

CAPRISNP

Détection et sélection de SNPs pour la
construction d'une puce 50K caprine

Contexte

Towards a 50K International Goat SNP chip

[Home](#)[Contacts](#)

As it was announced in July 2010, the International Goat Genome Consortium federates existing projects on SNP detection in order to produce a high density SNP chip. The deadline for ordering this 50K SNP is September the 30th, 2011 and the chip will be available at the end of year 2011. This first effort to give access to a high density genotyping tool may be completed in the future by other contributions and the tool may evolve regarding the needs of the International Community. If you are interested in ordering this tool, please contact Gwenola Tosser-Klopp (IGGC co-coordinator, INRA), Wenguang Zhang (IGGC co-coordinator, IMAU/KIZ/BGI) and Cindy Lawley (Illumina Scientist/Agriculture Consortia Manager Illumina, Inc.)



Data Contributions

- ✓ University of Utrecht, Netherlands (H. Heuven / M. Groenen)
Reduced Representation Libraries of 17 Goats, 100,000 SNPs



Data Processing

1. Essential Steps
 - ✓ Centralisation of the data at INRA, France



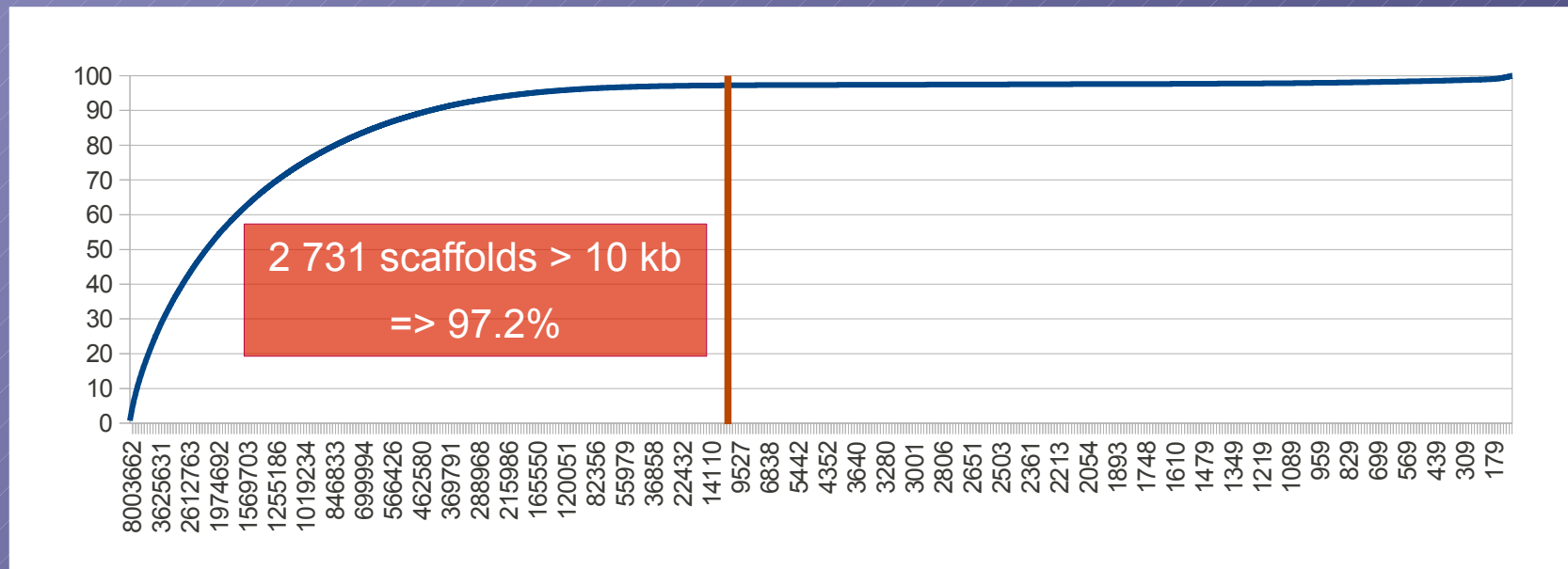
Timeline / Pricing

- ✓ Illumina chips available by beginning of December 2011
Milestones: First of July SNP
Submission letter from all consortia

Génome de référence

Assemblage produit par les chinois :

- 285 375 scaffolds/contigs => longueur totale = 2 662 629 614 pb
- avec des longueurs de 100 pb à 18 937 182 pb



Beijing Genome Institute / Kunming Institute of Zoology / Inner Mongolia Agricultural University,
China (W. Zhang, W. Wang)

Les données

INRA (Ph. Mulsant / R. Rupp / C.Moreno) :

RRL of 6 goats (454) and whole genome sequencing (HiSeq) of 13 Alpine, Saanen and Creole

Malaysian Agricultural Research and Development Institute,
Malaysia & DNA Landmarks, Canada (JJ Abdullah / TP Yu) :

Whole genome sequencing of 3*20 Boer, Savanna and Kacang meat and indigenous goats

University of Utrecht, Netherlands (H. Heuven / M. Groenen)

RRL of 17 Saanen dairy goats. 120 millions of 32 bp paired-ends sequences

Autres : EST + gènes d'intérêt

Le pipeline – données INRA (et hollandaises)

PROJETS RUNS DOWNLOAD

recherche OK

Projets > CAPRISNP

Projet CAPRISNP :

Etude de polymorphisme entre différents individus de chèvre après digestion et séquençage d'ADNg. 41 run(s) et 2 analyse(s) ont été réalisées sur le projet CAPRISNP. L'ensemble des données brutes et des résultats d'analyses occupent 315.57 Gb d'espace disque pour l'ensemble du projet.

**BWA aln
samtools sort**

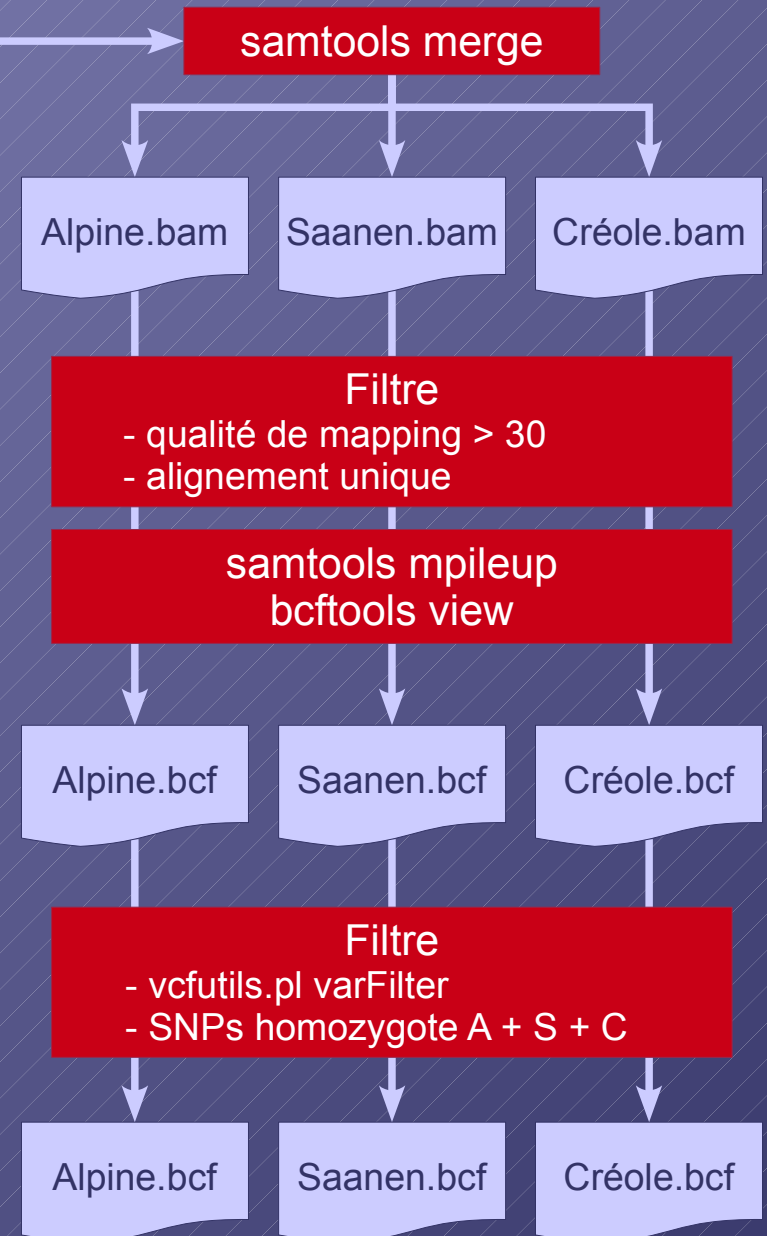
Runs réalisés :

