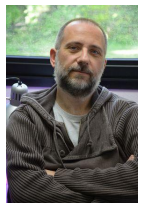
Session de formation 2018

SouthGreen bioinformatics platform

12 Mars	Guide de survie à Linux : les commandes de base pour débuter sur un serveur linux
13 Mars	Linux avancé : manipuler et filtrer des fichiers sans connaissance de programmation
15 Mars	Initiation à l'utilisation du cluster bioinformatique itrop
22 Mars	Initiation à git
23 Mars	Initiation aux gestionnaires de workflow South Green: Galaxy ou TOGGLe
26 Mars	Initiation aux analyses de données transcriptomiques
23 Avril	**Initiation aux analyses de données metabarcoding**

Bioversity International






Institut de Recherche pour le Développement

diade IPME









SouthGreen bioinformatics platform






agap UMR BGPI Biologie et Génétique des Interactions Plante-Parasite LSTM Laboratoire des Symbioses Tropicales et Méditerranéennes

INRA SCIENCE & IMPACT cirad



Bioversity International

Institut de Recherche pour le Développement

SouthGreen bioinformatics platform

diade IPME

UMI233 TransVIH-MI

MiVEGEC

Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle

plateau i-trop

agap UMR BGPI Biologie et Génétique des Interactions Plante-Parasite LSTM Laboratoire des Symbioses Tropicales et Méditerranéennes

INRA SCIENCE & IMPACT

cirad

www.southgreen.fr

https://github.com/SouthGreenPlatform

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

Session de formation 2018

- 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: Visualizing and plotting all sample results with Phinch

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

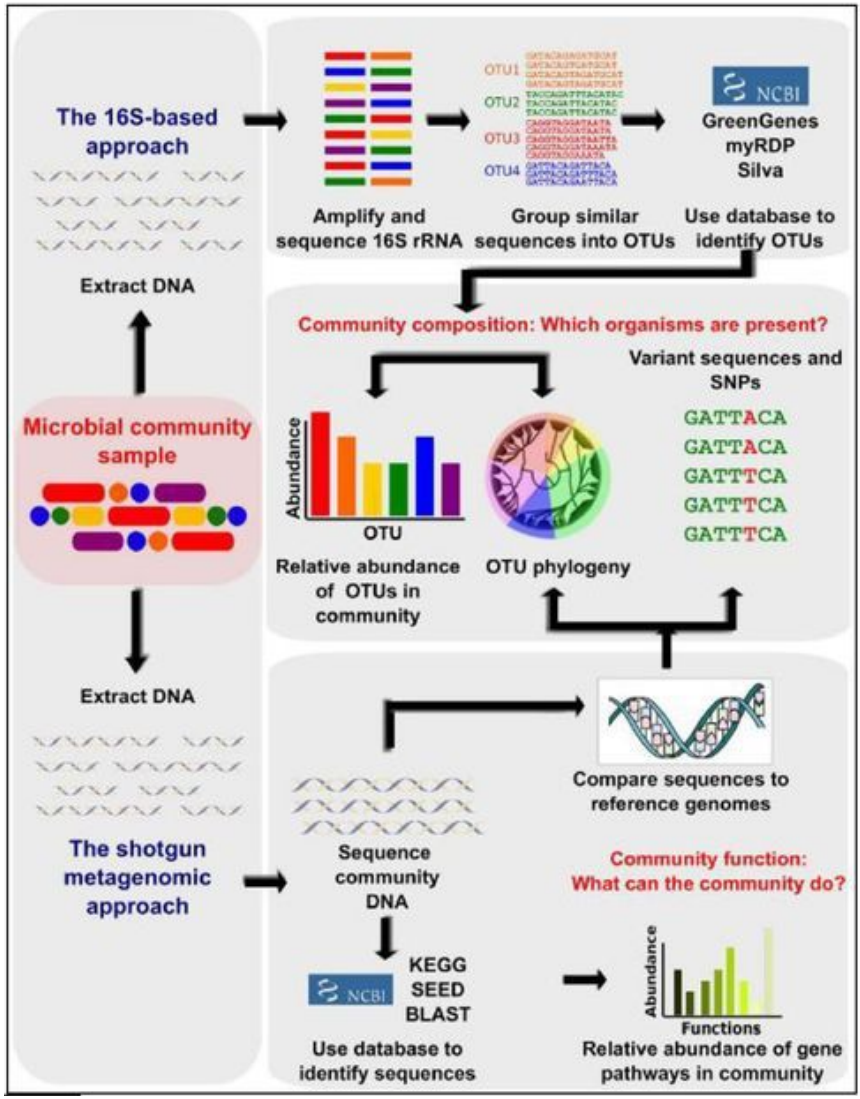
=> Organisms can be studied directly in their environments bypassing the need to isolate each species

=> 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

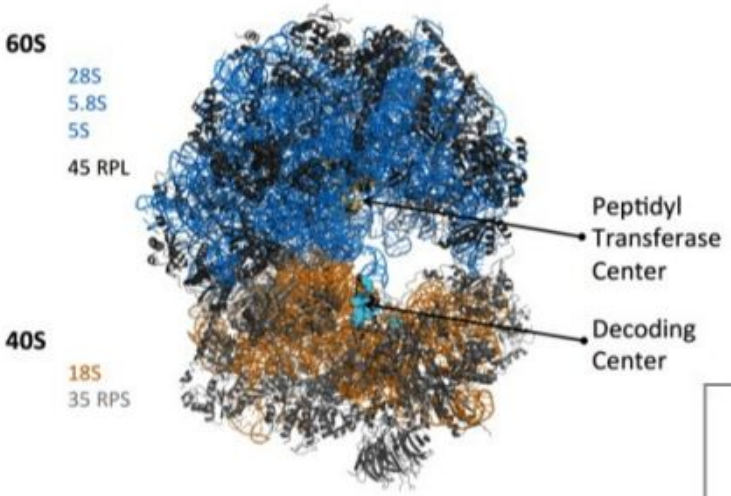
Strategies in Diversity Characterisation

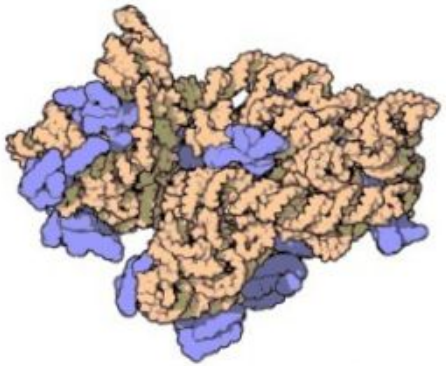
Technique	Advantages and challenges	Main applications
Metataxonomics using amplicon sequencing of the 16S or 18S rRNA gene or ITS	+ Fast and cost-effective identification of a wide variety of bacteria and eukaryotes – Does not capture gene content other than the targeted genes – Amplification bias – 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

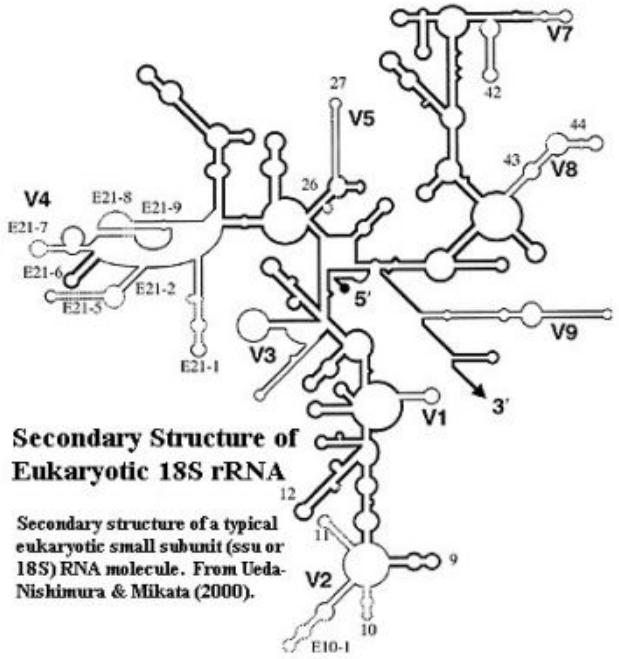
A universal gene: ribosomal RNA





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).

