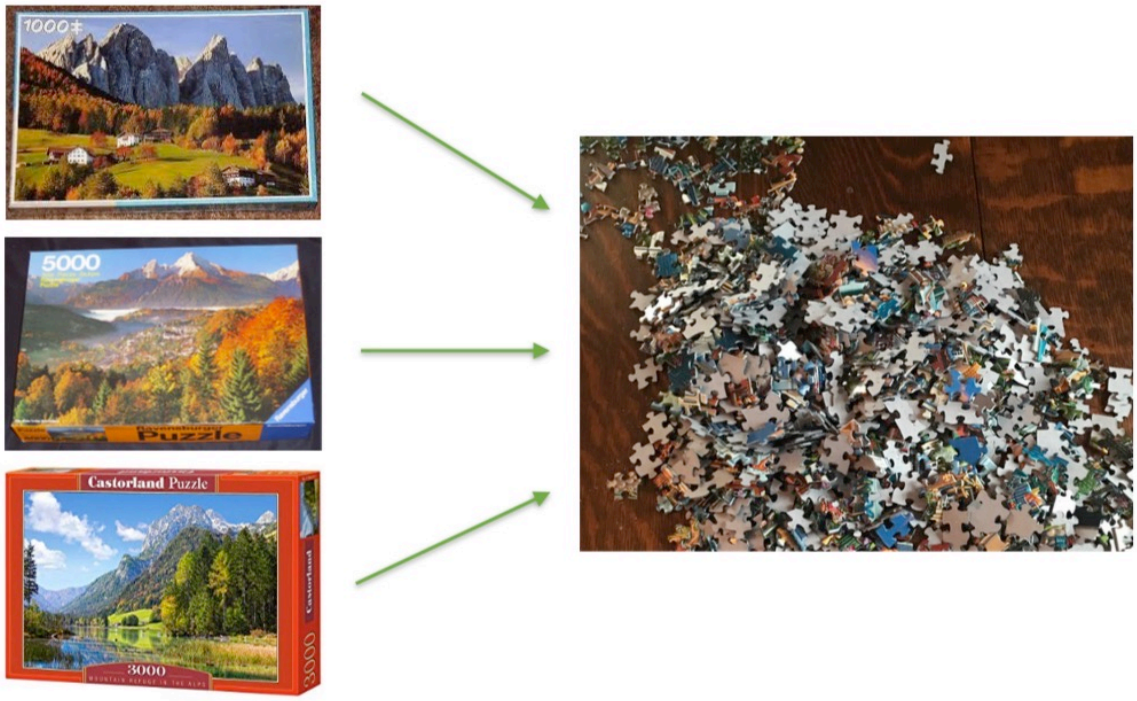


## Metagenomics via Whole Genome Shotgun (WGS)

### Difficulties:

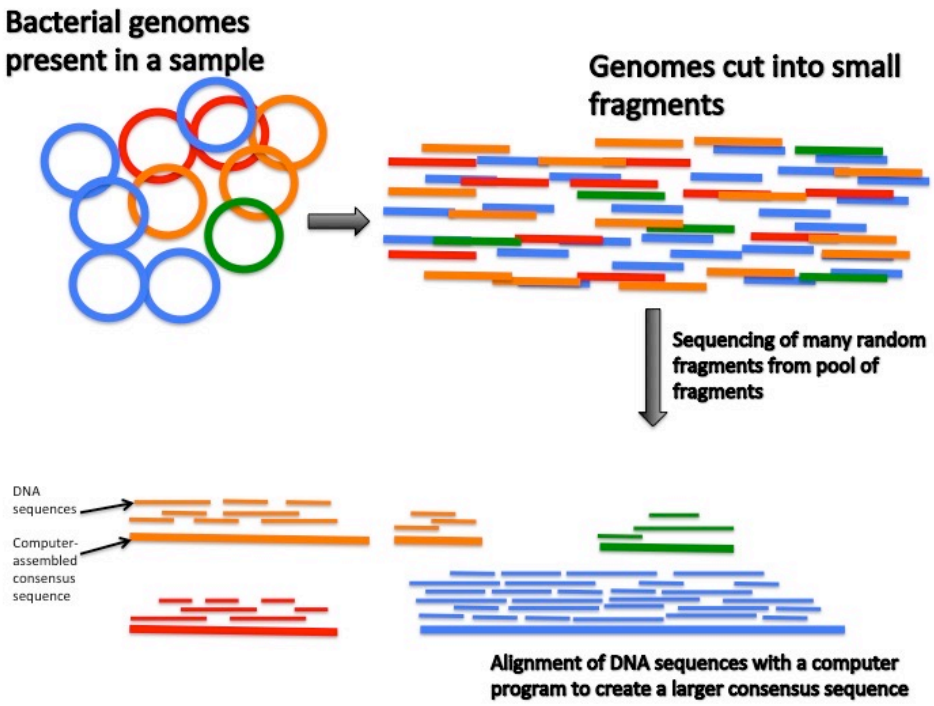
- **Requires *de novo* assembly**
- **Many genomes (hundreds to thousands)**
- **Often closely related**
- **Coverage very different (abundant vs rare)**
- **Large amounts of data**
- **Host contamination**

## Like thousands jigsaw puzzles mixed together!!






## Whole metagenome assembly



## After genome assembly

- Most metagenomics assemblers use de Bruijn graphs
- Even single genome assemblies are often still highly fragmented
- Metagenome assemblies are worse as many methods used for reducing the fragmentation cannot be applied in a mixture of genomes

