Goat genome assembly, Availability of an international 50K SNP chip and RH panel: an update of the International Goat Genome Consortium projects

Gwenola Tosser-Klopp on behalf of IGGC

Cattle and sheep workshop,
PAG meeting, 2012, January, the 15th

Outline

- IGGC presentation / history
- Goat Genome Assembly
- RH panel (discussed on monday)
- International goat SNP chip
- Next projects: Brian Sayre

IGGC

- International Goat Genome Consortium
- Created in March, 2010
- www.goatgenome.org
- Coordination: Wenguang Zhang & Gwenola Tosser-Klopp
- 3 ongoing projects
- Open meeting on Monday afternoon (3pm:5pm, Towne room) to discuss further projects

Capra hircus genome assembly

Yunnan black goat

Cashmere goat

Yunnan black goat (XX) lead to high quality reference genome of the domestic goat generated by combining Illumina new-generation short reads sequencing and the optical mapping technology of large DNA molecules which was used to generate the super-scaffolds.
Cashmere goat transcriptomes of primary and secondary fiber-growing follicles were generated

284,683 primary super-scaffolds

	Contig		Scaffold		Primary Super-scaffold	
	bp	Number	bp	Number	bp	Number
N90	4,410	141,869	440,999	1,348	582,523	976
N80	7,994	100,335	846,998	922	1,175,001	664
N70	11,323	73,948	1,253,003	664	1,739,998	481
N60	14,862	54,526	1,694,371	482	2,447,724	352
N50	18,720	39,408	2,212,139	344	3,057,189	254
Total	2,522,851,955	542,145	2,662,658,003	285,383	2,662,728,047	284,683

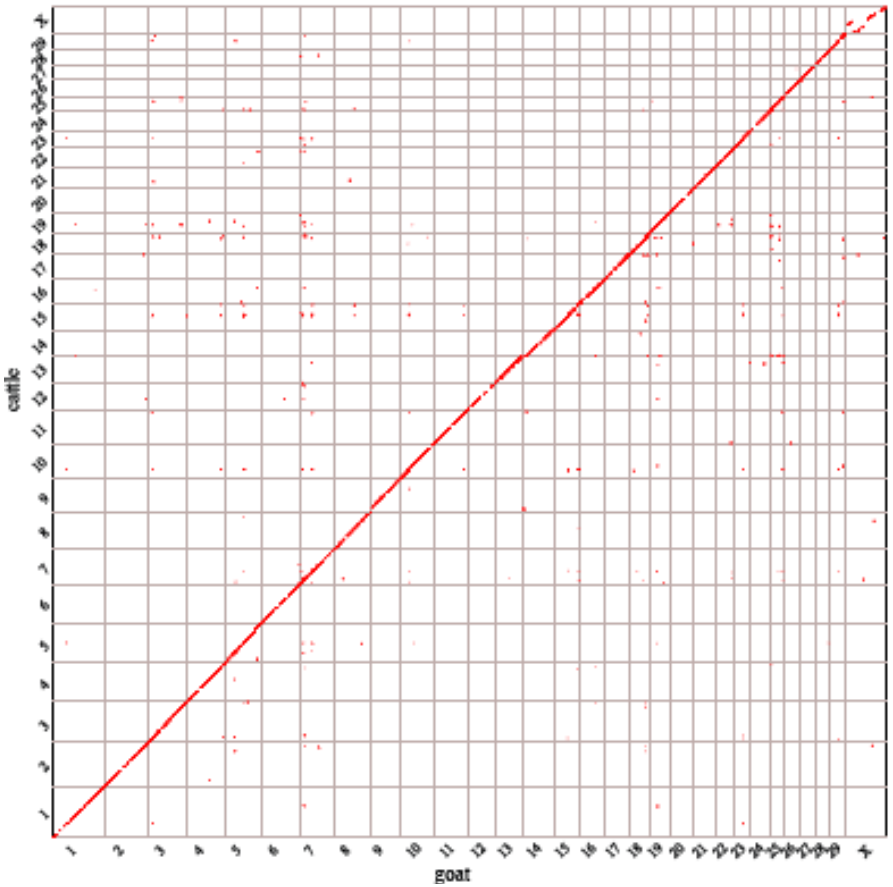
*Total number: the number of contig/scaffold sequences with length > 100bp

The assembled base pairs (SOAPdenovo software**) total 2.66 Gb, which is about 92% of the estimated goat genome size (~ 2.9 Gb)**

