Analyse de variants génétiques (SNPs, indels)

Alexis Dereeper – Christine Tranchant

Formation Bio-informatique IRD Ouagadougou 2016

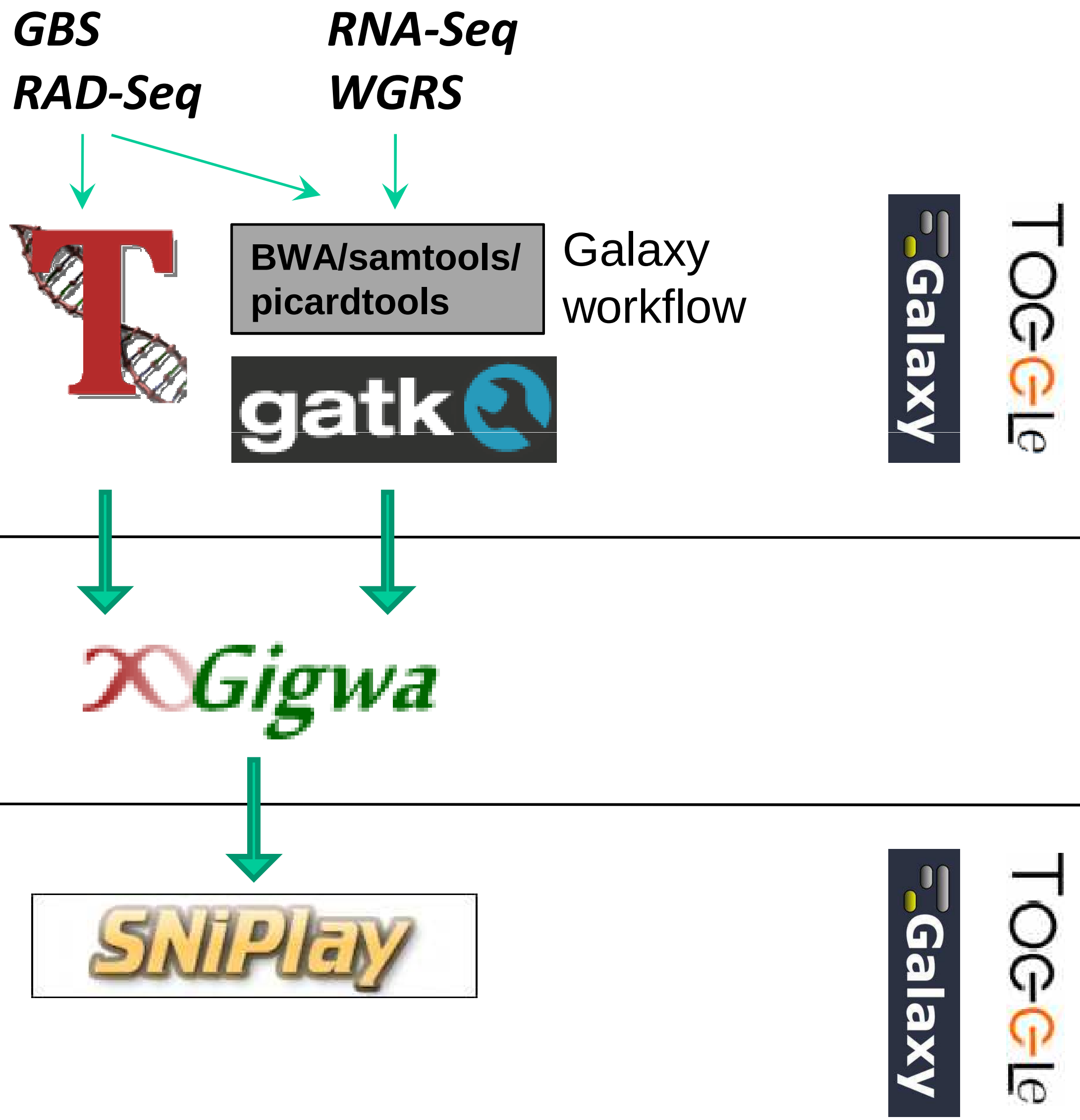
I- SNP calling and genotype assignation

II- SNP data analyses and visualization

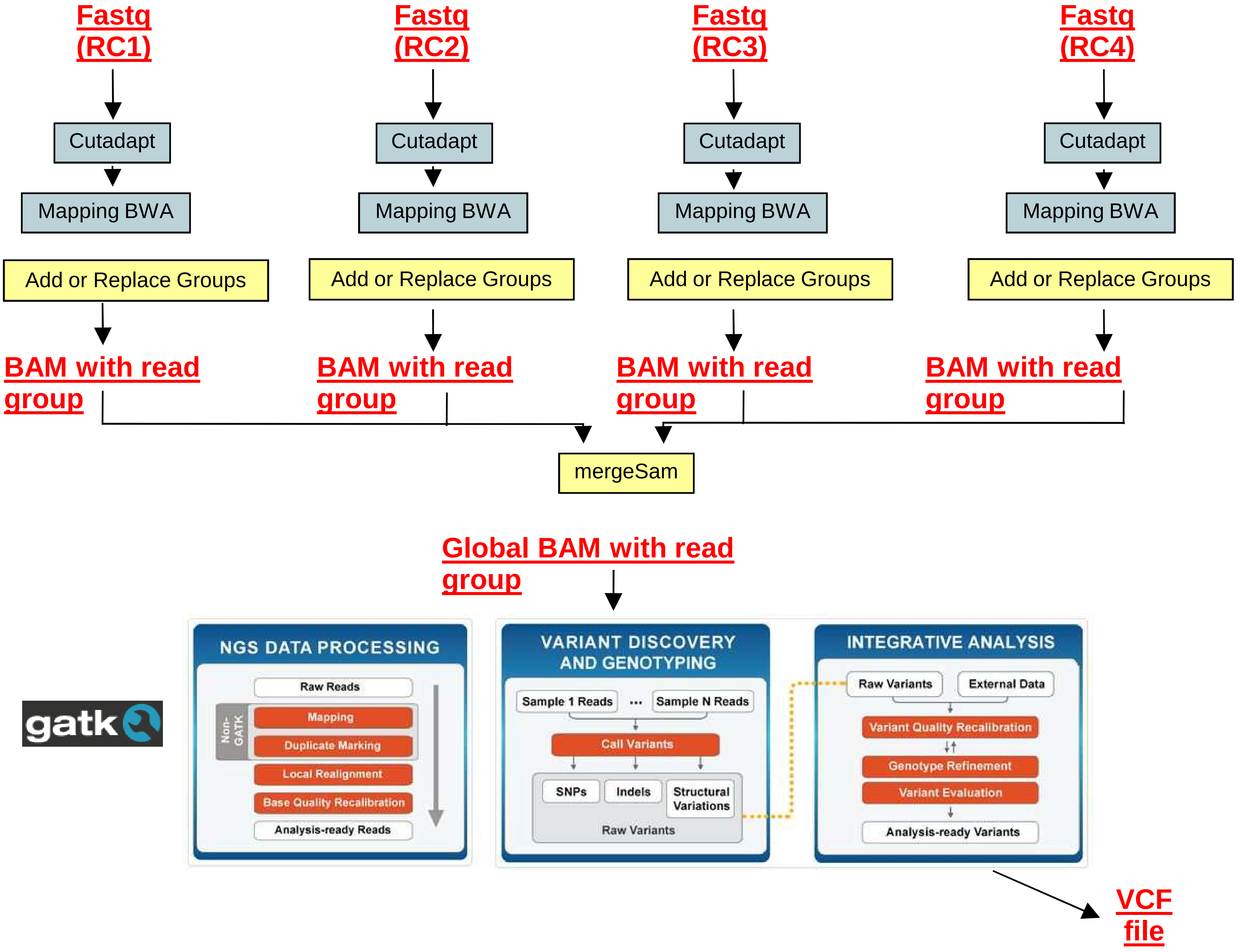
Reads pre-processing and mapping
+
SNP Calling and genotype assignation

Genotyping data
Storage and mining

Genotyping data analyses and visualization
(GWAS, diversity…)



I- SNP calling and genotype assignation



Fastq (RC1)

Cutadapt

Mapping BWA

Add or Replace Groups

BAM with read group

Fastq (RC2)

Cutadapt

Mapping BWA

Add or Replace Groups

BAM with read group

Fastq (RC3)

Cutadapt

Mapping BWA

Add or Replace Groups

BAM with read group

Fastq (RC4)

Cutadapt

Mapping BWA

Add or Replace Groups

BAM with read group

mergeSam

Global BAM with read group

gatk

VCF file

Format SAM/BAM

A header listing at least the different sequences in the reference and their length.

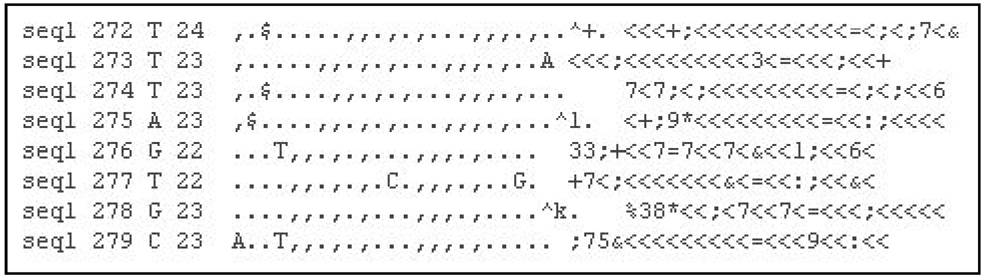
At least one line for each read. The different fields are separated by tabulation.

```
ILLUMINA-199BA2:90:FC:3:73:6197:17241 16    chr11       66288622  37    72M  *      0      0
TGGGTGTAAGTGAGGAGATTGGGACTTTAGACCAGGCACTCACTCAGAGGAGTTACTGACAGGGCA
GGGAGG ;BHHHIFHFIHIHIIHIHIIIIGDIIIIGIEIIHIIIHIIIFIHIIIIIIHIIIHIIHIIIIIIHIFIIIIII      XT:A:U      NM:i:0
      X0:i:1       X1:i:0       XM:i:0       XO:i:0       XG:i:0       MD:Z:72
```

Read name
Bitwise flag
Reference sequence
Start position of the alignment on the reference
Mapping quality (255 means unknown quality)
CIGAR String
Reference sequence and alignment start of the mate
Observed fragment length
Nucleotidic and quality sequence of the read
And optional fields

Format Pileup

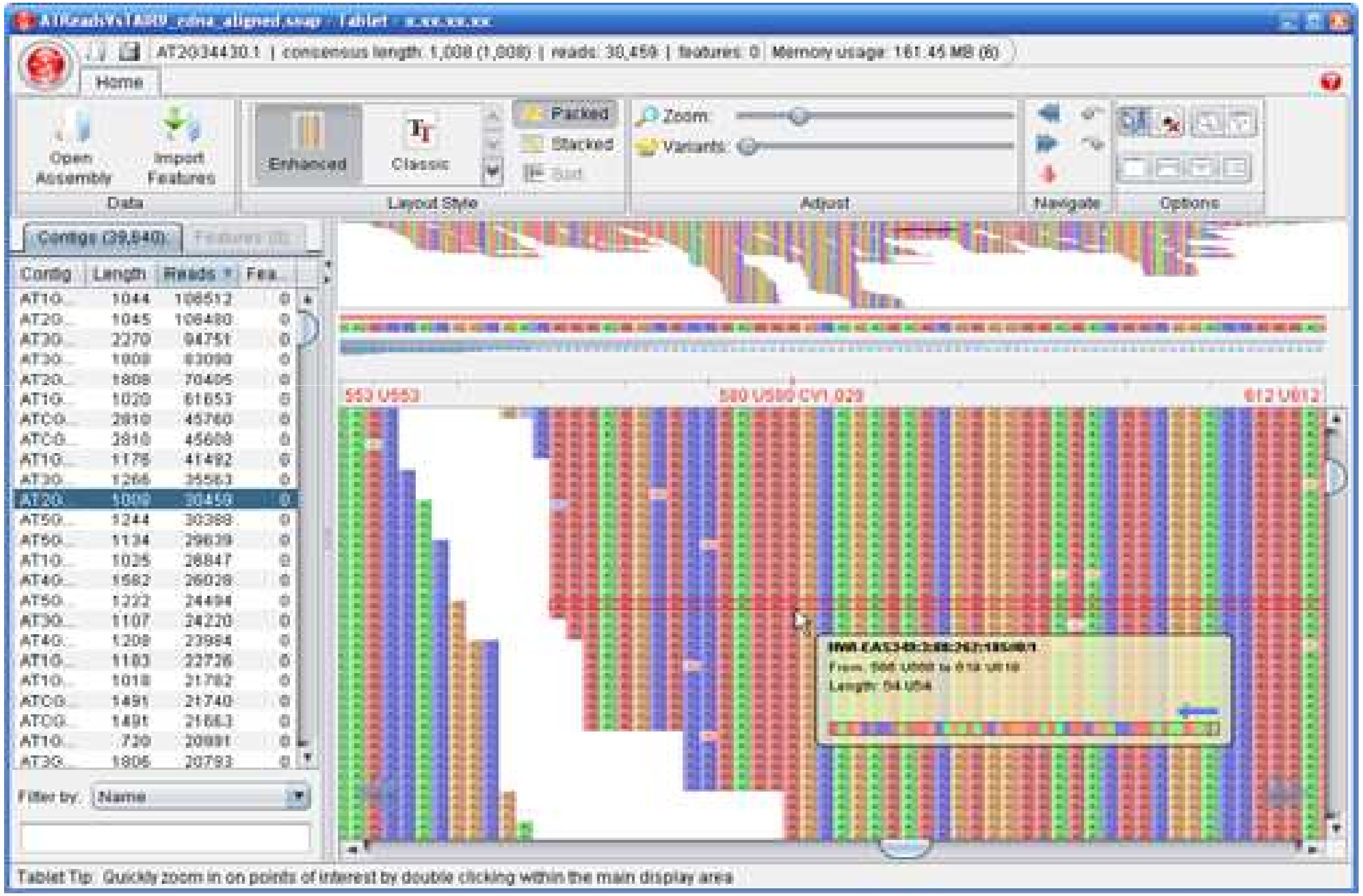
- Another format for variant calling (generated by samtools)
- Describe alignment row by row (not line by line like in SAM format)
- Used by softwares such as **Varscan** (varscan pileup2snp)
- Frequently used for rare variants, with a low frequency (e.g. viral pop)



```
seq1 272 T 24  ,.$.....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,.,..A <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,.,...    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,,.,.,...,,,.,....^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,..... ;75&<<<<<<<<<=<<<9<<:<<
```