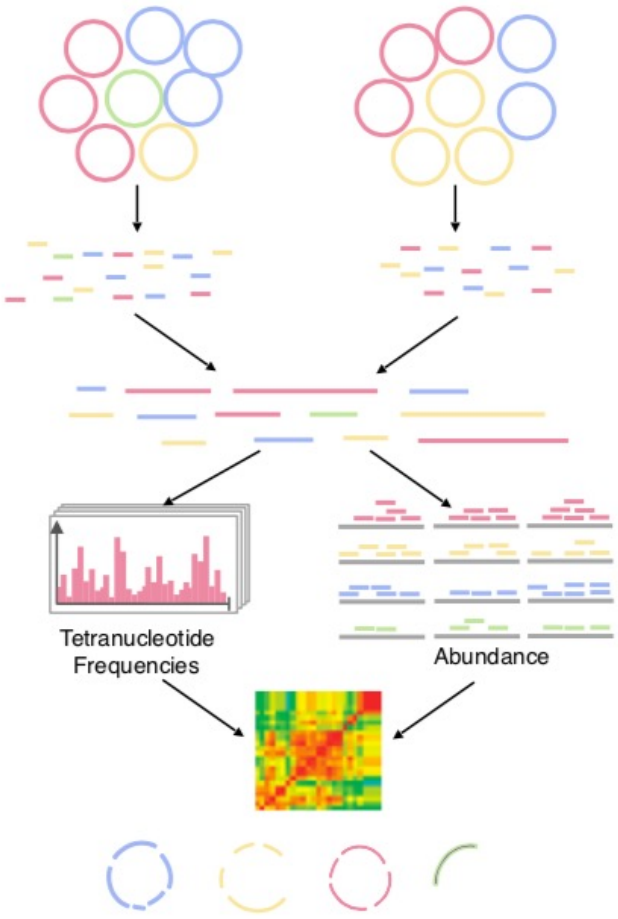
- The idea is to sort the data into smaller groups or « bins »




## Assembly and binning

**Binning can use various data to produce Metagenome Assembled Genomes (MAGs)**

1) Abundance (coverage)
2) Tetranucleotide frequencies

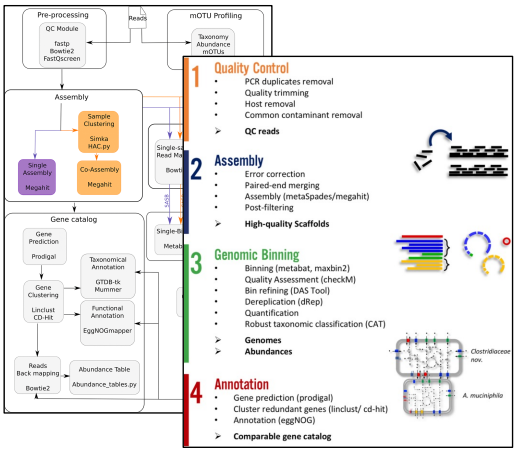


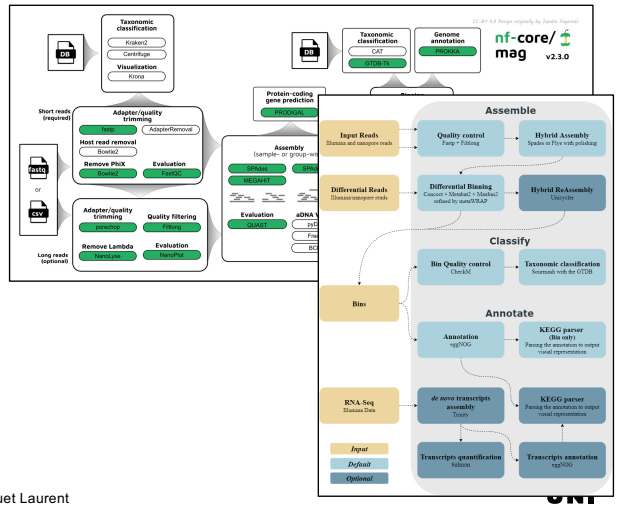
## Complete automated workflows and pipelines!

### Most use Snakemake or Nextflow or custom pipelines

snakemake

nextflow






## Automated workflows and pipelines differences

**Many pipelines are currently in development**
**They usually start from raw short reads or contigs and go to various levels of analysis (mainly to Binning into MAGs).**
**Some pipelines go even further to annotation of gene catalogs.**

| | |
|---|---|
| MGnify | https://www.ebi.ac.uk/metagenomics/pipelines/5 |
| MG-RAST : | https://www.mg-rast.org |
| MOCAT2 : | https://mocat.embl.de/about.html |
| Metagenome-Atlas | https://metagenome-atlas.readthedocs.io |
| nf-core/mag | https://nf-co.re/mag |
| MUFFIN | https://github.com/RVanDamme/MUFFIN |
| METAWRAP | https://github.com/bxlab/metaWRAP |
| MAGNETO | https://gitlab.univ-nantes.fr/bird_pipeline_registry/magneto |

**Many others…**

**The differences are mainly in the tools and databases used, but more importantly the strategies for assembly (co-assembly) or binning (co-binning) or both.**

**Most pipelines are designed for short reads!**

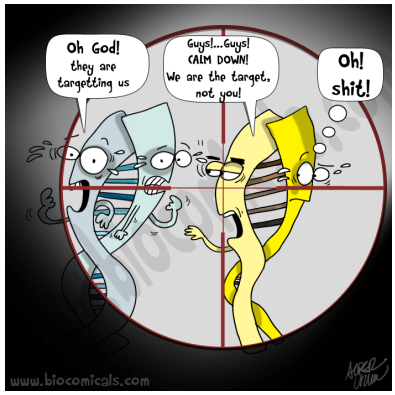
## Targeted vs WGS

| Amplicon (targeted) | WGS |
|---|---|
| Sequence alignment to a reference database is sufficient | Requires *de novo* assembly |
| | |
| PCR selection of a gene of interest | No selection |
| 16S or ITS or COI or other… | all DNA is sequenced (contaminants!) |
| Error correction (denoising) | QC/cleaning + *de novo* assembly |
| Clustering to ASV or OTU | Binning to MAG |
| Taxonomic assignment | Taxonomic assignment |
| No function available | Functional assignment |
| | |
| Downstream analysis | Downstream analysis |
| E.g., alpha and beta diversity | E.g., comparative genomics |



