SouthGreen bioinformatics platform

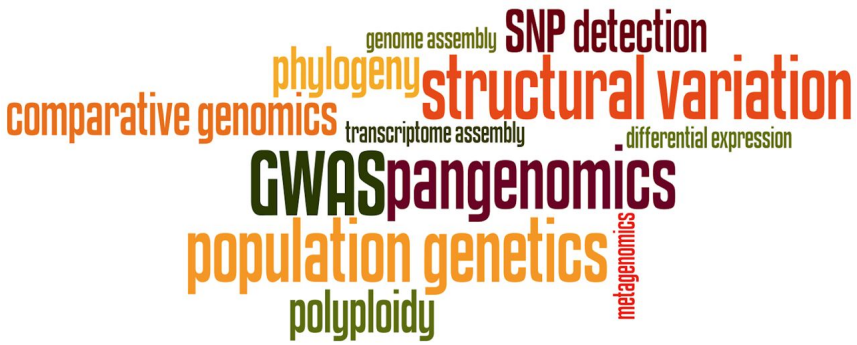
Trainings 2019

agap Bioversity International UMR BGPI Biologie et Génétique des Interactions Plante-Parasite diade IPME

ABiMS ASTRE Animal - Santé Territoires - Risques - Ecosystèmes BOREA Biologie des Organismes et Ecosystèmes Aquatiques intertryp Host-vector-parasite-environment interactions in neglected tropical diseases due to trypanosomatids MiVEGEC Maladies Infectieuses et Vecteurs Ecologie, Génétique, Évolution et Contrôle UMR QualiSud

SouthGreen bioinformatics platform

Bioinformatics platform dedicated to the genetics and genomics of tropical and Mediterranean plants and their pathogens




Rice

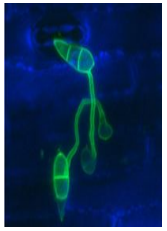

Banana


Palm


Sorghum


Coffee


Cassava


Magnaporthe

www.southgreen.fr

South Green bioinformatics platform

Charte

i-Trop bioinformatics

diade

Larmande Pierre
Sabot François
Tando Ndomassi
Tranchant-Dubreuil Christine

IPME

Comte Aurore
Dereeper Alexis

diade IPME

Orjuela-Bouniol Julie

agap

Bocs Stephanie
De Lamotte Frédéric
Droc Gaetan
Dufayard Jean-François
Hamelin Chantal
Martin Guillaume
Pitollat Bertrand
Ruiz Manuel
Sarah Gautier
Summo Marilyne

Bioversity International

Rouard Mathieu
Guignon Valentin
Catherine Breton

UMR BGPI

Mahé Frédéric
Ravel Sébastien

intertryp

Sempere Guilhem

Bioversity International

cirad

IRD Institut de Recherche pour le Développement FRANCE

INRA SCIENCE & IMPACT

SouthGreen bioinformatics platform

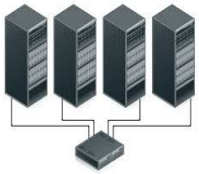
Workflow manager

TOGGLE
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

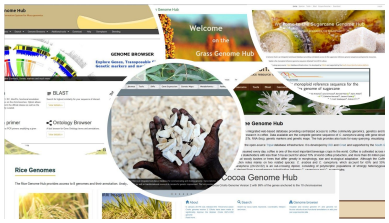
HPC and trainings….

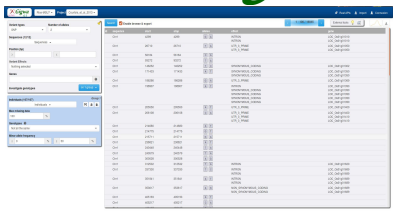


IRD Institut de Recherche pour le Développement

cirad

Genome Hubs & Information System

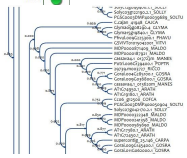




SNPs and Indels

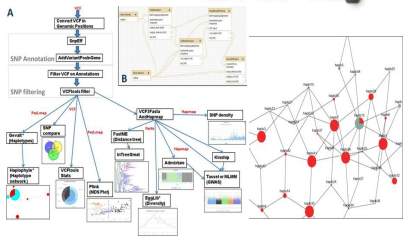
GreenPhyl

Family Id	Family Name	Number of sequences	Status
GP000010	Cytochrome P450 superfamily	6942	
GP000017	AP2/EREBP transcription factor family: ERF/DREB group (partial)	5142	
GP000020	NAC transcription factor family	4574	
GP000028	MADS transcription factor family		
GP000018	Haem peroxidase superfamily		
GP000066	General substrate transporter superfamily		
GP000022	Subtilisin-like Serine Proteases family		
GP000019	NPF, NRT1/PTR FAMILY		



Gene families

SNiPlay



https://github.com/SouthGreenPlatform

@green_bioinfo
@itropBioinfo

The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics, Current Plant Biology, 2016

South Green bioinformatics platform

ABiMS — Erwan Corre

IPME — Marie Simonin, Sébastien Cunnac

ASTRE Animal · Santé Territoires · Risques · Écosystèmes — Etienne Loire, Julie Reveillaud

UMR QualiSud — Florentin Constancias

diade — Valentin Klein

MiVEGEC Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle — Valérie Noël

i-Trop bioinformatics

And more collaborators !

SouthGreen bioinformatics platform

18-19/03	Guide de survie à Linux - IRD
21/03	Initiation à l'utilisation du cluster CIRAD – CIRAD
22/03	Initiation à l'utilisation du cluster itrop - IRD
15-16/04	Initiation au gestionnaires de workflow SG & Gigwa – IRD
18-19/04	Guide du Jedi en Linux & bash - CIRAD
13–16/05	Python - IRD
17/05	Initiation aux analyses de données transcriptomiques – IRD
21/05	Utilisation avancée du cluster IRD – IRD
23-24/05	Initiation aux analyses de données métagénomiques – IRD
6/06	Manipulation de données et figures sous R – CIRAD
25-27/09	Assemblage et annotation de transcriptomes - IRD

Session de formation 2019

- Toutes nos formations :
  https://southgreenplatform.github.io/trainings/
- Topo & TP : **Metagenomics**
- Environnement de travail : **Logiciels à installer**

Initiation aux analyses de données metabarcoding

www.southgreen.fr

https://southgreenplatform.github.io/trainings

Planning

1. **Introduction générale**

2. **Partie pratique**

Practice 1: Obtaining an OTU table with FROGS in Galaxy

Practice 2: Obtaining an OTU table with FROGS in Command Line

Practice 3: Handling and visualisation of OTU table using PhyloSeq R package

3. **Conclusions**

What metagenomics is ?

Metagenomics (Environmental Genomics or Community Genomics) is the study of genomes recovered from environmental samples without the need for culturing them

Metagenomics processes data using bioinformatics tools

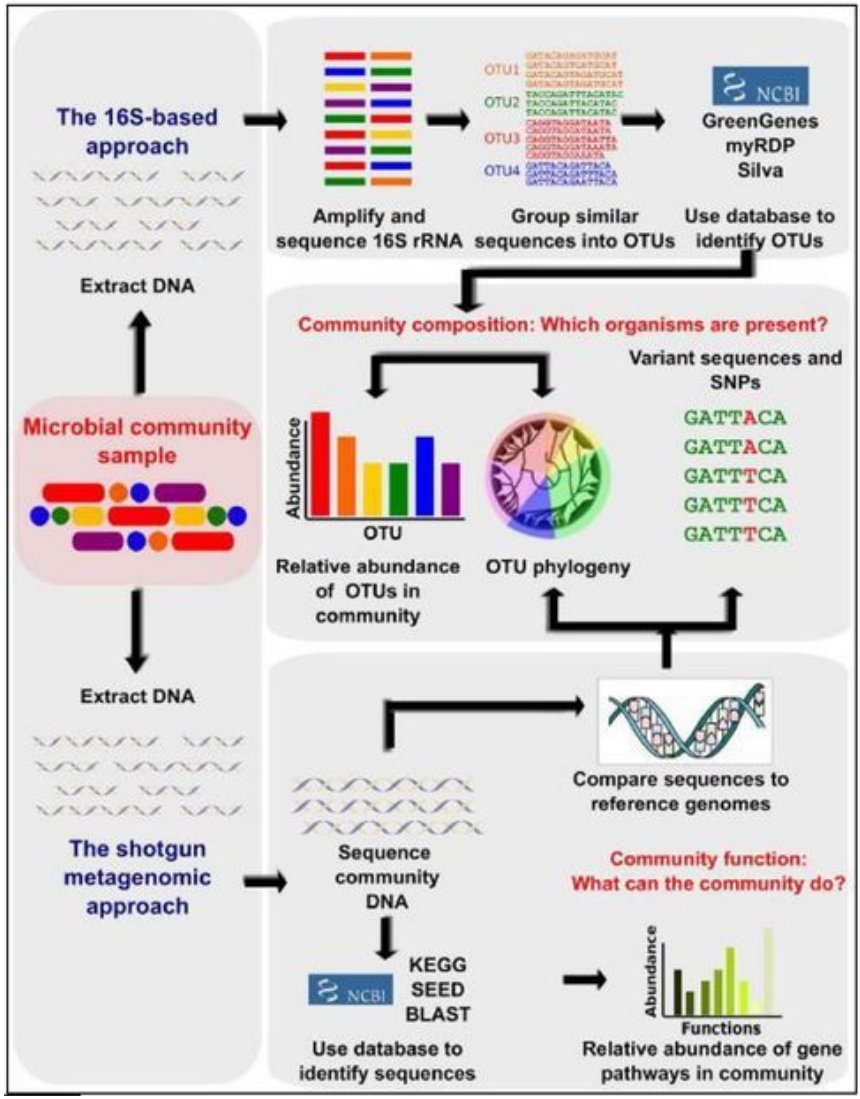
=> Organisms can be studied directly in their environments bypassing the need to isolate each species

=> There are significant advantages for viral metagenomics, because of difficulties cultivating the appropriate host

Two main strategies in metagenomics

We can distinguish targeted metagenomics or shot-gun metagenomics :

- 16S rRNA metabarcoding is used to characterize the bacterial communities of an environment
- Whole-genome sequencing when the goal is to identify gene functions and pathways, or reconstruct microbial genomes.



Morgan XC, Huttenhower (2012) Plos Comput Biol

Markers genes vs Shotgun metagenomics

Marker Gene Profiling	Shotgun Metagenomics Profiling
Less expensive (~$100 per sample)	Still very expensive (~$1000 per sample)
Computational needs can be met by desktop / small server computers	Usually requires huge computational resources (cluster of computers)
Provides mainly taxonomic profiling	Provides both taxonomic and functional profiling
For 16S, majority of genes can be assigned at least to phylum level	Many more unassigned gene fragments (“wasted” data)
Relatively free of host DNA contamination	Prone to host DNA contamination

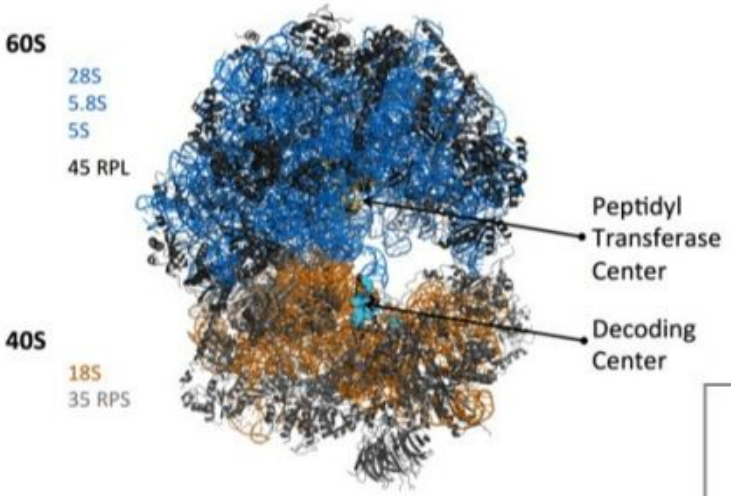
Strategies in Diversity Characterisation

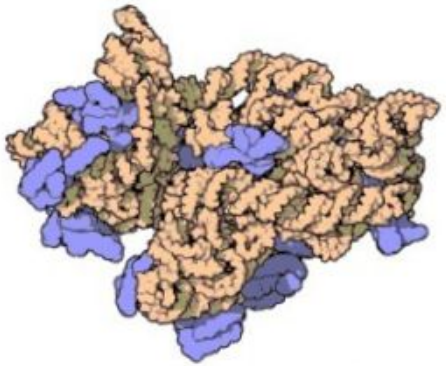
Technique	Advantages and challenges	Main applications
Metataxonomics using amplicon sequencing of the 16S or 18S rRNA gene or ITS	+ Fast and cost-effective identification of a wide variety of bacteria and eukaryotes – Does not capture gene content other than the targeted genes – Amplification bias – Viruses cannot be captured	* Profiling of what is present * Microbial ecology * rRNA-based phylogeny
Metagenomics using random shotgun sequencing of DNA or RNA	+ No amplification bias + Detects bacteria, archaea, viruses and eukaryotes + Enables *de novo* assembly of genomes – Requires high read count – Many reads may be from host – Requires reference genomes for classification	* Profiling of what is present across all domains * Functional genome analyses * Phylogeny * Detection of pathogens
Meta-transcriptomics using sequencing of mRNA	+ Identifies active genes and pathways – mRNA is unstable – Multiple purification and amplification steps can lead to more noise	* Transcriptional profiling of what is active

Breitwieser et al. 2016

Metabarcoding strategie

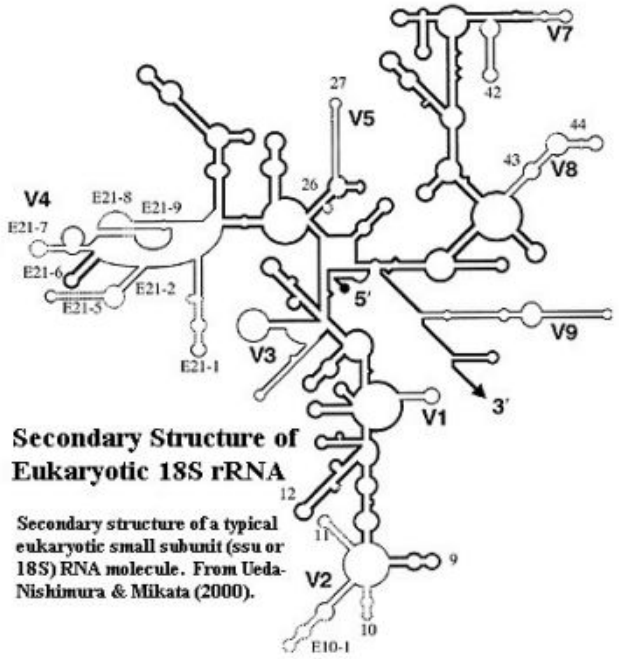
A universal gene: ribosomal RNA





Small Sub-Unit (SSU)

16S	Bacteria Archaea Mitochondria Chloroplasts
18S	Eukaryota



Secondary Structure of Eukaryotic 18S rRNA

Secondary structure of a typical eukaryotic small subunit (ssu or 18S) RNA molecule. From Ueda-Nishimura & Mikata (2000).

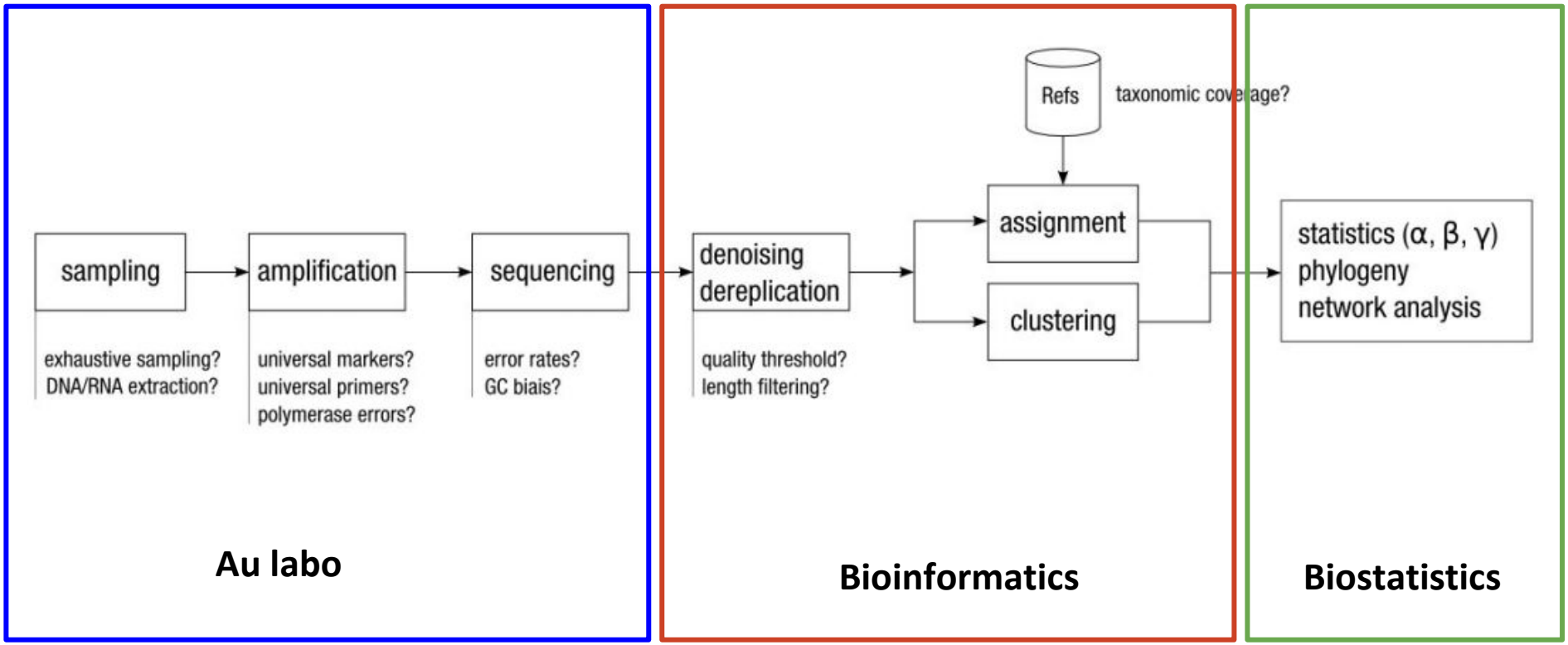
other markers can be used (e.g., ITS). Requirements are: conserved distal regions for primers, variable internal regions, and available sets of reference sequences.

