

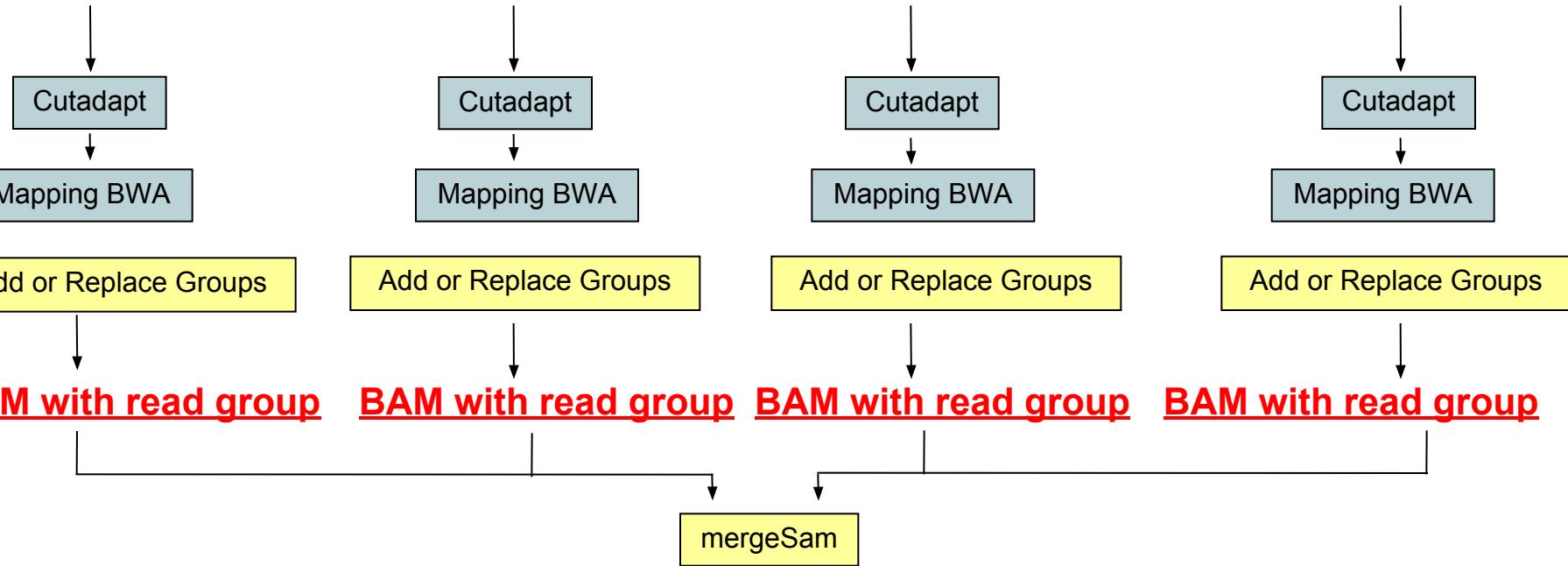
Analyse de variants génétiques (SNPs, indels)

Fastq (RC1)

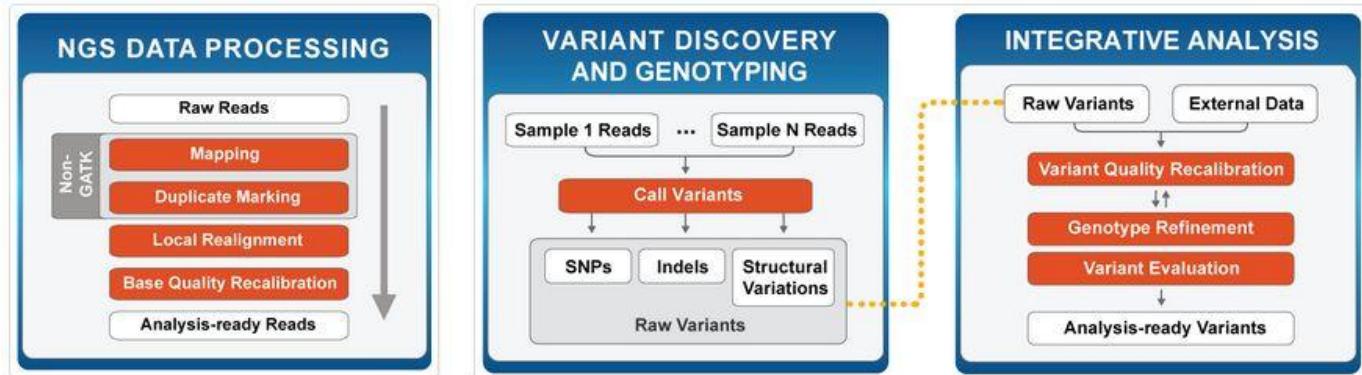
Fastq (RC2)

Fastq (RC3)

Fastq (RC4)



