GenomeHarvest
diversity, organization and dynamics

Formation sur les logiciels de reconstruction de génomes mosaïques sous galaxy

http://cc2-web1.cirad.fr/galaxydev et http://galaxy.southgreen.fr/galaxy/

Projet Genome Harvest

WP1 - Management - Federate the community around research actions

WP2

Characterize genome architecture of crops derived from multiple funder (sub)species

→ **mosaic genome structure** within crops genome

→ **large structural variations** within crops genome

WP3

Impact of genome architecture

→ on chromosome **recombinaison/segregation**

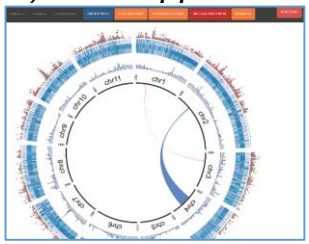
→ on allele (gene) **expression**

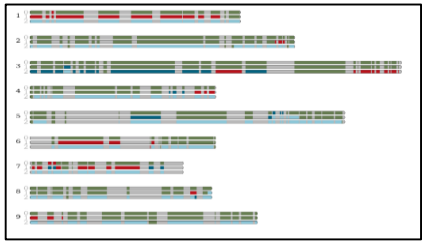
WP4

- Implement tools in platform
- Capacity building/Training
- Scientific events

Plusieurs niveaux d'intégration des outils

1) Développement d'outils de visualisation

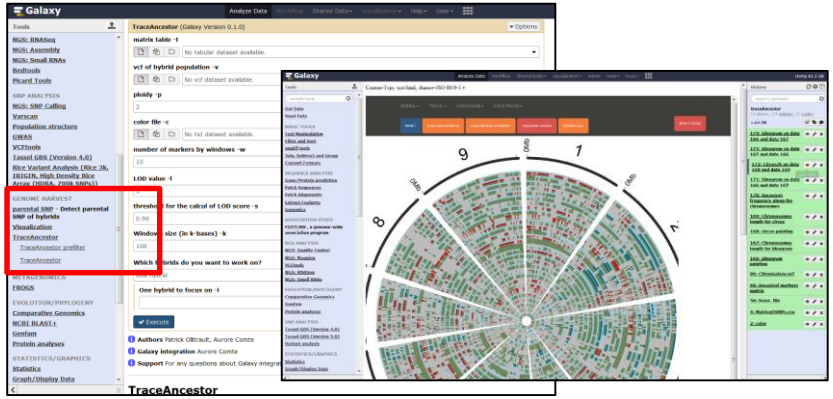




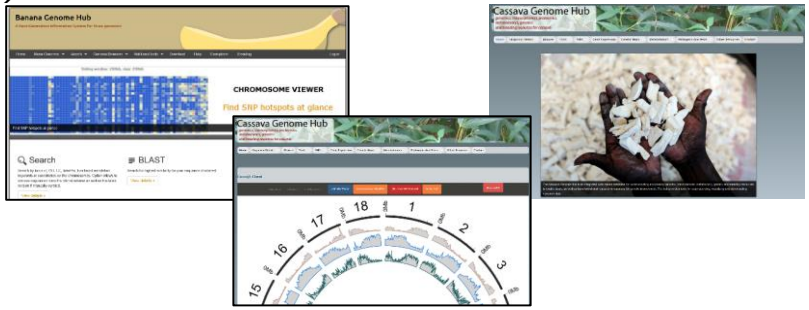
2) Accès aux outils par Galaxy.
Développement de wrappers/workflows

Galaxy

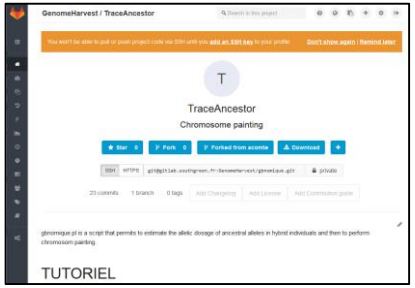
http://galaxy.southgreen.fr/galaxy/



3) Connexion des outils aux Genome Hub



4) Mise à disposition du code et documentation via GitHub/GitLab



github SOCIAL CODING

GitLab

WP2
(characterize inter(sub)specific mosaic genome structures)

WP3
(impact of genome structure on gene expression)

KDE classifier

python

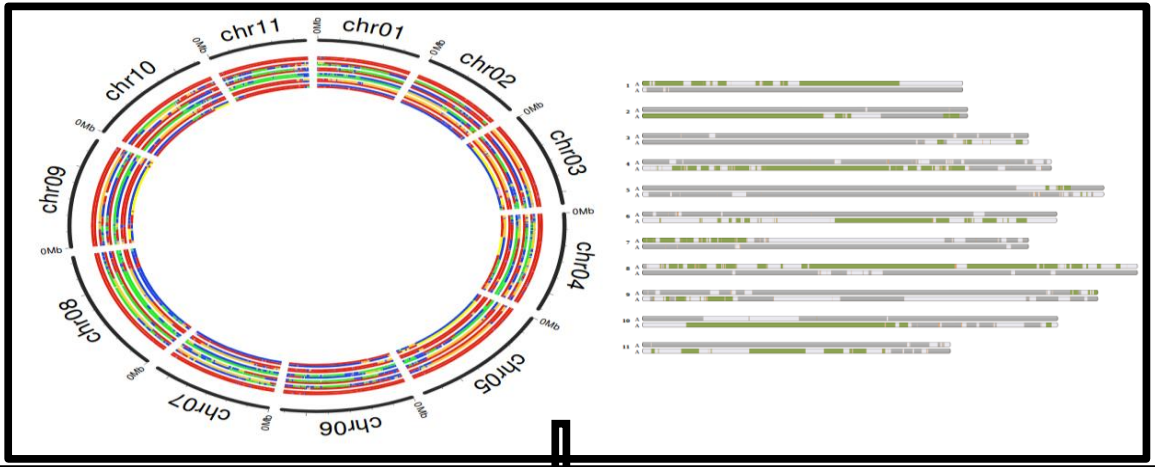
VCF Hunter

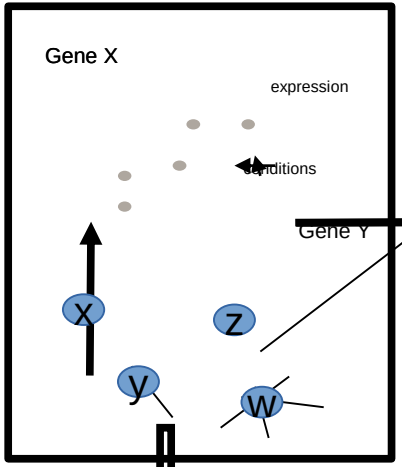
python

TraceAncestor

perl

ASE co-expression





Galaxy

Banana Genome Hub
A Next-Generation Information System for Musa genomics

GitLab

github
SOCIAL CODING

WP2

(characterize inter(sub)specific mosaic genome structures)

KDE classifier

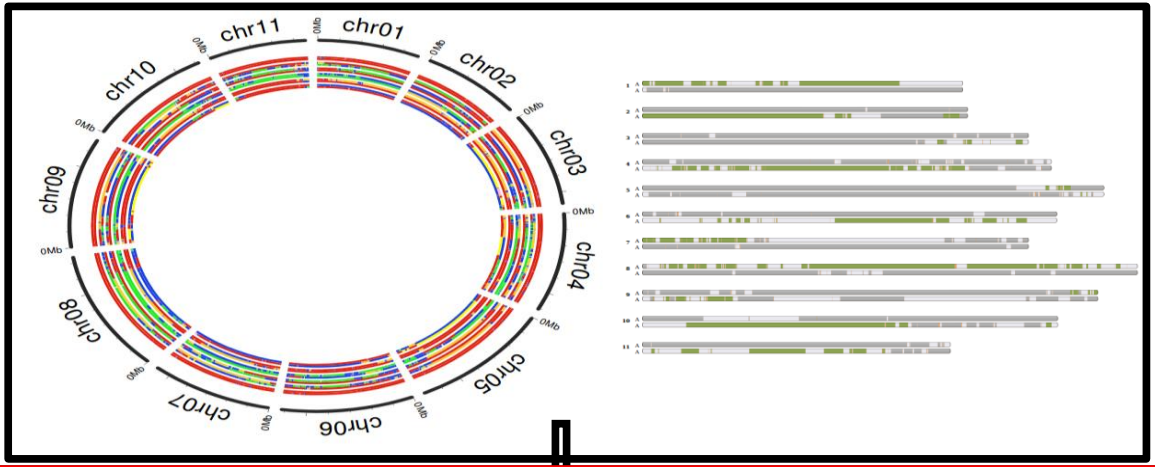
python

VCF Hunter

python

TraceAncestor

perl



Galaxy

Banana Genome Hub

A Next-Generation Information System for Musa genomics

GitLab

github
SOCIAL CODING

galaxy.southgreen.fr/galaxy/ 67 % google drive

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Tools

NGS: RNASeq
NGS: Assembly
NGS: Small RNAs
Bedtools
Picard Tools

SNP ANALYSIS
NGS: SNP Calling
Varscan
Population structure
GWAS
VCFtools
Tassel GBS (Version 4.0)
Rice Variant Analysis (Rice 3k, IRIGIN, High Density Rice

GENOME HARVEST
parental SNP - Detect parental SNP of hybrids
Visualization
TraceAncestor
Traceancestor
Markers Detector

METAGENOMICS
FROGS

EVOLUTION/PHYLOGENY
Comparative Genomics
NCBI BLAST+
Genfam
Protein analyses

STATISTICS/GRAPHICS
Statistics
Graph/Display Data

SOUTHGREEN PROJECTS
SNiPlay3

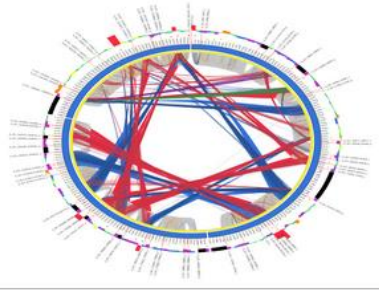
cirad | IRD | Bioversity | INRA | SupAgro

SouthGreen bioinformatics platform

Welcome to GALAXY

Our pre-configured and validated workflows

NGS analyses | SNP calling | SNP analysis | GWAS | Structural variations



Structural variations

Scaffremodeler can be used to detect large structural variations between a reference sequence and a resequenced genome (Martin et al, 2016)

Input: Fastq + FASTA

Access workflow

Chrom. reconstruction | Metagenomics | Gene families

These workflows as part of the services provided by South Green

How to load big datasets?

Core values

- **Accessibility**
  - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- **Reproducibility**
  - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- **Transparency**
  - Users can share or publish their analyses (histories, workflows, visualizations)
  - Pages: online Methods for your paper

Pages: interactive, web-based documents that describe a complete analysis.

=> Diffusion des wrappers via le Galaxy ToolShed

Les outils présentés aujourd’hui ne sont pas encore publiés et ne sont pas tous publics.

“Les outils n'étant pas encore publiés et la plupart étant destinés à être valorisés prochainement, l'équipe SEG émet quelques précautions et souhaite que les participants s'engagent à ne pas diffuser les outils développés par l'équipe ou publier sur banane et canne à sucre avec ces approches tant que les outils n'ont pas été publiés par l'équipe.”

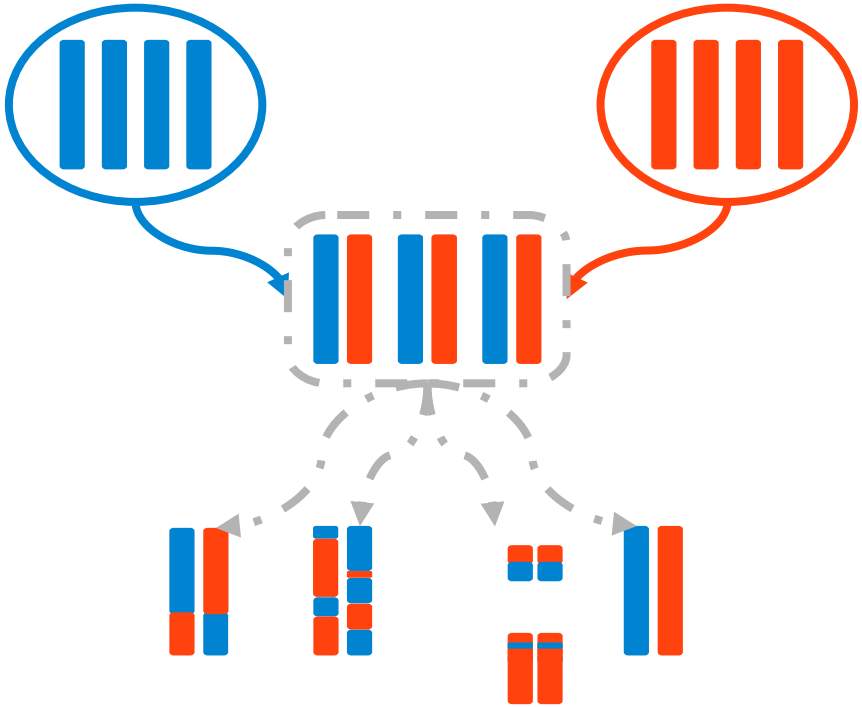
Introduction

●Les évènements d'hybridation entre espèces et sous-espèces sont largement répandus chez les plantes cultivées.

●Brassage génétique → **génomes mosaïques** ayant des origines ancestrales différentes

●**Intérêts ?**

−Histoire de la domestication des plantes cultivées
−Origines ancestrales de certains traits phénotypiques.



Introduction

- Trois modèles biologiques:
  - Nombre d'ancêtres différent (3 à 6)
  - 3 niveaux de structure
  - Génomes ancestraux plus ou moins bien connus

Eureka chr2

PTBA00008 chr4

IRIS 313 11030 chr8

Left to right:
Curk thesis, chapter 4 (2014)
Martin et al, in prep
Santos et al, in prep

Introduction

Outil	Organisme sur lequel il a été développé	Nombre d'haplotypes pris en charge	Permet le phasing ou prend en charge données phasées	Méthode	inputs
TraceAncestor	agrume	2 à 4	non	Estimation de la fréquence des allèles ancestraux identifiés par indice GST	VCF + Liste de SNPs diagnostiques
KDE_Classifier	riz	1	non	Kernel Density Estimation	VCF / géno + fichier structure
VCFHunter	banane	N	non	ACP + clusterisation	VCF

TraceAncestor

PATRICK
OLLITRAULT

-Les parents/ancêtres sont déjà identifiés (pamplemousse, mandarinier, cédrat, micrantha)

-**Méthode** : Pour un individu donné, mesurer fenêtre par fenêtre sur un chromosome la fréquence de présence de SNPs ancestraux.

