

So far:

* Using predicted breeding values as selection criteria to find parents of future generations using all available phenotypic information together with pedigree relationships.

* Used linear mixed effect models to get to predicted breeding values

no marker

Traditional approach

2006 / 7

Genomic Selection

Peter von Rohr

2022-12-02

use marker information on a large scale to predict breeding values

Making selection decisions to find parents of future generation based on genomic breeding values.

Shift in paradigm, mainly in cattle breeding

Introduction

* Meuwissen et al. (2001): How to use total genotypic values for prediction of breeding values.

* Genotypic values (V_{ij}) for a single locus model: with values

$$\begin{array}{l|l} & \overbrace{V_{ij}} \\ G_1G_1 & +a \\ G_1G_2 & +d \\ G_2G_2 & -a \end{array}$$

- ▶ Proposed in 2001
- ▶ Widely adopted in 2007/2008
- ▶ Costs of breeding program reduced due to shorter generation intervals
- ▶ In cattle: young sire selection versus selection based on sire proofs
- ▶ In pigs: early selection among full sibbs
- ▶ Inbreeding must be considered

By consequently basing selection decisions on genomics breeding values, costs of a cattle breeding program could be reduced by about 90%

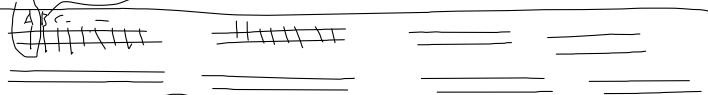
accurate predictions at very young ages

Cattle: As soon as calf is born, a hair sample taken and is sent to the lab and after 2-4 weeks, genomic breeding values are available. Reliabilities range between 30-50%

For allele frequencies that considered to be constant

Animal i:

$$V_1 + \dots + V_B + \dots + \mu_i^*$$



$f(G_1)$
 $f(G_2)$



$$BV_G = (p+T)\alpha + TV_H + BV_I + BV_J + BV_K + \dots + \alpha$$

$$\mu_i \sim N(0, (1+f_i)\sigma^2)$$

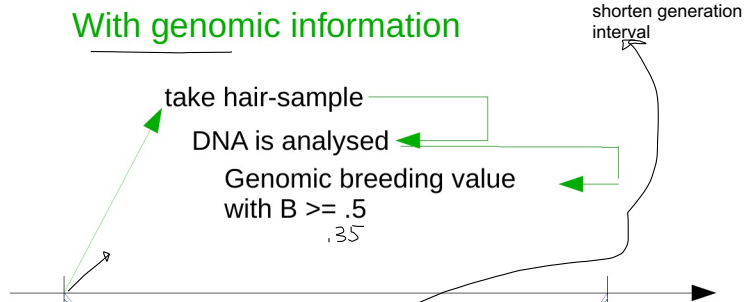
Reason for linear mixed effect models

Terminology

- ▶ **Genomic Selection:** use of genomic Information for selection decisions
- ▶ Genomic Information is used to predict **genomic breeding values**