Projets métagénomiques

4900 projets sur NCBI (avril 2018)

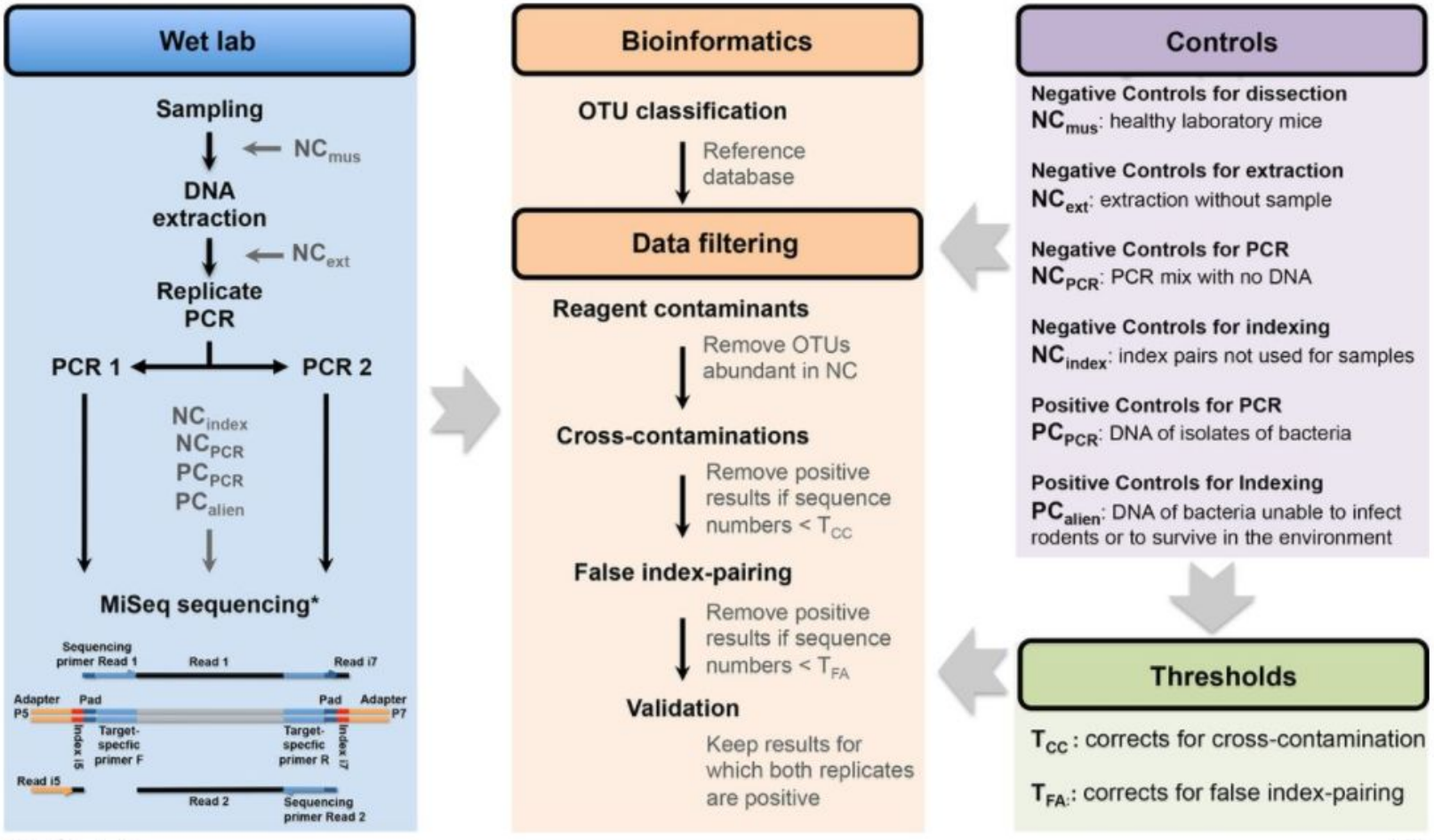
- Sable de plage
- Moustique
- Corail
- Glace
- Air de la ville de Singapour
- Surface de la cuvette des toilettes
- Fromages
- ...

Amplicon-based studies general pipeline



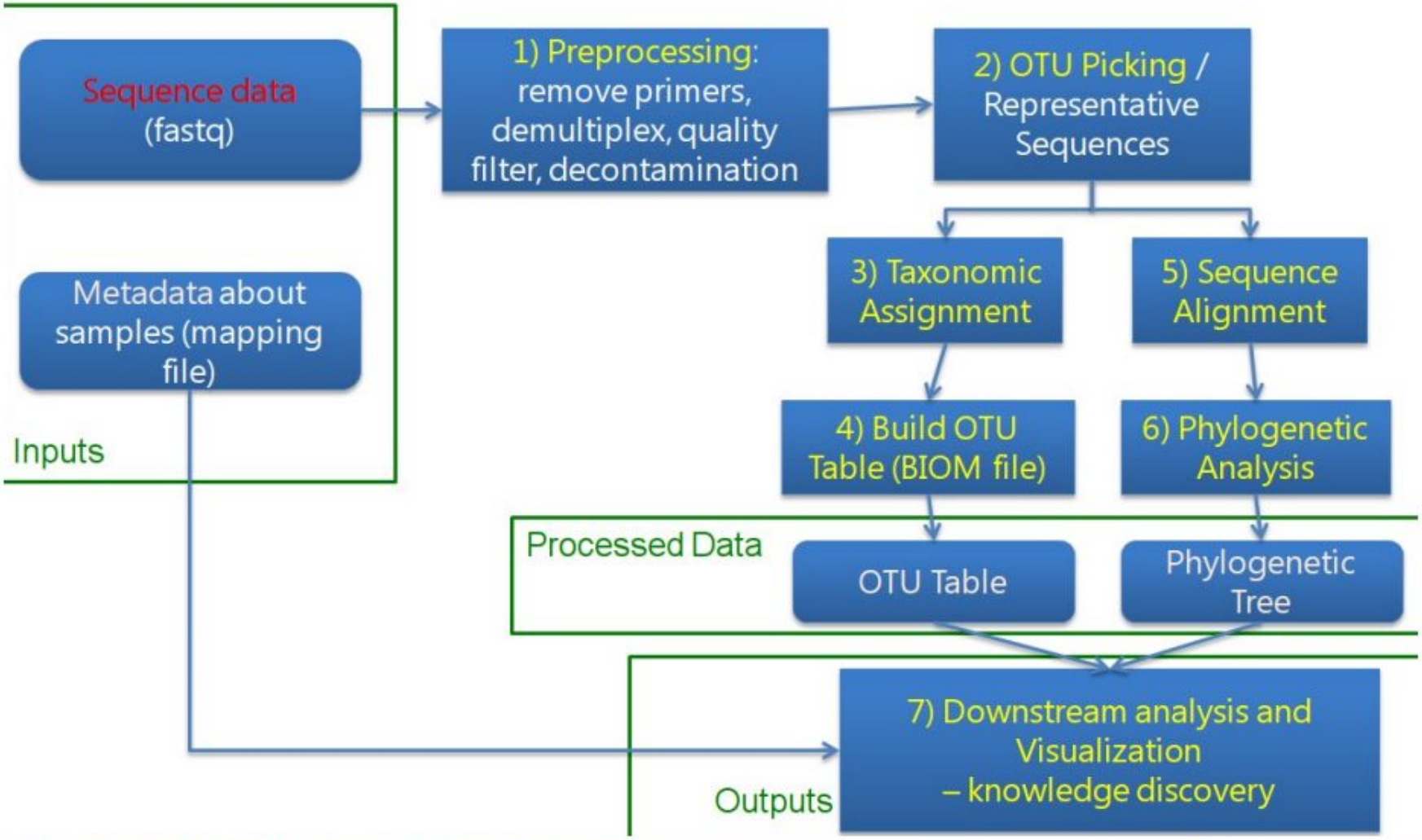
TODAY !

Workflow of the wet laboratory, bioinformatics, and data filtering procedures in the process of data filtering for 16S rRNA amplicon sequencing.



DOI: 10.1128/mSystems.00032-16

Overall bioinformatics workflow



Major metagenomics pipelines

targeted amplification

Mothur (2009)
Patrick Schloss
open-source
single piece
most cited
stats

Qiime

Qiime (2010)
Gregory Caporaso
open-source
python wrapper
most used(?)
stats

Uparse

Uparse (2013)
Robert Edgar
closed-source
usearch commands
popular
no stats

Emerging and new methods

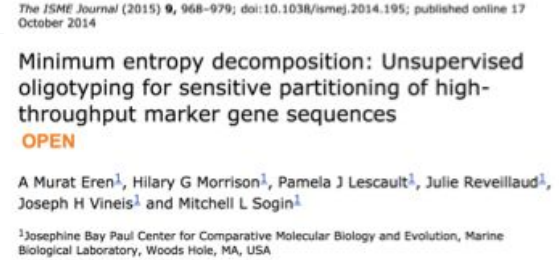
Clustering

–Oligotyping and Minimum Entropy Decomposition

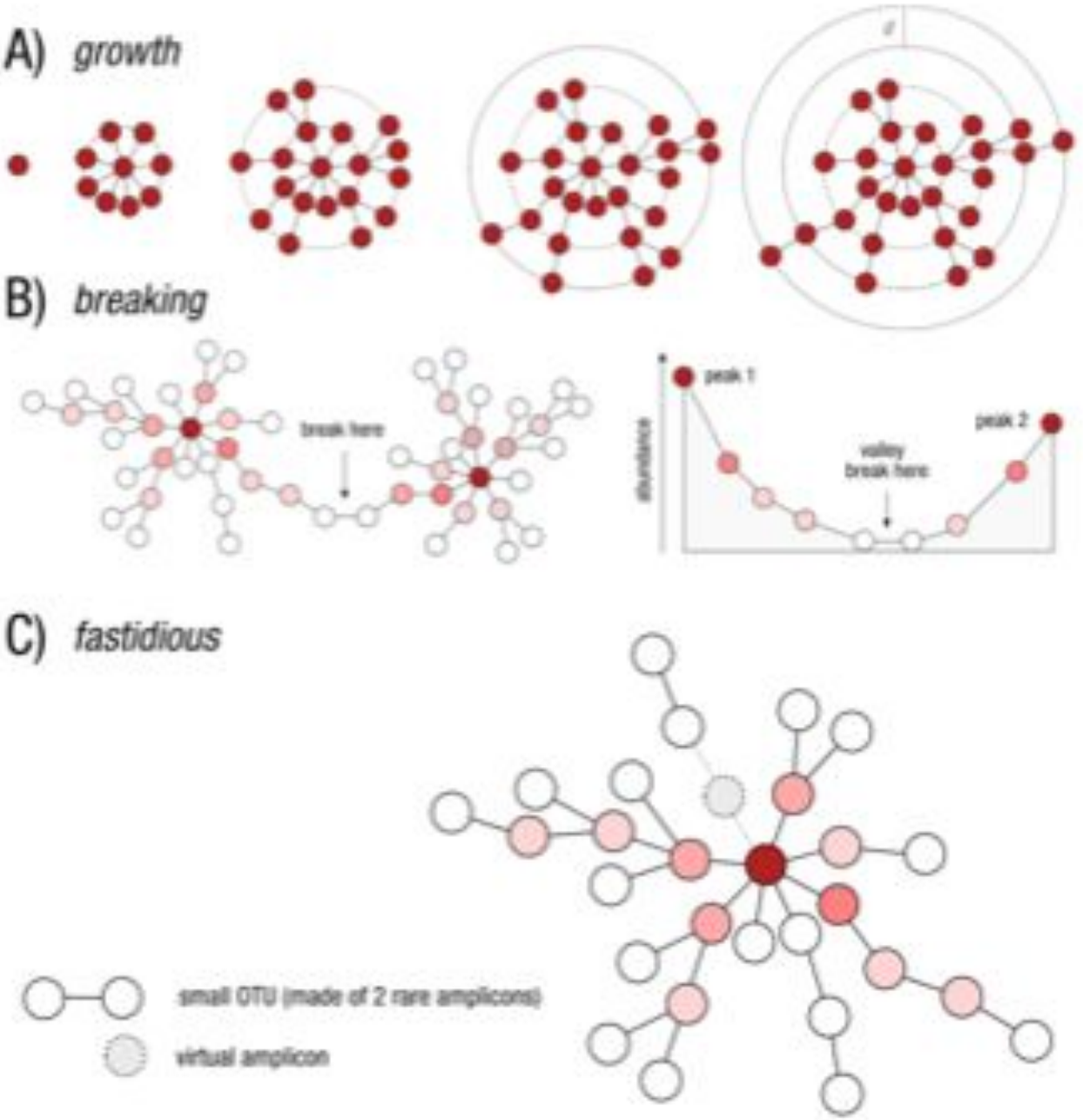
–SWARM-V2: exploration of OTU natural boundaries around most abundant sequences

Denoising

–DADA2 (probabilistic approach for sequencing error detection and correction)



SWARM2

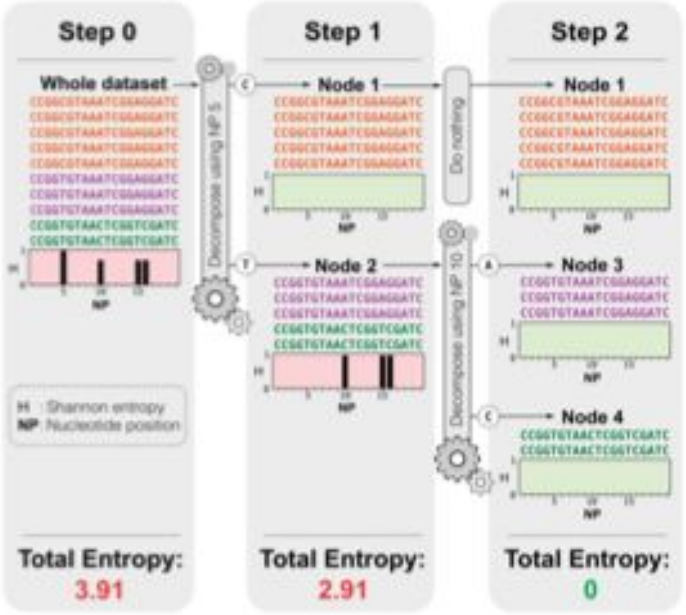

A) growth
B) breaking
break here
abundance
peak 1
peak 2
valley
break here
C) fastidious
small OTU (made of 2 rare amplicons)
virtual amplicon

