Genomic Breeding Programs and Validation

Ardidale Summer Course 2015
Ben Hayes and Hans Daetwyler
Course Outline

• Day 1
  – Introduction
  – Generation, quality control, alignment of sequence data
  – Detection of variants, quality control and filtering

• Day 2
  – Imputation from SNP array genotypes to sequence data

• Day 3
  – Genome wide association studies with SNP array and sequence variant genotypes

• Day 4 & 5
  – Genomic prediction with SNP array and sequence variant genotypes (BLUP and Bayesian methods)
  – Use of genomic selection in breeding programs
Day 5

- Validation – traps for young players!
- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
Validation of genomic selection

• Aim of genomic selection
  – predict (young) selection candidates without phenotypes

• How to test or validate predictions?

• Test predictions in a population sample that is similar to selection candidates

• Key principle of validation
  – Independence of reference and validation populations
Validation - Accuracy of genomic prediction

Estimate Genomic Predictions

Reference Population
With genotypes and phenotypes

Predict

Validation Population
With genotypes and phenotypes

Predict

Selection Candidates
without Phenotypes

Calculate accuracy as the correlation between genomic breeding values and highly accurate breeding values or phenotypes.
Accuracy and bias

• Most commonly used:
  – $r=\text{correlation}(\text{GEBV, phenotypes})$
  – Gives accuracy of a group of individuals

• Individual accuracy
  – Calculated using the prediction error variance from the diagonal of the coefficient matrix (GBLUP)

• Regression of phenotypes ($y$) on GEBV ($x$)
  – Deviation from expectation of the slope
  – Expectation is usually 1
  – If not close to expectation $\Rightarrow$ then biased
Standard error of a correlation

• Correlations have a standard error which depends on sample size and the magnitude of the correlation

• An approximation of this standard error was given by Fisher (see Fisher z transform)
  – $SE \sim \frac{1}{\sqrt{N-3}}$

• In our practical examples
  – 31 individuals
  – $SE = \frac{1}{\sqrt{31-3}} = 0.189$
Two main ways to (cross)-validate

1\textsuperscript{st} way: Highly accurate individuals
- Dairy bull progeny test (e.g. Daughter trait deviations)
- Very large progeny groups or many clones (plant replication)
- Step1: Estimate marker effects in reference population
- Step2: Predict highly accurate individuals and calculate accuracy

2\textsuperscript{nd} way: ‘Classic’ cross-validation
- Step 1: Divide dataset into \( n \) subsets of individuals
- Step 2: Predict each subset using all other subsets
- Step 3: Calculate accuracy in each subset and take mean across all subsets
Approximating the accuracy of true breeding value

- The upper limit of genomic selection accuracy is given by the accuracy of observations.

- Divide by accuracy of observation to approximate accuracy of additive genetic component (i.e. breeding value).

- If using DYD (daughter yield deviations, “daughter means”)
  - \( r/\text{accuracy(DYD)} \)

- If using phenotypes
  - \( r/\sqrt{h^2} \)
Validation - Independence

• Always ask question:
  – If the validation individuals were selection candidates what data would be available?
  – Then only use that data for reference!

• Independence of ‘data’, not independence in relationship
Independence

• Validation individuals are not used in the reference pop

• Validation phenotypes do not contribute to observed variables of reference pop
  – E.g. excluded when calculating estimated breeding values

• Validation individuals do not have contemporaries of same age in reference
Independence

• Choosing a subset of SNP with a GWAS
  – Only use reference population to choose SNP
  – If validation population is used for GWAS then you are overfitting (upward bias in accuracy)
Independence

Target of prediction

• Validation population should be similar to selection candidates
• Similar relationship to reference as selection candidates
  – Same number of generations removed
  – Same breeds
  – Same population
• Same SNP density
  – Consider imputation error
Day 5

- Validation – traps for young players!
- **Design of reference populations for Genomic selection**
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
Genomic Selection – Breeding Programs

Reference Population
- Genotypes
- Phenotypes

Selection Candidates
- Genotypes

Prediction Equation
Genomic Breeding Value = \( W_1X_1 + W_2X_2 + W_3X_3 \ldots \)

Selected Parents
- Estimated breeding values
Genomic Selection – Breeding Programs

Reference Population
- Genotypes
- Phenotypes

Prediction Equation
Genomic Breeding Value = $w_1x_1 + w_2x_2 + w_3x_3 \ldots \ldots$

Selection Candidates
- Genotypes

Selected Parents
- Estimated breeding values

Genetic Gain
$\Delta G = \frac{ir\sigma_g}{L}$

Selection Intensity
Accuracy
Genetic SD
Generation Interval
Cycle Time
Genomic Selection

- Useful for traits where variation is contributed by large number of loci, e.g. yield.

