Analyse de données métagénomiques 16S - FROGS

Module 20

Olivier Rué

MaIAGE - Migale

September 12, 2023

FROGS

FROGS team

- FROGS is a INRAE development project



Vincent Darbot Maria Bernard Olivier Ré

Developers


Lucas Auer Laurent Cauquil

Biology experts

Patrice Déhais

Galaxy support

Mahendra Mariadassou

Statistical expert

Géraldine Pascal

Coordinator

FROGS articles

Bioinformatics, 34(8), 2018, 1287–1294
doi: 10.1093/bioinformatics/btx791
Advance Access Publication Date: 7 December 2017
Original Paper

Sequence analysis

FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié[1,†], Lucas Auer[2,†], Maria Bernard[3], Mahendra Mariadassou[4], Laurent Cauquil[5], Katia Vidal[5], Sarah Maman[5], Guillermina Hernandez-Raquet[6], Sylvie Combes[5] and Géraldine Pascal[5,*]

[1]Bioinformatics platform Toulouse Midi-Pyrenees, MIAT, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex, France, [2]INRA, UMR 1136, Université de Lorraine, INRA-Nancy, 54280, Champenoux, France, [3]GABI, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France, [4]MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France, [5]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France and [6]Laboratoire d'ingénierie des Systèmes Biologiques et des Procédés-LISBP, Université de Toulouse, INSA, INRA, CNRS, Toulouse, France

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Bonnie Berger

Received on May 10, 2017; revised on December 1, 2017; editorial decision on December 4, 2017; accepted on December 5, 2017

Abstract

Motivation: Metagenomics leads to major advances in microbial ecology and biologists need user friendly tools to analyze their data on their own.
Results: This Galaxy-supported pipeline, called FROGS, is designed to analyze large sets of amplicon sequences and produce abundance tables of Operational Taxonomic Units (OTUs) and their taxonomic affiliation. The clustering uses Swarm. The chimera removal uses VSEARCH, combined with original cross-sample validation. The taxonomic affiliation returns an innovative multi-affiliation output to highlight databases conflicts and uncertainties. Statistical results and numerous graphical illustrations are produced along the way to monitor the pipeline. FROGS was tested for the detection and quantification of OTUs on real and *in silico* datasets and proved to be rapid, robust and highly sensitive. It compares favorably with the widespread mothur, UPARSE and QIIME.
Availability and implementation: Source code and instructions for installation: https://github.com/geraldinepascal/FROGS.git. A companion website: http://frogs.toulouse.inra.fr.
Contact: geraldine.pascal@inra.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The expansion of high-throughput sequencing of rRNA amplicons has opened new horizons for the study of microbial communities. By making it possible to study all micro-organisms from a given environment without the need to cultivate them, metagenomics has led to major advances in many fields of microbial ecology, from the study of the impact of microbiota on human and animal pathologies (Hess *et al.*, 2011; Hooper *et al.*, 2012; Jovel *et al.*, 2016) to the study of biodiversity in environmental ecosystems and the search for biomarkers of pollution (Andres and Bertin, 2016; de Vargas *et al.*, 2015). Determining the composition of a microbial ecosystem, at low cost and great depth, is still largely based on the amplification and sequencing of biodiversity marker genes, also called amplicons, such as rRNA genes and ITS. The clustering of sequences into

Briefings in Bioinformatics, 22(6), 2021, 1–6
https://doi.org/10.1093/bib/bbab318
Problem Solving Protocol

FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers

Maria Bernard†, Olivier Rué†, Mahendra Mariadassou and Géraldine Pascal

Corresponding author: Geraldine Pascal, GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France. Tel.: +33 (0)5 61 28 51 05; E-mail: geraldine.pascal@inrae.fr
†Maria Bernard and Olivier Rué are joint first authors.

Abstract

Fungi are present in all environments. They fulfil important ecological functions and play a crucial role in the food industry. Their accurate characterization is thus indispensable, particularly through metabarcoding. The most frequently used markers to monitor fungi are ITSs. These markers are the best documented in public databases but have one main weakness: polymerase chain reaction amplification may produce non-overlapping reads in a significant fraction of the fungi. When these reads are filtered out, traditional metabarcoding pipelines lose part of the information and consequently produce biased pictures of the composition and structure of the environment under study. We developed a solution that enables processing of the entire set of reads including both overlapping and non-overlapping, thus providing a more accurate picture of fungal communities. Our comparative tests using simulated and real data demonstrated the effectiveness of our solution, which can be used by both experts and non-specialists on a command line or through the Galaxy-based web interface.

Key words: fungi; ITS; metabarcoding; workflow; amplicon; metagenomics

Introduction

Using amplicon sequencing to describe the microbial composition of an environment is a time saving and cost-effective strategy and can be used even for very large-scale surveys [1]. Most studies currently focus on the bacterial fraction of microbial communities but the fungal fraction is equally important, as fungi are ubiquitous and provide several ecosystem services [2]. Unfortunately, studying the fungal fraction using metabarcoding has its own challenges. Indeed, in fungi, there is no equivalent of the 16S rRNA gene, which is widely used and highly suitable for bacteria. The best candidates are internal transcribed spacers (ITS), but these are more difficult to manipulate. The main problem with ITS is size polymorphism, with a size range of 361–1475 bases in UNITE 7.1 [3] (unlike 16S where 95% of the sequences have a length between 1205 and 1556 bases). Most studies describing ITS data analyses process either (i) paired-end reads but filter out non-overlapping, non-mergeable reads, thus systematically discarding taxa with longer ITS, or (ii) single-end reads, thus limiting taxonomic resolution and losing the benefit of information contained in longer sequences [4, 5].

Maria Bernard is a bioinformatics engineer. She is a member of a platform team conducting NGS sequence analysis and designing software. She specializes in workflow development in particular for metabarcoding analysis.
Olivier Rué is a bioinformatics engineer. He is in charge of data analysis at the Migale bioinformatics facility. He specializes in the analysis of metabarcoding and metagenomics data.
Mahendra Mariadassou has a PhD in statistics. He is involved in the development of new statistical methods and tools for metabarcoding analysis.
Géraldine Pascal has a PhD in bioinformatics and coordinates the FROGS project. She is currently involved in designing solutions for long read problems, workflow development and metagenomics analysis.
Submitted: 19 April 2021; **Received (in revised form):** 19 July 2021

Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, L[illegible]atia Vidal, Sarah Maman, Guillermina HernandezRaquet, Sylvie Combes, Géraldine Pascal. "FROGS: Find, Rapidly, OTUs with Galaxy Solution." Bioinformatics, , Volume 34

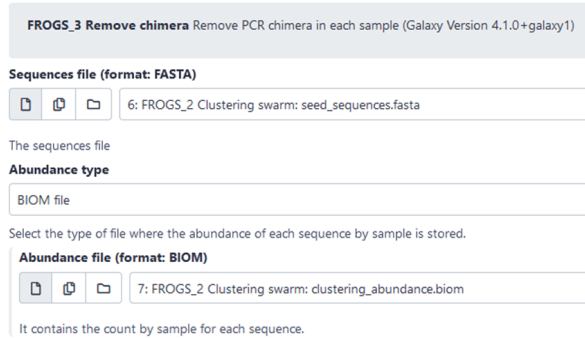
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

How to use FROGS

- Command line

```
remove_chimera.py
  --input-biom clustering.biom \
  --input-fasta clustering.fasta \
  --non-chimera remove_chimera.fasta \
  --out-abundance remove_chimera.biom \
  --summary remove_chimera.html
```

- Galaxy instances via web



- Migale - https://galaxy.migale.inrae.fr/
- Genotoul - https://vm-galaxy-prod.toulouse.inrae.fr/
- IFB - https://metabarcoding.usegalaxy.fr/

FROGS docs and help

- Website: https://frogs.toulouse.inrae.fr
- Github: https://github.com/geraldinepascal/FROGS.git
- Newsletter: subscription request at frogs-support@inrae.fr
- ? Need help
  - frogs-support@inrae.fr for generic questions
  - help-migale@inrae.fr for bugs/quotas/errors with Galaxy Migale instance

TP1: Introduction to Galaxy

Sequencing data

FASTQ format

```
@ST-E00114:1342:HHMGVCCX2:1:1101:3123:2012 1:N:0:TCCGGAGA+TCAGAGCC
CTTGGTCATTTAGAG
+
***<<*AEF???***
@ST-E00114:1342:HHMGVCCX2:1:1101:11556:2030 1:N:0:TCCGGAGA+TCAGAGCC
CATTGGCCATATCAT
+
AAAE??<<*???***
```

Meaning

```
@Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
@Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
```

Quality score encoding

```
                          33                        59   64       73                            104                   126
                           |                         |    |        |                              |                     |
                ASCII code !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
                           |                         |    |        |                              |                     |
Sanger          Phred+33   0.......................26...31.......40                              |                     |
Solexa          Solexa+64  |                        -5....0........9.............................40                     |
Illumina 1.3+   Phred+64   |                         |    0........9.............................40                     |
Illumina 1.5+   Phred+64   |                         |    |  3.....9...............................41                     |
Illumina 1.8+   Phred+33   0.2......................26...31.......41                              |                     |
PacBio          Phred+33   0..................20........30........40........50.........................................93
```

Quality score

Measure of the quality of the identification of the nucleobases generated by automated DNA sequencing

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

FASTQ compression

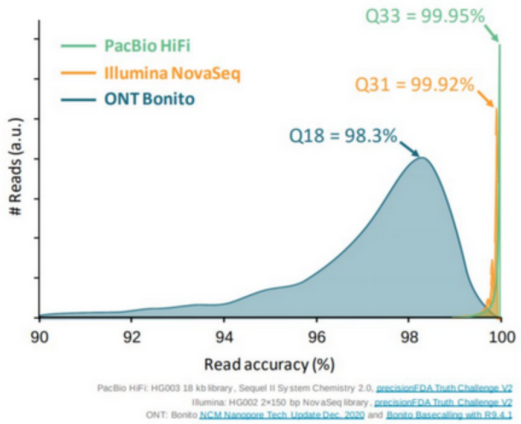
- Compression is essential to deal with FASTQ files (reduce disk storage)
- *extension*: `file.fastq.gz`
- Tools are (almost all) able to deal with compressed files 👍

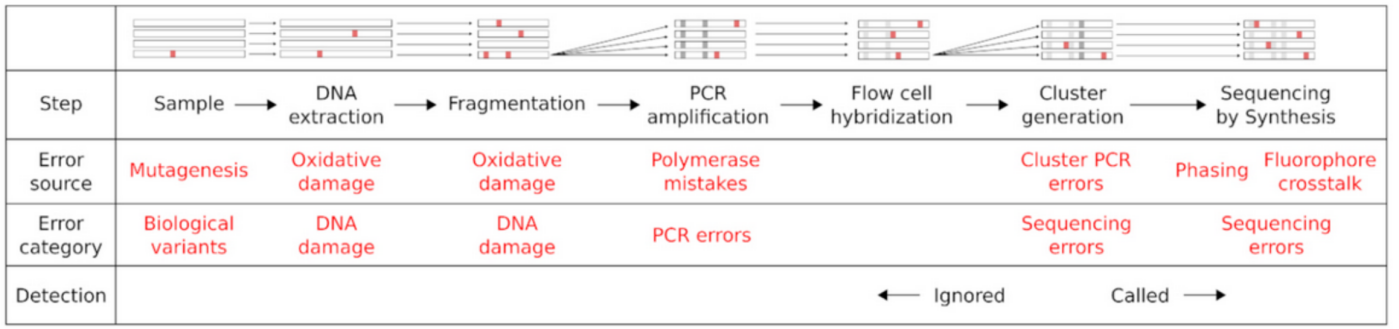
Quality control

Quality control

- One of the most easy step in bioinformatics …
- … but one of the most important
- check if everything is ok
- Indicates if/how to clean reads
- Shows possible sequencing problems
- The results must be interpreted in relation to what has been sequenced

Reads are not perfect





Sequencing error profiles of Illumina sequencing instruments [3]

Why QC'ing your reads?