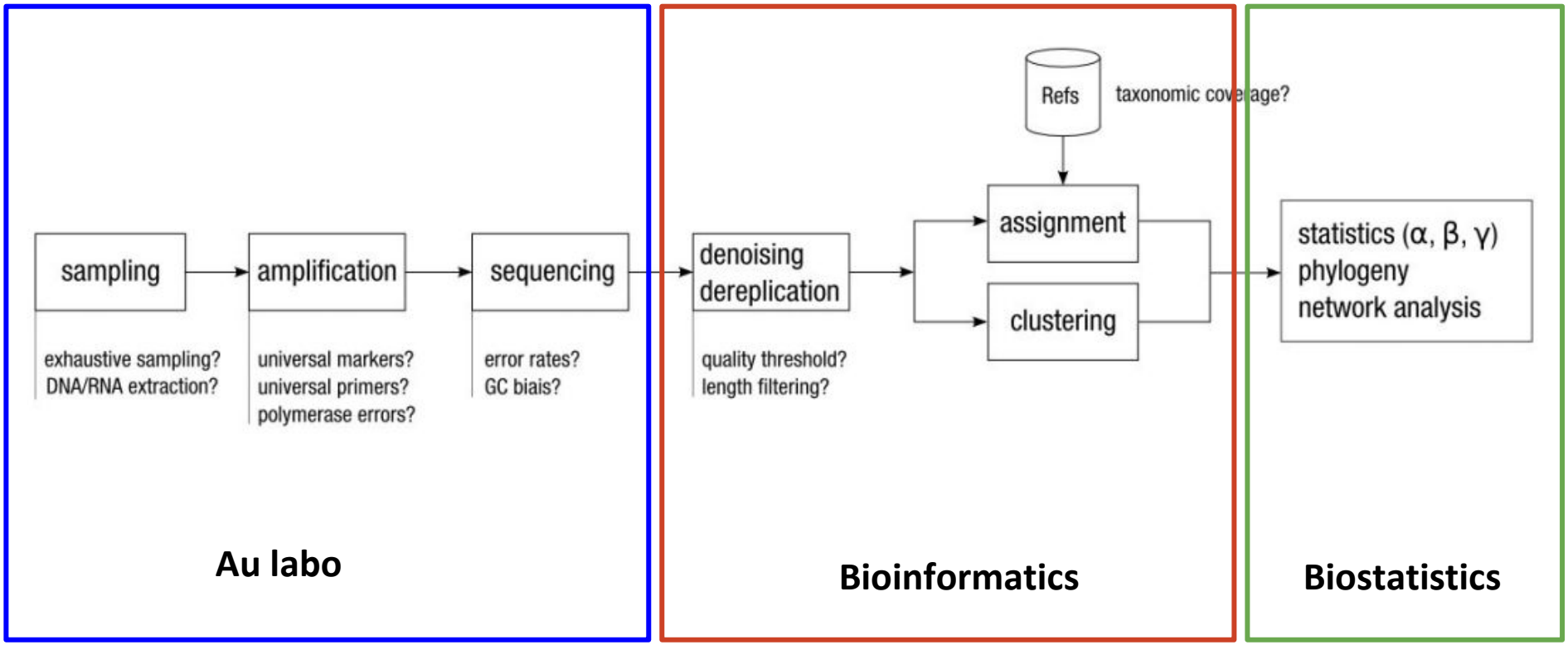
other markers can be used (e.g., ITS). Requirements are: conserved distal regions for primers, variable internal regions, and available sets of reference sequences.

Projets métagénomiques

4900 projets sur NCBI (avril 2018)

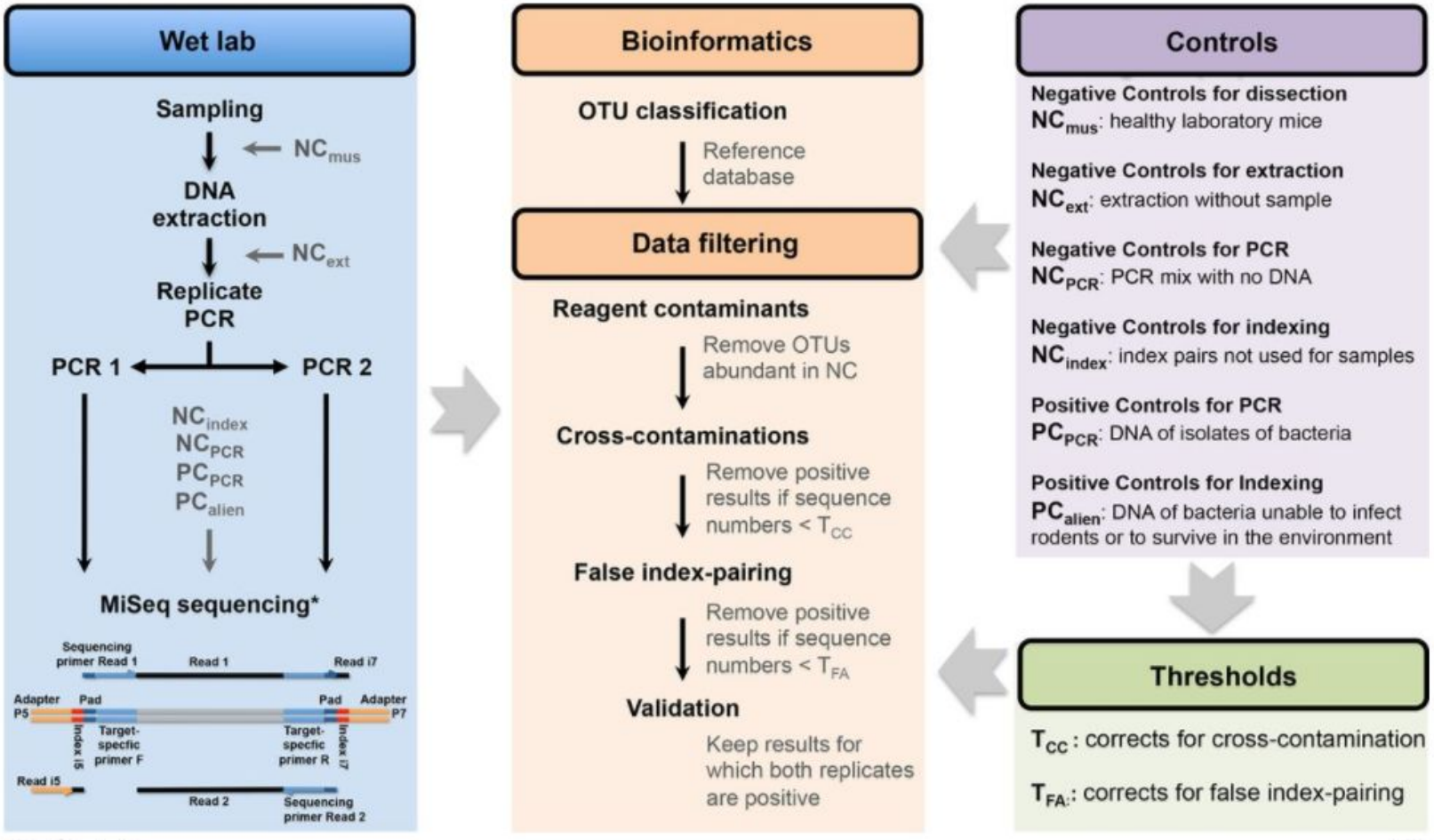
- Sable de plage
- Moustique
- Corail
- Glace
- Air de la ville de Singapour
- Surface de la cuvette des toilettes
- Fromages
- ...

Amplicon-based studies general pipeline



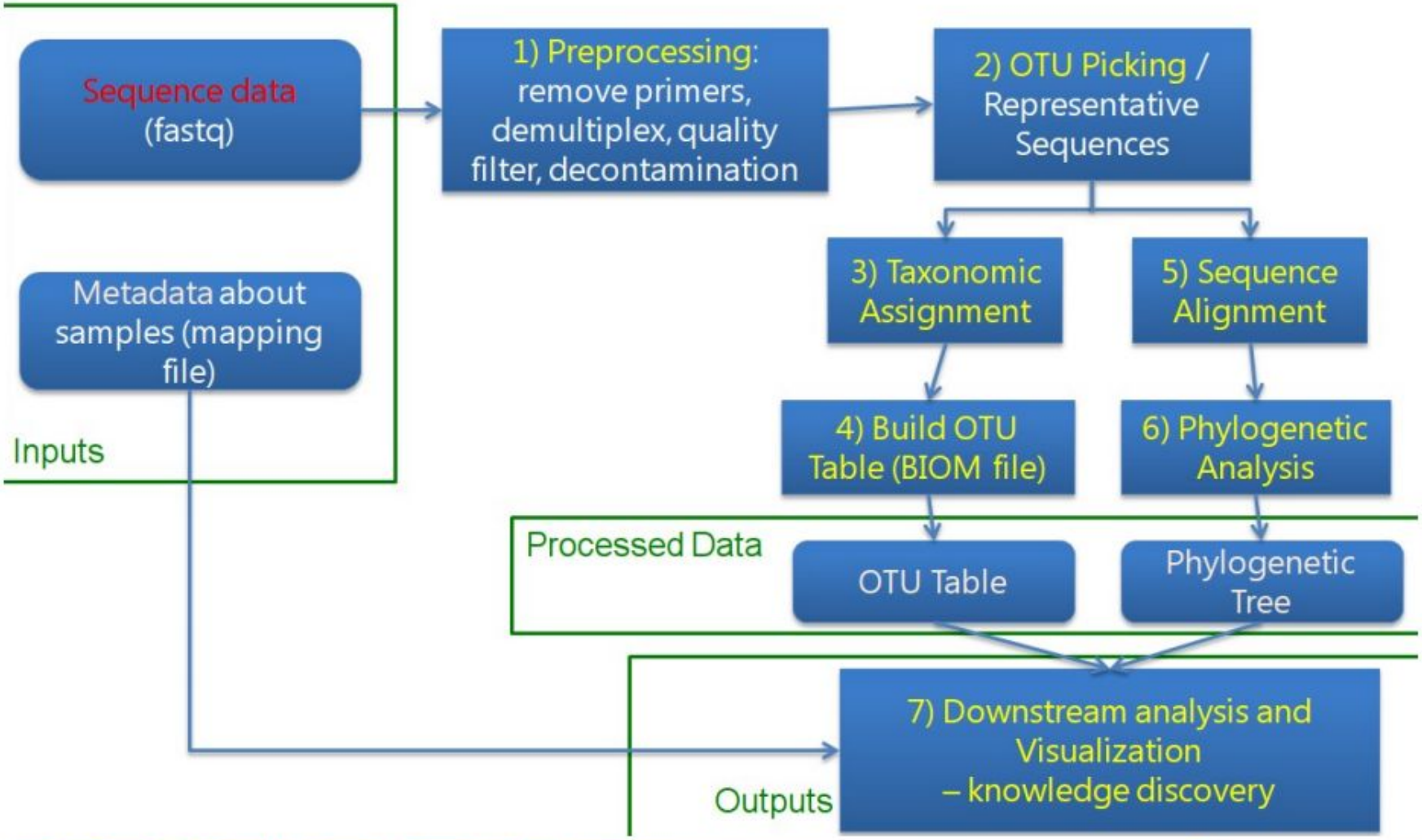
TODAY !