Tablet

- Graphical tool to visualize assemblies
- Accept many formats ACE, SAM, BAM



ATReadsVsTAIR9_cdna_aligned.soap - Tablet - x.xx.xx.xx

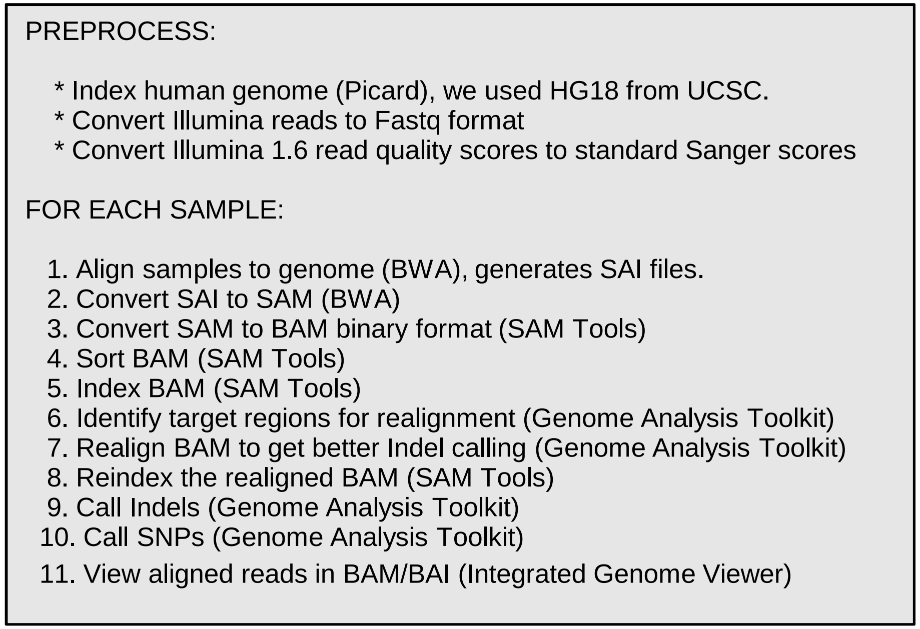
AT2G34430.1 | consensus length: 1,008 (1,008) | reads: 30,459 | features: 0 | Memory usage: 161.45 MB (6)

Contig	Length	Reads	Fea...
AT1G...	1044	108512	0
AT2G...	1045	106480	0
AT3G...	2270	94751	0
AT3G...	1009	83090	0
AT2G...	1808	70405	0
AT1G...	1020	61653	0
ATCG...	2810	45760	0
ATCG...	2810	45609	0
AT1G...	1176	41492	0
AT3G...	1266	35563	0
AT2G...	1008	30459	0
AT5G...	1244	30388	0
AT5G...	1134	29639	0
AT1G...	1025	26847	0
AT4G...	1582	26028	0
AT5G...	1222	24494	0
AT3G...	1107	24220	0
AT4G...	1208	23984	0
AT1G...	1183	22726	0
AT1G...	1018	21782	0
ATCG...	1491	21740	0
ATCG...	1491	21663	0
AT1G...	720	20891	0
AT3G...	1806	20793	0

Tablet Tip: Quickly zoom in on points of interest by double clicking within the main display area

GATK (Genome Analysis ToolKit)

- Software package to analyse NGS data.
- Implemented to analyse human resequencing data, for medical purpose (1000 genomes, The Cancer Genome Atlas)
- Includes depth analyses, quality score recalibration, SNP/InDel detection
- Complementary with other packages: SamTools, PicardTools, VCFtools, BEDtools



PREPROCESS:

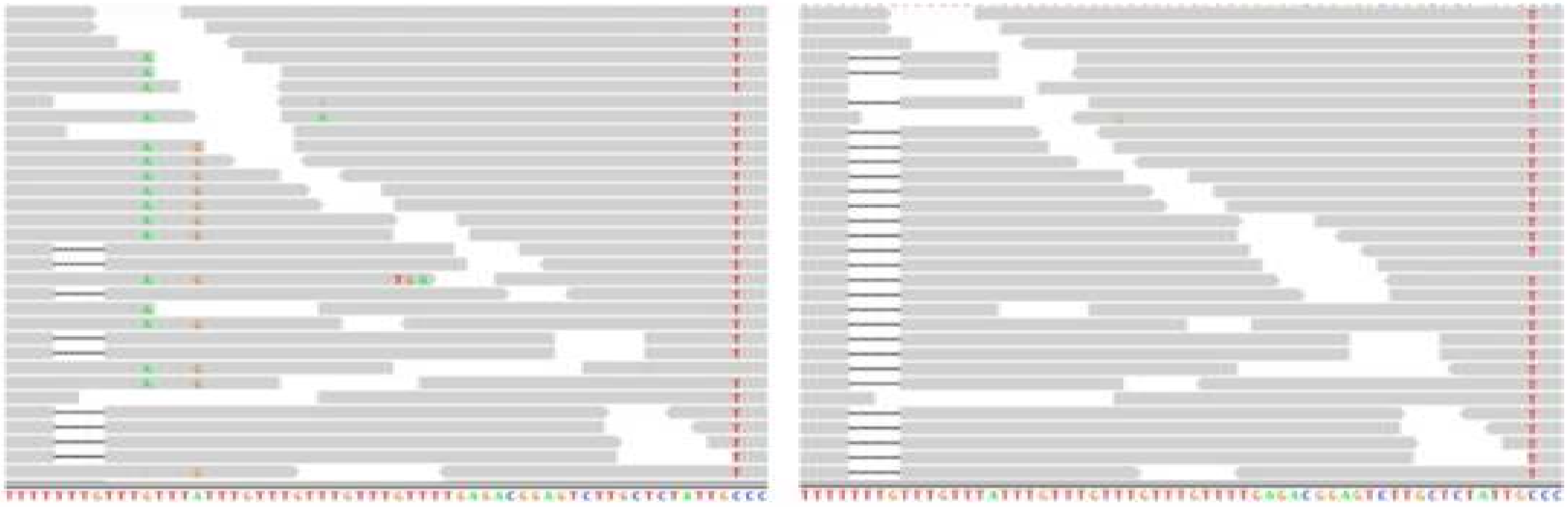
* Index human genome (Picard), we used HG18 from UCSC.
* Convert Illumina reads to Fastq format
* Convert Illumina 1.6 read quality scores to standard Sanger scores

FOR EACH SAMPLE:

1. Align samples to genome (BWA), generates SAI files.
2. Convert SAI to SAM (BWA)
3. Convert SAM to BAM binary format (SAM Tools)
4. Sort BAM (SAM Tools)
5. Index BAM (SAM Tools)
6. Identify target regions for realignment (Genome Analysis Toolkit)
7. Realign BAM to get better Indel calling (Genome Analysis Toolkit)
8. Reindex the realigned BAM (SAM Tools)
9. Call Indels (Genome Analysis Toolkit)
10. Call SNPs (Genome Analysis Toolkit)
11. View aligned reads in BAM/BAI (Integrated Genome Viewer)

GATK (Genome Analysis ToolKit)

- IndelRealigner module: realigns around indels in order to avoid false positive SNPs



HiSeq data, raw BWA alignments

HiSeq data, after MSA

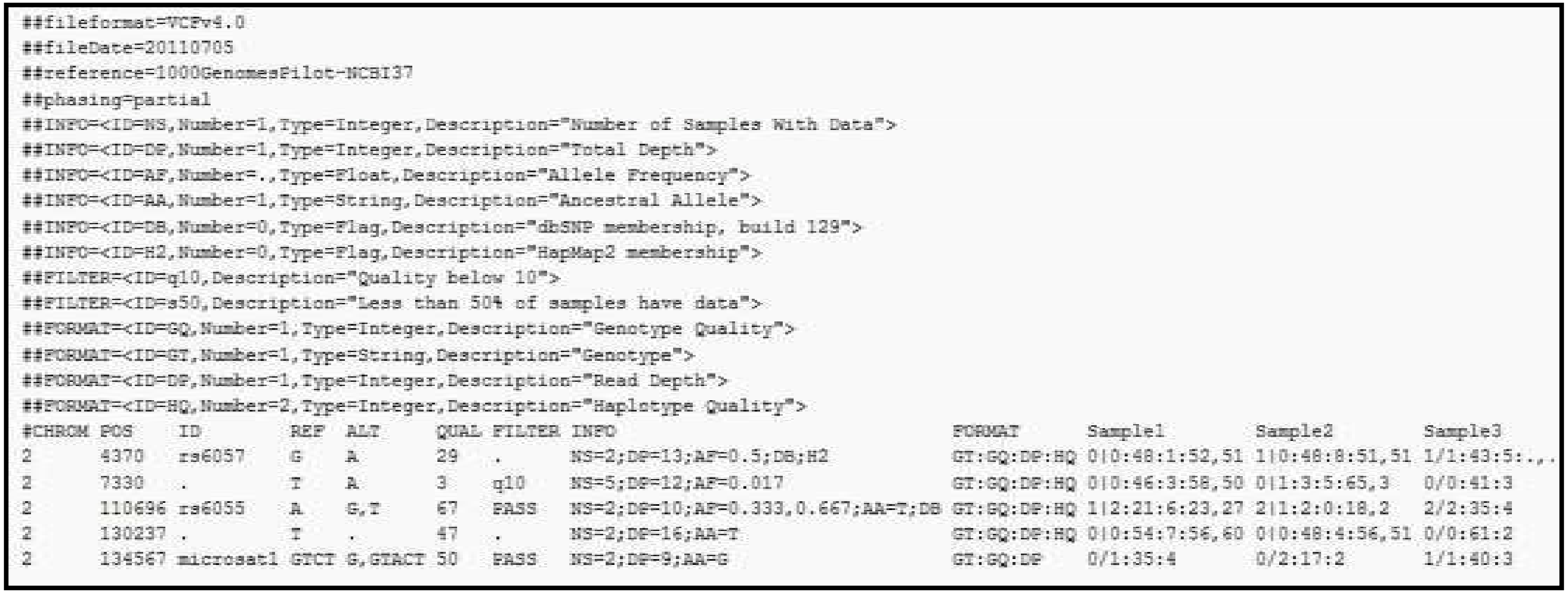
Format VCF (Variant Call Format)

→ Advantages:

Variation description for each position + genotype assignations

Indexed flat files.

