



TANDEM



TP :
Pipeline Galaxy d'analyse de données
haut débit d'ADN 16S : séquençage
454 / MiSeq

Objectif :

Cette formation a pour but de vous apprendre à analyser des données de séquençage d'ADN 16S provenant des plateformes Roche 454 ou bien Illumina MiSeq. Il s'agit ici de manipuler les données brutes pour les rendre exploitables par les logiciels spécialisés, d'identifier les populations bactériennes via deux stratégies de classification (direct des lectures ou par la construction d'« Operational Taxonomic Unit », OTU), de quantifier ces populations et de calculer les différents indices de diversité alpha.

Indications Pratiques :

Les cours théoriques en ligne :

- Sigenae sig-learning : <http://sig-learning.toulouse.inra.fr/>
métagénomique : analyse d'ADN 16S : séquençage 454 :
métagénomique : analyse d'ADN 16S : séquençage MiSeq :

Pour vous connecter à Galaxy :

- Instance Sigenae de Toulouse : <http://sigenae-workbench.toulouse.inra.fr/>
Demande de compte : <http://bioinfo.genotoul.fr/index.php?id=81>

Sous l'instance Galaxy de Sigenae Toulouse, tous les outils concernant l'analyse des données 16S sont classés dans la section « Metagenomics Mothur 454 and MiSeq », dans le menu de gauche. D'autres outils généraux vous seront très souvent utiles pour manipuler vos fichiers, n'hésitez pas à naviguer dans ces catégories.

Les logiciels utilisés par le pipeline :

mothur : http://www.mothur.org/wiki/Main_Page

Schloss, P.D., et al. Appl Environ Microbiol, 2009. 75(23):7537-41

swarm : <https://github.com/torognes/swarm>

Krona : <http://sourceforge.net/projects/krona/>

Ondov BD, et al. BMC Bioinformatics. 2011 Sep 30; 12(1):385.

Les données

Le projet Patho-ID, étude des pathobiomes des rongeurs et des tiques, financé par le métaprogramme INRA-MEM, étudie les zoonoses portées par les rats et tiques de divers endroits dans le monde, les systèmes de co-infections et les interactions entre pathogènes.

Dans ce but ils ont extrait des centaines d'échantillons de rongeurs et de tiques et en ont extrait l'ADN 16S puis les ont séquencés sur plateforme Roche 454 dans un premier temps et en Illumina MiSeq ensuite.

Pour cette formation, ils nous ont autorisés à mettre à disposition un sous ensemble de ces données.

Etape 1 : chargement des données

Puisque nous avons 2 jeux de données, nous vous proposons de créer 2 historiques qui vont correspondre aux deux analyses.

Dans chaque historique vous aurez besoin d'une base de données de référence :

- un fichier fasta de séquences d'ADN 16S alignées (format ARB)
- un fichier tabulé des taxonomies correspondantes

Ici nous utiliserons la base de données Silva préformatée par Mothur : <http://www.mothur.org/w/images/9/98/Silva.bacteria.zip>

Vous aurez également besoin d'un fichier de configuration décrivant les différentes constructions séquencées.

Ainsi que le fichier fasta de lectures pour le séquençage 454 et des fichiers fastq pour le séquençage MiSeq.

Pour cette formation, nous avons constitué 2 jeux de données :

- PathoID 454 : http://genoweb.toulouse.inra.fr/~formation/5_ADN16S/Patho_ID_454/
- PathoID MiSeq : http://genoweb.toulouse.inra.fr/~formation/5_ADN16S/Patho_ID_MiSeq/

Historique : PathoID 454

→ Créez un nouvel historique et renommez le PathoID_454

→ Chargez les données correspondantes à ce jeu de données (cf TP initiation Galaxy)

Exercice :

Combien d'échantillons avons nous ?

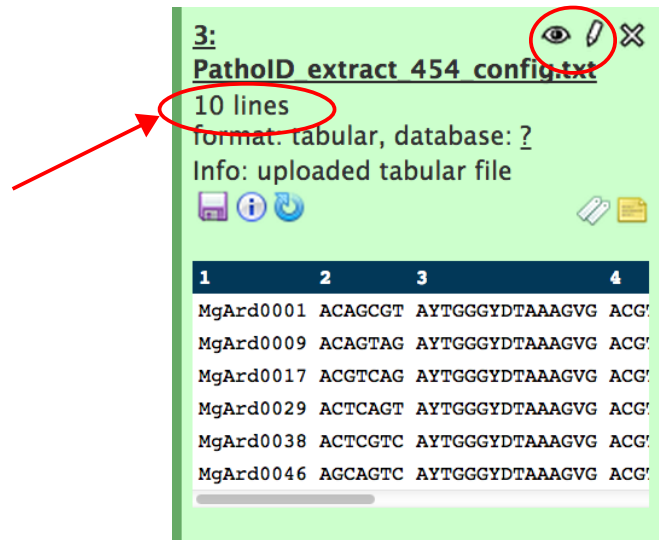
Combien de séquences ont été générées ?

En visualisant le fichier de configuration, quel type de construction d'amplicon a été utilisé ?

Réponses :

- Combien d'échantillons avons nous ?

Vous pouvez bien sûr visualiser le fichier de configuration en cliquant sur « l'oeil » du dataset correspondant, ou simplement cliquer sur le nom du dataset et visualiser les information relative à ce fichier.

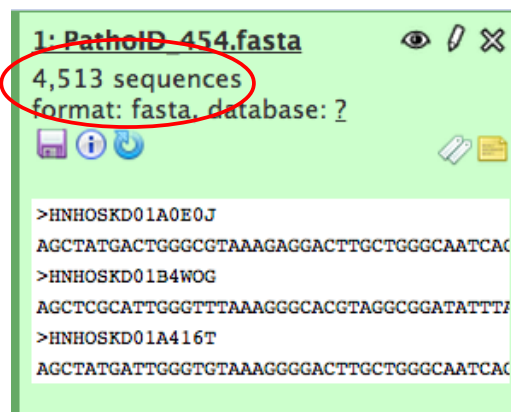


The screenshot shows a dataset viewer interface. At the top, the dataset name is '3: PathoID extract 454 config.txt'. Below the name, the text '10 lines' is circled in red, with a red arrow pointing to it from the left. To the right of the name, there are icons for 'eye', 'pencil', and 'X'. Below the line count, it says 'format: tabular, database: ?' and 'Info: uploaded tabular file'. There are also icons for a folder, information, and download. The main content area shows a table with 4 columns and several rows of data.

