Accuracy of genomic prediction using whole-genome sequence data in white egg layer chickens

Marzieh Heidaritabar, Mario Calus, Hendrik-Jan Megens, Addie Vereijken, Martien Groenen, John Bastiaansen
Acknowledgements

www.hendrix-genetics.com

www.breed4food.com
Introduction
Why sequence data?

- **Causative mutation** are assumed to be in data
- Expectation: **Higher accuracy** of genomic prediction
- Understand **biology**
Why sequence data?

- **Causative mutation** are assumed to be in data
- Expectation: **Higher accuracy** of genomic prediction
- Understand **biology**

Why imputation?

- Sequencing is still expensive!
- To get a large population of sequenced animals:
  - Sequence **key ancestors**
  - **Imputation** of genotyped animals
Objectives

Can **accuracy** of genomic prediction be **improved** by:

- Sequencing **key ancestors** of population and imputation of genotyped animals?
- **Pre-selecting SNPs** that are more likely to affect the phenotype?
Material and methods
Data

62 ancestors

G0

G1

G2

G3

2,140 white egg layers (60K)
Sequenced 22 key ancestors
~ 40% genetic variation

2,140 white egg layers (60K)

1,244 females (60K)
Quality control

60K

~ 25K

Called SNPs: ~ 9 million

~ 6.8 million SNPs
MAF distribution

MAF distribution_60K

Percent of Total

Minor Allele Frequency (MAF)

MAF distribution_SEQ

Percent of Total

Minor Allele Frequency (MAF)
MAF Distribution after MAF filtration (< 0.025)
MAF Distribution after MAF filtration (< 0.025)

~ 4.6 million SNPs remained for further analyses
Imputation to whole-genome sequence

- Beagle V4

22 sequenced key ancestors

Imputation

1,244 genotyped animals (60K)
Genomic prediction

Trait: Number of eggs in the first production period
1,244 females

Reference: 1,004 animals
Hatched Apr, 2009-June, 2011

Validation: 240 animals
Hatched Oct-Nov, 2011

Accuracy = \frac{r_{BV, Phen}}{\sqrt{h^2}}
Prediction models

- **BLUP** (pedigree-based model)

- **GBLUP**
  - **Assumption:** equal variances for all SNPs

- **BayesC**
  - Gibbs sampling
  - 60,000 iterations (10,000 burn-in)
  - **Assumption:**
    - 95% of SNPs with zero effect
    - 5% of SNPs with large effect
Genomic prediction using biological information

<table>
<thead>
<tr>
<th>Annotation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>41,010</td>
</tr>
<tr>
<td>Non-synonymous</td>
<td>15,516</td>
</tr>
<tr>
<td>Non-coding</td>
<td>4,539,701</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4,596,227</strong></td>
</tr>
</tbody>
</table>
Genomic prediction using biological information

<table>
<thead>
<tr>
<th>Annotation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>41,010</td>
</tr>
<tr>
<td>Non-synonymous</td>
<td>15,516</td>
</tr>
<tr>
<td>Non-coding</td>
<td>4,539,701</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4,596,227</strong></td>
</tr>
</tbody>
</table>

- **Synonymous SNPs**: do not alter amino acid sequence of a protein
- **Non-synonymous SNPs**: alter amino acid sequence of a protein
Results
Prediction accuracies

- BLUP
- GBLUP_60K
- GBLUP_SEQ
- BayesC_60K
- BayesC_SEQ

Prediction accuracy: 0.59
Prediction accuracies

Prediction accuracy

- BLUP: 0.59
- GBLUP _60K: 0.74
- BayesC _60K
- BayesC SEQ
Prediction accuracies

- BLUP: 0.59
- GBLUP 60K: 0.74
- BayesC 60K: 0.75
- GBLUP_SEQ: 0.75
- BayesC_SEQ: 0.75
Prediction accuracies

<table>
<thead>
<tr>
<th>Method</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLUP</td>
<td>0.59</td>
</tr>
<tr>
<td>GBLUP 60K</td>
<td>0.74</td>
</tr>
<tr>
<td>GBLUP_SEQ</td>
<td>0.75</td>
</tr>
<tr>
<td>BayesC 60K</td>
<td>0.72</td>
</tr>
<tr>
<td>BayesC_SEQ</td>
<td></td>
</tr>
</tbody>
</table>
Genomic prediction using biological information

Prediction accuracy:
- All SNPs: 0.75
- Coding non-synonymous
Genomic prediction using biological information

- Prediction accuracy for All_SNPs: 0.75
- Prediction accuracy for Coding: 0.68
- Prediction accuracy for non-synonymous: (not shown in the image)
Genomic prediction using biological information

- Non-coding regulatory regions have an affect on phenotype (Koufariotis et al. 2014)
- QTL and selection signals were found in non-coding regions (Bird et al. 2006; Drake et al. 2005)
Conclusions

- **Whole-genome sequence** hardly improved genomic prediction accuracy compared with **60K**
- **GBLUP** and **BayesC** produced **similar** prediction accuracies
- Selection of **non-synonymous SNPs** did not improve accuracy of genomic prediction
Thank you

Email: marzieh.heidaritabar@wur.nl
Which animals to sequence?

Proportion of diversity (Druet et al. 2014)

- **G matrix** for 2,140 genotyped animals
  \[ P_n = G_n^{-1} c_n \]
  - \( G_n \) : a subset of G matrix (62 ancestors)
  - \( c_n \) : average genomic relationship of 62 animals with whole population
  - \( P_n \) : proportion of genetic diversity
Which animals to sequence?

Proportion of diversity (Druet et al. 2014)