The use of multi-breed reference populations and multi-omic data to maximize accuracy of genomic prediction

M. E. Goddard¹,², I.M. MacLeod², K.E. Kemper³, R. Xiang¹, I. Van den Berg¹, M. Khansefid², H. D. Daetwyler² & B.J. Hayes⁴

¹Faculty of Veterinary & Agricultural Science, University of Melbourne, ²Agriculture Victoria, Bundoora, ³Institute for Molecular Bioscience, and ⁴Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, QLD 4067, Australia.
This talk

Introduction

Do QTL segregate across breeds?

Why are multi-breed GEBVs hard?

Solutions
Introduction

GEBV accuracy is low if
reference population is small, or
target populations is distantly related to training population

Training populations within breed are too small
numerically small breed
hard to measure traits eg FCE

Therefore, use multi-breed training population

Training on a different breed to target $\rightarrow$ low accuracy

Aim = Accurate GEBVs for a breed with a small training population based on a multi-breed training population
Do QTL segregate across breeds? (Kath Kemper)
Do QTL segregate across breeds?

Across 11 QTL, length of conserved haplotype (0.4kb-55kb) around mutation suggest age of QTL mutations varies ~ 2,000 to 50,000 generations old

Prior to breed formation

QTL can and do segregate across breeds, although drift and selection can result in fixation.
Age of myostatin mutations (50 – 10 gen) (O’Rourke et al)
Why are multi-breed GEBVs hard?

SNP x breed interactions
  differences in LD phase between breeds
QTL x breed interactions
  Due to non-additive gene action
    typically small variances
  equivalent to sire x breed interactions
    typically small
  Low accuracy even in simulation

Differences in allele frequency
  $F_{ST}$ is low
  QTL segregate across breeds
Why are multi-breed GEBVs hard?

LD phase differs between breeds

Within breed GEBVs estimate the effect of large chromosome segments

This works due to LD within a breed

Effective number of chromosome segments = 5000

That is, segments 600 kb long
Why are multi-breed GEBVs hard?

Within breed GEBVs estimate the effect of large chromosome segments

This works due to LD within a breed

Effective number of chromosome segments = 5000

That is, segments 600 kb long

Across breeds conserved segments are much smaller (x10 smaller)
Solutions

Increase size of training population

Include target breed in training population
Aussie Reds

Holstein 4000 bulls, 10023 cows

Jersey 1044 bulls, 4232 cows

Aussie Reds 114 Bulls

Real or imputed 630K SNP for all individuals
Accuracy of Bayes R (Irene van den Berg)
Solutions

Increase size of training population

Include target breed in training population

Use denser SNP panels or sequence
Variance explained by SNPs and sequence (Iona Macleod)

Proportion of Total Genetic Variance Explained by SNP and Pedigree: BayesR (Mixed Hol & Jer)

- % Genetic Var - SNP
- % Genetic Var - Ped

- Temperament
- Stature
- Milk Yield
- Protein Yield

Aust Bull & Cow: Holstein & Jersey
Danz Bulls Only: Holstein & Jersey
Harnessing the power of whole-genome sequence: first global report of improved genomic prediction accuracy using sequence data in sheep

Iona MacLeod, Bolormaa Sunduimijid, Majid Khansefid, Andrew Swan, Julius van der Werf & Hans Daetwyler
Validation sets - low relationships with Ref.:
1. Merino
2. Merino x Border Leicester F1
GWAS – Carcass Fat Depth (ccfat)
Meat Traits:

GBLUP Accuracy - Merino x Border Leicester

![Graph showing prediction accuracy for different traits (ccfat, cemd, imf, pemd, sf5, pwt) comparing 50K and 50K+Top Seq (2) with different bars for each trait.]
Solutions

Increase size of training population

Include target breed in training population

Use denser SNP panels or sequence

Use Bayesian statistical method not GBLUP
Accuracy $r(DGV, DTD)$ in Aussie Red Bulls

(Iona MacLeod)
Wool Traits:
Prediction Accuracy in Merinos

Prediction Accuracy

- **TRAIT**
  - Breech Wrinkle
  - Clean Fleece Wt
  - Fibre Diameter

- **Methods**
  - 50K
  - BayesR 50K+Top Seq
  - BayesRC 50K+Top Seq

Graph showing prediction accuracy for different traits and methods.
BayesR vs BLUP (BTA11)

O = BayesR
O = GBLUP
Solutions

Increase size of training population

Include target breed in training population

Use denser SNP panels or sequence

Use Bayesian statistical method not GBLUP

Use multiple traits
Multi-trait GWAS (Ruidong Xiang)

Previously reported calving difficulty locus
(Purfield et al 2015)
Validation of lead pleiotropic SNPs
(Ruidong Xiang)

Select 21 lead pleiotropic SNPs and confirmed by conditional analysis in bulls

Linear index validation of lead pleiotropic SNPs in cows:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>SNPs no.</th>
<th>SNP no. with the same effect directions</th>
<th>Percent</th>
<th>SNPs no. P&lt;0.05 in validation GWAS</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>21</td>
<td>21</td>
<td>100%</td>
<td>17</td>
<td>81%</td>
</tr>
<tr>
<td>PC</td>
<td>21</td>
<td>21</td>
<td>100%</td>
<td>18</td>
<td>86%</td>
</tr>
<tr>
<td>CT</td>
<td>21</td>
<td>21</td>
<td>100%</td>
<td>17</td>
<td>81%</td>
</tr>
</tbody>
</table>
The effects of lead SNPs across independent traits

Cluster of SNPs

Effect of SNPs

Graph showing the correlation and effects of SNPs across different traits.
Solutions

