Travaux pratiques Formation

« transcriptomique & analyse du polymorphisme »

Les pipelines sont accessibles depuis l'adresse :

https://bbric-pipelines.toulouse.inra.fr/galaxy/

Transcriptomique

TP Mesure de l'expression

1- Récupérer les données dans Galaxy

On souhaite étudier l'expression des gènes de la bactérie *S bbric*.

Sur BBRIC::Reference

Récupérez La séquence génomique et son annotation :

* S. bbric genomic sequence
* S. bbric structural annotation

Sur BBRIC::Archive

Récupérez les Librairies RNAseq (au choix) :

Librairie Sb-2404-33-R1 (single end orienté) (50nt)

*S.bbric-RbmLong-GGK33.fastq.gz

Ou

Librairie Sb-2404-21-R1 (paired end orienté, le read 1 qui donne l'orientation)

*S.bbric-RbmLong-GGK21.ope.1.fastq.gz
*S.bbric-RbmLong-GGK21.ope.2.fastq.gz

Ou

Librairie Sb-2404-36-R1 (paired end orienté, le read 1 donne l'orientation, enrichie en petits ARN)

* S.bbric-RbmSmall-GGK36.ope.1.fastq.gz
* S.bbric-RbmSmall-GGK36.ope.2.fastq.gz

2- Lancement de la mesure de l'expression

Accéder à l'outil de mesure de l'expression, soit à partir de bbric.toulouse.inra.fr, soit en navigant dans Galaxy (tools BBRIC Protocols)

Régler les paramètres en fonction des librairies que vous souhaitez analyser :

Maximal distance between paired reads (si pair) = 70nt pour les Small ; 130nt pour les Long

☐ Faire tourner le pipeline avec 2 valeurs différentes pour le nombre de mismatches autorisé (0 et 2)

3- Analyse des fichiers de résultats :

Fichier de stats : Quelle différence constatez-vous entre 0 et 2 mismatchs ?

Fichier de count : Quelle différence constatez-vous entre 0 et 2 mismatches ? Est-ce ce à quoi vous vous attendiez ?

Contrôles Qualité sur Fichiers de sortie « Mesure de l'expression »

1) Allez dans Shared Data → Data Libraries

2) Cliquez sur BBRIC Protocols

3) Sélectionnez “QC RNASeq”

4) Sélectionnez “Import to current history” et cliquez sur GO : les fichiers d'entrée pour tester ce pipeline sont maintenant dans votre historique

5) Cliquez sur “Analyse Data”

6) Dans le panel de gauche, cliquez sur “QC Tools” puis sur “RNAseq libraries”

7) Remplissez le formulaire avec les fichiers importés précédemment (avec le fichier de comptage au format BBRIC) et cliquez sur Execute

8) Au bout de quelques minutes, deux fichiers résultats s'afficheront : QC_RNASeq_stats et QC_RNASeq_count. Parcourez les fichiers pdf générés et reportez vous au document de formation pour l'interprétation des différents graphiques

Analyse du polymorphisme

Détection des SNPs (samtools/VarScan) pour analyse de leurs effets

1 Récupération du jeu de test dans share data

Shared Data ☮ Data Librairies ☮ BBRIC protocols > cocher 'X.bbric' (sélectionne toutes les données) ☮ 'Import to current history' Go

Xbbric_A = séquences de l'échantillon A (2 fastq, 100nt)

Xbbric_B = séquences de l'échantillon B (2 fastq, 100nt)

Xbbric_C = séquences de l'échantillon C (2 fastq, 76nt)

Xbbric genomic sequence = séquence du génome (fasta)

Xbbric structural annotation = annotation du génome (gff)

2 Charger le workflow

BBRIC protocols ☮ POLYMORPHIMS > SNP detection and effect prediction

3 Paramétrer l'analyse

- Cocher 'Paired-end library'
- 'Maximum distance between paired reads (nt)' : 300
- Cliquer 2 fois sur 'Add new Paired-end library'
- Sélectionner les reads des 3 individus et mettre A, B, et C dans les 'Sample name'
- Sélectionner 'Xbbric genomic sequence' pour 'Reference genome file (fasta)'
- Sélectionner 'Xbbric structural annotation' pour 'Genome annotation file (GFF3)'
- 'Minimum hit length' :50 (car échantillon C a une longueur de reads de 76nt)
- 'Maximum number of mismatches': 4
- 'Minimum position coverage': 20
- 'Minimum variant coverage': 10
- 'Minimum position frequency': 0.2
- 'Minimum homozygous frequency': 0,75
- 'Effect results filter': Select All
- Cocher 'Upper cas nucleotides'

4 Analyse des résultats

Nous obtenons trois fichiers de résultats :

SNP effect by gene : Nombre de chaque catégorie d'effets trouvés pour chaque gène de l'annotation

SNP effect report : Rapport HTML produit par SnpEff avec des statistiques sur le type et le nombre d'effets prédits.

