



Select_SNP_4_Imputation (SS4I) : Un outil simple de sélection de SNP en fonction du DL.

Frédéric Héroult, Jeremy Yon, Florian Herry, Sophie Allais, Pascale Le Roy



Héroult et al. / [Selection de SNP pour l'Imputation: SS4I](#)



Séminaire INCoMINGS La Rochelle 10-11 octobre 2016

Select_SNP_4_Imputation (SS4I)

- ✓ Outil développé en Python
- ✓ Sélection d'un panel réduit de SNP:
 - Puce basse densité
 - Imputation
- ✓ Sélection basé sur le DL entre chaque paire de SNP au sein d'un chromosome.



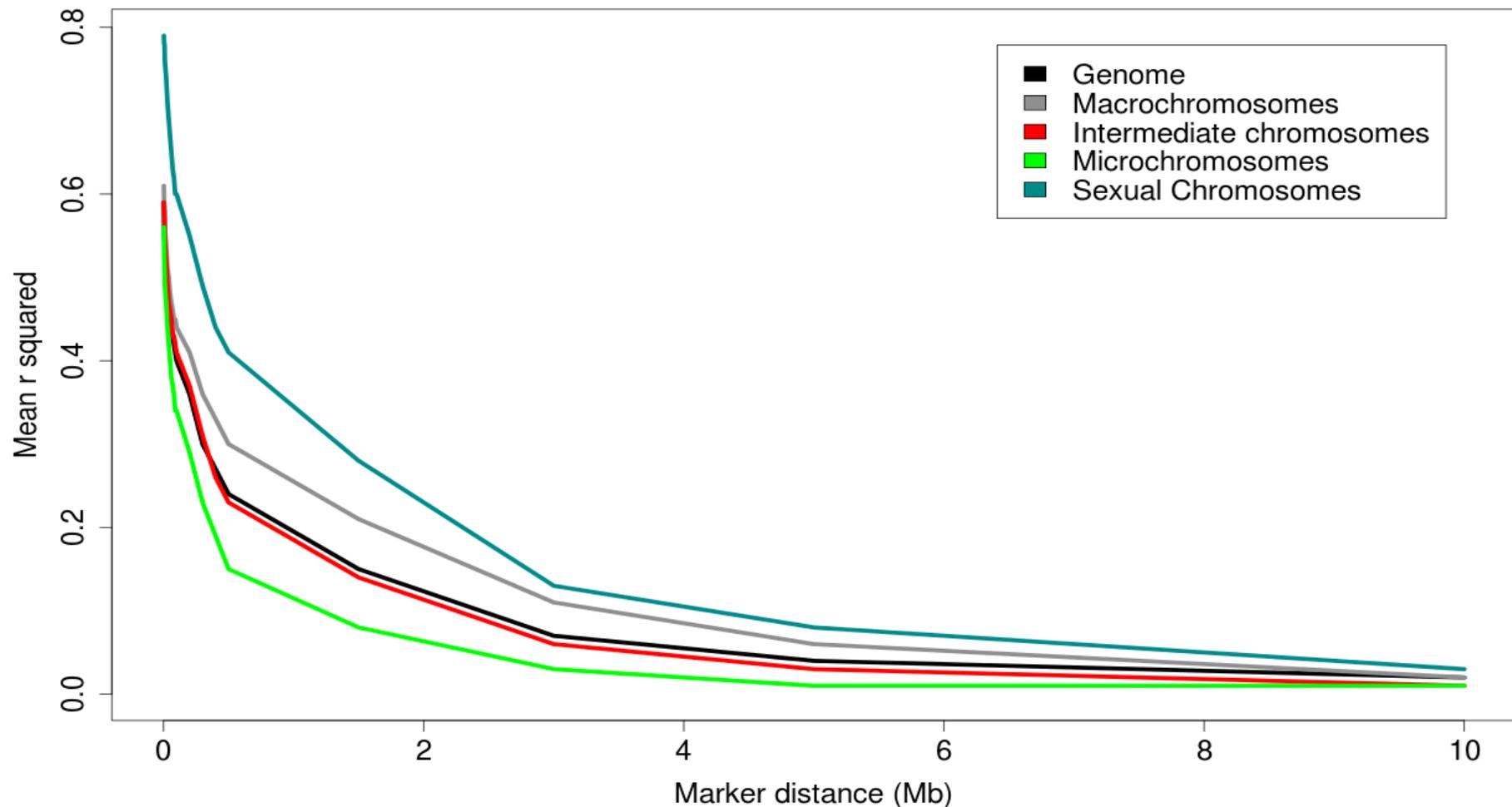
Structure particulière du génome Gallus gallus.

Structure et DL du génome *Gallus gallus*

- ✓ 39 paires de chromosomes.
 - Macrochromosomes (GGA1 to GGA5)
 - Chromosomes intermédiaires (GGA6 to GGA10)
 - Microchromosomes (GGA11 to GGA 38)
 - Sexual chromosomes (Z & W).

