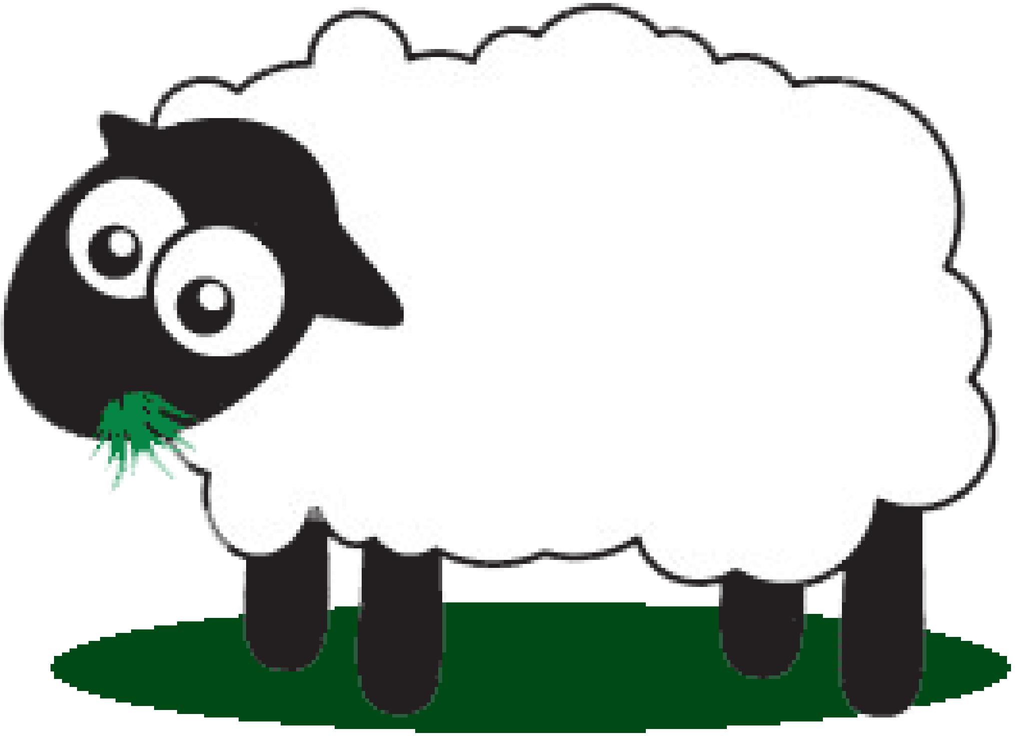


High quality reference genome of the domestic sheep (Ovis aries)

Yu Jiang and Brian P. Dalrymple

CSIRO Livestock Industries
on behalf of the International Sheep Genomics Consortium

Outline of presentation

- Evaluation of the draft assembly of a large animal genome (Oar v2.0) generated by next-generation sequencing platform
- Pipeline for producing Oar v3.0
- Draft strategy for completed reference assembly Oar v4.0

- Application of sheep genome sequence
  - Identification of inprinted genes by screening allelic imbalance expression

ISGC

Sheep genomics resources (ISGC)

Genomic Resource	Description	Data
International Mapping Flock	crossing five breeds	Crawford *et al*., 1995
Linkage Map	genotyping of the IMF	Maddox *et al*., 2001
BAC library	a male Texel, CHORI-243	In early 2002
5000-rad RH	US Suffolk ram	Eng *et al*., 2004
BAC ends sequence	More than 10,000 pair of BACs	In 2005
1200-rad RH	INRA	Laurent *et al*., 2007
Virtual Genome	Using BACs to reorder human genome	Dalrymple *et al*., 2008
67,000 SNPs	Mainly from six animals	In May 2008
Oar v1.0	~43% genome base on cattle genome	In Jan 2009
SNP50 BeadChip	49,034 genotypes post QC	In Aug 2008
Genotyping	1536 SNP chips	Kijas *et al*., 2009
HapMap	Over 70 breeds	Kijas *et al*., 2012
Oar v2.0	~95% genome from two Texel animals	In Mar 2011
HapMap	Over 70 breeds	Kijas *et al*., 2012
Oar v3.0	~99% genome from two Texel animals	In Mar 2012?