Afficher 10 éléments par page Rechercher :

Nom du run	Nom du projet	Date	Espèce	Nature des données	Type	Nombre de séquences	Taille totale des séquences	Description	Séquenceur
5126	CAPRISNP	12-07-11	Chèvre	gDNA	1/8 FlowCell A - Lane 6	174 131 888	17 587 320 688	Concentration 7pM - Taille insert 265pb	HiSeq 2000
0026	CAPRISNP	12-07-11	Chèvre	gDNA	1/8 FlowCell A	122 178 002	12 440 078 202	Concentration 7pM - Taille	HiSeq 2000

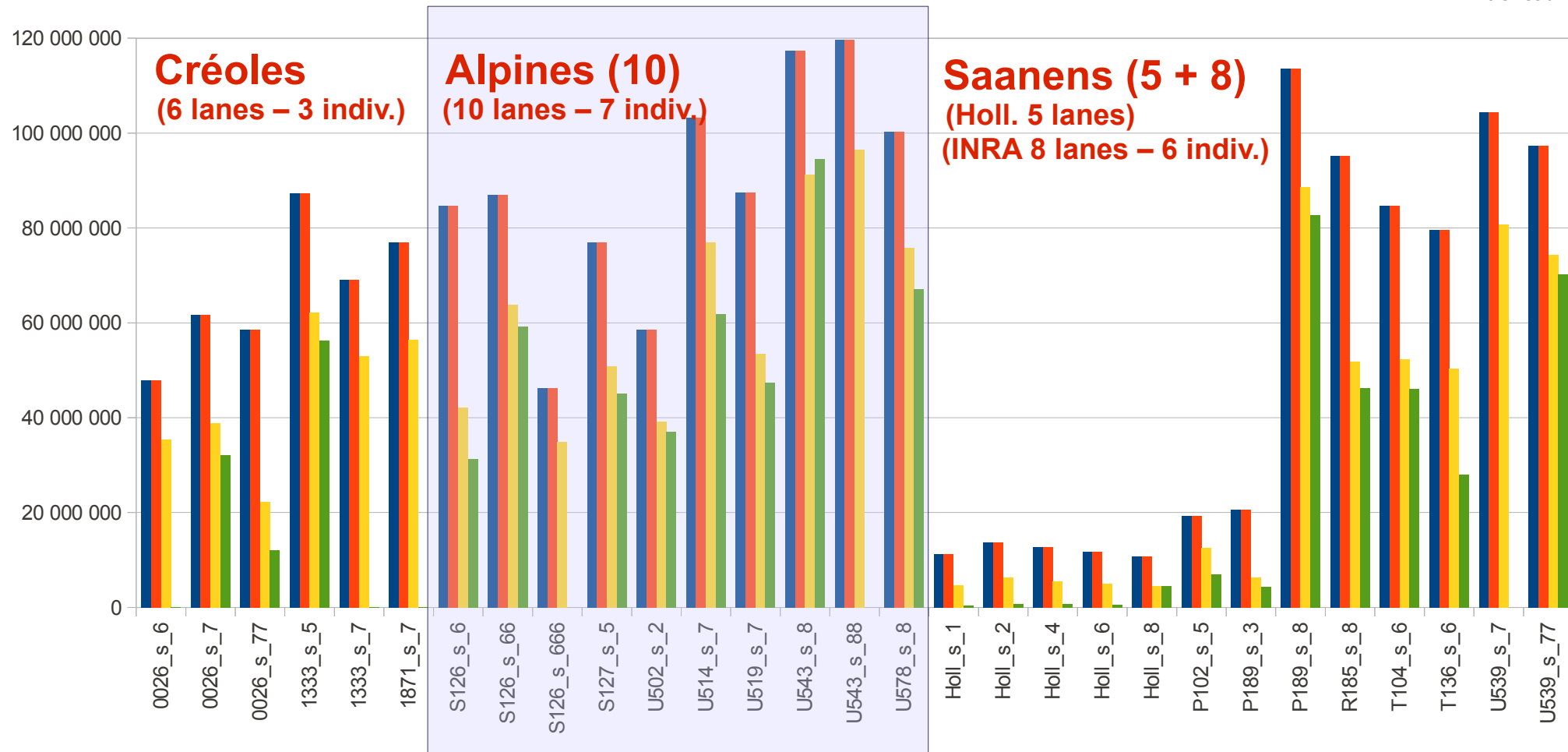
varFilter :