Structure et DL du génome *Gallus gallus*

Niveau et étendue du DL entre ≠ catégorie de chromosomes

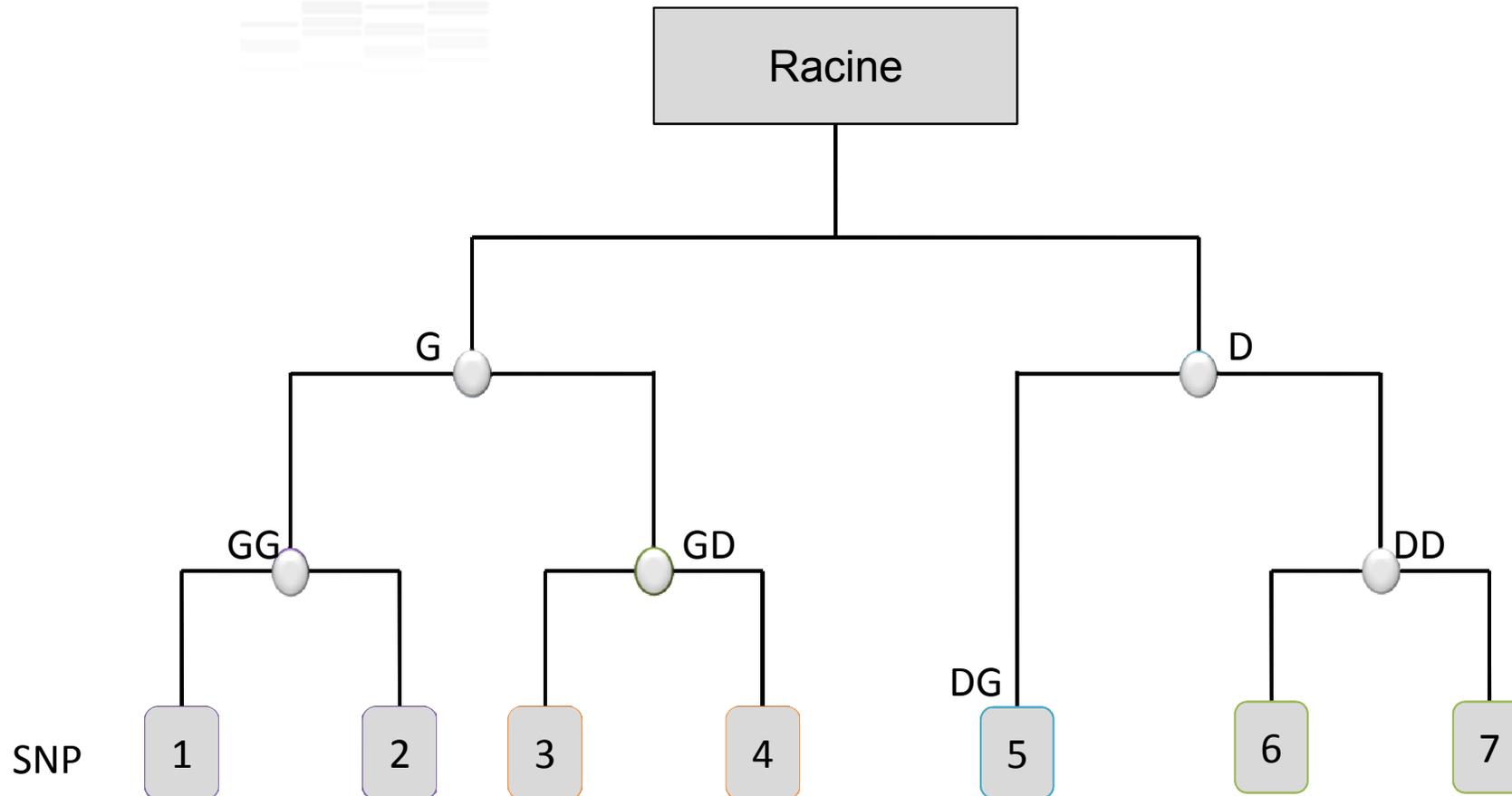


Principe de fonctionnement

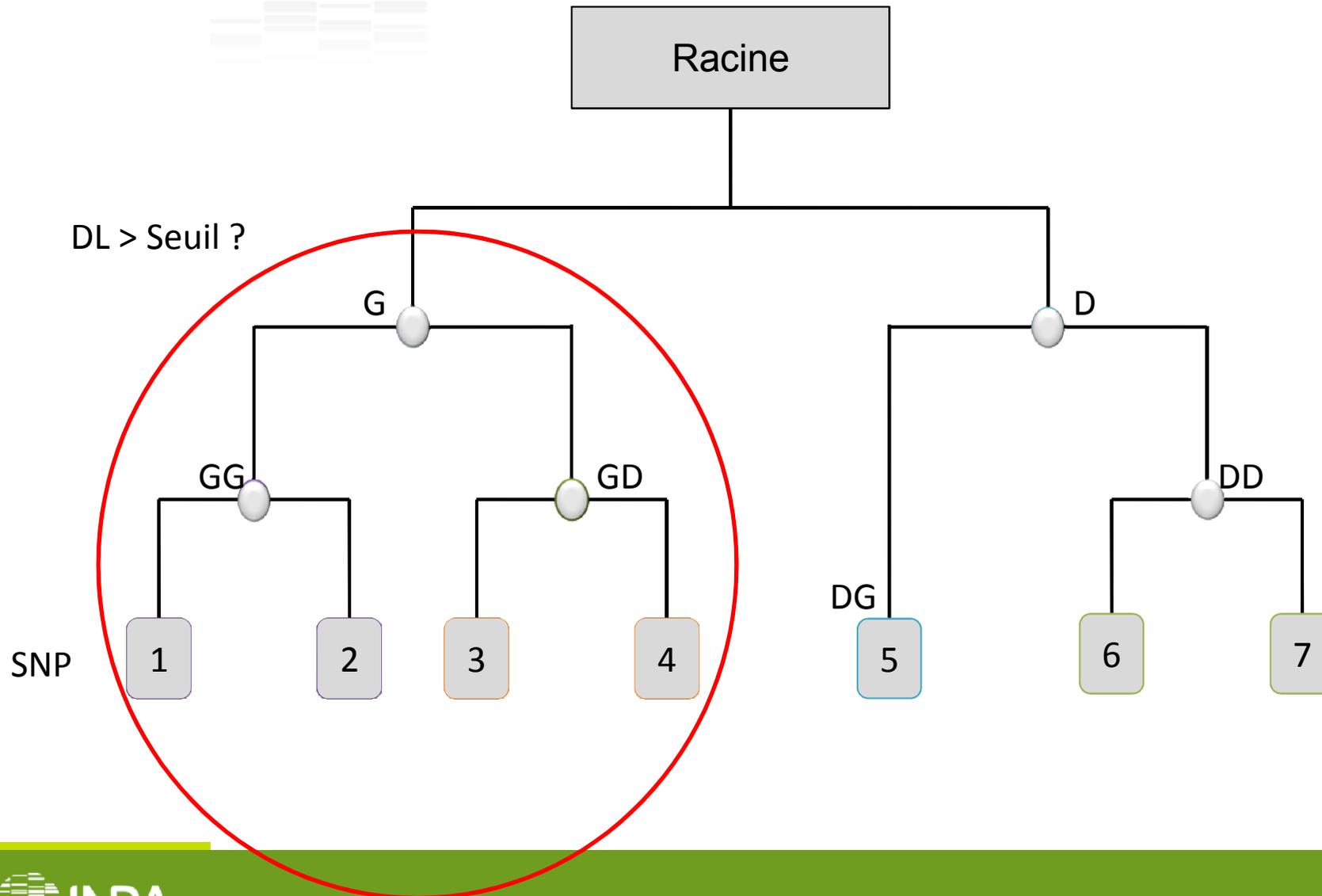
Pour chaque chromosome

- ✓ Calcul du DL entre chaque paire SNP (PLINK).
- ✓ Clustering Hiérarchique Ascendant basé sur le DL.
- ✓ Création dendrogramme.
- ✓ Sélection d'ensemble de SNP répondant au critère de seuil de DL.
- ✓ Sélection d'un SNP représentatif de chaque ensemble.

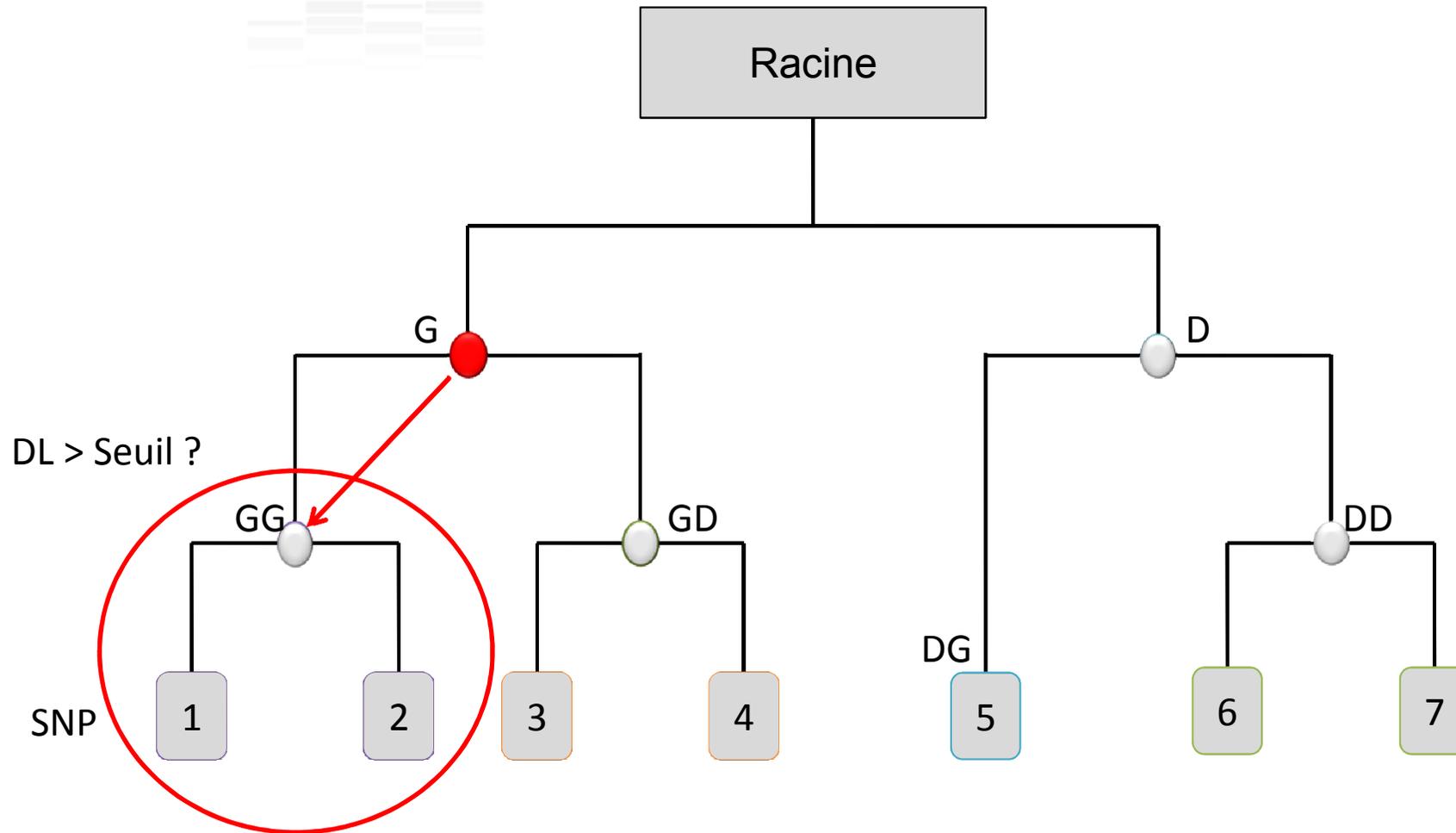
Sélections des ensembles de SNP / DL.