Binary files also exist: BCF format



```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS    ID        REF  ALT    QUAL FILTER INFO                              FORMAT      Sample1        Sample2        Sample3
2      4370   rs6057    G    A      29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330   .         T    A      3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696 rs6055    A    G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237 .         T    .      47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567 microsat1 GTCT G,GTACT 50  PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

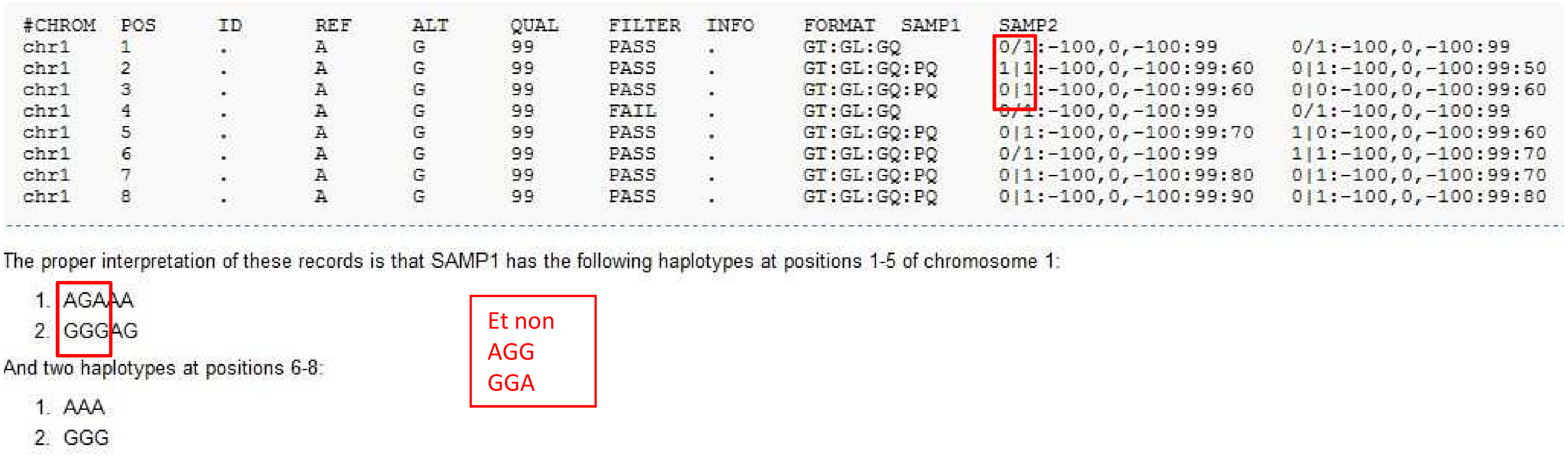
Other GATK functionalities

- Module DepthOfCoverage:

Allows to get sequencing depth for each gene, each position and each individual

- Module ReadBackedPhasing:

Allows to set, if possible, associations between alleles (phase and haplotypes) when we are in an heterozygote situation.



