Accuracy of Genomic Prediction

Julius van der Werf
and Sang Hong Lee
Genomic Prediction: basic idea

To predict a trait EBV at a young age,

- good for:
  - late traits
  - hard to measure traits
Genomic prediction accuracy

- Derive from the model, e.g. PEV from GBLUP mixed model equations

- Validate with other EBVs or phenotypes
  - Validation population
  - Cross-validation

- Predict in advance based on theory and assumptions about population
Genomic Prediction: basic idea

1) Somebody (else) measures lots of sheep, and their DNA → Reference population

2) A breeder tests DNA on young rams

Illustrating (dis-)similarity of chromosome segments
Genotype information

Father

101001110111001101110011
010100111100011000110010

Mother

000100111001010100110111
1010111011111111111110

Chromosome segments are passed on

Progeny

101001110111001101110011
000100111100011000110011

Within a population, members will share chromosome segments
We can follow inheritance via SNPs
Degree of sharing can be represented in a genomic relationship (= observed based on SNPs)
(similar to genetic relationship = expected based on pedigree)
Genomic Prediction: basic idea

1) Somebody (else) measures lots of sheep, and their DNA → Reference population

2) A breeder tests DNA on young rams

Large diversity of segments → less accuracy
populations of haplotypes

Holstein Friesian, a pig/poultry nucleus

Limited diversity
Long segment sharing

Smaller $N_e$, longer segment sharing, fewer “effective loci”

Merino sheep, humans

More diversity
Short segment sharing
Sub populations

Not only recent $N_e$ but also historic $N_e$ is relevant
Genomic prediction accuracy

Design parameters

- Effective population size ($N_e$)
- Effective # chromosome segments ($M_e$)
- Sample size in reference data ($N$)
- Heritability ($h^2$)
Genomic prediction accuracy \textit{Using Daetwyler et al, 2008}

Accuracy$^2$ of estimating a random effect $= n / (n + \lambda)$ \hspace{1cm} $\lambda = \frac{V_e}{V_a}$

If genome exists of $M_e$ independently segregating ‘effective chromosome segments’

And each segment has variance $VA / M_e$, then accuracy$^2$ of estimating each segment

$$\frac{N}{N + V_e / (V_a / M_e)} = \frac{NV_a}{NV_a + V_e M_e} = \frac{h^2}{h^2 + \frac{M_e}{N}}$$

$$r_{g, \hat{g}} = \sqrt{\frac{h^2}{h^2 + \frac{M_e}{N}}}$$

$N$ = nr observations
$M_e$ = effective nr loci

Valid if “all genetic variance is captured by markers”
See also Dekkers 2007 (Path coefficient method)

\[ P \leftarrow h \rightarrow G \leftarrow \sqrt{1-h^2} \rightarrow E \]
\[ G \leftarrow q \rightarrow Q \leftarrow \sqrt{1-q^2} \rightarrow R \]
\[ Q \leftarrow r_Q \rightarrow \hat{Q} \leftarrow \sqrt{1-r_Q^2} \rightarrow e \]

Trait heritability = $h^2$

G = total BV
Q = genetic effects captured by marker(s)
R = residual polygenic effects

Model for phenotype: $P = G + E$
Model for BV: $G = Q + R$
Genomic prediction accuracy Using Goddard et al, 2011

Depends on

i) Proportion of genetic variance at QTL captured by markers $q^2$

ii) Reliability of estimating marker effects $r^2_{Qhat}$

Accuracy = $\sqrt{q^2 \cdot r^2_{Qhat}}$

= $q \cdot r_{Qhat}$
Genomic prediction accuracy  
Using Goddard et al, 2011

Depends on

i) Proportion of genetic variance at QTL captured by markers

\[ q^2 = \frac{M}{M_e + M} \]

Depends on marker-QTL LD

Depends on

M = # markers

M_e = ‘effective number of chromosome segments’

i) Accuracy of estimating marker effects
Genomic prediction accuracy

Using Goddard et al, 2011

Depends on

i) Proportion of genetic variance at QTL captured by markers

\[ q^2 = \frac{M}{(M_e + M)} \]

