Sequencing and assembly of the sheep genome reference sequence

Yu Jiang

Kunming Institute of Zoology, CAS, China

the International Sheep Genomics Consortium (ISGC)

ISGC Presentations

Yu Jiang, Kunming Institute of Zoology
(jiangyu96@163.com)
"Sequencing and assembly of the sheep genome reference sequence"
http://sheep.genomics.org.cn/

Brian Dalrymple, CSIRO Livestock Industries
(brian.dalrymple@csiro.au)
"from scaffolds to chromosomes"
http://www.livestockgenomics.csiro.au/perl/gbrowse.cgi/vsheep2/

Outline

- Sampling and sequencing
- De novo assembling and QC
- Gene annotation and QC
- SNP calling and other data

***Ovis aries* : (26+XY)**
genome size ~3Gb

The data generation phase of sheep genome project commenced at two sequencing facilities in late 2009.

Texel is a popular terminal sire breed in several countries, and served as the paternal grandsire breed of the sheep international mapping population.

The Texel female was 6 months old (provided by Jacob B. Hansen, University of Copenhagen)---------- KIZ-BGI
DNA: Liver RNA: 7 tissues
The Texel ram was used previously as the DNA source for CHORI-243 BAC library (Dalrymple *et al.,* 2007).----Roslin

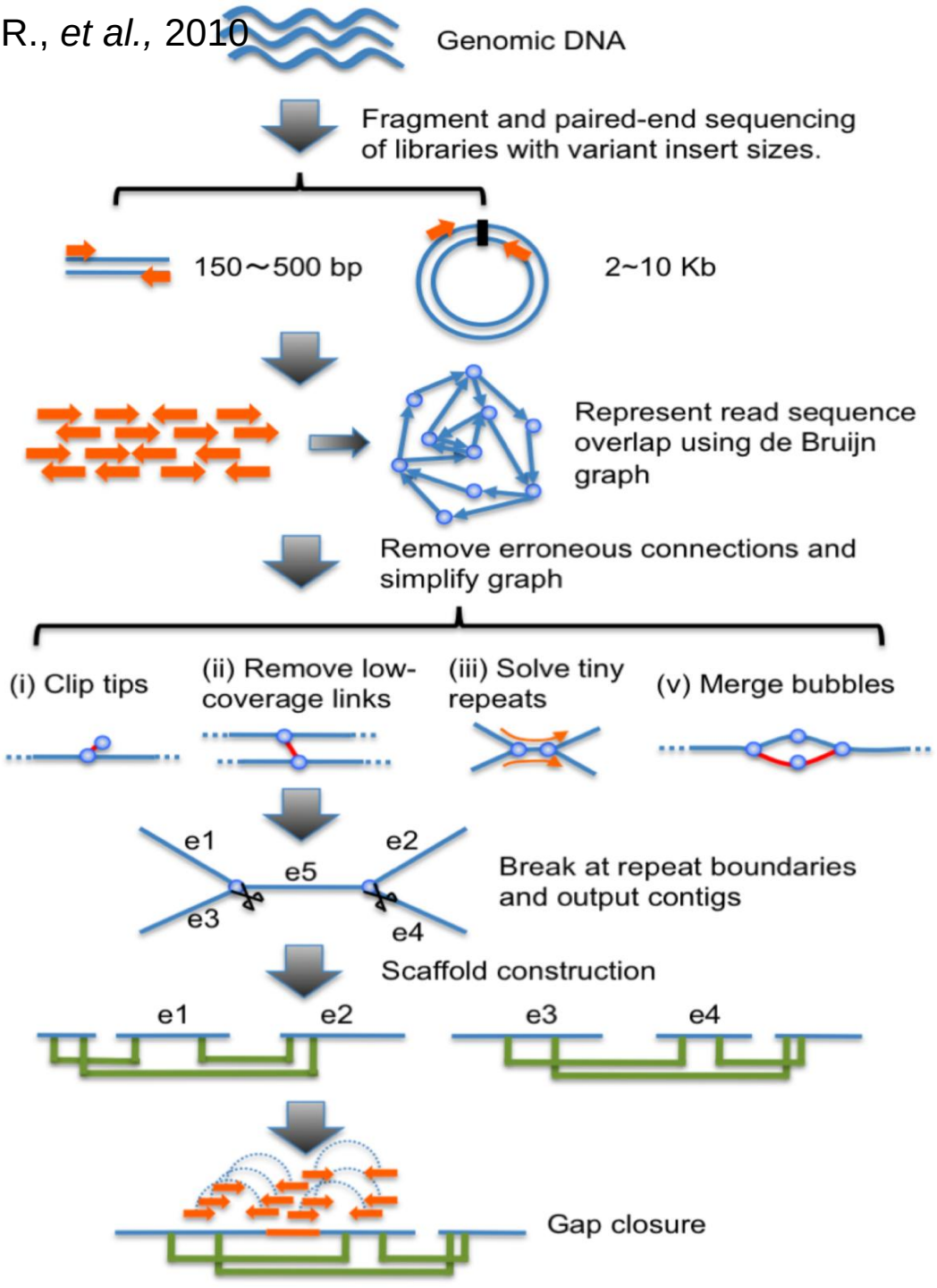
Sequences included in the genome reference sequence

Sample	Purpose	Sequence method	Paired-end libraries	Insert size (bp)	Lib-raries	GA Lanes	Total length (Gb)	Reads Length (bp)	Coverage (X)
Female	assembly	Illumina	180bp	150-210	**1**	4	23.8	101	7.93
Female	assembly	Illumina	350bp	280-420	**4**	21	105.0	101	35.00
Female	assembly	Illumina	800bp	650-950	**2**	6	32.0	101	10.67
Female	assembly	Illumina	2kbp	1.6-2.4k	**2**	11	35.7	45	11.90
Female	assembly	Illumina	5kb	4.5-5.5k	**2**	6	18.5	45	6.17
Female	assembly	Illumina	10kb	8.5-10.5k	**1**	3	8.3	45	2.77
Female	assembly	Illumina	20kb	15-22k	**1**	1	1.8	45	0.60
Male	fill gap	Illumina	200bp	120-280	**1**	16	77	101	24.0
Male	fill gap	Illumina	500bp	400-700	**1**	24	72	101	25.5
Male	for check	454	8kb				3.3		1.10
Male	for check	454	20kb				1.5		0.50
Male	for check	Sanger	184kb				0.3	687	0.09

The 75X Texel ewe reads was used to assembly into contigs and scaffolds. Then, all Illumina reads both from the ram and ewe were applied for gap filling to improve the assembly.

The 454(BCM-HGSC) and Sanger sequences were also mapped to the revised genome reference sequence.

