20K Grapevine Illumina SNP chip

The **GrapeReSeq_Illumina_20K_SNP_chip** contains 15022 SNPs from *Vitis vinifera* genotypes and 4978 SNPs from *Vitis* species that were chosen with the following process.

I - PART 1 : 15K Grapevine Illumina SNP chip (*Vitis vinifera*)

computed by IGA (http://www.appliedgenomics.org/) and URGI (http://urgi.versailles.inra.fr/).

1/ List of genotypes

1.1/ The following genotypes were paired-end sequenced using an Illumina GAII or HighSeq plateform and the sequences were analyzed at URGI in the frame of the KBBE GrapeReSeq EU project (see Vitis_Sequences_Stats.pdf):

Araklinos
Cabernet franc
Carignan noir
Castellana Blanca
Chouchillon
Colorino
Espadeiro tinto
Jaén
Lambrusque Campmarcel
Hebén
Lambrusque E
Lambrusque-psl2
Listàn Prieto
Malvasia
Maska
Medouar
Vitis sylvestris Dirmstein mâle
Orlovi nogti
Savagnin
Sultanine
Tsolikoouri
Teulere pied sauvage

1.2/ The following genotypes were paired-end sequenced using an Illumina GAII or HighSeq plateform and the sequences were analyzed at IGA (see Vitis_Sequences_Stats.pdf):

Aglianico
Barbera
Bovale
Cannonau
Carignano
Cesanese daffile
Enantio
Pinot Noir ENTAV 115
Grignolino
Lambrusco grasparossa
Lambrusco sorbara
Montepulciano
Nebbiolo
Negro amaro
Nero_davola
Nieddu
Petit rouge
Primitivo
Raboso piave
Sagrantino
Sangiovese
Schiava grossa
Schioppettino
Tocai R5
Vermentino

2/ Quality trimming

At URGI :
The short reads were trimmed and filtered when transmitted to URGI (see README_EPGV_DataTransfer_Illumina_Sequencing.pdf)
The chloroplast reads were not filtered out before mapping.

At IGA :
Quality trimming and chloroplast filtering with
http://iga-rna.sourceforge.net/ tool
A typical common command is as follow:
a) rNA --create --fasta chloroplast.fa --reference chloroplast.rNA
b) rNA --filter-for-assembly --query1 $READ1 --query2 $READ2 --reference chloroplast.rNA --threads 8 --min-mean-phred-quality 20 --min-size 50 --output $TRIM_OUT

where $READ1 and $READ2 are paired fastq files and $TRIM_OUT is the output filename.

Quality trimming is based on the modified-Mott's algorithm (e.g. see http://www.clcbio.com/manual/genomics/Quality_trimming.html).

3/ Alignments with BWA
=> each group (URGI and IGA) did his own alignments on the reference genome PN40024 (12X version) and SNP call, one for **each accession.**

Alignments were done with BWA with -n 0.01 ("distribution"parameter) at URGI and default parameters (-n 0.04) at IGA.
(see http://biostar.stackexchange.com/questions/16267/what-does-bwas-n-parameter-mean)

4/ SNP detection
SNP calling have been run **on each single accession separately** and on reads aligning uniquely along the genome to avoid misplacements, usually due to repeats

⇨ input: only uniquely aligned reads, with flag XT:A:U after filtering of the SAM format.

IGA used VarScan2.2.3 and URGI VarScan2.2.8.

5/ First SNP filtering

- **min-coverage** : Minimum read depth at a position to make a call : **4**
- min-reads2 : Minimum supporting reads at a position to call variants : 2 (default)
- min-avg-qual : Minimum base quality at a position to count a read : 15 (default)
- min-var-freq : Minimum variant allele frequency threshold : 0.01 (default)

- p-value : p-value threshold for calling variants : 99e-02 (default)
- min Freq. to call homozygote: 0.75 (default)
- Ignore variants with >90% support on one strand: Yes

Additional filters were applied only by URGI on its sequences:

- Removing the positions with 3 or 4 SNPs for one individual that probably correspond to paralogous positions.
- Removing SNPs corresponding to sequencing errors based on PN40024 SNPs calling.

IGA further filtered all VarScan results as follows:

- Both SNPs datasets were filtered out according to their coverage compared to the average genome coverage. SNPs were kept if their coverage was above a minimum coverage corresponding to half the average and below a maximum coverage corresponding to the double of the average coverage. This was done to avoid errors derived from low coverage regions and from possible remaining mapping on repeated regions.
- Both SNP datasets were filtered out with a DIP (Deletion/Insertion Polymorphisms, i.e. few bp Indel) and structural variation (large Indel) database built by IGA on the Italian set of varieties under the assumption that most of them would be also found in the french set of varieties. DIPs were computed with VarScan (same parameters as for SNPs detection) and redundant DIPs were removed.
- For both datasets, SNP in repetitive regions were filtered out: for that purpose, IGA combined in a single track :
    - computationally detected repeats with ReAS (IGA), RepeatMasker (Genoscope), Tandem Repeats Finder (TRF) (Genoscope)
    - manually curated transposable elements (IGA)
    - microsatellites and 100 surrounding bp (IGA)
- Adjacent SNPs i.e in a 50 bp window around the chosen SNP were filtered out to comply with ISelect design constraints. 50 bp regions around the SNP were extracted and blasted against the genome to check for unicity (either of the two primers with more than one Blast hit at 1e-10).

