Analyse de variants génétiques (SNPs, indels)

Gautier Sarah - Alexis Dereeper – Christine Tranchant

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

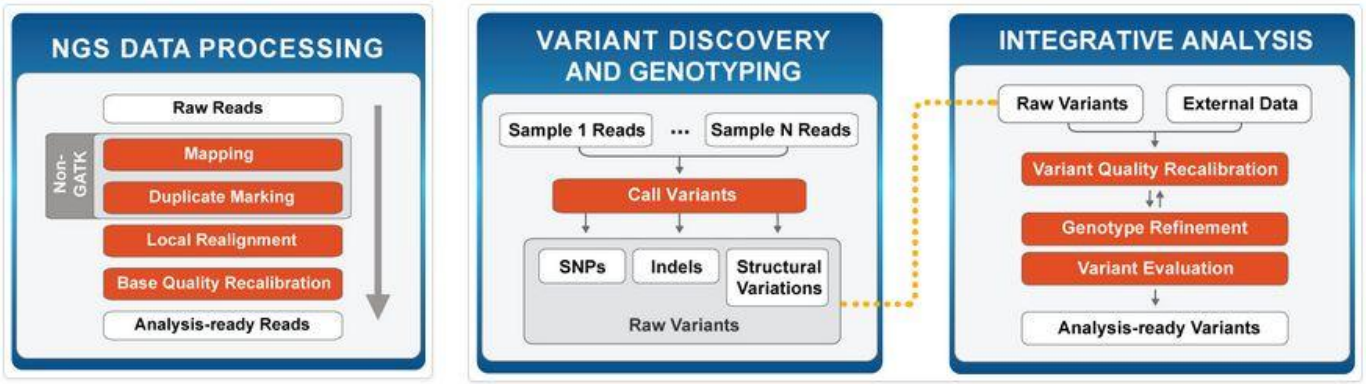
Mapping BWA

Add or Replace Groups

BAM with read group

mergeSam

Global BAM with read group



VCF file

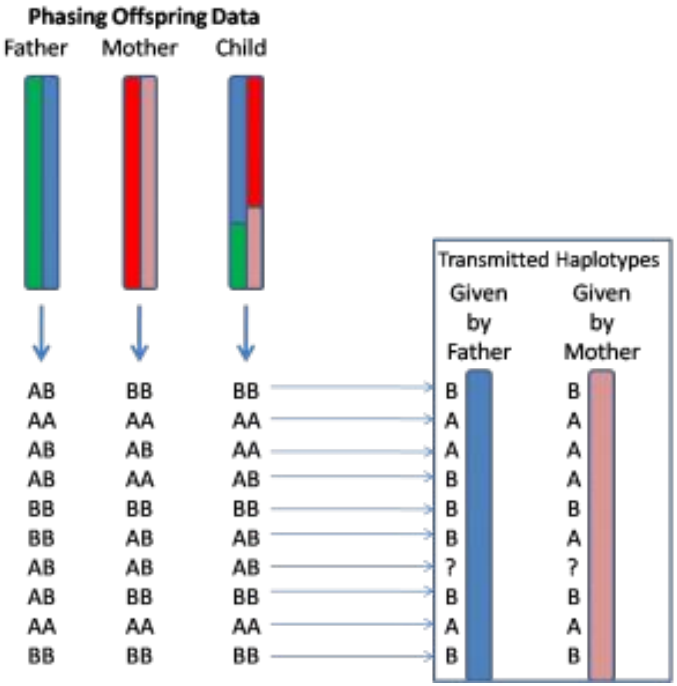
Haplotypes and phasing

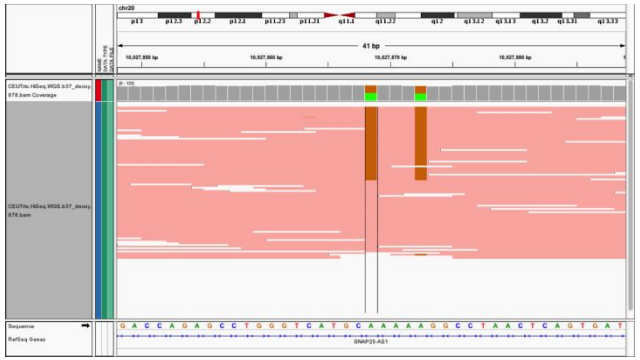
- **Haplotype**: Specific groups of genes or alleles that progeny inherited from one parent

- **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)





SNP annotation using SnpEff

SnpEff

Genetic variant annotation and effect prediction toolbox.

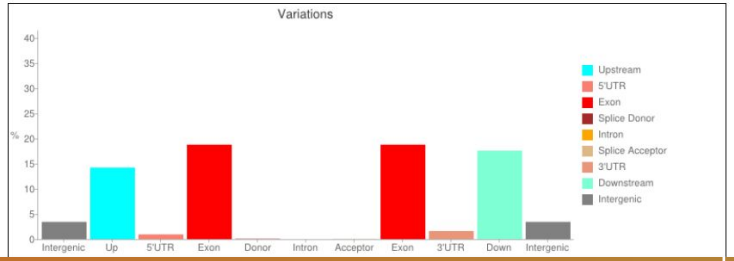
- It annotates and predicts the effects of variants on genes (amino acid changes…)
- Uses as input GFF annoation file and VCF

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%



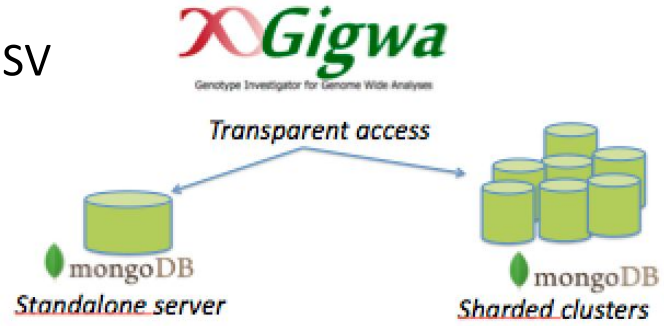
	feature	effect	codon_change	amino_acid_change	MAF	missing_data
2860:	exon	SYNONYMOUS_CODING	caC/caT	H/H	28.1%	0.0%
	#	#	#	#	27.5%	0.0%
2870:	intron	#	#	#	28.1%	0.0%
2870:	exon	NON_SYNONYMOUS_CODING	cAg/cTg	Q/L	28.1%	0.0%
2870:	exon	NON_SYNONYMOUS_CODING	aAc/aTc	N/I	28.1%	0.0%
2870:	exon	NON_SYNONYMOUS_CODING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