Li, R., *et al.*, 2010



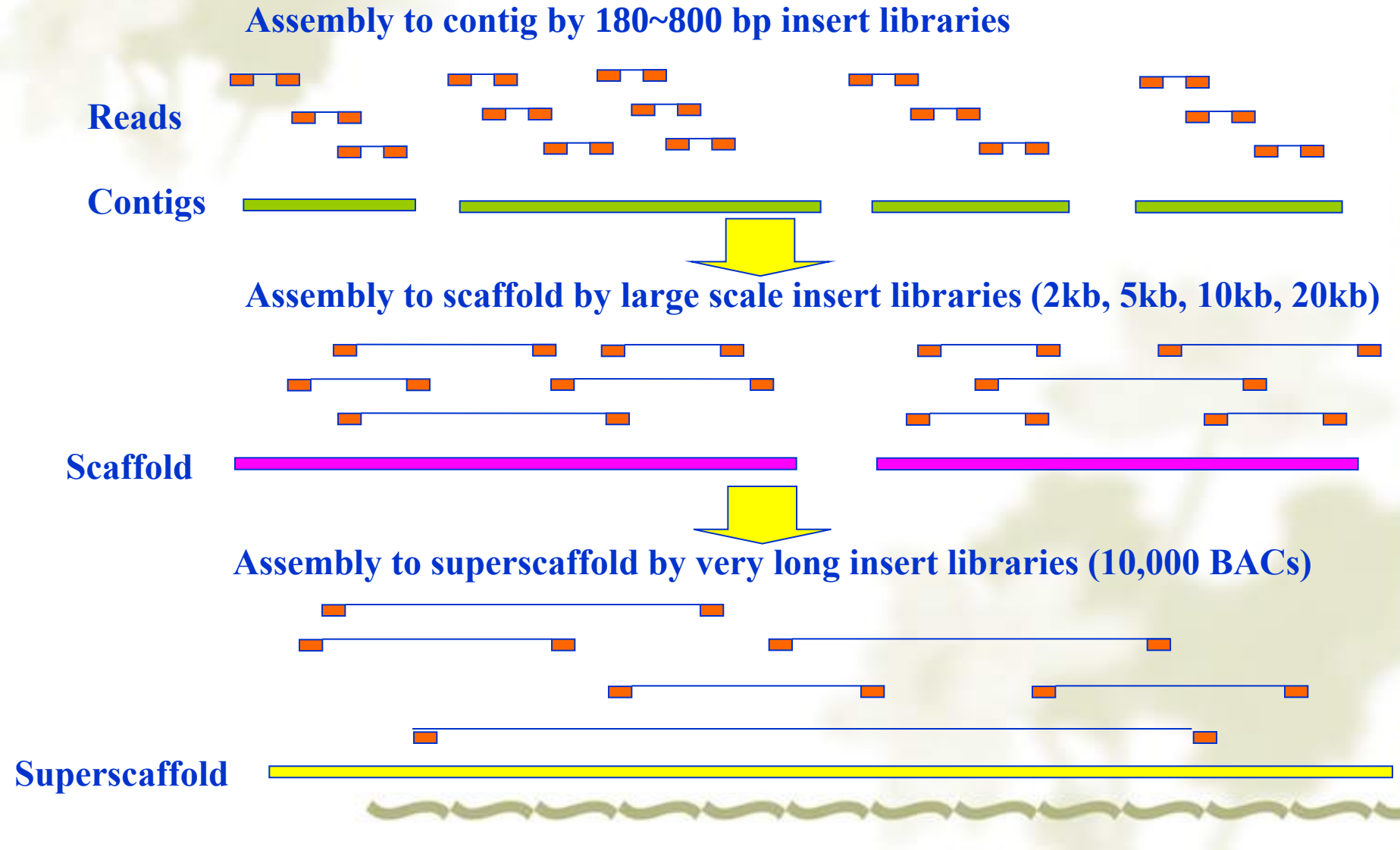
Step0. high-quality reads.

Step1. contig

Step2. scaffold

Step3. gap filling

De novo assembling strategy



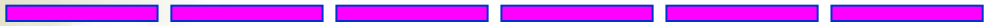
De novo assembly result (scaffold)

	Contig Size(bp)	Contig Number	Scaffold Size(bp)	Scaffold Number
N90	4,525	142,185	202,439	2823
N80	7,718	101,384	406,941	1900
N70	10,872	74,723	618,490	1358
N60	14,205	55,011	841,456	983
N50	**17,951**	39,663	**1,079,158**	696
Total Size	2.45Gb		**2.71Gb**	
Total Number(>100bp)		3,811,758		490,776
Total Number(>2Kb)		241,657		**8,115**

N90 size: the length such that 90% of the assembled genome lies in blocks of N50 size or longer; 6.9% is “N” (gap).

Fill in intra-scaffolds and get the final contigs

Scaffold



The gaps in scaffold will been filled by all the available reads other rest gaps are still marked by multiple"N"

Final Scaffold

The scaffold will been divided into the final contigs

Final contigs

Checking

RH makers or BAC ends, all reads are mapped back to scaffolds

106 scaffolds need to be splited into 210 segments

1000 small tandem duplications(total size is 2 Mb) have be masked.

Mapping scaffolds to chromosomes

Combined Anchor Strategic in sheep physical mapping

- 10,000 BAC pair ends (CHORI-243)

- Using the 59,000 SNP markers
- (i)Thomas Faraut (INRA, France) provided the sheep radiation hybrids (RH) maps (about 39,000).
- (ii) Jill Maddox (Melbourne university) about one thousand linkage markers

- Synteny relationships with close related species
- (iii)Cattle: 17,482 markers→ place 90% of the assembly (Btau_4.0)
- (iv)Goat (N50 17Mb) and Tibetan Antelope (N50 2Mb) (on going for OarX)

Super scaffold on chromosomes

Super scaffold	Size (bp)	Number
N50	37,056,980	23
N60	28,759,665	30
N70	21,637,475	41
N80	14,866,675	55
N90	6,068,238	83
Total Number		349
Total Size	2,569,509,652 (95%)	

Integrality Quality control

- 91% (2.71Gb/2.97Gb)

Genome size=kmer_num/kmer_dpeth

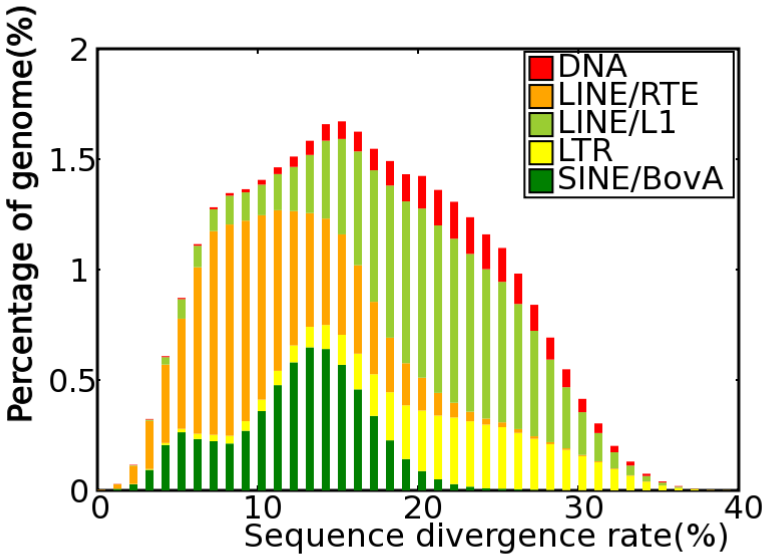
(Distribution of 17-mer frequency in the raw sequencing reads)

- Single copy regions

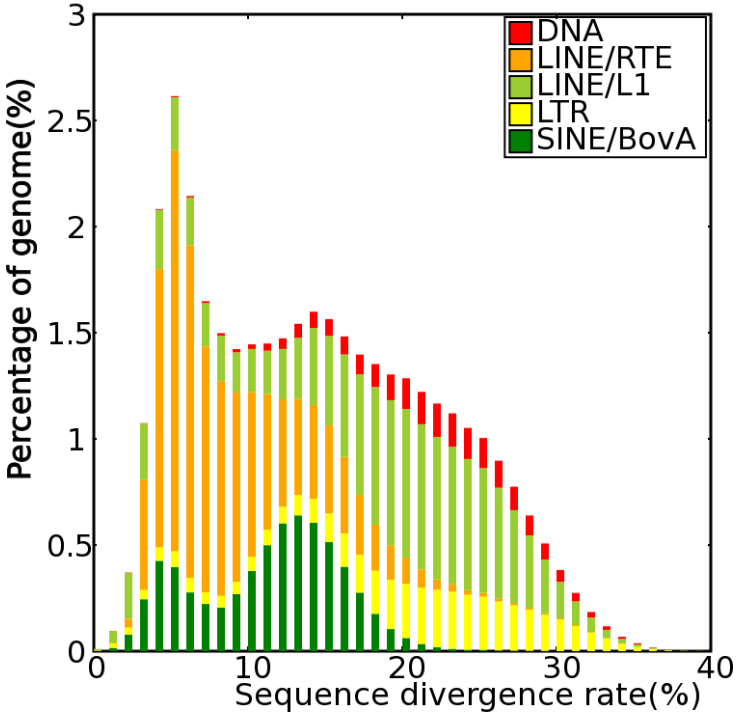
94% (51275 of the 54590 SNPs in Ovine SNP50K BeadChip have good hit)