Oar v3.0 will be a combinative version of v2.0 and v1.0

Main issues with Oar v2.0

- Oar v2.0
  - the length of scaffold set is 2.71Gb, N50 is 1.1Mb, 6.9% is gap. 95% (2.57 Gb) of the scaffold sequence was placed onto the chromosomes, Using sheep BACs, RH map and linkage markers.
  - Independently assembly, generally very high conserved synteny between Oarv2.0 and sheep BACs, goat genome or bovine genome, suggests the assembly is good.
- Gaps in the assembly
  - True gaps
    - Repeat elements not present in their entirety
    - High GC regions, in particular around transcription start sites
  - Artificial gaps
    - Localised assembly issues, in particular at the ends of scaffolds
- Assembly err: incorrectly duplicated sequence
  - Missed segmental duplicates
  - Artificial tandem duplicates (which lead to artificial gaps)

Sequences included in the genome reference sequence

Sample	Purpose	Sequence method	Paired-end libraries	Insert size (bp)	Libraries	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**				0.7		0.60
Male	for check	454	**20kb**				0.4		0.30
other	for check	454					9.0		3.00
Male	for check	Sanger	**184kb**				0.3	687	0.09

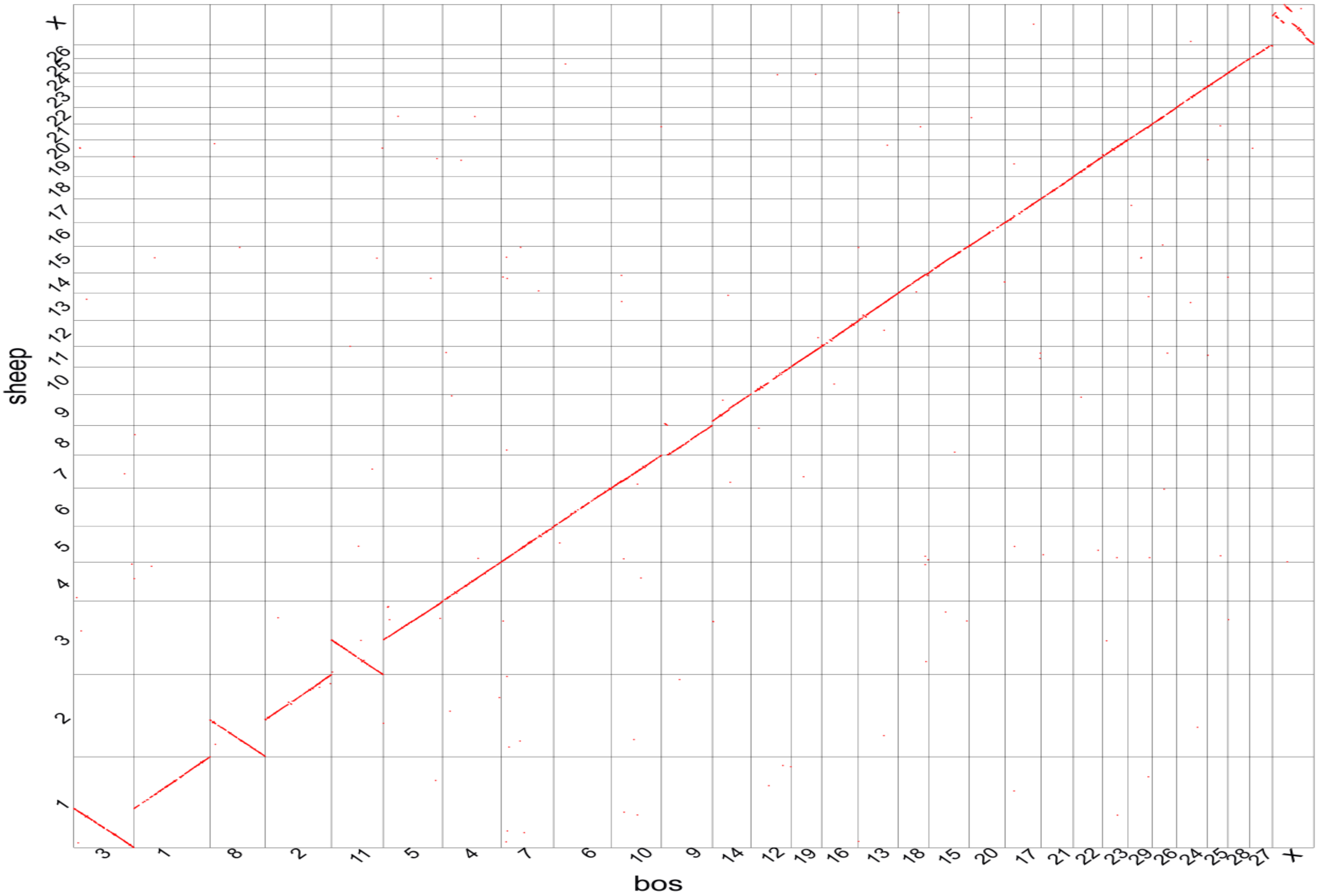
The sheep long insert paired-reads were used for assembly error correction.

