Formation "Cycle NGS": Module 5 Métagénomique

Les différentes parties du TP:

- Contexte de l'étude
- Analyse primaire avec les OTU
- Analyse primaire avec les ASV
- Analyse secondaire
- Liens utiles

Contexte de l'étude

Dans ce TP, vous allez apprendre à effectuer une analyse microbienne 16S complète à l'aide de la suite d'outils QIIME2 dans Galaxy. Nous utiliserons l'instance Galaxy France usegalaxy.fr. Pour le TP nous utiliserons ce lien, ressources propores à la formation.

Ce TP se base sur le tutoriel de QIIME2 Moving Pictures.

Les jeux de données utilisés dans ce TP proviennent d'une étude sur le microbiome humain [Caporaso et al. (2011)]. Les données ont été générées sur un Illumina HiSeq sur la région 4 (V4) de l'ADNr 16S. Cette étude représente l'une des premières études à grande échelle du microbiome humain. C'est une série temporelle (de 15 mois et de 6 mois) qui comprend 2 sujets sains, 4 sites de prélèvement différents (intestin, langue, paume gauche et paume droite) représentant au total 396 points. Le premier point (T=0) correspond au point directement prélevé après un traitement antibiotique.

Dans ce TP, pour réduire le temps d'exécution de certains programmes, nous allons travailler uniquement sur 34 échantillons provenant de cette étude, mais couvrant les différentes variables.

Analyse primaire avec les OTU

L'analyse primaire a pour but de générer la table d'OTU à partir des données de séquençage haut-débit. Elle nécessite une étape de pre-processing des reads, une étape de clustering et une étape d'assignation taxonomique.

Import des données

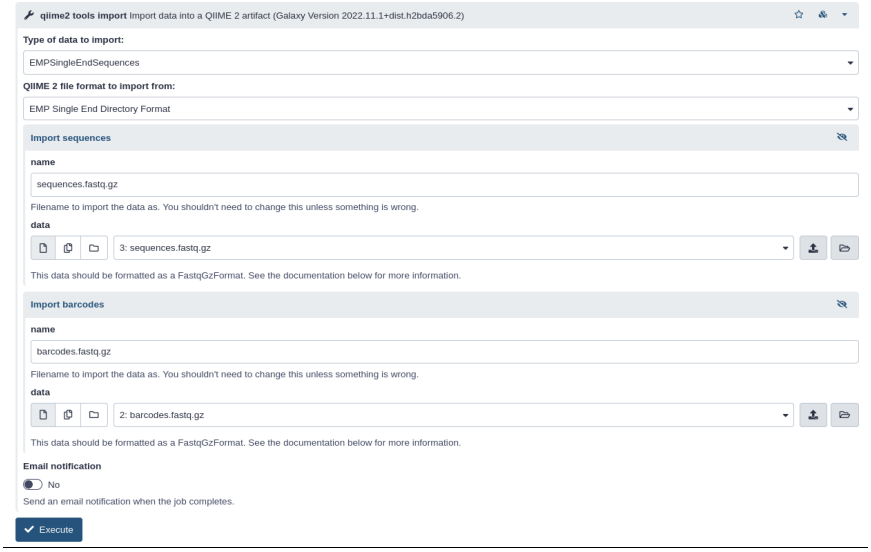
Téléchargez les fichiers:

- sequences.fastq.gz: ce fichier contient les reads
- barcodes.fastq.gz: ce fichier contient les code-barres associés à chaque read du fichier *sequences.fastq.gz*
- sample-metadata.tsv: ce fichier contient les métadonnées associées à l'étude

Dans le menu de gauche de Galaxy, tout en haut, cliquez sur le bouton *Upload Data*. Chargez ensuite les 3 fichiers précédents et cliquez sur *start*.

Une fois les fichiers chargés dans votre historique, en cliquant sur l'icône "crayon" pour éditer les attributs, renommez le fichier de métadonnées en *sample-metadata.tsv*.

Nous allons maintenant créer le QZA, unique format de fichier pris en charge par QIIME2. Pour ce faire, utilisez l'outil *qiime2 tools import* avec ces paramètres:



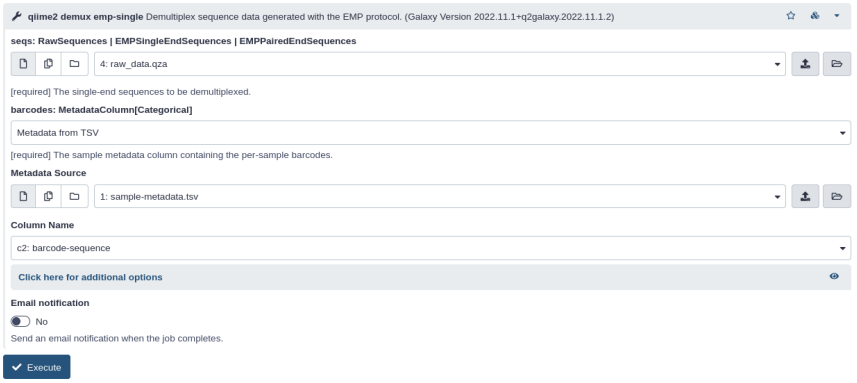
Pour finir renommez ce fichier *raw_data.qza*.

> Visualisez le **fichier contenant les métadonnées**.
> Quelles sont les données contenues dans ce fichier?
> Quel est le type de données pour l'année? Quelle est le type de données pour *bodySite*?
> Combien y a-t-il de sites différents dans cette étude?
> De quel site provient l'échantillon L5S174?
> De quel sujet provient l'échantillon L3S378?
> Quel jour de l'expérience a été prélevé l'échantillon L1S105?

Démultiplexage des données

Nous allons démultiplexer les données, c'est-à-dire répartir les lectures par échantillons à l'aide des code-barres. Pour ce faire nous allons utiliser la fonction **demux** de QIIME2, qui nécessite en entrée les lectures et les code-barres ainsi que la séquence des code-barres pour chaque échantillon, cette information se trouvant dans le fichier de métadonnées.

Utilisez l'outil *qiime2 demux emp-single* avec ces paramètres:



Deux fichiers sont produits: le fichier *demux* qui contient les données démultipléxées et un fichier contenant les détails de la correction d'erreurs.
Renommez le fichier demux en *demux.qza*.

Afin de visualiser le fichier *demux* produit à cette étape, nous allons générer le fichier QZV associé.
Utilisez l'outil *qiime2 demux summarize* avec ces paramètres:

```
data: demux.qza
```

En cliquant sur la visualisation, le fichier QZV se télécharge sur votre machine. Vous pouvez maintenant le charger dans l'interface de visualisation QIIME2 view pour l'explorer.

> Combien a-t-on de reads au total dans les 34 échantillons?
> Quel est le nombre de reads minimal? Pour quel échantillon?
> Quel est le nombre de reads maximal? Pour quel échantillon?
> Combien y-a-t-il de reads pour l'échantillon L1S76?
> Que pensez-vous de la qualité des reads?

Pre-processing

Nous allons commencer par nettoyer les reads en fonction de la qualité. Pour cela, nous allons utiliser l'outil *qiime2 quality-filter q-score* avec ces paramètres:

```
demux: demux.qza
min-quality: 20
```

> Que signifie le paramètre *min-length-fraction* dont la valeur par défaut est de 0.75?

Deux fichiers sont produits:

- filtered-sequences.qza : qui contient les séquences filtrées
- filter-stats.qza: qui contient les statistiques du filtrage

Pour visualiser le fichier *filter-stats.qza* nous allons utiliser l'outil *qiime2 metadata tabulate* avec ces paramètres:

```
input: Metadata : Metadata from Articfact
Metadata Source: filter-stats.qza
```

> Visualisez le fichier QZV sur QIIME2 view pour répondre aux questions sur l'échantillon L1S76:
> Combien de reads restent-ils à la fin de l'étape de filtrage?
> Combien des reads ont été tronqués?
> A quoi correspond la colonne *Reads-to-short-after-truncation*?
> Combien de reads ont été écartés car ils contenaient trop de bases ambiguës?

Renommez le fichier *filtered-sequences.qza* en *demux_filtered.qza*.
Nous allons maintenant visualiser ce fichier en utilisant l'outil *qiime2 demux summarize* avec ces paramètres:

```
data: demux_filtered.qza
```

> Combien de reads restent-ils au total après filtrage?
> Combien de reads restent-ils pour l'échantillon L1S76?
> Observez le plot de qualité. Que remarquez-vous?