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

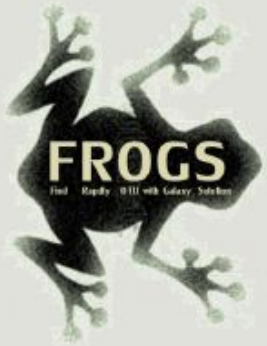
Qiime (2010)
Gregory Caporaso
open-source
python wrapper
most used(?)
stats

Uparse

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

Which bioinformatics solutions?

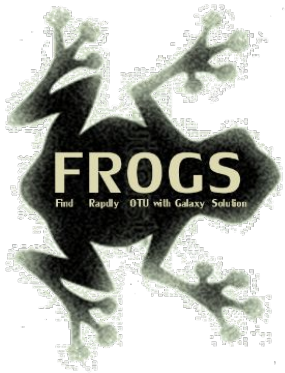
	Disadvantages
QIIME	Installation problem Command lines
UPARSE	Global clustering command lines
MOTHUR	Not MiSeq data without normalization Global hierarchical clustering Command lines
MG-RAST	No modularity No transparence



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

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

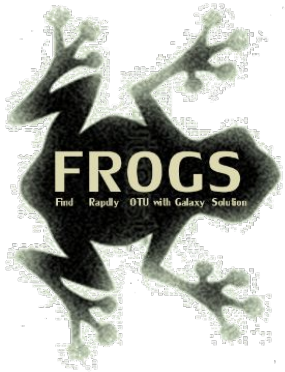
GitHub

Galaxy

Practice 1:

Obtaining an OTU table with FROGS in Galaxy

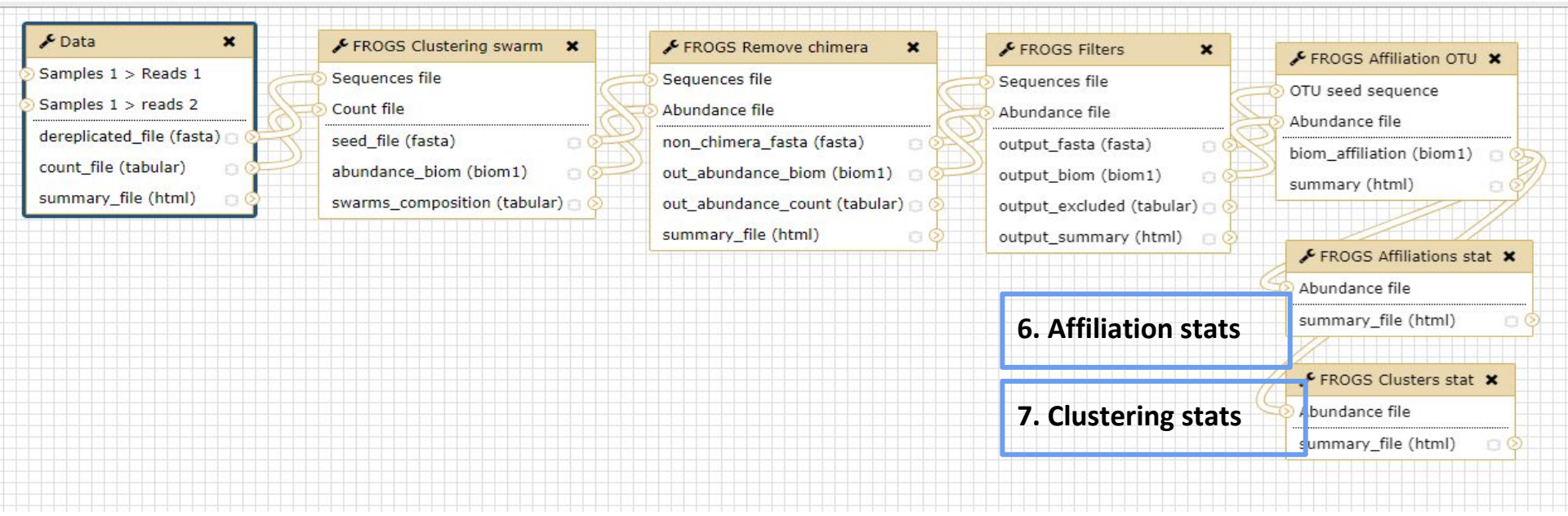
FROGS: Find, Rapidly, OTUs with Galaxy Solution



- 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



1.Pre-process

A preprocessing tool :

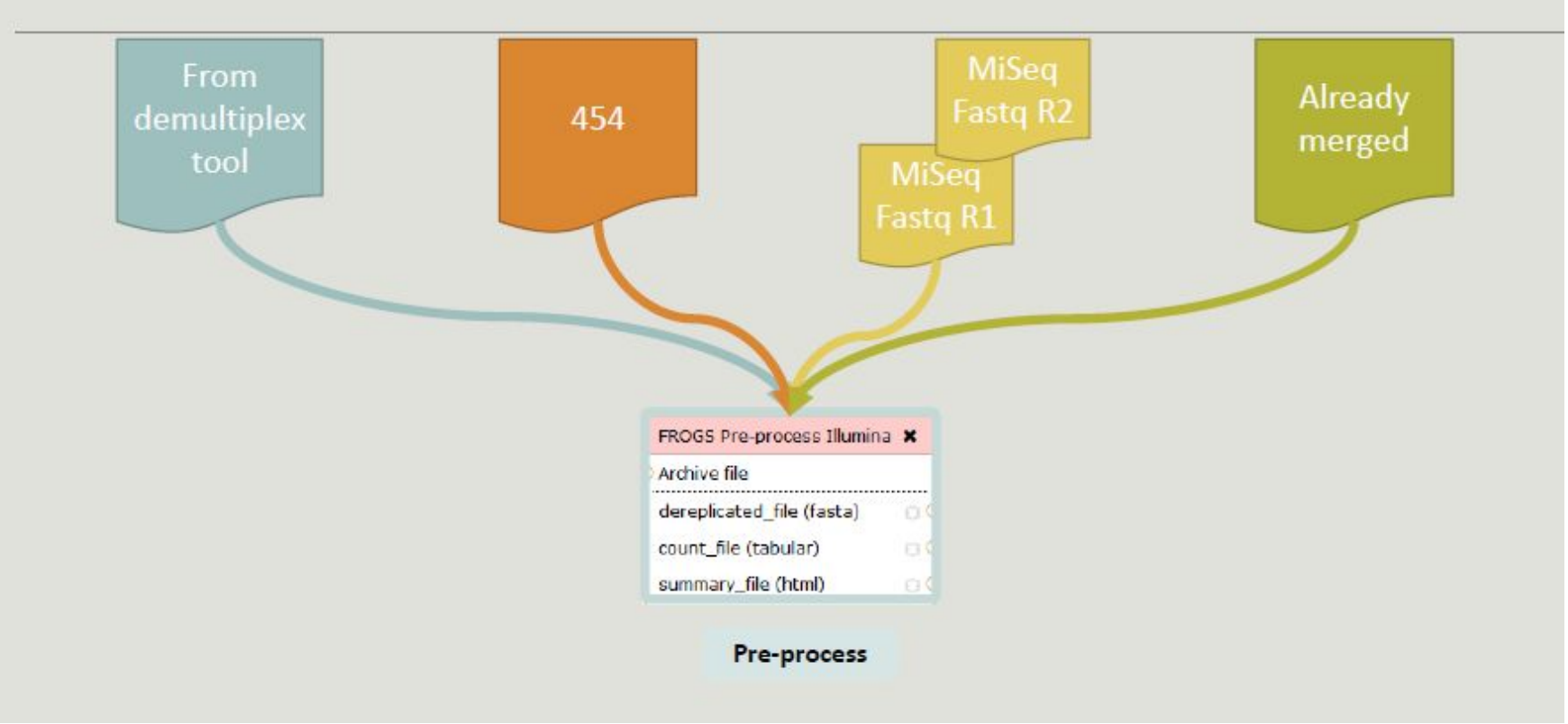
- merges paired sequences into contigs with flash,
- cleans the data with cutadapt,
- deletes the chimeras with VSEARCH combined with a cross-validation method and
- dereplicates sequences with a home-made python script.