- Segmental duplication or repeat regions are underestimated (maybe 10~20% is missing)

Repeat divergence distribution



Sheep



Cattle(Btau4)

Assembly quality control

- using RH makers or BAC ends

  when assembly, from 3,811,758 contigs (>100bp) to **8,115** scaffolds(>2kb)

  106 scaffolds need to be splited into 210 segments (adjusted already)

- using complete BAC sequences
  - NCBI (15 complete BAC sequences from CH243)
  - The sheep Major Histocompatibility Complex region

Quality control using BAC sequences

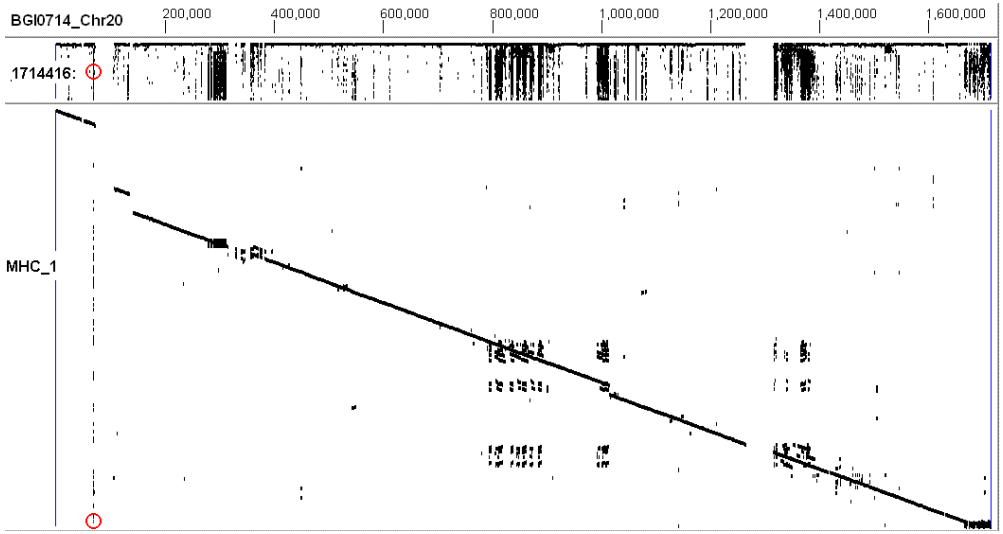
Ref-Oar4	ref-start	ref-end	BAC_name	BAC length	match-coverage	all-coverage	Identity
scaffold1489	34347	188387	gi_145651874	153498	0.902	1.004	0.996
scaffold1489	185786	385157	gi_66275383	198973	0.966	1.002	0.997
scaffold1489	380635	409967	gi_134032077	30087	0.977	0.975	0.996
scaffold1489	408126	551322	gi_119360798	143308	0.953	0.999	0.996
scaffold1489	525666	606538	gi_146134567	80882	0.981	1.000	0.999
scaffold1489	569237	744461	gi_139093431	176451	0.958	0.993	0.995
scaffold1489	738250	889820	gi_62954874	151538	0.978	1.000	0.997
scaffold1489	885369	983017	gi_118142661	95794	0.897	1.019	0.995
scaffold1489	978215	1108973	gi_116517344	131638	0.951	0.993	0.997
scaffold1489	1106968	1296975	gi_119433720	190282	0.960	0.999	0.996
scaffold1489	1282306	1421623	gi_119360797	139569	0.990	0.998	0.998
scaffold1489	1408536	1570307	gi_134032078	162012	0.969	0.999	0.996
scaffold1489	1566870	1753428	gi_62420187	187152	0.985	0.997	0.999
scaffold1489	1742232	1905943	gi_119372332	156221	0.820	1.048	0.994
scaffold1489	1890834	2078175	gi_117557604	186963	0.936	1.002	0.993
sum				**2184368**	**0.943**	**1.003**	**0.996**

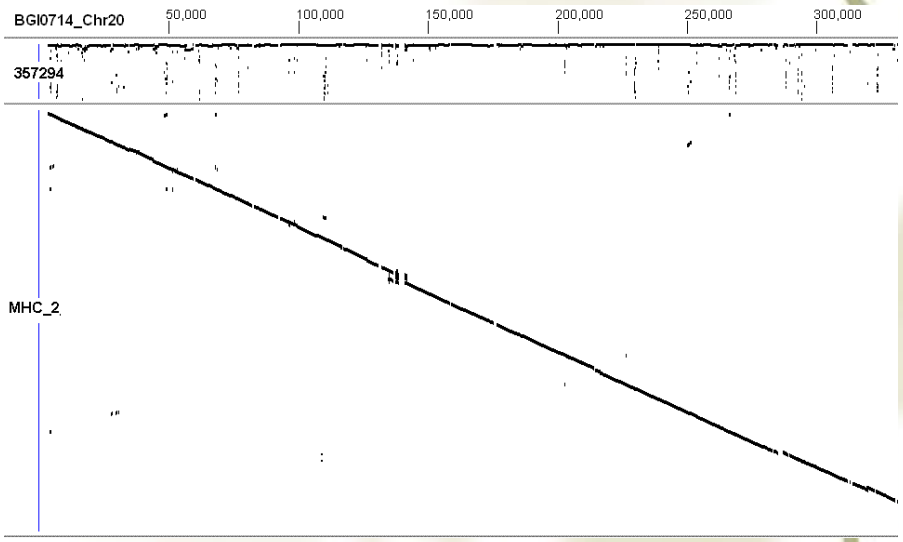
15 complete BAC sequences from CH243, the difference between the male and female Texel is about 0.002 (5Mb/2.7Gb).

The sheep Major Histocompatibility Complex region (assembled by 26 overlapping BACs) (BMC Genomics 2010, 11:466)

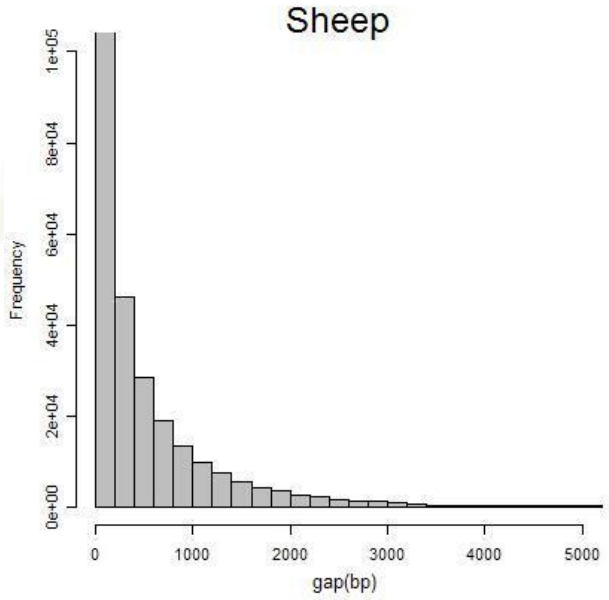
Comparing the MHC contigs with the two loci of ref Oar20, there are no significant difference except several gaps on Oar20 with a total length of 363kb (30%).

