Analyse de la diversité codante et non codante le long du génome de la vigne

Bacilieri R., Canaguier A., Guichard C., Thareau V., Bounon R , Le Paslier M.-C., Le Cunff L., Nicolas S., Dereeper A., Peros J.-P., Adam-Blondon A.-F., P. This

Projet AIP Bioressources 2007-2008

Introduction

Reséquençage de 1000 régions du génome de la vigne

- cartographie de la diversité SNP et indel
  (identification SNP et carte génétique)

- évolution du polymorphisme le long du génome
  (étude évolution génome et trace sélection)

- comparaison polymoprhisme entre espèces
- alimentation de projets à venir :
  - DL-Vitis
  - GrapeReSeq
  - Génétique d'association
  - Trace sélection

INRA

UNIVERSITÉ PARIS-SUD 11

cnrs

27 V. Vinifera, 7 sylvestris et 12 espèces utilisées en amélioration

Introduction

Equipes

UMR 1097 Diversité et adaptation des plantes cultivées (Dia-PC)
Montpellier

UMR 1131 Vigne et Vins Santé de la Vigne et Qualité du Vin
Colmar

UMR 1165 Génomique Végétale, URGV
Evry

UMR 1287 Physiologie et Génomique fonctionnelle de la vigne
Bordeaux

UR 1164 Génomique – Info (URGI)
Evry

UMR 8618 FAMEVO – IBP Université Paris-Sud / CNRS
Orsay

UR 1279 Étude du polymorphisme des génomes végétaux
Evry

M&M

A.1. Critères pour le choix des gènes

400 gènes candidats

(Evry, Montpellier, Bordeaux, Colmar)

(métabolisme du fer, résistance, taille de la baie, tanins…)

110 gènes dans deux régions QTL pour étude DL

(QTL tanins et taille de la baie)

500 gènes bien repartis – 19 chr + random

100 gènes pour remplir les trous physiques

(>750M pb)

M&M

A.2. Filtres utilisés pour la recherche automatique des amorces (V. Thareau, SPADS-Primer3) :

- tm homogène (haut débit)
- évitement auto-hybridation amorces
- taille amplicon < 1300 puis 700
- spécificité de l'amplification
- match Unigenes (manuelle)
- maximiser le % d'exon
- position

M&M

B. Cartographie

135 gènes dans Chr Un-random

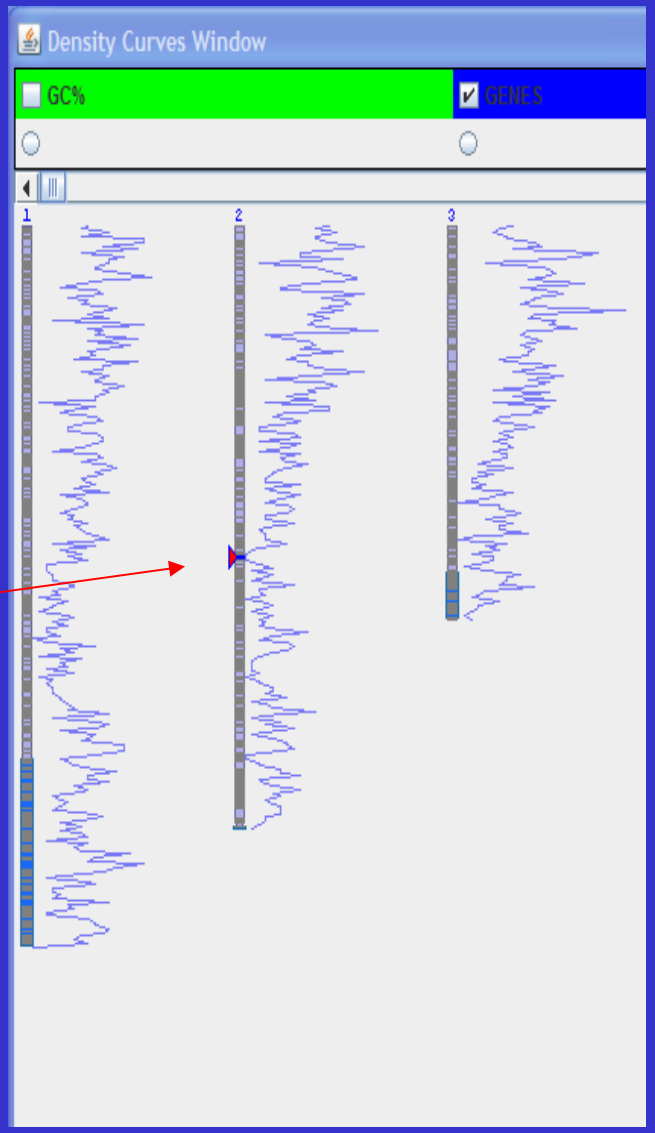
Environ 42 gènes par chromosome

Distance min = 3k pb (2 gènes candidats)

Distance max = 3379k pb
(chr 2, désert génique)

Distance moyenne = 500k pb

(taille génome de la vigne = 460M pb)

