

Programme de la journée

Sarah Maman, Ibouniyamine Nabihoudine

- **Initiation à l'utilisation de Galaxy**

Maria Bernard, Laurent Cauquil

- **Etape 1 : Analyse de séquences 16S :
Preprocessing des lectures**
- **Etape 2 : Alignement des lectures et filtre**
- **Etape 3 : Analyse taxonomique des séquences**
- **Etape 4 : Construction des OTUs et analyse
taxonomique**



Analyse d'ARN 16S bactériens

Séquençage 454 et MiSeq

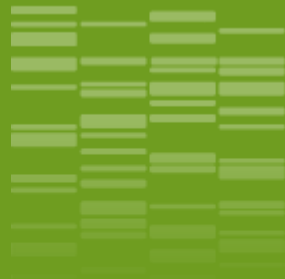


Maria Bernard
Laurent Cauquil



geno
toul
bioinfo





01 Introduction



Introduction

Objectifs : Etude de la communauté bactérienne par séquençage haut-débit

- Mise en évidence du polymorphisme d'un gène entre les espèces microbiennes
- Gène choisi : gène codant pour l'ARNr 16S
 - Molécule ubiquitaire
 - Régions conservées (production d'amplicons)
 - Régions hypervariables (permet de différencier les espèces)
- Calcul des courbes de raréfaction par échantillon
- Indices de diversité (Ace, Shannon...), coverage

Comparison of the microbial community structures of untreated wastewaters from different geographic locales.

Shanks OC, Newton RJ, Kelty CA, Huse SM, Sogin ML, McLellan SL.

Appl Environ Microbiol. 2013 May;79(9):2906-13. doi: 10.1128/AEM.03448-12. Epub 2013 Feb 22.

Méthode



Prélèvement d'échantillons

Extraction d'ADN

Amplification par PCR

Séquençage

Nettoyage des séquences

Affiliation taxonomique

Regroupement en OTU
(Operational Taxonomic Unit)

Objectifs de la formation

A partir de données de séquençage 454/Miseq sur une région d'intérêt de l'ARN 16S :

- Identification des taxonomies de chaque lecture
- Quantification des populations bactériennes pour chaque échantillon
- Recherche des différents « Operational Taxonomic Unit » (OTU)
- Quantification des OTUs par échantillon
- Calcul des courbes de raréfaction par échantillon
- Indices de diversité (Ace, Shannon...), estimation de la couverture de Good.

Outils

- Galaxy : interface web
 - Le projet initial : <http://galaxyproject.org/>
 - L'instance de Toulouse, Sigenaie/Genotoul <http://sigenaie-workbench.toulouse.inra.fr/>
- Mothur / Swarm: analyse métagénomique
 - http://www.mothur.org/wiki/Main_Page
Schloss P.D. et al. Appl Environ Microbiol, 2009. 75(23):7537-41
 - <https://github.com/torognes/swarm>
Mahé F et al. PeerJ PrePrints 2:e386v1
- Krona : « Hierarchical data browser », pour la visualisation des populations et de leur quantité
 - <http://sourceforge.net/projects/krona/>
Ondov BD et al. BMC Bioinformatics. 2011 Sep 30; 12(1):385.

Pré requis de la formation

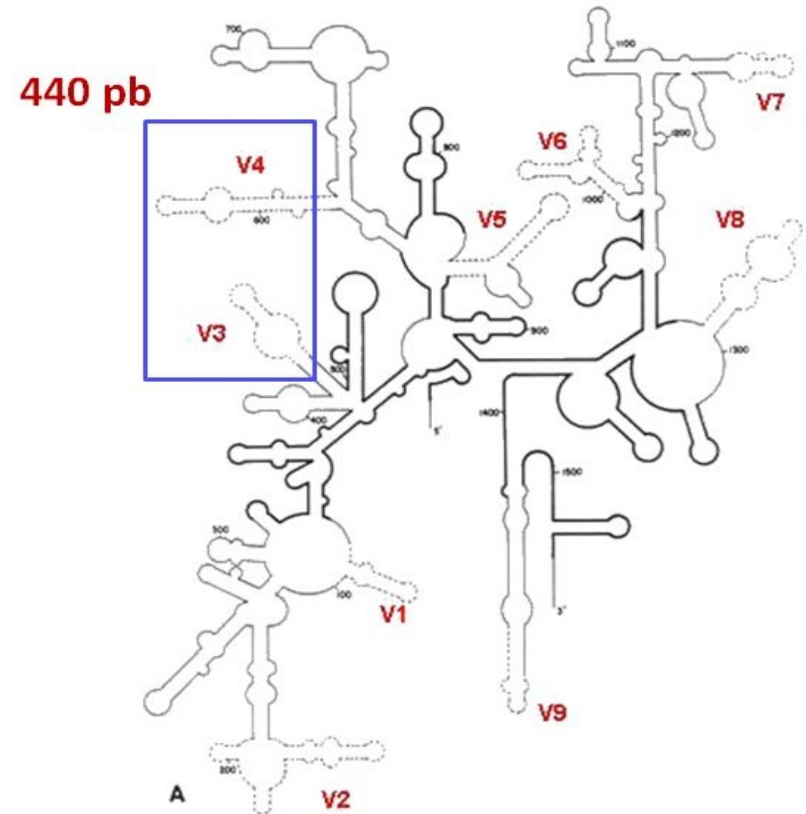
- Base pour l'utilisation de Galaxy
 - Formation initiation à Galaxy
- ou
- Suivez la formation «Galaxy» disponible sur le e-learning: <http://sig-learning.toulouse.inra.fr/>

Données - Construction des amplicons

Détermination des populations bactériennes présentes dans un échantillon:

Sélection d'une ou plusieurs régions variables grâce à un couple de primer

Chakravorty 2007,
"A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria"
Klindworth 2012,
"Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies"



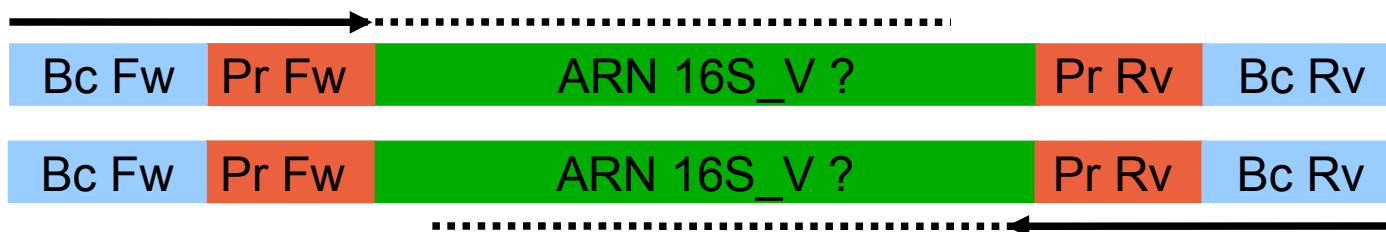
Données - Construction des amplicons

- Critère de définition des régions:
 - Taille des séquences obtenues au séquençage
 - Communauté étudiée
 - ❖ bactéries, archaea, eucaryotes
 - ❖ groupe taxonomique particulier (phylum, genre, espèce...)

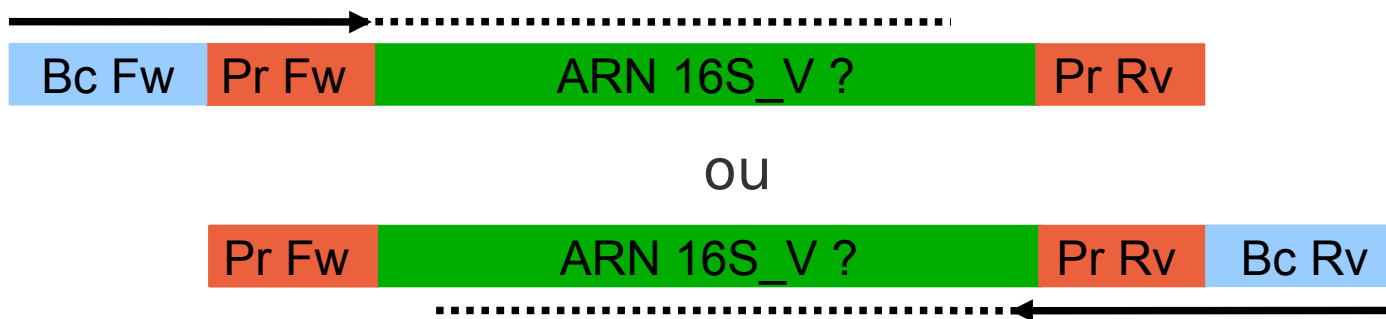
Données - Construction des amplicons (454)

On peut ensuite multiplexer les séquences (utilisation de barcodes) pour traiter plusieurs échantillons dans un même run (diminution des coûts)

- Séquençage plate-forme Get



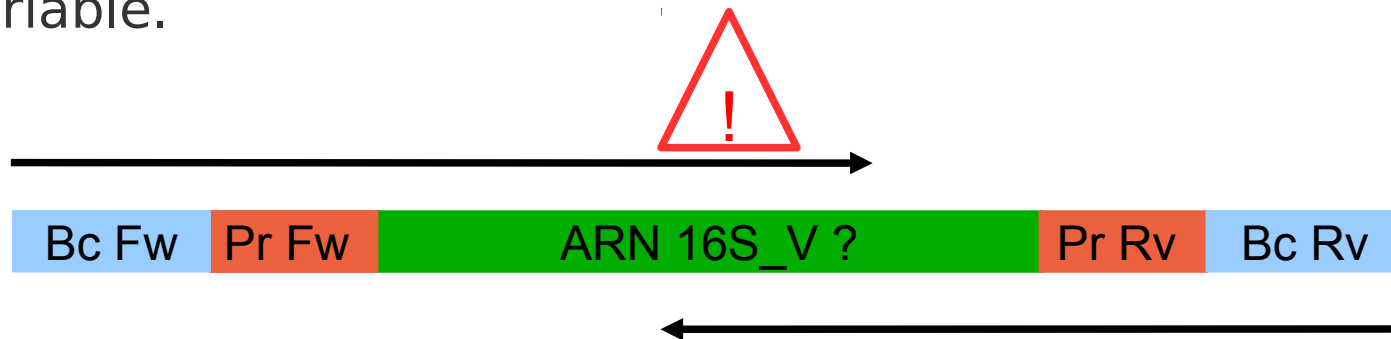
- Séquençage Texas (SCOT)



Données - Construction des amplicons (MiSeq)

Pour le séquençage MiSeq on utilise le séquençage paire-end 2 x 250 bases avec une zone de chevauchement entre les 2 lectures (au moins 10 bases)

Cette zone de chevauchement ne doit pas avoir de région variable.



Données - Construction des amplicons

- Bilan du séquençage :
 - 454 : 2 fichiers par run.

Fichier de lectures au format fasta : run_name.fasta	>F11Fcsw_92 AGAGAGCAAGTGCATGCTGCCTCCCGTAGG...
Fichier de qualité de séquençage au format fasta : run_name.qual	>F11Fcsw_92 40 40 40 40 40 40 40 40 40 40 40 40 37 37 37 37 37 37 40 40 40 40 40 40 40 40 40 40 40 40...

Un fichier fasta contient les informations de chaque lectures sur 2 lignes.

- La première ligne commence par « > » et contient l'identifiant de la séquence
- La seconde ligne contient la séquence proprement dite ou la qualité associée à chaque base

Données - Construction des amplicons

- Bilan du séquençage :
 - MiSeq : 2 fichiers par run et par échantillon.

Fichier de lectures au format fastq : sample_name_run_name_R1.fastq	@M00967:43:000000000-A3JHG:1:1101:18327:1699 1:N:0:188 NACGGAGGATGCGAGCGTTATCCGG.. + #>>AABABBFFFGGGGGGGGGGGGGG...
Fichier de lectures au format fastq : sample_name_run_name_R2.fastq	@M00967:43:000000000-A3JHG:1:1101:18327:1699 2:N:0:188 CCTGTTTGATCCCCGCACTTTCGTG.. + BABBBFFFFFFFFFGEggggggggGhg...