1.Pre-process

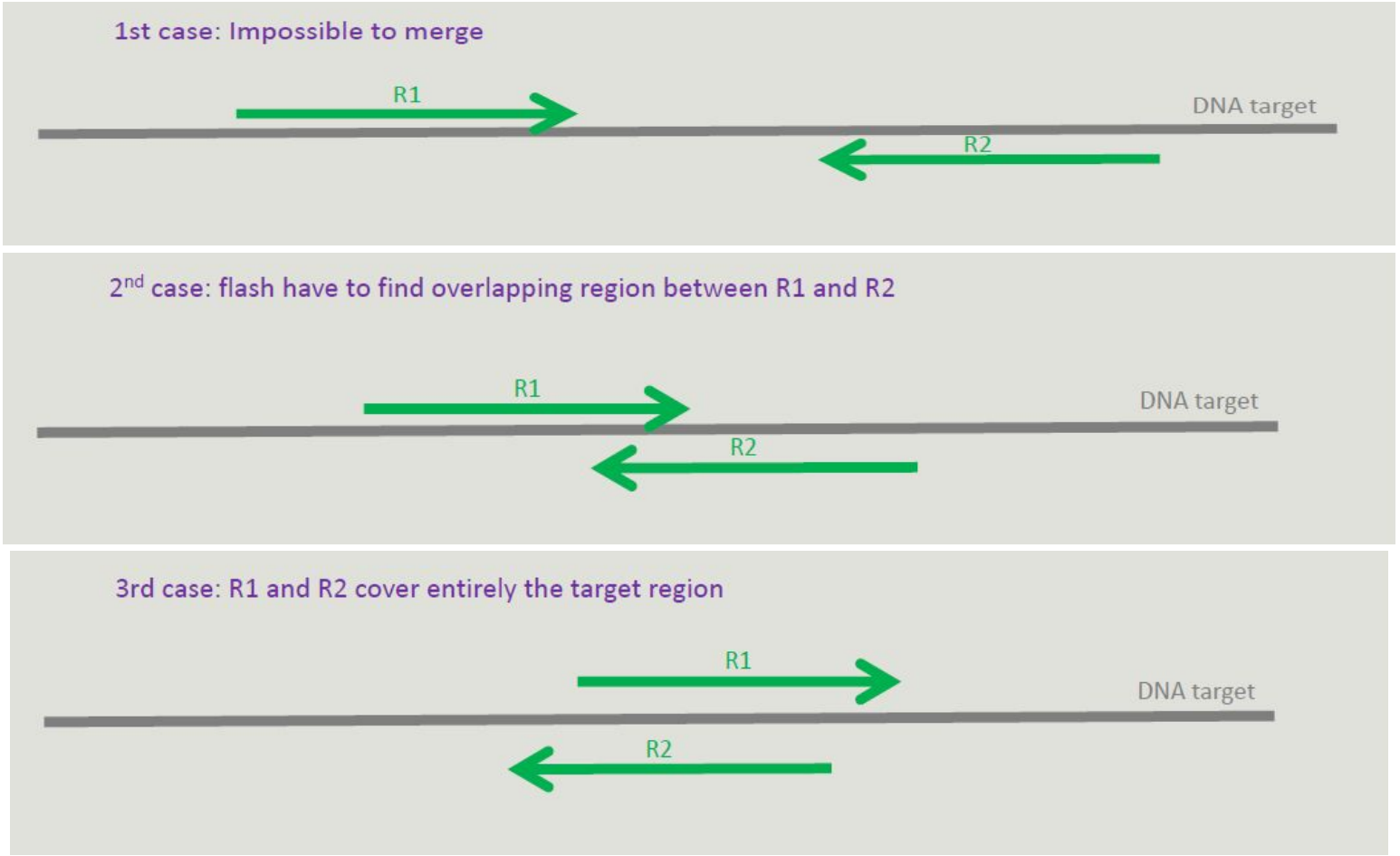
FROGs takes the reads (R1 and R2) from multiple samples and performs the following steps:

- 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)

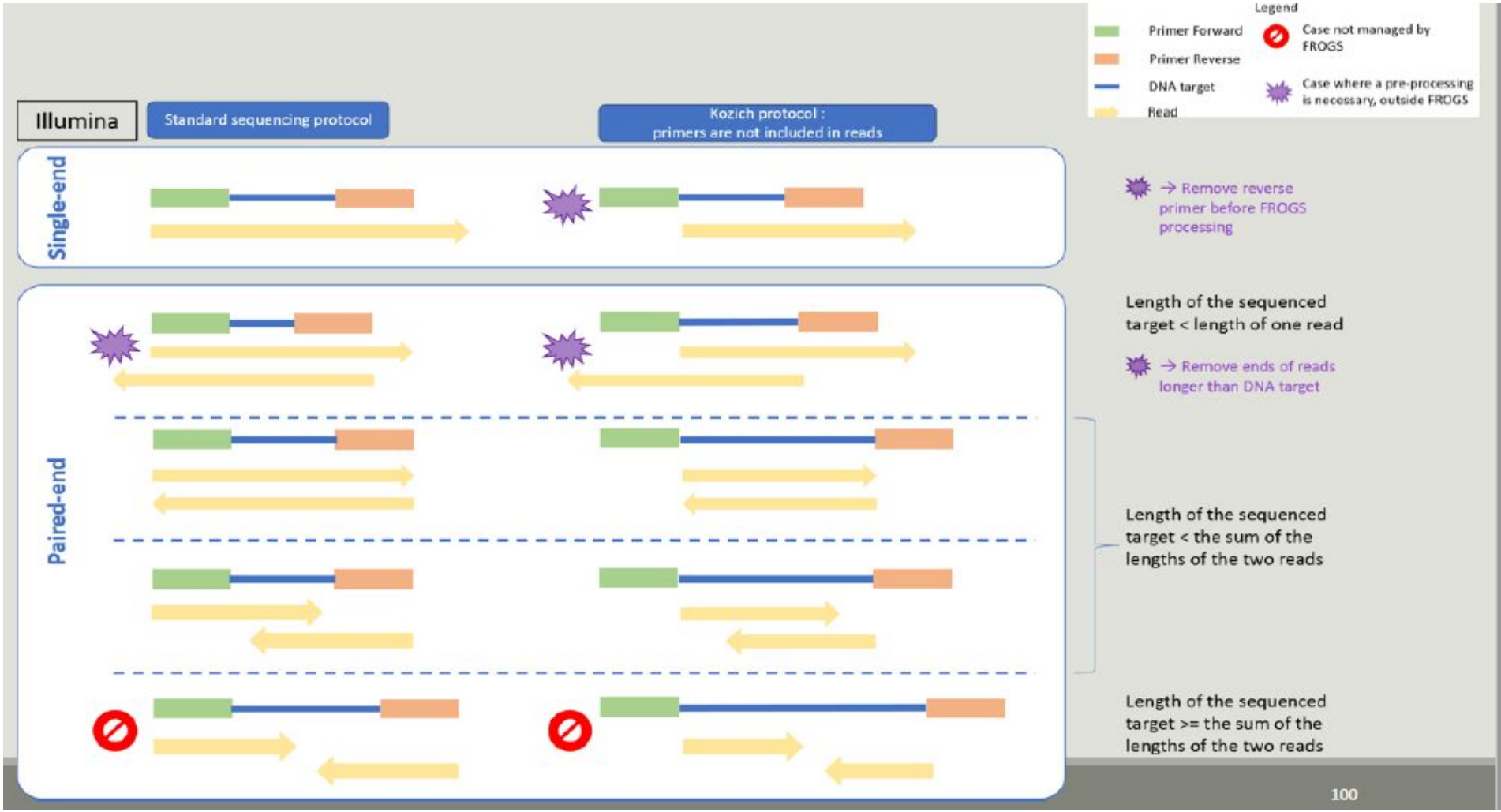


The goal of Flash (**Fast Length Adjustment of Short reads**) is to merge R1 and R2



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	✓

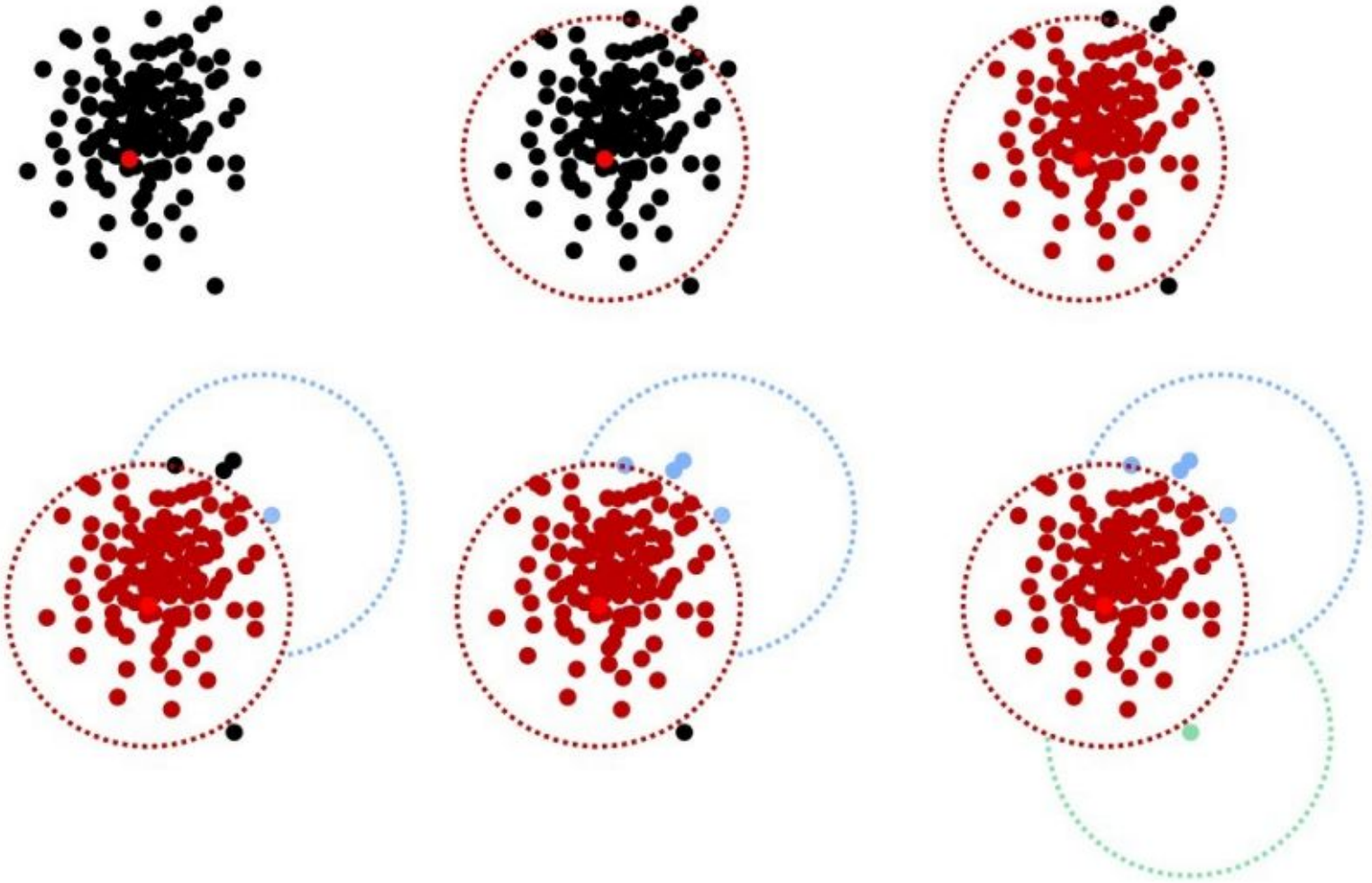
2. Clustering Swarm

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

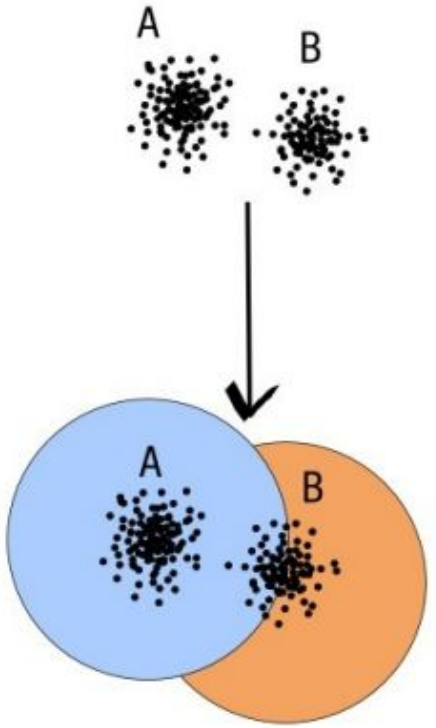
- Sorts reads by abundance.
- Clusters the reads into pre-clusters using Swarm and distance parameter of 1.
- Sorts these pre-clusters by abundance.
- Cluster the pre-clusters using Swarm and a user-specified distance.