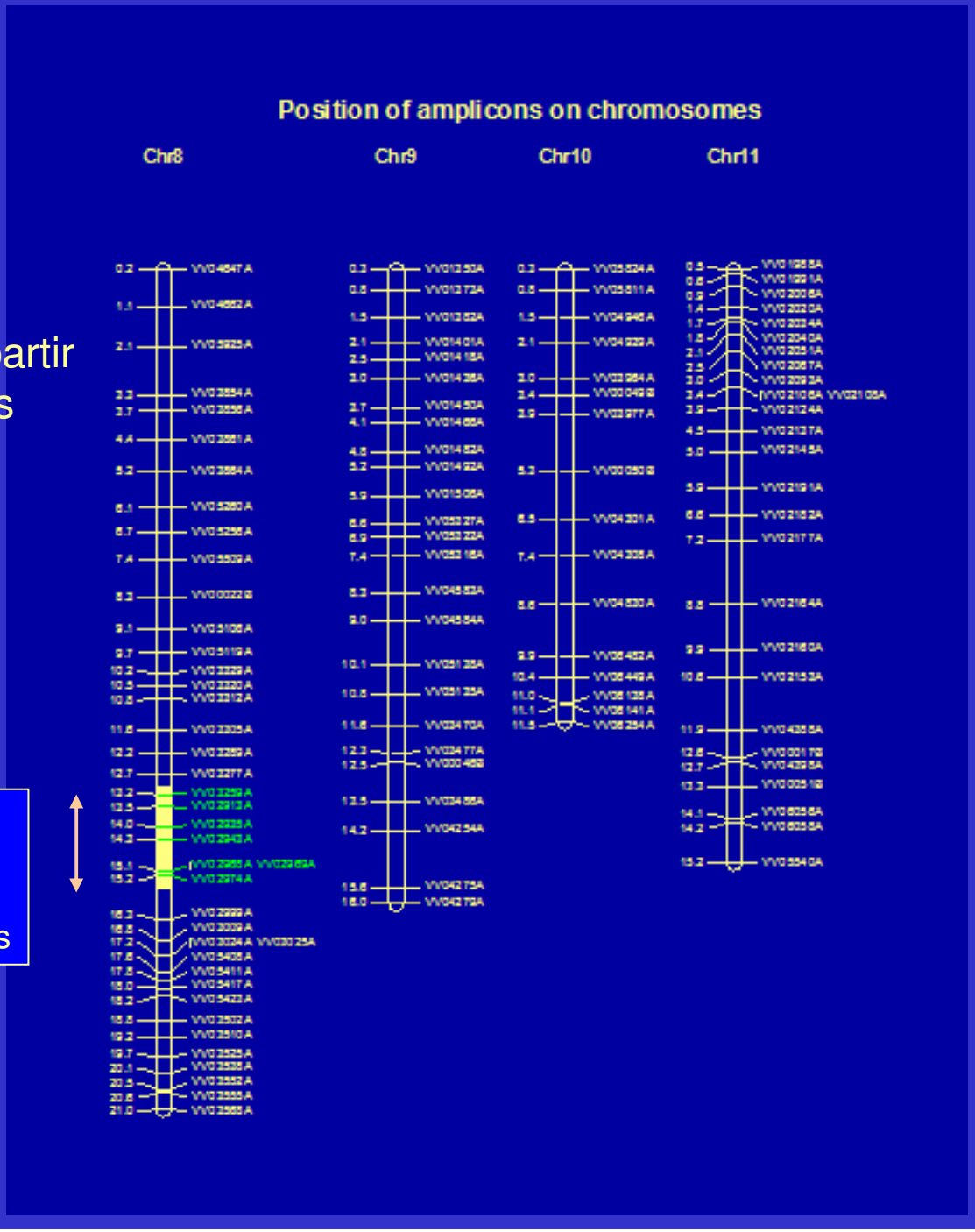


M&M

B. Cartographie

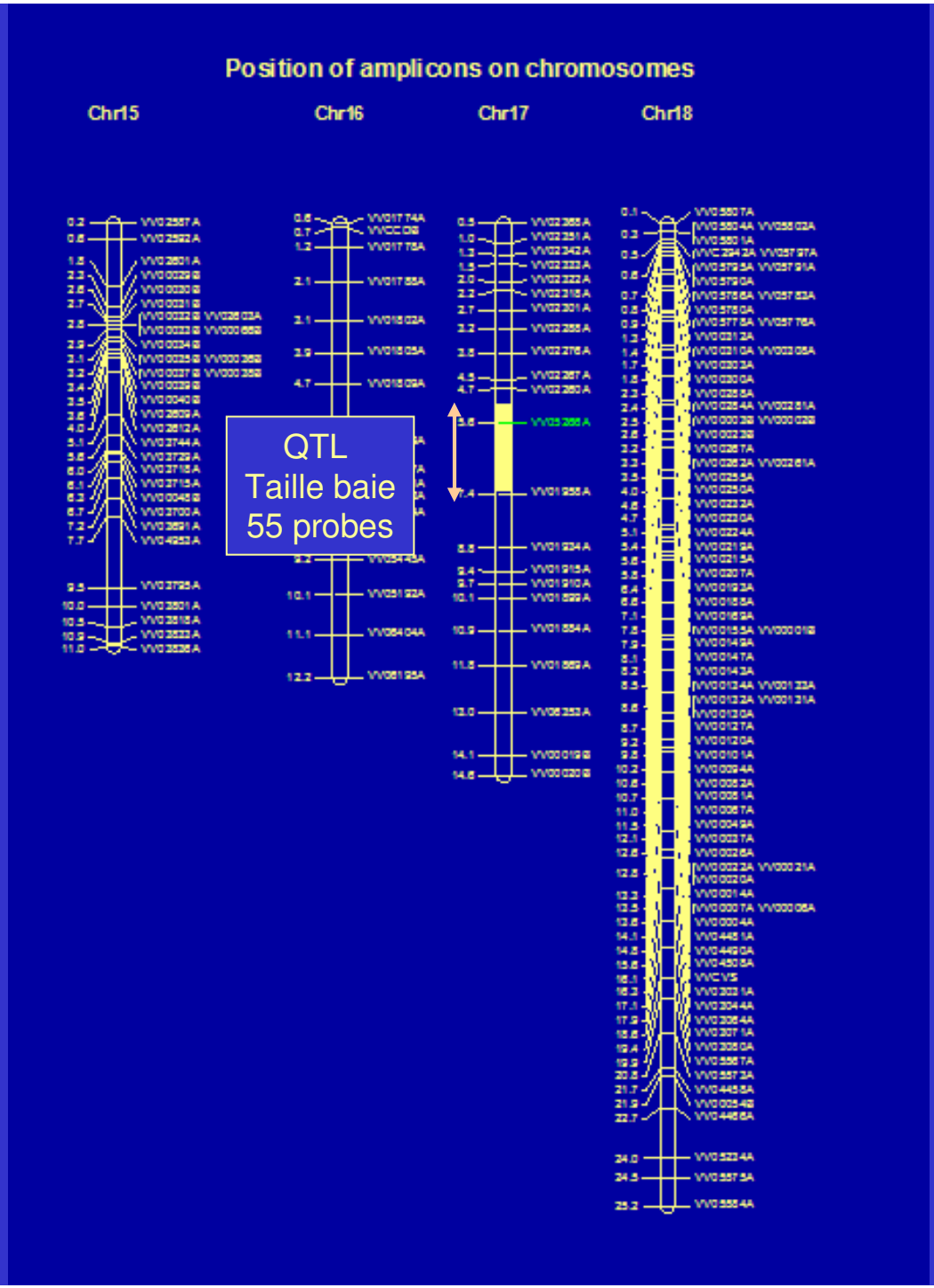
Cette carte a été construite à partir des positions de 920 amplicons (moins le Chr random)

(seulement quelques chrs sont representés)



M&M

B. Cartographie

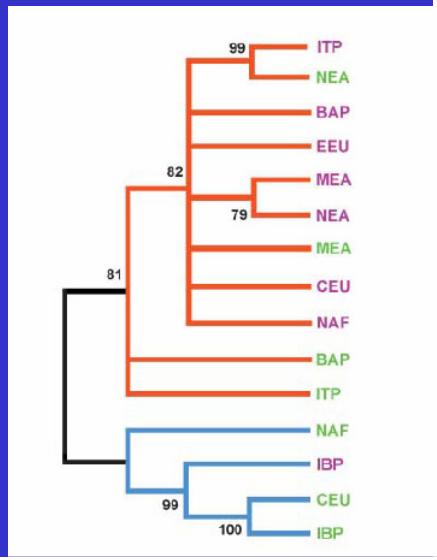


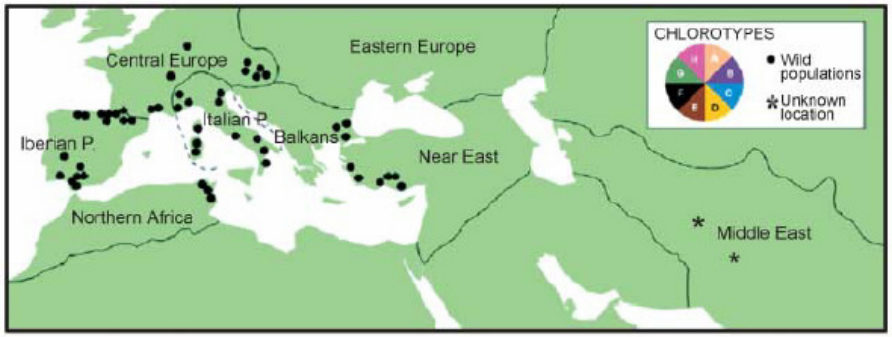
C. Echantillon étudié

M&M

CC24 - Est	17 + Sultanine
CC24 - Ouest	7 + Syrah
PN40024	1
CC24 - Cuve	*(15)*
CC24 - Table	*(9+1)*
V. sylvestris	7
Espèces Amérique	7
Espèces Asie	6
Total ADNs	*47*

Arroyo et al., Mol. Ecol. 2006





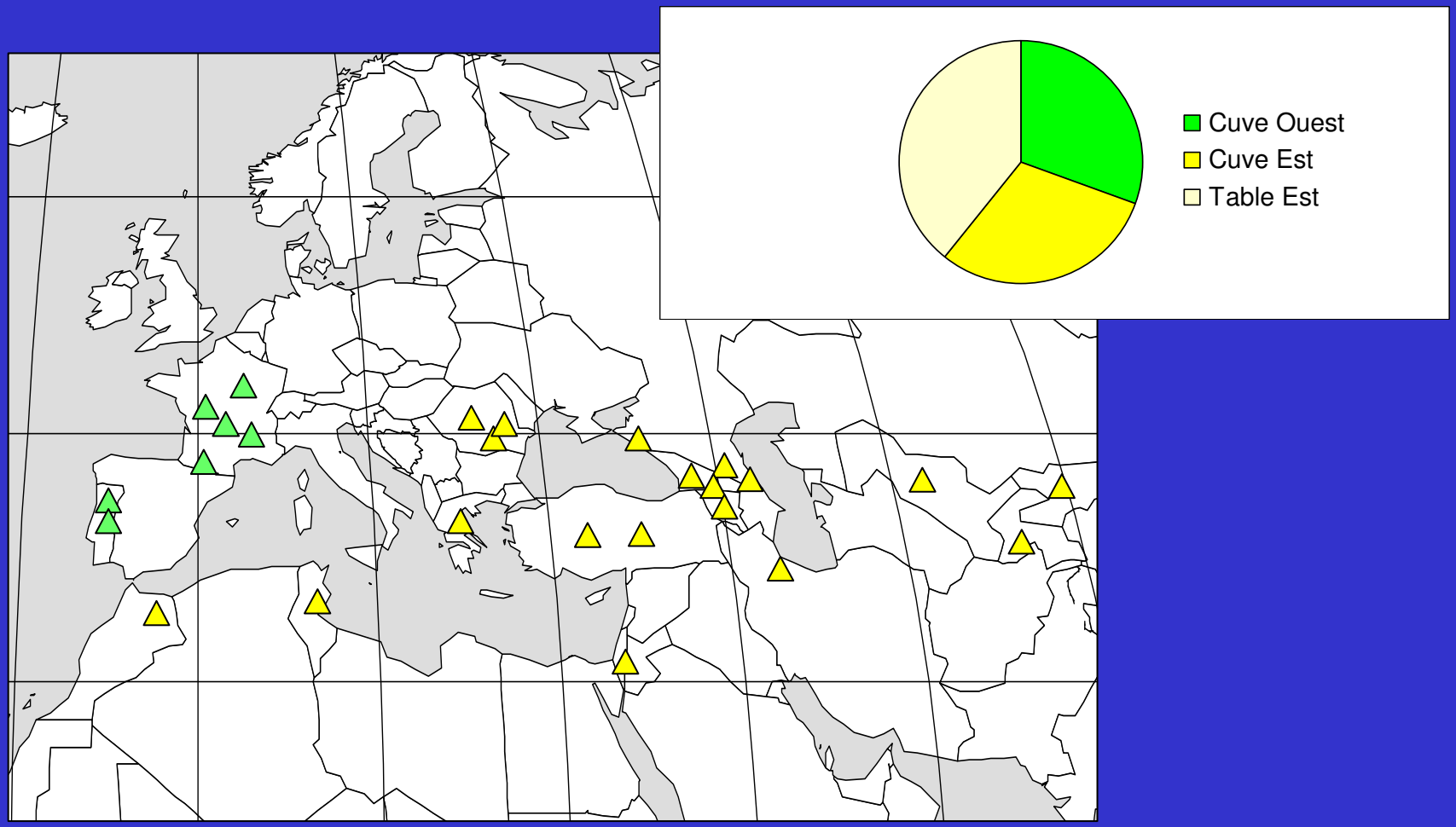
	IBP	CEU	ITP	NAF	BAP	EEU	NEA	MEA
Wild								
n	388	68	36	27	8		132	29
h±SD	0.13±0.05	0.03±0.01	0.43±0.05	0.34±0.05	0.40±0.08		0.36±0.05	0.44±0.09
Cultivated								
n	61	58	52	34	67	106	75	60
h±SD	0.27±0.04	0.39±0.06	0.39±0.03	0.41±0.07	0.43±0.07	0.40±0.06	0.41±0.07	0.36±0.07