1	2	3	4
MgArd0001	ACAGCGT	AYTGGGYDTAAAGVG	ACG
MgArd0009	ACAGTAG	AYTGGGYDTAAAGVG	ACG
MgArd0017	ACGTCAG	AYTGGGYDTAAAGVG	ACG
MgArd0029	ACTCAGT	AYTGGGYDTAAAGVG	ACG
MgArd0038	ACTCGTC	AYTGGGYDTAAAGVG	ACG
MgArd0046	AGCAGTC	AYTGGGYDTAAAGVG	ACG

- Combien de séquences ont été générées ?

De la même façon que précédemment, le fait de cliquer sur le nom du fichier « fasta » vous indique, non pas le nombre de lignes mais le nombre de séquences. Galaxy est capable d'identifier le type de données d'un fichier en fonction de son extension. Attention donc à ce détail, en pensant à vérifier le format de fichier identifié par Galaxy.



The screenshot shows a dataset viewer interface. At the top, the dataset name is '1: PathoID_454.fasta'. Below the name, the text '4,513 sequences' is circled in red. To the right of the name, there are icons for 'eye', 'pencil', and 'X'. Below the sequence count, it says 'format: fasta, database: ?'. There are also icons for a folder, information, and download. The main content area shows a list of sequence identifiers and their corresponding DNA sequences.

```
>HNHOSKD01A0E0J
AGCTATGACTGGGCGTAAAGAGGACTTGCTGGGCAATCAC
>HNHOSKD01B4WOG
AGCTCGCATTTGGGTTTAAAGGGCACGTAGCGGATATTT
>HNHOSKD01A416T
AGCTATGATTGGGTGTAAAGGGGACTTGCTGGGCAATCAC
```

- En visualisant le fichier de configuration, quel type de construction d'amplicon a été utilisé ?

MgArd0001	ACAGCGT	AYTGGGYDTAAAGVG	ACGTACA	TACCVGGGTATCTAATCC	16S_V4
Myodes glareolus		0	1		
MgArd0009	ACAGTAG	AYTGGGYDTAAAGVG	ACGTACA	TACCVGGGTATCTAATCC	16S_V4
Myodes glareolus		0	1		
MgArd0017	ACGTCAG	AYTGGGYDTAAAGVG	ACGTACA	TACCVGGGTATCTAATCC	16S_V4
Myodes glareolus		0	1		
MgArd0029	ACTCAGT	AYTGGGYDTAAAGVG	ACGTACA	TACCVGGGTATCTAATCC	16S_V4
Myodes glareolus		0	1		

Un fichier de configuration de données 454 est un fichier tabulé, qui contient à minima 5 colonnes et qui indique ce que nous devrions trouver si nous regardions les extrémités des lectures de chaque échantillon:

- colonne 1 : sample name
- colonne 2 : séquence barcode forward
- colonne 3 : séquence primer forward
- colonne 4 : séquence barcode reverse
- colonne 5 : séquence primer reverse
- colonne 6 : nom de la région 16S cible

Comme les colonnes forward et reverse sont remplies, nous pouvons supposer que les lectures ont été séquencées soit en forward ET en reverse, soit uniquement à partir d'une extrémité, mais que le fragment a été séquencé entièrement (dû à sa taille). Dans tous les cas cela indique au programme de regarder les 2 extrémités des lectures.

Si vous regardez la colonne barcode reverse, vous remarquerez qu'elle est identique pour tous les échantillons. En fait dans ce projet, pour pouvoir identifier un échantillon il faut obligatoirement identifier le barcode en forward ET le barcode en reverse sur chaque lecture, par conséquent on recherchera également les 2 primers et les deux barcodes sur chaque lecture.

Historique : PathoID MiSeq

→ De même que précédemment, créez un nouvel historique et renommez le en PathoID_MiSeq.

→ Chargez les données correspondant à cette analyse (cf TP Initiation à Galaxy).

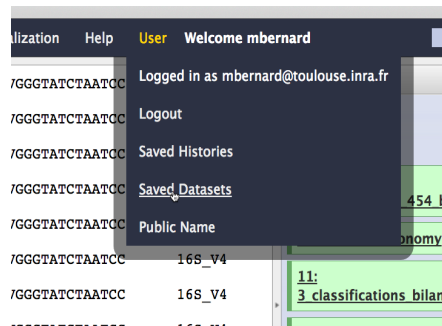
Pour ce jeu de données, vous ne récupérez qu'un fichier de config. En effet les données sont déjà démultiplexées et dans ce cas nous nous retrouvons avec un trop grand nombre de fichiers à charger dans Galaxy. Nous avons donc opté pour une autre solution qui consiste à charger les données sur le cluster

Genotoul, nous utiliserons ensuite le chemin du dossier pour charger les données dans Galaxy.

Nous n'avons pas remis les informations sur la base de données de référence car elles sont identiques à celles utilisées pour l'analyse en 454.

Ainsi au lieu de copier (et donc de prendre de l'espace sur votre quota) nous allons faire un lien vers les datasets de notre premier historique sur ce nouvel historique.

→ Allez dans « Saved DataSets » :



La liste de l'ensemble des datasets de tout vos historiques apparaît. Sélectionnez les 2 fichiers correspondants à la base de données Silva :

- silva.bacteria.fasta
- silva.bacteria.rdp.fasta

puis, tout en bas, cliquez sur « Copy to current history ».

Observations :

Les données étant déjà démultiplexées, le fichier de configuration ne contient cette fois-ci que la liste des noms des échantillons.

Etape 2 : Démultiplexage/contigage des lectures et recherche des séquences uniques.

