

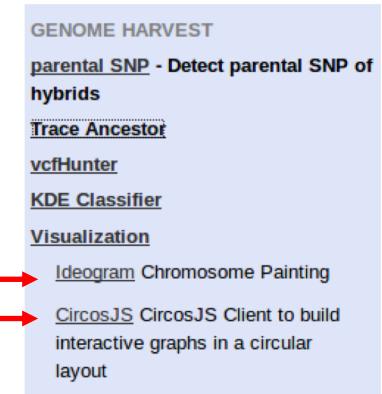
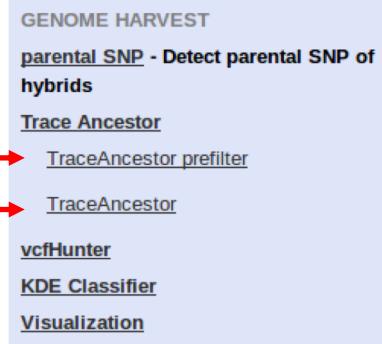
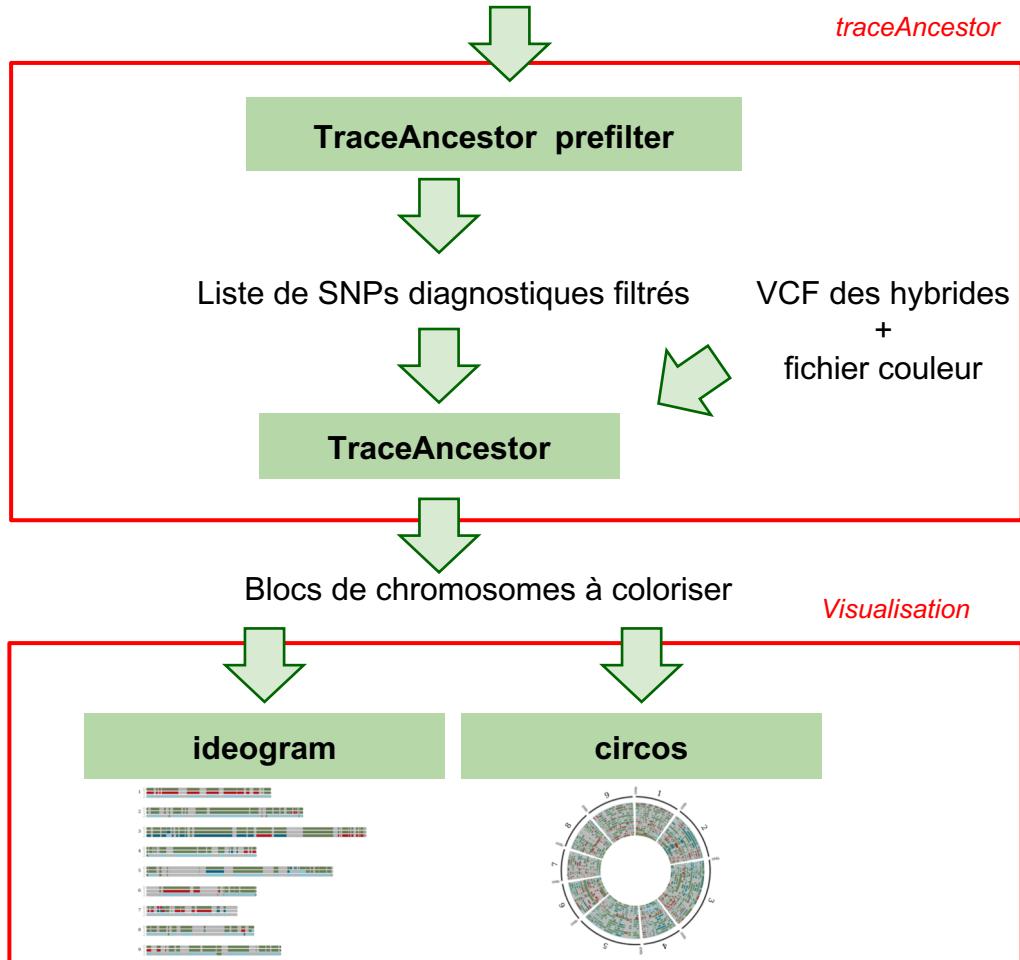


- Les parents/ancêtres sont déjà identifiés (pamplemousse, mandarinier, cédrat, micrantha)
- **Méthode** : Pour un individu donné, mesurer fenêtre par fenêtre sur un chromosome la fréquence de présence de SNPs ancestraux.