C. Échantillon étudié

M&M

Structure of the cultivated sample

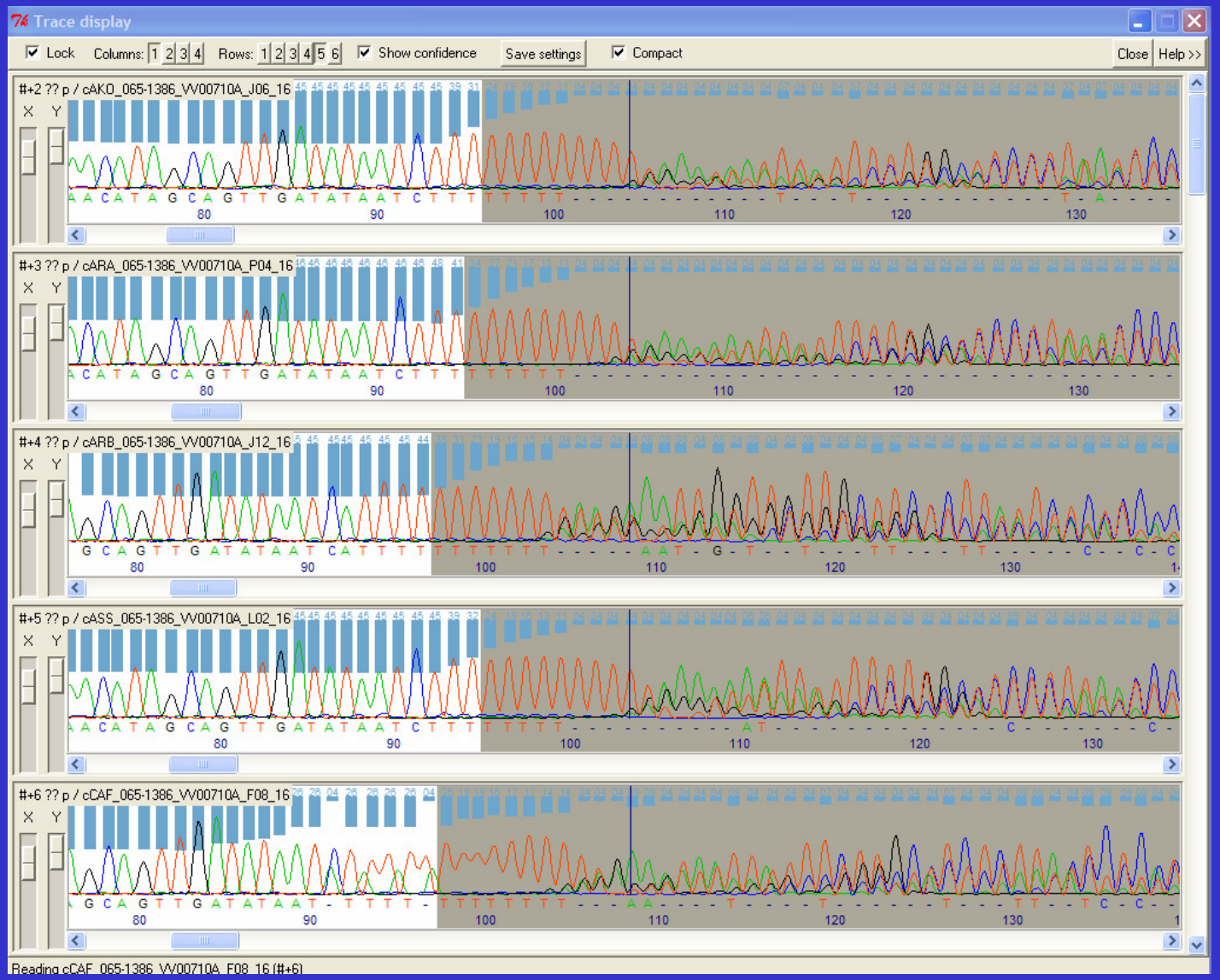
(Structure) : CC24 + Syrah + Sultanine



D'après Le Cunff et al., 2008

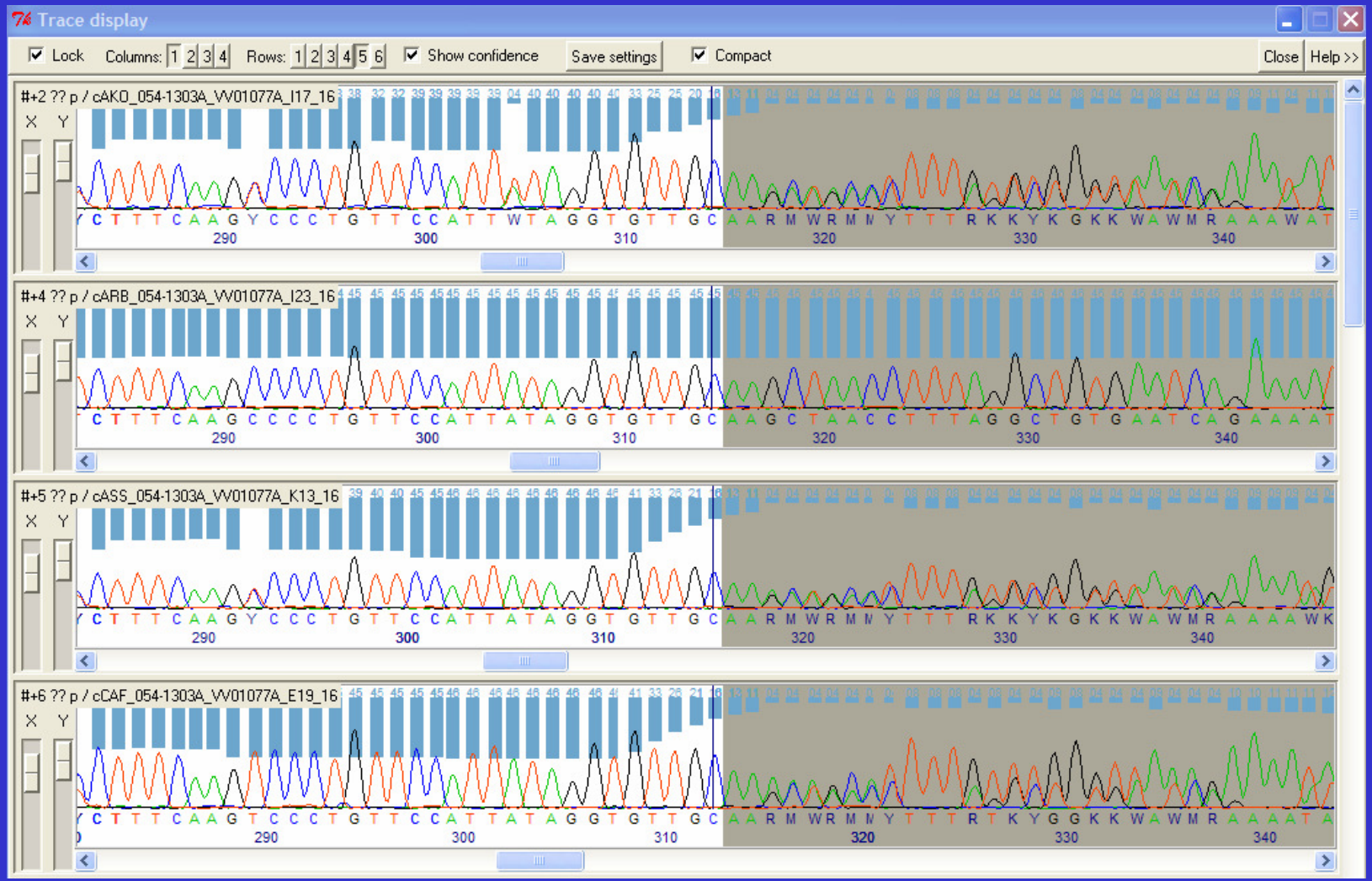
D. Qualité du reséquençage

Exemple d'un microsat hétérozygote



D. Qualité du réséquençage

Exemple d'un pseudogène qui co-amplifie

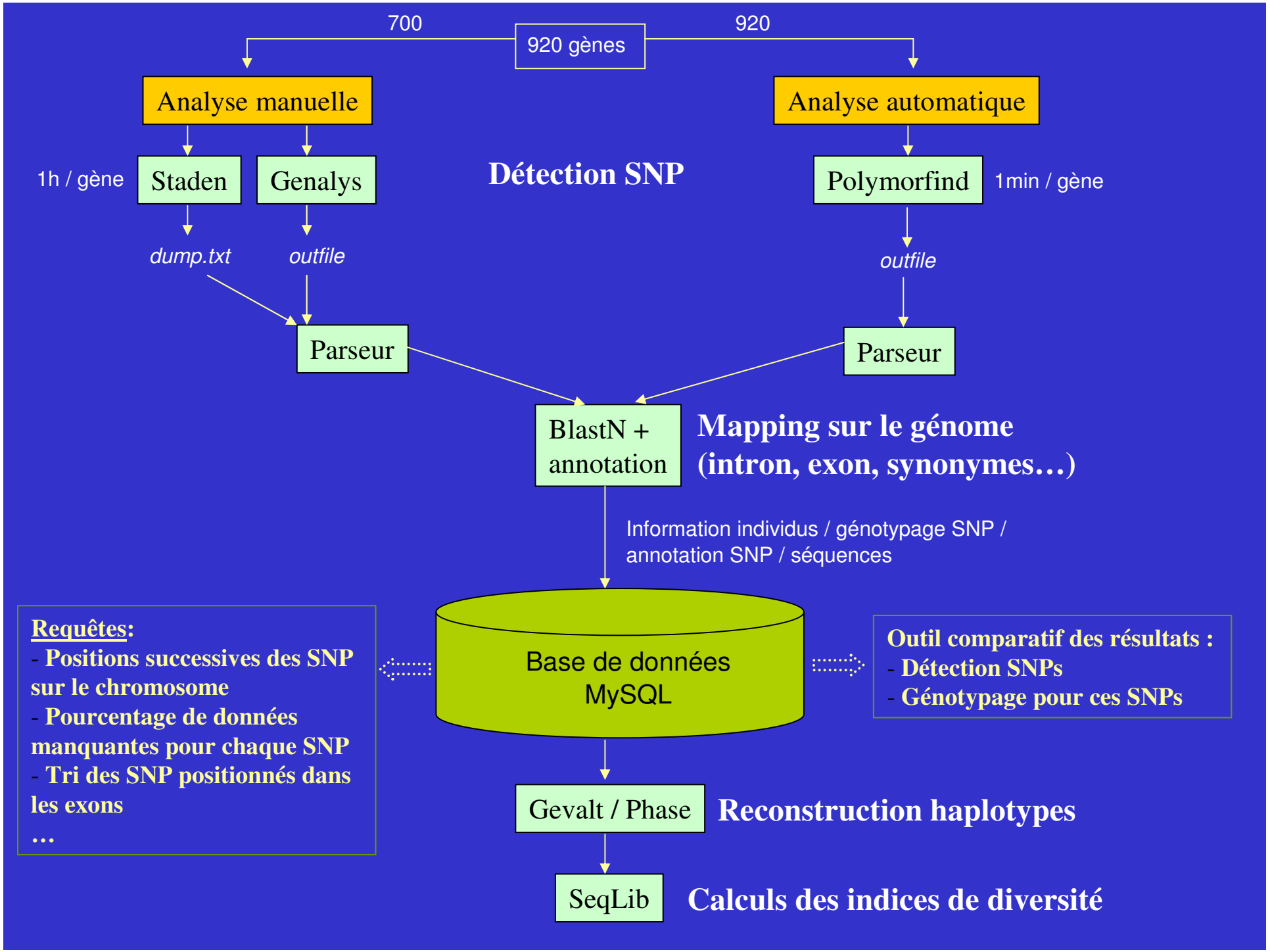


D. Qualité du reséquençage – récapitulatif

(paramètre TPAR Phred)