## Targeted or Amplicon sequencing

**Library preparation requires PCR amplification**
**QC**
**Clean reads**
optional: combine overlapping PE
**Align reads to reference database**
**Clusterize reads to OTUs**
**Assign OTUs to taxonomy**




**Tools (examples):**

Mothur
Qiime2
MapSeq
MEGAN
Metaphlan
PIPITS (for fungi)
...

## Choosing the target of PCR amplification

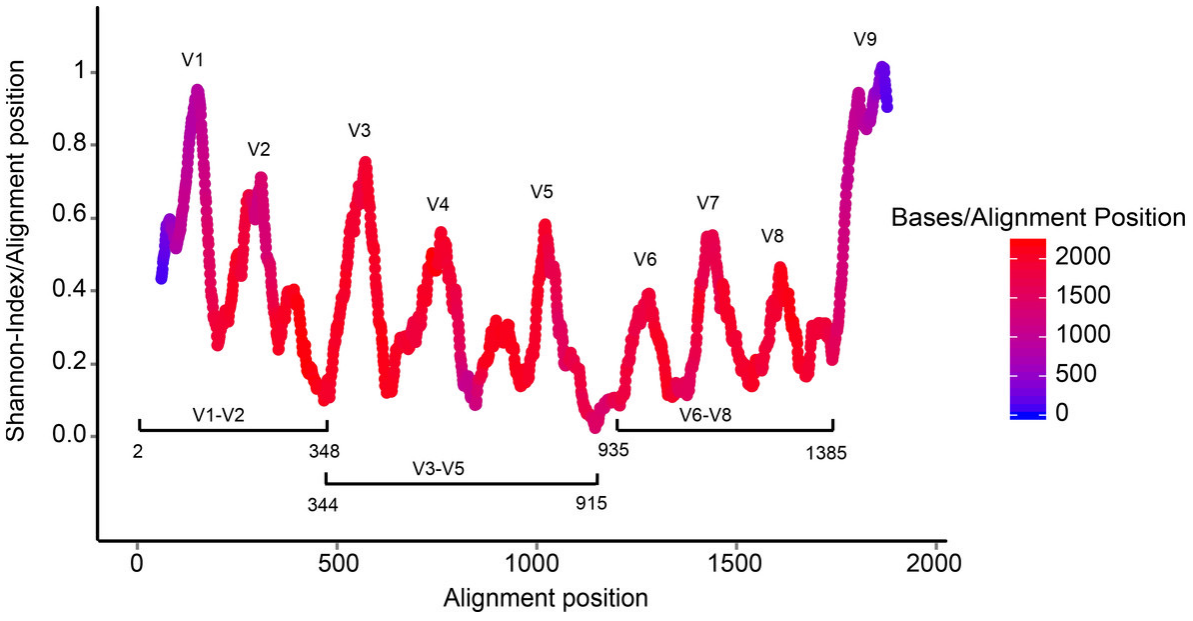
**Usually this is a gene with some variable sequences in all species surrounded by conserved sequences:**