Clustering

Maintenant que les données sont prêtes, nous allons générer les OTU, ainsi que l'arbre phylogénétique qui seront utilisés dans l'analyse secondaire. Nous allons utiliser VSEARCH.

Dé-réplication

La première étape est de dé-répliquer les séquences, le but ici étant de grouper les séquences identiques afin de réduire le temps d'exécution. Nous allons utiliser *qiime2 vsearch dereplicate-sequences* avec ces paramètres:

```
sequences: demux_filtered.qza
```

Deux fichiers sont générés: *dereplicated-sequences.qza* qui contient les séquences uniques et *dereplicated-table.qza* qui contient le nombre d'occurences de chaque séquence unique.

Il est possible d'obtenir des informations sur l'exécution du job en cliquant sur le pictogramme **i** *dataset details*. Cette page donne différents types d'information: les jeux de données d'entrée, les paramètres utilisés par le programme, les sorties, le temps d'exécution ou encore les ressources utilisées. Dans la partie *Job Information*, vous pouvez voir les sorties de l'outils. Dans la partie *Tool Standard Error*, vous avez des informations sur l'étape de dé-réplication.

> Quelle est la taille moyenne des séquences d'entrée?
> Combien de séquences uniques ont été identifiées?

Afin de visualiser la table de fréquence, lancez ensuite *qiime2 feature-table summarize* avec ces paramètres:

```
table: dereplicate-table
1: sample_metadata: Metadata: Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Combien de reads sont encodés dans cette table?
Combien de reads pour l'échantillon L1S76? Cela vous semble-t-il normal?
Combien de fois apparaît la séquence la plus fréquente?
Dans combien d'échantillons différents apparaît-elle?

Note: Cette étape ne modifie pas les données, c'est juste de la mise en forme pour la suite!

Clustering *closed-reference*

Ce type de clustering se base sur une banque de référence. Nous allons charger la banque 97_otus de GreenGenes ainsi que la taxonomie associée.

Téléchargez puis chargez les fichiers suivant (*upload*):

- 97_otus.fasta
- 97_otu_taxonomy.tsv

Il faut ensuite importer ces fichiers en QZA. Utilisez l'outil *qiime2 tools import* avec ces paramètres:

```
Type of data to import: FeatureData[Sequence]
QIIME 2 file format to import from: DNAFASTA Format
data: 97_otus.fasta
```

Renommez le fichier généré en *97_otus.qza*.

Utilisez l'outil *qiime2 tools import* avec ces paramètres:

```
Type of data to import: FeatureData[Taxonomy]
QIIME 2 file format to import from: HeaderlessTSVTaxonomyFormat
data: 97_otu_taxonomy.tsv
```

Renommez le fichier généré en *97_otu_taxonomy.qza*.

Nous allons maintenant lancer le clustering. Pour ce faire utilisez l'outil *qiime2 vsearch cluster-features-closed-reference* avec ces paramètres:

```
sequences: dereplicated-sequences.qza
table: dereplicated-table.qza
reference_sequences: 97_otus.qza
perc_identity: 0.97
```

3 fichiers sont générés:

- *clustered-table.qza* est la table d'OTU
- *clustered-sequences.qza* contient les séquences représentatives pour chaque OTU (à partir des centroïdes)
- *unmatched-sequences.qza* contient les séquences qui n'ont pas été clusterisées

Nous allons visualiser la table d'OTU en utilisant *qiime2 feature-table summarize* avec ces paramètres:

```
table: clustered-table.qza
sample_metadata: Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Combien de reads sont répartis en OTU?
Où se trouvent les reads manquants?
Combien y-a-il de reads pour l'échantillon L1S76?
Combien d'OTU ont été formés?
Combien de reads contient l'OTU le plus gros?
Dans combien d'échantillons cet OTU est-il retrouvé?

Clustering *de novo*

Nous allons maintenant utiliser un clustering *de novo*, ce type de clustering n'utilise aucune banque de référence pour la formation des OTU.

Utilisez l'outil *qiime2 vsearch cluster-features-de-novo* avec ces paramètres:

```
--i-sequences: dereplicated-sequences.qza
--i-table: dereplicated-table.qza
--p-perc-identity: 0.97
```

Nous allons visualiser la table d'OTU en utilisant *qiime2 feature-table summarize* avec ces paramètres:

```
table: clustered-table.qza
sample_metadata: Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Combien de reads sont répartis en OTU?
Est-ce attendu?
Combien d'OTU ont été formés?
Combien de reads contient l'OTU le plus gros?
Dans combien d'échantillons cet OTU est-il retrouvé?

Clustering *open-reference*

Pour finir, nous allons réalisé un clustering de type *open-reference*, qui est un mélange des deux méthodes précédentes. Dans un premier temps, les OTU sont formés par comparaison à une banque, les reads non-clusterisés à cette étape sont soumis à un clustering *de novo*.

Nous allons utilisé l'outil *qiime2 vsearch cluster-features-open-reference* avec ces paramètres:

```
sequences: dereplicated-sequences.qza
table: dereplicated-table.qza
reference-sequences: 97_otus.qza
perc-identity: 0.97
```

Nous allons visualiser la table d'OTU en utilisant *qiime2 feature-table summarize* avec ces paramètres:

```
table: clustered-table.qza
sample_metadata: Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Combien de reads sont répartis en OTU?
Est-ce normal?
Combien d'OTU ont été formés?
Combien de reads contient l'OTU le plus gros?
Dans combien d'échantillons cet OTU est-il retrouvé?

BILAN: comparez le nombre de reads et d'OTU pour chaque type de clustering. Quelles observations pouvez-vous faire?

ATTENTION: pour la suite nous allons utiliser uniquement les résultats du clustering *open-reference*. Les mêmes étapes devraient être réalisées pour les autres méthodes.

Filtrage des chimères

Une fois les OTU formés ils nous faut identifier et filtrer les chimères.

Utilisez l'outil *qiime2 vsearch uchime-denovo* avec ces paramètres:

```
sequences: clustered-sequences.qza
table: clustered-table.qza
```

Cette étape génère 3 fichiers:

- chimeras.qza: contient les reads chimériques
- non-chimeras.qza: contient les reads non-chimériques
- stats.qza: contient les statistiques de l'étape d'identification des chimères (scores, statistiques internes au programme...)

En cliquant sur le pictogramme **i**, dans la partie *Tool Standard Error*, vous avez les statistiques concernant l'identification des séquences chimériques.

Combien de séquences chimériques ont été identifiées?

Nous allons maintenant filtrer les chimères de la table d'OTU générée précédemment.

Utilisez l'outil *qiime2 feature-table filter-features* avec ces paramètres:

```
table: clustered-table.qza
metadata: Metadata from Artifact
Metadata Source: chimeras.qza
exclude_ids: Bool: YES
```

Cet outil permet de filtrer une table d'OTU selon différents critères (métadonnées, sequences à exclure ou à conserver...).

Il nous faut également filtrer les séquences représentatives pour éliminer celles correspondant aux chimères. Pour ce faire utilisez l'outil *qiime2 feature-table filter-seqs* avec ces paramètres:

```
data: clustered-sequences.qza
metadata: Metadata : Metadata from Artifact
Metadata Source: chimeras.qza
exclude_ids: Bool: YES
```

Enfin nous allons visualiser la table d'OTU filtrée, c'est à dire sans les chimères. Utilisez l'outil *qiime2 feature-table summarize* avec ces paramètres:

```
table: filtered_table.qza
sample_metadata: Metadata : Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Combien de reads restent-ils dans la table après filtrage des chimères?
A combien d'OTU cela correspond-il?

Filtrage des singletons

Le clustering mène souvent à la formation de singletons, c'est-à-dire des OTU composés d'un seul read. Il est conseillé de filtrer les singletons pour le reste des analyses, ils sont souvent considérés comme du bruit plutôt que provonant d'organismes rares.

Utilisez l'outil *qiime2 feature-table filter-features* avec ces paramètres:

```
table: filtered_table.qza
min-frequency:2
```

Renommez le fichier généré en *noSingleton_table.qza*.

Pour visualiser la table sans singletons, utilisez *qiime2 feature-table summarize* avec ces paramètres:

```
table: noSingleton_table.qza
sample_metadata: Metadata : Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Combien de reads restent-ils dans la table après filtrage des singletons?
Combien y-a-il il de reads pour l'échantillon L1S76?
Combien d'OTU ont été formés?
Commentez ces chiffres.