“the length and depth will change in problem regions”

ISGC

Ruminants have very conserved synteny relationships

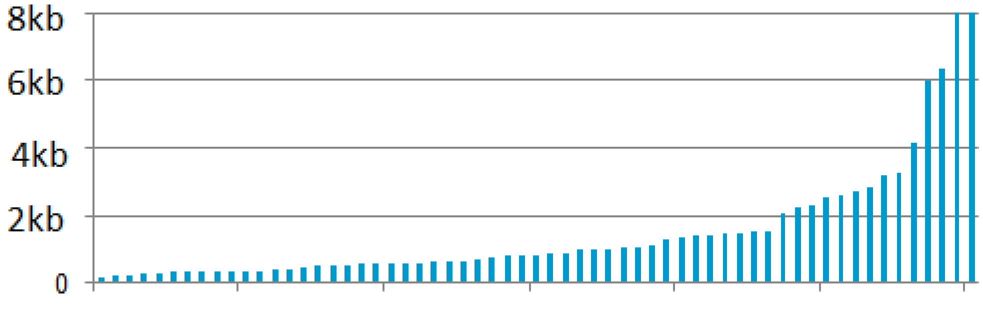
the problem regions will be double check



Intra-scaffold gaps and errs:
Compare assembly sets with sequenced BACs (CHORI-243)

- 61 gaps (4.7% of total length)
  - 3.9% is known repeat sequence
  - 0.3% of is unique sequence which could be filled from existing 454 reads

15 BACs (2.2Mb) VS Scaffold1489



the length distribution of 61 gaps

- errors
  - the length of gap is wrong : 3

	Correction (bp)	Length in Scaffold (bp)	Length in BAC (bp)
gap	-12944	13972	1028
gap	-1752	7762	6010
gap	-1154	1810	656

  - artificial tandem duplications: 4

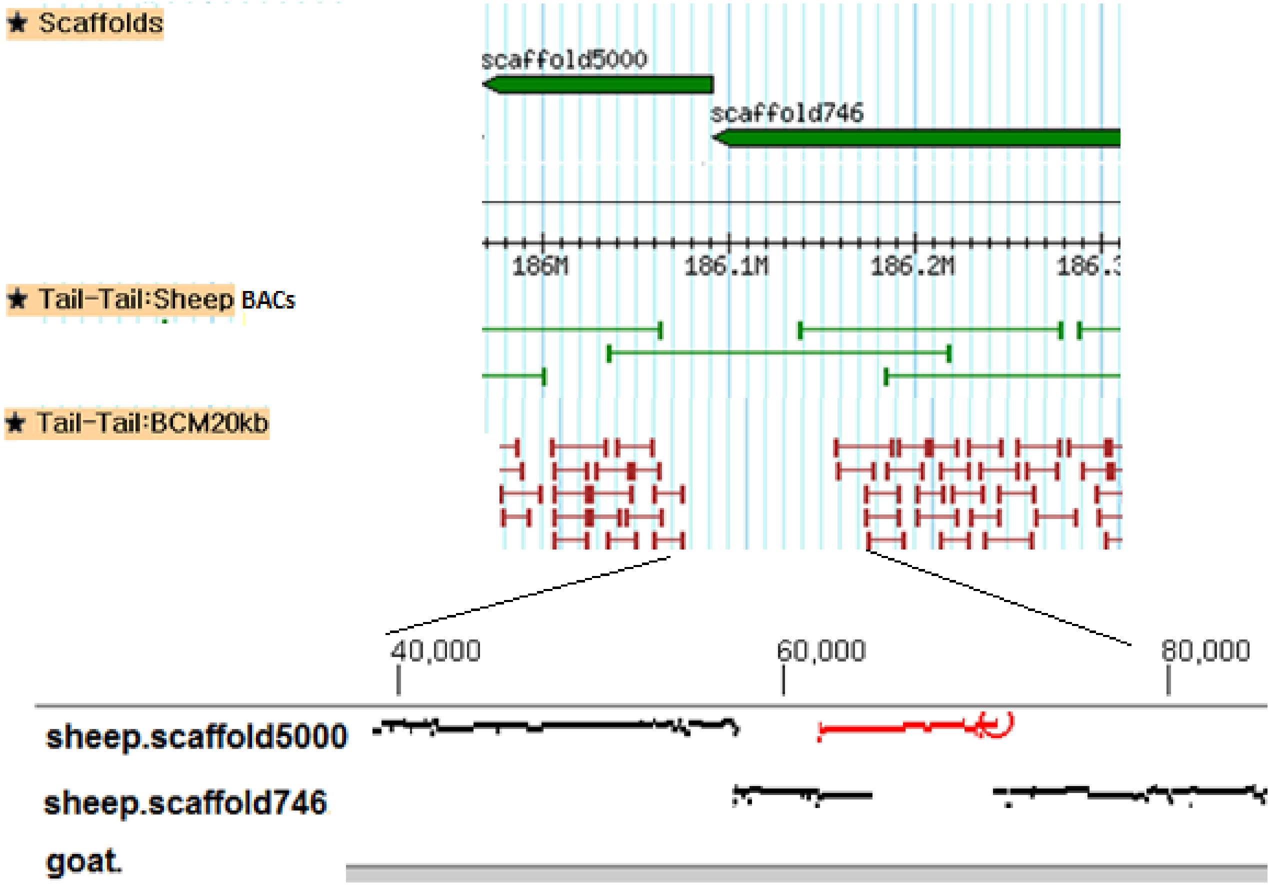
tandem dup	-1517
tandem dup	-1448
tandem dup	-973
tandem dup	-580

