Analyse du génome à partir de la ligne de commande

Dans ce document, nous avons fourni des instructions pour une première analyse, y compris la construction d'un arbre phylogénétique, des séquences de agents pathogènes bactériens. Toutes ces analyses, et bien d'autres, nécessitent de générer et d'assembler des génomes consensus à partir de données de séquençage brutes.

Notes importantes pour suivre ce tutoriel :

- Le texte sur fond gris en `police monospace` représente les commandes à saisir. En général, les commandes occupent une ligne, cependant, dans ce document, les commandes peuvent visuellement s'étendre jusqu'à la ligne suivante. Nous ajouterons une ligne vide entre les commandes pour indiquer la présence de plusieurs commandes.
- Le texte en gras entouré de **< >** est quelque chose que vous devrez remplacer par votre propre nom de dossier, de chemin d'accès ou d'échantillon.
- Ce tutoriel suppose que vous avez configuré le répertoire de pipeline bioinformatique sur votre ordinateur et que vous avez assemblé des séquences consensus au format FASTA. Si ce n'est **pas** le cas et que vous disposez de données de séquençage brutes non assemblées, suivez les instructions de Configuration de la ligne de commande et d'Assemblage du génome sur la ligne de commande avant de continuer.
- L'analyse phylogénétique nécessite un ensemble de génomes de contrôle auxquels comparer vos séquences. Les génomes de contrôle peuvent vous aider à identifier les lignées auxquelles vos séquences appartiennent et à déterminer si les séquences nouvellement générées sont similaires à celles qui ont été publiées précédemment. Il convient d'accorder une attention particulière à l'inclusion de génomes présentant une diversité temporelle, géographique et génétique appropriée.

Ce tutoriel vous guidera dans une première analyse des séquences génomiques des agents pathogènes. Le processus comprend deux étapes principales :

1. Typer les séquences nouvellement générées.
2. Construire un arbre phylogénétique des séquences d'agents pathogènes nouvellement générées et publiquement disponibles.

Ce tutoriel s'appuie sur le pipeline bioinformatique BACPAGE disponible à l'adresse https://github.com/CholGen/bacpage. Les instructions pour la configuration de ce pipeline et un répertoire de projet peuvent être trouvées dans Command Line Setup et Genome Assembly on the Command Line. Toutes les commandes de ce tutoriel doivent être exécutées à partir du répertoire du pipeline bioinformatique (généralement ~/bacpage) et font référence au fichier de configuration YAML édité dans Genome Assembly on the Command Line (généralement situé dans **<project-path>**/`config.yaml`).

ÉTAPE 0 : Configuration du répertoire de projets

Cette analyse suppose que vous avez terminé l'assemblage du génome en ligne de commande et que vous avez créé un répertoire de projets dans le répertoire du pipeline. En outre, vous devez avoir généré des séquences consensus à partir des reads de séquençage bruts pour chacun de vos échantillons.

1. Naviguez jusqu'à votre répertoire de projets :

```
cd ~/bacpage/<project-path>
```

2. Confirmez qu'il y a des fichiers FASTQ démultiplexés pour chaque échantillon (c'est-à-dire deux fichiers FASTQ par échantillon) dans votre répertoire de projets en visualisant le contenu du répertoire `input/` de votre répertoire de projets.

```
ls input
```

3. Confirmez la présence de séquences consensus pour chaque échantillon (c'est-à-dire un fichier FASTA par échantillon) dans le répertoire de votre projets en consultant le contenu du répertoire `results/consensus_sequences/` :

```
ls results/consensus_sequences/
```

ÉTAPE 1 : Typage bactérien

La première étape de nombreux protocoles d'analyse consiste à caractériser largement les isolats à l'aide de leurs informations moléculaires. Pour ce faire, nous utiliserons le typage de séquences multilocus (MLST), qui détecte les variants alléliques de sept gènes domestiques omniprésents, et le sérotypage, qui détecte la présence ou l'absence de six gènes de facteurs de virulence.

Note : Par défaut, le pipeline BACPAGE est configuré pour analyser les génomes de *Vibrio cholerae*. Si vous étudiez un autre agent pathogène, vous pouvez modifier les paramètres « *mlst_profiling/scheme* » dans <p**roject-path**>/`config.yaml` en faveur de n'importe quel schéma de typage publié par PubMLST (exécutez `mlst -list` pour afficher les schémas de typage disponibles).