Minumum Entropy Decomposition

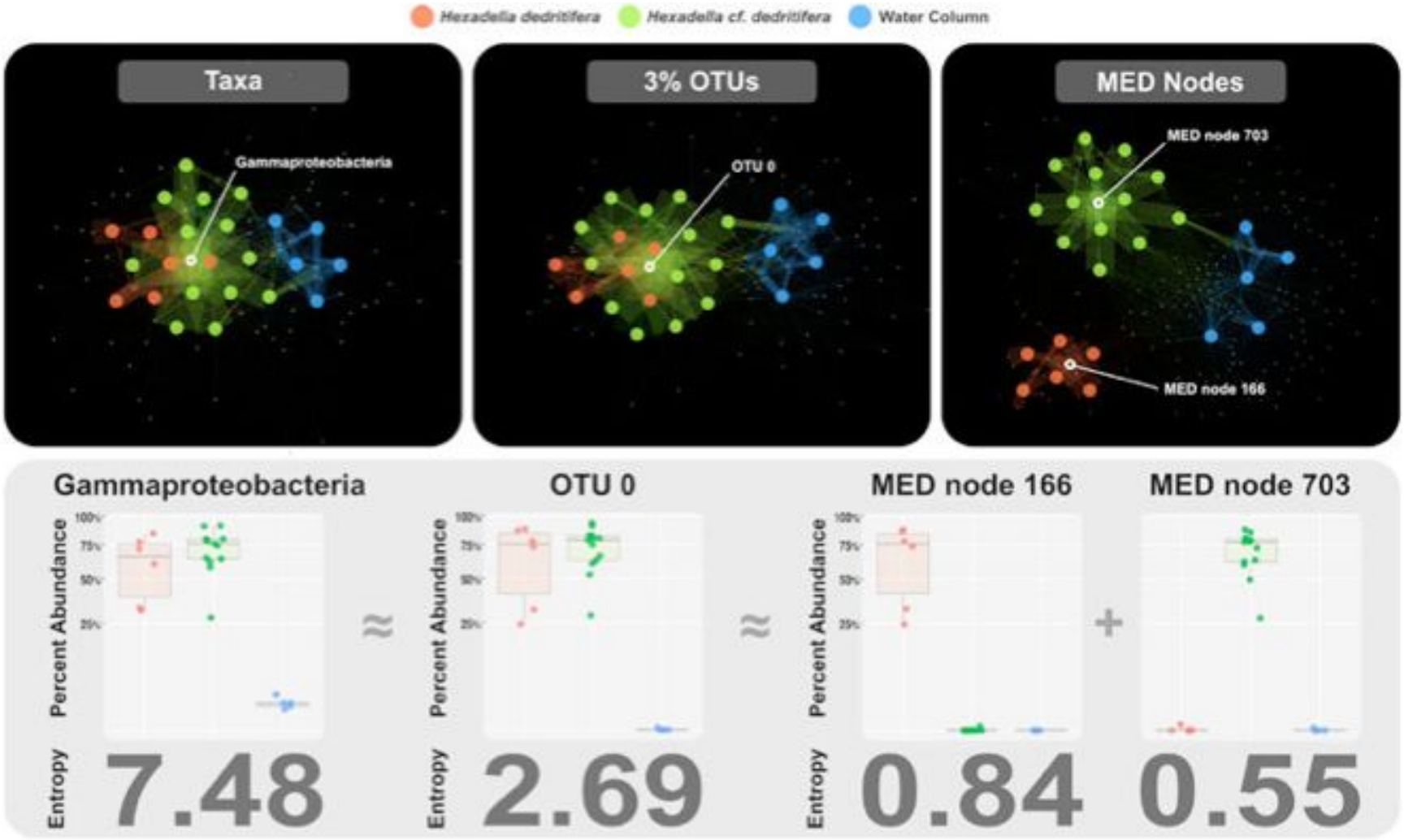
–The entire dataset is split according of the position of highest entropy => 4 nodes (A, T, C or G)

–Each of these nodes is in turn split according to its highest entropy position and so on.

–Decomposition stops when all nodes reach minimum entropies.



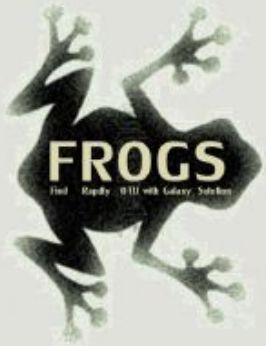
Minimum Entropy Decomposition



MED node **166** vs. MED node **703**:
99.2% sequence identity

Which bioinformatics solutions?
Today Clustering Example: FROGS

	Disadvantages
QIIME	Installation problem Command lines
UPARSE	Global clustering command lines
MOTHUR	Not MiSeq data without normalization Global hierarchical clustering Command lines
MG-RAST	No modularity No transparence



QIIME allows analysis of high-throughput community sequencing data
J Gregory Caporaso et al, Nature Methods, 2010; doi:10.1038/nmeth.f.303
Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.
Schloss, P.D., et al., Appl Environ Microbiol, 2009, doi: 10.1128/AEM.01541-09

UPARSE: Highly accurate OTU sequences from microbial amplicon reads
Edgar, R.C. et al, *Nature Methods*, 2013, dx.doi.org/10.1038/nmeth.2604
The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes
F Meyer et al, BMC Bioinformatics, 2008, doi:10.1186/1471-2105-9-386

Which bioinformatics solutions?
Today Denoising example:

Fast and accurate sample inference from amplicon data with single-nucleotide resolution

FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié Lucas Auer Maria Bernard Mahendra Mariadassou Laurent Cauquil Katia Vidal Sarah Maman Guillermina Hernandez-Raquet Sylvie Combes Géraldine Pascal

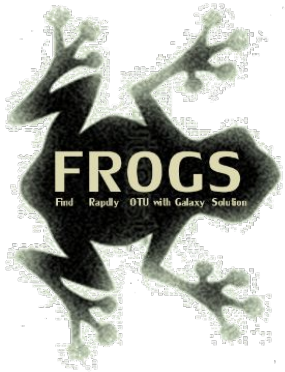
Bioinformatics, Volume 34, Issue 8, 15 April 2018, Pages 1287–1294,https://doi.org/10.1093/bioinformatics/btx791

https://github.com/geraldinepascal/FROGS.git

Practice 1:

Obtaining an OTU table with FROGS in Galaxy

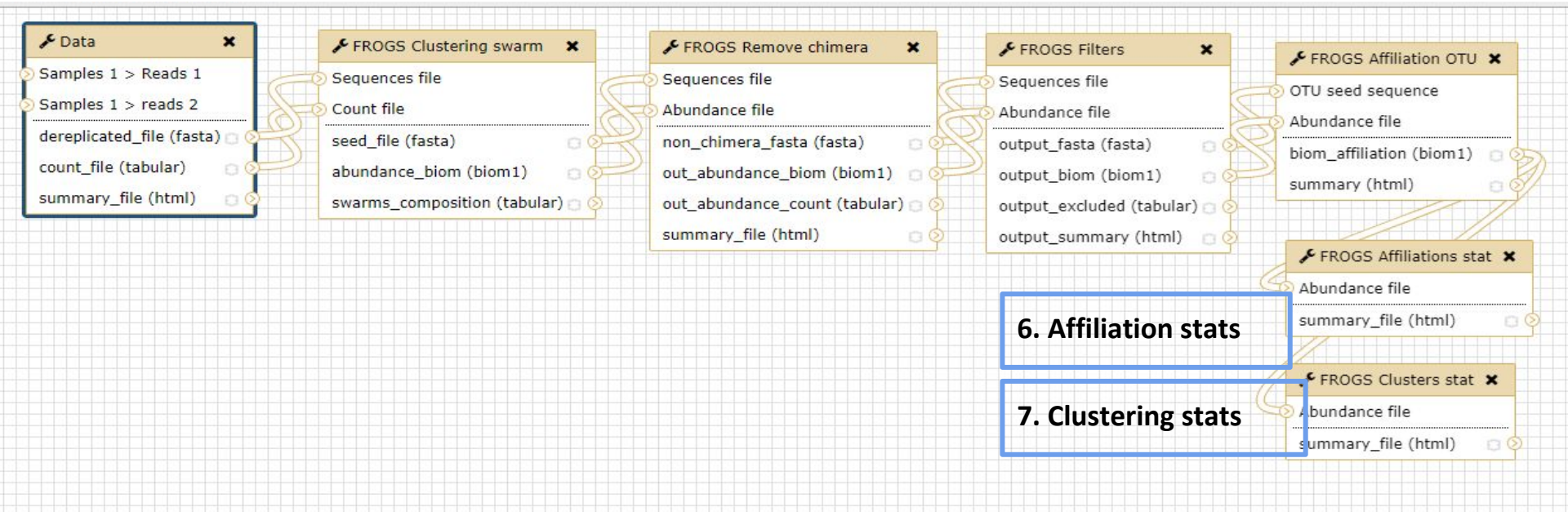
FROGS: Find, Rapidly, OTUs with Galaxy Solution



- Use platform Galaxy
- Set of modules= Tools to analyze your “big” data
- Independent modules
- Run on Illumina/454 data 16S, 18S, and 23S
- New clustering method
- Many graphics for interpretation
- User friendly, hiding bioinformatics infrastructure/complexity

FROGS Pipeline on Galaxy

1.Pre-process | 2. Clustering | 3. Remove Chimera | 4. Filtering | 5. Affiliation OTU



Pre-process

A preprocessing tool :

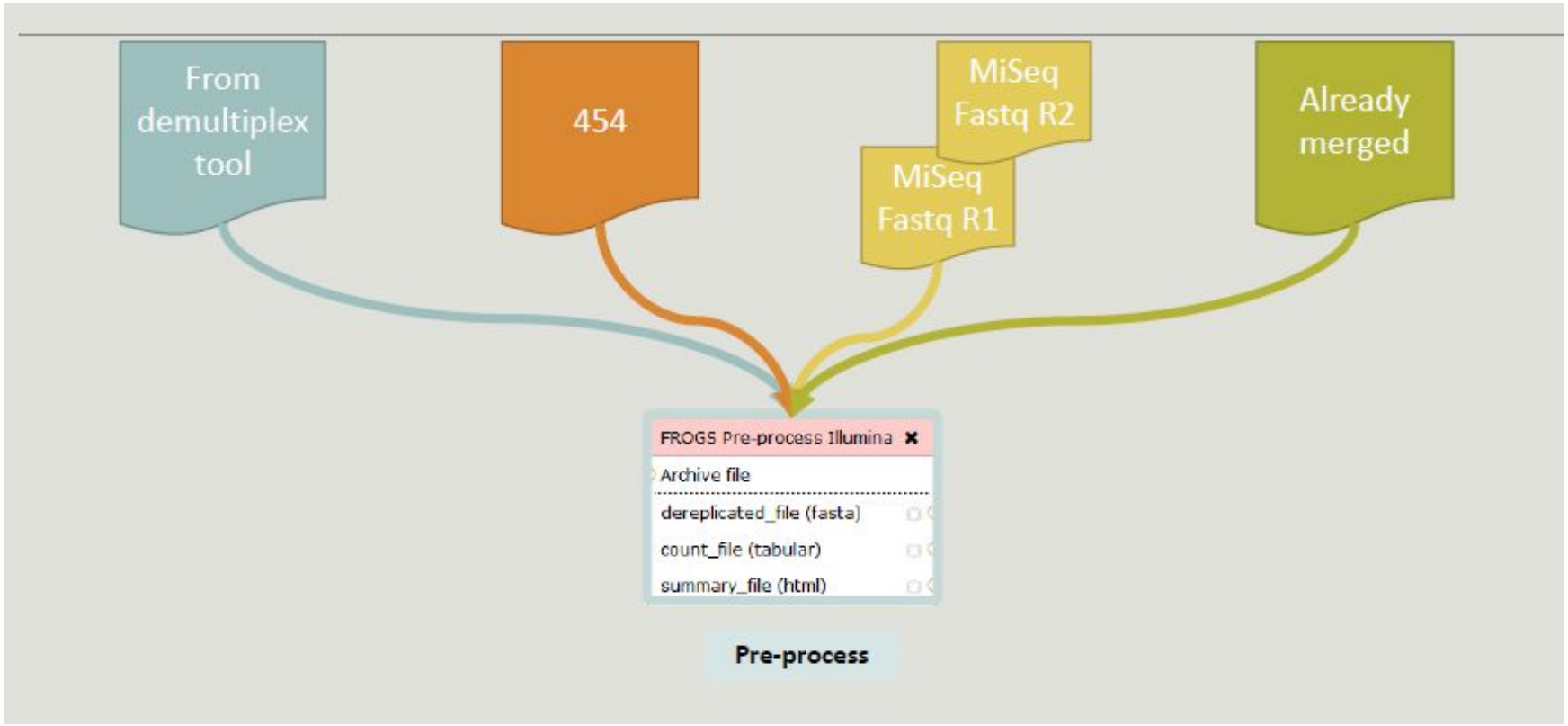
- merges paired sequences into contigs with flash or vsearch
- cleans the data with cutadapt,
- deletes the chimeras with VSEARCH combined with a cross-validation method and
- dereplicates sequences with a home-made python script.

Pre-process

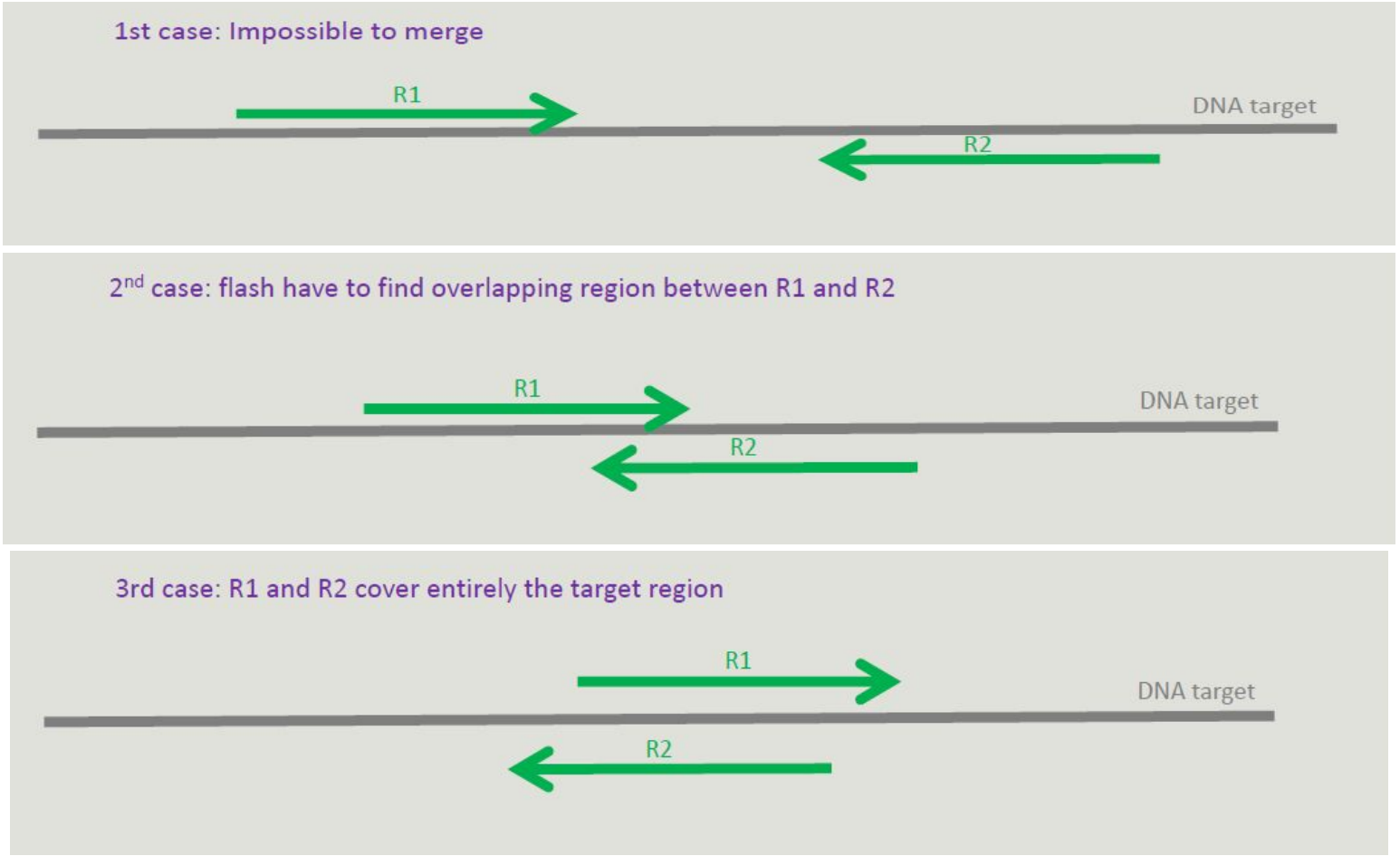
FROGs takes the reads (R1 and R2) from multiple samples and performs the following steps:

- If the data is not in contigs, R1 et R2 will be overlapped
- Contigs that are too big or too small will be filtered out.
- Sequences that are too small or of poor quality will be filtered out.
- Sequences will be de-replicated: duplicates will be removed but the number of duplicates will be recorded.