N sequences totales	52 542	100%
N. sequences illisibles	4 229	8%
N. sequences mauvaise qualité	6 658	13%
N. sequences lisibles	41 655	79%
N genes sequencés	1 116	100%
N genes exploitables	**920**	82%
N gènes lus manuellement	700	63%
N gènes lus automatiquement (polymorfind)	920	82%
N séq illisibles à cause de l'hétérozygotie	9212	18%
N sequences illisibles pour d'autres causes	1 675	3%

Exploitation bioinformatique et stockage des données


700
920 gènes
920
Analyse manuelle
Analyse automatique
1h / gène
Staden
Genalys
Détection SNP
Polymorfind
1min / gène
dump.txt
outfile
outfile
Parseur
Parseur
BlastN + annotation
Mapping sur le génome (intron, exon, synonymes…)
Information individus / génotypage SNP / annotation SNP / séquences
Requêtes:
- Positions successives des SNP sur le chromosome
- Pourcentage de données manquantes pour chaque SNP
- Tri des SNP positionnés dans les exons
…
Base de données MySQL
Outil comparatif des résultats :
- Détection SNPs
- Génotypage pour ces SNPs
Gevalt / Phase
Reconstruction haplotypes
SeqLib
Calculs des indices de diversité

Extraction des informations à partir d'un alignement

SNP Analysis

1. Load your data | 2. Individuals selection | 3. Run

Input format: *dump.txt* (from Staden) FASTA alignment

Number of files: collection unique

Upload a *dump.txt* :

Browse...

Or paste it here (load example)

Blast for mapping and annotation

Database: Vitis

Gevalt pre-processing

Multiallelic locus treatment:

treat as missing data

replace minority allele by majority allele

Calculation of diversity indexes

SeqLib pre-processing:

Discard sequences having more than 5 % of marker with missing data

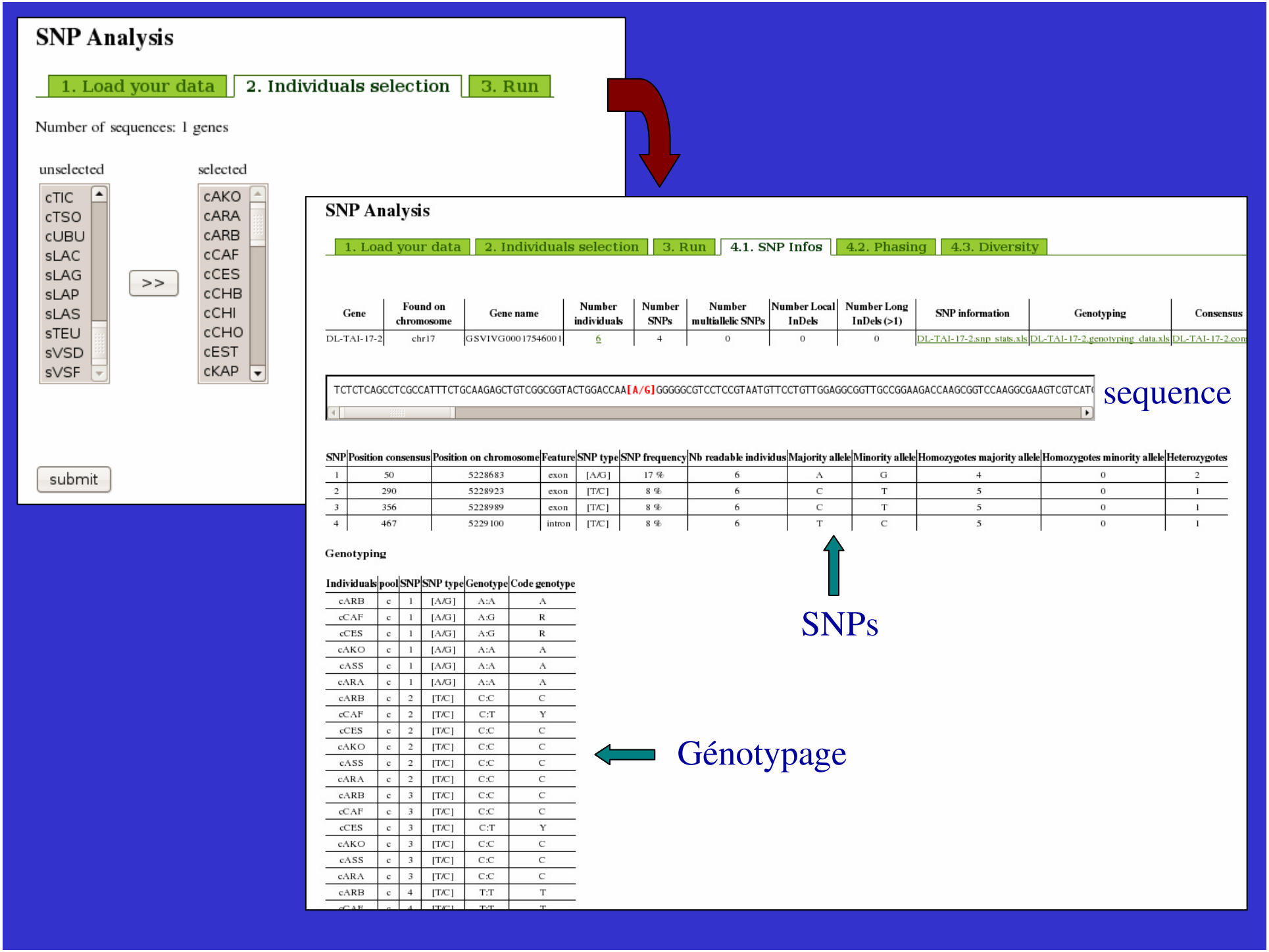
Network analysis

submit

DL-TAI-17-03cs

```
                                                                10        20        30        40        50        60        70        80        90       100
  +2 cAKO_0411-1065A_DL-TAI-17-3_J18_6                  .....T.............................M.........T...................A.......A.....................
  +3 cARA_0411-1065A_DL-TAI-17-3_P16_6  .............T.............................C.........T...................A.......A.....................
  +4 cARB_0411-1065A_DL-TAI-17-3_J24_6  .............Y.............................M.........T...................A.......A.....................
  +5 cCAF_0411-1065A_DL-TAI-17-3_F20_6  .............Y.............................M.........T...................A.......A.....................
  +6 cCES_0411-1065A_DL-TAI-17-3_H24_6  .............Y.............................M.........T...................A.......A.....................
  +7 cCHB_0411-1065A_DL-TAI-17-3_B16_6  .............Y.............................M.........T...................A.......R.....................
  +8 cCHI_0411-1065A_DL-TAI-17-3_L16_6          ....Y.............................M.........T...................A.......A.....................
  +9 cCHO_0411-1065A_DL-TAI-17-3_P20_6  .............Y.............................M.........T...................A.......A.....................
 +10 cEST_0411-1065A_DL-TAI-17-3_N22_6  .............Y.............................M.........T...................A.......A.....................
 +11 cKAP_0411-1065A_DL-TAI-17-3_F22_6                .............................M.........T...................A.......A.....................
 +12 cKIC_0411-1065A_DL-TAI-17-3_J14_6  .............Y.............................M.........T...................A.......R.....................
 +13 cLAE_0411-1065A_DL-TAI-17-3_N16_6                ..............................M.........T...................A.......A.....................
 +14 cLAR_0411-1065A_DL-TAI-17-3_H20_6  .............Y.............................M.........T...................A.......A.....................
 +15 cMAK_0411-1065A_DL-TAI-17-3_B24_6  .............Y.............................M.........T...................A.......A.....................
 +16 cMED_0411-1065A_DL-TAI-17-3_B18_6  .............T.............................C.........T...................A.......A.....................
 +17 cMEH_0411-1065A_DL-TAI-17-3_F16_6                ...............-.............M.........T...................A.......A.....................
 +18 cMUF_0411-1065A_DL-TAI-17-3_L24_6  .............Y.............................M.........T...................A.......A.....................
 +19 cOA7_0411-1065A_DL-TAI-17-3_D20_6          .....T.............................M.........T...................A.......A.....................
 +20 cORL_0411-1065A_DL-TAI-17-3_D24_6  .............T.............................C.........T...................A.......A.....................
 +21 cPER_0411-1065A_DL-TAI-17-3_D18_6  .............Y.............................M.........T...................A.......A.....................
 +22 cPLE_0411-1065A_DL-TAI-17-3_P22_6  .............Y.............................M.........T...................A.......A.....................
 +23 cPNI_0411-1065A_DL-TAI-17-3_L22_6          .....Y.............................M.........T...................A.......R.....................
 +24 cSUL_0411-1065A_DL-TAI-17-3_F24_6  .............Y.............................M.........T...................A.......A.....................
 +25 cSYR_0411-1065A_DL-TAI-17-3_N18_6                                                                        .......M.--...A.....................
 +26 cTIC_0411-1065A_DL-TAI-17-3_H14_6                                                                  .....................A.......A.....................
 +27 cTSO_0411-1065A_DL-TAI-17-3_F14_6                                                                      ....A.......A.....................
 +28 cUBU_0411-1065A_DL-TAI-17-3_L20_6            ..T.............................M.........T...................A.......R.....................
 +29 sLAC_0411-1065A_DL-TAI-17-3_N24_6                                            .A.........-T...................A.......A.....................
 +30 sLAG_0411-1065A_DL-TAI-17-3_J22_6  .............Y.............................M.........T...................A.......A.....................
 +31 sLAP_0411-1065A_DL-TAI-17-3_F18_6                                                                  .C.....................A.......A.....................
 +32 sLAS_0411-1065A_DL-TAI-17-3_H16_6
 +33 sTEU_0411-1065A_DL-TAI-17-3_B20_6  .............Y.............................M.........T...................A.......A.....................
 +34 sVSD_0411-1065A_DL-TAI-17-3_B22_6  .............Y.............................M.........T...................A.......A.....................
  +1 sVSF_0411-1065A_DL-TAI-17-3_L18_6  .............Y.............................M.........T...................A.......A.....................
     CONSENSUS                          CATCCTCAGACTTC-GGGCAAGTTGAGCTAGCTAACAGGGAAAT-AAGAACATCTYGATGAAAGTGGTGAATTCC-ACAGAAA-GATTGGTCTATTAGGCTTCA
```