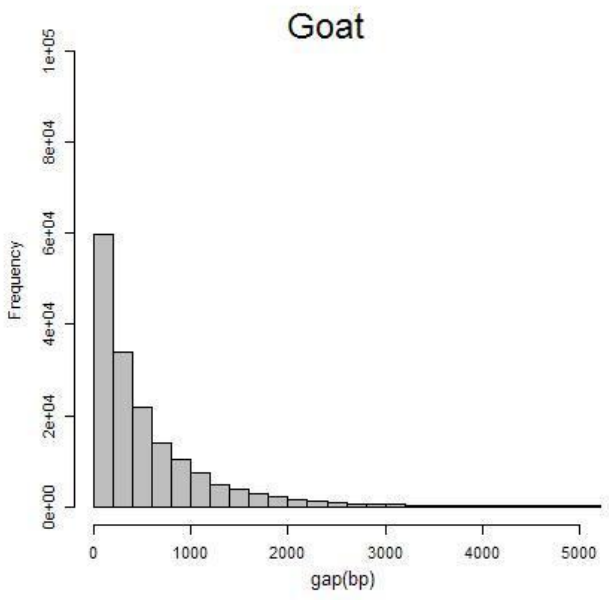
	Length	Mapped
MHC_contig1	363 kb	357kb
MHC_contig2	2071 kb	1714kb

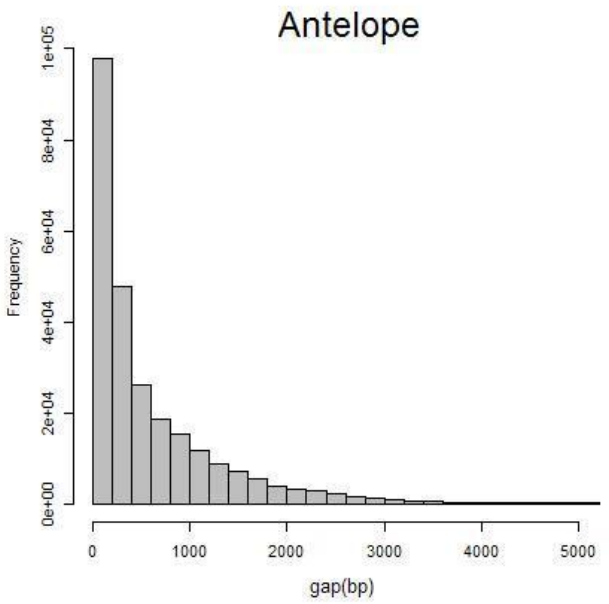


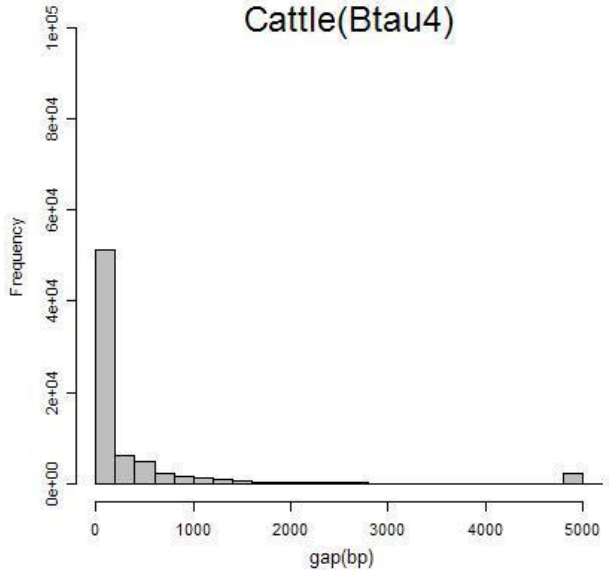


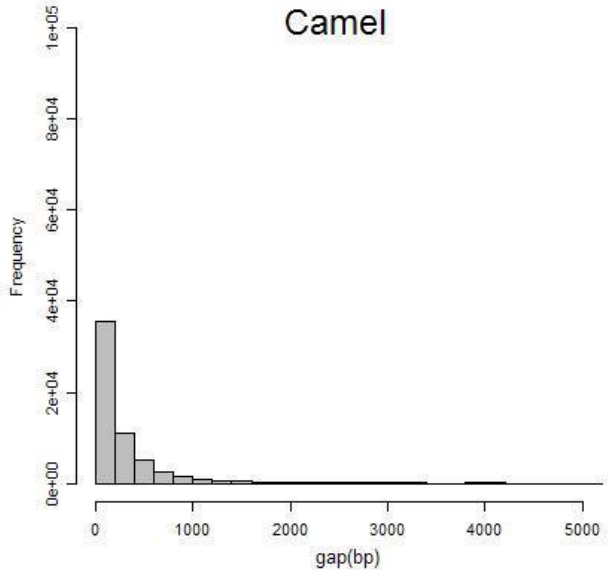
Gap length distribution (a new round gap filling for sheep)

