Février_2017_Formation Métagénomique avec le logiciel FROGS
Géraldine Pascal, Laurent Cauquil, Maria Bernard

To cite this version:
Géraldine Pascal, Laurent Cauquil, Maria Bernard. Février_2017_Formation Métagénomique avec le logiciel FROGS. 2017. hal-02785427

HAL Id: hal-02785427
https://hal.inrae.fr/hal-02785427
Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Training on Galaxy: Metagenomics
February 2017

Find Rapidly OTU with Galaxy Solution

FRÉDÉRIC ESCUDIÉ*, and LUCAS AUER*, MARIA BERNARD, LAURENT CAUQUIL, KATIA VIDAL, SARAH MAMAN, MAHENDRA MARIADASSOU, SYLVIE COMBES, GUILLERMINA HERNANDEZ-RAQUET, GÉRALDINE PASCAL

*These authors have contributed equally to the present work.
Feedback:

What are your needs in “metagenomics”? 454 / MiSeq?

Your background?
9 am to 5 pm

2 short coffee breaks morning and afternoon

Lunch
12.30 to 2.00 pm
Overview

- Objectives
- Material: data + FROGS
- Data upload into galaxy environment
- Demultiplex tool
- Preprocessing
- Clustering + Cluster Statistics
- Chimera removal

- Filtering
- Affiliation + Affiliation Statistics
- Normalization
- Tool descriptions
- Format transformation
- Workflow creation
- Download data
- Some figures
Objectives

Analyses of bacterial communities

High-throughput sequencing of 16S/18S RNA amplicons

Illumina data, sequenced at great depth

Bioinformatics data processing

WITH operational taxonomic units (OTUs) and their taxonomic affiliation.

Abundance table
OTUs for ecology

Operational Taxonomy Unit: a grouping of similar sequences that can be treated as a single « species »

Strengths:
- Conceptually simple
- Mask effect of poor quality data
  - Sequencing error
  - In vitro recombination (chimera)

Weaknesses:
- Limited resolution
- Logically inconsistent definition
Objectives

<table>
<thead>
<tr>
<th>OTU</th>
<th>Affiliation</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
<th>Sample 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU1</td>
<td>Species A</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>45</td>
<td>75</td>
<td>18645</td>
</tr>
<tr>
<td>OTU2</td>
<td>Species B</td>
<td>741</td>
<td>0</td>
<td>456</td>
<td>4421</td>
<td>1255</td>
<td>23</td>
</tr>
<tr>
<td>OTU3</td>
<td>Species C</td>
<td>12786</td>
<td>45</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OTU4</td>
<td>Species D</td>
<td>127</td>
<td>4534</td>
<td>80</td>
<td>456</td>
<td>756</td>
<td>108</td>
</tr>
<tr>
<td>OTU5</td>
<td>Species E</td>
<td>8766</td>
<td>7578</td>
<td>56</td>
<td>0</td>
<td>0</td>
<td>200</td>
</tr>
</tbody>
</table>
Why we have developed FROGS

The current processing pipelines struggle to run in a reasonable time.

The most effective solutions are often designed for specialists making access difficult for the whole community.

In this context we developed the pipeline FROGS: «Find Rapidly OTU with Galaxy Solution».
Material
Sample collection and DNA extraction
« Meta-omics » using next-generation sequencing (NGS)

DNA

- Metagenomics
  - Amplicon sequencing
  - Shotgun sequencing

RNA

- Metatranscriptomics
  - RNA sequencing

Who is here?
- Wolfe et al., 2014

What can they do?
- Almeida et al., 2014

What are they doing?
- Dugat-Bony et al., 2015
The gene encoding the small subunit of the ribosomal RNA

The most widely used gene in **molecular phylogenetic** studies

Ubiquist gene: **16S rDNA** in prokaryotes; **18S rDNA** in eukaryotes

**Gene encoding a ribosomal RNA**: non-coding RNA (not translated), part of the small subunit of the ribosome which is responsible for the translation of mRNA in proteins

Not submitted to lateral gene transfer

Availability of databases facilitating comparison
(Silva 2015: >22000 type strains)
Secondary structure of the 16S rRNA of *Escherichia coli*

In red, fragment R1 including regions V1 and V2; in orange, fragment R2 including region V3; in yellow, fragment R3 including region V4; in green, fragment R4 including regions V5 and V6; in blue, fragment R5 including regions V7 and V8; and in purple, fragment R6 including region V9.

*Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences*

Pablo Yarza, et al.  
*Nature Reviews Microbiology* 12, 635–645 (2014) doi:10.1038/nrmicro3330
The gene encoding the small subunit of the ribosomal RNA
Steps for Illumina sequencing

- 1\textsuperscript{st} step: one PCR
- 2\textsuperscript{nd} step: one PCR
- 3\textsuperscript{rd} step: on flow cell, the cluster generations
- 4\textsuperscript{th} step: sequencing
Amplification and sequencing

« Universal » primer sets are used for **PCR amplification** of the phylogenetic biomarker

The primers contain **adapters** used for the sequencing step and **barcodes** (= tags = MIDs) to distinguish the samples (multiplexing = sequencing several samples on the same run)

exemple: V3
desired sequence
exemple: V4
Cluster generation

Prepare Genomic DNA Sample

Randomly fragment genomic DNA and ligate adaptors to both ends of the fragments.

Attach DNA to Surface

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Bridge Amplification

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Cluster generation

- Fragments Become Double Stranded
- Denature the Double-Stranded Molecules
- Complete Amplification

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Denature the double-stranded molecule

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Cycle of new strand synthesis and denaturation to make multiple copies of the same sequence (amplification).
Reverse strands are washed.
Sequencing by synthesis

**Determine First Base**

The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

Light signal is more strong in cluster

**Image First Base**

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

**Determine Second Base**

The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.
Sequencing by synthesis

Image Second Chemistry Cycle

Sequencing Over Multiple Chemistry Cycles

After laser excitation, the image is captured as before, and the identity of the second base is recorded.

The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time. Barcode is read, so cluster is identified.

After first sequencing (250 or 300 nt of Reverse strand), fragment form bridges again and Forward strand can be sequenced also.
Identification of bacterial populations may be not discriminating.
Amplification and sequencing

Sequencing is generally performed on **Roche-454** or **Illumina MiSeq** platforms.

Roche-454 generally produce ~ 10 000 reads per sample

MiSeq ~ 30 000 reads per sample

Sequence length is >650 bp for pyrosequencing technology (Roche-454) and 2 x 300 bp for the MiSeq technology in paired-end mode.
Methods
### Which bioinformatics solutions?

<table>
<thead>
<tr>
<th></th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| 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.  

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  
FROGS?

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

Data acquisition
- Output files from Genotoul: bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtd, yd, sam, fasta, pdf, xsq, tangu, bw, png, gff, pileup, pileup.gz, zip

Pre-process
- Archive file
  - dereplicated_file (fasta)
  - count_file (tabular)
  - summary_file (html)

Clustering
- Sequences file
  - Count file
  - seed_file (fasta)
  - abundance_biom (biom1)
  - swarms_composition (tabular)

Chimera
- Sequences file
  - Abundance file
  - non_chimera fasta (fasta)
  - out_abundance_biom (biom1)
  - out_abundance_count (tabular)
  - summary_file (html)

Affiliation
- OTU seed sequence
- Abundance file
- biom_affiliation (biom1)
- summary (html)
**Home made script**

- Data acquisition
- Pre-process
- Clustering
- Chimera
- Demultiplexing
- Normalization
- Cluster Statistics
- Filter
- Convert to TSV
- Convert to standard Biom
- Convert TSV to Biom

**Tools:**

- flash (1.2.11)
- cutadapt (1.8.3)
- Swarm (v2.1.1)
- VCHIME of VSEARCH package (1.1.3)
- RDPClassifier and NCBI Blast+ (2.2.29) on Silva SSU

**New for Galaxy**

- tar.gz format
- Home made script

**Affiliation Statistics**

- Home made script
Together go to visit FROGS

In your internet browser (Firefox, chrome, Internet explorer):

http://sigenae-workbench.toulouse.inra.fr/

Enter your email address and password from GenoToul
Available Tools

FROGS Pre-process Step 1 in metagenomics analysis: denoising and denoising. (Galaxy Version 1.5.0)

Sequence

Input type

Files by samples

Samples: files can be provided in single archive or with two files (K1 and R2) by sample.

