Recherche et analyse de polymorphismes SNP

1- Tablet : Détection visuelle de SNP avec Tablet

Tablet est un outil graphique de visualisation d'assemblage et d'alignement de séquences issues de NGS (Next Generation Sequencing).

Lancer le logiciel Tablet via le Java Web Start : http://bioinf.hutton.ac.uk/tablet/faq.shtml.
Charger le fichier SAM issu du mapping par BWA, fichier rassemblant l'ensemble des individus.
(Depuis Galaxy: Data libraries => Formation => SNP => V2 => MergedSAM_sorted)

Sélectionner le gène Os01g62920. Sur ce gène :

- Pouvez-vous repérer des SNP ?
- Repérez un SNP où tous les individus séquencés différent de la référence
- Repérez un SNP où seul RC5 diffère de la référence
- Repérez une position hétérozygote
- Intuitivement, comment estimeriez-vous la fiabilité d'un SNP ? Comment feriez-vous la différence entre une position hétérozygote et une erreur de séquençage ?
- Quels peuvent-être les problèmes rencontrés du fait de la présence de pseudogènes ou gènes répétés ?

Notes techniques :
Tablet propose de nombreuses options pour la visualisation des reads assemblés, notamment la colorisation selon les « Read Groups » ou encore la mise en évidence des variants.
Ce logiciel va permettre de confirmer à l'œil des SNP détectés de manière automatique mais ne sera pas utilisé pour une détection des variants à large échelle.

2- Détection de SNP avec GATK via Galaxy :

GATK (Genome Analysis Tool Kit) est une librairie logicielle permettant l'analyse de données NGS. GATK va détecter les SNP et indels et affecter à chacune de ces positions un génotype aux individus. GATK fournit en sortie un fichier au format VCF (Variant Call Format).
Spécification du format VCF

Ouvrir Galaxy.
Convertir le fichier SAM en fichier BAM en utilisant *SAM-to-BAM*.
Créer le workflow GATK suivant :
IndelRealigner =>UnifiedGenotyper

- A quoi correspondent les différents champs du fichier VCF ?
- Observer les SNPs obtenus et vérifier avec Tablet
- Identifiez dans le fichier VCF une position hétérozygote ?
- Vérifier l'effet de *IndelRealigner* sur l'appel de SNP en relançant directement *UnifiedGenotyper*. Observer la différence de SNPs obtenus entre les 2 traitements sur le gène Os12g32240.1

Lancer le module *DepthOfCoverage* pour obtenir la profondeur de séquençage par position par individu.

- Observer le fichier de sortie.

Lancer le module *ReadBackedPhasing* permettant de retrouver les haplotypes.

- Pour RS4 présentant plusieurs positions hétérozygotes consécutives, tentez de reconstituer manuellement les 2 haplotypes du gène Os12g32240.1. La technique de séquençage a-t-elle permis d'identifier directement les haplotypes (avec *ReadBackedPhasing*)? Pour RS3 ?

Notes techniques :

- Note1 : Le module IndelRealigner permet un réalignement local autour des insertions/délétions afin d'éviter des erreurs d'appel de SNP.

- Note2 : Le module UnifiedGenotyper propose 2 options appelées « Stand call conf » et « Stand emit conf » permettant de fixer les bornes de qualité de SNP à partir desquels les SNP doivent être appelés et taggués comme étant « PASS » et « LowQual ».

- Note3 : En pratique, il a été observé que les erreurs d'appels de SNP sont plus fréquentes sur les transversions que sur les transitions.

- Note4 : D'autres technologies générant des reads plus longs (ex : PacBio) permettent théoriquement d'améliorer la résolution des haplotypes.

3- Utilisation du pipeline SNiPlay pour l'exploitation des SNP

SNiPlay est une application Web dédié à la recherche et l'analyse de SNP à partir de données de séquençage.

Notes techniques :

SNiPlay offre la possibilité d'associer des informations externes aux individus. Typiquement, il est possible d'associer l'appartenance aux compartiments cultivés ou sauvages.
Par ailleurs, plusieurs options de SNiPlay sont disponibles :

- *une option consiste à filtrer sur les données manquantes par position et à supprimer les portions de séquences ayant plus d'un certain % de données manquantes*
- *une option consiste à écarter de l'analyse ou recoder en SNP bialléliques les SNP multi-alléliques et les insertions/délétions.*
- *une option consiste à baser l'annotation des SNP (positions génomiques, synonymes/non synonymes…) non plus sur un BLAST des séquences mais directement sur l'annotation du génome si le mapping a été fait sur les CDS ou les mRNA de ce génome.*

Aller sur le pipeline SNiPlay : http://sniplay.cirad.fr.

Cocher le type d'entrée VCF puis charger 3 fichiers :

- le fichier VCF phasé.
- le fichier FASTA de référence.
- le fichier de profondeur de séquençage.

Intégrer les informations externes sur les individus analysés comme donné en exemple ci-dessous

```
Accession,compartiment
reference,cultivated
RC1,cultivated
RC2,cultivated
RC3,cultivated
RC4,cultivated
RC5,cultivated
RC6,cultivated
RC7,cultivated
RC8,cultivated
RC9,cultivated
RC10,cultivated
RS1,wild
RS2,wild
RS3,wild
RS4,wild
RS5,wild
RS6,wild
RS7,wild
RS8,wild
RS9,wild
RS10,wild
```

Choisir le riz comme génome de référence pour positionner les SNP, et indiquez que la référence correspond aux mRNA.
Sélectionner les étapes « Network analysis » et « Distance tree ».
Laisser le reste par défaut, et lancer l'analyse.

3-1- SNP et statistiques

- Observer les SNP obtenus et les statistiques associées.
- Observer les alignements qui ont été reconstruits à partir du VCF et de la référence.
- Quelle information fournie en entrée a permis de définir où la séquence de chaque individu commence ?
- Quel gène possède un INDEL ?

3-2- Partage de SNP entre groupes

- Observer les SNP partagés entre individus cultivés et sauvages.
- Combien de SNP permettent de discriminer entre les compartiments cultivés et sauvages ?

3-3- Design de puces Illumina

- Quel fichier serait-il possible de soumettre à Illumina pour designer des puces SNP (technologie VeraCode) sur l'ensemble des gènes? Que contient ce fichier?
- Que se passe-t-il si une insertion/délétion se trouve juste en amont d'un SNP ?

3-4- Fichiers alléliques

SNiPlay génère des fichiers de génotypage en différent format accepté par plusieurs logiciels d'analyse : STRUCTURE, DARwin, Phase, TASSEL

- Observer les différents formats de fichiers alléliques disponibles.

3-5- Annotation des SNP

SNiPlay permet de mapper les séquences sur un génome de référence pour annoter les SNP.

- Combien de SNPs sont retrouvés dans les exons ? Quelle est la part de SNPs non-synonymes sur ces derniers?
- Quel est le ratio Transition/Transversion ?

3-6- Reconstruction d'haplotypes

SNiPlay offre la possibilité de reconstruire les haplotypes des individus, c'est-à-dire les groupes d'allèles situés sur les mêmes chromosomes homologues.

- Combien y a-t-il d'haplotypes distincts pour le gène Os01g62920 ?

3-7- Réseau d'haplotypes

Haplophyle est un outil permettant de générer des réseaux d'haplotypes.

- Pour le gène Os01g62920.1, existe-t-il des haplotypes spécifiques du compartiment cultivé ? Observer le réseau d'haplotypes pour ce gène.

3-8- Arbre de distance

- Observer l’arbre de distance généré pour l’ensemble des positions polymorphes.