Selecting animals for whole genome sequencing: methods optimization and comparison

Adrien Butty¹, M. Sargolzaei¹,², F. Miglior¹,³, B. Gredler⁴, C. Baes¹

¹CGIL - University of Guelph; ²Semex Alliance, Guelph; ³Canadian Dairy Network, Guelph, Ontario, Canada; ⁴Qualitas AG, Zug, Switzerland
The Efficient Dairy Genome Canada Project

- International project aiming to develop genomic evaluation for Feed Efficiency and Methane Emissions of dairy cattle

- Genome Canada Project will sequence 48 animals to improve imputation of the Holstein population at whole-genome sequence density.

- Currently 451 Holstein bulls from the 1000 Bulls Genome Project are sequenced and available to us.

⇒ The task: create a list of the animals to sequence.
Genotyping is getting cheaper and cheaper...
- Routine genotyping is practised for AI bulls everywhere
- Mostly use of medium or high density chips
  - for genotyping and for genomic evaluation

Sequencing is also getting cheaper, but still expensive for large scale sequencing
- Cost per genome: > $1,000
- Imputation of medium or high density genotypes to whole-genome sequence (WGS) density feasible, but

→ Good reference population needed to more accurately impute rare variants!
Selection methods

• Without genotypic information
  – pedigree based; aim for high genetic contribution

• Within genotyped population
  – based on genomic relationship matrix
  – relying on haplotype frequencies
    • target common haplotypes
    • target rare haplotypes

Currently, sequenced animals are mostly key ancestors; common haplotypes are thus expected to be sequenced.
Pedigree-based method

- Also called "Key Ancestors" method

- Aim to find the \( n \) animals explaining the most of the genetic variance of a population with:

\[
p_n = A^{-1}_n c_n
\]

\( p_n \) = vector of the proportion of gene pool captured by the \( n \) animals

\( A^{-1}_n \) = subset of the numerator relationship matrix

\( c_n \) = average relationships of \( n \) selected animals with the entire population

\( n \) = number of selected animals

- True proportion of genetic variance mostly overestimated
- The \( A \) matrix can be replaced by the \( G \) matrix.
Haplotype-based methods

- Druet et al. (2014) ⇒ maximize haplotype coverage / sum of the haplotype frequency at every SNP

\[ \text{Sample Weight} = \sum_{i=1}^{NHAP} f_i \]

- Bickhart et al. (2016) ⇒ prioritize sequencing of rare haplotypes

\[ \text{Sample Weight} = \sum_{i=1}^{NHAP} f_i^2 - 2f_i + 1 \]
Optimize the genetic diversity of the sequenced animals

- use of probabilistic algorithm; simulated annealing
- optimization based on a group of animals and not on individuals
  - already sequenced and newly selected animals accounted for
  - genetic diversity = sum of count of unique haplotypes

Select animals that will enable higher accuracy of imputation of rare variants within the Holstein North-American population
Optimize the genetic diversity of the sequenced animals.
  - use of probabilistic algorithm; simulated annealing
  - optimization based on a group of animals and not on individuals
    - already sequenced and newly selected animals accounted for
    - genetic diversity = sum of count of unique haplotypes

Select animals that will enable higher accuracy of imputation of rare variants within the Holstein North-American population
### Index of genetic diversity

<table>
<thead>
<tr>
<th>Haplotype A</th>
<th>Haplotype B</th>
<th>Haplotype C</th>
<th>Haplotype D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal 1 H1</td>
<td>10010001210010010011001</td>
<td>00101000101011001110</td>
<td>1011000011100111010</td>
</tr>
<tr>
<td>Animal 1 H2</td>
<td>10001000101000010011001</td>
<td>0001000011001111110</td>
<td>1011000011100111010</td>
</tr>
<tr>
<td>Animal 2 H1</td>
<td>01000100010100001110011</td>
<td>00101000000010011010</td>
<td>0000000001001111111</td>
</tr>
<tr>
<td>Animal 2 H2</td>
<td>00101000010100101100110</td>
<td>00010000100001001110</td>
<td>1010000001001111110</td>
</tr>
<tr>
<td>Animal 3 H1</td>
<td>01000001010010100011001</td>
<td>00001000100000010011</td>
<td>1000000100000111110</td>
</tr>
<tr>
<td>Animal 3 H2</td>
<td>01000100101000101100110</td>
<td>00010000100000010011</td>
<td>1000000100000111110</td>
</tr>
<tr>
<td>Animal 4 H1</td>
<td>10000000010100001110011</td>
<td>00010000100000010011</td>
<td>1000000100000111110</td>
</tr>
<tr>
<td>Animal 4 H2</td>
<td>00000011100100100011001</td>
<td>01001000010000011110</td>
<td>1000000100000111110</td>
</tr>
<tr>
<td>Animal 5 H1</td>
<td>00000001100100010011001</td>
<td>01001000010000011110</td>
<td>1000000100000111110</td>
</tr>
<tr>
<td>Animal 5 H2</td>
<td>00000111100100010011001</td>
<td>01001000010000011110</td>
<td>1000000100000111110</td>
</tr>
<tr>
<td></td>
<td>10000011000000010011001</td>
<td>01001000010000011110</td>
<td>1000000100000111110</td>
</tr>
</tbody>
</table>

**Total diversity index:** 18
Simulated annealing

- Find the global optimum when several local optima exist
- Accept “bad” steps when these are probabilistically near to another optimum (the good one, who knows...?)
  - high temperature $\rightarrow$ big “bad” steps accepted
  - low temperature $\rightarrow$ only small “bad” steps accepted

Source: Wiki User Kingpin13
Animals and genotypes

• Candidates
  – Holstein bulls born after 01/01/2011 in Canada or USA and genotyped with 50K or higher density
    35,706 animals
  – After filtering possible crossbreds
    32,000 animals

• Sequenced animals
  – HOL or RED animals from Run5 of the 1’000 Bulls Genome Project
    451 animals

• All genotypes (back-) imputed to 50K panel with FImpute
  – Pedigree contained 151,436 animals, up to 48 generations
  – 44,347 autosomal SNP
PCA distributions

![PCA diagram]

- PC1
- PC2
- Candidates
- Sequenced
PCA distributions
Genetic Diversity Index

- All animals: 100%
- A matrix: 33.3%
- G matrix: 32.8%
- Sim. Ann.: 40.2%
Proportion of rare haplotype alleles

Method of selection

- A matrix
- G matrix
- Sim. Ann.

Proportion of allele with maf < 0.05 covered by the sequenced & selected animals (%)

0 5 10 15 20 25 30 35 40 45
• Methods of selection for sequencing have shifted from pedigree or SNP-based to haplotype-based methods

• Aiming discovery and coverage of rare variants at sequencing should improve their imputation accuracy

• Our method identifies animals representing the whole population as well as previous methods

• Using our approach, we hope to better cover rare haplotypes in future sequenced populations
We very gratefully acknowledge support from our funders and collaborators:
more about the Efficient Dairy Genome Project:

« An international initiative to decrease the environmental footprint of dairy cattle using genomics »

by Filippo Miglior,
Wednesday at **14:45** in room **3B**