Comparative assessment of methods for estimating genomic relationships and their use in predictions in an admixed population

M.L. Makgahlela¹,², I. Strandén¹,², U.S. Nielsen³, M.J. Sillanpää¹,⁴, J. Juga¹ & E.A. Mäntysaari²

¹University of Helsinki, Finland
²MTT Agrifood Research Finland, 31600 Jokioinen
³Danish Agricultural Advisory Service, Udkaersvej 15, Denmark
⁴University of Oulu, Finland
Introduction

- Accurate estimation of relationships between animals is an important step in any routine genetic evaluations

- Relationships were previously based on pedigree information only

- Conversely, most current evaluations use both marker-derived relationship matrix (\(G\)) and pedigree-based relationships (\(A\))

- \(G\) estimators are more accurate than \(A\) because they have more variation between closely related individuals
Introduction

- The accuracy of $G$ estimators may be even higher
  - If founder population allele frequencies were available

- In the absence, current population allele frequencies are used to make $G$ and that defines the founder population

- The use of observed allele frequencies in structured populations however, may lead to biased estimation of $G$
Objectives

- To estimate $A$ and $G$ matrices
  - Different $G$ matrices were estimated using either observed allele frequencies across breeds or breed allele means

- To estimate breeding values (EBV) and direct genomic values (DGV) using different $G$ matrices
  - Estimated coefficients and their respective DGVs were compared
The population

- The Nordic Red dairy cattle (RDC) is a combined population
  - 3 sub-populations from DNK, SWE & FIN
  - 2nd largest breeding population, with $Ne$ larger than Holsteins
  - Most animals in the data (~98%) are composites of breeds
- Absence of pure breed animals remains a major limiting factor for the estimation of breed-specific allele frequencies
Materials and Methods

- Data were genotypes of 38194 SNP markers for 4106 bulls
- Breed proportions for bulls were estimated from the full Nordic RDC pedigree (>4m animals)
  - 3 main breeds defined with mean BP>10% were,
    - SRB, FAY & NRF
    - Remaining breeds with mean BP<10% -> breed “OTHER”
- Phenotypes were cow IDDs for milk, protein & fat, from 2010 NAV routine evaluations
Estimation of relationships

- Genomic relationships (\(G\)) were estimated following methods 1 and 2 by VanRaden (2008)

- \(G\) estimated using observed allele frequencies (\(GOF\))

\[
\text{GOF} = \frac{\text{ZZ}'}{k}
\]

- \(Z_{i,j} \leftarrow (0 - 2p_j); (1 - 2p_j); (2 - 2p_j)\)

- \(p_j\) is the frequency for the 2\(^{nd}\) allele & \(k = 2\sum_j p_j(1 - p_j)\)

Number of “second” alleles
Estimation of relationships

- **G** matrices estimated using *breed allele means* (**GBM** and **GBM2**)

\[
\text{GBM} = \frac{\text{MM}'}{k}
\]

- \(M_{i,j} \leftarrow (0 - 2p_{ij}); (1 - 2p_{ij}); (2 - 2p_{ij})\)

- \(p_{ij}\) is the expected allele frequency of marker \(j\) for bull \(i\) given it’s base breed proportions

✓ computed by multiple regression of genotypes on BP
Estimation of relationships

- Modification of VanRaden method II
- There,

\[ G_{PvRII} = ZDZ' / m = ZD^{0.5} D^{0.5} Z' / m \]

- \( m \) is the number of markers
Estimation of relationships

Following the same, we define:

\[ \text{GBM2} = \frac{M^* M^{*'}}{m} \]

\[ M_{i,j}^* \leftarrow \frac{-2p_{ij}}{\sqrt{2p_{ij}(1-p_{ij})}}, \frac{1-2p_{ij}}{\sqrt{2p_{ij}(1-p_{ij})}}, \frac{2-2p_{ij}}{\sqrt{2p_{ij}(1-p_{ij})}} \]

- \( m \) is the number of markers
Pedigree relationships (A) were estimated for genotyped bulls only, using RelaX2 computer program.

GOF and GBM2 were combined with 20% weight on A to yield GAOF and GABM2.

\[ G^* = wG + (1-w)A \]
Variance components, EBVs & DGVs were estimated separately for each matrix, using a GBLUP model.

\[ y = Xb + Za + e, \]

- \( y \) is a vector of cow IDD
- \( X \) and \( Z \) are design matrices allocating records to \( b \) and \( a \)
- \( b \) is a vector of fixed mean and breed regression effects
- \( a \) is a vector of breeding values
- \( e \) is a vector of residuals

Breed regression effects were used only for predictions with GBM and GABM2.

✓ Predicted values included fixed regression solutions
Statistics of (diagonals)-1 by estimator

<table>
<thead>
<tr>
<th></th>
<th>Across populations</th>
<th>Within Swedish bulls</th>
<th>Within Danish bulls</th>
<th>Within Finnish bulls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>0.012</td>
<td>0.000</td>
<td>0.135</td>
<td>0.008</td>
</tr>
<tr>
<td>GOF</td>
<td>0.019</td>
<td>-0.129</td>
<td>0.379</td>
<td>0.006</td>
</tr>
<tr>
<td>GBM</td>
<td>-0.051</td>
<td>-0.254</td>
<td>0.310</td>
<td>-0.043</td>
</tr>
<tr>
<td>GBM2</td>
<td>-0.242</td>
<td>-0.387</td>
<td>0.093</td>
<td>-0.238</td>
</tr>
</tbody>
</table>
Histograms of (diagonals)-1 for A and G matrices

Across populations

Within Danish population

Within Swedish population

Within Finnish population
Correlations between A and G matrices for pair-wise relationships

<table>
<thead>
<tr>
<th></th>
<th>Across populations</th>
<th>Within Swedish bulls</th>
<th>Within Danish bulls</th>
<th>Within Finnish bulls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.702</td>
<td>0.661</td>
<td>0.789</td>
<td>0.781</td>
</tr>
<tr>
<td>GOF</td>
<td>0.537</td>
<td></td>
<td></td>
<td>0.784</td>
</tr>
<tr>
<td>GBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.644</td>
<td>0.856</td>
<td>0.819</td>
<td>0.759</td>
</tr>
<tr>
<td>GOF</td>
<td>0.625</td>
<td></td>
<td></td>
<td>0.876</td>
</tr>
<tr>
<td>GBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Correlations between EBV & DGV from different estimators for validation bulls

<table>
<thead>
<tr>
<th></th>
<th>GOF</th>
<th>GBM</th>
<th>GBM2</th>
<th>GAOF</th>
<th>GABM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.67</td>
<td>0.67</td>
<td>0.66</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>GOF</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>GBM</td>
<td></td>
<td>1.00</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>GBM2</td>
<td></td>
<td></td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>GAOF</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

EBV = i.e. Parent average
Conclusions

- The use of simple observed allele frequencies across breeds over-estimate values in $G$ for:
  - Populations with the least number of animals in the combined data and/or,
  - Individuals from distantly related populations

- Estimated breed allele means reduced country differences in coefficients, similarly, but shifted them too much towards zero or less
The prediction of DGV converged to similar solutions regardless of allele frequencies used.

- Inclusion of breed regressions for GBM & GBM2 brought breed means back into the DGV.

The validation accuracy slightly increased when A and G matrices were combined.

- A single-step GBM2 and A including non-genotyped animals could increase the prediction of DGV even more.
😊 Thank you for your attention
Questions !!!