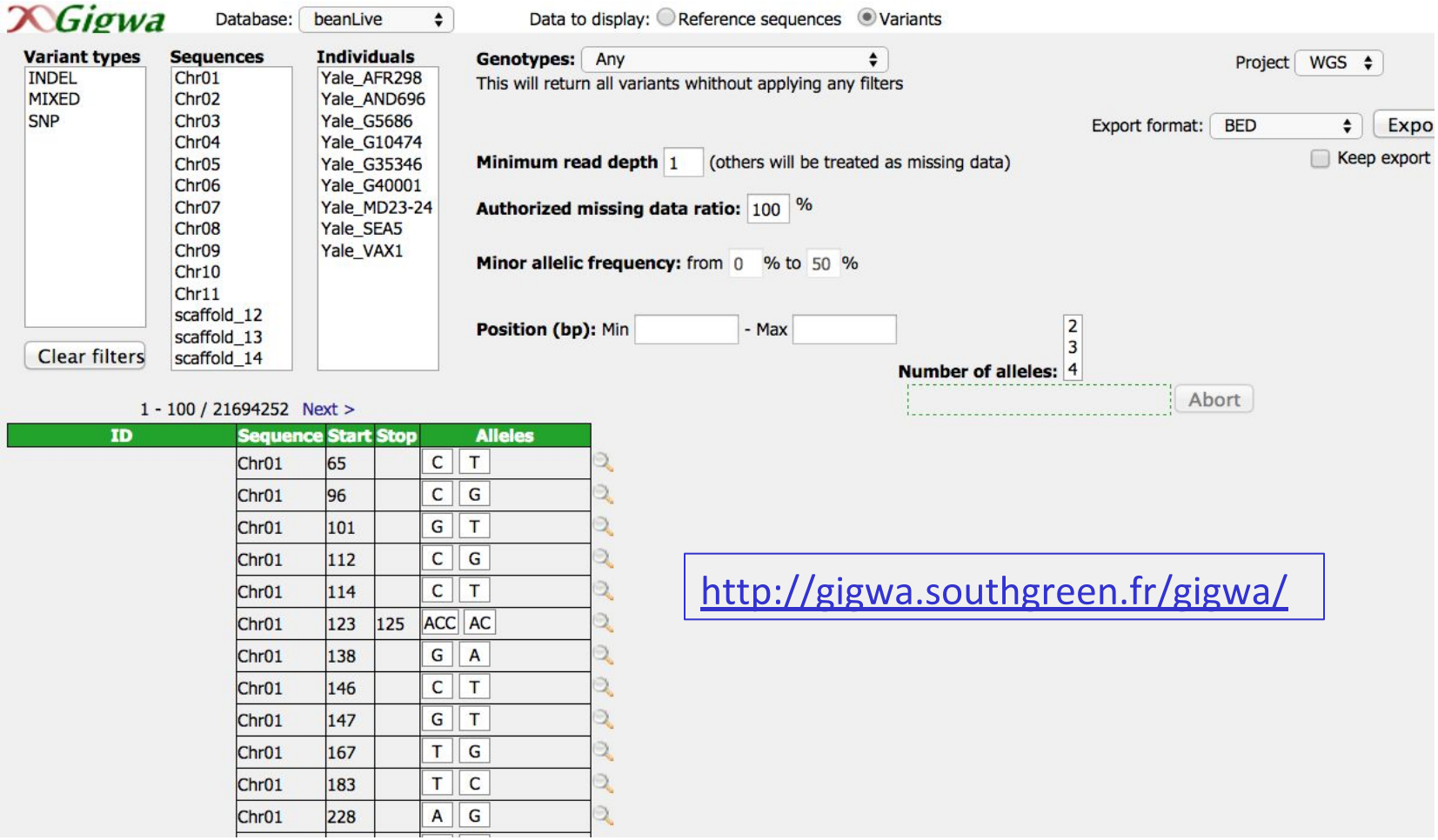
Previous 1 2 3 4 5 … 30 Next

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

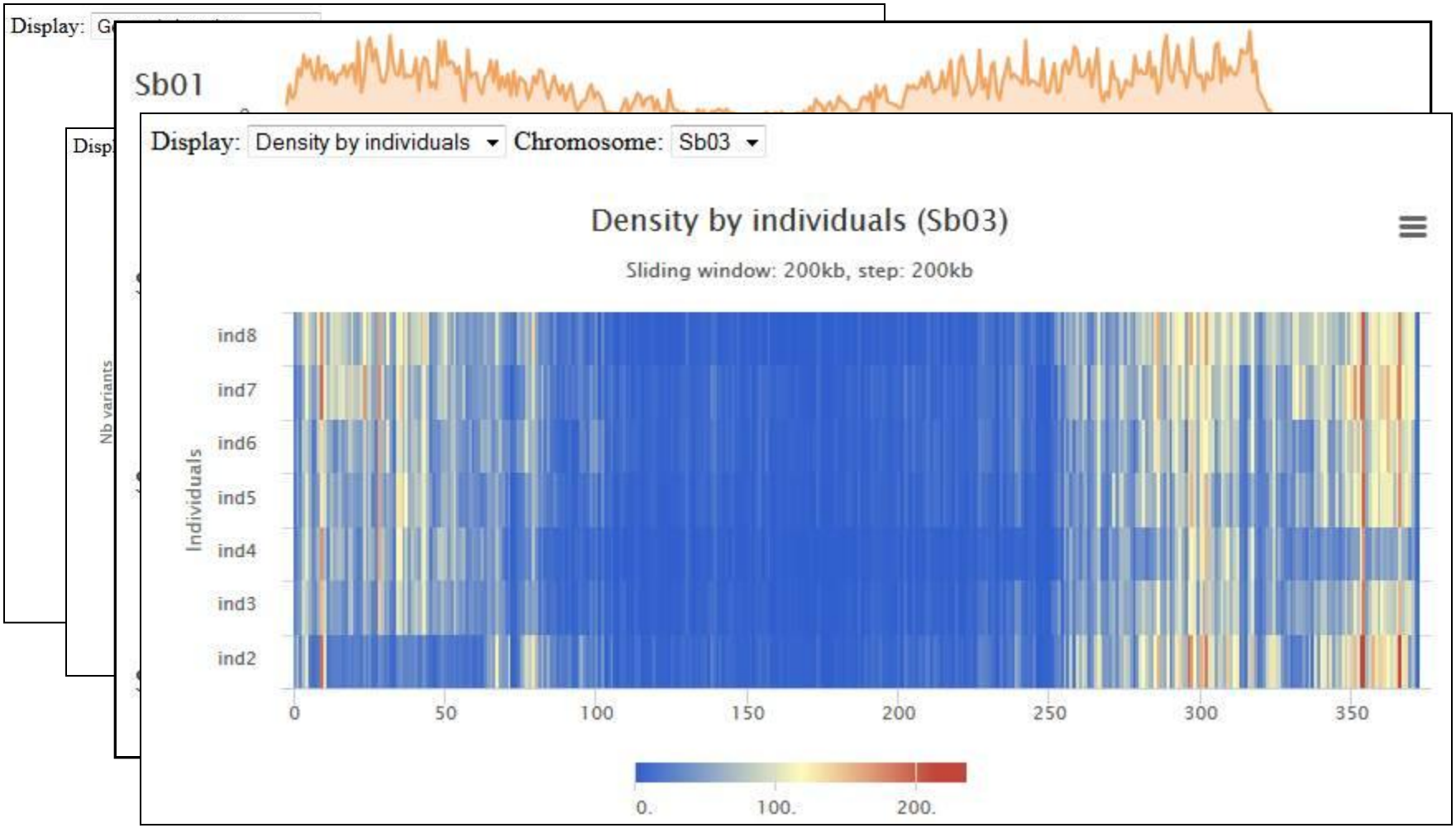
« With NGS arise serious computational challenges in terms of storage, search, sharing, analysis, and data visualization, that redefine some practices in data management. »

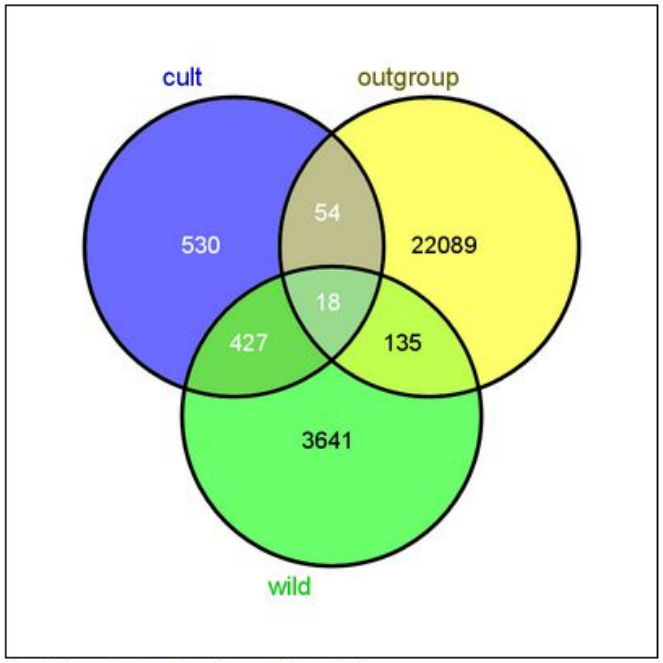
- Based on NoSQL technology mongoDB
- Handles VCF files (Variant Call Format) and annotations
- Supports multiple variant types: SNPs, InDels, SSRs, SV
- Powerful genotyping queries
- Easily scalable with MongoDB sharding