FROGS was designed to support multiplexed and demultiplexed sequences (Run FROGS Demultiplexing before Pre-process)

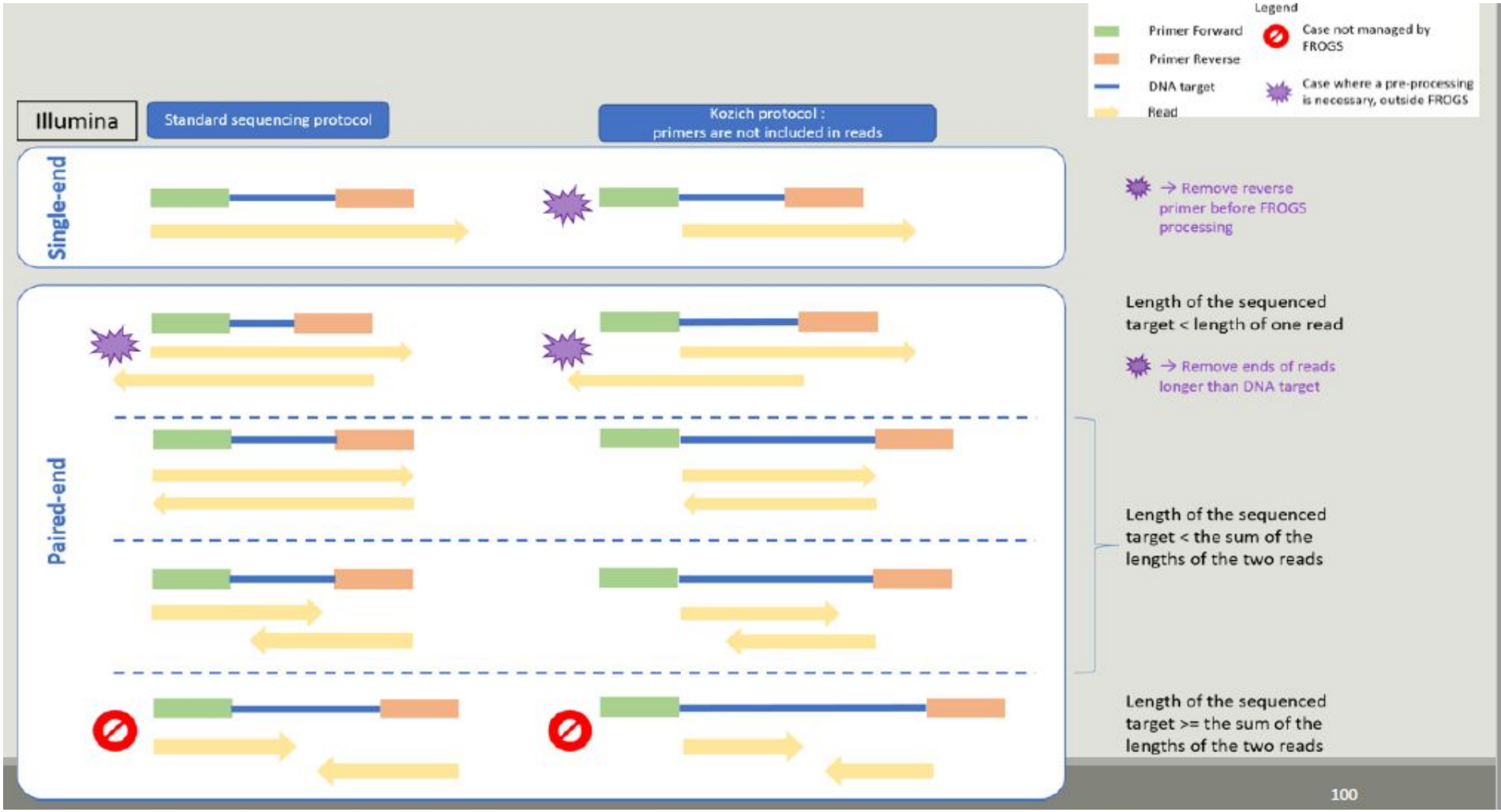


The goal of Flash (**Fast Length Adjustment of Short reads**) is to merge R1 and R2



https://ccb.jhu.edu/software/FLASH/
INRA Genotouls communication

Standard vs Kozich protocol



Preprocess tool in bref

	Take in charge
Illumina	✓
454	✓
Merged data	✓
Not merged data	✓
Without primers	✓
Only R1 or only R2	⊘
Too distant R1 and R2 to be merged	soon
On-overlapping R1 R2	⊘

	Take in charge
Archive .tar.gz	✓
Fastq	✓
Fasta	⊘
With only 1 primer	⊘
Multiplexed data	⊘
Demultiplexed data	✓

Practice

1.1

Aller sur le Practice 1 du github et lancez Preprocess

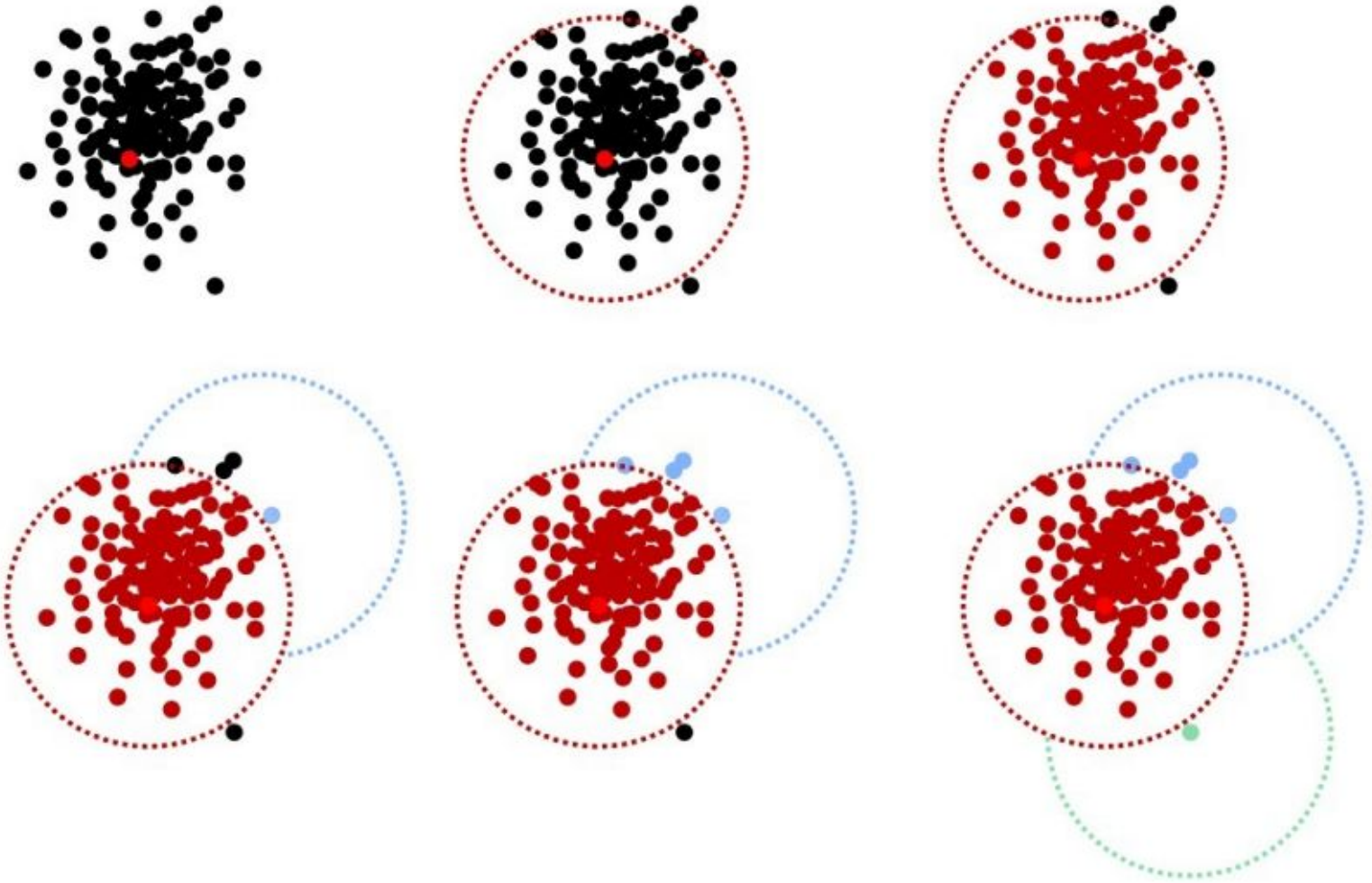
Clustering Swarm

In this step, sequences are clustered into groups using Swarm. This takes the pre-processed fasta and counts files and does the following:

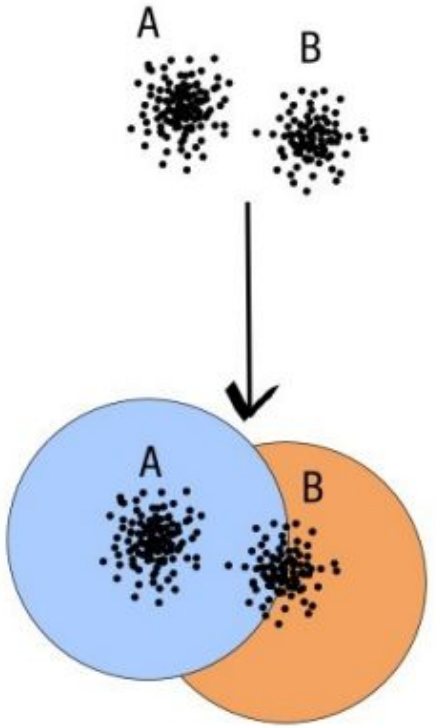
- Sorts reads by abundance.
- Clusters the reads into pre-clusters using Swarm and distance parameter of 1.
- Sorts these pre-clusters by abundance.
- Cluster the pre-clusters using Swarm and a user-specified distance.

https://github.com/torognes/swarm.git

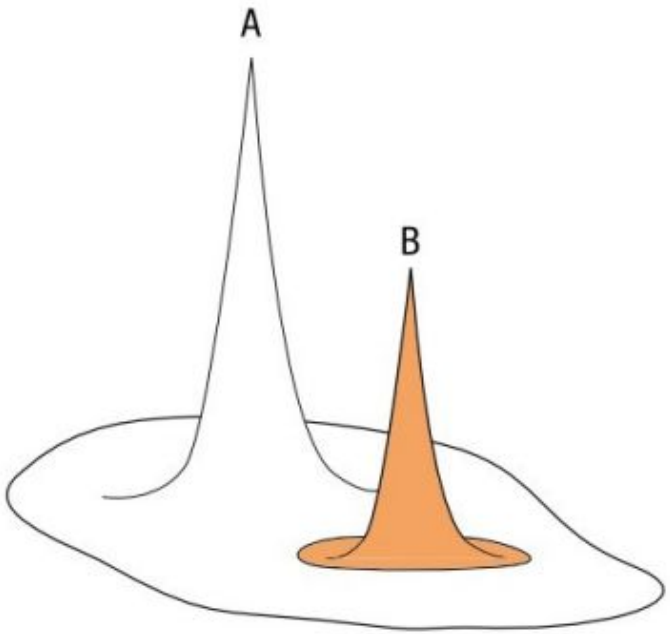
How traditional clustering works?



Swarm: fast, exact and high-resolution clustering



clustering threshold (often 97%) is most of the time unadapted and can mask diversity.



swarm uses abundance values and a new clustering strategy to delineate natural high-quality OTUs.

Swarm uses local clustering threshold, not a global clustering threshold

Frederic Mahé communication

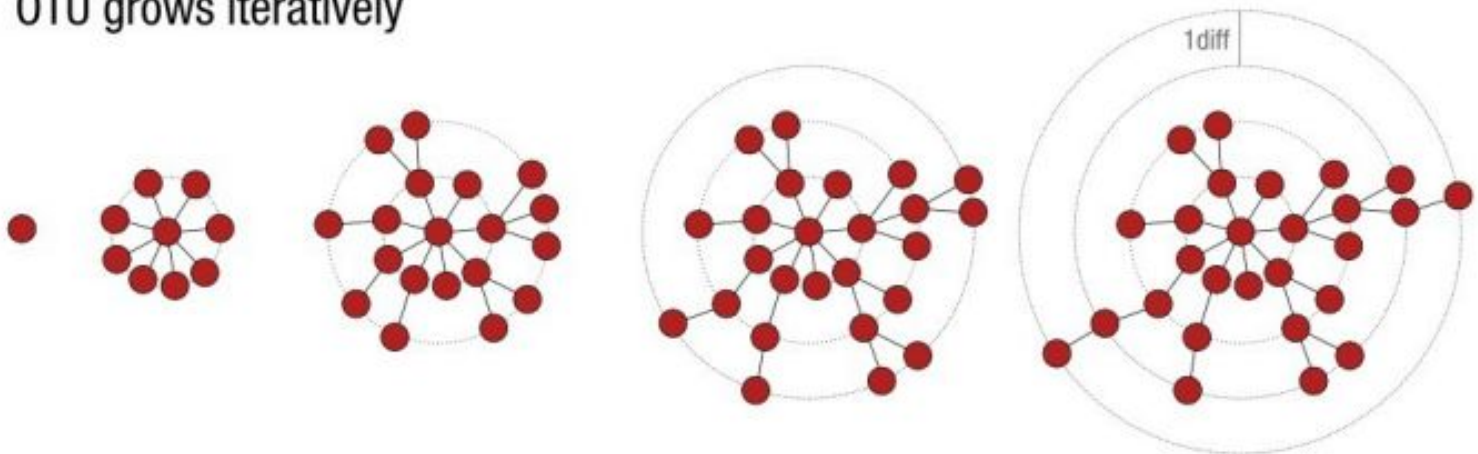
Swarm clustering method

growth phase

ACGT	ACGT	ACGT	
AGGT	A - GT	A - - T	
differences	1	1	2

Avoid & speed-up comparisons
- composition-based prefiltering
- memoization
- fast Needleman-Wunsch

OTU grows iteratively



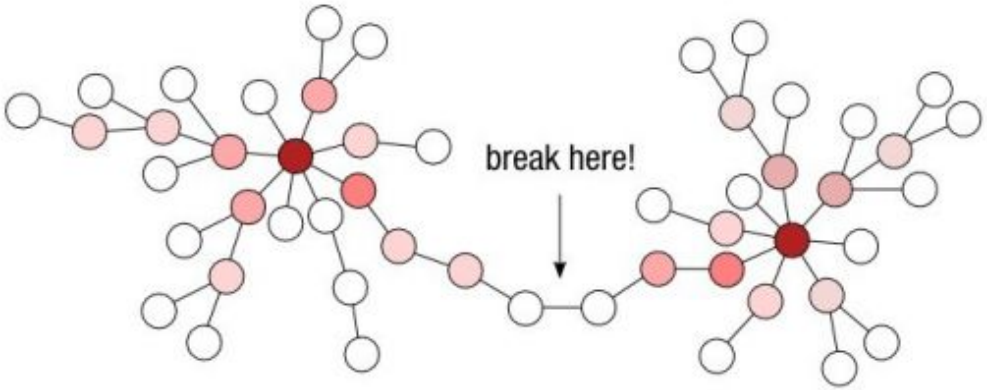
initial seed (randomly picked from amplicon dataset)

no more closely related amplicons, the process stops

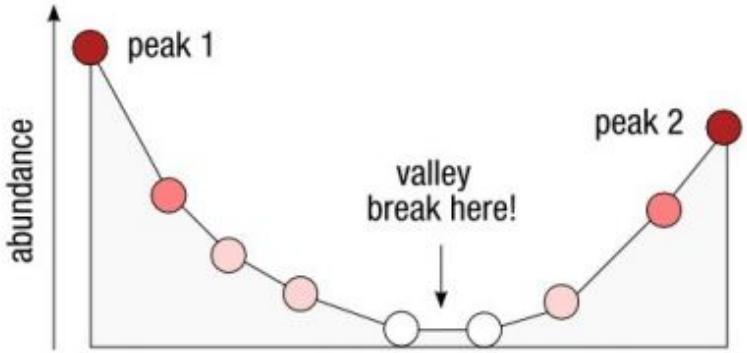
Frederic Mahé communication

Swarm clustering method

breaking phase



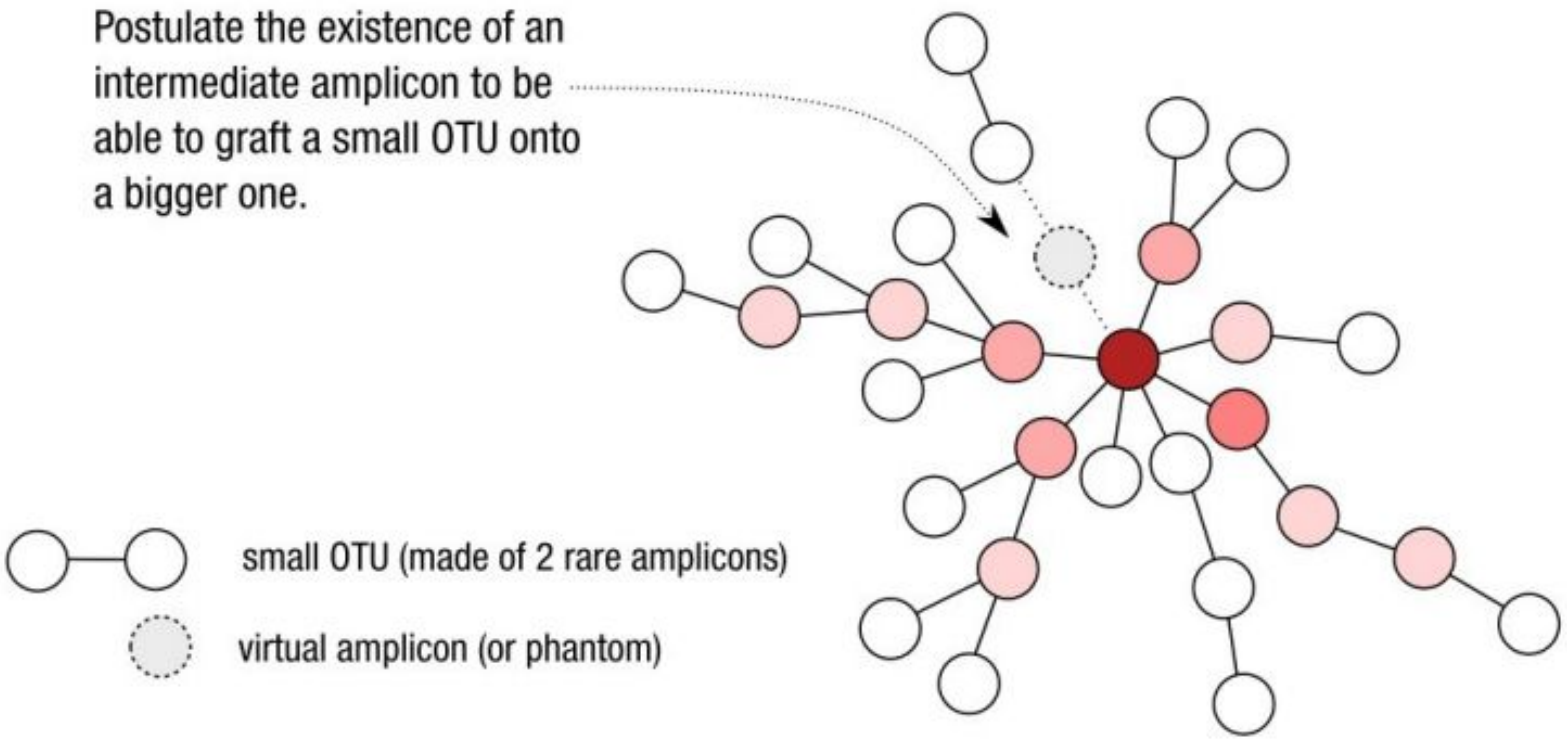
Take into account the abundance of amplicons to produce higher-resolution clusters.