Historique Pathoid_454

Revenez sur l'historique 454 via « User » → « Saved Histories ».

Exercice

Lancez le module 1 de démultiplexage des données 454 : *454*_1_Pre_process, et remplissez les champs comme demandé.

Pour la taille **minimum** des lectures après trimming, elle dépend de la région séquencée et des primers/barcode utilisés. Ici indiquez 200pb.

Rappelez vous bien le type de construction des amplicons et le type de multiplexage utilisés.

Réponses :

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

FASTA file of 454 reads:

Configuration file (tabular format):

Number of barcode per read mandatory to recognize a sample:

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):

Number of mismatch authorized in the barcode:

Number of mismatch authorized in the primer:

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

Trois nouveaux datasets apparaissent dans votre historique. En attendant la fin d'exécution du programme, lançons le contigage des données MiSeq.

Historique PatholD_MiSeq

Revenez sur l'historique PatholD_MiSeq.

Lancez le module 1 de démultiplexage des données MiSeq: *MiSeq*_1_Pre_process, et remplissez les champs comme demandé.

Rappelez vous que les données sont déjà démultiplées.

Pour la taille **maximum**, la région 16S V4 fait environ 250bp, mais cela peut être variable selon les espèces bactériennes. Le séquençage MiSeq produit tellement de données qu'en plus de réduire les chances d'erreur, nous sommes également intéressés par le fait de réduire le nombre de séquences. Indiquez

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

Configuration file (tabular format):

Extension of your input files:

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

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

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

Maximum number of ambiguous contigs bases:

Maximum contigs length (MANDATORY):

ici 280.

De nouveaux trois datasets apparaissent, revenons sur l'historique PathoID_454 pour regarder ces sorties.

Exercice :

Historique PathoID_454

Explorez les résultats et caractérisez les informations que vous donne chaque dataset.

Combien de séquences ont été identifiées comme appartenant à un échantillon et quelle taille font ces séquences?

Combien de séquences uniques récupérons nous ?

Historique PathoID_MiSeq

Et en ce qui concerne les résultats de contigage des données MiSeq :

Combien de contigs au total avons nous généré, et quelle est la taille du plus grand contig ?










En moyenne combien y a t il de contigs par échantillon ?

Combien de séquences uniques conservons nous après sélection des contigs ?

Réponses :

454

- Explorez les résultats et caractérisez les informations que vous donne chaque dataset.

7: 1 preprocess 454 bilan.html   
6: uniques.reads.count.table   
5: unique.reads.fasta   

- Les **résultats détaillés** dans une page HTML
- Un **tableau de comptage** du nombre d'occurrence des **séquences uniques** dans les échantillons
- Un fichier **fasta** des **séquences uniques** tout échantillon confondu

- Combien de séquences ont été identifiées comme appartenant à un échantillon et quelle taille font ces séquences?

Ces informations sont dans le fichier HTML. Cliquez sur l'oeil de ce dataset pour visualiser les statistiques d'exécution du module 1.

Sur les 4 513 lectures 454, 4 013 appartiennent à un de nos 10 échantillons et elles font en moyenne 208,68pb.

- Combien de séquences uniques récupérons nous ?

Cette information est également indiquée dans le fichier HTML, ou bien dans le descriptif du dataset « unique.reads.fasta » : 1,222 séquences.

MiSeq

- Combien de contigs au total avons nous généré, et quelle est la taille du plus grand contigs ?

statistiques générales sur le contigage

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	247	247	0	3	1
2.5%-tile:	1	251	251	0	4	5303
25%-tile:	1	251	251	0	4	53022
Median:	1	251	251	0	5	106044
75%-tile:	1	251	251	1	5	159065
97.5%-tile:	1	253	253	7	6	206784
Maximum:	1	501	501	60	241	212086
Mean:	1	252.619	252.619	0.751209	4.77519	

Taille maximum

Total

- En moyenne combien y a t il de contigs sélectionnés par échantillon ?

Nombre= 10 Somme= 153819 **Moyenne= 15381.90** SD= 6295.02 max= 27605.00 min= 6317.00
Mediane= 16159.00

- Combien de séquences uniques conservons nous après sélection des contigs ?

Comme précédemment l'information est indiquée dans le fichier HTML ou bien dans la description du dataset uniques_contigs.fasta

Etape 3 : Alignement des séquences uniques sur la base de données de référence et sélection des séquences représentant la région d'intérêt.

Que l'on ait des données MiSeq ou bien 454, en fin d'exécution du premier module on se retrouve avec :

- 1 fichier fasta de séquences uniques tout échantillon confondu
- 1 fichier tabulé représentant le nombre d'occurrences des séquences uniques dans chaque échantillon.

Nous allons donc passer à l'étape 3 qui consiste à aligner ces séquences sur une base de référence, de sélectionner les séquences qui correspondent effectivement à la région 16S d'intérêt, de filtrer les lectures qui seraient dues à des erreurs de séquençage ou bien à la formation de chimères pendant l'amplification PCR, et finalement sélectionner les alignements uniques.

Pour se faire nous allons utiliser l'outil « *2_Alignment », aussi bien sur les données 454 que MiSeq.

La seule différence (outre la méthode de séquençage) entre notre analyse 454 et notre analyse MiSeq sont les primers qui ont été utilisés pour amplifier la région 16S-V4.

- Séquençage 454 (les séquences des primers sont disponibles dans le fichier de configuration) :
 - primer forward : AYTGGGYDTAAAGVG
 - primer reverse : TACCVGGGTATCTAATCC

- Séquençage MiSeq (ici les séquences ne sont pas indiquées dans le fichier de configuration car nous n'avons pas le démultiplexage à faire)
 - primer forward : CAGCMGCCGCGGTAA
 - primer reverse : GGACTACHVGGGTWTCTAATCC

La région 16SV4 fait environ 250pb. Pour conserver un maximum d'informations sans perdre trop d'informations nous vous proposons d'indiquer un nombre minimum de bases alignées à 200. En fonction des résultats que vous observerez vous pouvez relancer le module en faisant varier cette taille. Par exemple en MiSeq, vu la quantité de données nous pouvons nous permettre d'être plus stringent pour réduire encore le nombre de séquences.