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

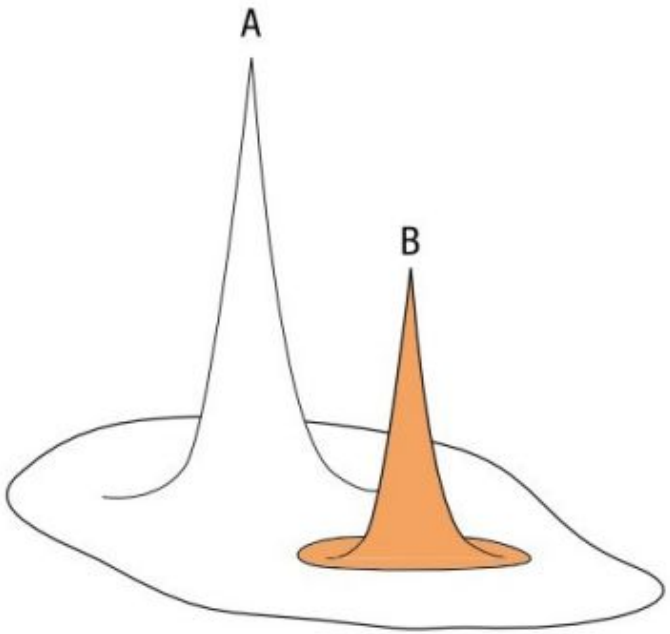
How traditional clustering works?



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

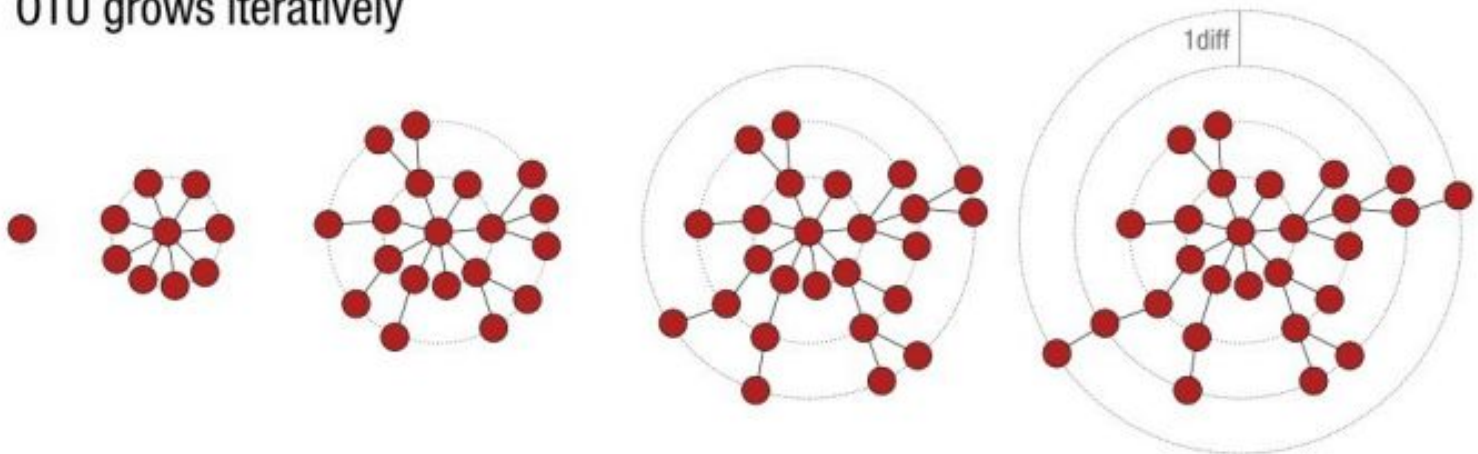
Swarm clustering method

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



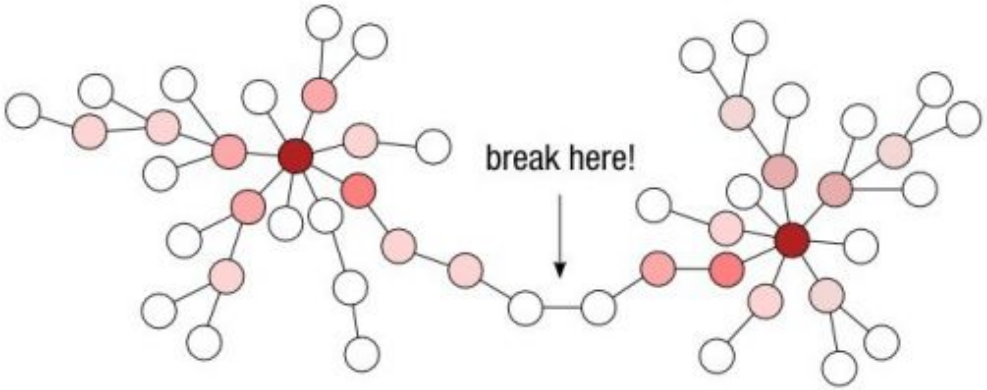
initial seed (randomly picked from amplicon dataset)

no more closely related amplicons, the process stops

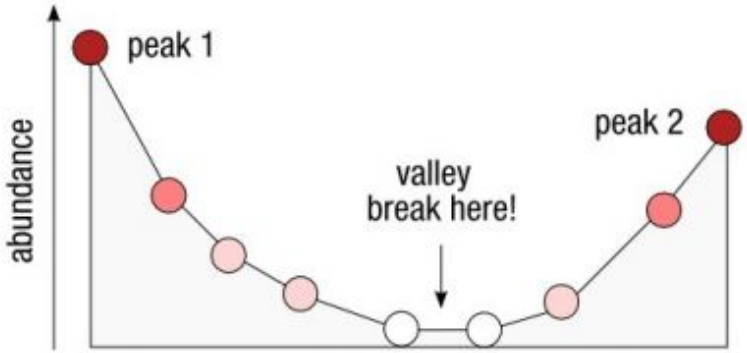
Frederic Mahé communication

Swarm clustering method

breaking phase



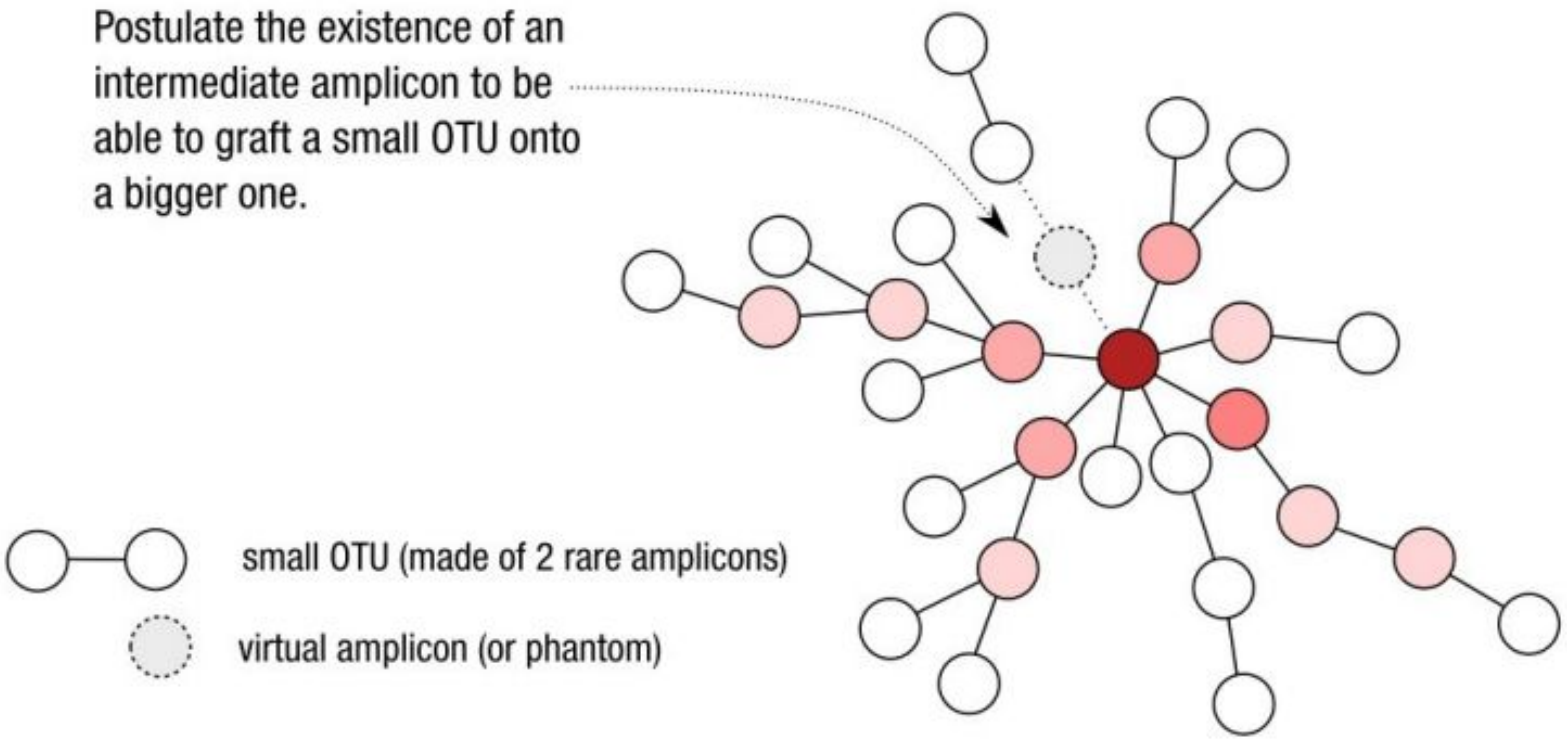
Take into account the abundance of amplicons to produce higher-resolution clusters.



