TD : Detect polymorphism data from VCF file

1 - Using the Application SNiPlay exploit SNP s obtained by RNASeq

SNiPlay is a Web application dedicated to research and analysis of SNP s from sequencing data.

Go SNiPlay pipeline : http://sniplay.cirad.fr : Pipeline for SNP analysis => Pipeline v3

1 -1 SNP and Statistics 3 genes

Load obtained VCF file. " Add files "Then" Start upload »
Choose R iz as a reference genome for annotating SNPs s, and specify the option to recalculate the genomic positions.

Annotation of SNPs is achieved by SnpEff software based on the genomic positions of the variants and the annotation of the genome. One option is to reposition the variants on chromosomes if the mapping has been done on the CDS or the mRNA, such that SnpEff to operate. This feature uses the GFF annotation (positions of genes and transcripts) associated with the selected genome.

After selecting individuals left and chromosomes right, observe the SNP obtained and associated statistics.

Check that the genes found match those expected.
Why do some variants are found in UTRs ?
What part of synonyms SNP on all SNPs located in the CDS genes?

Observe the export format HapMap.

Now export variants in VCF format to calculate various general statistics. Observe the different information. What individual has the highest rate of heterozygosity ?

1 -2- analysis on the complete transcriptome (Grown dataset and wild)

For the remainder of TD, we use a data set corresponding to the complete transcriptome of wild and cultivated *individuals O.glaberrima* (Nabholz et al, 2014). This data set is accessible from the application.
Click on " Query databases "And then choose the rice and the project" RNASeq_Nabholz_et_al_2014 ".

How yt he discovered SNPs in total complete transcriptome ?

1 - 1- 2- density SNPs

Send variants to a density of SNPs analysis along chromosomes. Leave the size of the sliding window by default. You Qu'observez ?
Observe the SNP density for each sample taken independently. It reflects the percentage change compared to the reference for each individual.

The analysis of polymorphisms is from RNASeq thus the SNP density data will be affected by the density of genes along chromosomes.
This mode of representation is preferred for genomic data.

1 -2 -2 - flanking sequences for d esign Illumina chips

Export File potentially usable flanking sequences for designer SNP chips (Veracode technology) on the set of genes of chromosome 2?
What's in this file?
What would happen if an insertion / deletion was located just upstream of a SNP?

1 -2-3- SNP sharing between groups

It is possible to combine external information to individuals. Typically, it is possible to associate membership in cultivated or wild compartments.

Define two groups of individuals, cultivated and wild, in order to take into account for the analysis to follow.

RC1; cultivated
RC2, cultivated
RC3, cultivated
RS1; wild
RS2 ; Wild

Select export VCF and observe the SNP shared between cultivated and wild individuals.
How SNP possible to distinguish between cultivated and wild compartments ?

1 -2-4- distance Tree

U n distance tree can be reconstructed from the SNPs and allèèles, using DNADIST FastME or software.

Generating a distance shaft (change FASTA export) for all chromosome2 variants. Do you see a separation of the two groups ?
What do you observe about the branch lengths ?

1 - 2-5- Ha plotypes and LD

SNiPlay offers the possibility of reconstru ire haplotypes of individuals, that is to say the groups of alleles located on the same homologous chromosome. Caution, this haplotype inference there by lociel Gevalt) not using the phasing information VCF. Haplotypes of networks are generated by the software Haplophyle.

Switch to gene analysis mode (Level = genes).
Search for 3 previously studied genes and variants to export their PED format to send to a haplotype analysis.

How many does he separate haplotypes for the gene loc_ Os01g62920?
Observe haplotype networks. Are there specific haplotypes grown compartment for this gene ?

1 -2 to 6 - Analysis of a sub-sample and filter applications

Search Now only variants of chromosome 2 in the wild group. How many has he of non-synonymous SNPs on chromosome2, showing no missing data, with a MAF less than 10% ?

2 - GWAS analysis

This part of the TD will resume some aspects of a genetic association study (GWAS) conducted in an article published by *Courtois et al* in 2013. The study, based on data GBS (By Genotyping Sequencing) obtained on a panel *O. sativa japonica* individuals, aims to set markers involved in the control of different root traits.
The idea is to conduct a comprehensive analysis GLM and to verify the effect of a correction by the structure and the kinship (MLM analysis).
This data set is accessible from the application.
Click on " Query databases "And then choose the rice and the project" GBS_Courtois_et_al_2013 ".
Recover from the Galaxy game phenotypic data for this study, and save it on your machine.

2 -1 GLM Analysis

Collect all the markers and send them to GWAS pipeline following a GLM model. Load phenotypic data. Observe the loaded file, waits u input.
Observe some statistics provided after verification of data and start the analysis.
Observe the Manhattan plot. Which chromosomes can you observe potentially associated markers ?
Observe qqplot. In view of this curve, the model is it suitable for data ? Keep the window open for future comparisons.

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- population structure

Remove multi-allelic markers dataset and the S end an analysis of population structure.
Test different possible values of K between two and six.
What appears to be the optimal value for K ? Observe the representation of clusters of individuals for this value of K. Collect the percentage corresponding admixture file and save it.

Analysis of populations e structure is formed here by the software Admixture.
Admixture will test the plausibility of each value of K (number of ancestral populations) and return the admixture percentages of each individual in the population.
For an analysis of structure that is time-consuming calculation, it is possible to pre-filter variants to keep only those that are quite distant from each other (eg : Minimum interval between markers) (those high in DL may contain the same information).

2-3 - Kinship Analysis

Start a kinship analysis of IBD then retrieve the relatedness matrix.

The analysis is performed by IBD kinship TASSEL. It provides a matrix of relatedness between each individual 2-2.

2 -4- GWAS analysis corrected by the structure and the kinship

Start an analysis of GWAS with the GLM and return the structure to correction by the structure.
Observe the correction at the qqplot.
Start a 3rd analysis with MLM model to make a correction to both the structure and the kinship.
Observe the correction at the qqplot. There remains significant markers ?

GWAS analyzes can sometimes give errors (false positives) due to population structure or apparentements between individuals.
One of the limitations of the approach " Association mapping "Is the high risk of false positives in terms s association in structured panels. T here are models to correct analyzes the structural information and kinship, to control potential false positives.