The inputs contain 1 file by sample: Reads 1 and Reads 2 are already contiged by pair.

Available Tools

Tool Configuration and Execution

Datasets

History
Upload data  

Go to demultiplexing tool
What kind of data?

4 Upload → 4 Histories

Multiplexed data
Pathobiomes rodents and ticks
multiplex.fastq
barcode.tabular

454 data
Freshwater sediment metagenome
454.fastq.gz
SRA number
SRR443364

MiSeq R1 fastq + R2 fastq
Farm animal feces metagenome
sampleA_R1.fastq
sampleA_R2.fastq

MiSeq contiged fastq in archive tar.gz
Farm animal feces metagenome
100spec_90000seq_9samples.tar.gz
1ST CONNEXION

RENAME HISTORY

- click on Unnamed history,
- Write your new name,
- Tap on Enter.
History gestion

- Keep all steps of your analysis.
- Share your analyzes.
- At each run of a tool, a new dataset is created. The data are not overwritten.
- Repeat, as many times as necessary, an analysis.
- All your logs are automatically saved.
- Your published histories are accessible to all users connected to Galaxy (Shared Data / Published Histories).
- Shared histories are accessible only to a specific user (History / Option / Histories Shared With Me).
- To share or publish a history: User / Saved histories / Click the history name / Share or Publish.
Your turn! - 1

LAUNCH UPLOAD TOOLS
Accounts:

- anemone
- arome
- aster
- bleuet
- camelia
- capucine
- chardon
- clematite
- cobee

- coquelicot
- cosmos

Password: f1o2r3!
Your turn: exo 1

Create the 1st history multiplexed
Import files « multiplex.fastq » and « barcode.tabular » present in the Genotoul folder /work/formation/FROGS/

Create the 2nd history 454
Import file « 454.fastq.gz » present in the Genotoul folder /work/formation/FROGS/
(datatype fastq or fastq.gz is the same !)

Create the 3rd history MiSeq R1 R2
Import files « sampleA_R1.fastq » and « sampleA_R2.fastq » present in the Genotoul folder /work/formation/FROGS/

Create the 4th history MiSeq contiged
Import archive file « 100spec_90000seq_9samples.tar.gz » present in the Genotoul folder /work/formation/FROGS/
History creation

Your history is empty. Click 'Get Data' on the left pane to start.
Upload data: different methods

Default method, your files are on your computer or accessible on the internet, they are copied on your Galaxy account.

You can only upload one local file at a time → 10 samples ≥ 10 uploads
You can upload multiple files using URLs but only smaller than 2Go

Each uploaded file will consume your Galaxy’s quota!
Upload data: different methods

Specific SIGENAE GENOTOUL method. It allows you to access to your files in your work account on the Genotoul without consuming your Galaxy quota.

And if you have multiple samples?
See How_to_create_an_archiveTAR.ppt

How to transfer files on /work of Genotoul?
See How_to_transfert_to_genotoul.ppt
If you have an archive on your own computer and smaller than 2Go, you may use this specific FROGS tool to upload your samples archive instead of the default « Upload File » of Galaxy.
Upload data: different methods

New functionality in latest Galaxy version:
http://147.99.108.167/galaxy/

You can only upload multiple files at a time but only smaller than 2Go
Demultiplexing tool
Data acquisition

- Upload File from Galaxy
  - out1 (bam, txt, tabular, fastq, sam,fasta, pdf, xsl, tar.gz, bw, png, tiff, pileup, pileup.gz, zip)

Pre-process

- FROGS Pre-process
  - Archive file
    - depiliated_file (fasta)
    - count_file (tabular)
    - summary_file (html)

Demultiplexing

- FROGS Demultiplex reads
  - Barcode file
    - Select_fasta_dataset
      - demultiplexed_archive (data)
      - undemultiplexed_archive (data)
      - summary (tabular)

Clustering

- FROGS Clustering swarm
  - Sequences file
    - Count file
      - count_file (tabular)
      - abundance_file (tab)
      - swarms_composition (tabular)

- FROGS Remove chimer
  - Sequences file
    - Abundance file
      - non_chimera.fasta (fasta)
      - out_abundance_biom (biom1)
      - out_abundance_count (tabular)
      - summary_file (html)

Affiliation

- FROGS Affiliation OTU
  - OTU seed sequence
    - Abundance file
      - biom_affiliation (biom1)
      - summary (html)
Barcoding ?

ATGGCTG
CTTTGCTA
TTGGGAC
GCAGCTG
Demultiplexing

Sequence demultiplexing in function of barcode sequences:
- In forward
- In reverse
- In forward and reverse

Remove unbarcoded or ambiguous sequences
Demultiplexing forward

Single end sequencing

Paire end sequencing

R1

R2
Demultiplexing reverse

Adapter A Primer Fwd Amplicon sequence targeted Primer Rv Barcode Rv Adapter B

Single end sequencing

Paire end sequencing

R1

R2
Demultiplexing forward and reverse

Single end sequencing

Pair end sequencing

R1

R2
Your turn! - 2

LAUNCH DEMULTIPLEX READS TOOL
Accounts:

- anemone
- arome
- aster
- bleuet
- camelia
- capucine
- chardon
- clematite
- cobee
- coquelicot
- cosmos
- cyclamen
- Password: f1o2r3!
The tool parameters depend on the input data type.
Exercise 2

In multiplexed history launch the demultiplex tool:

« The Patho-ID project, rodent and tick’s pathobioms study, financed by the metaprogram INRA-MEM, studies zoonoses on rats and ticks from multiple places in the world, the co-infection systems and the interactions between pathogens. In this aim, they have extracted hundreds of or rats and ticks samples from which they have extracted 16S DNA and sequenced them first time on Roche 454 plateform and in a second time on Illumina Miseq plateform. For this courses, they authorized us to publicly shared some parts of these samples. »

Exercise 2

In **multiplexed** history launch the demultiplex tool:

Data are single end reads  
→ only 1 fastq file

Samples are characterized by an association of two barcodes in forward and reverse strands  
→ multiplexing « both ends »

```
2: /work/frogs/Formation/multiplex.fastq
1: /work/frogs/Formation/barcode.txt
```
Exercise 2

Demultiplex tool asks for 2 files: one « fastq » and one « tabular »

1. Play with pictograms
2. Observe how is built a fastq file.
3. Look at the stdout, stderr when available (in the pictogram )
What it does

Classify single or paired end reads in function of barcode forward or reverse in the first or both reads.

Command line:

demultiplex.py --input-R1 'FQ_INPUT1' --input-R2 'FQ_INPUT2' --input-barcode 'TXT1'
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.

- If you have Roche 454 sequences in sff format, you must convert them with some program like sff2fastq
Results

A tar archive is created by grouping one (or a pair of) fastq file per sample with the names indicated in the first column of the barcode tabular file.

With barcode mismatches >1 sequence can corresponding to several samples. So these sequences are non-affected to a sample.

Sequences without known barcode. So these sequences are non-affected to a sample.
Format: Barcode

BARCODE FILE is expected to be tabulated:

- 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)

Take care to indicate sequence barcode in the strand of the read, so you may need to reverse complement the reverse barcode sequence. Barcode sequence must have the same length.

Example of barcode file.
The last column is optional, like this, it describes sample multiplexed by both fragment ends.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Forward Barcode</th>
<th>Reverse Barcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>MgArd00001</td>
<td>ACAGCGT</td>
<td>ACGTACA</td>
</tr>
</tbody>
</table>
Format: FastQ

FASTQ: Text file describing biological sequence in 4 lines format:

- first line start by "@" correspond to the sequence identifier and optionally the sequence description. "@Sequence_1 description1"
- second line is the sequence itself. "ACAGC"
- third line is a "+" following by the sequence identifier or not depending on the version
- fourth line is the quality sequence, one code per base. The code depends on the version and the sequencer

@HNHOSKD01ALD0H
ACAGCGTGACAGGGGTACCCAGTCAGCCATGACGTAGCAGTACAA
+
CCCFFFFFFFFHHHJJJJJJHHHFF@DEDDDDDDD@CDDDDACDD
How it works?

