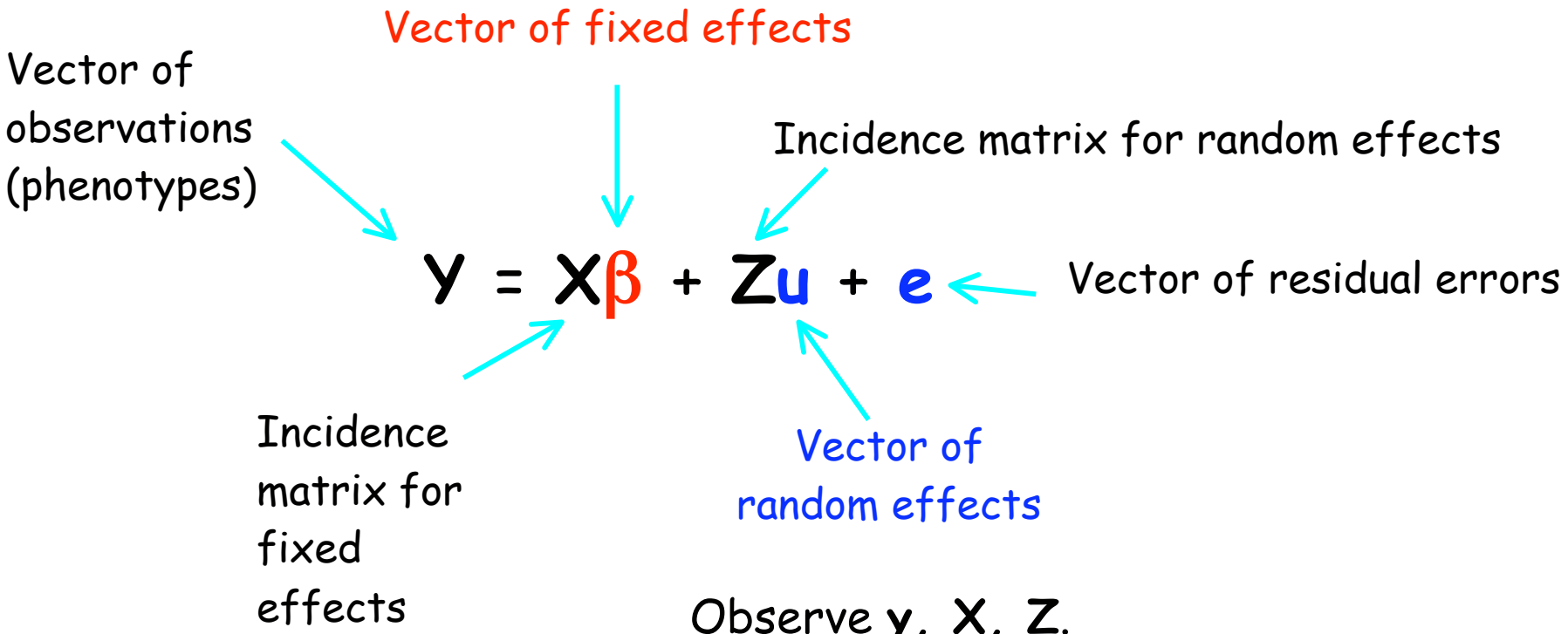
Lecture 28:
BLUP and Genomic Selection

Bruce Walsh lecture notes
Synbreed course
version 11 July 2013

BLUP Selection

- The idea behind BLUP selection is very straightforward:
- An appropriate mixed-model is constructed (such as the animal model) to estimate individual breeding values
  - These are called EBVs (estimated breeding values) or PBVs (predicted breeding values). The later because statisticians often speak of estimating fixed effects and predicting random effects.
  - Individuals with the largest EBVs are chosen for the next generation
  - The predicted response is simply the average of the EBVs in the selected parents.

Brief review: The general mixed model



Observe **y**, **X**, **Z**.

Estimate fixed effects β

Predict random effects **u**, **e**

Henderson's Mixed Model Equations

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$, $\mathbf{u} \sim (0, \mathbf{G})$, $\mathbf{e} \sim (0, \mathbf{R})$, $\text{cov}(\mathbf{u}, \mathbf{e}) = 0$,

If X is n x p and Z is n x q

p x p p x q

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

q x pq q x q

The whole matrix is (p+q) x (p+q)

Easier to numerically work with than BLUP/BLUE equations

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)$$

$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$

Inversion of an n x n matrix

Standard Errors

A relatively straightforward extension of Henderson's mixed-model equations provides estimates of the standard errors of the fixed and random effects. Let the inverse of the leftmost matrix in Equation 26.5 be

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix} \tag{26.6}$$

where $\mathbf{C}_{11}$, $\mathbf{C}_{12}$, and $\mathbf{C}_{22}$ are, respectively, $p \times p$, $p \times q$, and $q \times q$ submatrices. Using this notation, Henderson (1975) showed that the sampling covariance matrix for the BLUE of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\sigma}(\hat{\boldsymbol{\beta}}) = \mathbf{C}_{11} \tag{26.7a}$$

that the sampling covariance matrix of the prediction errors $(\hat{\mathbf{u}} - \mathbf{u})$ is given by

$$\boldsymbol{\sigma}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{22} \tag{26.7b}$$

Matrix of PEV's

and that the sampling covariance of estimated effects and prediction errors is given by

$$\boldsymbol{\sigma}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{12} \tag{26.7c}$$

(We consider $\hat{\mathbf{u}} - \mathbf{u}$ rather than $\hat{\mathbf{u}}$ as the latter includes variance from both the prediction error and the random effects $\mathbf{u}$ themselves.)

The Animal Model, $y_i = \mu + a_i + e_i$

Here, the individual is the unit of analysis, with y_i the phenotypic value of the individual and a_i its BV

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \mu, \qquad \mathbf{u} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} \qquad \mathbf{G} = \sigma_A^2 \, \mathbf{A},$$

Where the additive genetic relationship matrix **A** is given by $A_{ij} = 2\theta_{ij}$, namely twice the coefficient of coancestry

Assume R = σ^2_e*I, so that $\mathbf{R^{-1}} = 1/(\sigma^2_e)*\mathbf{I}$.