349 super-scaffolds, using optical mapping

Super scaffold	Size (bp)	Number
N10	37,201,881	6
N20	31,8,804	13
N20	26,309,660	22
N40	22,009,453	32
N50	**18,182,911**	**45**
N60	14,726,152	60
N70	10,332,770	81
N80	6,331,148	112
N90	2,857,184	171
Total Number		**349**
Total Size 2,525,731,503		

30 pseudo-chromosomes



Based on the high colinearity between bovine and goats (Cribiu, E. P. *et al. Cytogenet Cell Genet* 2001), we used bovine genome to assemble the **315 super-scaffolds together with 422 extra scaffolds which were not included in super-scaffolds into 30 pseudo-chromosomes for the goat.**

SNP chip project

Cattle and sheep workshop,
PAG meeting, 2012, January, the 15th

Data

– INRA, France :

- 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 :

- Whole genome sequencing of 64 Boer, Savanna and Kacang meat and indigenous goats

–University of Utrecht, Netherlands

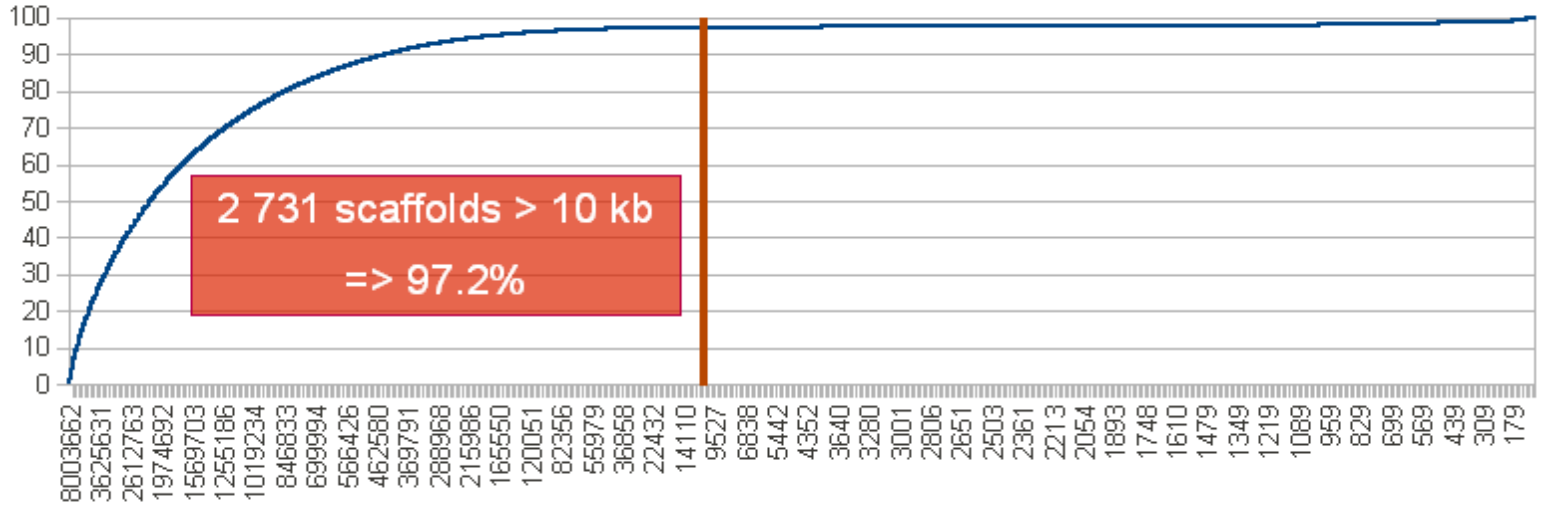
- RRL of 17 Saanen dairy goats. 120 millions of 32 bp paired-ends sequences

–Italy, Spain, USA : ESTs + genes

Goat scaffolds were used to map the « SNP » sequences

BGI scaffolds :