2 outils dans galaxy :

-**TraceAncestor prefilter**

-**TraceAncestor**

Matrice contenant les indices de différenciation par SNP (GST) par ancêtre

traceAncestor

TraceAncestor prefilter

Liste de SNPs diagnostiques filtrés

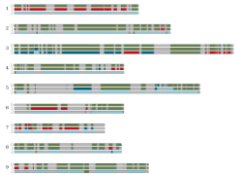
VCF des hybrides
+
fichier couleur

TraceAncestor

Blocs de chromosomes à coloriser

Visualisation

ideogram



circos

GENOME HARVEST

parental SNP - **Detect parental SNP of hybrids**

Trace Ancestor

- TraceAncestor prefilter
- TraceAncestor

vcfHunter

KDE Classifier

Visualization

GENOME HARVEST

parental SNP - **Detect parental SNP of hybrids**

Trace Ancestor

vcfHunter

KDE Classifier

Visualization

- Ideogram Chromosome Painting
- CircosJS CircosJS Client to build interactive graphs in a circular layout

TraceAncestor prefilter

→ Sélection des marqueurs diagnostiques

Données manquantes

Indices de différenciation (GST) par ancêtre (chacun en comparaison aux 3 autres)

Fréquence de l'allele ALT

#CHROM	POS	REF	ALT	%Nref	GSTA1	GSTA2	GSTA3	GSTA4	FA1	FA2	FA3	FA4
1	85524	A	G	0.3103448276	0.2	0.2	0.2	1	0	0	0	1
1	108710	A	T	0.6206896552	0.2	1	0.2	0.2	0	1	0	0
1	108741	T	A	0.2413793103	0.2	0.2	1	0.2	0	0	1	0
1	109226	A	T	0	0.2	0.2	0.2	1	0	0	0	1
1	109661	A	G	0.3448275862	0.2	0.2	1	0.2	0	0	1	0
1	110915	A	C	0.3448275862	1	0.2	0.2	0.2	0	1	1	1

-Tri en fonction des données manquantes (< 0.3 par défaut)
-Tri en fonction des valeurs de GST (> 0.9 par défaut). Si GST fort pour un ancêtre→ la diversité allélique totale à cette position est majoritairement expliquée par cet ancêtre

TraceAncestor prefilter

→ Sélection des marqueurs diagnostiques

Données manquantes

Indices de différenciation (GST) par ancêtre (chacun en comparaison aux 3 autres)

Fréquence de l'allele ALT

#CHROM	POS	REF	ALT	%Nref	GSTA1	GSTA2	GSTA3	GSTA4	FA1	FA2	FA3	FA4
1	85524	A	G	0.3103448276	0.2	0.2	0.2	1	0	0	0	1
1	108710	A	T	0.6206896552	0.2	1	0.2	0.2	0	1	0	0
1	108741	T	A	0.2413793103	0.2	0.2	1	0.2	0	0	1	0
1	109226	A	T	0	0.2	0.2	0.2	1	0	0	0	1
1	109661	A	G	0.3448275862	0.2	0.2	1	0.2	0	0	1	0
1	110915	A	C	0.3448275862	1	0.2	0.2	0.2	0	1	1	1

-Définition de la valeur de l'allèle ancestral -> REF ou ALT?

-Si F > 0.8 → ALT

-Si F < 0.2 → REF

TraceAncestor prefilter

→ Sélection des marqueurs diagnostiques

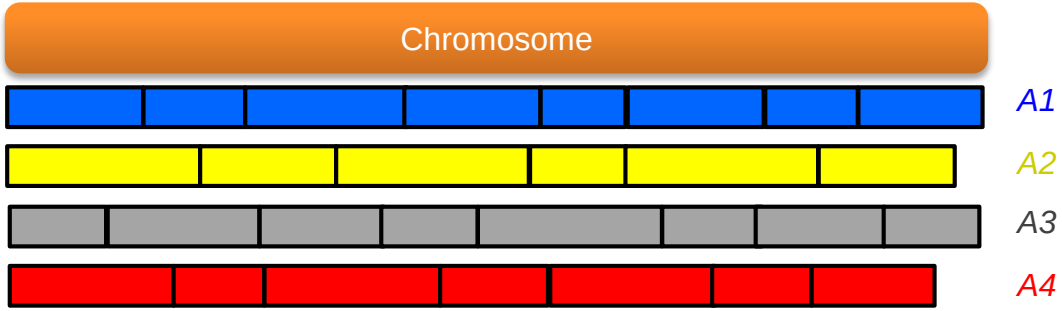
#CHROM	POS	REF	ALT	%Nref	GSTA1	GSTA2	GSTA3	GSTA4	FA1	FA2	FA3	FA4
1	85524	A	G	0.3103448276	0.2	0.2	0.2	1	0	0	0	1
1	108710	A	T	0.6206896552	0.2	1	0.2	0.2	0	1	0	0
1	108741	T	A	0.2413793103	0.2	0.2	1	0.2	0	0	1	0
1	109226	A	T	0	0.2	0.2	0.2	1	0	0	0	1
1	109661	A	G	0.3448275862	0.2	0.2	1	0.2	0	0	1	0
1	110915	A	C	0.3448275862	1	0.2	0.2	0.2	0	1	1	1

ancestor	chromosome	position	allele
A3	1	108741	A
A4	1	109226	T

TraceAncestor

Etape 1

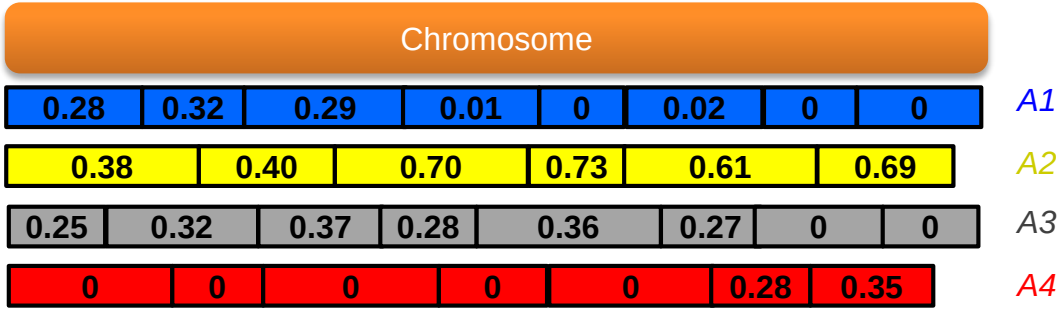
Découpage du chromosome en fenêtres non chevauchantes de 10 SNPs



TraceAncestor

Etape 2

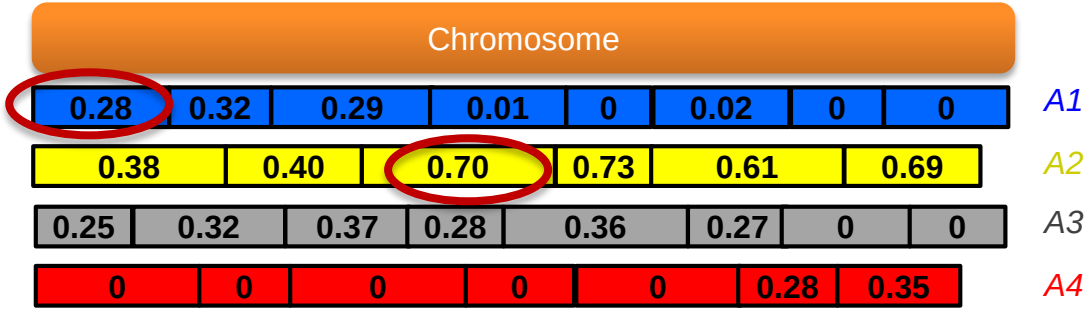
Calcul de la fréquence des reads ancestraux par ancêtre et par fenêtre de 10 SNPs.



TraceAncestor

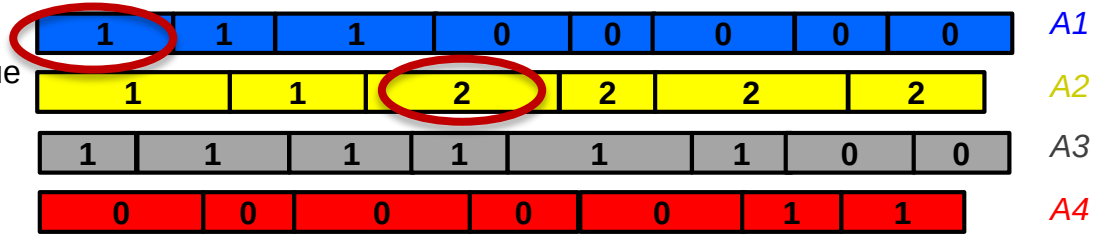
Etape 2

Calcul de la fréquence des reads ancestraux par ancêtre et par fenêtre de 10 SNPs.



Etape 3

Estimation du dosage allélique de chaque ancêtre par fenêtre de 10 SNPs



Test de vraisemblance (LOD) des différentes hypothèses 2 à 2, entre la fréquence observée et théorique pour triploïde

Diploid: 0.05 / 0.5 / 0.95

Triploid: 0.05 / 0.33 / 0.66 / 0.95

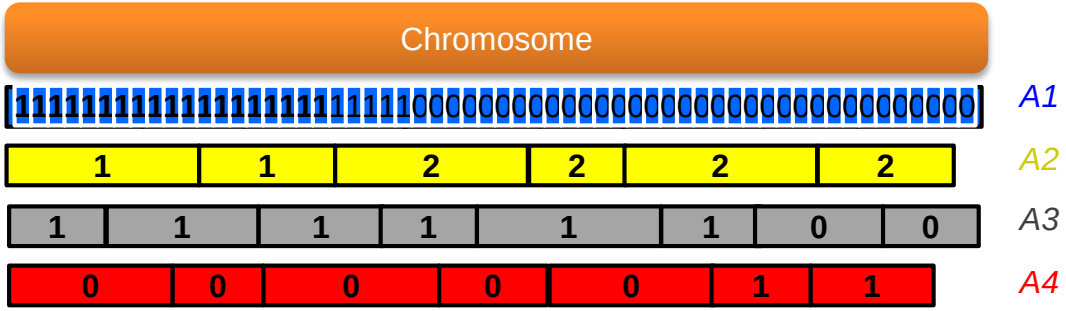
Tetraploid: 0.05 / 0.25 / 0.5 / 0.75 / 0.95

Si (-3 < LOD < 3) => indétermination

TraceAncestor

Etape 3

Estimation du dosage allélique de chaque ancêtre par fenêtre de 10SNP



Etape 4

Division du chromosome en sous-fenêtres non chevauchantes de 100kb. Le dosage allélique des fenêtres de 10 SNP est reporté dans les fenêtres de 100Kb

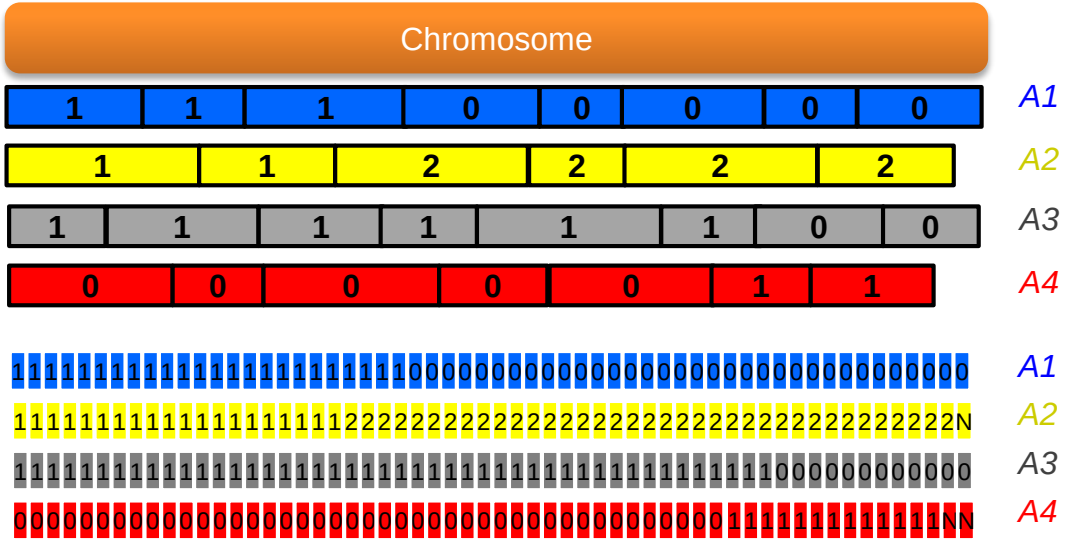
TraceAncestor

Etape 3

Estimation du dosage allélique de chaque ancêtre par fenêtre de 10SNP

Etape 4

Division du chromosome en sous-fenêtres non chevauchantes de 100kb. Le dosage allélique des fenêtres de 10SNP est reporté dans les fenêtres de 100Kb



TraceAncestor

Etape 3

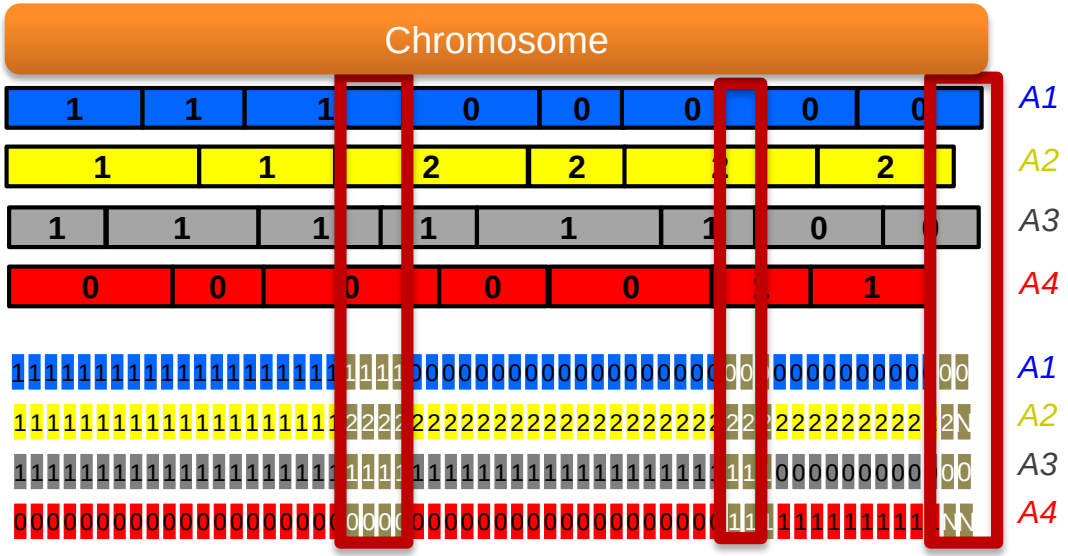
Estimation du dosage allélique de chaque ancêtre par fenêtre de 10SNP

Etape 4

Division du chromosome en sous-fenêtres non chevauchantes de 100kb. Le dosage allélique des fenêtres de 10SNP est reporté dans les fenêtres de 100Kb

Si la somme du dosage allélique de tous les ancêtres pour une fenêtre est différente de la ploïdie:

→ indetermination



TraceAncestor

Etape 3

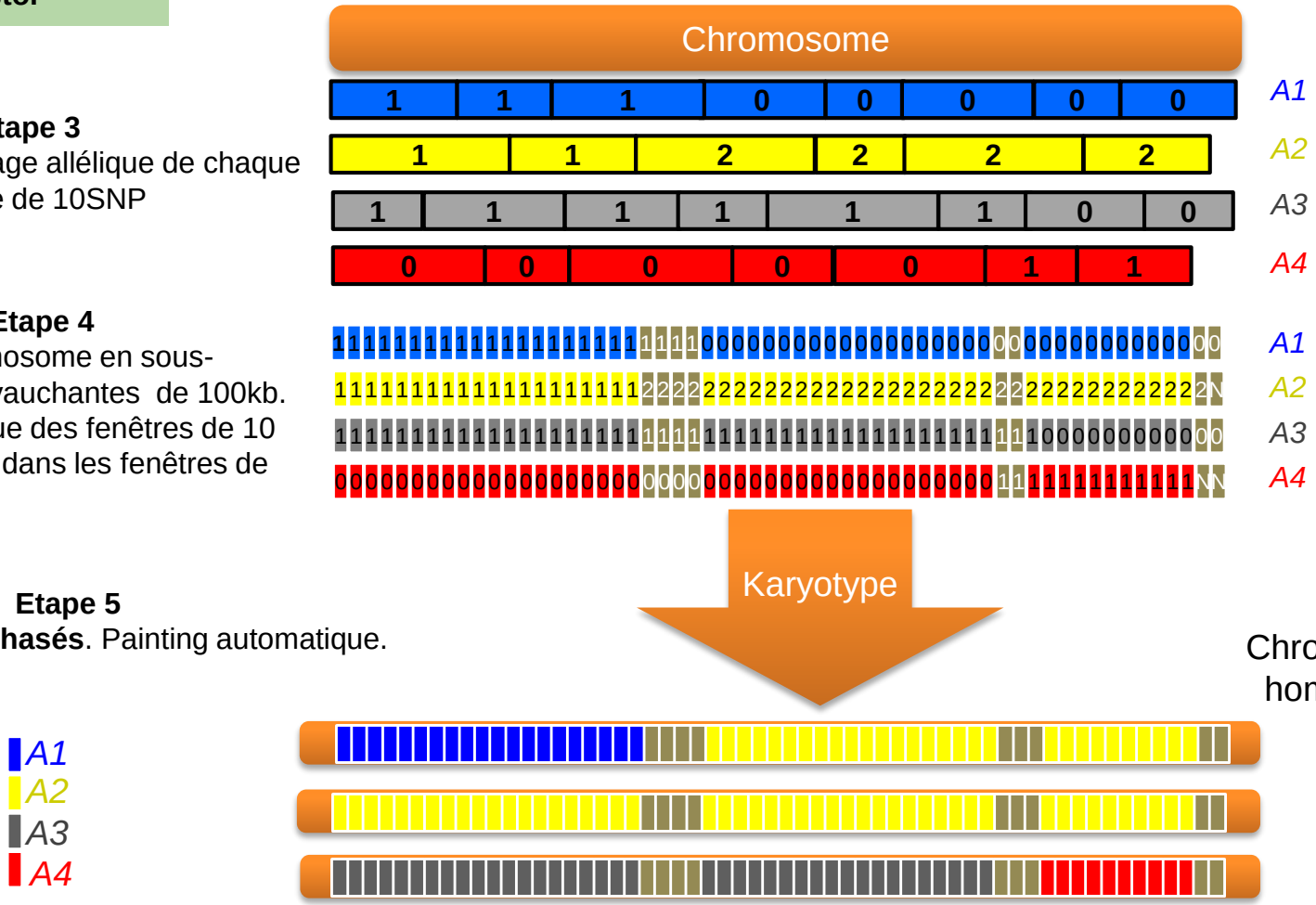
Estimation du dosage allélique de chaque ancêtre par fenêtre de 10SNP

Etape 4

Division du chromosome en sous-fenêtres non chevauchantes de 100kb. Le dosage allélique des fenêtres de 10 SNPs est reporté dans les fenêtres de 100Kb

Etape 5

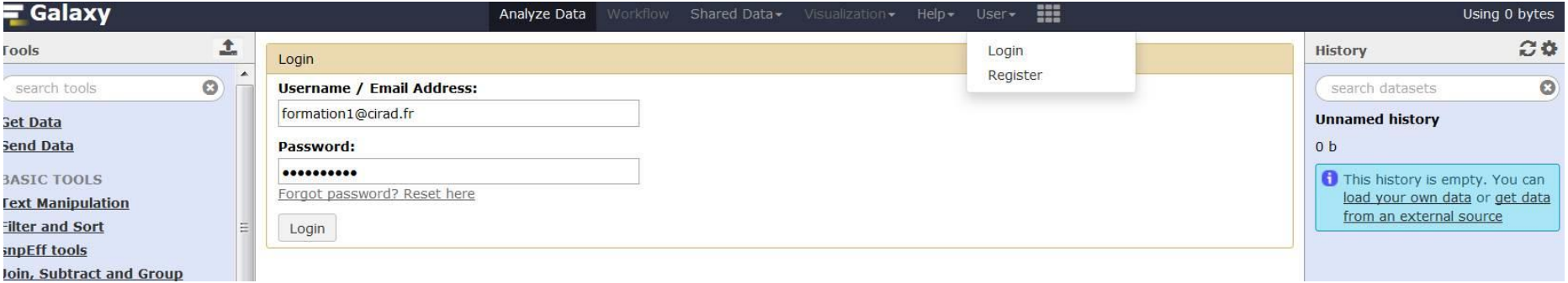
Caryotypes non-phasés. Painting automatique.



UTILISATION DE TRACE ANCESTOR SOUS GALAXY

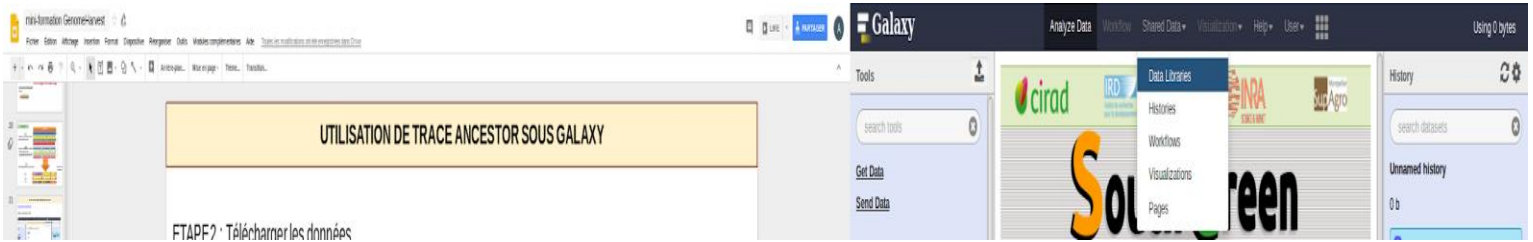
http://galaxy.southgreen.fr/galaxy/

ETAPE 1 : se connecter à galaxy

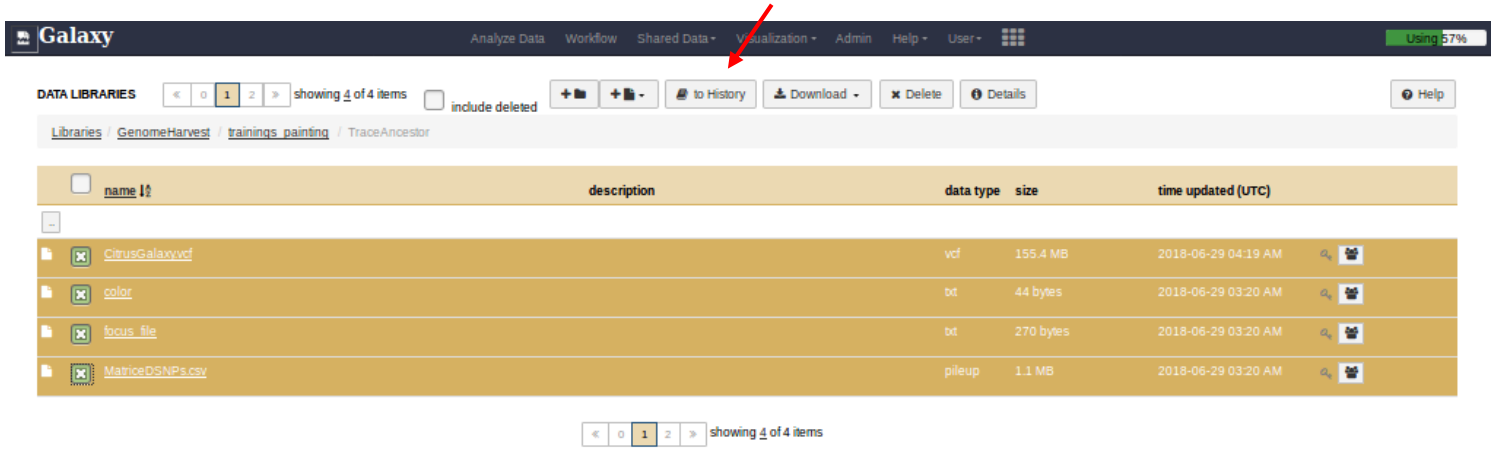


UTILISATION DE TRACE ANCESTOR SOUS GALAXY

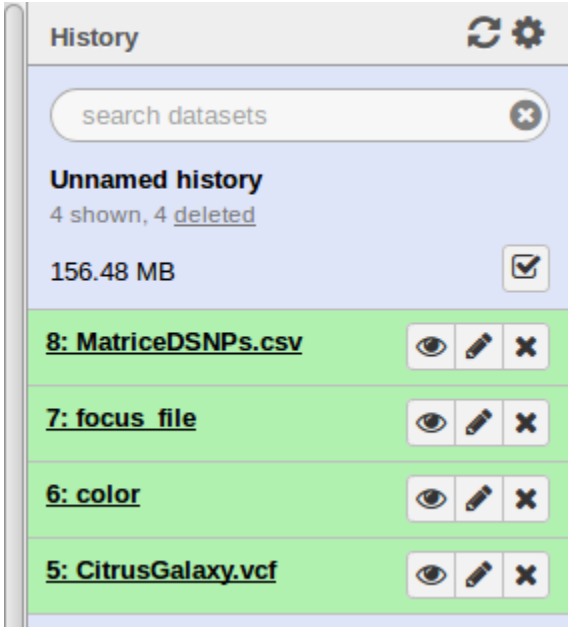
ETAPE 2 : Charger les données tests de la librairie partagée “TraceAncestor” vers l’historique



DataLibrary → GenomeHarvest → trainings_painting → TraceAncestor



UTILISATION DE TRACE ANCESTOR SOUS GALAXY



Matrice contenant les données GST

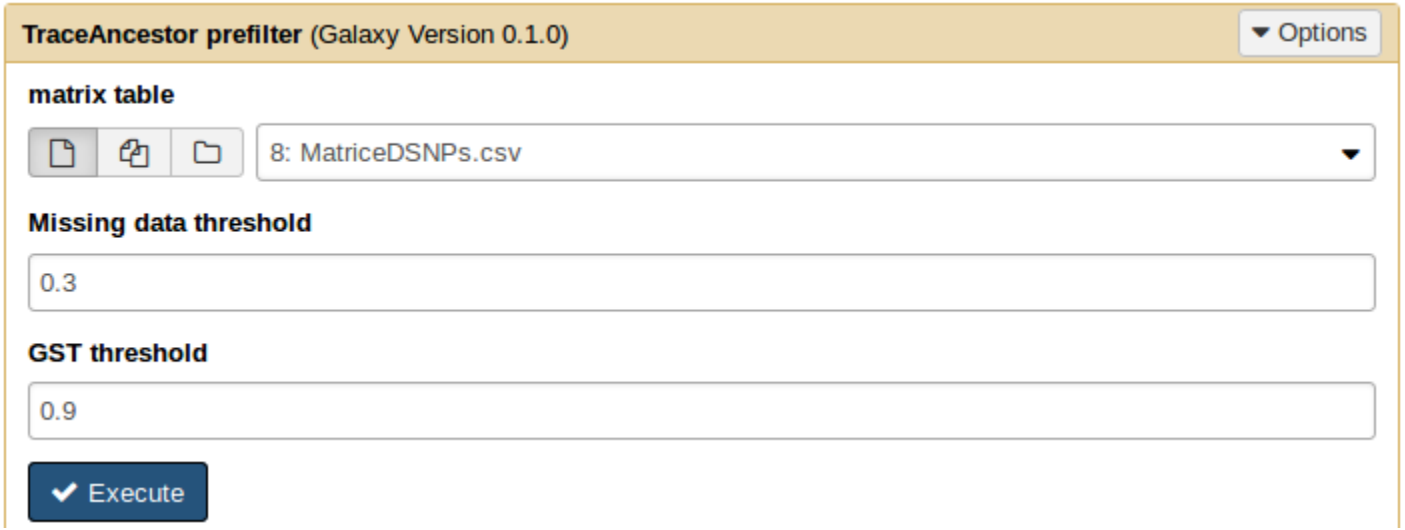
Fichier contenant des noms d'hybrides spécifiques sur lesquels on veut réaliser le painting

Association couleur - ancêtre

VCF des hybrides

UTILISATION DE TRACE ANCESTOR SOUS GALAXY

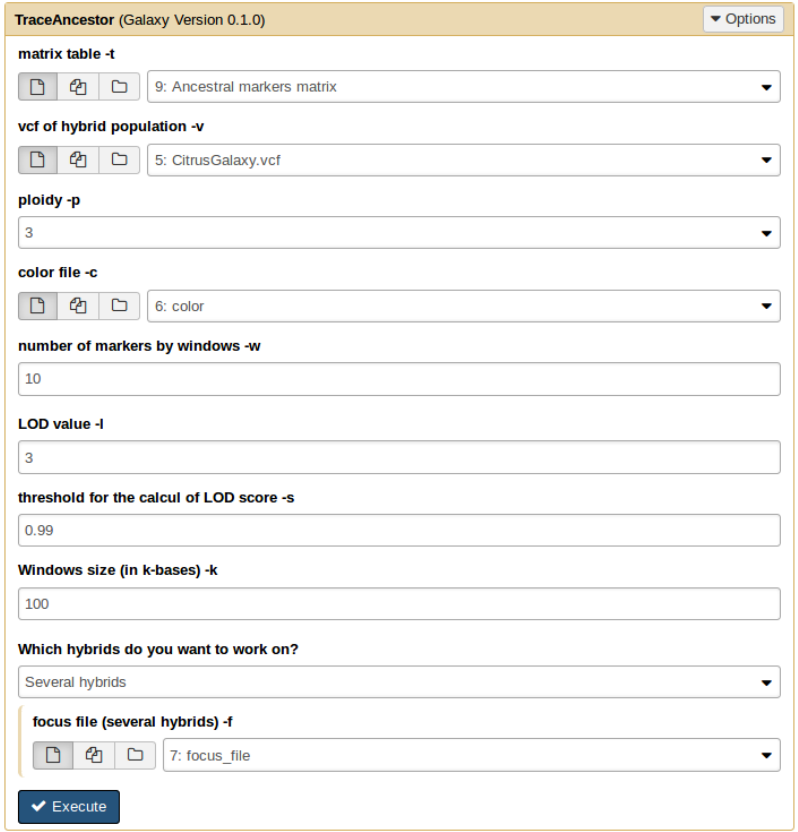
ETAPE 3 : Lancer TraceAncestor prefilter pour obtenir la matrice des marqueurs diagnostiques



→ Output : *Ancestral markers matrix* (matrice contenant les marqueurs diagnostiques filtrés)

UTILISATION DE TRACE ANCESTOR SOUS GALAXY

ETAPE 4 : Lancer TraceAncestor pour obtenir les fichiers de blocs de chromosomes à coloriser



- Nombre de marqueurs par fenêtres
- Valeur du LOD à partir de laquelle une hypothèse est acceptée
- Taux d'erreurs acceptée
- Taille des sous-fenêtres
- Choix du focus pour le painting:
  - -Un individu
  - -Plusieurs individus
  - -Tous les individus

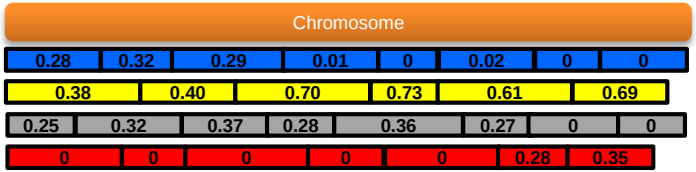
UTILISATION DE TRACE ANCESTOR SOUS GALAXY

OUTPUTS

→ Différents outputs si on choisit de faire le focus sur un seul individu

Ancestors frequency along the chromosomes

Hybrid	Ancestry	Chromosome	Position_Start	Position_End	Frequence
Sample61	A1	1	1	1029641	0.3038