- profondeur min 6
- profondeur max 100
- min. allèle minoritaire 2



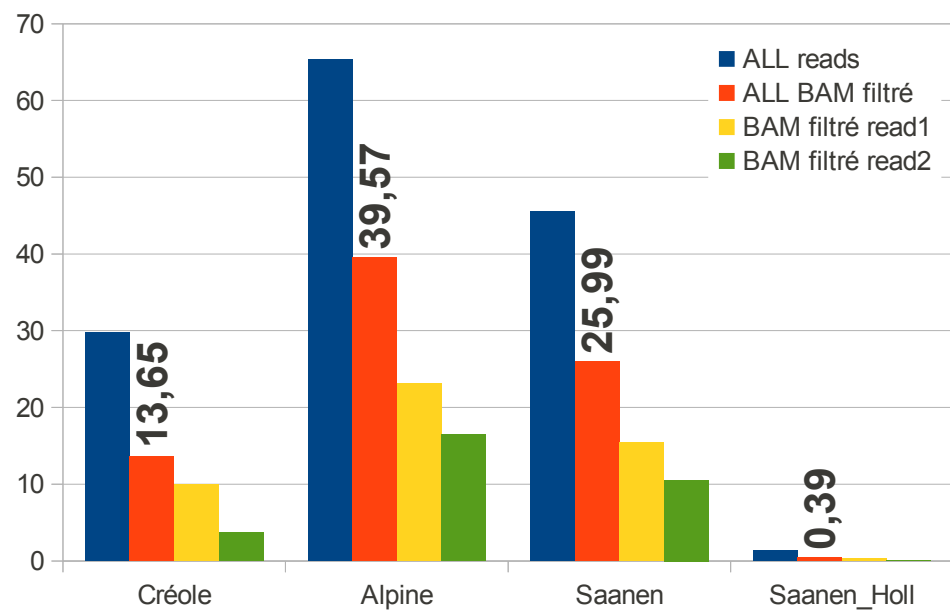
Les données -1-

■ Nb read1
 ■ Nb read2
 ■ BAM filtré read1
 ■ BAM filtré read2

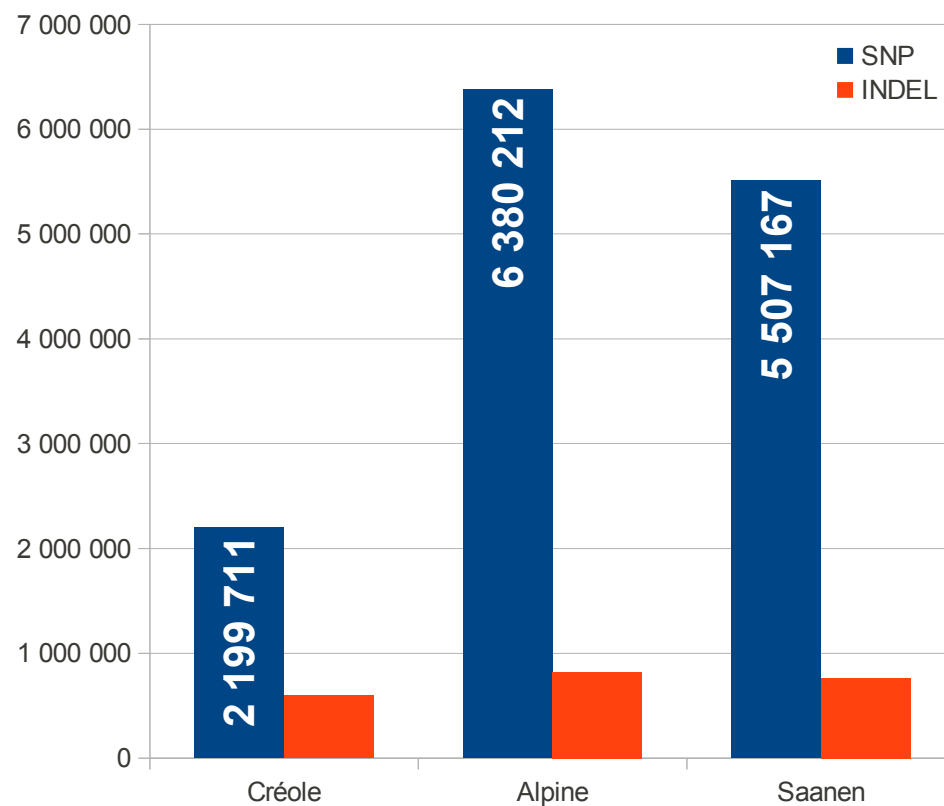


Les données -2-

Taux de couverture par race



Nombre de SNP/INDEL par race



Le pipeline – données externes

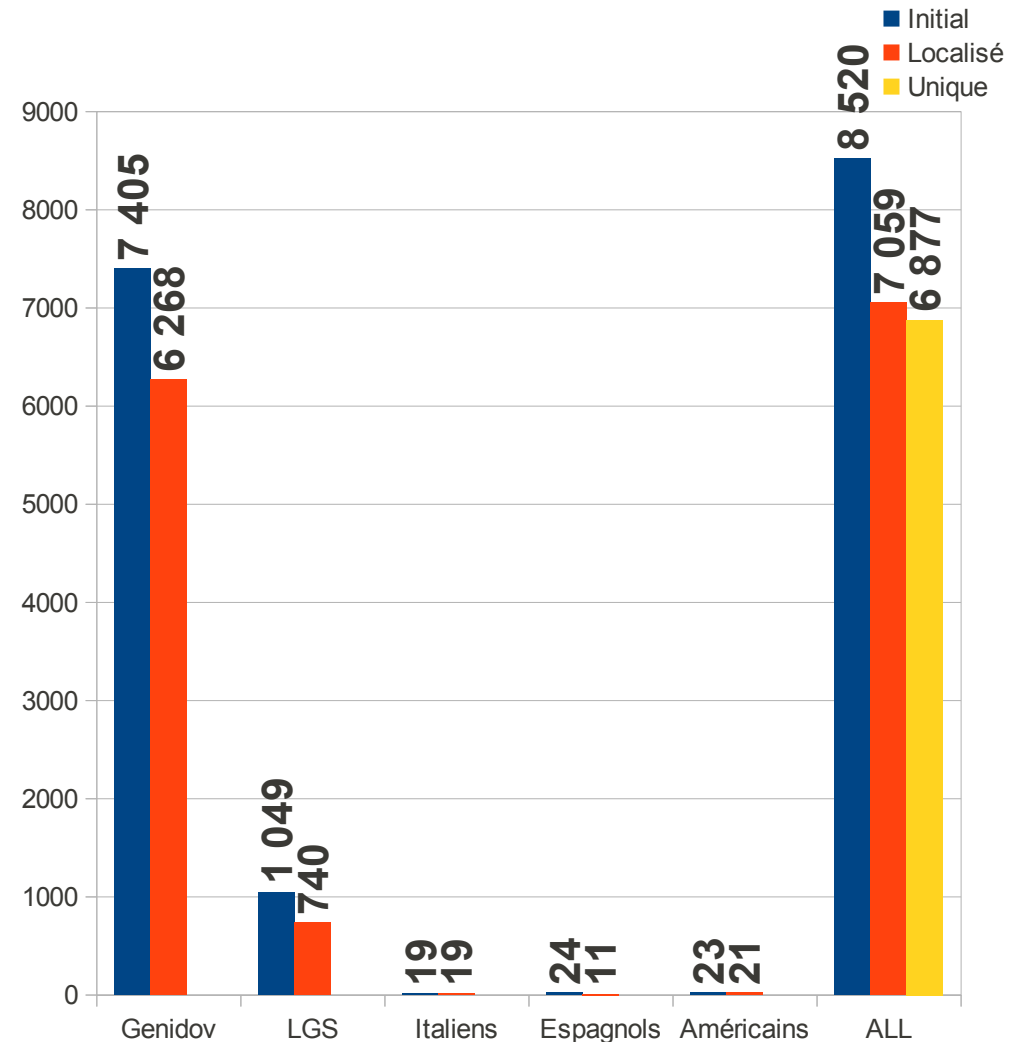
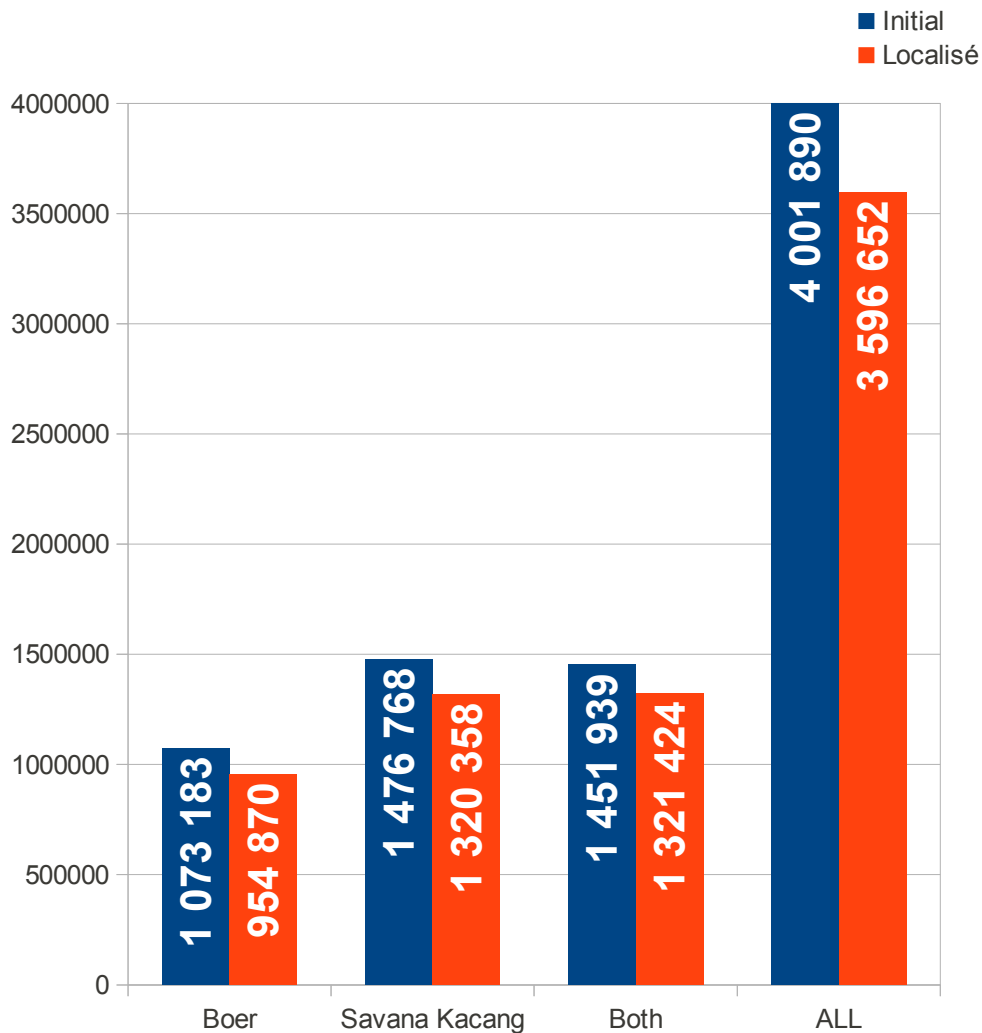
Données malésiennes

- × Alignement BWA (bwasw)
- × Extraction de la position des SNPs
- × Construction du fichier pour chargement dans la base de données

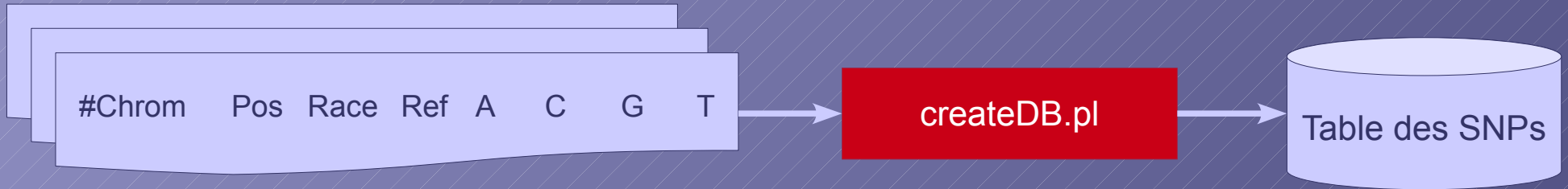
EST

- × Alignement BWA (bwasw)
- × Validation par alignement SIM4
- × Extraction de la position des SNPs
- × Construction du fichier pour chargement dans la base de données