6/ Second SNP Filtering

353K filtered SNPs were submitted to Illumina (in ADT format) to be scored by the Assay Design Tool (https://icom.illumina.com).

- SNPs with Illumina score < **0.9** were filtered out. Only SNP for Infinium II array (Type2) were kept (*Infinium I SNPs, A/T, C/G transversions, are not used because they necessitate two beads per SNP to type them*).
- IGA provided the number of accessions/chromosomes supporting a SNP and the following rules were applied to further select the SNPs:
    - ⇨ 90% of the SNP with MAF (Minor Allele Frequency) above 0.1 (85463 SNPs)
    - ⇨ 10% of the SNP with MAF between 0.05 and 0.1 (27631 SNPs)

URGI finally selected equidistant SNPs (see Vvinifera_SNPnumber_per_chr.pdf).

8/ Illumina genotyping array SNP distribution

=> **15022 SNPs** have been chosen on 47 *Vitis vinifera* genotypes.
with MAF > 0.1 : 13347 SNPs
with 0.05 < MAF < 0.1 : 1470 SNPs
with SNPs from ICVV team (J. Zapater) : 205 SNPs

II - PART 2 : 5K Grapevine Illumina SNP chip (*Vitis* species)

computed by URGI (http://urgi.versailles.inra.f).

1/ List of genotypes

Vitis aestivalis
- *Vitis aestivalis 1*
- *Vitis aestivalis 2*
- *Vitis aestivalis 3*
- *Vitis aestivalis sauvage*

Vitis berlandierii
- *Vitis berlandierii 10594*
- *Vitis berlandierii planchon*
- *Vitis berlandierii resseguier*

Vitis labrusca
- *Vitis labrusca concord*
- *Vitis labrusca fredonia*
- *Vitis labrusca labrusca*

Vitis cinerea
- *Vitis cinerea 1*
- *Vitis cinerea 2*
- *Vitis cinerea 3*

Vitis lincecumï
- *Vitis lincecumii-gross*

Muscadinia rotundifolia

2/ Quality trimming

The short reads were trimmed and filtered when transmitted to URGI (see README_EPGV_DataTransfer_Illumina_Sequencing.pdf)
The chloroplast reads were not filtered out before mapping.

3/ Alignments with BWA

Alignments were done with BWA with -n 0.01 (“distribution”parameter; see http://biostar.stackexchange.com/questions/16267/what-does-bwas-n-parameter-mean)

4/ SNP detection

SNP calling have been run **on each single accession separately** and on reads aligning uniquely along the genome (to avoid misplacements, usually due to repeats).
=> input: only uniquely aligned reads, with flag XT:A:U after filtering of the SAM format.

VarScan2.2.8 was used with the following parameters :

- **min-coverage** : Minimum read depth at a position to make a call : **10**
- min-reads2 : Minimum supporting reads at a position to call variants : 4
- min-avg-qual : Minimum base quality at a position to count a read : 30
- min-var-freq : Minimum variant allele frequency threshold : 30%
- p-value : p-value threshold for calling variants : 0.001

5/ SNP filtering

Additional filters were applied on the detected SNPs:

- Removing the positions with 3 or 4 SNPs for one individual that probably correspond to paralogous positions.

- Removing SNPs corresponding to sequencing errors based on PN40024 SNPs calling.
- Filtering out of clustered SNPs (group of SNPs closer than **60** bp to each other) for ISelect design.

6/ Illumina submission

The filtered SNP was submitted to Illumina (in ADT format) for scoring by the Assay Design Tool of Illumina (https://icom.illumina.com).

7/ Final filters

- Only SNP for Infinium II array (Type2) were kept (*InfiniumI SNPs, A/T, C/G transversions, are not used because they necessitate two beads per SNP to type them*)
- SNP with ADT scoring < **1** were filtered out (<0.7 for *M. rotundifolia*).
- Redondant SNPs between species have been manually curated to avoid redundancies.
- Equidistant SNPs along the *V. vinifera* genome sequence were finally chosen.

8/ Illumina genotyping array SNP distribution

=> **4978 SNPs** have been chosen on 15 *Vitis* species.

Vitis aestivalis : 1000 SNPs (ae)
Vitis berlandierii : 1000 SNPs (be)
Vitis labrusca : 1000 SNPs (lb)
Vitis cinerea : 1000 SNPs (cn)
Vitis lincecumï : 400 SNPs (li)
Muscadinia rotundifolia : 578 SNPs (mu)