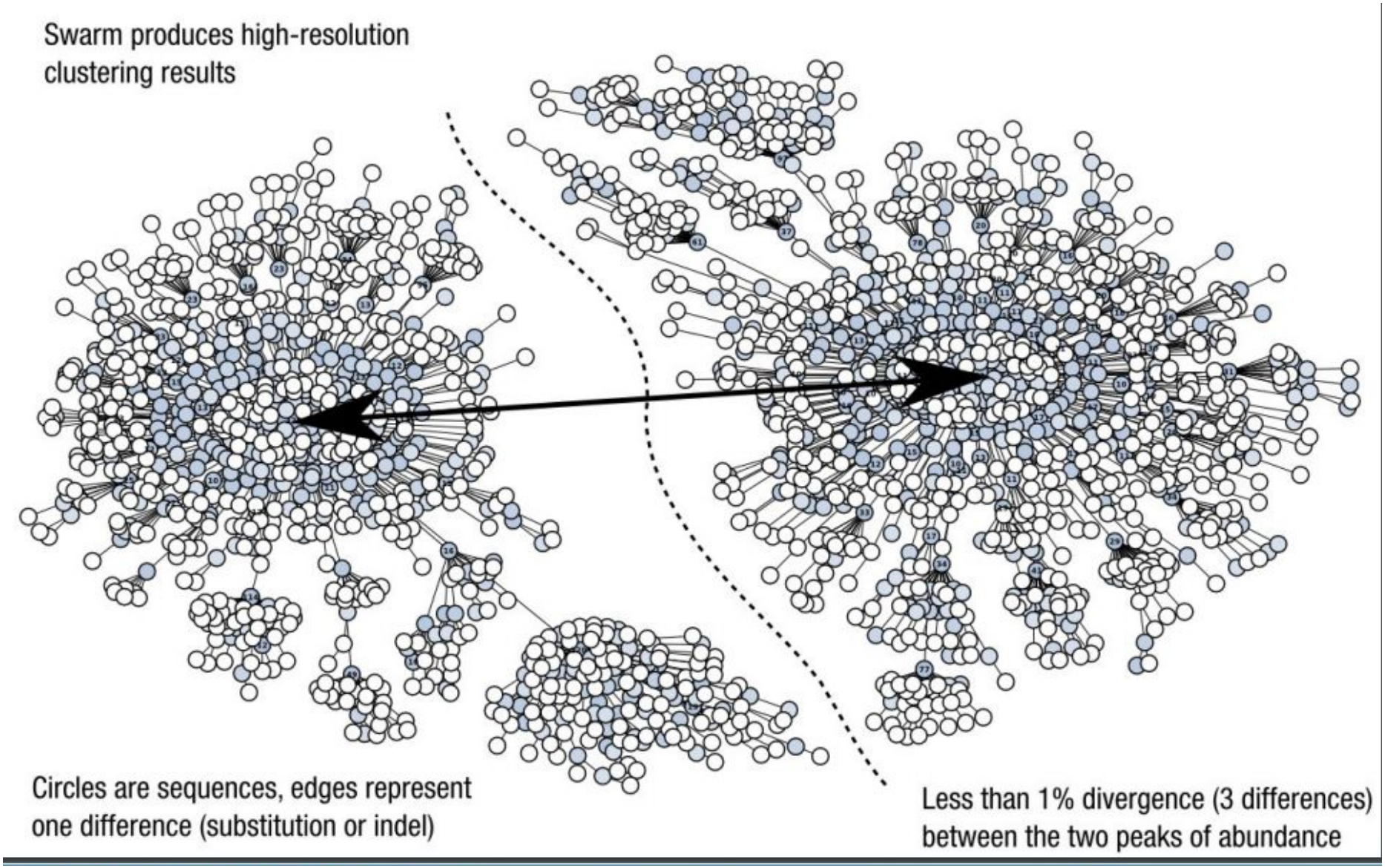
Assuming that original sequences are more abundant than erroneous copies.

Swarm clustering method

grafting phase



Frederic Mahé communication



Frederic Mahé communication

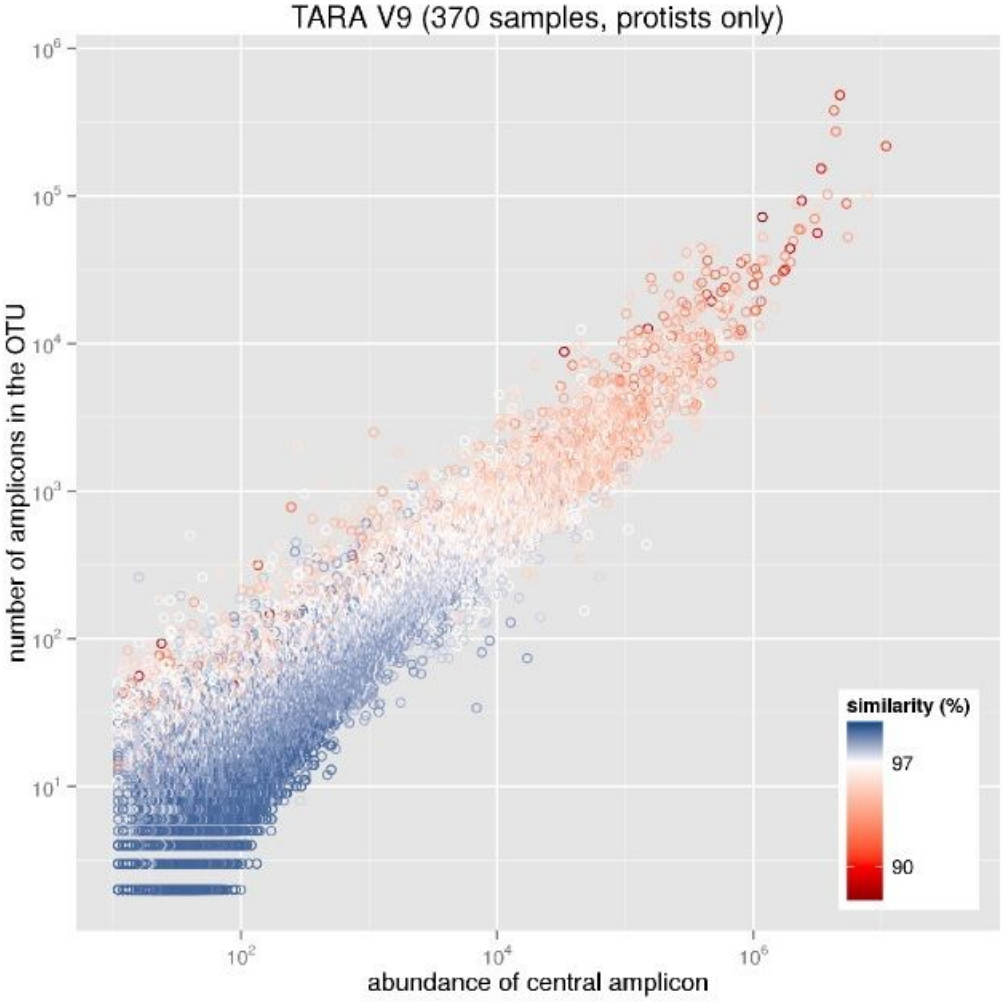
Swarm 2.0 is a highly scalable denoising-clustering method

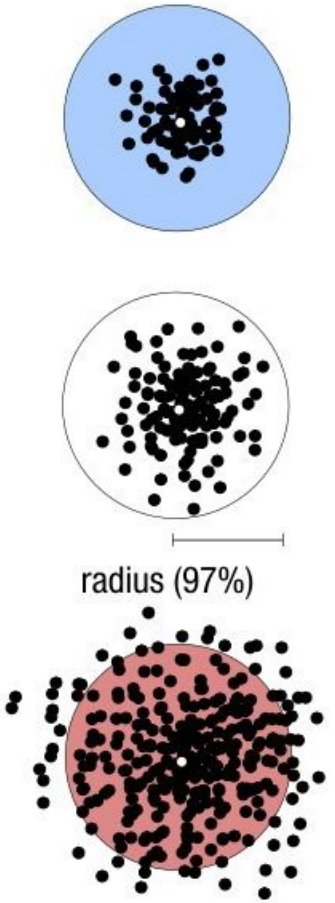
28,275 samples
2.3 billion reads

swarm: 5 hours
usearch: >150 days

Frederic Mahé communication

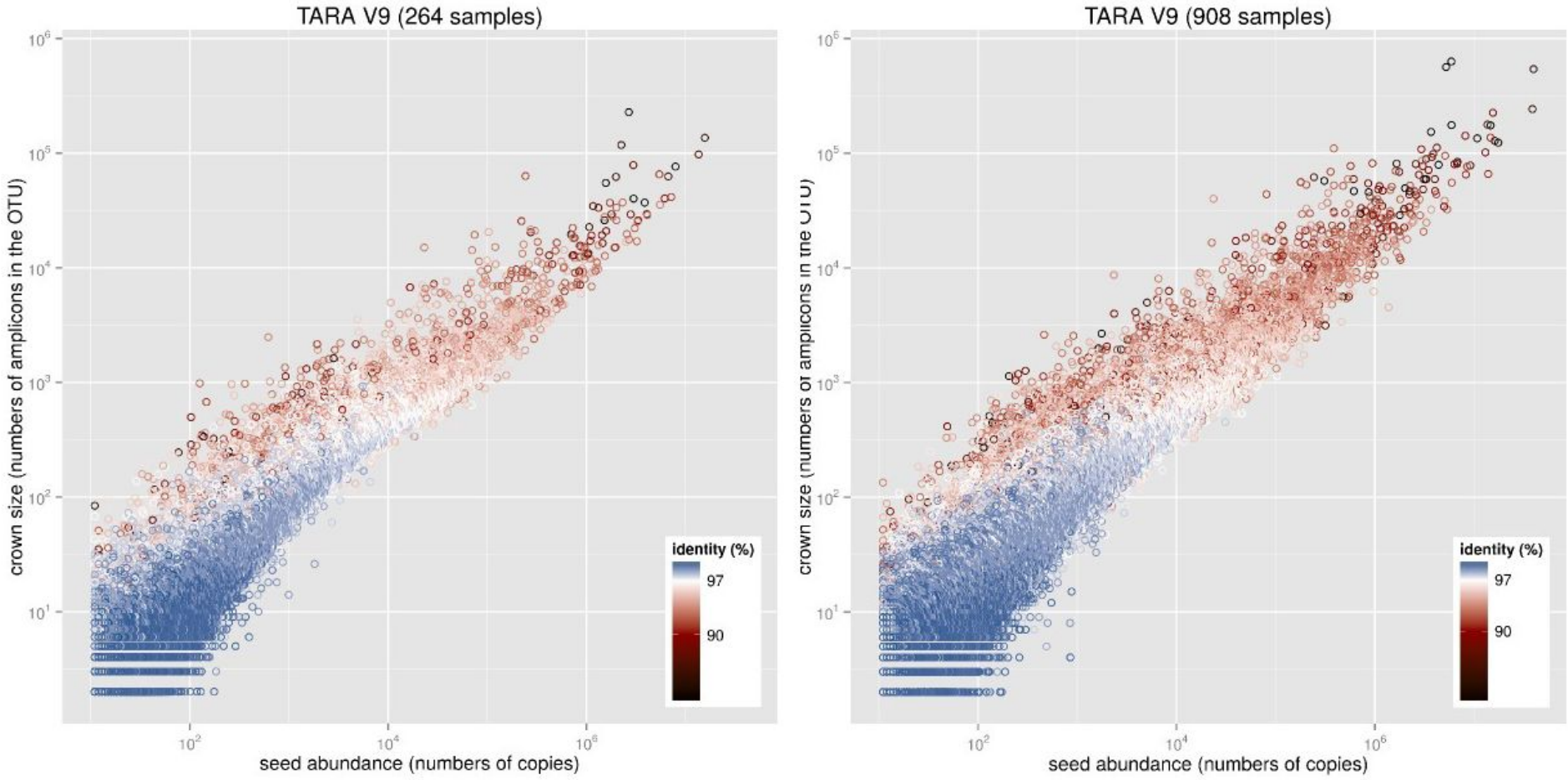
What if we'd used a fix 97%-clustering threshold?





Frederic Mahé communication

Seed abundance vs cloud vs cluster radius shows 97%-threshold inadequacy



clusters produced with swarm using d = 1

Frederic Mahé communication

Practice

1.2
1.3

Aller sur la partie 1.2 Clustering *and* 1.3 stats clustering *du github*

Remove chimera

PCR-generated chimeras are typically created when an aborted amplicon acts as a primer for a heterologous template. Subsequent chimeras are about the same length as the non-chimeric amplicon and contain the forward and reverse primer sequence at each end of the amplicon.