- Transparent access
- Takes phasing information into account when importing/exporting in VCF format





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


Display:
Sb01
Display: Density by individuals
Chromosome: Sb03
Density by individuals (Sb03)
Sliding window: 200kb, step: 200kb
Individuals
ind8
ind7
ind6
ind5
ind4
ind3
ind2
0
50
100
150
200
250
300
350
0.
100.
200.



Specific and shared polymorphisms between groups

Comparison between individuals

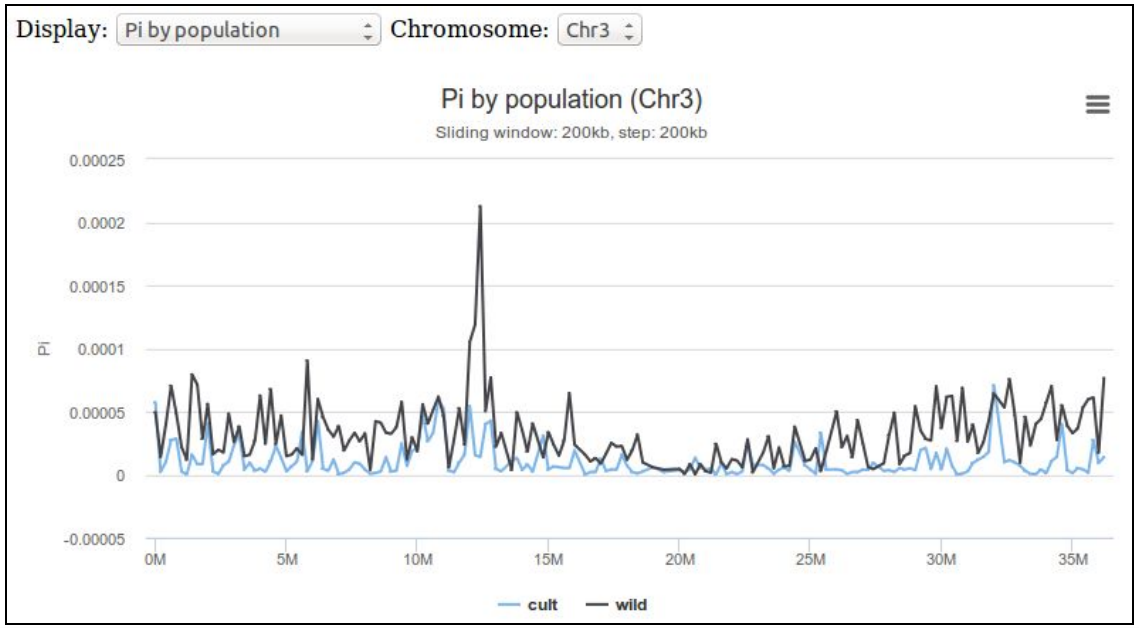
Fst: Fixation index: measure of population differentiation due to genetic structure.