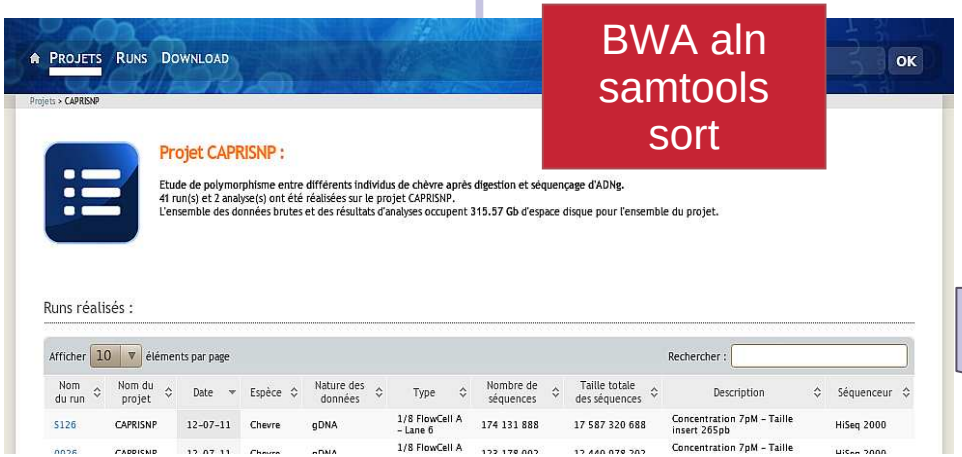
- 285 375 scaffolds/contigs => total length = 2, 662 Gpb
- length from 100 pb to 19 Mpb



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

Pipeline – INRA + NL data

12 million SNPs + INDELs detected



BWA aln
samtools sort

samtools merge

Alpine.bam Saanen.bam Creole.bam

Filter
- mapping quality> 30
- unique alignment

samtools mpileup
bcftools view

Alpine.bcf Saanen.bcf Créole.bcf

Filter
- vcfutils.pl varFilter
- SNPs homozygous A + S + C

Alpine.bcf Saanen.bcf Créole.bcf

varFilter :
- min depth= 6
- max depth=100
- Min 2 reads for 1 allele

Malaysian data & ESTs

Malaysian data

BWA alignment (bwasw)

SNP position extraction

File to implement the database

3.5 million SNPs

EST

BWA alignment (bwasw)