Chimera: from 5 to 45% of reads (Schloss 2011)

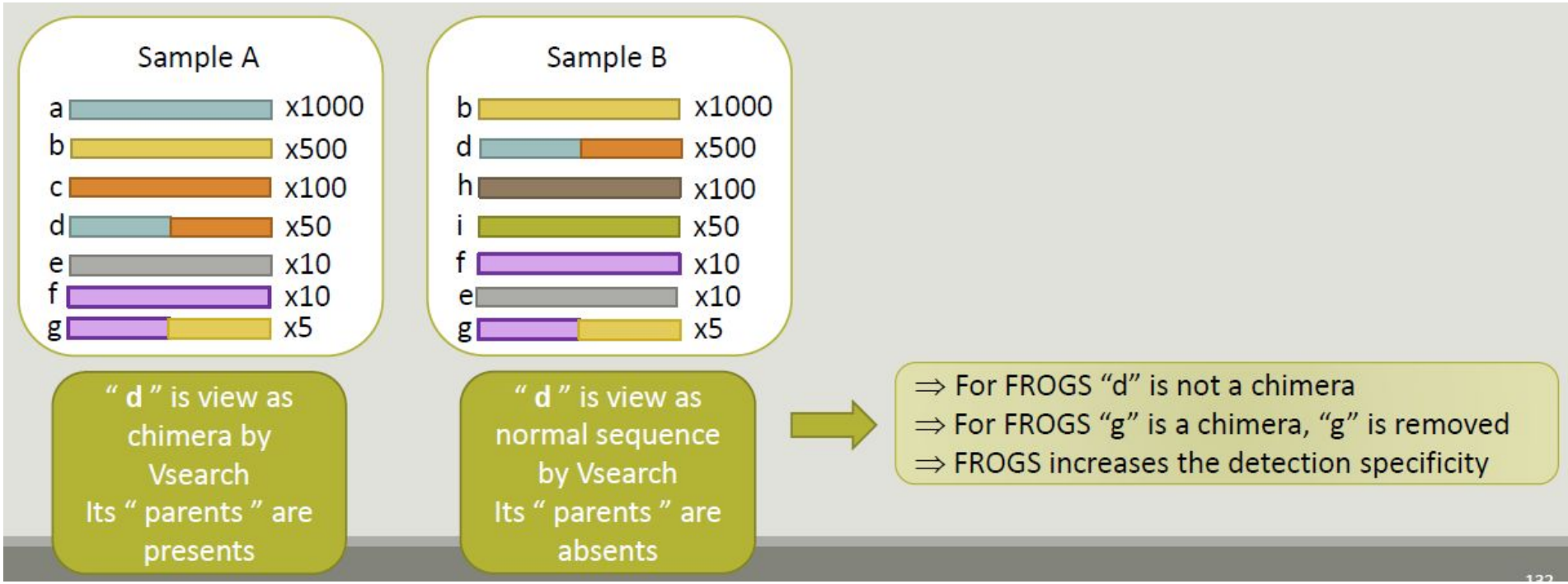
Remove chimera

Closely-related sequences may form chimeras (mixed sequences) during PCR (libray prep). This step removes these sequences by the following method:

- Splits input data into samples
- Uses **vsearch** to find chimeras in each sample
- Removes chimeras

Remove chimera

Chimera removal tool uses VSEARCH combined with an innovative chimera cross-validation.



vsearch: open-source alternative for usearch

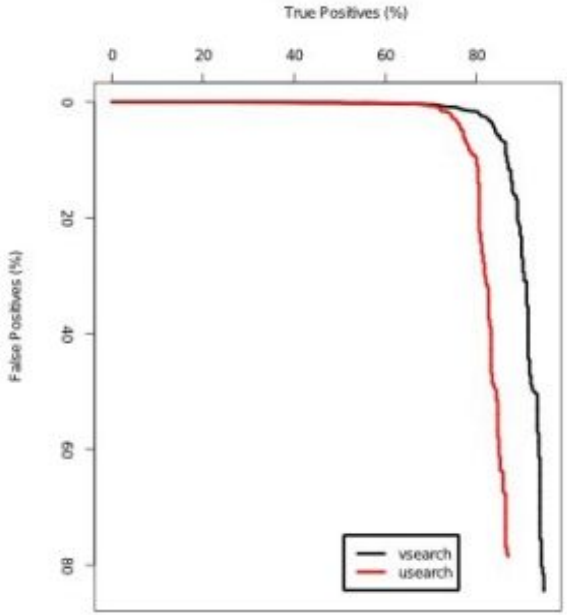
clustering, chimera detection, dereplication, searching, sorting, masking and shuffling

usearch (Rob Edgar):
- very important for metagenomics,
- 1,000 citations,
- fundation for QIIME,
- closed-source & costly

Quantitative Insights Into Microbial Ecology
qiime

growing success:
- many happy users,
- faster and improved,
- fundation for QIIME 2.0

vsearch:
- free and open-source,
- fast,
- documented,
- revive the research field



Torbjørn Rognes
Oslo University

https://github.com/torognes/vsearch.git

Practice

1.4

Aller sur la partie 1.4 remove chimera *du github*

OTU Filters

- The OTUs (Operational Taxonomic Units) have now been clustered. A filtering tool allows to remove noisy data. In this step, we will filter out some of the OTUs that are either not in at least 2 samples, and contain at least 2 sequences. Last allows eliminate singletons.
- Filters can be also done after affiliation taxonomy.

Practice

1.5
1.6

Aller sur la partie 1.5 OTU filtering *et* 1.6 Clustering stats *du github*

Affiliation OTU

- An OTU is a cluster of sequences. This step adds the taxonomy to the abundance file. It uses the SILVA database for rRNA.
- Affiliation tool returns taxonomic affiliation for each OTU using two methods with a unique multi-affiliation output

rdp

silva
high quality ribosomal RNA databases

Affiliation Strategy of FROGS

Double Affiliation with :

1. RDPClassifiers
2. Blastn+ : all identical Best Hits with the tag "**Multi-affiliation**".

V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | 16S unknown species
V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | 16S Butyrivibrio fibrisolvens
V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | 16S rumen bacterium 8 | 9293-9
V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | 16S Pseudobutyrivibrio xylanivorans
V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | 16S Pseudobutyrivibrio ruminis
FROGS Affiliation:	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | **Multi-affiliation**

Steps	Description
1	RDPClassifier is used with database to associate to each OTU a taxonomy and a bootstrap (example: *Bacteria;(1.0);Firmicutes;(1.0);Clostridia;(1.0);Clostridiales;(1.0);Clostridiaceae 1;(1.0);Clostridium sensu stricto;(1.0);*).
2	blastn+ is used to find alignment between each OTU and the database. Only the bests hits with the same score has reported.
3	For each OTU with several blastn+ results a consensus is determined on each taxonomic level. If all the taxa in a taxonomic rank are identical the taxon name is reported otherwise *Multi-affiliation* is reported. By example, if you have an OTU with two corresponding sequences, the first is a *Bacteria;Proteobacteria;Gamma Proteobacteria;Enterobacteriales*, the second is a *Bacteria;Proteobacteria;Beta Proteobacteria;Methylophilales*, the consensus will be *Bacteria;Proteobacteria;Multi-affiliation;Multi-affiliation*.

Affiliation Stats

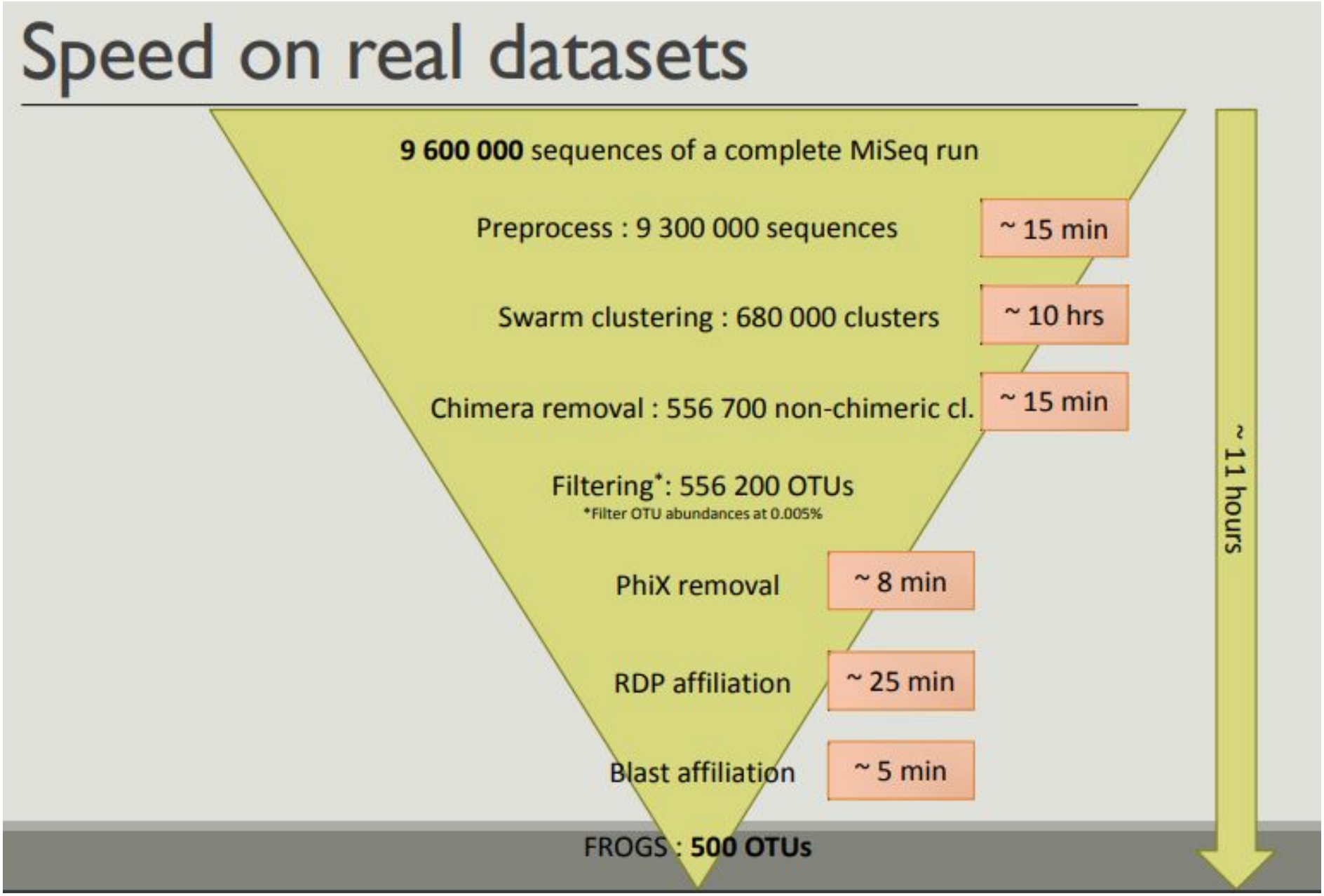
This step computes some statistics from the analysis and generates a report of the OTUs/taxonomy found.

Practice

1.7
1.8

Aller sur la partie 1.7 Affiliation taxonomique *et* 1.8 Affiliation stats *du github*

Speed on real datasets



BIOM Format

The BIOM format was motivated by several goals. First, to facilitate efficient handling and storage of large, sparse biological contingency tables; second, to support encapsulation of core study data (contingency table data and sample/observation metadata) in a single file; and third, to facilitate the use of these tables between tools that support this format (e.g., passing of data between QIIME, MG-RAST, and VAMPS.).

The FROGS biom format contains:

- OTU count tables (required)
- OTU description : taxonomy

BIOM Format


```
{"rows": [{"id": "Cluster_1", "metadata": {"comment": "na", "blast_taxonomy": "Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Mitochondria;unknown genus;Multi-affiliation",
 "rdp_bootstrap": "1.0;1.0;1.0;1.0;1.0;1.0;1.0", "taxonomy": ["Bacteria", "Proteobacteria", "Alphaproteobacteria", "Rickettsiales", "Mitochondria", "unknown genus", "Multi-affiliation
"], "seed_id": "SRR2107428.352", "rdp_taxonomy": "Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Mitochondria;unknown genus;Oryza sativa Indica Group (long-grained rice)"}}
, {"id": "Cluster_2", "metadata": {"comment": "na", "blast_taxonomy": "Bacteria;Multi-affiliation;Multi-affiliation;Multi-affiliation;Multi-affiliation;Multi-affiliation;Multi-affilia
tion", "rdp_bootstrap": "1.0;1.0;1.0;1.0;1.0;1.0;0.97", "taxonomy": ["Bacteria", "Multi-affiliation", "Multi-affiliation", "Multi-affiliation", "Multi-affiliation", "Multi-affiliation
", "Multi-affiliation"], "seed_id": "SRR2107428.552", "rdp_taxonomy": "Bacteria;Cyanobacteria;Oxyphotobacteria;Chloroplast;unknown family;unknown genus;Oryza sativa Japonica Group (Ja
panese rice)"}}, {"id": "Cluster_3", "metadata": {"comment": "na", "blast_taxonomy": "Bacteria;Firmicutes;Clostridia;Clostridiales;Family XVIII;unknown genus;unknown species", "rdp_bo
otstrap": "1.0;1.0;1.0;1.0;1.0;1.0;1.0", "taxonomy": ["Bacteria", "Firmicutes", "Clostridia", "Clostridiales", "Family XVIII", "unknown genus", "unknown species"], "seed_id": "SRR2107
428.5215", "rdp_taxonomy": "Bacteria;Firmicutes;Clostridia;Clostridiales;Family XVIII;unknown genus;unknown species"}}, {"id": "Cluster_4", "metadata": {"comment": "na", "blast_taxono
my": "Bacteria;Chloroflexi;Anaerolineae;SBR1031;A4b;unknown genus;unknown species", "rdp_bootstrap": "1.0;1.0;1.0;1.0;1.0;1.0;1.0", "taxonomy": ["Bacteria", "Chloroflexi", "Anaeroline
ae", "SBR1031", "A4b", "unknown genus", "unknown species"], "seed_id": "SRR2107428.1388", "rdp_taxonomy": "Bacteria;Chloroflexi;Anaerolineae;SBR1031;A4b;unknown genus;metagenome"}}, {
"id": "Cluster_5", "metadata": {"comment": "na", "blast_taxonomy": "Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Anaerosinus;Multi-affiliation", "rdp_bootstrap":
"1.0;1.0;1.0;1.0;1.0;1.0;0.75", "taxonomy": ["Bacteria", "Firmicutes", "Negativicutes", "Selenomonadales", "Veillonellaceae", "Anaerosinus", "Multi-affiliation"], "seed_id": "SRR21074
57.742", "rdp_taxonomy": "Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Anaerosinus;Massilibacillus massiliensis"}}, {"id": "Cluster_6", "metadata": {"comment": "n
a", "blast_taxonomy": "Bacteria;Chloroflexi;Chloroflexia;Thermomicrobiales;JG30-KF-CM45;unknown genus;unknown species", "rdp_bootstrap": "1.0;1.0;1.0;1.0;1.0;1.0;0.88", "taxonomy": ["
Bacteria", "Chloroflexi", "Chloroflexia", "Thermomicrobiales", "JG30-KF-CM45", "unknown genus", "unknown species"], "seed_id": "SRR2107428.6857", "rdp_taxonomy": "Bacteria;Chloroflexi
;Chloroflexia;Thermomicrobiales;JG30-KF-CM45;unknown genus;unknown species"}}, {"id": "Cluster_7", "metadata": {"comment": "na", "blast_taxonomy": "Bacteria;Proteobacteria;Gammaproteo
bacteria;Betaproteobacteriales;Rhodocyclaceae;Multi-affiliation;Multi-affiliation", "rdp_bootstrap": "1.0;1.0;1.0;1.0;1.0;0.89;0.85", "taxonomy": ["Bacteria", "Proteobacteria", "Gamma
proteobacteria", "Betaproteobacteriales", "Rhodocyclaceae", "Multi-affiliation", "Multi-affiliation"], "seed_id": "SRR2107428.9270", "rdp_taxonomy": "Bacteria;Proteobacteria;Gammaprot
eobacteria;Betaproteobacteriales;Rhodocyclaceae;Azonexus;unknown species"}}, {"id": "Cluster_8", "metadata": {"comment": "na", "blast_taxonomy": "Bacteria;Firmicutes;Clostridia;Clostr
idiales;Clostridiaceae 1;Clostridium sensu stricto 1;Multi-affiliation", "rdp_bootstrap": "1.0;1.0;1.0;1.0;1.0;1.0;0.69", "taxonomy": ["Bacteria", "Firmicutes", "Clostridia", "Clostri
diales", "Clostridiaceae 1", "Clostridium sensu stricto 1", "Multi-affiliation"], "seed_id": "SRR2107428.48665", "rdp_taxonomy": "Bacteria;Firmicutes;Clostridia;Clostridiales;Clostrid
```

```
ity": 100.0, "taxonomy": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae", "Pseudomonas", "Pseudomonas sp."], "evalue": "2.04e-130", "aln_le
ngth": 253, "perc_query_coverage": 100.0, "subject": "EU037285.1.1481"}, {"perc_identity": 100.0, "taxonomy": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales",
"Pseudomonadaceae", "Pseudomonas", "Pseudomonas sp."], "evalue": "2.04e-130", "aln_length": 253, "perc_query_coverage": 100.0, "subject": "EU043333.1.1393"}, {"perc_identity": 100.0,
"taxonomy": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae", "Pseudomonas", "Pseudomonas sp."], "evalue": "2.04e-130", "aln_length": 253,
perc_query_coverage": 100.0, "subject": "EU043332.1.1393"}, {"perc_identity": 100.0, "taxonomy": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonada
ceae", "Pseudomonas", "Pseudomonas mendocina"], "evalue": "2.04e-130", "aln_length": 253, "perc_query_coverage": 100.0, "subject": "GU227613.1.1465"}, {"perc_identity": 100.0, "taxono
my": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae", "Pseudomonas", "Pseudomonas sp."], "evalue": "2.04e-130", "aln_length": 253, "perc_qu
ery_coverage": 100.0, "subject": "EF550156.1.1385"}, {"perc_identity": 100.0, "taxonomy": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae",
"Pseudomonas", "Pseudomonas oleovorans"], "evalue": "2.04e-130", "aln_length": 253, "perc_query_coverage": 100.0, "subject": "GQ387664.1.1453"}, {"perc_identity": 100.0, "taxonomy":
"Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae", "Pseudomonas", "Pseudomonas sp."], "evalue": "2.04e-130", "aln_length": 253, "perc_query_co
verage": 100.0, "subject": "EF198249.1.1532"}, {"perc_identity": 100.0, "taxonomy": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae", "Pseu
omonas", "Pseudomonas sp."], "evalue": "2.04e-130", "aln_length": 253, "perc_query_coverage": 100.0, "subject": "EF197741.1.1336"}, {"perc_identity": 100.0, "taxonomy": ["Bacteria",
Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae", "Pseudomonas", "Pseudomonas sp."], "evalue": "2.04e-130", "aln_length": 253, "perc_query_coverage": 100.
0, "subject": "EF175872.1.1441"}, {"perc_identity": 100.0, "taxonomy": ["Bacteria", "Proteobacteria", "Gammaproteobacteria", "Pseudomonadales", "Pseudomonadaceae", "Pseudomonas", "Pse
udomonas sp."], "evalue": "2.04e-130", "aln_length": 253, "perc_query_coverage": 100.0, "subject": "EF198250.1.1514"}, {"perc_identity": 100.0, "taxonomy": ["Bacteria", "Proteobacteri
```


