Reference Population Design
Retraining
etc.

Jack Dekkers
The Promise of Genomic Selection

Reduce requirement to get phenotypes on selection candidates and on close relatives in order to estimate EBV

Traditional EBV

Genomic Selection

Estimates of marker effects from training population

Candidate
How to build a Reference Population for a Closed Breeding Population?

Accuracy of GEBV is greater if

• More individuals are genotyped and phenotyped

• Heritability of phenotype is higher
  → genotype individuals that have high EBV accuracy
  + use deregressed EBV

• Selection candidates are more related to training data
  → include parental generation in training

• How important is it to achieve a ‘target’ accuracy of genomics right from the start?
  • You’re going to need to retrain anyway?
  • Build-up training data during GS implementation
To Retrain or Not to Retrain
Results from Stochastic Simulation

Hong-hua Zhao, Jennifer Young, David Habier, Rohan Fernando, Jack Dekkers
(unpublished)
Response from Genomic Selection - Simulation

Generation

0
. . .
. . .
. . .
1000
1001
1002
. . .
1012

20 chr of 150 cM
100,000 SNPs (freq. = ½ , LE)

Random mating, $N_e = 100$
LD generated by drift and mutation

Allocate 100 loci with MAF>0.1 as QTL and 2,000 as SNPs
Expand pop.size to 1,000 – phenotype - $h^2 = 0.3$
Estimate marker effects by Bayes-B
Mate random 20 males to random 60 females

Select 20/240 males 60/240 females

Strategies
BLUP-1 = PBLUP – last phenotypes collected in G1001
BLUP-all= PBLUP – continuous phenotyping
GS-1 = Bayes-B GS – without retraining – no P after G1001
Accuracy of EBV

- 100 QTL GS-1
- 100 QTL GS-all
- 100 QTL BLUP-all
- 100 QTL BLUP-1
- 100 QTL GS-1 no selection
- 100 QTL GS-all no selection

Generation

BLUP

GS

GS-all

BLUP-all

GS-1

BLUP-1

Accuracy of EBV
The graph illustrates the Genetic Variance over generations for different methods:

- **100 QTL GS-1**
- **100 QTL GS-all**
- **100 QTL BLUP-1**
- **100 QTL BLUP-all**

Each line represents the genetic variance decrease over generations, with the y-axis showing the variance from 0.40 to 0.00 and the x-axis representing generations from 1001 to 1013.

Accuracies averaged over 16 traits

Size of training and validation data

<table>
<thead>
<tr>
<th>Generation</th>
<th># geno-typed</th>
<th># with own record</th>
<th># progeny with genotyped parents Early</th>
<th>Validation data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1</td>
<td>777</td>
<td>295</td>
<td>2443</td>
<td>322</td>
</tr>
<tr>
<td>2</td>
<td>1215</td>
<td>618</td>
<td>4892</td>
<td>295</td>
</tr>
<tr>
<td>3</td>
<td>1628</td>
<td>913</td>
<td>7562</td>
<td>357</td>
</tr>
<tr>
<td>4</td>
<td>2108</td>
<td>1273</td>
<td>9319</td>
<td>274</td>
</tr>
<tr>
<td>5</td>
<td>2708</td>
<td>1563</td>
<td>11486</td>
<td>262</td>
</tr>
</tbody>
</table>

Graph showing accuracies over generations for G-GLUP and P-GLUP models.
### Size of training and validation data

<table>
<thead>
<tr>
<th>Generation</th>
<th># genotyped</th>
<th># with own record</th>
<th># progeny with genotyped parents Early</th>
<th>Validation data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1</td>
<td>777</td>
<td>295</td>
<td>2443</td>
<td>322</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>295</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>357</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>274</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>262</td>
</tr>
</tbody>
</table>

#### Need for Retraining

Wolc et al. (GSE, 2011)

---

**Train on data prior to generation 1**

---

![Graph showing accuracy over generations for different models: Bayes-A, BayesC-pi, G-BLUP-bivar, G-BLUP, P-BLUP-bivar, P-BLUP. The graph highlights the expected accuracy for P-BLUP and G-BLUP models.](graph.png)
How to build a Reference Population for a Closed Breeding Population?

Accuracy of GEBV is greater if

- More individuals are genotyped and phenotyped
- Heritability of phenotype is higher
  → genotype individuals that have high EBV accuracy
    + use deregressed EBV
- Selection candidates are more related to training data
  → include parental generation in training

- How important is it to achieve a ‘target’ accuracy of genomics right from the start?
  - You’re going to need to retrain anyway?
  - Build-up training data during GS implementation
Strategy for Implementation of Genomic Selection within a breed/line (for Pigs & Poultry?)

1. Genotype $\geq 3$ generations of parents with HD panel
   - Use for initial training

2. Genotype selection candidates with ELD panel
   - Impute HD genotypes and compute G-EBV for selection

3. Re-genotype selected sires (and dams) with HD panel

4. Retrain with new data on LD/HD-genotyped animals
Reference Population for ‘New Traits’
when # phenotypes is limited and genotyping is not
Genotype individuals with phenotype, rather than parents
Grevenhof, Bijma, van Arendonk GSE 2012

\[ N_E = 100, L = 30, \text{ and } h^2 = 0.3. \]
Bastiaansen et al. GSE ‘12

Deep vs. Shallow Reference pop.

N=500 in reference
- In 1 generation
- or across 5 gens.