Depends on marker-QTL LD

M = # markers

M_e = ‘effective number of chromosome segments’

Depends on

i) Accuracy of estimating marker effects

\[ r^2_{Qhat} = \frac{V_{qhat}}{V_q} = \frac{N}{(N+\lambda)} \]

\[ \lambda = \frac{M_e}{(q^2 \cdot h^2)} \]

Accuracy = \( \sqrt{(q^2 \cdot r^2_{Qhat})} \)

= \( q \cdot r_{Qhat} \)

With very many markers

i) Proportion of genetic variance at QTL captured by markers

\[ q^2 = 1 \]

ii) Accuracy of estimating marker effects

\[ r_{Qhat}^2 = \frac{V_{qhat}}{V_q} = \frac{N}{N + \lambda} = \frac{h^2}{h^2 + M_e/N} \]

\[ \lambda = \frac{M_e}{h^2} \]

same as Daetwyler

Accuracy = \[ \sqrt{r_{Qhat}^2} = r_{Qhat} \]
\[ M_e \text{ is a function of } N_e \]

- \[ M_e = 2N_e \ln N_{\text{chr}} / \ln(4N_e L) \] (Goddard 2009)

- \[ M_e = 2N_e \ln N_{\text{chr}} / \ln(N_e L) \] (Goddard et al. 2011)

- \[ M_e = 2N_e \ln N_{\text{chr}} / \ln(2N_e) \] (Meuwissen et al. 2013)
Difference among the formulas

- $N_e = 500$, $L=1M$, $h^2 = 0.5$ and $N = 5000$,
- accuracy = 0.62, 0.58, 0.60
Validating ‘Effective number of segments’

Can use actual data on A and G to test this

Compare G and A matrices \[ G - A = D + E \]

\( D = \) deviation in relationship at QTL

\[ \text{Var}(D) = 1/M_e \]

\( E = \) error

\[ \text{Var}(E) = 1/nr \text{ Markers} \]

\[ M_e = 1/\text{var}(A_{ij}) \]

Given genomic relationships (after collecting data), it is possible to empirically get \( M_e \) from the data.
Simulation

- Coalescence gene dropping
  - \( N_e = 500 \) for 500 generations
  - \( L = 1 \) Morgan
  - \( N_{\text{chr}} = 30 \)
  - Recombination according to \( L \)
  - Mutation rate = \( 10^{-08} \)
  - \( N = 3000 \) in the last generation

- Estimate \( A_{ij} \) and obtain empirical \( M_e \)
Difference from empirical $M_e$

$h^2 = 0.5$ and $N = 5000$, accuracy = 0.62, 0.58, 0.60 vs. 0.82 (simulation)
Revisit the theory

\[ M_e = \frac{N_{chr}}{\left[ \ln(4N_eL +1) + 4N_eL(\ln(4N_eL +1) \quad 1)) \right] / (8N_e^2L^2) + \left(1/3N_e \right) \times (N_{chr} \quad 1) \] 

Assuming LD \( r^2 = 1 / (1 + 4N_e \times c) \)

\[ M_e = \frac{N_{chr}}{\left[ \ln(2N_eL +1) + 2N_eL(\ln(2N_eL +1) \quad 1)) \right] / (4N_e^2L^2) + \left(1/3N_e \right) \times (N_{chr} \quad 1) \] 

Assuming LD \( r^2 = 1 / (2 + 4N_e \times c) \)

For more detail, see a bioRxiv paper Lee et al, 2016
doi: http://dx.doi.org/10.1101/054494
Empirical $M_e$ and new formula

Agreed well

Main difference with previous work is due to accounting for covariance between chromosomes
Genomic prediction accuracy

\[
N_e = 1,000
\]

Expect very little improvement with denser markers
What effective population size?

*Hanwoo? ~ 94 (Gondro)*

Populations not homogeneous.

Within and between breed/line accuracies

Some accuracy due to population structure
How do we validate accuracy?

- Validation population
  - EBV (based on progeny test)
  - Phenotype
  - Is it a homogeneous group?