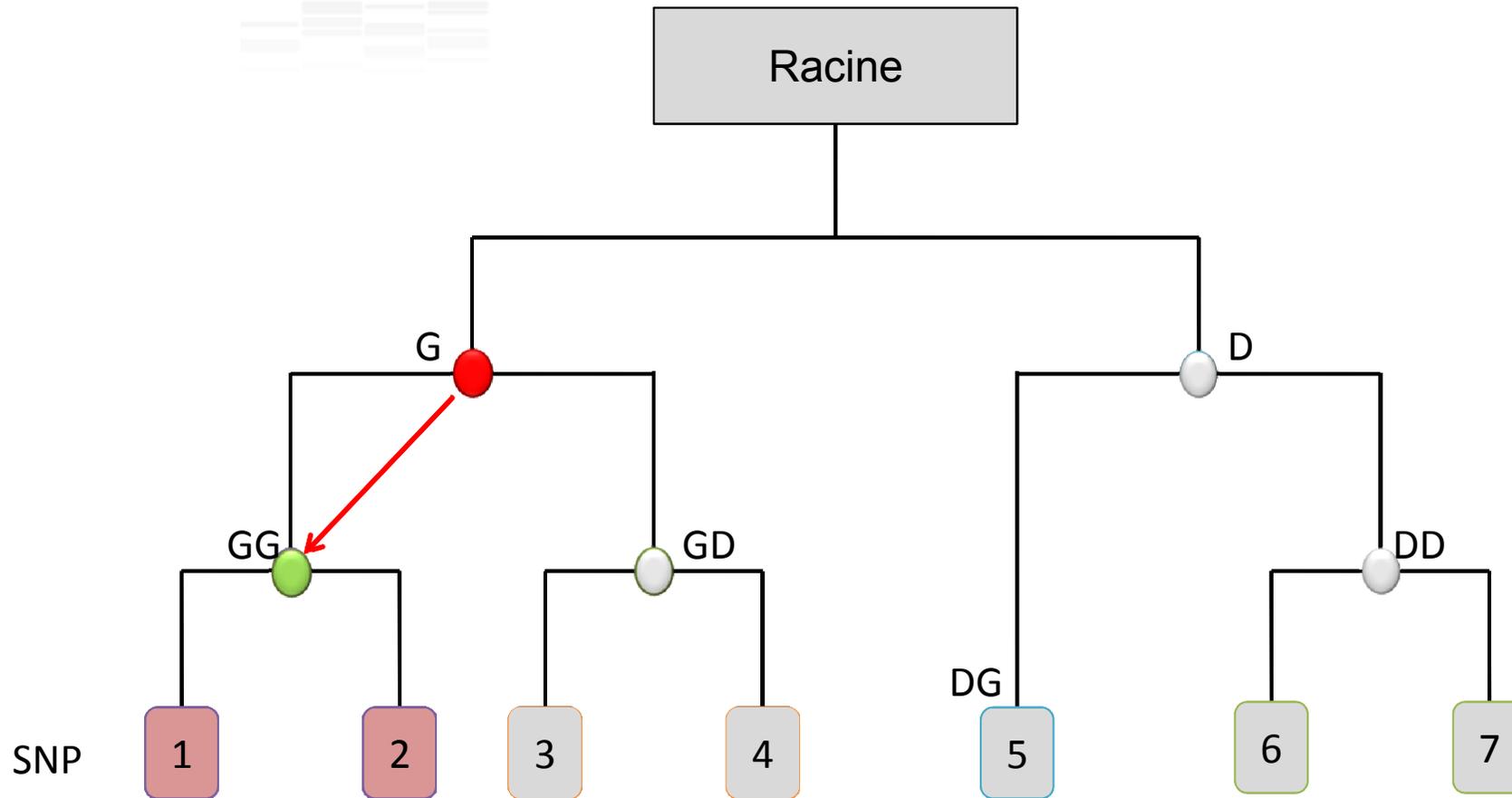
Sélections des ensembles de SNP / DL.



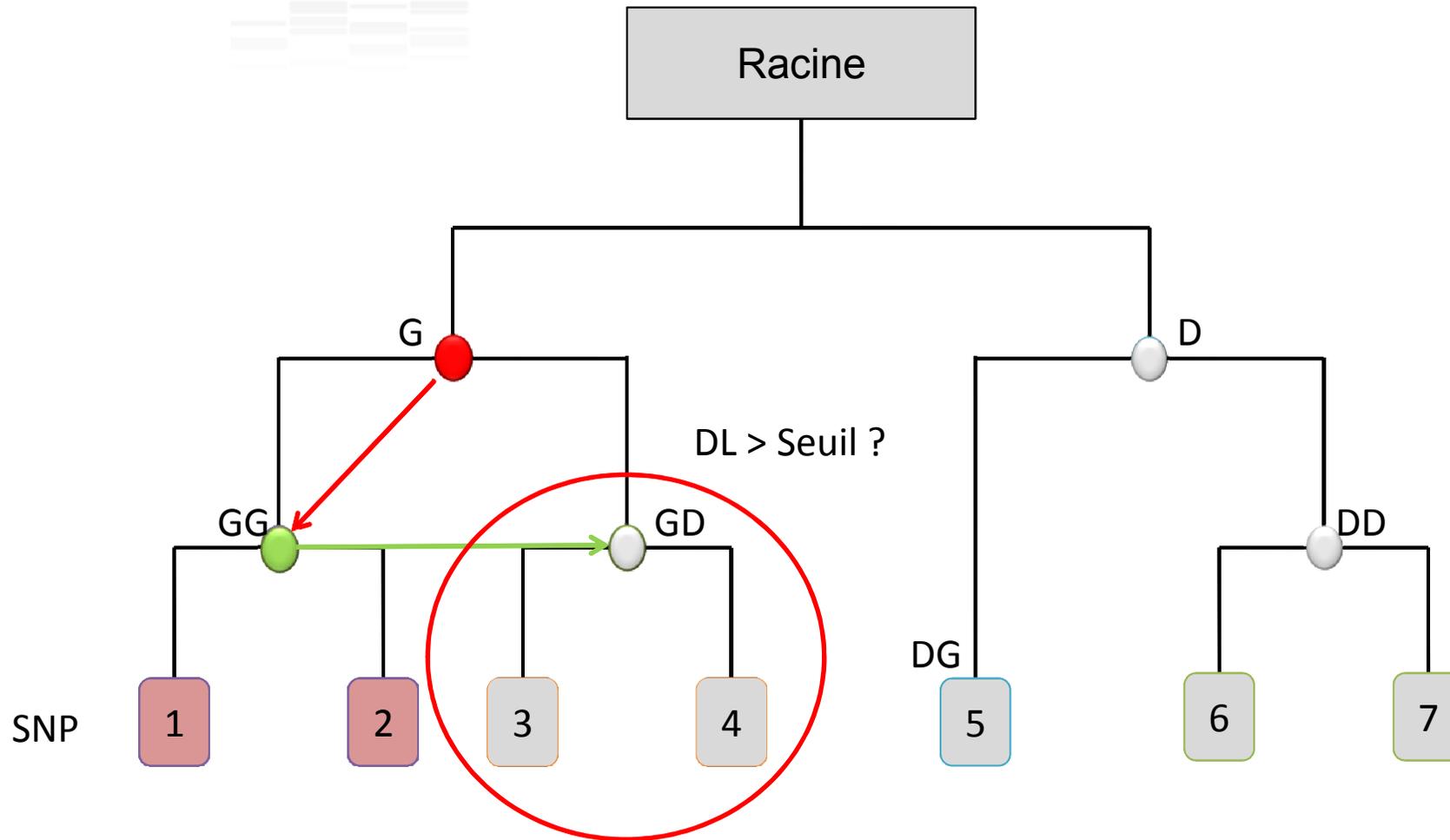
Sélections des ensembles de SNP / DL.