- Large benefit for traits that are difficult/expensive to measure, measured late in breeding cycle

- Accelerate genetic gain by reducing generation interval, increasing accuracy of selection and selection intensity
Factors affecting genomic prediction accuracy

• Reference population size \((N_p)\)
  – Genotyped and phenotyped

• Heritability \((h^2)\)

• Genetic diversity - number of effective chromosome segments \((M_e)\)
  – Effective population size \((N_e)\)
    • Linkage disequilibrium (LD)
  – Genome length

• Proportion of genetic variance captured by markers (\(~\text{marker density}\))

\[
r = \sqrt{\frac{N_P h^2}{N_P h^2 + M_e}}
\]
Reference populations for GS

• Also called “training sets”

• Two principles for design

  1) Make it large -> QTL effects are small!

  2) Make it close to candidates for selection
Reference populations for GS

• How large?
Reference populations for GS

Parameters affecting accuracy of genomic breeding values

- \( N_p \)  Size of reference population
- \( h^2 \)  Heritability of trait
- \( Me \)  Number of independent chromosome segments
  
  • Daetwyler et al. (2008, 2010), Goddard (2008)

\[
r = \sqrt{\frac{N_p h^2}{N_p h^2 + Me}}
\]
Reference populations for GS

• Parameters affecting accuracy of genomic breeding values
  – $N$  Size of reference population
  – $h^2$  Heritability of trait
  – $Me$  Number of independent chromosome segments
    • Also called number of effective loci affecting the trait
Reference populations for GS

- Number of loci affecting the trait
  - Conservative assumption - quantitative traits affected by very large number of loci, normal distribution of effects
  - = number of independent chromosome segments
Reference populations for GS

- Number of loci affecting the trait
  - Conservative assumption - quantitative traits affected by very large number of loci, normal distribution of effects
  - = number of independent chromosome segments
Reference populations for GS

• Number of loci affecting the trait
  – Conservative assumption - quantitative traits affected by very large number of loci, normal distribution of effects
  – = number of independent chromosome segments

\[ Me = 2N_e L \]

• \( N_e \) = effective population size, \( L \) is genome length in Morgans
Reference populations for GS

• accuracy of genomic breeding values

\[ r = \sqrt{\frac{N_P h^2}{N_P h^2 + M_e}} \]

• Number of chromosome segments
  – \( M_e = 2N_e L \)
Real Data

• Dairy cattle (Holsteins)
  ➢ USA results (N=1000-6700) for Net Merit Index (VanRaden et al. 2009)
  ➢ Australian results (N=1100-3300) for Australian Profit Ranking
  ➢ $h^2 = 0.9$
  ➢ $N_e = 100$

• Accuracies $r(\text{GEBV}, \text{EBV})$ in validation data sets
Deterministic prediction vs. Holstein data

Accuracy of genomic breeding value vs. Number of bulls in reference population

- Predicted Daetwyler et al. (2008)
- US Holstein data
- Australian Holstein data

Hayes et al., 2009, AAABG
Deterministic prediction
Reference populations for GS

• Which individuals/lives?

• The relationship of the reference population to the selection candidates affects accuracy of GEBV
Influence of relationships on GS accuracy

- Relationship of validation to reference important contributor to accuracy
Relationship Effects on Accuracy

• Being highly related to at least 1 line in reference pop is more important than having a high mean relationship to reference

Daetwyler et al., 2014. TAG
Reference populations for GS

• Which individuals/lines?

• The relationship of the reference population to the selection candidates affects accuracy of GEBV

• Need individuals close to those being predicted in reference

• At the same time, as diverse as possible so that many individuals/lines can be accurately predicted
Population relatedness/diversity has great effect on accuracy ($N_p = 1000, h^2=0.5$)

Lin et al. 2014, Crop and Pasture Science
Ryegrass Cultivar versus Livestock Diversity?

- Calculate Fixation Indices:
  - Ryegrass
    - Italian and perennial
  - Cattle
    - Dairy and beef
  - Sheep
    - Wool, meat and maternal breeds
Across Cultivar/Breed Prediction?