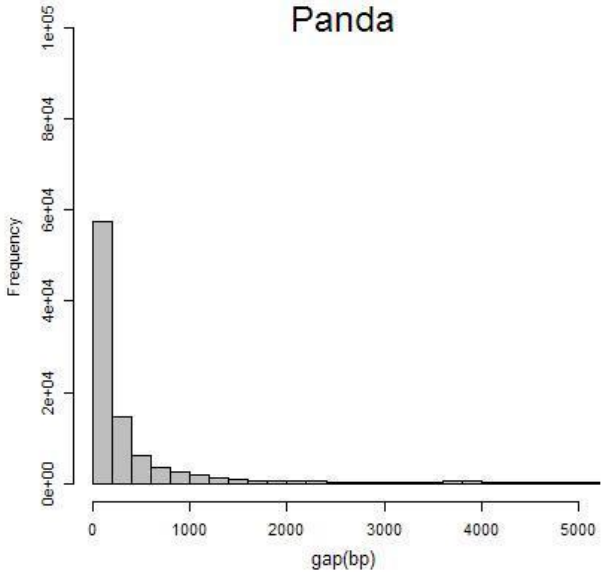




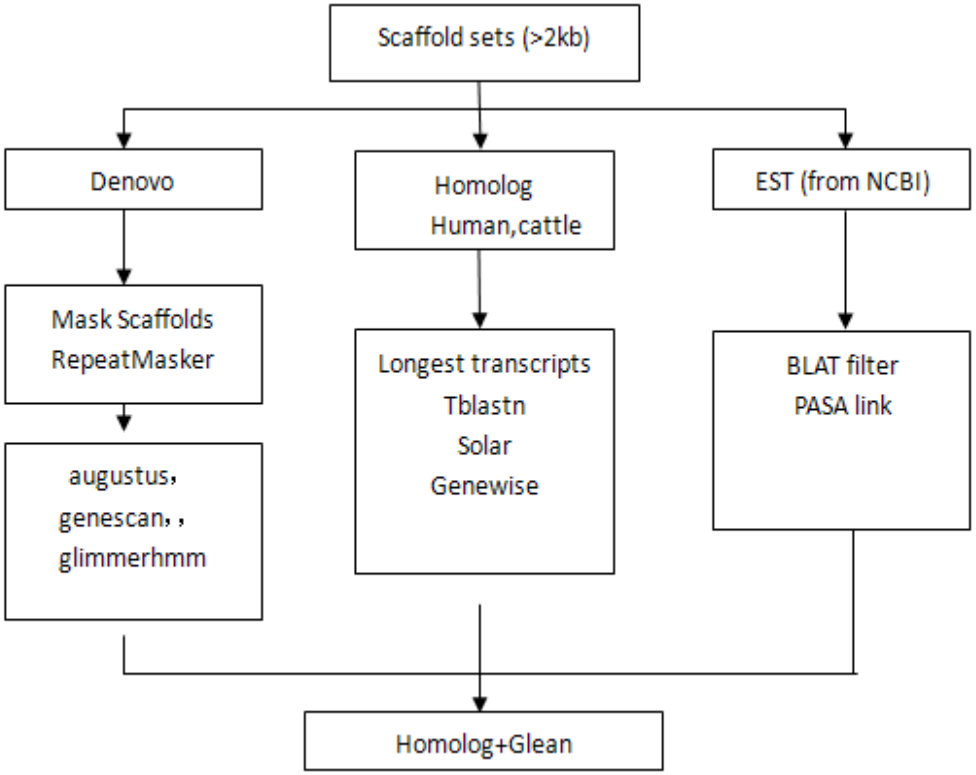


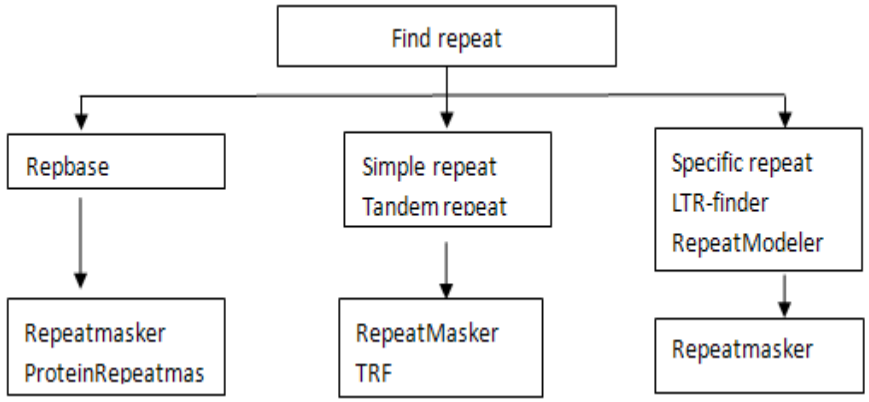


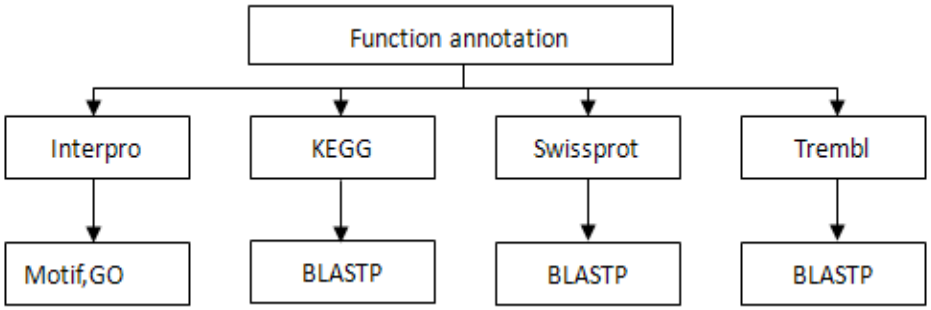


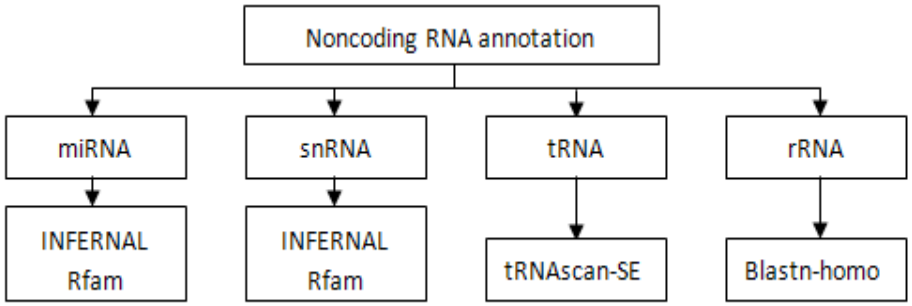


Gene Annotation Pipeline (the core is homology annotation using human and cattle ensemble gene sets)









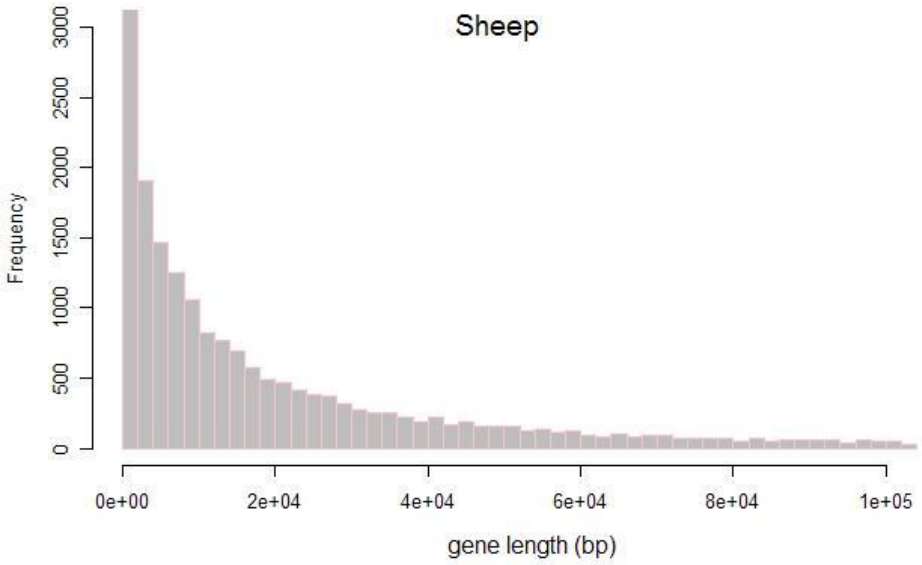
Protein-coding gene: loss one exon for one gene on average

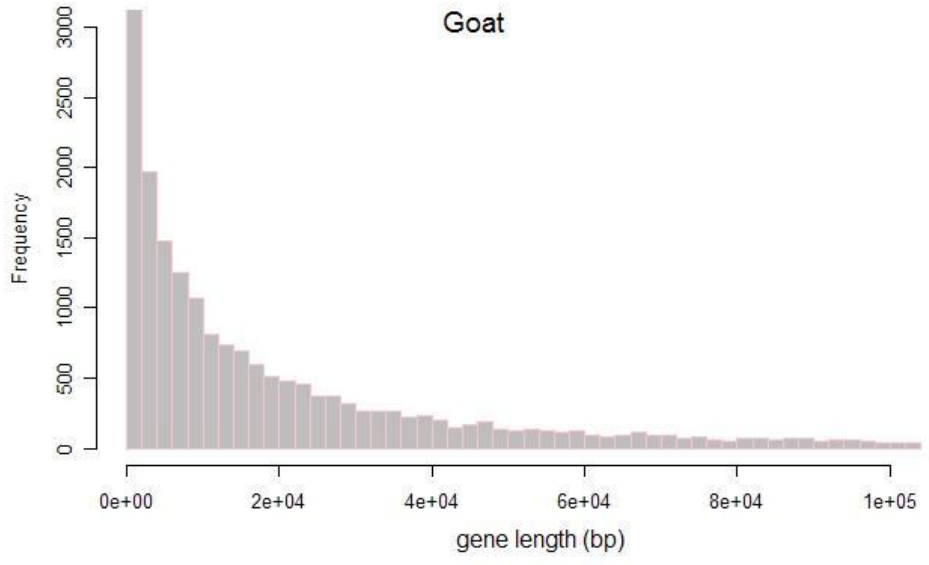
	Gene (kb)	Pep (aa)	exons	Total number
Human	49.7	578.3	10.3	20074
Cattle	32.9	510.6	9.3	20892
Sheep	28.5	445.8	8.1	21982
Goat	30.0	461.4	8.2	22175