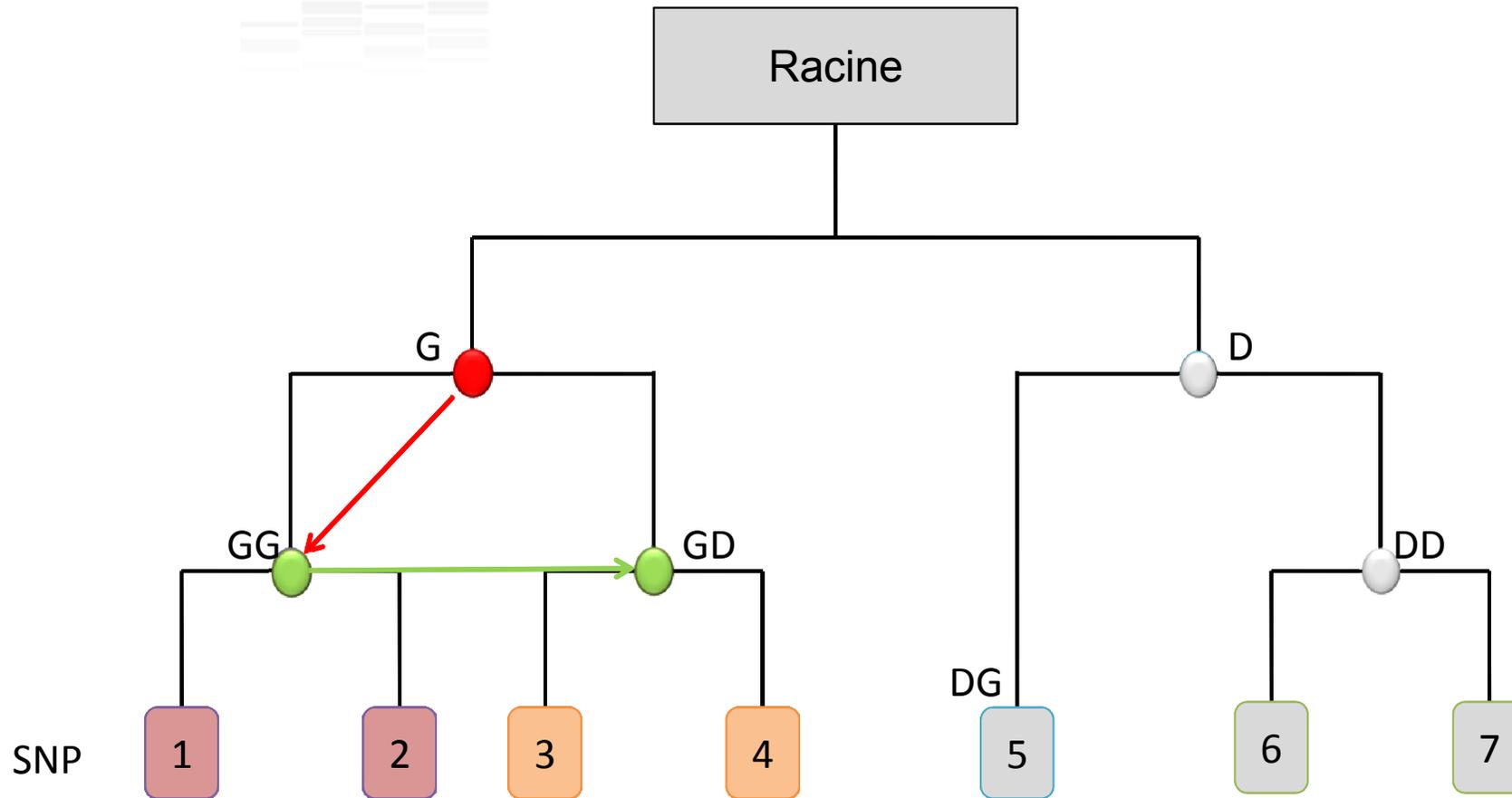
Sélections des ensembles de SNP / DL.



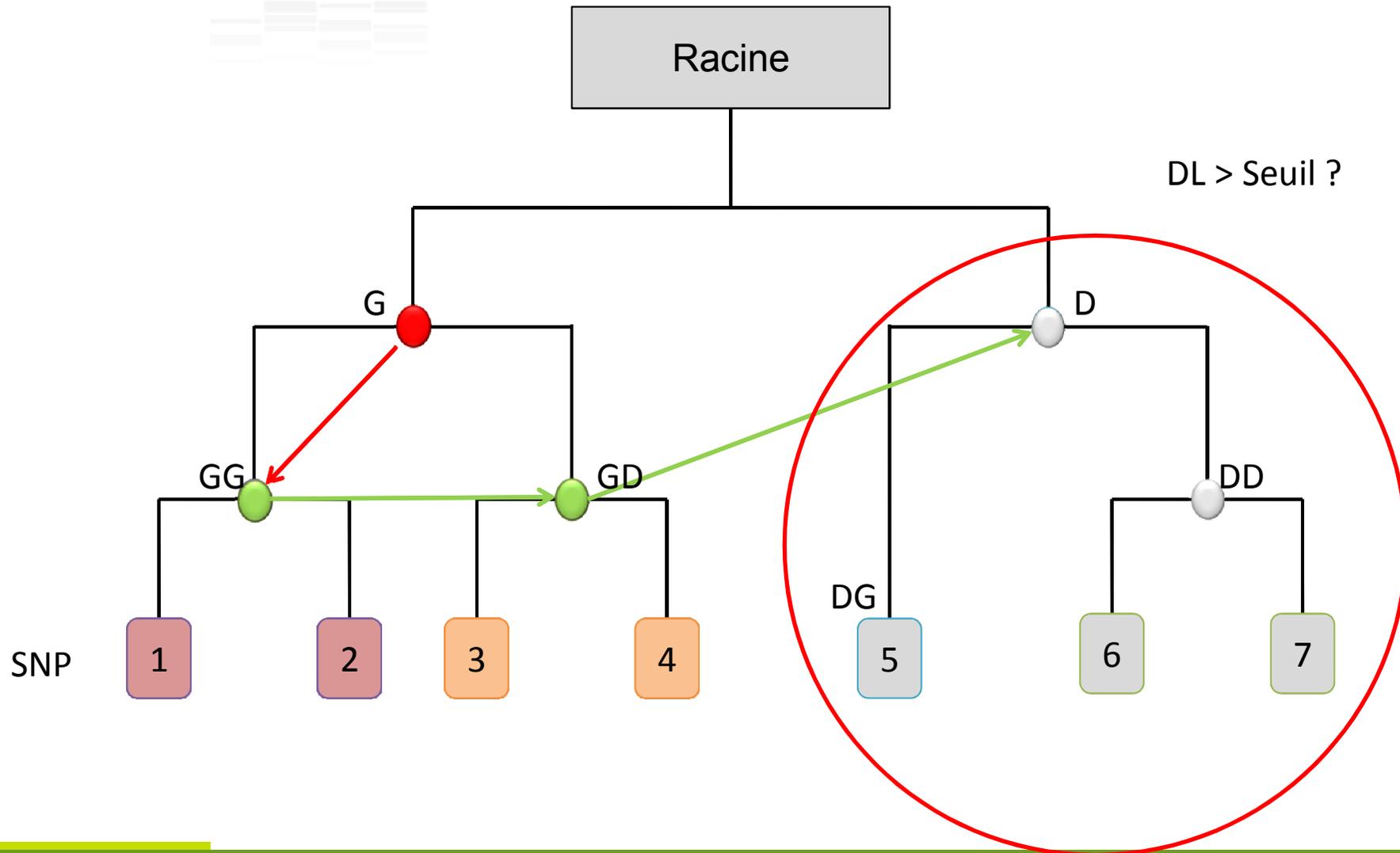
Sélections des ensembles de SNP / DL.