Un fichier fastq contient les informations de chaque lectures sur 4 lignes.

- La première ligne commence par « @ » et contient l'identifiant de la séquence
- La seconde contient la séquence proprement dite
- La troisième est un « + »
- La quatrième contient les scores de qualité de chaque base

Données : échantillon de données test

- Séquençage 454 :

Costello et al, ont publié en 2009, <http://www.sciencemag.org/content/326/5960/1694>, une étude sur les différences de populations bactériennes entre différents tissus du corps humain et entre différents individus. Pour cela ils ont prélevé 27 tissus différents chez 9 individus et ont séquencé les régions V1-V2 (27f-338r) de l'ADN 16S sur une plateforme Roche 454.

Données test : 24 échantillons d'ADN 16S V1-V2 (4 tissus de 3 femmes et 3 hommes)

- Séquençage MiSeq :

Schloss et al (développeur de mothur) est en cours de publication, <http://aem.asm.org/content/early/2013/06/17/AEM.01043-13>, d'un protocole d'analyse d'amplicon sur plateforme de séquençage Illumina MiSeq. Dans cette étude ils s'intéressent à l'analyse des effets du microbiome de l'intestin sur la santé. Pour cela ils ont prélevé les fèces de souris à différents temps après sevrage et ont analysé les régions V3-V4, V4, V4-V5 de l'ADN16S.

Données test : 10 échantillons d'ADN 16S V4 (5 temps juste après sevrage et 5 temps en fin d'expérience) d'une souris.

Données : un fichier de configuration

- Un fichier de configuration tabulé décrivant la construction des amplicons:
 - Multiplexage simple

sample_name	tag_f	primer_f	tag_r	primer_r	16S_domain	extra
F11Fcsw	NONE	NONE	AGAGAGCAAGTG	CATGCTGCCTCCCGTAGGAGT	16S_V ?	...
F12Fcsw	NONE	NONE	AGTAGTATCCTC	CATGCTGCCTCCCGTAGGAGT	16S_V ?	...
...

- Multiplexage double

sample_name	tag_f	primer_f	tag_r	primer_r	16S_domain	extra	...
MgArd0001	ACAGCGT	AYTGGGYDTAAAGV G	ACGTACA	TACCVGGGTATCTAATCC	16S_V4	Myodes	...
MgArd0002	ACAGCGT	AYTGGGYDTAAAGV G	ACGTCAG	TACCVGGGTATCTAATCC	16S_V4	Myodes	...
...

Données : les bases de référence

- Base de données de référence

Cette référence est constituée de deux fichiers :

- Un alignement d'ARN 16S (contenant à minima votre région d'intérêt). Le fichier doit être au format fasta, l'alignement au format ARB (<http://www.arb-home.de/home.html>)
- Les taxonomies associées à chacune des séquences de références :

Sequence ID	Taxonomy
U87775.1	Bacteria;Alphaproteobacteria;Rhizobiales;Azorhizobium_et_rel.;Methylobacterium_et_rel.;Bosea;

Données : les bases de référence

Limite des bases de données :

On ne peut identifier que ceux qu'on a déjà vu !

Suivant la base de référence utilisée l'affiliation taxonomique pourra être plus ou moins précise suivant les groupes (classe, ordre, famille, genre, espèce) et les embranchements (Firmicutes, Bacteroidetes...)

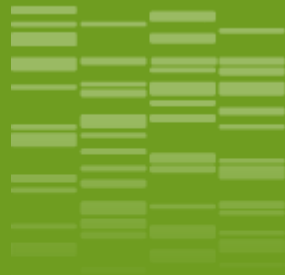
Base disponibles :

Silva, RDP, Greengenes, NCBI (Mothur fourni une version des séquences Silva correctement formatée avec les assignations taxonomiques des différentes bases, [ici](#))

Possibilité d'utiliser une bases personnalisée

Données : bilan des entrées

- Vous avez donc à fournir :
 - Un/des fichiers de séquences (fasta ou fastq) multipléxées pour le 454, multipléxées ou non mais non contiguées pour le MiSeq
 - Un fichier de configuration décrivant la construction des amplicons
 - 2 fichiers d'ADN 16S de référence : les séquences alignées et les taxonomies associées.



02

Pipeline d'analyse mothur Théorie



Pipeline mothur



1_Pre-process

Démultiplexage et/ou Contigage des lectures
Sélection des séquences uniques

Pipeline mothur



1_Pre-process

Démultiplexage et/ou Contigage des lectures
Sélection des séquences uniques



2_Alignment

Alignement des séquences sur une base 16S de référence
Sélection des séquences sur la région d'intérêt
Filtres des erreurs de séquençage
Sélection des alignements uniques

Pipeline mothur



1_Pre-process

Démultiplexage et/ou Contigage des lectures
Sélection des séquences uniques



2_Alignment

Alignement des séquences sur une base 16S de référence
Sélection des séquences sur la région d'intérêt
Filtres des erreurs de séquençage
Sélection des alignements uniques



3_Sequences_analysis

Assignation taxonomique de
chaque séquence

Pipeline mothur



1_Pre-process

Démultiplexage et/ou Contigage des lectures
Sélection des séquences uniques



2_Alignment

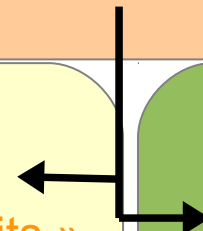
Alignement des séquences sur une base 16S de référence
Sélection des séquences sur la région d'intérêt
Filtres des erreurs de séquençage
Sélection des alignements uniques

4_OTUs_analysis

Construction des « Operational Taxonomic Units »
Calcul de la diversité alpha
Assignation taxonomique des OTUs

3_Sequences_analysis

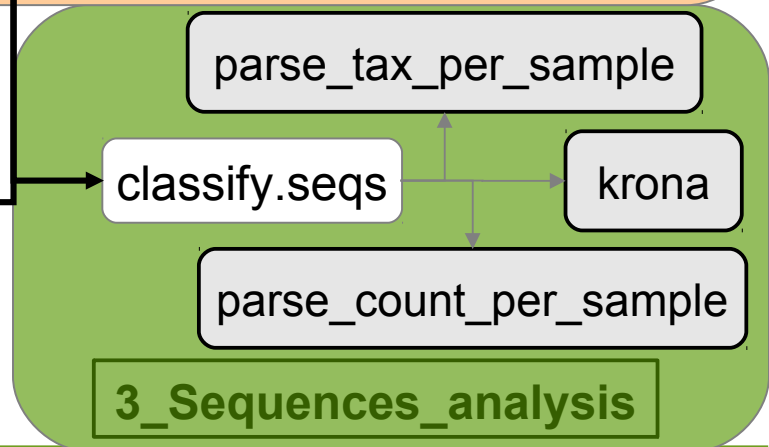
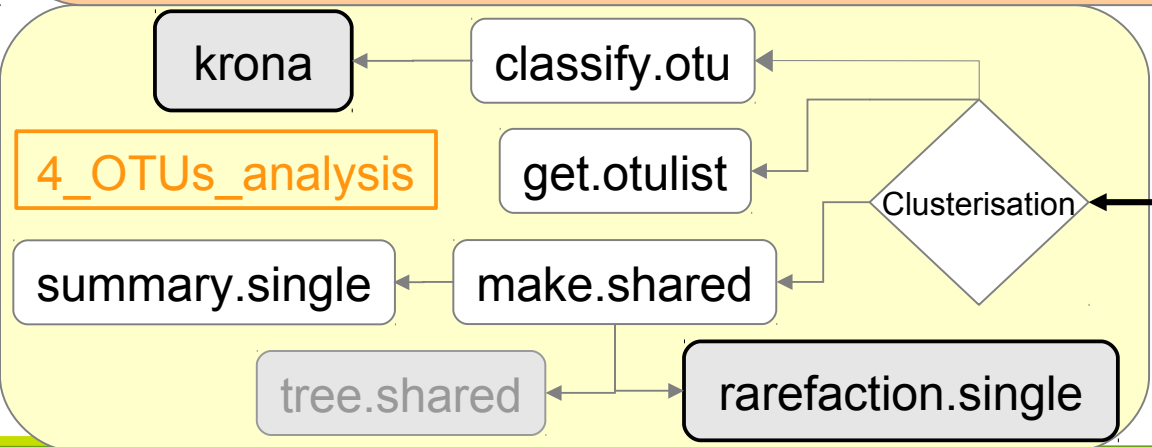
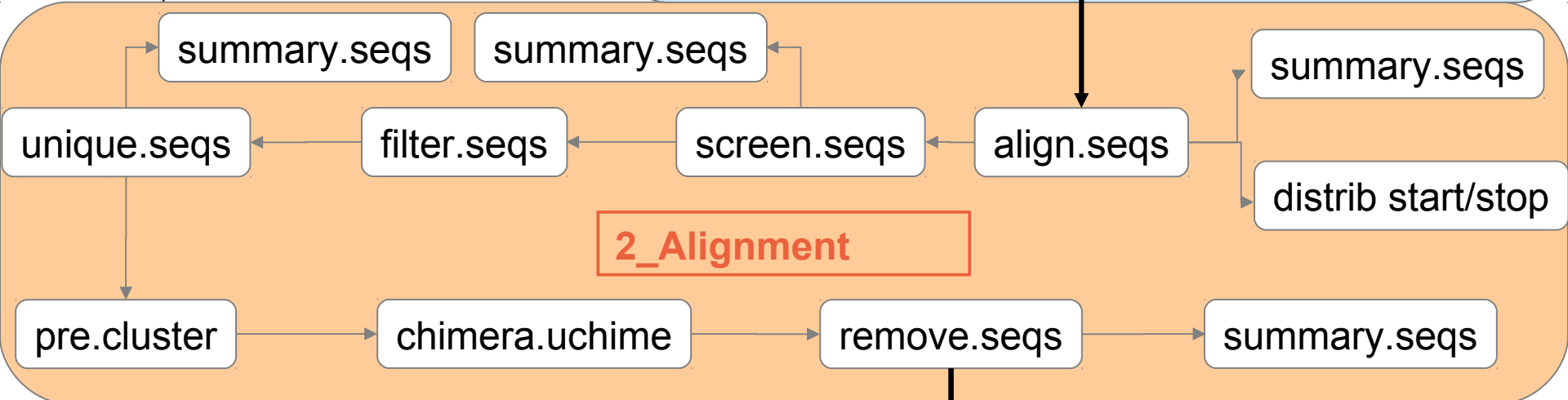
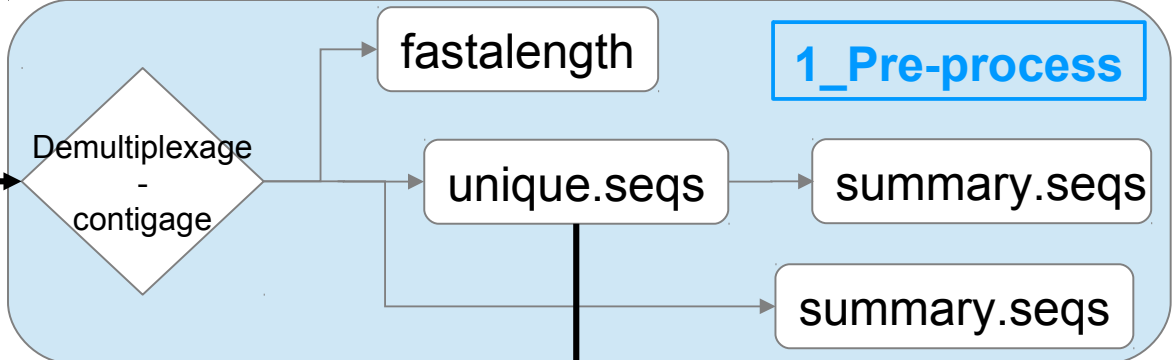
Assignation taxonomique de
chaque séquence



Pipeline mothur



Fasta
Fastq



Pipeline mothur

Vous avez vu qu'il y a un grand nombre d'étapes pour arriver à générer les assignations taxonomiques d'un ensemble d'échantillons.

