Identification of sheep SNPs using Illumina sequencing and design of the Ovine SNP50 Beadchip

Brian Dalrymple on behalf of the ISGC
Three sources of SNPs used for the chip design

• Pilot 1,536 SNP chip SNPs
  • Sanger resequencing of regions amplified using primers designed from BAC-end sequences
    • Nine diverse animals
  • Validated using Illumina GoldenGate on 403 animals
    • >90% validation rate
  • 1142/1536 able to be positioned on sheep assembly v1.0
  • Use only validated SNPs
  • Aim to include as many of the sheep parentage panel as possible

• Illumina GA SNPs
  • 76,044 high confidence SNPs

• 454-FLX SNPs
  • ~270,000 high confidence SNPs called from a total of 3 X 454-FLX sequencing from six different animals
How did we discover the SNPs using the Illumina GA?
Reduced representational sequencing

- DNA pooled from 60 diverse animals – primarily female
- HaeIII digestion, 3 size fractions
  - 75-90 bp
  - 100-120 bp
  - 130-155 bp

Three fractions of genomic DNA
Sequence to ~20 x depth with Illumina GA (undersampled)
Illumina GA sequencing of samples

- 112 million reads from 3 Illumina GA runs
- 33 bases / read
- 3.7 Gb in total

<table>
<thead>
<tr>
<th>Total Reads</th>
<th>Non CC count</th>
<th>N Read Count</th>
<th>Ave qual &lt;25</th>
<th>Reads Passed</th>
<th>Different seqs</th>
<th>Non singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>112,075,672</td>
<td>22,965,875</td>
<td>830,817</td>
<td>8,308,027</td>
<td>84,675,240</td>
<td>9,048,701</td>
<td>2,942,112</td>
</tr>
</tbody>
</table>

- Total different sequences 0.3 Gb
- Different sequences represented more than once = 2.9 million
  - Singletons likely to be sequencing errors
- At 33 bases per read, just under 0.1 Gb non singleton different sequence
- Added back the GG to make 35 bases of sequence prior to aligning back the genome assembly

Reads discarded

The final working set
**Calling Illumina GA SNPs**

- Aligned all different non singleton sequences to the sheep genome assembly v1.0
- Identified all differences between overlapping Illumina GA reads
  - Due to the previous filtering all SNPs have two or more reads for each allele

**HaeIII fragments with sequences from both ends**

Number of sequences

**Solexa SNPs (A1 A3 B2 combined)**

**Solexa A1 A3 B2 Stocks**

Graphical representation of the alignment and identification of SNPs.
Excluded all cases with more than one putative SNP in overlapping Illumina GA reads
Filtering putative SNPs on quality score

- Required the putative SNP base to have a quality score of $\geq 27$ in at least one read for each allele.
- 76,044 SNPs called.

Illumina GA SNP

Discarded

Kept
Calling Illumina GA SNPs

- Number of SNPs depends on cut offs used
  - Reducing the cut offs significantly substantially increases the number of SNPs predicted
  - Prior to final quality cut off 149,405 putative SNPs had been called
- We decided to filter hard to maximise our confidence that our pool of putative SNPs for chip design were real SNPs
  - A similar philosophy had been applied to the 454 SNP calling to generate the set for inclusion in the chip design process
SNP distribution

- *HaeIII* sites are not randomly distributed in the sheep genome
  - perhaps not the best enzyme to have chosen!
- SNPs from Illumina GA sequencing follow *HaeIII* site distribution
- SNPs from 454-FLX sequencing more evenly distributed across the genome
  - more like the distribution of the sequence coverage
SNP validation

• So how likely are the SNPs called from the sequences to be real SNPs vs sequencing errors?
  • 64 randomly selected Illumina SNPs
  • 112 randomly selected 454 SNPs
  • tested on 63 DNA samples including those used for discovery and the International mapping flock using a Sequenom iPLEX system
  • Only two SNPs, one each type were not polymorphic
  • More stringently >80% passed QC (>85% genotype calls, HW equilibrium test, MAF>0.05 in the 63 selected animals)
  • Predicts > 85% validation on the Illumina Infinium system
    • Expect higher on the actual chip
  • Overall little difference between 454 and Illumina-derived SNPs
Designing the Ovine SNP50 beadchip
SNP chip design I

- **Spacing**
  - 60K SNPs evenly distributed across the genome
    - Ruled out favouring genes and genome regions very early on

- **Quality**
  - Assumed Validated Sanger 1536 SNPs > Illumina GA > 454-FLX
  - Probability that variation is a real SNP, not a sequencing artefact
    - However based on gut feeling rather than sound data for Illumina v. 454