Exercices :

historique PathoID_454

L'exécution de ce module génère également 3 nouveaux datasets. Quel est la différence entre `uniques_reads.fasta` et `uniques_reads_aln.fasta` ?

Caractériser l'alignement des séquences sur la référence ? Vérifiez que les filtres de sélection et de tri ont eu un effet positif sans pour autant perdre trop d'alignements ou de positions.

historique PathoID_MiSeq

Les données MiSeq subissent le même genre de traitements que pour le 454.

Combien y a-t-il de séquences détectées comme provenant d'une erreur de séquençage ? Combien ont été détectées comme chimériques ?

Réponses :

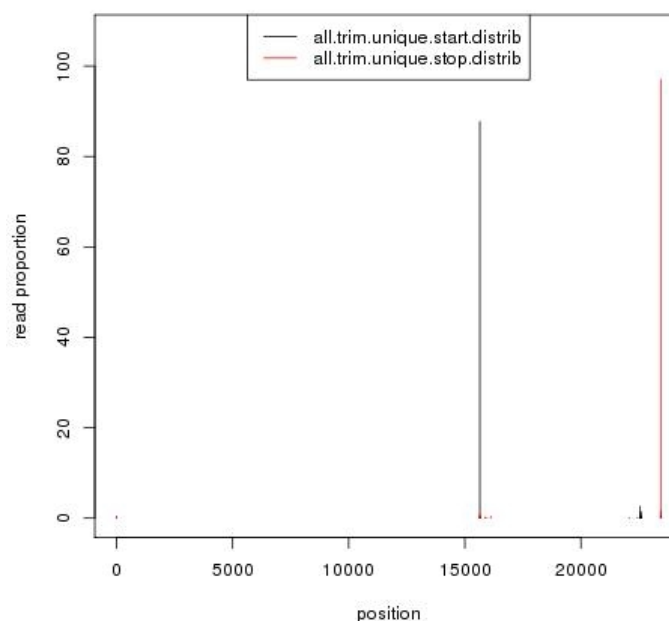
454 : Caractériser l'alignement des séquences sur la référence ?

Dans le fichier HTML vous disposez de 2 types de présentations sur la distribution des positions start et stop et d'alignement

• **Alignment statistics**

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	0	0	0	0	1	1
2.5%-tile:	15647	23440	5	0	2	101
25%-tile:	15647	23440	207	0	4	1004
Median:	15647	23440	207	0	4	2007
75%-tile:	15647	23440	207	0	4	3010
97.5%-tile:	22581	23440	207	0	5	3913
Maximum:	23440	23440	212	0	8	4013
Mean:	16301.2	23314.7	185.686	0	3.92225	

alignment position distribution



Ces deux résultats vous indiquent que les positions stop sont très restreintes sur une position de l'ADN 16S : 23 440 malgré quelques artéfacts à 0, alors que les positions start sont plus variables entre 0 et 23440. Dans le tableau vous avez également la description de la distribution du nombre de bases alignées. On observe que 2,5% des alignements contiennent entre 0 ou 5 bases.

Après sélection, vous retrouvez le même tableau que précédemment, sur les alignements sélectionnés sur des seuils de position et sur une taille minimum d'alignement. On voit que cette fois ci 100% des alignements commencent à la

position 15 647 de l'ADN 16S et se terminent à la position 23 440 avec au minimum 203pb alignées et jusqu'à 212 pb.

Nous avons 1 222 séquences uniques représentant 4 013 lectures, après sélection, nous avons sélectionné 1 040 alignements uniques de 287 positions (gaps compris), représentant 3 561 lectures, et après recherche des erreurs de séquençage on conserve 530 alignements représentant 3 451 lectures.

MiSeq :

Combien y a-t-il de séquences détectées comme provenant d'une erreur de séquençage ? Combien ont été détectées comme chimériques ?

L'alignement des séquences MiSeq est très « propre » comparé à celui des données 454, un peu moins de 200 séquences uniques ont été filtrées lors de l'étape de sélection. Sur les 36 327 séquences uniques 26 235 séquences ont été détectées comme provenant potentiellement d'une erreur de séquençage. Ces séquences sont caractérisées par une faible abondance et sont différentes d'une séquence fortement abondante d'au maximum 2 mismatches, elles représentent 70% de nos séquences. Ces séquences ne sont pas supprimées, elles fusionnent avec une séquence unique proche et la table de comptage est mise à jour.

Sur les 10 092 séquences uniques restantes, 1 865 séquences chimériques ont été détectées.

Au départ on avait donc 38 906 séquences uniques représentant 153 819 contigs, après sélection et filtre nous avons réduit de presque 80% le nombre de séquences uniques tout en conservant 97% des contigs.

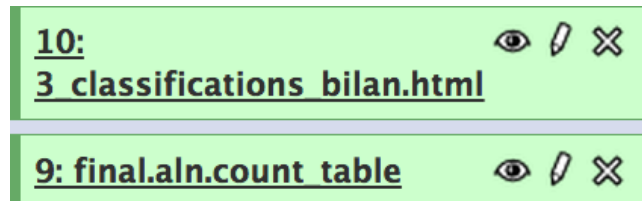
Etape 4 : Assignment taxonomique des séquences uniques alignées.

Cette étape compose la première stratégie de recherche des populations bactériennes composant nos échantillons. L'idée de ce module est de rapprocher chaque séquence à une séquence de référence et de lui conférer sa taxonomie. Pour estimer la confiance que l'on porte à cette taxonomie on assigne une valeur de bootstrap à chaque niveau taxonomique. L'idée est de savoir si sur 100 tentatives, nous retrouvons la même assignation taxonomique à chaque niveau taxonomique.

→ Appliquons l'outil *3_Sequence_Analysis sur nos données 454 et sur nos données MiSeq.