Accuracy of EBV averaged over 30 replicates

No Retraining

Low uneq = low # QTL (30) unequal variance

Low eq. = low # QTL (30) equal variance

High uneq = high # QTL (300) unequal variance

High eq. = high # QTL (300) equal variance

Shallow has advantage only in first generation

Reference population: Deep Shallow

Bayes
Which individuals should be entered into central test stations?

Potential bull dams? Konig and Swalve JDS 2009

Limited gain in accuracy EBV of bulldams with addition of own record

Figure 2. Correlation between index and aggregate genotype ($r_{TI}$) for scenario I by altering the heritability of the trait and the correlation between the true breeding value and genomic EBV ($r_{mg}$). Dotted line: index without genomic information; solid line with open diamonds: $r_{mg} = 0.5$; solid line with solid squares: $r_{mg} = 0.7$; solid line with open triangles: $r_{mg} = 0.9$. 

$r_{MBV} = 0.8$
$r_{MBV} = 0.7$
$r_{MBV} = 0.5$
$r_{MBV} = 0$
Low density genotyping and Imputation

Jack Dekkers

Animal Breeding & Genetics
Department of Animal Science
Iowa State University
Implementing GS in Pig/Poultry Programs

Problem
High cost of genotyping \(\leftrightarrow\) value of an individual
Very large numbers of selection candidates

Impossible to implement genomic selection based on high density genotyping in cost efficient manner

Solution
Combination of strategic genotyping and imputation
Information used for imputation

• LD across the population
  – To impute from medium density (>10,000 SNP) to high density – up to sequence

• Linkage within families
  – To impute from very low density (<1000 SNP) to high density
Imputation using population-wide LD

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD-genotype</td>
<td>ACAAGGATTCCGAT</td>
</tr>
<tr>
<td>HD-genotype</td>
<td>GCTATCATGCTAT</td>
</tr>
<tr>
<td>LD-genotype</td>
<td>--T-------A-------T</td>
</tr>
</tbody>
</table>
Imputation using population-wide LD

Haplotype

HD-genotype

ACGAGGATTCGAT

HD-genotype

GCTATCATGCTCTAT

LD-genotype

--T---A-----T--
Imputation using population-wide LD

Haplotypes

HD-genotype

HD-genotype

LD-genotype

AC\textcolor{Green}{A}G\textcolor{Orange}{G}G\textcolor{Aqua}{A}T\textcolor{Red}{T}C\textcolor{Blue}{C}G\textcolor{Magenta}{A}T

G\textcolor{Red}{C}T\textcolor{Orange}{A}T\textcolor{Green}{C}A\textcolor{Blue}{T}G\textcolor{Magenta}{C}C\textcolor{Orange}{T}AT

G\textcolor{Red}{C}T\textcolor{Orange}{A}T\textcolor{Green}{C}A\textcolor{Blue}{T}G\textcolor{Magenta}{C}C\textcolor{Orange}{T}AT
Information used for imputation

• LD across the population
  – To impute from medium density (>10,000 SNP) to high density – up to sequence

• Linkage within families – as explained before
  – To impute from very low density (<1000 SNP) to high density
Imputation based on Linkage Information

**Requirements:**

- Ordered/phased HD SNP genotypes of parents
- Imputation of HD SNP genotypes on progeny

\[
\text{HD-GS} \Rightarrow \text{EBV}_i = \sum_{\text{SNP } k} \left( g^m_{ik} + g^p_{ik} \right)
\]

\[
\text{LD-GS} \Rightarrow \text{EBV}_i = \sum_{\text{SNP } k} \left( p^{md}_{ik} g^m_{dk} + p^{pd}_{ik} g^p_{dk} + p^{ms}_{ik} g^m_{sk} + p^{ps}_{ik} g^p_{sk} \right)
\]

- Probability that \( i \) received dam’s maternal allele at SNP \( k \)
- \( \text{PDM} = \text{Probability of Descent of Marker allele} \)
Accuracy of G-EBV based on High- vs Low-Density SNP genotyping

Simulation (Habier et al. 2009 Genetics)

With re-genotyping of parents using HD panel

Training
Imputation results in HyLine data

Neil O’Sullivan, Janet Fulton, Petek Settar and Jesus Arango

HY-LINE INTERNATIONAL

Anna Wolc, David Habier, John Hickey, Mehdi Sargolzaei, Dorian Garrick, Rohan Fernando, Nathan Bowerman, Chunkao Wang, Jack Dekkers,

IOWA STATE UNIVERSITY, USA
POZNAN UNIVERSITY OF LIFE SCIENCES, POLAND
UNIVERSITY OF NEW ENGLAND, AUSTRALIA
UNIVERSITY OF GUELPH, CANADA
Hy-Line data

- 8 generations of HD sires and dams
- Selection candidates: 544 individuals from generation 9

- **High Density genotypes** - 4,893 segregating SNPs on chromosome 1

- **Low Density genotypes** – Simulated panel of 73 ~equally spaced SNPs (equivalent to ~400 SNP across the genome)
Accuracy of imputed genotypes in generation 9 with 8 generations of sires and various generations of dams HD genotyped, and the remaining generations of dams Low Density genotyped.
Genomic Selection using Low-Density SNPs

Conclusions

GS can be implemented by genotyping selection candidates for <400 SNPs spread across the genome

- Loss in accuracy limited: < 5 % - if parents re-genotyped HD - sufficient to genotype only sires
- Cost effectiveness depends on cost of Low- vs. High-density genotyping
  $20 \leftrightarrow ?? \rightarrow $150
- Loss in accuracy ~ independent of # QTL and # traits
- LD-genotyped individuals can also be used for training
- Allows imputing to higher densities / sequence from founders