<table>
<thead>
<tr>
<th>Species</th>
<th>Mean $F_{st}$</th>
<th>Min $F_{st}$</th>
<th>SD $F_{st}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian Ryegrass</td>
<td>0.06</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Perennial Ryegrass</td>
<td>0.13</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Italian to Perennial</td>
<td>0.27</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td>Cattle</td>
<td>0.08</td>
<td><strong>0.02</strong></td>
<td>0.03</td>
</tr>
<tr>
<td>Sheep</td>
<td>0.06</td>
<td><strong>0.02</strong></td>
<td>0.02</td>
</tr>
</tbody>
</table>
Expectations of Accuracy Depending on Reproductive System in Plants

• Inbreeding plants
  • Higher levels of linkage disequilibrium (LD)
  • Lower $N_e$
  • $\rightarrow$ population level genomic selection

• Outbreeding plants
  • Low LD in populations, but still high LD within family
  • High $N_e$ in populations $\rightarrow$ need extremely large reference populations
  • but low $N_e$ within family
  • $\rightarrow$ ‘family’ genomic selection designs
Day 5

• Validation – traps for young players!

• Design of reference populations for Genomic selection

• How many markers?

• How often to re-estimate SNP effects?

• Optimal breeding program design with genomic selection

• Genomic selection and inbreeding
How many markers?

• $10 \times Ne \times L$
  - $Ne = \text{effective population size}$
  - $L = \text{length of genome (Morgans)}$
  - Meuwissen et al. (2009) GSE

  - E.g. Holsteins
  - $10 \times 100 \times 30 = 30,000$
Genomic Predictions Residual Feed Intake

- Collaboration DPI Vic, Livestock Improvement Corporation and Dairy NZ (Richard Spelman, Kevin MacDonald, et al.)
- 1000 heifers each
- Genotyped 800,000 SNPs (Illumina Bovine HD)
Genomic predictions

Prediction Equation: \[ \text{RFI} = x_1 + x_2 + x_3 + x_4 + \ldots + x_{800,000} \]
Genomic Predictions Residual Feed Intake

• To derive prediction equation

• GBLUP -> all markers have small, non zero effect

• BayesR -> proportion of markers have zero effect, rest have small to moderate effects
<table>
<thead>
<tr>
<th>Trait</th>
<th>Marker Panel</th>
<th>GBLUP</th>
<th>BayesR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liveweight</td>
<td>50K</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>800K</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>Residual Feed Intake</td>
<td>50K</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>800K</td>
<td>0.29</td>
<td>0.41</td>
</tr>
</tbody>
</table>
Day 5

- Validation – traps for young players!
- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Use of sequence data in genomic prediction
Genomic selection

• How often to re-estimate SNP effects?
  – If the markers used in genomic selection were actually the underlying mutations causing the QTL effects, the estimation of SNP could be performed once in the reference population.
    • Promise of sequence data

  – GEBVs for all subsequent generations could be predicted using these effects.
Genomic selection

- How often to re-estimate SNP effects?
  - In practise will be markers with low to moderate levels of $r^2$ with the underlying mutations (QTL)
  - Do not capture all of QTL variance
  - Over time, recombination between the markers and QTL will reduce the accuracy of the GEBV using SNP effects from the original reference population.
  - We need to re-estimate SNP effects
  - How often?
Genomic selection

• How often to re-estimate SNP effects?

Table 4.3. The correlation between estimated and true breeding values in generations 1003–1008, where the estimated breeding values are obtained from the BayesB marker estimates in generations 1001 and 1002. From Meuwissen et al. (2001).

<table>
<thead>
<tr>
<th>Generation</th>
<th>$r_{TBV:EBV}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1003</td>
<td>0.848</td>
</tr>
<tr>
<td>1004</td>
<td>0.804</td>
</tr>
<tr>
<td>1005</td>
<td>0.768</td>
</tr>
<tr>
<td>1006</td>
<td>0.758</td>
</tr>
<tr>
<td>1007</td>
<td>0.734</td>
</tr>
<tr>
<td>1008</td>
<td>0.718</td>
</tr>
</tbody>
</table>

The generations 1004–1008 are obtained in the same way as 1003 from their parental generations.
Genomic selection

- Depending on trait and population, the decay of accuracy may be partly dependant on genomic selection method .....  
  - Mainly true for traits with few QTL

Habier et al., 2007, Genetics
Genomic selection

• In practice, breeders re-estimate marker effects often

• Whenever new phenotypes or genotypes can be added to reference population
Day 5

• Validation – traps for young players!

• Design of reference populations for Genomic selection

• How many markers?

• How often to re-estimate SNP effects?

• Optimal breeding program design with genomic selection

• Genomic selection and inbreeding
Optimal breeding program design

• With genomic selection, we can potentially predict GEBV with an accuracy of 0.8 for selection candidates at birth

• How does this change the optimal breeding program design?
Optimal breeding program design

• With genomic selection, we can potentially predict GEBV with an accuracy of 0.8 for selection candidates at birth

• How does this change the optimal breeding program design?