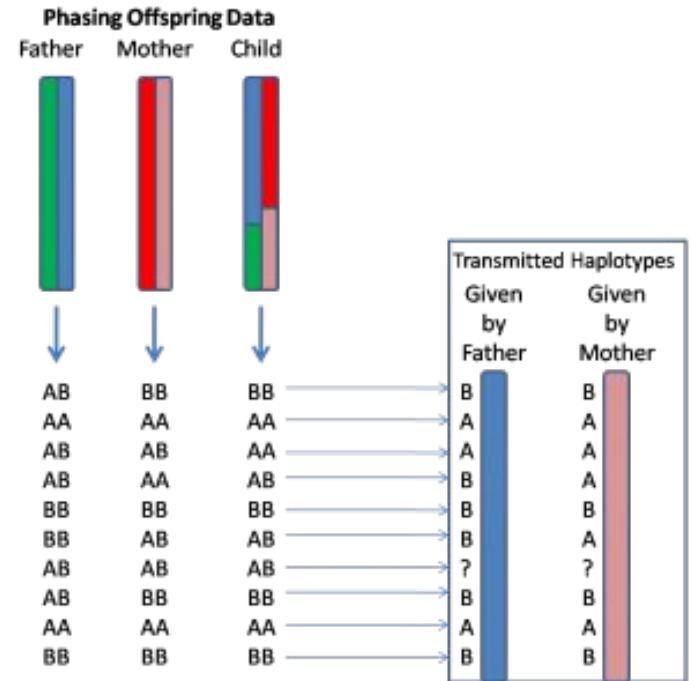
Global BAM with read group



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 inferred by statistics methods using non-ambiguous 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

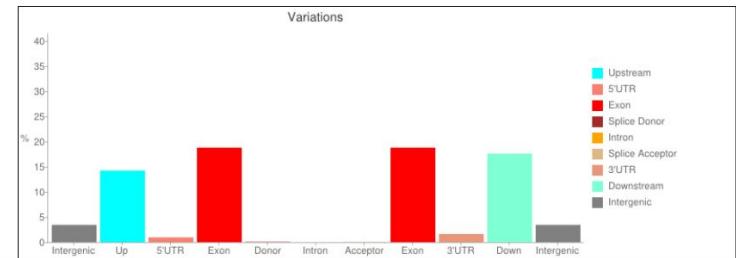
SnpEff

Genetic variant annotation and effect prediction toolbox.

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

Type	Count	Percent
Type (alphabetical order)		
3_prime_UTR_variant	2	0.002%
5_prime_UTR_premature_start_codon_gain_variant	1,320	1.615%
5_prime_UTR_variant	87	0.106%
downstream_gene_variant	685	0.338%
intergenic_region	14,390	17.607%
intragenic_variant	2,804	3.431%
intron_variant	1	0.001%
missense_variant	34,019	41.613%
non_coding_exon_variant	4,326	5.293%
splice_acceptor_variant	5,328	6.519%
splice_donor_variant	32	0.039%
splice_region_variant	46	0.056%
start_lost	1,355	1.668%
stop_gained	7	0.009%
stop_lost	69	0.084%
stop_retained_variant	5	0.006%
synonymous_variant	2	0.002%
upstream_gene_variant	5,627	6.885%
	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	34,019	41.613%
NONE	3	0.004%
SPlice_SITE_ACCEPTOR	32	0.039%
SPlice_SITE_DONOR	46	0.056%
SPlice_SITE_REGION	1,355	1.668%
TRANSCRIPT	7	0.009%
UPSTREAM	11,634	14.235%
UTR_3_PRIME	1,320	1.615%
UTR_5_PRIME	772	0.945%



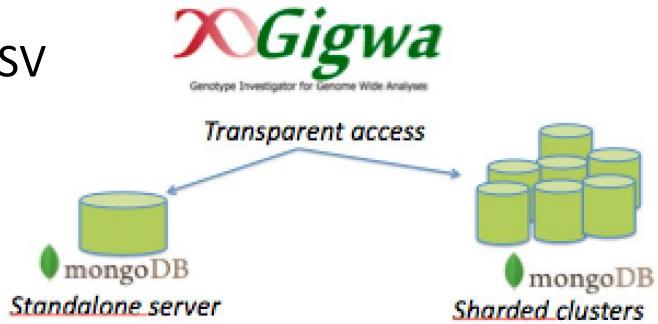
feature	effect	codon_change	amino_acid_change	MAF	missing_data
:860: exon	SYNONYMOUS_CODING	caC/caT	H/H	28.1%	0.0%
#	#	#	#	27.5%	0.0%
:870: intron	#	#	#	28.1%	0.0%
:870: exon	NON_SYNONYMOUS_CODING	cAg/cTg	Q/L	28.1%	0.0%
:870: exon	NON_SYNONYMOUS_CODING	aAc/aTc	N/I	28.1%	0.0%
:870: exon	NON_SYNONYMOUS_CODING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

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
- 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