En outre, vous pouvez fournir votre propre répertoire de gènes facteurs de virulence au format FASTA plutôt que les gènes spécifiques à *Vibrio cholerae* fournis par le pipeline. Pour ce faire, changez la valeur de « *reference_genes* » dans <**project-path**>/`config.yaml` pour le répertoire contenant les fichiers FASTA pour ces gènes.

1. Naviguez jusqu'au répertoire du pipeline bioinformatique :

```
cd ~/bacpage
```

2. Pour réaliser le MLST, exécutez la commande suivante :

```
snakemake --configfile <project-path>/config.yaml --cores
<number-of-processors> --keep-going --until mlst_profiling
```

Cette commande génère un fichier unique **<project-path>**/`results/reports/mlst_types.csv`. Ce fichier est un CSV contenant une ligne pour chaque échantillon, décrivant les variants alléliques qu'il possède pour chacun des gènes domestiques (*adk, gyrB, mdh, metE, pntA, purM, pyrC*). À l'aide de ces allèles, MLST assigne un type de séquence à chaque échantillon.

3. Les isolats peuvent également être classés selon la présence ou l'absence de cinq gènes de facteurs de virulence. Exécutez la commande suivante pour détecter ces gènes pour chacun de vos échantillons :

```
snakemake --configfile <project-path>/config.yaml --cores
<number-of-processors> --keep-going --until
virulence_factor_profiling
```

Cette commande générera un fichier unique **<project-path>**/`results/reports/typing_information.csv`, qui ressemblera à l'exemple ci-dessous pour trois génomes du choléra. Chaque ligne indique la fraction du gène (indiquée par la colonne) couverte par les reads d'un échantillon individuel.

sample	ctxA	tcpA_classical	tcpA_eltor	toxR	wbeO1	wbfO139
échantillon1	1.00	0.32	1.00	0.98	1.00	0.22
échantillon2	1.00	0.31	1.00	0.98	1.00	0.22
échantillon3	1.00	0.33	1.00	0.98	1.00	0.22

Les gènes trouvés dans un échantillon auront une fraction couverte proche de 1,00, ce qui indique que les reads mappent l'intégralité du gène. Les gènes non trouvés dans cet échantillon auront une fraction couverte inférieure. Dans l'exemple ci-dessus, nous pouvons dire que le premier échantillon contient les gènes *ctxA, wbeO1* et *tcpA El Tor*, mais pas les gènes *wbfO139* ou *tcpA classical*. Nous pouvons donc affirmer qu'il s'agit d'un *Vibrio cholerae* O1 et qu'il appartient probablement au biotype El Tor.

ÉTAPE 2 : Alignement de séquences multiples

Nous allons maintenant générer une phylogénie incluant vos génomes nouvellement générés. Bien qu'il soit parfois utile de réaliser un arbre phylogénétique en utilisant uniquement les séquences nouvellement générées, il est généralement plus utile de combiner les séquences nouvellement générées avec un ensemble de séquences précédemment publiées, appelé « ensemble de données de contrôle ».

L'analyse phylogénétique nécessite un alignement de séquences multiples en entrée. Étant donné que toutes vos séquences nouvellement générées et l'ensemble de données de base ont été assemblées en s'alignant sur une référence commune, nous pouvons facilement générer un alignement en concaténant leurs fichiers individuels. Si toutes vos séquences (y compris les séquences publiées antérieurement que vous incluez dans votre ensemble de

données de contrôle) n'ont pas été assemblées en s'alignant sur le *même* génome de référence, vous devrez utiliser un logiciel d'alignement à forte intensité de calcul pour aligner les séquences les unes par rapport aux autres.

Note : Nous recommandons que seules les séquences qui couvrent au moins 90% du génome soient incluses dans l'inférence phylogénétique. Par défaut, le pipeline n'inclura les séquences dans l'alignement que si elles atteignent ce seuil de couverture. Si vous souhaitez une rigueur différente, modifiez la valeur *« tree_building/required_coverage »* dans <**project-path**>/`config.yaml` en lui donnant la valeur souhaitée. En outre, vous pouvez envisager d'utiliser les résultats du typage pour déterminer les séquences à inclure (par exemple, n'inclure que les séquences classées comme un certain sérotype).