Likewise, G = σ^2_A*A, so that $\mathbf{G^{-1}} = 1/(\sigma^2_A)*\mathbf{A^{-1}}$.

The "animal" model estimates the breeding value for each individual, even for a plant or tree! Same approach also works to estimate line (genotypic) values for inbreds.

The PEV and Accuracy

Recall that the q x q submatrix $\mathbf{C}_{22}$ in

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix}$$

Has as its diagonal elements the Predictor error variances (PEV) for each EBV

Hence, $PEV_{ii} = Var(EBV_i - A_i)$, the variance of i-th predicted value breeding value around its true value.

The smaller PEV_{ii}, the more accurate the estimate of its BV.

i's accuracy, ρ_i, is the correlation between the EBV and the true BV (recall accuracy of phenotype in predicting BV is h).

Reliability of EBVs

- The reliability of the EBV for individual i is just ρ_i^2. Recalling that h^2 is the reliability of phenotype alone as a predictor of breeding value, the extent to which the reliability exceeds h^2 is a measure of how much information is added by relatives.
- PEV and ρ are connected by
  - $PEV_{ii} = (1- \rho_i^2)Var(A)$. Hence, can easily compute the reliability (and accuracy) for any EBV
  - $\rho_i^2 = 1 - PEV_{ii}/Var(A)$.

Advantages of BLUP selection

- Easily accommodates fixed factors
- The relationship matrix A fully accounts
  - for all different types of relatives
  - Age-structure
  - Drift
  - Selection, assortative mating generated LD
- Unbalanced designs trivially handled
- Prediction of response (given EBVs of chosen parents)

Pitfalls of BLUP

- Strictly speaking, true BLUP assumes variance components are known without error
- Typically, use REML to estimate variances, and then use these in BLUP = "empirical BLUP". This does not account for the error introduced into EBVs by error in the variance estimate.
  - Using a fully Bayesian framework fully accommodates this concern.
- BLUP selection increases inbreeding relative to mass selection.

Extensions

- A number of extensions of the basic mixed model were examined in previous notes, e.g.,
  - repeated records
  - common family effect
  - genetic maternal effect
  - associate effects
- The other major extension is **multivariate BLUP**, where a vector of traits is considered for each individual.
- The key for multivariate BLUP is that we form a single vector of random effects by simply "stacking" the individual vector of BVs for each trait.

For trait j ($1 \leq j \leq k$), the mixed model becomes

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\mathbf{a}_i + \mathbf{e}_j$$

$$\begin{pmatrix}\mathbf{a}_j \\ \mathbf{e}_j\end{pmatrix} \sim \begin{pmatrix}\mathbf{0} \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\sigma^2_{A_j}\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{e_j}\mathbf{I}\end{pmatrix}$$

We can write this as **y** = **X**β + **Z**a+ **e**, where

$$\begin{pmatrix}\mathbf{y}_1 \\ \vdots \\ \mathbf{y}_k\end{pmatrix} = \begin{pmatrix}\mathbf{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{X}_k\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k\end{pmatrix} + \begin{pmatrix}\mathbf{Z}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Z}_k\end{pmatrix}\begin{pmatrix}\mathbf{a}_1 \\ \vdots \\ \mathbf{a}_k\end{pmatrix} + \begin{pmatrix}\mathbf{e}_1 \\ \vdots \\ \mathbf{e}_k\end{pmatrix}$$

Again, the BLUP for the vector of all EBVs is given by

$$\widehat{\mathbf{a}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

With **V** the covariance structure for this model

Covariance structure for EBVS

The resulting covariance structure for the stacked vector of breeding values is

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_k\end{pmatrix} = \begin{pmatrix}\sigma^2(A_1)\mathbf{A} & \cdots & \sigma(A_1,A_k)\mathbf{A}\\ \vdots & \ddots & \vdots\\ \sigma(A_k,A_1)\mathbf{A} & \cdots & \sigma^2(A_k)\mathbf{A}\end{pmatrix} = \mathbf{G}\otimes\mathbf{A}$$

where $\otimes$ denotes the Kronecker (or direct) product (LW Chapter 26) and

$$\mathbf{G} = \begin{pmatrix}\sigma^2(A_1) & \cdots & \sigma(A_1,A_k)\\ \vdots & \ddots & \vdots\\ \sigma(A_k,A_1) & \cdots & \sigma^2(A_k)\end{pmatrix}$$

is the matrix of genetic covariances of interest.

The genetic variance-covariance matrix ***G*** accounts for the genetic covariances among traits. ***G*** has k variances and k(k-1)/2 covariances, which must be estimated (REML) from the data.

Covariance structure for residuals

Similarly, the covariance structure for the stacked vectors of residuals is

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{e}_1\\ \vdots\\ \mathbf{e}_k\end{pmatrix} = \mathbf{E}\otimes\mathbf{I}, \quad \text{where} \quad \mathbf{E} = \begin{pmatrix}\sigma^2(e_1) & \cdots & \sigma(e_1,e_k)\\ \vdots & \ddots & \vdots\\ \sigma(e_k,e_1) & \cdots & \sigma^2(e_k)\end{pmatrix}$$

Finally, we need to specify any covariances between **a** and **e**. By construction $\sigma(a_z, e_z) = \sigma(a_w, e_w) = 0$, while the standard assumption is $\sigma(A_z, e_w) = \sigma(A_w, e_z) = 0$, giving the covariance structure as

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_k\\ \mathbf{e}_1\\ \vdots\\ \\ \mathbf{e}_k\end{pmatrix} = \begin{pmatrix}\mathbf{G}\otimes\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{E}\otimes\mathbf{I}\end{pmatrix}$$

Here the matrix E accounts for within-individual correlations in the environmental (or residual) values.