The FROGS biom format contains:

- OTU count tables (required)
- OTU description : taxonomy

BIOM standardization

You can retrieve a standardize biom file using FROGS BIOM to std BIOM

Others solutions to obtain a standard BIOM :

Example
You can eliminate OTUs from Chloroplast or/and Mitochondria :

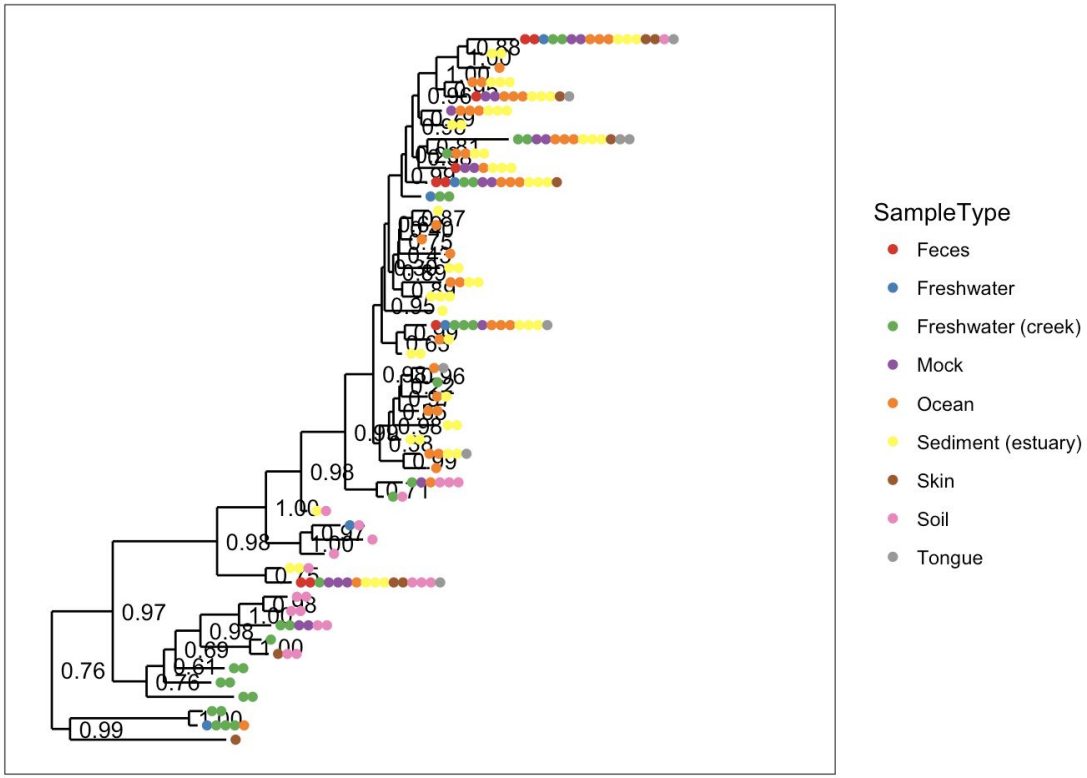
- If you want to remove these OTUs, you can convert your BIOM to TSV.
- Eliminate these OTUs and reconvert your TSV to BIOM using *FROGS TSV to BIOM and BIOM to std BIOM*

Practice

1.9

Aller sur la partie 1.9 BIOM format standardization *du github*

Building a Tree



Alignement fait avec MAFFT
Tree construit avec FastTree

Practice

1.10

Aller sur la partie 1.10 Building a tree *du github*

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

FROGSSTATS with Phyloseq R package

Goals

- Exploratory Data Analysis
  - α-diversity: how diverse is my community?
  - β-diversity: how different are two communities?
  - Visual assessment of the data
    - Bar plots: what is the composition of each community?
    - Multidimensional Scaling: how are communities related?
    - Heatmaps: are there interactions between species and (groups of) communities?
  - Use a distance matrix to study structures:
    - Hierarchical clustering: how do the communities cluster?
    - Permutational ANOVA: are the communities structured by some known environmental factor (pH, height, etc)?

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

FROGSSTAT with Phyloseq R package

- R package (McMurdie and Holmes, 2013) to analyse community composition data in a phylogenetic framework

It uses other R packages:

- Community ecology functions from vegan, ade4, picante
- Tree manipulation from ape
- Graphics from ggplot2
- (Differential analysis from DESeq2)

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

MODULES

- **FROGSSTAT Phyloseq Import**

  - Convert biom and sample information in a R objet.
  - Needs phylogenetic tree.
  - Metadata order (in each sample variable) are used to organised graphics. So take extra care when you construct your sample_metadata file.

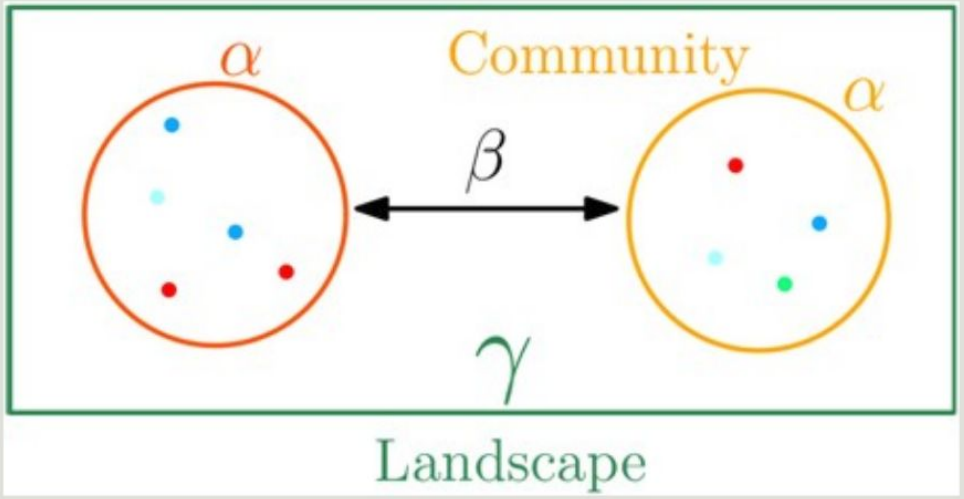
MODULES

- **FROGSSTAT Phyloseq Composition Visualisation** :
  - Exploring biodiversity => Explore the sample raw count according to a sample variable (env_material, geography ...) to organise data graphics.

Exploring biodiversity : statistical indices

Second Step, compute and compare diversity indices. Three flavors of diversity:

- **α-diversity**: diversity within a community;
- **β-diversity**: diversity between communities;
  - β-dissimilarities/distances
    - Dissimilarities between pairs of communities
    - Often used as a first step to compute-diversity
- γ-diversity: diversity at the landscape scale (blurry for bacterial communities);



http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

Diversity indices and metrics

- alpha diversity vs beta-diversity

	PnG.AD3.5	PnGa.N3.1	PnGa.N2.2	PnG.A1.1	PnGa.N2.5	PnGa.C1.5	PnG.A2.5	PnGa.C1.3	Kingdom	Phylum	Class	Order	Family	Genus	Species
OTU92316380387813	24	1	0	1	0	4	5	1	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Mesorhizobium	unknown
OTU73918700542926	101	0	0	0	0	6	1	12	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	unknown	unknown
OTU48893143140721	0	0	0	0	0	2	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium	unknown
OTU86008798522388	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	A0839	unknown	unknown
OTU95550865589689	11	0	0	0	0	0	0	0	Bacteria	Acidobacteria	Blastocatellia	Blastocatellales	Blastocatellaceae (Subgroup 4)	Blastocatella	unknown
OTU80396560526700	0	0	0	0	0	2	0	2	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	unknown	unknown
OTU46199850751507	0	0	0	0	0	0	0	0	Bacteria	Actinobacteria	Acidimicrobiia	Acidimicrobiales	unknown	unknown	unknown
OTU11038366021679	0	0	0	3	0	0	1	0	Bacteria	Proteobacteria	Alphaproteobacteria	unknown	unknown	unknown	unknown
OTU2893239483436	0	11	0	4	7	0	71	0	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	unknown	unknown
OTU79805388854995	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	unknown	unknown
OTU93544040077386	9	0	0	0	0	0	0	0	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Nocardioides	unknown
OTU99287582337946	0	4	12	0	10	0	0	6	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Polyangiaceae	Sorangium	unknown
OTU42963829278873	0	0	0	0	0	0	0	0	Bacteria	Actinobacteria	Thermoleophilia	Gaiellales	unknown	unknown	unknown
OTU3845850732916	0	0	0	4	0	0	1	0	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	unknown	unknown	unknown
OTU41491064711811	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Lysobacter	unknown
OTU58528220400457	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Alphaproteobacteria Incertae Sedis	Unknown Family	unknown	unknown
OTU10265322223188	0	0	0	0	0	0	1	0	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Propionivibrio	unknown
OTU39090004064502	2	13	1	23	1	22	21	20	Bacteria	Saccharibacteria	unknown	unknown	unknown	unknown	unknown
OTU82925522100430	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium	unknown
OTU63057105915481	0	0	0	3	0	0	5	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Solimonadaceae	unknown	unknown
OTU24880888488557	0	0	0	0	0	0	4	1	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Lysobacter	unknown
OTU86133997633085	0	0	0	0	0	2	0	5	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	unknown
OTU21709764964738	11	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	unknown	unknown
OTU95889076292545	0	0	0	74	10	2	62	1	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Cloacibacterium	unknown
OTU22275543613597	0	2	0	4	0	5	17	0	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	unknown
OTU41875855341992	105	0	0	1	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas	unknown
OTU46740806112089	61	21	3	4	7	7	0	1	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	unknown	unknown
OTU13716996697572	0	1	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	unknown	unknown
OTU75707377710417	6	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	unknown	unknown
OTU78095271946327	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Solimonadaceae	unknown	unknown
OTU64875474337174	0	0	0	1	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Thermomonas	unknown

Diversity indices and metrics

- alpha diversity vs beta-diversity

	PnG.AD3.5	PnGa.N3.1	PnGa.N2.2	PnG.A1.1	PnGa.N2.5	PnGa.C1.5	PnG.A2.5	PnGa.C1.3	Kingdom	Phylum	Class	Order	Family	Genus	Species
OTU92316380387813	4	1	0	1	0	4	5	1	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Mesorhizobium	unknown
OTU73918700542926	1	0	0	0	0	6	1	12	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	unknown	unknown
OTU48893143140721	0	0	0	0	0	2	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium	unknown
OTU86008798522388	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	A0839	unknown	unknown
OTU95550865589689	1	0	0	0	0	0	0	0	Bacteria	Acidobacteria	Blastocatellia	Blastocatellales	Blastocatellaceae (Subgroup 4)	Blastocatella	unknown
OTU80396560526700	0	0	0	0	0	2	0	2	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	unknown	unknown
OTU46199850751507	0	0	0	0	0	0	0	0	Bacteria	Actinobacteria	Acidimicrobiia	Acidimicrobiales	unknown	unknown	unknown
OTU11038366021679	0	0	0	3	0	0	1	0	Bacteria	Proteobacteria	Alphaproteobacteria	unknown	unknown	unknown	unknown
OTU2893239483436	0	11	0	4	7	0	71	0	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	unknown	unknown
OTU79805388854995	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	unknown	unknown
OTU93544040077386	9	0	0	0	0	0	0	0	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Nocardioidaceae	Nocardioides	unknown
OTU99287582337946	0	4	12	0	10	0	0	6	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Polyangiaceae	Sorangium	unknown
OTU42963829278873	0	0	0	0	0	0	0	0	Bacteria	Actinobacteria	Thermoleophilia	Gaiellales	unknown	unknown	unknown
OTU3845850732916	0	0	0	4	0	0	1	0	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	unknown	unknown	unknown
OTU41491064711811	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Lysobacter	unknown
OTU58528220400457	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Alphaproteobacteria Incertae Sedis	Unknown Family	unknown	unknown
OTU10265322223188	0	0	0	0	0	0	1	0	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Propionivibrio	unknown
OTU39090004064502	2	13	1	23	1	22	21	20	Bacteria	Saccharibacteria	unknown	unknown	unknown	unknown	unknown
OTU82925522100430	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium	unknown
OTU63057105915481	0	0	0	3	0	0	5	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Solimonadaceae	unknown	unknown
OTU24880888488557	0	0	0	0	0	0	4	1	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Lysobacter	unknown
OTU86133997633085	0	0	0	0	0	2	0	5	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	unknown
OTU21709764964738	1	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	unknown	unknown
OTU95889076292545	0	0	0	74	10	2	62	1	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Cloacibacterium	unknown
OTU22275543613597	0	2	0	4	0	5	17	0	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	unknown
OTU41875855341992	5	0	0	1	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas	unknown
OTU46740806112089	1	21	3	4	7	7	0	1	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	unknown	unknown
OTU13716996697572	0	1	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	unknown	unknown
OTU75707377710417	6	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	unknown	unknown
OTU78095271946327	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Solimonadaceae	unknown	unknown
OTU64875474337174	1	0	0	1	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Thermomonas	unknown

alpha- diversity

Exploring biodiversity : α-diversity

α-diversity indices available in phyloseq :

- Species **richness** : number of observed otus
- **Chao1** : number of observed otu + estimate of the number of unobserved otus
- **Shannon** entropy / **Jensen** : the width of the otu relative abundance distribution. Roughly, it reflects our (in)ability to predict the otu of a randomly picked bacteria.
- **Simpson** : 1 - probability that two bacteria picked at random in the community belong to different otu.
- **Inverse Simpson** : inverse of the probability that two bacteria picked at random belong to the same otu.

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

MODULES

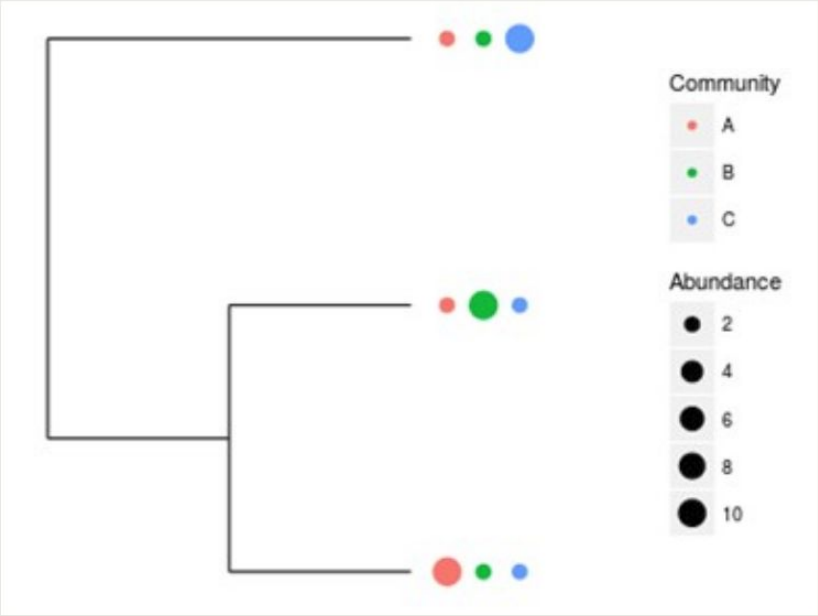
- **FROGSSTAT Phyloseq Alpha Diversity** :
  - Choose which α-diversity indices you want to compute according to sample variable

Exploring biodiversity : β-diversity

Many diversities (both compositional and phylogenetic) offered by Phyloseq through the generic distance function.

Different dissimilarities capture different features of the communities:

- qualitatively, communities are very similar
- quantitatively they are very different
- phylogeneticaly two communities seems to be closer than the third one.



http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

Exploring biodiversity : β-diversity

- In general, qualitative diversities are more sensitive to factors that affect presence/absence of organisms (such as pH, salinity, depth, etc) and therefore useful to study and define bioregions (regions with little of no flow between them)...
- ... whereas quantitative distances focus on factors that affect relative changes (seasonal changes, nutrient availability, concentration of oxygen, depth, etc) and therefore useful to monitor communities over time or along an environmental gradient.

Different distances capture different features of the samples.

There is no "one size fits all"

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

FROGSSTATS with Phyloseq R package

MODULES

- **FROGSSTAT Phyloseq beta Diversity** :
  - Choose which β-diversity indices you want to compute according to sample variable
  - Phyloseq supports currently 43 beta diversity distance methods :

"unifrac" "wunifrac" ”dpcoa" "jsd" "manhattan" "euclidean" "canberra" "bray" "kulczynski " "jaccard" "gower" "altGower" "morisita" "horn" "mountford" "raup" "binomial" "chao" "cao" ...