**Several genes have been used : 16S, 18S, ITS, gyrB, rpoB, recA etc.**

**The most popular being the 16S for bacteria.**



## Selection of primers: reminder about 16S rRNA variable regions

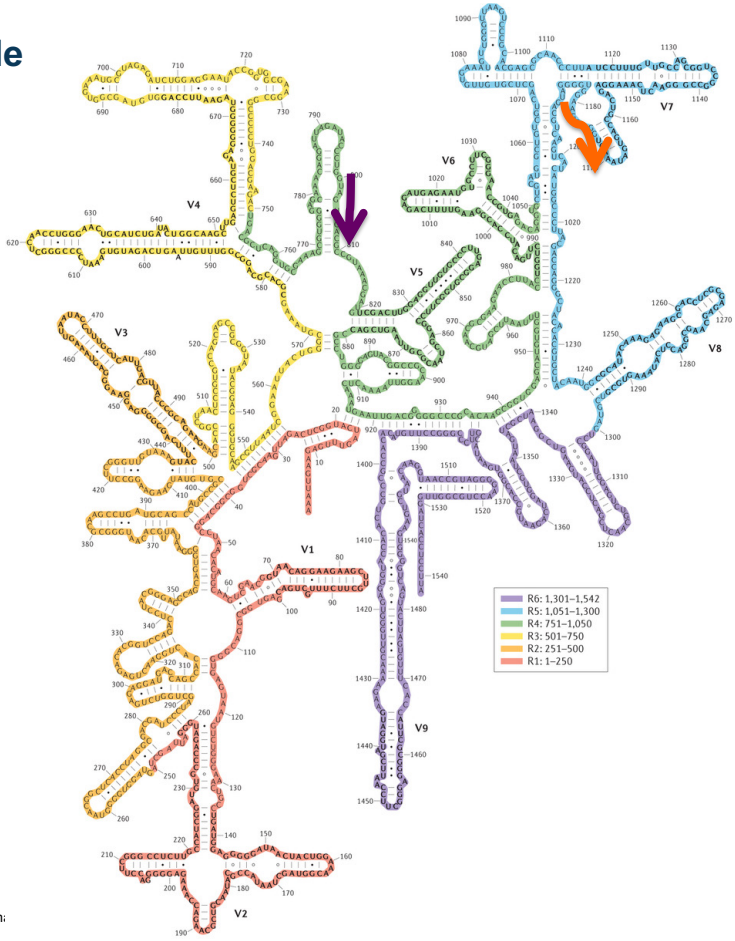


## Selection of the primers example

### 16S rRNA

### primers:
### 799F
### 1193R
### distance ~400bp

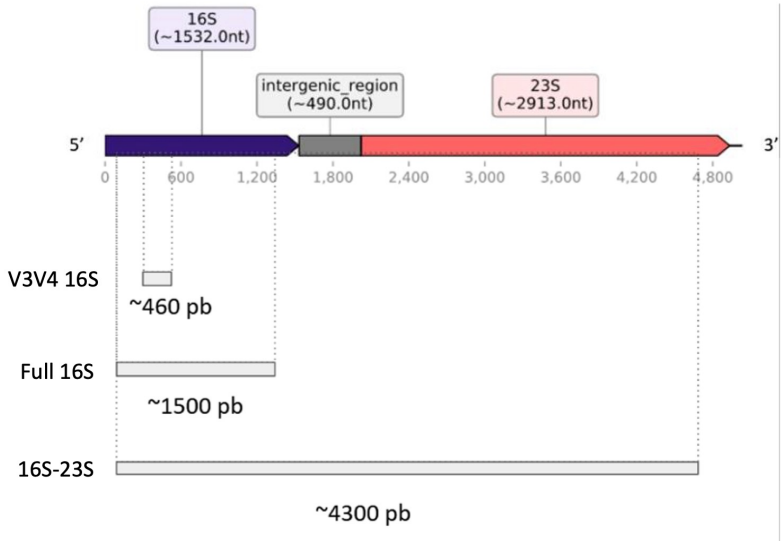
### Sequenced by MiSeq
### 2x 300bp -> overlap!



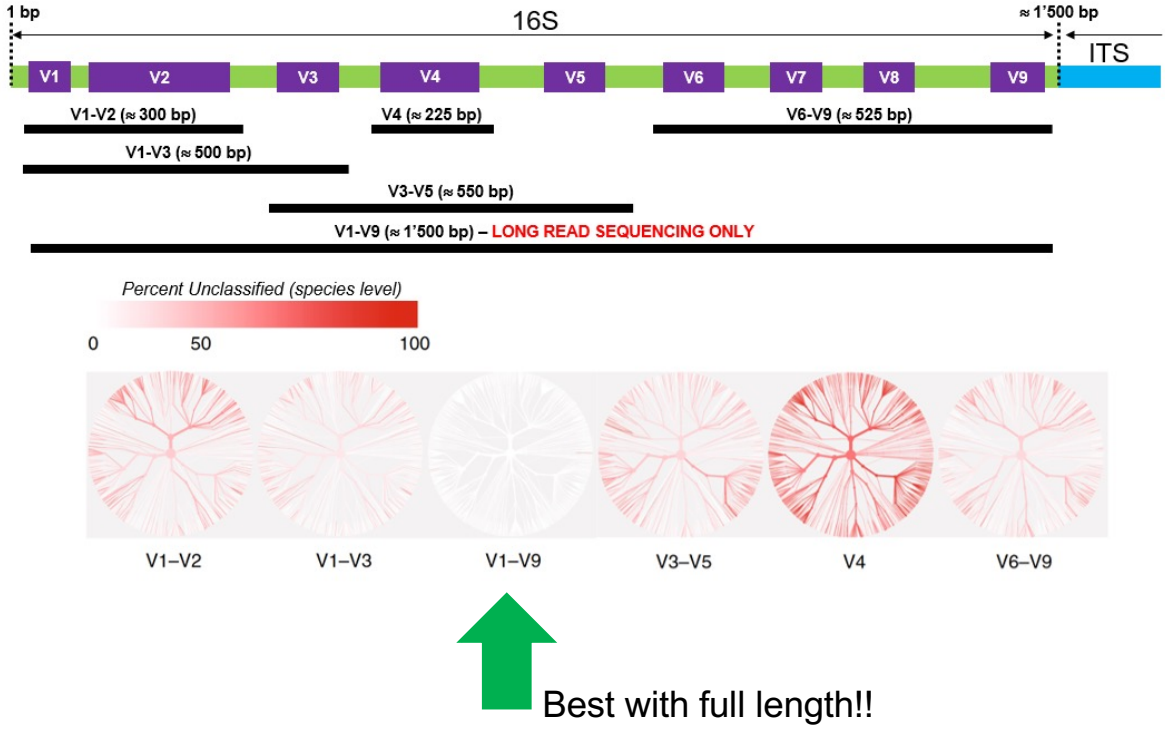

Nature Reviews | Microbiology

## Full length amplicon sequencing

### With the improvement of long read sequencing and the development of other long read techniques like LoopSeq, it becomes possible to sequence full length 16S or even larger amplicons



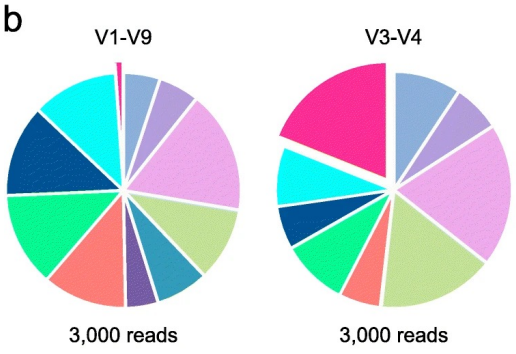
# Improved classification with full length 16S



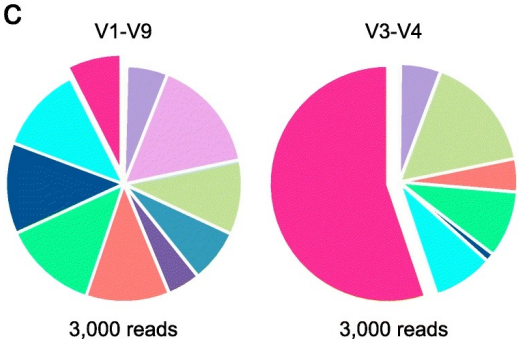

Best with full length!!



