Using additional SNPs selected from whole genome sequence (WGS) data for genomic prediction in Danish Jersey

Aoxing Liu, Mogens Sandø Lund, Didier Boichard, Sebastien Fritz, Emre Karaman, Yachun Wang, Guosheng Su

Feb 12, 2018
• Introduction
• Material and methods
• Results and discussion
• Conclusion
Introduction
**Background**

High throughput genotyping

- LD (7 K)
- MD (54 K)
- HD (777 K)
- WGS (~26,700 K)

**Hypothesis:** Higher SNP density -> better LD -> higher reliability
**Background**

---

**Hypothesis:** Higher SNP density -> better LD -> higher reliability

**Real data:** HD ≈ 54K (Su et al., 2012) & Imputed WGS ≈ HD (Van Binsbergen et al., 2015)

- Only causative mutations or variants very close to causative mutations can improve reliability
  
  (van den Berg et al., 2016)

- non-causative mutations bring noise
Objectives

(1) 54 K + SNPs selected from WGS

(2) Joint reference

(3) Models

Reliable genomic prediction

- Investigate effects of additional WGS SNPs on genomic prediction
- Effects of using additional WGS SNPs in a joint reference
- Assessed models on their efficiency to use information of additional WGS SNPs
Material and methods
Workflow

Experience from large scale use of the EuroGenomics custom SNP chip in cattle (Boichard et al., WCGALP, 2018)

- NOR SNPs (Brondum et al., 2015)
  - peaks of QTL from Nordic Holsteins, Nordic Red and Danish Jersey

- FR SNPs
  - literature
  - a strong variant effect predictor annotation (e.g. non-synonymous substitution)
  - regulatory regions of genes
  - peaks of QTL
  - breakpoints of structural SNPs
Imputation

Animal
- DK bulls: ~1,300
- US bulls: ~1,200
- DK cows: ~31,000

Genotype
- 54K chip
- standard LD chip
- customized LD chip
  - standard LD chip
  - NOR SNPs
  - FRA SNP

Pedigree
- 6,100 males
- 66,000 females

Two-step imputation
(Fimpute)

54K

54K + NOR SNPs + FRA SNPs

(1)

(2)
**Prediction: SNP**

**Quality control**

- Minor allele frequency > 0.01
- Imputation accuracy
  - correlation > 0.8
  - concordance rate > 0.8

<table>
<thead>
<tr>
<th>SNPs</th>
<th>No. of SNPs</th>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>54K</td>
<td>40,452</td>
<td></td>
<td>33,166</td>
</tr>
<tr>
<td>NOR SNPs</td>
<td>1,754</td>
<td></td>
<td>1,270</td>
</tr>
<tr>
<td>FRA SNPs</td>
<td>4,325</td>
<td></td>
<td>2,427</td>
</tr>
</tbody>
</table>
**Prediction: GBLUP model**

- **One-component model**
  \[ y = 1\mu + X \mathbf{g} + e \]
  54K/ 54K+selected WGS SNPs

- **Two-component model**
  \[ y = 1\mu + X_{54K}\mathbf{g}_{54K} + X_{WGS}\mathbf{g}_{WGS} + e \]
  54K Selected WGS SNPs

---

**Scenarios**

<table>
<thead>
<tr>
<th>Component_One</th>
<th>Component_Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>54K_NOR</td>
<td>NOR</td>
</tr>
<tr>
<td>54K_FRA</td>
<td>FRA</td>
</tr>
<tr>
<td>54K_NOR_FRA</td>
<td>NOR+FRA</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Component_One</th>
<th>Component_Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>54K_NOR</td>
<td>54K</td>
</tr>
<tr>
<td>54K_FRA</td>
<td>54K</td>
</tr>
<tr>
<td>54K_NOR_FRA</td>
<td>54K</td>
</tr>
</tbody>
</table>
Reference

- DK: ~1,000 DK bulls born before 2005
- Joint DK-US: ~1,000 DK bulls born before 2005
  ~1,200 US bulls

Validation

- ~300 DK bulls born after 2005
Prediction: Model comparison

Compare reliabilities from different models/scenarios:

- **SE of reliability:**
  
  Non-parametric Bootstrap with 10,000 samples

- **Significant test**
  
  Two-tailed paired t-test with p-value = 0.05
Results and discussion
Results - Reliability of milk

- Inclusion of additional WGS SNPs significantly improved reliability (11.4-17.0%)
- Inclusion of all additional WGS SNPs achieved highest reliabilities
Results-Reliability of milk

A joint DK-US reference significantly better than a DK reference (20%)
Results - Reliability of milk

- Two-component model
- One-component model

- Additional WGS SNPs improved reliabilities of a joint reference (11.5-13.6%)
A two-component model improved reliabilities (4.8%)
Results - Reliability of protein & fat

- Similar to milk

Protein

- One-component model
- Two-component model

Fat

- One-component model
- Two-component model
No significant difference between 54K and 54K + selected WGS SNPs
Dose the improvement of reliabilities come from increase of SNP density?

- **54K**
  - ~33 K SNPs
  - **✓ 54K + NOR + FRA**

**VS.**

- Randomly remove ~4K SNPs from 54K

- **54K**
  - ~33 K SNPs
  - **✓ 54Kminus + NOR + FRA**

No. of SNPs is equal to 54K chip!
### Results

#### Reliability (54Kminus + NOR + FRA) – Reliability (54K + NOR + FRA)

<table>
<thead>
<tr>
<th>Trait</th>
<th>Reference</th>
<th>One-component</th>
<th>Two-Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk</td>
<td>DK</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>DKUS</td>
<td>0</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>DKUS</td>
<td>-0.002</td>
<td>-0.003</td>
</tr>
<tr>
<td>Protein</td>
<td>DK</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>DKUS</td>
<td>-0.002</td>
<td>-0.003</td>
</tr>
<tr>
<td>Fat</td>
<td>DK</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DKUS</td>
<td>-0.001</td>
<td>-0.003</td>
</tr>
</tbody>
</table>

- No difference between 54K + NOR + FRA and 54Kminus + NOR + FRA
- Improvement of reliabilities using additional WGS SNPs not from increase of SNP density
Conclusion
Conclusion

- Additional WGS SNPs improved reliabilities for milk production, not for fertility and mastitis
- The inclusion of all additional WGS SNPs achieved highest reliabilities
- A joint DK-US reference better than a DK reference for all traits
- Additional WGS SNPs further improved reliabilities of a joint DK-US reference
- A two-component model improved reliabilities for milk production
• Gert Pedersen Aamand, NAV
• Esa Mantysaari, Luke
• Per Madsen, Aarhus University
• Goutam Sahana, Aarhus University
• Ulrik Sander Nielsen, SEGES
• Han Mulder, Wageningen University & Research
• Xiaowei Mao, Cornell University
• Peipei Ma, Shanghai Jiao Tong University
Imputation accuracy

- Correlation = COR (TRUE, IMPUTED)

- Concordance rate = \(\frac{\text{No. of animals with correctly imputed genotypes}}{\text{No. of animals with imputed genotypes}}\)
Non-parametric Bootstrap

1) Read data of 269 bulls in validation population
2) Randomly sample 269 rows with replacement
3) Calculate R2 for SCE1 and SCE2 for each bootstrap sample
4) Repeat this process 10,000 times
5) Differences between reliabilities among scenarios: CI and paired t-test

<table>
<thead>
<tr>
<th>ID</th>
<th>DRP</th>
<th>R2_DRP</th>
<th>SCE1</th>
<th>SCE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>104.1</td>
<td>99</td>
<td>1.92</td>
<td>-7.13</td>
</tr>
<tr>
<td>2</td>
<td>88.9</td>
<td>93</td>
<td>-1.38</td>
<td>-11.89</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>269</td>
<td>113.0</td>
<td>99</td>
<td>22.40</td>
<td>16.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round</th>
<th>R2_SCE1</th>
<th>R2_DRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10,000</td>
<td>0.39</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Two-tailed paired t-test with df=10,000-1

\[
\text{mean} (R2_{SCE1}) - \text{mean} (R2_{SCE2}) = \frac{\text{mean} (R2_{SCE1}) - \text{mean} (R2_{SCE2})}{\text{se}}
\]