```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMP1   SAMP2
chr1    1       .       A       G       99      PASS    .       GT:GL:GQ        0/1:-100,0,-100:99      0/1:-100,0,-100:99
chr1    2       .       A       G       99      PASS    .       GT:GL:GQ:PQ     1|1:-100,0,-100:99:60   0|1:-100,0,-100:99:50
chr1    3       .       A       G       99      PASS    .       GT:GL:GQ:PQ     0|1:-100,0,-100:99:60   0|0:-100,0,-100:99:60
chr1    4       .       A       G       99      FAIL    .       GT:GL:GQ        0/1:-100,0,-100:99      0/1:-100,0,-100:99
chr1    5       .       A       G       99      PASS    .       GT:GL:GQ:PQ     0|1:-100,0,-100:99:70   1|0:-100,0,-100:99:60
chr1    6       .       A       G       99      PASS    .       GT:GL:GQ:PQ     0/1:-100,0,-100:99      1|1:-100,0,-100:99:70
chr1    7       .       A       G       99      PASS    .       GT:GL:GQ:PQ     0|1:-100,0,-100:99:80   0|1:-100,0,-100:99:70
chr1    8       .       A       G       99      PASS    .       GT:GL:GQ:PQ     0|1:-100,0,-100:99:90   0|1:-100,0,-100:99:80
```

The proper interpretation of these records is that SAMP1 has the following haplotypes at positions 1-5 of chromosome 1:

1. AGAAA
2. GGGAG

And two haplotypes at positions 6-8:

1. AAA
2. GGG

Et non
AGG
GGA

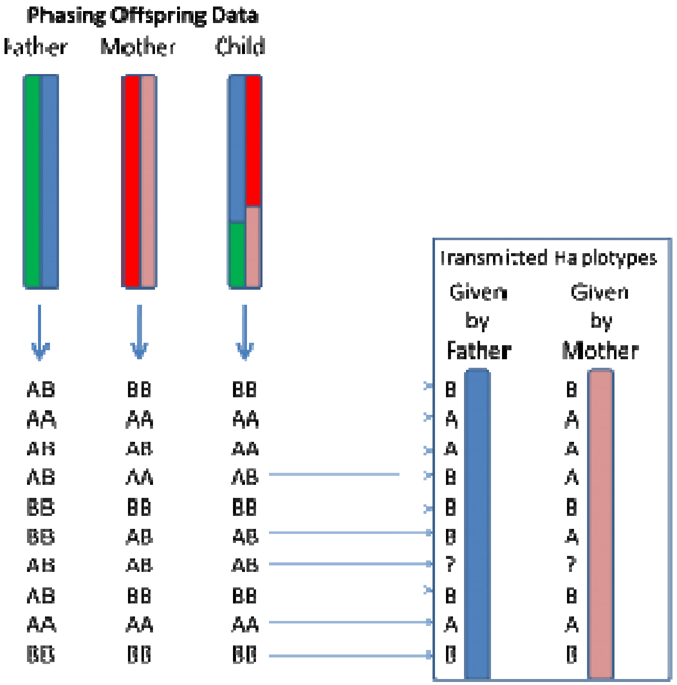
Haplotypes and phasing

- **Haplotype**: Specific groups of genes or alleles that progeny inherited from one parent

- **Phasing**: Determination of haplotype phase. Process of statistical estimation of haplotypes from genotype data.

- Can be infered by statistics methods using non-ambigous haplotypes present in the dataset (Gevalt, ShapeIT, Phase)

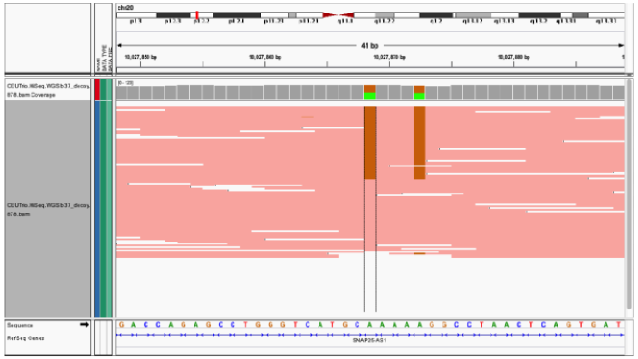
- Can be resolved using physical association of alleles within the reads (GATK ReadBackedPhasing, GATK HaplotypeCaller)



Phasing Offspring Data

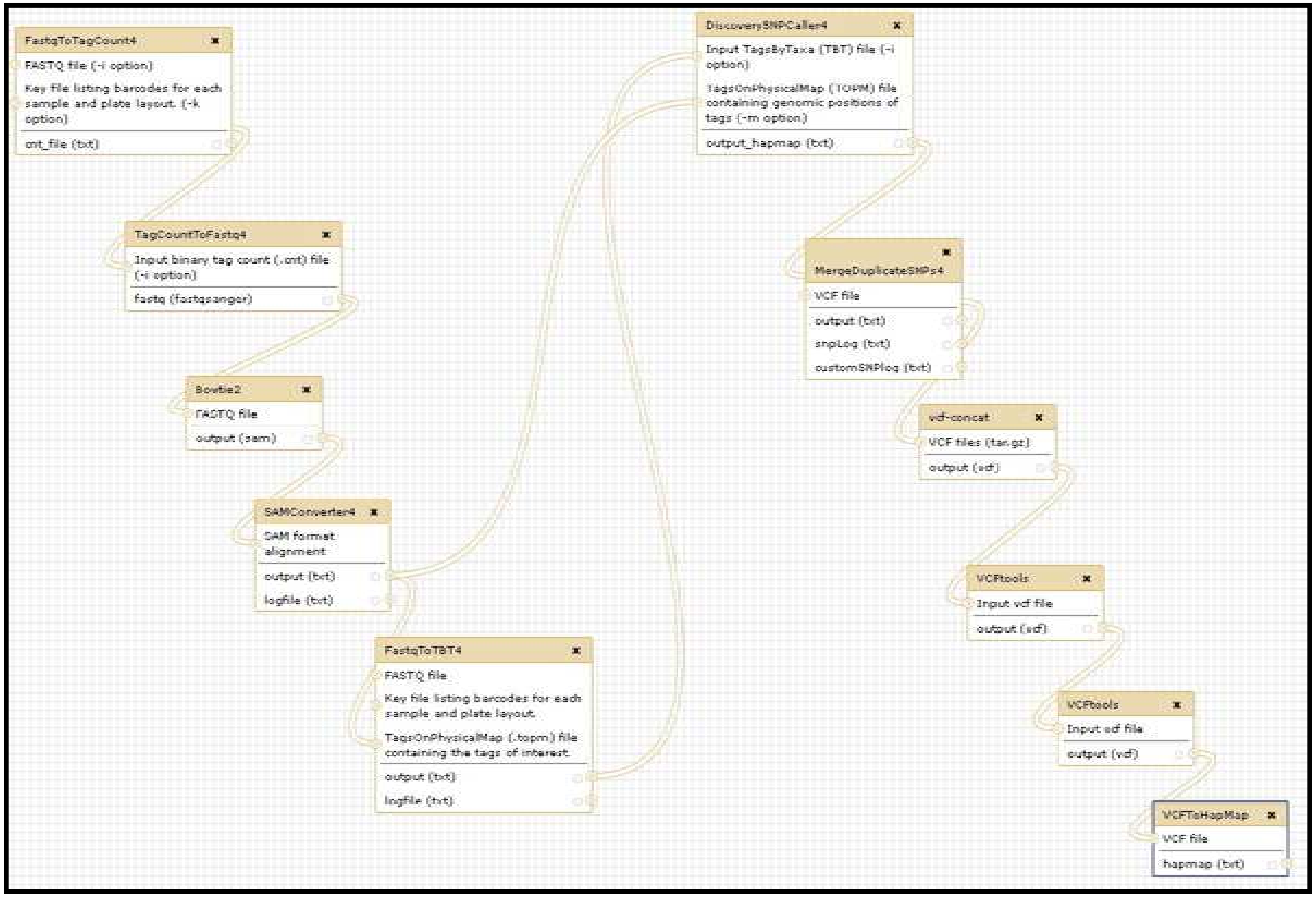
Transmitted Haplotypes

Father	Mother	Child	Given by Father	Given by Mother
AB	BB	BB	B	B
AA	AA	AA	A	A
AB	AB	AA	A	A
AB	AA	AB	B	A
BB	BB	BB	B	B
BB	AB	AB	B	A
AB	AB	AB	?	?
AB	BB	BB	B	B
AA	AA	AA	A	A
BB	BB	BB	B	B


chr20
41 bp
G A C C A G A G C C T G G G T C A T G C A A A A A G G C C T A A C T C A G T G A T

For GBS data

Tassel pipeline
Version 5

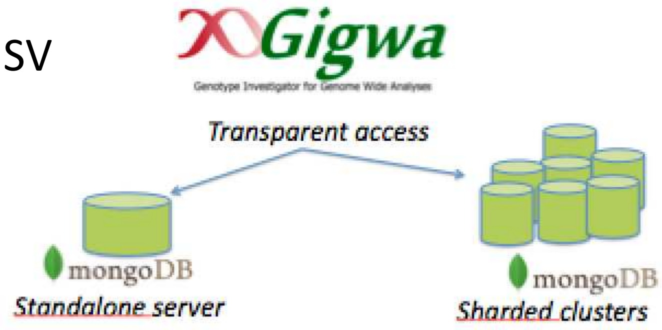


II- SNP data analyses and visualization

Projet Gigwa, pour la gestion des données massives de variants (GBS, RADSeq, WGRS)

> « With NGS arise serious computational challenges in terms of storage, search, sharing, analysis, and data visualization, that redefine some practices in data management. »

- Based on NoSQL technology mongoDB
- Handles VCF files (Variant Call Format) and annotations
- Supports multiple variant types: SNPs, InDels, SSRs, SV
- Powerful genotyping queries
- Easily scalable with MongoDB sharding