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_SEAS5
Yale_VAX1

Project: WGS

Export format: BED

Exp

 Keep export**Genotypes:** Any

This will return all variants without applying any filters

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: _____

2

3

4

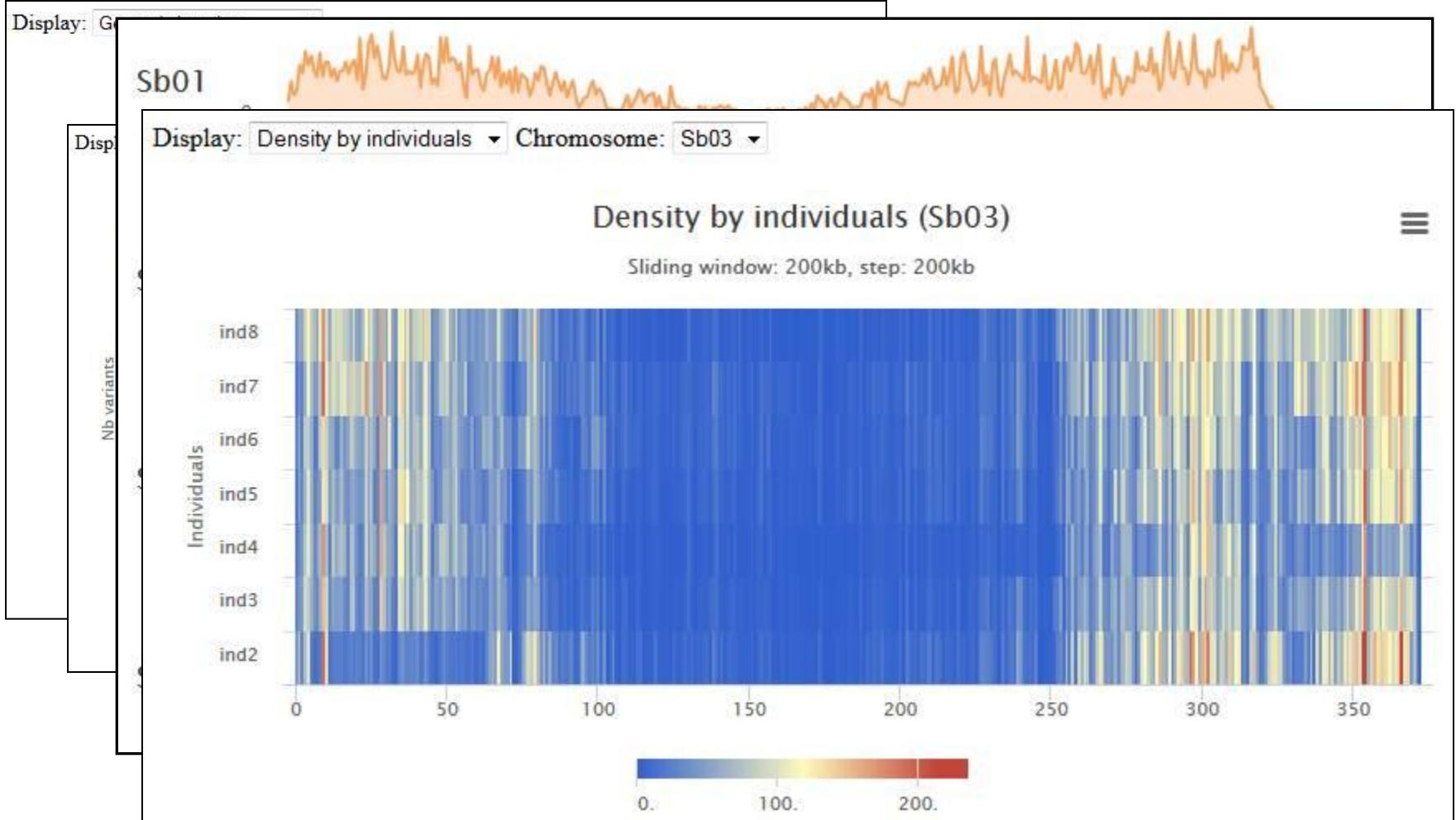
Number of alleles: _____

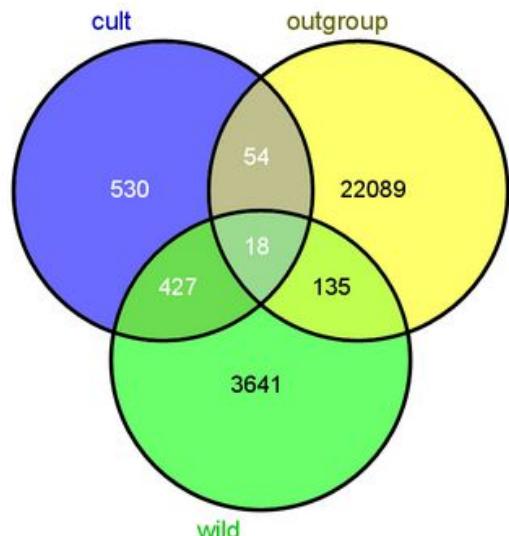