Try to answer to (not always) simple questions:

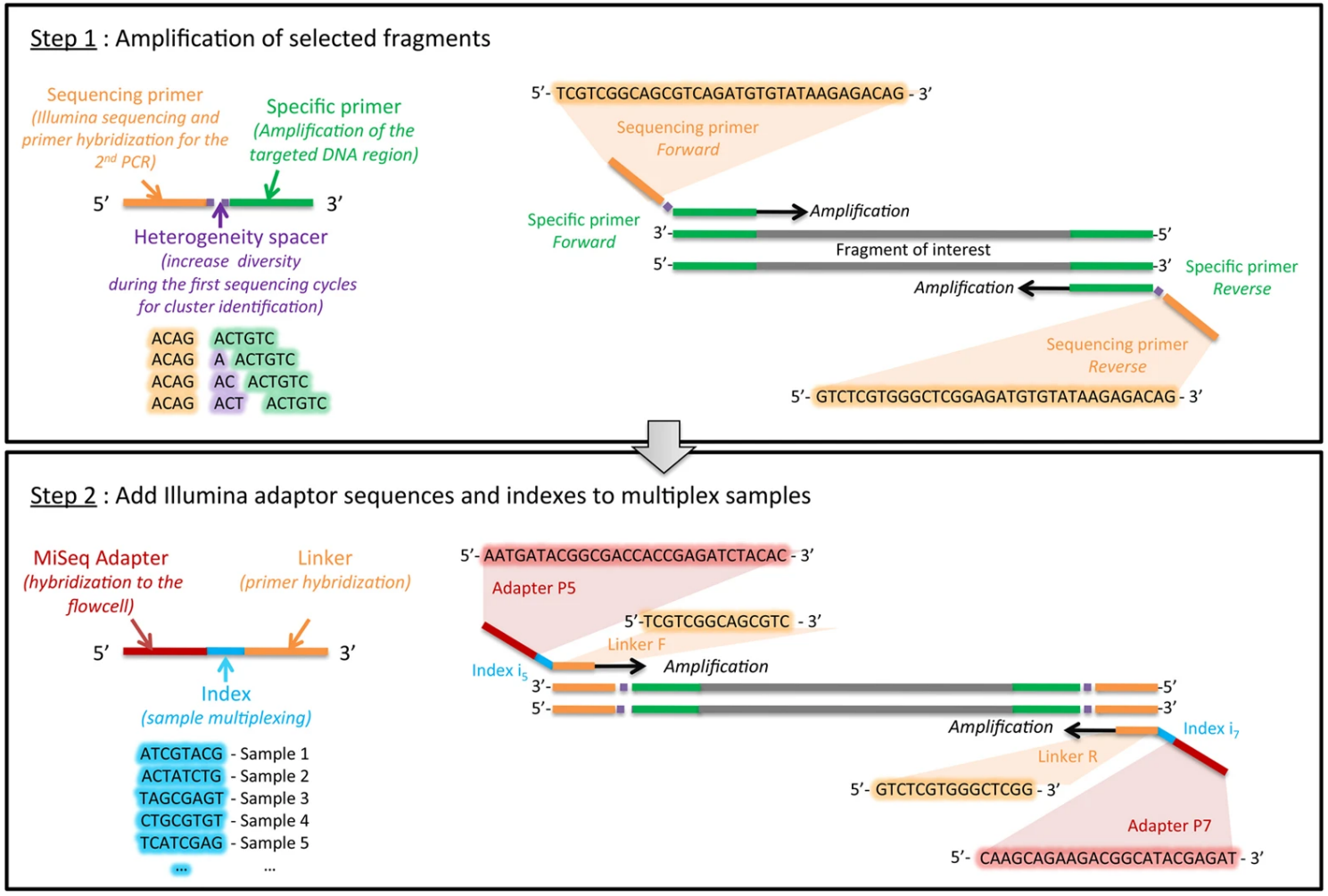
- Are data conform to the expected level of performance?
  - Size / Number of reads / Quality
- Residual presence of adapters or indexes?
- (Un)expected techincal biases?
- (Un)expected biological biases?

> **Warning**
>
> QC without context leads to misinterpretation!

TP2: Quality control

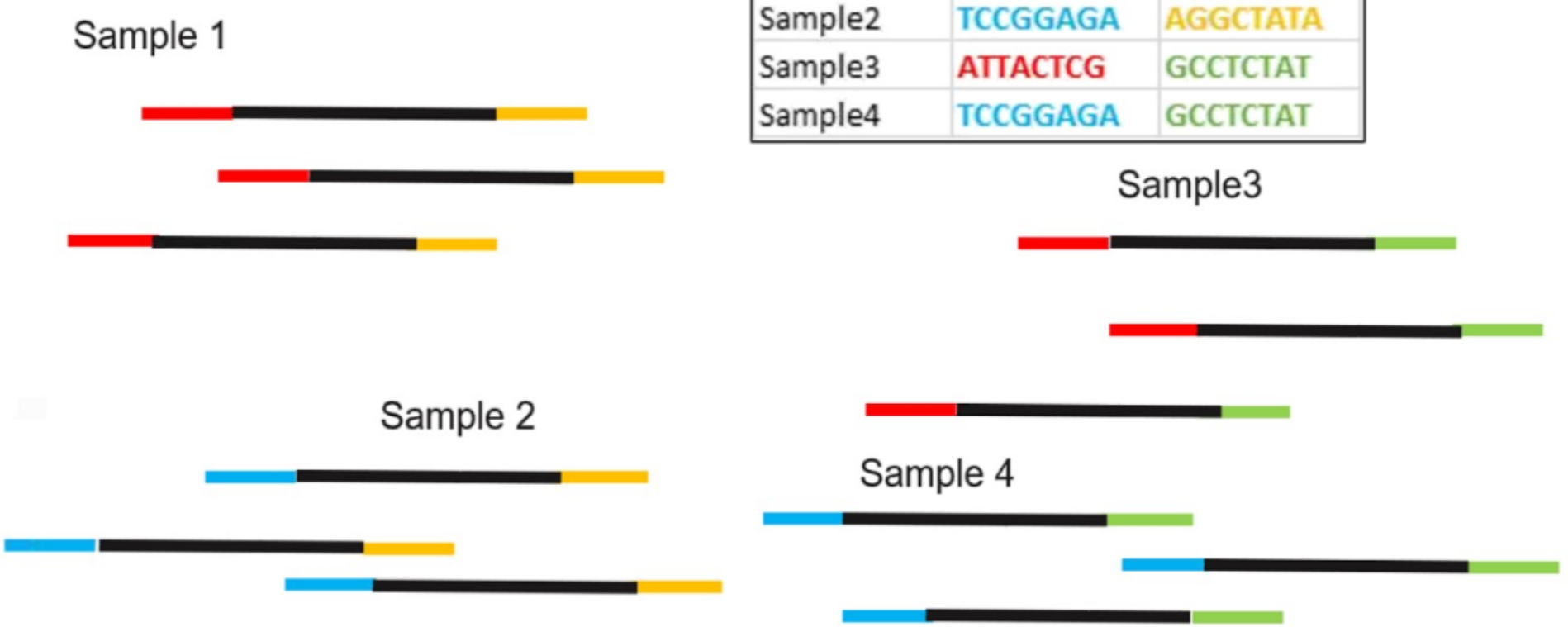
Demultiplexing

Multiplexing principle



Demultiplexing by bioinformatics

[Data]		
Sample_ID	index	index2
Sample1	ATTACTCG	AGGCTATA
Sample2	TCCGGAGA	AGGCTATA
Sample3	ATTACTCG	GCCTCTAT
Sample4	TCCGGAGA	GCCTCTAT



Credits: https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/processing-raw-scrna-seq-data.html

Demultiplexing tool

- Assign each read to FASTQ files depending on barcode found
- BARCODE FILE is expected to be tabular:
  - first column corresponds to the sample name (unique, without space)
  - second to the forward sequence barcode used (None if only reverse barcode)
  - optional third is the reverse sequence barcode (optional)

TP DEMULTIPLEXING

Advices

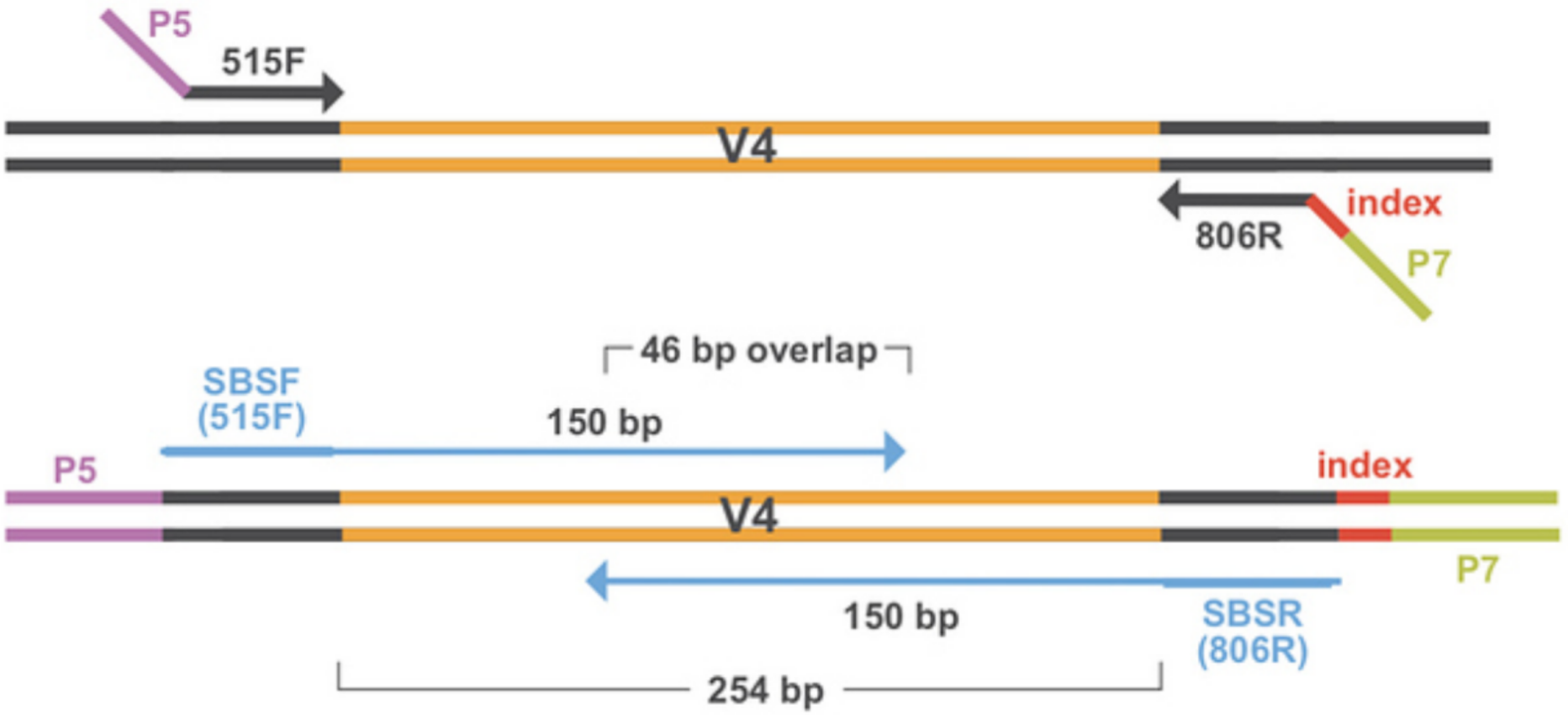
- Do not forget to indicate barcode sequence as they are in the fastq sequence file, especially if you have data multiplexed via the reverse strand.
- For the mismatch threshold, we advised you to let the threshold to 0, and if you are not satisfied by the result, try with 1. The number of mismatch depends on the length of the barcode, but often those sequences are very short so 1 mismatch is already more than the sequencing error rate.
- If you have different barcode lengths, you must demultiplex your data in different times beginning by the longest barcode set and used the “unmatched” or “ambiguous” sequence with smaller barcode and so on.

