Genomic prediction and GWAS with sequence information versus HD or 50k SNP chips

Roel Veerkamp, Rianne van Binsbergen, Mario Calus, Chris Schrooten, Aniek Bouwman
Background

- Whole genome sequence data expected to perform better in GWAS and GS, more persistent across generations / breeds
  - Causal mutation (QTN) is included
  - No dependency on LD between SNP and QTL
Identifying QTN with GS?

Prediction reliability

- Pedigree
- BLUP
- BovineHD
- BSSVS
- Sequence
- BSSVS

HD SNP: 631,428
SEQ SNP: 12,590,056

Genomic Prediction with 12.5 Million SNPs for 5503 Holstein Friesian Bulls

Objective

- The potential benefit of using sequence data for
  - QTL detection
  i.r.t.
  - Explaining total genetic variation
  - Prediction accuracy GS
Method (1): Imputation to sequence

5556 Holstein Friesian bulls CRV
777K SNP genotypes (Illumina BovineHD BeadChip)

1147 animals (multiple breeds)
28M SNP (whole-genome sequence data)

1000 bull genomes project Run 4

5556 Holstein Friesian bulls with phenotype (PY) and imputed sequence

3469 bulls used for discovery and training & 2287 bulls used for validation

Minimac v2

Aniek Bouwman
Method (2): statistical models

- **GWAS using single SNP regression (GCTA)**
  - GWAS using single SNP regression
  - Include GRM based on HD SNP set
  - MAF >0.01

- **Conditional and joint GWAS (COJO)**
  - Stepwise selection of SNP explaining additional variance
Method (3): Two validation methods

SNP set selection from GWAS:

1. Estimate heritability in validation animals using GRM based on selected sets of SNP
2. Train GRM (based on selected SNP sets) on discovery animals, back solve SNP and predict DGV for 2287 validation animals. Correlate DGV with phenotypes.
Results GWAS: 50K
Results GWAS : HD
Results GWAS Sequence + cojo5
Results GWAS: Cojo5 on Chr14 (DGAT)
### Results: SNP set selection GWAS

Number of SNP selected in the different SNP sets:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>HD</th>
<th>50k</th>
<th>COJO</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>13,789,029</td>
<td>656,044</td>
<td>49,580</td>
</tr>
<tr>
<td>-log(p)&gt;3</td>
<td>24,387</td>
<td>1,238</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>119</td>
</tr>
<tr>
<td>-log(p)&gt;5</td>
<td>2,194</td>
<td>159</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>49</td>
</tr>
</tbody>
</table>
## Results: Heritability

$h^2$, phenotypic variance explained by GRM using selected SNP sets

<table>
<thead>
<tr>
<th>GRMs</th>
<th>Sequence</th>
<th>HD</th>
<th>50k</th>
<th>COJO</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.83</td>
<td>0.82</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>-log(p)&gt;3</td>
<td>0.53</td>
<td>0.40</td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>-log(p)&gt;5</td>
<td>0.60*</td>
<td>0.43*</td>
<td>0.22*</td>
<td>0.16</td>
</tr>
</tbody>
</table>

*Scale problems with GRM when estimating variances*
## Results: Genomic prediction

Correlation between genomic breeding value and phenotype

<table>
<thead>
<tr>
<th>GRMs</th>
<th>Sequence</th>
<th>HD</th>
<th>50k</th>
<th>COJO</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>-log(p)&gt;3</td>
<td>0.58</td>
<td>0.56</td>
<td>0.42</td>
<td>0.38</td>
</tr>
<tr>
<td>-log(p)&gt;5</td>
<td>0.39</td>
<td>0.30</td>
<td>0.28</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Conclusions

- Simply using sequence within Holstein population, unlikely to improve GS, but helps QTL detection.

- Subsets of selected SNP always poorer $h^2$ and GS
  - Full seq. accuracy GS of 0.68 and $h^2 = 0.83$
  - 51 SNPs accuracy GS of 0.31 and $h^2 \approx 0.16$ (DGAT!)

- Good way to get realistic expectations from QTL.
Acknowledgements

1000 bull genomes consortium
www.1000bullgenomes.com