Index selection

- Index selection with multivariate BLUPs is easy.
- If the merit function is $H = \Sigma a_i A_i$, then the BLUP of H for individual k is
  - $BLUP(H_k) = \Sigma a_i$*EBV(trait i for individual k)
  - Note that since we are using estimated of breeding values, the index weights for the EBVs are the same as for the merit function.
  - With phenotypic index selection, we were using phenotypic values, and hence the usual need for different weights on the selected index to maximize gain in the merit function (index of breeding values).

G-BLUP

- A key feature with BLUP is obtaining the relationship matrix **A**. This is typically done from the pedigree.
- However, pedigree-based **A** values are based on the expected relatedness between individuals, not their actual relationships.
  - For example, 2θ for full-sibs has an expected value of 1/2, but there is variation around this value, so that with two pairs of sibs, one may have a realized 2θ of 0.38, the other 0.58.
  - Clearly, want to weight the second pair more, but using a pedigree-based **A** weights both sets equally (at 0.5).
  - With sufficiently-dense genetic markers can actually estimate the realized value
  - Using an **A** based on these genomic estimated of relationship is called G-BLUP (genomic BLUP), and **A** computed this way the genomic relationship matrix.

Example 20.6. While dense SNP Chips (platforms scoring hundreds of thousands to millions of SNPs in a single pass) at the time of this writing exist only for humans and a few important domesticated and model species, they are expected to become widespread (or replaced by other techniques such as whole-genome sequencing). SNP chips (or similar high-density polymorphism data) offer a very simple approach for obtaining the relationship matrix **A**. Given their very low mutation rates, two SNP alleles that are alike in state (AIS) can be viewed as being identical by descent (IBD), allowing us to compute the coefficient of coancestry θ_{ij} (LW Chapter 7) directly from the SNP data, and hence the entry $A_{ij} = 2\theta_{ij}$ in the relationship matrix. Recall that θ_{ij} is simply the probability that a randomly-drawn allele from individual i and a randomly-drawn allele from individual j are IBD, or in our case alike in state at a given SNP. Coding the two alleles at a SNP locus as **0**/**1**, if (at a given SNP locus) individual i is **11** while j is **10**, then all random draws from i are allele **1**, while half the random draws from j are also **1**, giving (for that locus) $\theta_{ij} = 1/2$. If i is **11** and j is **00**, then θ_{ij} (at this locus) is zero. Likewise if both are **10**, then with probability (1/2)(1/2) = 1/4, SNP allele **1** is drawn from both, while with probability 1/4, SNP allele **0** is drawn from both, while all other draws do not match, giving 1/4 + 1/4 = 1/2 as the coefficient of coancestry. The resulting coefficients of coancestry for all possible combinations for a biallelic SNP becomes

COC values for a given SNP

	Genotype of i		
Genotype of j	**11**	**10**	**00**
11	1	0.5	0
10	0.5	0.5	0.5
00	0	0.5	1

One computes the coefficient of coancestry for each SNP, taking the average value over all loci as the coefficient of coancestry for that pair of individuals. Toro et al. (2002) refer to this as **molecular coancestry**. Note that we can compare an individual with itself ($i = j$), which returns 1 for each homozygous locus and 1/2 for each heterozygous loci.

Genomic selection

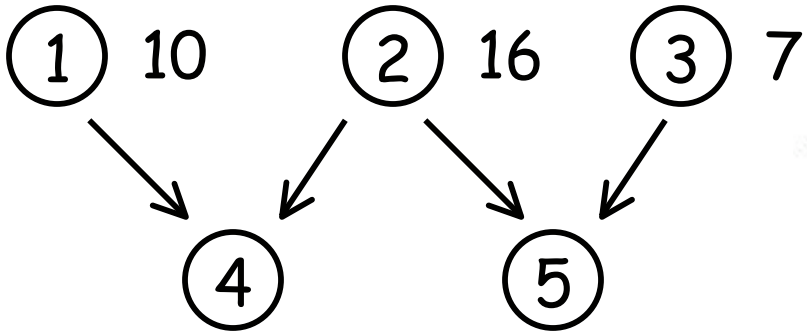
- G-BLUP is an example of genomic selection, using very dense marker information to make inferences on breeding values. An extension of MAS, but now using very many markers (thousands +). Why do this?
  - Predict BV in the absence of phenotype
    - Such early-generation scoring can increase rate of response
  - Improve estimate of BVs

EBVs for unmeasured individuals

- Before proceeding into genomic selection, we note that standard BLUP machinery allows us to estimate the breeding value of an unmeasured individual (i.e., an individual with no phenotypic record).
- GS also allows us to predict the breeding value for an unmeasured individual (no phenotype) for whom we also have genetic marker information.

Example

Suppose individuals 1 - 3 are measured, 4 & 5 are not.
Assume only a single fixed effect, the mean μ.



$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 1 & 1/4 \\ 0 & 1/2 & 1/2 & 1/4 & 1 \end{pmatrix}$$

Model becomes

$$\begin{pmatrix} 10 \\ 16 \\ 8 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

Here $\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, $\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$

Letting Var(A) = 100, Var(e) = 100, $\mathbf{V = ZGZ^T + R = 200^* I}$

$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ Gives BLUE(μ) = 11

$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)$ returns

$$\widehat{\mathbf{a}} = \begin{pmatrix} -0.50 \\ 2.50 \\ -2.00 \\ 1.00 \\ 0.25 \end{pmatrix}$$

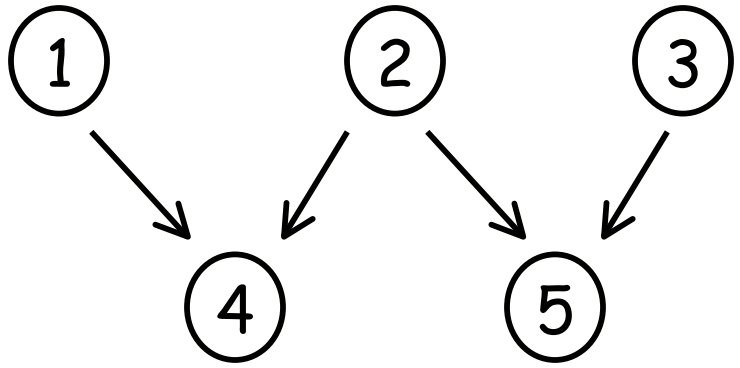
Average base pop EBVs = 0

EBVs for individuals (4,5) with no phenotypic records

Key: Information from relatives provides estimates for BV of unmeasured relatives.