FROGS preprocess

What FROGS preprocess does?

- Merging of R1 and R2 reads with vsearch [4], flash [5] or pear [6] (only in command line)
- Deletes sequences without good primers
- Finds and removes adapter sequences with cutadapt
- Deletes sequence with not expected lengths
- Deletes sequences with ambiguous bases (N)
- Dereplication
- *removing homopolymers (size = 8) for 454 data*
- *quality filter for 454 data*

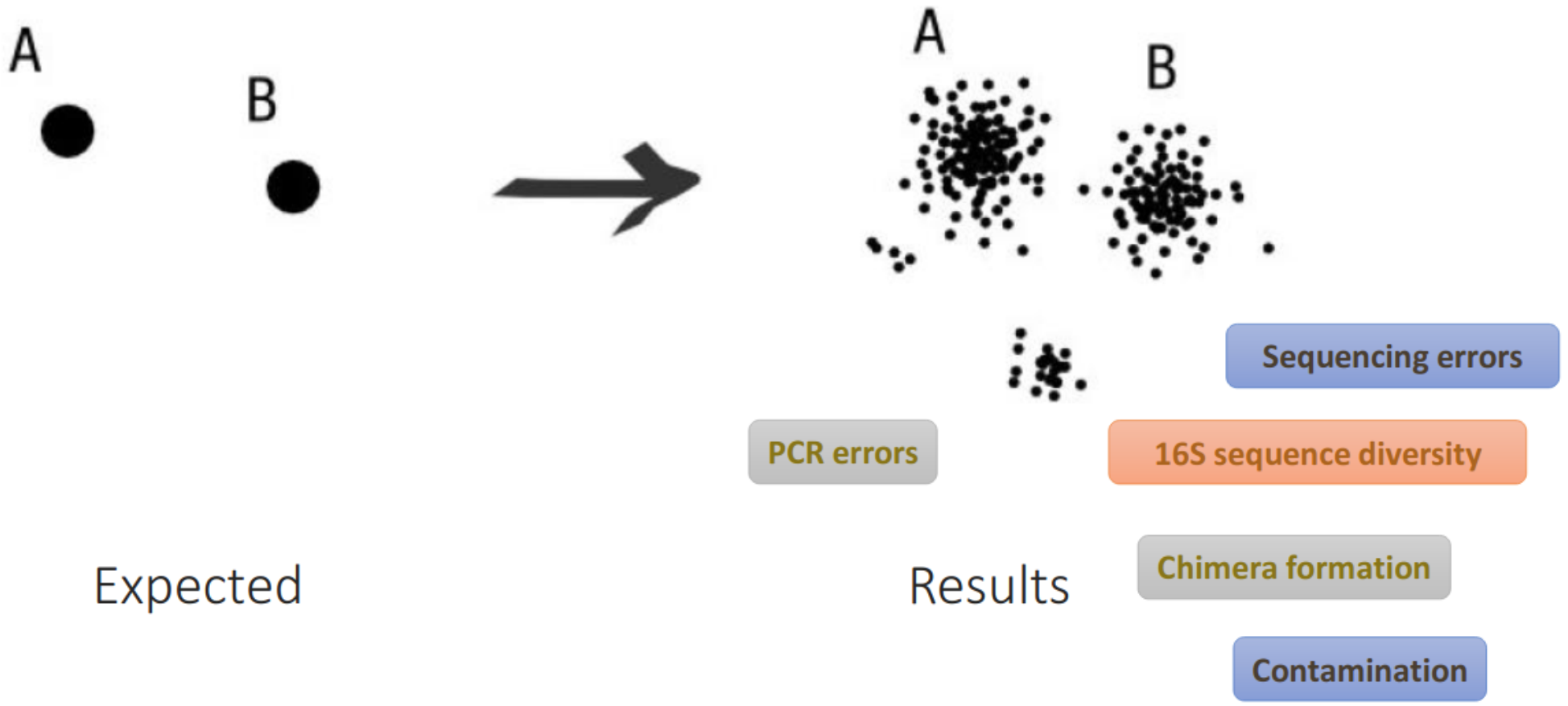
Merging of paired-end reads



TP FROGS preprocess

Clustering

Sequencing data are noised



How to deal with these noised sequences?

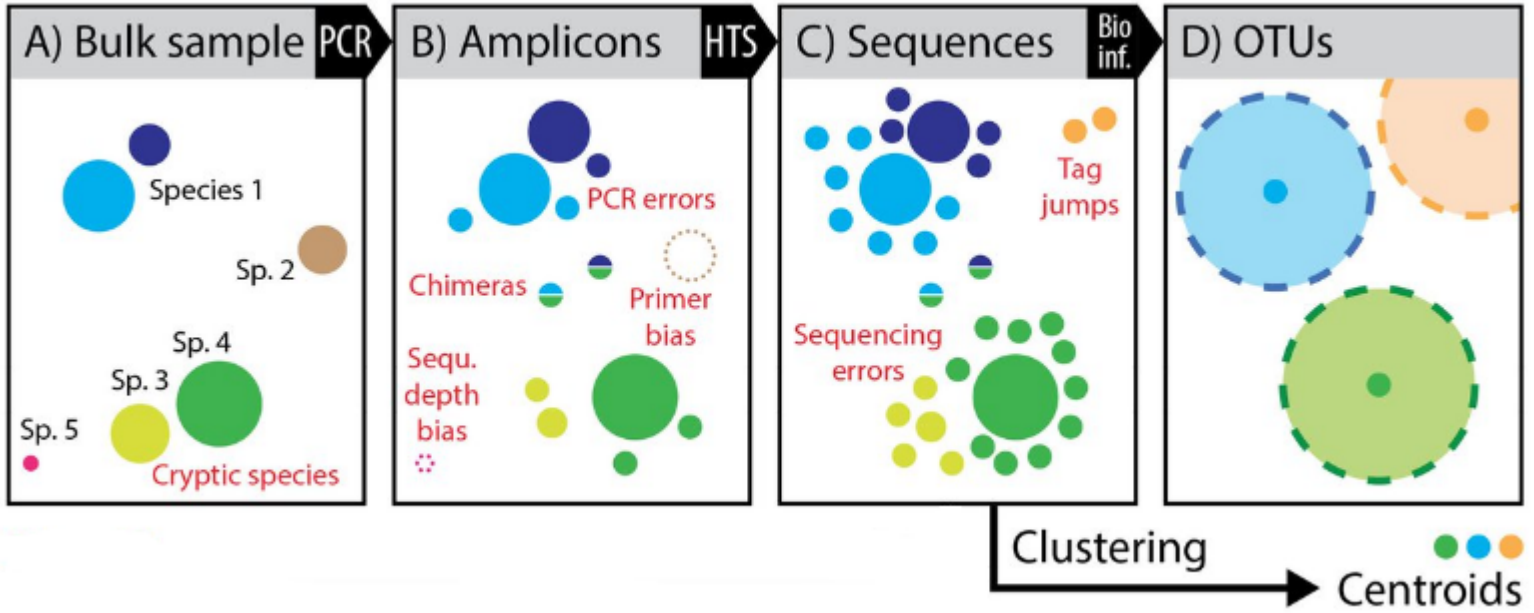
- Comparison all against all
  - Very accurate
  - Requires a lot of memory and/or time
- Clustering
  - closed-reference / open-reference
  - de novo
- Denoising

Vocabulary

- A lot of terms for features built by softwares
  - OTUs, zOTUS, ASVs, ESVs…
- A recent review establishes the vocabulary [7]
  - OTUs / ASVs / swarm clusters
- *ASVs are identical denoised reads with as few as 1 base pair difference between variants, representing an inference of the biological sequences prior to amplification and sequencing errors*
- OTUs are formed with a % threshold clustering
- Swarm clusters are a third feature type