Le seul champs libre auquel vous avez accès est le seuil de bootstrap que vous autorisez pour conserver ou non une assignation taxonomique. Ce seuil est en fait utilisé lors d'une seconde exécution du module, vous aurez donc les résultats pour les taxonomies « brutes » et pour celles restreintes à ce seuil.

L'exécution de cet outil va générer 2 nouveaux datasets.



Exercices :

historique PathoID_454

Quels sont les 2 échantillons qui diffèrent le plus des autres ?

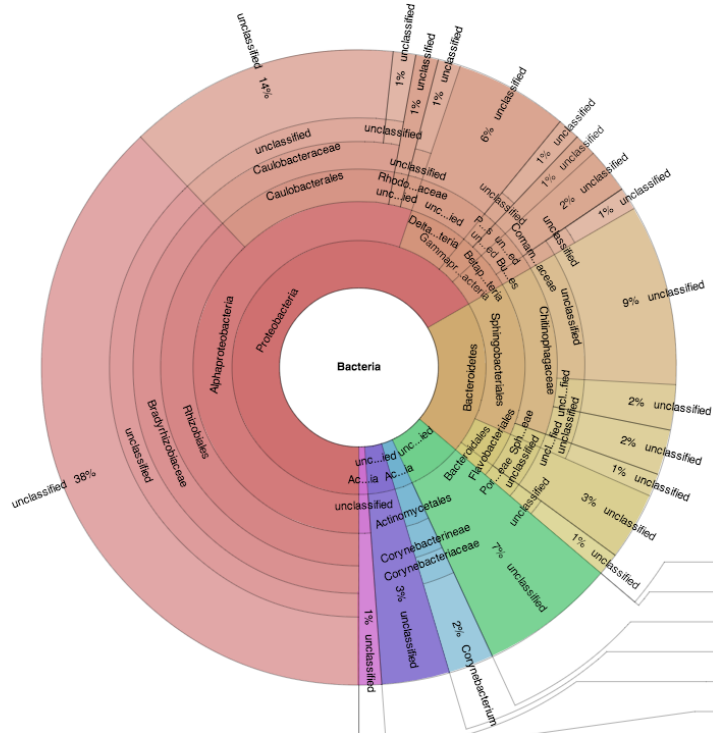
historique PathoID_MiSeq

Les données MiSeq subissent le même genre de traitements que pour le 454.

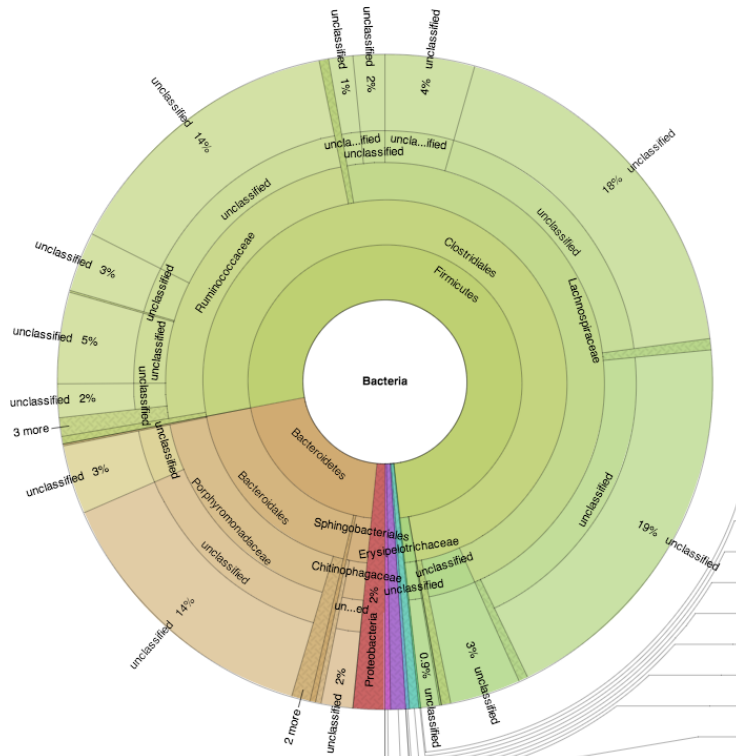
Dans l'échantillon MgArd0073, quelle est la proportion de Nitrobacter parmi les Bradyrhizobiaceae, si nous effectuons l'analyse taxonomique sans valeur seuil de bootstrap ou si au contraire nous regardons l'analyse avec seuil de bootstrap à 50 ?

Réponses :

454 : En explorant le fichier de comptage des taxonomies, on peut voir que l'échantillon MgArd0038 est celui qui a deux populations bactériennes qui lui sont relativement spécifiques : les Firmicutes et les Bacteroidetes. Cette observation se visualise très bien en regardant le Krona et en affichant successivement les échantillons.



