Selection of sequence variants to improve dairy cattle genomic predictions

J. R. O’Connell,¹ M.E. Tooker,² D.M. Bickhart,² and J.B. Cole,² and P. M. VanRaden²

¹University of Maryland School of Medicine, Baltimore, MD, USA
²Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD, USA

joconnel@medicine.umaryland.edu
Strategies to choose from millions of imputed sequence (SEQ) variants

- O’Connell and VanRaden
- Based on simulated data
2015 simulated data

- 26,984 HOL bulls in U.S. reference population
- 30 million simulated variants; 10,000 QTLs
- 30 equal-length chromosomes (100 Mbases)
- 3 different chip densities (HD, MD, LD)
- 5 independent traits (same QTL locations)
### Simulation: REL from 1M, 60K+1M subset

<table>
<thead>
<tr>
<th>Trait</th>
<th>600K</th>
<th>60K+25K</th>
<th>Difference</th>
<th>All 1M</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.3</td>
<td>85.4</td>
<td>5.1</td>
<td>86.7</td>
<td>6.4</td>
</tr>
<tr>
<td>2</td>
<td>80.1</td>
<td>85.3</td>
<td>5.2</td>
<td>87.7</td>
<td>7.6</td>
</tr>
<tr>
<td>3</td>
<td>80.4</td>
<td>84.9</td>
<td>4.5</td>
<td>86.1</td>
<td>5.7</td>
</tr>
<tr>
<td>4</td>
<td>78.6</td>
<td>83.5</td>
<td>4.9</td>
<td>84.8</td>
<td>6.2</td>
</tr>
<tr>
<td>5</td>
<td>81.2</td>
<td>86.0</td>
<td>4.8</td>
<td>87.6</td>
<td>6.4</td>
</tr>
<tr>
<td>Avg.</td>
<td>80.1</td>
<td>85.0</td>
<td>4.9</td>
<td>86.4</td>
<td>6.3</td>
</tr>
</tbody>
</table>
1000 Bulls Genome Project

- 1000 Bulls Genome Project is an international SEQ project that seeks to pool resources in order to impute SEQ-derived genetic variants across a wide range of cattle breeds.

- To join the project required a minimum of 25 animals sequenced at 10.5× coverage and approval by the project’s steering committee.

  - USDA contributed 76 bulls (26 Holstein)
1000 Bulls Genome Project (continued)

- SEQ alignment map created according to set specifications and collected from partners

- SAMtools used to identify SNPs and indels and produce genotype probabilities

- Beagle used for imputation

- Project data heavily processed, filtered, and imputed
  - 10% of 60K and HD SNPs missing
SNP vs. SEQ variants

- **SNPs**
  - At least 2 different nucleotides (A, C, G, or T) observed
  - Previous SNP chips only include these

- **SEQ data has many insertions/deletions (indels)**
  - Indels can range in length (up to 50 bases)
  - Not easily captured by chip technology
  - Calls have lower quality than for SNPs

- Other more complex variant classes (such as copy number variants) were not identified from the raw data
Methods for HD+, HD+indels (HD+I), 77K

- Current HD chip has 312,614 usable SNPs after removing more than half due to high LD

- HD+: 481,904 candidate SEQ SNPs added
  - 107,471 exonic
  - 9,422 splice variants (same gene, different protein)
  - 35,242 untranslated regions at beginning and end of genes
  - 329,769 SNPs 2kb upstream or 1kb downstream of genes

- HD+I: Also added 249,966 indels in or near genes to HD+

- 77K: Add 17K to current 60K evaluation chip to compare with Wiggans’ 77K selected from HD
## Edits to 39 million variants

<table>
<thead>
<tr>
<th>Edit</th>
<th>Number removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remove MAF &lt; 0.01</td>
<td>20M</td>
</tr>
<tr>
<td>Remove for LD &gt; 0.95</td>
<td>13M</td>
</tr>
<tr>
<td><strong>Total removed</strong></td>
<td><strong>33M</strong></td>
</tr>
<tr>
<td><strong>Total remaining</strong></td>
<td><strong>6M</strong></td>
</tr>
</tbody>
</table>

**Imputation**

- Remove for imputation accuracy          3M
Methods for imputation

- Imputation quality assessment
  - Select 40 of 440 SEQ Holsteins
  - Reduce to HD
  - Impute to SEQ
  - Compare with original SEQ
- HD imputed genotypes for 26,970 progeny-tested Holstein bulls
- Findhap designed for equally spaced markers, but SEQ-selected markers are bunched near genes
Imputation accuracy