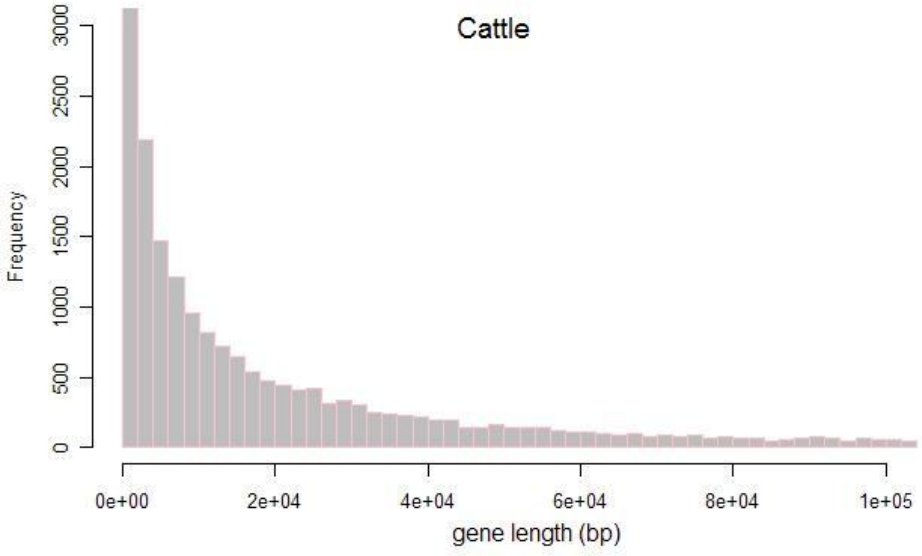
Sheep	Number	Percent(%)
Annotated	20199	91.9
Swissprot	19717	89.7
TrEMBL	19972	90.9
InterPro	16761	76.2
KEGG	14438	65.7
GO	12475	56.8
Unannotated	1783	8.1

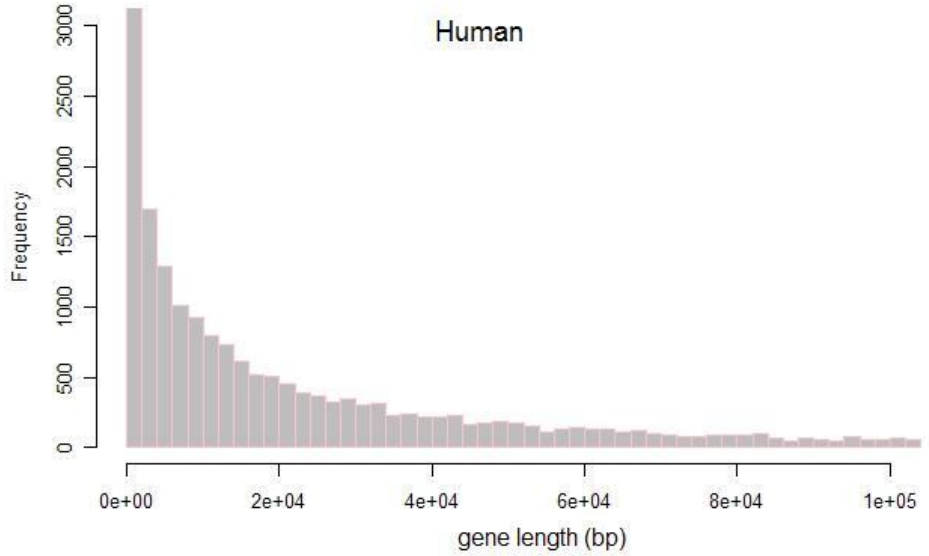
Homology	Gene family	Sheep gene
Bos & Hom :	14252	
Bos & Sheep :	14765	16604
Hom & Sheep :	14082	15967