Dans le pipeline Galaxy que l'on vous propose, ces étapes sont résumées en 4 grands modules que nous allons maintenant détailler.

Sous Galaxy, ces modules sont présents dans la section « 8 - Trainings », « Metagenomics Mothur » .

8 - TRAININGS

Reads alignment and SNP calling

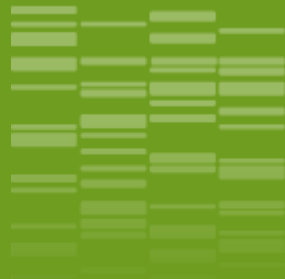
RNA-Seq

sRNAseq

SNP annotation

Metagenomics Mothur 454 and MiSeq

- *454* 1 Pre process (e-learning available)
- *MiSeq* 1 Pre process
- *2 Alignment (e-learning available)
- *3 Sequences analysis (e-learning available)
- *454* 4 OTUs analysis (e-learning available)
- *MiSeq* 4 OTUs analysis (e-learning available)



03

**Pipeline d'analyse mothur
Galaxy / Etape 1
Preprocessing des lectures**

Pipeline d'analyse mothur : Données 454

- Chargement des fichiers d'entrées













Proposition de base de référence fourni par mothur ici :

<http://www.mothur.org/w/images/9/98/Silva.bacteria.zip>

1 - UPLOAD YOUR DATA

Get Data

- * [Upload File](#)
- [UCSC Main](#) table browser
- [UCSC Test](#) table browser
- [UCSC Archaea](#) table browser
- * [Upload local file from filesystem path](#) Upload data to history without copying on server (e-learning available)
- [BioMart](#) Central server
- [Get Microbial Data](#)
- * [TESTS](#) [Save my data](#) on Genotoul
- * [EBI SRA](#) [ENA](#) [SRA](#)

4: stool.forward.fasta	  
3: costello.config.txt	  
2: silva.bacteria.rdp.tax	  
1: silva.bacteria.fasta	  

Les lectures et le fichier de configuration

Les fichiers des références

Pipeline d'analyse mothur : Données MiSeq

- Chargement des fichiers d'entrées

1 - UPLOAD YOUR DATA

Get Data

- * [Upload File](#)
- [UCSC Main](#) table browser
- [UCSC Test](#) table browser
- [UCSC Archaea](#) table browser
- * [Upload local file from filesystem path](#) Upload data to history without copying on server (e-learning available)
- [BioMart](#) Central server
- [Get Microbial Data](#)
- * [TESTS](#) [Save my data](#) on Genotoul
- * [EBI SRA](#) ENA SRA

3: Schloss.config.txt	👁️ ✎ ✕
2: silva.bacteria.fasta	👁️ ✎ ✕
1: silva.bacteria.rdp.tax	👁️ ✎ ✕

le fichier de configuration

Les fichiers des références

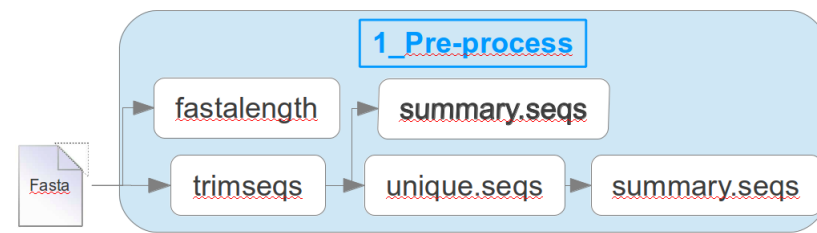
Les lectures seront chargés en indiquant le chemin du dossier genotoul dans lequel vous les avez stockées.



Outil « 1_Pre-process »

Pipeline mothur

Etape 1 : 454 preprocess



Les lectures sont sous la forme :



La première étape du pipeline consiste à démultipléxer les données, et à supprimer les primers : outil « `*454*_1_Pre-process` »

Metagenomics Mothur 454 and MiSeq

- ***454* 1 Pre process** (e-learning available)
- *MiSeq* 1 Pre process
- *2 Alignment (e-learning available)
- *3 Sequences analysis (e-learning available)
- *454* 4 OTUs analysis (e-learning available)
- *MiSeq* 4 OTUs analysis (e-learning available)

***454* 1 Pre process (version 1.0.0)**

FASTA file of 454 reads:
7: 1_preprocess_454_bilan.html ▾

Configuration file (tabular format):
7: 1_preprocess_454_bilan.html ▾

Number of barcode per read mandatory to recognize a sample:
1 ▾

Number of primer mandatory per read : 1 per read (sequencing process may be not complete over the 16S RNA) or 2 (the fragment is completely sequenced):
1 ▾

Number of mismatch authorized in the barcode:
0

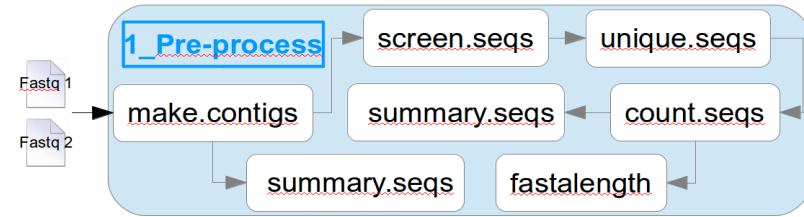
Number of mismatch authorized in the primer:
0

Minimum length of read after trimming barcode and primer (MANDATORY):
150

Execute

Pipeline mothur

Etape 1 : MiSeq preprocess



Les lectures sont sous la forme :



La première étape du pipeline consiste à démultiplexer les données, et à contiguer les lectures: outil « ***MiSeq* 1_Pre-process** »

Metagenomics Mothur 454 and MiSeq

- *454* 1 Pre process (e-learning available)
- ***MiSeq* 1 Pre process**
- *2 Alignment (e-learning available)
- *3 Sequences analysis (e-learning available)
- *454* 4 OTUs analysis (e-learning available)
- *MiSeq* 4 OTUs analysis (e-learning available)

MiSeq 1 Pre process (version 1.0.0)

Configuration file (tabular format):

1: config.txt

Extension of your input files:

fastq

Perform barcode trimming ? You data are multiplexed : Yes, else No:

No

Your path to access to your demultiplexed read files on Genotoul (example : /work/user/projet/):

/home/mbernard/wc

Perform pair assembling ? (for now only unassembled read are accepted. So Yes):

Yes

Maximum number of ambiguous contigs bases:

0

Maximum contigs length (MANDATORY):

280

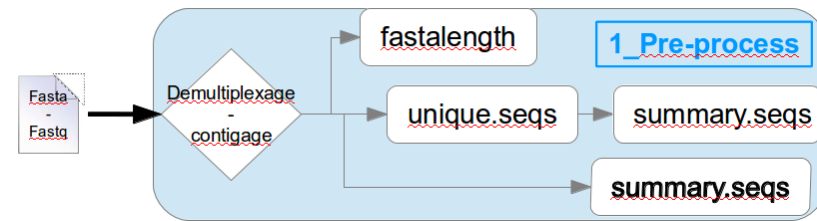
Execute





Résultats de « 1_Pre-process »

Pipeline mothur

Etape 1 : preprocessing



Résultats

7:	  
<u>1_preprocess_454_bilan.html</u>	
6:	  
<u>uniques.reads.count.table</u>	
5:	  
<u>unique.reads.fasta</u>	

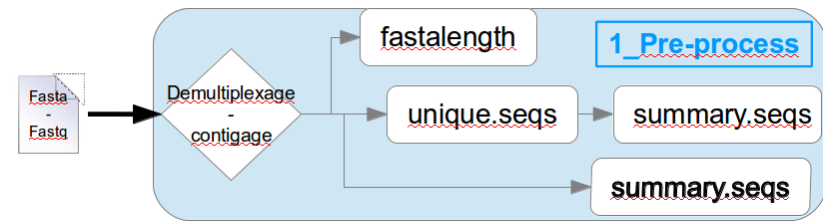
- Fichier HTML: fichier html présentant les résultats du module
- Fichier « count » : fichier tabulé contenant le nombre d'occurrences de chaque séquence unique dans chaque échantillon

Representative_Sequence	Total	F11Fcsw	F12Fcsw
F11Fcsw_6529	1568	38	17
F21Fcsw_12128	1764	0	0

- Fichier fasta : fichier fasta contenant l'ensemble des séquences uniques de tous échantillons confondus

Pipeline mothur

Etape 1 : 454 preprocess

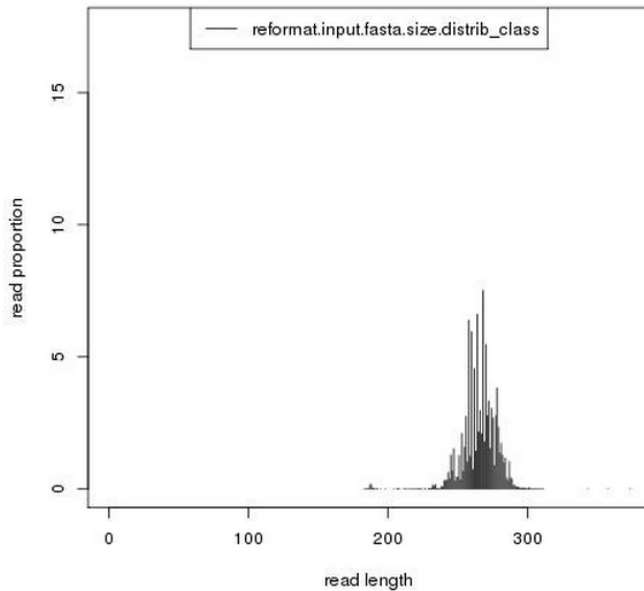


Résultats : le fichier HTML

• statistique de taille

Nombre= 37126 Somme= 9866886 Moyenne= 265.77 SD= 12.21 max= 373.00 min= 183.00 Mediane= 267.00

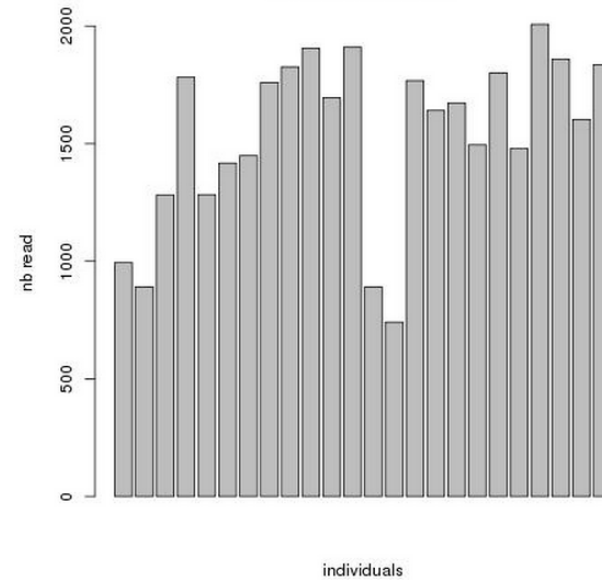
read length distribution of reformat.input.fasta



Statistiques sur le nombre de lectures démultiplexées par echantillon

Nombre= 24 Somme= 36986 Moyenne= 1541.08 SD= 352.67 max= 2007.00 min= 741.00 Mediane= 1641.00

nb readdistribution

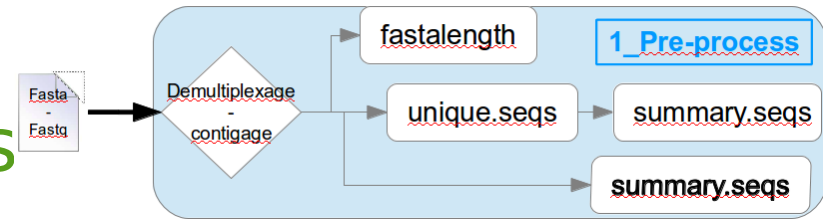


[count details](#)

nombre total de séquences uniques: 17577

Pipeline mothur

Etape 1 : MiSeq preprocess



Résultats : le fichier HTML

Le contigage

statistiques générales sur le contigage

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	248	248	0	3	1
2.5%-tile:	1	252	252	0	3	3810
25%-tile:	1	252	252	0	4	38091
Median:	1	252	252	0	4	76181
75%-tile:	1	253	253	0	5	114271
97.5%-tile:	1	253	253	6	6	148552
Maximum:	1	503	502	249	243	152360
Mean:	1	252.811	252.811	0.697867	4.44854	

Sélection des contigs

les contigs sont sélectionnés si leur taille est < 280 bases et s'il contiennent au maximum 0 bases ambiguës

Statistiques générales sur les contigs filtrés

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	250	250	0	3	1
2.5%-tile:	1	252	252	0	3	3227
25%-tile:	1	252	252	0	4	32265
Median:	1	252	252	0	4	64530
75%-tile:	1	253	253	0	5	96794
97.5%-tile:	1	253	253	0	6	125832
Maximum:	1	270	270	0	12	129058
Mean:	1	252.462	252.462	0	4.36663	

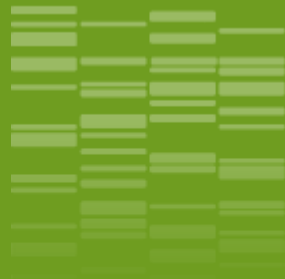
Les tailles des contigs avant filtre étaient de 248 à 503 pb et contenaient jusqu'à 249 bases ambiguës.