# Example MOCK analysis with full length 16S *vs* V3-V4 (10 strains even mix)




| Genus | V1-V9 | V3-V4 |
|---|---|---|
| *Bacillus* | 5.07% | 9.30% |
| *Bifidobacterium* | 5.63% | 6.60% |
| *Clostridium* | 16.97% | 19.70% |
| *Deinococcus* | 10.23% | 16.10% |
| *Enterococcus* | 7.30% | 0.13% |
| *Escherichia* | 4.50% | 0.00% |
| *Lactobacillus* | 11.63% | 5.77% |
| *Rhodobacter* | 12.97% | 9.17% |
| *Staphylococcus* | 12.73% | 5.97% |
| *Streptococcus* | 11.87% | 8.30% |
| mis- or unclassified | 1.10% | 18.97% |

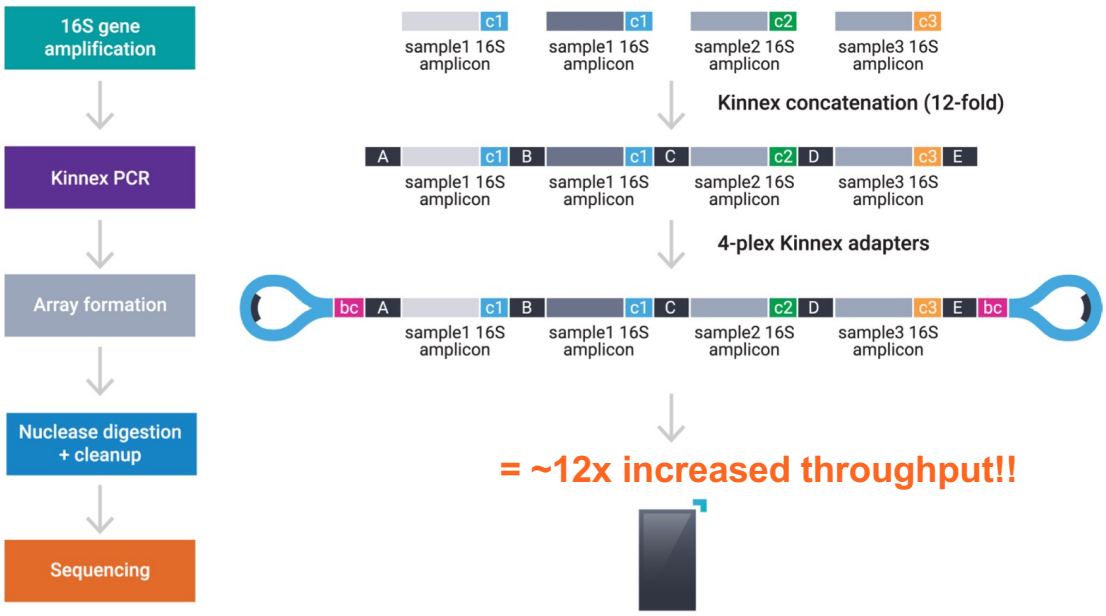



| Species | V1-V9 | V3-V4 |
|---|---|---|
| *B. cereus* | 0.40% | 0.17% |
| *B. adolescentis* | 5.63% | 5.53% |
| *C. beijerinckii* | 15.67% | 0.03% |
| *D. radiodurans* | 10.23% | 16.07% |
| *E. faecalis* | 7.30% | 0.00% |
| *Escherichia* spp. | 4.50% | 0.00% |
| *L. gasseri* | 11.47% | 4.57% |
| *R. sphaeroides* | 12.97% | 9.07% |
| *S. epidermidis* | 12.50% | 1.20% |
| *S. mutans* | 11.87% | 8.20% |
| mis- or unclassified | 7.47% | 55.17% |

Matsuo et al. BMC Microbiology (2021) 21:35 https://doi.org/10.1186/s12866-021-02094-5

## With PacBio: Kinnex 16S rRNA library

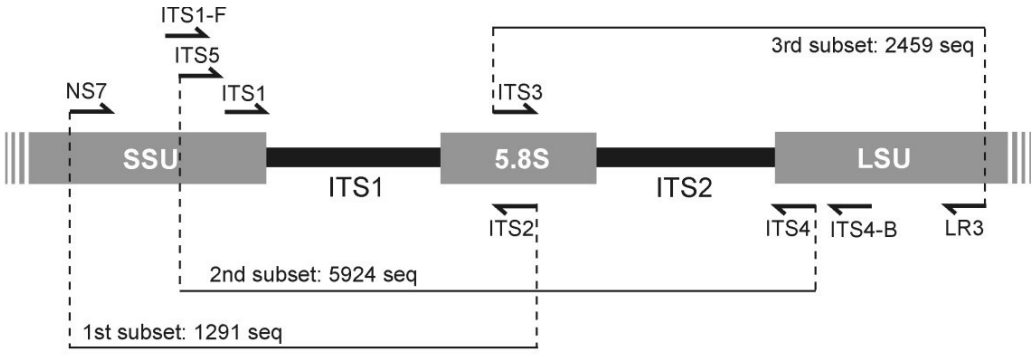
### Combining multiple full length 16S (up to 12) in one HiFi read!