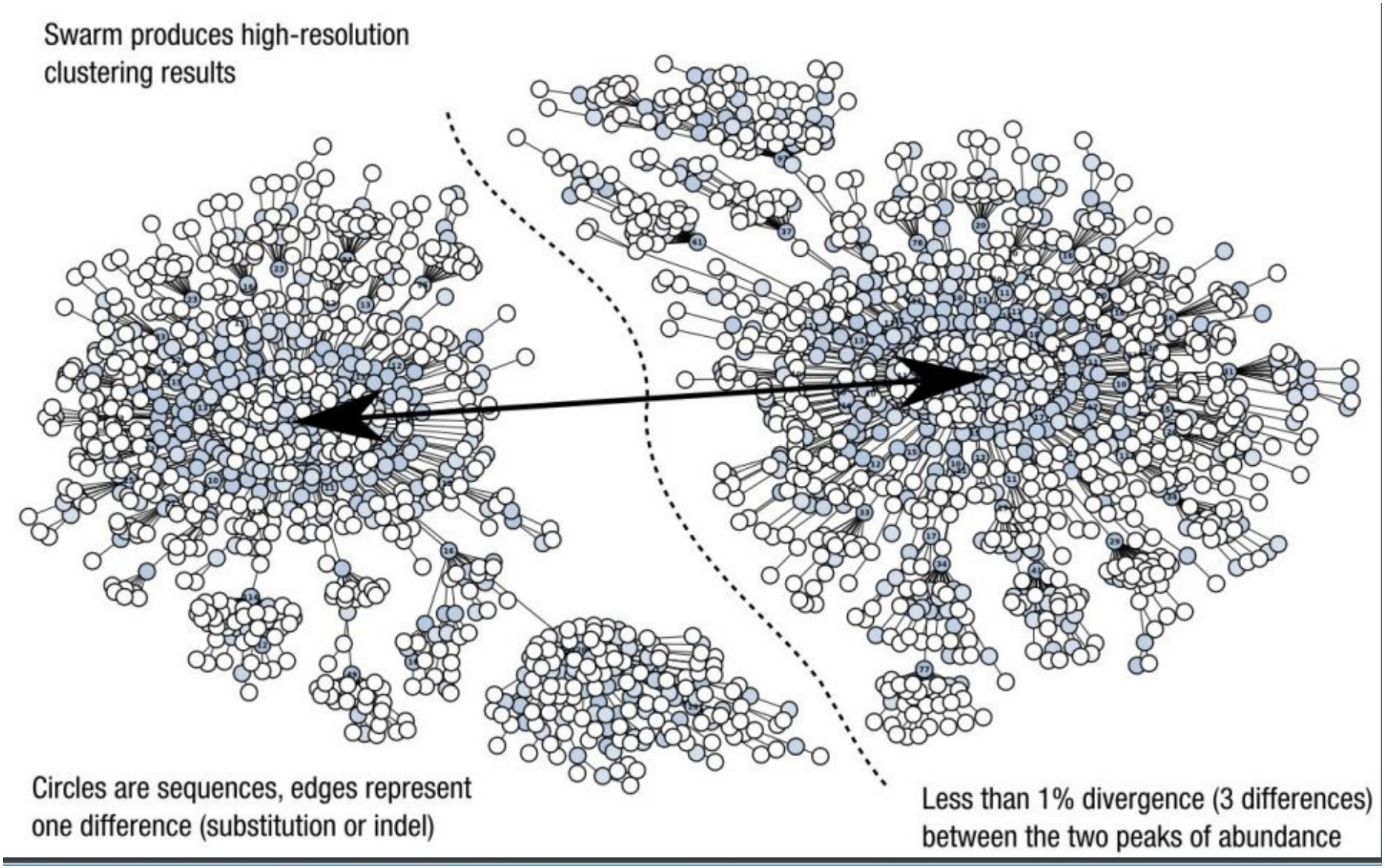
Assuming that original sequences are more abundant than erroneous copies.

Swarm clustering method

grafting phase



Frederic Mahé communication



Frederic Mahé communication

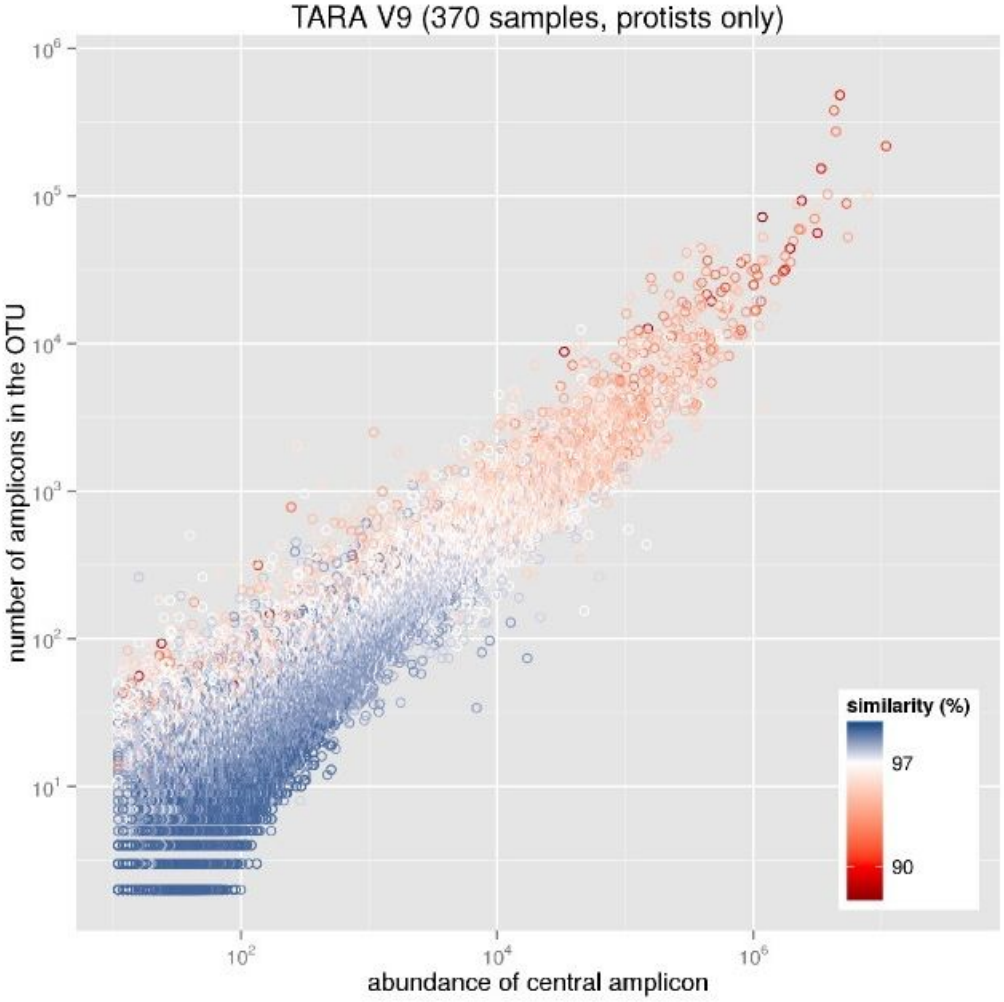
Swarm 2.0 is a highly scalable denoising-clustering method