Genome-wide adjustment in 2,600 Mb:
5,124 artificial tandem duplicates were removed (-15 Mb)
6,000 the length of gaps were changed (-20 Mb)

Inter-scaffold gaps on chromosomes:

real gaps, unmapped scaffolds and wrong assembly tips

Example of overlapping assembly between two scaffolds



4403 inter-scaffold gaps screening: (incorrect ends ratio is ~10%)

657 overlapping (10bp~3kb) and even 118 skips(3kb~100kb)

From Oar v2.0 to Oar v3.0 (Scaffold level)

Step1. Gap filling using Illumina reads

dataset	gaps	length	% genome
OAR v2.0	306,000	190 Mb	6.90%
after filling	158,000	123 Mb	4.70%

Short gaps (0~2kp) 144,348
primarily repeats and very high-GC sequences
Estimate that at least 75% and as many as 95% may be repeats

Long gaps (2kb~20kb) 13,357
May not be real gaps (Masked duplicated sequence, or unmapped contigs)

Step2. Correction of assembly errs

- 5,124 artificial tandem duplicates were removed (-15 Mb)
- 6,000 the length of gaps were changed (-20 Mb)

From Oar v2.0 to Oar v3.0 (chromosome level)

Step1. 4403 Inter-scaffold gaps estimation

92.9% (4089/4403) gaps were evaluated by syntenic sequence and the gaps have a median length of 760 bp, and a total length of 42 Mb.

Step2. merge overlapped scaffolds

657 overlapping (10bp~3kb) and even 118 skips(3kb~100kb)

Step3. anchoring unmapped scaffolds using syntenic relationship

> \> 20kb 1,242 scaffolds (57 Mb) are considered;
>
> < 2kb 480,000 scaffolds(70 Mb) are removed (duplicates or reads)

~300 extra scaffolds (35 Mb) were mapped onto chromosomes

The left unmapped scaffolds (30 Mb):

It means 99% of scaffold sequence (by length) is now assigned to a position on a chromosome (2.6 Gb assembly length)

Before release : Oar 3.0

Step1. another round of gap filling

- Gap filling using BAC-end sequences and 454 reads
- Gap filling using remaining unassigned Illumina scaffolds

Step2. reads re-mapped for assembly checking

- Duplications
- Errs

pre Oar v3.0 mainly covered ~99% single copy region and ~96% transcript region

- Mapping 59,042 SNPs onto pre Oarv3.0
  - 96.9%(57217/59042) when mapping identities>=98%
  - 99.4%(58677/59042) when mapping identities>=95%

    means ~99% single copy region is covered

- Mapping 330kb cDNA onto pre Oarv3.0
  - For the single hit cDNA, 95.8% (487bp/508bp) is mapped

    means ~96% transcript region is covered (5’ ends of genes prefer high GC)

Using Oarv1.0 (assembled 454 data sets) close 3,040 (1.2 Mb length) 5’ end gaps

- average length of 391 bp
- GC content is 63.6% (compare GC content 42% in whole genome).