Abort

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/>





Specific and shared polymorphisms between groups

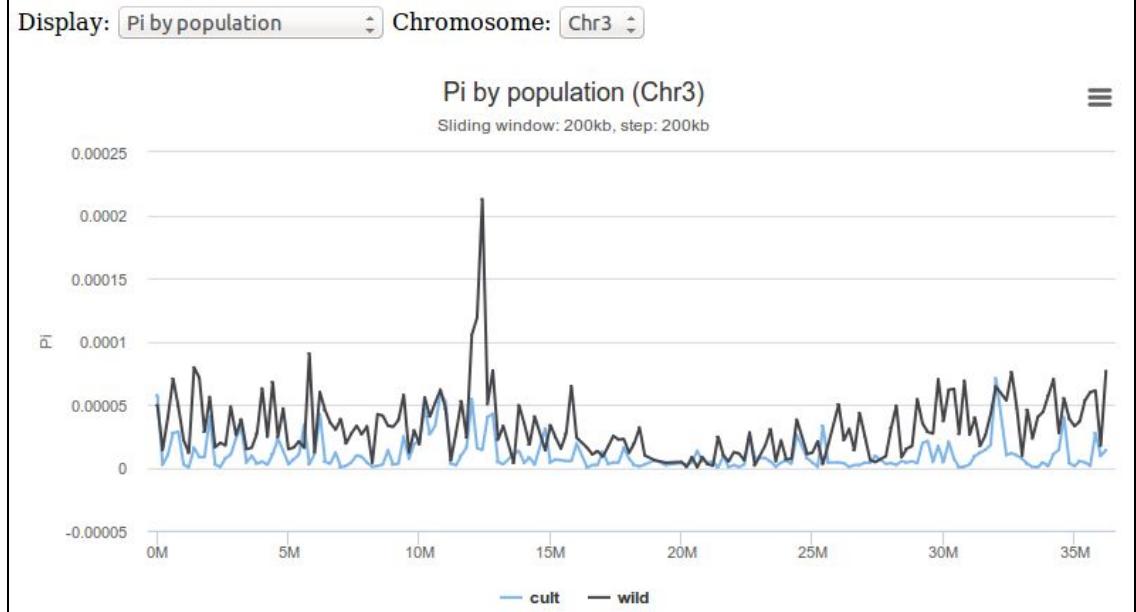
Comparison between individuals

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

Diversity analysis



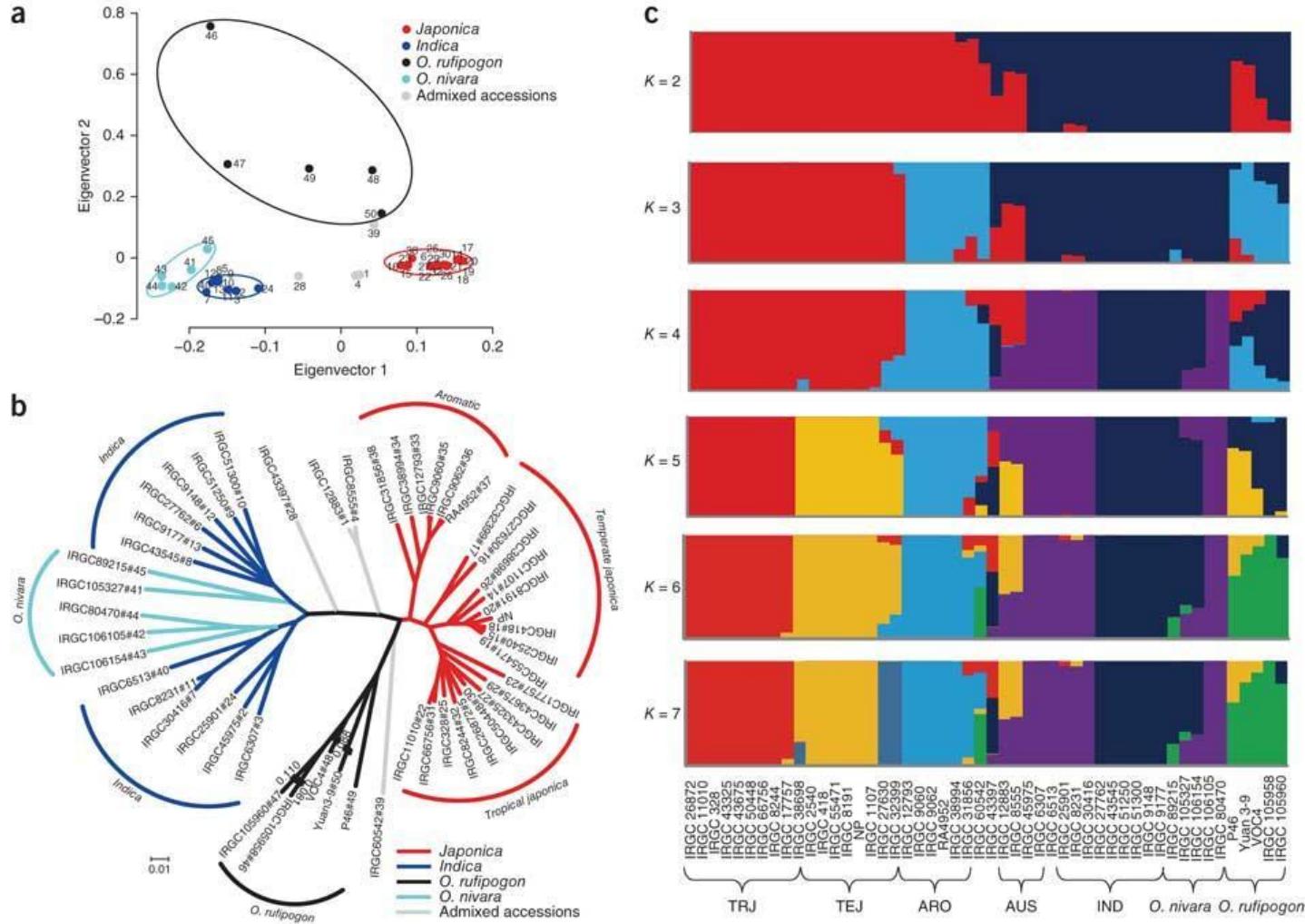
SNP density by individuals can allow the detection of introgression event.