Circos Painting

1	1	1700000	#DF0101	Sample61
1	1700001	2200000	#B9B9B9	Sample61

Chromosomes length for circos

1	28919326
2	36354460

Ideogram Painting

1	0	1	1700000	#DF0101
1	0	1700001	2200000	#B9B9B9

Chromosomes length for ideogram

Sample61	305908623	012
Sample62	305908623	012

UTILISATION DE TRACE ANCESTOR SOUS GALAXY

ETAPE 5 : Visualisation

App web circos : http://genomeharvest.southgreen.fr/visu/circosJS/demo/index.php

Circos

CircosJS CircosJS Client to build interactive graphs in a circular layout (Galaxy Version 0.0.1)

Values for Chromosome Length

169: Chromosomes length for circos

track

1: track

Track name

stack

Track type

Stack

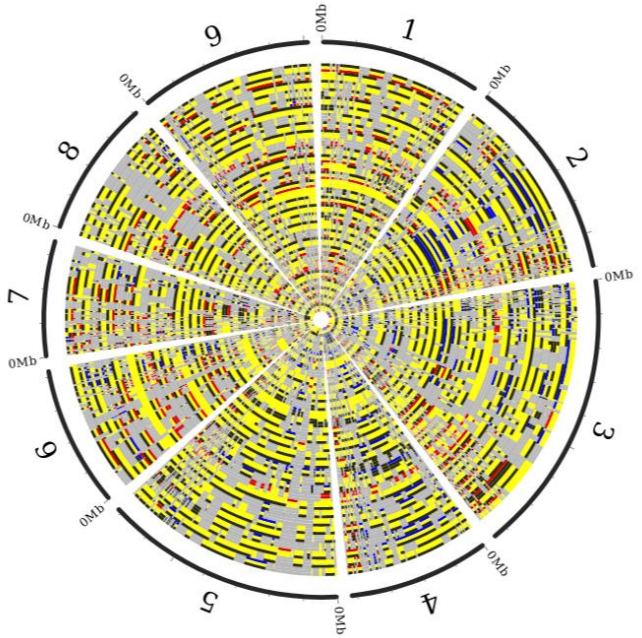
Track data

168: circos painting

Insert track

Track goes here

Execute

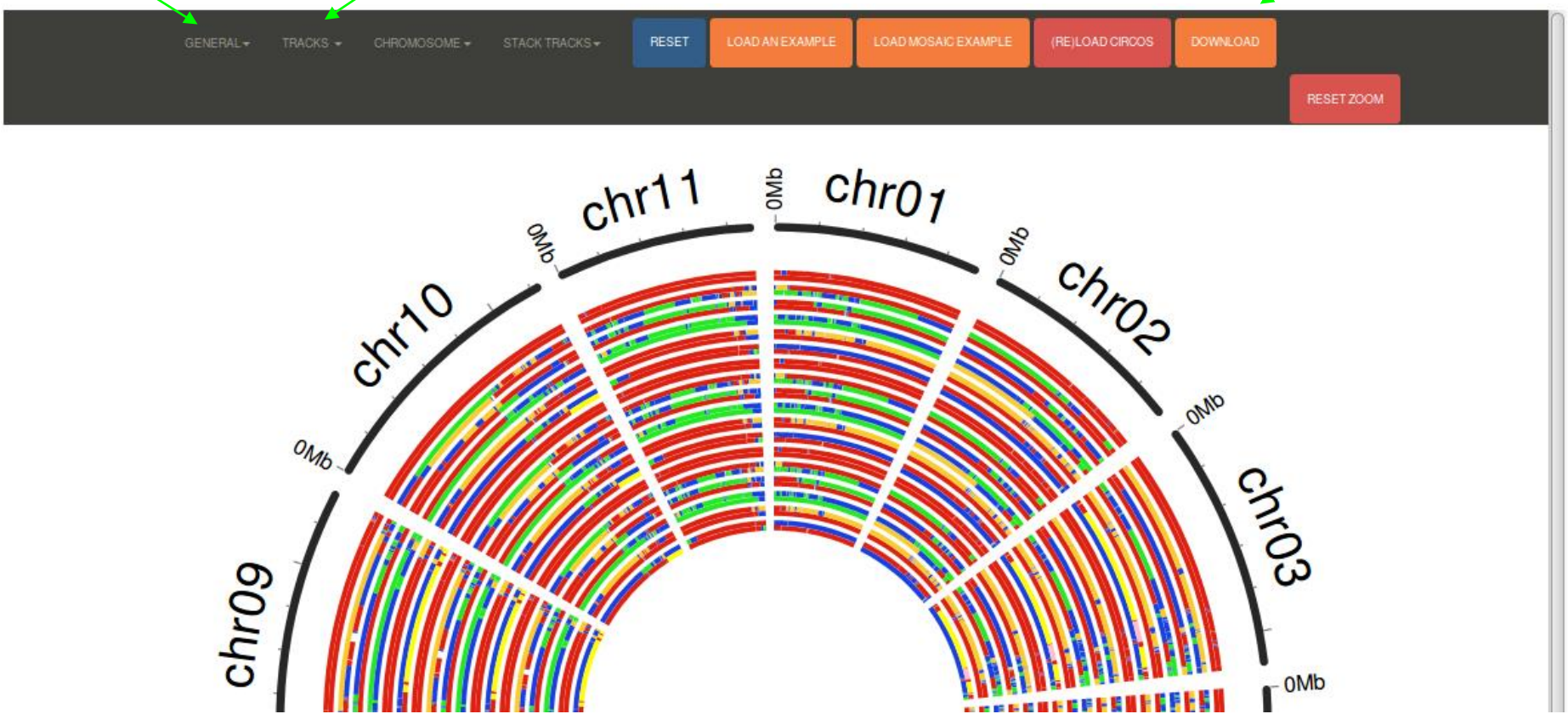


UTILISATION DE TRACE ANCESTOR SOUS GALAXY

Longueur des chromosomes

Chro - début - fin - couleur - individu

Télécharger en png



UTILISATION DE TRACE ANCESTOR SOUS GALAXY

ETAPE 5 : Visualisation

App web ideogram: http://genomeharvest.southgreen.fr/visu/ideogram/newindex.php

Ideogram

Ideogram Chromosome Painting (Galaxy Version 0.0.1)

Values for Chromosome Length

167: Chromosomes length for ideogram

chr length

Ancestral Blocks

166: Ideogram painting

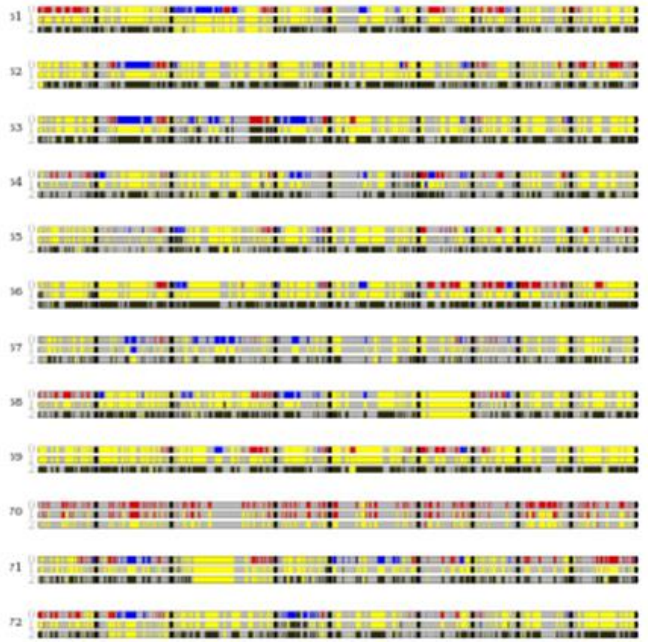
chr haplotype start end #color

Ploidy

3

ploidy

Execute

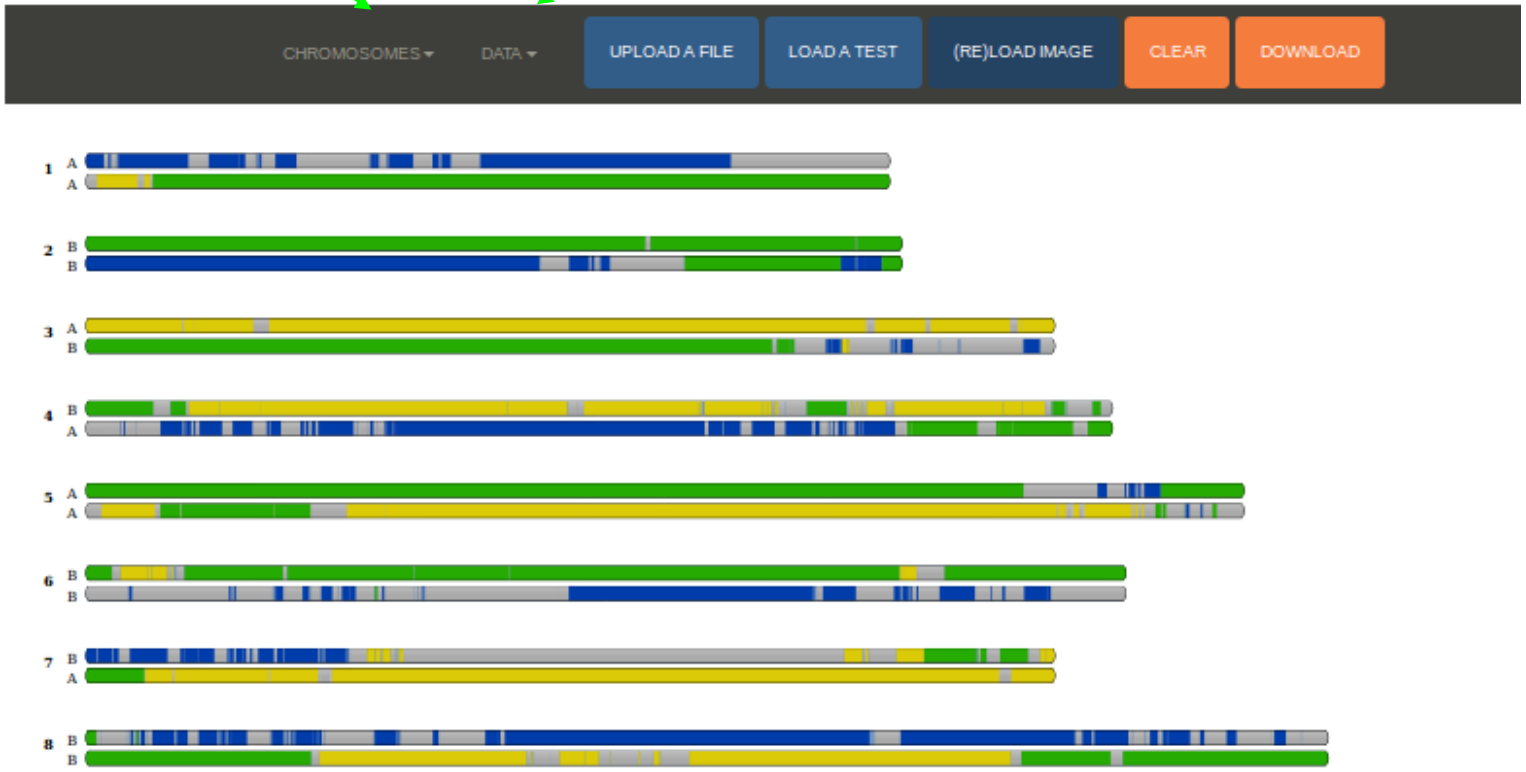


UTILISATION DE TRACE ANCESTOR SOUS GALAXY

Longueur des chromosomes

Chro - haplo - début - fin - couleur

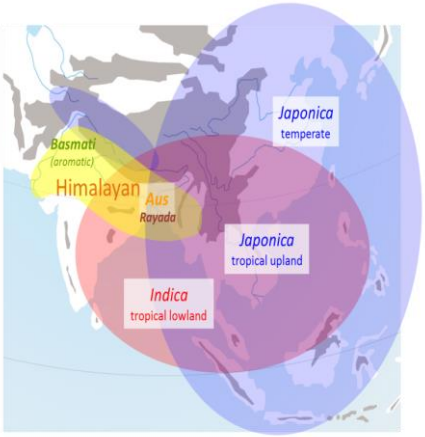
Télécharger en png

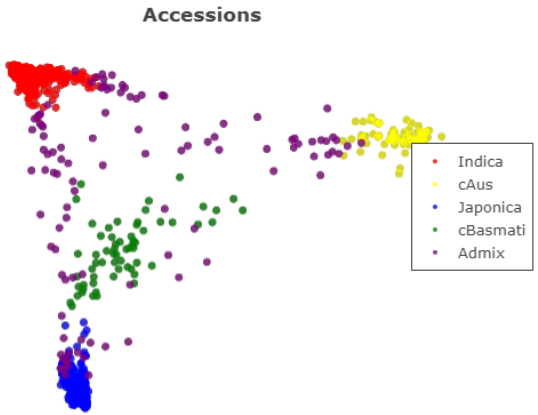


KDE_Classifier

JOAO SANTOS

-Etude de la structure mosaïque des riz aromatiques = hybrides entre variétés japonica / aus / autres
-Quelles régions du génome ont une origine aus, japonica, autres groupes ou outliers?



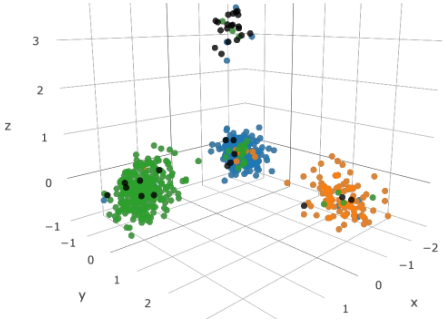


-**Méthode** : basée sur une estimation par noyaux (Kernel Density Estimation (KDE)) avec classe intermédiaire et classe outlier par fenêtre le long du génome

-Taux d'hétérozygotie très bas

1) 3 entrées:
Fichier génotypage (VCF ou geno+bim+fam)
Liste individus de référence
Liste des individus en admixture.

