SouthGreen bioinformatics platform

TD: Explore SNP polymorphism data from VCF file

1 - Using the SNiPlay application to explore SNPs obtained by RNA-Seq

SNiPlay is a Web application dedicated to detection and analysis of SNPs from sequencing data.

Go to the SNiPlay online pipeline: http://sniplay.southgreen.fr:
Pipeline for SNP analysis => Pipeline v3

1 -1 - SNP Statistics based on 3 genes

Load the VCF file previously obtained: “Add files” then “Start upload”
Choose “Rice_MSU6” as a reference genome for annotating SNPs, and specify the option that allows to recalculate genomic positions.

Annotation of SNPs is achieved by the SnpEff software by using both genomic positions of variants and the structural annotation of the genome. One option consists of preliminarily recalculating chromosomal positions of variants if the mapping has been done on CDS or mRNA, so that SnpEff can operate. This operation uses the selected genome together with its GFF annotation (positions of genes and transcripts).

After having selected individuals on the left and chromosomes on the right, observe the SNPs obtained and associated statistics.

Check that the genes found match those expected. Why are some variants located in UTRs?

What is the part of non-synonymous SNPs among those located in CDS?
Observe the export format HapMap.

Now export variants in BCF format to be sent to VCFtools that calculate various general statistics. Observe the different information.
What individual has the highest rate of heterozygosity?

1 - 2 - Analysis of the complete transcriptome (Cultivated + wild dataset)

To continue, we will use a dataset corresponding to the complete transcriptome sequenced from wild (*O.barthii*) and cultivated (*O.glaberrima)* African Rice accessions (Nabholz et al, 2014).
This dataset is accessible from the application.
Click on “SNP Database” => "SNP Queries v3".
Select Rice_MSU6 => RNASeq_Nabholz_et_al_2014 project

How many SNPs have been discovered in total using the complete transcriptome?

1 - 2 - 1 - Population structure

sNMF and entropy

After having discarded the meridionalis sample (outgroup), compute a population structure analysis with the sNMF software.
What is the best value of K (number of populations)?
Observe the admixture values for the different groups for K=3.
Can we observe a separation between our two groups?

Cultivated and wild => different level of diversity

A SNP-based distance tree has been generated using FastMe. Do you see a separation of the two groups?
What do you observe about branch lengths?
Confirm the postulate by displaying the MDS plot of individuals and colorization for K=3.

1 - 2 - 2 - Comparison cultivated/wild

It is possible to combine external information to individuals. Typically, it is possible to associate affiliation to cultivated or wild compartments, to be taken in consideration for subsequent analyses.

Discard sativa and meridionalis for this task.
By using the toolbox “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
```

Run the analysis for comparison of SNPs (Venn diagram). How many variants are shared between cultivated and wild individuals? How many SNPs can be considered as good markers to distinguish between cultivated and wild compartments?

Perform a diversity analysis that calculates and plots in sliding windows various diversity indexes.
Can we localize regions in the genome that show high level of differentiation between cultivated and wild (high FST values)?
Display the nucleotide diversity Pi for each population (“Pi by population”), plotted along the chromosomes. Note the difference of nucleotide diversity.

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?

1 - 2 - 3 - Density of SNPs

Close inspection of this genomic region reveals that most of the variation comes from one cultivated individual.

Export in Hapmap format to send variants to an analysis of SNP density along chromosomes. Leave the size of the sliding window by default.
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?
Confirm that excluding this individual leads to a reduction in *O. glaberrima*'s genetic diversity in this region.

Note: The analysis of polymorphisms is from RNA-Seq 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 - 4 - Distance tree

Reconstruct the phylogeny using all loci positioned between position 4 and 6 Mb of chromosome 5. This reveals that RC3 is distinctively an out-group of *sativa+barthii/glaberrima* clade.

1 - 2 - 5 – Haplotype network

The haplotype is the association of alleles for each homologous chromosome. For a region containing less than 200 markers, SNiPlay allows haplotype blocks to be estimated using the Gevalt software in order to then calculate haplotype network using Haplophyle. Each individual is assigned to a haplotype and network

Assign your individuals to 5 groups: sativa, meridionalis, RC3, cultivated and wild.
Export genotyping data in PED for haplotype analysis.
Reconstruct a haplotype network for the 100kb region from 5.5Mb to 5.6Mb of chr5.
Does this reveal a separate haplotype for RC3?

1 - 2 - 6 – Get flanking sequences for the design of Illumina chips

Export a file reporting flanking sequences that can be potentially usable for the design of SNP chips (Veracode technology). To do so, you can first specify a minimum distance between two sites so that markers are homogenously distributed along the genome.
What is in this file?
What would happen if an insertion / deletion is located just upstream of a SNP?

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 prefilter
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 structure and 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.