For each sequence or sequence pair the sequence fragment at the beginning (forward multiplexing) of the (first) read or at the end (reverse multiplexing) of the (second) read will be compared to all barcode sequence.

If this fragment is equal (with less or equal mismatch than the threshold) to one (and only one) barcode, the fragment is trimmed and the sequence will be attributed to the corresponding sample.

Finally fastq files (or pair of fastq files) for each sample are included in an archive, and a summary describes how many sequence are attributed for each sample.
Pre-process tool
Pre-process

Data acquisition

Upload File from Genotoul

out1 (bam, txt, tabular, fastq, sanger, esfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, proq, sff, pileup, pileup.gz, zip)

Pre-process

FROGS Pre-process

Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Demultiplexing

FROGS Demultiplex reads

Barcode file
Select fasta dataset

demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)

Clustering

FROGS Clustering swarm

Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

Chimera

FROGS Remove chimera

Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Affiliation

FROGS Affiliation OTU

OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

71
Pre-process

From demultiplex tool

454

MiSeq Fastq R1

MiSeq Fastq R2

Already contiged
Amplicon-based studies general pipeline

Pre-process

- sampling
  - exhaustive sampling?
  - DNA/RNA extraction?

- amplification
  - universal markers?
  - universal primers?
  - polymerase errors?

- sequencing
  - error rates?
  - GC bias?

- denoising
dereplication
  - quality threshold?
  - length filtering?

- assignment

- clustering

- refs
  - taxonomic coverage?

- statistics (α, β, γ)
  - phylogeny
  - network analysis

chemistry, physics and randomness
bioinformatics (preparation)
bioinformatics (actual analyses)
Pre-process

- Delete sequence with not expected lengths
- Delete sequences with ambiguous bases (N)
- Delete sequences do not contain good primers
- Dereplication

- + removing homopolymers (size = 8) for 454 data
- + quality filter for 454 data
Example for:

- Illumina MiSeq data
- 1 sample
- Non joined

Parameters for the merging

**Pre-process example 1**
Pre-process example 1

[V3 – V4] 16S variability

Primer sequences
Example for:

- Sanger 454 data
- 1 sample
- Joined

**Pre-process example 2**

- [V3 – V4] 16S variability
- Primer sequences

**FROGS Pre-process** Step 1 in metagenomics analysis: denoising and dereplication. (Galaxy Version 1.5.0)

- **Sequencer**: 454

  Select the sequencer family used to produce the sequences.

- **Input type**: One file by sample

  Samples files can be provided in single archive or with one file by sample.

- **Samples**

  1: Samples

  - **Name**: my_sample
    
    The sample name.

  - **Sequence file**: 1:/work/formation/FROGS/454.fastq.gz

  FASTQ file of sample.

- **Minimum amplicon size**

  380

  The minimum size for the amplicons (with primers).

- **Maximum amplicon size**

  500

  The maximum size for the amplicons (with primers).

- **5' primer**

  ACGGGAGGCACCGAGC

  The 5' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primer parameters'.

- **3' primer**

  AGGATTAGATACCTGGTA

  The 3' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primer parameters'.

- **Execute**
Example for:
- Illumina MiSeq data
- 9 samples in 1 archive
- Joined
- Without sequenced PCR primers (Kozich protocol)
Your turn! - 3

GO TO EXERCISES 3
Exercise 3.1

Go to « 454 » history

Launch the pre-process tool on that data set

→ objective : understand the parameters

1- Test different parameters for « minimum and maximum amplicon size »

2- Enter these primers: Forward: ACGGGAGGCAGCAG  Reverse: AGGATTAGATACCTGGTA
Size range of 16S V3-V4: [380 – 500]

Primers used for sequencing V3-V4:
Forward: ACGGGAGGCAGCAG
Reverse: AGGATTAGATACCCCTGGTA

Sample name is required
Exercise 3.1

What do you understand about amplicon size, which file can help you?

What is the length of your reads before preprocessing?

Do you understand how enter your primers?

What is the « FROGS Pre-process: dereplicated.fasta » file?

What is the « FROGS Pre-process: count.tsv » file?

Explore the file « FROGS Pre-process: report.html »

Who lose a lot of sequences?
To be kept, sequences must have the 2 primers.

<table>
<thead>
<tr>
<th>Samples</th>
<th>before process</th>
<th>with the two primers</th>
<th>with expected length</th>
<th>without N</th>
<th>without large homopolymer</th>
<th>without nearest poor quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample_454</td>
<td>28,009</td>
<td>20,227</td>
<td>6,806</td>
<td>8,677</td>
<td>6,875</td>
<td>6,872</td>
</tr>
</tbody>
</table>

To adjust your filtering, check the distribution of sequence lengths.
Cleaning, how it work?

Filter contig sequence on its length which must be between min-amplicon-size and max-amplicon-size

use cutadapt to search and trim primers sequences with less than 10% differences
Cleaning, how it work?

dereplicate sequences and return one uniq fasta file for all sample and a count table to indicate sequence abundances among sample.

In the HTML report file, you will find for each filter the number of sequences passing it, and a table that details these filters for each sample.
Exercise 3.2

Go to « MiSeq R1 R2 » history

Launch the pre-process tool on that data set

→ objective: understand flash software
MiSeq
R1 R2

Primers used for this sequencing:
Forward: CCGTCAATTC
Reverse: CCGCNNGCTGCT
Lecture 5’ → 3’

Real 16S sequenced fragment

Size with primers

>&ERR619083.M00704
CCGTCAATTCCATGAGATTTGCAACCTTGCAGGCGATCGTT
TATCGCGTTAGCTTCGCCAAGCAGCACATCTCTCGCCCTAGCCCAACGTACATCG
TTTAGGGTGTGACACTCCCGGTATATCAATCTCTGCTACCCCAACGTACATCG
AGCTCCAGGCTAGTGAACAGACAGAGAGACGCTTTCGCCACTGGTGTTCCTC
CATATACTACGGATTTCACCGGTACACATGGAATTCCACTCTCCCCTTAC
TCCAGACTTCCGAGGCTCACTCGCTTTACGCCCAATAATCCGGACAA
CGCTTGCCACCTACGTATTA

Expected amplicon size:
410
The expected size of the amplicon.

Minimum amplicon size:
340
The minimum size for the amplicon.

Maximum amplicon size:
450
The maximum size for the amplicon.

Sequencing protocol:
Illumina standard
The protocol used for sequencing: standard or custom with PCR primers as sequencing primers.

5’ primer:
CCGTCAATTCC
The 5’ primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters'.

3’ primer:
CCGCNGCTGCT
The 3’ primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters'.
Flash, how it works?

To contig read1 and read2 with FLASh with:

- A **minimum overlap** equals to 
  
  \[ \text{minimum overlap} = (R1\text{-size} + R2\text{-size}) - \text{expected-amplicon-size} \]

- And a **maximum overlap** equal to \[ \text{expected-amplicon-size} \] with a maximum of 10% mismatch among this overlap

**Ex:** minimum overlap \((250+250) - 410 = 50\)

maximum overlap 450

90% of the amplicon are smaller than \[ \text{expected-amplicon-size} \]
Exercise 3.2

Exercise 3.3

Go to « MiSeq contiged » history
Launch the pre-process tool on that data set
→ objective: understand output files
Exercise 3.3

3 samples are **technically replicated** 3 times: 9 samples of 10,000 sequences each.