2 outils dans galaxy :

- TraceAncestor prefilter
- TraceAncestor

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



## Genotyping by Sequencing (GBS)

Papedas

*Citrus micrantha*



Cédratiers

*Citrus medica*



Mandariniers

*Citrus reticulata*



Pamplemoussiers

*Citrus maxima*



Diagnostic markers of the **four basic taxa** was estimated based on the **G<sub>ST</sub>** parameter (coefficient of gene differentiation which measure the differentiation among subpopulations; Nei, 1973).

We considered **two sub-populations:** **1- the taxa concerned** and **2- a theoretical population of the 3 other basic taxa.**  
The **higher the value is, the more differentiated the taxa are.**

G<sub>st</sub> is defined as the ratio between the inter population diversity and the total diversity:

$$G_{st} = \frac{H_e_{Tot} - H_s}{H_e_{Tot}} = \frac{H_e_{Tot} - \sum h_e/n}{H_e_{Tot}}$$



#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



#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  $F > 0.8 \rightarrow$  ALT
  - Si  $F < 0.2 \rightarrow$  REF

## TraceAncestor prefilter

→ Sélection des marqueurs diagnostiques



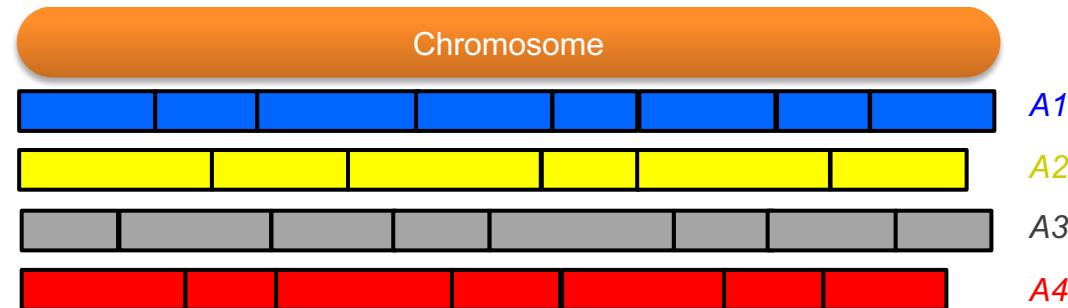
#CHROM	POS	REF	ALT	%Nref	GSTA1	GSTA2	GSTA3	GSTA4	FA1	FA2	FA3	FA4
1	85524	A	G	0.3103448276	0.2	0.2	0.2	1	0	0	0	1
1	108710	A	T	0.6206896552	0.2	1	0.2	0.2	0	1	0	0
1	108741	T	A	0.2413793103	0.2	0.2	1	0.2	0	0	1	0
1	109226	A	T	0	0.2	0.2	0.2	1	0	0	0	1
1	109661	A	G	0.3448275862	0.2	0.2	1	0.2	0	0	1	0
1	110915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

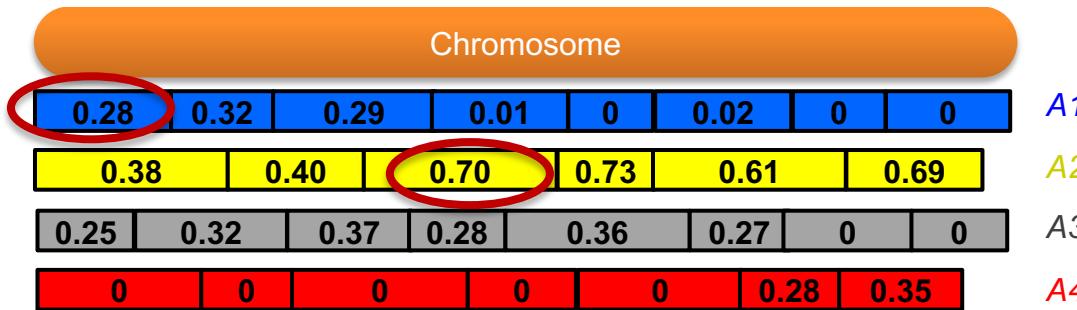
Chromosome

0.28	0.32	0.29	0.01	0	0.02	0	0	A1
0.38	0.40	0.70	0.73	0.61	0.69			A2
0.25	0.32	0.37	0.28	0.36	0.27	0	0	A3
0	0	0	0	0	0.28	0.35		A4