Les données externes

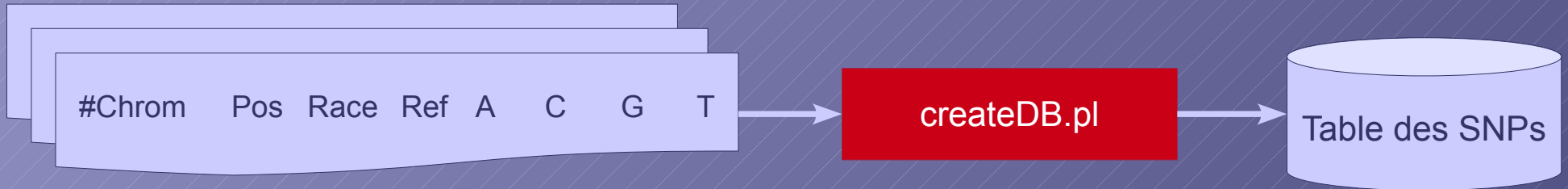


Chargement en base de données



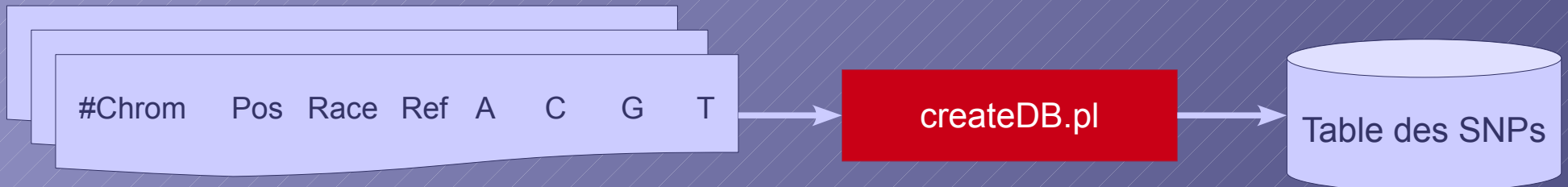
chrom	pos	ref_base	alpine	boer	creole	est	saanen	savana_kacang	illumina	category	infinium	maf	alleles	allele_type	dist_prev	dist_next	frame
scaffold1	54	T	A				AT			13	1	0.500	AT	2	54	76	40
scaffold1	130	C	CT		C					20	2	0.167	TC	2	76	1	-1
scaffold1	131	A	AG	AG	G			AG		11	2	0.500	AG	2	1	21	-1
scaffold1	152	A						AG		20	2	0.125	AG	2	21	1	-1
scaffold1	153	T	GT	GT				GT		11	2	0.300	TG	2	1	47	-1
scaffold1	200	G	A				AG	AG		11	2	0.500	AG	2	47	51	40
scaffold1	251	C	T	CGT	CT		CT	CGT		11	-1	0.167	TCG	3	51	71	40
scaffold1	322	A		AC				G		20	2	0.125	AC	2	71	117	60
scaffold1	439	A		AG				C		20	2	0.125	AG	2	117	67	60
scaffold1	506	G	AG	AG				AG		11	2	0.300	AG	2	67	348	60
scaffold1	854	A	G		G		AG	GTA		20	2	0.167	AG	2	348	40	40
scaffold1	894	T	CT							13	2	0.167	TC	2	40	11	-1

Chargement en base de données



chrom	pos	ref_base	alpine	boer	creole	est	saanen	savana_kacang	illumina	category	infinium	mafo	alleles	allele_type	dist_prev	dist_next	frame
sc	chrom	pos	ref_base	alpine	boer	creole	est	saanen	savana_kacang				AT	2	54	76	40
sc													TC	2	76	1	-1
sc	scaffold1	54	T	A				AT					AG	2	1	21	-1
sc	scaffold1	130	C	CT		C							AG	2	21	1	-1
sc	scaffold1	131	A	AG	AG	G			AG				TG	2	1	47	-1
sc	scaffold1	152	A						AG				AG	2	47	51	40
sc	scaffold1	153	T	GT	GT				AG				TCG	3	51	71	40
sc	scaffold1	200	G	A				AG	AG				AC	2	71	117	60
sc	scaffold1	251	C	T	CGT	CT		CT	CGT				AG	2	117	67	60
sc	scaffold1	322	A		AC								AG	2	67	348	60
sc	scaffold1	439	A		AG								AG	2	348	40	40
sc	scaffold1	506	G	AG	AG				AG				TC	2	40	11	-1
sc	scaffold1	854	A	G		G		AG									
sc	scaffold1	894	T	CT													

Chargement en base de données



chrom	pos	ref_base	alpine	boer	creole	est	saanen	savana_kacang	illumina	category	infinium	maf	alleles	allele_type	dist_prev	dist_next	frame
scaffold1	54	T	A						illumina	category	infinium	maf	alleles	allele_type	dist_prev	dist_next	frame
scaffold1	130	C	CT														
scaffold1	131	A	AG														
scaffold1	152	A					13		1	0.500	AT		2		54	76	40
scaffold1	153	T	GT				20		2	0.167	TC		2		76	1	-1
scaffold1	200	G	A				11		2	0.500	AG		2		1	21	-1
scaffold1	251	C	T				20		2	0.125	AG		2		21	1	-1
scaffold1	322	A					11		2	0.300	TG		2		1	47	-1
scaffold1	439	A					11		2	0.500	AG		2		47	51	40
scaffold1	506	G	AG				11		-1	0.167	TCG		3		51	71	40
scaffold1	854	A	G				20		2	0.125	AC		2		71	117	60
scaffold1	894	T	CT				20		2	0.125	AG		2		117	67	60
							11		2	0.300	AG		2		67	348	60
							20		2	0.167	AG		2		348	40	40
							13		2	0.167	TC		2		40	11	-1

Infinium

- II : autres => 1 sonde
- I : A/T et C/G => 2 sondes

Nombre de SNP par catégorie avant filtrage

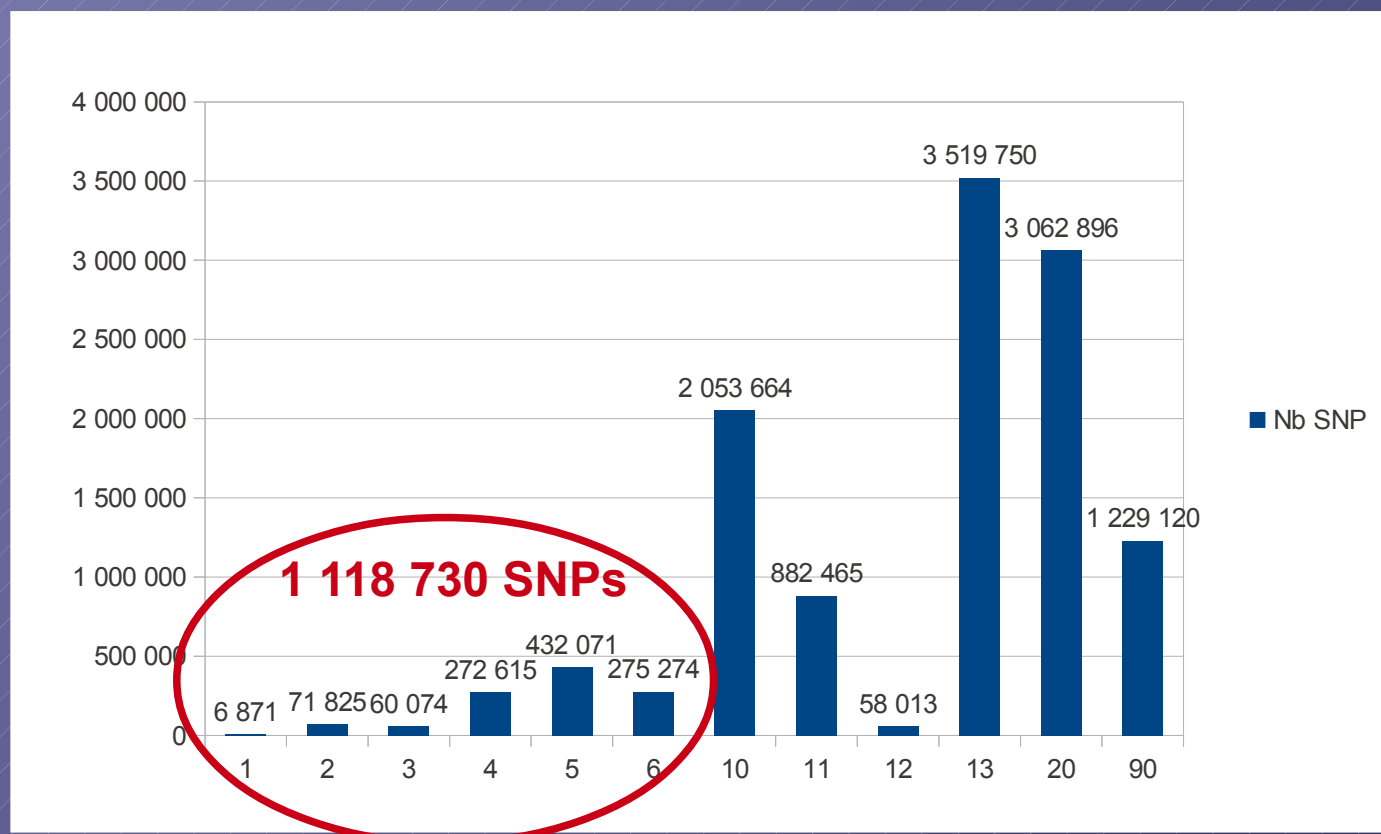
11 924 638 entrées dans la table avec :