dump.txt
issu de Staden


SNP Analysis
1. Load your data
2. Individuals selection
3. Run
Number of sequences: 1 genes
unselected
selected
cTIC
cTSO
cUBU
sLAC
sLAG
sLAP
sLAS
sTEU
sVSD
sVSF
>>
cAKO
cARA
cARB
cCAF
cCES
cCHB
cCHI
cCHO
cEST
cKAP
submit
SNP Analysis
1. Load your data
2. Individuals selection
3. Run
4.1. SNP Infos
4.2. Phasing
4.3. Diversity
Gene
Found on chromosome
Gene name
Number individuals
Number SNPs
Number multiallelic SNPs
Number Local InDels
Number Long InDels (>1)
SNP information
Genotyping
Consensus
DL-TAI-17-2
chr17
GSVIVG00017546001
6
4
0
0
0
DL-TAI-17-2.snp_stats.xls
DL-TAI-17-2.genotyping_data.xls
DL-TAI-17-2.cons
TCTCTCAGCCTCGCCATTTCTGCAAGAGCTGTCGGCGGTACTGGACCAA[A/G]GGGGGCGTCCTCCGTAATGTTCCTGTTGGAGGCGGTTGCCGGAAGACCAAGCGGTCCAAGGCGAAGTCGTCAT
sequence
SNP
Position consensus
Position on chromosome
Feature
SNP type
SNP frequency
Nb readable individus
Majority allele
Minority allele
Homozygotes majority allele
Homozygotes minority allele
Heterozygotes
1 50 5228683 exon [A/G] 17 % 6 A G 4 0 2
2 290 5228923 exon [T/C] 8 % 6 C T 5 0 1
3 356 5228989 exon [T/C] 8 % 6 C T 5 0 1
4 467 5229100 intron [T/C] 8 % 6 T C 5 0 1
SNPs
Genotyping
Individuals
pool
SNP
SNP type
Genotype
Code genotype
cARB c 1 [A/G] A:A A
cCAF c 1 [A/G] A:G R
cCES c 1 [A/G] A:G R
cAKO c 1 [A/G] A:A A
cASS c 1 [A/G] A:A A
cARA c 1 [A/G] A:A A
cARB c 2 [T/C] C:C C
cCAF c 2 [T/C] C:T Y
cCES c 2 [T/C] C:C C
cAKO c 2 [T/C] C:C C
cASS c 2 [T/C] C:C C
cARA c 2 [T/C] C:C C
cARB c 3 [T/C] C:C C
cCAF c 3 [T/C] C:C C
cCES c 3 [T/C] C:T Y
cAKO c 3 [T/C] C:C C
cASS c 3 [T/C] C:C C
cARA c 3 [T/C] C:C C
cARB c 4 [T/C] T:T T
Génotypage

Polymorfind

1. Load your data | 2. Run

Upload an *archive (.zip, .tar.gz) containing all the chromatograms* :

Browse...

Note:The archive must be in this form:

```
zip, tar.gz ==> +-- individual1_protocolId_geneName_xxx.ab1
                -- individual2_protocolId_geneName_xxx.ab1
                -- individual3_protocolId_geneName_xxx.ab1
                ...
```

Optionnaly, you can filter on individuals.
Enter the list of individuals to consider (individuals separat

submit

Analyse automatique par *Polymorfind*

Polymorfind

1. Load your data | 2. Run | 3. SNP Infos

Gene	Number individuals	Number SNPs	Number multiallelic SNPs	Number InDels	SNP information	Genotyping	Consensus	fasta alignment
DL-TAI-17-2	27	9	0	1	DL-TAI-17-2.snp_stats.xls	DL-TAI-17-2.genotyping_data.xls	DL-TAI-17-2.cons.fas	DL-TAI-17-2.align.fas

```
AAATTTCTGCTACTATAACACTACAACCTCTCTCAGCCTCGCCATTTCTGCAAGAGCTGTCGGCGGTACTGGACCAA[A/G]GGGGGCGTCCTCCGTAATGTTCCTGTTGGAGGCGGTTGCCGGAAG
```

SNP	Position consensus	SNP type	SNP frequency	Nb readable individus	Majority allele	Minority allele	Homozygotes majority allele	Homozygotes minority allele	Heterozygotes
1	78	[A/G]	20 %	27	A	G	20	4	3
2	150	[A/G]	4 %	27	A	G	25	0	2
3	246	[A/G]	4 %	27	G	A	25	0	2
4	273	[A/G]	7 %	27	G	A	23	0	4
5	289	[T/C]	6 %	27	T	C	24	0	3
6	318	[T/C]	2 %	26	C	T	25	0	1
								0	3
								0	1
								0	1
								0	3

Gene : DL-TAI-17-2

SNP	Position	Software	SNP type	Readable ind	Homozygotes minority	Heterozygotes	Software	SNP type	Readable ind	Homozygotes minority	Heterozygotes	Difference SNP	Difference missing data	Difference genotyping
1	chr17 : 5228683	Staden	[A/G]	31	7	4	Polymorfind	[A/G]	31	8	3	no	yes (4)	0 / 29
2	chr17 : 5228755	Staden	[A/G]	31	0	3	Polymorfind	[A/G]	31	0	3	no	yes (4)	0 / 29
3	chr17 : 5228851	Staden	[A/G]	31	0	3	Polymorfind	[A/G]	31	1	2	no	yes (4)	1 / 29 (display)
4	chr17 : 5228878	Staden	[A/G]	31	0	3	Polymorfind	[A/C/G]	32	1	4	no	yes (3)	0 / 29
5	chr17 : 5228894	Staden	[T/C]	31	0	5	Polymorfind	[T/C]	31	1	4	no	yes (4)	1 / 29 (display)
6	chr17 : 5228923	Staden	[T/C]	31	0	2	Polymorfind	[T/C]	31	0	1	no	yes (4)	0 / 28
7	chr17 : 5228989	Staden	[T/C]	31	0	3	Polymorfind	[T/C]	31	1	3	no	yes (4)	0 / 29
8	chr17 : 5229053	Staden		0	0	0	Polymorfind	indel	33	0	1	yes	#	#
9	chr17 : 5229094	Staden	[A/G]	30	0	1	Polymorfind	[A/G]	30	0	1	no	yes (6)	0 / 27
10	chr17 : 5229100	Staden	[T/C]	30	0	3	Polymorfind	[T/C]	30	0	1	no	yes (6)	0 / 27
11	chr17 : 5229130	Staden		0	0	0	Polymorfind	indel	32	0	3	yes	#	#
12	chr17 : 5229193	Staden	[T/C]	27	0	3	Polymorfind		0	0	0	yes	#	#

Difference in the list of analyzed individuals : cPER, cCAF, cKIC, cLAE

Comparatif des résultats

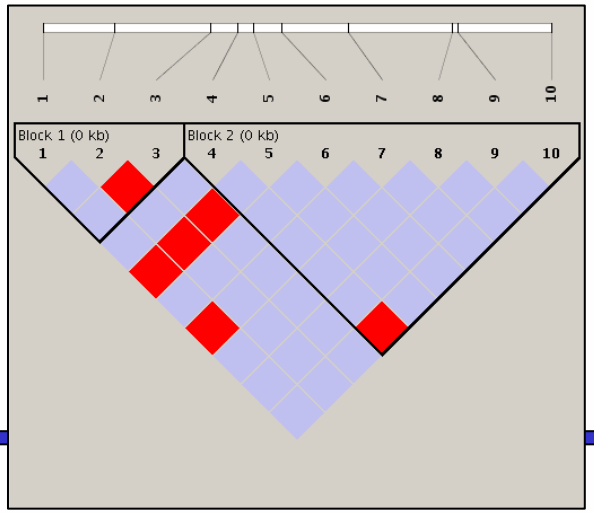
Exploitation des données : haplotypes, analyse de diversité

SNP Analysis

1. Load your data | 2. Individuals selection | 3. Run | 4.1. SNP Infos | **4.2. Phasing** | 4.3. Diversity | 4.4. Network analysis

Gene	Number individuals	Number SNPs	.ped file	.var file	tassel input file	haplotypes	distinct haplotypes	number distinct haplotypes	linkage disequilibrium	phased genotypes
DL-TAI-17-2	9	6	DL-TAI-17-2.ped	DL-TAI-17-2_var	DL-TAI-17-2.tassel_input.txt	DL-TAI-17-2.haplotypes	DL-TAI-17-2.distinct_haplo.fas	5	DL-TAI-17-2.ld	DL-TAI-17-2.phased_genotype.fas

Family	Individual	Child	1	2	3	4	5	6
cAKO	1	Singleton	A	G	C	C	C	T
cAKO	1	Singleton	A	G	C	C	T	T
cARA	2	Singleton	A	G	C	C	T	T
cARA	2	Singleton	A	G	C	C	T	T
cARB	3	Singleton	A	G	C	C	T	T
cARB	3	Singleton	A	G	C	C	T	T
cASS	4	Singleton	A	G	C	C	T	T
cASS	4	Singleton	A	G	C	C	T	T
cCAF	5	Singleton	G	G	T	C	T	T
cCAF	5	Singleton	A	G	C	C	T	T
cCES	6	Singleton	G	G	C	T	T	T
cCES	6	Singleton	A	G	C	C	T	T
cCHI	8	Singleton	A	G	C	C	T	T
cCHI	8	Singleton	A	G	C	C	T	T
cCHO	9	Singleton	A	A	C	C	T	C
cCHO	9	Singleton	A	G	C	C	T	T



SNP Analysis

1. Load your data | 2. Individuals selection | 3. Run | 4.1. SNP Infos | 4.2. Phasing | **4.3. Diversity** | 4.4. Network analysis