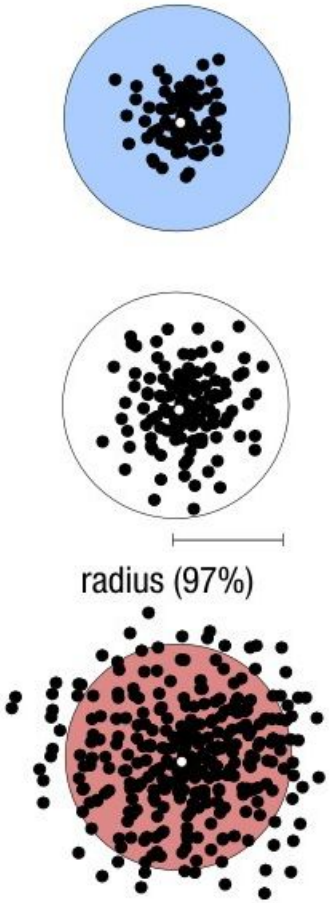
28,275 samples
2.3 billion reads

swarm: 5 hours
usearch: >150 days

Frederic Mahé communication

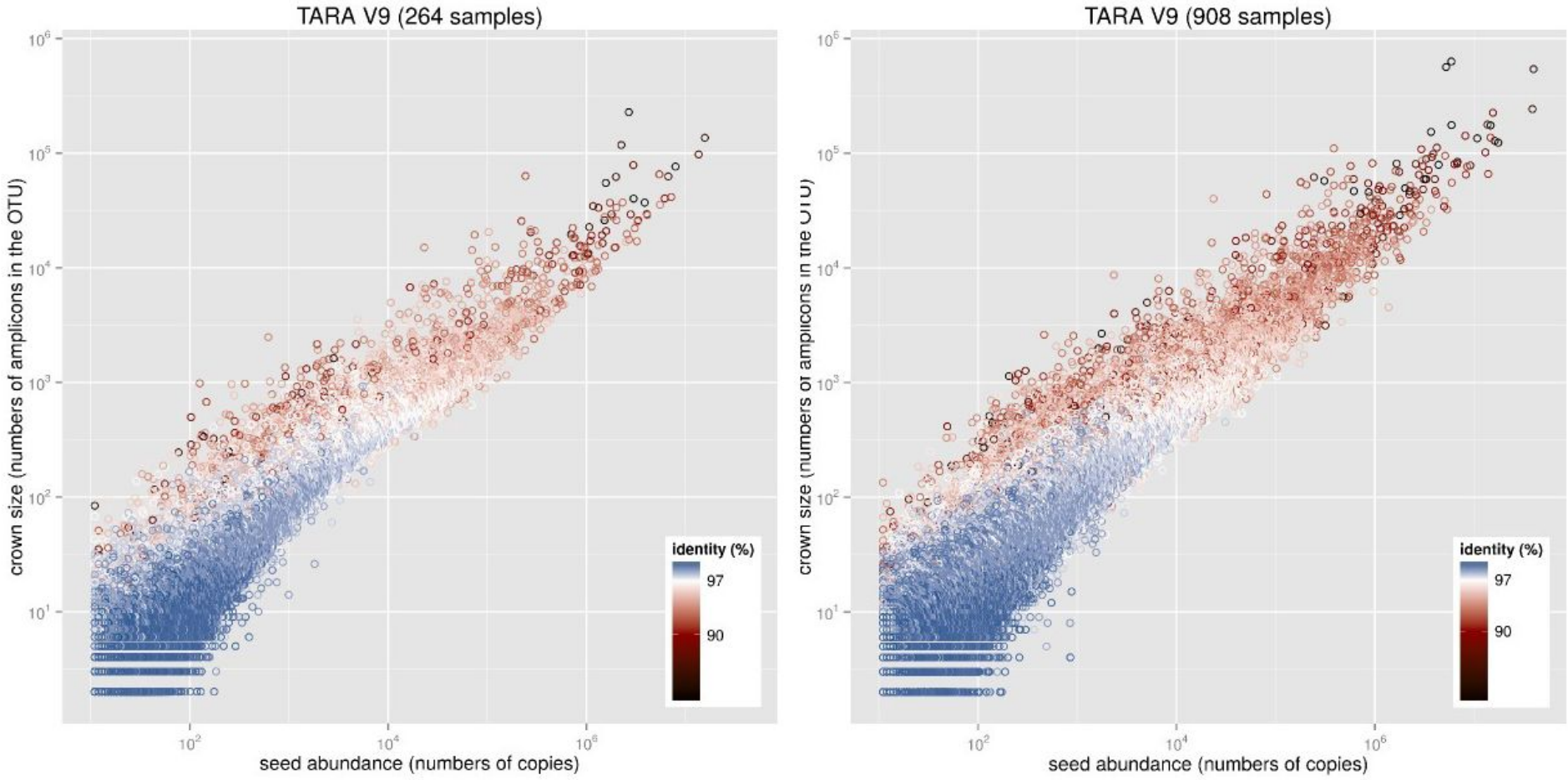
What if we'd used a fix 97%-clustering threshold?





Frederic Mahé communication

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



clusters produced with swarm using d = 1

Frederic Mahé communication

3. 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)

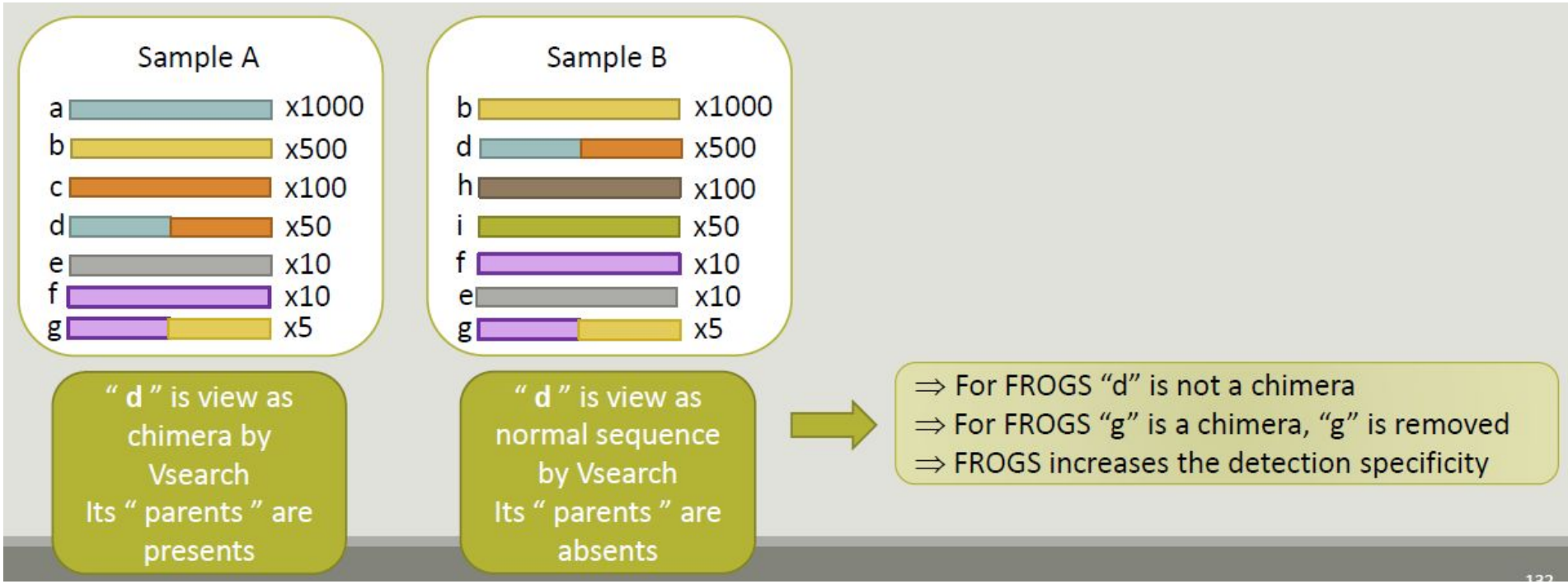
3. Remove chimera

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

3. Remove chimera

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



vsearch: open-source alternative for usearch

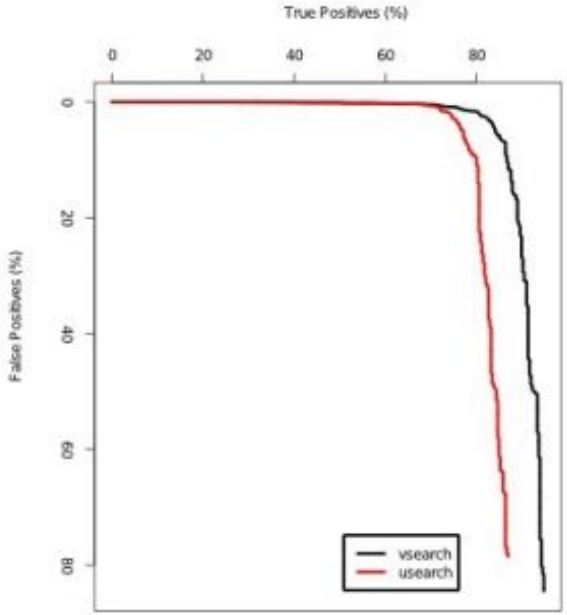
clustering, chimera detection, dereplication, searching, sorting, masking and shuffling

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

4. 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.

5. 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**".

