Technologies, resources and tools for the exploitation of the sheep and goat genomes.

The plan

• The current state of the genomes

• The current resources and tools

• The relationships between the genome assembly and the tools and resources and utility

• What do the next two years hold?

• Looking further ahead
THE STATE OF THE GENOME
Goat

• ~2.66-Gb draft genome assembly of a female Yunnan black goat
• Dong et al., Nat Biotechnol. 2013 Feb;31(2):135-41
• Super scaffolding using optical mapping

<table>
<thead>
<tr>
<th></th>
<th>Contig</th>
<th>Scaffold</th>
<th>Scaffold with fosmid sequences</th>
<th>Super-scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bp</td>
<td>Number</td>
<td>bp</td>
<td>Number</td>
</tr>
<tr>
<td>N90</td>
<td>4,410</td>
<td>141,869</td>
<td>440,999</td>
<td>1,348</td>
</tr>
<tr>
<td>N80</td>
<td>7,994</td>
<td>100,335</td>
<td>846,998</td>
<td>922</td>
</tr>
<tr>
<td>N70</td>
<td>11,323</td>
<td>73,948</td>
<td>1,253,003</td>
<td>664</td>
</tr>
<tr>
<td>N60</td>
<td>14,862</td>
<td>54,526</td>
<td>1,694,371</td>
<td>482</td>
</tr>
<tr>
<td>N50a</td>
<td>18,720</td>
<td>39,408</td>
<td>2,212,139</td>
<td>344</td>
</tr>
<tr>
<td>Totalb</td>
<td>2,522,851,955</td>
<td>542,145</td>
<td>2,662,658,003</td>
<td>285,383</td>
</tr>
</tbody>
</table>

\[a\] N50 includes fosmids.
\[b\] Total numbers are not exact due to rounding.
\[c\] Number includes additional scaffolds.
Sheep Oar v3.1

• Texel ewe (BGI), ram (Roslin, BCM)
  • Some additional data generation and data analysis from Oar v 2.0
• Primary objective to refine draft assembly of the sheep reference genome
  • Halved number of gaps - doubled contig N50
  • Scaffold order and orientation refined on the basis of sheep BACs and linkage and RH maps
  • Removed 12,000 false duplicates with length 28 Mb
  • covered ~99% of the unique genome
• Released September 2012 GenBank
  • Annotated by NCBI, Ensembl (coming very soon), UCSC (?)

The Texel female was 6 months old (provided by Jacob B. Hansen, University of Copenhagen) sequenced BGI DNA: Liver RNA: 7 tissues.

The Texel ram was used previously as the DNA source for CHORI-243 BAC library DNA: Blood (Dalrymple et al., 2007).-----Roslin
## Sequence used for Oar v2.0 and additional for Oar v3.1