Gene	Number individuals	Number SNPs	phased genotypes	number individuals for seqlib	sequence length	sequence length efficient	number polymorphic sites	tW	Pi	D	number haplotypes	heterozygoty
DL-TAI-17-1	34	6	DL-TAI-17-1.phased_genotype.fas	30	632	522	6	0.00246	0.00219	-0.27766	7	0.609
DL-TAI-17-2	33	10	DL-TAI-17-2.phased_genotype.fas	27	667	593	9	0.00333	0.00187	-1.21310	8	0.576
DL-TAI-17-3	34	45	DL-TAI-17-3.phased_genotype.fas	27	613	536	44	0.01801	0.02007	0.38705	53	0.981
DL-TAI-17-4	33	8	DL-TAI-17-4.phased_genotype.fas	30	633	633	6	0.00203	0.00189	-0.17087	10	0.629
DL-TAI-17-5	28	6	DL-TAI-17-5.phased_genotype.fas	25	632	621	6	0.00216	0.00117	-1.16194	6	0.452
DL-TAI-17-6	30	25	DL-TAI-17-6.phased_genotype.fas	26	672	618	23	0.00824	0.00533	-1.13671	16	0.769
DL-TAI-17-7	27	25	DL-TAI-17-7.phased_genotype.fas	25	715	625	25	0.00893	0.00704	-0.69162	32	0.943
DL-TAI-17-8	23	18	DL-TAI-17-8.phased_genotype.fas	11	582	580	10	0.00473	0.00469	-0.03058	5	0.595
DL-TAI-17-9	23	13	DL-TAI-17-9.phased_genotype.fas	11	511	511	10	0.00537	0.00357	-1.14455	4	0.479

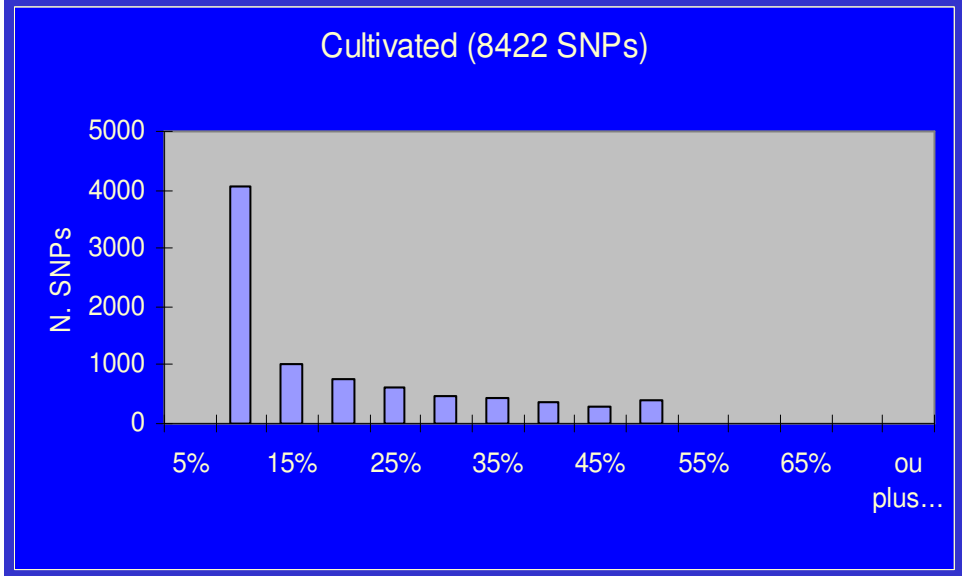
E. Premiers résultats

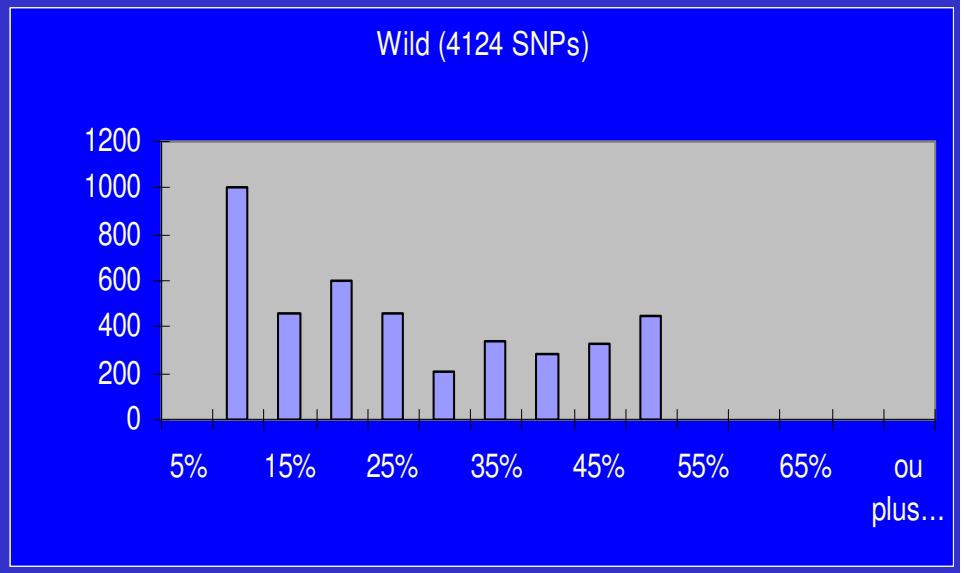
	N. Genes utilised	Average N. available inds	Missing genotypes	Average sequence length	Average N. polymor-phic sites	Average N. multialleli c SNPs	Average N. Long InDels (>1)
CULTIVARS	609	24.7	9%	671.9	11.0	0.43	0.91
WILD	533	6.3	10%	671.2	6.3	0.12	0.91
ASIAN SPECIE	518	5.2	14%	672.8	12.3	0.31	1.00
AM. SPECIES	476	5.8	18%	675.9	14.1	0.46	1.03

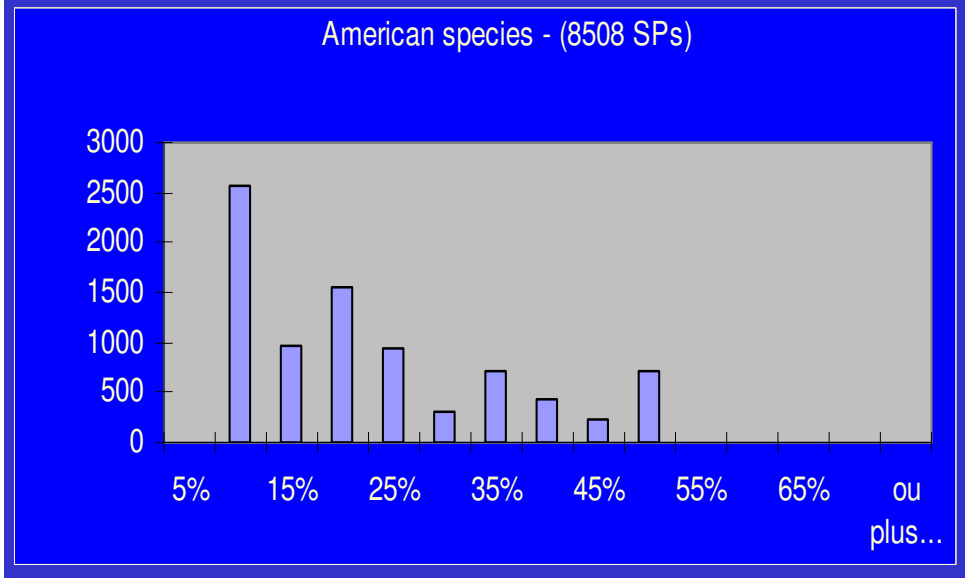
	% SNP dans exon	Average number haplotypes	Average heterozygo-sity	tW	Pi	D
CULTIVARS	54%	7.48	0.55	0.011	0.011	0.030
WILD	55%	3.29	0.45	0.016	0.017	0.103
ASIAN SPECIES		5.38	0.68	0.022	0.021	-0.111
AM. SPECIES		5.98	0.71	0.020	0.020	-0.149