Benefits in Cattle

With genomic information



5-7 years

birth

$$\hat{u}_i = \frac{1}{2}(\hat{u}_s + \hat{u}_d)$$

Predicted breeding value based on parents

$$B = \frac{1}{4}(B_s + B_d)$$

80% 20% → 30%

25%

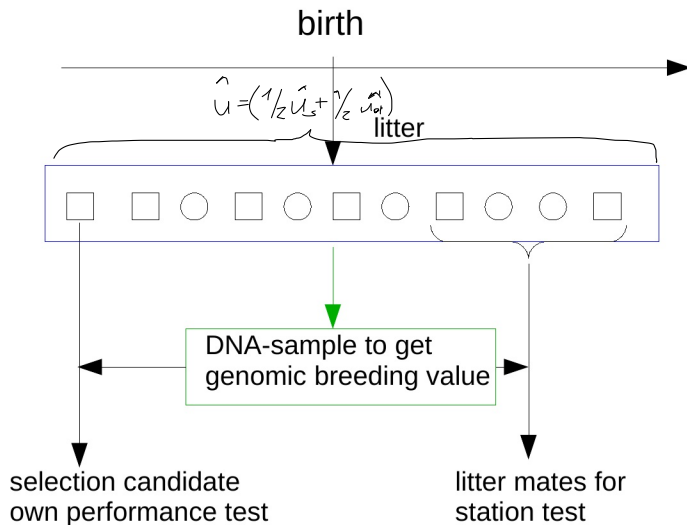
Progeny test results $B \geq 0.5$

selection

traditional

Without genomic information

Benefits in Pigs

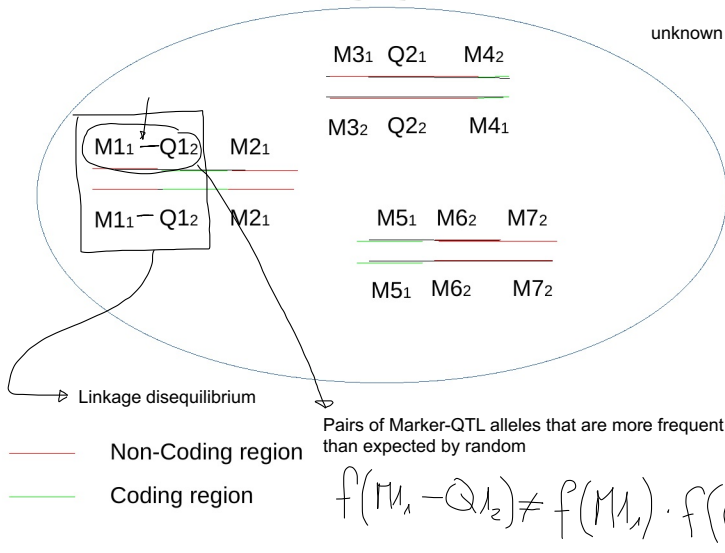


Genetic Model

- ▶ Recall: BLUP animal model is based on infinitesimal model
- ▶ Prediction of genomic breeding values is based on **polygenic model**
- ▶ In polygenic model: **Single Nucleotide Polymorphisms** (SNP) are used as markers
- ▶ Marker genotypes are expected to be associated with genotypes of **Quantitative Trait Loci** (QTL)

Polygenic Model

Distribution of SNP (M) and QTL (Q)



Statistical Models

A_{ni}	$SM_1 \dots$	$\dots SM_k$	y_i
1			y_1
2			y_2
\vdots			\vdots
N			y_N

Two types of models are used

1. marker-effect models (MEM)
2. genomic-breeding-value based models (BVM)

$$k = 150,000$$

$$800,000 - H$$

$$\text{Seq: } 2 \cdot 10^7$$

MEM

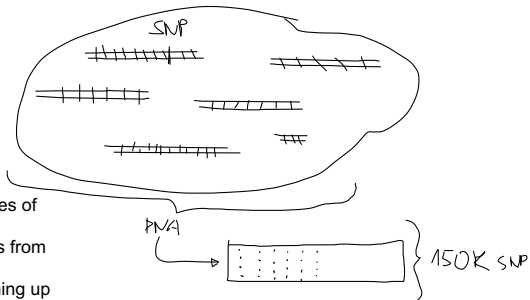
$$y_i = \mu + \underbrace{\beta_1 a_1 + \beta_2 a_2 + \dots + \beta_k a_k}_{\text{fixed effect}} + \epsilon_i$$

- ▶ marker effects (a -values) are fitted using
 - ▶ a simple linear model \rightarrow marker effects are fixed
 - ▶ a linear mixed effects model \rightarrow marker effects are random
- ▶ Problem of finding which markers are associated to QTL
- ▶ With high number of SNP compared to number of genotyped animals: very large systems of equations to solve

Recap 2022-12-09:

* Genomic Selection: Selection process based on predicted breeding values using genomic information

* Genomic (as opposed to genetic) is used when marker information that is used is evenly spread across the whole genome.