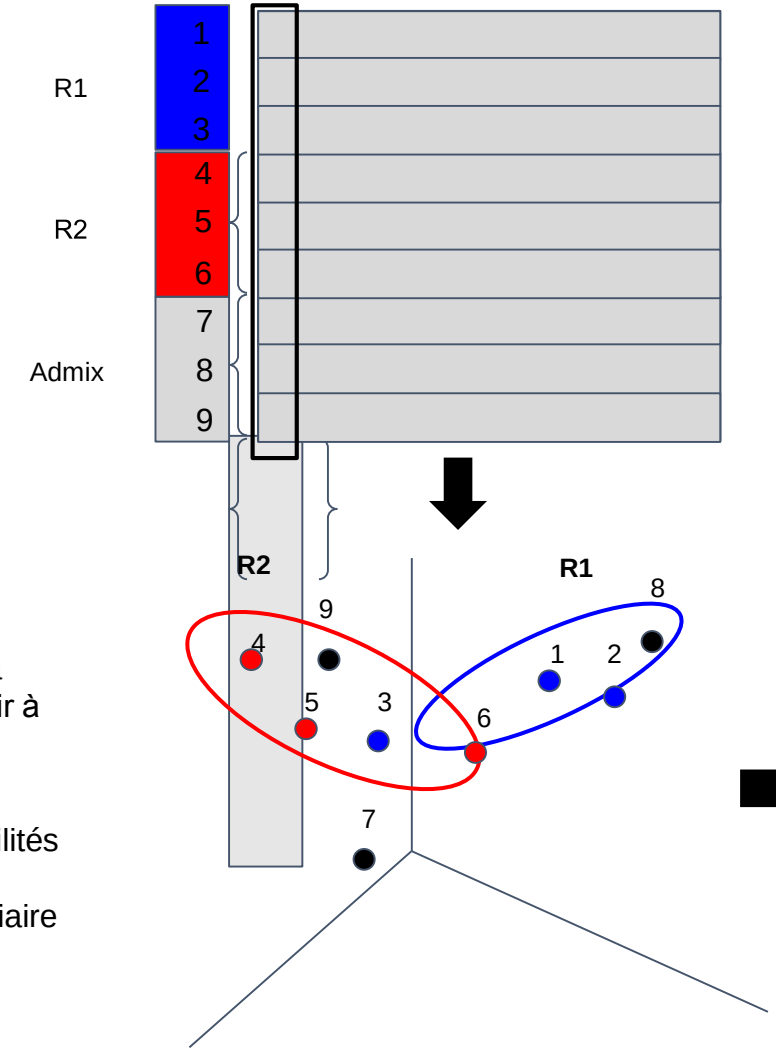
2) ACP par fenêtre le long du génomes.



3) Kernel Density Estimation => calcul de la probabilité de chaque haplotype d'appartenir à chaque distribution reference

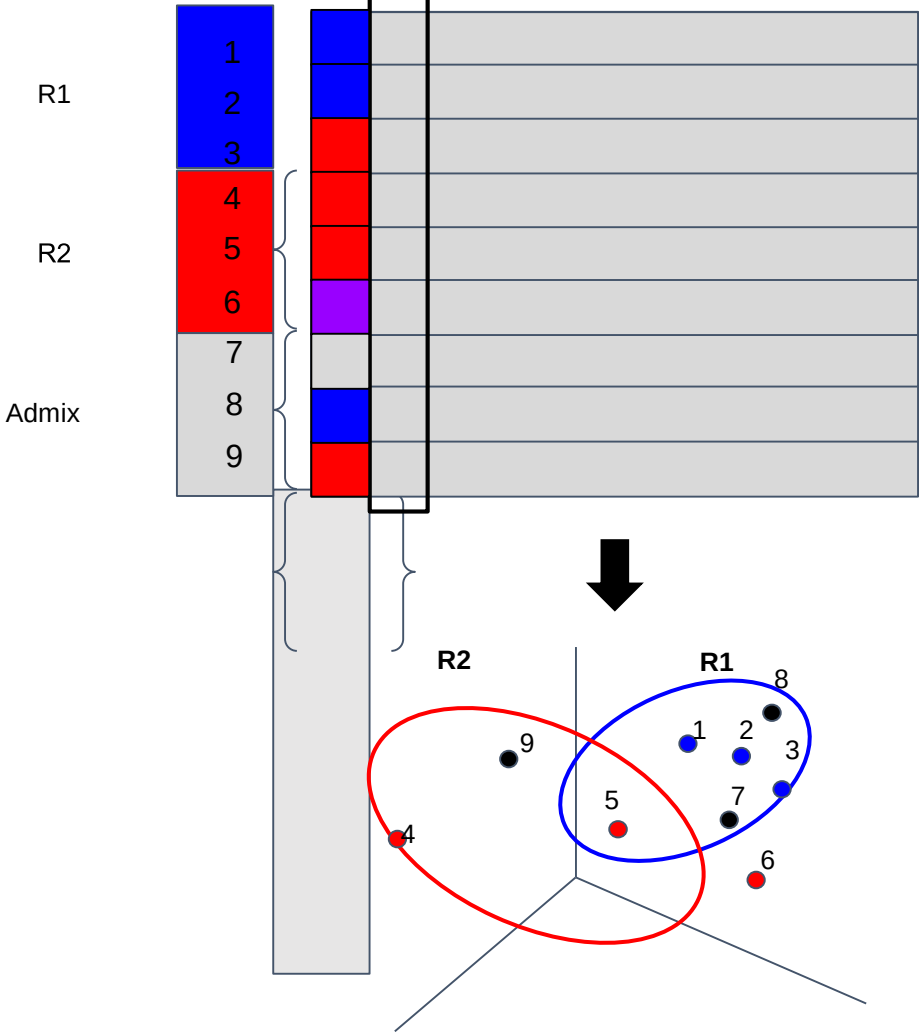
4) Pour un individu on compare ces probabilités pour chaque référence deux à deux
Comparaison > seuil -> pur sinon intermédiaire

DIfférences global vs local

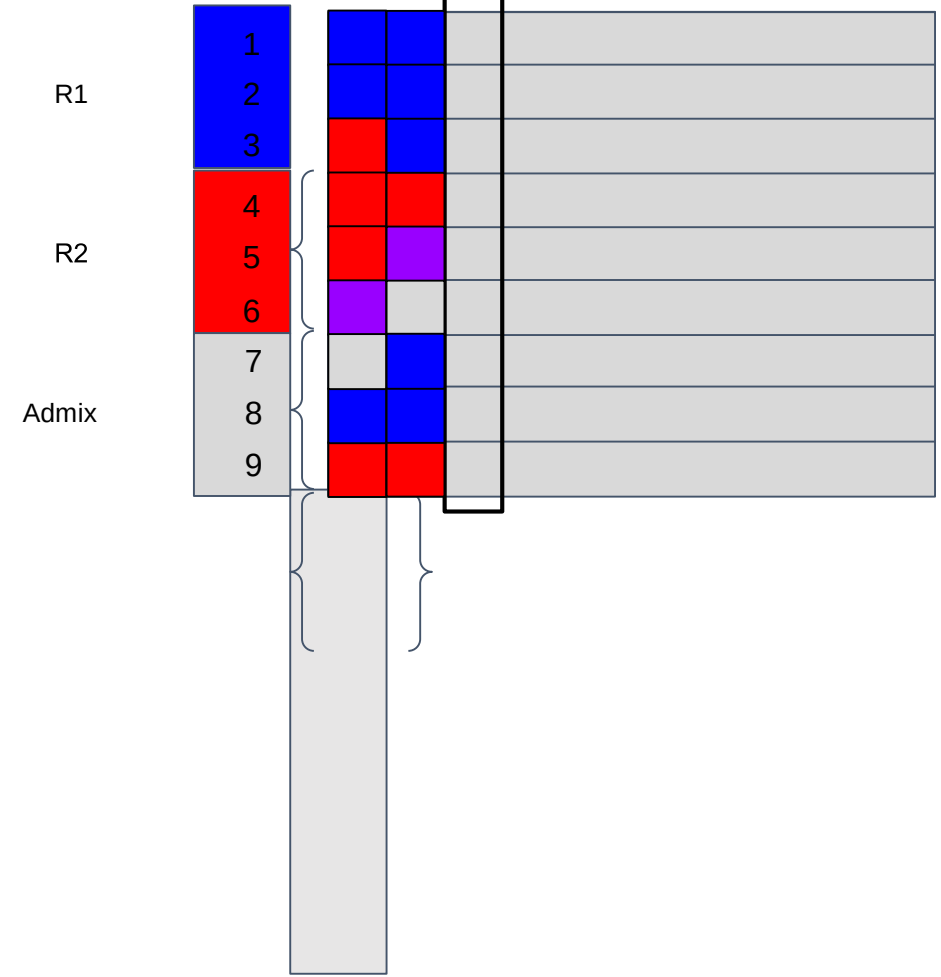


	R1	R2
1	+	-
2	+	-
3	-	+
4	-	+
5	-	+
6	+	+
7	-	-
8	+	-
9	-	+

R1
R2
Admix



		R1	R2
R1	1	+	-
	2	+	-
	3	+	-
R2	4	-	+
	5	+	+
	6	-	-
Admix	7	+	-
	8	+	-
	9	-	-



		R1	R2
R1	1	+	-
	2	+	-
	3	+	-
R2	4	-	+
	5	+	+
	6	-	-
Admix	7	+	-
	8	+	-
	9	-	-

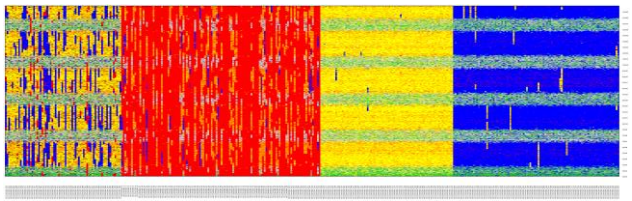
1 chromosome

Parallélisation par blocs (multi-threading)

Vcf ou Geno + structure population

KDE

ideogram



Plusieurs chromosomes

Parallélisation par blocs (multi-threading) +
Parallélisation par chromosome (collection Galaxy)

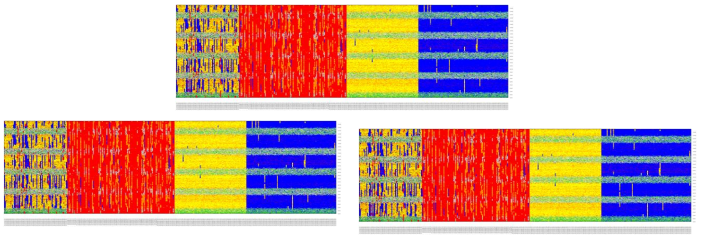
Vcf

structure population

vcfSpliter

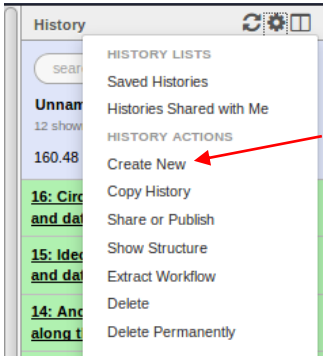
KDE

ideogram

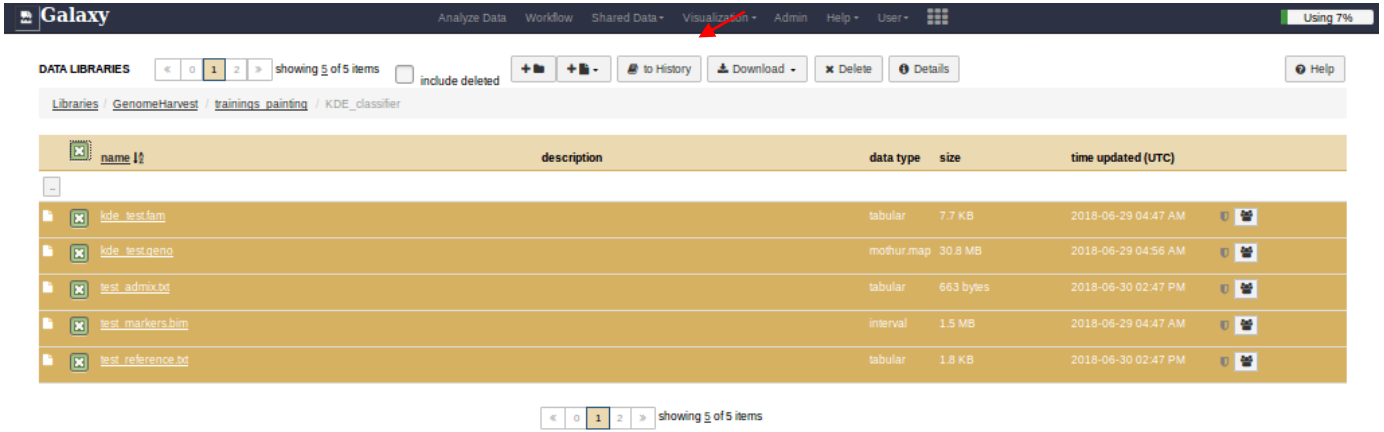


UTILISATION DE KDE_CLASSIFIER SOUS GALAXY

ETAPE 1 : Créer un nouvel historique

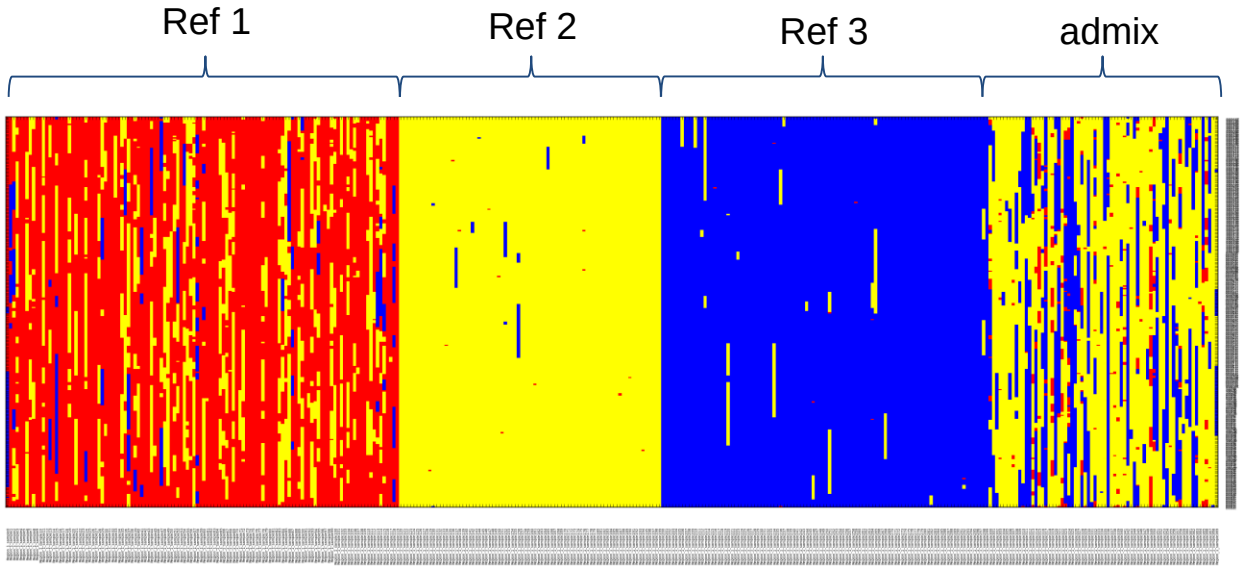


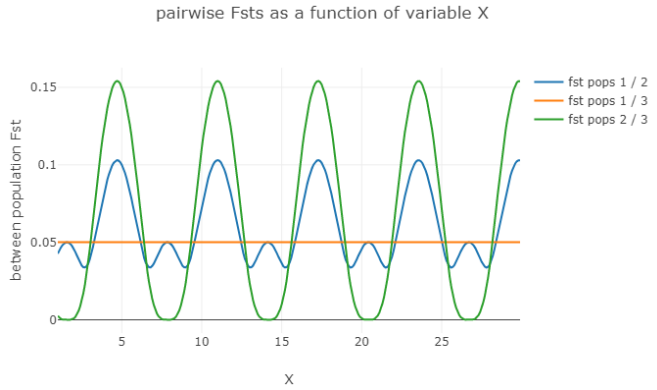
ETAPE 2 : Charger les données tests de la librairie partagée “KDE_Classifier” vers l’historique



UTILISATION DE KDE_CLASSIFIER SOUS GALAXY

Données simulées:
-3 refs (dont une compliquée très admixée -> R1) + admix
-Variations de la distance génétique entre deux populations
→ 2 distances proches = Intermédiaire





Format du fichier de génotypage (geno ou VCF)

Fichiers de génotypage

Focus sur un chromosome en cas de chromosomes multiples

Fichiers de structure

Taille de la fenêtre pour l'ACP (en SNPs)