- 1 229 120 Indels
- 10 695 518 SNPs

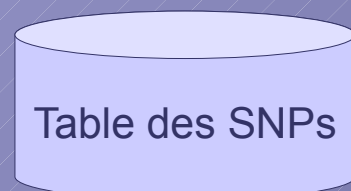
1 : EST
2 : Hétéroz. dans 5 races
3 et 4 : Hétéroz. dans 4 races
5 et 6 : Hétéroz. dans 3 races

10 : Hétéroz. S et A
11 : Hétéroz. (A ou S)
et (C ou B ou KS)
12 : Hétéroz. C et (B ou KS)
13 : Hétéroz. A ou S

20 : autres
90 : INDEL



Sélection des SNPs



Sélection + Tri

SNPs triés
par catégorie
et MAF

Répartition



Sélection des SNPs :

- ✓ Présent sur un scaffold > 10 000 pb
- ✓ Bi-allélique
- ✓ « Frame » minimum de 20 pb
- ✓ MAF > 0.2
- ✓ Illumina > 0.8
- ✓ Infinium II

Tri de la sélection :

- ✓ Catégorie croissante
- ✓ MAF décroissante

537 145 SNPs

Sélection des SNPs



Sélection + Tri

SNPs triés
par catégorie
et MAF

Répartition



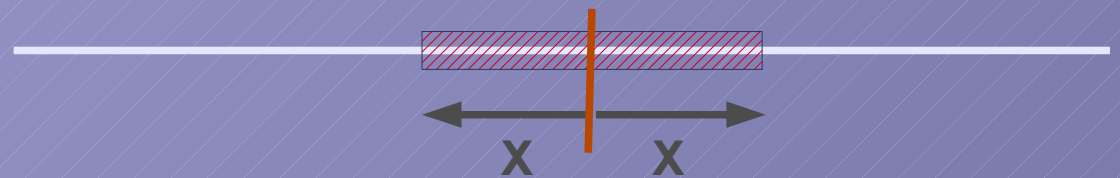
Sélection des SNPs :

- ✓ Présent sur un scaffold > 10 000 pb
- ✓ Bi-allélique
- ✓ « Frame » minimum de 20 pb
- ✓ MAF > 0.2
- ✓ Illumina > 0.8
- ✓ Infinium II

Tri de la sélection :

- ✓ Catégorie croissante
- ✓ MAF décroissante

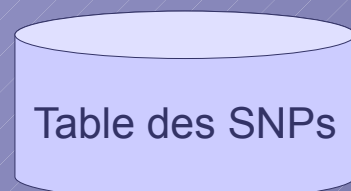
537 145 SNPs



Avec $X = 28\,500$ pb

$2\,700\,000\,000 / 60\,000 = 45\,000$ pb

Sélection des SNPs



Sélection + Tri

SNPs triés
par catégorie
et MAF

Répartition



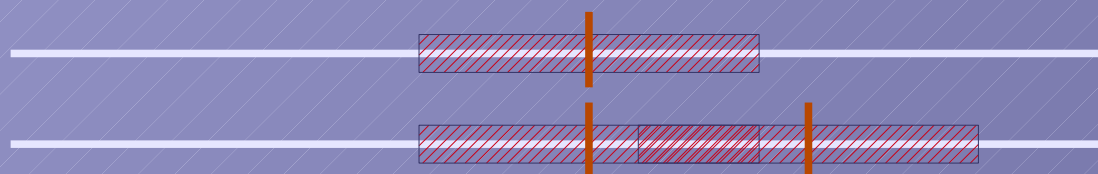
Sélection des SNPs :

- ✓ Présent sur un scaffold > 10 000 pb
- ✓ Bi-allélique
- ✓ « Frame » minimum de 20 pb
- ✓ MAF > 0.2
- ✓ Illumina > 0.8
- ✓ Infinium II

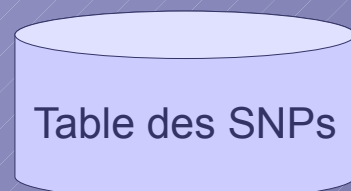
Tri de la sélection :

- ✓ Catégorie croissante
- ✓ MAF décroissante

537 145 SNPs



Sélection des SNPs



Sélection + Tri

SNPs triés
par catégorie
et MAF

Répartition



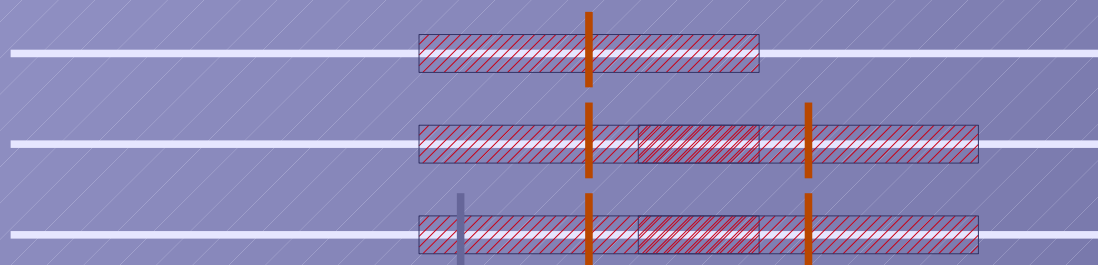
Sélection des SNPs :

- ✓ Présent sur un scaffold > 10 000 pb
- ✓ Bi-allélique
- ✓ « Frame » minimum de 20 pb
- ✓ MAF > 0.2
- ✓ Illumina > 0.8
- ✓ Infinium II

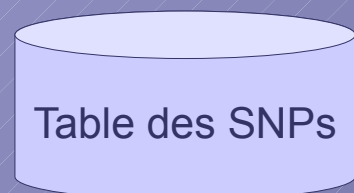
Tri de la sélection :

- ✓ Catégorie croissante
- ✓ MAF décroissante

537 145 SNPs



Sélection des SNPs



Sélection + Tri

SNPs triés
par catégorie
et MAF

Répartition



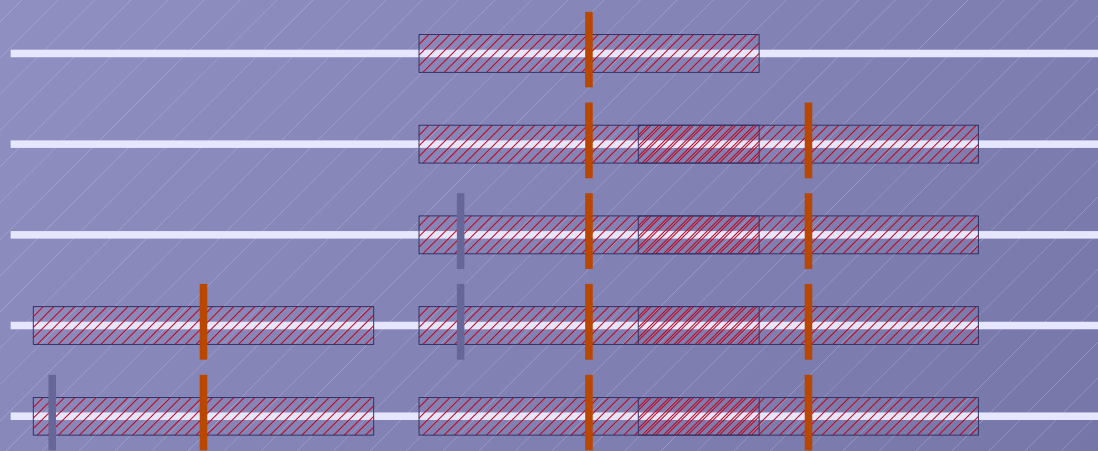
Sélection des SNPs :