Statistical Models

* Fixed linear effect model ==> estimates of "Marker-effects"

* Marker-effects correspond to a-values from single locus model

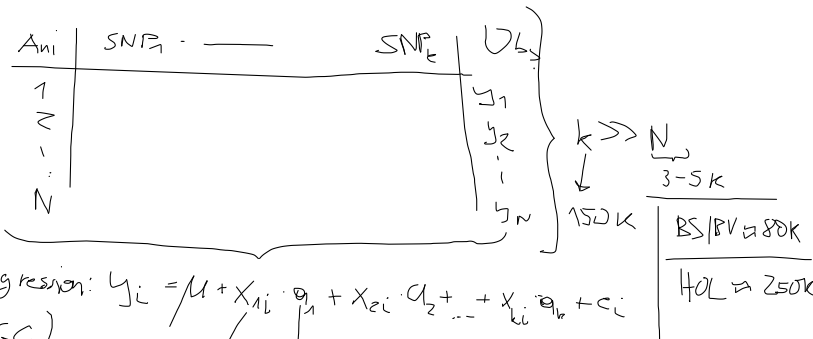
* Genomic breeding values from summing up the marker effects that correspond to the genotypes of a given animal.

Problem: Too many effects to use least squares in a simple Regression model.

Problem: Too many effects to use least squares in a simple Regression model.

* Already Meuwissen et al. (2001) already realized that problem, they proposed:

> First run a single marker GWAS



-1: G₂G₂
 0: G₁G₂
 1: G₁G₁
 d=0

Code for the Genotype of animal i at SNP1

Marker-effect (a-value) for SNP 1

Genotypic values in 1-locus model

G₁G₁ : a
 G₁G₂ : d
 G₂G₂ : -a

Least Squares:

$$\hat{a} = (X^T X)^{-1} X^T y$$

Because, $k \gg N$, the Matrix : $(X^T X)^{-1}$

Possible solutions:

*Replace least square with LASSO

*Use mixed linear effects models with the marker effects as random effects.

$$y = \underbrace{X\beta}_{\text{fix}} + \underbrace{Wa}_{\text{random}} + e \quad \begin{matrix} \nearrow \text{residuals} \\ \downarrow \mathcal{N}(0, I \cdot \sigma_a^2) \end{matrix}$$

Solutions are obtained with MME

BVM

Breeding Value based model
* Mixed linear effect model

$$y = X\beta + Zg + e$$

genomic BV

- ▶ genomic breeding values as random effects
- ▶ similar to animal model
- ▶ genomic relationship matrix (G) instead of numerator relationship matrix (A)

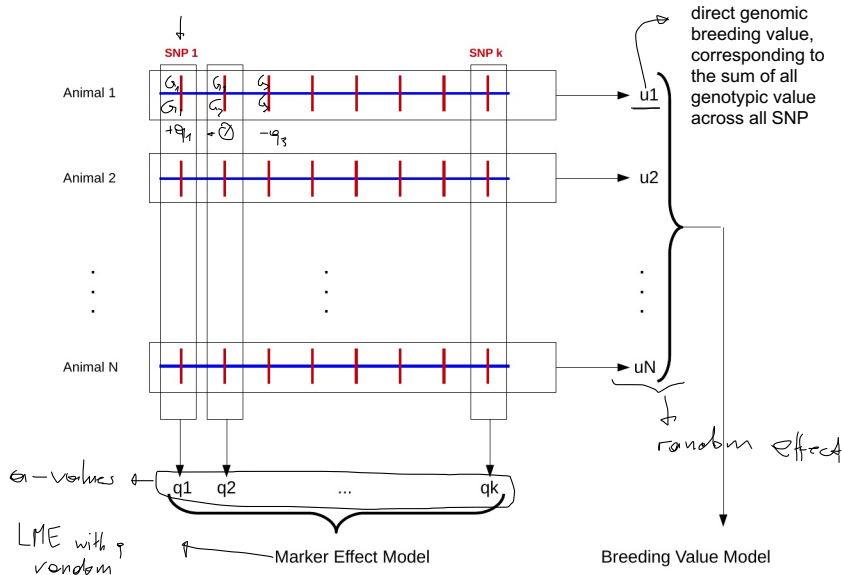
Solutions are obtained via MME

Traditional BLUP Animal model, vector of breeding values: $u \rightarrow \text{var}(u) = A \cdot \sigma_u^2$