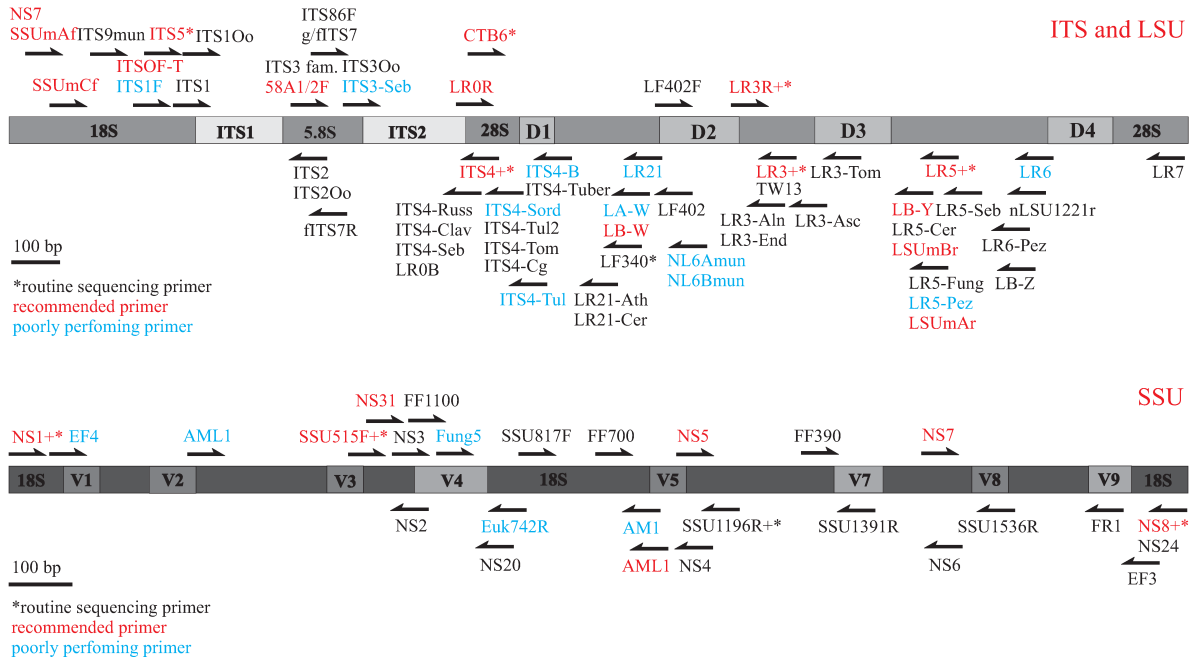



## ITS

### For fungi, the most widely used fungal genetic markers are the internal transcribed spacers (ITS1 and ITS2) of the nuclear ribosomal subunit, which are located between the small and large subunit gene (SSU/18S and LSU/28S, respectively).