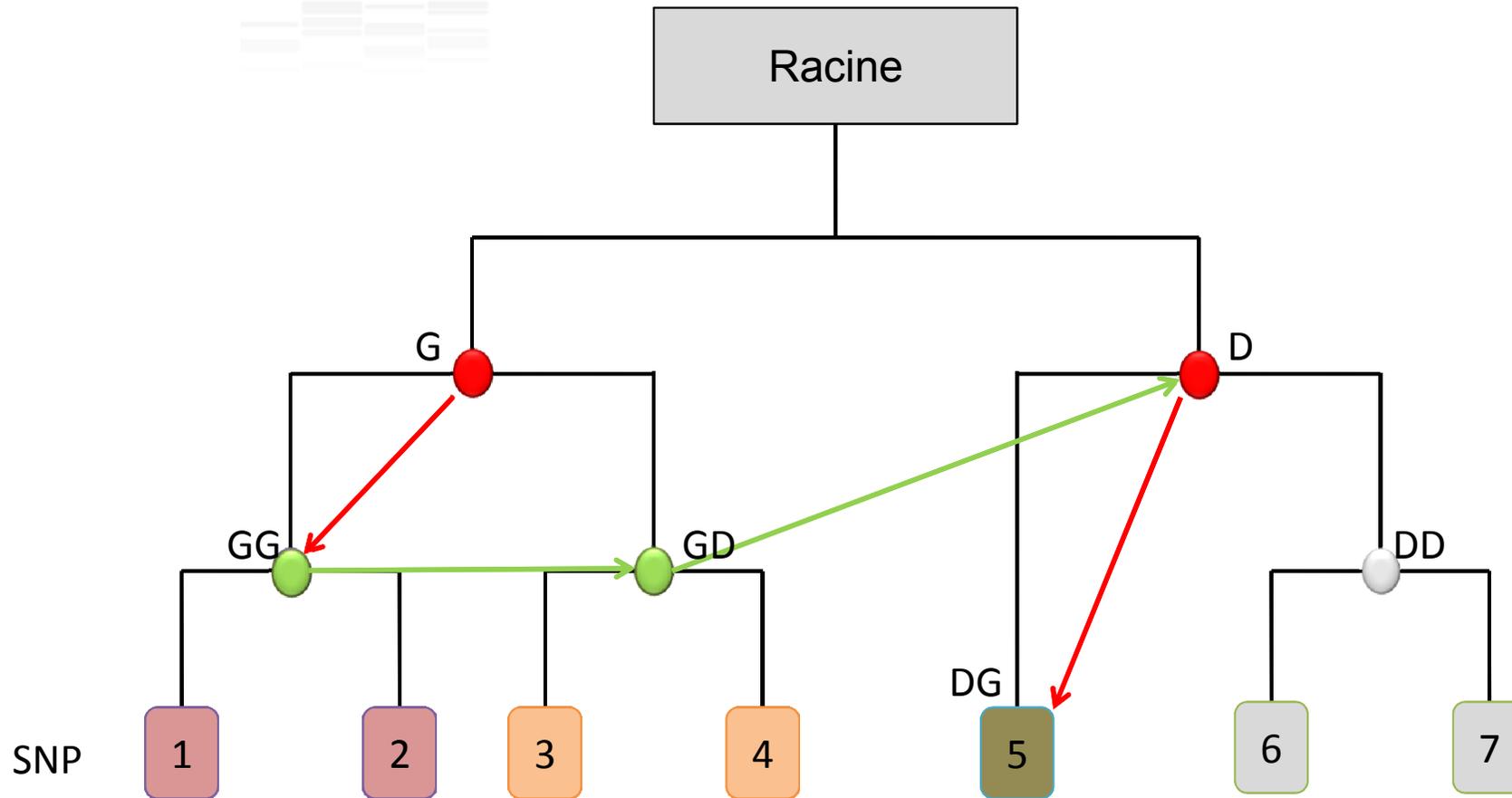
Sélections des ensembles de SNP / DL.



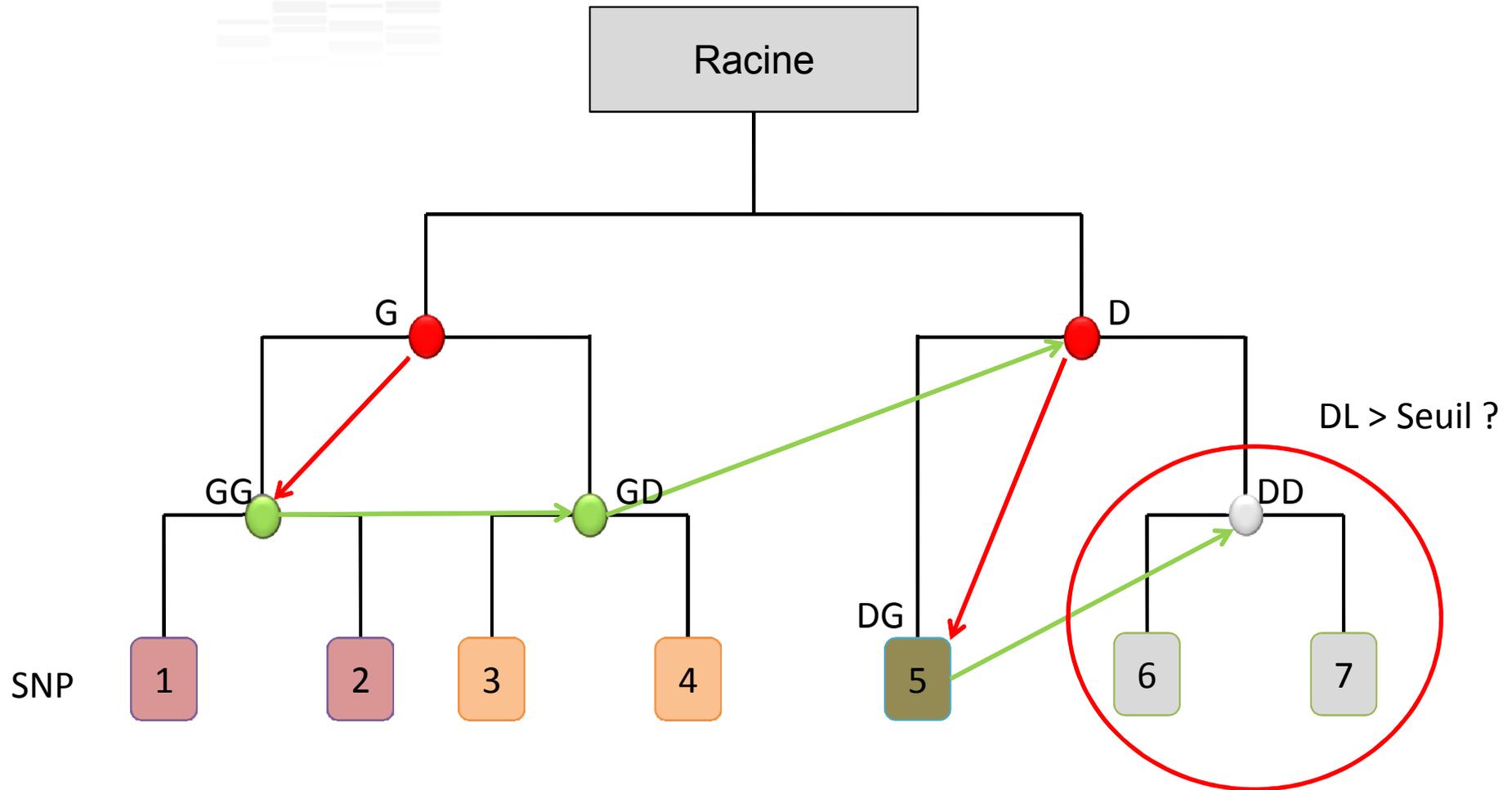
Sélections des ensembles de SNP / DL.