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

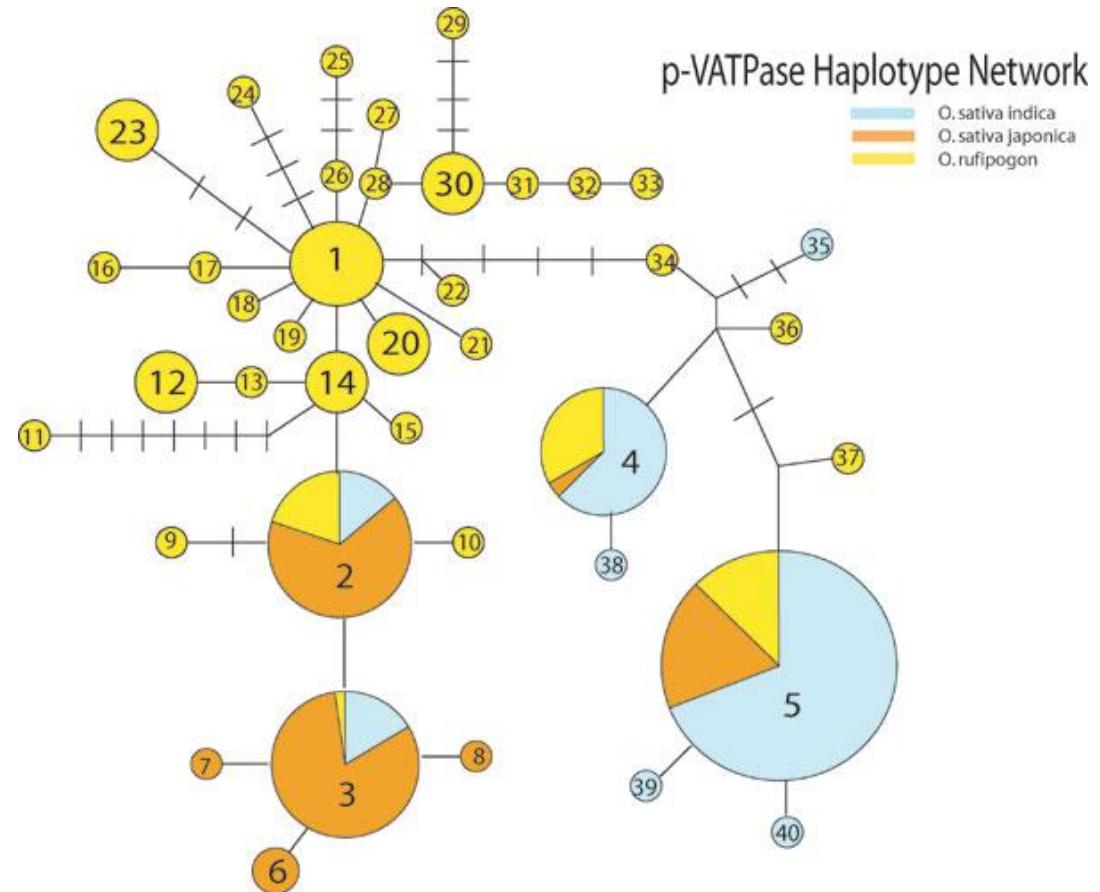
Population structure

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



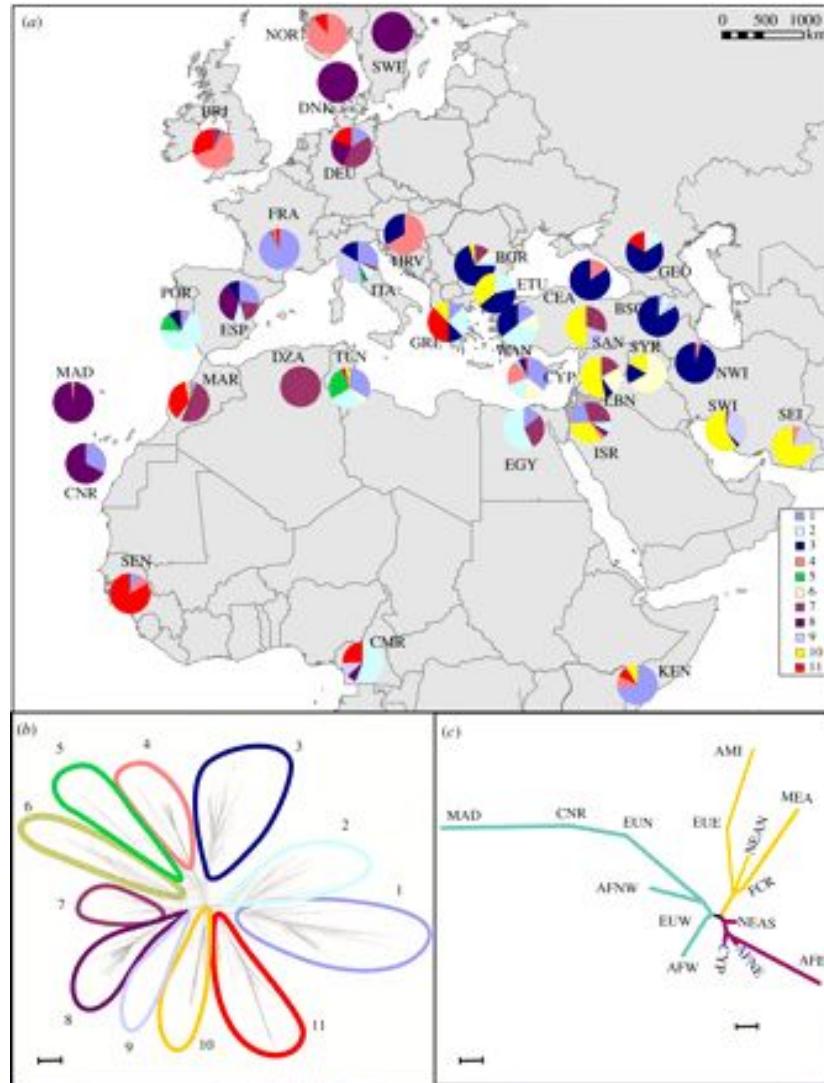
Haplotype network

Exemple d'une région génomique chez le Riz

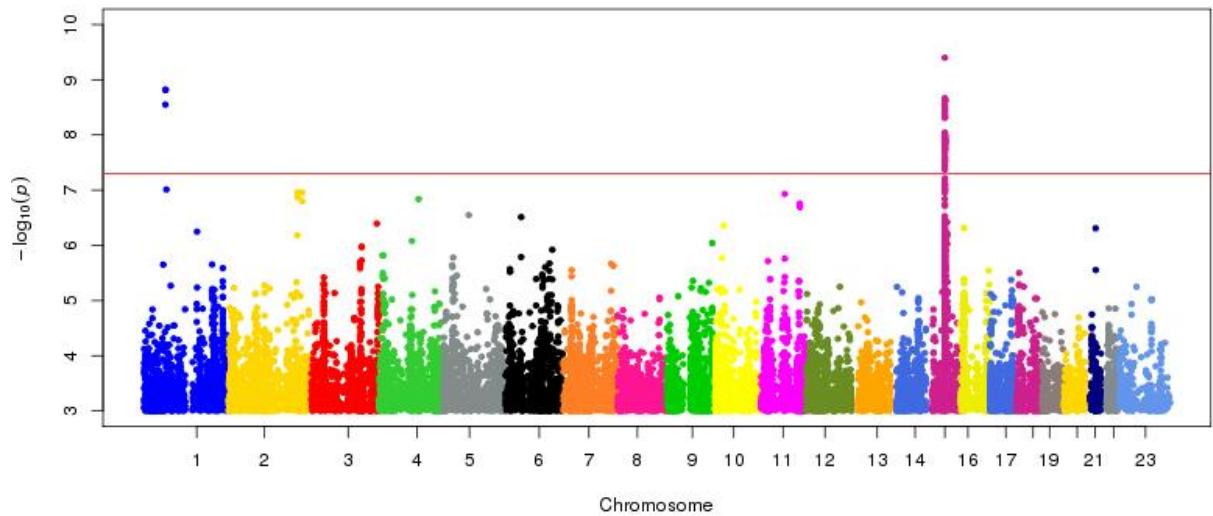


Haplotype and geographical distribution