- ✓ Présent sur un scaffold > 10 000 pb
- ✓ Bi-allélique
- ✓ « Frame » minimum de 20 pb
- ✓ MAF > 0.2
- ✓ Illumina > 0.8
- ✓ Infinium II

Tri de la sélection :

- ✓ Catégorie croissante
- ✓ MAF décroissante

537 145 SNPs

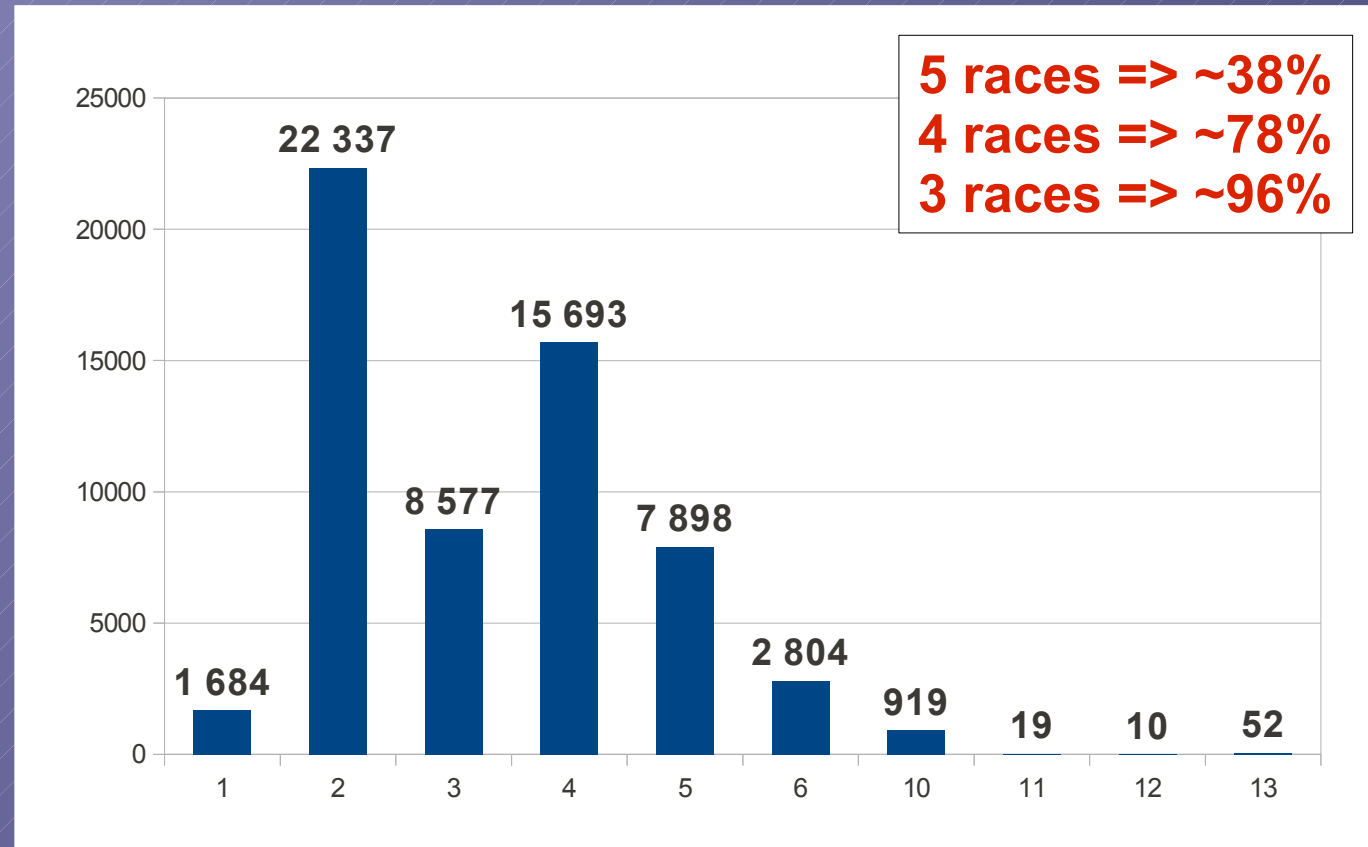


60 000 SNPs - Par catégorie

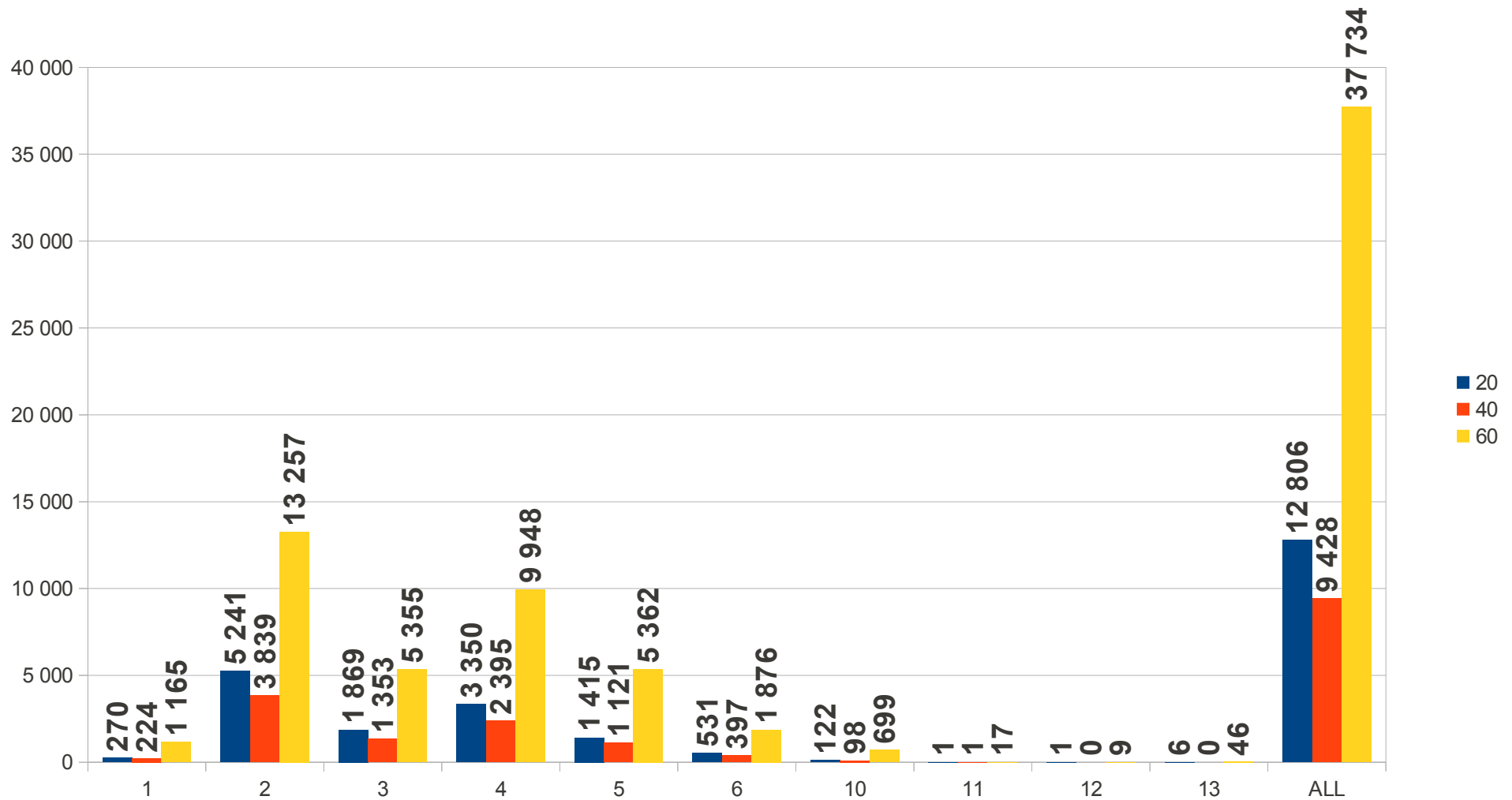
1 : EST
2 : Hétéroz. dans 5 races
3 et 4 : Hétéroz. dans 4 races
5 et 6 : Hétéroz. dans 3 races

