Discovery and analysis of SNP polymorphisms

1- Tablet: Manual detection of SNPs with Tablet

Tablet is a graphical viewer for NGS (Next Generation Sequencing) assemblies and alignments.

Retrieve the SAM assembly (*selection.sam*) from Galaxy Shared libraries.
Import the SAM file and the reference file into your history.
Look at the SAM file, and try to guess the signification of each column.
How is coded the alignment of a read?

Launch the Tablet software via Java Web Start: http://bioinf.hutton.ac.uk/tablet/faq.shtml.
Ensure that the assembly has already been sorted with SortSam and load the assembly file (gathering all individuals) in the SAM format.
Select the gene *GSMUA_Achr7G15140_001*.

- Can you identify SNPs?
- Identify a SNP for which GrandNainHolland differ from the reference sequence.
- Identify a SNP for which all sequenced individuals differ from the reference sequence (except *PahangHD*).
- Can you identify a deletion?
- Identify a SNP resulting from a difference between all sequenced individuals.
- Identify a SNP showing a heterozygous position.
- Intuitively, how do you estimate the reliability of a SNP? How can we distinguish between a heterozygous position and a sequencing mistake?
- What problems can we encounter when analyzing SNPs on multigenic families or pseudogenes?

Technical notes:
Tablet proposes many options for the visualization of assembled reads, notably the « Read Groups » coloration or the highlight of variants.
This software enables to visually confirm automatically detected SNPs but cannot be used for a large scale detection of variants.

2- SNP detection using GATK in Galaxy

GATK (Genome Analysis Tool Kit) « is a structured software library that makes writing efficient analysis tools using next-generation sequencing data ».
GATK detects SNP and indels and assign to each position a genotyping value to individuals. GATK provides an output file in the VCF format (Variant Call Format).

Open Galaxy.
Convert the SAM file into a binary alignment with *SAM-to-BAM.*

Launch the *IndelRealigner* module from the BAM file.

- Observe the different intervals the program has targeted to realign the reads locally? Verify with Tablet for the gene *GSMUA_Achr7G15140_001*.

Launch the *UnifiedGenotyper* module.

- What is the significance of the different fields of the VCF format?
- Observe the SNPs obtained and verify the reliability with Tablet.
- Run this module again with more stringent parameters to filter out SNPs, and remove SNP marked "LowQuality"
- Identify in VCF a heterozygous position?

Launch the *DepthOfCoverage* module to obtain depth of coverage for each position and individual. Observe the output.

Launch the *ReadBackedPhasing* module enabling to get haplotypes.

- For individuals showing heterozygote positions, can haplotypes be directly identified using sequencing techniques (with *ReadBackedPhasing*)? For this question, you can open the phased VCF file and try to reconstitute manually the two haplotypes of gene *GSMUA_Achr7G15140_001* for individual *GrandNainHolland*.
- Can you identify two consecutive heterozygote positions for which the phasing has been resolved using the reads?

Restart the analysis by making a complete GATK workflow.

Technical notes:

- *Note1: The IndelRealigner module enables a local realignment around insertions/deletions, in order to avoid errors in SNP calling.*

- *Note2: The UnifiedGenotyper module proposes 2 options called « Stand call conf » and « Stand emit conf » to fix SNP quality thresholds for which SNPs can be called and tagged as « PASS » or « LowQual ».*

- *Note3: It was observed that errors of SNP calling occur more frequently on transversions than on transitions.*

3- Use of the SNiPlay pipeline for exploring SNPs

SNiPlay is a Web-based application dedicated to detection and analysis of SNP from sequencing data.

Go to the SNiPlay pipeline: http://sniplay.cirad.fr.
Select the VCF format input.
Rename on your computer input files as required by the application:

- polymorphisms.vcf
- reference.fa
- depth.txt

Load the 3 input files.
Enter 3 for minimum depth coverage.
Choose the Banana reference genome to anchor SNPs in the genome, and check the option « Reference sequences correspond to CDS ».
For haplotypes reconstruction, choose the program Gevalt.
Select the step "distance tree".
Select all individuals.