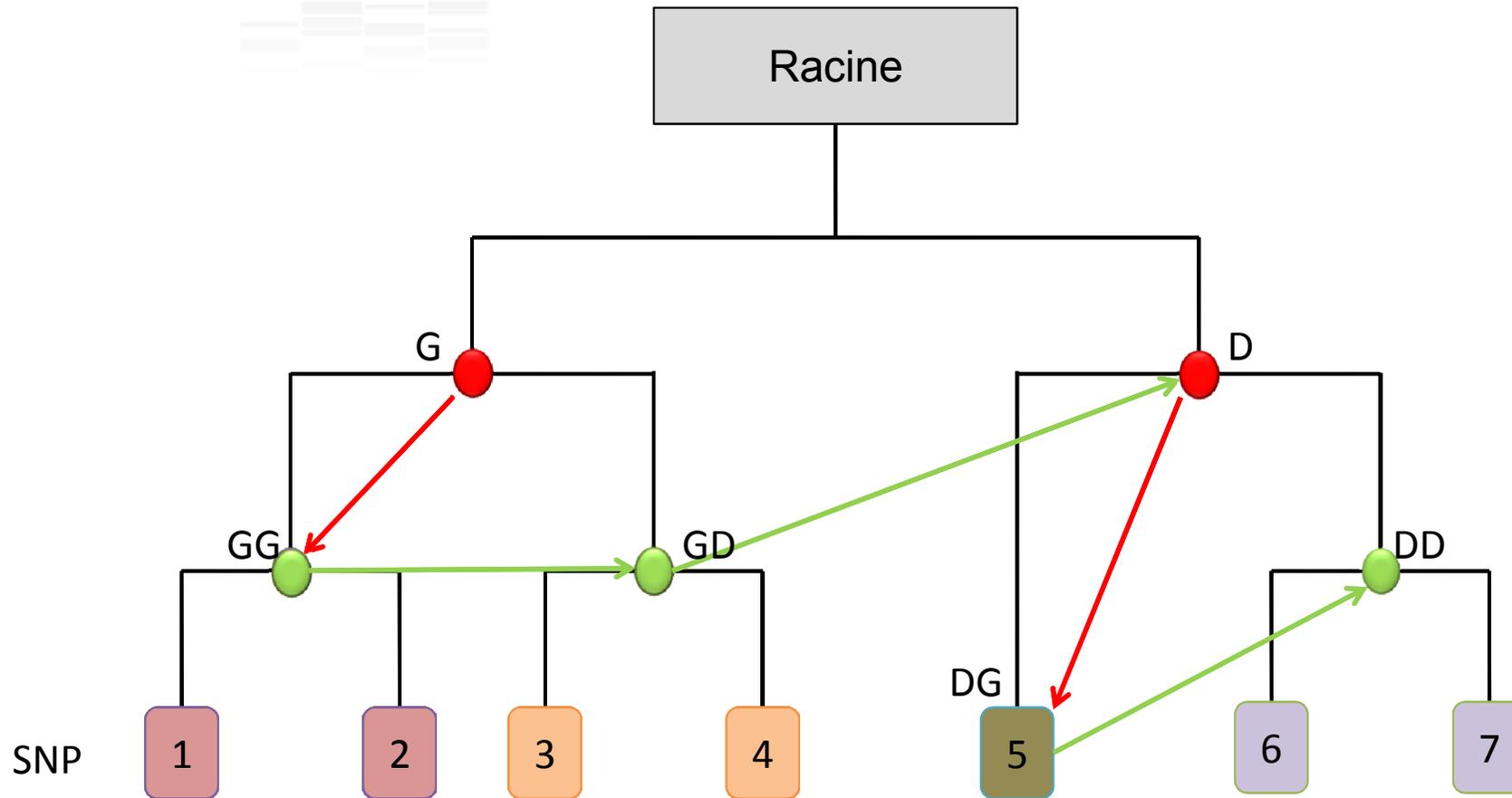
Sélections des ensembles de SNP / DL.



Sélections des ensembles de SNP / DL.



Sélections des ensembles de SNP / DL.



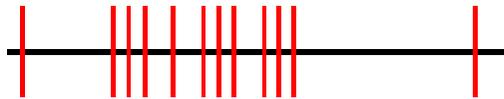


Sélection d'un SNP dans chaque ensemble.

- ✓ Au sein de chaque groupe:
 - MAF du SNP : la plus forte
 - Position du SNP / autres SNP du groupe.

Sélection d'un SNP dans chaque ensemble.

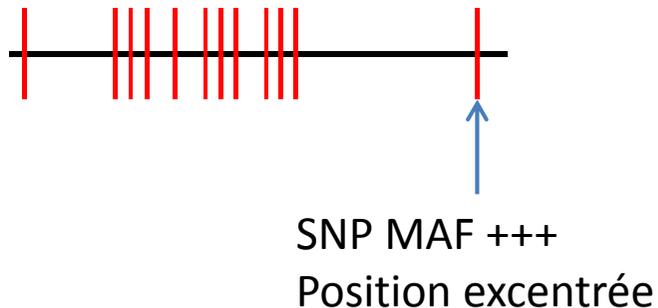
- ✓ Au sein de chaque groupe:
 - MAF du SNP : la plus forte
 - Position du SNP / autres SNP du groupe.



Sélection d'un SNP dans chaque ensemble.

✓ Au sein de chaque groupe:

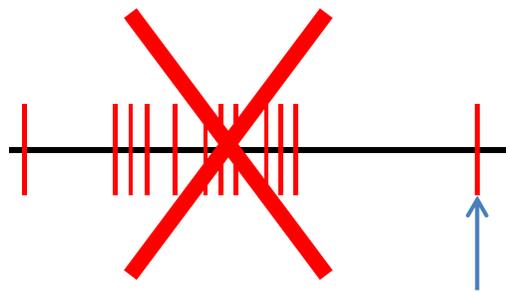
- MAF du SNP : la plus forte
- Position du SNP / autres SNP du groupe.