- Cross-validation
  - Across families
  - Random (also within families)
### Relationship with reference population

*Clark et al 2011*

<table>
<thead>
<tr>
<th>Method</th>
<th>Close</th>
<th>Distant</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ped 0 - 0.25</td>
<td>0 - 0.125</td>
<td>0 - 0.05</td>
</tr>
<tr>
<td></td>
<td>Genom 0.08 – 0.35</td>
<td>0.08 – 0.26</td>
<td>0.08 – 0.16</td>
</tr>
<tr>
<td>BLUP-Shallow pedigree</td>
<td>0.39</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>BLUP-Deep Pedigree</td>
<td>0.42</td>
<td>0.21</td>
<td>0.04</td>
</tr>
<tr>
<td>gBLUP</td>
<td><strong>0.57</strong></td>
<td>0.41</td>
<td><strong>0.34</strong></td>
</tr>
</tbody>
</table>

Additional accuracy from family info

'baseline accuracy': graphs predict 0.36 for Ne=100, N=1750, h²=0.3
Relatedness matters more if the reference population is smaller.
Using a stratified Reference population - populations are not homogeneous

Direct Relatives
Ne = 10
N = 50

Own Herd
Ne = 100
N = 500

Wider population
Ne = 1000
Relative importance

- $h^2 = 0.25$
- Data from smaller $N_e$ is more important

Upper: $N_e = 1000 + N_e = 10$ (N=500)
Middle: $N_e = 1000 + N_e = 100$ (N=500)
Lower: $N_e = 1000$ only

GBV accuracy vs Total # in reference population
Sample availability

- $h^2 = 0.25$
- $N_e = 10$ would have $< N = 100$ (maximum acc. = 0.73)
- $N_e = 100$ would have $< N = 1,000$ (maximum acc. = 0.81)
- $N_e = 1,000$ can have $N = 20,000$ (acc. = 0.83)
Composite design

- $h^2 = 0.25$
- Smaller $N_e$ is important with smaller total $N$
- Benefit from large $N_e$ too (0.78 to 0.89)

Upper: $N_e=1000 + N_e=100$ (N=500) + $N_e=10$ (N=50)
Lower: $N_e=1000$ only
Implication

- **Marker density**
  - For beef cattle or sheep, very dense markers (e.g. 600K) may not be cost-effective, compared to 50K
  - For $N_e = 1000$, accuracy is similar between 50K and 600K

- **Marker density is not a critical design parameter**
  - $> 50K$ with $N_e = 1000$ (livestock)
  - $> 200K$ with $N_e = 10,000$ (human)

- **But, it may matter with very large $N_e$**
  - Multi-breeds or multi-ethnicities
Implication

To maximise prediction accuracy

- give a priority to genotype reference sample of smaller $N_e$,
- e.g. close relatives > flocks (local, village) > states > country > ...
- When $h^2$ is lower, reference sample of smaller $N_e$ is more important

Note that $N_e$ can be changed, depending on the target sample
Implication

- To maximise prediction accuracy
  - Sample availability is much higher for larger $N_e$ (in terms of sample size)
  - e.g. close relatives $<$ flocks (local, village) $<$ states $<$ country $<$ ...

- Heterogeneous stocks are important as well
  - Unlimited source
    - Common SNP chips across breeds or ethnicities
    - Getting cheaper
Implication

- To maximise prediction accuracy
  - Composite design would be desirable
    - \( N_e=1000 \) (\( N=10,000 \)) + \( N_e=100 \) (\( N=500 \)) + \( N_e=10 \) (\( N=50 \))

- It may be useful if one can get the expected prediction accuracy before conducting an experiment. For example,
  - When adding a bunch of heterogeneous stocks to your data, how much the accuracy can be increased?
  - When adding a number of newly genotyped individuals, what accuracy can you expect?
  - And, what is the power?
Implication

MTG2
https://sites.google.com/site/honglee0707/mtg2

Given design parameters, MTG2 can provide the expected accuracy and power

See section 7 and 9 in the manual