G-BLUP

Suppose we have maker data.
How does this change EBVs?



Suppose marker data gives A as

2 slightly inbred

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0.5 & 0 \\ 0 & 1.2 & 0 & 0.5 & 0.5 \\ 0 & 0 & 1 & 0 & 0.5 \\ 0.5 & 0.5 & 0 & 1 & 0.2 \\ 0 & 0.5 & 0.5 & 0.2 & 1 \end{pmatrix}$$

4 & 5 slightly less related than 1/2 sibs

G-BLUP

$$\hat{\mathbf{a}} = \begin{pmatrix} -0.69 \\ 2.62 \\ -1.59 \\ 0.80 \\ 0.30 \end{pmatrix}$$

Pedigree-BLUP

$$\hat{\mathbf{a}} = \begin{pmatrix} -0.50 \\ 2.50 \\ -2.00 \\ 1.00 \\ 0.25 \end{pmatrix}$$

Background issues for GS

- Before proceeding, some quick refreshers in
  - QTL mapping and its limitations
  - Linkage disequilibrium (LD)
  - Association mapping
  - The "missing heritability" problem

QTL mapping

- Marker-trait associations within a family, close pedigree, or (most powerfully) a line cross
- Relatively low marker density (~ 5-10 cM/marker) sufficient
- Relies on an excess of parental gametes to generate marker-trait association
- Widely used 1980's ~ today, although ideas go back to 1917 and 1923
- Power a function of differences in QTL allelic effects, marker-trait recombination frequency c
  - Power for detection scales roughly as $2a(1-c)^2$
  - Dependence of power on freq(Q) is entirely through whether the sampled pedigree contains Q

Limitations of QTL mapping

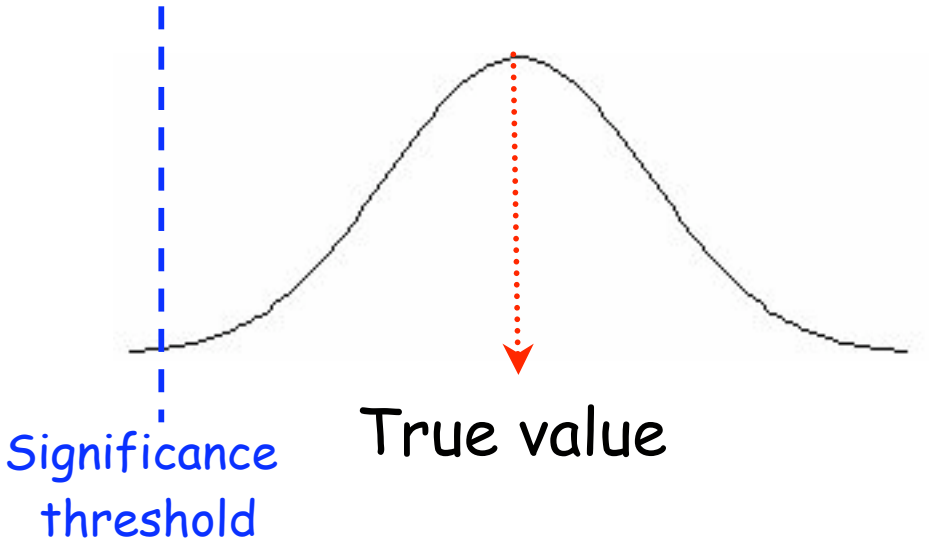
- Poor resolution (~20 cM or greater in most designs with sample sizes in low to mid 100's)
  - Detected "QTLs" are thus large chromosomal regions
- Fine mapping requires either
  - Further crosses (recombinations) involving regions of interest (i.e., RILs, NILs)
  - Enormous sample sizes
    - If marker-QTL distance is 0.5cM, require sample sizes in excess of 3400 to have a 95% chance of 10 (or more) recombination events in sample
    - 10 recombination events allows one to separate effects that differ by ~ 0.6 SD

Limitations of QTL mapping (cont)

- "Major" QTLs typically fractionate
  - QTLs of large effect (accounting for > 10% of the variance) are routinely discovered.
  - However, a large QTL peak in an initial experiment generally becomes a series of smaller and smaller peaks upon subsequent fine-mapping.
- The Beavis effect:
  - When power for detection is low, marker-trait associations declared to be statistically significant significantly overestimate their true effects.
  - This effect can be very large (order of magnitude) when power is low.