Après filtres, on conserve 129 058 contigs < 280 pb et sans aucune base ambiguë



TP : Preprocessing des données

Séquençage 454 vs MiSeq



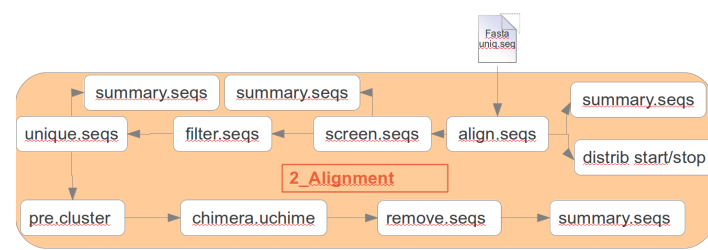
04

**Pipeline d'analyse mothur
Galaxy / Etape 2
Alignement des séquences**



Pipeline mothur

Etape 2 : alignement



Les séquences uniques précédentes vont être alignées sur un alignement multiple de référence : outil « **2_Alignment** ». Cette étape est commune aux deux types de séquençage.

Metagenomics Mothur 454 and MiSeq

- *454* 1 Pre process (e-learning available)
- *MiSeq* 1 Pre process
- ***2 Alignment** (e-learning available)
- *3 Sequences analysis (e-learning available)
- *454* 4 OTUs analysis (e-learning available)
- *MiSeq* 4 OTUs analysis (e-learning available)

***2 Alignment (version 1.0.0)**

Fasta file of unique sequences (barcode and primer removed) for all the sample. (fasta output from *454/MiSeq*_1_Pre_process):
13: uniques_contigs.fasta

Tabular file of unique sequences occurrences in each sample (« count table » output from *45/MiSeq*_1_Pre_process):
14: uniques_contigs.count_table

Sequence of forward primer for 16S region of interest:
NONE

Sequence of reverse primer for 16S region of interest:
NONE

Fasta file of 16S aligned (ARB format) reference sequences:
2: silva.bacteria.fasta

Taxonomies of your reference fasta file:
3: silva.bacteria.rdp.tax

Minimum number of aligned bases:
200

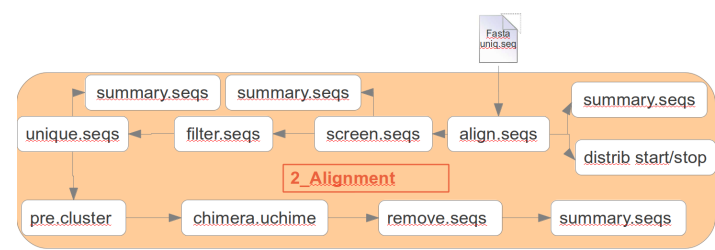
Execute

Fichiers de sortie du module 1

Fichiers de référence

Pipeline mothur

Etape 2 : alignement



L'outil « **2_Alignment** » va, une fois les lectures alignées, procéder à différentes étapes de filtre et de sélection

```
>HNHOSKD01A95T8
.....*.....CCCTGAAAGATCTTCACTGA
>HNHOSKD01B2K7O
.....
>HNHOSKD01A09WD
.....TGCGTAGATGGTAATAAAGT*CTGACATTGAGGCACGAAAGCGTGGGGAGCAAACA.....
>HNHOSKD01A0YKZ
.....CACGTAGGGCGGATATTTAAGTCAAGA*CTGACGCTGAGGTGCGAAAGCGTGGGGAGCAAACA.....
.....AGCGTAGACGGCATTGCAAGTCTGAAGTGA*CTGACGTTGAGGCTCGAAAGCGTGGGGAGCAAACA.....
```

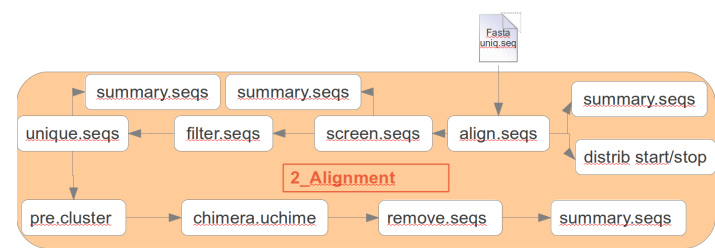
Étape 1) *screen.seqs* :
Sélection des lectures
dans une région d'intérêt

Étape 2) *filter.seqs* :
Sélection des positions d'alignement de façon
avoir tous les alignements de même taille

Etape 3) *unique.seqs* : Puisque les séquences ont été
trimmées aux extrémités, il est possible de réduire encore
le nombre de séquences uniques.

Pipeline mothur

Etape 2 : alignement



Après avoir sélectionné les alignements correspondants à notre région d'intérêt, « **2_Alignment** » va supprimer les lectures dues à des erreurs de séquençage selon 2 stratégies utilisées consécutivement

- *Pre.cluster* annote les séquences uniques dues à des erreurs de séquençage, et met à jour le fichier «count»

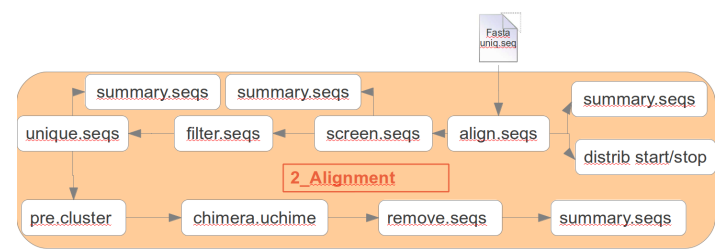
- *Chimera.uchime* recherche les séquences chimériques, *remove.seq*s les supprime



Résultats de « 2_Alignment »

Pipeline mothur

Etape 2 : alignement

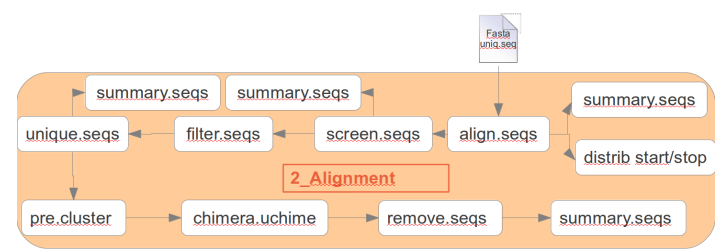


10: final.aln.count_table	👁️ ✎ ✕
9: final.aln.fasta	👁️ ✎ ✕
8: 2_align_bilan.html	👁️ ✎ ✕

- Fichier «count» :
le fichier est mis à jour avec les séquences alignées et filtrées
- Fichier fasta :
fichier fasta contenant l'ensemble des alignements des séquences uniques tous échantillons confondus
- Fichier HTML :
fichier html présentant les résultats du module

Pipeline mothur

Etape 2 : alignement



- Résultats : le fichier HTML

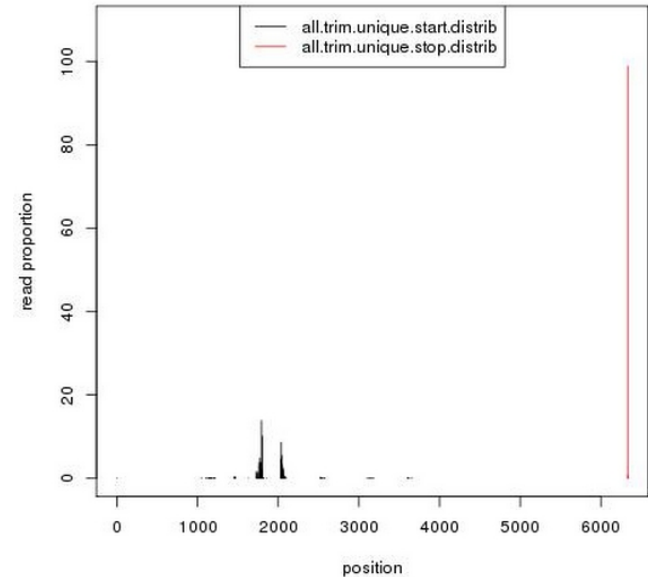
Les statistiques d'alignement dans le fichier HTML.

Avant screen.seqs

- Alignment statistics

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1044	6332	150	0	3	1
2.5%-tile:	1726	6333	210	0	4	925
25%-tile:	1773	6333	226	0	5	9247
Median:	1795	6333	234	0	5	18494
75%-tile:	2037	6333	241	0	5	27740
97.5%-tile:	2080	6333	254	0	6	36062
Maximum:	3660	6334	310	0	6	36986
Mean:	1881.84	6333	232.762	0	4.96201	

alignment position distribution

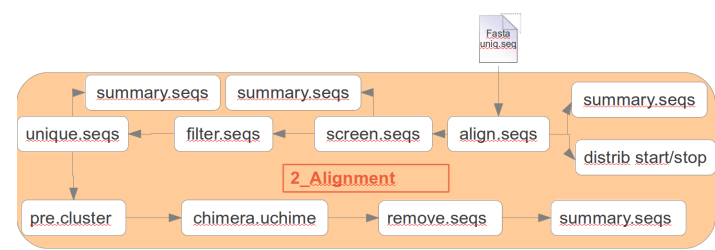


Comment analyser ces tableaux ?

- L'idée est de sélectionner la plus grande région couverte par le plus grand nombre de séquences
- Pour ce jeu de donnée les séquences se terminent assez précisément à la même position ~6 333, alors que la position start est plus variable entre 1044 et 3660. De même le nombre de bases alignées est assez variable.

Pipeline mothur

Etape 2 : alignement



• Résultats : le fichier HTML

Les statistiques d'alignement dans le fichier HTML.

Après *screen.seqs*

• Sélection des alignement

Les lectures dont les start est < à 2042 et le stop > à 6333 sont sélectionnées.

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1044	6333	210	0	3	1
2.5%-tile:	1726	6333	223	0	4	793
25%-tile:	1773	6333	229	0	5	7926
Median:	1793	6333	235	0	5	15851
75%-tile:	2032	6333	242	0	5	23776
97.5%-tile:	2042	6333	254	0	6	30909
Maximum:	2042	6334	310	0	6	31701
Mean:	1837.76	6333	235.895	0	5.00864	

Après *unique.seqs*, *filer.seqs*, *pre.cluster*, *chimera.uchime*, *remove.seqs*

alignements sélectionnés	Pre clustering	chimere	Alignement final	Nombre d'échantillons conservés	nombre moyen de lectures par échantillon (min - max)
12206 alignements uniques de 456 positions (31701 lectures)	4403	2416	5 387 alignements uniques (28 518 lectures)	24	14.5 (3.00 - 26.00)

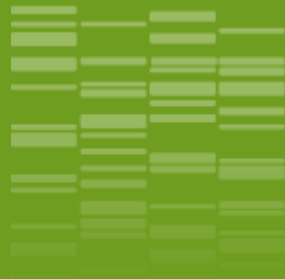
Comment analyser ces tableaux ?

