How do imputation errors affect genomic breeding values?

Eduardo Pimentel, C. Edel, R. Emmerling, K.-U. Götz

Institute of Animal Breeding
Motivation / Objective

- Most studies about the effects of imputation report:
  - overall correlations between GEBV
  - comparisons between software

- Further investigate the causes and patterns underlying the bias in GEBV due to imputation errors
Material and Methods

- DEA-System, December 2013
  - 3494 BSW candidates 50k

- Data set 1
  - 3494 animals with 6k

- Data set 2
  - 3494 animals with 50k

masked

Routine
Material and Methods

- Data set 1
  - 3494 animals with 6k
  - findhap / FImpute

- Data set 1
  - 3494 animals with 50k

- Data set 2
  - 3494 animals with 50k

- Prediction of GEBV for 37 traits

- Comparison between GEBV from observed 50k with GEBV from imputed 50k
### Changes in ranking within TOP 50 candidates

<table>
<thead>
<tr>
<th>Trait</th>
<th>Rank correlation</th>
<th>Also top 50 in imputed set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>findhap</td>
<td>FlImpute</td>
</tr>
<tr>
<td>Milk (kg)</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>Fat (kg)</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>Protein (kg)</td>
<td>0.82</td>
<td>0.91</td>
</tr>
<tr>
<td>SCS</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td>Workability</td>
<td>0.71</td>
<td>0.88</td>
</tr>
<tr>
<td>Udder depth</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>Feet &amp; legs</td>
<td>0.89</td>
<td>0.93</td>
</tr>
<tr>
<td>Udder</td>
<td>0.80</td>
<td>0.84</td>
</tr>
<tr>
<td>Overall score</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>Average (n=37)</td>
<td>0.84</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Impact of Imputation errors on GEBV

Protein (kg) / FImpute

All Candidates

$r = 0.99$
Impact of Imputation errors on GEBV

Protein (kg) / Flimpute

All Candidates

$r = 0.99$

TOP 50 Candidates

$r = 0.91$
Impact of Imputation errors on GEBV

Protein (kg) / FImpute

All Candidates

GEBV from 50k vs GEBV from imputed 50k

r = 0.99

TOP 50 Candidates

r = 0.91

BOTTOM 50 Candidates

r = 0.93
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.
A possible explanation

TOP 50 Candidates

Mean proportion of changes for the most frequent allele

Proportion of all changed alleles (FImpute)

One bar for each of the 37 traits
A possible explanation

FImpute

TOP 50 Candidates

One bar for each of the 37 traits

Mean proportion of changes for the most frequent allele

Proportion of all changed alleles

0.40 0.45 0.50 0.55 0.60 0.65 0.70

BOTTOM 50 Candidates

One bar for each of the 37 traits

Proportion of all changed alleles

0.40 0.45 0.50 0.55 0.60 0.65 0.70

findhap

TOP 50 Candidates

One bar for each of the 37 traits

Proportion of all changed alleles

0.40 0.45 0.50 0.55 0.60 0.65 0.70

BOTTOM 50 Candidates

One bar for each of the 37 traits

Proportion of all changed alleles

0.40 0.45 0.50 0.55 0.60 0.65 0.70
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.
A possible explanation

Assuming Hardy-Weinberg equilibrium:

\[
\begin{align*}
\text{f}(A_1A_1) &= p^2 \\
\text{f}(A_1A_2) &= 2pq \\
\text{f}(A_2A_2) &= q^2
\end{align*}
\]
A possible explanation

Assuming Hardy-Weinberg equilibrium:

\[
\begin{align*}
f(A_1A_1) &= p^2 \\
f(A_1A_2) &= 2pq \\
f(A_2A_2) &= q^2
\end{align*}
\]

**Genotype**

<table>
<thead>
<tr>
<th></th>
<th>$A_2A_2$</th>
<th>$A_1A_2$</th>
<th>$A_1A_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypic value</td>
<td>-a</td>
<td>0</td>
<td>+a</td>
</tr>
</tbody>
</table>

Population mean: \( M = a(p - q) + 2pqd \)

*Average effect of $A_1$: \( \alpha_1 = q[a + d(q - p)] \)

*Average effect of $A_2$: \( \alpha_2 = -p[a + d(q - p)] \)

*as deviation from the population mean*
A possible explanation

Assuming Hardy-Weinberg equilibrium:
\[
\begin{align*}
    f(A_1A_1) &= p^2 \\
    f(A_1A_2) &= 2pq \\
    f(A_2A_2) &= q^2
\end{align*}
\]

Population mean: \( M = a(p - q) + 2pqd \)

*Average effect of \( A_1 \): \( \alpha_1 = q[a + d(q - p)] \)

*Average effect of \( A_2 \): \( \alpha_2 = -p[a + d(q - p)] \)

*as deviation from the population mean
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, **bottom animals benefit and top animals are penalized.**
A possible explanation

TOP 50 Candidates

- Mean proportion of changes for an allele with a positive effect on the trait
- Mean proportion of changes for an allele with a negative effect on the trait

Proportion of all changed alleles (PipeR)

One red and one blue bar for each of the 37 traits
A possible explanation
A possible explanation

**FImpute**

TOP 50 Candidates

Mean proportion of changes for an allele with a positive effect on the trait
Mean proportion of changes for an allele with a negative effect on the trait

Proportion of all changed alleles

One red and one blue bar for each of the 37 traits

**findhap**

TOP 50 Candidates

Proportion of all changed alleles

One red and one blue bar for each of the 37 traits

**BOTTOM 50 Candidates**

Proportion of all changed alleles

One red and one blue bar for each of the 37 traits

**BOTTOM 50 Candidates**

Proportion of all changed alleles

One red and one blue bar for each of the 37 traits
Conclusions

- Bias in GEBV due to imputation errors
  - downwards in top and upwards in bottom segment
Conclusions

- Bias in GEBV due to imputation errors
  - downwards in top and upwards in bottom segment
- Imputation algorithms usually suggest haplotypes with higher frequency and more neutral effects
  - disadvantage for top and advantage for bottom animals
Conclusions

- Bias in GEBV due to imputation errors
  - downwards in top and upwards in bottom segment
- Imputation algorithms usually suggest haplotypes with higher frequency and more neutral effects
  - disadvantage for top and advantage for bottom animals
- Might have implications, especially for mixed pools of candidates genotyped at different densities
Aknowledgements

- Paul VanRaden (findhap)
- Mehdi Sargolzaei (FImpute)
- interOgenomics
Thanks!