Imputation accuracy (%)

HD+  HD+I  3Million

Chromosome

1  3  5  7  9  11  13  15  17  19  21  23  25  27  29  31

91  92  93  94  95  96  97  98  99
Selecting the best SEQ variants

- Developing field – no “gold” standard as to best way to select variants

- 77K chip
  - HD+ results used to choose 1,000 variants with large effects for each of the 33 traits
  - Reduce 33,000 to 17,000
    - SNPs near *DGAT1* and other QTL
    - 60K chip
    - Duplicate SNPs that effect multiple traits
Chr 5 net merit SNP selection example

Before edits: 1,719 SNPs
Chr 5 net merit SNP selection example

After edits: 693 SNPs
## Gains in REL

<table>
<thead>
<tr>
<th>Trait</th>
<th>HD only</th>
<th>HD + 482K Difference</th>
<th>HD + indels Difference</th>
<th>60K only</th>
<th>60K+ 17K Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk</td>
<td>34.1</td>
<td>33.9</td>
<td>-0.2</td>
<td>33.9</td>
<td>-0.2</td>
</tr>
<tr>
<td>Fat</td>
<td>33.7</td>
<td>34.0</td>
<td>0.3</td>
<td>33.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>Protein</td>
<td>27.9</td>
<td>27.0</td>
<td>-0.9</td>
<td>26.7</td>
<td>-1.2</td>
</tr>
<tr>
<td>Fat %</td>
<td>49.2</td>
<td>52.7</td>
<td>3.5</td>
<td>52.4</td>
<td>3.2</td>
</tr>
<tr>
<td>Protein %</td>
<td>42.1</td>
<td>41.6</td>
<td>0.5</td>
<td>43.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Interbull annual meeting, Puerto Veras, Chile – October 25, 2016 (15)
## Gains in REL (continued)

<table>
<thead>
<tr>
<th>Trait</th>
<th>HD only</th>
<th>HD + 482K</th>
<th>Difference</th>
<th>HD + indels</th>
<th>Difference</th>
<th>60K only</th>
<th>60K+ 17K</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL</td>
<td>36.1</td>
<td>33.9</td>
<td>−0.3</td>
<td>36.4</td>
<td>0.3</td>
<td>35.6</td>
<td>38.2</td>
<td>2.6</td>
</tr>
<tr>
<td>SCS</td>
<td>35.9</td>
<td>34.0</td>
<td>0.2</td>
<td>37.1</td>
<td>1.2</td>
<td>35.1</td>
<td>37.0</td>
<td>1.9</td>
</tr>
<tr>
<td>DPR</td>
<td>30.8</td>
<td>27.0</td>
<td>−0.8</td>
<td>31.2</td>
<td>0.4</td>
<td>29.0</td>
<td>33.0</td>
<td>4.0</td>
</tr>
<tr>
<td>CCR</td>
<td>28.7</td>
<td>52.7</td>
<td>−0.6</td>
<td>28.8</td>
<td>0.1</td>
<td>28.9</td>
<td>31.8</td>
<td>2.9</td>
</tr>
<tr>
<td>HCR</td>
<td>19.0</td>
<td>41.6</td>
<td>1.3</td>
<td>19.7</td>
<td>0.7</td>
<td>20.5</td>
<td>21.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>
### Gains in REL (continued)