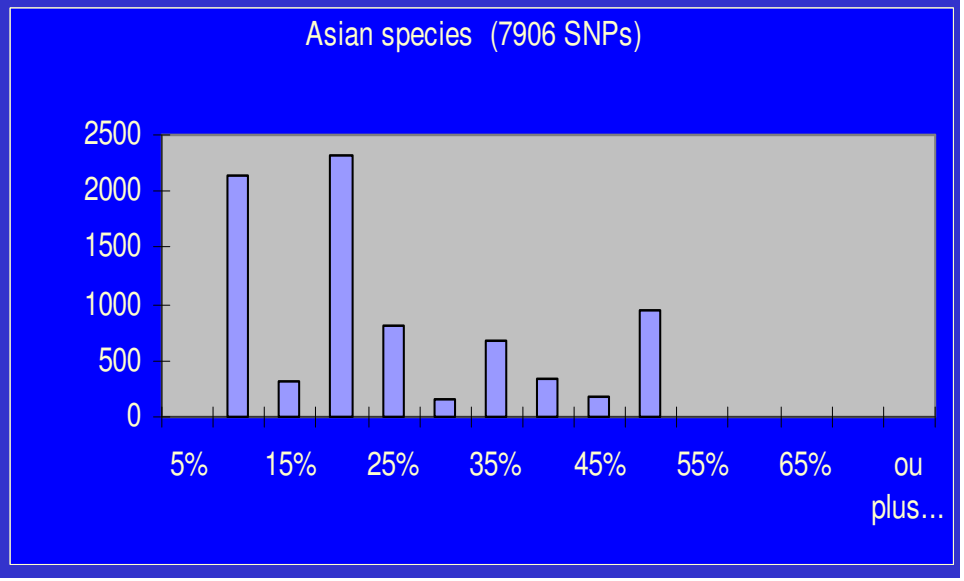
E. Premiers résultats

Distribution of frequency of SNP mutations





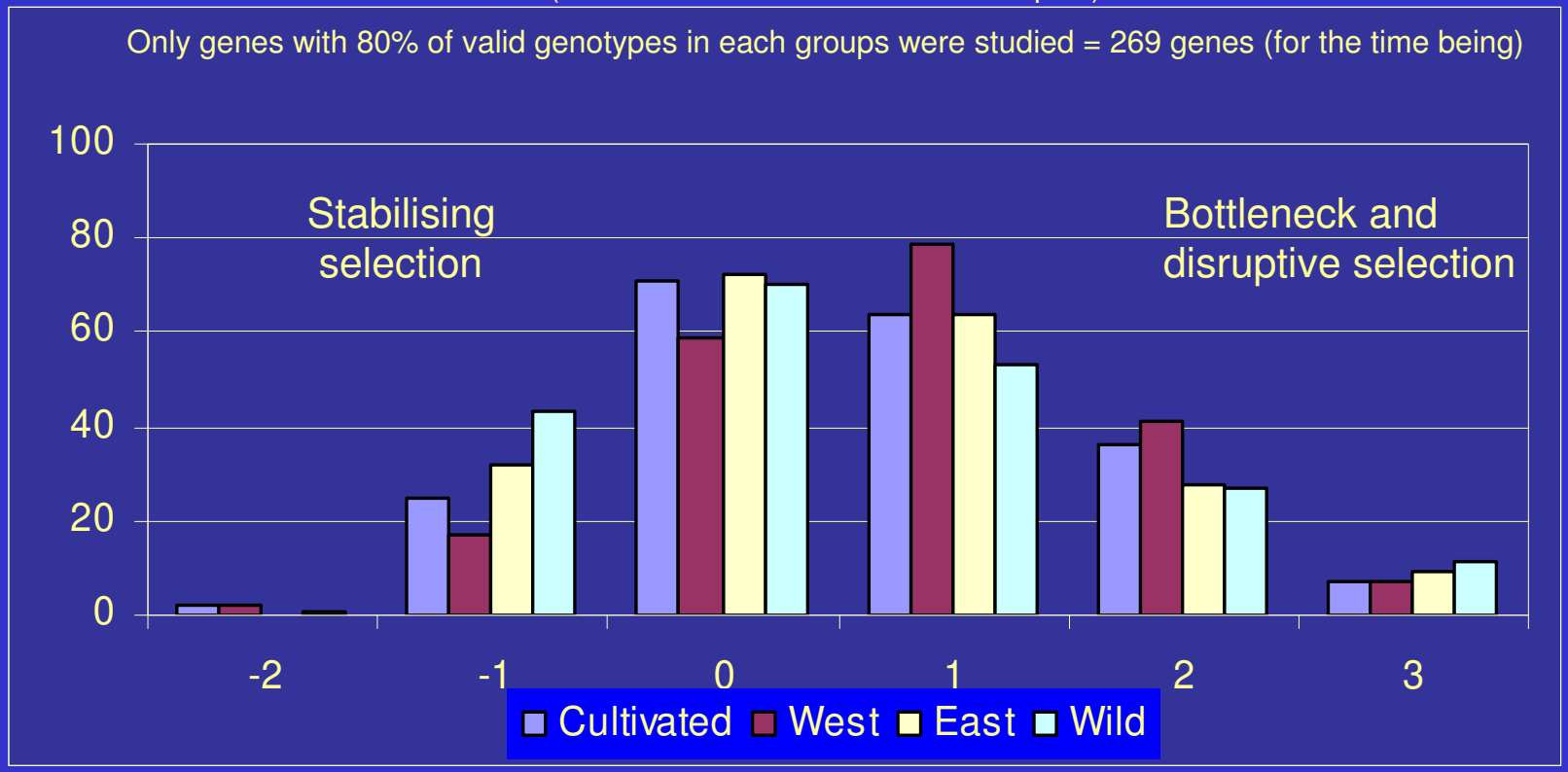




E. Premiers résultats

Distribution of the Tajima's D values (p=0.004)

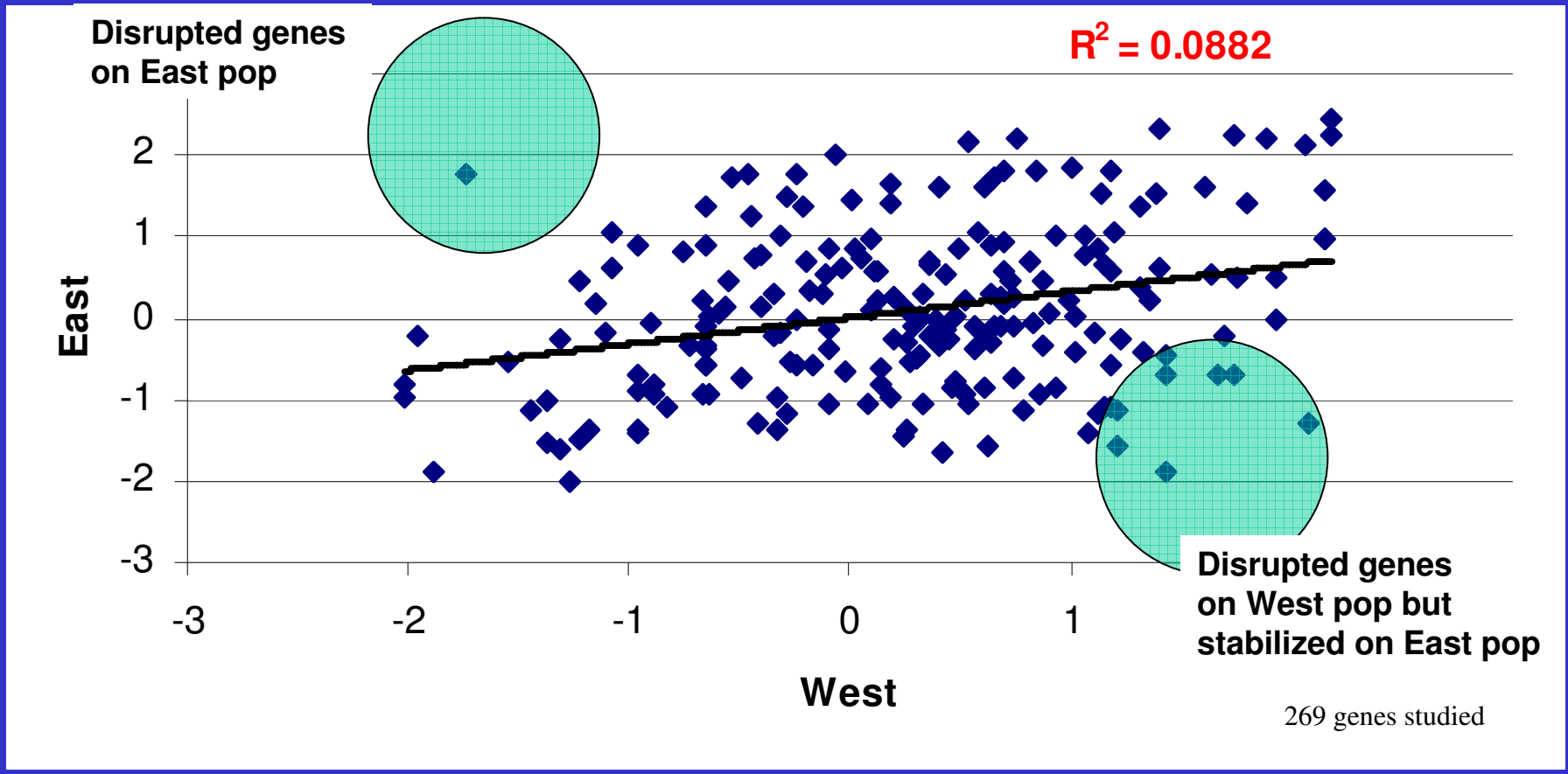
(Wild and structured cultivated samples)



	Average	Standard deviation	Mediane
Cultivated	0.1102	1.0053	0.0412
East	0.0861	1.0192	-0.0336
West	**0.2568**	0.9367	**0.3210**
Wild	-0.0425	1.1219	-0.2481

E. Premiers résultats

Scanner plot of Tajima's D values in West and East populations

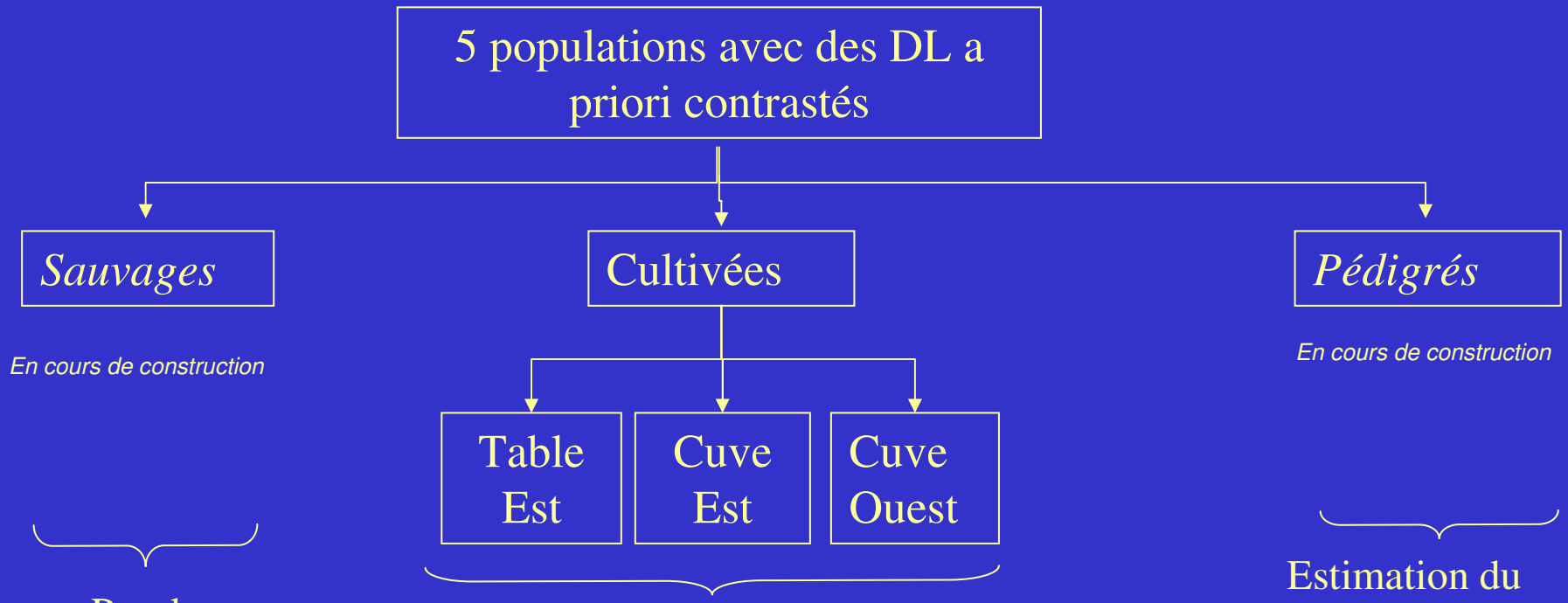


Selection acts on different genes in each group

Are these genes involved in QTLs, functions ?