- Increase size of training population
- Include target breed in training population
- Use denser SNP panels or sequence
- Use Bayesian statistical method not GBLUP
- Use multiple traits
- Use gene expression
Number of cis eQTL in cattle (Ben Hayes)

<table>
<thead>
<tr>
<th>-log10Pvalue</th>
<th>Milk</th>
<th>Blood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNP</td>
<td>FDR</td>
</tr>
<tr>
<td>1</td>
<td>10,019,870</td>
<td>0.958</td>
</tr>
<tr>
<td>2</td>
<td>1,150,197</td>
<td>0.835</td>
</tr>
<tr>
<td>3</td>
<td>173,662</td>
<td>0.553</td>
</tr>
<tr>
<td>4</td>
<td>40,601</td>
<td>0.237</td>
</tr>
<tr>
<td>5</td>
<td>15,299</td>
<td>0.063</td>
</tr>
<tr>
<td>6</td>
<td>6,831</td>
<td>0.014</td>
</tr>
<tr>
<td>7</td>
<td>3,340</td>
<td>0.003</td>
</tr>
<tr>
<td>8</td>
<td>2,201</td>
<td>0.000</td>
</tr>
</tbody>
</table>
eQTL and QTL (meat quality) comparison within 50kb of calpastatin (Majid Khansefid)
eQTL and QTL (meat quality, PW hip height and multi-trait) overlap
<table>
<thead>
<tr>
<th></th>
<th>Effect</th>
<th>P-value</th>
<th>Prop. $\sigma^2_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Additional traits</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phosphorus conc.</td>
<td>41.8</td>
<td>$1.10 \times 10^{-11}$</td>
<td>0.107</td>
</tr>
<tr>
<td>eSLC37A1</td>
<td>0.160</td>
<td>$3.55 \times 10^{-18}$</td>
<td>0.224</td>
</tr>
<tr>
<td><strong>Key production trait, milk yield</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>milk yield – Holstein cows</td>
<td>-37.6</td>
<td>$2.19 \times 10^{-3}$</td>
<td>0.001</td>
</tr>
<tr>
<td>milk yield – Holstein bulls</td>
<td>-40.3</td>
<td>$3.17 \times 10^{-3}$</td>
<td>0.003</td>
</tr>
<tr>
<td>milk yield – Jersey cows</td>
<td>-45.2</td>
<td>$3.26 \times 10^{-3}$</td>
<td>0.002</td>
</tr>
</tbody>
</table>

That is the allele that *increases* expression of SLC27A1 (an antiporter):

1. *Increases* phosphorus concentration
2. *Decreases* milk yield

(Kemper et al)
Solutions

Gene expression data
  gene cis eQTL
  splicing cis eQTL
  exon cis eQTL
Phenotypic differences due to splicing

• Human Tau gene splicing related to the Alzheimer’s disease

• Many genome variants affecting gene splicing, sQTL contribute to human diseases

Li et al., 2016
Overlap between eQTL and milk QTL (Ruidong Xiang)
Example: FUK, chr 18, fat yield (Irene van den Berg)

Local GEBV variance

Correlation local GEBV & gene expression
Solutions

Include target breed in training population

Use denser SNP panels or sequence

Use Bayesian statistical method not GBLUP

Multi-trait analysis e.g. gene expression data

Use functional annotation of genome
SNP effects at cellular level

- Quantify the impact of a mutation on gene expression levels

Schaub et al. 2012
Genomic prediction – Milk (Iona MacLeod)

- BayesR

<table>
<thead>
<tr>
<th>Total SNP</th>
<th>Zero</th>
<th>Tiny</th>
<th>Small</th>
<th>Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>905,813</td>
<td>99.3%</td>
<td>0.69%</td>
<td>0.004%</td>
<td>0.001%</td>
</tr>
</tbody>
</table>

- BayesRC

<table>
<thead>
<tr>
<th>SNP Class</th>
<th>No. SNP</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lact genes + NSC</td>
<td>3768 (0.4%)</td>
<td>95.0%</td>
<td>4.3%</td>
<td>0.58%</td>
<td>0.12%</td>
<td>11%</td>
</tr>
<tr>
<td>Lact other</td>
<td>57722 (6%)</td>
<td>99.3%</td>
<td>0.7%</td>
<td>0.05%</td>
<td>0.004%</td>
<td>12%</td>
</tr>
<tr>
<td>All others</td>
<td>847905</td>
<td>99.5%</td>
<td>0.5%</td>
<td>0.01%</td>
<td>0.000%</td>
<td>77%</td>
</tr>
</tbody>
</table>
Cattle stature (Aniek Bouwman, Ben Hayes et al)

<table>
<thead>
<tr>
<th>Annotation class</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>intergenic_variant</td>
<td>83</td>
</tr>
<tr>
<td>upstream_gene_variant</td>
<td>11</td>
</tr>
<tr>
<td>5_prime_UTR_variant</td>
<td>1</td>
</tr>
<tr>
<td>intron_variant</td>
<td>55</td>
</tr>
<tr>
<td>missense_variant</td>
<td>5</td>
</tr>
<tr>
<td>downstream_gene_variant</td>
<td>8</td>
</tr>
<tr>
<td>ChiP-SEQ peaks*</td>
<td>8</td>
</tr>
<tr>
<td>WBC eQTL</td>
<td>10</td>
</tr>
</tbody>
</table>
The bad news

Accuracy only improves a little

You need to capture a high proportion of total variance
Conclusion

Data from the target breed is the most useful

But, training data from other breeds helps

Advantage to use sequence data and Bayesian method

Sequence imputation loses accuracy

Identify near perfect markers and genotype them directly

Expression data and functional annotation helps select best variants