Taille des chevauchements de fenêtre (en SNPs)

Méthode de clustering

Fréquence d'allèles hétérozygotes acceptée par locus

Choix de la technique de réduction de dimensions (ACP, NMF) + paramètres correspondants

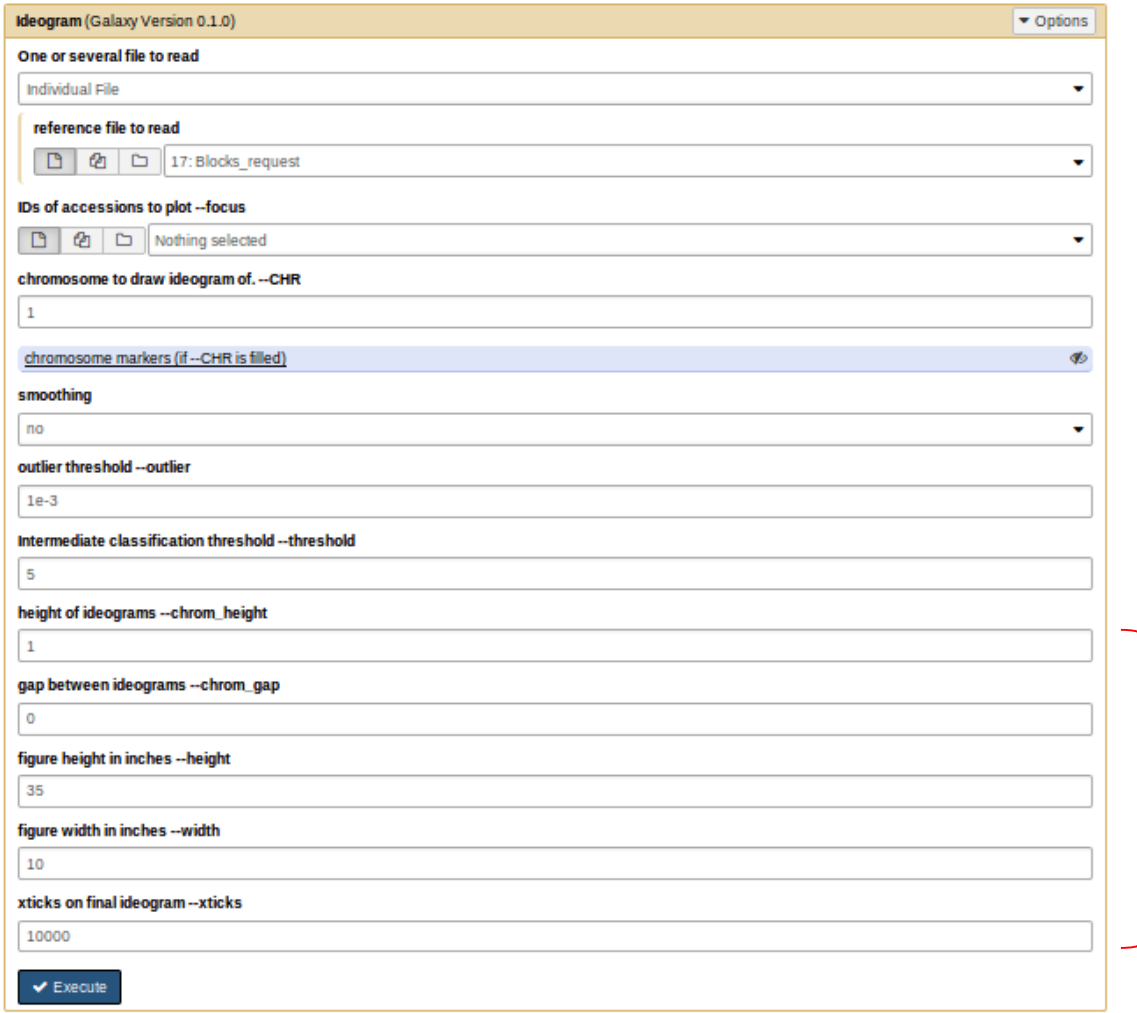
UTILISATION DE KDE_CLASSIFIER SOUS GALAXY

OUTPUTS

Blocks_request

CHR	In	Out	Ref	Sample_300
1	10000	10664	0	2.59e-13
1	10000	10664	1	0.56
1	10000	10664	2	0.65

→ Résultats de KDE
-Valeurs de dispersions par distribution : p-value pour chaque individu vis à vis de chacune des distributions références à chaque fenêtre analysée
-La comparaison puis le painting se font dans ideogram



1 fichier ou 1 collection

Sortie de KDE

Fichier de focus

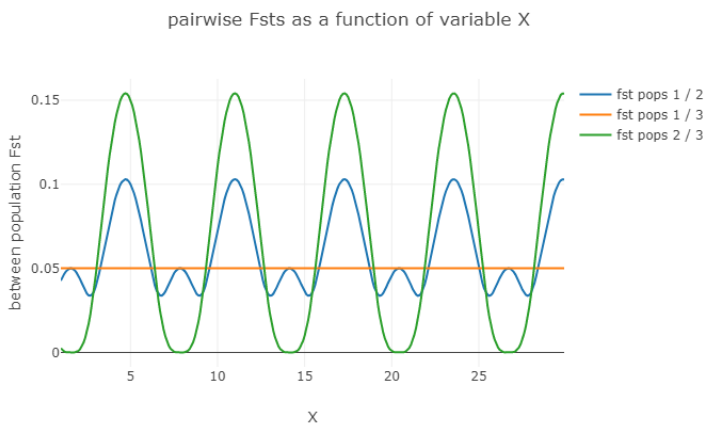
Chromosome à visualiser

Début + fin de la zone de painting d'un chromosome

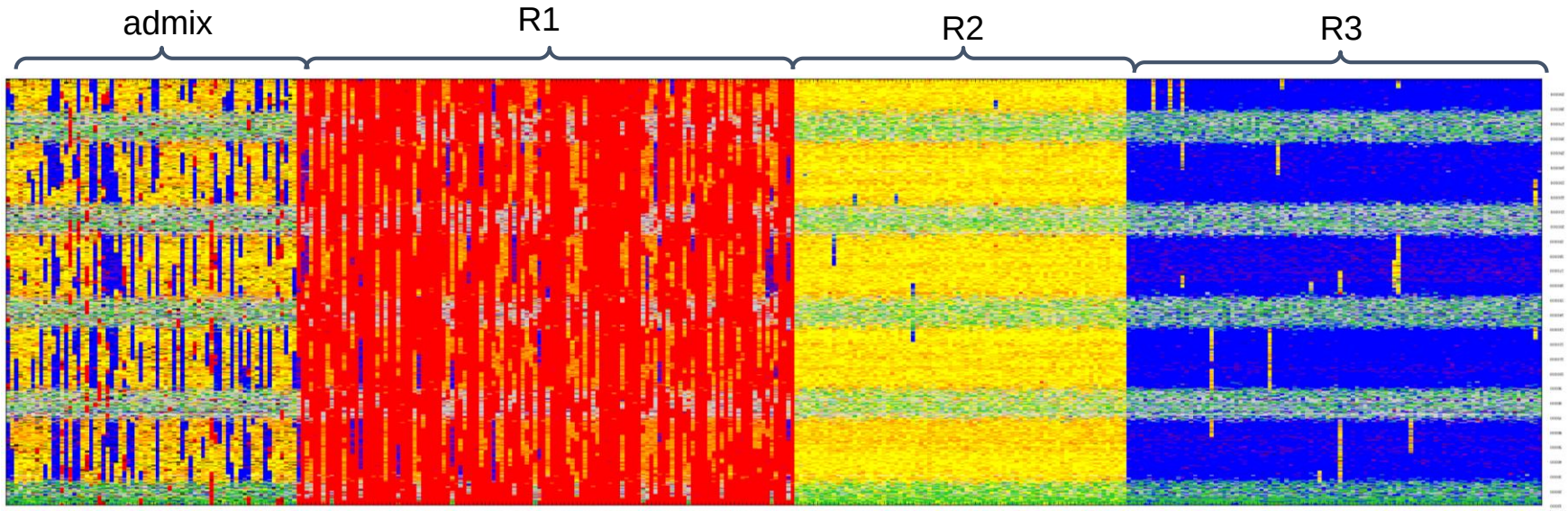
Seuil pour déterminer si un individu est outlier pour une fenêtre

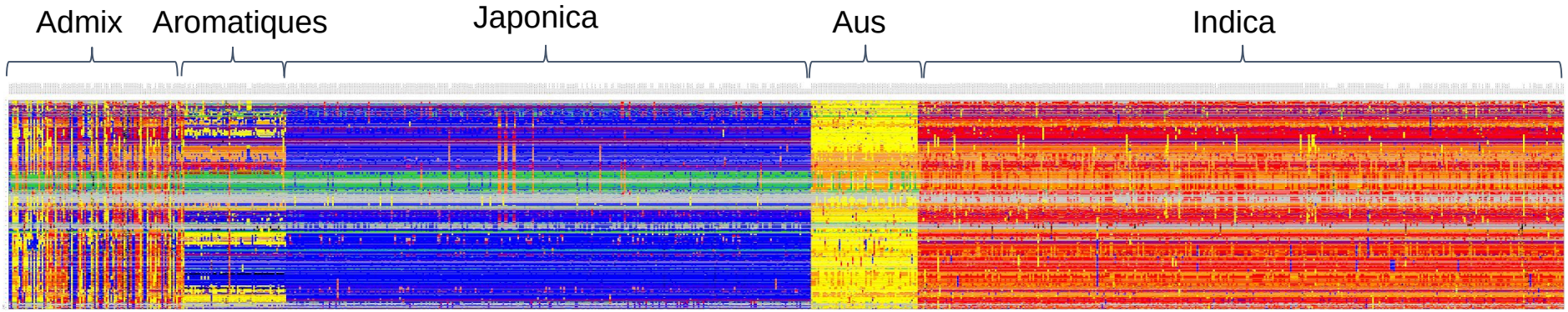
Seuil pour déterminer si un individu est intermédiaire ou pur pour une fenêtre

Paramètres de dessin



→ Variation de la distance génétique des populations simulées le long du chromosome artificiel




Admix
Aromatiques
Japonica
Aus
Indica

VCFHunter

GUILLAUME MARTIN

-Étudier la structure mosaïque du génome
-Approche exploratoire
-Développé pour travailler sur un jeu de données banane

-**Contraintes** : Hybrides hétérozygotes et difficulté à phaser.
Peu de représentants ancestraux actuels (calcul de gst difficile).

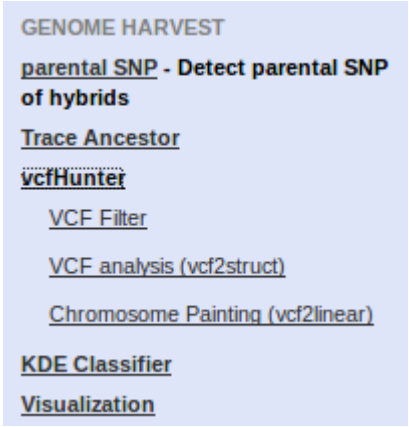
-**Méthode** : Approche par analyses multivariées (ACP) puis clusterisation (attribution des allèles à des groupes ancestraux)

-3 outils dans galaxy:

-**VCF Filter** : Filtre un vcf en fonction de différents paramètres

-**VCF Analysis** (vcf2struct): Analyses multiples sur un vcf (statistiques, ...)

-**Chromosome Painting** (vcf2linear)



1) VCFHunter FILTER

VCF Filter → **VCF Analysis (STAT)**

Filtrage du VCF et analyse statistique générale des variations génétiques

↓

VCF Analysis (FACTORIAL)

2) ACP

ACP : Projection des allèles (variables) et des individus

↓

3) VCFHunter CLUSTERING

VCF Analysis (SNP_CLUST-MeanShift) → **VCF Analysis (ViSUALIZE_VAR_2D)**

Clusterisation des allèles en fonction de leur position sur les axes
→ assignation des allèles à des groupes ancestraux

↓

4) VCFHunter PAINTING

Chromosome Painting → **Circos**

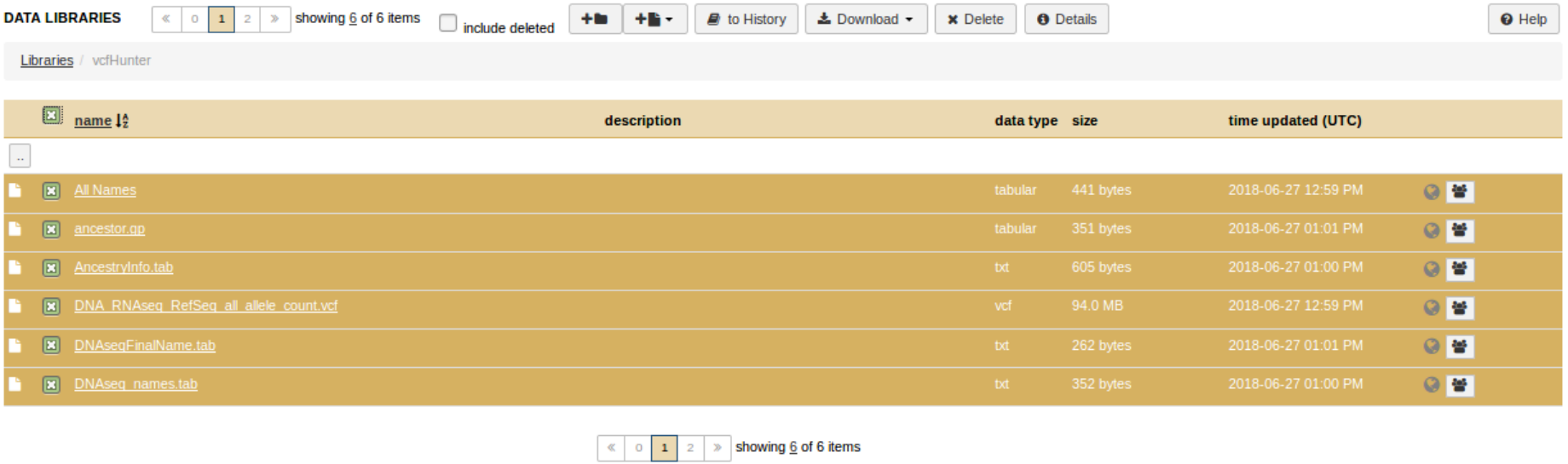
Reconstruction de la structure mosaïque
Painting et visualisation