- Screen.seq sélectionne sur une position start < à 85%-tile des séquences et une position stop > à 25%-tile ainsi que sur un nombre minimum de bases (paramètre fourni par l'utilisateur). Sur cet exemple le critère de nombre de bases alignées est celui qui a le plus d'effet.
- On vérifie également le nombre de positions des alignements après *filter.seq* (1^{ère} colonne du tableau bilan) et après *pre.cluster*, et *chimera.uchime* le nombre d'alignements/lectures conservées (4^e colonne).



TP : Alignement des données

Séquençage 454 vs MiSeq

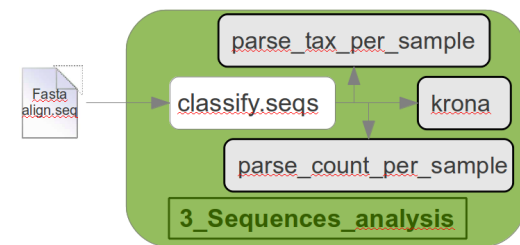


05

**Pipeline d'analyse mothur :
Galaxy / Etape 3 : classification
des séquences**

Pipeline mothur

Etape 3 : classification des séquences



Cette 3^e étape, via l'outil « 3_Sequences_analysis » va rechercher l'assignation taxonomique de chaque séquence unique alignée et calculer des tableaux de comptage pour chaque taxonomie dans chaque échantillon. Cette étape est commune aux deux types de séquençage.

Metagenomics Mothur 454 and MiSeq

- *454* 1 Pre process (e-learning available)
- *MiSeq* 1 Pre process
- *2 Alignment (e-learning available)
- *3 Sequences analysis (e-learning available)
- *454* 4 OTUs analysis (e-learning available)
- *MiSeq* 4 OTUs analysis (e-learning available)

***3 Sequences analysis (version 1.0.0)**

Fasta file of unique alignments for all the sample. (fasta output from 2_Alignment):
27: final.aln.fasta

Tabular file of unique sequences occurrences in each sample (« count table » output from 2_Alignment):
28: final.aln.count_table

Fasta file of 16S aligned (ARB format) reference sequences:
8: silva.bacteria.fasta

Taxonomies of your reference fasta file:
9: silva.bacteria.rdp.tax

Bootstrap threshold to keep or not taxonomy assignment at each taxon level:
50

Execute

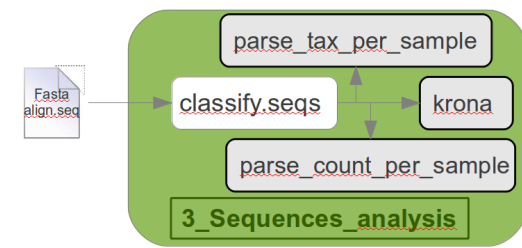
Fichiers de sortie du module 2



Résultats «3_Sequences_analysis»

Pipeline mothur

Etape 3 : classification des séquences



- Fichier de taxonomies :

résultats d'assignation taxonomique pour chaque séquence unique sous la forme :



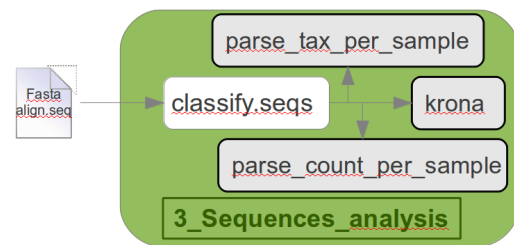
Identifiant d'une séquence unique	Taxonomie détaillée avec valeur de bootstrap
F11Fcsw_455	Bacteria(100);Firmicutes(74);Clostridia(73);Clostridiales(66);unclassified_Clostridiales(60);unclassified;unclassified;unclassified;

- Fichier HTML :

fichier html présentant les résultats du module

Pipeline mothur

Etape 3 : classification des séquences



Résultats : le fichier HTML

- Un tableau de comptage des lectures par échantillon et par taxonomie

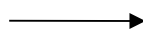
Taxlevel	rankID	taxon	daughterlevels	total	F11Fcsw	F12Fcsw	F12Fcsw	...
0	0	Root	1	28599	688	648	860	...
1	0.1	Bacteria	8	28599	688	648	860	...
2	0.1.1	Bacteroidetes	4	17639	305	133	310	...
3	0.1.1.1	Bacteroidia	1	17508	299	131	306	...
...

- Une archive contenant un tableau de comptage par échantillon :

[individuals_taxonomy.tgz](#)



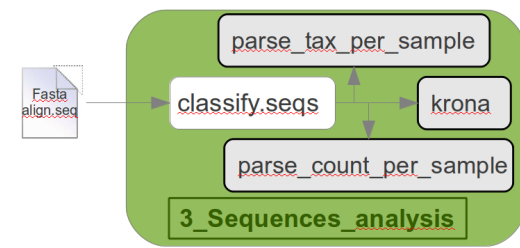
F11Fcsw.tax.summary



taxlevel	rankID	taxon	daughterlevels	total	uniqSeq	mean_bootstrap
0	0	Root	1	688		
1	0.1	Bacteria	8	688	226	100
2	0.1.5	Bacteroidetes	4	305	105	96.3619
3	0.1.5.2	Bacteroidia	1	299	103	93.6408
4	0.1.5.2.1	Bacteroidales	5	299	103	93.6408

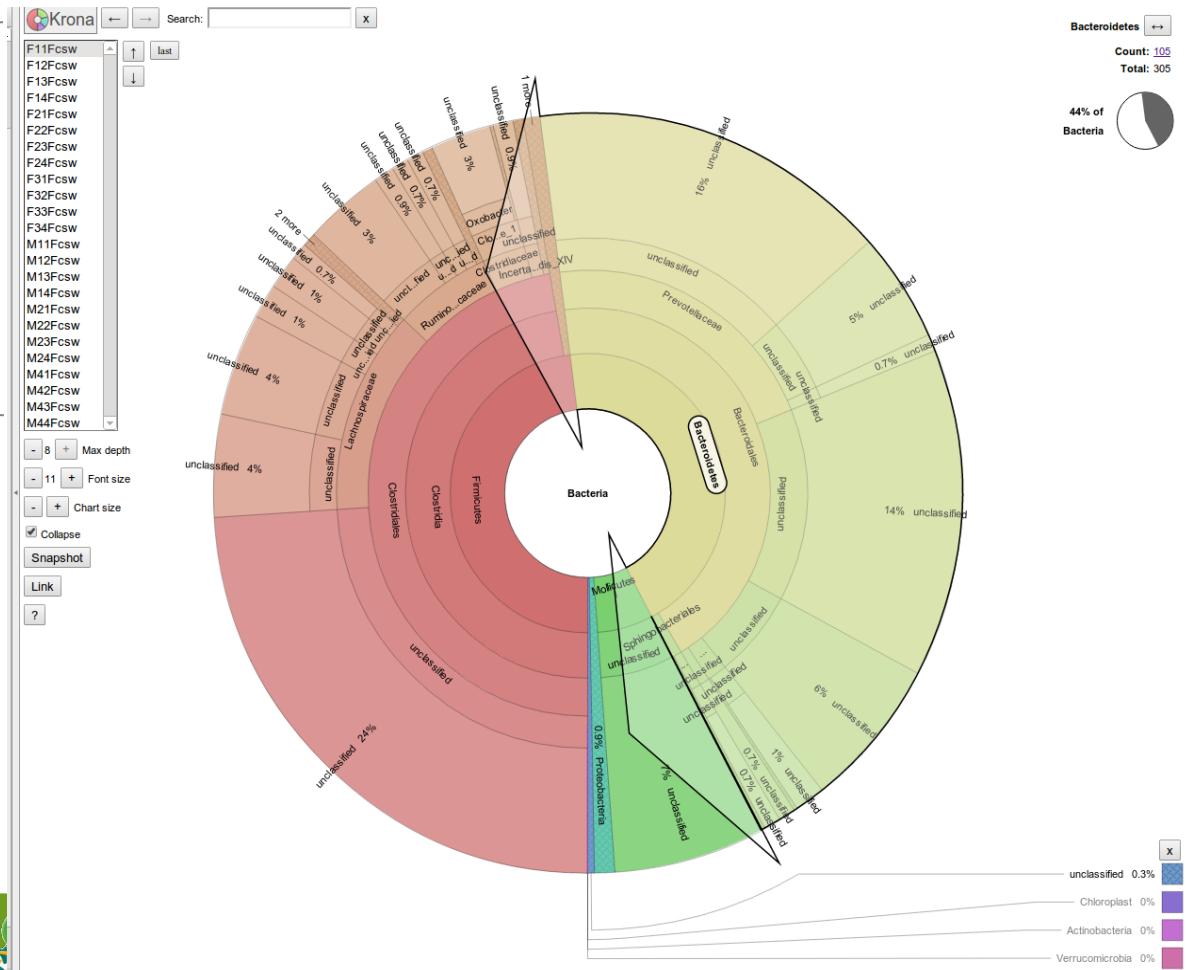
Pipeline mothur

Etape 3 : classification des séquences



- Des visualisateurs « Krona »

Liste des échantillons



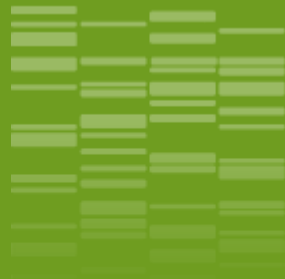
Proportion ou comptage des lectures composants la taxonomie sélectionnée





TP : classification des lectures

Séquençage 454 vs MiSeq



06

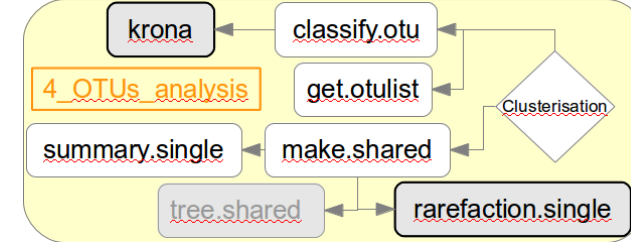
**Pipeline d'analyse mothur
Galaxy / Etape 4
Classification des OTUs**



Outil «4_OTUs_analysis»

Pipeline mothur

Etape 4 : classification des OTUs (454)



L'analyse des OTUs permet une classification taxonomique plus fine que lecture par lecture. Ce dernier module, « ***454*** **4_OTUs_analysis** », va permettre de construire ces « Operational Taxonomic Units »

Metagenomics Mothur 454 and MiSeq

- ***454*** 1 Pre process (e-learning available)
- ***MiSeq*** 1 Pre process
- ***2 Alignment** (e-learning available)
- ***3 Sequences analysis** (e-learning available)
- ***454*** 4 OTUs analysis (e-learning available)
- ***MiSeq*** 4 OTUs analysis (e-learning available)

***454* 4 OTUs analysis (version 1.0.0)**

Fasta file of unique alignments for all the sample. (fasta output from 2_Alignment):
27: final.aln.fasta

Tabular file of unique sequences occurrences in each sample (« count table » output from 2_Alignment):
28: final.aln.count_table

Taxonomies of your unique sequence aligned. (« taxonomy » output from 3_Sequences_analysis):
32: final.aln.taxonomy

Bootstrap threshold to keep or not taxonomy assignment at each taxon level:
51

Maximal distances authorized between sequences to construct an OTU (0.03 equal to 3%):
0.03