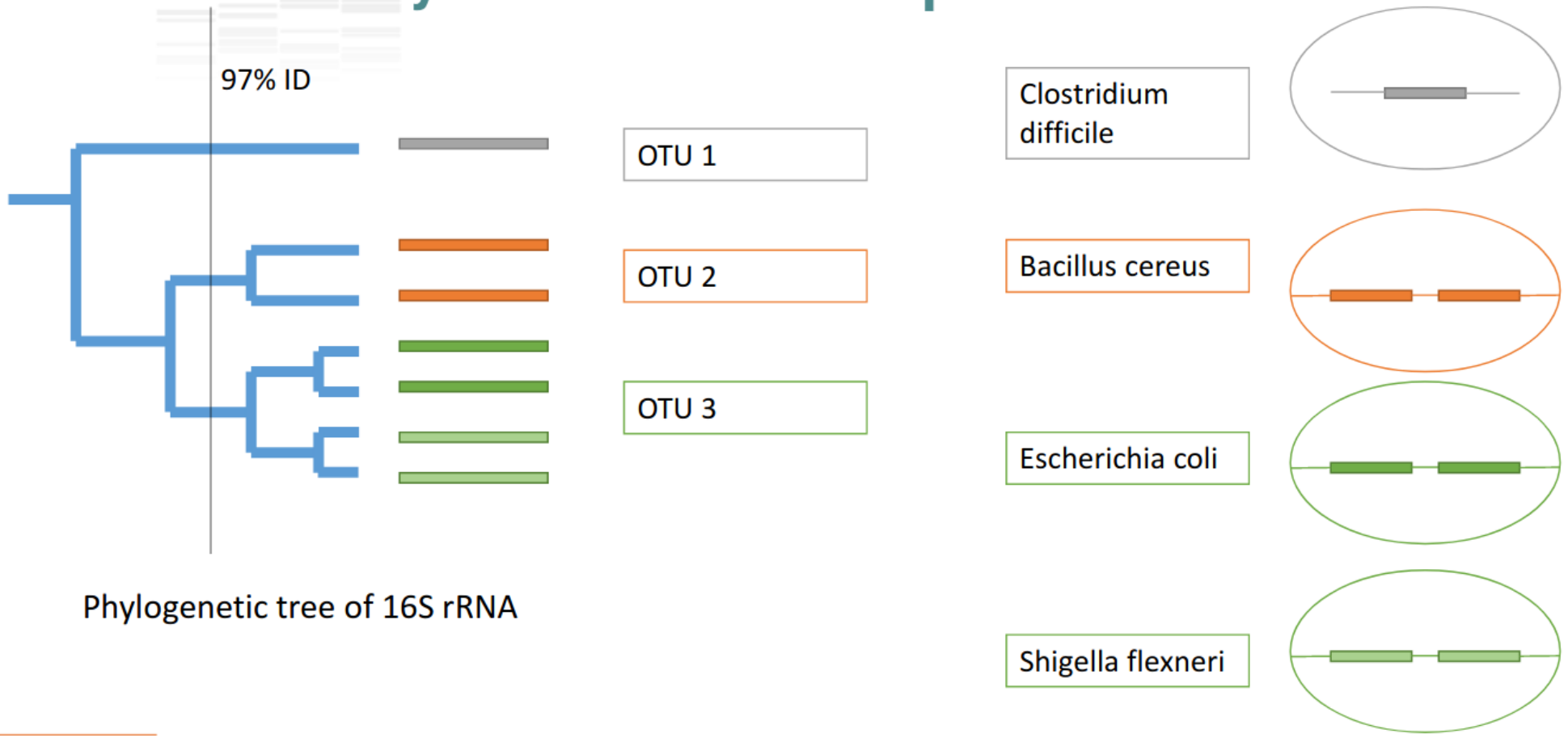
OTU paradigm

- Operational Taxonomic Unit



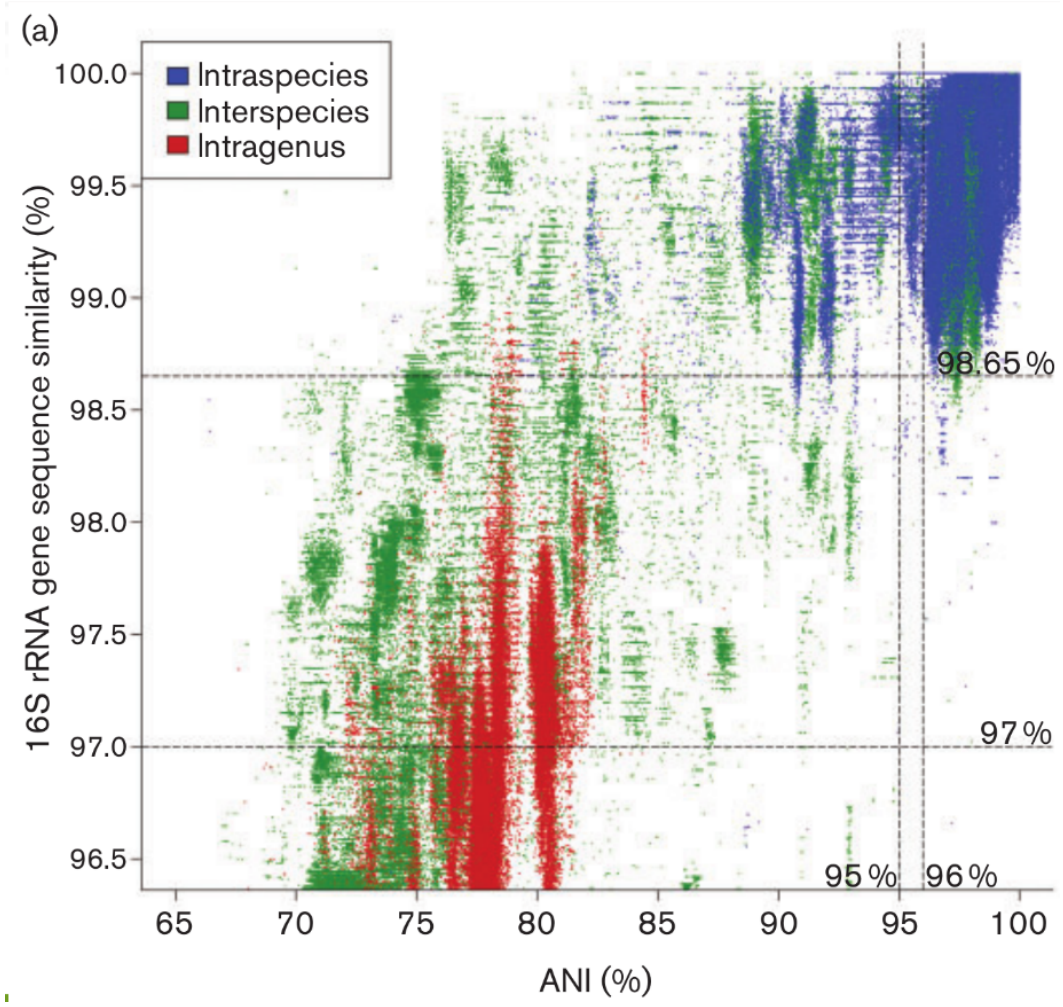
Operational Taxonomic Units

OTUs: a Proxy for « Bacterial Species »