- 100_10000seq_sampleA1.fastq
- 100_10000seq_sampleA2.fastq
- 100_10000seq_sampleA3.fastq
- 100_10000seq_sampleB1.fastq
- 100_10000seq_sampleB2.fastq
- 100_10000seq_sampleB3.fastq
- 100_10000seq_sampleC1.fastq
- 100_10000seq_sampleC2.fastq
- 100_10000seq_sampleC3.fastq
Exercise 3.3

• 100 species, covering all bacterial phyla
• Power Law distribution of the species abundances
• Error rate calibrated with real sequencing runs
• 10% chimeras
• 9 samples of 10 000 sequences each (90 000 sequences)
Exercise 3.3

“Grinder (v 0.5.3) (Angly et al., 2012) was used to simulate the PCR amplification of full-length (V3-V4) sequences from reference databases. The reference database of size 100 were generated from the LTP SSU bank (version 115) (Yarza et al., 2008) by

1. filtering out sequences with a N,
2. keeping only type species
3. with a match for the forward (ACGGRAGGCAGCAG) and reverse (TACCAGGGTATCTAATCCTA) primers in the V3-V4 region and
4. maximizing the phylogenetic diversity (PD) for a given database size. The PD was computed from the NJ tree distributed with the LTP.”
Primers used for this sequencing:
5’ primer: ACGGGAGGCACGAG
3’ primer: TAGGATTAGATACCCCTGGTA
Lecture 5’ → 3’
Exercise 3.3 - Questions

1. How many sequences are there in the input file?
2. How many sequences did not have the 5’ primer?
3. How many sequences still are after pre-processing the data?
4. How much time did it take to pre-process the data?
5. What can you tell about the sample based on sequence length distributions?
Clustering tool
Why do we need clustering?

Amplification and sequencing are not perfect processes.

Error rates?

polymerase errors?

chemistry, physics and randomness

Frédéric Mahé communication
Expected Results

A B

To have the best accuracy:

Method: All against all

- Very accurate
- Requires a lot of memory and/or time

=> Impossible on very large datasets without strong filtering or sampling
How traditional clustering works?
Input order dependent results

- Decreasing length, decreasing abundance, external references
Single a priori clustering threshold
Swarm clustering method

![Diagram of swarm clustering method](image)
Comparison Swarm and 3% clusterings

Radius expressed as a percentage of identity with the central amplicon (97% is by far the most widely used clustering threshold)
Comparison Swarm and 3% clusterings

More there is sequences, more abundant clusters are enlarged (more amplicon in the cluster). More there are sequences, more there are artefacts
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.

1st run for denoising:
Swarm with $d = 1$ - high clusters definition
linear complexity

2nd run for clustering:
Swarm with $d = 3$ on the seeds of first Swarm
quadratic complexity

Gain time!
Remove false positives!
Cluster stat tool
FROGS Clusters stat Process some metrics on clusters. (Galaxy Version 1.4.0)

**Abundance file**

File field:
- 5: FROGS Clustering swarm: abundance.biom

Clusters abundance (format: BIOM).

- [ ] Execute
Your Turn! - 4

LAUNCH CLUSTERING AND CLUSTERSTAT TOOLS
Exercise 4

Go to « MiSeq contiged » history

Launch the Clustering SWARM tool on that data set with aggregation distance = 3 and the denoising objectives:

- understand the denoising efficiency
- understand the ClusterStat utility
Exercise 4

1. How much time does it take to finish?
2. How many clusters do you get?
Exercise 4

3. Edit the biom and fasta output dataset by adding d1d3.

4. Launch FROGS Cluster Stat tools on the previous abundance biom file.
Exercise 4

5. Interpret the boxplot: **Clusters size summary**
6. Interpret the table: **Clusters size details**
7. What can we say by observing the **sequence distribution**?
8. How many clusters share “sampleB3” with at least one other sample?
9. How many clusters could we expect to be shared?
10. How many sequences represent the 550 specific clusters of “sampleC2”?
11. This represents what proportion of “sampleC2”? 
12. What do you think about it?
13. How do you interpret the « Hierarchical clustering »?

The « Hierarchical clustering » is established with a Bray Curtis distance particularly well adapted to abundance table of very heterogenous values (very big and very small figures).
Most of clusters are singletons
Clusters size summary

After filtering little clusters

Clusters size distribution

Clusters size distribution (decile)

<table>
<thead>
<tr>
<th>Decile</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
</tr>
<tr>
<td>Median</td>
<td>112</td>
</tr>
<tr>
<td>6</td>
<td>145</td>
</tr>
<tr>
<td>7</td>
<td>225</td>
</tr>
<tr>
<td>8</td>
<td>412</td>
</tr>
<tr>
<td>9</td>
<td>994</td>
</tr>
<tr>
<td>Max</td>
<td>13,337</td>
</tr>
</tbody>
</table>
Most of clusters are singletons

<table>
<thead>
<tr>
<th>Cluster size</th>
<th>Number of clusters</th>
<th>% of all clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4,595</td>
<td>77.36</td>
</tr>
<tr>
<td>2</td>
<td>886</td>
<td>14.58</td>
</tr>
<tr>
<td>3</td>
<td>155</td>
<td>2.61</td>
</tr>
<tr>
<td>4</td>
<td>83</td>
<td>1.40</td>
</tr>
<tr>
<td>5</td>
<td>42</td>
<td>0.71</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>0.49</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>0.37</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>0.22</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>0.10</td>
</tr>
</tbody>
</table>

After clustering
Most of sequences are contained in big clusters

The small clusters represent few sequences
58% of the specific clusters of sampleA1 represent around 5% of sequences. Could be interesting to remove if individual variability is not the concern of user.

### Sequences

<table>
<thead>
<tr>
<th>Sample</th>
<th>Shared clusters</th>
<th>Own clusters</th>
<th>Shared sequences</th>
<th>Own sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>100_10000seq_sampleA1</td>
<td>367</td>
<td>513</td>
<td>9,447</td>
<td>528</td>
</tr>
<tr>
<td>100_10000seq_sampleA2</td>
<td>365</td>
<td>490</td>
<td>9,476</td>
<td>503</td>
</tr>
<tr>
<td>100_10000seq_sampleA3</td>
<td>384</td>
<td>483</td>
<td>9,478</td>
<td>494</td>
</tr>
<tr>
<td>100_10000seq_sampleB1</td>
<td>395</td>
<td>548</td>
<td>9,397</td>
<td>572</td>
</tr>
<tr>
<td>100_10000seq_sampleB2</td>
<td>375</td>
<td>508</td>
<td>9,455</td>
<td>515</td>
</tr>
<tr>
<td>100_10000seq_sampleB3</td>
<td>376</td>
<td>562</td>
<td>9,388</td>
<td>579</td>
</tr>
<tr>
<td>100_10000seq_sampleC1</td>
<td>372</td>
<td>539</td>
<td>9,413</td>
<td>562</td>
</tr>
<tr>
<td>100_10000seq_sampleC2</td>
<td>389</td>
<td>550</td>
<td>9,408</td>
<td>567</td>
</tr>
<tr>
<td>100_10000seq_sampleC3</td>
<td>361</td>
<td>516</td>
<td>9,442</td>
<td>525</td>
</tr>
</tbody>
</table>

367 clusters of sampleA1 are common at least once with another sample.
Hierarchical clustering

Hierarchical classification on Bray Curtis distance

Newick tree available too

Samples distribution tab
Chimera removal tool
Our advice: Removing Chimera after Swarm denoising + Swarm d=3, for saving time without sensitivity loss
What is 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 (for.) and reverse (rev.) primer sequence at each end of the amplicon.

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

Fichot and Norman *Microbiome* 2013 1:10
doi:10.1186/2049-2618-1-10
A smart removal chimera to be accurate

We use a sample cross-validation

Sample A
- a: x1000
- b: x500
- c: x100
- d: x50
- e: x10
- f: x10
- g: x5

“d” is view as chimera by Vsearch
Its “parents” are presents

Sample B
- b: x1000
- d: x500
- h: x100
- i: x50
- f: x10
- e: x10
- g: x5

“d” is view as normal sequence by Vsearch
Its “parents” are absents

⇒ For FROGS “d” is not a chimera
⇒ For FROGS “g” is a chimera, “g” is removed
⇒ FROGS increases the detection specificity
Your Turn! - 5

LAUNCH THE REMOVE CHIMERA TOOL
Exercise 5

Go to « MiSeq contiged » history

Launch the « FROGS Remove Chimera » tool

Follow by the « FROGS ClusterStat » tool on the swarm d1d3 non chimera abundance biom

→ objectives :
  ▪ understand the efficiency of the chimera removal
  ▪ make links between small abundant OTUs and chimeras
FROGS Remove chimera Step 3 in metagenomics analysis: Remove PCR chimera in each sample. (Galaxy Version 1.3.0)

**Sequences file**
- 5: FROGS Clustering swarm: seed_sequences.fasta
  - The sequences file (format: fasta).

**Abundance file**
- BIOM file
  - Select the type of file where the abundance of each sequence by sample is stored.

**Abundance file**
- 6: FROGS Clustering swarm: abundance.biom
  - It contains the count by sample for each sequence.

Execute
Exercise 5

1. Understand the « FROGS remove chimera : report.html»
   a. How many clusters are kept after chimera removal?
   b. How many sequences that represent? So what abundance?
   c. What do you conclude?
Exercise 5

2. Launch « FROGS ClusterStat » tool on non_chimera_abundanced1d3.biom
3. Rename output in summary_nonchimera_d1d3.html
4. Compare the HTML files
   a. Of what are mainly composed singleton ? (compare with precedent summary.html)
   b. What are their abundance?
   c. What do you conclude?