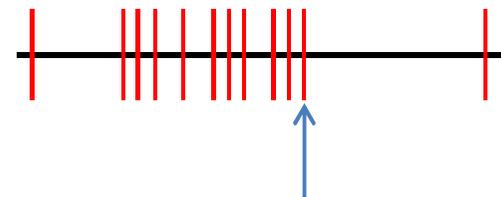
Sélection d'un SNP dans chaque ensemble.

✓ Au sein de chaque groupe:

- MAF du SNP : la plus forte
- Position du SNP / autres SNP du groupe.



SNP MAF +++
Position excentrée



SNP MAF ++
Position représentative du groupe

Fichier « input »

- ✓ Trois fichiers sont nécessaires au fonctionnement du programme.
 - Un fichier de génotypage au format PLINK «.ped»:
6 colonnes + génotype, 1 ligne par individu

Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype & Genotype

```
Uto_124 Uto_124 0 0 0 -9 G G T T T T
Uto_13 Uto_13 0 0 0 -9 G G T T C T
Uto_243 Uto_243 0 0 0 -9 A G T T T T
Uto_244 Uto_244 0 0 0 -9 G G T T C T
Uto_273 Uto_273 0 0 0 -9 G G T T T T
```

Fichier « input »

- ✓ Trois fichiers sont nécessaires au fonctionnement du programme.
 - Un fichier de génotypage au format PLINK «.ped»
 - Un fichier de carte au format PLINK «.map»:
4 colonnes, 1 ligne par SNP

Chromosome, SNP identifier, Genetic distance, Base-pair position

25	AX-77283174	0	5208
24	AX-80751587	0	6369
28	AX-76388168	0	6695
28	AX-76388358	0	7122
25	AX-80909995	0	7629

Fichier « input »

- ✓ Trois fichiers sont nécessaires au fonctionnement du programme.
 - Un fichier de génotypage au format PLINK «.ped»
 - Un fichier de carte au format PLINK «.map»
 - Un fichier paramètre.

Fichier « input »

```
# PARAMETERS FILE

#----- INPUT FILES -----
#in_geno : path to the input genotype file
in_geno =

#in_map : path to the input map file
in_map =

#----- ANALYSIS PARAMETERS -----
#LD (r2) threshold :threshold for analysis (must be a decimal number, use a point, not a comma)
threshold = 0.5

#chr : chromosome to analyse
#for several chromosomes : chr = 7,8,Y
#for all chromosomes : chr = all
chromosome = all

#SNP_window: Maximum number of consecutive SNP to be consider for the selection.
#for chromosome with a large number of SNP, SNP selection will be realized in several steps.
SNP_window = 35000

#----- OUTPUT -----
#path_out : Location for output files
path_out =
```

Fichier « input »

```
# PARAMETERS FILE

#----- INPUT FILES -----
#in_genotype : path to the input genotype file
in_genotype =

#in_map : path to the input map file
in_map =

#----- ANALYSIS PARAMETERS -----
#LD (r2) threshold :threshold for analysis (must be a decimal number, use a point, not a comma)
threshold = 0.5

#chr : chromosome to analyse
#for several chromosomes : chr = 7,8,Y
#for all chromosomes : chr = all
chromosome = all

#SNP_window: Maximum number of consecutive SNP to be consider for the selection.
#for chromosome with a large number of SNP, SNP selection will be realized in several steps.
SNP_window = 35000

#----- OUTPUT -----
#path_out : Location for output files
path_out =
```

Si SNP > 40 k SNP sur 1 chromosome => problème mémoire !

Fichier « output »

- ✓ Deux fichiers sont générés en sortie.
 - Un fichier log:
 - ✓ rappel des paramètres de l'analyse
 - ✓ déroulement de l'analyse.

```
#### Paramètres de l'analyse:

Fichier.carte:...../Geno_Compar_ScriptR_Python.map
Fichier.genotype:....../Geno_Compar_ScriptR_Python.ped
Seuil.DL:.....0.5
Chromosome:.....all
Fenêtre.d'analyse.maximum.35000.SNP.

#####

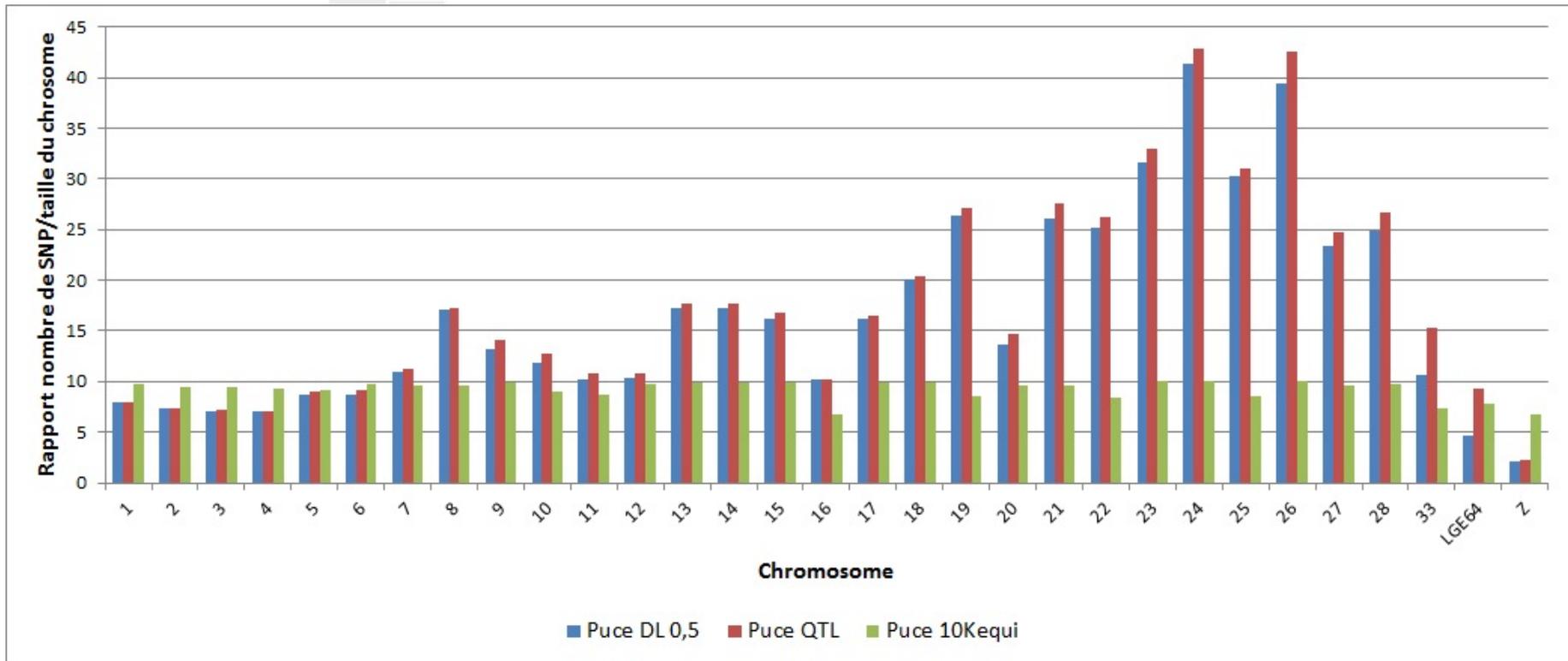
# Chromosome.Z
Le.chromosome.sera.analysé.en.une.seule.fois.
Plink.(fréquence).à.10h48.le.06/10
Plink.(r2).à.10h48.le.06/10
Importation.des.données.MAF.et.DL.à.10h55.le.06/10
SNP.supprimé(s).de.l'analyse.['AX-80820745','AX-77260073','AX-80915400']
Création.du.dendrogramme.à.10h56.le.06/10
Sélection.des.clusters.et.du.snp.représentatif.à.11h00.le.06/10
L'analyse.du.chromosome.Z.s'est.terminée.à.11h00.le.06/10
```