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

+ 2186 polymorphisms inter-group

Diversity analysis



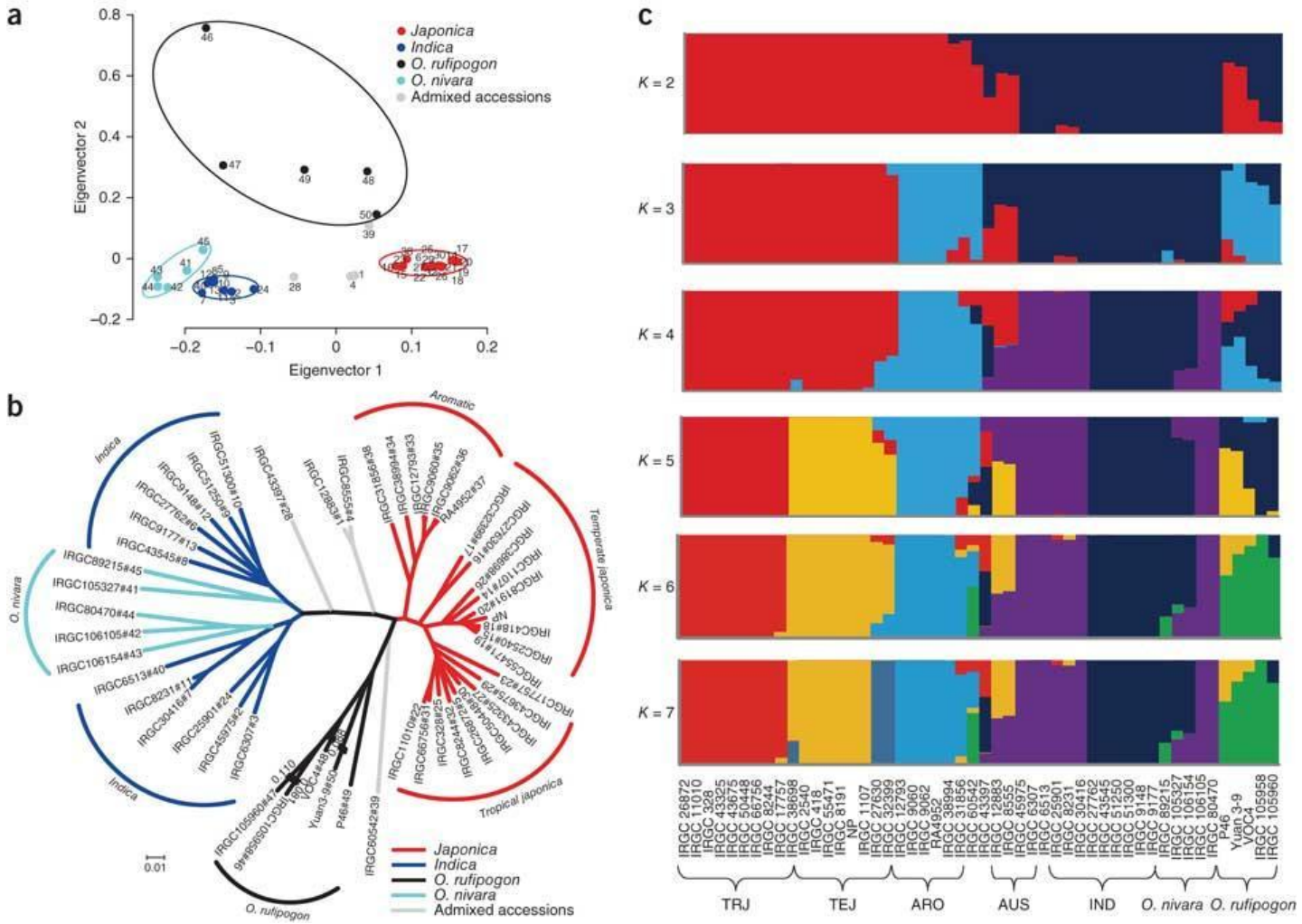
SNP density by individuals can allow the detection of introgression event.