Genomic breeding values: $g \rightarrow \text{var}(g) = \underbrace{G}_{\text{genomic relationship matrix}} \cdot \sigma_g^2$

genomic relationship matrix

MEM versus BVM



Marker effect model:

* Result will be marker effect for every SNP ==> vector

$$\hat{q} = \begin{bmatrix} \hat{q}_1 \\ \hat{q}_2 \\ \vdots \\ \hat{q}_k \end{bmatrix}$$

$$y = X\beta + Wq + e$$

* Goal: Genomic breeding (\hat{g}) value from Marker Effects \hat{q} :

$G_n G_i: 1$
 $G_1 G_2: 0$
 $G_2 G_2: -1$

{

Δ_{ni}	SNP ₁	...	SNP _k
1	$G_1 G_1$	0	-1
2	$G_2 G_2$...	1
...			
N			

$$\Rightarrow \hat{g}_i = 1 \cdot \hat{q}_1 + 0 \cdot \hat{q}_2 + (-1) \cdot \hat{q}_3 + \dots + 1 \cdot \hat{q}_k$$

Logistic Procedures

A_n	SMP_1	...	SMP_k	a_{e_i}
1				y_1
2				y_2
...				...
N				y_N

Currently used in cattle breeding

markers

▶ Two Step:

- ▶ use reference population to get marker effects using MEM
- ▶ use marker effects to get to genomic breeding values

▶ Single Step

- ▶ MEM or BVM in a single evaluation
- ▶ difficulty how to combine animals with and without genotypes

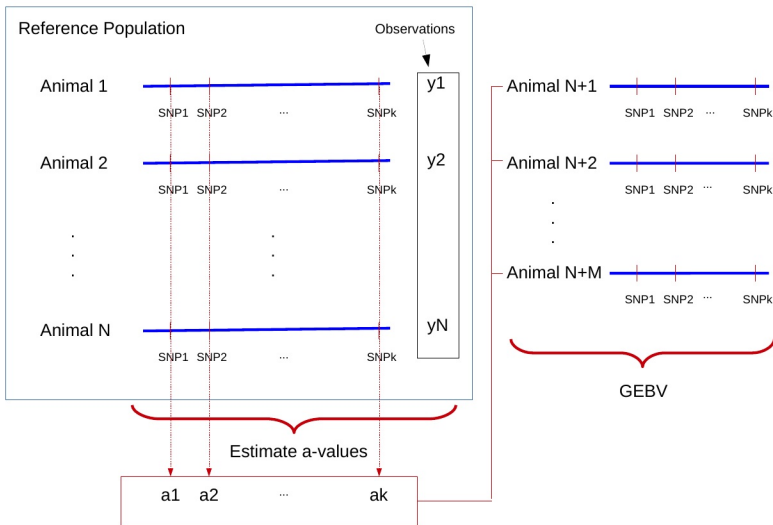
Two Step in practice:

- * Reference population of animals with predicted breeding values using BLUP animal model with a high reliability ($> 0.35 - 0.5$)
- * For animals in the reference population, we have marker information and observations are available
- * Since the reference population consists of mostly male animals, the observations are based on de-regressed traditional breeding values
- * De-regression is the transformation of the variability from the scale of the predicted breeding values back to the scale of phenotypic observations.
- * Three times a year (April, August and December) marker effects are estimated using the reference population data.

- * For new-born animals, hair samples are sent to the Lab
- * DNA is extracted and the genotypes at all marker positions are determined
- * The marker genotypes together with marker effects are used to predict direct genomic breeding values, as shown before
- * Done every 2 weeks

- + Procedure works well.
- + As a consequence many selection decisions can be taken base on young animals.
- + This allows to shorten the generation interval (from 5-7 years down to 2 years)
- +/- In the long run, not so many bulls are going to be progeny tested anymore ==> reference population does not grow anymore
- animals in reference population are getting older and further away from the current breeding population with a negative effect on the accuracy of marker effects estimates.
- heavily dependent on the reference population.
- As soon as all male animals are genotyped, new data can only come from cow genotypes. But since their reliability is seldom high, they are not considered in the reference population.