Gene length distribution: no difference





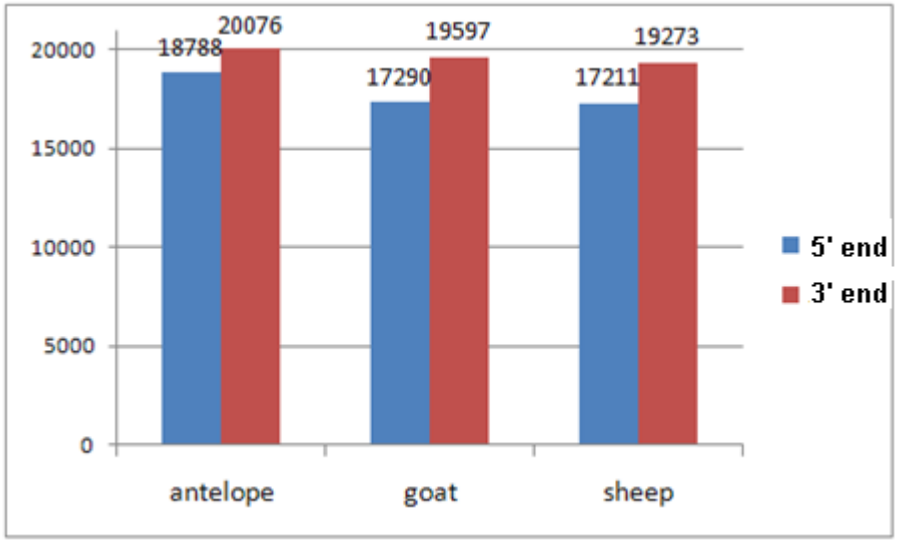


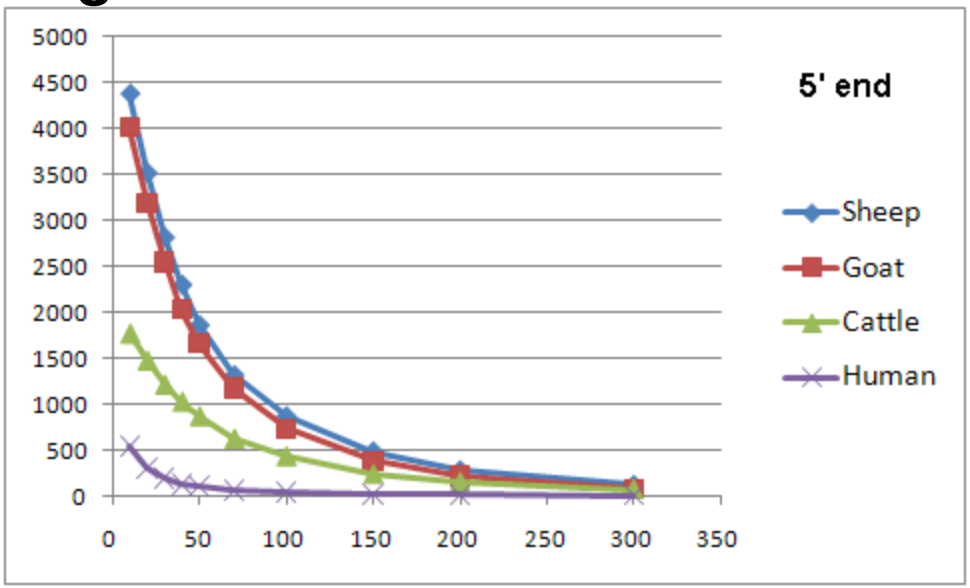


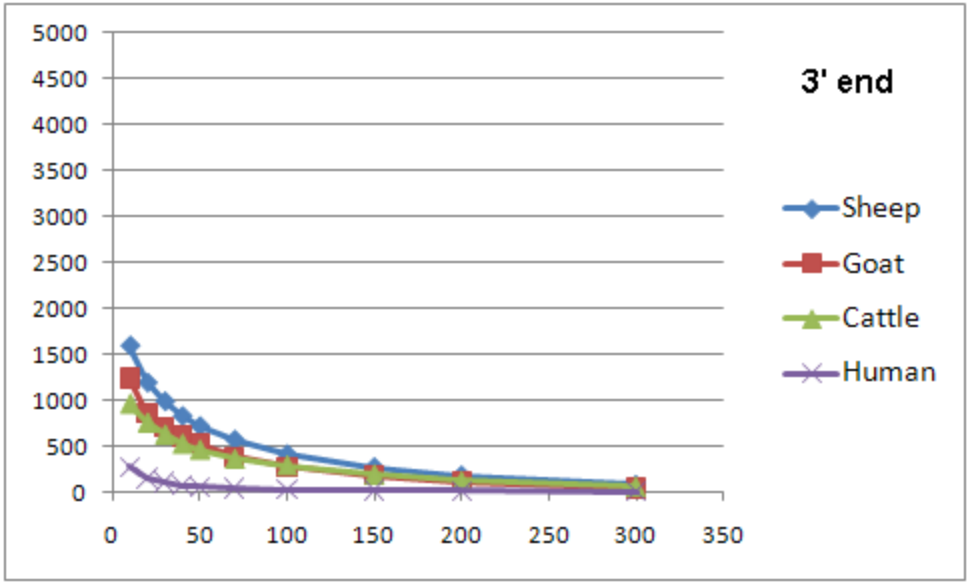
The beginning of 2000-3000 genes are gaps
The left 1000 questionable genes located on inter-scaffolds or tandem repeat regions

11000 single-copy orthologous genes were aligned, and the missing gene number on head and tail is shown:

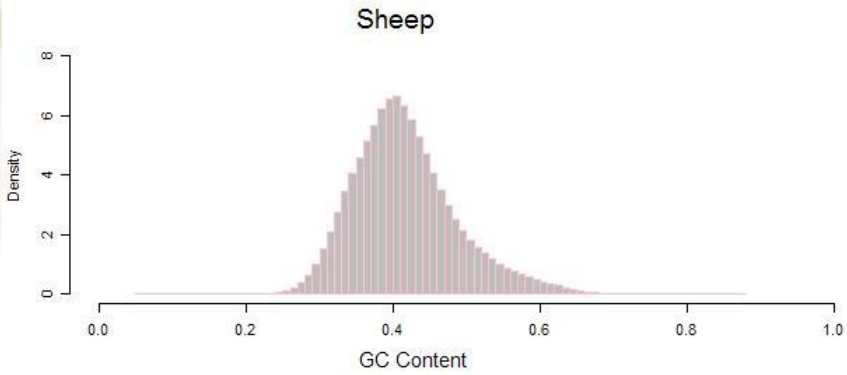
The 5’ end and 3’ end 20 aa of the 20892 cattle genes were used to Tblastn the sheep, goat and antelope genome, the hit number is shown:

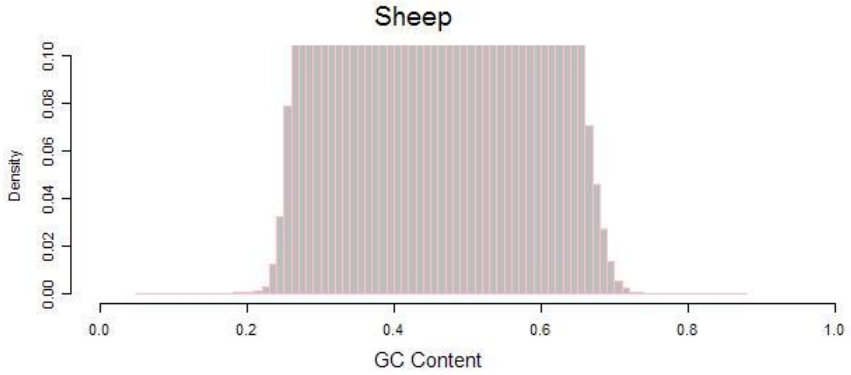


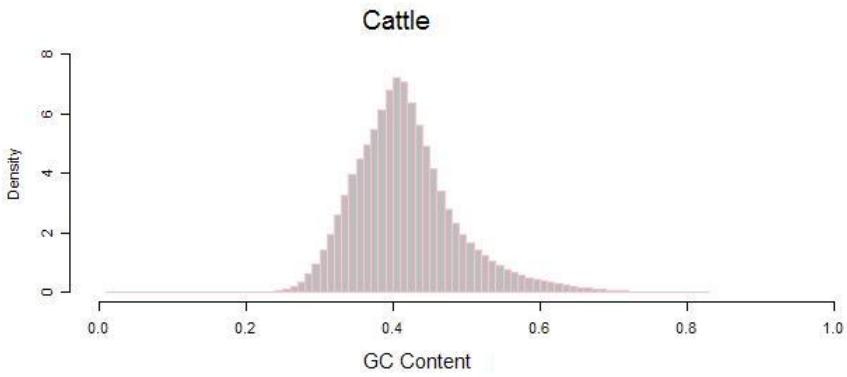


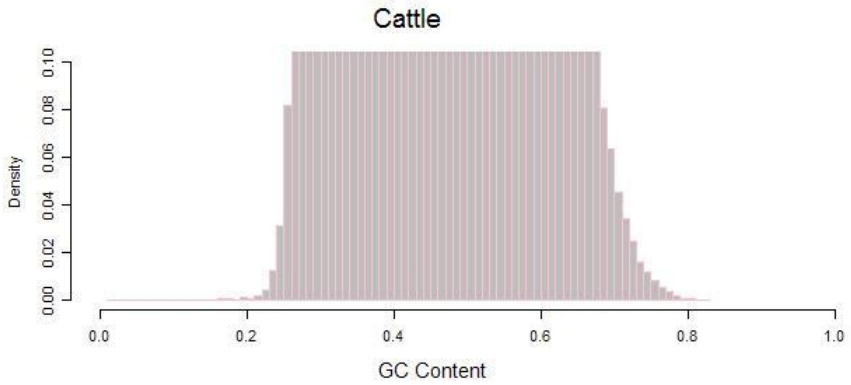


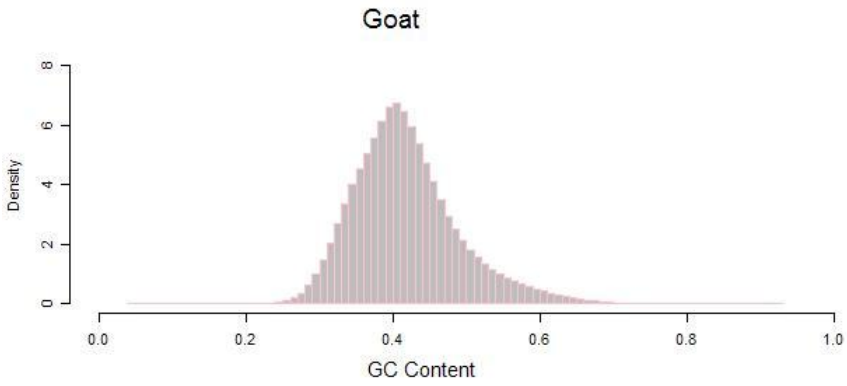
Genome GC content distribution: a tiny part of high GC region is missing during Illumina sequencing, when adjust y-axis scale from 8 to 0.1