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

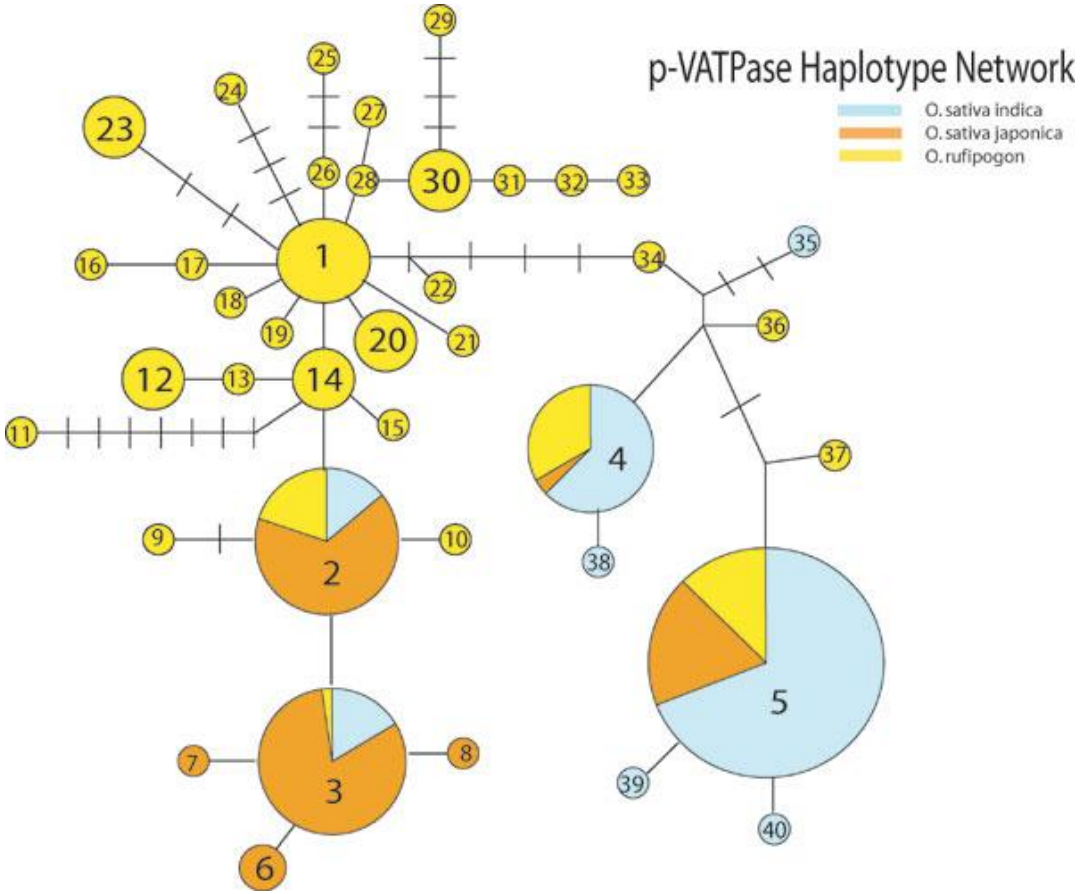
Population structure

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



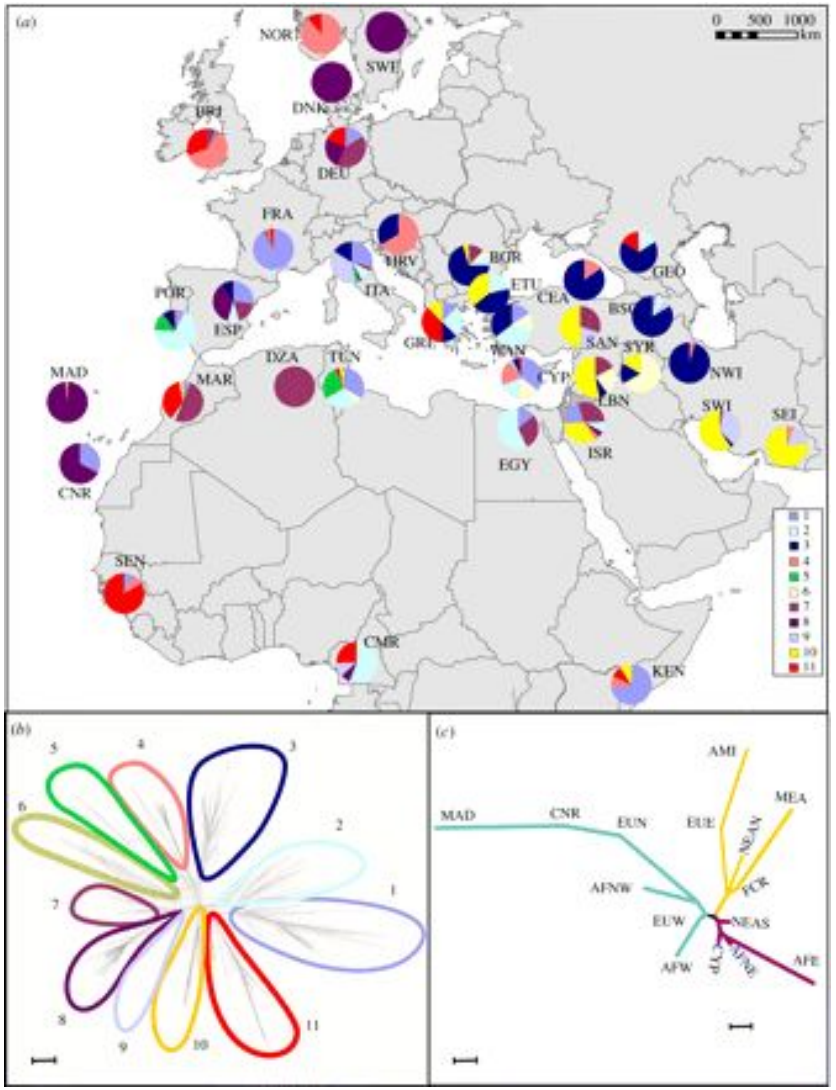
Haplotype network

Exemple d'une région génomique chez le Riz

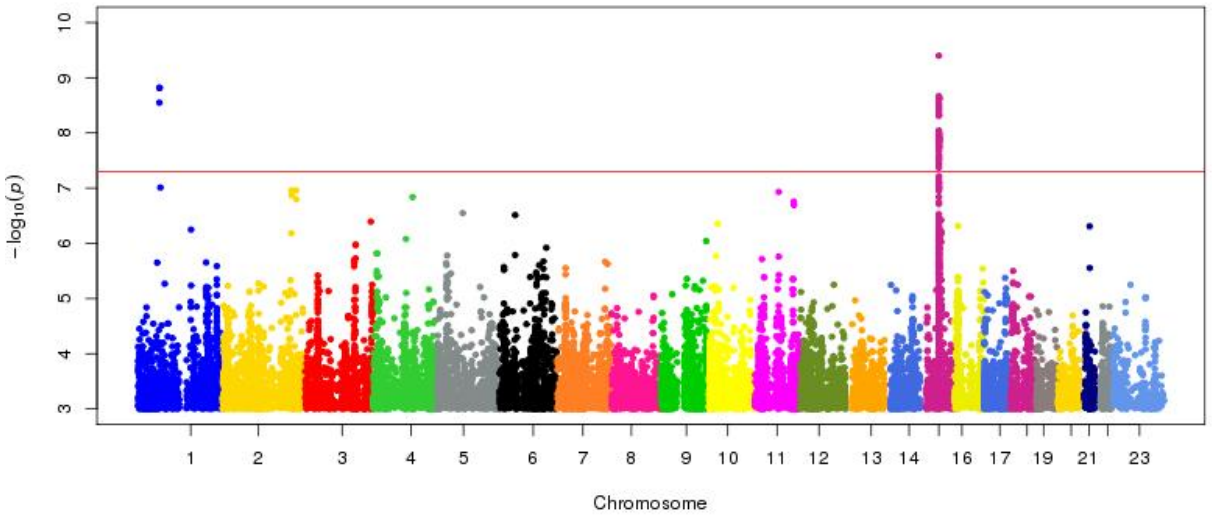


Haplotype and geographical distribution

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.