• Breed from individuals as early as possible
Optimal breeding program design

• In dairy cattle, recent structure was
  – Each year select a team of calves to form a progeny test team
  
    – At one years of age these bulls are mated to ‘random’ cows from the population
  
    – At four years of age the daughters of the bulls start lactating
Optimal breeding program design

• In dairy cattle, recent structure was
  – Each year select a team of calves to form a progeny test team
  – At one year of age these bulls are mated to random cows from the population
  – At four years of age the daughters of the bulls start lactating
  – At five years of age the bulls receive a progeny test “proof” based on the performance of their daughters
  – The bulls are then selected on the basis of these proofs to be “breeding bulls”
    • Semen sold to commercial farmers
Optimal breeding program design

• In dairy cattle with genomic selection..
  – Genotype a large number of bull calves from the population
  – Calculate GEBVs for these calves
    • Accuracy ~ 0.7 = accuracy of progeny test
  – Select team based on GEBV
  – Sell semen from these bulls as soon as they can produce it
Optimal breeding program design

• In dairy cattle with genomic selection.
  – Genotype a large number of bull calves from the population
  – Calculate GEBVs for these calves
    • Accuracy ~ 0.7 = accuracy of progeny test
  – Select team based on GEBV
  – Sell semen from these bulls as soon as they can produce it

  – Generation interval reduced from ~4 yrs to ~ 2 yrs
    • $\Delta G = ir\sigma_g / L$
  – Double rate of genetic gain
Optimal breeding program design

- In dairy cattle with genomic selection..
  - Genotype a large number of bull calves from the population
  - Calculate GEBVs for these calves
    - Accuracy ~ 0.7 = accuracy of progeny test
  - Select team based on GEBV
  - Sell semen from these bulls as soon as they can produce it
    - Generation interval reduced from ~4 yrs to ~ 2 yrs
      - $\Delta G = \frac{ir\sigma_g}{L}$
    - Double rate of genetic gain
    - Save the cost of progeny testing!
      - Reduce costs by 92% (Schaeffer et al. 2006)
Genomic selection: Dairy cattle

\[ \Delta G = \frac{ir\sigma_g}{L} \]
Genomic selection: Meat sheep

Gains can be made by selection for breeding objective traits directly, e.g. Lean meat yield vs. scanned eye muscle area

\[ \Delta G = \frac{ir\sigma_g}{L} \]
### Increased genetic gain from genomic selection

<table>
<thead>
<tr>
<th>Industry</th>
<th>Potential increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dairy Cattle</td>
<td>60-120% (Pryce et al. 2011)</td>
</tr>
<tr>
<td>Meat sheep</td>
<td>21% (van der Werf 2011)</td>
</tr>
<tr>
<td>Wool sheep</td>
<td>38% (van der Werf 2011)</td>
</tr>
<tr>
<td>Beef cattle</td>
<td>29-158% (Van Eenennaam 2011)</td>
</tr>
<tr>
<td>Layers</td>
<td>40% (Dekkers et al 2009)</td>
</tr>
<tr>
<td>Broilers</td>
<td>20% (Dekkers et al. 2009)</td>
</tr>
</tbody>
</table>
Optimal breeding program design

- Synergy with reproductive technologies
- If we can predict genetic gain accurately at birth, genetic gain depends on generation interval
- Reproductive technologies to reduce this
  - Juvenile in-vitro embryo transfer?
  - Extreme technologies like in-vitro meosis
- Must manage inbreeding!!
Optimal breeding program design

Exploring Genomic Selection in Perennial Ryegrass Breeding Programs Using Simulation

**Aim**
Investigate genetic gain achieved with genomic selection in perennial ryegrass breeding programs using computer simulations:

1. Simulate a conventional ryegrass breeding program
2. Replace phenotypic selection with genomic selection
3. Compare the genetic gain and inbreeding between the two breeding strategies
Breeder visual preference ($\text{BVP, } h^2=0.2$)

Flowering time ($h^2=0.6$)

Persistency ($h^2=0.1$) & yield ($h^2=0.3$) in plots
Genomic Breeding of Ryegrass

- Genomic prediction on persistency & yield
  - Reference from plots

- Flowering time
  - Reference from clonal rows

- Genomic prediction on persistency & yield
  - Reference from plots
Genomic Selection: Simulation Strategy