## TraceAncestor

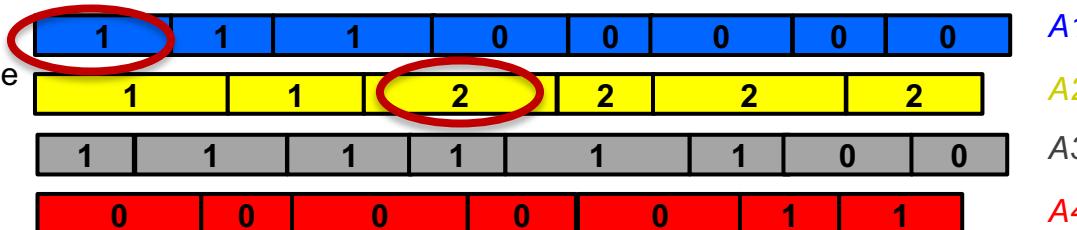
### Etape 2

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

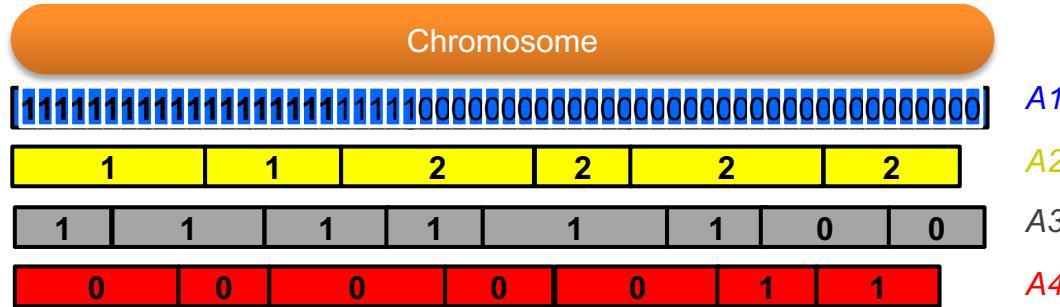
Tetraploid: 0.05 / 0.25 / 0.5 / 0.75 / 0.95