<u>Technical notes</u>:

Several options are available in SNiPlay:

- an option consists of filtering out SNPs on missing data position by position, in other term to remove sequence sections having more than a specified percentage of missing data

- an option consists of discarding multiallelic SNPs or inserions/deletions from the analysis, or of recoding them into biallelic SNPs.

- an option consists of using directly GFF genome annotation for SNP annotation (genomic positions, synonym/non-synonym) instead of using Blast (if the mapping has been performed against the predicted CDS).

3-1- SNP and statistics

- Observe SNPs and associated statistics.
- Observe alignments reconstructed from VCF file and reference. What kind of information given as input enabled to define the position where each individual sequence must begin?

3-2- Design of Illumina genotyping chip

- Find the file that you will be able to submit to Illumina to design SNP chips (VeraCode technology) for the whole genes.
- What does this file contain?
- What would be happened if an insertion/deletion is located 20 bases before a SNP?

3-3- Allelic files

SNiPlay generates genotyping files in different format specific to recognized analysis softwares: STRUCTURE, DARwin, Phase, TASSEL

- Observe the different available genotyping formats.

3-4- Annotation des SNP

SNiPlay is able to map sequences on a reference genome and to annotate the SNPs.
In our case, the reference sequence used for the mapping corresponds to CDS from the genome annotation, we have specified as an option and SNiPlay locate SNPs on the genome so that they can be annotated by SnpEff.

- Verify that gene names correspond to expected ones.
- What is the proportion of synonymous SNPs?
- How many transversions are there?

3-5- Haplotype reconstruction

Gevalt has the ability to reconstruct haplotypes for each individual, i.e. the combination of alleles at adjacent locations (loci) on each homologous chromosome.

- How many distinct haplotypes are there for the gene *GSMUA_Achr7G15140_001*?
- How does Gevalt manage missing data?

3-6- Haplotype networks

Haplophyle is a pipeline for the analysis of genotyping data and includes haplotype network analysis.

Select the « Network analysis » step to visualize these networks.
You can also launch the program independently http://haplophyle.cirad.fr.

3-7- Distance tree

- Observe the distance tree generated for each gene.

3-8- Diversity indexes

EggLib calcultaes various diversity indexes and SNiPlay provides series of diplays/plots to show the distribution of the different values. This functionality can be used for large scale analysis (when analyzing a complete transcriptome dataset for example).

- Look at the signification of the different diversity indexes provided.
- Observe the different graphical outputs.

4- Analysis of a subset of samples

Restart the analysis by changing the sample of analyzed individual. For instance, keep only the reference, Pahang and PahangHD to detect variations within this subset.

- Observe new results.
- How many SNP are remaining? Compare with your initial manual predictions.
- Can you also verify that PahangHD (doubled haploid) correspond to the reference.

5- SNP sharing between groups

SNiPlay has the ability to associate external informations to sequences. Typically, it is possible to link geographic origin or genetic group assessment (ex: cultivated or wild compartments).

Restart the analysis (with all individuals) by adding some fictive external information linked to analyzed individuals such as this example below

```
Accession,ploidy
GrandNain,triploid
GrandNainHolland,triploid
PisangMadu,diploid
PisangPipit,diploid
Pahang,diploid
PahangHD,diploid
reference,diploid
```

- Observe SNP sharing between groups.
- Can you identify SNP positions able to distinguish between diploid and triploid accessions?
- Can you identify a haplotype specific to diploid accession?

Pahang is supposed to be heterozygote, with one haplotype corresponding to PahangHD (and the reference).

- Can you verify this assumption by using groups and haplotypes?

6- SNPs in the Banana Genome Hub

A SNP calling has been performed after the mapping of RNA-Seq reads originated from 2 individuals (Pahang and PahangHD) against the predicted CDS. Resulting SNPs have been integrated into the SNiPlay database, then exported in GFF to be integrated the Banana Genome Browser. All these information are available from the Banana Genome Hub.

Go to http://banana-genome.cirad.fr, in the section Tools => SNP.

- How many SNPs are referenced in the database?
- How many SNPs are referenced between PahangHD and the reference? Between Pahang and the reference? How many heterozygote positions are found in Pahang and PahandHD?
- Which gene(s) in chromosome 11 contains more than 3 non-synonymous SNPs with no missing data?
- Look at the SNPs at positions 3518436 and 3518450 in the Genome Browser.