- Transparent access



Gigwa

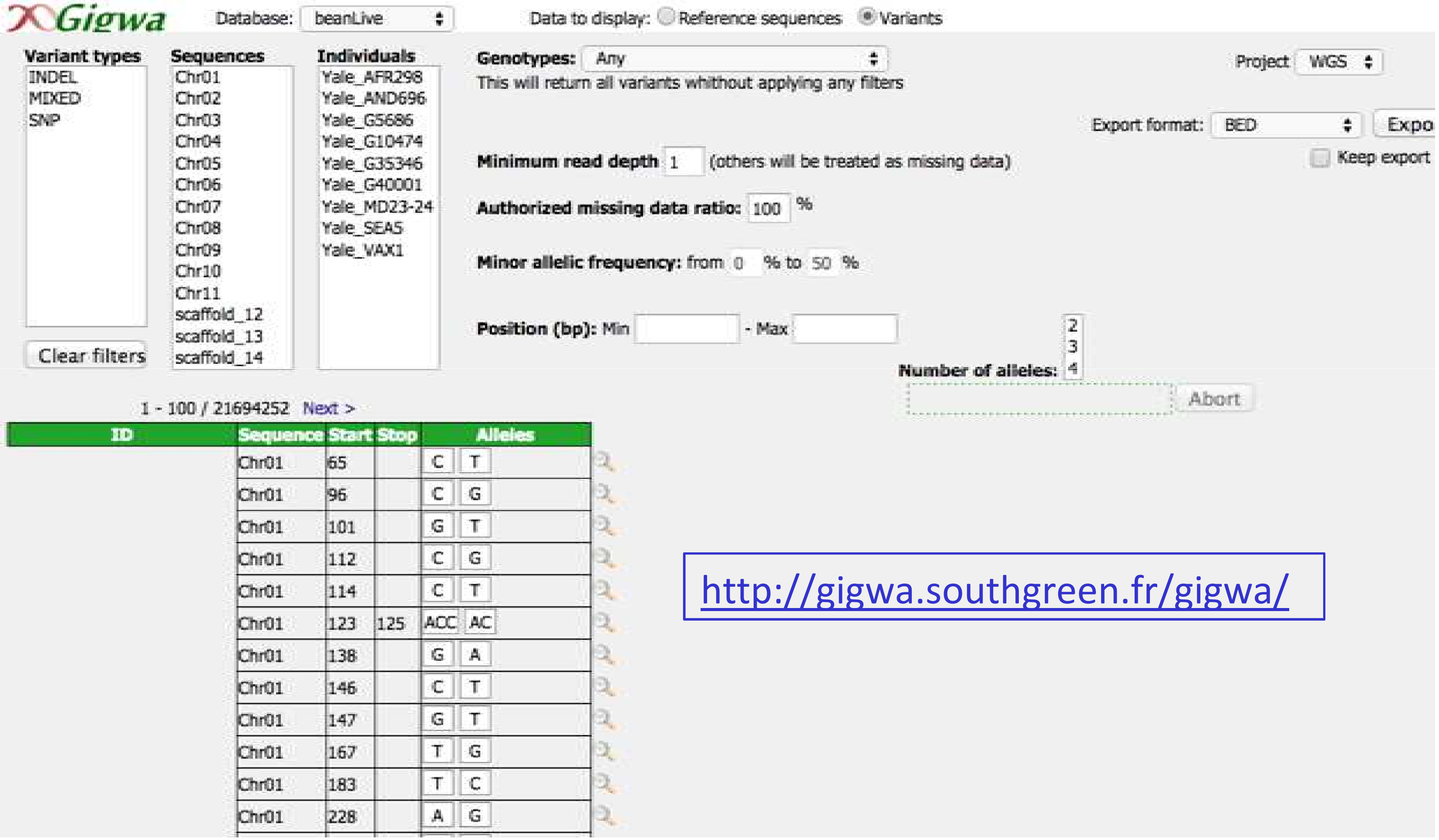
Genotype Investigator for Genome Wide Analyses

Transparent access

mongoDB Standalone server

mongoDB Sharded clusters

- Takes phasing information into account when importing/exporting in VCF format



Gigwa   Database: beanLive   Data to display: ○ Reference sequences ● Variants

Variant types: INDEL, MIXED, SNP

Sequences: Chr01, Chr02, Chr03, Chr04, Chr05, Chr06, Chr07, Chr08, Chr09, Chr10, Chr11, scaffold_12, scaffold_13, scaffold_14

Individuals: Yale_AFR298, Yale_AND696, Yale_G5686, Yale_G10474, Yale_G35346, Yale_G40001, Yale_MD23-24, Yale_SEA5, Yale_VAX1

Clear filters

Genotypes: Any
This will return all variants whithout applying any filters

Project WGS

Export format: BED   Expo

□ Keep export

Minimum read depth 1 (others will be treated as missing data)

Authorized missing data ratio: 100 %

Minor allelic frequency: from 0 % to 50 %

Position (bp): Min - Max

Number of alleles: 2 3 4

Abort

1 - 100 / 21694252 Next >

ID	Sequence	Start	Stop	Alleles	
	Chr01	65		C	T
	Chr01	96		C	G
	Chr01	101		G	T
	Chr01	112		C	G
	Chr01	114		C	T
	Chr01	123	125	ACC	AC
	Chr01	138		G	A
	Chr01	146		C	T
	Chr01	147		G	T
	Chr01	167		T	G
	Chr01	183		T	C
	Chr01	228		A	G

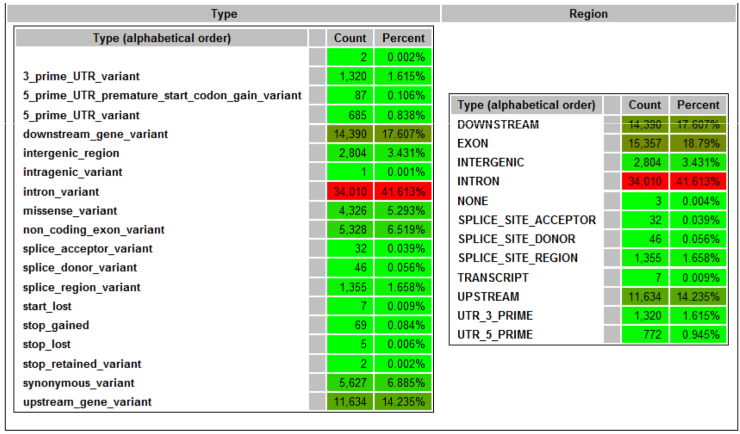
http://gigwa.southgreen.fr/gigwa/

SNP annotation using SnpEff

SnpEff

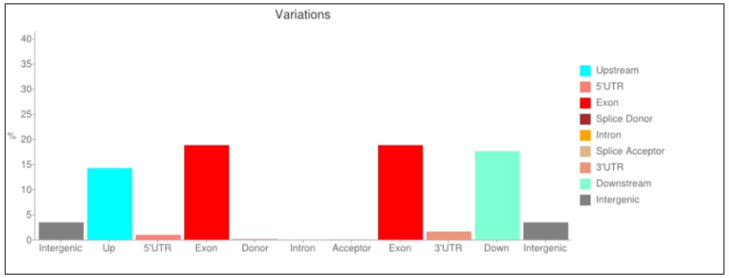
Genetic variant annotation and effect prediction toolbox.

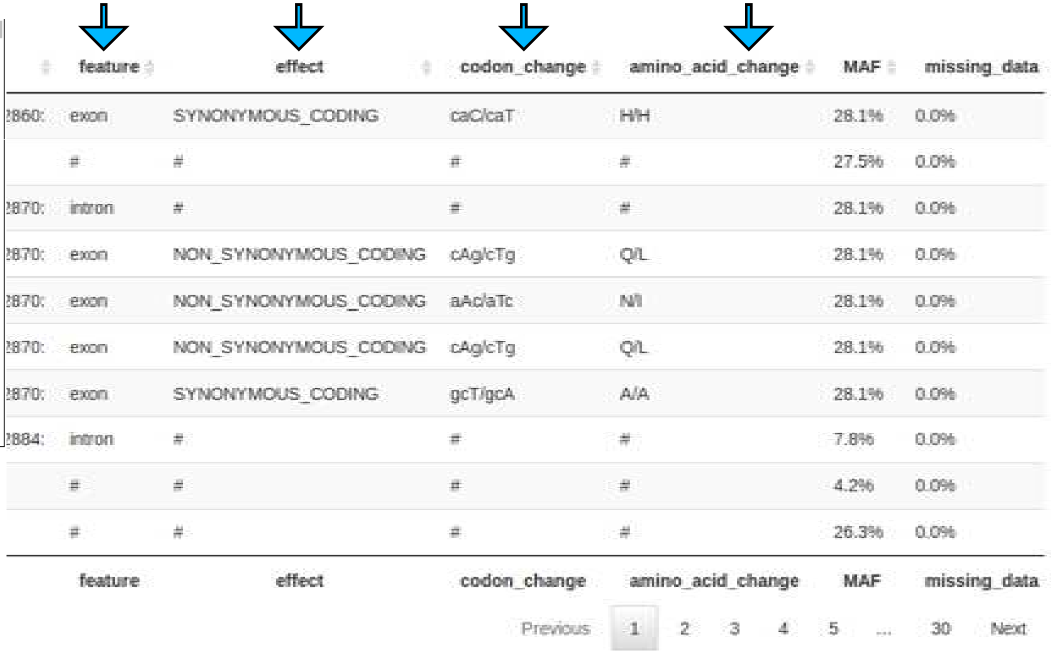
- It annotates and predicts the effects of variants on genes (amino acid changes…)
- Uses as input GFF annoation file and VCF



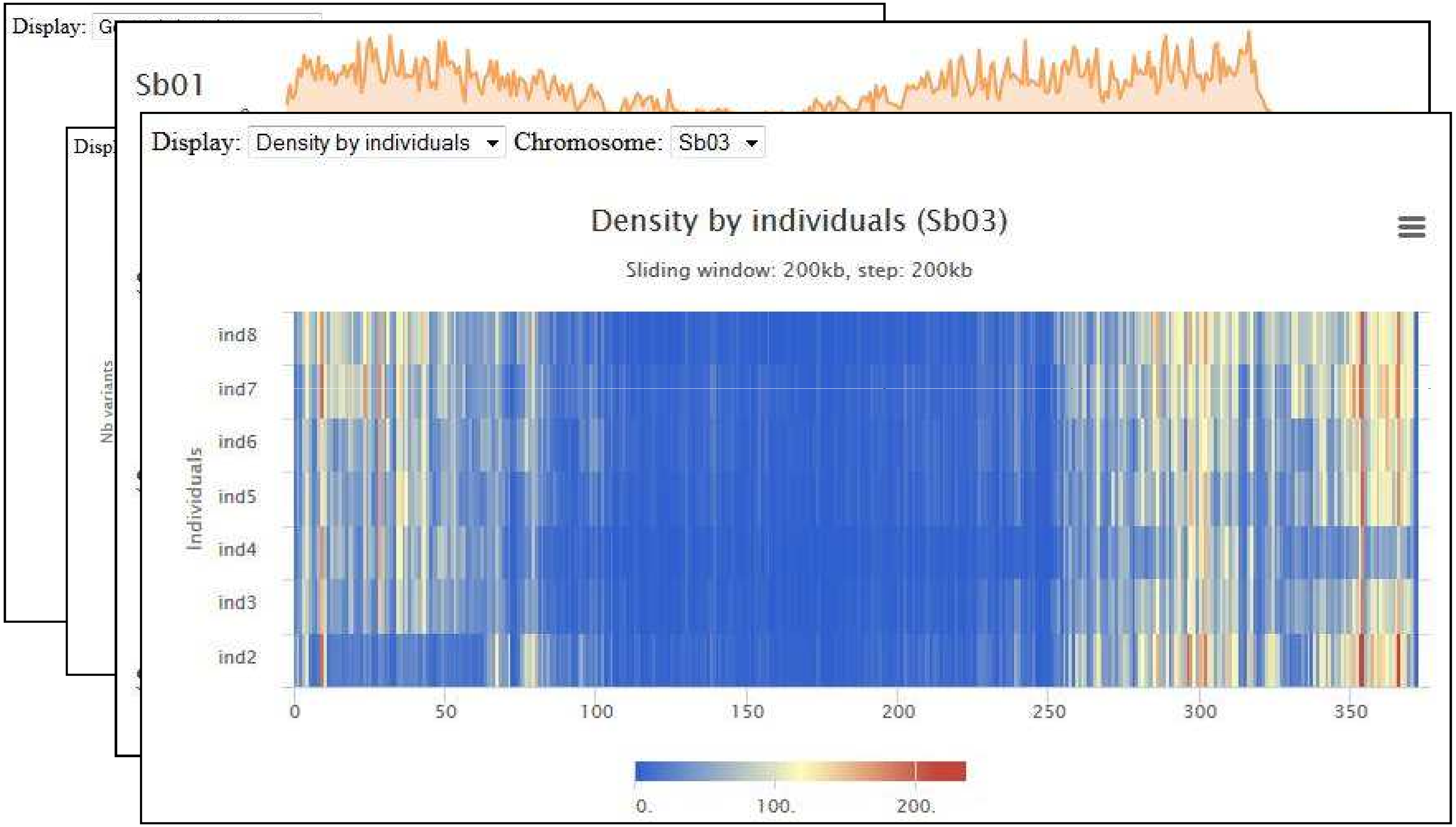
Type			
Type (alphabetical order)		Count	Percent
		2	0.002%
3_prime_UTR_variant		1,320	1.615%
5_prime_UTR_premature_start_codon_gain_variant		87	0.106%
5_prime_UTR_variant		685	0.838%
downstream_gene_variant		14,390	17.607%
intergenic_region		2,804	3.431%
intragenic_variant		1	0.001%
intron_variant		34,010	41.613%
missense_variant		4,326	5.293%
non_coding_exon_variant		5,328	6.519%
splice_acceptor_variant		32	0.039%
splice_donor_variant		46	0.056%
splice_region_variant		1,355	1.658%
start_lost		7	0.009%
stop_gained		69	0.084%
stop_lost		5	0.006%
stop_retained_variant		2	0.002%
synonymous_variant		5,627	6.885%
upstream_gene_variant		11,634	14.235%

Region			
Type (alphabetical order)		Count	Percent
DOWNSTREAM		14,390	17.607%
EXON		15,357	18.79%
INTERGENIC		2,804	3.431%
INTRON		34,010	41.613%
NONE		3	0.004%
SPLICE_SITE_ACCEPTOR		32	0.039%
SPLICE_SITE_DONOR		46	0.056%
SPLICE_SITE_REGION		1,355	1.658%
TRANSCRIPT		7	0.009%
UPSTREAM		11,634	14.235%
UTR_3_PRIME		1,320	1.615%
UTR_5_PRIME		772	0.945%





	feature	effect	codon_change	amino_acid_change	MAF	missing_data
2860:	exon	SYNONYMOUS_CODING	caC/caT	H/H	28.1%	0.0%