Si (-3 < LOD < 3) => indétermination

## TraceAncestor

### Etape 3

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



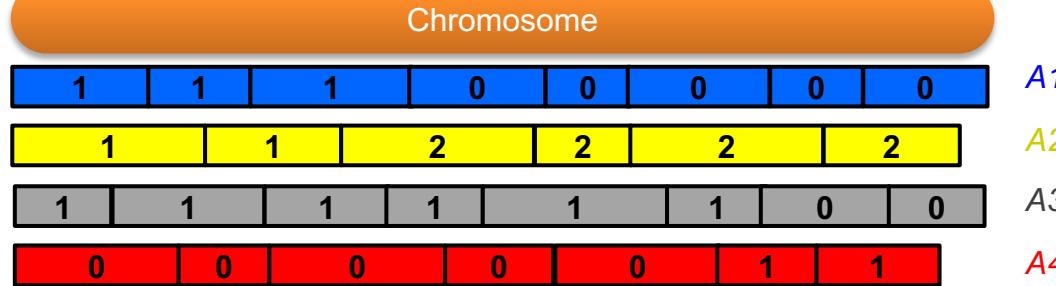
### Etape 4

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

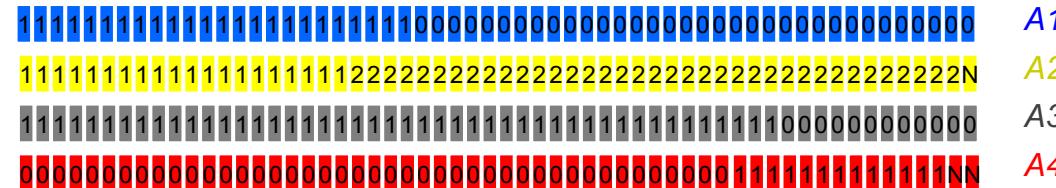
### Etape 3

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

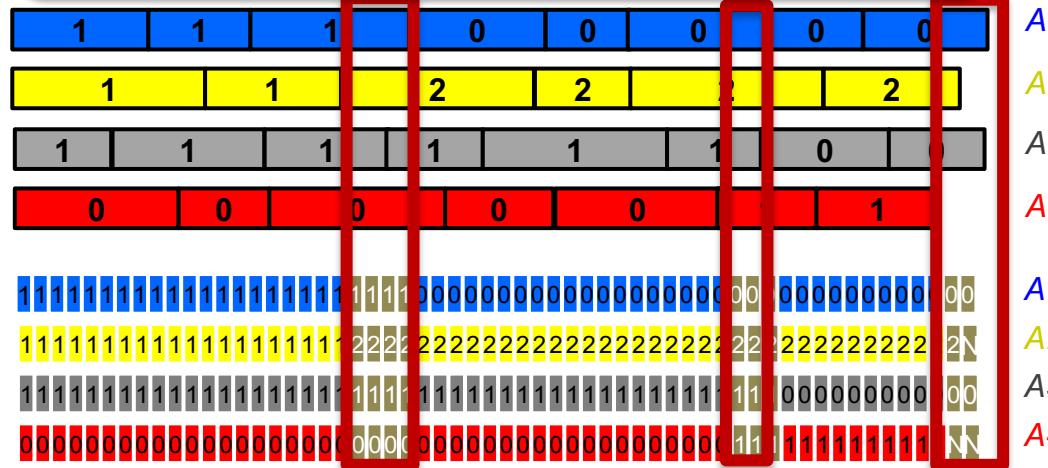
Estimation du dosage allélique de chaque ancêtre par fenêtre de 10SNP

### 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:

Chromosome

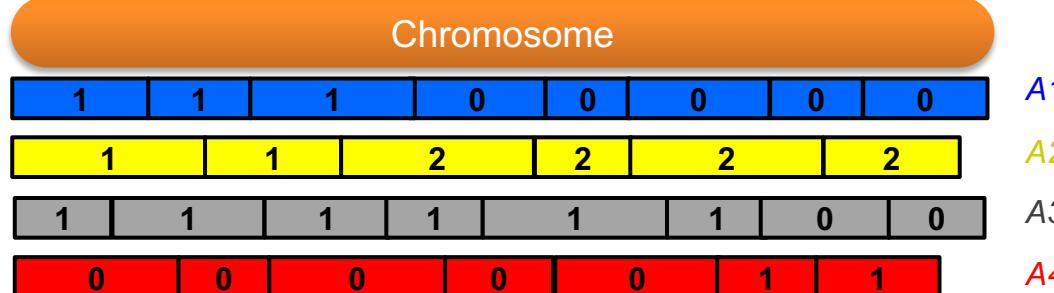


→ indetermination

## TraceAncestor

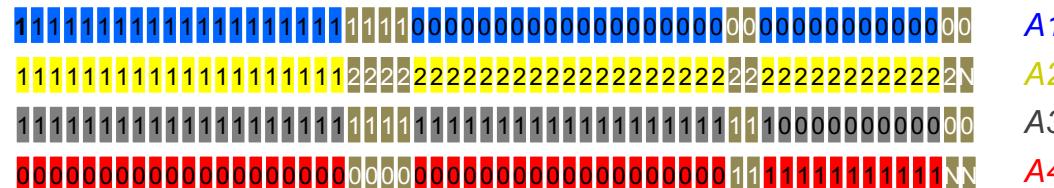
### Etape 3

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



### Karyotype

### Etape 5

Caryotypes non-phasés. Painting automatique.