10 : Hétéroz. S et A
11 : Hétéroz. (A ou S)
et (C ou B ou KS)
12 : Hétéroz. C et (B ou KS)
13 : Hétéroz. A ou S

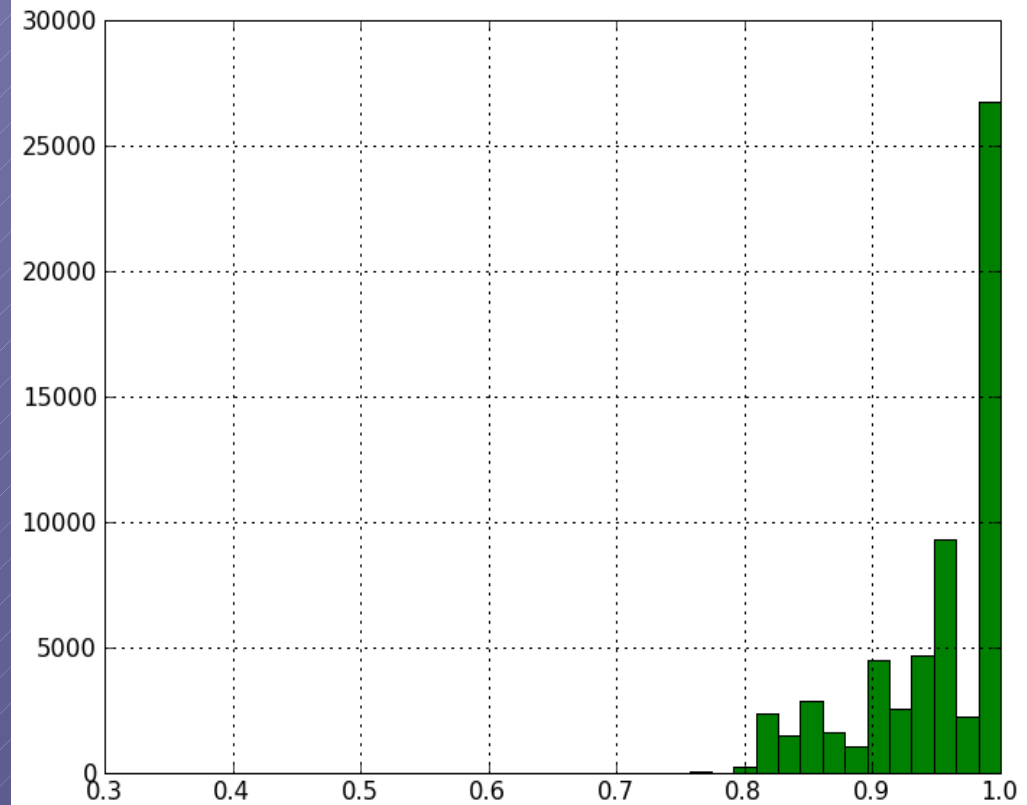
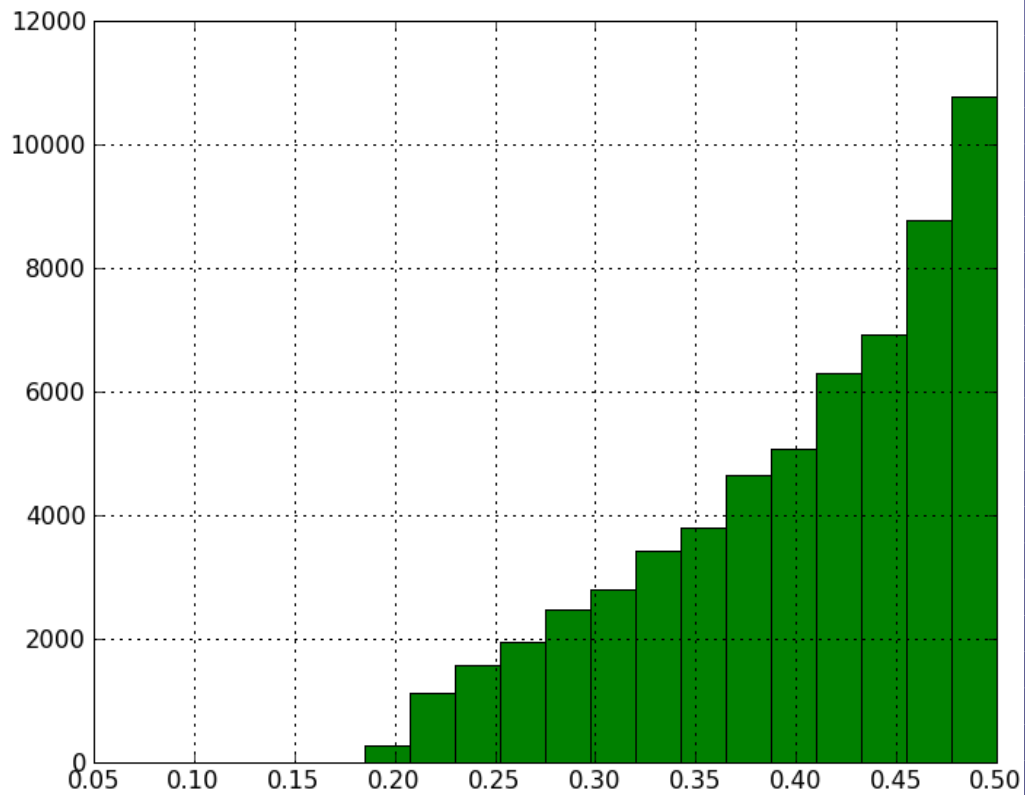
20 : autres
90 : INDEL



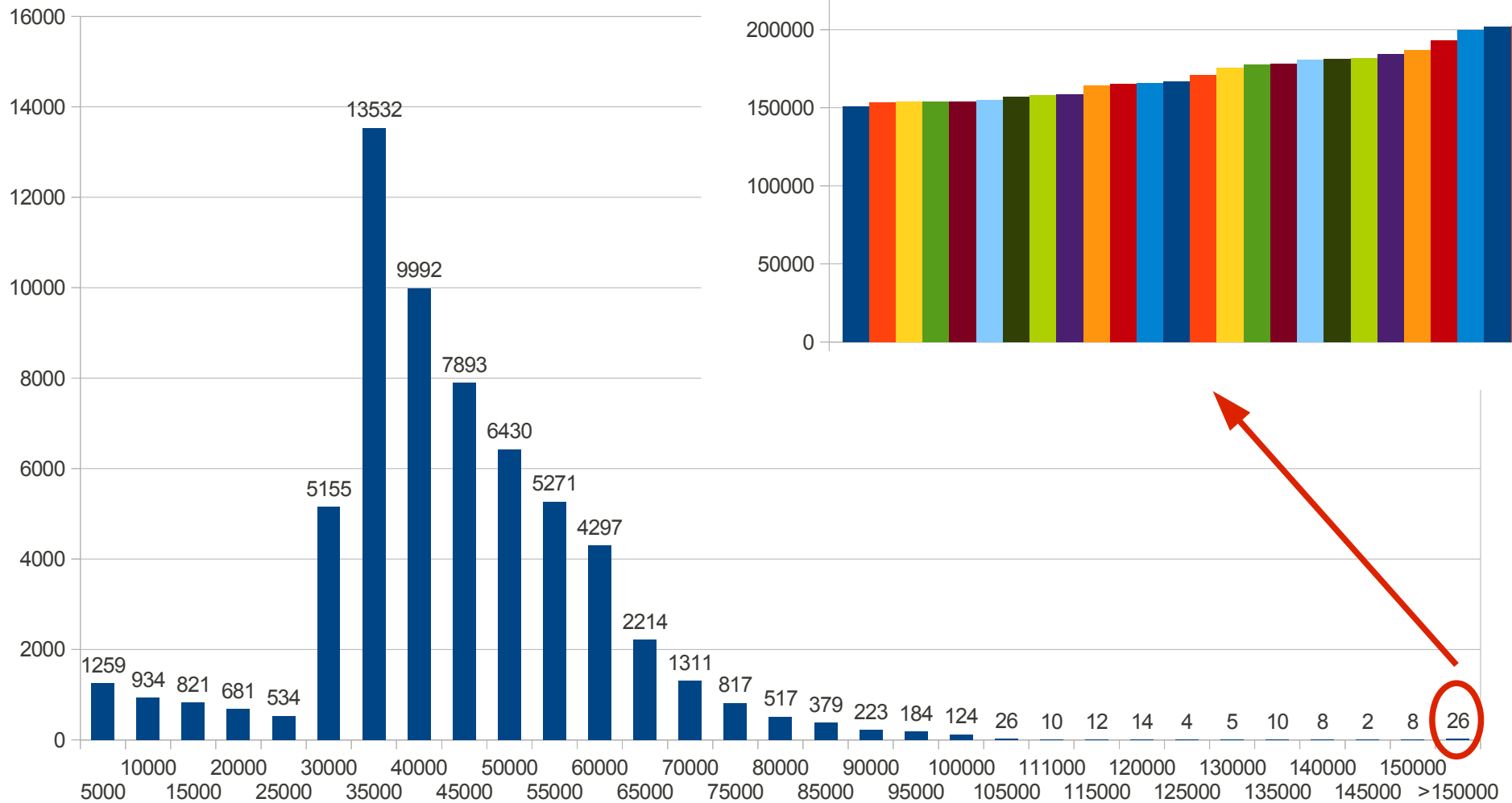
60 000 SNPs - Par catégorie et « frame »



