TD : Exploitation des données de polymorphismes SNP à partir de fichier VCF

1- Utilisation de l'application SNiPlay pour explorer les SNPs obtenus par RNASeq

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

Aller sur le pipeline SNiPlay : http://sniplay.southgreen.fr: Pipeline for SNP analysis => Pipeline v3

1-1- SNP et statistiques sur 3 gènes

Charger le fichier VCF obtenu. : « Add files » puis « Start upload »
Choisir le Riz comme génome de référence pour annoter les SNPs, et spécifier l'option qui permet de recalculer les positions génomiques.

L'annotation des SNPs est réalisée par le logiciel SnpEff en se basant sur les positions génomiques des variants et l'annotation du génome. Une option consiste à préalablement repositionner les variants sur les chromosomes si le mapping a été fait sur les CDS ou les mRNA, de telle sorte que SnpEff puisse fonctionner. Cette fonctionnalité utilise l'annotation GFF (positions des gènes et transcrits) associée au génome sélectionné.

Après avoir sélectionné les individus à gauche et les chromosomes à droite, observer les SNP obtenus et les statistiques associées.

Vérifier que les gènes trouvés correspondent bien à ceux attendus.
Pourquoi certains variants se retrouvent dans les UTRs ?
Quelle est la part de SNP non-synonymes parmi ceux situés dans les parties codantes (CDS) des gènes?

Envoyer le fichier obtenu vers une analyse VCFtools qui permet de calculer différentes statistiques générales. Observer les différentes informations. Quel individu présente le taux d'hétérozygotie le plus important ?

1-2- Analyse sur le transcriptome complet (jeu de données Cultivé et Sauvage)

Pour la suite du TD, nous utiliserons un jeu de données correspondant au transcriptome complet obtenus à partir d'individus cultivés (*O.barthii*) et sauvages (*O.glaberrima*) de Riz africains (Nabholz et al, 2014). Ce jeu de données est accessible depuis l'application.
Cliquer sur « SNP Database » => « SNP queries v3 ».
Choisir le Riz (Rice_MSU6) et le projet « RNASeq_Nabholz_et_al_2014 ».

Combien y-a-t-il de SNPs au total découvert le transcriptome complet ?

1-2-1- Structure de population

L'analyse de structure de populations est réalisée ici par le logiciel sNMF.
sNMF va tester la vraisemblance de chaque valeur de K (nombre de populations ancestrales) et renvoyer les pourcentages d'admixture de chaque individu aux populations.

Après avoir écarté l'individu *meridionalis* (outgroup), réaliser une analyse de structure de population par le logiciel sNMF.
Quelle est la meilleure valeur de K (nombre de populations le plus vraisemblable) ?
Observez les valeurs d'admixture pour les différents groupes pour K=3.
Peut-on observer une séparation entre nos 2 groupes ?

Un arbre de distance basé sur les SNPs a été généré par fastME. Pouvez-vous voir une séparation des 2 groupes Qu'observez-vous quant aux longueurs de branches ? Confirmez cette hypothèse en affichant le plot MDS des individus pour K=3.

1-2-2- Comparaison cultivé/sauvage

Il est possible d'associer des informations externes aux individus. Typiquement, il est possible d'associer l'appartenance aux compartiments cultivés ou sauvages, pour être prise en compte dans les analyses.

Ecarter *sativa* et *meridionalis* pour cette partie de l'analyse.
En utilisant l'option « Assign individuals to groups/populations », définir deux groupes d'individus, cultivés et sauvages, comme ceci :

```
RC1;cultivated
RC2;cultivated
RC3;cultivated
RS1;wild
RS2 ;wild
```

Réaliser l'analyse de comparaison de SNPs (diagramme de Venn). Combien de variants sont partagés entre les individus cultivés et sauvages ? Combien de SNPs peuvent être considérés comme des marqueurs permettant de discriminer entre les compartiments cultivé et sauvage ?

Réaliser une analyse de diversité qui calcule et représente différents indices de diversité en fenêtre glissante.
Pouvez-vous localiser dans le génome des régions montrant un fort niveau de différentiation entre cultivés et sauvages (fortes valeurs de FST) ?

Afficher la diversité nucléotidique Pi pour chaque population « Pi by population » le long des chromosomes. Vous pourrez noter la différence de diversité entre les 2 groupes, les sauvages ayant une diversité plus forte que les cultivés.
Pouvez-vous identifier une région qui présente un profil anormal de Pi chez les cultivés (avec une diversité génétique particulièrement haute) ?
Par quelle hypothèse pouvez-vous expliquer cette observation ?

1-2-3- Densité de SNPs

Une inspection plus fine de cette région révèle que la plupart des variations provient d'un des individus cultivés.

Exporter les données au format Hapmap pour envoyer les variants vers une analyse de densité le long des chromosomes. Laisser la taille de la fenêtre glissante par défaut. Observer la densité de SNPs pour chaque échantillon pris indépendamment. Cela reflète le pourcentage de variations par rapport à la référence, pour chaque individu.
Quel individu est à l'origine de la diversité génétique exceptionnellement forte retrouvée chez les cultivés dans cette région ?
Est-ce que cela supporte votre première hypothèse ?
Vous pouvez éventuellement confirmer cela par la suppression de cet individu dans le jeu de données conduit à une réduction de la diversité génétique *d'O.glaberrima*

Note : L'analyse de polymorphismes est issue de données RNASeq donc la densité en SNPs va être impactée par densité en gènes le long des chromosomes.
Ce mode de représentation est à privilégier pour des données génomiques.