F. Perspectives

Comparer le DL entre 5 populations ayant des histoires évolutives différentes

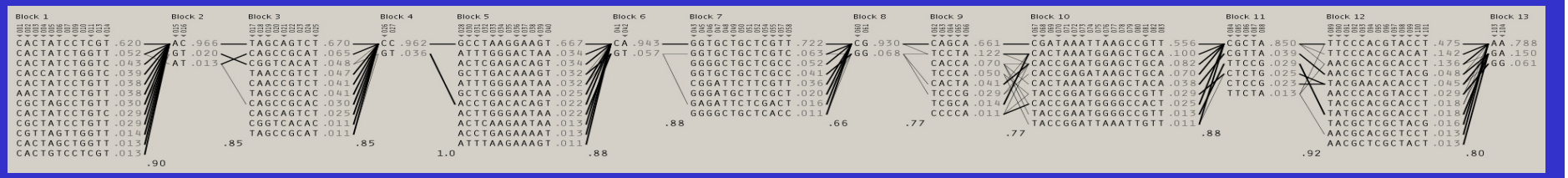


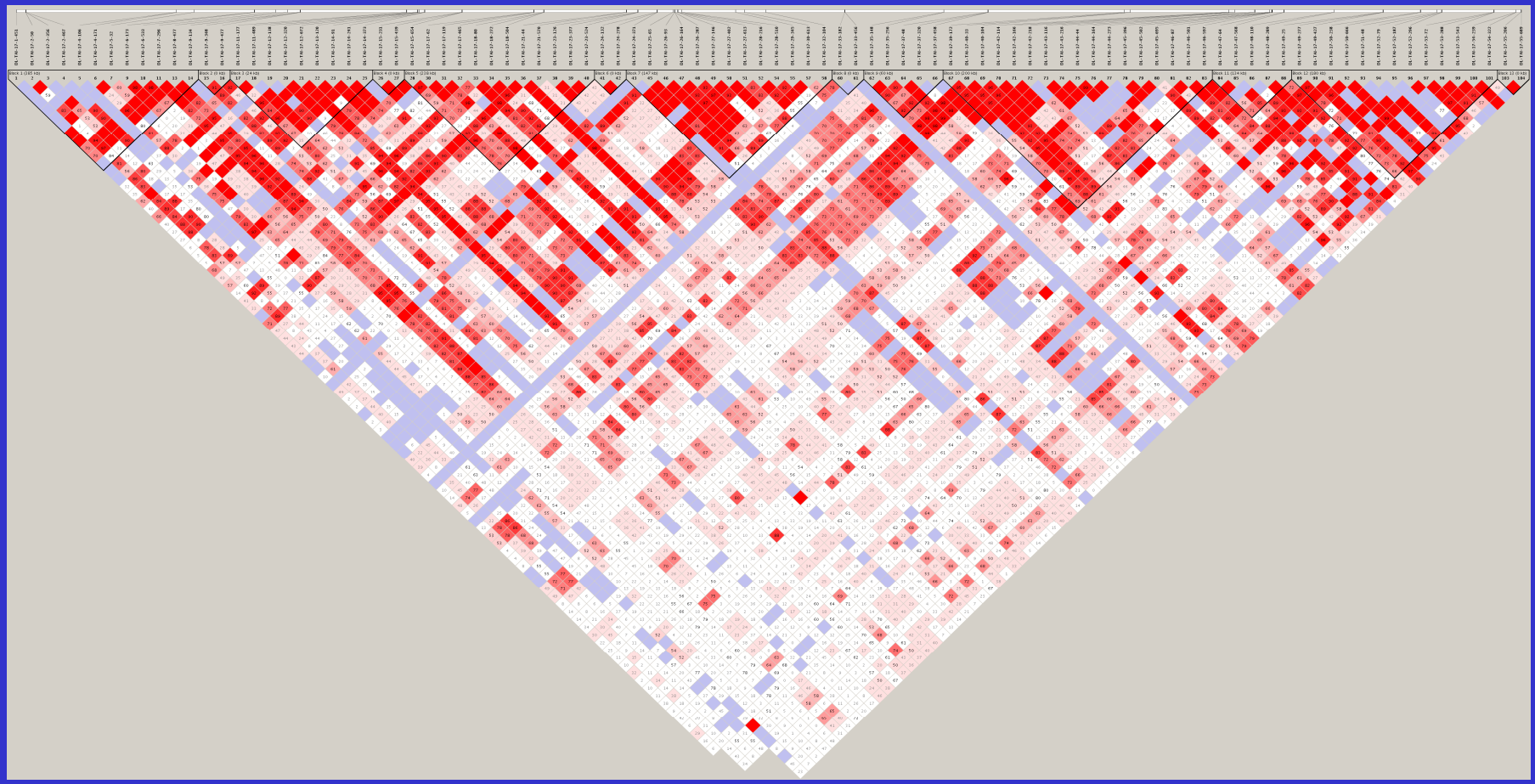
⇒Modélisation variation du DL le long du génome?

⇒Reconstitution histoire évolutive des régions?

F. Perspectives

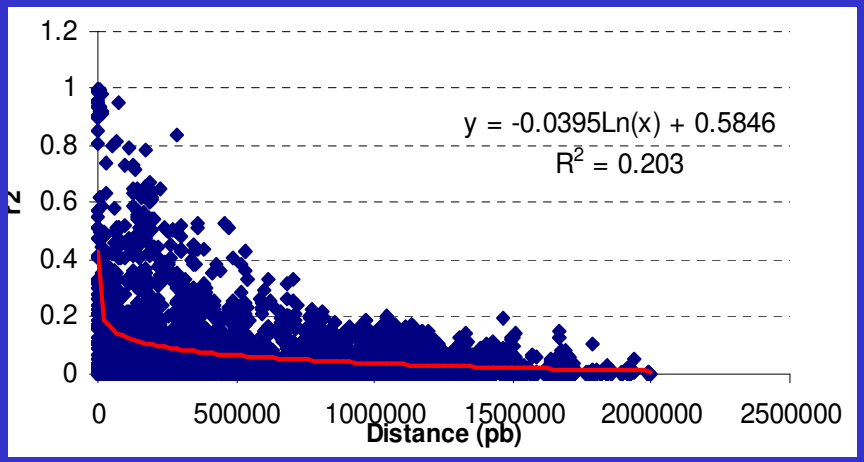
Structure du DL, différentiation et trace de sélection

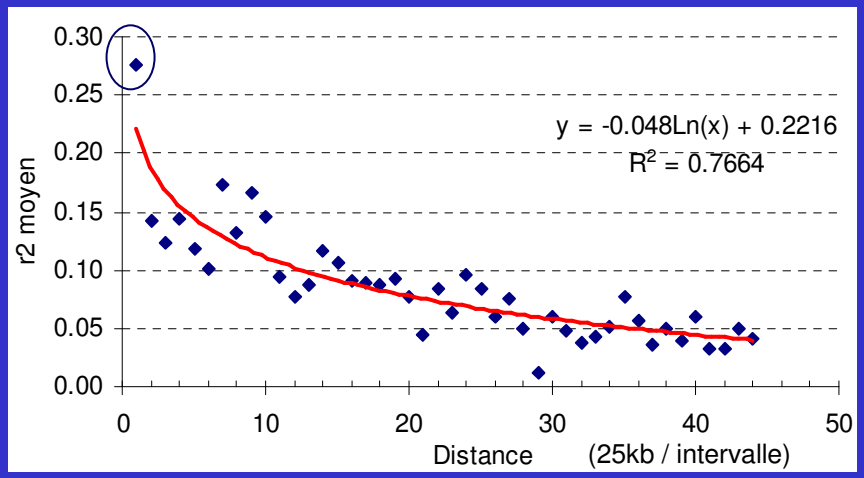




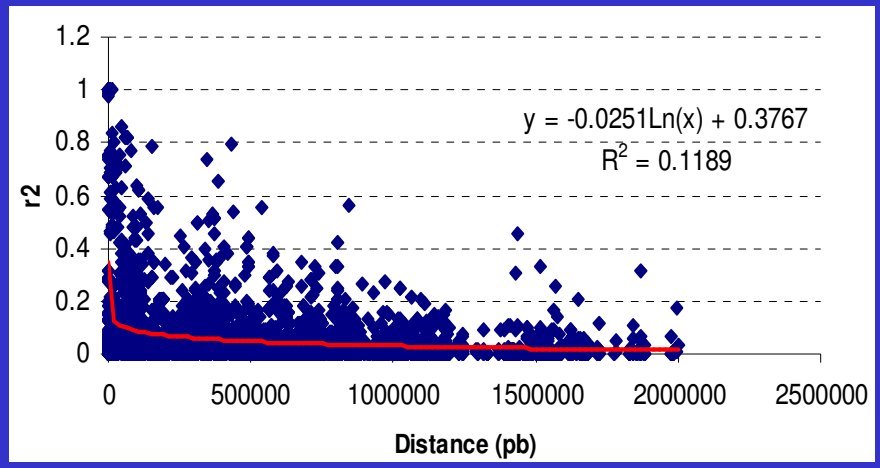
F. Perspectives

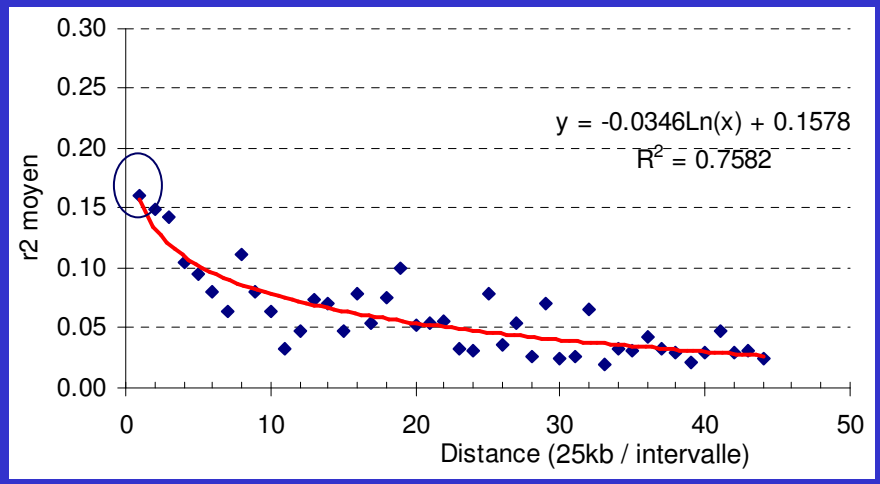
Région Taille (3 pop)





Région Tanins (3 pop)





Le r2 décroît lentement mais différemment dans les deux régions

Il semble décroître moins vite dans la région tanins / taille MAIS le r2 moyen est moins élevé notamment pour des distances courtes (<25kb)