## List of primers for ITS and 18S: see UNITe web site https://unite.ut.ee/primers.php

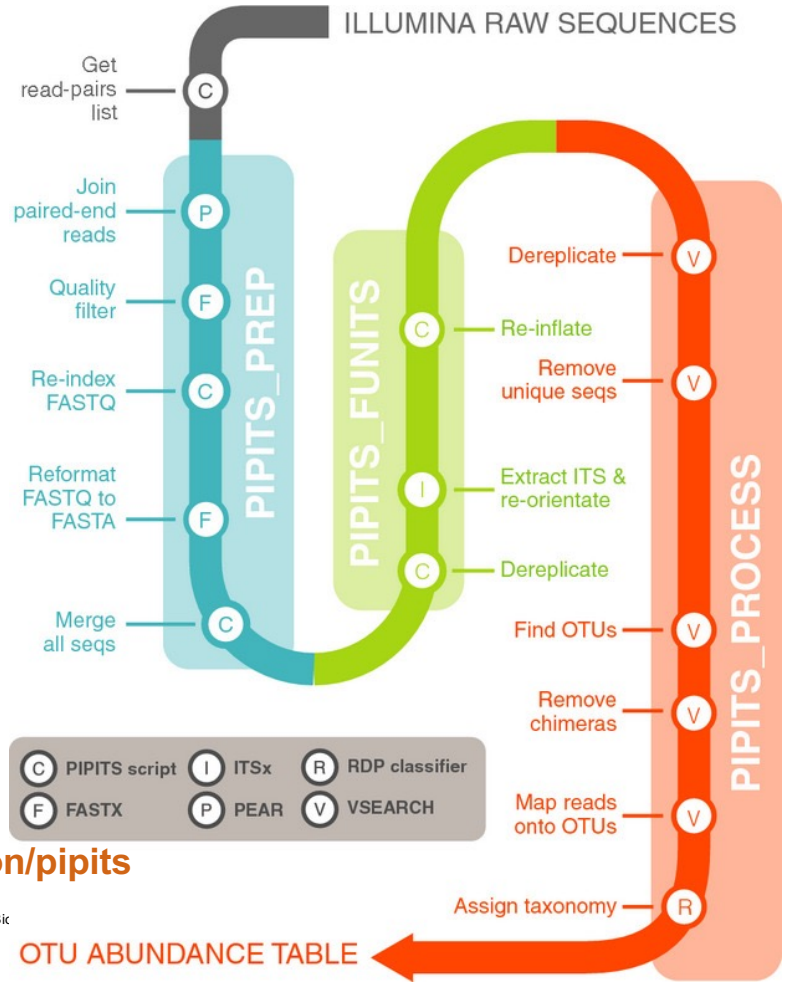





## ITS pipelines

### PIPITS3

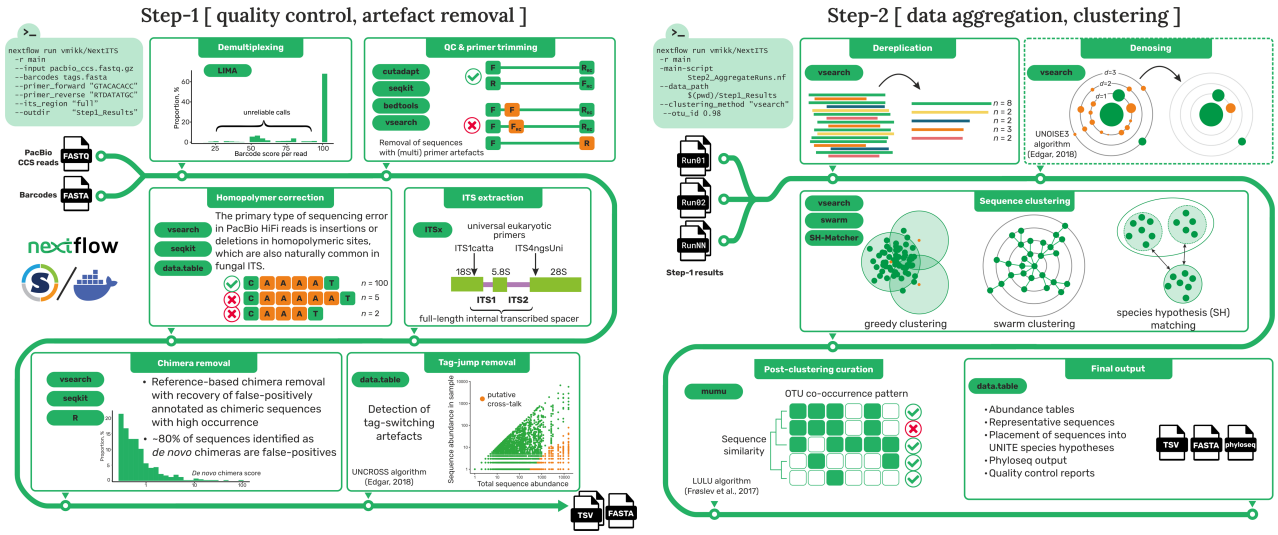
**A 3 steps pipeline generating OTUs abundance table and their classification**




**https://github.com/hsgweon/pipits**

# NextITS https://next-its.github.io

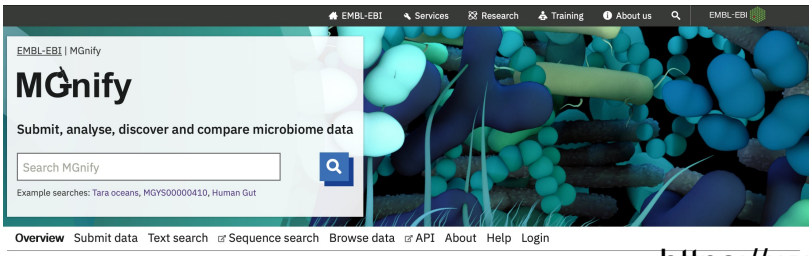
**NextITS** is an automated pipeline for metabarcoding fungi and other eukaryotes with full-length ITS sequenced with PacBio.
(clustering to OTU)



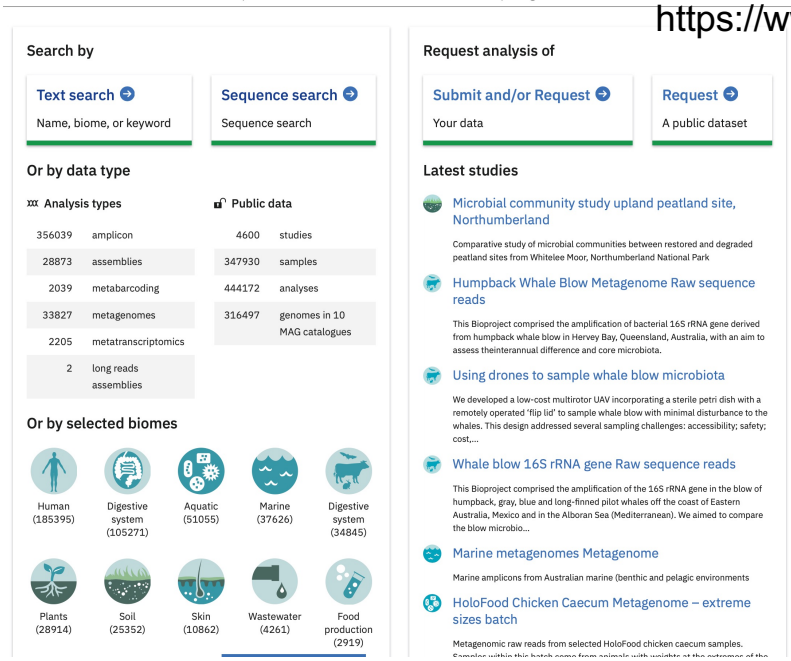

**UNIVERSITÉ DE FRIBOURG / UNIVERSITÄT FREIBURG** | Biochemistry/Bioinformatics Unit | Falquet Laurent

Mikryukov V., Anslan S., Tedersoo L. NextITS: a pipeline for metabarcoding fungi and other eukaryotes with full-length ITS sequenced with PacBio. https://github.com/vmikk/NextITS