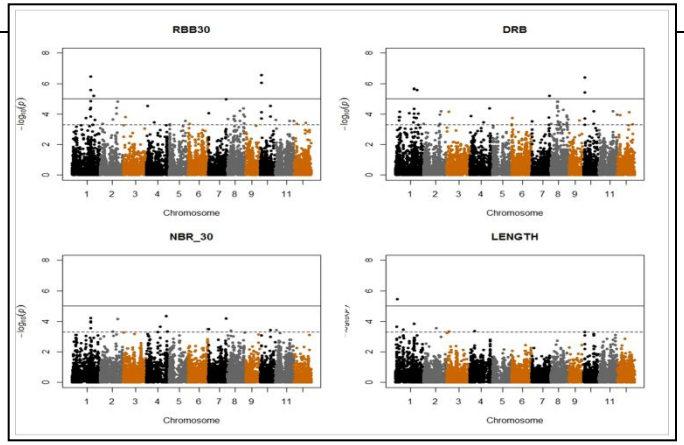
Study of root characters using GWAS in Oryza sativa japonica. Influence of a correction using structure and kinship

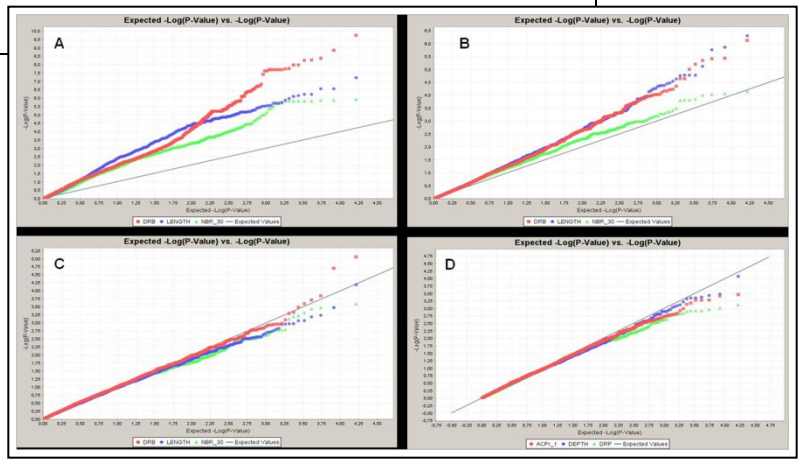
PLOS ONE

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

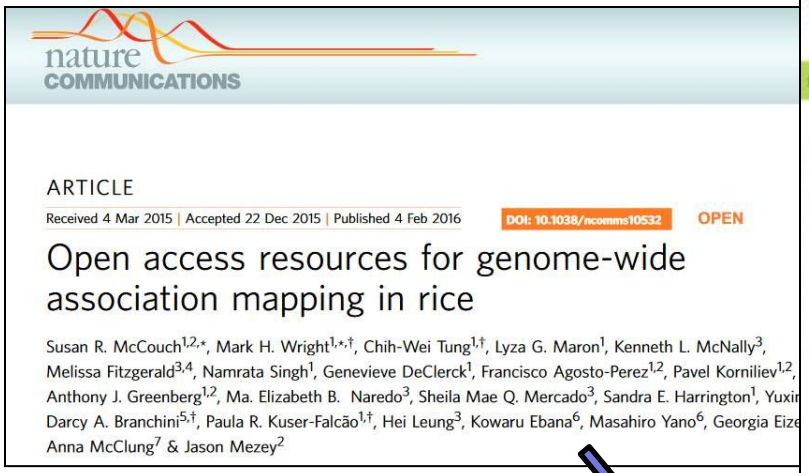
Brigitte Courtois , Alain Audebert, Audrey Dardou, Sandrine Roques, Thaura Ghneim- Herrera, Gaëtan Droc, Julien Frouin, Lauriane Rouan, Eric Gozé, Andrzej Kilian, Nourollah Ahmadi, Michael Dingkuhn

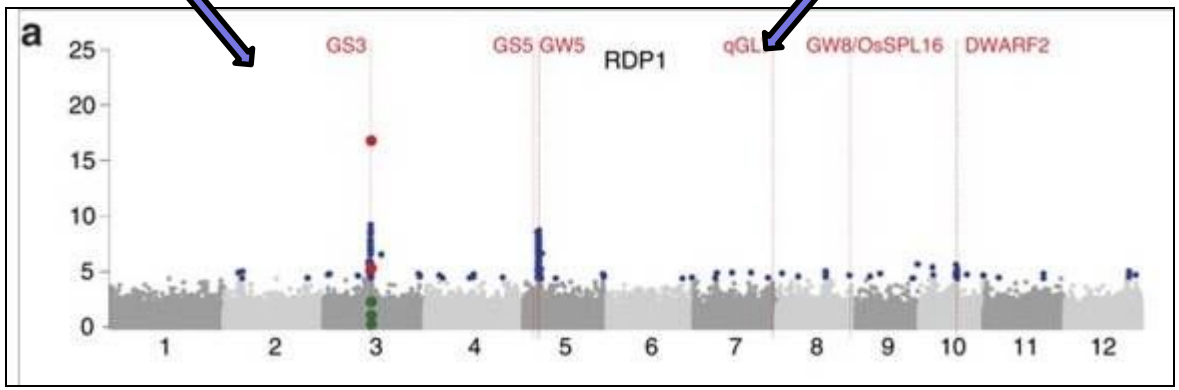
Published: November 5, 2013 • DOI: 10.1371/journal.pone.0078037