Beavis Effect

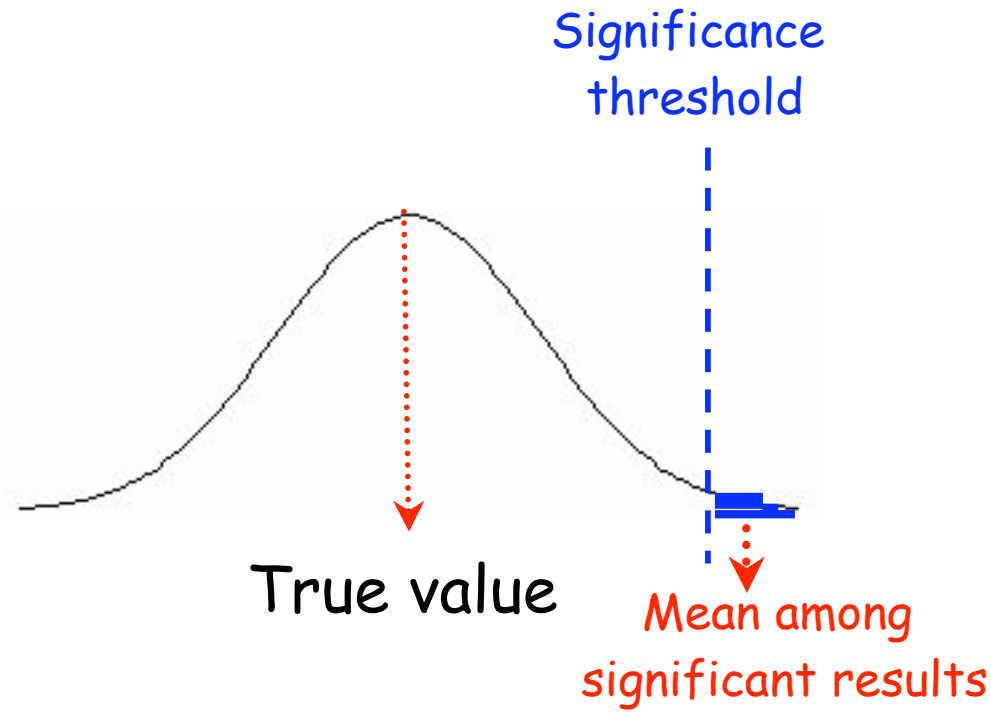
Also called the "winner's curse" in the GWAS literature



Distribution of the realized value of an effect in a sample

High power setting: Most realizations are to the right of the significance threshold, and the average value of these approaches the true value

In low power settings, most realizations are below the threshold, hence most of the time the effect is scored as being nonsignificant



However, the mean of those declared significant is much larger than the true mean

Background:
LD: Linkage disequilibrium

$D(AB) = freq(AB) - freq(A)*freq(B)$.

LD = 0 if A and B are independent. If LD not zero, correlation between A and B in the population

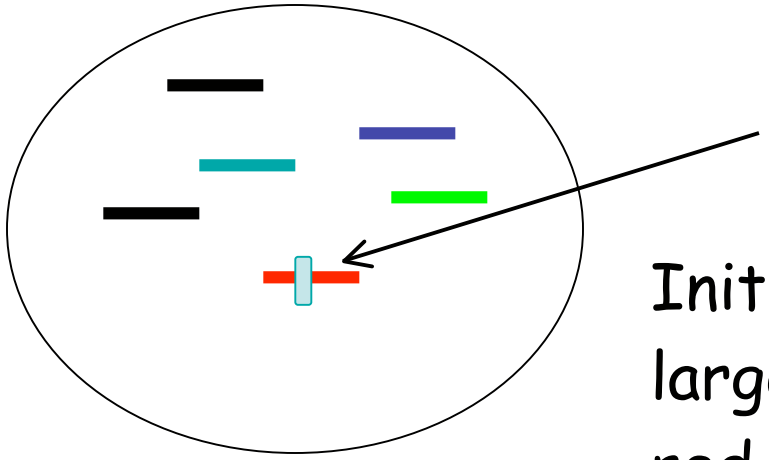
If a marker and QTL are linked, then the marker and QTL alleles are in LD in close relatives, generating a marker-trait association.

The decay of D: $D(t) = (1-c)^t D(0)$

here c is the recombination rate. Tightly-linked genes (small c) initially in LD can retain LD for long periods of time

Fine-mapping genes

Suppose an allele causing an effect on the trait arose as a single mutation in a closed population



New mutation arises on red chromosome

Initially, the new mutation is largely associated with the red haplotype

Hence, markers that define the red haplotype are likely to be associated (i.e. in LD) with the mutant allele

Background: Association mapping

- If one has a very large number of SNPs, then new mutations (such as those that influence a trait) will be in LD with very close SNPs for hundreds to thousands of generation, generating a marker-trait association.
  - Association mapping looks over all sets of SNPs for trait-SNP associations. GWAS = genome-wide association studies.
  - This is also the basis for genomic selection
- Main point from extensive human association studies
  - Almost all QTLs have very small effects
  - Marker-trait associations do not fully recapture all of the additive variance in the trait (due to incomplete LD)
  - This has been called the "missing heritability problem" by human geneticists, but not really a problem at all (more shortly).

Association mapping

- Marker-trait associations within a population of unrelated individuals
- Very high marker density (~ 100s of markers/cM) required
  - Marker density no less than the average track length of linkage disequilibrium (LD)
- Relies on very slow breakdown of initial LD generated by a new mutation near a marker to generate marker-trait associations
  - LD decays very quickly unless very tight linkage
  - Hence, resolution on the scale of LD in the population(s) being studied (1 ~ 40 kB)
- Widely used since mid 1990's. Mainstay of human genetics, strong inroads in breeding, evolutionary genetics
- Power a function of the genetic variance of a QTL, not its mean effects

Association mapping (cont)

Q/q is the polymorphic site contributing to trait variation, M/m alleles (at a SNP) used as a marker

Let p be the frequency of M, and assume that Q only resides on the M background (complete disequilibrium)

Haloptype	Frequency	effect
QM	rp	a
qM	(1-r)p	0
qm	1-p	0

Haloptype	Frequency	effect
QM	rp	a
qM	(1-r)p	0
qm	1-p	0

Effect of m = 0

Effect of M = ar

Genetic variation associated with Q = $2(rp)(1-rp)a^2$
~ $2rpa^2$ when Q rare. Hence, little power if Q rare