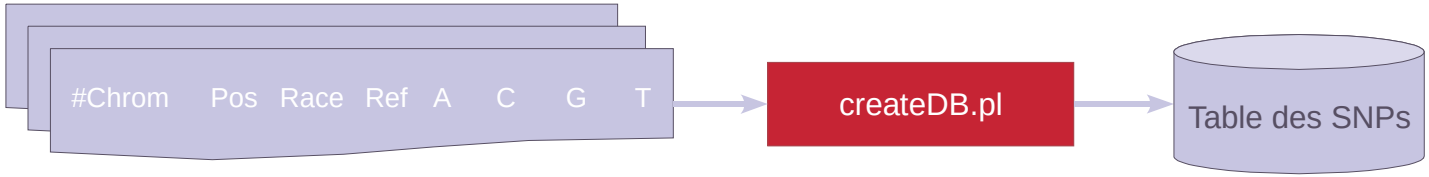
SIM4 alignment validation

SNP position extraction

File to implement the database

7000 SNPs

Data base filling



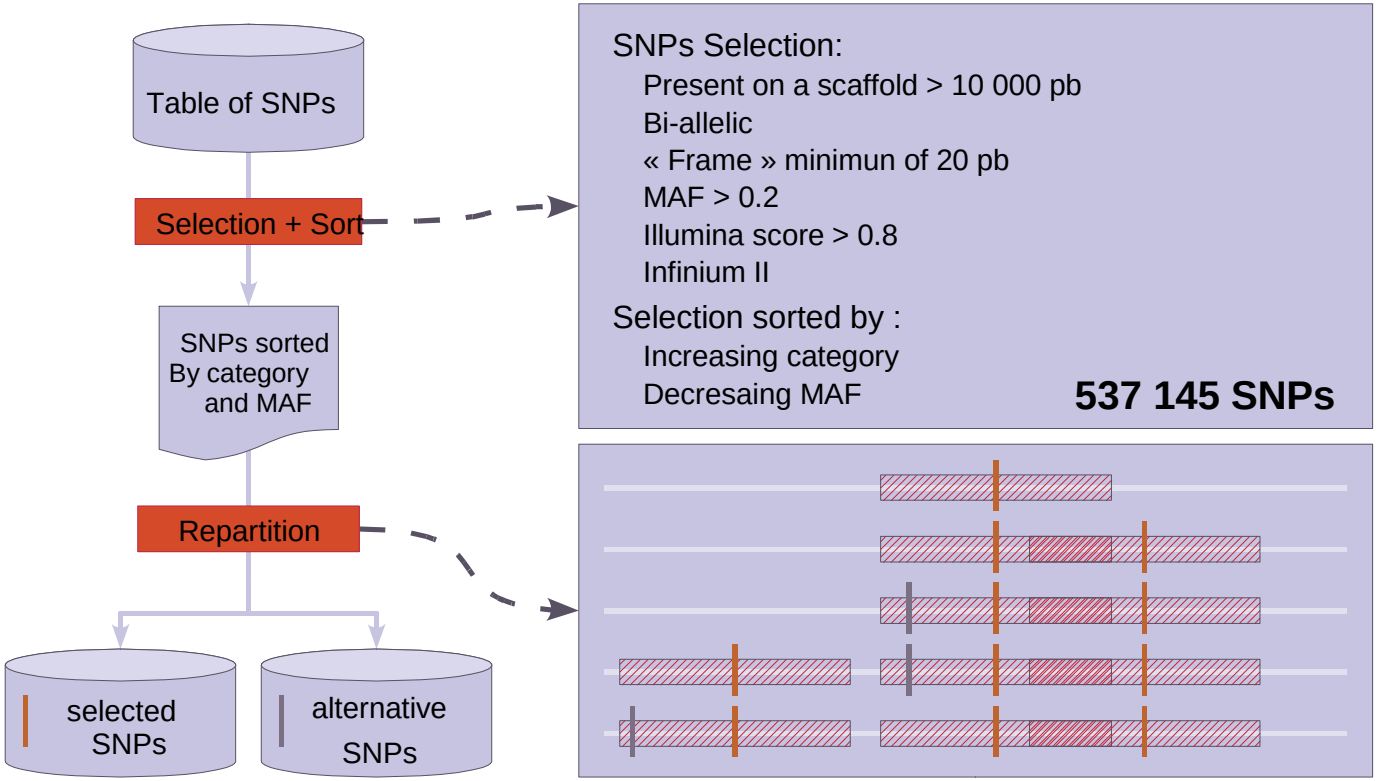
chrom	pos	ref_base	alpine	boer	creole	est	saanen	savana_kacang	illumina	category	infinium	maf	alleles	allele_type	dist_prev	dist_next	frame
scaffold1	54	T	A				AT			13	1	0.500	AT	2	54	76	40
scaffold1	130	C	CT		C					20	2	0.167	TC	2	76	1	-1
scaffold1	131	A	AG	AG	G			AG		11	2	0.500	AG	2	1	21	-1
scaffold1	152	A						AG		20	2	0.125	AG	2	21	1	-1
scaffold1	153	T	GT	GT				GT		11	2	0.500	AG	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						20	2	0.125	AG	2	71	117	60
scaffold1	439	A		AG						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			20	2	0.167	AG	2	348	40	40
scaffold1	894	T	CT	CT						13	2	0.167	TC	2	40	11	-1
scaffold1	905	A	AC							13	2	0.167	AC	2	11	189	-1
scaffold1	1094	T	CT		C		CT			20	2	0.333	TC	2	189	17	-1
scaffold1	1111	G	AG				AG			10	2	0.333	AG	2	17	144	-1
scaffold1	1255	G	A				AG			13	2	0.500	AG	2	144	113	60
scaffold1	1368	I								90	-1	0.000		0	113	200	60
scaffold1	1568	C	CT				CT			10	2	0.333	TC	2	200	20	20
scaffold1	1588	A	AG				AG			10	2	0.333	AG	2	20	9	-1

11 924 638 SNPs :
- 1 229 120 Indels
- 10 695 518 SNPs

Infinium
I : A/T et C/G => 2 probes
II : other => 1 probe

Category
1 : EST
2 : Heteroz. in 5 breeds
3 & 4 : Heteroz. in 4 breeds
5 & 6 : Heteroz. in 3 breeds…