ISGC

Draft strategy for completion of the sheep reference genome assembly – Oarv4.0

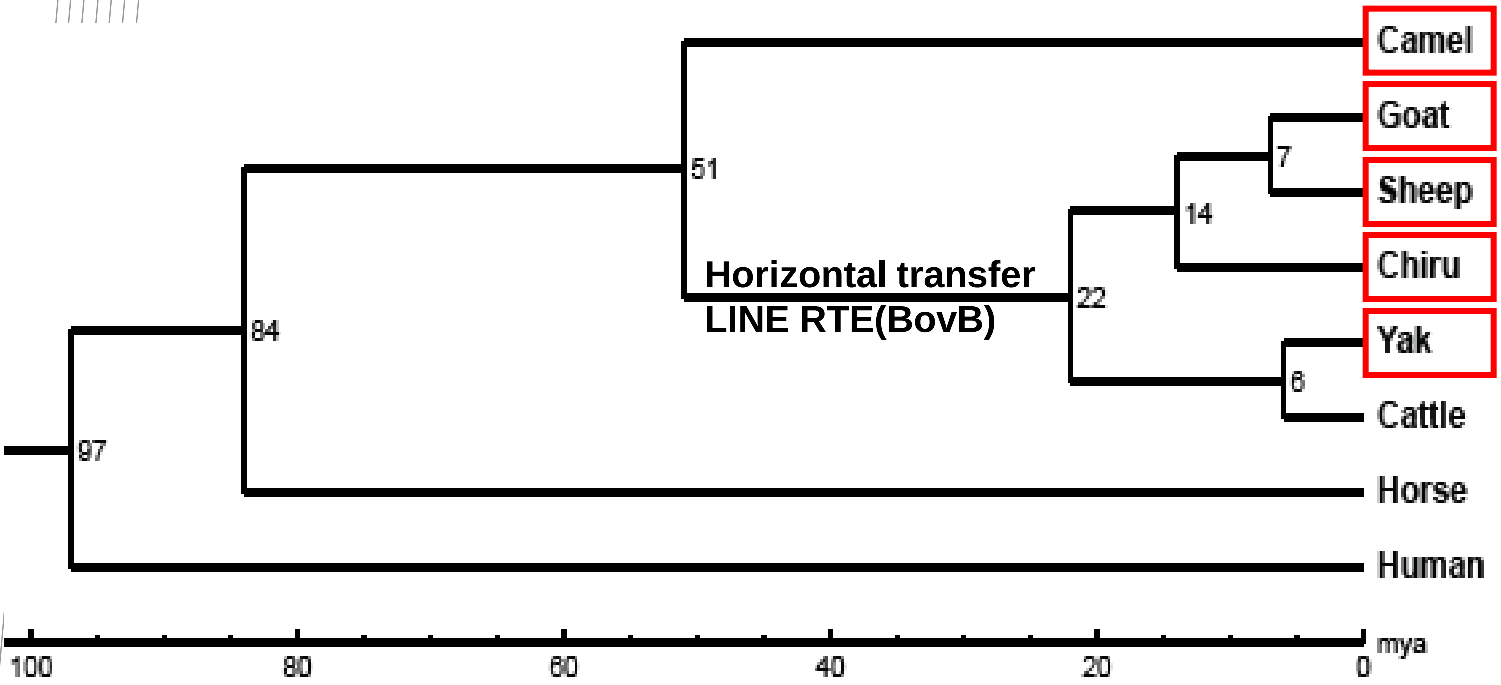
- New data required
  - Sequence data to fill gaps and resolve remaining ambiguities in the assembly
    - High GC Illumina sequencing?
    - More long insert mate pair sequence for Repeat gaps?
    - Re-sequence for fixed segmental duplications?
  - Map BACs to X and Y chromosomes cytogenetically
  - Fill the remaining big gaps and part of Oar Y with BAC sequencing,

- Planned date for release PAG2013

- To be discussed at the ISGC workshop
  - Monday 16th Jan 11:30 am – 3:00 pm
  - Towne Room in the Meeting House

ISGC

We need a good sheep genome on behalf of ruminants or small ruminants clade



Horizontal transfer
LINE RTE(BovB)

	Contig N50	Scaffold N50
Camel	54Kb	4.1Mb
Goat	18Kb	2.2Mb
Sheep	31Kb	1.1Mb
Chiru	19Kb	2.7Mb
Yak	20Kb	1.4Mb
Cattle	49Kb	2.0Mb
Horse	112Kb	46Mb