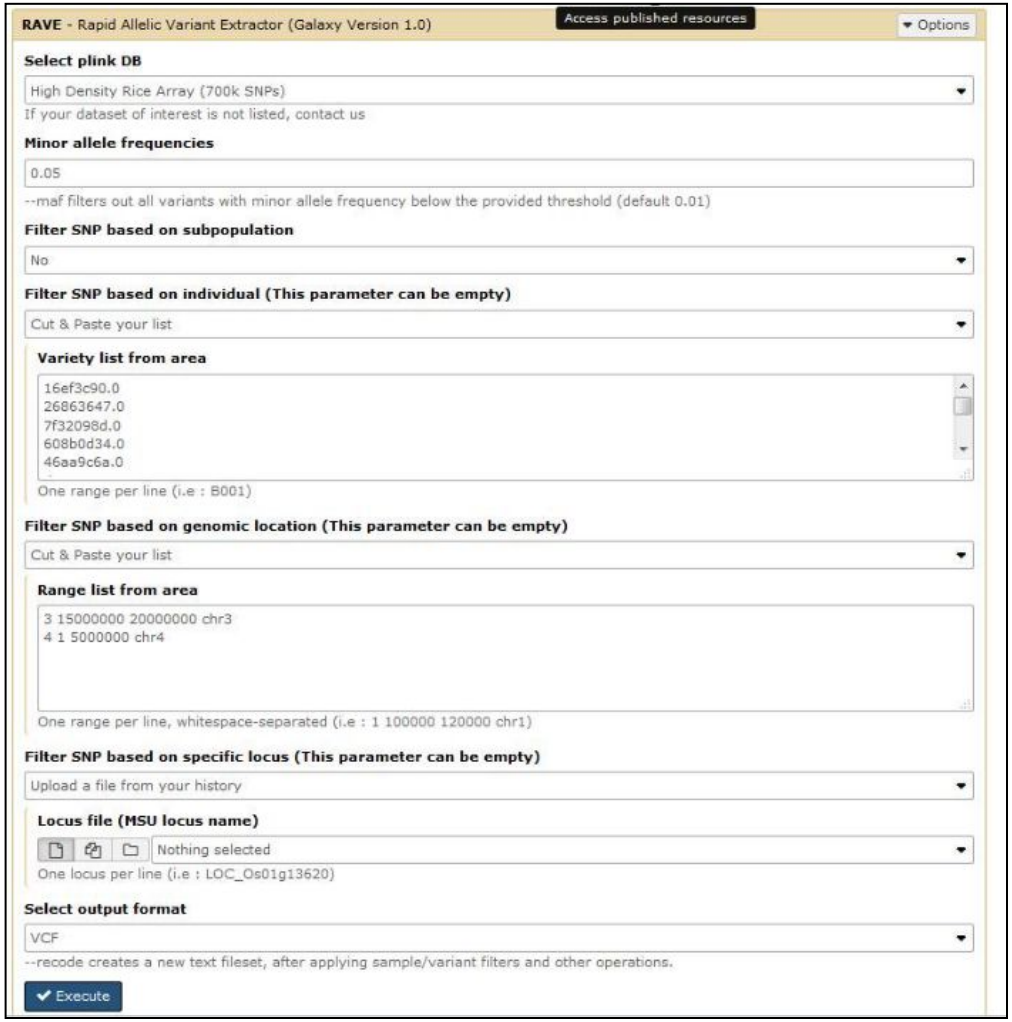
Exemple du Riz: 3000 genomes+ HDRA (High density Rice Array)





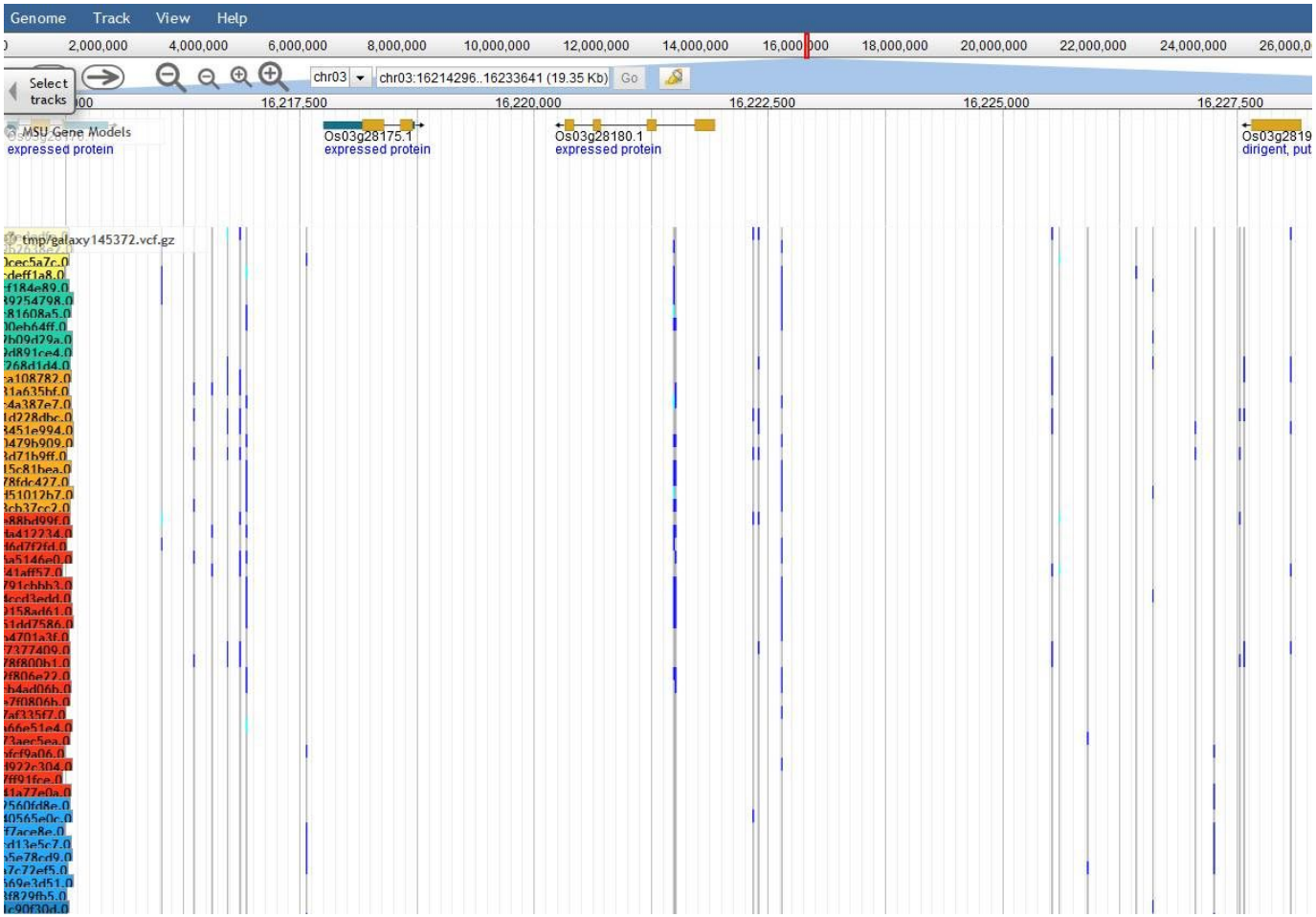
Exemple du Riz: 3000 genomes+ HDRA (High density Rice Array)