SNP selection



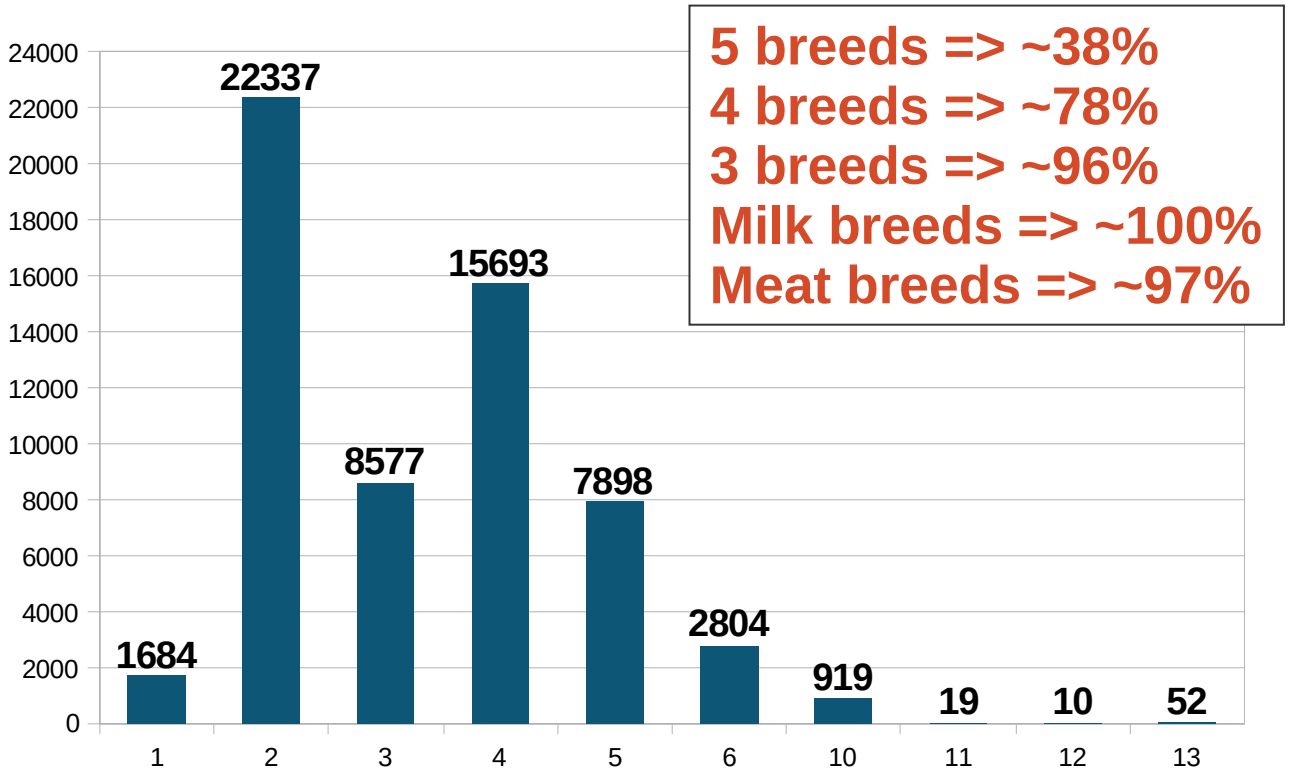
28,500 pb minimum between 2 SNPs

60 000 selected SNPs

1 : EST
2 : Heteroz. in 5 breeds
3 et 4 : Heteroz. in 4 breed
5 et 6 : Heteroz. in 3 breed

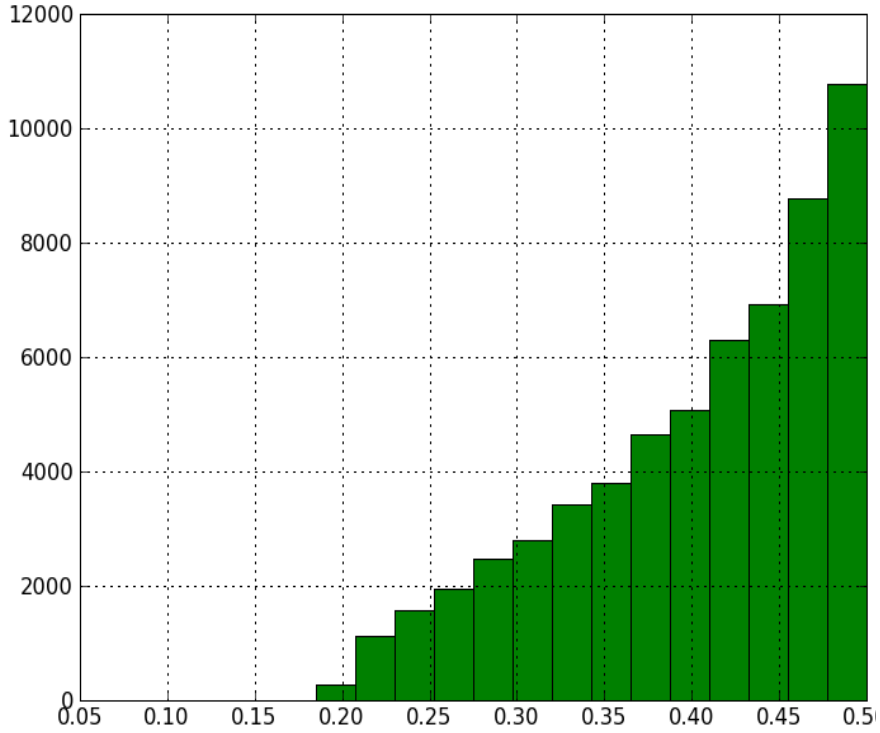
10 : Heteroz. S & A
11 : Heteroz. (A or S
&(C or B or KS)
12 : Heteroz. C et (B or KS
13 : Heteroz. A or S