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

FROGS

Exploring the structure : ordination

- Each community is described by otus abundances
- Otus abundance maybe correlated
- PCA finds linear combinations of otus that
  - are uncorrelated
  - capture well the variance of community composition

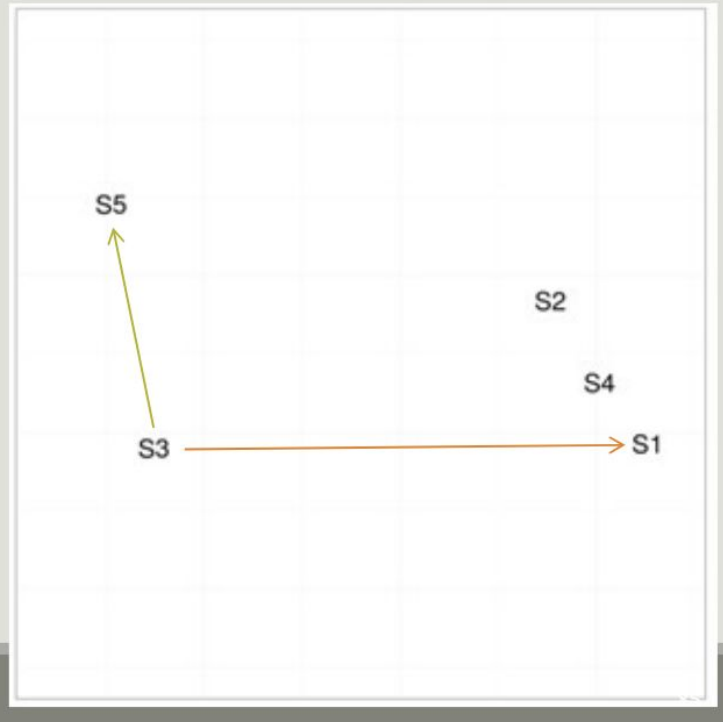
But variance is not a very good measure of β-diversity

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

Exploring the structure : ordination

The Multidimensionnal Scaling (MDS or PCoA) is equivalent to a principal component Analysis (PCA) but preserves the β-diversity instead of the variance.
The MDS try to represent samples in two dimensions
➔ The samples ordination.

	Distance Matrix				
	S1	S2	S3	S4	S5
S1	0.00	2.21	6.31	0.99	7.50
S2	2.21	0.00	5.40	1.22	5.74
S3	6.31	5.40	0.00	5.75	3.16
S4	0.99	1.22	5.75	0.00	6.64
S5	7.50	5.74	3.16	6.64	0.00



http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

Exploring the structure : Heatmap

- Heatmap is an other representation of the abundance table.
- It tries to reveal if there is a structure between a group of OTUs and a group of samples.
- It
  - Finds a meaningful order of the samples and the OTUs
  - Allows the user to choose a custom order (in R)
  - Allows the user to change the colour scale (in R)
  - Produces a ggplot2 object, easy to manipulate and

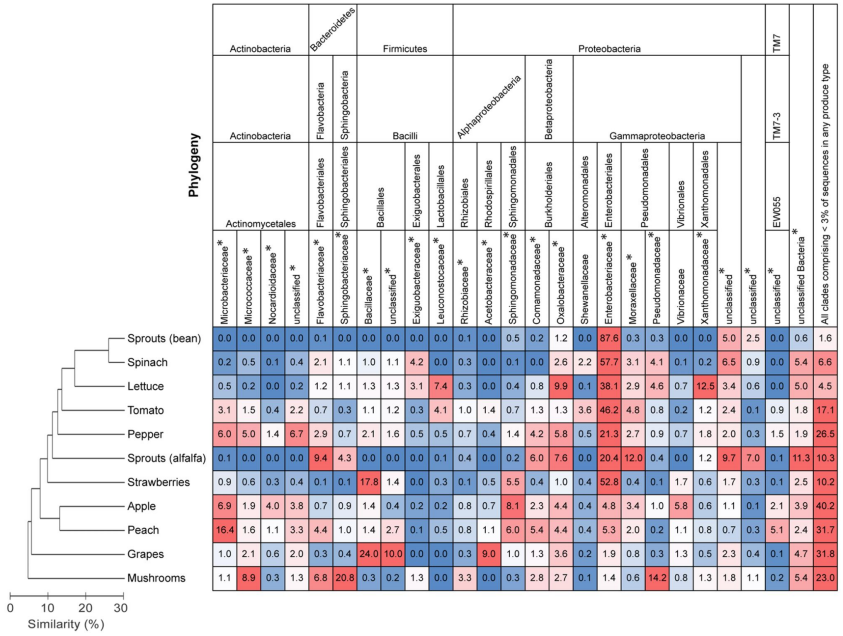


Figure 2. Relationships between bacterial communities on each produce type and relative abundances of bacterial families. The dendrogram is based on mean Bray-Curtis dissimilarities and shows differences among produce types in the overall composition of the bacterial communities. The heatmap shows mean relative abundances (%) of bacterial families on produce types. Only families and unclassified groupings representing at least three percent on any produce type are represented.
doi:10.1371/journal.pone.0059310.g002

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

MODULES

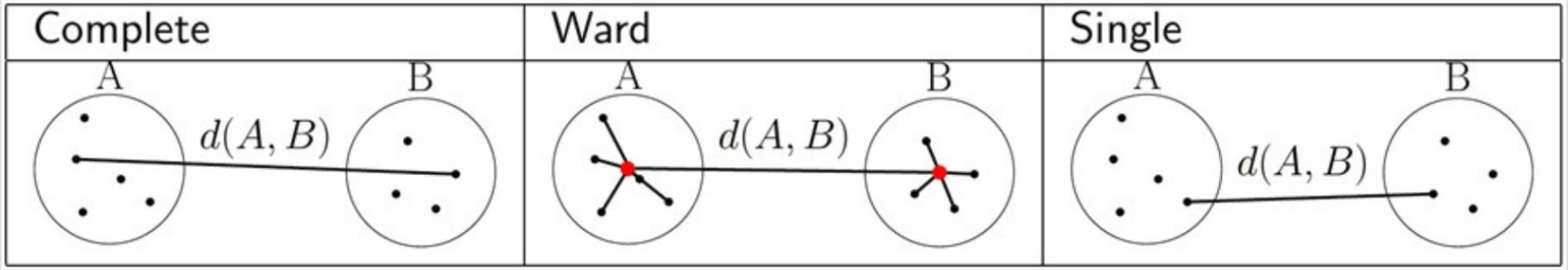
- **FROGSSTAT Phyloseq Structure Visualisation**

- head map plot and ordination
  - Explore the sample normalised count using a sample variable
  - Use the beta diversity distance matrix
  - Ordination method (most commonly used is MDS/Pcoa)

Exploring the structure : clustering

The clustering aims to represent samples in a tree based on a distance matrix and a linkage function:

- Complete linkage: tends to produce compact, spherical clusters and guarantees that all samples in a cluster are similar to each other.
- Ward: tends to also produces spherical clusters but has better theoretical properties than complete linkage.



http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

MODULES

- **FROGSSTAT Phyloseq Sample Clustering**

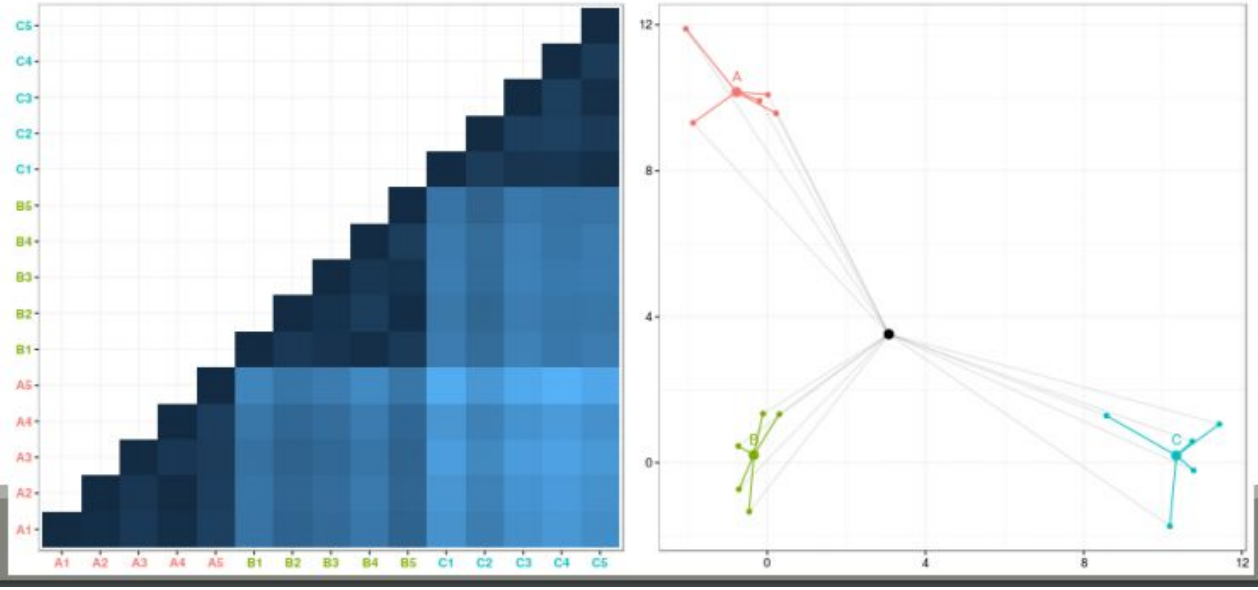
  Explore the sample normalised count using a sample variable
  Use the beta diversity distance matrix
  The tree different linkage functions will be used, generating three different trees

Diversity partitioning

Are the structures seen linked to metadata ? Have the metadata got an effect on our communities composition ?

To answer these questions, **multivariate analyses** that :

- tests composition differences of communities from different groups using a distance matrix
- compares within group to between group distances



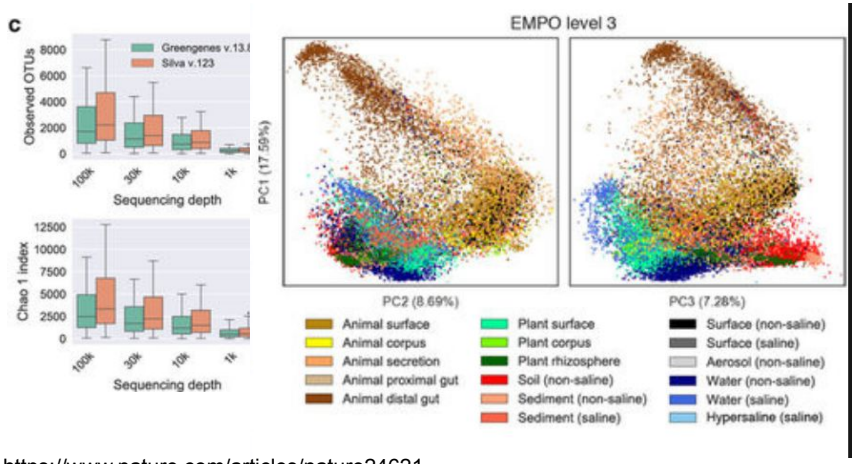
http://genoweb.toulouse.inra.fr/~formation/15_FROGS/11-April2018/FROGS_phyloseq_April_2018.pdf

MODULES

- **FROGSSTAT Phyloseq Anova**

  Which variable influence the diversity ?
  Does a given variable have an influence on the beta diversity variance ?

Common plots in microbiome studies



https://www.nature.com/articles/nature24621

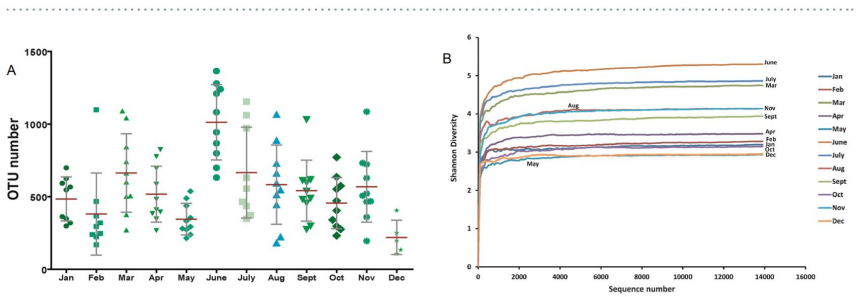


Figure 1. (**A**) Scatter plots showing OTU numbers with respect to total sequences obtained from 16 S rRNA genes associated with months. The error bars were the mean with SD. (**B**) Shannon diversity curve of bacterial community derived from 12 months.

http://dx.doi.org/10.1038/s41598-018-20862-8

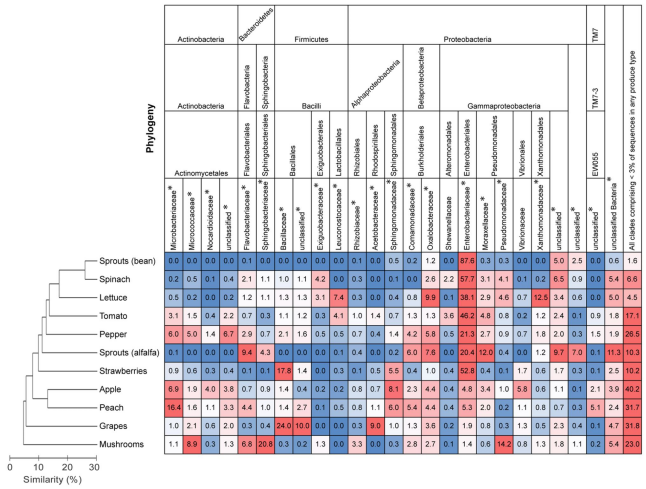
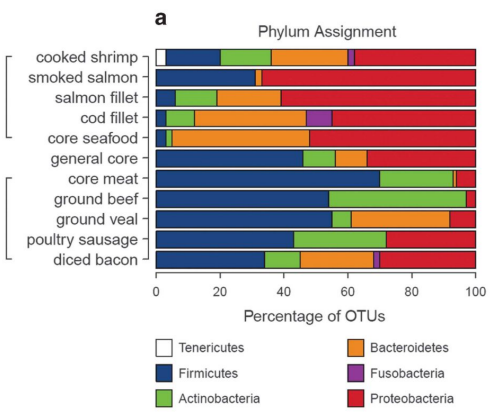
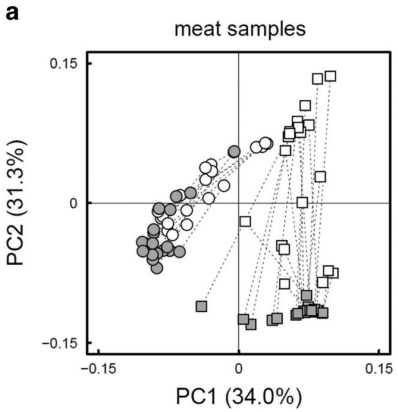


Figure 2. Relationships between bacterial communities on each produce type and relative abundances of bacterial families. The dendrogram is based on mean Bray-Curtis dissimilarities and shows differences among produce types in the overall composition of the bacterial communities. The heatmap shows mean relative abundances (%) of bacterial families on produce types. Only families and unclassified groupings representing at least three percent on any produce type are represented.
doi:10.1371/journal.pone.0059310.g002





http://www.nature.com/doifinder/10.1038/ismej.2014.202

… to describe alpha/beta-diversity and taxonomy of the communities and link patterns to covariables

Practice

1.11 *Aller sur la partie* Phyloseq stats in FROGSTAT *du github*

FROGS en Ligne de commande sur le cluster i-Trop IRD

Practice

2

Aller sur la partie FROGS in Command Line *du github*

Phyloseq sur R : Hands on!

Practice

3

Aller sur la practice 3 [Phyloseq stats in R](#) *du github*

Fast and accurate sample inference from amplicon data with single-nucleotide resolution

- DADA2 uses a probabilistic error model from the data itself
- Results in a dataset partitioning into Amplicon Sequence Variants (e.g. unique sequences) instead of clusters of similar sequences into OTUs

https://benjjneb.github.io/dada2/tutorial.html

Fast and accurate sample inference from amplicon data with single-nucleotide resolution

Processing with DADA2 in R

Setting up our working environment

Quality trimming/filtering

Generate an error model of our data

Dereplication (or not)

Inferring ASVs

Merge forward and reverse reads

Generate a count table

Chimera identification

Overview of counts throughout

Assigning taxonomy

Extracting the standard goods from R

Contributeurs pour cette formation








- Alexis Dereeper
- Julie Orjuela-Bouniol
- Florentin Constancias
- Julie Reveillaud
- Marie Simonin
- Frederic Mahé
- Aurore Compte
- Hans Schrieke

Bonnes pratiques

Comment citer Galaxy?

"The authors acknowledge the South Green Platform (http://www.southgreen.fr) for providing the galaxy instance (http://bioinfo-inter.ird.fr:8080/ or http://galaxy.southgreen.fr/galaxy/) that have contributed to the research results reported within this paper."

Comment citer les clusters?

"The authors acknowledge the IRD i-Trop HPC at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: http://bioinfo.ird.fr/ "

"The authors acknowledge the CIRAD UMR-AGAP HPC (South Green Platform) at CIRAD montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.southgreen.fr"

→ **N'oubliez pas de citer aussi les outils utilisés !**

SUIVEZ NOUS SUR TWITTER !

South Green : @green_bioinfo

I-Trop : @ItropBioinfo

Merci pour votre attention !