Exploitation des données de polymorphismes SNP à partir de fichier VCF

1- Exploration des données de génotypage à l'aide de Gigwa

Pour le TD, nous utiliserons un jeu de données correspondant au transcriptome complet obtenus à partir d'individus cultivés (*O.glaberrima*)(RC) et sauvages (*O.barthii*)(RS) de Riz africains (Nabholz et al, 2014).

Se connecter sur le site de Gigwa: http://gigwa.southgreen.fr/
Choisir la base Rice-MSU6_1
Lancer la requête sans filtre

Combien obtenez vous de SNPs?
Déterminer combien de SNP non synonymes sont présents dans la base.

Nous allons nous intéresser maintenant au SNPs entraînant l'apparition d'un codon stop prématuré.
Filtrer votre recherche afin de n'obtenir que les variants entraînant un codon stop prématuré.
Regarder le SNP position sur le 6517616 sur le chromosome 1.
Que pouvez vous en dire?

2- Analyse de SNPs dérivés de données RNA-Seq chez le Riz Africain (Individus Cultivés et Sauvages)

Le jeu de données que nous avons regardé dans Gigwa est accessible depuis Galaxy :

http://galaxy.southgreen.fr/galaxy/

« Shared Data => Data Libraries => Galaxy4SNiPlay => snp.vcf »

Charger le fichier VCF dans votre historique.

Importer le workflow « SNiPlay3_without_annotation" et observer l'organisation des briques logicielles.

1-1- Statistiques générales

Lancer le workflow « SNiPlay3_without_annotation » en écartant l'individu *meridionalis* (outgroup).

Combien y-a-t-il de SNPs au total découvert le transcriptome complet ? Combien de SNPs y a-t-il entre les individus africains ?

Observer les sorties graphiques montrant les statistiques générales.
Quel individu présente le niveau d'hétérozygotie le plus fort ?
Quel est le ratio Ts/Tv global ?
Observer la distribution de densité de SNPs le long des chromosomes. Que peut-on observer ?

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- Analyse de la 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.

Observer les sorties de 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 en fin de workflow par fastME. Pouvez-vous voir une séparation des 2 groupes à ce niveau ? Qu'observez-vous quant aux longueurs de branches ? Confirmez cette hypothèse en affichant le plot MDS des individus pour K=3.

1-3- 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.

Relancer un workflow en écartant cette fois *sativa* et *meridionalis* et en renseignant les informations de groupes « cultivés » et « sauvages ». Pour cela, vous pouvez créer un nouveau fichier en vous basant sur celui-ci généré par Snmf:

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

Observer les sorties d'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) ?
Combien de SNPs peuvent être considérés comme des marqueurs permettant de discriminer entre les compartiments cultivé et sauvage (FST=1) ?

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 ?

1-4- Densité de SNPs

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

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 pourriez 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*

1-5- Arbre de distance

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

Relancer le workflow sur l'ensemble des individus en ciblant spécifiquement les loci situés entre 4 et 6 Mb du chromosome 5. Visualiser l'arbre de distance.
Ceci révèle que RC3 est distinctement un outgroup du clade *sativa+barthii/glaberrima*

3- Analyse GWAS

Cette partie du TD va reprendre quelques aspects d'une étude de génétique d'association (GWAS) menées dans un article publié par *McCouch et al* en 2015. L'étude, basée sur des données de puces haute densité (HDRA : High density Rice Array) obtenues sur un panel d'individus *O.sativa japonica* et *indica*, vise à définir des marqueurs impliqués dans le contrôle de la taille du grain.
L'idée est de procéder à une analyse globale GLM, avec une correction par la structure.
Importer depuis la librairie partagée le VCF « chr1_2_3_4.gwas.vcf ».
Importer également le jeu de données phénotypiques correspondant à cette étude.
Lancer le workflow GWAS.
Observer le Manhattan plot. Sur quels chromosomes pouvez-vous observer des marqueurs potentiellement associés ?

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.

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.