	#	#	#	#	27.5%	0.0%
2870:	intron	#	#	#	28.1%	0.0%
2870:	exon	NON_SYNONYMOUS_CODING	cAg/cTg	Q/L	28.1%	0.0%
2870:	exon	NON_SYNONYMOUS_CODING	aAc/aTc	N/I	28.1%	0.0%
2870:	exon	NON_SYNONYMOUS_CODING	cAg/cTg	Q/L	28.1%	0.0%
2870:	exon	SYNONYMOUS_CODING	gcT/gcA	A/A	28.1%	0.0%
2884:	intron	#	#	#	7.8%	0.0%
	#	#	#	#	4.2%	0.0%
	#	#	#	#	26.3%	0.0%
	feature	effect	codon_change	amino_acid_change	MAF	missing_data

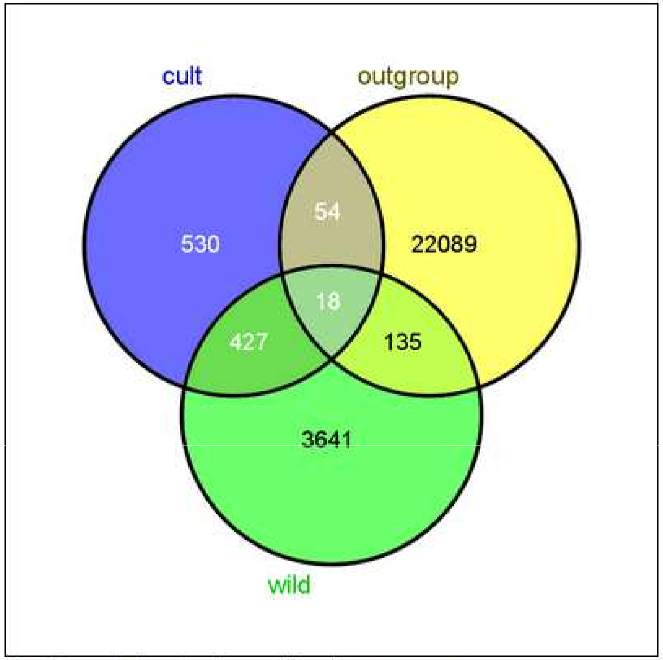


Display: Density by individuals

Chromosome: Sb03

Density by individuals (Sb03)

Sliding window: 200kb, step: 200kb



Specific and shared polymorphisms between groups

+ 2186 polymorphisms inter-group

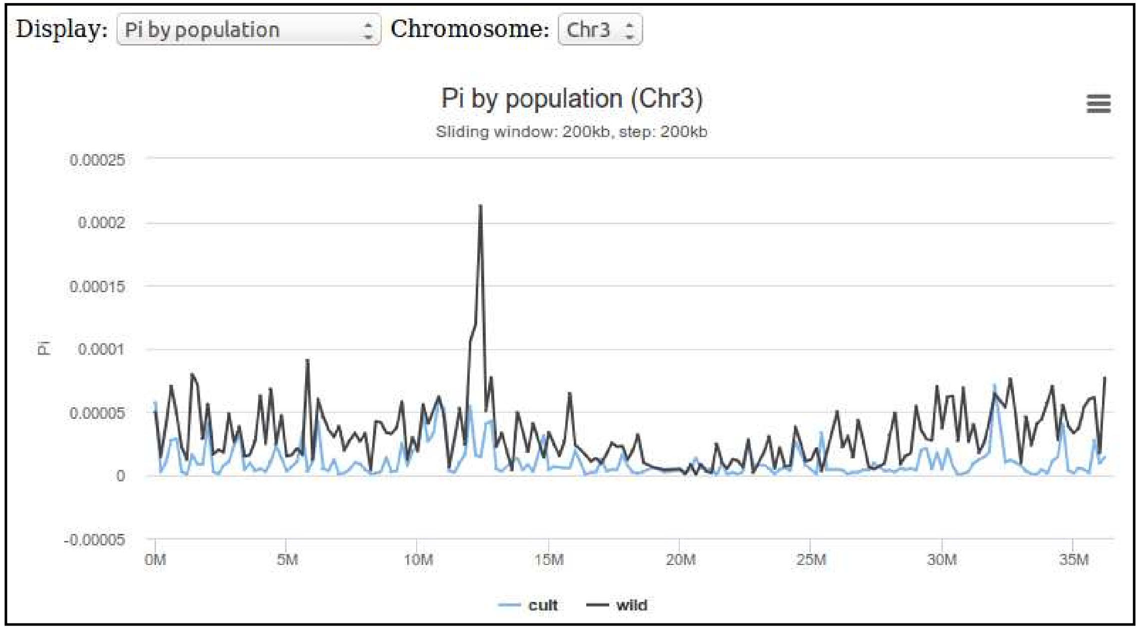
Comparison between individuals

Fst: Fixation index: measure of population differentiation due to genetic structure.

Pi: Nucleotide diversity: Average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population

Used to measure the degree of polymorphism within a population

Diversity analysis



Display: Pi by population Chromosome: Chr3

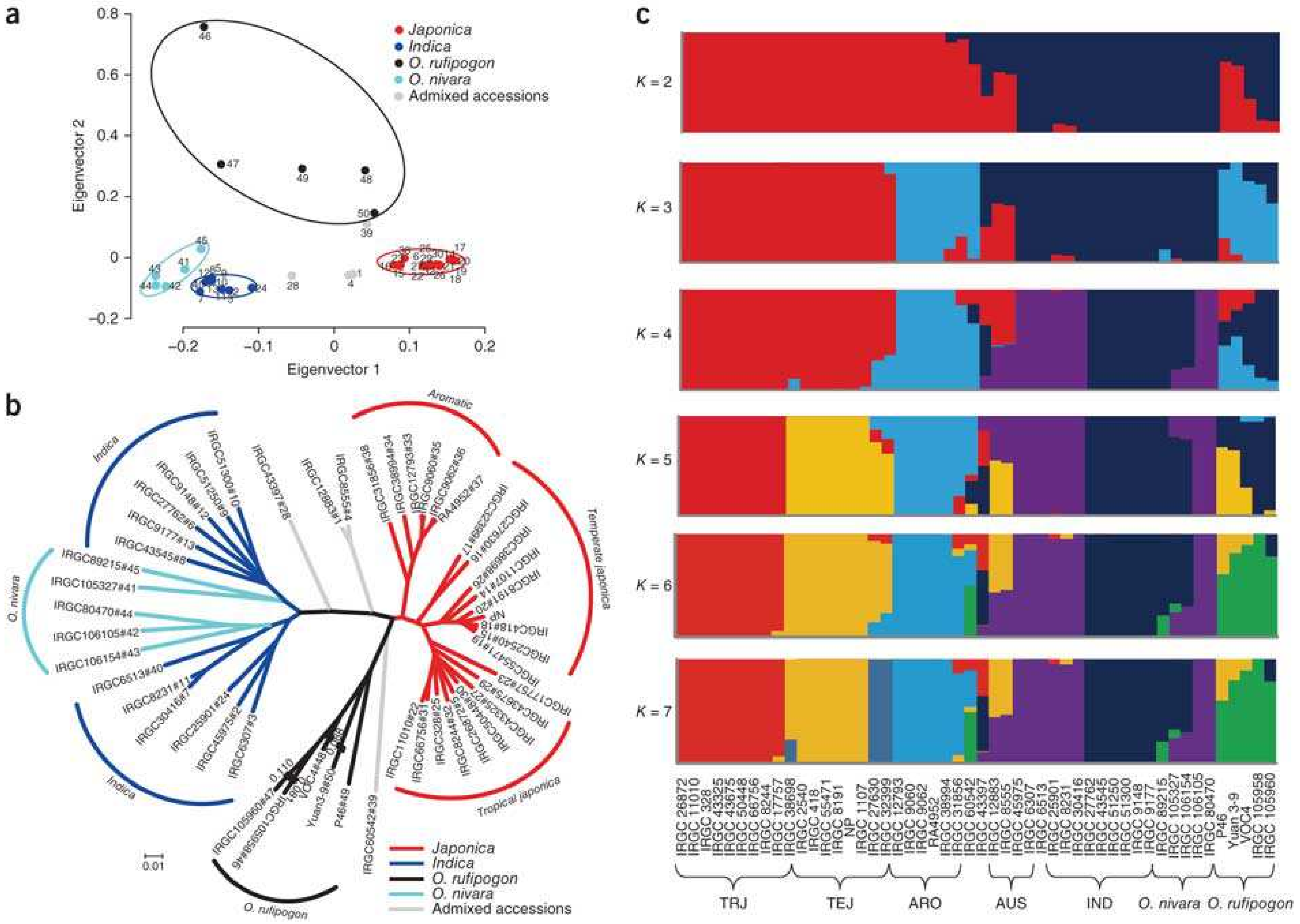
SNP density by individuals can allow the detection of introgression event.

Introgression = Movement of a exogene region (gene flow) from one species into the gene pool of another by the repeated backcrossing of an interspecific hybrid with one of its parent species

Widely used in agronomy obtained but can occurs naturally

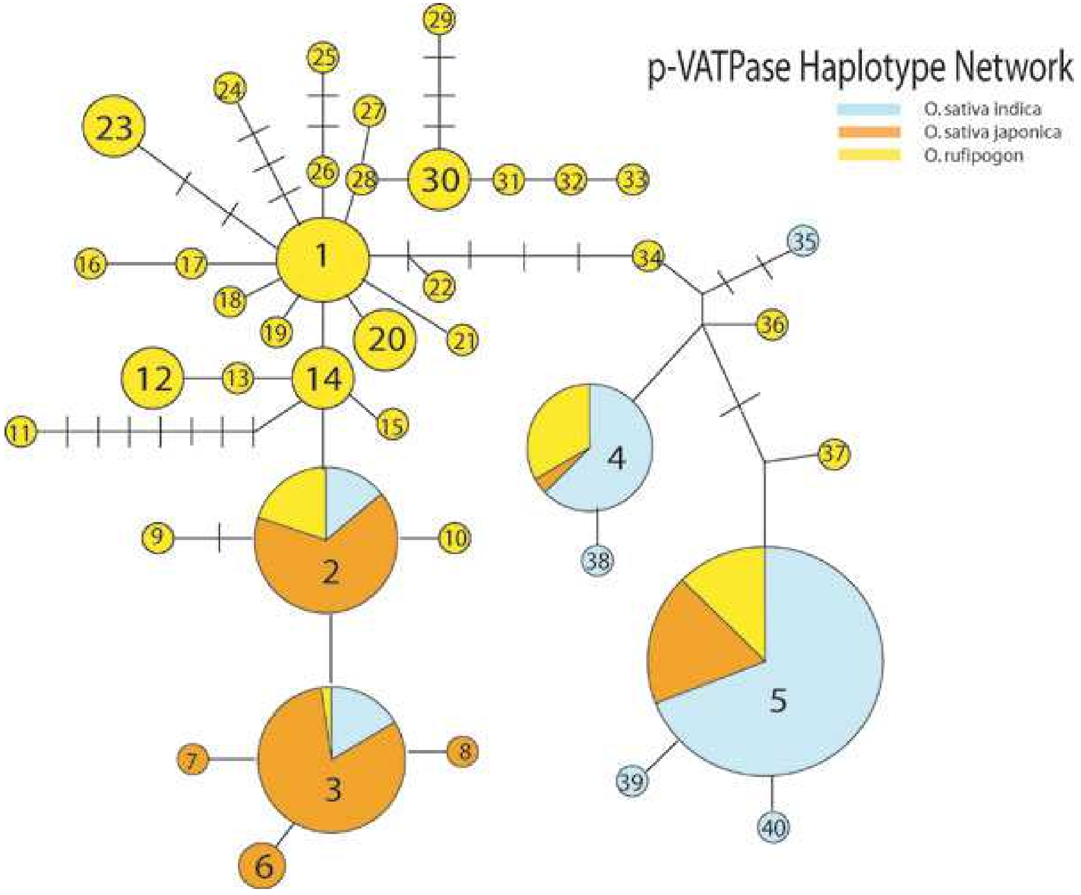
Population structure

Ex: Riz asiatique après re-séquençage (Xun et al, Nature, 2011)



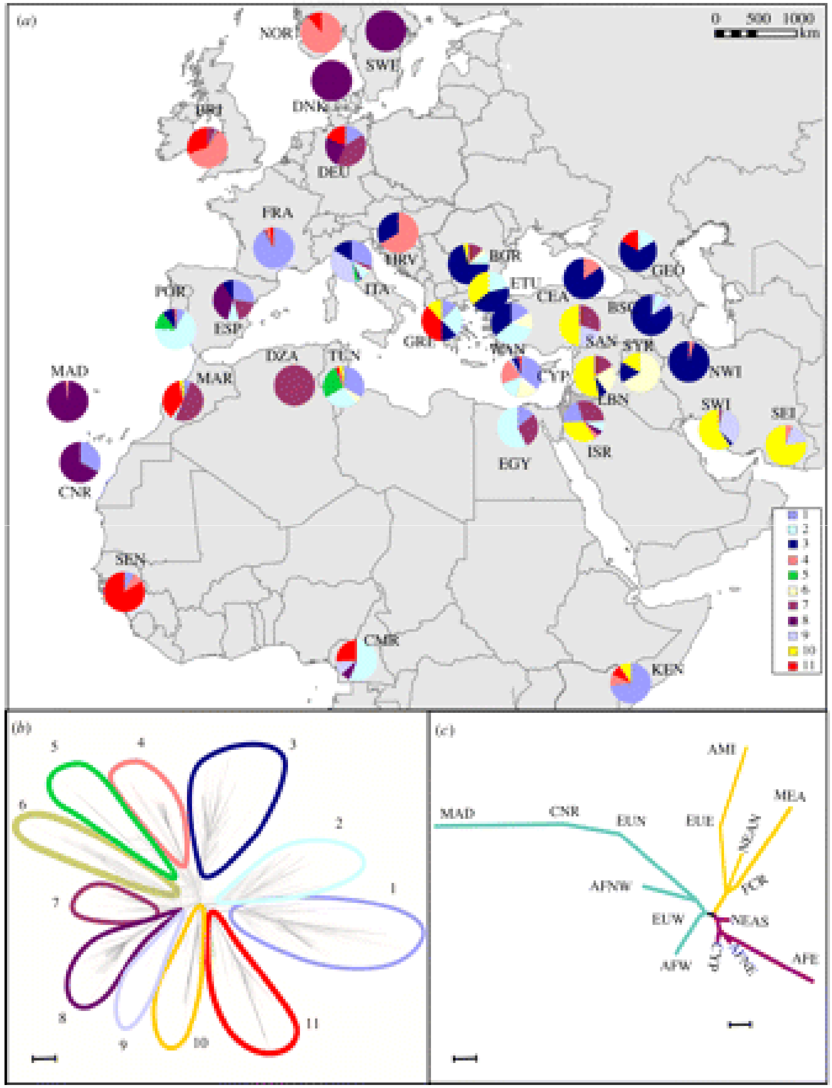
Haplotype network

Exemple d’une région génomique chez le Riz

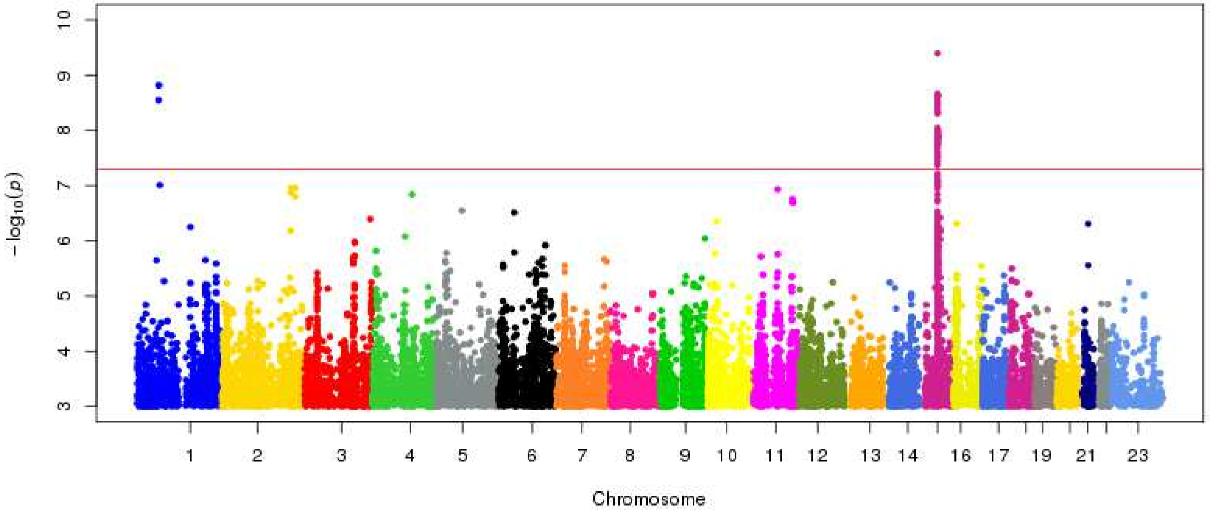


Haplotype and geographical distribution

Différenciation génétique de la souris domestique (Bonhomme et al, 2010)



GWAS (Genome-Wide Association Studies)



- Estimate association between a marker and a phenotypic character

- Manhattan plots: displays GWAS statistical tests (-log10 pvalue) along chromosomes

- TASSEL, MLMM sofwares

- False positives because of the studied structuration panel

=> correction using structure population et and kinship

GWAS issues

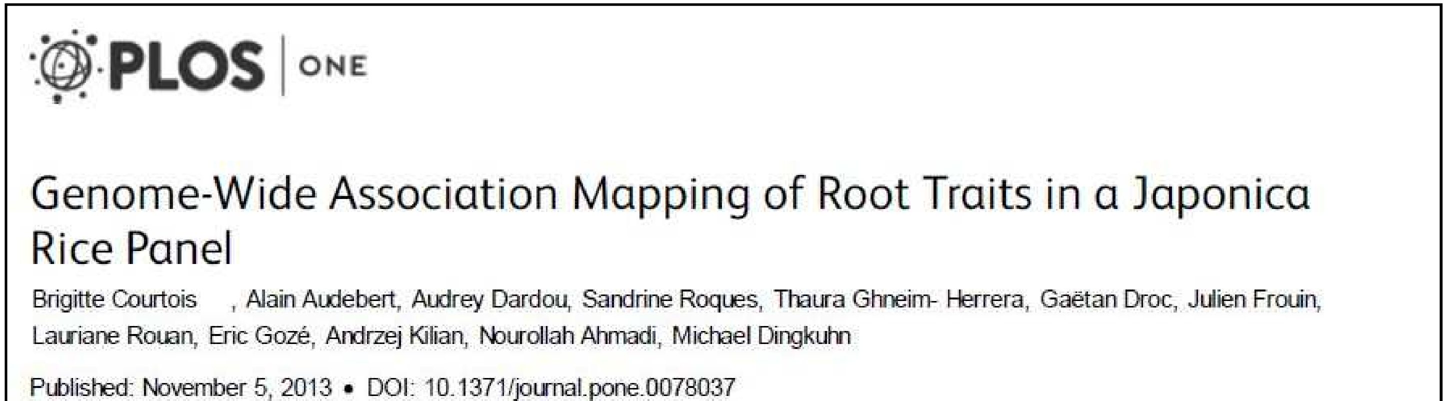
- **Choice of genotypic panel**: phenotypic diversity for target traits must be sufficient (core-collection, MAGIC lines, NAM…)

- **Population structure** induces high rates of false associations (false positives)
- Correction using structure population et and kinship. Mixed models:
    - Q
    - K (widely used)
    - Q+K (widely used)

- **Density of markers** must be enough to provide a good genome cover. Density can be also highly variable.

- **Linkage disequilibrium (LD) landscape**: level of intra- and inter-chromosomal LD (number of loci in LD with loci from other chromosomes). Ideally, LD profile must be flat to avoid distorsion in association patterns.