The weakly abundant OTUs are mainly false positives, our data would be much more exact if we remove them
Filters tool
Advise:

Apply filters between “Chimera Removal” and “Affiliation”. Remove OTUs with weak abundance and non-redundant before affiliation.

You will gain time!
Filters

Filters allows to filter the result thanks to different criteria and may be used after different steps of pipeline:

- On the abundance
- **On RDP affiliation**
- **On Blast affiliation**
- On phix contaminant
4 filter sections

Abundance filters
RDP affiliation filters
BLAST affiliation filters
Contamination filter
Input

Fasta sequences and its corresponding abundance biom files

Filter 1: abundance

*** THE FILTERS ON OTUS IN SAMPLES, OTUS SIZE and SEQUENCE PERCENTAGE

Apply filters

If you want to filter OTUs on their abundance and occurrence.

Minimum number of samples

3
Fill the field only if you want this treatment. Keep OTU present in at least this number of samples.

Minimum proportion/number of sequences to keep OTU

0.00005
Fill the field only if you want this treatment. Use decimal notation for proportion (example: 0.01 for keep OTU with at least 1% of all sequences); use integer notation for number of sequence (example: 2 for keep OTU with at least 2 sequences, so remove single singleton).

N biggest OTU

100
Fill the fields only if you want this treatment. Keep the N biggest OTU.
Filter 2 & 3: affiliation
Filter 4: contamination

Soon, several contaminant banks
Your Turn! - 6

LAUNCH DE LA TOOL FILTERS
Exercise 6

Go to history « MiSeq contiged »

Launch « Filters » tool with non_chimera_abundance_1d3.biom, non_chimerad1d3.fasta

Apply 2 filters:

- Minimum proportion/number of sequences to keep OTU: 0.00005*
- Minimum number of samples: 3

→ objective: play with filters, understand their impacts on false-positives OTUs
If Filters fields are « Apply » so you have to fill at one field. Otherwise, galaxy become red!
Exercise 6

1. What are the output files of “Filters”? 
3. How many OTUs have you removed?
4. Build the Venn diagram on the two filters.
5. How many OTUs have you removed with each filter “abundance > 0.005%”, “Remove OTUs that are not present at least in 3 samples”? 
6. How many OTUs do they remain?
7. Is there a sample more impacted than the others?
8. To characterize these new OTUs, do not forget to launch “FROGS Cluster Stat” tool, and rename the output HTML file.
Removing little OTUs (conservation rate =0.005%) and non shared OTU (in less than 2 samples)

On simulated data, singleton are:
~99.9% are chimera and
~0.1% are sequences with sequencing errors, non clustered
Affiliation tool
FROGS Affiliation OTU

Step 4 in metagenomics analysis: Taxonomic affiliation of each OTU's seed by RDPtrools and BLAST (Galaxy Version 0.8.0)

Using reference database

- silva123 16S
- silva123 16S
- silva128 16S
- silva128 16S
- silva128 16S
- silva122 16S
- silva122 16S
- silva122 16S
- greengenes13_5
- midas_S123_2.1.3
- midas_S119_1.20
- pr2_gb203_4.5

Select reference from the list

Also perform RDP assignment?

- Yes
- No

Taxonomy affiliation will be perform thanks to Blast. This can also be done by also with RDP classifier (default No)

OTU seed sequence

- 17: FROGS Filters: sequences.fasta

Abundance file

- 18: FROGS Filters: abundance.biom

OTU abundances (format: BIOM)
1 Cluster = 2 affiliations

Double Affiliation vs SILVA 123 (for 16S, 18S or 23S), SILVA 119 (for 18S) or Greengenes with:

1. RDPClassifier* (Ribosomal Database Project): one affiliation with bootstrap, on each taxonomic subdivision.

Bacteria(100);Firmicutes(100);Clostridia(100);Clostridiales(100);Lachnospiraceae(100);Pseudobutyrivibrio(80); Pseudobutyrivibrio xylanivorans (80)

2. NCBI Blastn+**: all identical Best Hits with identity %, coverage %, e-value, alignment length and a special tag “Multi-affiliation”.

Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Pseudobutyrivibrio;Pseudobutyrivibrio ruminis; Pseudobutyrivibrio xylanivorans

Identity: 100% and Coverage: 100%

Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Qiong Wang, George M.Garrity, James M. Tiedje and James R. Cole

BLAST+: architecture and applications
Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer and Thomas L Madden
### Affiliation Strategy of FROGS

Blastn+ with "**Multi-affiliation**" management

<table>
<thead>
<tr>
<th>V3 – V4</th>
<th>Bacteria</th>
<th>Firmicutes</th>
<th>Clostridia</th>
<th>Clostridiales</th>
<th>Lachnospiraceae</th>
<th>Pseudobutyrivibrio</th>
<th>16S unknown species</th>
</tr>
</thead>
<tbody>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S Butyrivibrio fibrisolvens</td>
</tr>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S rumen bacterium 8</td>
</tr>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S Pseudobutyrivibrio xylanivorans</td>
</tr>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S Pseudobutyrivibrio ruminis</td>
</tr>
</tbody>
</table>

5 identical blast best hits on SILVA 123 databank
Affiliation Strategy of FROGS

Blastn+ with “Multi-affiliation” management

<table>
<thead>
<tr>
<th>V3 – V4</th>
<th>Bacteria</th>
<th>Firmicutes</th>
<th>Clostridia</th>
<th>Clostridiales</th>
<th>Lachnospiraceae</th>
<th>Pseudobutyrivibrio</th>
<th>16S unknown species</th>
</tr>
</thead>
<tbody>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S Butyrivibrio fibrisolvens</td>
</tr>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S rumen bacterium 8</td>
</tr>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S Pseudobutyrivibrio xylanivorans</td>
</tr>
<tr>
<td>V3 – V4</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
<td>Lachnospiraceae</td>
<td>Pseudobutyrivibrio</td>
<td>16S Pseudobutyrivibrio ruminis</td>
</tr>
</tbody>
</table>

FROGS Affiliation: Bacteria | Firmicutes | Clostridia | Clostridiales | Lachnospiraceae | Pseudobutyrivibrio | Multi-affiliation
Your Turn! – 7

LAUNCH THE « FROGS AFFILIATION » TOOL
Exercise 7.1

Go to « MiSeq contiged » history

Launch the « FROGS Affiliation » tool with

- SILVA 123 or 128 16S database
- FROGS Filters abundance biom and fasta files (after swarm d1d3, remove chimera and filter low abundances)

→ objectives :
  - understand abundance tables columns
  - understand the BLAST affiliation
FROGS Affiliation OTU

OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation
Exercise 7.1

1. What are the « FROGS Affiliation » output files?
2. How many sequences are affiliated by BLAST?
3. Click on the « eye » button on the BIOM output file, what do you understand?
4. Use the Biom_to_TSV tool on this last file and click again on the “eye” on the new output generated. What do the columns? What is the difference if we click on case or not? What consequence about weight of your file?
Exercise 7.1

5. Understand Blast affiliations - Cluster_2388 (affiliation from silva 123)

<table>
<thead>
<tr>
<th>blast_subject</th>
<th>blast_evalue</th>
<th>blast_len</th>
<th>blast_perc_query_coverage</th>
<th>blast_perc_identity</th>
<th>blast_taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>JN880417.1.1422</td>
<td>0.0</td>
<td>360</td>
<td>88.88</td>
<td>99.44</td>
<td>Bacteria; Planctomycetes; Planctomycetacia; Planctomycetales; Planctomycetaceae; Telmatocola; Telmatocola sphagniphila</td>
</tr>
</tbody>
</table>
Blast JN880417.1.1422 vs our OTU

OTU length: 405

Excellent blast but no matches at the beginning of OTU.
Blast columns

OTU_2 seed has a best BLAST hit with the reference sequence AJ496032.1.1410.