Phylogeny of 5 recently sequenced ruminant species:
next generation sequence
Sanger sequence

From genotype to phenotype

Genome-wide exploration of the Imprinted genes

- Gene imprinted is an epigenetic modification to inactivate one allele of a gene in a parent -of-origin manner (Fowden, 2011). There may be more than 1000 loci with parent-of-origin allelic effects in mouse brain (Gregg, 2010)
  - One example is the famous *DLK*1 imprinted gene cluster which resides in a 220 kb region of Oar18.(Charlier, Genome Res. 2001), which is related with the Polar overdominance of the callipyge phenotype (economic trait).
- In general, many known inprinted genes are related with developmental or epigenetic regulation. It can be used for economic trait or disease studies, to correct the inconsistency between genotypes and phenotypes.
- We identified 636 putative imprinted genes in 5492 highly expressed sheep genes. More than 600 of them are novel.

ISGC

Identified inprinted genes by screening allelic expression

- Genome-wide and transcriptome-wide identification of Allelic imbalanced SNPs and genes
  - The genome assembly
  - The largest number of SNPs
  - Deep RNA-seq data from multiple tissues
- We identified 5 Mb heterozygous SNPs
  - Based on the checking of 50K SNP CHIP experiment, the false positive rate of heterozygous SNPs is less than 0.33%.
- 15Gb of RNA-seq from seven tissues from the same female sheep was sequenced.

8.9% of SNPs show strong allelic imbalance (90:10)

- Cutoff: To detect strong allelic imbalance (90:10), which is defined as inprinted genes (Amada,2011), the statistical test show strong power at 20-fold coverage (>97% correctly) (Nothnagel, 2011).

- 41,669 SNPs >=20-fold coverage
  - 8.9% (3693/41669) SNPs show strong allelic imbalance (90:10)
  - 11.3%(4699/41669) SNPs show weak allelic imbalance (2-9 fold change or ratio between 66.7:33.3 to 90:10).
- Verified by two known examples:
  - *DLK*1 imprinted gene cluster: 198 adjacent SNPs in 238 kb region of Oar18 show strong allelic imbalance .
  - *IGF2* locus: our data support the known biallelic expression in liver and mono-allelic expression in the other sampled tissues

636 putative inprinted genes (strong allelic imbalance)

- The strong allelically imbalanced SNPs are clustered in genes

332 genes are support by one SNP;
304 genes are support by multiple (avg=3.5) SNPs

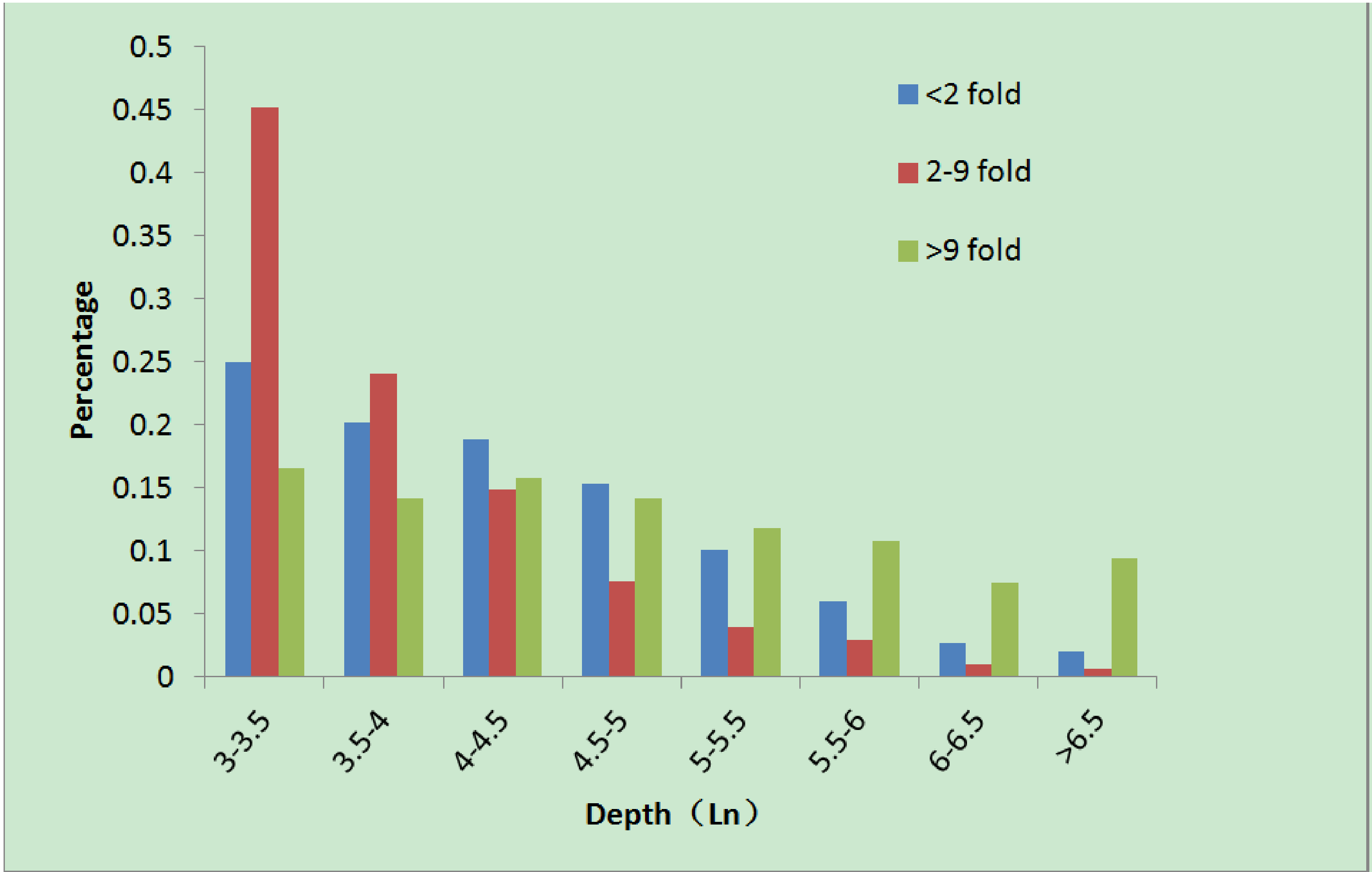
  - 636 genes show strong allelic imbalance (putative inprinted genes) from 5492 inspected sheep genes