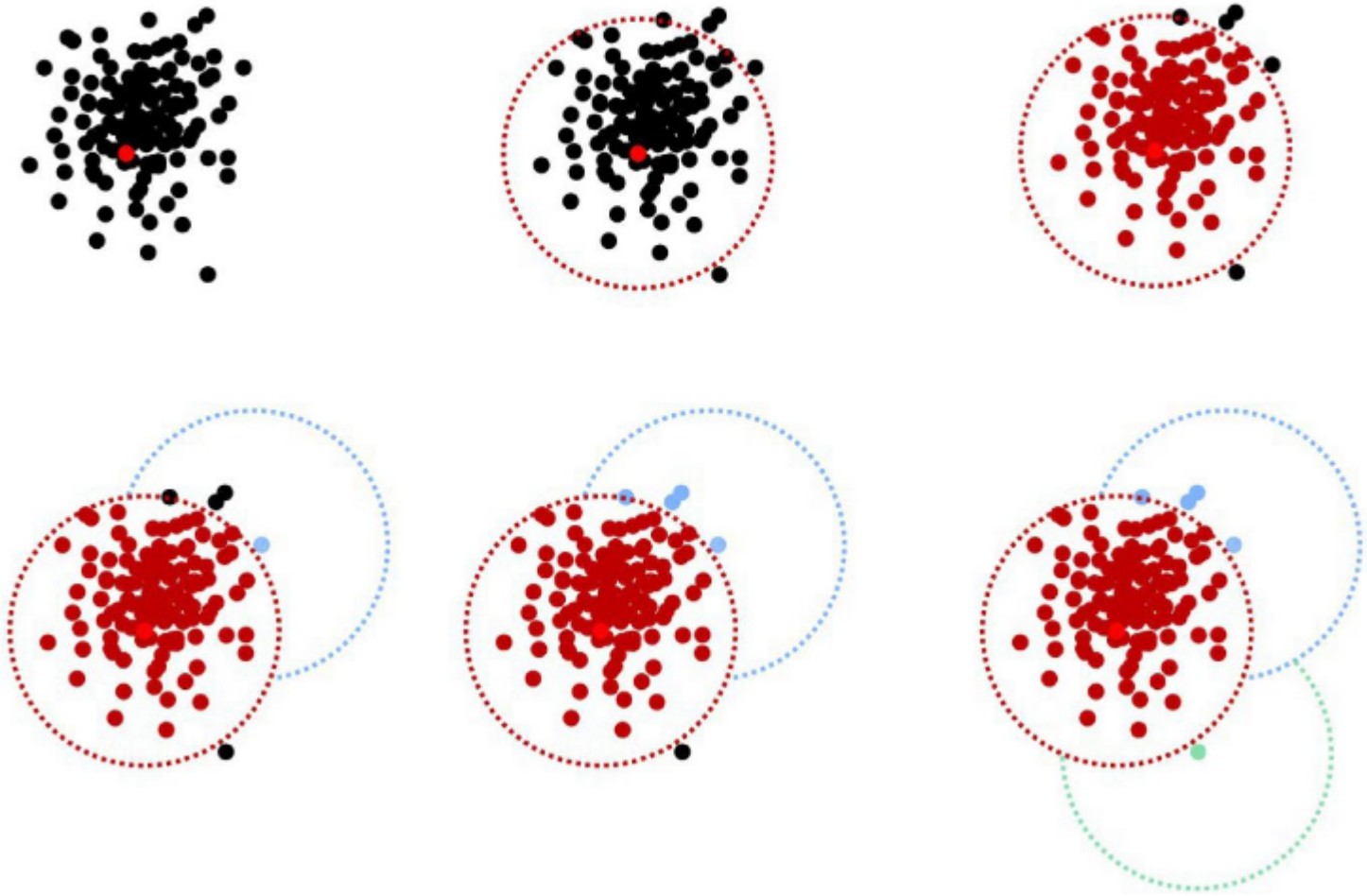


Phylogenetic tree of 16S rRNA

Operational Taxonomic Units

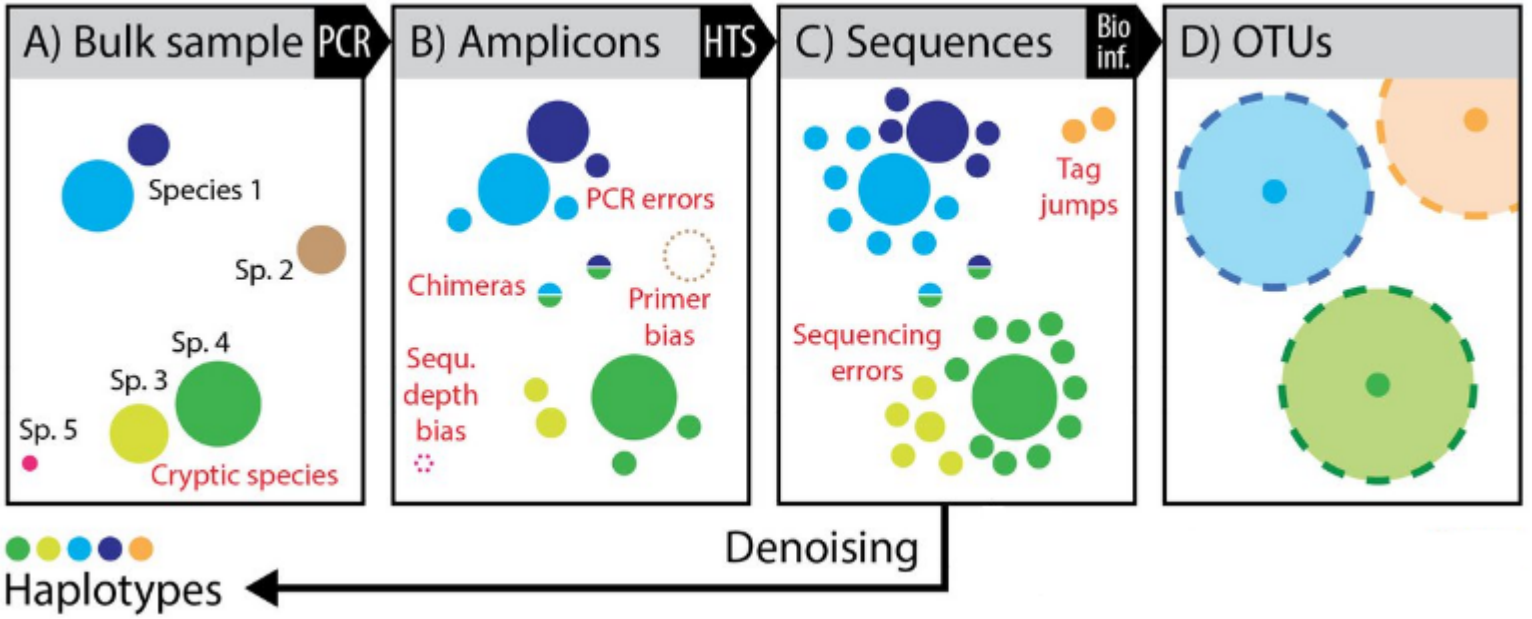


Operational Taxonomic Units



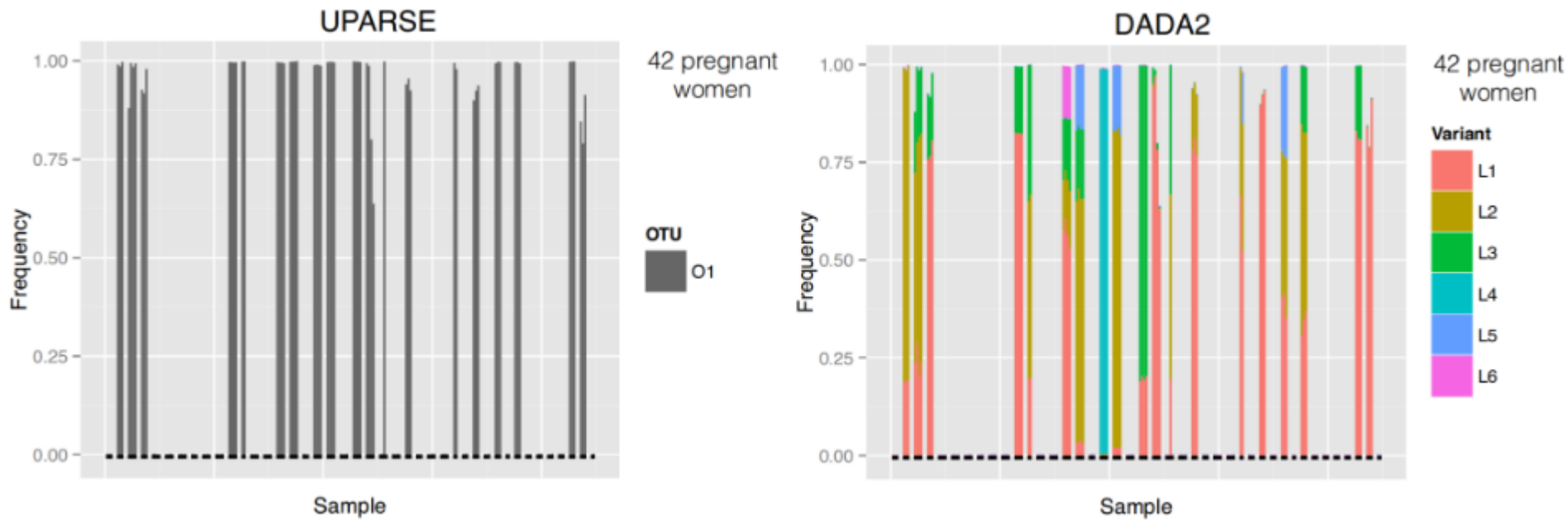
ASV paradigm

- Amplicon Sequence Variants



ASV are inferred by a de novo process in which biological sequences are discriminated from errors on the basis of the expectation that biological sequences are more likely to be repeatedly observed than are error-cont uences

ASV resolution



- ASV resolution changes the composition for these samples

Credits: https://benjjneb.github.io/dada2/SMBS_DADA2.pdf

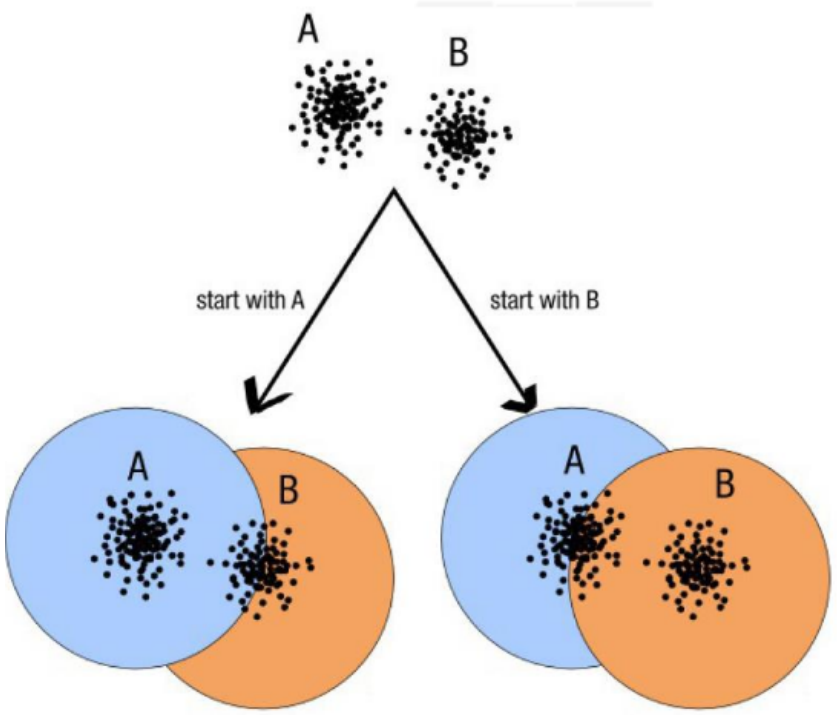
Swarm

Swarm [8] is a notably different sequence clustering approach, which, while technically a clustering algorithm, may also be considered a denoising method when using the fastidious method with *d=1*. It relies on the maximum number of differences between reads (local linking threshold) and forms clusters that are resilient to input-order changes, thus creating stable, high-resolution features (herein referred to as swarm-clusters). When using the fastidious method with *d=1*, swarm aims to produce clusters centered around real biological sequences, where clusters represent sequence variants.

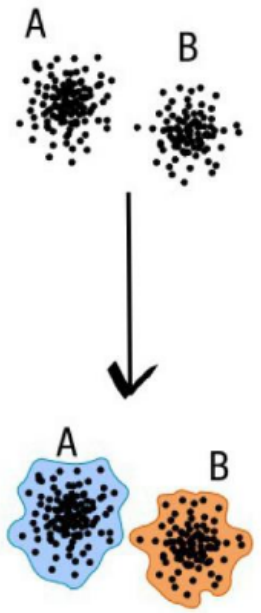
Since FROGS uses swarm (with the fastidious method with d=1) and strongly promotes denoising by chimera removal and cluster filtering, FROGS produces ASVs.

Why Swarm?

- Fixed clustering threshold is a real problem
- OTUs construction is input-order depenent



decreasing length,
decreasing abundance,
external references

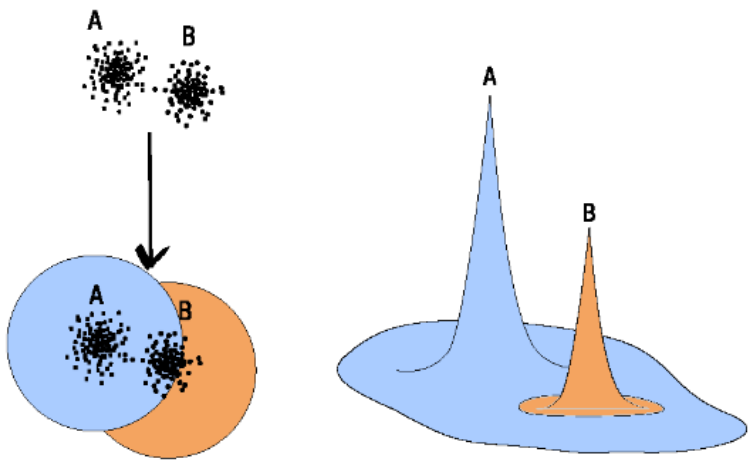


natural limits of clusters

Swarm: A smart idea

swarm

large-scale 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.

agglomeration rather than division

PeerJ

Torbjørn Rognes

Oslo University

Submitted 4 August 2015
Accepted 31 October 2015
Published 10 December 2015

Corresponding authors
Frédéric Mahé,
mahe@rhrk.uni-kl.de
Torbjørn Rognes,
torognes@ifi.uio.no

Academic editor
Gilles van Wezel

Additional Information and Declarations can be found on page 10

DOI 10.7717/peerj.1420

Swarm v2: highly-scalable and high-resolution amplicon clustering

Frédéric Mahé[1], Torbjørn Rognes[2,3], Christopher Quince[4], Colomban de Vargas[5,6] and Micah Dunthorn[1]

[1] Department of Ecology, Technische Universität Kaiserslautern, Kaiserslautern, Germany
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Department of Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway
[4] Warwick Medical School, University of Warwick, Warwick, United Kingdom
[5] UMR 7144, EPEP–Évolution des Protistes et des Écosystèmes Pélagiques, Station Biologique de Roscoff, CNRS, Roscoff, France
[6] UMR 7144 Station Biologique de Roscoff, Sorbonne Universités, UPMC Univ Paris 06, Roscoff, France

ABSTRACT

Previously we presented Swarm v1, a novel and open source amplicon clustering program that produced fine-scale molecular operational taxonomic units (OTUs), free of arbitrary global clustering thresholds and input-order dependency. Swarm v1 worked with an initial phase that used iterative single-linkage with a local clustering threshold (d), followed by a phase that used the internal abundance structures of clusters to break chained OTUs. Here we present Swarm v2, which has two important novel features: (1) a new algorithm for $d = 1$ that allows the computation time of the program to scale linearly with increasing amounts of data; and (2) the new fastidious option that reduces under-grouping by grafting low abundant OTUs (e.g., singletons and doubletons) onto larger ones. Swarm v2 also directly integrates the clustering and breaking phases, dereplicates sequencing reads with $d = 0$, outputs OTU representatives in fasta format, and plots individual OTUs as two-dimensional networks.

Subjects Biodiversity, Bioinformatics, Environmental Sciences, Microbiology, Molecular Biology
Keywords Environmental diversity, Barcoding, Molecular operational taxonomic units

INTRODUCTION

Traditional *de novo* amplicon clustering methods that can handle large high-throughput sequencing datasets (e.g., *Edgar, 2010*; *Ghodsi, Liu & Pop, 2011*; *Fu et al., 2012*) suffer from two fundamental problems. First, they rely on an arbitrary fixed global clustering threshold to group amplicons into molecular operational taxonomic units (OTUs). Global clustering thresholds have rarely been justified and are not applicable to all taxa and marker lengths (e.g., *Caron et al., 2009*; *Nebel et al., 2011*; *Dunthorn et al., 2012*; *Brown et al., 2015*). Second, there is variability in the clustering results due to amplicon input order (*Koeppel & Wu, 2013*; *Mahé et al., 2014*).

To solve these problems, we previously introduced the open source Swarm v1 program that implemented an initial clustering phase written in C++, then a breaking phase written in Python (*Mahé et al., 2014*). Swarm's clustering phase (Fig. 1A) was novel in its approach to single linkage clustering in that, instead of using a global clustering (e.g., *Hartmann et al., 2012*; *Huse et al., 2010*), amplicons were iteratively added together using a

How to cite this article Mahé et al. (2015), Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3:e1420; DOI 10.7717/peerj.1420

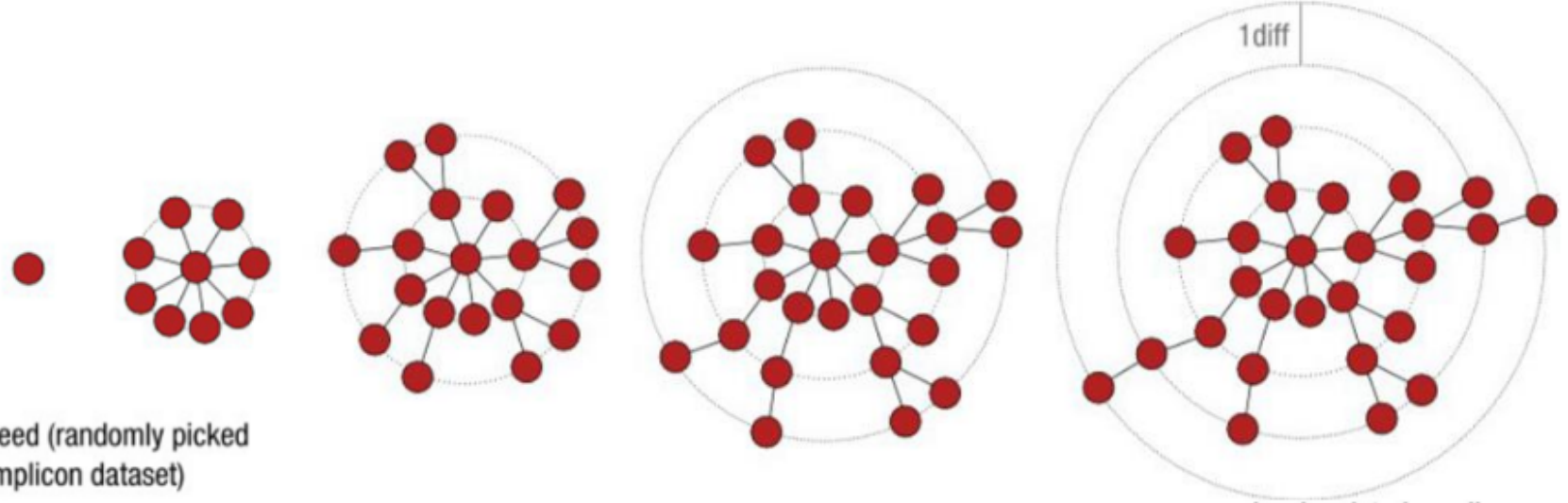
Mahé et al., 2015: Swarm v2: highly-scalable and high-resolution amplicon clustering [8]