Nous allons également filtrer les séquences correspondant aux singletons au sein des séquences représentatives. Utilisez l'outil *qiime2 feature-table filter-seqs* avec ces paramètres:

```
data: filtered-data.qza
table: noSingleton_table.qza
```

Renommez le fichier généré en *noSingleton_sequences.qza*.

Arbre phylogénétique

Nous allons générer un arbre phylogénétique à partir des séquences de nos échantillons qui nous sera utile pour l'analyse secondaire. Les séquences représentatives vont être alignées et un arbre phylogénétique sera construit à partir de cet alignement.

Réalisez l'alignement multiple des séquences représentatives avec *qiime2 alignment mafft* avec ces paramètres:

```
sequences: noSingleton_sequences.qza
```

Il faut ensuite masquer les parties de l'alignement qui ne sont pas correctes (gaps) avant de créer l'arbre phylogénétique. Utilisez *qiime2 alignment mask*

avec ces paramètres:

```
alignment: alignment.qza
```

Nous allons maintenant généré l'arbre phylogénétique. Utilisez *qiime2 phylogeny fasttree* avec ces paramètres:

```
alignment: masked_alignment.qza
```

Enfin, il faut enraciner l'arbre en utilisant *qiime2 phylogeny midpoint-root* avec ces paramètres:

```
tree: tree.qza
```

Note: L'arbre phylogénétique peut être exporté pour être visualisé avec un logiciel adapté.

Assignation taxonomique

Une fois les OTU formés, l'étape suivante est d'assigner une taxonomie à chaque OTU. Pour ce faire, une séquence représentative pour chaque OTU est annotée en la comparant à une banque de référence, cette annotation étant étendue à tous les reads appartenant à l'OTU.

Ici, nous allons utilisé le classifier de VSEARCH. Il compare la séquence à une banque et assigne la taxonomie consensus des différents hits. Utilisez l'outil *qiime2 feature-classifier classify-consensus-vsearch* avec ces paramètres:

```
query: noSingleton_sequences.qza
reference-reads: 97_otus.qza
reference_taxonomy: 97_otus_taxonomy.qza
```

Pour visualiser le fichier produit utilisez *qiime2 metadata tabulate* avec ces paramètres:

```
input: Metadata: Metadata from Artifact
Metadata Source: classification.qza
```

> Que contient le fichier *classification.qza*?

Il est possible de générer des barplots pour visualiser la composition de chaque échantillon. Utilisez l'outil *qiime2 taxa barplot* avec ces paramètres:

```
table: FeatureTable: noSingleton_table.qza
taxonomy: classification.qza
sample_metadata: Metadata : Metadata from TSV
Metadata Source: sample-metadata.tsv
```

> Quel est le phylum (level 2) majoritaire dans les échantillons provenant de l'intestin?
> Observez-vous des différences de composition entre les différentes parties du corps?
> Observez-vous des différences significatives entre les deux sujets?
> Observez-vous des différences significatives entre les points T=0 de l'expérience et les autres points?

Nous voici à la fin de l'analyse primaire avec QIIME2. Nous disposons d'une table d'OTU, de la taxonomie associée et d'un arbre phylogénétique qui seront utilisés lors de l'analyse secondaire.

Analyse primaire avec les ASV

Création des ASV

Nous allons maintenant réaliser l'analyse avec les ASV. Pour ce faire nous allons utiliser DADA2. Cet algorithme étant un *denoiser* nous n'avons pas besoin des étapes de pre-processing. L'un des paramètres très important de DADA2 est *trunc_len* qui correspond à la position à laquelle les séquences doivent être tronquées à cause de la baisse de la qualité. Lorsque ce paramètre est mal ajusté, par exemple pas assez grand, la formation des ASV est biaisées car DADA2 ne parvient pas à identifier correctement les erreurs en fin de reads. Il faut cependant garder un maximum d'information et donc ne pas trop tronquer les reads.

Nous allons créer un nouvel historique. Pour ce faire, cliquez sur le signe **+** en haut dans le menu de droite. Créez un nouvel historique et nommez le *primaire_ASV*. En cliquant sur *Show hitories side-by-side* tous les historiques s'affichent. Vous pouvez ainsi copier les fichiers *sample-metadata.tsv*, *demux.qza* et son fichier de visualisation à partir de l'historique précédent.

> Reprenez les statistiques du démultiplexage. A partir du plot interactif de qualité, déterminez la valeur à laquelle les reads doivent être tronqués.

Nous allons maintenant procéder à la formation des ASV. Pour ce faire, utilisez l'outil *qiime2 dada2 denoise-single* avec ces paramètres:

```
demultiplexed_seqs: demux.qza
trunc-len: 120
```

> A quelle taille les reads ont-ils été tronqués?

DADA2 génère 3 fichiers:

- table.qza: la table d'ASV
- representative-sequences.qza: les séquences représentatives pour chaque ASV
- denoising-stats.qza: statistiques du denoising contenant le nombre de reads restant à chaque étape de l'algorithme

Nous allons étudier le fichier contenant les statistiques du denoising. Utilisez l'outil *qiime2 metadata tabulate* avec ces paramètres:

```
input: Metadata: Metadata from Artifact
Metadata Source: denoising-stats.qza
```

> Observez le nombre de reads restants à chacune des étapes. Que remarquez-vous? Quelle est l'étape qui élimine le plus de reads?

Nous allons maintenant étudier la table d'ASV générés par DADA2. Utilisez l'outil *qiime2 feature-table summarize* avec ces paramètres:

```
table: table.qza
sample_metadata: Metadata: Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Combien de reads sont répartis en ASV?
Quel est le nombre d'ASV générés par DADA2?
Combien y-a-il il de reads pour l'échantillon L1S76?
Combien de reads contient l'ASV le plus gros?
Dans combien d'échantillons cet ASV est-il retrouvé?
Comparer ces chiffres aux résultats obtenus avec les OTU.

Nous allons maintenant étudier les séquences représentatives. Utilisez l'outil *qiime2 feature-table tabulate-seqs* avec ces paramètres:

```
data: representative-sequences.qza
```

Que contient ce fichier?
De quel organisme semble provenir l'ASV f023384b8f989d014dd2ead7f10db307?

Arbre phylogénétique

De la même façon que pour la partie précédente nous allons générer l'arbre phylogénétique qui sera utilisé dans certaines analyses secondaires.

Utilisez *qiime2 alignment mafft* avec ces paramètres:

```
sequences: representative-sequences.qza
```

Utilisez *qiime2 alignment mask* avec ces paramètres:

```
alignment: alignment.qza
```

Utilisez *qiime2 phylogeny fasttree* avec ces paramètres:

```
alignment: masked_alignment.qza
```

Utilisez *qiime2 phylogeny midpoint-root* avec ces paramètres:

```
tree: tree.qza
```

Assignation taxonomique

Nous allons utiliser le classifier *classify-sklearn* pour réaliser l'assignation taxonomique.

Télécharger et charger le fichier gg-13-8-99-515-806-nb-classifier.qza.

Utilisez l'outil *qiime2 feature-classifier classify-sklearn* avec ces paramètres:

```
reads: representative-sequences.qza
classifier: gg-13-8-99-515-806-nb-classifier.qza
```

Pour la visualisation utilisez l'outil *qiime2 metadata tabulate* avec ces paramètres:

```
input: Metadata: Metadata from Artifact
Metadata Source: classification.qza
```

Que contient le fichier?
Quelle est l'assignation de l'ASV f023384b8f989d014dd2ead7f10db307?
Comparer avec l'assignation taxonomique obtenue avec BLAST de la partie précédente.
A quel niveau taxonomique divergent-elles (espèce, genre, famille...)?

Nous allons maintenant visualiser la composition taxonomique à l'aide d'un barplot. Utilisez l'outil *qiime2 taxa barplot* avec ces paramètres:

```
table: table.qza
taxonomy: classification.qza
metadata: Metadata: Metadata from TSV
Metadata Source: sample-metadata.tsv
```

Quel est le phylum (level 2) majoritaire dans les échantillons provenant de l'intestin?
Observez-vous des différences de composition entre les différentes parties du corps?
Observez-vous des différences significatives entre les deux sujets?
Observez-vous des différences significatives entre les points T=0 de l'expérience et les autres points?
Comparez avec les résultats obtenus avec les OTU.

Nous voici à la fin de l'analyse primaire avec DADA2. Nous disposons d'une table d'ASV, de la taxonomie associée et d'un arbre phylogénétique qui seront utilisés lors de l'analyse secondaire.

Analyse secondaire - Fichier ASV

Analyse de diversité alpha et bêta

Les analyses de diversité de QIIME 2 sont disponibles via le plugin q2-diversity, qui prend en charge le calcul des mesures de diversité alpha et bêta, l'application des tests statistiques associés et la génération de visualisations interactives.
Nous allons appliquer la méthode core, qui raréfie votre FeatureTable[Fréquence] à une profondeur spécifiée par l'utilisateur, calcule plusieurs mesures de diversité alpha et bêta, et génère des tracés d'analyse des coordonnées de principe (PCoA) en utilisant **Emperor** pour chaque mesure de diversité bêta.
Les métriques calculées par défaut sont :

- Diversité alpha
  - Indice de diversité de Shannon: une mesure quantitative de la richesse de la communauté
  - observed-OTU: mesure qualitative de la richesse de la communauté
  - La pd-faith_tree: une mesure qualitative de la richesse de la communauté qui intègre un arbre phylogénétique
  - Evenness: une mesure de l'uniformité de la communauté
- Diversité bêta
  - Distance Jaccard: une mesure qualitative de la dissimilitude de la communauté
  - Distance de Bray-Curtis: mesure quantitative de la dissimilitude de la communauté
  - Distance UniFrac non-pondérée: une mesure qualitative de la dissimilarité de la communauté qui intègre un arbre phylogénétique
  - Distance pondérée UniFrac: une mesure quantitative de la dissimilarité de la communauté qui intègre un arbre phylogénétique

Le paramètre important qui doit être fourni par l'utilisateur est

`--p-sampling-depth`

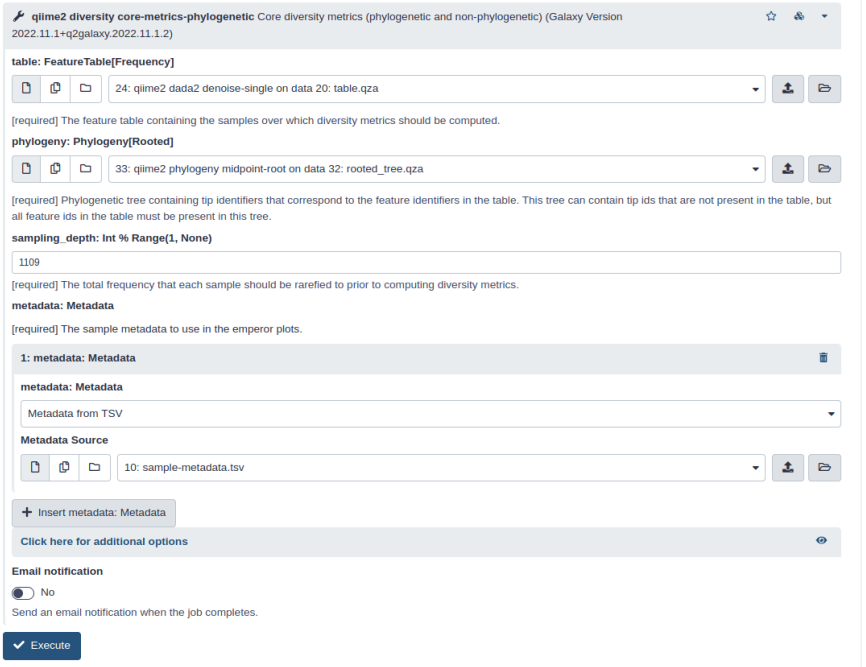
C'est la profondeur d'échantillonnage (c'est-à-dire la raréfaction). Comme la plupart des mesures de diversité sont sensibles à une variation de profondeur dans les différents échantillons, ce script sous-échantillonnera au hasard les comptages de chaque échantillon à la valeur fournie pour ce paramètre. Par exemple, si vous fournissez *--p-sampling-depth* à 500, cette étape sous-échantillonnera les comptages dans chaque échantillon sans remplacement de sorte que chaque échantillon dans le tableau résultant ait un total de 500.

Attention, si pour certains échantillons le nombre de comptage est inférieur à cette valeur, ces échantillons seront retirés de l'analyse de diversité. Le choix de cette valeur est délicat. Nous vous recommandons de faire votre choix en consultant les informations présentées dans le fichier table.qzv qui a été créé durant l'analyse primaire et en choisissant une valeur aussi élevée que possible (afin de conserver plus de séquences par échantillon) tout en excluant aussi peu d'échantillons que possible.

> Visualisez la table.qzv QIIME2, et en particulier l'onglet "Interactive Sample Detail". Quelle valeur choisiriez-vous pour la profondeur d'échantillonnage ("--p-sampling-depth")?
> Combien d'échantillons seront exclus de votre analyse en fonction de ce choix ? Combien de séquences totales allez-vous analyser avec cette profondeur?

Nous allons maintenant estimer les diversités alpha et beta. Utilisez l'outil *qiime diversity core-metrics-phylogenetic* avec ces paramètres:

```
--i-phylogeny rooted-tree.qza
--i-table table.qza
--p-sampling-depth 1109
--m-metadata-file sample-metadata.tsv
```



Ici, nous réglons le paramètre *--p-sampling-depth* depth à **1109**.
Cette valeur a été choisie en fonction du nombre de séquences de l'échantillon L3S341 parce qu'elle est proche du nombre de séquences des quelques échantillons suivants qui ont un nombre de séquences plus élevé, et parce qu'elle est plus élevée que le nombre de séquences de l'échantillon précédent qui a moins de séquences. Cela nous permettra de conserver la plupart de nos échantillons. L'échantillon qui a le moins de séquences sera éliminé des analyses *core-metrics-phylogenetic* et de tout ce qui utilise ces résultats.
Note
La profondeur d'échantillonnage de **1109** a été choisie en fonction de DADA2. Si vous utilisez une méthode comme Deblur (ou Vsearch sur les otus), vous voudrez peut-être choisir une profondeur d'échantillonnage différente. Appliquez la logique du paragraphe précédent pour vous aider à choisir une profondeur d'échantillonnage.

Note
Dans de nombreux runs Illumina, vous observerez quelques échantillons dont le nombre de séquences est très faible. Vous voudrez généralement les exclure de l'analyse en choisissant une valeur plus grande pour la profondeur d'échantillonnage.

Représentation graphique de l'alpha rarefaction

Dans cette section, nous explorerons la diversité alpha en fonction de la profondeur d'échantillonnage à l'aide de *qiime alpha-rarefaction*.
Cet outil calcule une ou plusieurs mesures de diversité alpha à des profondeurs d'échantillonnage multiples, par étapes comprises entre 1 (optionnellement contrôlé avec *--p-min-depth*) et la valeur fournie comme *--p-max-depth*.
A chaque étape de profondeur d'échantillonnage, 10 tables raréfiées seront générées, et les mesures de diversité seront calculées pour tous les échantillons dans les tables. Le nombre d'itérations (tables raréfiées calculées à chaque profondeur d'échantillonnage) peut être contrôlé avec *--p-itérations*.
Les valeurs moyennes de diversité seront tracées pour chaque échantillon à chaque profondeur d'échantillonnage, et les échantillons peuvent être regroupés en fonction des métadonnées dans la visualisation résultante si les métadonnées de l'échantillon sont fournies avec le paramètre *--m-metadata-file*.

Nous allons maintenant generer ces courbes de rarefaction. Utilisez l'outil *qiime diversity alpha-rarefaction* avec ces paramètres:

```
--i-table table.qza
--i-phylogeny rooted-tree.qza
--p-max-depth 4000
--m-metadata-file sample-metadata.tsv
```

La visualisation aura deux tracés.
le premier graphique sert principalement à déterminer si la richesse des échantillons a été entièrement observée ou séquencée.
Si les lignes de la courbe semblent s'égaliser (c.-à-d. s'approcher d'une pente de zéro) à une certaine profondeur d'échantillonnage le long de l'axe des x, cela laisse à penser que la collecte de séquences supplémentaires au-delà de cette profondeur d'échantillonnage n'entraînerait probablement pas l'observation d'autres caractéristiques.
Si les lignes d'un graphique ne s'égalisent pas, c'est peut-être parce que la richesse des échantillons n'a pas encore été entièrement observée (parce que trop peu de séquences ont été générées).

Le graphe du bas de cette visualisation est important pour le regroupement des échantillons par métadonnées.
Il illustre le nombre d'échantillons qui restent dans chaque groupe lorsque le tableau des caractéristiques est raréfié à chaque profondeur d'échantillonnage.
Si une profondeur d'échantillonnage donnée **d** est supérieure à la fréquence totale d'un échantillon **s** (c.-à-d. le nombre de séquences obtenues pour un échantillon **s**), il n'est pas possible de calculer la mesure de la diversité pour les échantillons **s** à la profondeur d'échantillonnage **d**.
Si plusieurs des échantillons d'un groupe ont des fréquences totales inférieures à **d**, la diversité moyenne présentée pour ce groupe à **d** dans la courbe supérieure sera peu fiable car elle aura été calculée sur quelques échantillons. Lorsque l'on regroupe des échantillons par métadonnées, il est donc essentiel d'examiner le graphique du bas pour s'assurer que les données présentées dans le graphique du haut sont fiables.

Note
La valeur que vous fournissez pour *--p-max-depth* doit être déterminée en examinant l'information *Frequency per sample* présentée dans le fichier *table.qzv* qui a été créé durant l'analyse primaire.
En général, le choix d'une valeur qui se situe quelque part autour de la fréquence médiane semble bien fonctionner, mais vous voudrez peut-être augmenter cette valeur si les lignes de la courbe de raréfaction résultante ne semblent pas s'égaliser, ou diminuer cette valeur si vous semblez perdre plusieurs de vos échantillons en raison de fréquences totales faibles plus près de la profondeur d'échantillonnage minimale que la profondeur maximale.

> Lors du regroupement des échantillons par *BodySite* et de l'affichage de la courbe de raréfaction alpha pour la mesure "observed_otus", quels *BodySite* (le cas échéant) semblent présenter une couverture de diversité suffisante (c'est-à-dire que leurs courbes de raréfaction se stabilisent) ?
>
> En groupant les échantillons par *BodySite* et en visualisant la courbe de raréfaction alpha pour la métrique "observed_otus", la ligne pour les échantillons de la "right palm" semble se stabiliser à environ 40, mais elle saute ensuite à environ 140. Quelle peut etre la raison ? (Indice : assurez-vous de regarder à la fois les tracés du haut et du bas.)

Analyse differentielle - Alpha diversité

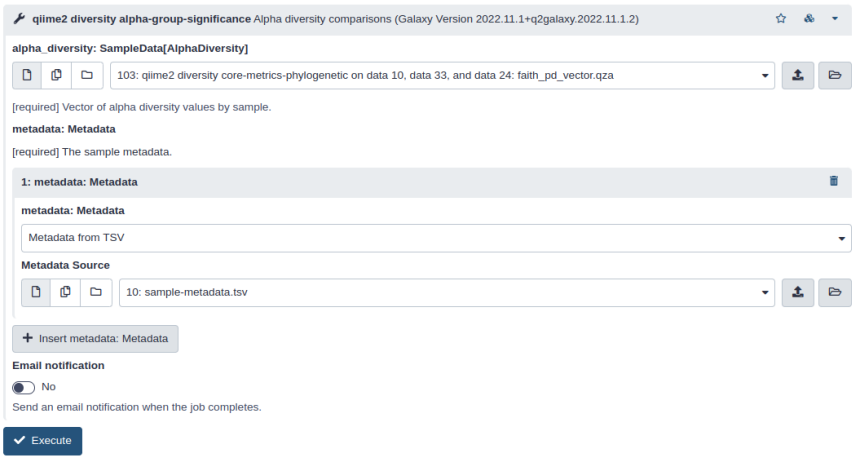
Après avoir calculé les paramètres de diversité, nous pouvons commencer à explorer la composition microbienne des échantillons dans le contexte des métadonnées de l'échantillon. Cette information est présente dans l'exemple de fichier de métadonnées qui a été chargé plus tôt dans Galaxy.
Nous allons d'abord tester les associations entre les colonnes de métadonnées catégorielles et les données sur la diversité alpha. C'est ce que nous allons faire ici pour la diversité phylogénétique faith_pd_vector, eveness_vector et les meta-données.

Nous allons maintenant tester ces associations. Utilisez l'outil *qiime diversity alpha-group-significance* avec ces paramètres:

```
--i-alpha-diversity core-metrics-results/faith_pd_vector.qza
--m-metadata-file sample-metadata.tsv
```

Deuxieme test d'association :

```
--i-alpha-diversity core-metrics-results/evenness_vector.qza
--m-metadata-file sample-metadata.tsv
```



> Quelles colonnes de métadonnées d'échantillons catégorielles sont les plus fortement associées aux différences en terme de diversité phylogénétique (faith'pd) de la communauté microbienne ? Ces différences sont-elles statistiquement significatives ?
>
> Quelles colonnes de métadonnées d'échantillons catégorielles sont les plus fortement associées aux différences en terme d'équitabilité (evenness) de la communauté microbienne ? Ces différences sont-elles statistiquement significatives ?

Dans cet ensemble de données, aucune colonne de métadonnées d'échantillonnage continu (p. ex., DaysSinceExperimentStart) n'est corrélée à la diversité alpha ; nous ne ferons donc pas de test pour ces associations ici. Si vous souhaitez effectuer ces tests (pour cet ensemble de données ou pour d'autres), vous pouvez utiliser la commande *qiime diversity alpha-correlation*.

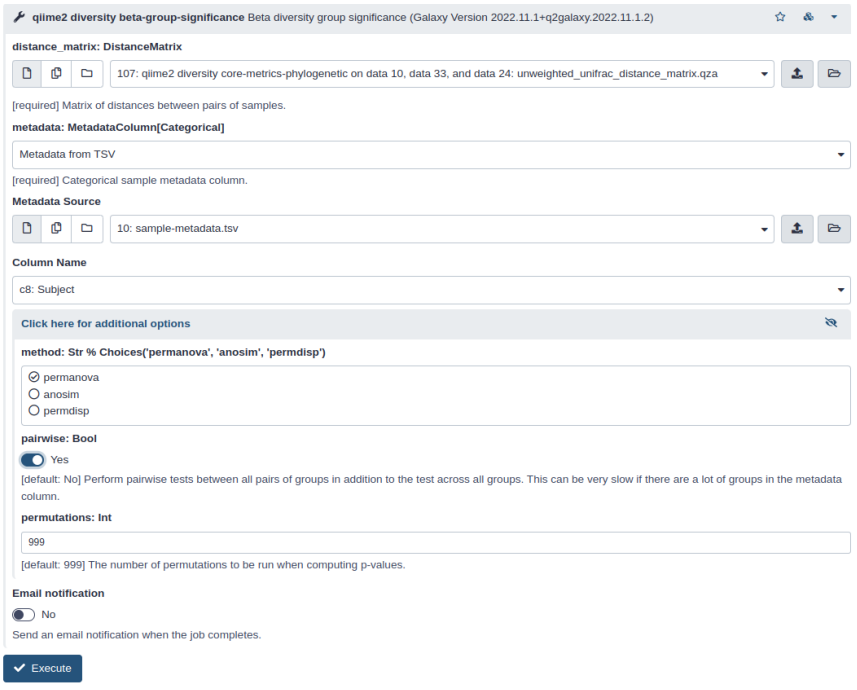
Analyse de diversité bêta

Ensuite, nous analyserons la composition de l'échantillon dans le contexte des métadonnées catégorielles à l'aide de PERMANOVA (décrite pour la première fois dans Anderson (2001)) à l'aide de la commande *qiime beta-groupe-significance*.
Les commandes suivantes permettent de vérifier si les distances entre les échantillons au sein d'un groupe, comme les échantillons prélevés au même endroit du corps (p. ex., l'intestin), sont plus semblables entre elles qu'entre les échantillons des autres groupes (p. ex., langue, paume gauche et paume droite).
Nous allons l'exécuter sur des colonnes spécifiques de métadonnées que nous souhaitons explorer, plutôt que sur toutes les colonnes de métadonnées auxquelles il est applicable.
Ici, nous allons l'appliquer à nos distances UniFrac non-pondérées, en utilisant deux colonnes de métadonnées d'échantillonnage, comme suit.
Nous allons maintenant tester ces associations. Utilisez l'outil *qiime diversity beta-group-significance* avec ces paramètres:

```
--i-distance-matrix core-metrics-results/unweighted_unifrac_distance_matrix.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column BodySite
--p-pairwise
```

Deuxieme test d'association:

```
--i-distance-matrix core-metrics-results/unweighted_unifrac_distance_matrix.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column Subject
--p-pairwise
```



> Les associations entre les sujets et les différences de composition microbienne sont-elles statistiquement significatives ? Et les *BodySite* ? Quelles

paires spécifiques de *BodySite* sont significativement différentes les unes des autres ?

Encore une fois, aucune des métadonnées de l'échantillon continu (variables qualitatives : *Month*, *Days*,*DaysSinceExperimentStart*) que nous avons pour cet ensemble de données n'est corrélée à la composition de l'échantillon ; nous ne ferons donc pas de test pour ces associations ici. Si vous souhaitez effectuer ces tests, vous pouvez utiliser *qiime metadata distance-matrix* en combinaison avec les commandes *qiime diversity mantel* et *qiime diversity bioenv*.

Enfin, l'ordination est une approche populaire pour explorer la composition des communautés microbiennes dans le contexte des métadonnées d'échantillonnage. Nous pouvons utiliser l'outil **Emperor** pour explorer les tracés de coordonnées principales (PCoA) dans le contexte des métadonnées échantillons.
Alors que notre commande *core-metrics-phylogenetic* a déjà généré quelques tracés **Emperor**, nous voulons passer un paramètre optionnel, *--p-custom-axes*, qui est très utile pour explorer les données des séries temporelles.
Les résultats du PCoA qui ont été utilisés dans *core-metrics-phylogeny* sont également disponibles, ce qui permet de générer facilement de nouvelles visualisations avec **Emperor**.

Nous allons explorer les tracés **Emperor** pour les distances **UniFrac non pondéré** et **Bray-Curtis** afin que le tracé résultant contienne des axes pour la coordonnée principale 1, la coordonnée principale 2 et la variable **DaysSinceExperimentStart**. Nous utiliserons ce dernier axe pour explorer comment ces échantillons ont changé au fil du temps.

Nous allons maintenant générer les plot emperor. Utilisez l'outil *qiime emperor plot* avec ces paramètres:

```
--i-pcoa core-metrics-results/unweighted_unifrac_pcoa_results.qza
--m-metadata-file sample-metadata.tsv
--p-custom-axes DaysSinceExperimentStart
```

Retrouve-t-on les mêmes tendances qu'avec les graphiques de beta diversité généré par l'outil core-diversity-phylogenetic ? (Indice : Expérimentez avec des points de coloration par différentes métadonnées.)

Tests d'abondance différentielle avec ANCOM

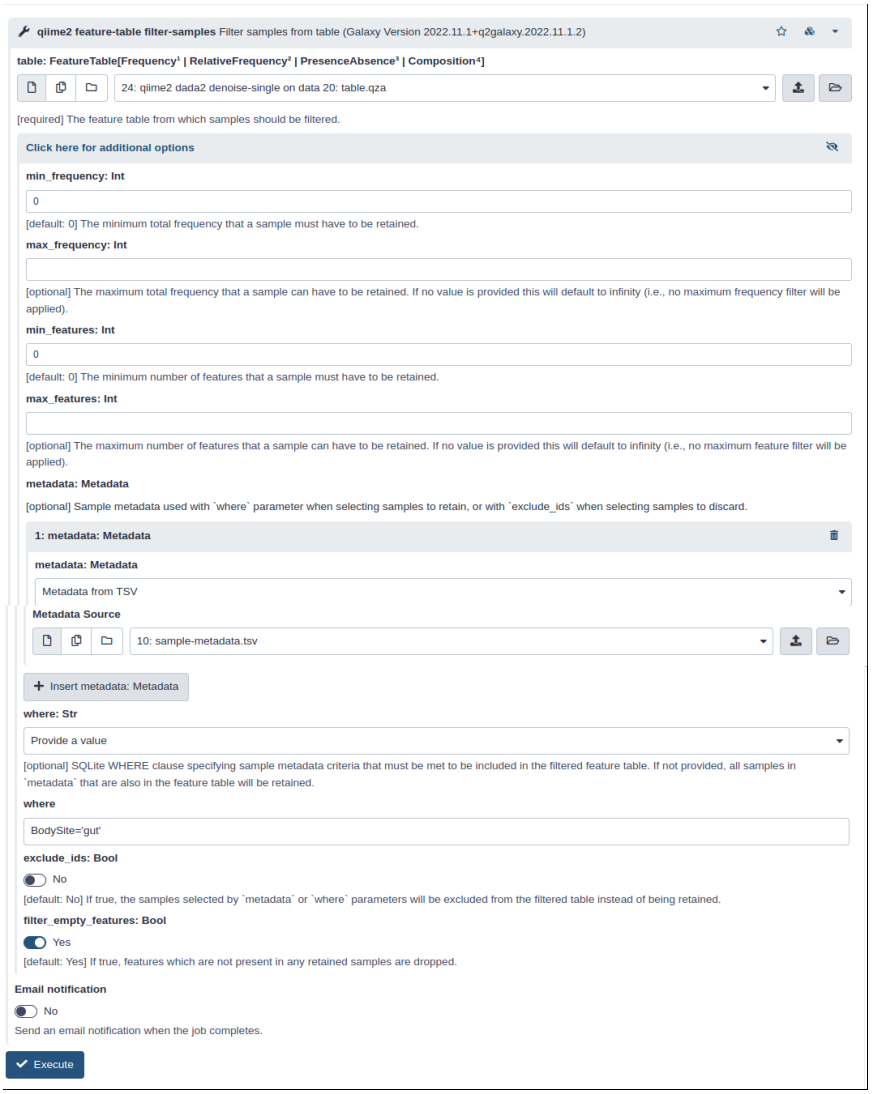
L'ANCOM peut être utilisé pour identifier les ASVs qui sont différentiellement exprimés (c.-à-d. présentes en différentes abondances) dans les groupes d'échantillons.
Attention, comme pour toute méthode statistique, vous devez connaître les hypothèses et les limites de l'ANCOM avant de l'utiliser.

Note
L'analyse de l'abondance différentielle dans l'analyse microbiologique est un domaine de recherche actif. Il y a deux plugins QIIME 2 qui peuvent être utilisés pour cela : q2-gneiss et q2-composition. Cette section utilise q2-composition.

ANCOM est implémenté dans le plugin q2-composition.
ANCOM suppose que peu d'ASVs (moins de 25 % environ) changent d'un groupe à l'autre.
Si vous vous attendez à ce qu'un plus grand nombre d'ASVs changent entre vos groupes, vous ne devriez pas utiliser ANCOM car il sera plus sujet aux erreurs (une augmentation des erreurs de type I et de type II est possible).
Parce que nous nous attendons à ce que beaucoup d'ASVs changent en abondance sur les *BodySite*, dans ce tutoriel, nous allons filtrer notre table de départ (table.qza) pour ne contenir que des échantillons de *gut*.
Nous appliquerons ensuite l'ANCOM pour déterminer quelles variantes de séquence (ASV) et quels genres, s'il y en a, sont différentiellement exprimés dans les échantillons de *gut* de nos deux sujets.

Nous allons maintenant filter nos échantillons. Utilisez l'outil *qiime feature-table filter-samples* avec ces paramètres:

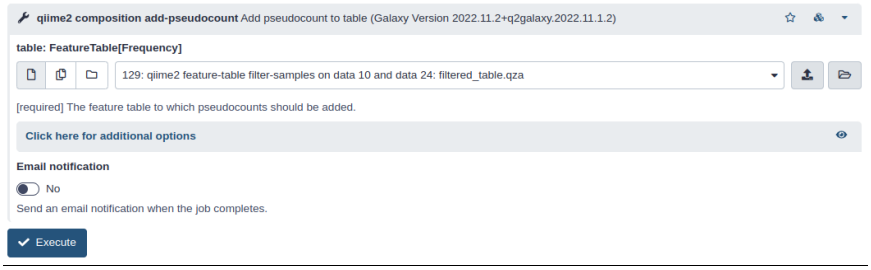
```
--i-table table.qza
--m-metadata-file sample-metadata.tsv
--p-where "BodySite='gut'"
```



ANCOM fonctionne sur le dataset FeatureTable[Composition] QIIME 2, qui est basé sur des fréquences des ASVs par échantillon, mais ne peut tolérer des fréquences égales à zéro. Pour construire le dataset de composition, un dataset FeatureTable[Frequency] doit être fourni pour add-pseudocount (méthode d'imputation), qui produira le dataset FeatureTable[Composition].

Nous allons maintenant filter nos échantillons. Utilisez l'outil *qiime composition add-pseudocount* avec ces paramètres:

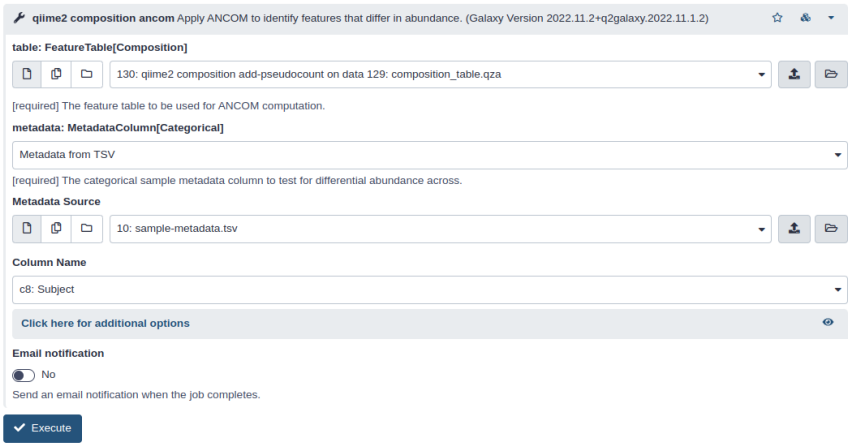
```
--i-table filtered-gut-table.qza
```



Nous pouvons ensuite exécuter ANCOM sur la colonne *Subject* pour déterminer quelles ASVs diffèrent en abondance dans les échantillons de *gut* des deux sujets.

Nous allons maintenant executer le test ANCOM. Utilisez l'outil *qiime composition ancom* avec ces paramètres:

```
--i-table comp-gut-table.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column Subject
```

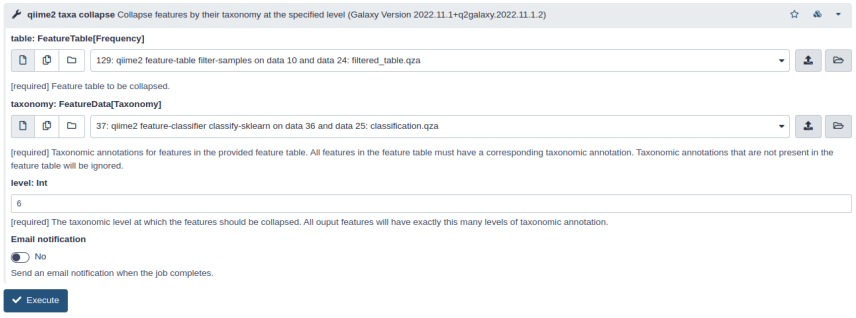


> Quelles ASVs diffèrent dans l'abondance selon le sujet ?
> Quelles sont les taxonomies des ASVs ? (Pour répondre à la dernière question, vous devrez vous référer à une autre visualisation qui a été générée dans ce tutoriel.)

Nous sommes aussi souvent intéressés à effectuer un test d'abondance différentielle à un niveau taxonomique spécifique. Pour ce faire, nous pouvons réduire notre FeatureTable[Fréquence] au niveau taxonomique d'intérêt, puis réexécuter les étapes ci-dessus. Dans ce tutoriel, nous réduisons notre table au niveau du genre (c'est-à-dire au niveau 6 de la taxonomie de Greengenes).

Nous allons maintenant executer le test ANCOM au niveau taxonomique 6. Utilisez l'outil *qiime taxa collapse* avec ces paramètres:

```
--i-table filtered-gut-table.qza
--i-taxonomy classification.qza
--p-level 6
```



Puis l'outil *qiime composition add-pseudocount* avec ces paramètres:

```
--i-table gut-table-l6.qza
```

Enfin l'outil *qiime composition ancom* avec ces paramètres:

```
--i-table comp-gut-table-l6.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column Subject
```

> Quels sont les genres différentiellement abondants en fonction du sujet ? Dans quel sujet chaque genre est-il le plus abondant ?

Analyse secondaire - Fichier OTU

Ouvrir l'historique contenant l'analyse primaire OTU

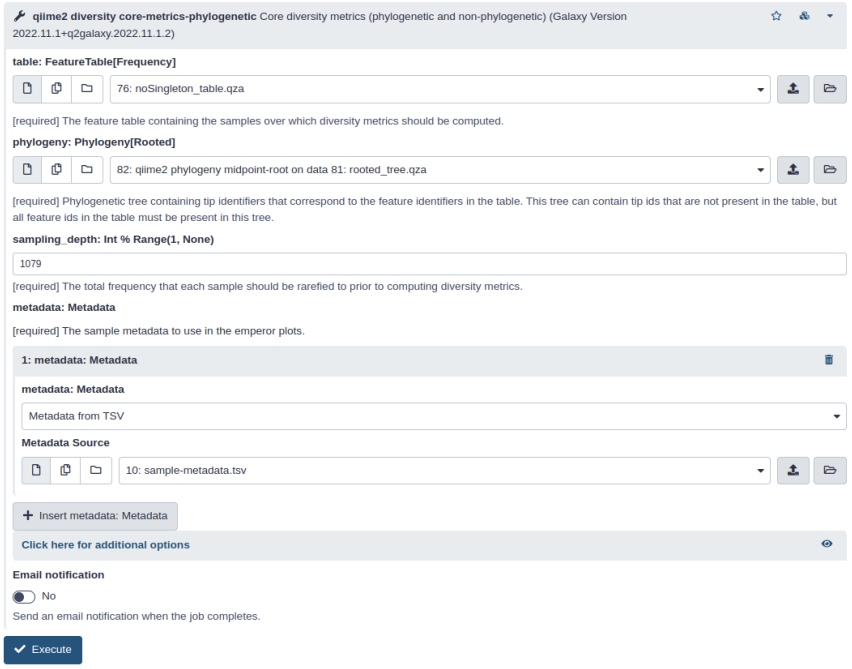
Analyse de diversité alpha et bêta

Diversité alpha

> Visualisez noSingleton_table.qzv, et en particulier l'onglet "Interactive Sample Detail". Quelle valeur choisiriez-vous de passer pour la profondeur ("--p-sampling-depth") d'échantillonnage ? Combien d'échantillons seront exclus de votre analyse en fonction de ce choix ? Ces échantillons sont-ils identiques aux échantillons supprimés lors de l'analyse par ASV?

Nous allons maintenant estimer les diversités alpha et beta. Utilisez l'outil *qiime diversity core-metrics-phylogenetic* avec ces paramètres:

```
--i-phylogeny rooted-tree.qza
--i-table noSingleton_table.qza
--p-sampling-depth 1079
--m-metadata-file sample-metadata.tsv
```



Représentation graphique de l'alpha rarefaction

Nous allons maintenant generer ces courbes de rarefaction Utilisez l'outil *qiime diversity alpha-rarefaction* avec ces paramètres:

```
--i-table noSingleton_table.qza
--i-phylogeny rooted-tree.qza
--p-max-depth 4700
--m-metadata-file sample-metadata.tsv
```

Lors du regroupement des échantillons par *BodySite* et de l'affichage de la courbe de raréfaction alpha pour la mesure *observed_otus*, quels *BodySite* (le cas échéant) semblent présenter une couverture de diversité suffisante (c'est-à-dire que leurs courbes de raréfaction se stabilisent) ? Les résultats sont-ils les mêmes qu'avec les ASV?

Analyse differentielle - Alpha diversité

Nous allons maintenant tester ces associations. Utilisez l'outil *qiime diversity alpha-group-significance* avec ces paramètres:

```
--i-alpha-diversity core-metrics-results/faith_pd_vector.qza
--m-metadata-file sample-metadata.tsv
```

Deuxieme test d'association :