- **Minor Allele Frequency**
  - Only known for Illumina GA and 1536 SNPs
    - But due to low coverage of sequencing a 454-FLX SNP is more likely to have a higher MAF
  - SNPs with MAF< 0.2 discarded

- **Use of chip real estate**
  - Infinium I v. II (SNPs)
    - Infinium II uses one position on chip v. two positions for Infinium I
    - Infinium I assay for AT and GC SNPs, ~17% of all SNPs
Not all SNPs can be converted to assays

- **Probe score**
  - Probability that assay will work on chip – we are using Illumina Infinium platform
    - All SNPs scored and those with a probe score < 0.8 discarded
SNP chip design II

- Region containing SNP should be unique
  - Checked using the 200 base flanking sequence used to design assay
  - But it is not a complete genome sequence
- Assay oligonucleotide should be unique in the genome
  - Rescreened assay oligonucleotides
  - But it is not a complete genome sequence
- Assay oligonucleotide should not contain other SNPs
  - Do not know all the SNPs that occur in the population of sheep
  - And it is not a complete genome sequence
Taking account of flanking SNPs

- All SNPs were positioned on the assembly
- All SNPs positioned within the assembly were replaced by their IUPAC codes
  - A/C → M, A/G → R, A/T → W
- Flanking sequence for each SNP was extracted from the assembly and were sent to Illumina for assay primer scoring
  - Illumina assay primer choice and scoring will reflect the existence of known SNPs and N’s in the flanking sequence
SNP Selection (SNPs in Range)

- From the last selected SNP:
  - The next SNP is selected within a specified range
  - The range is offset from the selected SNP by a specified step size
  - The “best” SNP is selected within this region
  - If there is more than one “best” SNP, the one closest to the center of the range is selected
- Parameters for step size and range were selected by titrating against the target number of SNP required for the chip design

<table>
<thead>
<tr>
<th>OAR1</th>
<th>Sanger 1536</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Illumina</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>454</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Last selected SNP
- Selectable SNPs
- Unselectable SNPs

60kb range
35kb step size
SNP Selection (SNPs not in Range)

- **From the last selected SNP:**
  - The next available SNP is located
  - The range is centered on this located SNP
  - The “best” SNP is selected within this region
  - If there is more than one “best” SNP, the one closest to the center of the range is selected
Selected SNPs

- 59388 total SNPs automatically selected
- 60717 total beads were filled
- SNPs manually added
  - 53 parentage SNPs (86/139 already selected)
    - 14 sequences had to be manually fixed to yield good Illumina scores
    - 1 sequence would not pass Illumina design
  - 16 SNPs had duplicate mappings to the genome, replaced with nearby SNPs
  - 15 mitochondrial SNPs added
  - Last 2 remaining SNPs were selected to lie within the largest remaining gaps in the genome assembly
The Ovine SNP50K beadchip final design

- Used all 60,800 available beads – 59,454 SNPs
- Mode spacing 35kb
  - still a few long gaps without any SNPs
- Average assay design score 0.975

<table>
<thead>
<tr>
<th>Source</th>
<th>Infinium I</th>
<th>Infinium II</th>
<th>Total</th>
<th>Percent of available SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger</td>
<td>29</td>
<td>571</td>
<td>600</td>
<td>43%</td>
</tr>
<tr>
<td>454</td>
<td>1,049</td>
<td>39,125</td>
<td>40,174</td>
<td>14%</td>
</tr>
<tr>
<td>Illumina</td>
<td>268</td>
<td>18,401</td>
<td>18,669</td>
<td>25%</td>
</tr>
<tr>
<td>mtDNA</td>
<td>0</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,346</td>
<td>58,108</td>
<td>59,454</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>2.26%</td>
<td>97.74%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Distribution of gaps between SNPs

- Step size

Spacing [bp]

Frequency

Range size
Preliminary Ovine SNP50 beadchip parameters

- ~55K SNPs assayable on the chip
  - In line with illumina standards
- Based on sheep Hapmap project using a very diverse range of animals
  - ~54K SNPs producing genotype calls
  - ~53K Polymorphic in at least one animal (>98%)
  - ~52K MAF >0.05
  - Average MAF 0.3
- Overall very similar parameters to the horse and cow chips
- Selecting for MAF >0.2 appears to have generated a chip with a significantly higher average SNP MAF than horse and cow more in the 0.2 range
  - Although there may be other reasons for this
The sheep Hapmap project sample set