<table>
<thead>
<tr>
<th>Sample</th>
<th>Purpose</th>
<th>Sequence method</th>
<th>Paired-end libraries</th>
<th>Libraries</th>
<th>GA Lanes</th>
<th>Total length (Gb)</th>
<th>Reads Length (bp)</th>
<th>Coverage (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>assembly</td>
<td>Illumina</td>
<td>180bp</td>
<td>1</td>
<td>4</td>
<td>23.8</td>
<td>101</td>
<td>7.93</td>
</tr>
<tr>
<td>Female</td>
<td>assembly</td>
<td>Illumina</td>
<td>350bp</td>
<td>4</td>
<td>21</td>
<td>105.0</td>
<td>101</td>
<td>35.00</td>
</tr>
<tr>
<td>Female</td>
<td>assembly</td>
<td>Illumina</td>
<td>800bp</td>
<td>2</td>
<td>6</td>
<td>32.0</td>
<td>101</td>
<td>10.67</td>
</tr>
<tr>
<td>Female</td>
<td>assembly</td>
<td>Illumina</td>
<td>2kbp</td>
<td>2</td>
<td>11</td>
<td>35.7</td>
<td>45</td>
<td>11.90</td>
</tr>
<tr>
<td>Female</td>
<td>assembly</td>
<td>Illumina</td>
<td>5kb</td>
<td>2</td>
<td>6</td>
<td>18.5</td>
<td>45</td>
<td>6.17</td>
</tr>
<tr>
<td>Female</td>
<td>assembly</td>
<td>Illumina</td>
<td>9kb</td>
<td>1</td>
<td>3</td>
<td>8.3</td>
<td>45</td>
<td>2.77</td>
</tr>
<tr>
<td>Female</td>
<td>assembly</td>
<td>Illumina</td>
<td>17kb</td>
<td>1</td>
<td>1</td>
<td>1.8</td>
<td>45</td>
<td>0.60</td>
</tr>
<tr>
<td>Female*</td>
<td>fill gap</td>
<td>Illumina</td>
<td>200bp</td>
<td>1</td>
<td>1</td>
<td>3.0</td>
<td>45</td>
<td>1.00</td>
</tr>
<tr>
<td>Male</td>
<td>fill gap</td>
<td>Illumina</td>
<td>200bp</td>
<td>1</td>
<td>16</td>
<td>77</td>
<td>101</td>
<td>24.0</td>
</tr>
<tr>
<td>Male</td>
<td>fill gap</td>
<td>Illumina</td>
<td>500bp</td>
<td>1</td>
<td>24</td>
<td>72</td>
<td>101</td>
<td>25.5</td>
</tr>
<tr>
<td>Male**</td>
<td>fill gap</td>
<td>Illumina</td>
<td>554bp</td>
<td>8</td>
<td>1</td>
<td>36</td>
<td>101</td>
<td>12.0</td>
</tr>
<tr>
<td>Male**</td>
<td>fill gap</td>
<td>Illumina</td>
<td>1.3kb</td>
<td>1</td>
<td>1</td>
<td>27</td>
<td>101</td>
<td>9.00</td>
</tr>
<tr>
<td>Six***</td>
<td>fill gap</td>
<td>454</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>9.0</td>
<td>240</td>
<td>3.00</td>
</tr>
<tr>
<td>Male</td>
<td>check</td>
<td>454</td>
<td>8kb</td>
<td>8</td>
<td></td>
<td>0.7</td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>Male</td>
<td>check</td>
<td>454</td>
<td>20kb</td>
<td>10</td>
<td></td>
<td>0.4</td>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td>Male</td>
<td>Check/fill gap</td>
<td>Sanger</td>
<td>184kb</td>
<td>0.3</td>
<td></td>
<td>687</td>
<td></td>
<td>0.09</td>
</tr>
</tbody>
</table>

* MeDIP-seq for high GC content sequence
** New Illumina protocol for GC content unbiased sequence
*** Six animal from breeds: AW, ROM, TEX, MER, SBF and PD
How does Oar v3.1 compare with other species?

- using next generation sequence and assembly platform
- Sanger sequence

*41 kb contigs assigned to chromosomes, 40 kb, all contigs
CURRENT PUBLIC DOMAIN TOOLS AND RESOURCES
Goats

• Genome assembly
  • NCBI annotation – EST-based and predicted RefSeqs and gene models
  • Sequence searches – NCBI
  • RH panel
  • Optical Map

• 50 K SNP chip

• 25 animals resequencing
  • ADAPTmap project
    – Coordinating genotyping and sequencing goat breeds
    – http://www.goatadaptmap.org/
SNP discovery

- 6 breeds:
  - Alpine, Saanen, Creole
  - Savanna, Katjang, Boer

- Data from several laboratories
  - Genomic sequences
    - INRA, France
    - MARDI, Malaysia
    - University of Utrecht, Netherlands
  - ESTs
    - Italy, Spain, USA...
60 000 selected SNPs

1 : EST
2 : Het. in 5 breeds
3 and 4 : Het. in 4 breeds
5 and 6 : Het. in 3 breeds
10 : Het. S & A
11 : Het. (A or S) and (C or B or KS)
12 : Het. C and (B or KS)
13 : Het. A or S
20 : other
90 : INDEL

Alpine, Saanen, Creole, Savanna, Katjang, Boer

5 breeds => ~38%
4 breeds => ~78%
3 breeds => ~96%
Milk breeds => ~100%
Meat breeds => ~97%
60,000 SNPs - Spacing

*median interval => ~ 40kb
Chip manufacturing and characteristics

• 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
• 53,348 synthesized loci
• 52,295 successful loci
• 8,000 ordered samples in September 2011
• Cluster files (.egt) available: Gwenola.Tosser@toulouse.inra.fr
• SNP sequences and annotation published in dbEST
  Information on www.goatgenome.org
A chip useful for many breeds

