Identification of Genetic Variation Regulating Gene Expression in Dairy Cattle with RNA Sequence Data

Amanda Chamberlain
Introduction

• Variants in cis-regulatory regions can cause variation in gene expression between individuals and between tissues

• Many mutations affecting complex traits reside in regulatory regions of genome

• Such mutations have been shown to increase the accuracy of genomic predictions

• This study uses an RNA sequence dataset to map:
  – Expression QTL (eQTL) – change total gene expression
  – Allele specific expression QTL (aseQTL) – change the allelic imbalance of expression
  – Splice QTL (sQTL) – change isoform expression
3 Analyses

112 Holstein MC and WBC
29 Jersey MC (RIN>6)

RNA sequenced
(SureSelect Stranded + HiSeq3000)

Paired reads aligned (TopHat2, STAR)

Gene counts
(HTSeq-count, DESeq)

Allele counts
(samtools)

Exon/Intron counts
(FeatureCount, Leafcutter)

109 Holstein +
29 Jersey BovineSNP50

1000 bull genomes
sequence genotypes
(run 5)

Imputed whole genome sequence data
(Fimpute) Phased with Eagle

Association testing with SNP within 1MB

eQTL analysis
(EMMAX)

aseQTL analysis
(EMMAX)

sQTL analysis
(Matrix eQTL and leafcutter)

Gene expression
Allelic imbalance
 Isoform expression
Gene expression

edgeR MDS Plot

- Milk
- Blood

Leading logFC dim 1

Leading logFC dim 2
Results

<table>
<thead>
<tr>
<th></th>
<th>eQTL</th>
<th></th>
<th>aseQTL</th>
<th></th>
<th>sQTL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC</td>
<td>WBC</td>
<td>MC</td>
<td>WBC</td>
<td>MC</td>
<td>WBC</td>
</tr>
<tr>
<td>Variants tested</td>
<td>10.9M</td>
<td>10.4M</td>
<td>19.3M</td>
<td>17.8M</td>
<td>14.3M</td>
<td>14.3M</td>
</tr>
<tr>
<td>Genes/Positions/Exons</td>
<td>12,772</td>
<td>11,577</td>
<td>311,815</td>
<td>291,638</td>
<td>109,571</td>
<td>108,486</td>
</tr>
<tr>
<td>Sig variants (P&lt;1x10^{-4})</td>
<td>15,299</td>
<td>98,340</td>
<td>1.6M</td>
<td>2.9M</td>
<td>28,907</td>
<td>138,907</td>
</tr>
<tr>
<td>FDR</td>
<td>6%</td>
<td>0.8%</td>
<td>2.3%</td>
<td>1.7%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Genes</td>
<td>361</td>
<td>554</td>
<td>6,314</td>
<td>5,085</td>
<td>283</td>
<td>929</td>
</tr>
<tr>
<td>Gene overlap</td>
<td>60</td>
<td>3,701</td>
<td></td>
<td></td>
<td>107</td>
<td></td>
</tr>
</tbody>
</table>
A

SNP effect on expression of gene in milk (log transformed read counts)

SNP effect on expression of gene in blood (log transformed read counts)

\[ y = 0.26x + 0.01 \]
\[ R^2 = 0.39 \]

B

eQTL effect

\[ y = 0.6226x - 0.2579 \]
\[ R^2 = 0.288 \]

C

sQTL effect value in blood

sQTN effect value in milk

\[ r = 0.676 \]
\[ p < 2.2e-16 \]

D

White blood cells

Milk cells

Percentage of significant SNP

Distance from TSS (kb)
sQTL in milk and blood

https://www.biorxiv.org/content/early/2017/11/16/220251
Chr6:87392580, within the 5th exon & the splice site
Allele effects

- A allele: t value on 5th exon expression ratio
  - 0

- B allele: t value on 5th exon expression ratio
  - 4

Allele frequency

- Holstein:
  - A allele frequency: 0.2
  - B allele frequency: 0.8

- Jersey:
  - A allele frequency: 0.8
  - B allele frequency: 0.2
Bovine topologically associated domains

Map TADs from human, mouse and dog to cattle

Dixon et al. (2012)

https://www.biorxiv.org/content/early/2018/01/04/242792
## Input TAD dataset from human ESC, Significance threshold $p \leq 10^{-8}$

<table>
<thead>
<tr>
<th>QTL type</th>
<th>Cell type</th>
<th># significant SNP</th>
<th># within target gene TAD</th>
<th># expected</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>aseQTL</td>
<td>WBC</td>
<td>13302</td>
<td>6375</td>
<td>0.0013302</td>
<td>0.0000%</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>9553</td>
<td>3862</td>
<td>0.0009553</td>
<td>0.0000%</td>
</tr>
<tr>
<td>eQTL</td>
<td>WBC</td>
<td>327</td>
<td>114</td>
<td>0.0000327</td>
<td>0.0000%</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>343</td>
<td>24</td>
<td>0.0000343</td>
<td>0.0001%</td>
</tr>
</tbody>
</table>

Same result for all 6 input TAD datasets and significance levels from $p\leq10^{-5}$ to $p\leq10^{-8}$
Conclusions

- There are many eQTL, aseQTL and sQTL that affect total gene expression, allelic imbalance or isoform expression.
- Many occur in both cell types, with similar effect.
- Most occur within 100 kb of TSS, but some are up to 1Mb away.
- Most occur within the same bovine TAD as their gene targets.
- TADs provide a better search space than an arbitrary distance from the gene.
- Many of the potential regulatory variants are now being genotyped in large populations to determine their affect on increasing the accuracy of genomic prediction.
Acknowledgements

Sample collection, genotyping and library making
Dr Coralie Reich, Brett Mason, Claire Prowse Wilkins, Josie Garner, Meaghan Douglas, Dr Amanda Chamberlain

Sequencing
Dr Tracie Webster

Sequence imputation
Prof Ben Hayes, Dr Iona Macleod

aseQTL, eQTL and sQTL analysis
Prof Ben Hayes, Dr Ruidong Xiang, Dr Christy Vander Jagt, Dr Amanda Chamberlain, Prof Mike Goddard

TAD analysis
Min Wang, Dr Tim Hancock, Prof Ben Hayes, Dr Amanda Chamberlain, Prof Mike Goddard

Funding
Dairy Futures CRC, Dairy Bio