Study of root characters using GWAS in Oryza sativa japonica. Influence of a correction using structure and kinship

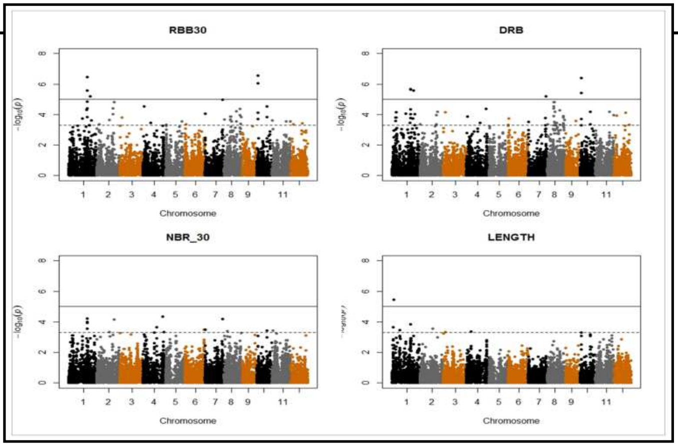


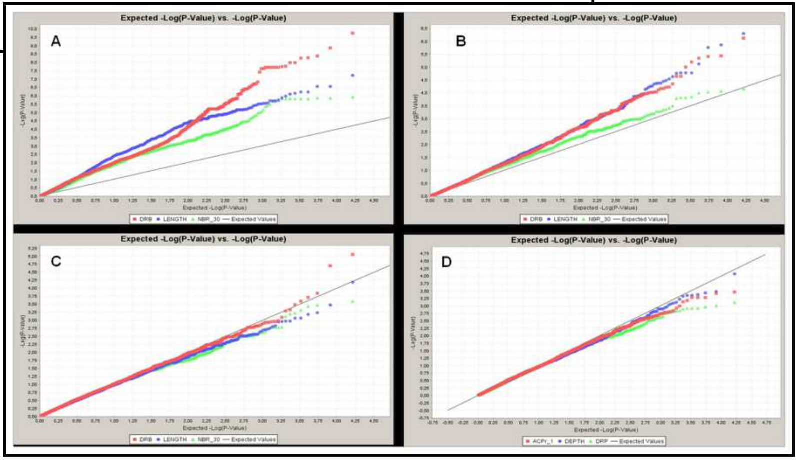
PLOS | ONE

Genome-Wide Association Mapping of Root Traits in a Japonica Rice Panel

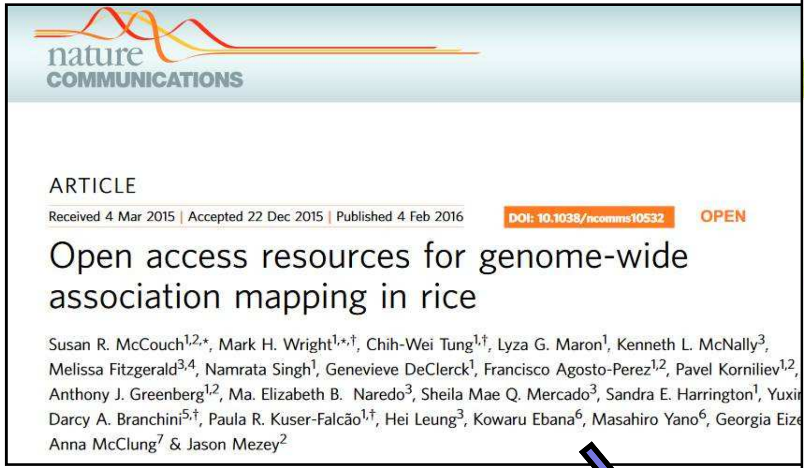
Brigitte Courtois , Alain Audebert, Audrey Dardou, Sandrine Roques, Thaura Ghneim- Herrera, Gaëtan Droc, Julien Frouin, Lauriane Rouan, Eric Gozé, Andrzej Kilian, Nourollah Ahmadi, Michael Dingkuhn

Published: November 5, 2013 • DOI: 10.1371/journal.pone.0078037





Exemple du Riz: 3000 genomes+ HDRA (High density Rice Array)



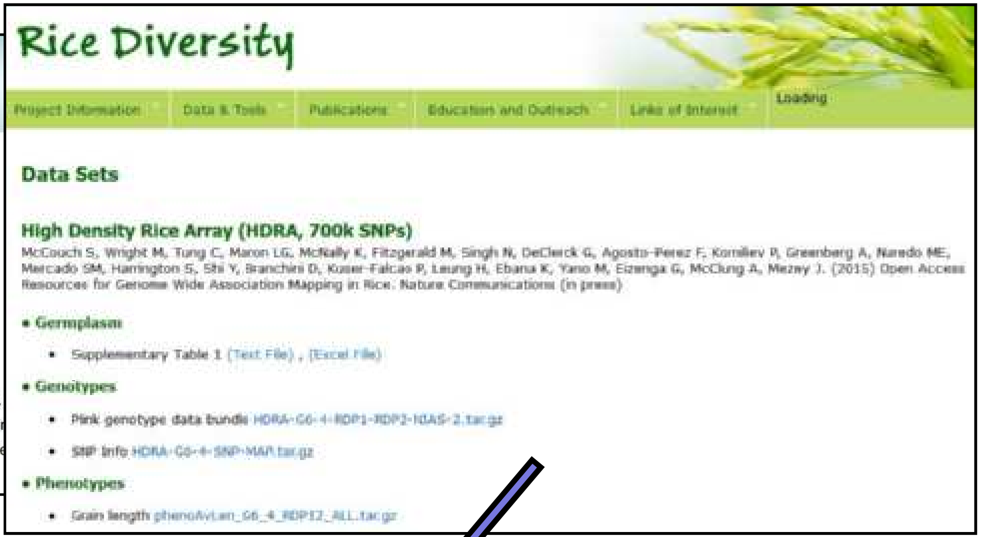
nature COMMUNICATIONS

ARTICLE

Received 4 Mar 2015 | Accepted 22 Dec 2015 | Published 4 Feb 2016 DOI: 10.1038/ncomms10532 OPEN

Open access resources for genome-wide association mapping in rice

Susan R. McCouch[1,2,*], Mark H. Wright[1,*,†], Chih-Wei Tung[1,†], Lyza G. Maron[1], Kenneth L. McNally[3], Melissa Fitzgerald[3,4], Namrata Singh[1], Genevieve DeClerck[1], Francisco Agosto-Perez[1,2], Pavel Korniliev[1,2], Anthony J. Greenberg[1,2], Ma. Elizabeth B. Naredo[3], Sheila Mae Q. Mercado[3], Sandra E. Harrington[1], Yuxin Darcy A. Branchini[5,†], Paula R. Kuser-Falcão[1,†], Hei Leung[3], Kowaru Ebana[6], Masahiro Yano[6], Georgia Eize Anna McClung[7] & Jason Mezey[2]



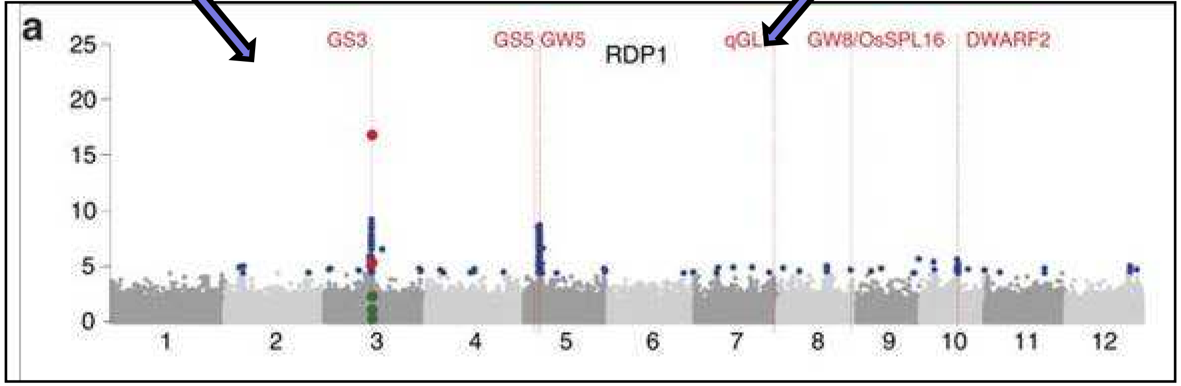
Rice Diversity

Data Sets

High Density Rice Array (HDRA, 700k SNPs)

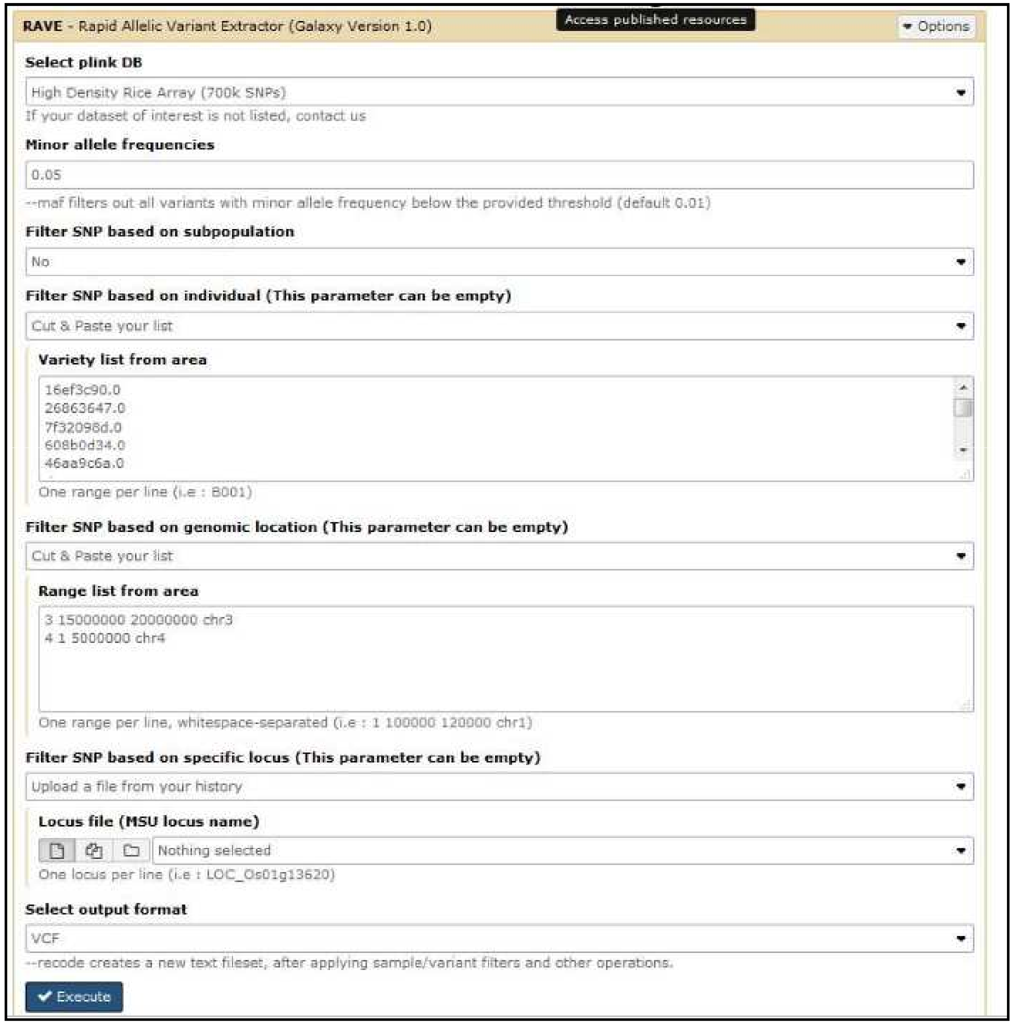
McCouch S, Wright M, Tung C, Maron LG, McNally K, Fitzgerald M, Singh N, DeClerck G, Agosto-Perez F, Korniliev P, Greenberg A, Naredo ME, Mercado SM, Harrington S, Shi Y, Branchini D, Kuser-Falcao P, Leung H, Ebana K, Yano M, Eizenga G, McClung A, Mezey J. (2015) Open Access Resources for Genome Wide Association Mapping in Rice. Nature Communications (in press)

- Germplasm
  - Supplementary Table 1 (Text File) , (Excel File)
- Genotypes
  - Pink genotype data bundle HDRA-G6-4-RDP1-RDP2-NIAS-2.tar.gz
  - SNP info HDRA-G6-4-SNP-MAP.tar.gz
- Phenotypes
  - Grain length phenoAvLen_G6_4_RDP12_ALL.tar.gz



Exemple du Riz: 3000 genomes+ HDRA (High density Rice Array)

Extraction rapide des variants après sélection d'une region / population donnée.



RAVE - Rapid Allelic Variant Extractor (Galaxy Version 1.0)

Select plink DB

High Density Rice Array (700k SNPs) ▾

If your dataset of interest is not listed, contact us

Minor allele frequencies

0.05

--maf filters out all variants with minor allele frequency below the provided threshold (default 0.01)

Filter SNP based on subpopulation

No ▾

Filter SNP based on individual (This parameter can be empty)

Cut & Paste your list ▾

Variety list from area