<table>
<thead>
<tr>
<th>Breed</th>
<th>Samples</th>
<th>SNPs MAF&gt;0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpine</td>
<td>53</td>
<td>51339</td>
</tr>
<tr>
<td>Angora</td>
<td>26</td>
<td>47195</td>
</tr>
<tr>
<td>Boer</td>
<td>30</td>
<td>48494</td>
</tr>
<tr>
<td>Creole</td>
<td>38</td>
<td>50216</td>
</tr>
<tr>
<td>Jinlan</td>
<td>13</td>
<td>45648</td>
</tr>
<tr>
<td>Katjang</td>
<td>13</td>
<td>33873</td>
</tr>
<tr>
<td>Saanen</td>
<td>57</td>
<td>51689</td>
</tr>
<tr>
<td>Savanna</td>
<td>20</td>
<td>46629</td>
</tr>
<tr>
<td>Skopelos</td>
<td>27</td>
<td>50908</td>
</tr>
<tr>
<td>Yunling</td>
<td>1</td>
<td>17335</td>
</tr>
</tbody>
</table>
Sheep

- Reference genome assembly
  - NCBI annotation – EST-based and predicted RefSeqs and gene models
  - Ensembl annotation – RNA-Seq-based and predicted gene models
  - Genome browsers and sequence searches – NCBI and Ensembl
  - Genes with allelically imbalanced expression

- SNP chips/panels
  - Parentage SNP panel
  - Small SNP panel- 7k
  - Ovine SNP 50K beadchip – Illumina
  - Sheep HD beadchip – Illumina

- Hapmap project SNP 50 genotypes
  - ~2800 animals diversity fo breeds, multiple animals per breed
Sheep

• Genome resequencing
  • >75 individuals diversity, mainly single individuals per breed
  • ~32 million SNPs
• CNV region list - preliminary
• Tissue gene expression survey ~40 tissues
  • Texel trio and fetus with same parents
The sheep variome/resequencing

• ∼10 X Illumina read coverage was generated from each of 68 diverse domestic sheep, 3 *Ovis canadensis* (bighorn) and 2 *Ovis dalli* (thinhorn) individuals.

• The reads from each animal were aligned to the reference sheep genome to identify
  • SNPs
  • Insertions
    – Reference genome does not contain all sequence present in any sheep
  • Deletions
    – All sheep do not contain all the sequence present in the reference genome
  • variable coverage between individuals and hence significant CNVs.

• Sequence, BAM and vcf files are available for all of the animals by contacting James Kijas

• Sequence data also available from GenBank
Sheep CNVs – ASIP locus, sheep colour

One copy

Two copies

Multicopy adjacent seq

One copy

Two copies
### HD chip design

<table>
<thead>
<tr>
<th>SNP type</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equally spaced ~5kb</td>
<td>569,029</td>
</tr>
<tr>
<td>Literature SNPs and Indels</td>
<td>289</td>
</tr>
<tr>
<td>Ovine SNP50 Beadchip SNPs</td>
<td>49,044</td>
</tr>
<tr>
<td>Functional SNPs maf002</td>
<td>30,347</td>
</tr>
<tr>
<td>Functional SNPs maf01</td>
<td>25,180</td>
</tr>
<tr>
<td>Genotyping by sequencing (Redrep)</td>
<td>21,262</td>
</tr>
<tr>
<td>Chromosome unknown</td>
<td>1,088</td>
</tr>
<tr>
<td>TOTAL Design</td>
<td>685,734</td>
</tr>
<tr>
<td>TOTAL pass</td>
<td>606,006</td>
</tr>
<tr>
<td>Call rate &gt;= 0.98</td>
<td>589,630</td>
</tr>
</tbody>
</table>

- Did not screen out SNPs in duplicated non repetitive DNA (aka CNVs and seg dup) and these can be detected (and used)
- Illumina iSelect
- Mean sample call rate = 0.9932 +/- 0.0006
- Mean SNP call rate 0.9931 +/- 0.0695, that includes all the snps with no calls
Tissue gene expression library

- 3-SR and Roslin
- Texel ram and ewe, their lamb and their embryo
- Paired end, stranded, 150 base reads
- ~40 tissues
- ~1.1 TB of data
- Forms the core of Ensembl annotation of the sheep genome
THE RELATIONSHIPS BETWEEN THE GENOME AND TOOLS
Cost v. quality v. utility

- HD SNP chip
- SNP50/parentage SNP chips
- Oar v 1
- Oar v 2
- Oar v 3.1
- Tissue exp. survey
- Oar v 4
- 75 genomes