Swarm

- A robust and fast clustering method for amplicon-based studies
- The purpose of swarm is to provide a novel clustering algorithm to handle large sets of amplicons
- swarm results are resilient to input-order changes and rely on a small local linking threshold d, the maximum number of differences between two amplicons
- swarm forms stable high-resolution clusters, with a high yield of biological information
- Default: forms a lot of low-abundant OTUs that are in fact artifacts and need to be removed
- **Swarm (fastidious method + *d=1*) clusters + filters → ASVs**

d: the small local linking threshold

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

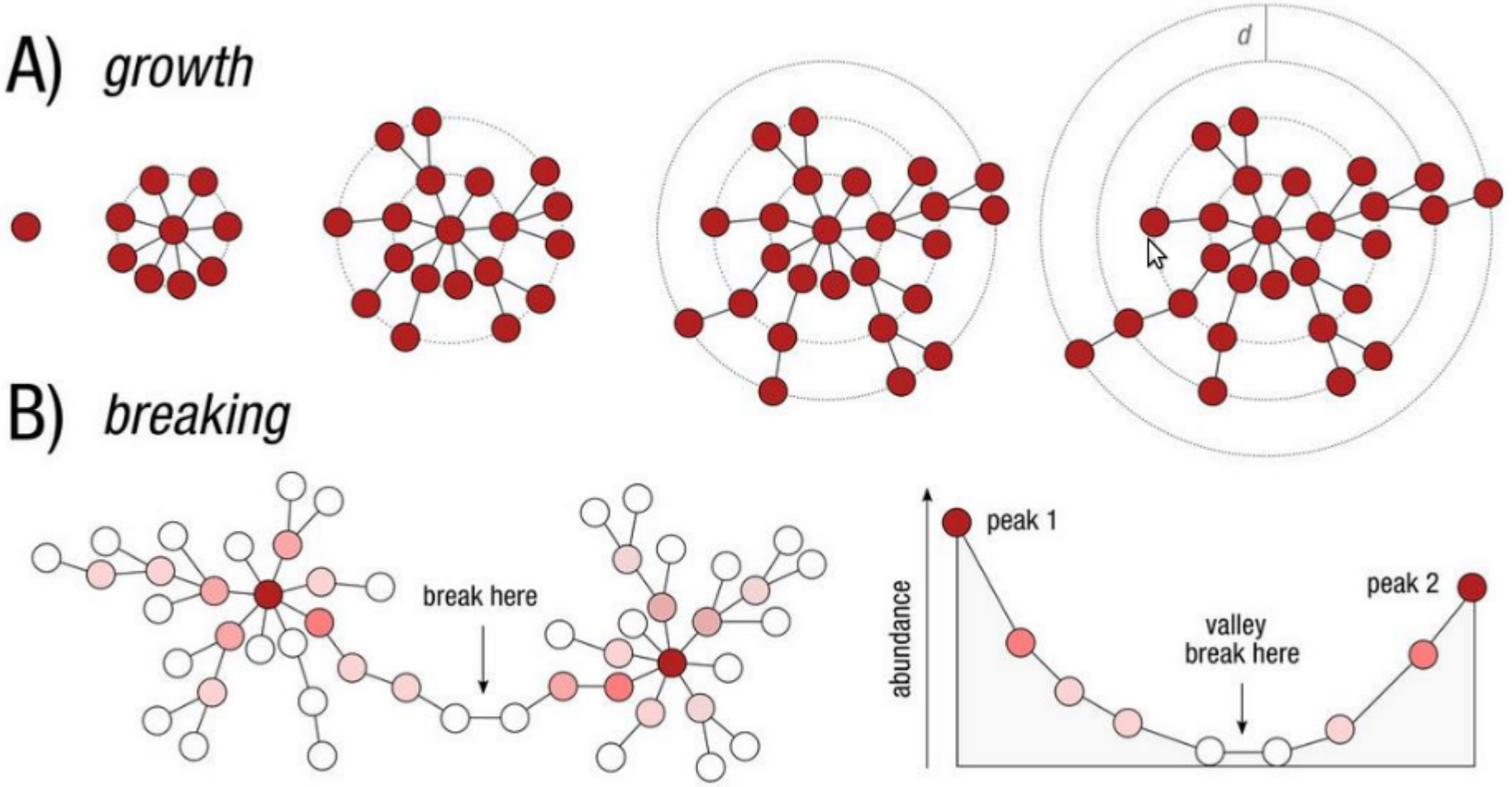


initial seed (randomly picked from amplicon dataset)

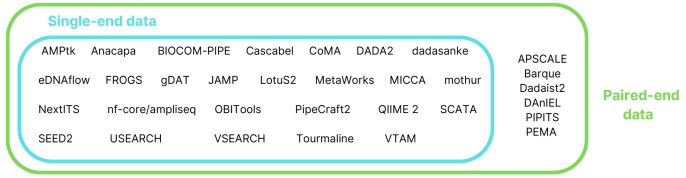
explore the amplicon space

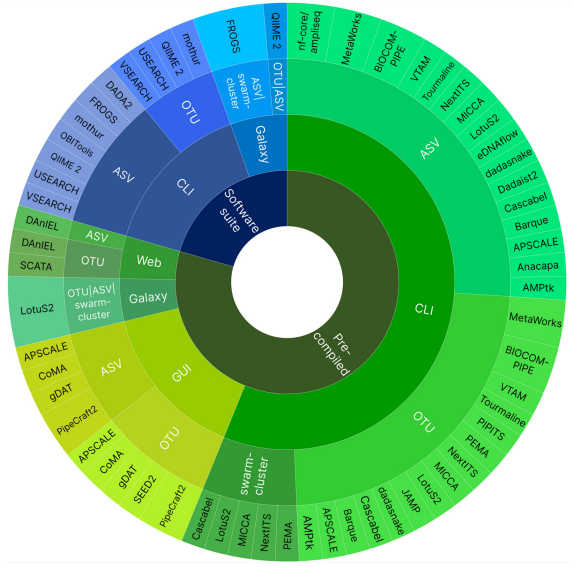
no more closely related amplicons, the process stops (equivalent to the Kruskal algorithm when d = 1)

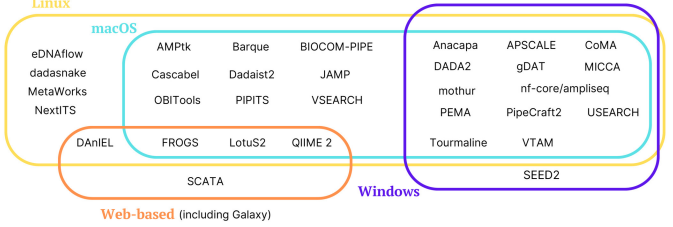
Swarm steps



Which method to choose?







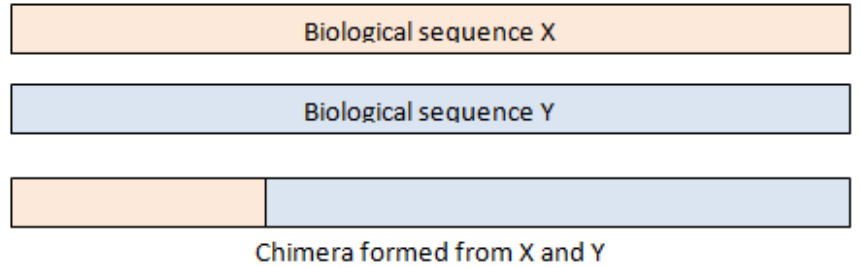
Advantages and inconvenients

- *Which type of features to prefer may be context-dependent, and both may even be used in the same study*
- ASV demonstrate a biologically informative fine-scale resolution [9]
- But difficult to separate noise from a real signal in low abundant reads [10]
- ASVs represent stable and reproducible units across studies whereas OTUs are dataset-specific features (swarm clusters are not ⚠)
  - problematic for longitudinal and very big studies

📢 FROGS will soon offer the choice between swarm and dada2 for ASV creation

TP FROGS clustering

Chimera removal

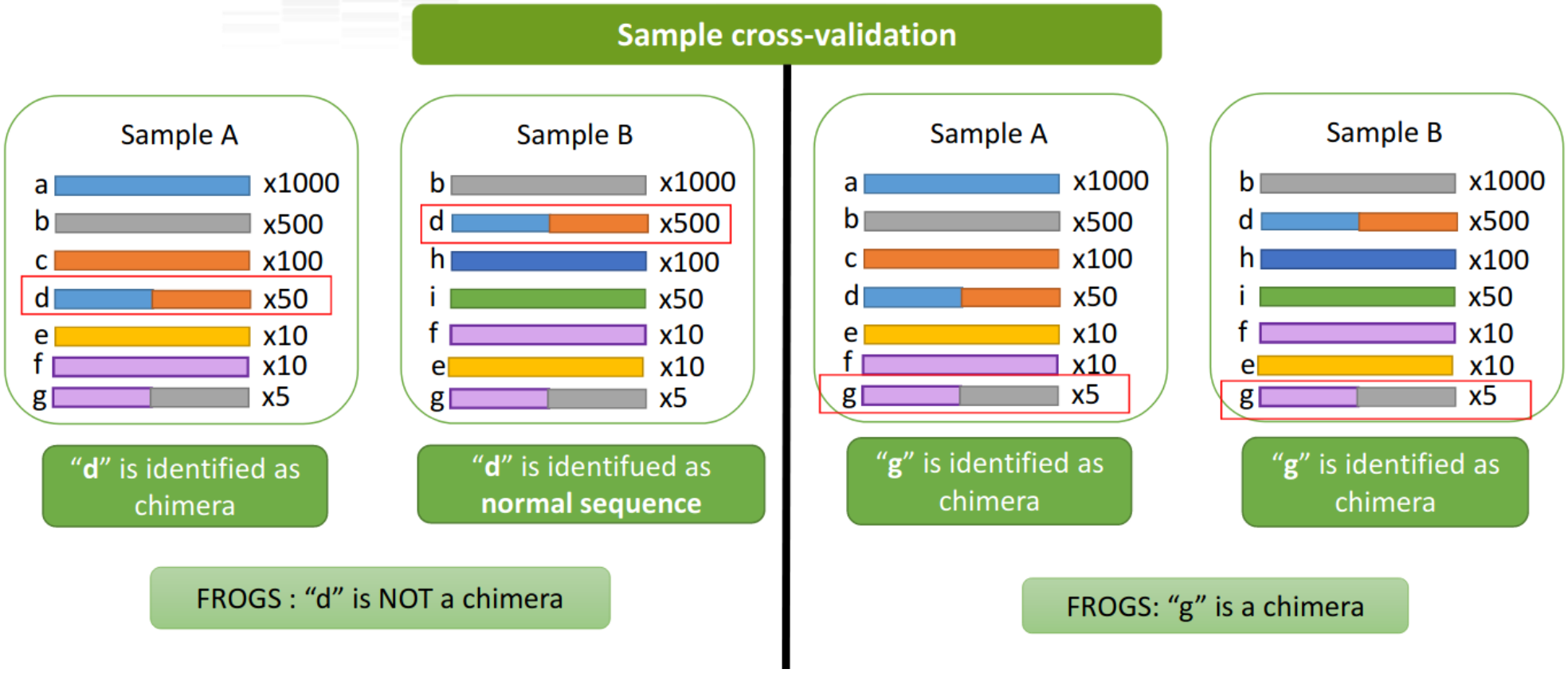


Chimera detection strategies

- Reference based: against a database of «genuine» sequences
  - dependant of the references used
- De novo: against abundant sequences in the samples 👍
- FROGS uses vsearch [4] as chimera removal tool

