Travaux pratiques Formation

« annotation fonctionnelle & analyse du polymorphisme »

Vendredi 16 Octobre 2015

LIPM

Les pipelines sont accessibles depuis l'adresse :

https://bbric.toulouse.inra.fr

Analyse du polymorphisme

Genome browser

https://frama.link/Xbbric

Détection des SNPs à partir de données de re-séquençage (GATK)

1/Charger le jeu de données dans galaxy

- Utiliser Get Data pour charger les fichiers suivants :
  - GL349685 (fichier fasta génomique)
  - GL349685_JF191_1.fastq (banque de séquence PE Illumina)
  - GL349685_JF191_2.fastq (banque de séquence PE Illumina)
  - GL349685_JF975_1.fastq (banque de séquence PE Illumina)
  - GL349685_JF975_2.fastq (banque de séquence PE Illumina)
  - GL349685.gff3 (fichier d'annotation format gff 3)

2/Choisir dans BBRIC Protocole "Variant detection and allele frequency GATK based"

Charger le fichier fasta et les deux librairies paired-ends

3/Analyse du résultat

Regarder le fichier .vcf
Caractéristique du SNP à la position :
45 404
41 048
23798

4/lancement des SNPeff

Charger le fichier .vcf résultats de la précédente étape.
Le fichier d'annotation GL349685.gff3
Lancer l'application sans option.

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

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)

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'

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 building an allelic 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)

B. Allelic count

Mesure de l'expression

TP Mesure de l'expression

1- Récupérer les données : séquence génomique, annotations et librairies RNAseq

Allez dans Shared Data → Data Libraries

Cliquez sur BBRIC Protocols

Dans la partie Genome Annotation/Bacteria, sélectionnez les données souhaitées :

-- Reference genome file **:**
* S. bbric genome

-- Genome annotation file :
* Result S. bbric annotation

--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, c'est 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 petit ARN)
* S.bbric-RbmSmall-GGK36.ope.1.fastq.gz
* S.bbric-RbmSmall-GGK36.ope.2.fastq.gz

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

2- Lancement de la mesure de l'expression

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

Paramètres par défaut

Nb of mismatches =_faire tourner le pipeline avec 2 valeurs différentes de mismatch (0 et 2)

3- Analyse des fichiers de résultats :

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

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 “Analyze 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) 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 de protéomes

TP OrthoMCL-Companion

Analyse comparative de protéomes d'espèces apparentées Outil

1. Recherchez le pipeline sur le portail BBRIC :
https://bbric.toulouse.inra.fr

2. Téléchargez le jeu de données test :
3 proteomes of about 400 proteins each annotated with InterPro 51.0

..			Dossier de fichiers
Bacsu.faa	148 286	85 717	Fichier FAA
Bacsu.iprscan	832 267	99 749	Fichier IPRSCAN
Bacsu.txt	22	22	Document texte
Ecoli.faa	147 958	78 982	Fichier FAA
Ecoli.iprscan	925 147	114 354	Fichier IPRSCAN
Ecoli.txt	18	18	Document texte
Xbbric.faa	169 909	84 557	Fichier FAA
Xbbric.iprscan	615 987	75 754	Fichier IPRSCAN
Xbbric.txt	19	19	Document texte

3. Connectez-vous via le bouton login, et préparez-vous à lancer une nouvelle analyse.

4. Lancez les 3 analyses suivantes :
Analyse1 = charger les 3 fichiers fasta avec leur fichier d'annotation, laisser les paramètres orthomcl par défaut et ne mettez aucune espèce en référence ; lancer
Analyse2 = à présent mettez « Ecoli » en espèce de référence ; lancer
Analyse3 = ici mettez « Ecoli et Bacsu » en espèces de référence ; lancer

5. Allez à l'onglet « List of analyses » pour visualiser/ récupérer vos résultats.

6. Naviguez dans vos résultats, et trouvez :

a. De combien de protéines sont constitués les protéomes ?
R :

b. Combien de in-paralogues sont identifiés dans l'espèce Xbbric ?
R :

c. Combien de groupes orthomcl sont partagés par ces 3 espèces ?
R :

d. Donner les 4 groupes OrthoMCL partagés par Ecoli et Bacsu seulement ?
R :

e. Chercher les espèces qui se retouvent dans le groupe ORTHOMCL0 ? ORTHOMCL168 ?
R :

7. Ouvrez le fichier orthomcl.out et cherchez le groupe ORTHOMCL1.
Que pouvons nous déjà dire sur ce groupe ?

ORTHOMCL1(14 genes,3 taxa): BG10168(Bacsu) BG10169(Bacsu) BG10170(Bacsu) BG10929(Bacsu) BG10970(Bacsu) BG10971(Bacsu) BG10972(Bacsu) BG11243(Bacsu) BG11961(Bacsu) BG11962(Bacsu) BG12652(Bacsu) b0586(Ecoli) b0586__ENTF-MONOMER(Ecoli) entF(Xbbric)

R :

8. Chercher dans le fichier fasta PanProteome.fasta, et ce dans les 3 analyses, le gène qui a été conservé pour représenter ce groupe.
Que remarquez-vous ? Est ce le résultat attendu ?

Analyse1 : La référence est =>
ORTHOMCL1(14 genes,3 taxa): BG10168(Bacsu) BG10169(Bacsu) BG10170(Bacsu) BG10929(Bacsu) BG10970(Bacsu) BG10971(Bacsu) BG10972(Bacsu) BG11243(Bacsu) BG11961(Bacsu) BG11962(Bacsu) BG12652(Bacsu) b0586(Ecoli) b0586__ENTF-MONOMER(Ecoli) entF(Xbbric)

Analyse2 : La référence est =>
ORTHOMCL1(14 genes,3 taxa): BG10168(Bacsu) BG10169(Bacsu) BG10170(Bacsu) BG10929(Bacsu) BG10970(Bacsu) BG10971(Bacsu) BG10972(Bacsu) BG11243(Bacsu) BG11961(Bacsu) BG11962(Bacsu) BG12652(Bacsu) b0586(Ecoli) b0586__ENTF-MONOMER(Ecoli) entF(Xbbric)

Analyse3 : La référence est =>
ORTHOMCL1(14 genes,3 taxa): BG10168(Bacsu) BG10169(Bacsu) BG10170(Bacsu) BG10929(Bacsu) BG10970(Bacsu) BG10971(Bacsu) BG10972(Bacsu) BG11243(Bacsu) BG11961(Bacsu) BG11962(Bacsu) BG12652(Bacsu) b0586(Ecoli) b0586__ENTF-MONOMER(Ecoli) entF(Xbbric)

R :