Genetic variation associated with marker M is
$2p(1-p)(ar)^2 \sim 2pa^2r^2$

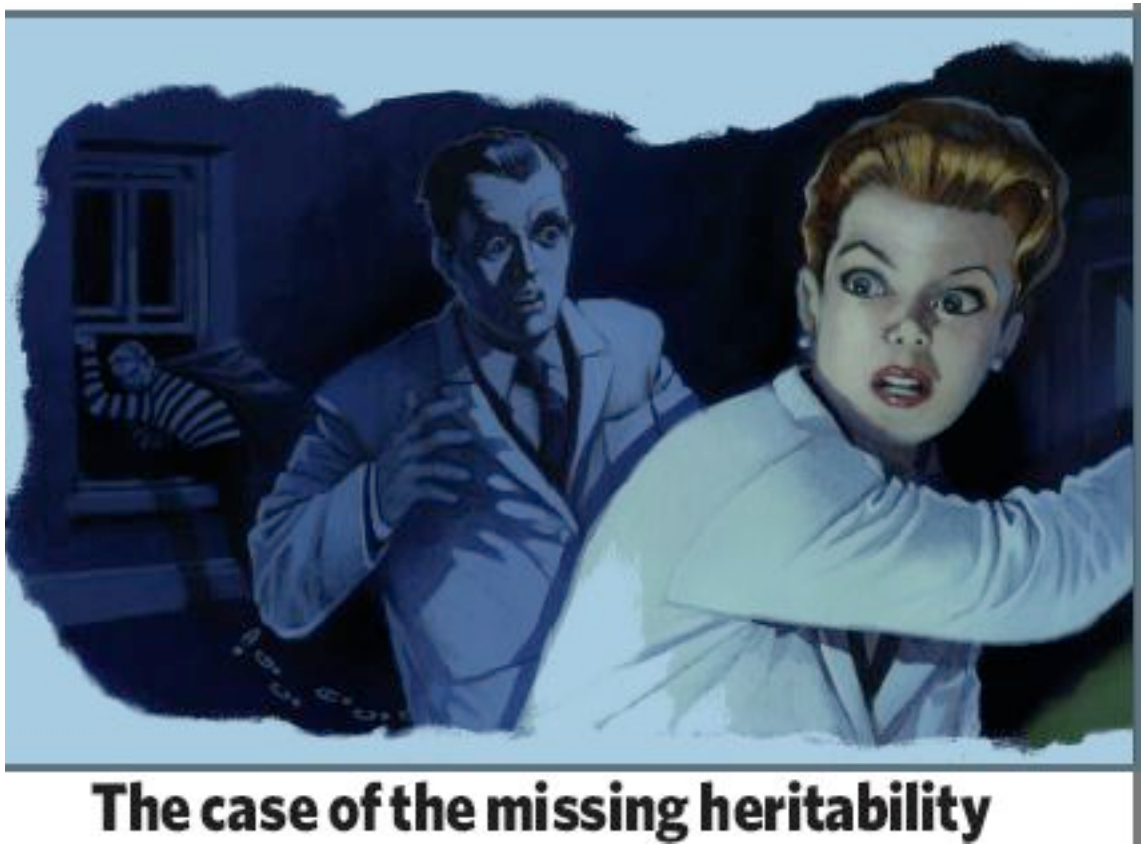
Ratio of marker/true effect variance is ~ r

Hence, if Q rare within the A class, even less power, as M only captures a fraction of the associated QTL.

Common variants

- Association mapping is only powerful for common variants
  - freq(Q) moderate
  - freq (r) of Q within M haplotypes modest to large
- Large effect alleles (a large) can leave small signals.
- The fraction of the actual variance accounted for by the markers is no greater than ~ ave(r), the average frequency of Q within a haplotype class
- Hence, don't expect to capture all of Var(A) with markers, esp. when QTL alleles are rare but markers are common (e.g. common SNPs, p > 0.05)
- Low power to detect G x G, G x E interactions

"How wonderful that we have met with a paradox. Now we have some hope of making progress" -- Neils Bohr



Infamous figure from *Nature* on the angst of human geneticists over the finding that all of their discovered SNPs still accounted for only a fraction of relative-based heritability estimates of human disease.

The "missing heritability" pseudo paradox

- A number of GWAS workers noted that the sum of their significant marker variances was much less (typically 10%) than the additive variance estimated from biometrical methods
- The "missing heritability" problem was birthed from this observation.
- Not a paradox at all
  - Low power means small effect (i.e. variance) sites are unlikely to be called as significant, esp. given the high stringency associated with control of false positives over tens of thousands of tests
  - Further, even if all markers are detected, only a fraction ~ r (the frequency of the causative site within a marker haplotype class) of the underlying variance is accounted for.

From MAS to GS

- The idea behind MAS, which grew out of QTL mapping, was the thought that first QTLs could be detected, and then using marker tags, MAS selected on these QTLS to improve response
- Several problems
  - "QTLs" really large chromosome regions ~ 40cM
  - QTLs of large effect fractionate into smaller and smaller effects upon fine mapping
  - Detected QTL effects are overestimated (Beavis effect)
  - Human Association studies: most QTL of small effect
- Key paper: Meuwissen, Hayes & Goddard (2001)
  - Skip trying to find QTLs altogether, use regressions involving ALL of the markers at once, use a training set to find the marker weights, and then use this to predict breeding values
  - Problem of more marker genotypes than scored phenotypes (shrinkage methods, random models)

While today there is a huge and complex literature on genomic selection, all of the basic ideas were clearly defined in Meuwissen's et al, classic paper:

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes† and M. E. Goddard†,‡

*Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, †Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and ‡Institute of Land and Food Resources, University of Melbourne, Parkville 3052, Victoria, Australia

Manuscript received August 17, 2000
Accepted for publication January 17, 2001

Key concern: finding weights for all markers when number of markers >> number of scored (phenotyped) individuals. A lot of different approaches to do this have been proposed. Bottom line: GBLUP hard to beat, easy to do!