The reference sequence taxonomic affiliation is this one.

Evaluation variables of BLAST

<table>
<thead>
<tr>
<th>blast_perc_identity</th>
<th>blast_perc_query_coverage</th>
<th>blast_evalue</th>
<th>blast_aln_length</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>411</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>419</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>427</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>426</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>427</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>419</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>401</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>421</td>
</tr>
<tr>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>404</td>
</tr>
</tbody>
</table>

Convert to TSV

FROGS BIOM to TSV

- Abundance file
- Sequences file
- tsd_file (tabular)
- multi_affi_file (tabular)
Focus on “Multi-”

Cluster 1 has 5 identical blast hits, with different taxonomies as the species level

Observe line of Cluster 1 inside abundance.tsv and multi_hit.tsv files, what do you conclude?

Cluster 1 has 5 identical blast hits, with different taxonomies as the species level.
Focus on “Multi-”

*(affiliation from silva 123)*

Observe line of Cluster 11 inside abundance.tsv and multi_hit.tsv files, what do you conclude?

Cluster_11 has 2 identical blast hits, with identical species but with different strains (strains are not written in our data)
Focus on “Multi-”

(affiliation from silva 123)

Observe line of Cluster 43 inside abundance.tsv and multi_hit.tsv files, what do you conclude?

Cluster 43 has 2 identical blast hits, with different taxonomies at the genus level
### Back on Blast parameters

#### Evaluation variables of BLAST
Blast variables : e-value

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.

The lower the E-value, or the closer it is to zero, the more "significant" the match is.
Blast variables: blast_perc_identity

Identity percentage between the Query (OTU) and the subject in the alignment (length subject = 1455 bases)

Query length = 411
Alignment length = 411
0 mismatch
-> 100% identity
Blast variables: blast_perc_identity

Identity percentage between the Query (OTU) and the subject in the alignment (length subject = 1455 bases)

Query length = 411
Alignment length = 411
26 mismatches (gaps included) -> 94% identity
Blast variables:

**blast_perc_query_coverage**

Coverage percentage of alignment on query (OTU)

<table>
<thead>
<tr>
<th>Score</th>
<th>Expect</th>
<th>Identities</th>
<th>Gaps</th>
<th>Strand</th>
<th>Query</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>750 bits(411)</td>
<td>0.0</td>
<td>411/411(100%)</td>
<td>0/411(0%)</td>
<td>0/+1</td>
<td>Query 1</td>
<td>Subject 1</td>
</tr>
<tr>
<td>100% coverage</td>
<td>169</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example alignment:

- **Query 1**: TGGGAAATATTGCACAATTGGGGAACCCGATGCTCAGGCACGAGCCTGGGATGACGG
- **Subject 1**: TGGGAAATATTGCACAATTGGGGAACCCGATGCTCAGGCACGAGCCTGGGATGACGG

Query length = 411

100% coverage
Blast variables : blast-length

Length of alignment between the OTUs = “Query” and “subject” sequence of database

<table>
<thead>
<tr>
<th>OTU</th>
<th>Coverage %</th>
<th>Identity %</th>
<th>Length alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU1</td>
<td>100</td>
<td>98</td>
<td>400</td>
</tr>
<tr>
<td>OTU2</td>
<td>100</td>
<td>98</td>
<td>500</td>
</tr>
</tbody>
</table>

More mismatches/gaps
Who have already used RDP previously?

Optional and not in our guideline
How works RDP?

Query

Words 8 letters
Words frequency

Databank

Words 8 letters
Words frequency

Compare words frequencies

Affiliation
How works RDP?

**Result:**
Bacteria(100) ; Genus_A(50) ; Sp1(25)
The dysfunctions of RDP?

Databank

Root

Bacteria

Genus_A

Sp 1

Genus_B

Sp 4

Sp 5

Eukaryota

Genus_C

Sp 7

Species 2 is removed

OTU query

Result:

?
The dysfunctions of RDP n°1?

Databank

Root

Bacteria

Genus_A

Genus_B

Eukaryota

Genus_C

Sp 1

Sp 4

Sp 7

Sp 5

OTU query

Result:

Bacteria(100); Genus_A(33); sp1(33) OR Bacteria(100); Genus_B(66); sp5(33)

Genus (50) was expected. If seq order was inversed in banks, so we got the second statistics.
The dysfunctions of RDP n°2?

Databank
- Root
  - Bacteria
    - Genus_A: 900 species
  - Eukaryota
    - Genus_B: 100 species
    - Genus_C
      - Sp 7

Result:
- Bacteria(100); Genus_A(90); spX(0.1) OR Bacteria(100); Genus_B(10); spX(0.1)

Many species in one genus and little in the other: So, RDP can give very different results

Influenced by heterogeneity in last ranks
The dysfunctions of RDP n°3?

Databank

Root

Bacteria

Genus_A

Genus_B

Eukaryota

Genus_C

Sp 1
Sp 2
Sp 3

Sp 4
Sp 5
Sp 6

Sp 7

OTU query

Result:?

Mismatches are at different places.
The dysfunctions of RDP n°3?

Si le mismatch se fait sur un mot très "significatif" dans le profil de k-mers, RDP ne tombera que rarement sur l'espèce lors du bootstrap. Avec une même distance d'édition (2 mismatches) on peut donc avoir une grande différence de bootstrap pour peu que le mot affecté soit important dans le profil.
Report on abundance table, the multiple identical affiliations

Divergence on the composition of microbial communities at the different taxonomic ranks

<table>
<thead>
<tr>
<th>Taxonomic ranks</th>
<th>Average divergence of the affiliations of the 10 samples (%) 500setA</th>
<th>Average divergence of the affiliations of the 10 samples (%) 100setA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Phylum</td>
<td>0.46</td>
<td>0.41</td>
</tr>
<tr>
<td>Class</td>
<td>0.64</td>
<td>0.50</td>
</tr>
<tr>
<td>Order</td>
<td>0.94</td>
<td>0.68</td>
</tr>
<tr>
<td>Family</td>
<td>1.18</td>
<td>0.78</td>
</tr>
<tr>
<td>Genus</td>
<td>1.76</td>
<td>1.30</td>
</tr>
<tr>
<td>Species</td>
<td>23.87</td>
<td>34.80</td>
</tr>
</tbody>
</table>

Reliable?

Identical V3-V4 solution

Divergence on the composition of microbial communities at the different taxonomic ranks

Report on abundance table, the multiple identical affiliations
<table>
<thead>
<tr>
<th>Taxonomic ranks</th>
<th>Average divergence of the affiliations of the 10 samples (%) 500setA</th>
<th>Average divergence of the affiliations of the 10 samples (%) 100setA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Phylum</td>
<td>0.46</td>
<td>0.41</td>
</tr>
<tr>
<td>Class</td>
<td>0.64</td>
<td>0.50</td>
</tr>
<tr>
<td>Order</td>
<td>0.94</td>
<td>0.68</td>
</tr>
<tr>
<td>Family</td>
<td>1.18</td>
<td>0.78</td>
</tr>
<tr>
<td>Genus</td>
<td>1.76</td>
<td>1.30</td>
</tr>
<tr>
<td>Species</td>
<td>23.87</td>
<td>34.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taxonomic ranks</th>
<th>Median divergence of the affiliations of the 10 samples (%) 500setA</th>
<th>Median divergence of the affiliations of the 10 samples (%) 100setA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Phylum</td>
<td>0.46</td>
<td>0.41</td>
</tr>
<tr>
<td>Class</td>
<td>0.64</td>
<td>0.50</td>
</tr>
<tr>
<td>Order</td>
<td>0.93</td>
<td>0.68</td>
</tr>
<tr>
<td>Family</td>
<td>1.17</td>
<td>0.78</td>
</tr>
<tr>
<td>Genus</td>
<td>1.60</td>
<td>1.00</td>
</tr>
<tr>
<td>Species</td>
<td>6.63</td>
<td>5.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taxonomic ranks</th>
<th>Median divergence of the affiliations of the 10 samples (%) 500setA filter: 0.005% - 505 OTUs</th>
<th>Median divergence of the affiliations of the 10 samples (%) 100setA filter: 0.005% - 100 OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Phylum</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>Class</td>
<td>0.57</td>
<td>0.48</td>
</tr>
<tr>
<td>Order</td>
<td>0.81</td>
<td>0.64</td>
</tr>
<tr>
<td>Family</td>
<td>1.08</td>
<td>0.74</td>
</tr>
<tr>
<td>Genus</td>
<td>1.43</td>
<td>0.76</td>
</tr>
<tr>
<td>Species</td>
<td>1.53</td>
<td>0.78</td>
</tr>
</tbody>
</table>