# UTILISATION DE TRACE ANCESTOR SOUS GALAXY

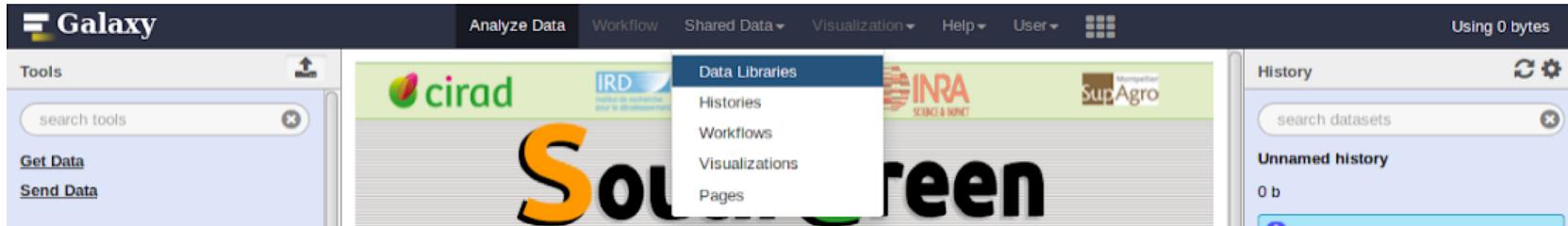
<http://galaxy.southgreen.fr/galaxy/>

## ETAPE 1 : se connecter à galaxy

The screenshot shows the Galaxy web interface. On the left, there's a sidebar with 'Tools' (search tools, Get Data, Send Data), 'BASIC TOOLS' (Text Manipulation, Filter and Sort, grepEff tools, Join, Subtract and Group). The main area has a 'Login' form with 'Username / Email Address' (formation1@cirad.fr) and 'Password' (redacted). Below the password field are links for 'Forgot password?' and 'Reset here'. A 'Login' button is at the bottom of the form. At the top, there's a navigation bar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User' (with a dropdown menu showing 'Login' and 'Register'), and a grid icon. On the right, there's a 'History' panel titled 'Unnamed history' showing '0 b'. A message in a blue box says: 'This history is empty. You can load your own data or get data from an external source'.

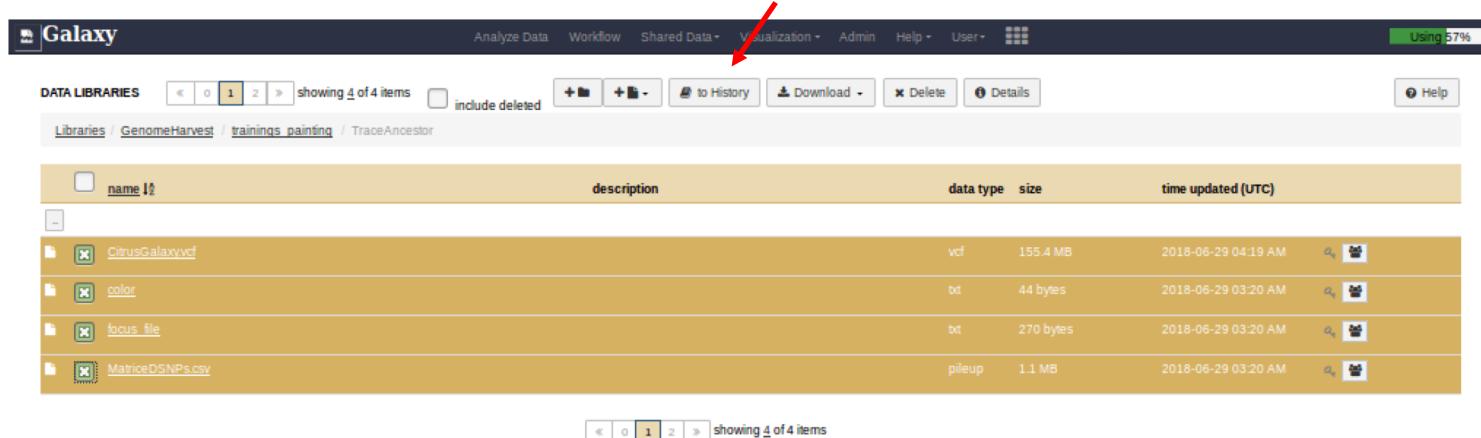
# UTILISATION DE TRACE ANCESTOR SOUS GALAXY

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



The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User', and a grid icon. A sub-menu 'Data Libraries' is open, containing 'Histories', 'Workflows', 'Visualizations', and 'Pages'. The 'Histories' option is highlighted. On the left, there's a 'Tools' panel with 'search tools' and 'Get Data/Send Data' buttons. The main workspace features logos for 'cirad', 'IRD', 'INRA', and 'SupAgro', with the word 'Source' partially visible. The right side shows a 'History' panel titled 'Unnamed history' with a single item '0 b'.