Execute

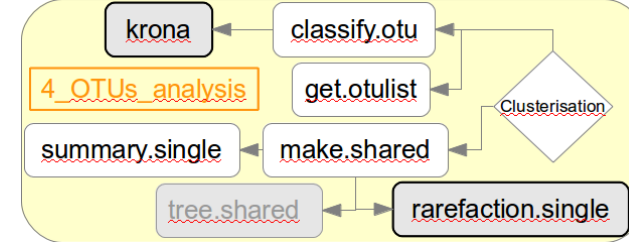
Fichiers de sortie du **module 2**

Fichier de sortie du **module 3**

Mothur clustering

Pipeline mothur

Etape 4 : classification des OTUs (MiSeq)



L'analyse des OTUs permet un classification taxonomique plus fine que lecture par lecture. Ce dernier module, « ***MiSeq*** **4_OTUs_analysis** », va permettre de construire ces Operational Taxonomic Units »

Metagenomics Mothur 454 and MiSeq

- ***454*** 1 Pre process (e-learning available)
- ***MiSeq*** 1 Pre process
- ***2*** Alignment (e-learning available)
- ***3*** Sequences analysis (e-learning available)
- ***454*** 4 OTUs analysis (e-learning available)
- ***MiSeq*** 4 OTUs analysis (e-learning available)

***MiSeq* 4 OTUs analysis (version 1.0.0)**

Fasta file of unique contigs for all the sample. (fasta output from 1_preprocess):
13: uniques_contigs.fasta ▾

Tabular file of unique sequences occurrences in each sample (« count table » output from 2_Alignment):
18: final.aln.count_table ▾

Taxonomies of your unique sequence aligned. (« taxonomy » output from 3_Sequences_analysis):
20: final.aln.taxonomy ▾

Bootstrap threshold to keep or not taxonomy assignment at each taxon level:
51

Maximal distances authorized between sequences to construct an OTU (0.03 equal to 3%):
0.03

Execute

Sortie du module 1

Sortie du module 2

Sortie du module 3

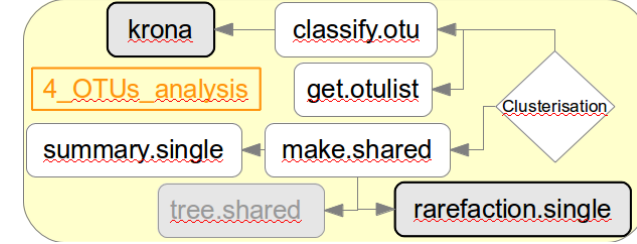
Swarm clustering



Résultats «4_OTUs_analysis»

Pipeline mothur

Etape 4 : classification des OTUs (454/MiSeq)



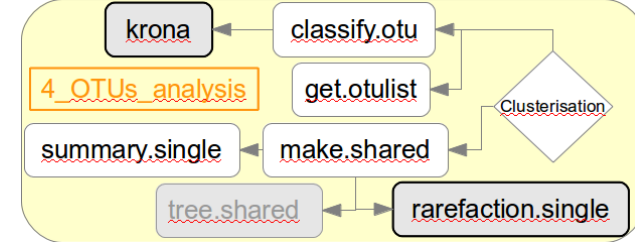
15:    [4_otus_analysis_Miseq_bilan.html](#)

- Fichier HTML

seul le fichier bilan des résultats est retourné ici.

Pipeline mothur

Etape 4 : classification des OTUs (454)



- Résultats : le fichier HTML
- Constitution des OTUs

distance moyenne intra OTU	nombre OTU
unique*	5387
0.01	4282
0.02	2755
0.03	1943
0.04	1460
0.05	1112
0.06	879
0.07	724

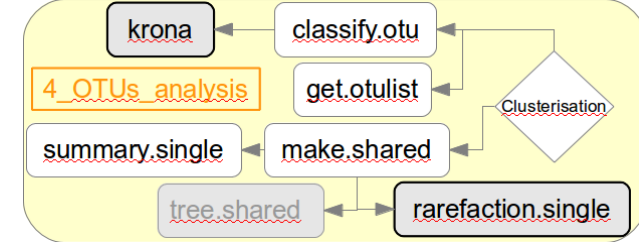
Plus la distance intra-groupe est grande, plus on agrège de séquences, plus le nombre d'OTUs est restreint.

Statistique du nombre de lectures composant les OTUs : distance de 3 %

Nombre= 1943 Somme= 5387 Moyenne= 2.77 SD= 9.88 max= 237.00 min= 1.00 Mediane= 1.00

Pipeline mothur

Etape 4 : classification des OTUs



- Résultats sur fichier HTML
- Diversité alpha : couverture des OTUs, estimateurs de richesse/diversité

label	Group	numOtus	Otu0001	Otu0002	Otu0003
0.03	F11Fcsw	1943	74	12	0
0.03	F12Fcsw	1943	19	14	0
0.03	F13Fcsw	1943	75	18	2
0.03	F14Fcsw	1943	56	24	22

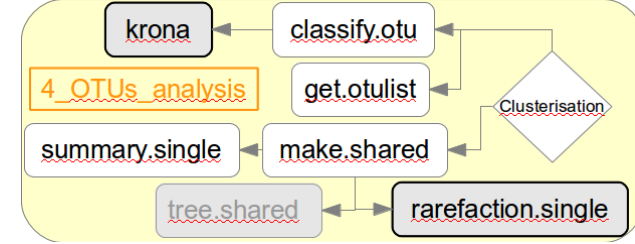
Répartition des OTUs selon la contrainte dans les différents échantillons

Tableaux d'estimateurs de richesse (chao/ace/jakknife) et de diversité (shannon/npshannon/simpson) et de leur intervalle de confiance pour chaque contrainte de distance dans chaque échantillon.

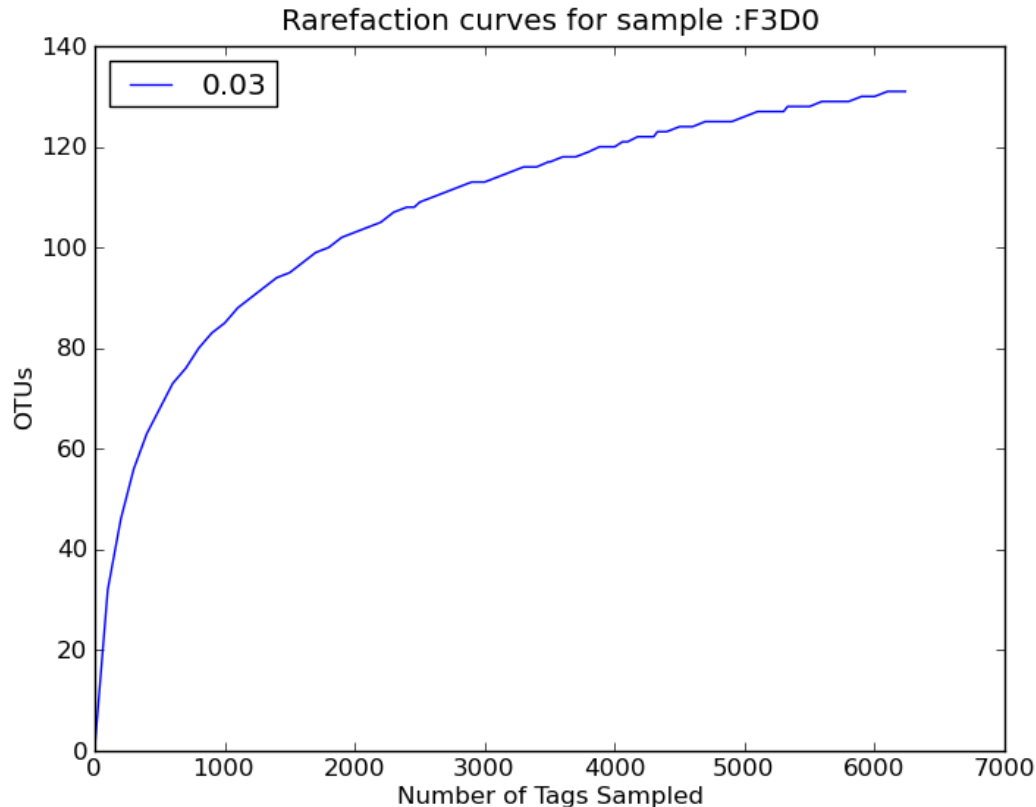
label	group	sobs	chao	chao_lci	chao_hci
0.03	F11Fcsw	159.000000	333.840000	258.221837	467.087681
0.03	F12Fcsw	152.000000	284.840000	225.534816	391.974296
0.03	F13Fcsw	155.000000	385.052632	281.633347	572.932674
0.03	F14Fcsw	166.000000	586.428571	391.849040	948.647486

Pipeline mothur

Etape 4 : classification des OTUs

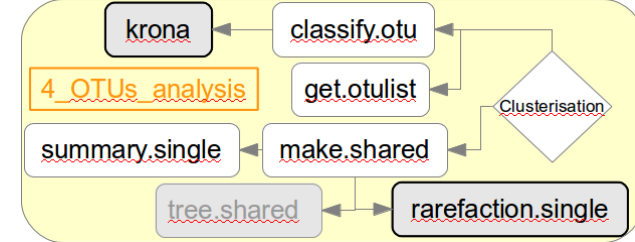


- Résultats sur fichier HTML
- Diversité alpha : courbe de raréfaction



Pipeline mothur

Etape 4 : classification des OTUs



- Résultats sur fichier HTML
- Classification taxonomique des OTUS
(un tableau par contrainte de distance)

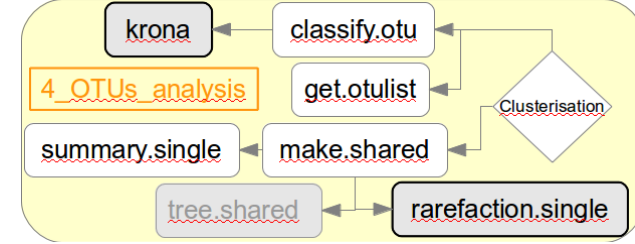
OTU	Size	Taxonomy
OTU0001	1466	Bacteria(100);Firmicutes(100);Clostridia(100);Clostridiales(100);Lachnospiraceae(100);Syntrophococcus(100);unclassified(100);unclassified(100);

- Tableaux de comptages en OTUs
(un tableau par contrainte de distance)

taxlevel	rankID	taxon	daughterlevels	total	F3D0	F3D1	F3D141	...
0	0	Root	1	269	131	121	119	...
1	0.1	Bacteria	10	269	131	121	119	...
2	0.1.1	Acidobacteria	1	1	1	1	1	...
...