SNP matrix file : Fichier VCF zippé

TP pour la « Détection des SNPs et calcul des fréquences alléliques pour les positions bi-alléliques »

Plan :

##

I. 1ere Partie (optionnelle): Chargement des banques dans l'historique courant

1- Importer les banques de données à assembler dans son historique courant, à partir du menu Shared data/ Data Libraries/BBRIC protocols
2- Sélectionner le dossier X.bbric pour importer les datasets Xanthomanas bbric
   a. les 3 banques Paired-End :
      i. Xbbric_A
      ii. Xbbric_B
      iii. Xbbric_C
   b. Xbbric genomic sequence : le génome de référence au format fasta
   c. Xbbric structural annotation : le gff3 contenant l'annotation structurale
3- Cliquez sur GO pour importer ces jeux de données dans votre historique courant.
4- Après confirmation de l'import, vous pouvez retourner à l'accueil pour analyser les données : à partir du menu Analyze Data

Cette partie est optionnelle, car elle aura sans doute était vu et réalisé dans le TP pour le pipeline « Détection des SNPs et de leurs effets » qui utilisent les mêmes jeux de données.

II. 2eme Partie: Exécution du pipeline

1- Chargement du pipeline

1. Aller à l'accueil pour analyser les données : à partir du menu Analyze Data
2. Dans le menu Tools de gauche, cliquez sur « BBRIC protocols » pour dérouler la liste des outils et pipelines
3. Allez à la section « POLYMORPHISM », et cliquez sur le pipeline « SNP detection for pooled/one allele frequency matrix »

2- Chargement des données dans le pipeline

1. Cliquez sur Paired-end library, pour renseigner la banque paired-end à charger
2. Renseigner le champ « **Maximum distance between paired reads (nt):** » avec pour valeur 500 ou 600 au choix (ici on a des banques paired-end 2x100 ou 2x75 pb avec des tailles d'inserts estimés autour

de 300 ; cet intervalle de valeurs tient compte de la variabilité/divergence entre la souche à mapper et celle de référence ; en principe, elle devrait être estimé après avoir mappé les données une première fois)

3. Dans le menu déroulant « **Read file 1:** », sélectionner le fichier paired-end R1, par exemple « **Xbbric_A.pe.1.250k.fastq.gz** » pour **la souche A**, identifié avec **pe.1**
4. Dans le menu déroulant « **Read file 2**», sélectionner le fichier paired-end R2, par exemple « **Xbbric_A.pe.2.250k.fastq.gz** » pour la **souche A**, identifié avec **pe.2**
5. Renseigner le nom de l'échantillon à associer à la banque paired-end : par exemple, Xbbric_A pour la souche A.
6. Pour ajouter une nouvelle banque, cliquez sur « Add new Paired-end library »
7. Puis rejouer les étapes 3. à 5. pour ajouter les banques associées aux **souches B et C**.

3- Paramétrage du mapping et du calling de variant SNP

1. Dans le menu déroulant « **Reference genome file (fasta) :** », sélectionner le génome de référence « **Xbbric genomic sequence** »
2. Paramétrer les options du mapping de Glint :
   a. Dans le champ « **Minimum hit length:** », laisser la valeur à **50** ou bien renseigner une valeur convenable, supérieur pour être stringent
   b. Dans le champ « **Maximum number of mismatches :** », laisser la valeur à **4** ou bien renseigner une valeur convenable tenant compte de la divergence entre la souche à mapper et la souche de référence, diminuer cette valeur pour être stringent
3. Paramétrer les options de calling de variant de VarScan :
   a. Dans le champ « **Minimum position coverage :** », laisser la valeur à **10** ou bien renseigner une valeur convenable, si supérieure il faut que la profondeur de séquençage soit suffisante
   b. Dans le champ « **Minimum variant coverage :** », laisser la valeur à **2** ou bien renseigner une valeur convenable, si supérieure il faut que la profondeur de séquençage soit suffisante
   c. Dans le champ « **Minimum variant frequency:** », laisser la valeur à **0.01** ou bien renseigner une valeur convenable, si inférieure il faut que la profondeur de séquençage soit suffisante, pour être sur de capter un variant très peu fréquent

Enfin cliquez « **Execute** » pour lancer le pipeline.

III. Les fichiers de sorties

A. SNP matrix file (VCF)

Allelic count