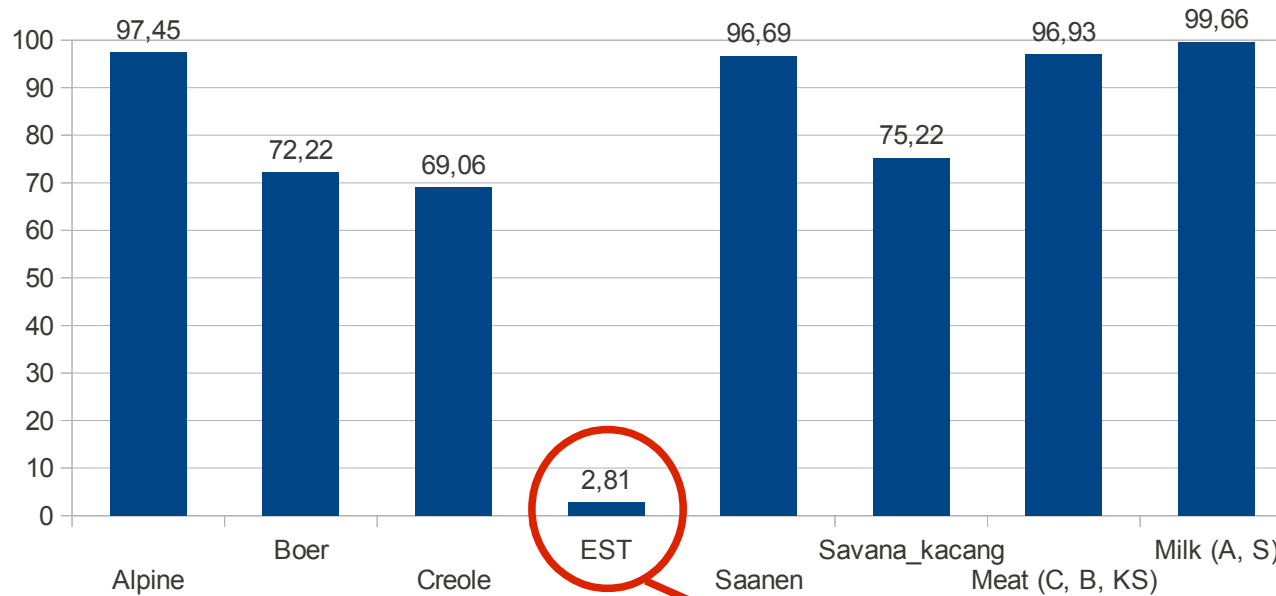
60 000 SNPs - MAF - Score illumina



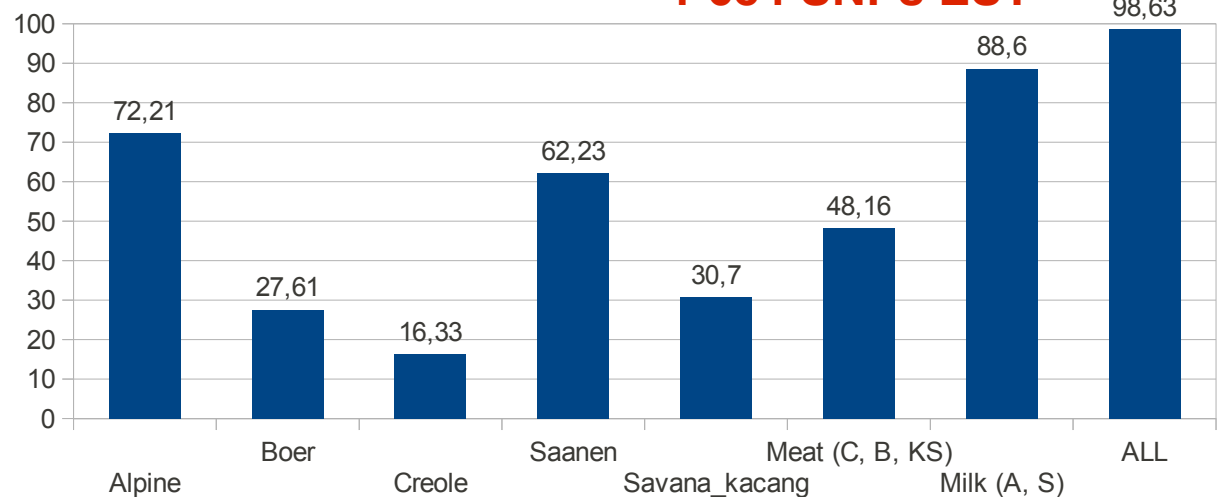
60 000 SNPs - Répartition



60 000 SNPs - Pourcentage par race

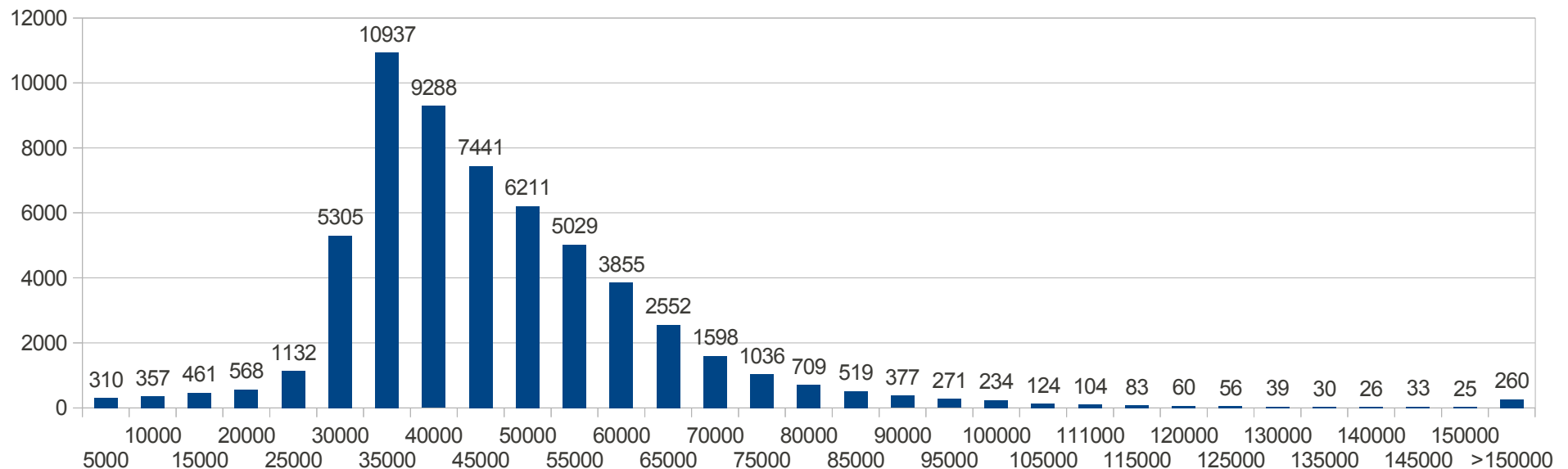


1 684 SNPs EST



60 000 SNPs - Répartition vs UMD3

- × 3 316 scaffold/chr. => Longueur totale = 2 670 405 961 pb
- × 30 chromosomes => 2 660 906 405 pb
- × Localisation bovine des 60 000 SNPs :
 - 59 084 SNPs localisés avec bwa bwasw (150 / SNP / 150)
 - **59 001 SNPs localisés sur un chr bovin**



Merci pour votre attention ...

260 regions > 150 000pb