1. Localisez l'ensemble de données de contrôle que vous souhaitez comparer à vos séquences nouvellement générées. Cet ensemble de données doit être un fichier FASTA unique contenant des entrées distinctes pour chaque séquence de votre ensemble de données de contrôle, également appelé « multi-FASTA ». Nous recommandons de placer le fichier multi-FASTA dans les `ressources/` répertoire du pipeline bioinformatique (typiquement `~/bacpage/resources`). Vous pouvez le faire en utilisant la ligne de commande ou en déplaçant simplement le fichier en utilisant le navigateur de fichiers sur votre ordinateur.

2. Une fois que vous avez placé votre ensemble de données de contrôle FASTA dans le répertoire de ressources, déterminez le chemin absolu de l'ensemble de données de contrôle. S'il a été placé à l'emplacement recommandé, vous pouvez trouver le chemin absolu en naviguant jusqu'au répertoire de ressources et en vérifiant la sortie de `pwd` :

   ```
   cd ~/bacpage/resources

   pwd
   ```

   Le chemin absolu de l'ensemble de données de contrôle sera la sortie de `pwd` plus le nom du fichier de l'ensemble de données de contrôle (se terminant par .fasta).

3. Retournez dans le répertoire du pipeline bioinformatique :

   ```
   cd ~/bacpage
   ```

4. Ajoutez le chemin d'accès absolu de votre ensemble de données de contrôle à <**project-path**>/`config.yaml`. Ouvrez le fichier de configuration dans un éditeur de texte, modifiez la valeur de <**background-dataset-path**> à la ligne 8 (voir ci-dessous), remplacez la valeur de *« generate/phylogeny »* à la ligne 15 par « True » et enregistrez le fichier.

   ```
   background_dataset: "<background-dataset-path>"
   ```

5. Générer un alignement de séquences multiples en exécutant la commande suivante :

   ```
   snakemake --configfile <project-path>/config.yaml --cores
   ```

```
<number-of-processors> --keep-going --until concatenate_sequences
```

Cette étape concatène simplement vos génomes nouvellement générés, et l'ensemble de données de contrôle s'il est présent, dans un multi-FASTA situé à : **<project-path>**/intermediates/illumina/phylogeny/complete_alignment.fasta

ÉTAPE 3 : Construire un arbre phylogénétique

Après avoir généré l'alignement des séquences multiples, nous pouvons procéder à l'inférence d'un arbre phylogénétique.

1. Exécutez l'étape d'inférence phylogénétique du pipeline en lançant la commande suivante :

```
snakemake --configfile <project-path>/config.yaml --cores
<number-of-processors> --keep-going --until generate_rooted_tree
```

En bref, cette étape génère une phylogénie en utilisant iqtree. Par défaut, cette étape utilise le modèle de substitution GTR et calcule le support des branches en effectuant 1000 bootstraps. Ces options peuvent être changées en modifiant la valeur de *« tree_building/iqtree_parameters »* dans **<project-path>**/config.yaml. Ce processus est intensif en termes de calcul et peut prendre quelques heures en fonction de la taille et du nombre de séquences analysées ainsi que de la vitesse de votre ordinateur.

La sortie de cette commande est une phylogénie au format Newick, <**project-path**>/results/<**project-directory-name**>.ml.tree.

2. Bien que le fichier d'arbre soit un fichier texte qui peut être ouvert et lu dans un éditeur de texte, il n'est généralement pas interprétable dans ce format. Nous utiliserons un visualiseur d'arbre GUI appelé FigTree pour visualiser les fichiers d'arbres. Dans le répertoire des applications de votre ordinateur, ouvrez FigTree. Cliquez sur Fichier -> Ouvrez, et sélectionnez la phylogénie nouvellement générée dans le navigateur de fichiers qui s'ouvre. Vous pouvez également ouvrir le fichier **<project-directory-name>**.ml.tree directement à partir du navigateur de fichiers.

3. La phylogénie devrait maintenant apparaître dans la fenêtre principale de FigTree. Recherchez vos échantillons dans l'arbre et déterminez les échantillons dont ils sont les plus proches.