DataLibrary → GenomeHarvest → trainings\_painting → TraceAncestor



The screenshot shows the Galaxy Data Libraries page. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', 'User', and a grid icon. A red arrow points to the 'to History' button in the toolbar. The toolbar also includes buttons for creating new items, deleting, and viewing details. Below the toolbar, the URL path is shown as 'Libraries / GenomeHarvest / trainings\_painting / TraceAncestor'. The main content area displays a table of datasets:

<input type="checkbox"/>	name	description	data type	size	time updated (UTC)	Actions
<input type="checkbox"/>	CitrusGalaxyvcf		vcf	155.4 MB	2018-06-29 04:19 AM	
<input type="checkbox"/>	color		txt	44 bytes	2018-06-29 03:20 AM	
<input type="checkbox"/>	focus file		txt	270 bytes	2018-06-29 03:20 AM	
<input type="checkbox"/>	MatriceDSNPs.csv		pileup	1.1 MB	2018-06-29 03:20 AM	

At the bottom, a pagination control shows 'showing 4 of 4 items'.

# UTILISATION DE TRACE ANCESTOR SOUS GALAXY

The screenshot shows the Galaxy History interface with the following details:

- History**: The main title.
- search datasets**: A search bar.
- Unnamed history**: The name of the history.
- 4 shown, 4 deleted**: The count of datasets.
- 156.48 MB**: The total size of the datasets.
- 8: MatriceDSNPs.csv**: A CSV file containing GST data.
- 7: focus\_file**: A file containing hybrid names.
- 6: color**: A file associating colors with ancestors.
- 5: CitrusGalaxy.vcf**: A VCF file for hybrids.

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

TraceAncestor prefilter (Galaxy Version 0.1.0) ▾ Options

**matrix table**

8: MatriceDSNPs.csv

**Missing data threshold**

0.3

**GST threshold**

0.9

Execute

→ 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

TraceAncestor (Galaxy Version 0.1.0) ▼ Options

**matrix table -t**  
9: Ancestral markers matrix

**vcf of hybrid population -v**  
5: CitrusGalaxy.vcf

**ploidy -p**  
3

**color file -c**  
6: color

**number of markers by windows -w**  
10

**LOD value -l**  
3

**threshold for the calcul of LOD score -s**  
0.99

**Windows size (in k-bases) -k**  
100

**Which hybrids do you want to work on?**  
Several hybrids

**focus file (several hybrids) -f**  
7: focus\_file

Execute

- ← 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

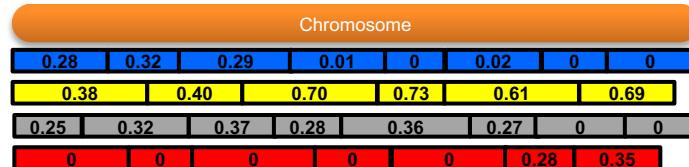
## OUTPUTS



→ 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
Sample6 1	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