Diversité « classique » des échantillons avec des proportions variables

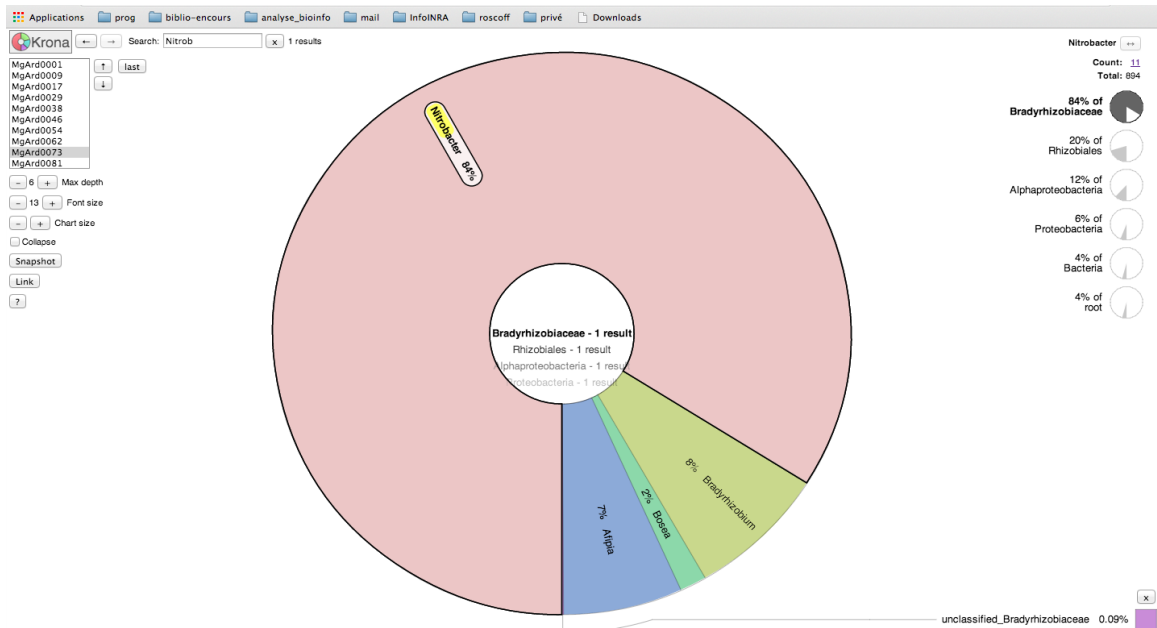


Diversité de l'échantillon MgArd0038

Si on continue l'exploration des échantillons, l'échantillon **MgArd0062** est caractéristique par la faible diversité de populations bactériennes. Il est à 99% composé de Bartonella.

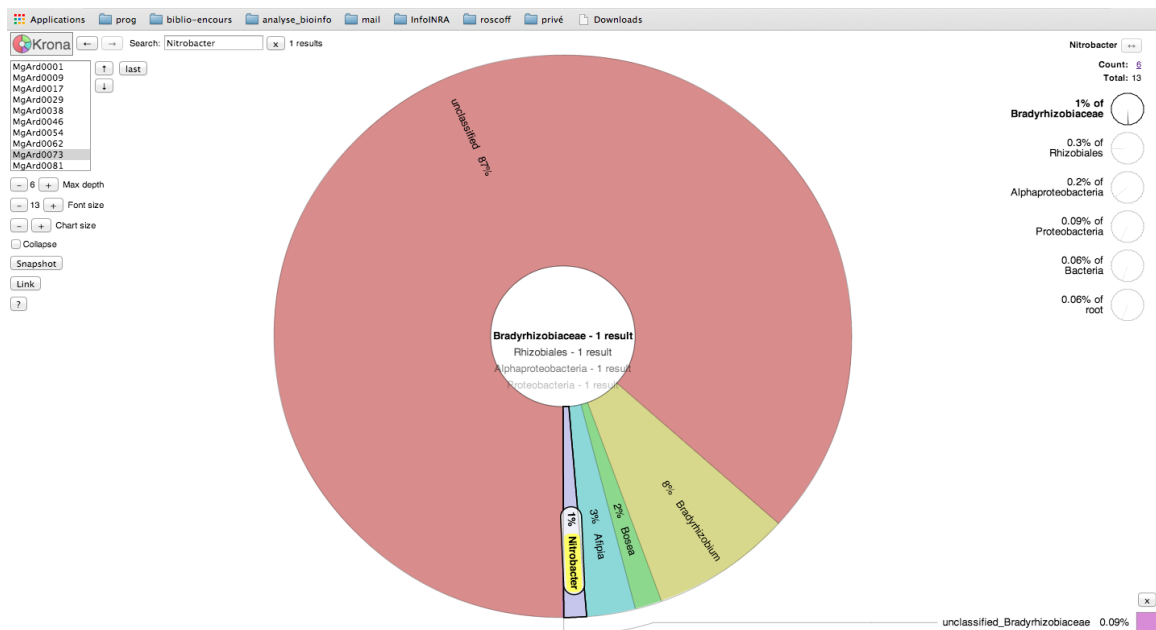
MiSeq :

Ouvrez le Krona fait sur l'analyse taxonomique sans seuil. Sélectionnez l'échantillon MgArd0073, décochez l'option « Collapse » , et tapez dans la section « search » « Nitrobacter ». Double cliquer sur l'étiquette Nitrobacter qui apparaît.



Nitrobacter est représenté par 11 séquences uniques présentes dans MgArd0073 ce qui représente 894 contigs. Cela représente 83% des Bradyrhizobiaceae.

Si on fait la même chose sur le Krona restreint aux taxonomies ayant un bootstrap > 50, on obtient :



Nitrobacter représente maintenant 1% des Bradyrhizobiaceae avec seulement 6 séquences uniques représentant 13 contigs.

Etape 5 : Construction d' « Operational Taxonomic Unit », calcul d'indice de diversité et assignation taxonomique.

Cette étape correspond à la seconde stratégie d'analyse de données 16S. Les séquences uniques sont clusterisées par similarité de séquences pour constituer des OTUs. Les échantillons sont ensuite caractérisés par leur composition en OTU. A partir de ces données de composition nous allons pouvoir caractériser les échantillons du point de vue de leur diversité, de leur couverture de séquençage....

Comme pour le premier outil, cette étape est spécifique du type de données de séquençage. Malgré tout l'interface de l'outil et les paramètres à indiquer sont quasiment identiques.

Lancez les outils `*454*_4_OTUs_analysis`, `*MiSeq*_4_OTUs_analysis`.

Un paramètre important de ces deux outils est la distance que nous autorisons entre séquences pour générer un OTU. Classiquement on trouve une distance autorisée de 3%, soit 0,03.

Question :

Quel est le paramètre qui change entre ces deux outils ?

Nous avons vu précédemment que l'échantillon MgArd0062 était essentiellement constitué de Bartonella. Dans l'analyse des données 454, regardez ce qui caractérise cet échantillon.

Est ce que vous obtenez les mêmes observations sur l'analyse MiSeq ?

Réponses :

454 4 OTUs analysis (version 1.0.0)

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

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

Taxonomies of your unique sequence aligned. (« taxonomy » output from 3_Sequences_analysis):
11: 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 2 : les reads alignés.