<table>
<thead>
<tr>
<th>Trait</th>
<th>HD only</th>
<th>HD + 482K Difference</th>
<th>HD + indels Difference</th>
<th>HD + candidate SNPs</th>
<th>HD + candidate SNPs Difference</th>
<th>HD + candidate SNPs Difference</th>
<th>HD + candidate SNPs Difference</th>
<th>HD + candidate SNPs Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final score</td>
<td>24.7</td>
<td>25.5</td>
<td>0.8</td>
<td>25.8</td>
<td>1.1</td>
<td>25.8</td>
<td>1.1</td>
<td>24.6</td>
</tr>
<tr>
<td>Stature</td>
<td>30.4</td>
<td>32.4</td>
<td>2.0</td>
<td>32.8</td>
<td>2.4</td>
<td>32.8</td>
<td>2.4</td>
<td>30.3</td>
</tr>
<tr>
<td>Strength</td>
<td>29.9</td>
<td>31.8</td>
<td>1.9</td>
<td>31.8</td>
<td>1.9</td>
<td>31.8</td>
<td>1.9</td>
<td>29.9</td>
</tr>
<tr>
<td>Dairy form</td>
<td>33.8</td>
<td>35.3</td>
<td>1.5</td>
<td>35.8</td>
<td>2.0</td>
<td>35.8</td>
<td>2.0</td>
<td>35.0</td>
</tr>
<tr>
<td>Net merit</td>
<td>23.8</td>
<td>24.3</td>
<td>0.5</td>
<td>24.4</td>
<td>0.6</td>
<td>24.4</td>
<td>0.6</td>
<td>23.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60K + selected</td>
<td></td>
<td>60K + selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>60K only</td>
<td>60K+ 17K Difference</td>
<td></td>
<td>60K + selected</td>
<td></td>
<td>60K + selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final score</td>
<td>24.6</td>
<td>27.8</td>
<td>3.2</td>
<td>25.8</td>
<td>3.2</td>
<td>25.8</td>
<td>3.2</td>
<td>24.6</td>
</tr>
<tr>
<td>Stature</td>
<td>30.3</td>
<td>34.7</td>
<td>4.3</td>
<td>32.8</td>
<td>4.3</td>
<td>32.8</td>
<td>4.3</td>
<td>30.3</td>
</tr>
<tr>
<td>Strength</td>
<td>29.9</td>
<td>34.5</td>
<td>4.6</td>
<td>31.8</td>
<td>4.6</td>
<td>31.8</td>
<td>4.6</td>
<td>29.9</td>
</tr>
<tr>
<td>Dairy form</td>
<td>35.0</td>
<td>38.2</td>
<td>3.2</td>
<td>35.8</td>
<td>3.2</td>
<td>35.8</td>
<td>3.2</td>
<td>35.0</td>
</tr>
<tr>
<td>Net merit</td>
<td>23.4</td>
<td>24.7</td>
<td>1.3</td>
<td>24.4</td>
<td>1.3</td>
<td>24.4</td>
<td>1.3</td>
<td>23.4</td>
</tr>
</tbody>
</table>
Overall gains in REL

<table>
<thead>
<tr>
<th>Trait group</th>
<th>HD + candidate SNPs</th>
<th>HD + 482K</th>
<th>HD + indels</th>
<th>60K + 17K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production</td>
<td>0.6</td>
<td>0.5</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>−0.1</td>
<td>0.5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Calving</td>
<td>−0.6</td>
<td>−1.8</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>1.0</td>
<td>0.8</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>All traits</td>
<td>0.6</td>
<td>0.5</td>
<td>2.7</td>
<td></td>
</tr>
</tbody>
</table>
Summary

- **39M** sequenced genotypes from **444** Holsteins edited to **6M**
- Imputed **6M** to **26,970** reference bulls then edited to **3M**
- Added gene-centric loci to HD chip to create HD+ and HD+I
- Estimated effect sizes using 2012 data
- Selected **17K** SNPs to add to **60K**
- Compared HD+ and HDII to HD and 77K to 60K using 2016 data
Summary

- HD+ and HD+I candidate approach
  - Negative REL differences
  - Prior variance spread thinner
  - Indels have less accurate calls

- 77K chip selection approach
  - Difference in REL always positive
  - Average REL gain of 2.7 percentage points across traits
  - Best performance
Sequence data – the future?

- The 1000 Bulls Genome Project run 5 – 1500 bulls
  - Unfiltered data on 70M variants available
- The 1000 Bulls Genome Project run 6 – 3000 bulls
  - Number of Holsteins?
  - Release date?
- Additional independent SEQ projects underway
- Better reference assembly
- Resources to collect data and generate independent call sets
SNP selection – the future?

- Fixed or variable number for each trait
- GWA, multiple regression or other methods to estimate effect size
- Bioinformatics
  - Gene expression, proteomics, methylation, chromatin structure to find [e,m,me,p]QTLs
  - Prioritize non-genic SNPs and SNPs in LD groups
- Functional data
  - Difficult and expensive to go from correlation to causality
Integration of SNP selection into genomics
Integration of SNP selection into genomics

- Different chips designers will choose different SNPs
- Low density chips will not able to include all SNPs
- SEQ SNPs may not perform on chip
- Need sufficient number of chips to power imputation
- Timing of sequencing, SNP selection and chip design
- Evaluating performance of SNPs for future designs
Wrapping it up

- Acknowledgments
  - JRO supported by USDA SCA 58-45-14-070-1

- Slides available at https://aipl.arsusda.gov/publish/present.htm

- Stay tuned for updates next year!