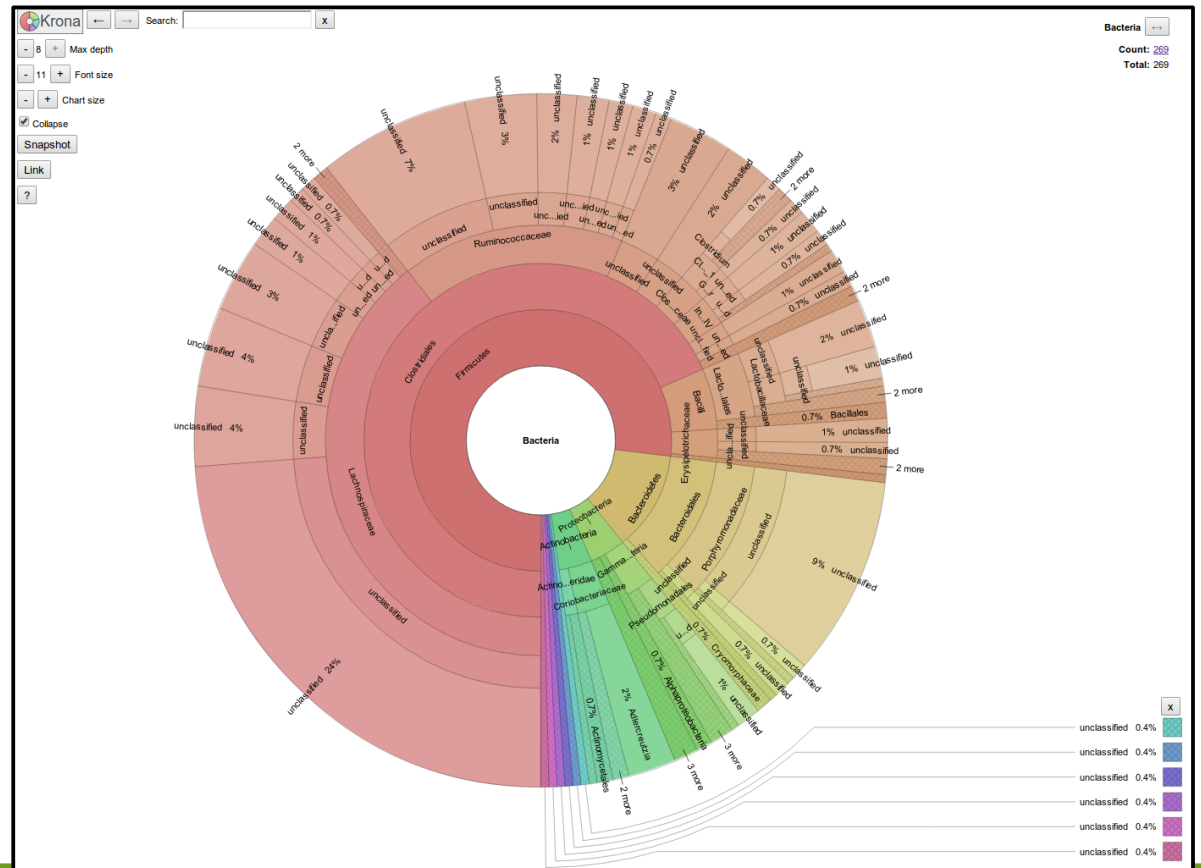
Pipeline mothur

Etape 4 : classification des OTUs



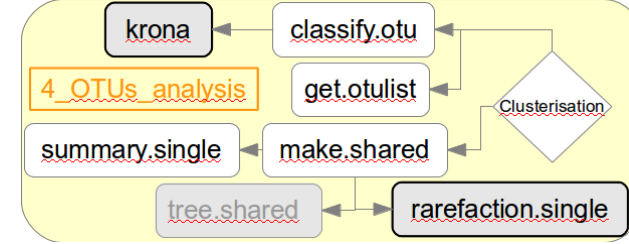
• Résultats sur fichier HTML

- Les visualisations « Krona »
- **Tous échantillons confondus**
- En fonction des échantillons, sur une contrainte donnée
- En fonction de la contrainte sur un échantillon donné



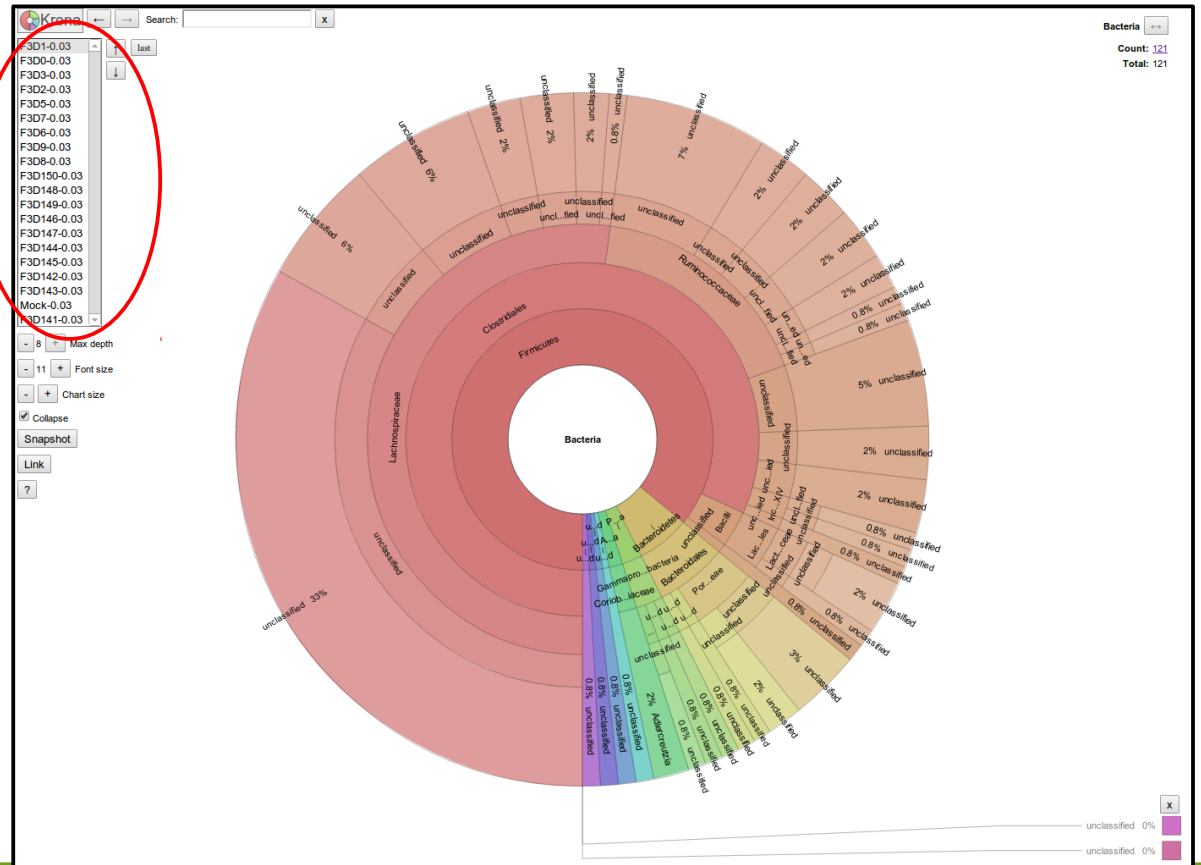
Pipeline mothur

Etape 4 : classification des OTUs



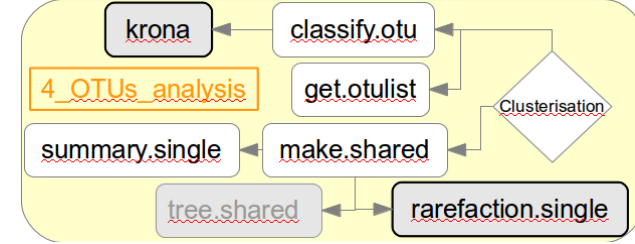
• Résultats sur fichier HTML

- Les visualisations « Krona »
- Tout échantillon confondu
- **En fonction des échantillons, sur une contrainte donnée**
- En fonction de la contrainte sur un échantillon donné



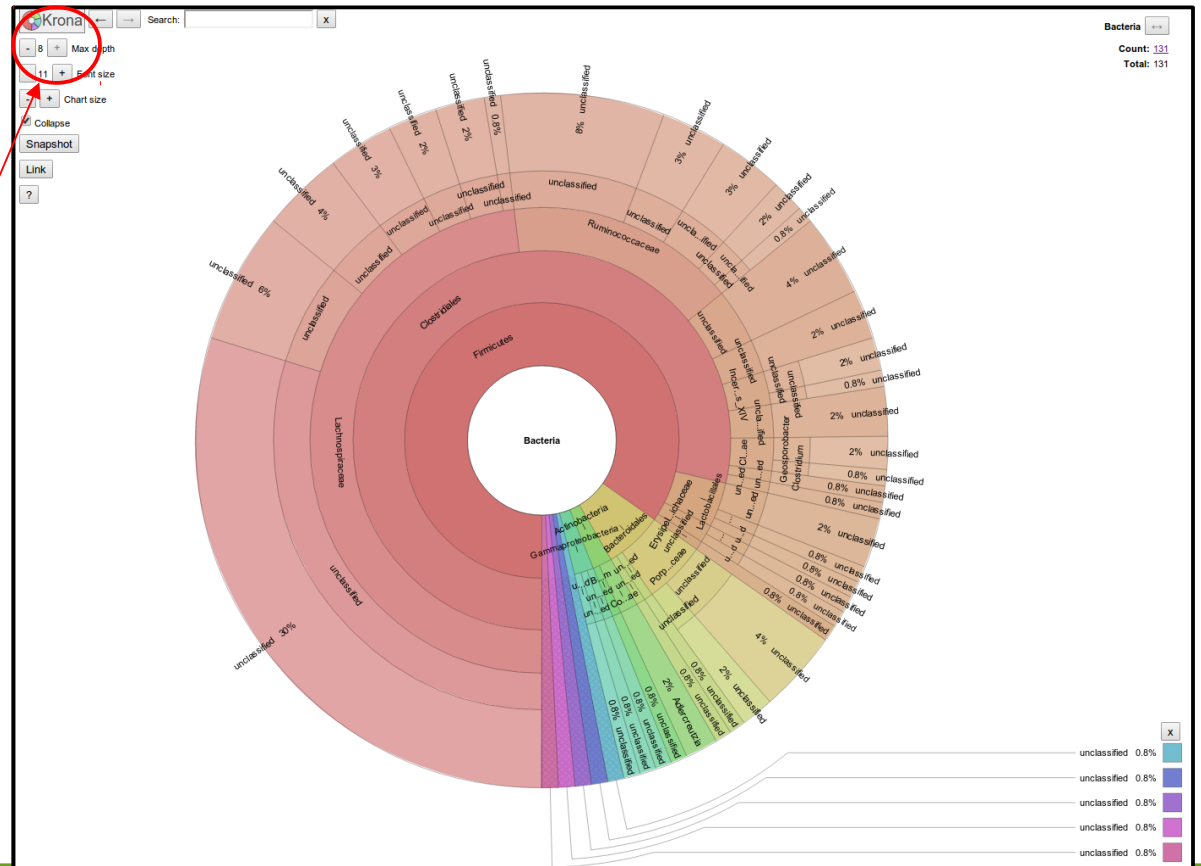
Pipeline mothur

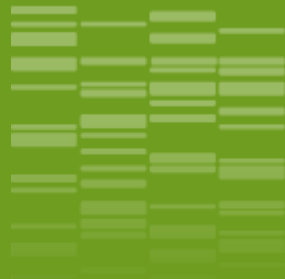
Etape 4 : classification des OTUs



• Résultats sur fichier HTML

- Les visualisations « Krona »
- Tout échantillon confondu
- En fonction des échantillons, sur une contrainte donnée
- **En fonction de la contrainte sur un échantillon donné** (en 454 plusieurs contraintes peuvent être listées)