Two Step Procedure



- ▶ Use a mixed linear effect model
- ▶ Genomic breeding values g are random effects

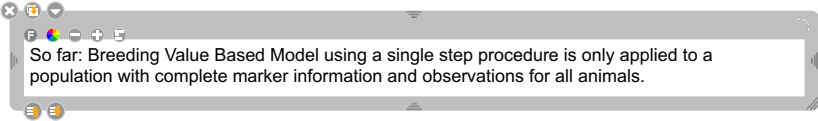
$$y = Xb + Zg + e$$

with

- ▶ $E(e) = 0, \text{var}(e) = I * \sigma_e^2$
- ▶ $E(g) = 0, \text{var}(g) = G * \sigma_g^2$
- ▶ Genomic relationship matrix G

Solutions for \hat{g} will be obtained by
Mixed Model Equations

Solution Via Mixed Model Equations

A screenshot of a presentation slide with a grey border and standard window controls (close, maximize, zoom, search) at the top. The text inside the slide reads: "So far: Breeding Value Based Model using a single step procedure is only applied to a population with complete marker information and observations for all animals."

So far: Breeding Value Based Model using a single step procedure is only applied to a population with complete marker information and observations for all animals.

- ▶ All animals have genotypes and observations

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \lambda * G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix}$$

with $\lambda = \sigma_e^2 / \sigma_g^2$.

Animals Without Observations

- ▶ Young animals do not have observations
- ▶ Partition \hat{g} into
 - ▶ \hat{g}_1 animals with observations and \rightarrow reference population
 - ▶ \hat{g}_2 animals without observations \rightarrow young animals
- ▶ Resulting Mixed Model Equations are (assume $\lambda = 1$)

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$$

$$\begin{bmatrix} X^T X & X^T Z & 0 \\ Z^T X & Z^T Z + G^{(11)} & G^{(12)} \\ \underline{0} & \underline{\quad} G^{(21)} & \underline{\quad} G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \\ \rightarrow 0 \end{bmatrix}$$

Predicted Genomic Breeding Values

young animals

- ▶ Last line of Mixed model equations

$$0 \cdot \hat{b} + G^{(21)} \cdot \hat{g}_1 + G^{(22)} \cdot \hat{g}_2 = 0$$

Same partitioning for vector g

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$$

$$\begin{aligned} \text{var}(g) &= \text{var} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \\ &= \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} = G \end{aligned}$$

$$\text{var}(g_1) = G_{11}$$

$$\text{var}(g_2) = G_{22}$$

$$\text{cov}(g_1, g_2) = G_{12} = G_{21}^T$$

$$G^{-1} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}^{-1} = \begin{bmatrix} G^{(11)} & G^{(12)} \\ G^{(21)} & G^{(22)} \end{bmatrix}$$

Solutions

- ▶ Solving for \hat{g}_2

$$\hat{g}_2 = -(G^{(22)})^{-1} \cdot G^{(21)} \cdot \hat{g}_1$$

In Summary, so far:

* Direct genomic breeding values based on MEM:

1. estimate Marker effects $\Rightarrow \hat{q}$

2. Use \hat{q} together with marker genotypes to get genomic breeding values

\Rightarrow Two step procedure with an additional advantage: not dependent on genomic relationship

* Breeding value based model in a single step procedure

* Linear mixed effect models to predict directly predict genomic breeding values for animals with and without observations.

* Solutions were obtained from Mixed Model equations which depend on Genomic Relationship Matrix G

\Rightarrow How to compute G ?

