Analyse bioinformatique de séquences - Mise en autonomie

1. Analyse de données NGS, assemblage et mapping

Les données peuvent être récupérées à partir de
Shared Data / Data Libraries / Formation NGS / Mise en autonomie /

Vous allez partir de données déjà nettoyées issues de séquençage RNASeq.
Nous avons extrait une zone du génome du bananier et les gènes présents sur cette zone.

Récupérer les reads provenant de Pahang et de PahangHD et mappés les sur les gènes de la région. Pour celà vous avez besoin de savoir que le séquençage a été fait en single-end.

2. Détection de polymorphismes

Les données peuvent être récupérées à partir de
Shared Data / Data Libraries / Formation / Mise en autonomie 2013/Mapping

A partir des reads RNASeq mappés sur les séquences codantes (CDS) d'une région de 2 Mb du génome du Bananier, recherchez des SNP entre les 2 génotypes Pahang et PahangHD.
PahangHD étant un haploïde doublé, nous nous attendons à ne pas trouver de variation en intra. Pouvez-vous le vérifier?
Par ailleurs, y a-t-il des positions hétérozygotes au sein du génotype Pahang?
A l'aide de l'outil SNiPlay, étudiez s'il existe des SNP non synonymes sur cette région (Pour rappel, le mapping est basé sur les CDS).
Quel est le gène de cette région présentant la diversité nucléotidique la plus élevée?

3. Annotation

Nous avons vu l'annotation d'éléments transposables et de gènes sur une région du chromosome 1 du génome du riz. Dans l'exercice ici proposé, l'analyse sera faite sur un génome différent, avec l'exemple du BAC MaC088K20 du Génome de la Banane.

En pratique, connectez-vous
* sur Galaxy pour récupérer les fichiers d'aide à l'annotation. Celles-ci sont stockées dans les *Shared Data / Data Libraries / Formation / Mise en autonomie 2013/Annotation.*
* sur la *sandbox2* du site du projet GNPAnnot pour configurer votre bac à sable.

- Aller sur le site web de GNPAnnot.
http://www.gnpannot.org/
Puis dans ressources cliquer sur Sandbox 2.
http://www.gnpannot.org/content/gnpannot-sandbox-form-2-without-access-restriction
Générer la sandbox (**ne cliquez qu'une fois**).
- Launch your GBrowse!
Cliquez sur l'exemple : MaC088K20.

Aller dans l'onglet « Select Tracks » pour choisir les pistes d'analyse à afficher

Revenez sur l'onglet « Browser » pour visualiser les prédictions affichées.

Cliquer sur un gène de la piste « Protein Coding Gene Model » et sur le lien « Edit with Artemis » du menu de la fenêtre Popup ou pour avoir toute la région la sélectionne au préalable et cliquer sur « Launch Artemis »

Ouvrir avec Java Web Start (défaut) -> OK

« MaC088K20_1..151999.jnlp est une application provenant d'un téléchargement depuis Internet. Voulez vous vraiment l'ouvrir » -> Ouvrir

Voulez vous exécuter l'application -> cocher « J'accepte le risque et je souhaite exécuter l'application » puis Exécuter

« Set Working Directory » -> OK

« Enter Database Address » -> remplissez l'identifiant « Password » puis cliquer sur OK

Artemis connecting -> Sequence loaded

- Le guide pratique d'Artemis connecté à Chado se trouve à l'adresse :
https://www.sanger.ac.uk/resources/software/artemis/#chado
ftp://ftp.sanger.ac.uk/pub/resources/software/artemis/database/chado.practical.guide.pdf

Une fois Artemis lancée, chargez les fichiers d'aide à l'annotation comme présenté lors des TD de mercredi, puis annotez la séquence comme montré mercredi.

4. Modélisation moléculaire

Ce module permet de s'intéresser aux protéines, en tant que résultats de l'expression des gènes et porteurs de fonctions biologiques. Il permet de se familiariser avec les bases de données et les outils de la modélisation moléculaire.
En partant de deux versions d'un gène (qui diffèrent par quelques SNPs), nous rechercherons la ou les protéines correspondantes et tenterons d'en créer un modèle tridimensionnel puis de faire le lien entre la position des SNPs et leur implication éventuelle dans une différence de fonction entre ces deux versions du gène.

5. Phylogénie moléculaire

L'objectif de cet exercice est de créer un workflow complet d'analyse, depuis un multifasta protéique, jusqu'à un arbre phylogénétique raciné et annoté.

Commencer par télécharger le fichier de séquences protéiques:
http://gohelle.cirad.fr/phylogeny/formation/zinc_finger.fasta
et le charger dans Galaxy.
Créer un nouveau workflow, l'éditer:Ajouter l'ensemble des étapes nécessaires:

1. **la donnée d'entrée multifasta,**
2. **l'alignement par MAFFT,**
3. **le nettoyage par GBlocks,**

4. **la conversion de format de Fasta à Phylip,**
5. **la phylogénie par PhyML,**
6. **la réconciliation par RAP-Green, et la donnée d'entrée "arbre des espèces".**

Sélectionnez l'arbre des gènes de RAP-Green comme sortie, et éventuellement d'autres à votre convenance.

Enfin, exécuter le workflow sur le fichier "zinc_finger.fasta".

Notes techniques:

La configuration de MAFFT a besoin d'être enrichie du choix d'une matrice de similarité entre acides aminés, pour des données protéiques.

PhyML sera configuré avec une loi LG+Gamma, dans cet exemple.