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 ($-\log_{10}$ pvalue) along chromosomes
- TASSEL, MLMM softwares
- 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 distortion in association patterns.

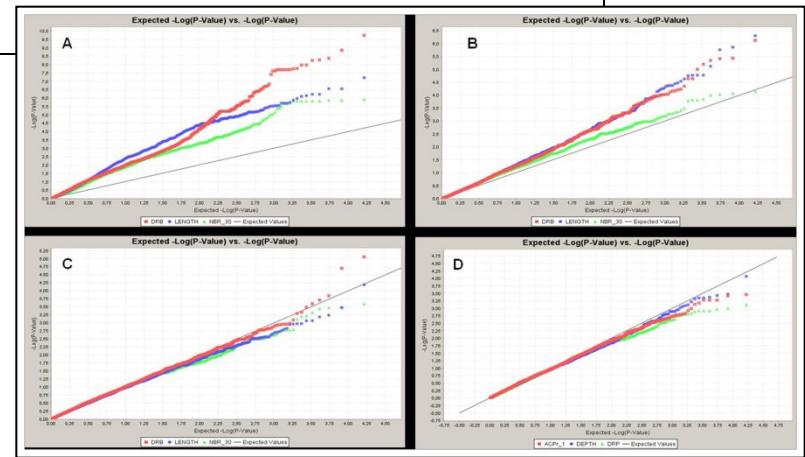
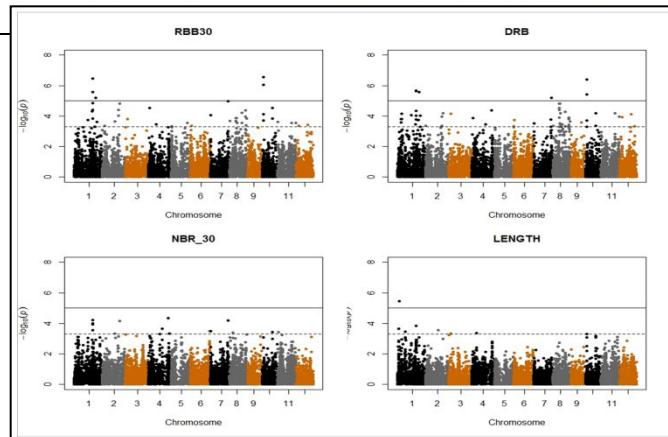
Study of root characters using GWAS in *Oryza sativa japonica*. Influence of a correction using structure and kinship