A little extra: the sample-cross validation

- FROGS adds a sample-cross validation



Chimera rates in samples

- From 5 to 40% in 16S data
- Few with ITS (<10%)

Samples	% Observed Chimera content			
	ABI3730	454 FLX Titanium		
	V1–V9	V1–V3	V3–V5	V6–V9
MC	5.99±3.07	14.26±10.34	14.75±9.45	13.49±8.52
gut	7.71±6.46	22.90±8.56	16.03±2.86	17.76±3.76
oral	7.22±6.35	20.55±11.73	10.98±4.01	9.10±5.02
skin	3.49±5.77	11.15±1.36	7.51±2.49	5.73±1.69
vaginal	6.31±6.64	12.60±6.70	6.62±3.51	3.00±1.65

*Values are averages ± STDEV calculated from multiple replicates of MC, and from replicates of multiple clinical samples originating from different body sites.
doi:10.1371/journal.pone.0039315.t001

Ward et al., 2012: Evaluation of 16S rDNA-based community profiling for human microbiome research [11]

TP Frogs remove chimera

Abundance/Prevalence filters

How to filter clusters?

- Low abundant sequences
- Clusters not shown in few replicates
- Contamination

TP Frogs cluster filters

Taxonomic affiliation

Comparison of approaches

Table 1 Number of taxonomic groups identified by each classifier among Illumina 16S rRNA gene sequences (SRR3225706) from a mock microbiome sample [33]. Counts are provided with and without including any sequences in the RDP training set that are labeled as belonging to the 20 expected genera

From: IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences

		Classified to genus level[α] (%)	Groups present in the mock community							Absent from mock community[β]		
			Root	Domain	Phylum	Class	Order	Family	Genus	Order	Family	Genus
Using the RDP training set	BLAST	97.9	1	0	0	0	0	0	17	0	0	24
	IDTAXA	94.2	1	0	1	1	2	5	14	0	1	2
	MAPSeq	96.5	1	0	0	0	0	4	15	0	2	6
	QIIME	95.4	1	0	0	0	0	0	16	0	0	7
	RDP Classifier	93.3	1	1	2	3	6	8	15	0	2	6
	SINTAX	94.2	1	1	1	4	3	3	14	1	0	3
	SPINGO	96.5	1	0	0	0	0	0	17	0	0	3
With expected genera excluded from training data	BLAST	17.3	1	0	0	0	0	0	0	0	0	65
	IDTAXA	0.01	1	1	1	2	3	4	0	0	2	2
	MAPSeq	24.6	1	0	0	2	5	11	0	1	8	20
	QIIME	13.5	1	0	0	0	0	0	0	0	0	16
	RDP Classifier	3.83	1	1	2	3	6	9	0	0	3	12
	SINTAX	8.76	1	1	1	7	5	6	0	1	1	9
	SPINGO	26.7	1	0	0	0	0	0	0	0	0	15

[α]Percent of total sequences from the mock community that were classified to the genus rank

[β]Other rank levels (root, domain, phylum, and class) all had counts of zero

RDP problems

- Depends too much on the databank used!
- Gives one affiliation for each feature with bootstrap, on each subdivision

```
Bacteria;(1.0);Actinobacteriota;(1.0);Actinobacteria;
(1.0);Propionibacteriales;(1.0);Propionibacteriaceae;(1.0);Cutibacterium;
(1.0);Cutibacterium acnes;(0.57);
```

The FROGS recommandation

- Use Blast and not RDP
- Check Blast metrics to avoid concluding too fast
- Take care of the reference databank used!

```
Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriacea
```

The FROGS databanks

- Command line: you can use your own databank
- Galaxy
  - You have access to several databanks
  - Admins have to add your databank
- The file must be well formated, we can do it for you
- ⚠ For private databanks, contact us!

The FROGS extra: the multi-affiliations

- FROGS gives all identical hits

```
Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;S
xylosus
Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;S
saprophyticus
```

Strictly identical (V1-V3 amplification) on 499 nucleotides

- FROGS can’t decide if it’s one or another
- You have to check if you can choose between multi-affiliations

To help you

- https://shiny.migale.inrae.fr/app/affiliationexplorer
- a very user-friendly Shiny web app, allowing users to modify very simply the affiliations from a FROGS abundance file

Filter ASVs on their affiliation

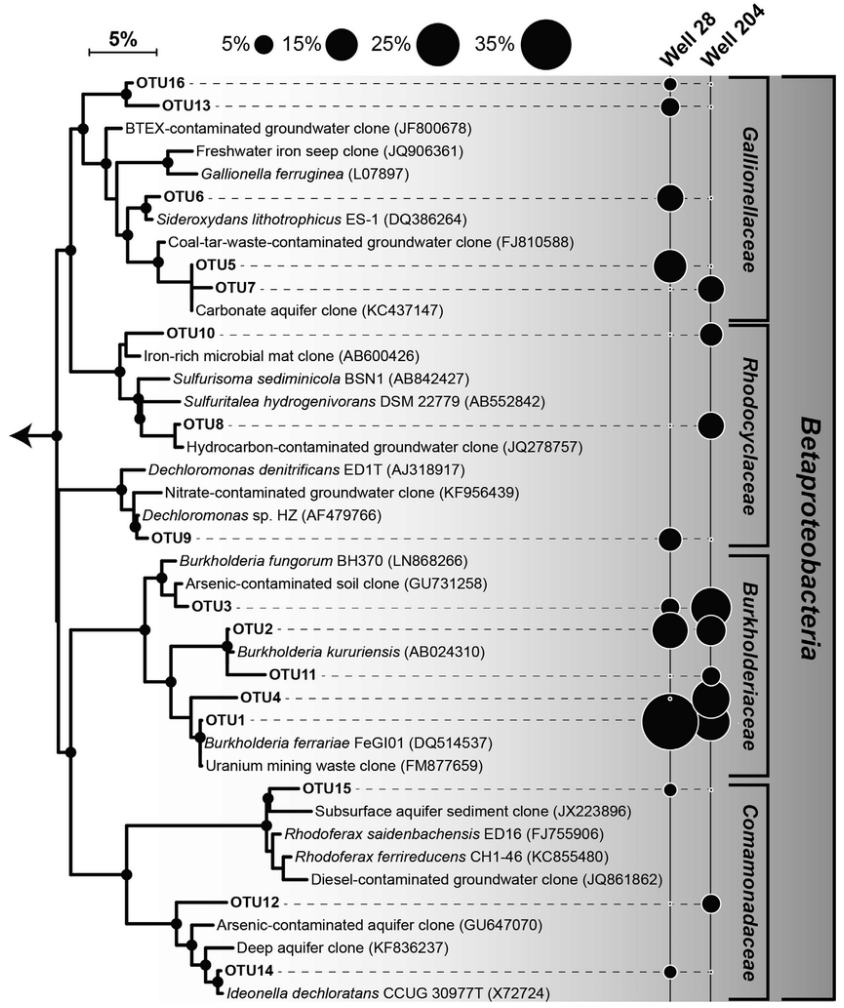
Affiliation filters

- Remaining contamination?
- Want to analyse only the `Firmicutes`?
- 2 modes
  - *Deleting*: remove ASVs
  - *Hiding*: only the affiliation is modified, not the abundance

Phylogenetic tree

FROGS tree

- This tool builds a phylogenetic tree thanks to affiliations of ASVs contained in the BIOM file
- Needed to compute beta-diversity indices based on phylogenetic distances
- Interesting to explore poor-characterized environments

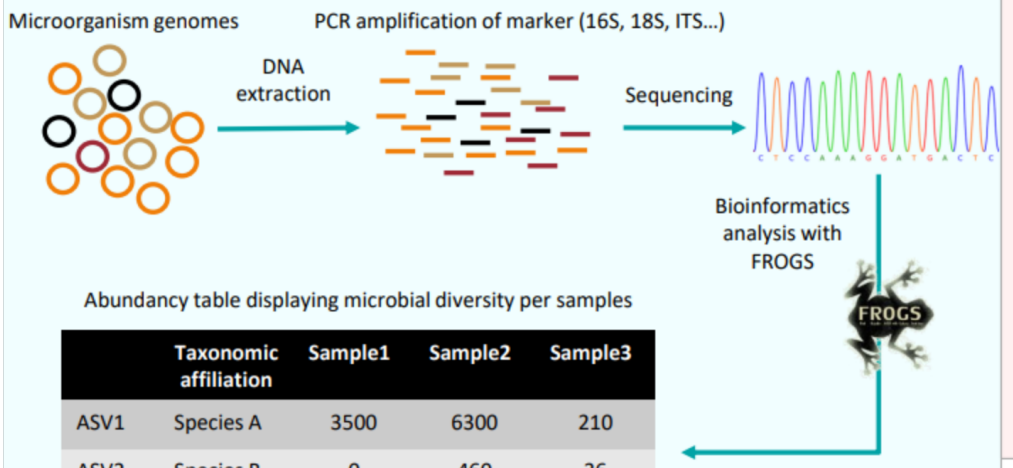


FROGSfunc: function inference

Concepts

Metabarcoding principle

Determine the diversity of a environment by amplification and sequencing of a genetic marker.

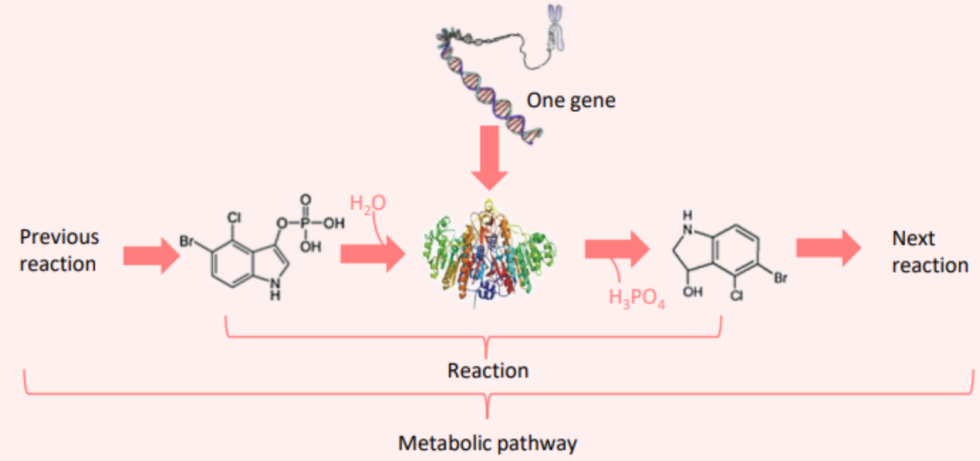


Abundancy table displaying microbial diversity per samples

	Taxonomic affiliation	Sample1	Sample2	Sample3
ASV1	Species A	3500	6300	210
ASV2	Species B	0	460	36
ASV3	Species C	400	700	500

Functional inference

Assuming that an organism has a metabolic function by the presence in the organism's genome of a known sequence having that function.