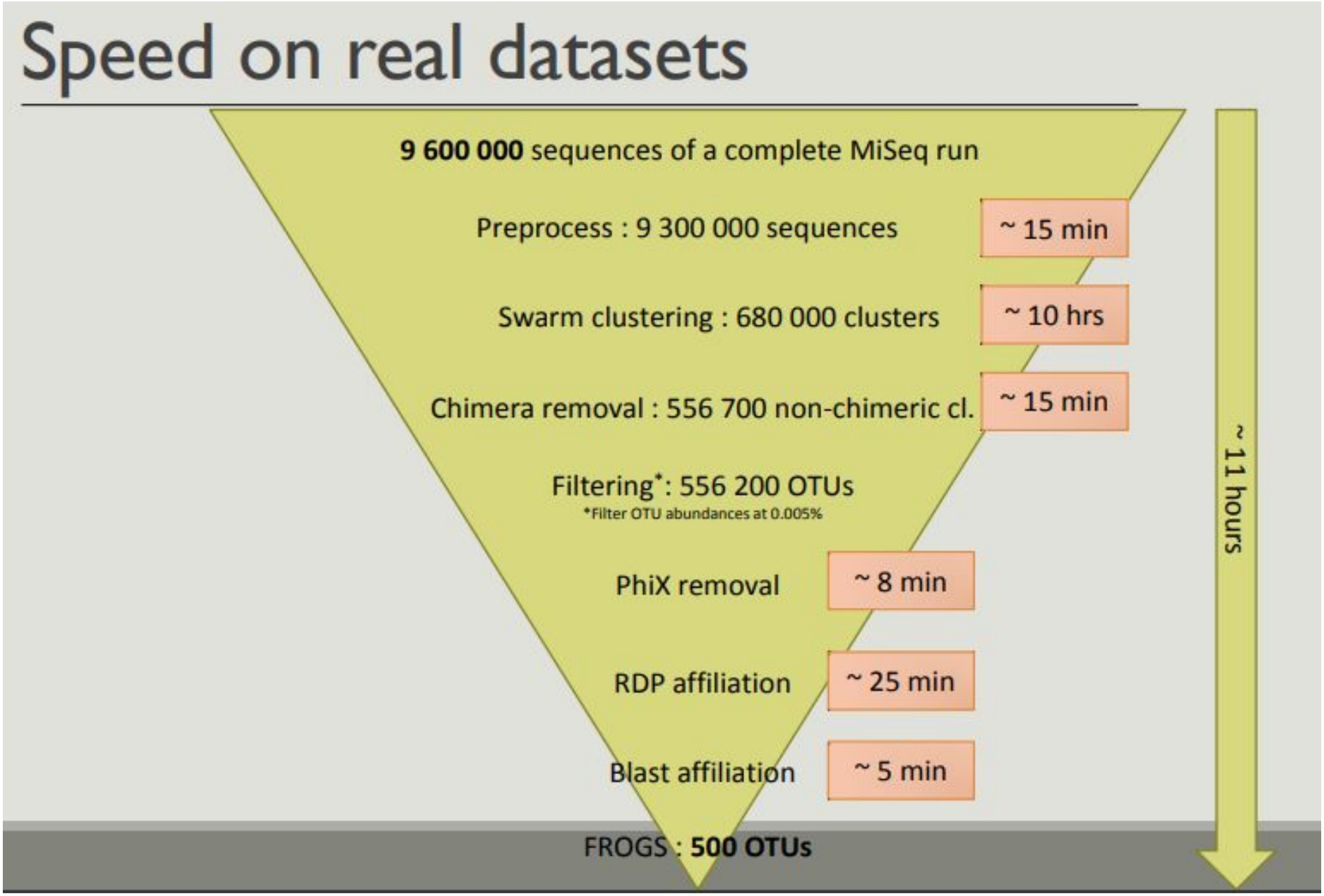
V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | 16S unknown species
V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | 16S Butyrivibrio fibrisolvens
V3 – V4	Bacteria | Firmicutes | Clostridia | Clostridiales | 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*.

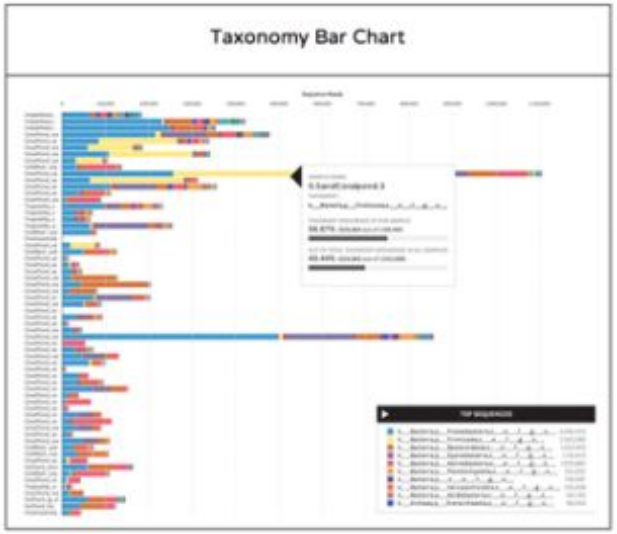
6. Affiliation Stats

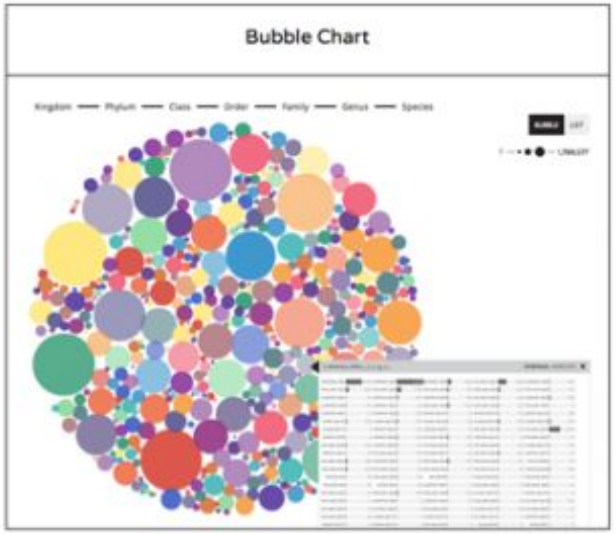
This step computes some statistics from the analysis and generates a report of the OTUs/taxonomy found.

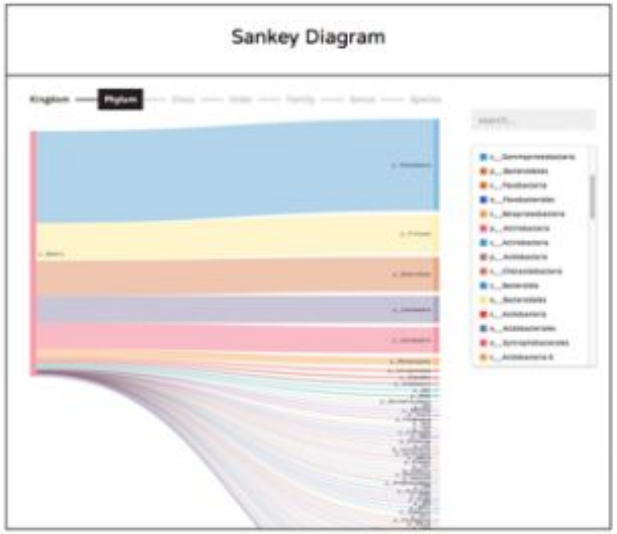
Speed on real datasets

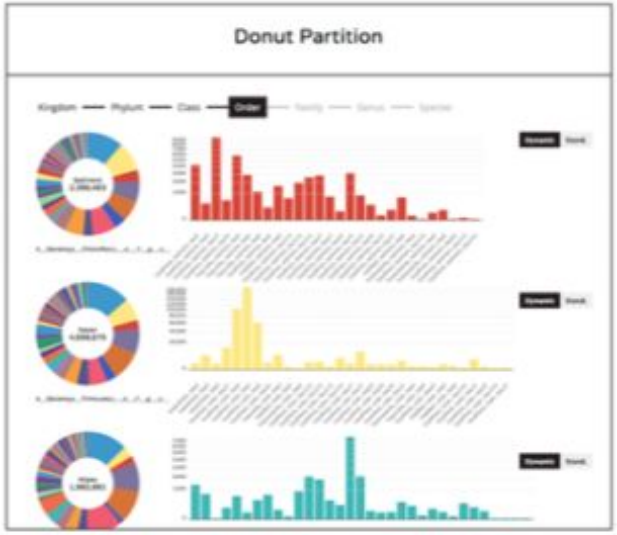


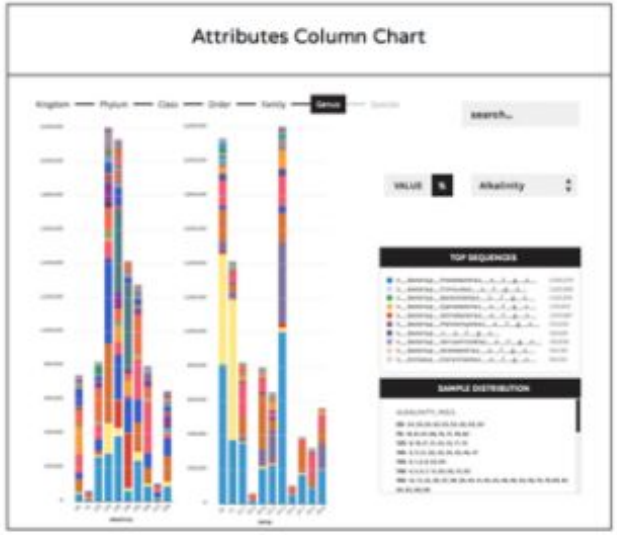
Practice 2: Visualizing and plotting all sample results with Phinch









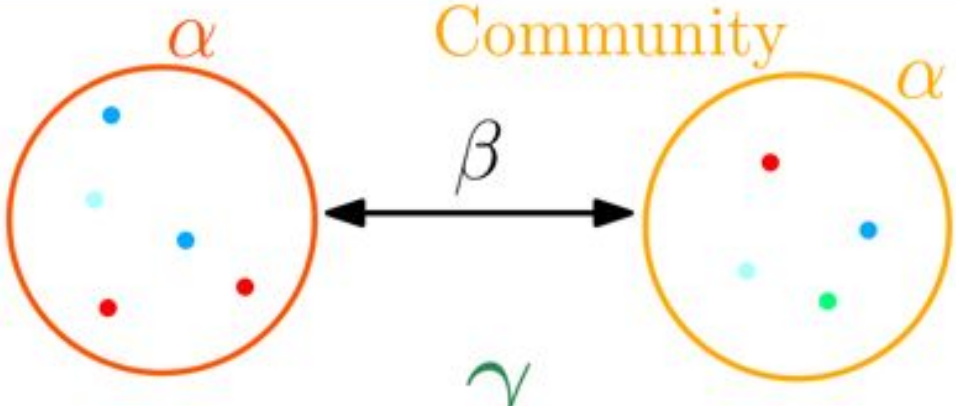


https://doi.org/10.1101/009944

Practice 3: Handling and visualisation of OTU table using PhyloSeq

Community analysis

- diversity indices and metrics
- metrics of diversity



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

Formateurs itrop / South Green
+ Collaborateurs UMR QualiSud CIRAD

- Alexis Dereeper
- Julie Orjuela-Bouniol
- Florentin Constancias



Merci pour votre attention !