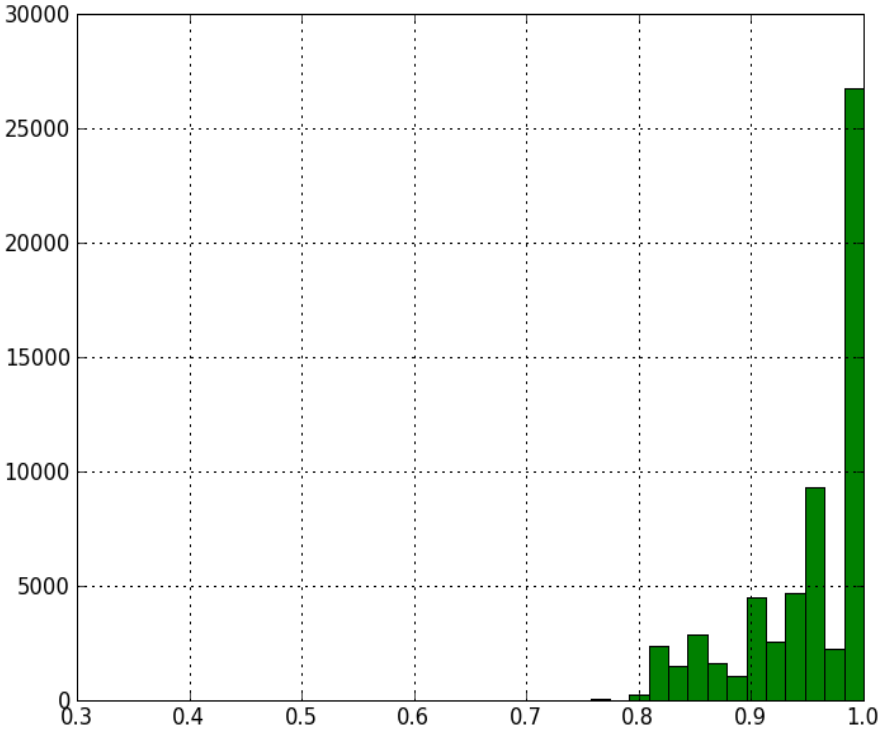
20 : other
90 : INDEL



5 breeds => ~38%
4 breeds => ~78%
3 breeds => ~96%
Milk breeds => ~100%
Meat breeds => ~97%