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



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

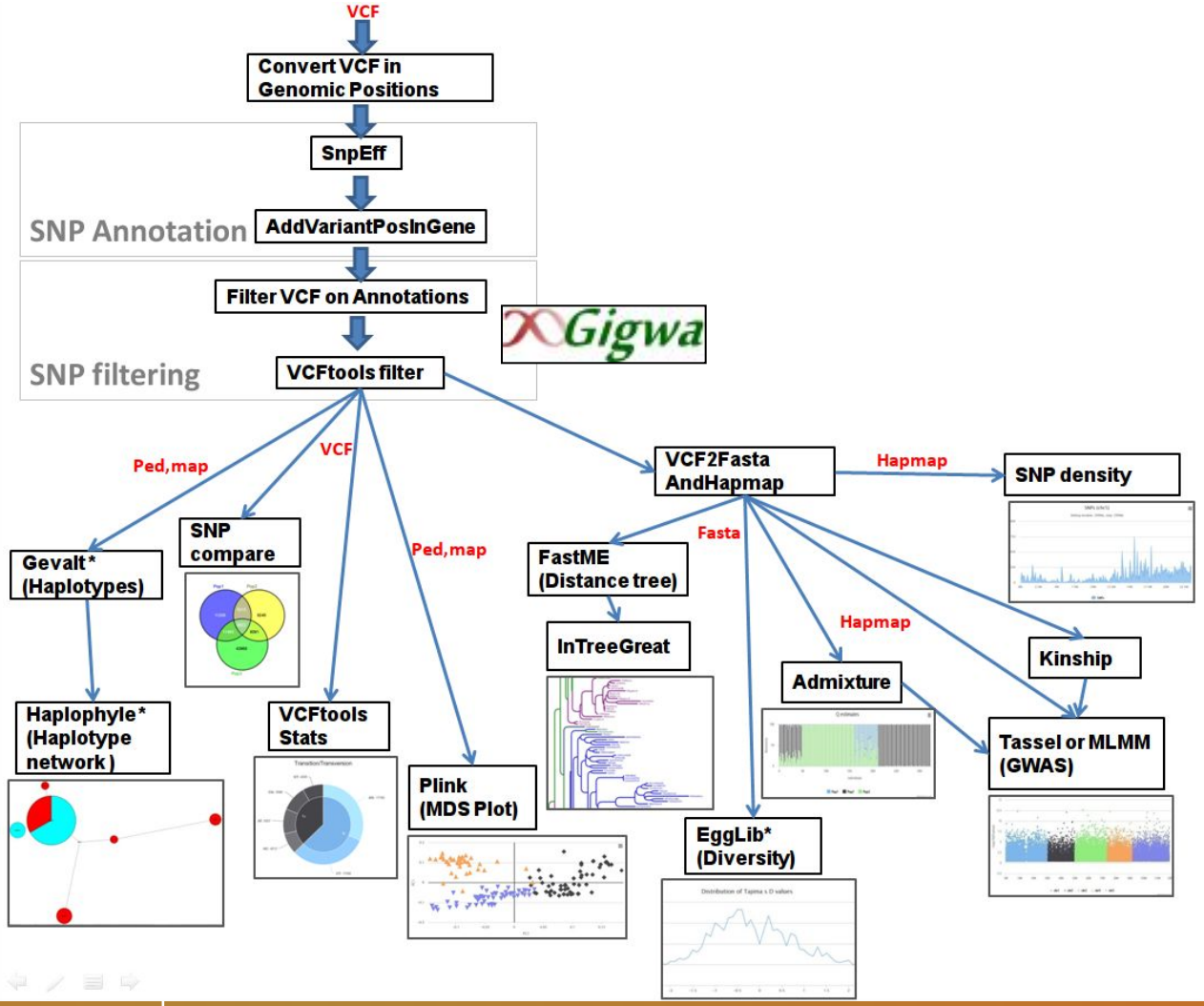
Visualisation du contexte génomique dans un génome browser (plugin Jbrowse)



SNiPlay

SNiPlay: Web application for polymorphism analyses

http://sniplay.southgreen.fr



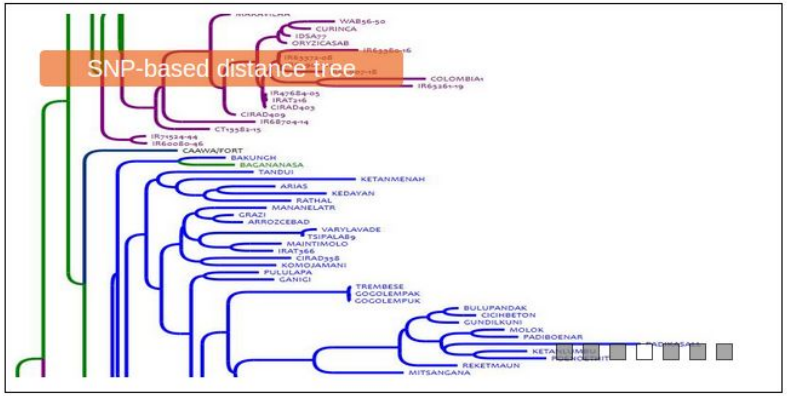
SNiPlay Site web

SNiPlay

http://sniplay.southgreen.fr

Home | Pipeline for SNP analysis | Tools | SNP Database | Documentation | How to cite | Login

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...
Start now



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