Fichier « output »

- ✓ Deux fichiers sont générés en sortie.
 - Un fichier log
 - Un fichier « Selected_SNP »

```
Nom_SNP : nom du SNP sélectionné dans le clusters
Chromosome : numéro du chromosome sur lequel l'analyse a été réalisé
Position : position physique du SNP sélectionné sur le chromosome
MAF : MAF du SNP sélectionné
Nbre_SNP_cluster : Nombre de SNP présent dans le cluster de SNP
R2_min : R2 minimum au sein du cluster de SNP
R2_max : R2 maximum au sein du cluster de SNP
R2_moyenne : moyenne des R2 au sein du cluster de SNP
R2_écart_type : écart type des R2 au sein du cluster de SNP
Distance_cluster_min : distance minimum entre les SNP du cluster.
Distance_cluster_max : distance maximum entre les SNP du cluster.
Distance_cluster_moyenne : moyenne des distances entre les SNP du cluster.
Distance_cluster_écart_type : écart type à la moyenne des distances entre les SNP du cluster.
SNP_min_distance : distance minimum entre le SNP sélectionné et les autres SNP du cluster
SNP_max_distance : distance maximum entre le SNP sélectionné et les autres SNP du cluster
SNP_moyenne_distance : moyenne des distances entre le SNP sélectionné et les autres SNP du cluster
SNP_écart_type_distance : écart type à la moyenne des distances entre le SNP sélectionné et les autres SNP du cluster
```

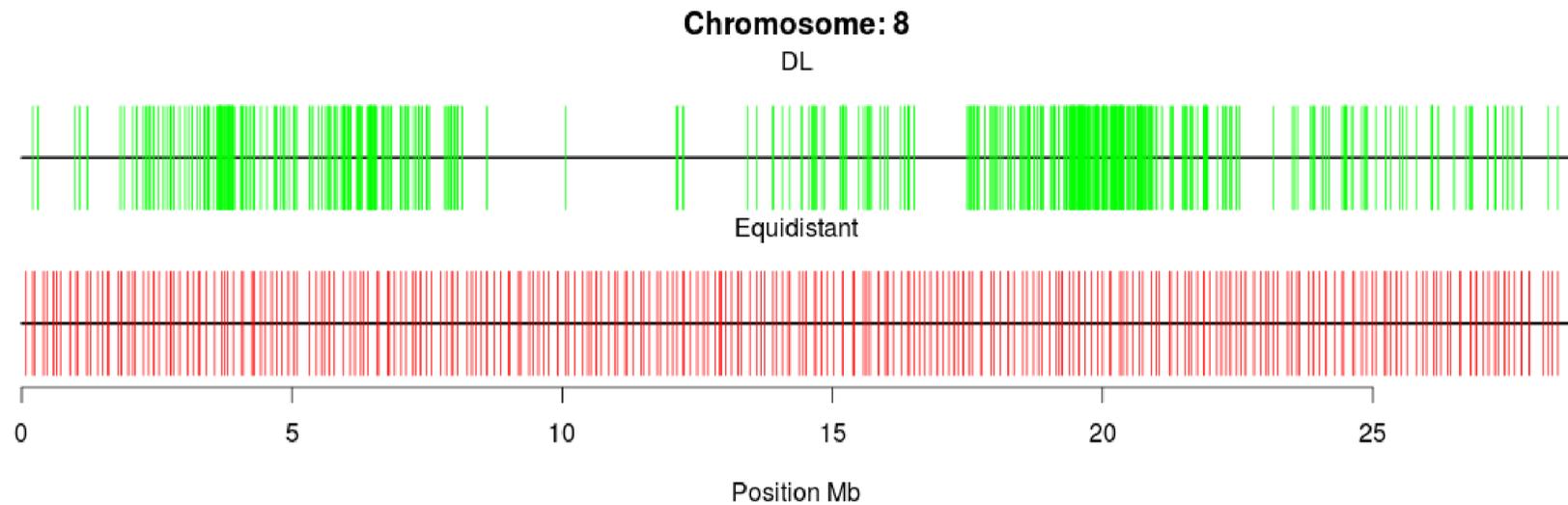
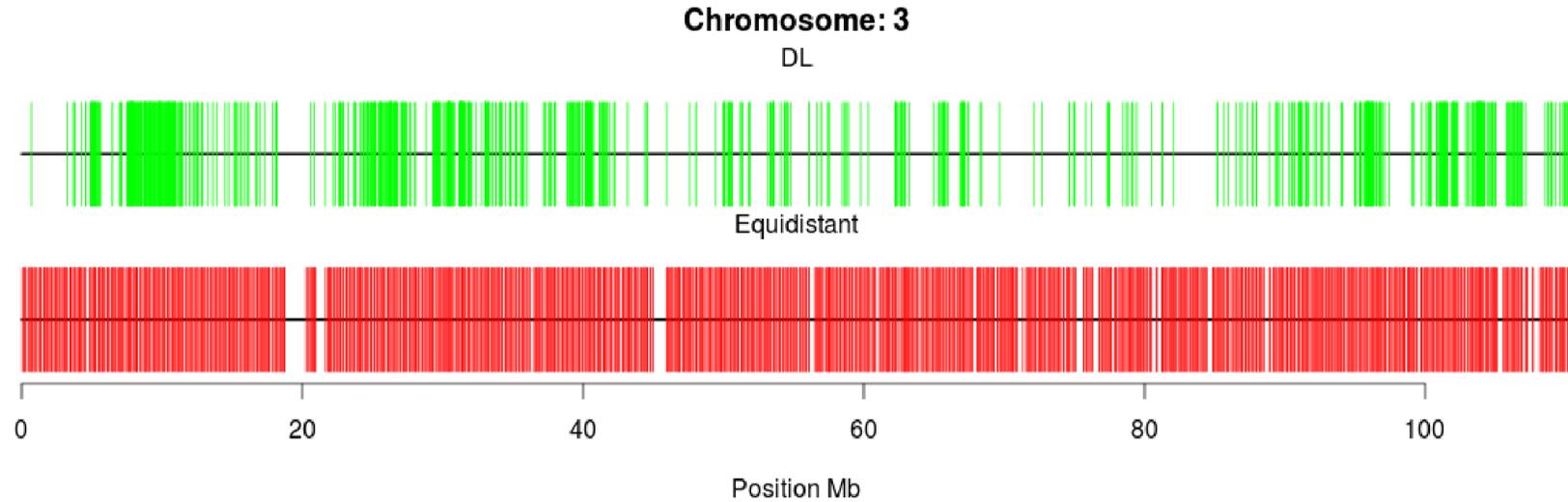
Résultats: Nombre SNP / Chromosome



Florian Herry, mémoire de fin d'études Septembre 2016

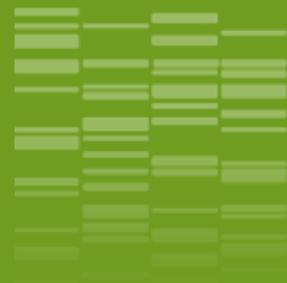
- ✓ Optimisation du nombre de SNP sur les macrochromosomes
- ✓ Densification du nombre de SNP sur les microchromosomes

Résultats: Répartition des SNPs.



Exécution du programme:

- ✓ Exécution du programme: bigmem 40 GT
Select_SNP_4_Imputation.py Parameters.txt
- ✓ Analyse de 280 k SNPs sur 31 Chromosomes.
- ✓ Temps d'exécution: 20h.



Merci de votre attention.