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. Maron¹, Kenneth L. McNally³, Melissa Fitzgerald^{3,4}, Namrata Singh¹, Genevieve DeClerck¹, Francisco Agosto-Perez^{1,2}, Pavel Korniliev^{1,2}, Anthony J. Greenberg^{1,2}, Ma. Elizabeth B. Naredo³, Sheila Mae Q. Mercado³, Sandra E. Harrington¹, Yuxin Darcy A. Branchini^{5,†}, Paula R. Kuser-Falcão^{1,†}, Hei Leung³, Kowaru Ebana⁶, Masahiro Yano⁶, Georgia Eize Anna McClung⁷ & Jason Mezey²

Rice Diversity

Project Information Data & Tools Publications Education and Outreach Links of Interest Leading

Data Sets

High Density Rice Array (HDRA, 700k SNPs)

McCook 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-Falcão 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**
 - Plink 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

a

GS3 GS5 GW5 RDP1 qGL GW8/OsSPL16 DWARF2

1 2 3 4 5 6 7 8 9 10 11 12

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) Access published resources ▾ Options

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
26663647.0
7f3209bd.0
608bb0d34.0
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

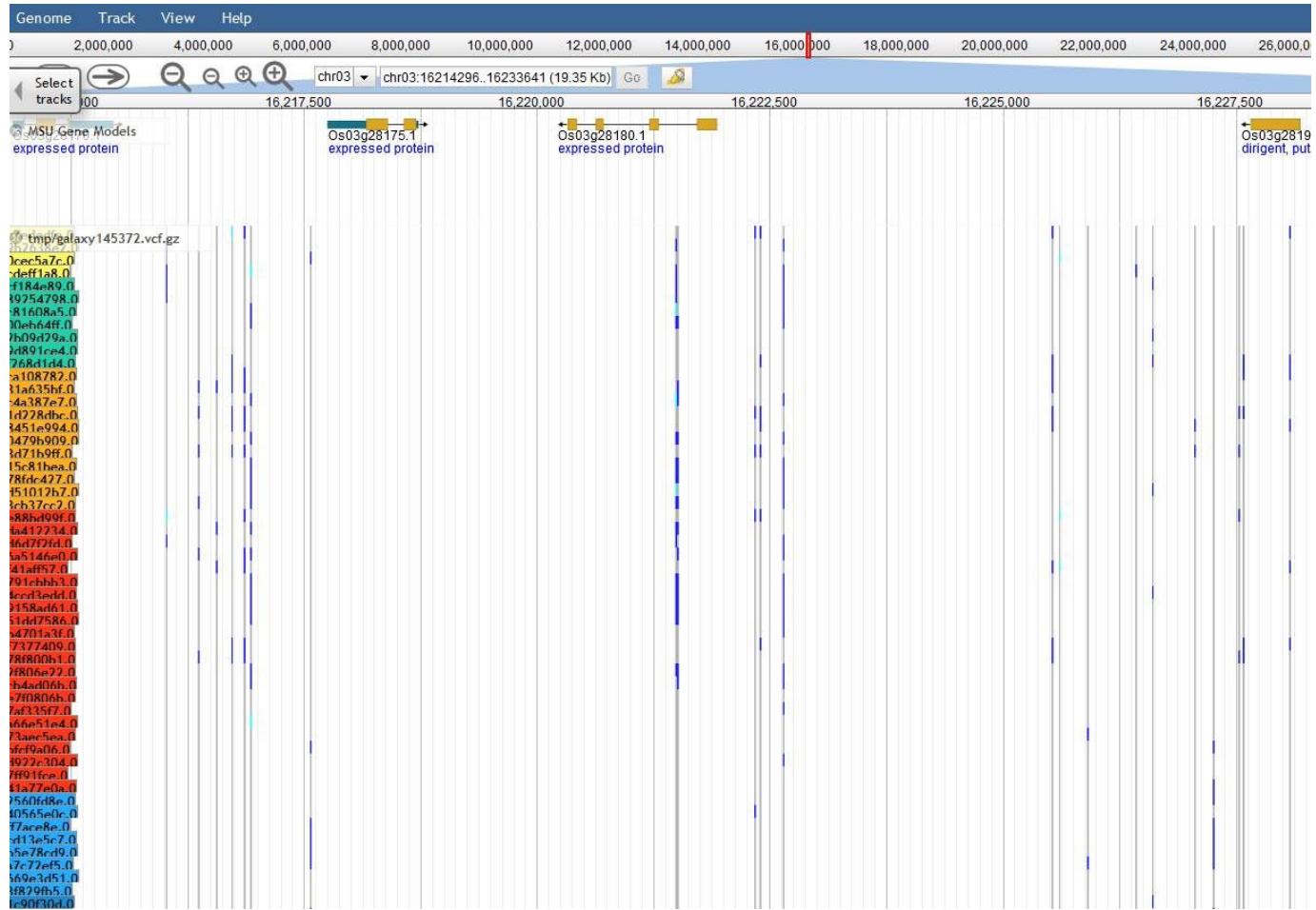
Locus file (MSU locus name)
Nothing selected
One locus per line (i.e : LOC_Os01g13620)