With the FROGS guideline

Only one best hit

Multiple best hit
Careful: Multi hit blast table is non exhaustive!

- Chimera (multiple affiliation)
- V3V4 included in others
- Missed primers on some 16S during database building
Affiliation Stat
Exercise 7.2

FROGS Affiliations stat (version 1.1.0)

Abundance file:
17: FROGS Affiliation OTU: affiliation.biom
OTUs abundances and affiliations (format: BIOM).

Rarefaction ranks:
Class Order Family Genus Species
The ranks that will be evaluated in rarefaction. Each rank is separated by one space.

Affiliation processed:
FROGS blast
Select the type of affiliation processed. If your affiliation has been processed with an external tool: use 'Custom'.

Execute

FROGS Affiliations stat (version 1.1.0)

Abundance file:
17: FROGS Affiliation OTU: affiliation.biom
OTUs abundances and affiliations (format: BIOM).

Rarefaction ranks:
Class Order Family Genus Species
The ranks that will be evaluated in rarefaction. Each rank is separated by one space.

Affiliation processed:
FROGS rdp
Select the type of affiliation processed. If your affiliation has been processed with an external tool: use 'Custom'.

Execute

Is it adequate on our data? Why?

23: FROGS
Affiliations stat: summary.html
Exercise 7.2

→ objectives:

understand rarefaction curve and sunburst

1. Explore the Affiliation stat results on FROGS blast affiliation.
2. What kind of graphs can you generate? What do they mean?
Samples size ~8500 sequences

The curve continues to rise

The number of sequences per sample is not large enough to cover all of the bacterial families

Available only after AFFILIATION TOOL

Rarefaction tab
Samples size ~85 000 sequences

The curve slows to rise with ~50 000 sequences

With 60 000 sequences, we catch almost all genus of bacteria
### Number of OTUs among their alignment results

<table>
<thead>
<tr>
<th>Coverage</th>
<th>[0% - 50%]</th>
<th>[50% - 80%]</th>
<th>[80% - 90%]</th>
<th>[90% - 95%]</th>
<th>[95% - 100%]</th>
<th>[100%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[100%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>[95% - 100%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>[90% - 95%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[80% - 90%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[50% - 80%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[0% - 50%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

by OTUs

by sequences
### Number of sequences among their alignment results

<table>
<thead>
<tr>
<th>Coverage</th>
<th>0% - 50%</th>
<th>50% - 80%</th>
<th>80% - 95%</th>
<th>90% - 95%</th>
<th>95% - 100%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>[100%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1657</td>
<td>74495</td>
</tr>
<tr>
<td>[95% - 100%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>[50% - 95%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[50% - 90%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[50% - 80%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[0% - 50%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Identity:** 4%
- **Coverage:** 5%
- **Nb sequences:** 1657

**by OTUs**

**by sequences**
TSV to BIOM
After modifying your abundance TSV file you can again:

- generate rarefaction curve
- sunburst

Careful:

- do not modify column name
- do not remove column
- take care to choose a taxonomy available in your multi_hit TSV file
- if deleting line from multi_hit, take care to not remove a complete cluster without removing all "multi tags" in you abundance TSV file.
- if you want to rename a taxon level (ex : genus "Ruminiclostridium 5;" to genus "Ruminiclostridium;"), do not forget to modify also your multi_hit TSV file.
TSV to BIOM

FROGS TSV to BIOM Converts a TSV file in BIOM file. (Galaxy Version 1.0.0)

Abundance TSV File
- 29: FROGS BIOM to TSV: abundance.tsv
  Your FROGS abundance TSV file. Take care to keep intact column name.

Multi_hits TSV File
- 30: FROGS BIOM to TSV: multi_hits.tsv
  TSV file describing multi blast hit.

Extract seed FASTA file
- Yes
- No
  If there is a 'seed_sequence' column, you can extract seed sequence in a separated FASTA file.

Execute
Your Turn! – 8

PLAY WITH TSV_TO_BIOM
Exercise 8

→ objectives: Play with multi-affiliation and TSV_to_BIOM

1. Observe in Multi_hit.tsv and abundance.tsv cluster_8 annotation
2. Observe le diversity diagramm
Exercise 8

3. How to change affiliation of cluster 8 ????
Exercise 8

4. Modify multi_hit.tsv and keep only:

Careful, no quotes around text !!!

5. Upload the new multihit file.
6. Create a new biom with a TSV_to_BIOM tool
7. Launch again the affilation_stat tool on this new biom
8. Observe the diversity diagram
Normalization
Normalization

Conserve a predefined number of sequence per sample:
- update Biom abundance file
- update seed fasta file

May be used when:
- Low sequencing sample
- Required for some statistical methods to compare the samples in pairs
Your Turn! – 9

LAUNCH NORMALIZATION TOOL
Exercise 9

Launch Normalization Tool

1. What is the smallest sequenced samples?
2. Normalize your data from Affiliation based on this number of sequence
3. Explore the report HTML result.
4. Try other threshold and explore the report HTML result
   What do you remark?
FROGS Abundance normalisation (Galaxy Version 1.1.1)

Sequences file
- 17: FROGS Filters: sequences.fasta
  Sequences file to normalize (format: fasta).

Abundance file
- 22: FROGS Affiliation OTU: affiliation.biom
  Abundances file to normalize (format: BIOM).

Number of reads
- 9088
  The final number of reads per sample.

Execute
Or, this number can be chosen according to the rarefaction curve. For example, we can choose the smallest number of sequences that still retain all the genus.
## Composition summary

### Nb OTUs before normalisation

<table>
<thead>
<tr>
<th>Sample</th>
<th>Nb OTU before normalisation</th>
<th>Nb OTU after normalisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>100_10000seq_sampleA1_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleA2_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleA3_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleB1_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleB2_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleB3_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleC1_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleC2_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>100_10000seq_sampleC3_cutadapt</td>
<td>144</td>
<td>135</td>
</tr>
</tbody>
</table>

Showing 1 to 9 of 9 entries

---

Nb OTUs after normalisation

All samples: 135

---
Filters on affiliations
Do not forget, with filter tool we can filter the data based on their affiliation.

- **Abundance filters**
- **RDP affiliation filters**
- **BLAST affiliation filters**
- **Contamination filter**
Exercise 10

1. Apply filters to keep only data with perfect alignment.
2. How many clusters have you keep?
## FROGS Filters

Filters OTUs on several criteria. (Galaxy Version 1.2.0)

### Sequences file

The sequence file to filter (format: fasta).

### Abundance file

The abundance file to filter (format: BIOM).

### Filters on OTUs in Samples, OTU Size and Sequence Percentage

---

#### *** THE FILTERS ON OTUS IN SAMPLES, OTU SIZE and SEQUENCE PERCENTAGE ***

- **No filters**
- **Apply filters**

If you want to filter OTUs on their abundance and occurrence.

If you want to filter OTUs on their taxonomic affiliation produced by RDP.

If you want to filter OTUs on their taxonomic affiliation produced by Blast.

### Maximum e-value (between 0 and 1)

Fill the field only if you want this treatment.