Genomic selection

- Meuwissen, Hayes & Goddard (MHG) noted that MAS does not work because the markers don't account for enough genetic variation
  - Too few markers are used
  - Markers used likely overestimate QTL effects (they were chosen because they had a significant effect) = Beavis effect
  - Most QTLs likely have a very small effect
- Their solution: Include all of the markers into the analysis and then use statistical methods that shrink their effects. Random effects and Bayesian methods allow for number of markers to be >> than number of phenotypes scored
- Basic idea: Use a training set of individuals with marker information and high-quality estimates of breeding value to "train" a model (find regression parameters), then use this model to predict BVs of individuals with only marker info

- WHG started with a simulated data set of ~ 50,000 markers in a population run for 1000 generations to reach mutation/drift equilibrium. Roughly ~ 2000 individuals where then phenotyped in generations 1001 and 1002 and used to train the model. ~2000 generation 1003 individuals were generated and their breeding values predicted using the model fit from gen 1001 & 1002 data
  - The problem they faced was fitting ~ 50,000 marker effects with ~ 2000 data points (Breeding values)
- WHG's first model was standard least squares (LS) where each marker was tested separately, with those whose marker-trait effect exceeded a multiple-testing threshold chosen. The selected markers where then jointly fit in a multiple regression.
- GEBVs (genomic estimated breeding values) given by
  - $GEBVi = \Sigma_k a_k g_{i,k}$, where a_k = weight in SNP marker k, $g_{i,k}$ = genotypic score at marker k for individual I.

- Their three other models used a random-effects approach. Recall that under this framework, one estimates the variance of some underlying distribution from which individual realizations (here, the BV variance explained by a given SNP) then have their values predicted. This allows the ability to predict p >> n effects.
- Model one: BLUP. This assumes the SNP variance is the same for each marker (the expected variance is the same over all sites), so that a particular realization for a given marker is drawn from this distribution. Basically, this assumes the infinitesmal model, and is just GBLUP.
- Model two: BayesA. This assumes that QTLs at different SNPs may have a distribution of different values, so that for a given marker the expected value for the variance (which is used to generate the particular realization) is itself drawn from a distribution. MHG assumed this distribution for the expected variance at a marker was an inverse chi-square distribution, which has most effects small, but a few rare effects.

- Model 3: BayesB. The problem with Bayes A is that all SNPs are assumed to have some nonzero marker variance (albeit very small). A potentially more realistic model is that a fraction π of sites have no variance, while the remainer (1- π) have their expected values drawn from some distribution, and (given this drawn value), a realization from for that site. Was computationally faster than Bayes A.
- Results: LS did very poorly, while the random effects models generally did well

TABLE 2

Comparing estimated *vs.* true breeding values in generation 1003

	$r_{TBV;EBV}$ + SE	$b_{TBV.EBV}$ + SE
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827
BayesB	0.848 + 0.012	0.946 + 0.018

Benchmark: r ~ 0.4 for missing record with pedigree-BLUP,
r ~ 0.8 for progeny test.

Different assumptions regarding the distribution of effects at the underlying (and unknown) QTLs leads to the many different models used for GS

Name	Reference	Assumed distribution of SNP effects
BLUP	Meuwissen et al. (2001)	Normal
BayesA	Meuwissen et al. (2001)	*t* distribution
BayesB	Meuwissen et al. (2001)	Mixture distribution of zero effects and *t* distribution of effects
Bayesian LASSO	Yi and Xu (2008)	Double exponential distribution of effects
BayesSSVS	Verbyla et al. (2009)	Mixture distribution of zero effects and *t* distribution of effects

Hayes & Goddard Genome 53: 876 (2010)

Name	Assumed distribution of SNP effects	Implication
BLUP	Normal	A very large number of QTL of small effect
BayesA	*t* distribution	A large number of QTL of small effect and a small proportion with moderate to large effect
BayesB	Mixture distribution of zero effects and *t* distribution of effects	A large number of genome regions with zero effect, a small proportion of QTL with moderate effects
Bayesian LASSO	Double exponential distribution of effects	Very large proportion of SNP with effect of close to zero, small proportion of moderate to large effect
BayesSSVS	Mixture distribution of zero effects and *t* distribution of effects	A large number of genome regions with almost zero effect, a small proportion of QTL with moderate effects

Hayes & Goddard Genome 53: 876 (2010)

A number of other methods, based on different assumptions around the distribution of QTL effects. A number are "machine learning" (semiparametric regression) approaches that make few assumptions about this underlying distribution, but are more in the form of taking a training set with some pattern (molecular data) with breeding values to generate some predictive function. Examples of such methods include support vector machines, semiparametric kernel regressions, and reproducing kernel Hilbert space regression.

Which version of GS to use?

- Different assumptions about these underlying distributions lead to different GEBVs estimators
- Generally, the differences are often small
- GBLUP is not only easy and robust, it is also often the best. Hence, recommendation is to use it unless only information suggests otherwise
  - This is a model-fitting issue, as predictability of the model in the testing set provides some indication of which method is best.
- Hayes & Goddard suggest that if the aim of GS is to select across populations, that using a model assuming most SNPs have zero effect and just a few have moderate/large may be best, as this will located those QTLs segregating across lines/breeds

Accuracy improves with more records, closer marker spacing

TABLE 3

Correlations between true and estimated breeding values when the number of phenotypic records is varied

	No. of phenotypic records		
	500	1000	2200
LS	0.124	0.204	0.318
BLUP	0.579	0.659	0.732
BayesB	0.708	0.787	0.848

TABLE 4

Correlations between true and estimated breeding values when the density of the marker map is varied and effective population size is 100

	Marker spacing (cM)		
	1	2	4
LS	0.318	0.354	0.363
BLUP	0.732	0.708	0.668
BayesB	0.848	0.810	0.737

Accuracy declines quickly over generations

TABLE 5

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

Generation	$r_{TBV;EBV}$
1003	0.848
1004	0.804
1005	0.768
1006	0.758
1007	0.734
1008	0.718

A closely-related issue is that a model trained (i.e., marker weights estimated) in one population does not translate to other populations. Models must also be retrained frequently.

