Can multi-subpopulation reference sets improve the genomic predictive ability for pigs?

Anna Fangmann¹, Sarah Bergfelder-Drüing², Ernst Tholen², Henner Simianer¹, Malena Erbe¹,³

¹Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University Goettingen, Germany
²Institute of Animal Science, Group of Animal Breeding and Husbandry, University of Bonn, Germany
³Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, Poing-Grub, Germany
Introduction

Cattle breeding
• Implementation of genomic prediction successful

Pig breeding
• Possible advantage of genomic prediction: increasing the accuracy of breeding values at the time point of selection
• For decades: separate breeding work of different pig breeding organizations in Germany, Switzerland and Austria
  → stratified subpopulations within breed German Large White
• Limiting factor: size of the training set within a breeding organization
Aim of this study

- Evaluation of a genomic breeding value prediction in the breed German Large White for the trait ‘number of piglets born alive’

- Assessment of the usefulness of multi-subpopulation reference sets based on data from different commercial pig breeding organizations

http://www.bayerfarm.de/static/media/images/upload/2_schwein.jpg
Material and Methods: Data

- Data from individuals of five different commercial pig breeding organizations → different subpopulations

- 2’251 individuals genotyped with Illumina Porcine 60k SNP Chip

- Conventional breeding values for ‘number of piglets born alive’ (NBA) → deregressed following Garrick et al. (2009)
• Quality control: Callrate per SNP > 97 %
  Callrate per individual > 98 %
  > 10 observations of an allele per marker

→ Finally: 2’053 individuals with 46’064 SNPs

• Genotypic data:

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>Total number of animals</th>
<th>Born between</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>187</td>
<td>2002 – 2011</td>
</tr>
<tr>
<td>2</td>
<td>140</td>
<td>1997 – 2011</td>
</tr>
<tr>
<td>3</td>
<td>155</td>
<td>2001 – 2011</td>
</tr>
<tr>
<td>4</td>
<td>821</td>
<td>1993 – 2011</td>
</tr>
<tr>
<td>5</td>
<td>540</td>
<td>2002 – 2011</td>
</tr>
</tbody>
</table>
Material and Methods: Subpopulation stratification

Assessment based on principal component analysis and calculated $F_{ST}$ values between subpopulation1 and another subpopulations
Material and Methods: Subpopulation stratification

Multi-subpopulation reference sets for validation set:

Subpopulation 1

**close**

**distant**
Material and Methods: GBLUP model

Genomic Predictions with ASReml (Gilmour et al., 2009):

\[ y = Xb + Wg + e \]

- \( y \) = vector of DRPs for NBA
- \( X \) = design matrix for fixed effects
- \( b \) = vector containing the fixed effects
  - a) within subpopulation: overall mean
  - b) multi-subpopulation: general mean and subpopulation
- \( W \) = design matrix for the random genomic effects
- \( g \) = vector of random genomic effects (DGV)
- \( e \) = vector of random residual effects

with \( g \sim N(0, G_x \sigma^2_g) \) and \( G_x \) = Genomic relationship matrix according to different approaches
Material and Methods: random five-fold cross validation

Assessment of predictive ability of DGV prediction

a) Within subpopulation:

80% reference set

20% validation set

80% of the animals were used as reference set

20% of the animals as validation set

Comparison of DGVs with DRP
Material and Methods: random five-fold cross validation

Assessment of predictive ability of DGV prediction

b) Multi-subpopulation:

80% reference set  20% validation set

Subp. 1  →  80%  →  80% of Subp. 1

Subp. 2  →  80%  →  80% of Subp. 2

Subp. 3  →  80%  →  80% of Subp. 3

Comparison of DGVs with DRP

20% Subp. 1
Results: random five-fold cross validation

Predictive ability with DRP and $G_{\text{VanRaden}}$ exemplary for subpopulation 1
Material and Methods: Forward Prediction

Assessment of predictive ability of DGV prediction

1. Within and multi-subpopulation:

Reference set: born before 2010

Validation set: born in 2010 and 2011
Material and Methods: Forward Prediction

Assessment of predictive ability of DGV prediction

1. Within and multi-subpopulation

2. Effect of different G matrices:
   
   • G introduced by VanRaden (2007)
     • with actual allele frequencies over total set of individuals
     • with founder allele frequencies (Gengler et al., 2007) per subpopulation

   • G introduced by Zhou et al. (2014)
     • accounting for substructure by including information of marker effects (estimated from reference set) and linkage disequilibrium
## Results: Forward Prediction

### Validation set: Subp. 1 (n = 87)

<table>
<thead>
<tr>
<th>Ref(n=)</th>
<th>100</th>
<th>222</th>
<th>255</th>
</tr>
</thead>
</table>

### Subp. 4 (n = 257)

<table>
<thead>
<tr>
<th>Cor(DGV,DRP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
</tr>
</tbody>
</table>

- **VanRaden**
- **Gengler**
- **Zhou**
Conclusions

✓ 5-fold CV: Decrease (slight decrease) in predictive ability for distantly (closely) related multi-subpopulation reference sets

✓ Forward prediction:
  ✓ Slight increase in predictive ability, especially by adding subpopulation 2 to the reference set
  ✓ Slight increase in predictive ability when using different G matrices, especially when accounting for substructures

✓ Forming a multi-subpopulation reference population generally did not lead to a better predictive ability for individuals within a specific subpopulation

✓ Necessity to creating more concurrent links between subpopulations, e.g. by using the same boars across populations
Thank you for your Attention!

Acknowledgment:
The author gratefully acknowledge the financial support of the project ‘pigGS’ by the Europäischen Fonds für regionale Entwicklung (EFRE), the state of North Rhine-Westphalia, the project management Jülich and the pig breeding organizations for providing the data.

We especially thank the EAAP for granting a scholarship for the EAAP Meeting 2015 in Warsaw.
References