VS

MiSeq 4 OTUs analysis (version 1.0.0)

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

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

Taxonomies of your unique sequence aligned. (« taxonomy » output from 3_Sequences_analysis):
11: 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 : les contigs non alignés.

454 :

Bilan HTML du module 4: « tableau de comptage du nombre de reads composant chaque OTU (contrainte 0.03) dans chaque échantillon ».

→ En regardant la répartition des lectures de l'échantillon MgArd0062 dans chaque OTU, on s'aperçoit que la très grande majorité des lectures sont incluses dans l'OTU002 : 1097.

Bilan HTML du premier module 454 :

→ nous avons accès au nombre total de lectures de l'échantillon : nous obtenons 1109 lectures. Nous voyons donc qu'il y a bien une très grande majorité des lectures qui représente une seule espèce.

Bilan HTML du module 4 : statistique de diversité alpha

→ Notre hypothèse se valide encore en regardant les statistiques de diversité. Comparé aux autres échantillons, l'indicateur « sobs » est à 5 pour notre échantillon alors qu'il est bien plus élevé chez les autres échantillons. De même pour les autres indicateurs, l'échantillon MgArd0062 est systématiquement supérieur ou inférieur à ceux des autres échantillons.

Bilan HTML du module 4 : les courbes de raréfaction

→ Néanmoins sur ce graphique, nous voyons que le séquençage n'est clairement pas suffisant pour couvrir l'ensemble des espèces détectables dans cet échantillon. Si nous regardons les autres échantillons, nous remarquons également que le séquençage semble insuffisant.

Bilan HTML du module 4 : visualisation Krona de l'échantillon MgArd0062

→ Au vu des différentes taxonomies présentes dans l'échantillon, nous retrouvons bien notre souche Bartonella.

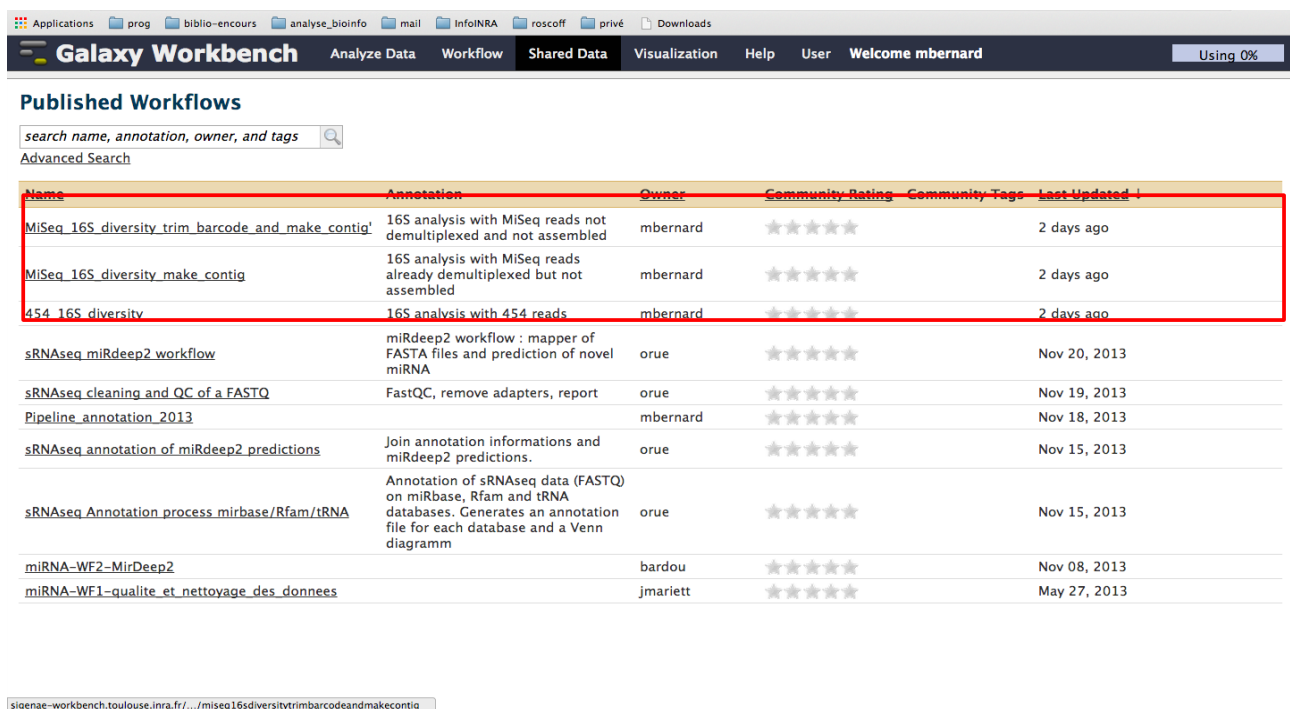
Attention cette version de Krona ne prend pas en compte l'abondance de chaque OTU dans l'échantillon visualisé.

MiSeq :

Si nous regardons l'analyse précédente, l'échantillon MgArd0062 est également essentiellement composé d'un seul OTU, OTU009 avec 27390 contigs sur les 27 605 contigs au départ. Les indicateurs de cet échantillon sont toujours extrêmes par rapport à ceux des autres échantillons, la courbe de raréfaction bien qu'améliorée par rapport à l'analyse 454 n'a toujours pas atteint un plateau. Dans le fichier de taxonomies des OTUs nous nous observons que l'OTU majoritaire est toujours Bartonella, par contre la visualisation Krona de l'échantillon MgArd0062 nous indique la présence d'autres phyla que nous n'avions pas détectés avant, tel que Acidobacteria ou Nitrospira.

Etape 6 (bonus) : Utilisation de workflow Galaxy.