What metabolic functions are presents in a microbial community ?

Based on PICRUSt2

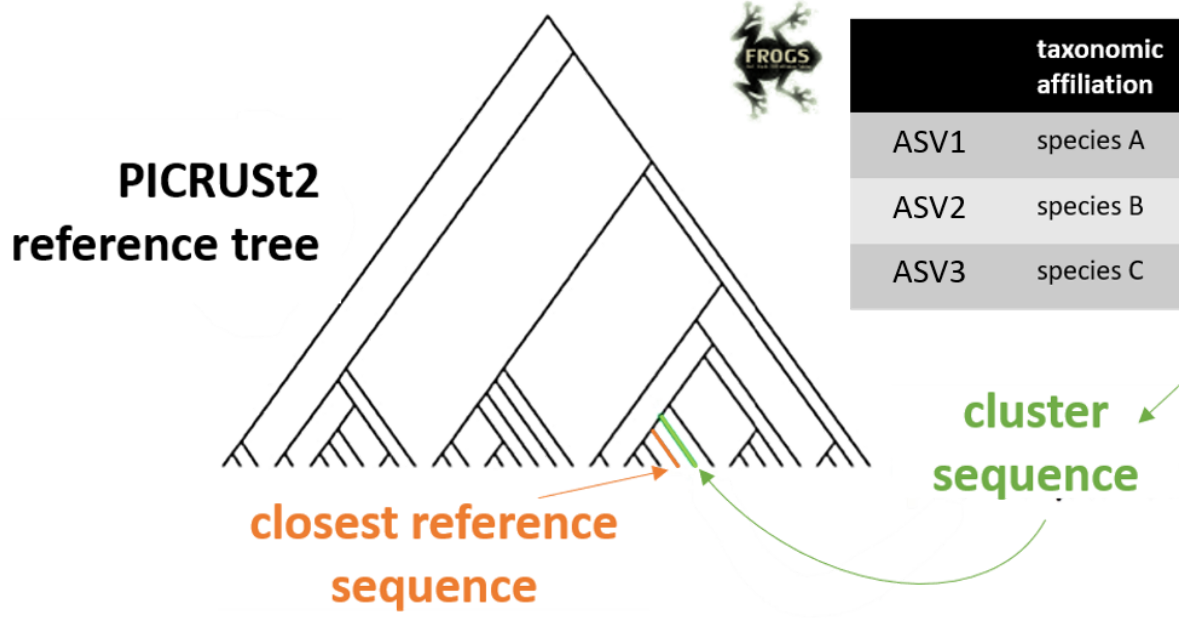
- PICRUSt [13] (Phylogenetic investigation of communities by reconstruction of unobserved states) is an open-source tool.
- It is a software for predicting functional abundances based only on marker gene sequences
- PICRUSt2 is composed of 4 python applications.
- No graphic interface exists to run PICRUSt2 for non-expert users.

Douglas et al., 2020: PICRUSt2 for prediction of metagenome functions [13]

How it works

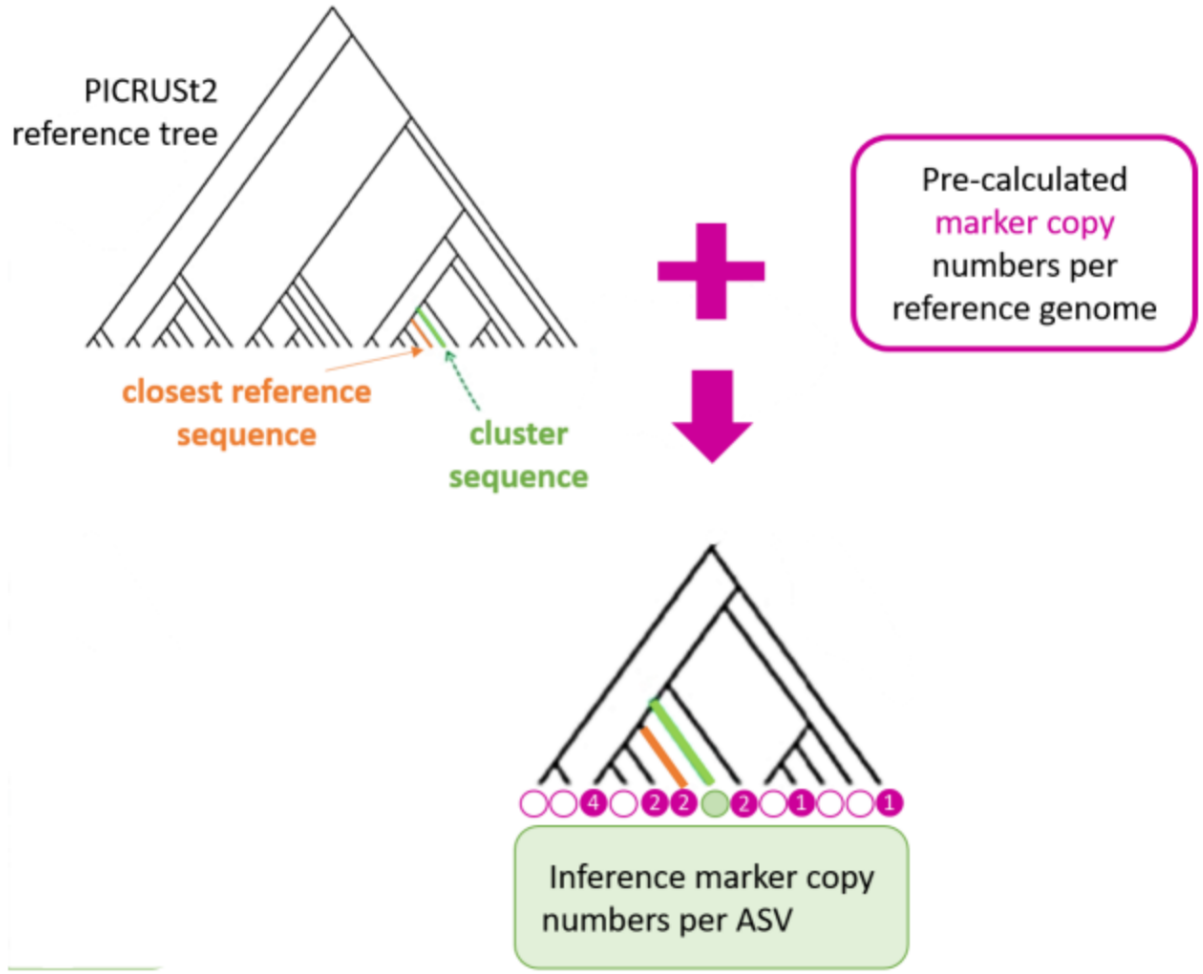
1. Places the ASVs into a reference phylogenetic tree and predicts of marker copy number in each ASV.
2. Predicts number of function copy number in each ASV and calculates functions abundances in each sample and ASV abundances according to marker copy number.
3. Calculates pathway abundances in each sample.

FROGSfunc placeseqs and copynumber



	taxonomic affiliation	fasta sequences	sample1	sample2	sample3
ASV1	species A	>AGGAG..	3500	6300	210
ASV2	species B	>ATGAG..	0	460	36
ASV3	species C	>AGGCT..	400	700	500

FROGSfunc placeseqs and copynumber



NSTI

- NSTI scores are simply the average branch length that separates each ASV in your sample from a reference bacterial genome, weighted by the abundance of that ASV in the sample.
- PICRUSt2 sets NSTI threshold to 2 per default. Some studies have shown that this threshold is permissive. Thus, it is important to see if the taxonomies between PICRUSt2 and FROGS are quite similar or not, in order to potentially choose a more stringent threshold afterwards.
  - 0 < Good < 0.5
  - 0.5 <= Medium < 1
  - 1 <= Bad < 2
  - To exclude >= 2

FROGSfunc functions

FROGSFUNC_1 abundance ASV

	S1	S2	S3
ASV1	3500	6300	210
ASV2	0	460	36
ASV3	400	700	500

÷

FROGSFUNC_1 marker.tsv

	16S copy number
ASV1	7
ASV2	4
ASV3	1

=

FROGSFUNC_2 marker_norm.tsv

	S1	S2	S3
ASV1	3500/7	6300/7	210/7
ASV2	0/4	460/4	36/4
ASV3	400/1	700/1	500/1

FROGSFUNC_2 marker_norm.tsv

	S1	S2	S3
ASV1	500	900	30
ASV2	0	115	9
ASV3	400	700	500

X

FROGSFUNC_2 EC_copynumber_predicted.tsv

	EC:1.1.1.1	EC:1.1.1.2	EC:1.1.1.3
ASV1	2	0	2
ASV2	1	0	0
ASV3	2	4	2

=

FROGSFUNC_2 functions_unstrat_EC.tsv

	S1	S2	S3
EC:1.1.1.1	1800	5460	1129
EC:1.1.1.2	1600	2800	2000
EC:1.1.1.3	1800	3200	1060

FROGSfunc pathways

FROGSFUNC_2 functions_unstrat_EC.tsv

	S1	S2	S3
EC:1.1.1.1	1800	5460	1129
EC:1.1.1.2	1600	2800	2000
EC:1.1.1.3	1800	3200	1060

+ PICRUST2 map of gene families to pathways →

FROGSFUNC_3 pathways_unstrat per sample and per reference

Pathways	S1	S2	S3
1CMET2-PWY	1289.7451	1485.2474	1233.5908
ANAEROFRUCAT-PWY	904.7455	1565.5453	1227.6231
ANAGLYCOLYSIS-PWY	1501.0804	1805.3271	1544.3206
ARG+POLYAMINE-SYN	0	49.3391	45.6559

References

1. Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, et al. FROGS: Find, rapidly, OTUs with galaxy solution. Bioinformatics. 2017;34:1287–94.

2. Bernard M, Rué O, Mariadassou M, Pascal G. FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. Briefings in Bioinformatics. 2021;22. doi:10.1093/bib/bbab318.

3. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics and Bioinformatics. 2021;3. doi:10.1093/nargab/lqab019.

4. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

5. Magoč T, Salzberg SL. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27:2957–63.

6. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast and accurate illumina paired-end reAd mergeR. Bioinformatics. 2013;30:614–20.

7. Hakimzadeh A, Abdala Asbun A, Albanese D, Bernard M, Buchner D, Callahan B, et al. A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. Molecular Ecology Resources. 2023.

8. Mahé F, Rognes T, Quince C, Vargas C de, Dunthorn M. Swarm v2: Highly-scalable and high-resolution amplicon clustering. PeerJ. 2015;3:e1420.

9. Couton M, Baud A, Daguin-Thiébaut C, Corre E, Comtet T, Viard F. High-thr
ethanol is effective at jointly examining infraspecific and taxonomic diversit

do not perform equally. Ecology and Evolution. 2021;11:5533–46.
10. De Santiago A, Pereira TJ, Mincks SL, Bik HM. Dataset complexity impacts both MOTU delimitation and biodiversity estimates in eukaryotic 18S rRNA metabarcoding studies. Environmental DNA. 2022;4:363–84.
11. Group JCHMPDGW. Evaluation of 16S rDNA-based community profiling for human microbiome research. PloS one. 2012;7:e39315.
12. Murali A, Bhargava A, Wright ES. IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. Microbiome. 2018;6:1–14.
13. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nature biotechnology. 2020;38:685–8.