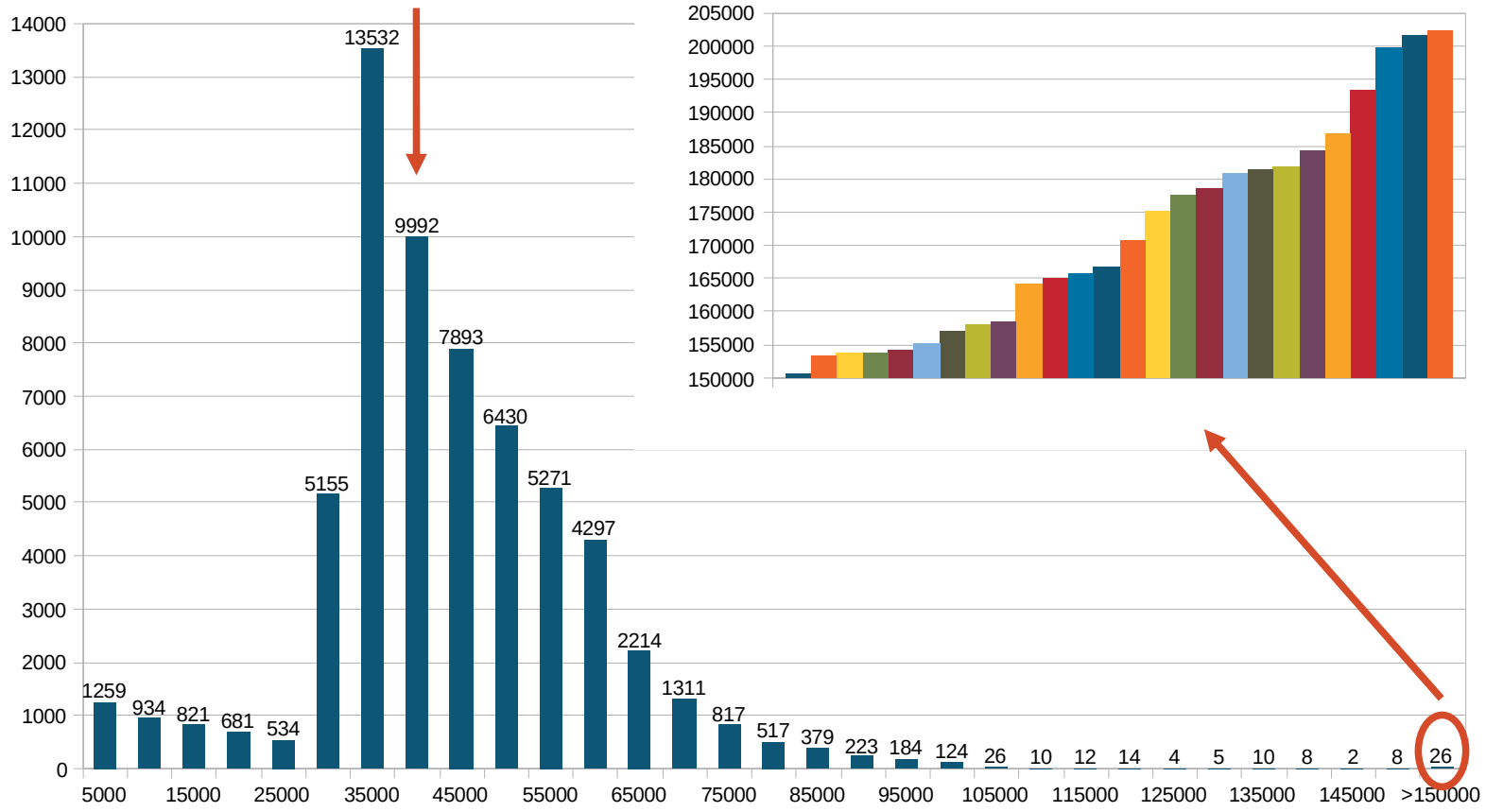
60 000 SNPs - MAF - Illumina Score





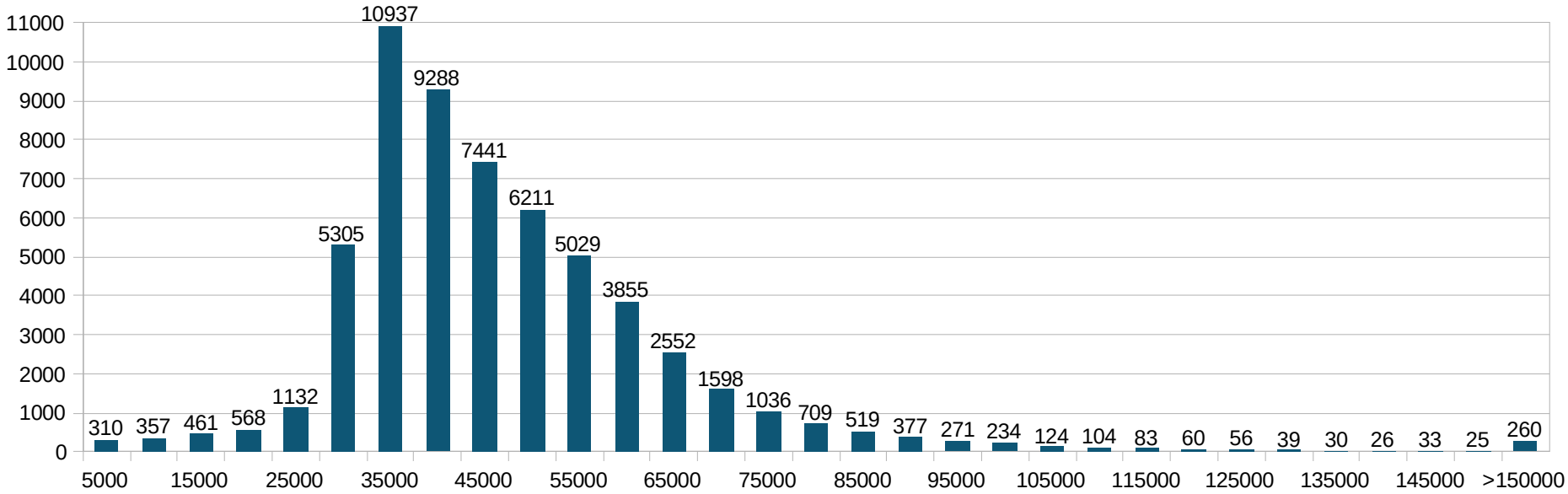
60 000 SNPs - Spacing

median interval => ~ 40kb



60 000 SNPs – Spacing on cattle genome (UMD3)

59 001 SNPs localised on a bovine chromosome



Chip manufacturing and cluster files

- Illumina iSelect design
- 288 animals were used for cluster file generation and quality control
- Includes the animals used for SNP discovery
- Breeds : Alpine, Saanen, Creole, Katjang, Savanna, Boer, Skopelos, Angora, Jinlan

SNP chip characteristics

- 53,348 synthesized loci
- 52,295 successful loci
- 8,000 ordered samples in September 2011
- Cluster files (.egt) available: Gwenola.Tosser@toulouse.inra.fr
- Annotation and publication of the loci coming soon

A chip useful for many breeds

Breed	Samples	SNPs MAF>0.05
Alpine	53	51339
Angora	26	47195
Boer	30	48494
Creole	38	50216
Jinlan	13	45648
Katjang	13	33873
Saanen	57	51689
Savanna	20	46629
Skopelos	27	50908
Yunling	1	17335

Upcoming projects of IGGC

- Hapmap project
- Resequencing
- Integration of RH and genome data
- …
- Open meeting on monday

Acknowledgements

- **SNP discovery:**
  - Henri Heuven
  - Saadiah Jamli, Tun-Ping Yu
  - Carole Moreno, Philippe Mulsant, Isabelle Palhière, Rachel Rupp, Gwenola Tosser-Klopp
  - Marcel Amills, Patrice Martin, Eric Pailhoux, Brian Sayre, Alessio Valentini, (ESTs)
  - Julien Sarry, Aurélie Tircazes
  - UNCEIA, Capgenes and Apis-gene (French breeding organizations)

- **Genome sequence:**
  - Jun Wang, Wen Wang, Wenguang Zhang

- **Bioinformatics:**
  - Philippe Bardou, Cédric Cabau, Thomas Faraut, Christophe Klopp,
  - Ibouniyamine Nabihoudine
  - Curt Van Tassell for testing his spacing software on the data

- **Illumina:**
  - André Eggen, Cindy Lawley, Karine Viaud