```
16ef3c90.0
26863647.0
7f32098d.0
608b0d34.0
46aa9c6a.0
```

One range per line (i.e : B001)

Filter SNP based on genomic location (This parameter can be empty)

Cut & Paste your list ▾

Range list from area

```
3 15000000 20000000 chr3
4 1 5000000 chr4
```

One range per line, whitespace-separated (i.e : 1 100000 120000 chr1)

Filter SNP based on specific locus (This parameter can be empty)

Upload a file from your history ▾

Locus file (MSU locus name)

Nothing selected ▾

One locus per line (i.e : LOC_Os01g13620)

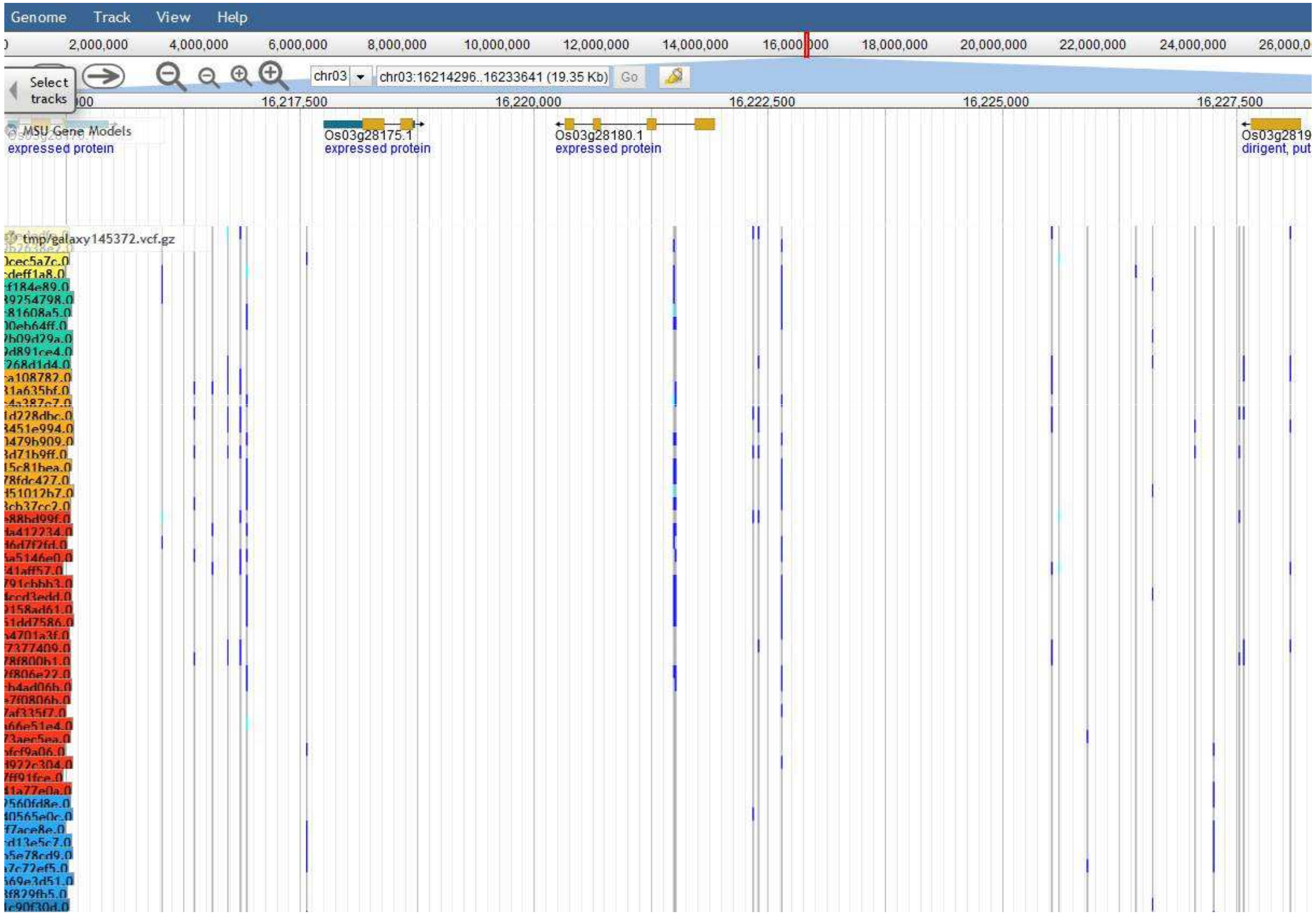
Select output format

VCF ▾

--recode creates a new text fileset, after applying sample/variant filters and other operations.

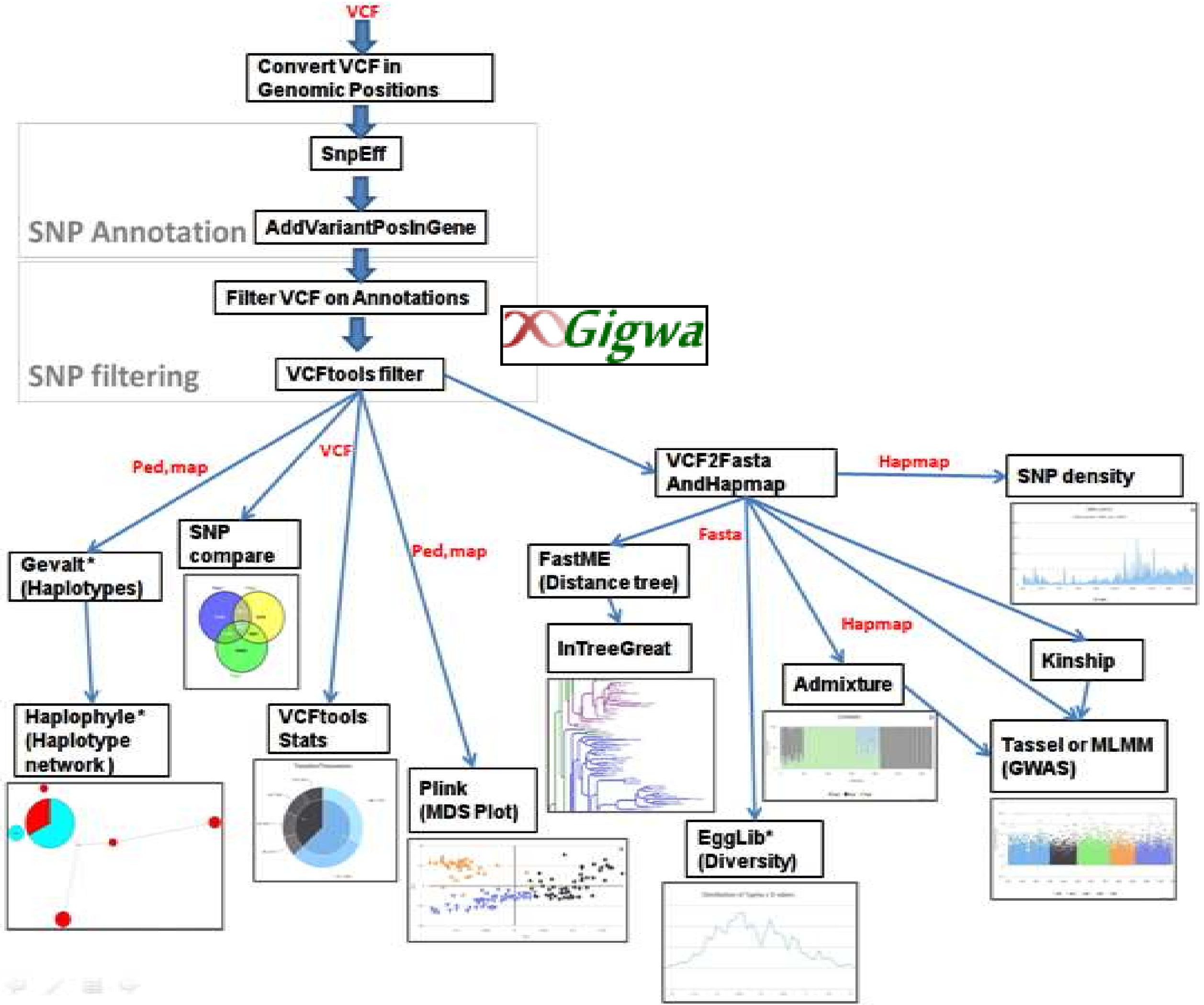
Exemple du Riz: 3000 genomes+ HDRA (High density Rice Array)

Visualisation du contexte génomique dans un génome browser (plugin Jbrowse)



SNiPlay: Web application for polymorphism analyses

http://sniplay.southgreen.fr



VCF

Convert VCF in Genomic Positions

SNP Annotation

SnpEff

AddVariantPosInGene

SNP filtering

Filter VCF on Annotations

VCFtools filter

Gigwa

Ped, map

VCF

Gevalt* (Haplotypes)

Haplophyle* (Haplotype network)

SNP compare

VCFtools Stats

Ped, map

Plink (MDS Plot)

VCF2Fasta AndHapmap

Hapmap

SNP density

Fasta

FastME (Distance tree)

InTreeGreat

EggLib* (Diversity)

Hapmap

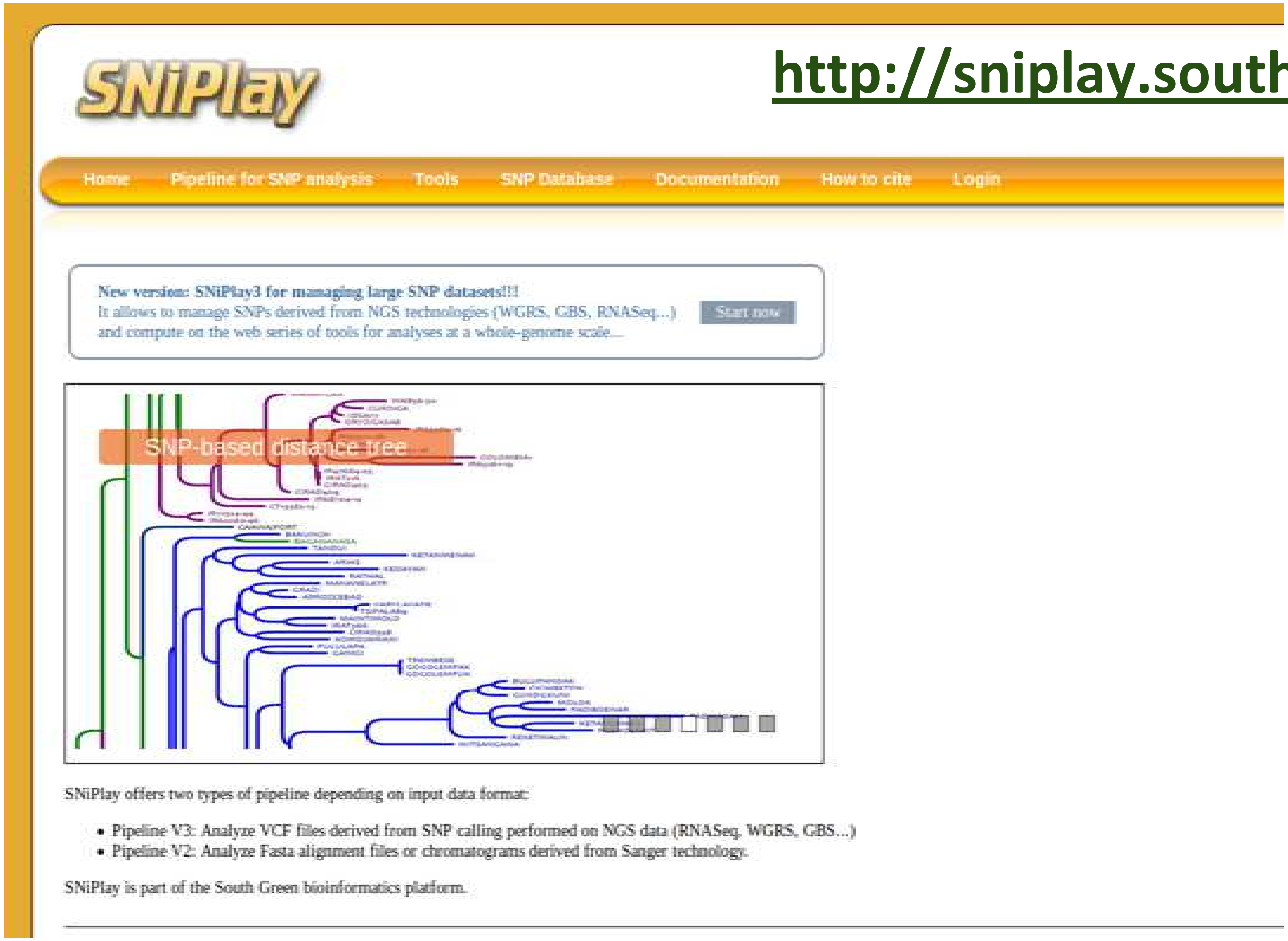
Admixture

Kinship

Tassel or MLMM (GWAS)

SNiPlay Site web

http://sniplay.southgreen.fr



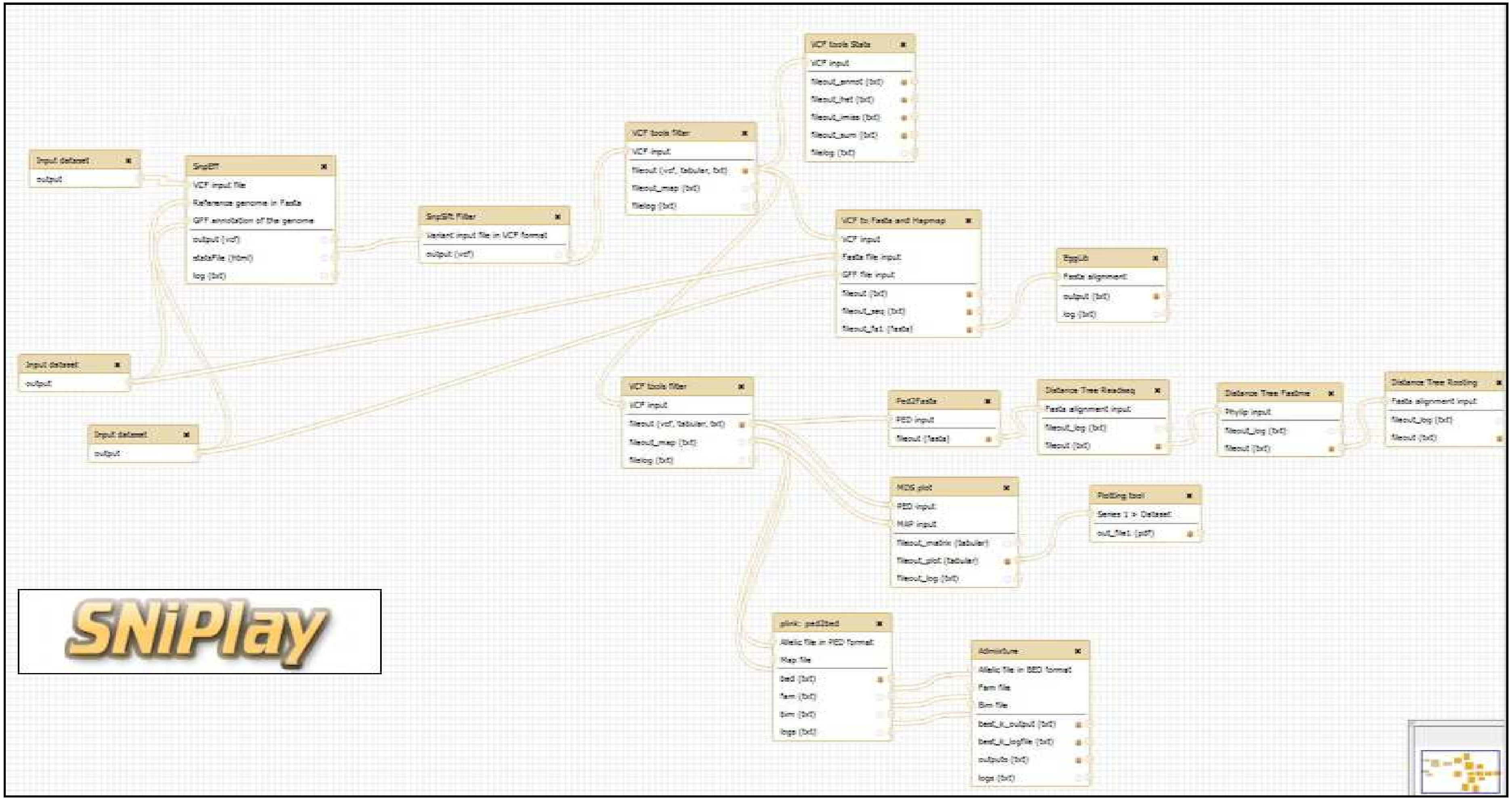
New version: SNiPlay3 for managing large SNP datasets!!!
It allows to manage SNPs derived from NGS technologies (WGRS, GBS, RNASeq...) and compute on the web series of tools for analyses at a whole-genome scale...

SNiPlay offers two types of pipeline depending on input data format:

- Pipeline V3: Analyze VCF files derived from SNP calling performed on NGS data (RNASeq, WGRS, GBS...)
- Pipeline V2: Analyze Fasta alignment files or chromatograms derived from Sanger technology.

SNiPlay is part of the South Green bioinformatics platform.

“Galaxy4Sniplay” : SNiPlay sous Galaxy


Input dataset
output
SnpEff
VCF input file
Reference genome in Fasta
GFF annotation of the genome
output (vcf)
statsFile (html)
log (txt)
SnpSift Filter
Variant input file in VCF format
output (vcf)
VCF tools filter
VCF input
fileout (vcf, tabular, txt)
fileout_map (txt)
filelog (txt)
VCF tools Stats
VCF input
fileout_annot (txt)
fileout_het (txt)
fileout_imiss (txt)
fileout_sum (txt)
filelog (txt)
VCF to Fasta and Hapmap
VCF input
Fasta file input
GFF file input
fileout (txt)
fileout_seq (txt)
fileout_fa1 (fasta)
EggLib
Fasta alignment
output (txt)
log (txt)
Input dataset
output
Input dataset
output
VCF tools filter
VCF input
fileout (vcf, tabular, txt)
fileout_map (txt)
filelog (txt)
Ped2Fasta
PED input
fileout (fasta)
Distance Tree Reading
Fasta alignment input
fileout_log (txt)
fileout (txt)
Distance Tree Fastme
Phylip input
fileout_log (txt)
fileout (txt)
Distance Tree Rooting
Fasta alignment input
fileout_log (txt)
fileout (txt)
MDS plot
PED input
MAP input
fileout_matrix (tabular)
fileout_plot (tabular)
fileout_log (txt)
Plotting tool
Series 1 > Dataset
out_file1 (pdf)
plink: ped2bed
Allelic file in PED format
Map file
bed (txt)
fam (txt)
bim (txt)
logs (txt)
Admixture
Allelic file in BED format
Fam file
Bim file
best_k_output (txt)
best_k_logfile (txt)
outputs (txt)
logs (txt)
SNiPlay