Advantages of genomic selection (GS)

- While theoretically possible that genomic selection returns higher accuracies than standard phenotype/pedigree based BLUP EBVs, this is not typically while GS is used.
- Main use: speed up generation time
  - Testing bulls in dairy cattle typically requires 6 - 7 years
  - With GS, generation intervals down to 3 years
  - Double rate of response, so if accuracy is at least half that of standard phenotype-based BLUP, increases the rate of response
- Concern is that the accuracy declines each generation
  - This requires constant updating of the model and hence the constant updating of phenotypic records.

GS impact greatest for:

- Sex-limited traits
- Traits that are expensive to measure
- Traits measured only by destruction of an individual
- Traits expressed late in life
- Traits expressed after individuals are selected

GS and inbreeding

- Simulations show that GS is expected to reduce inbreeding per generation relative to standard BLUP.
- Reason is that it exploits the Mendelian segregation variance (two sibs equally weighted with pedigree BLUP, differentially weighted with genome-based weights), hence full sibs less likely to be co-selected.
- However, because of decreasing generation interval, rate of inbreeding/year may be larger.

How many SNPs needed?

- SNP density (number & spacing of makers) depends on the amount of LD in the population
- Rough rule: For accurate genomic breeding need
  - LD between adjacent SNPs with $r^2 > 0.2$
- The expected LD between two markers at recombination frequency c under mutation-drift is
  - $r^2 \sim 1/(4N_e c + 1)$, or $c \sim 1/Ne$ (for $r^2 = 0.2$)
- Meuwissen (2009) $10N_eL$ markers need, where L = genome length
  - For Holstein cattle, $N_e \sim 100$, L = 30 Morgans (3000 cM), so that ~ 10*100*30 = 30,000 roughly equally-spaced SNPs
  - Likewise, for $N_e = 100$, $c \sim 1/100 = 0.01$ (for $r^2 = 0.2$). Number n of markers (given genomic length of 30 Morgans) again becomes ~ 30,000.

Accuracy of predicted values

Hayes & Goodard show the accuracy of genomic prediction depends on the number q of independent chromosome segments in a population, with $q \sim 2N_eL$

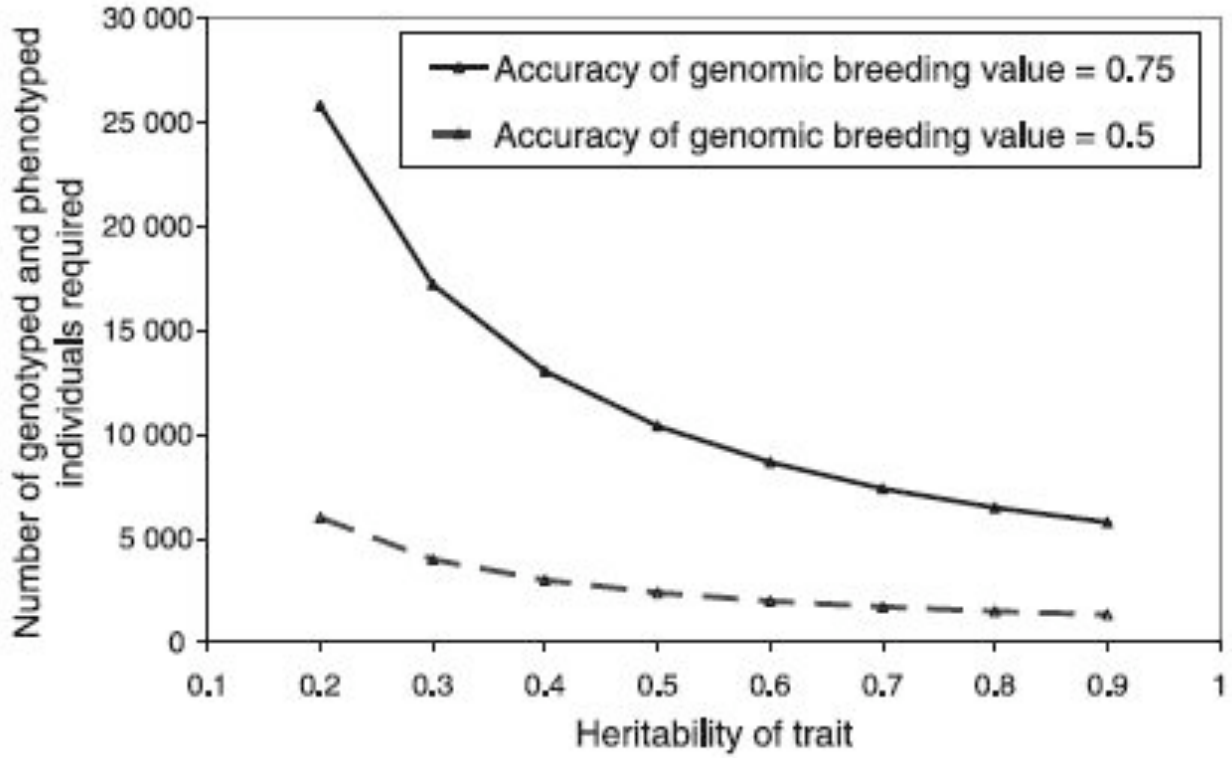
$$r^2 \simeq 1 - \frac{\lambda/(2N\sqrt{a}\,) \cdot \log(1 + a + 2\sqrt{a}\,)}{1 + a - 2\sqrt{a}}$$

$$a = 1 + \frac{2\lambda}{N}, \quad \lambda = \frac{qh^2}{\log(2N_e)}, \quad q = 2N_eL$$

N = number of phenotypic records in training population, h^2 = trait heritability

This is first-generation of response. Accuracy declines in subsequent generations.

Fig. 2. Number of genotyped and phenotyped individuals required in the reference population to reach a desired accuracy of genomic breeding value (for individuals that were not phenotyped) N_e was 100.



Hayes & Goddard Genome 53: 876 (2010)

Assuming N_e = 100