- >99.5% protein genes id and seq. correct genome
- >95%? protein genes id and seq. correct genome. Ensembl annot.
- ~85% protein genes ident and seq. correct genome. NCBI annotation
- GC-rich islands well covered
- GC-rich islands deficient
- CNV assessment accurate
- CNV assessment preliminary
- Y chromosome assembly
- True species differences > assembly errors

COST
QUALITY
THE NEXT TWO YEARS
Goat

• genome diversity database programme / ADAPTmap project will allow data sharing (50K SNP chip and resequencing) for genetic diversity studies, detection of selective sweeps, detection of CNV (using 50K chips)
• Interest in fibre production - collaboration between several countries China, Australia, Argentina..
• Seeking funding for resequencing of several breeds (with 20 individuals per breed including wild breed) and detect structural variations (mate-pair libraries)
• exon database for goat/RNAseq data (28 tissues of Cashmere goats and Angora)
Sheep

- Small number of localised reassembly patches
  - From BAC sequencing
  - Improving assembly prior to PacBio overlay
- Pacbio sequencing integrated with Illumina-based assembly
  - Oar v 4
    - Current gaps halved
    - Although low coverage should fill the unique sequence gaps as these are mainly short
    - Remaining gaps likely to be the big ones, especially if repeat rich in high copy number repeats
    - Most sequencing and assembly errors corrected
    - Very high quality protein coding gene annotations
    - Suitable for methylation studies
- Sheep genomes database
  - Many more genomes/exomes sequenced
# Pacbio sequencing

- Expecting better than
  - mean read length of 2.3 kb
  - half of the data is in reads longer than 4.2 kb
  - ten percent of the reads are longer than 6.2 kb
- Long reads definitely close gaps
- The better the base assembly the better the performance of PacBio in filling gaps

<table>
<thead>
<tr>
<th></th>
<th>D. pseudoobscura Fly</th>
<th>S. purpuratus Sea urchin</th>
<th>M. undulates Bird</th>
<th>R. norvegicus Rat</th>
<th>C. atys Monkey preliminary</th>
<th>C. atys Monkey complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Size</td>
<td>150 Mb</td>
<td>0.8 Gb</td>
<td>1.2 Gb</td>
<td>2.8 Gb</td>
<td>2.8 Gb</td>
<td>2.8 Gb</td>
</tr>
<tr>
<td>PacBio Coverage</td>
<td>24x</td>
<td>10.6x</td>
<td>5.1x</td>
<td>9.3x</td>
<td>6.8x</td>
<td>12.3x</td>
</tr>
<tr>
<td>Closed Gaps</td>
<td>68%</td>
<td>38%</td>
<td>21%</td>
<td>35%</td>
<td>64%</td>
<td>55%</td>
</tr>
<tr>
<td>Starting Gaps</td>
<td>6,010</td>
<td>142,764</td>
<td>49,376</td>
<td>109,263</td>
<td>186,841</td>
<td>115,698</td>
</tr>
<tr>
<td>Remaining Gaps</td>
<td>1,930</td>
<td>87,844</td>
<td>39,204</td>
<td>72,623</td>
<td>66,211</td>
<td>52,146</td>
</tr>
<tr>
<td>Contig N50 before</td>
<td>53 kb</td>
<td>14 kb</td>
<td>134 kb</td>
<td>60 kb</td>
<td>35 kb</td>
<td>57 kb</td>
</tr>
<tr>
<td>Contig N50 after</td>
<td>204 kb</td>
<td>24 kb</td>
<td>233 kb</td>
<td>118 kb</td>
<td>128 kb</td>
<td>166 kb</td>
</tr>
</tbody>
</table>
THE LONGER TERM
• Comparative genomics within and between species becomes a major force
  • Within a species, requires high quality reference genome and understanding of mechanisms of variation (CNVs etc.)
    – Pangeneome of sheep
  • Between species requires true differences >> assembly problems
• Linking phenotypes to molecular mechanisms
  • Predictive models based on biological processes
• The emergence of comparative systems biology within and between species
Funding sources

- International Science Linkages – Australian Government
- New Zealand Government
- Meat and Livestock Australia
- Australian Wool Innovations
- USDA
- BGI
- Genesis Faraday
- 3-SR
- UNCEIA
- Capgenes
- Apis-gene
- FarmIQ