- Sometimes several adjacent genes show allele-specific expression, suggesting that they are under control of common regulatory elements.

  - We identified more than 100 such loci, the top three are
    - known DLK1 region (Oar18). 198 SNPs in 238 kb
    - One sub-telomere region(Oar18). 132 SNPs in 913 kb
    - One sub-centromere region (Oar2). 85 SNPs in 19.8Mb

Pattern 1: Strong allelic imbalanced SNPs are enriched in highly expressed genes

•32.9% of the top 1000 expressed SNPs and 22.2% of the top 500 expressed genes are strong allelic imbalanced. The average ratio of strong allelic imbalanced SNPs and genes is 8.9% (3693/41669) and 11.6%(636/5492)



Pattern 2: Strong allelic imbalanced expression are basically conserved across tissues

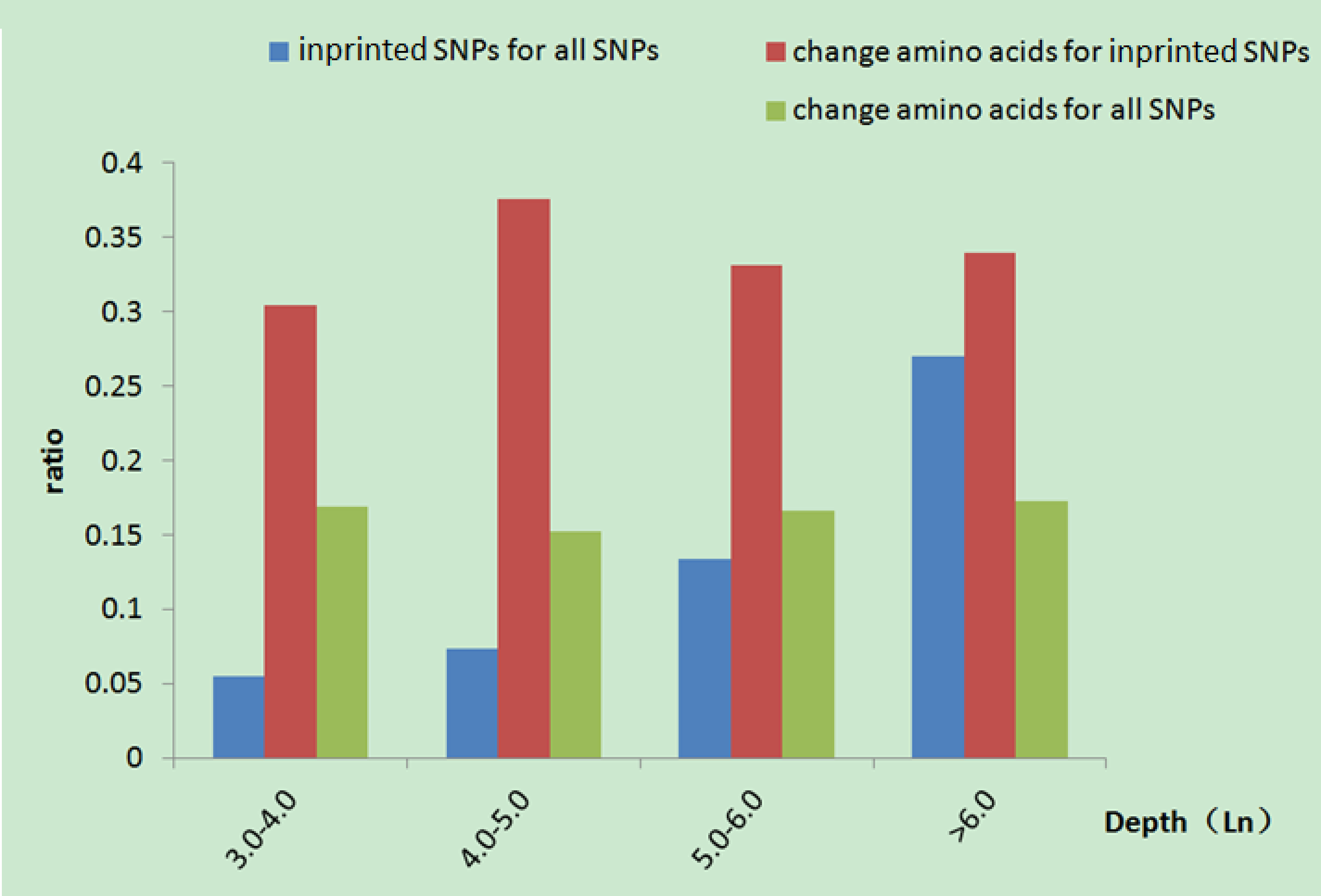
- 94.8% (3141/3314) of the strong allelic imbalanced SNPs are strongly specifically-expressed across all the inspected tissues (tissues with >=10-fold coverage and 90:10 imbalance).

- 4.6% (153/3314)% of the strong allelic imbalanced SNPs have both strongly specifically-expressed tissues and bi-allelic expressed tissues.

- Only 1 SNP is maternal allelic expressed in one tissue, but is paternal allelic expressed in the other tissue (maybe random err).

Pattern 3: SNPs leading to non-conservative amino acid changes prefer to be strong allelic imbalanced

1. Strong unbalanced SNPs change ratio : 35.6% (381/1070)
   All SNPs change ratio :16.3% (1965/12056)
