Explore SNP polymorphism data from a VCF file using SNiPlay Galaxy

1 – Analysis of SNPs coming from RNASeq data on African rice (cultivated vs wild individuals)

We will use a SNP dataset derived from the complete transcriptome sequenced from wild (*O.barthii*) and cultivated (*O. glaberrima*) African Rice accessions, as described by Nabholz et al, 2014.
This dataset is accessible from the South Green galaxy instance (http://galaxy.southgreen.fr/galaxy/)
"Shared Data => Data Libraries => Galaxy4SNiPlay => RNASeq => snp.vcf"

Load the file into your history.

Import the workflow "SNiPlay3_without_annotation" and take a look at it (Edit the imported workflow).

1 – 1 - General statistics

Run the workflow excluding the individual named *meridionalis* (outgroup).

How many SNPs have been initially discovered in total using the complete transcriptome? How many SNPs remain between African accessions after filtering?

Observe the different graphs showing general statistics. Which individual has the highest rate of heterozygosity? *(vcf_stats.het)*.
What is the global Ts/Tv ratio? *(vcf_stats.TsTv.summary, snp_density.TsTv)*
Take a look at the SNP density along the chromosomes. What can you observe? *(snp_density.snpden)*

Note: SNPs are coming from RNASeq data analysis. Consequently, the SNP density will be affected by the location of genes along the chromosomes. This mode of representation is preferred for genomic data.

1 – 2 - Population structure

Note: the analysis of population structure is performed here by the software sNMF. sNMF will test the likelihood of each value of K (number of ancestral populations) and return the admixture percentages of each individual in the population.

Observe the outputs from sNMF. What is the best value of K? *(Structure by sNMF, Best K Groups)*
Observe the admixture values for the different groups for K=3.
Can you see a separation between the two groups?

A SNP-based distance tree has been generated using FastME. *(Newick viewer)*
Can you see a separation of the two groups? What can you tell about branch lengths?
Try to confirm this postulate by displaying the MDS (MultiDimensional Scaling) plot of individuals for K=3. *(analyse.mds_plot.txt)*

1 – 3 - Comparison wild/cultivated

Note: It is possible to combine external information to individuals. Typically, we can associate affiliation to cultivated or wild compartments, to be taken in consideration for subsequent analysis.

Run the pipeline again with these following instructions:

- Discard *sativa* and *meridionalis* samples
- Assign individuals to groups/populations. Define two groups of individuals, cultivated and wild, as follows:

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

This group assignation is defined in the file named “groups” that can be imported from Shared library, and then given as input in the workflow.

Observe the different plots of various diversity indexes in sliding windows. Can you identify SNPs that could be considered as good markers to distinguish between cultivated and wild compartments? Can you pinpoint regions in the genome showing a high level of differentiation between cultivated and wild (high FST values)? *(snp_density.fst.by_marker.genes.txt, snp_density.fst.txt)*

Display the nucleotide diversity Pi for each population plotted along the chromosomes. Note the difference of nucleotide diversity between the two groups. Wild accessions show higher nucleotide diversity than the cultivated. Can you identify a region that shows abnormal Pi profile in cultivated (with an exceptionally high genetic diversity)?

By what hypothesis can we explain this observation? *(snp_density.combined.pi.txt)*

1 – 4 - Density of SNPs

Close inspection of this genomic region reveals that most of the variations come from one cultivated individual.
Observe the SNP density for each sample taken independently. It reflects the percentage change compared to the reference for each individual. Which cultivated individual can explain the exceptionally high genetic diversity in cultivated in this region? Does it support your first hypothesis? *(densities.by_sample)*
Confirm that excluding this individual leads to a reduction in *O. glaberrima*'s genetic diversity in this region.
Observe the overall representation of elements using CircosJS output (SNP density, density by accession, TsTv…). Relaunch the Circos by including Pi information by population (as a line track). *(CircosJS on data…)*

1 – 5 - Distance tree

Note: a distance tree can be reconstructed from alleles of SNPs using the software FastME.

Run the workflow on all the individuals by targeting specifically the region between 4 and 6 Mb from the chromosome 5. Visualize the distance tree. Which individual is an outgroup of sativa+barthii/glaberrima clade? *(Newick viewer)*

2 – GWAS analysis

This part of the training will resume some aspects of a genetic association study (GWAS) conducted in an article published by McCouch et al in 2015. The study, based on data HDRA (High Density Rice Array) obtained on a panel rice individuals (*indica*, *japonica*, *aus*…), aims to set markers involved in the control of grain length.
The idea is to conduct a comprehensive GWAS analysis using GWAS Galaxy workflow.

Note: Variant dataset has been reduced for the training to make it workable (only one marker every 1000 bp, only 256 accessions taken randomly).

2 – 1- GWAS using Tassel GLM and correction by structure

Datasets can be accessed from shared libraries:
"Shared Data => Data Libraries => Galaxy4SNiPlay => GWAS"
Load the files into your history.
Import the workflow "SNiPlay3_GWAS_2" and take a look at it (Edit the imported workflow).
Run the workflow in order to determine markers associated to grain length of rice.
Observe some statistics provided (structure)
(Structure by sNMF)

Observe the Manhattan plot. Which chromosomes are potentially associated markers? Observe QQplot. In view of this curve, is the model suitable for data?
(Tassel output, QQ plot)

The GLM (General Linear Model) is performed here by TASSEL software. A Manhattan plot allows to represent on the X axis markers according to their genomic coordinates and the Y axis the negative logarithm of the P-value of association with the trait studied. The markers with a strong association (with P- the lowest values) are the highest points.
A QQplot (quantile-quantile) assesses the adequacy of the fit of a given supply in a theoretical model. Comparing positions in the observed population in relation to the position in the theoretical.

2 – 2- Focus on the best predicted QTL

Run the vcf2jbrowse wrapper to visualize the variants in the Rice Jbrowse of South Green platform. Is the highest significant marker located within a gene? *(View on Jbrowse)*

We consider now the possibility to reconstruct haplotypes for all accessions in the main QTL region, and determine whether specific alleles (and haplotypes) are associated to a long rice grain.
By zooming the Manhattan plot, we can focus on the main QTL region, from the most significant marker to the next marker having a log10 pvalue higher than 10 (chromosome 3, 16733440 to 16753385).
Import and run the workflow “Haplotype analysis”:

- Adjust the filtering of VCF to keep only the QTL region
- Use the gathering information based on grain length for colorization

Observe the haplotype network. Is there any distinct haplotype(s) shared by all the accessions showing the phenotype “Long”? *(Cytoscape network)*