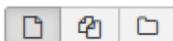
## ETAPE 5 : Visualisation

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

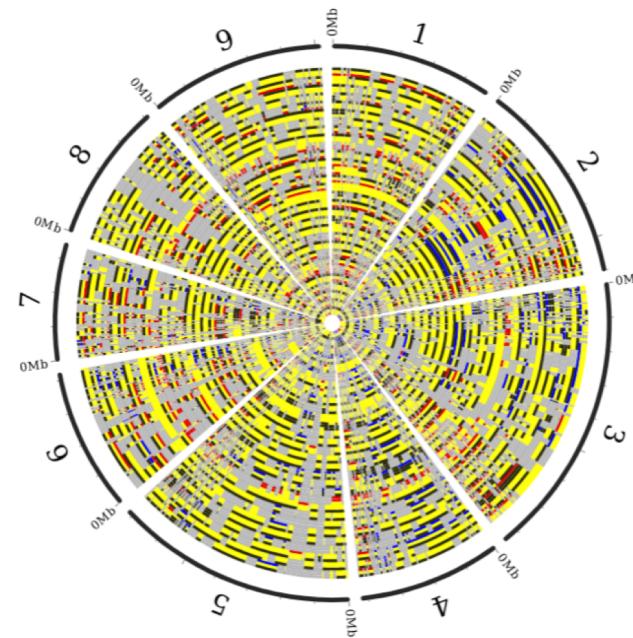


168: circos painting

+ Insert track

Track goes here

✓ Execute

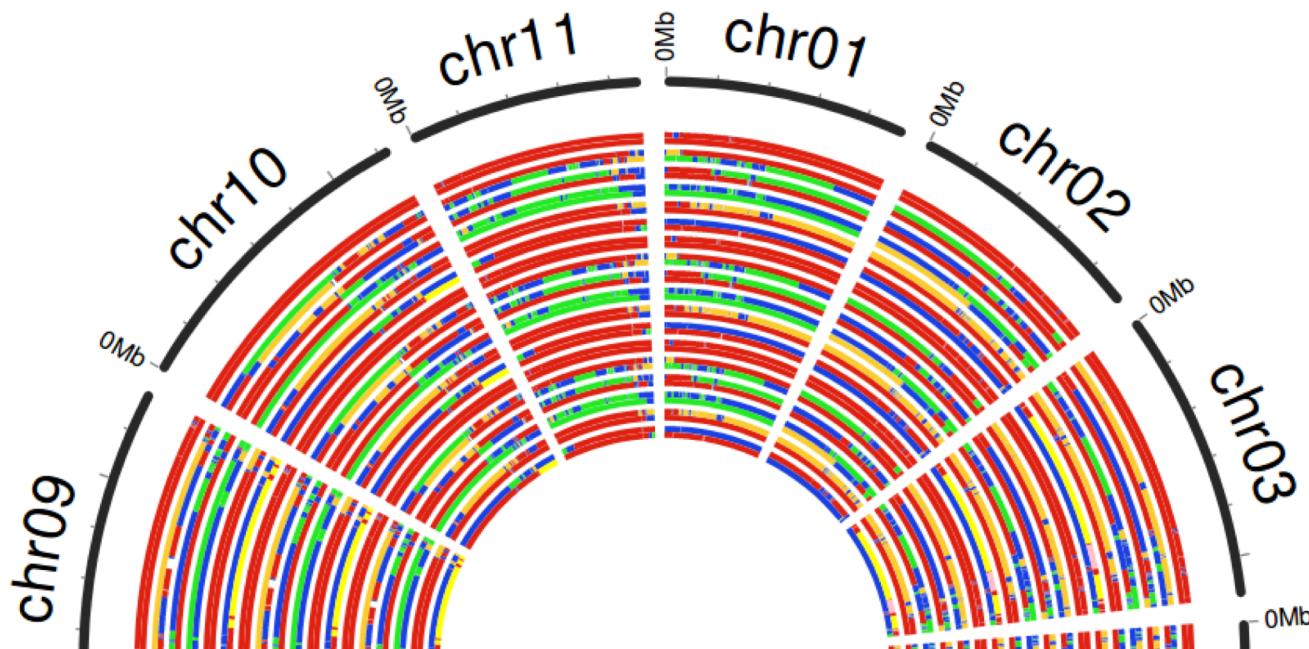


# UTILISATION DE TRACE ANCESTOR SOUS GALAXY

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

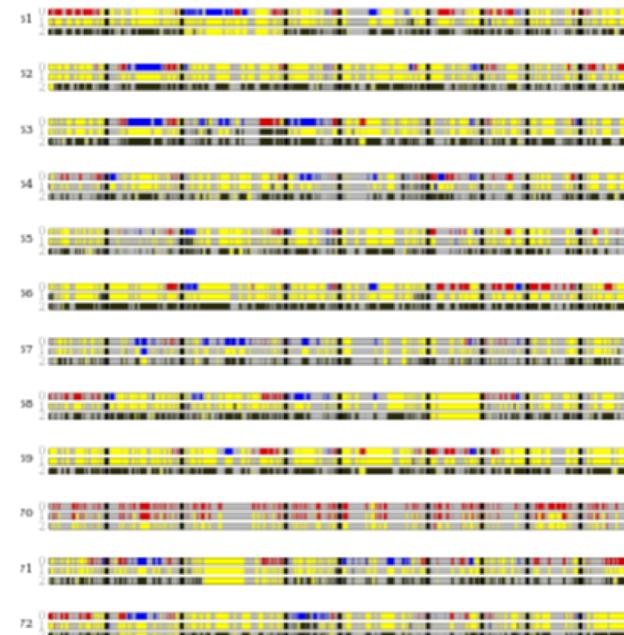
chr haplotype start end #color

**Ploidy**

3

ploidy

Execute



# UTILISATION DE TRACE ANCESTOR SOUS GALAXY

Longueur des chromosomes

Chro - haplo - début - fin - couleur

Télécharger en png

CHROMOSOMES ▾ DATA ▾ UPLOAD A FILE LOAD A TEST (RE)LOAD IMAGE CLEAR DOWNLOAD