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

Execute

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

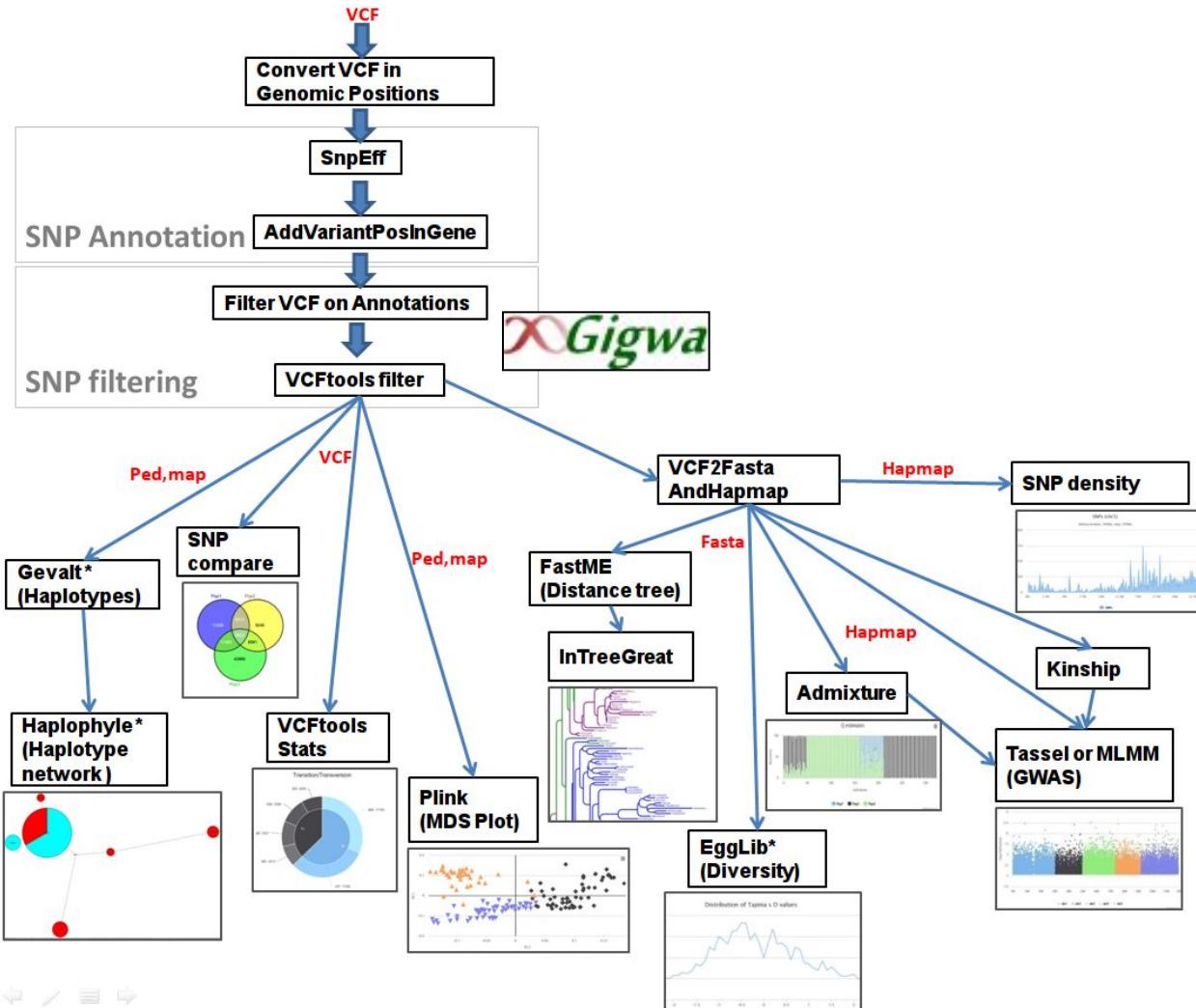
Visualisation du contexte génomique dans un génome browser (plugin Jbrowse)





SNiPlay: Web application for polymorphism analyses

<http://snipplay.southgreen.fr>



SNiPlay Site web

SNiPlay

<http://snipplay.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

