Alternative haplotype construction methods for genomic evaluation

D. Jónás¹,²,³, V. Ducrocq¹, M. N. Fouilloux⁴, P. Croiseau¹

¹INRA, UMR1313 GABI, Jouy-en-Josas, France
²AgroParisTech, Paris, France
³ALLICE, Paris, France
⁴Idele, UMR1313 GABI, Jouy en Josas, France
Genomic selection

- Improved breeding value estimation by exploiting knowledge on DNA sequence variation
- Bi-allelic SNP markers
- Requirement: large reference populations

An alternative to SNP: haplotypes

- A combination of N SNP
- Multi-allelic genetic marker
- Which haplotypes should be used?
Objectives

- Develop a method to *a priori construct* the “best” haplotypes for genomic selection
  - How to define the “best” haplotype?
    - Maximize the number of alleles
    - Consider the distribution of *allele frequencies*
Materials and Methods

- Haplotype construction
  - Definition: combination of $N$ consecutive SNP
  - An allele effect was considered “predictable”, if it had a sufficiently high allele frequency
  - Tested haplotype sizes: 3-5 SNP
  - Window size: 10 SNP
  - Two, slightly different criteria
    - Termed as Criterion-A and Criterion-B
    - Difference: Criterion-B includes a constraint on the equilibrium of allele frequencies

Flanking markers
The dataset

- Dairy cattle breed (Montbéliarde)
- n=2,235 individuals (Training: 1,666; Validation: 569)
- 5 dairy cattle production traits
  - Milk quantity, fat yield, fat content, protein yield, protein content
- 43,801 SNP from the 50K chip
- Assumed *a priori* information (QTL-SNP)
  - Represent *approximate* QTL positions
  - The most significant 1-, 3- and 6 thousand QTL-SNP
Analyses based on different sources of genomic information (**Bayes-C**)

- Only QTL-SNP
- Haplotypes built from:
  - Flanking markers
  - Criterion-A
  - Criterion-B

Comparison of the results

**Correlation** between phenotypes (DYD) and GEBV
### DYD-GEBV correlations observed with the flanking marker haplotypes

<table>
<thead>
<tr>
<th>Number of QTL-SNP</th>
<th>QTL-SNP</th>
<th>Flanking marker haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HS=3</td>
</tr>
<tr>
<td>1K</td>
<td>0.480</td>
<td>0.491</td>
</tr>
<tr>
<td>3K</td>
<td>0.499</td>
<td>0.523</td>
</tr>
<tr>
<td>6K</td>
<td>0.512</td>
<td>0.534</td>
</tr>
<tr>
<td>Optimal</td>
<td>0.512</td>
<td>0.534</td>
</tr>
</tbody>
</table>

- **Optimal number of QTL-SNP**: for each trait separately, the number of haplotypes leading to the highest correlation coefficient is considered.
- Average correlations of the 5 production traits are shown.
Frequency of the rare alleles decreased with either of the 2 proposed criteria

Similar results were observed with haplotypes of 3 and 5 SNP
Correlations observed with the flanking- and selected SNP

- Haplotypes built from the selected markers were superior in genomic selection
- Gain with the haplotype selection methods decreased as the haplotype size increased
Conclusions

- **Selection of SNP is beneficial** to build haplotypes for genomic selection
  - Reduced number of markers in the model
  - Larger number of alleles with proper frequencies
  - **Higher DYD-GEBV correlations**

- Decreasing gain with the increase of haplotype size

- Similar results with the **HD-chip**

- Similar results with **other breeds**
Thank you for your kind attention!