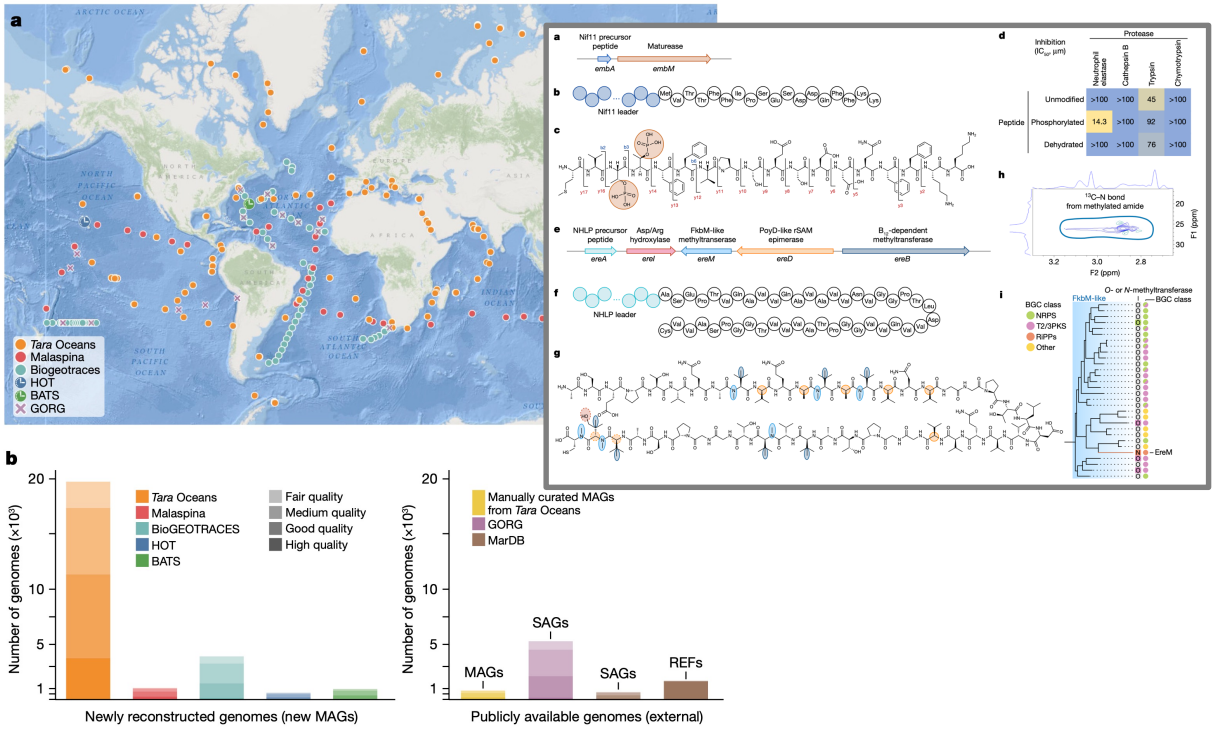
# MGnify the database available from EBI




https://www.ebi.ac.uk/metagenomics/



## Ocean metagenome & pathways https://microbiomics.io/ocean/

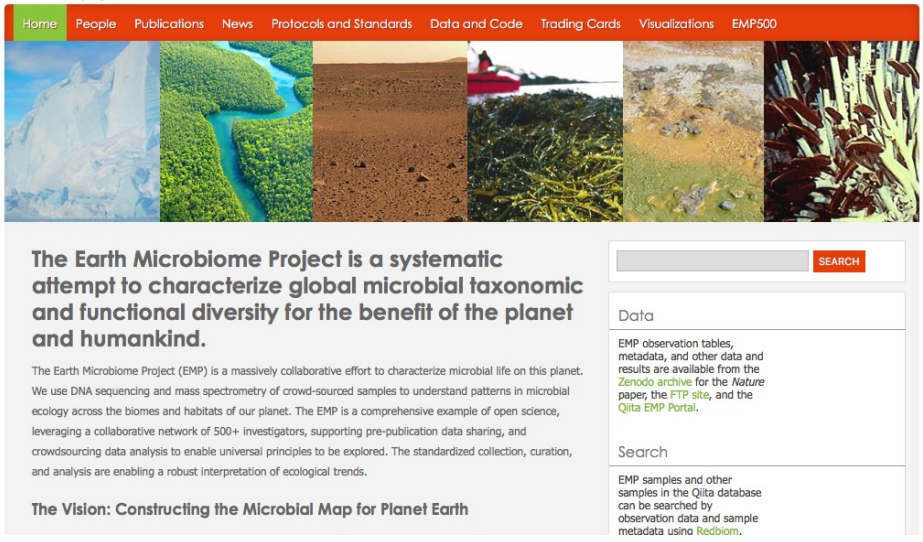





Paoli, L., Ruscheweyh, HJ., Forneris, C.C. et al. Biosynthetic potential of the global ocean microbiome. Nature 607, 111–118 (2022). https://doi.org/10.1038/s41586-022-04862-3

## Earth Microbiome Project

**The Earth Microbiome Project (EMP) is a systematic attempt to characterize global microbial taxonomic and functional diversity for the benefit of the planet and humankind. Most of the data generated to this point are from 16S rRNA amplicon sequencing, but the project also includes data from 18S and ITS amplicon sequencing, metagenomics, and metabolomics. https://www.earthmicrobiome.org**



## Summary

| Method | Targeted | WGS |
|---|---|---|
| Short reads | **Currently standard**<br>Many biaises | **Currently standard**<br>Several pipelines to choose from |
| Long reads (hifi) | Becomes competitive<br>Many advantages<br>Added throughput by Kinnex | Still expensive<br>But much easier to obtain MAGs |
| Hybrid (short+long) | Not used in practice | Rarely combined |



## Summary

**Importance of metagenomes for understanding our environment**

**Two main methods exist: Shotgun vs Targeted**
**Both methods extract different things from the data**
**e.g., Taxonomy assignment or functional assignment**

**Two sequencing techniques compete: short reads vs long reads**

**Many initiatives and databases gathering data**

**More details in the coming days !**

# Thank you for your attention. Questions?