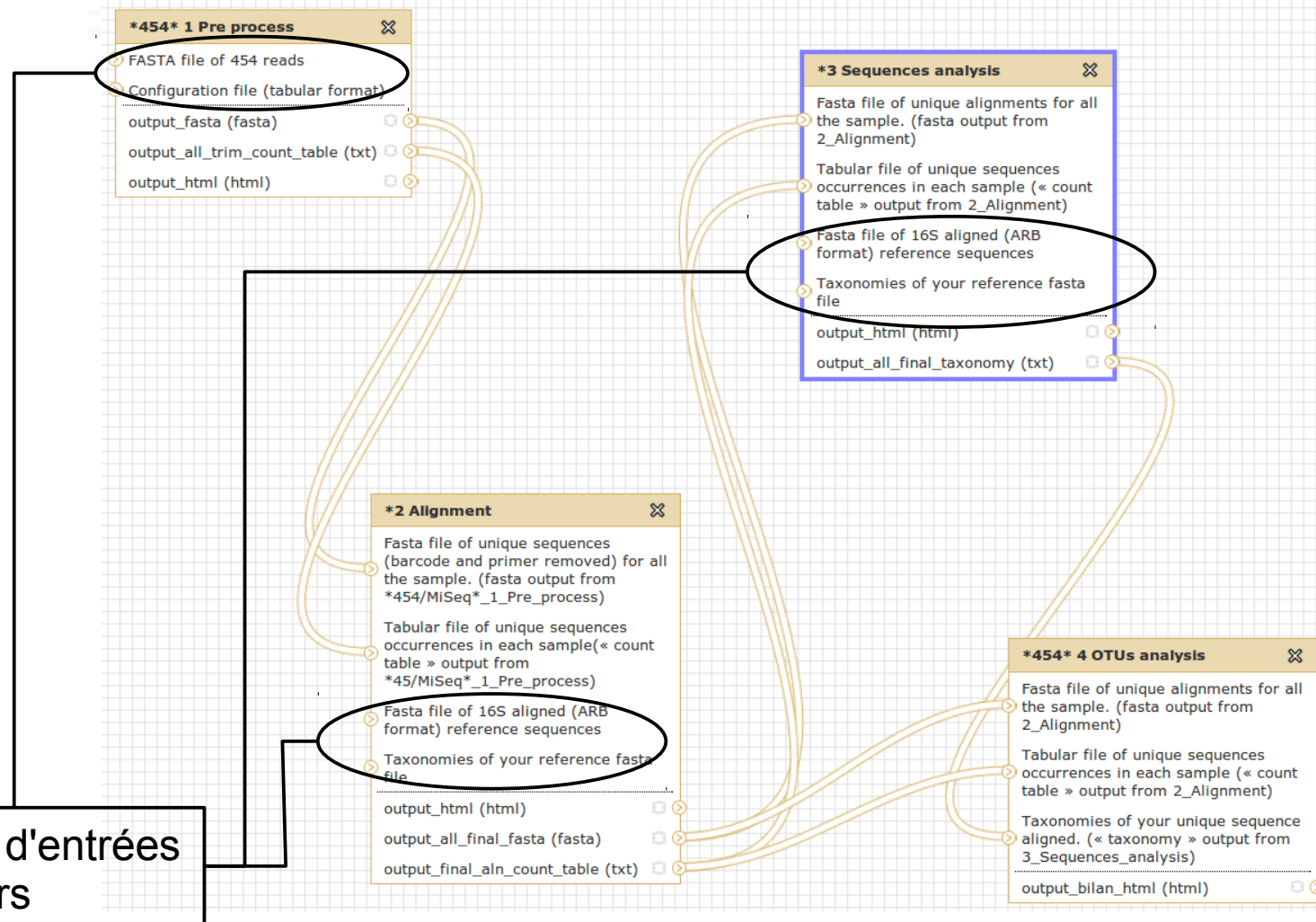




07

Pipeline d'analyse mothur : Galaxy / Utilisation du workflow

Pipeline mothur : 454 Utilisation du workflow



Pipeline mothur : MiSeq Utilisation du workflow

Sorties du module 1

***454* 1 Pre process**

- FASTA file of 454 reads
- Configuration file (tabular format)
- output_fasta (fasta)
- output_all_trim_count_table (txt)
- output_html (html)

Fasta uniq_seqs
Text count_table

Sorties du module 3

***3 Sequences analysis**

- Fasta file of unique alignments for all the sample. (fasta output from 2_Alignment)
- Tabular file of unique sequences occurrences in each sample (« count table » output from 2_Alignment)
- Fasta file of 16S aligned (ARB format) reference sequences
- Taxonomies of your reference fasta file
- output_html (html)
- output_all_final_taxonomy (txt)

Fasta uniq aln

Sorties du module 2

***2 Alignment**

- Fasta file of unique sequences (barcode and primer removed) for all the sample. (fasta output from *454/MiSeq*_1_Pre_process)
- Tabular file of unique sequences occurrences in each sample (« count table » output from *45/MiSeq*_1_Pre_process)
- Fasta file of 16S aligned (ARB format) reference sequences
- Taxonomies of your reference fasta file
- output_html (html)
- output_all_final_fasta (fasta)
- output_final_aln_count_table (txt)

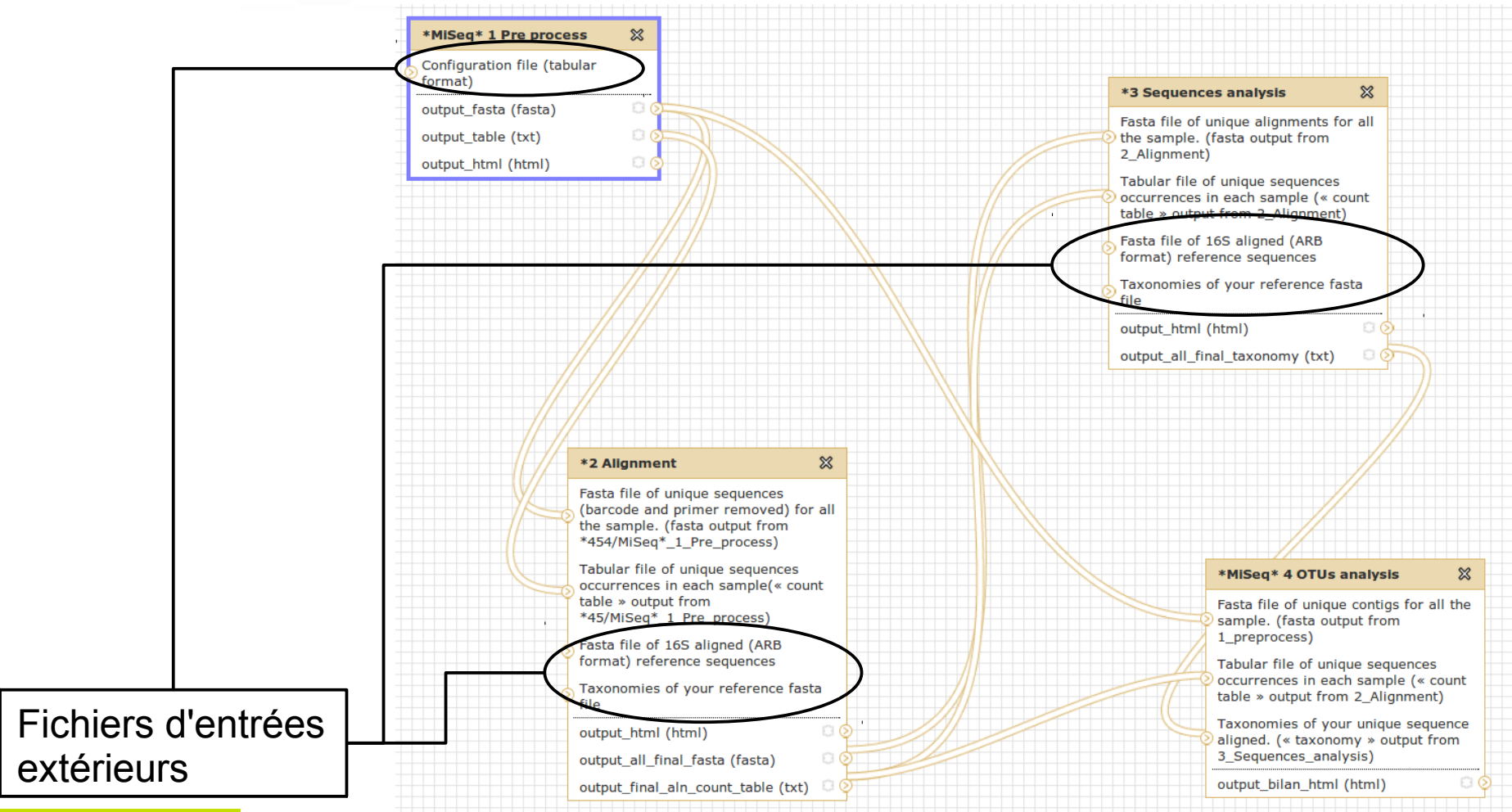
Fasta uniq aln
Text count_table

Sorties du module 4

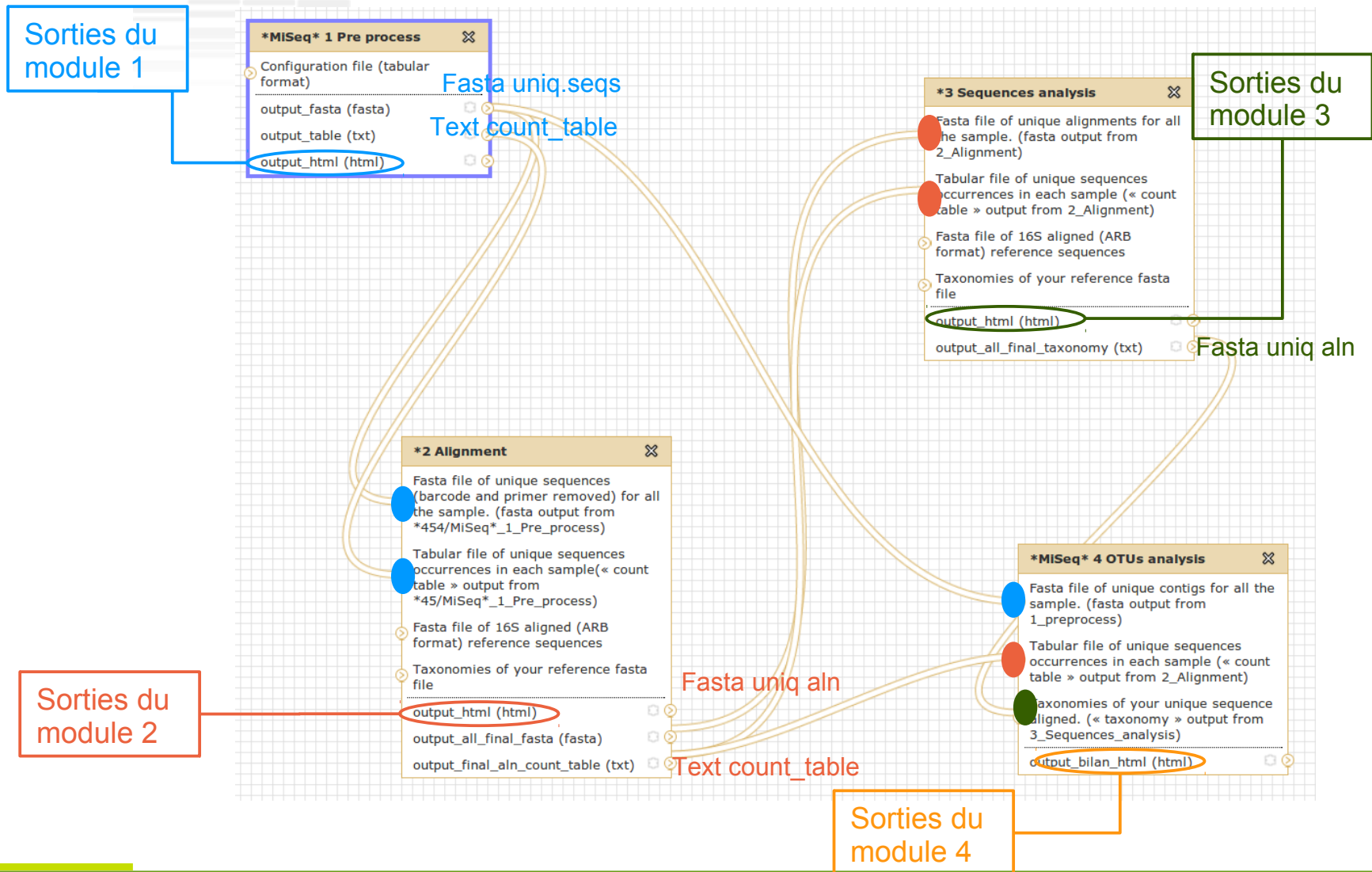
***454* 4 OTUs analysis**

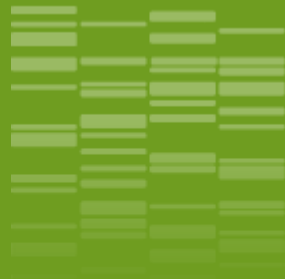
- Fasta file of unique alignments for all the sample. (fasta output from 2_Alignment)
- Tabular file of unique sequences occurrences in each sample (« count table » output from 2_Alignment)
- Taxonomies of your unique sequence aligned. (« taxonomy » output from 3_Sequences_analysis)
- output_bilan_html (html)

Pipeline mothur : MiSeq Utilisation du workflow



Pipeline mothur : MiSeq Utilisation du workflow





END

**Pipeline d'analyse mothur :
Galaxy / À vous de jouer !!**