Dans « Shared Data » → « Shared Workflow » vous trouverez les différents workflow public disponibles dont 3 workflows qui concernent l'analyse de données 16S :



Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
MiSeq 16S diversity trim barcode and make contig	16S analysis with MiSeq reads not demultiplexed and not assembled	mbernard	★★★★★		2 days ago
MiSeq 16S diversity make contig	16S analysis with MiSeq reads already demultiplexed but not assembled	mbernard	★★★★★		2 days ago
454 16S diversity	16S analysis with 454 reads	mbernard	★★★★★		2 days ago
sRNAseq miRdeep2 workflow	miRdeep2 workflow : mapper of FASTA files and prediction of novel miRNA	orue	★★★★★		Nov 20, 2013
sRNAseq cleaning and QC of a FASTQ	FastQC, remove adapters, report	orue	★★★★★		Nov 19, 2013
Pipeline_annotation_2013		mbernard	★★★★★		Nov 18, 2013
sRNAseq annotation of miRdeep2 predictions	Join annotation informations and miRdeep2 predictions.	orue	★★★★★		Nov 15, 2013
sRNAseq Annotation process mirbase/Rfam/tRNA	Annotation of sRNAseq data (FASTQ) on miRbase, Rfam and tRNA databases. Generates an annotation file for each database and a Venn diagram	orue	★★★★★		Nov 15, 2013
miRNA-WF2-MirDeep2		bardou	★★★★★		Nov 08, 2013
miRNA-WF1-qualite_et_nettoyage_des_donnees		jmariett	★★★★★		May 27, 2013

[signae-workbench.toulouse.inra.fr/.../miseq16sdiversitytrimbarcodeandmakecontig](https://www.genotoul.fr/signae-workbench.toulouse.inra.fr/.../miseq16sdiversitytrimbarcodeandmakecontig)

En sélectionnant sur le pipeline qui vous intéresse en fonction des données que vous avez, vous pouvez ensuite l'importer sur votre compte.

Analyse de données 454 :

Les données :

Formation Genotoul – Sigenae. Février 2014

Sarah Maman – Maria Bernard - Laurent Cauquil – Ibouniyamine Nabihoudine

Costello et al (<http://www.sciencemag.org/content/326/5960/1694>), ont publié en 2009, 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).

Téléchargez les données à partir des liens ci-dessous :

- Costello 454 :
http://genoweb.toulouse.inra.fr/~formation/5_ADN16S/Costello_extract_454/

Importez la base de référence Silva comme nous l'avons fait précédemment (« User » → « SaveDataSet » → Sélectionner les deux fichiers correspondants à cette référence, et cliquez sur « Copy to current history »)

N'oubliez pas de visualiser le fichier de configuration pour comprendre la structure des amplicons utilisés.

[Le workflow : 454 16S diversity](#)

Après avoir importé le workflow, allez dans le menu workflow et cliquez sur la flèche du workflow 454_16S_diversity puis sur « Edit ».

Vous arrivez sur une page qui présente le workflow et l'enchaînement des outils avec les liens vers les sorties d'un module vers les entrées du suivant.

Pour lancer le workflow, revenez sur l'onglet workflow et dans le menu du workflow d'intérêt cliquez sur « Run ». Vous arrivez sur la page de lancement avec 4 briques correspondant aux 4 outils que l'on a vu précédemment. Dans chaque brique vous avez les différentes options à remplir comme lorsqu'on lance les outils les uns après les autres.

Pour le jeu de données de Costello :

- brique 1 : en plus des fichiers d'entrée
 - indiquez la présence d'un seul barcode dans chaque lecture
 - indiquez la présence d'un seul primer dans chaque lecture
 - indiquez une taille minimum de 150pb
- brique 2 : en plus des fichiers de référence
 - indiquez une taille minimum d'alignement de : 150
 - indiquez la séquence de référence forward :
AGAGTTTGATCCTGGCTCAG
 - indiquez la séquence de référence reverse :
CATGCTGCCTCCCGTAGGAGT
- brique 3 : en plus des fichiers de référence
 - modifiez si vous le souhaitez le seuil de bootstrap

- brique 4 :
 - modifiez si vous le souhaitez le seuil de bootstrap
 - modifiez si vous le souhaitez la distance intra-cluster autorisée

Enfin cliquez sur run.

Toutes les sorties apparaissent en une seule fois dans votre historique. Elles vont devenir accessibles au fur et à mesure de l'exécution de chacun des modules automatiquement.

Analyse de données MiSeq:

Les données :

Schloss et al, développeur de mothur, ont publié un protocole d'analyse d'amplicon sur plateforme de séquençage Illumina MiSeq, <http://aem.asm.org/content/79/17/5112.long>. 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.

Téléchargez les données à partir des liens ci-dessous :

- Schloss MiSeq : http://genoweb.toulouse.inra.fr/~formation/5_ADN16S/Schloss_extract_MiSeq/

Le workflow : MiSeq 16S diversity_make_contig

Les données sont en effet déjà démultiplexées. Avant de pouvoir commencer vous devez charger les fichiers fastq dans un dossier sous genotoul comme vous l'avez vu lors de la formation initiation à Galaxy.

Après avoir importé le workflow, vous pouvez comme précédemment visualiser l'enchaînement des outils en cliquant sur « edit » du workflow ou le lancer en cliquant sur « run ».

Pour le jeu de données de Schloss :

- brique 1 : en plus des fichiers d'entrée
 - indiquez la présence d'un seul barcode dans chaque lecture
 - indiquez la présence d'un seul primer dans chaque lecture
 - indiquez une taille maximum de 275 pb
- brique 2 : en plus des fichiers de référence
 - indiquez une taille minimum d'alignement de 200
 - laissez la séquence de référence forward à NONE
 - laissez la séquence de référence reverse à NONE
- brique 3 : en plus des fichiers de référence

- modifiez si vous le souhaitez le seuil de bootstrap
- brique 4 :
 - modifiez si vous le souhaitez le seuil de bootstrap
 - modifiez si vous le souhaitez la distance intra-cluster autorisée

De même que précédemment les datasets de l'ensemble de l'analyse apparaissent dans votre historique. Ils seront visualisables lorsque l'exécution du module dont ils dépendent ce sera correctement déroulé.

Avec toutes les indications que nous avons vu précédemment, nous vous laissons explorer les résultats et analyser les populations bactériennes présentes dans ces échantillons.