Genomic Relationship Matrix

- ▶ Breeding value model uses genomic breeding values g as random effects
- ▶ Variance-covariance matrix of g are proposed to be proportional to matrix G

$$\text{var}(g) = G * \sigma_g^2$$

genetic variance explained
by all SNP markers



where G is called **genomic relationship matrix (GRM)**

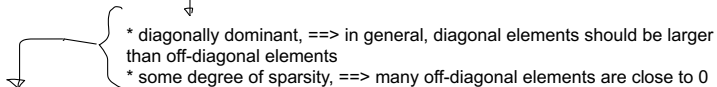
Properties of G

Genotype-code for animal i at SNP position 1, can either be -1, 0 or 1

$$g_i = X_{1i} \cdot q_1 + X_{2i} \cdot q_2 + \dots + X_{ki} \cdot q_k$$


genomic breeding value for animal i vector of marker effects

- ▶ genomic breeding values g are linear combinations of q
- ▶ g as deviations, that means $E(g) = 0$
- ▶ $var(g)$ as product between G and σ_g^2 where G is the genomic relationship matrix
- ▶ G should be similar to A



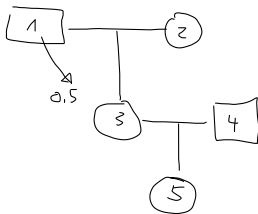
necessary conditions that G is non-singular, that means that an inverse of G exists

Change of Identity Concept

 numerator relationship matrix

- ▶ A is based on identity by descent (IBD)
- ▶ G is based on identity by state (including ibd), assuming that the same allele has the same effect
- ▶ IBS can only be observed with SNP-genotype data

Identify by descent (IBD): based on common ancestry defined in the pedigree



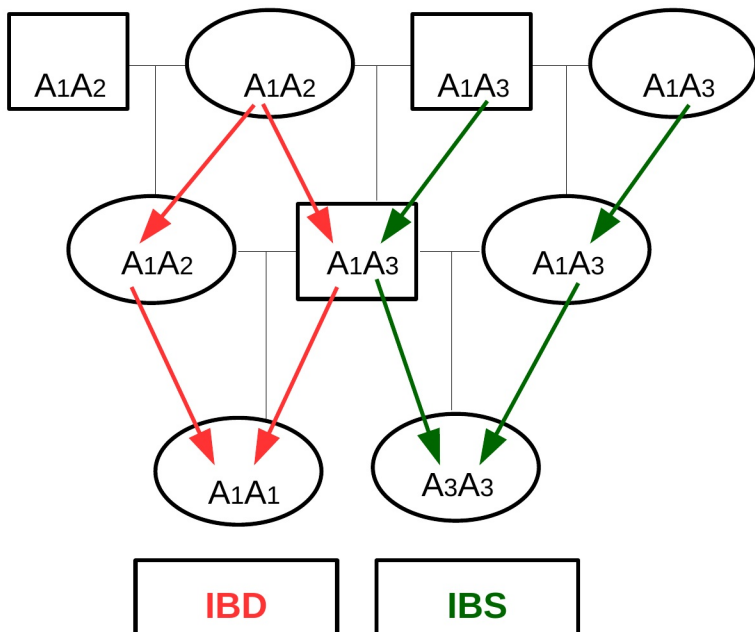
$$F = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & & \end{bmatrix}$$

Identify by state

$$G = \begin{bmatrix} 1 & 0.5 & 0 \\ & 1 & 0.5 \\ & & \end{bmatrix}$$

A_{ij}	SNP ₁	SNP ₂	...	SNP _k
1	$G_1 G_1$	-	-	-
2	$G_1 G_2$	-	-	-
3	$G_2 G_2$	-	-	-
⋮				

Identity



Linear Combination

- ▶ SNP marker effects (a values) from marker effect model are in vector q
- ▶ Genomic breeding values from breeding value model are determined by

$$g = U \cdot q$$

- ▶ Matrix U is determined by desired properties of g



is unknown, question is how should matrix U look like such that desired properties of g are fulfilled.