1-2-4- Arbre de distance

Un arbre de distance peut être reconstruit à partir des SNPs et allèles, en utilisant le logiciel Dnadist ou FastME.

Générer un arbre de distance (choisir l'export FASTA) en utilisant l'ensemble des loci situés entre 4 et 6 Mb du chromosome 5.
Ceci révèle que RC3 est distinctement un outgroup du clade *sativa+barthii/glaberrima*

1-2-5- Réseaux d'haplotypes

Un haplotype est l'association des allèles sur chaque chromosome homologue.
Pour une région contenant moins de 200 marqueurs, SNiPlay permet de reconstruire les blocks d'haplotypes par le logiciel Gevalt qui assigne ces haplotypes aux individus.
Dans un 2^{e} temps, un réseau d'haplotype peut être généré par le logiciel Haplophyle.

Note : Attention, il s'agit là d'inférence d'haplotypes par le lociel Gevalt), on n'utilise pas l'information de phasing du VCF.

Assigner les individus à 5 groupes : sativa, meridionalis, RC3, cultivés et sauvages.
Exporter les données de génotypage au format PED pour une analyse d'haplotypes.
Reconstruire un réseau d'haplotypes pour la région de 100kb située entre 5.5Mb et 5.6Mb du chr5.
Cela révèle-t-il un haplotype propre à RC3 ?

2- Analyse GWAS

Cette partie du TD va reprendre quelques aspects d'une étude de génétique d'association (GWAS) menées dans un article publié par Courtois *et al* en 2013. L'étude, basée sur des données GBS (Genotyping By Sequencing) obtenues sur un panel d'individus *O.sativa japonica*, vise à définir des marqueurs impliqués dans le contrôle de différents caractères racinaires.
L'idée est de procéder à une analyse globale GLM, puis de vérifier l'effet d'une correction par la structure et par la kinship (analyse MLM).
Ce jeu de données est accessible depuis l'application.
Choisir le Riz (Rice MSU7) et le projet « GBS_Courtois_et_al_2013 ».
Récupérer depuis Galaxy le jeu de données phénotypiques correspondant à cette étude, et l'enregistrer sur votre machine.

2-1- Analyse GLM

Collecter l'ensemble des marqueurs et les envoyer au pipeline GWAS en suivant un modèle GLM. Charger les données phénotypiques. Observer le fichier chargé, attendu en entrée.
Observer les quelques statistiques fournies après vérification des données puis lancer l'analyse.
Observer le Manhattan plot. Sur quels chromosomes pouvez-vous observer des marqueurs potentiellement associés ?
Observer le QQplot. Au vue de cette courbe, le modèle choisi est-il adapté aux données ? Garder la fenêtre ouverte pour les comparaisons ultérieures.

L'analyse GLM (General Linear Model) est réalisée ici par le logiciel TASSEL.
Un Manhattan plot permet de représenter sur l'axe X les marqueurs selon leur coordonnées génomiques et sur l'axe Y le logarithme négatif de la P-value d'association avec le trait étudié. Les marqueurs avec une forte association (avec les P-values les plus faibles) sont les points les plus élevés.
Un QQplot (quantile-quantile) permet d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique. On compare les positions dans la population observée par rapport à la position dans le théorique.

2-2- Structure de population

Supprimer les marqueurs multi-allèliques du jeu de données et envoyer les à une analyse de Structure de populations.
Tester différentes valeurs de K possibles entre 2 et 6.
Quel semble être la valeur optimale pour K ? Observer la représentation des clusters d'individus pour cette valeur de K. Récupérer le fichier de pourcentage d'admixture correspondant et enregistrer le.

Note : Pour une analyse de structure qui est gourmande en temps de calcul, il est possible de préalablement filtrer les variants pour ne garder que ceux qui sont assez distants les uns des autres (ex : intervalle minimum entre marqueurs) (ceux en fort DL peuvent contenir la même information).

2-3- Analyse de kinship

Lancer une analyse de kinship IBD puis récupérer la matrice d'apparentement.

L'analyse de kinship IBD est réalisée par TASSEL. Il fournit une matrice d'apparentement entre chaque individu 2 à 2.

2-4- Analyse GWAS corrigée par la structure et la kinship

Lancer une analyse de GWAS avec le modèle GLM et rentrer la structure pour une correction par la structure.
Observer la correction apportée au niveau du QQplot.
Lancer une 3^{e} analyse avec le modèle MLM afin de faire une correction par à la fois par la structure et par la kinship.
Observer la correction apportée au niveau du QQplot. Reste-t-il des marqueurs significatifs ?
Si oui, quels sont les allèles et haplotypes de ces marqueurs associés à un poids racinaire élevé ?
A proximité de quels gènes ces marqueurs sont-ils situés ?

Les analyses GWAS peuvent parfois donner des erreurs (faux-positifs) du fait de la structure des populations ou des apparentements entre individus.
Une des limitations de l'approche « association mapping » est le fort risque de faux-positifs en termes d'association dans les panels structurés. Il existe des modèles permettant de corriger les analyses par l'info de structure et kinship, afin de contrôler d'éventuels faux-positifs.