### Minimum identity % (between 0 and 1)

Fill the field only if you want this treatment.

### Minimum coverage % (between 0 and 1)

Fill the field only if you want this treatment.

### Minimum alignment length

Fill the field only if you want this treatment.
Tool descriptions
What it does

FROGS Pre-process filters and dereplicates amplicons for use in diversity analysis.

Inputs/Outputs

Inputs
By sample your sequences and their qualities.

Illumina inputs

Usage: The amplicons have been sequenced in paired-end. The amplicon expected length is inferior than the R1 and R2 length. R1 and R2 can be merge by the common region.
Files: One R1 and R2 by sample (format FASTQ)
Examples: spA_R1.fastq.gz, spA_R2.fastq.gz, spB_R1.fastq.gz, spB_R2.fastq.gz

OR

Usage: The single and sequencing cover all the amplicons or the R1 and R2 have already been overlapped.
Files: One sequence file by sample (format FASTQ).
Examples: spA.fastq.gz, spB.fastq.gz

454 inputs

Files: One sequence file by sample (format FASTQ)
Examples: spA.fastq.gz, spB.fastq.gz

These files must be added sample by sample or provide in an archive file (tar.gz).
Remark: In an archive if you use R1 and R2 files the names must end with _R1 and _R2.
Outputs

**Sequence file** (dereplicated.fasta):

Only one file with all samples sequences (format **FASTA**). These sequences are dereplicated: strictly identical sequence are represented only one and the initial count is kept in count file.

**Count file** (count.tsv):

This file contains the count of all uniq sequences in each sample (format **TSV**).

**Summary file** (excluded_data.html):

This file presents the ordered filters and the number of sequences passing these (format **HTML**).
### How it works

<table>
<thead>
<tr>
<th>Steps</th>
<th>Illumina</th>
<th>454</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>For uncorrected data: contig read1 and read2 with a maximum of 10% mismatch in the overlapped region (FLAG)</td>
<td>/</td>
</tr>
<tr>
<td>2</td>
<td>Filter contig sequence on its length which must be between Minimum amplicon size and Maximum amplicon size</td>
<td>/</td>
</tr>
<tr>
<td>3</td>
<td>Remove sequences where the two primers are not present and remove primers sequence (cutadapt). The primer search accept 10% of differences</td>
<td>Remove sequence where the two primers are not present, remove primers sequence and reverse complement the sequences with strand - (cutadapt). The primer search accept 10% of differences</td>
</tr>
<tr>
<td>4</td>
<td>Filter sequences on its length and with ambiguous nucleotides</td>
<td>Filter sequences on its length, with ambiguous nucleotides, with at least one homopolymer with size &gt;7nt and with distance between two poor qualities (&gt; 10) of ≤ 10 nt</td>
</tr>
<tr>
<td>5</td>
<td>Dereplicate sequences</td>
<td>Dereplicate sequences</td>
</tr>
</tbody>
</table>
Advises/details on parameters

Primers parameters
The primers must be provided in 5' to 3' orientation.
Example:

5' ATGCC GTCGTCGTTAAATGC ATTTCCG 3'
Value for parameter 5' primer: ATGCC
Value for parameter 3' primer: ATTTCCG

Amplicons sizes parameters
The two following images show two examples of perfect values for sizes parameters.
Workflow creation
Your Turn! – 11

CREATE YOUR OWN WORKFLOW!
Exercise 11

<table>
<thead>
<tr>
<th>Name</th>
<th># of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>formation workflow</td>
<td>9</td>
</tr>
<tr>
<td>demoNEM2015 workflow</td>
<td>9</td>
</tr>
<tr>
<td>PROBS_v1.0.06.05.2015</td>
<td>10</td>
</tr>
</tbody>
</table>

Workflows shared with you by others
No workflows have been shared with you.

Other options
Configure your workflow menu
Exercise 11

3. Workflow Name:
   Unnamed workflow

4. Workflow Annotation:
   A description of the workflow; annotation is shown alongside shared or published workflows.

Create
Exercise 11
For each tool, think to:

- Fixe parameter?
For each tool, think to:

- Fixe parameter?
- Automatically rename output files
For each tool, think to:
- Fixe parameter?
- Automatically rename output files
- Hide intermediate files?
For each tool, think to:

- Fixe parameter?
- Automatically rename output files
- Hide intermediate files?
For each tool, think to:

- Fixe parameter?
- Automatically rename output files
- Hide intermediate files?
Download your data
You have to download one per one your files
FROGS BIOM to Standard BIOM
FROGS biom to standard Biom

This step is required to run R
Some figures
Some figures - Fast

<table>
<thead>
<tr>
<th>NB SEQ</th>
<th>TIME with complete pipeline without Filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 000</td>
<td>40 min</td>
</tr>
<tr>
<td>400 000</td>
<td>4 hrs</td>
</tr>
<tr>
<td>3 500 000</td>
<td>2 days</td>
</tr>
<tr>
<td>10 000 000</td>
<td>5 days</td>
</tr>
</tbody>
</table>
Speed on real datasets

9,600,000 sequences of a complete MiSeq run

Preprocess: 9,300,000 sequences ~ 15 min

Swarm clustering: 680,000 clusters ~ 10 hrs

Chimera removal: 556,700 non-chimeric clusters ~ 15 min

Filtering*: 556,200 OTUs
  *Filter OTU abundances at 0.005%

PhiX removal ~ 8 min

RDP affiliation ~ 25 min

Blast affiliation ~ 5 min

FROGS: 500 OTUs ~ 11 hours
Simulated datasets, for testing FROGS’ Accuracy

- 500 species, covering all bacterial phyla
- Power Law distribution of the species abundances
- Error rate calibrated with real sequencing runs
- 20% chimeras
- 10 samples of 100 000 sequences each (1M sequences)

Simulated dataset : 1M sequences
↓
SWARM : 109 000 clusters
↓
VSEARCH: 21 000 clusters
↓
filters : 0.005%
505 OTUs
FROGS’ Accuracy

- $1.10^{+8}$ synthetic sequences were treated with FROGS, UPARSE and MOTHUR, QIIME, with their guidelines, to compare their performances
- 20, 100, 200, 500 or 1000 different species
- power law or a uniform distribution
- 5 to 20% of chimera

→ Divergence on the composition of microbial communities at the different taxonomic ranks
Divergence on the composition of microbial communities at genus rank

V3V4 Power Law

Affiliations divergence

Divergence (%)

0 25 50 75 100

20 100 200 500 1000

frogs uparse (SOP) mothur (SOP) qiime (SOP)
Affiliations divergence

Divergence on the composition of microbial communities at genus rank
FROGS performs as well as or better than UPARSE, MOTHUR and QIIME in most settings. The only condition in which FROGS does worse than UPARSE and MOTHUR is small community size (20).
The results of non-parametric paired tests (signed rank test) of Affiliation divergence on simulated data from UTAX QIIME (MA) with large communities (size > 200) with uniform abundance using the V4 region is better than FROGS. The differences, although significant, are small in that case: less than 2 percentage points in all cases and most marked at the Genus level where the divergences of both FROGS and QIIME (MA) are already quite moderate (6~10%).
Conclusions
Why Use FROGS?

- User-friendly
- Fast
- 454 data and Illumina data
  - sequencing methods change but same tool
  - easier for comparisons
- Clustering without global threshold and independent of sequence order
- New chimera removal method (Vsearch + cross-validation)
- Filters tool
- Multi-affiliation with 2 taxonomy affiliation procedures
- Cluster Stat and Affiliation Stat tools
- A lot of graphics
- Independant tools
How to cite FROGS

In waiting for the publication:

Pipeline FROGS on http://sigenae-workbench.toulouse.inra.fr/

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


To contact

FROGS:

frogs@toulouse.inra.fr

Galaxy:

sigenae-support@listes.inra.fr

Newsletter – demande d’abonnement:

mailto:sympa@listes.inra.fr?subject=sub%20frogs-newsletter
frogs-newsletter-request@listes.inra.fr
Next training sessions

$3^{rd}$ to $6^{th}$ July 2017 4 days

0.5 Galaxy day
2 FROGS days
1.5 Statistics phyloseq day (under R)