```
--i-alpha-diversity core-metrics-results/evenness_vector.qza
--m-metadata-file sample-metadata.tsv
```

> Quelles colonnes de métadonnées d'échantillons catégorielles sont les plus fortement associées en terme de diversité phylogénétique de la communauté microbienne ? Ces différences sont-elles statistiquement significatives ? Retrouve-t-on les même effets qu'avec les ASV?

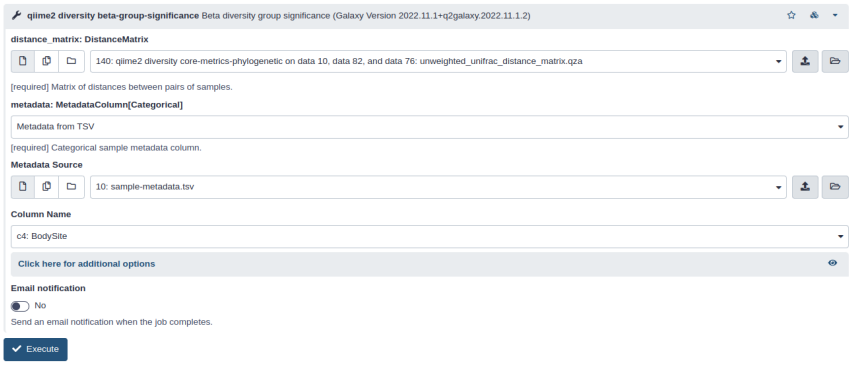
Analyse Beta-diversité

Nous allons tester les associations sur la béta-diversité:
Utilisez l'outil *qiime diversity beta-group-significance* avec ces paramètres:

```
--i-distance-matrix core-metrics-results/unweighted_unifrac_distance_matrix.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column BodySite
--p-pairwise
```

Deuxieme test d'association:

```
--i-distance-matrix core-metrics-results/unweighted_unifrac_distance_matrix.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column Subject
--p-pairwise
```

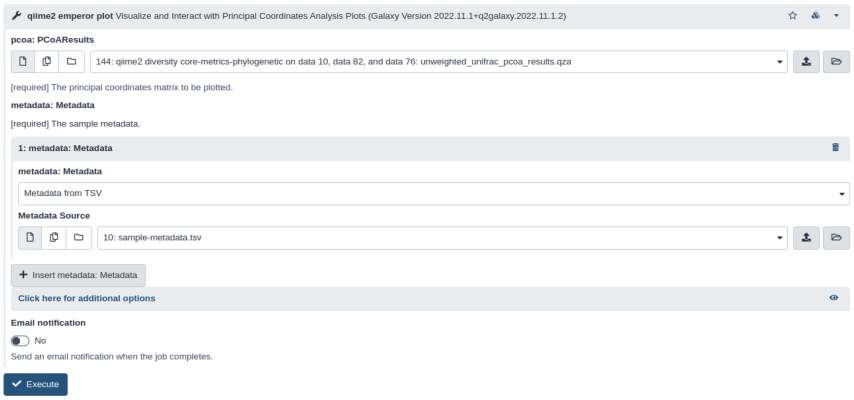


> Les associations entre les sujets et les différences de composition microbienne sont-elles statistiquement significatives ? Et les *BodySites* ? Quelles paires spécifiques de *BodySites* sont significativement différentes les unes des autres ? Retrouve-t-on les mêmes résultats qu'avec les ASV?

Visualisation sur Emperor

Nous allons maintenant générer les plot emperor. Utilisez l'outil *qiime emperor plot* avec ces paramètres:

```
--i-pcoa core-metrics-results/unweighted_unifrac_pcoa_results.qza
--m-metadata-file sample-metadata.tsv
--p-custom-axes DaysSinceExperimentStart
```

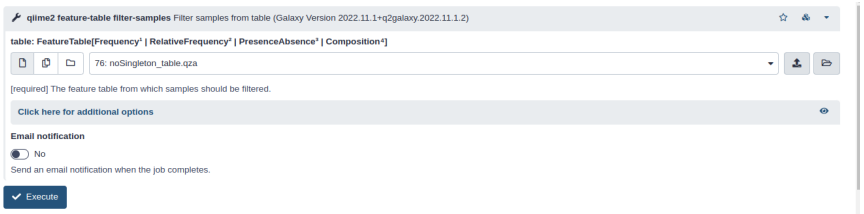


> Retrouve-t-on les mêmes tendances qu'avec les graphiques de beta diversité générés par l'outil core-diversity-phylogenetic ? (Indice : Expérimentez avec des points de coloration par différentes métadonnées.)

Tests d'abondance différentielle avec ANCOM

Nous allons d'abord filtrer nos échantillons. Utilisez l'outil *qiime feature-table filter-samples* avec ces paramètres:

```
--i-table noSingleton_table.qza
--m-metadata-file sample-metadata.tsv
--p-where "BodySite='gut'"
```



Comme pour les ASV, ANCOM ne peut tolérer des fréquences égales à zéro.

Utilisez l'outil *qiime composition add-pseudocount* avec ces paramètres:

```
--i-table filtered-gut-table.qza
```

Nous pouvons ensuite exécuter ANCOM sur la colonne *Subject* pour déterminer quelles taxons diffèrent en abondance dans les échantillons de *gut* des deux sujets.

Nous allons maintenant executer le test ANCOM. Utilisez l'outil *qiime composition ancom* avec ces paramètres:

```
--i-table comp-gut-table.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column Subject
```

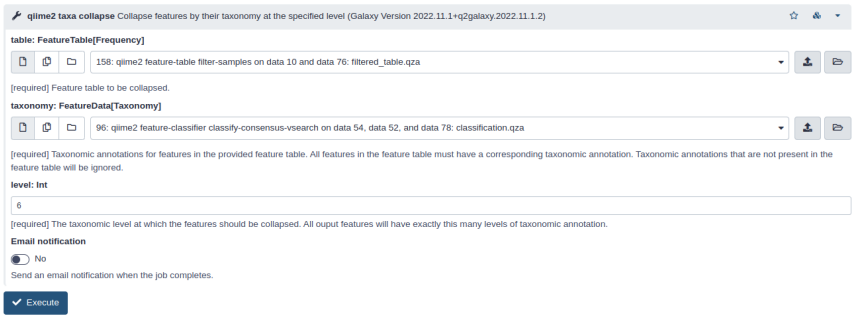
> Quelles ASVs diffèrent dans l'abondance selon le sujet ? Dans quel sujet chaque variante de séquence est-elle plus abondante ?
> Quelles sont les taxonomies de certaines de ces ASVs ? (Pour répondre à la dernière question, vous devrez vous référer à une autre visualisation qui a été générée dans ce tutoriel.).

> Retrouve-t-on les mêmes effets qu'avec les ASV?

Nous sommes aussi souvent intéressés à effectuer un test d'abondance différentielle à un niveau taxonomique spécifique. Pour ce faire, nous pouvons réduire notre FeatureTable[Fréquence] au niveau taxonomique d'intérêt, puis réexécuter les étapes ci-dessus. Dans ce tutoriel, nous réduisons notre table au niveau du genre (c'est-à-dire au niveau 6 de la taxonomie de Greengenes).

Nous allons maintenant executer le test ANCOM au niveau taxonomique 6. Utilisez l'outil *qiime taxa collapse* avec ces paramètres:

```
--i-table filtered-gut-table.qza
--i-taxonomy otu_taxonomy.qza
--p-level 6
```



Puis l'outil *qiime composition add-pseudocount* avec ces paramètres:

```
--i-table gut-table-l6.qza
```

Enfin l'outil *qiime composition ancom* avec ces paramètres:

```
--i-table comp-gut-table-l6.qza
--m-metadata-file sample-metadata.tsv
--m-metadata-column Subject
```

> Quels genres diffèrent en abondance en fonction du sujet ? Dans quel sujet chaque genre est-il le plus abondant ? Retrouve-t-on les mêmes résultats en ASV?

Liens utiles

Dans ce TP nous avons traité des données issues Illumina en single-end. Nous avons utilisé un nombre limité d'approches (DADA2 par exemple) mais de nombreux algorithmes sont disponibles au sein des différents pipelines. Quelques changements doivent être apportés pour traiter des reads paired-end et de nombreuses alternatives sont disponibles pour les différentes étapes. De nombreux tutoriels sont disponibles sur le site de QIIME2, ainsi qu'un forum et de la documentation.

Page réalisée par les membres de la plateforme PEGASE, mise à jour en avril 2023.