- **Advice and support:**
  - John McEwan

African Goat Production Value Chain Development Project

USDA-ARS and ILRI Sponsored Workshop

Nairobi, Kenya, November 2011

Cattle and sheep workshop,
PAG meeting, 2012, January, the 15th

African Goat Production Workshop

- A workshop was held in Nairobi, Kenya in November 2011 sponsored by the USDA-ARS and ILRI.
- The aims of this workshop were:

1. Bring together research experts for improvement of goat production in Africa

2. Determine the potential of applying genome-based tools to value chain development projects in goat production

3. Determine the current needs for characterization of goat populations and utilization of genome-based tools

General Project Development Concept

- Use emerging technologies to characterize and improve the adapted germplasm
- Development of a refined, high quality genome sequence and genome-based tools, if needed
- Development of genetic signatures for goat populations
- Determine the needs of the local producers for development of the goat production value chain
- Based on producer needs, develop improved germplasms using a genetic signature based approach

Outcomes

- Appears feasible to use high level genome-based tools for improved selection and sustainability of adapted germplasm
- Advantageous to have multiple independent genome sequence assemblies for the goat to improve the error checking and quality of the reference genome sequence
- Samples from African goat populations will be characterized with the current SNP panel to get an initial characterization of the populations
- Meet in October 2012, with support groups to develop local value chain assessments and determine methods for genetic signature utilization.

We thank you for your attention