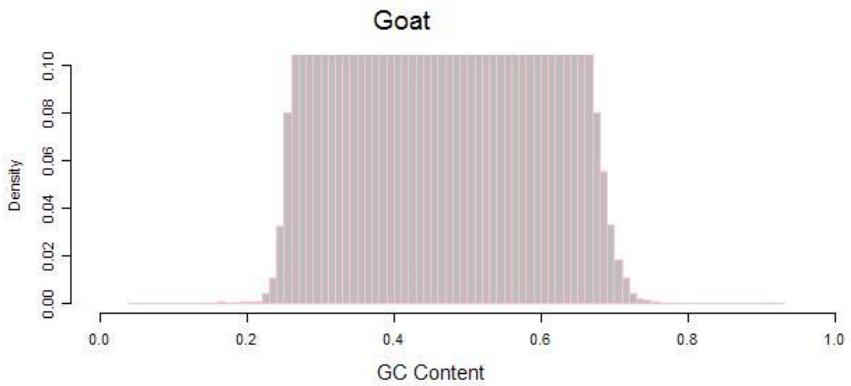












Non-coding RNA

Type		Copy (w)	Average length (bp)	Total length (bp)	% of genome
miRNA		**477**	88.83	**42,373**	**0.001564**
tRNA		**2,058**	**74.34**	**152,987**	**0.005645**
rRNA	**rRNA**	**292**	**97.26**	**28,401**	**0.001048**
	18S	53	111.51	5,910	**0.000218**
	28S	105	110.18	11,569	**0.000427**
	5.8S	4	90.00	360	**0.000013**
	5S	130	81.25	10,562	**0.000390**
snRNA	**snRNA**	**152**	**115.02**	**17,483**	**0.000645**
	CD-box	40	94.15	3,766	**0.000139**
	HACA-box	43	136.70	5,878	**0.000217**
	splicing	63	113.51	7,151	**0.000264**
	scaRNA	5	125.00	625	**0.000023**
	Unknown	1	63.00	63	**0.000002**

Calling heterozygotes and SNPs

SOAPsnp: (1) we used a Q20 quality cutoff; (2) we required at least 10 reads; (3) the overall depth, including randomly placed repetitive hits, had to be less than 100; (4) the approximate copy number of flanking sequences had to be less than 2, to avoid misreading by copy number variations. (basicly this method follows the panda and Yanhuang paper).

Using both the male and female Texel sequence separately identified about five million heterozygous SNPs

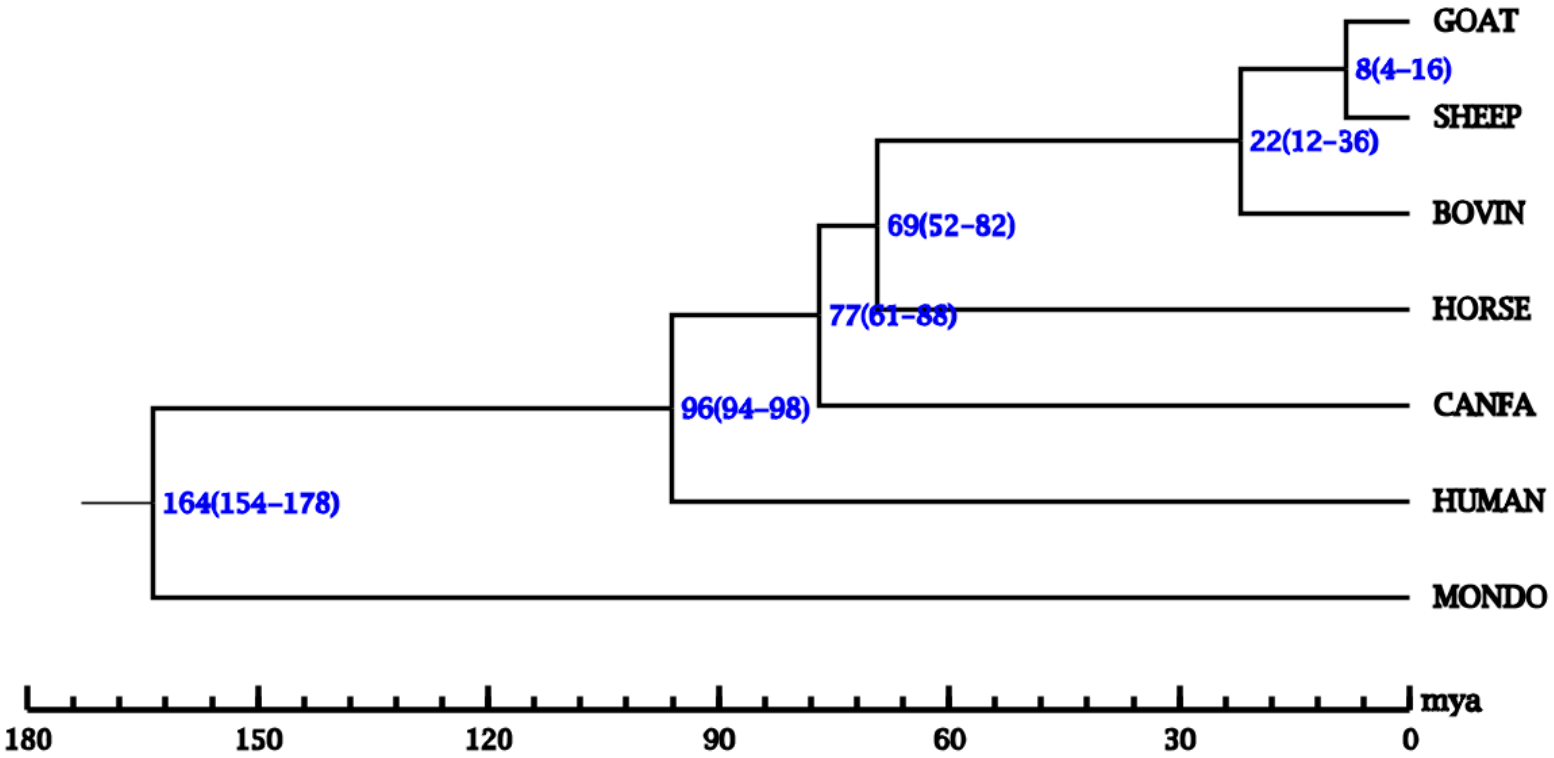
Ewe: 5,102,344
Ram: 5,247,132
Overlap:1,493,816

Checked by Ovine SNP50K BeadChip
47671 SNPs have clear results in the 50K CHIP and could be mapped onto reference genome. For these SNPs, **97.8% in the ewe and 97.9% in the ram** are matched.

The transcriptome data for gene annotation in near future

Sheep tissue	Total Length (Gb)
WAT per	2.4
Heart	2.1
Liver	2.4
Ovary	2.3
Kidney	2.1
Brain	2.1
Lung	1.9
total	15.3

Insert size: 200bp Read length:75bp



Phylogenetic tree using the 14,903 single-copy orthologous genes on 4-fold degenerate sites (using MCMC model, 97 Mya, the divergence time between human and dog97 as standard time)

Compare with other animal genomes

	Horse	Cattle	Panda	Sheep
sequence clone	Fosmid	BAC shotgun	150-500 2kb 5kb 10kb	150-800 2kb 5kb 10kb 20kb BAC
Coverage	6.8×	7.1×	56×	**75×**
Total contig length (ungapped)	2.43Gb	2.73Gb	2.25Gb	2.45Gb
Total scaffold length (gapped)	2.47Gb	2.87Gb	2.30Gb	2.71Gb
N50 contig	112Kb	49Kb	1.5kb 40Kb	**1.1kb** **18kb**
N50 scaffold	46Mb	2Mb	1.3Mb	**1.1Mb**
Heterozygosity Rate	1/2000	1/1700	1/750	**1/500**
Anchor to chromosome	96%	90%	---	**95%**
Repeat ratio	49.5%	46.5%	36.5%	40.5%

Acknowledge

BGI-Shenzhen
Jun Wang
Jian Wang
R&D Department
University of Copenhagen
Karsten Kristiansen
Jacob Hansen
INRA Toulouse
Thomas Faraut
University of Melb
Jillian Maddox
University of New England
Hutton Oddy
AgResearch
John E. McEwan
Kunming Institute of Zoology, CAS
Wen Wang
Wenguang Zhang
Yang Dong
CSIRO Livestock Industries
James Kijas
Brian Dalrymple
Wes Barris
The Roslin Institute
Richard Talbot
Utah State University
Noelle E. Cockett

All my colleagues in KIZ, BGI and ISGC