UTILISATION DE VCFHUNTER SOUS GALAXY

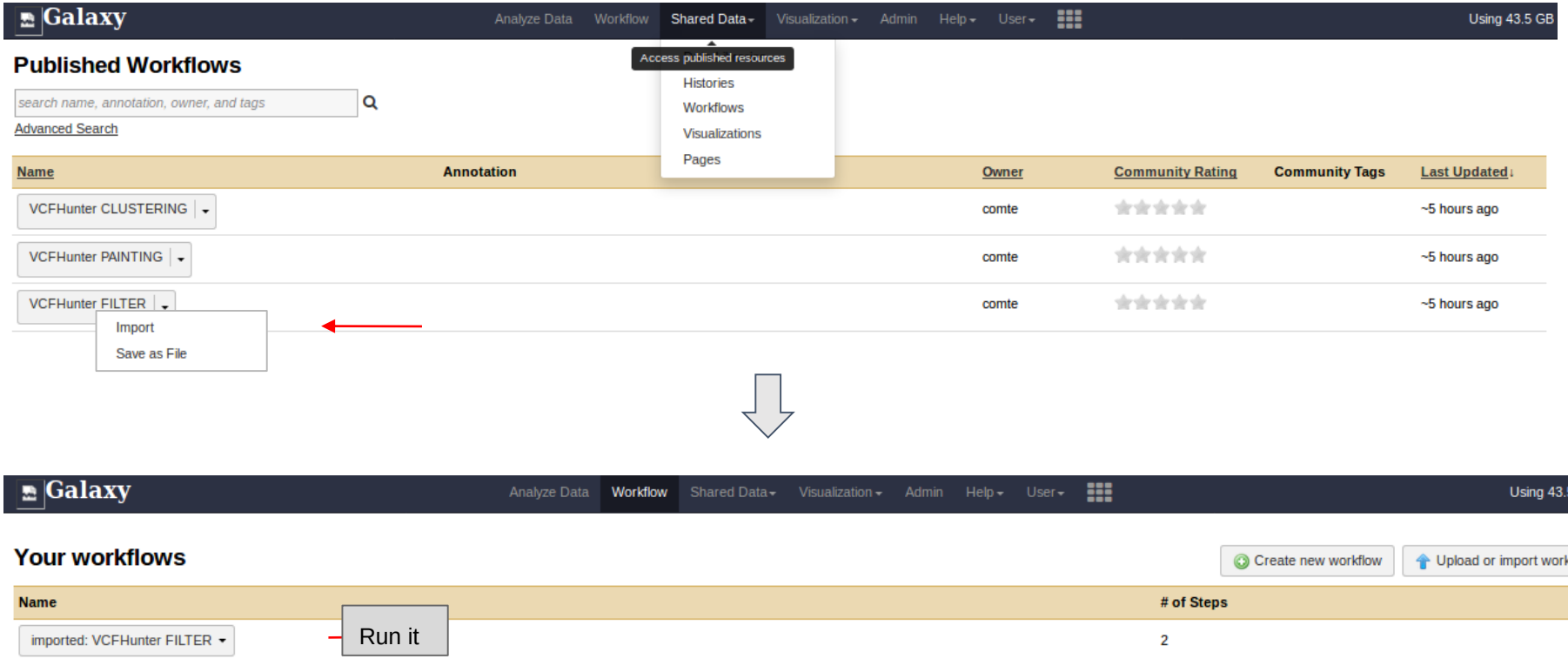
http://cc2-web1.cirad.fr/galaxydev/

ETAPE 1 : Créer un nouvel historique

ETAPE 2 : Charger les données de la library vcfHunter dans le nouvel historique



ETAPE 3 : Utiliser un workflow



1)VCFHunter FILTER

INPUTS:

Step 1: VCF Filter (version 0.1.0)

VCF file --vcf

3: filtered Vcf

names file

1: All Names

outgroup (optional) --outgroup

Selection is Optional

Minimal coverage by accession --MinCov

10

Maximal coverage to keep a genotype --MaxCov

1000

Minimal frequence to keep a genotype --MinFreq

0.05

Minimal allele coverage by accession --MinAl

3

Maximal number of missing genotype --nMiss

Not available.

Number of alleles to remove the site (optional) --RmAlAlt

Not available.

Variant status to filter out (Optional) --RmType

Nothing selected.

-VCF file

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample68	sample69
RefSeq	281	.	C	A,G	.	.	.	GT:AD:DP	0/0:31,2,0:33	0/0:19,0,1:20
RefSeq	282	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20
RefSeq	283	.	T	A,G	.	.	.	GT:AD:DP	0/0:33,1,0:34	0/0:19,1,0:20
RefSeq	284	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20

-Names files = un fichier contenant les accessions à filtrer

-Outgroup = un fichier contenant les accessions à ne pas filtrer mais à garder dans le vcf de sorti.

1)VCFHunter FILTER

Step 1: VCF Filter (version 0.1.0)

VCF file --vcf

3: filtered Vcf

names file

1: All Names

outgroup (optional) --outgroup

Selection is Optional

Minimal coverage by accession --MinCov

10

Maximal coverage to keep a genotype --MaxCov

1000

Minimal frequence to keep a genotype --MinFreq

0.05

Minimal allele coverage by accession --MinAl

3

Maximal number of missing genotype --nMiss

Not available.

Number of alleles to remove the site (optional) --RmAlAlt

Not available.

Variant status to filter out (Optional) --RmType

Nothing selected.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample68	sample69
RefSeq	281	.	C	A,G	.	.	.	GT:AD:DP	0/0:31,2,0:33	0/0:19,0,1:20
RefSeq	282	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20
RefSeq	283	.	T	A,G	.	.	.	GT:AD:DP	0/0:33,1,0:34	0/0:19,1,0:20
RefSeq	284	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20

On filtre les données :

Site coverage ≥ 10
Site coverage ≤ 1000

1)VCFHunter FILTER

Step 1: VCF Filter (version 0.1.0)

VCF file --vcf

3: filtered Vcf

names file

1: All Names

outgroup (optional) --outgroup

Selection is Optional

Minimal coverage by accession --MinCov

10

Maximal coverage to keep a genotype --MaxCov

1000

Minimal frequence to keep a genotype --MinFreq

0.05

Minimal allele coverage by accession --MinAl

3

Maximal number of missing genotype --nMiss

Not available.

Number of alleles to remove the site (optional) --RmAlAlt

Not available.

Variant status to filter out (Optional) --RmType

Nothing selected.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample68	sample69
RefSeq	281	.	C	A,G	.	.	.	GT:AD:DP	0/0:31,2,0:33	0/0:19,0,1:20
RefSeq	282	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20
RefSeq	283	.	T	A,G	.	.	.	GT:AD:DP	0/0:33,1,0:34	0/0:19,1,0:20
RefSeq	284	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20

On filtre les données:

Minor allele frequency (absolute ≥ 3 and relative ≥ 0.05)

1)VCFHunter FILTER

Step 1: VCF Filter (version 0.1.0)

VCF file --vcf

3: filtered Vcf

names file

1: All Names

outgroup (optional) --outgroup

Selection is Optional

Minimal coverage by accession --MinCov

10

Maximal coverage to keep a genotype --MaxCov

1000

Minimal frequence to keep a genotype --MinFreq

0.05

Minimal allele coverage by accession --MinAl

3

Maximal number of missing genotype --nMiss

Not available.

Number of alleles to remove the site (optional) --RmAlAlt

Not available.

Variant status to filter out (Optional) --RmType

Nothing selected.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample68	sample69
RefSeq	281	.	C	A,G	.	.	.	GT:AD:DP	0/0:31,2,0:33	0/0:19,0,1:20
RefSeq	282	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20
RefSeq	283	.	T	A,G	.	.	.	GT:AD:DP	0/0:33,1,0:34	0/0:19,1,0:20
RefSeq	284	.	A	C,T	.	.	.	GT:AD:DP	0/0:34,0,0:34	0/0:20,0,0:20

← Nombre maximal de génotypes manquants sur une ligne

← On peut retirer les sites mono - di - tri ou tetra … allélique

← Enlever des lignes selon des tags de GATK (PASS, LowQual, SnpCluster…)

1)VCFHunter FILTER → STATS

Step 2: VCF analysis (vcf2struct) (version 0.1.0)

Analysis type --type

STAT

VCF file --vcf

Output dataset 'filter' from step 1

names file --names

1: All Names

gff3 file (optional) --gff3

Selection is Optional

← Nom des accessions sur lesquelles réaliser les statistiques

General.stat : statistiques globales sur le VCF filtré

Accession.stat: Statistiques par accessions (données manquantes, allèles spécifiques, nombre de sites homozygotes et hétérozygotes).

→ *en fonction de ce qui est obtenu, l'utilisateur peut choisir de filtrer les données à nouveau.*

2) FACTORIAL (ACP)

VCF analysis (vcf2struct) (Galaxy Version 0.1.0) — Options

Analysis type --type

FACTORIAL

VCF file --vcf

1: filtered Vcf

names file --names

40: DNAseqFinalName.tab

A one column file containing accession names to treat

Axis number to keep for the factorial analysis --nAxes

6

Multivariate analysis type --mulType

coa

group file --group

4: AncestryInfo.tab

A file containing two sections: A section[group] with in col 1 accession name ; col 2 group (UN for unknown group). All group should be in capital letters. A section [color], that define for each group a color for pca drawing (in RGB+alpha percentage, ex: red=1:green=0:blue=0:alpha=0.1)

groups to delete --dGroup

Nothing selected

If passed, all alleles belonging to groups passed to this option will be removed.

Matrix of grouped alleles (with either a GROUP or a K-mean_GROUP column) --mat

Nothing selected

If a K-mean_GROUP column is found, the filter will be performed on this column, else it will be performed on the GROUP one.

Execute

Vcf filtré

Nom des accessions sur lesquelles travailler (accessions ancestrales choisies à partir d'une ACP préliminaire p/e)

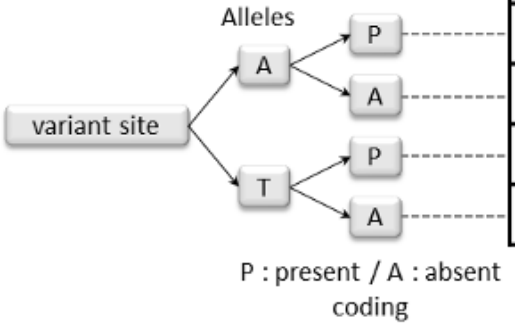
Nombre d'axes pour la représentation

Fichier de structure : Groupes / Couleurs (ne change pas le résultats de l'ACP)

2) FACTORIAL (ACP)

-Le VCF est transformé en une matrice présence(P)/absence(A), avec un allèle par ligne codé en 0,1

matrix4PCA



Variant	GROUP	Accession-1	...	Accession-n
Site1_A_P	UN	0	...	1
Site1_A_A	UN	1	...	0
Site1_T_P	UN	1	...	1
Site1_T_A	UN	0	...	0

T/T A/T

Genotypes

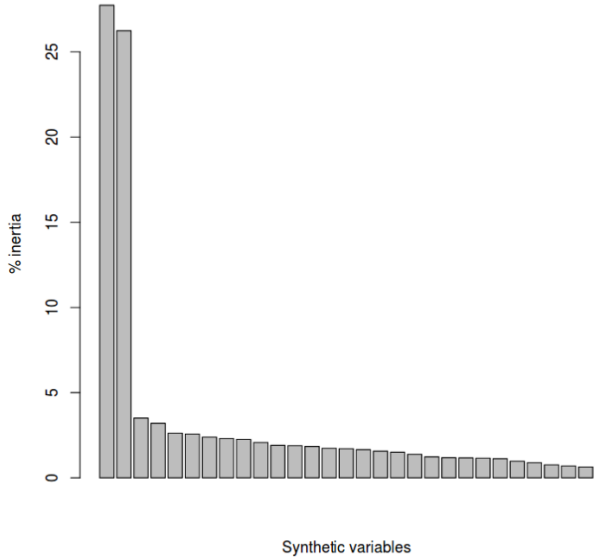
→ On réalise alors une ACP sur cette matrice transposée.

OUTPUTS

projection of accessions and alleles
→ Projection des individus et des allèles sur chaque couple d'axes (ex : 1vs2, 1vs3)

Allele coordinates
+
Individual coordinates

Axis inertia
→ calcul du pourcentage d'inertie pour chaque axe
→ choix des axes à conserver pour l'analyse de clustering
→ ici 1 et 2 explique la plupart de l'inertie



3) VCFHunter CLUSTERING

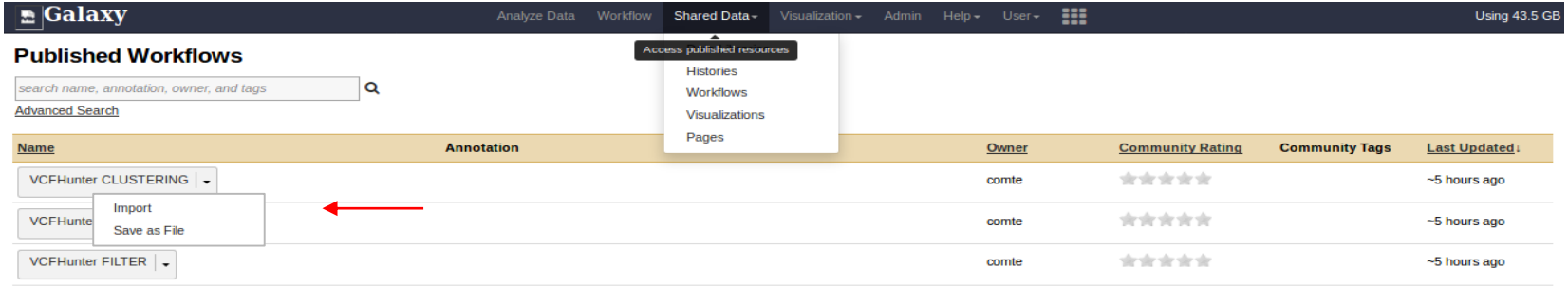
-On considère que la structure ancestrale = structure représentée par les axes de l'ACP

→ **Les allèles aux extrémités des nuages de points sont les allèles ancestraux**

-On cherche alors à les clusteriser.

→ 2 approches implémentées (MeanShift / Kmean). Ici on va utiliser l'approche MeanShift qui permet une détection automatique du nombre de groupes.

-Une fois la clusterisation réalisée, on visualise les groupes. Si un groupe est sous ou sur-représenté, il peut-être intéressant de refaire la clusterisation avec d'autres paramètres jusqu'à obtenir un nombre cohérent de groupes vraiment représentatifs de la structure ancestrale de l'espèce.



3) VCFHunter CLUSTERING

Step 1: VCF analysis (vcf2struct) (version 0.1.0)

Analysis type --type
SNP_CLUST-MeanShift

Allele coordinates --VarCoord
46: allele coordinates

Axes to use in mean shift clustering. Axis should be separated by ':' --dAxes
1:2

The quantile value to estimate de bandwidth parameters used in the MeanShift --quantile
0.15

The recoded vcf file --mat
45: matrix4PCA (recoded vcf)

Cluster all point in the MeanShift --MeanShiftAll
y

Bandwidth value used for mean shift --Bandwidth
Not available.

Cluster absent (A) and present (P) lines --AP
y

Coordonnées des allèles issues de l'ACP

Matrice recodée

Exploratoire. A faire varier pour obtenir la clusterisation optimale (dépend du jeu de données)

3) VCFHunter CLUSTERING

OUTPUTS

kMean allele grouping : même type de recodage de vcf que matrix4PCA mais avec une colonne K-mean_GROUP correspondant au groupe assigné à l'allèle après clusterisation.