2. Pattern 1 and Pattern 3 are independent, Pattern 3 are under selection



Non-conservative change means: amino acid property changes between neutral-base-acid or polar-nonpolar

Pattern 4: 52.2% of putative inprinted genes are conserved between goat and sheep (~ 7 Mya)

- From 3M SNPs and 10 Gb RNA-seq data from a female goat, we got a similar pattern1,2,3 to sheep
    - Have similar pattern1,2, 3 in goat; just a smaller number, because the total SNP number and sequence depth is smaller than sheep data.

- For the 5492 inspected sheep genes, 1399 of them also could be evaluated in their goat orthologs. 10% (140/1399) of them are allelic imbalanced in goat, comparing with the 8.9% in sheep.

- `For the 636 allelic imbalanced genes in sheep, 134 of them could be evaluated in their goat orthologs, and 52.2% (70/134) are also allelic imbalanced in goats.

Acknowledge

CSIRO Livestock Industries

James Kijas

AgResearch

John E. McEwan

Utah State University

Noelle E. Cockett

BGI-Shenzhen

Jun Wang

Xun Xu

R&D Department

The Roslin Institute

Alan Archibald

Richard Talbot

HGSC – BCM

Kim Worley

Richard Gibbs

INRA Toulouse

Thomas Faraut

Kunming Institute of Zoology, CAS

Wen Wang

Wenguang Zhang

Yang Dong

University of Copenhagen

Karsten Kristiansen

Jacob Hansen

University of Melbourne

Jillian Maddox

University of New England

Hutton Oddy

Welcome to ISGC workshop

Monday 16th Jan 11:30 am – 3:00 pm

Towne Room in the Meeting House