- 10 cycles, 50 independent replicates

Base Population → Varieties Pool → Breeding Program

Phenotypic Selection

C1 → C2 → C3 → C4

Equivalent to Historical Breeding Program

Phenotypic Selection

C5 → C6 → C7 → C8 → C9 → C10

Genomic Selection

C5 → C6 → C7 → C8 → C9 → C10
Accuracy of Genomic Selection for Ryegrass

- Increasing accuracy with increased reference population size
- Plot accuracy lower than spaced plants (due intense selection at spaced plant stage)
- Higher accuracy of genomic selection for traits measured in spaced plants

Accuracy persistency & yield (reference from plots)

Accuracy flowering time (reference from clonal rows)
Genetic Gain from Ryegrass Genomic Selection

- Genetic gain in base (cycle 4) genetic standard deviations for phenotypic and genomic selection over 30 years.
Optimal breeding program design

• In plants, it may be possible to quickly turn over generations of breeding in glasshouse
  – Use GEBVs to select without phenotyping

• Need to ensure that reference population is not too many generations removed from selection candidates

• Could select best, but still plant out a proportion of them for phenotyping to update reference population

• Accuracy of GEBVs needs to be high for such schemes (~ > 0.7)
Day 5

• Validation – traps for young players!

• Design of reference populations for Genomic selection

• How many markers?

• How often to re-estimate SNP effects?

• Optimal breeding program design with genomic selection

• Genomic selection and inbreeding
Components of a breeding value

• Breeding values of sire and dam
• Mendelian sampling term
  – Deviation due to sampling of alleles from parents
Using genetic markers

• Markers explain some within family variance
  – Give information on Mendelian sampling term

• Genomic selection can estimate Mendelian sampling term very accurately
  – Due to many markers
  – Expected versus ‘realised’ relationships

Methods’ utilisation of components

- Assume a juvenile without phenotypes
- Genomic selection uses Mendelian sampling term
- Selection can act on whole breeding value!
Measures of inbreeding/diversity

• Rate of inbreeding per generation
  – Appropriate for comparing methods
  – Counteracting forces also occur per generation
    • Mutation

• Rate of inbreeding per year
  – Relevant for breeding programs

• Diversity
  – Relationship among lines/individuals
  – Number of SNP
  – Genetic variance of traits
Effect of genomic selection on inbreeding

• Example: 4 young elite full brothers
  – Pedigree breeding values (BLUP) are the parent average → the same for all 4 bulls
    • Select all 4
  – GEBV will be different for all 4
    • Only select best

• Genomic selection results in less inbreeding *per generation* than BLUP
  – Selection on Mendelian sampling terms
  – Reduced co-selection
  – Breeding values of sibs less correlated
Example: GS and Diversity in Wheat

Rutkoski et al, 2015 Plant and Animal Genomes

• 2 year **in field** selection experiment in wheat
  - Two traits (Stem rust and pseudo-black chaff)
  - Compared phenotypic and genomic selection
  - Predict GEBV for 252 wheat lines from reference of 374 lines, select 5 and intermate

• Phenotypic selection slightly higher gain (within SE)

• Genomic selection reduced genetic variance more than phenotypic
Inbreeding of genomic selection breeding programs

• Genomic selection can drastically reduce generation intervals

• Increases genetic gain per year but also increases rate of inbreeding (loss of diversity) per year

• Need for controlling inbreeding:
  – Mate selection for less inbred progeny
  – Optimum contribution selection
    • Maximise genetic gain at a given level of inbreeding
  – Use a diverse set of sires
  – Highly likely that elite sires change every year
Genomic breeding in plants

Diversity in plant species greater than diversity in important agriculture animal species.

- Mammals and birds diploid

- Many plants polyploid, different ploidies in one species

- Different reproductive systems
  - Inbreeding versus outbreeding
  - Some outbreeding plants extremely diverse (Ne 10k +)
Plant reference genomes

If diploid and economically important: reference genome well progressed
  – maize

If polyploid and economically important: reference genomes progressing
  – Wheat, canola, ...

If diploid, polyploid and minor species: variable progress by few groups
  – Need to do it yourself!

*Status of reference genomes a limitation to plant genotyping technologies.*
Genomic selection in plant breeding programs

Want accurate estimated breeding values (GEBV) for selection

- Test many selection candidates
- Remove lowest GEBV from field trials
  - Achievable with intermediate accuracies (0.3 - 0.7)
- Pick elite parents for crossing early
  - Need higher accuracy > 70%
- Directed selection in resource population of crosses
- Deliver genetic gain in hard to measure trait

Another way of saying the same thing

- Reduce breeding cycle time
- Increase accuracy of breeding value
- Reduce phenotyping
- Increase number of selection candidates and selection intensity
Day 5

- Validation – traps for young players!
- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Use of sequence data in genomic prediction