<table>
<thead>
<tr>
<th>Sample Set</th>
<th>No Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic Sheep (64 breeds)</td>
<td></td>
</tr>
<tr>
<td>- Africa (6 breeds)</td>
<td>145</td>
</tr>
<tr>
<td>- Asia (9 breeds)</td>
<td>250</td>
</tr>
<tr>
<td>- Europe (29 breeds)</td>
<td>1796</td>
</tr>
<tr>
<td>- Middle East (7 breeds)</td>
<td>201</td>
</tr>
<tr>
<td>- South America (3 breeds)</td>
<td>98</td>
</tr>
<tr>
<td>- USA and the Caribbean (4 breeds)</td>
<td>248</td>
</tr>
<tr>
<td>- Australia and NZ (6 breeds)</td>
<td>153</td>
</tr>
<tr>
<td>total</td>
<td>2891</td>
</tr>
<tr>
<td>5 Species of Wild Sheep</td>
<td>122</td>
</tr>
<tr>
<td>9 Outgroup Species</td>
<td>52</td>
</tr>
<tr>
<td>Validation, Mapping, Parentage</td>
<td>283</td>
</tr>
<tr>
<td>Total</td>
<td>3348</td>
</tr>
</tbody>
</table>
Some other observations

- 1502 SNPs assigned to the X chromosome
  - Two blocks of atypical X clustering (PARs)
    - 5 SNPs in succession between 3.58 and 3.81 Mb
    - 87 SNPs in succession between 7.77 and 13.9 Mb
  - 11 other atypical X clustering SNPs sporadic across the X
- 35 SNPs with typical X clustering pattern, but not assigned to the X chromosome in genome build
  - 28 on previously unassigned fragments of genome
  - 7 on other chromosomes
- A number of other blocks with a significantly large number of successive problematic SNP
  - Some explanation in genome organisation?
Is a SNP is a SNP is a SNP?

<table>
<thead>
<tr>
<th></th>
<th>On final chip</th>
<th>zeroed</th>
<th>Intensity only</th>
<th>Nearby polymorphism/deletion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger</td>
<td>92.8%</td>
<td>0.35%</td>
<td>1.23%</td>
<td>8.11%</td>
</tr>
<tr>
<td>454</td>
<td>92.6%</td>
<td>1.5%</td>
<td>0.65%</td>
<td>6.98%</td>
</tr>
<tr>
<td>Illumina</td>
<td>92.2%</td>
<td>1.04%</td>
<td>0.51%</td>
<td>5.83%</td>
</tr>
</tbody>
</table>
And finally for now

- The number of SNPs with nearby polymorphisms that affect the assay in some way, around 6.6%, is higher than we expected, however many of these can still be used
  - Not clear if this is a downside of our sequencing, or SNP chip design approach, or some other issue
- The HapMap analysis is underway
- Hopefully some exciting association studies to come in the not too distant future.
## Genome and SNP chip sections of the International Sheep Genomics Consortium

- **AgResearch NZ**
  - John McEwan
  - Gemma Payne
  - Nessa O’Sullivan
  - Tracey Van Stijn
  - Theresa Wilson
  - Rudiger Brauning
  - Alan McCulloch
  - Russel Smithies
  - Benoit Auvray

- **University of Otago**
  - Jo-Ann Stanton
  - Chrissie
  - Mark

- **Illumina**
  - Marylin Munson
  - Kimberly Gietzen
  - Christian Haudenschild

- **Baylor College of Medicine**
  - Richard Gibbs
  - Donna Muzny
  - Michael E. Holder
  - Lynne Nazareth
  - Rebecca L. Thornton
  - Christie Kovar

- **CSIRO Livestock Industries**
  - Brian Dalrymple
  - James Kijas
  - David Townley
  - Abhirami Ratnakumar
  - Wes Barris
  - Sean McWilliam

- **Genesis Faraday**
  - Chris Warkup

- **The CORE snp center**
  - Roxann Ashworth

- **sheepGENOMICS**
  - Rob Forage
  - Terry Longhurst

- **TIGR**
  - Ewen Kirkness

- **Uni Melbourne**
  - Jill Maddox

- **USDA**
  - Tim Smith

- **UNE**
  - Hutton Oddy

- **Uni Sydney**
  - Frank Nicholas
  - Herman Raadsma

- **Utah State University**
  - Noelle Cockett
  - Chunhua Wu
Genome and SNP chip sections of the International Sheep Genomics Consortium

isgcdata.agresearch.co.nz
www.sheephapmap.org www.livestockgenomics.csiro.au