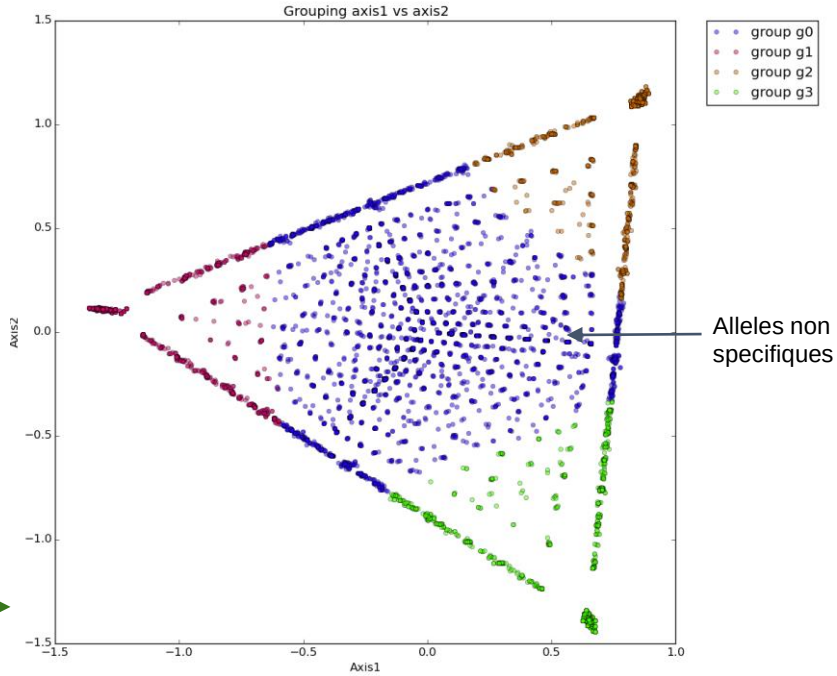
kMean allele gp prop: *“probabilité” de chaque allèle d’être dans chaque groupe*

Centroids group attribution: *association groupe -centroïdes*

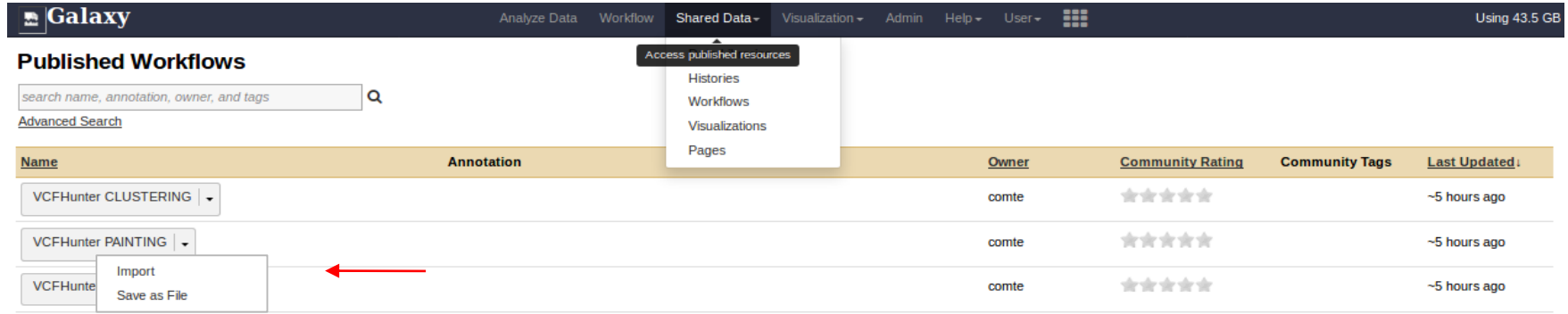
Coordinates of centroid: *coordonnées des centroïdes de chaque groupe*

Color File: *Couleur des groupes issus de la clusterisation*

Visu2D: visualisation des groupes



4) VCFHunter PAINTING



→ Ce programme réalise le painting des accessions le long des chromosomes en fonction des groupes ancestraux.

-Pour un individu, on compte par fenêtre le nombre d'allèles par groupe ancestral

-Comparaison de cette valeur avec une valeur attendue (estimée sur la base des quelques individus représentatifs des groupes ancestraux)

→ calcul du dosage par fenêtre pour cet ancêtre.

4) VCFHunter PAINTING

Running workflow "VCFHunter PAINTING"

Step 1: Chromosome Painting (vcf2linear) (version 0.1.0)

VCF file --vcf

590: filtered Vcf

Matrix file containing allele grouping --mat

586: kMean allele grouping

A one column file containing accession names to treat --names

588: DNAseq_names.tab

A two column file containing accession names used to simulate populations --namesH

589: ancestor.gp

use a reference genome --FormerFolder

no

Chromosomes names to work with --chr

Not available.

Allele number around a variant site to evaluate the structure at the site --win

25

Ploidy level --ploidy

2

Type of estimation performed --type

Binom

Estimator different from sd calculated as mean_value --prop.

0

group color --gcol

587: Color file

A 2 column file containing accession names used to simulate populations and their group --Ambiguous

Selection is Optional

Multiplicator of standard deviation for probability calculation of a segment to be of a group --sdMult

1

Vcf Filtré

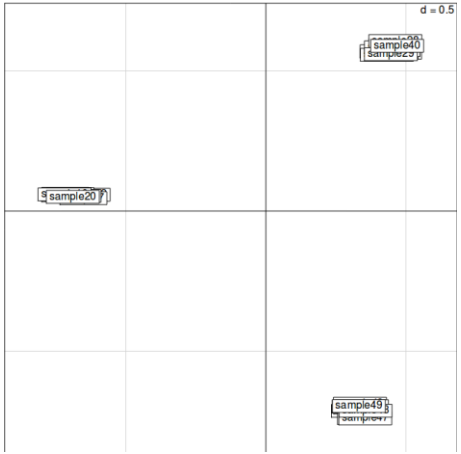
Matrice recodée

Noms d'accessions à étudier

Noms d'accessions pures par groupe servant à simuler une population puis calculer les valeurs attendues

Taille des fenêtre s / 2

Binom plus rapide que Simul



4) VCFHunter PAINTING

OUTPUTS

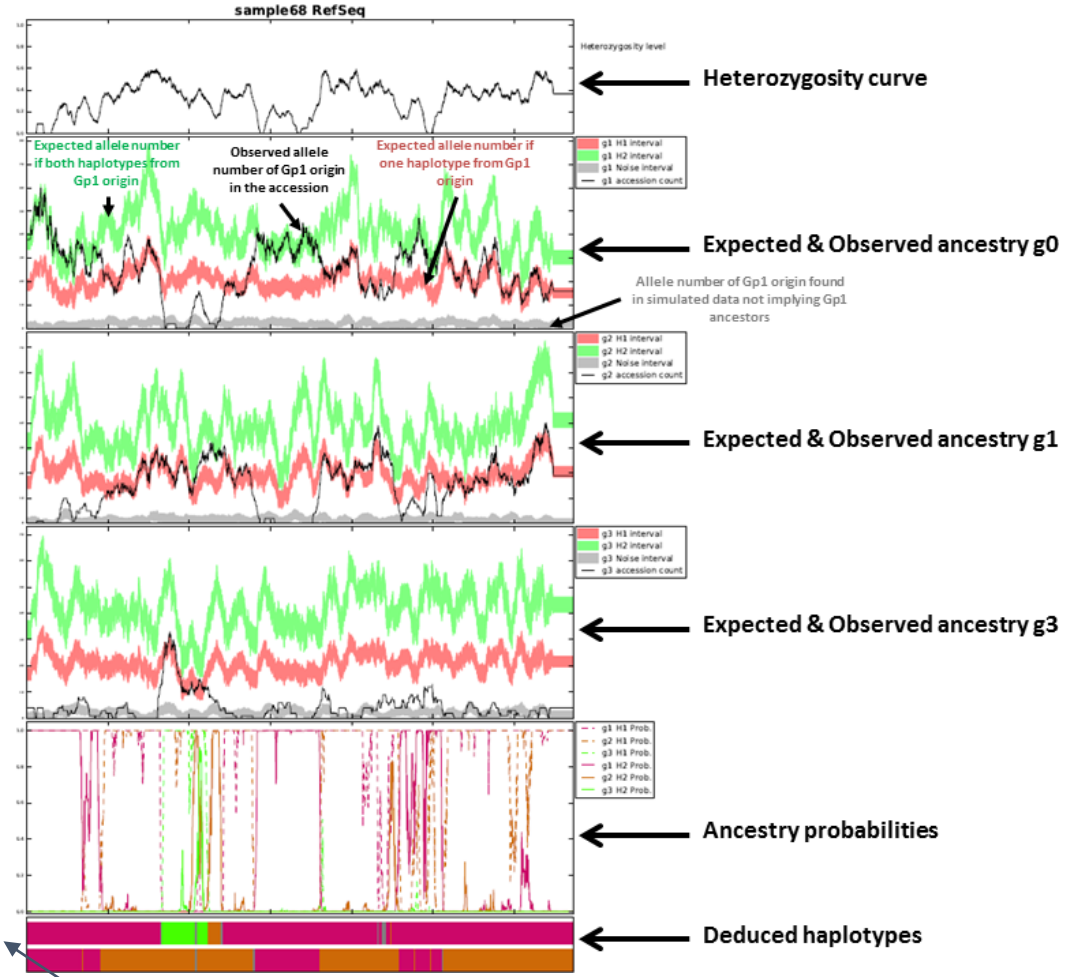
Pour chaque individu:

Chromosome Density
Pour les 3 ancêtres:
-Rose: 1 dose
-Vert: 2 doses
-Noir: observé

-**Chromosome tab**: informations servant à dessiner ces courbes ci-contre

-**Chromosome Haplotype**: positions des blocs mosaïque (x2)

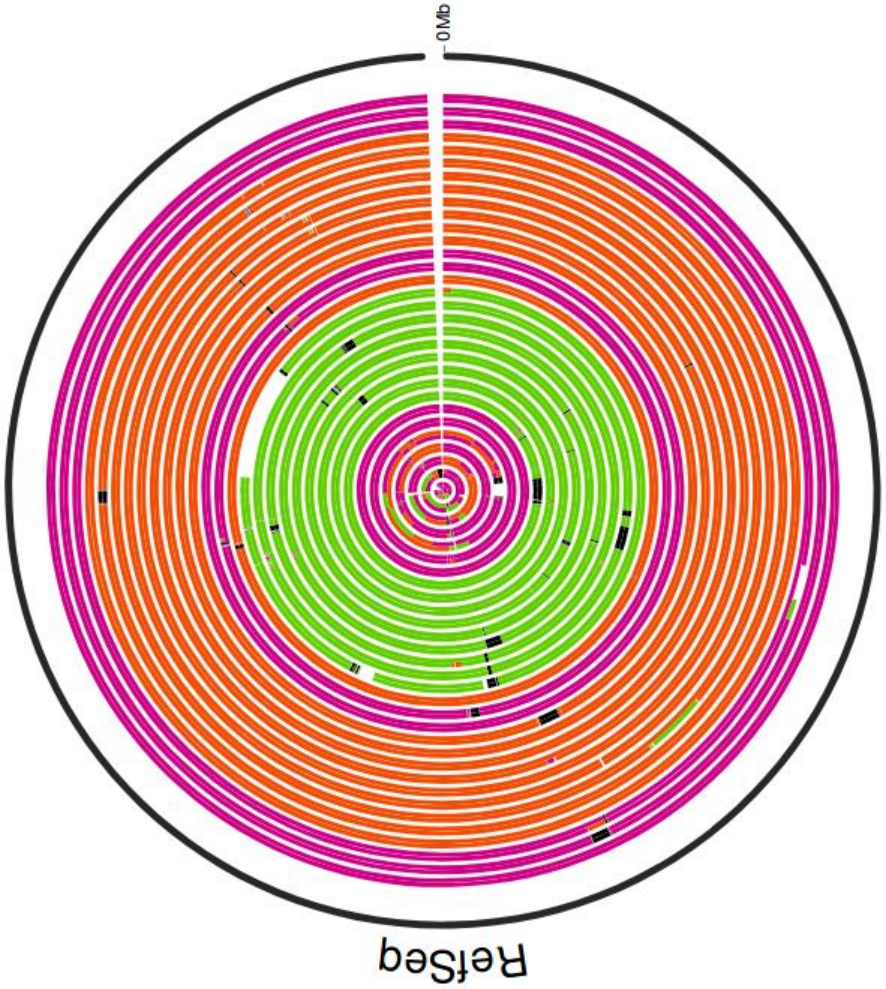
Haplotype resume to use for circos visualization
+
Chromosomes length



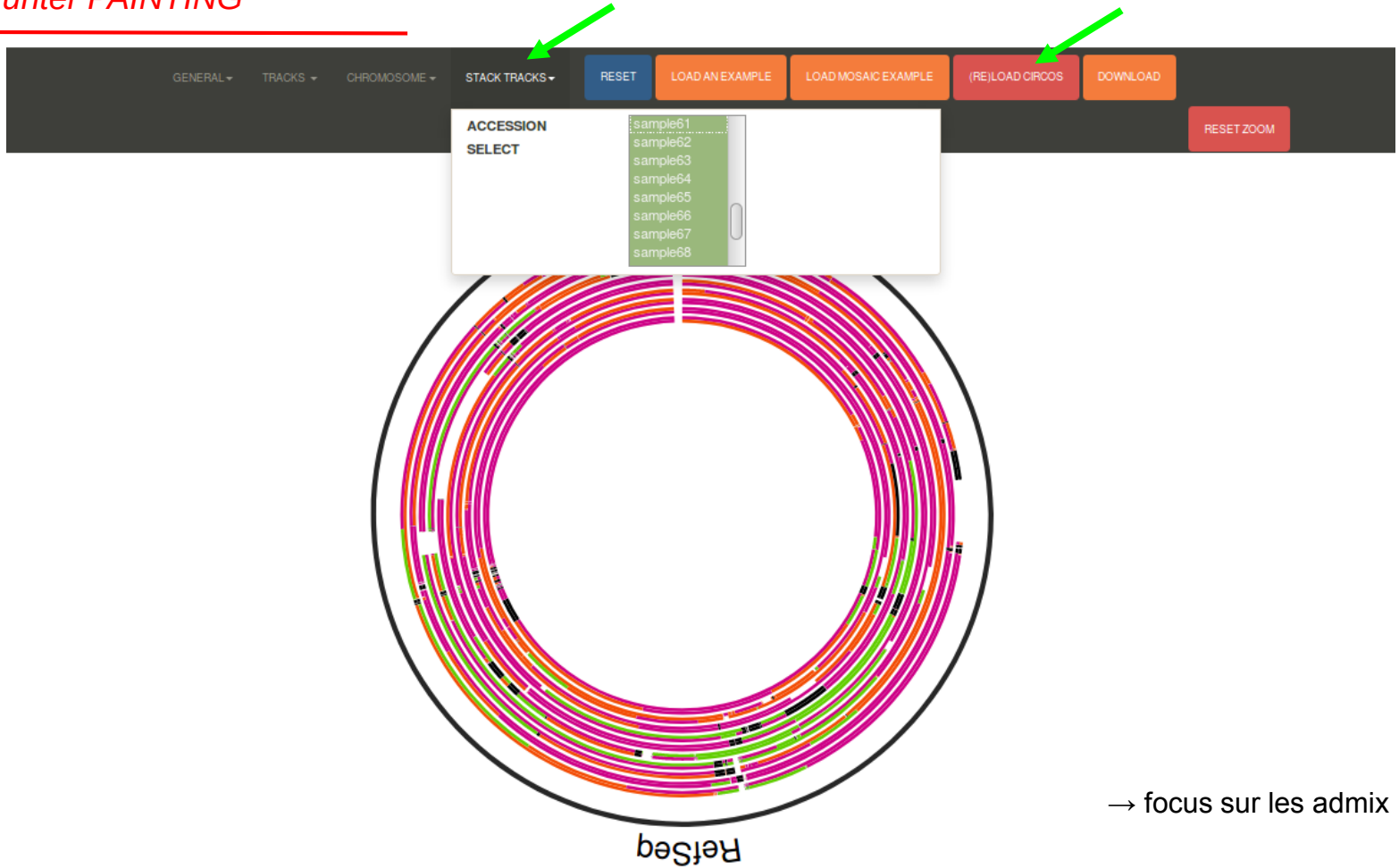
4) VCFHunter PAINTING

OUTPUTS

Pour tous les individus: **Circos**



4) VCFHunter PAINTING



→ focus sur les admix