Deviation

- ▶ Genomic breeding values are defined as deviation from a certain basis

$$\rightarrow E(g) = 0$$

- ▶ How to determine matrix U such that $E(g) = 0$

Equivalence Between Models

Decomposition of phenotypic observation y_i with

- ▶ Marker effect model

$$y_i = \underbrace{X_i^T \cdot \beta}_{*} + \underbrace{w_i^T \cdot q}_{*} + e_i$$

Genotype codes for animal i:
G1G1: 1
G1G2: 0
G2G2: -1

genomic components of observation y for animal i

- ▶ Breeding value model

$$y_i = \underbrace{(X_i^T \cdot \beta)}_{*} + g_i + e_i$$

- ▶ g_i and $w_i^T \cdot q$ represent the same genetic effects and should be equivalent in terms of variability

Expected Values

$$y_i = w_i^T q + e_i$$
$$y_i = \beta_i + e_i$$

- ▶ Required: $E(g_i) = 0$
- ▶ But: $E(\underline{w}_i^T \cdot q) = \underline{q}^T \cdot E(w_i)$
- ▶ Take q as constant SNP effects
- ▶ Assume w_i to be the random variable with:

$$w_i = \begin{cases} 1 & \text{with probability} \\ 0 & \text{with probability} \\ -1 & \text{with probability} \end{cases}$$

$$\begin{array}{|c|} \hline p^2 \\ \hline 2p(1-p) \\ \hline (1-p)^2 \\ \hline \end{array}$$

HW - $E(q)$

→ $E(w_i)$: For a single locus

$$E(w_i) = 1 * p^2 + 0 * 2p(1-p) + (-1)(1-p)^2 = p^2 - 1 + 2p - p^2 = \underline{2p - 1} \neq 0$$

Specification of g

- ▶ Set

$$g_i = (w_i^T - s_i^T) \cdot q$$

with $s_i = E(w_i) = 2p - 1$

- ▶ Resulting in

$$g = U \cdot q = (W - S) \cdot q \Rightarrow E(g) = 0$$

genotype codes
↑ ↘ $2r-1$

with matrix S having columns j with all elements equal to $2p_j - 1$ where p_j is the allele frequency of the SNP allele associated with the positive effect.

Genetic Variance

- ▶ Requirement: $\text{var}(g) = G * \sigma_g^2$
- ▶ Result from Gianola et al. (2009):

genetic variance based on genomic breeding values

$$\sigma_g^2 = \sigma_q^2 * \sum_{j=1}^k (1 - 2p_j(1 - p_j))$$

Variance of marker effects, assumed to be given

- ▶ From earlier: $g = U \cdot q$

$$\text{var}(g) = \text{var}(U \cdot q) = U \cdot \text{var}(q) \cdot U^T = UU^T \sigma_q^2$$

variance-covariance matrix of marker effects

$$\text{var}(q) = I * \sigma_q^2$$

- ▶ Combining

$$\text{var}(g) = G \cdot \sigma_g^2 = G \cdot \sigma_q^2 \cdot \sum_{j=1}^k (1 - 2p_j(1 - p_j)) = UU^T \sigma_q^2$$

$$\text{var}(g) = UU^T \sigma_q^2 = G * \sigma_q^2 * \sum_{j=1}^k (1 - 2p_j(1 - p_j))$$

$$\text{var}(g) = UU^T \underbrace{\sigma_q^2}_{> 0} = G * \underbrace{\sigma_q^2}_{> 0} * \sum_{j=1}^k (1 - 2p_j(1 - p_j))$$

$$\underbrace{UU^T}_{(W-S)} = G * \sum_{j=1}^k (1 - 2p_j(1 - p_j)) \quad \text{Allele frequency of SNP } j$$

Solve for G:


$$G = \frac{UU^T}{\sum_{j=1}^k (1 - 2p_j(1 - p_j))}$$

Genomic Relationship Matrix

$$G = \frac{UU^T}{\sum_{j=1}^k (1 - 2p_j(1 - p_j))}$$

How To Compute G

Genotype codes obtained from SNP data



- ▶ Read matrix W
- ▶ For each column j of W compute frequency p_j
- ▶ Compute matrix S and $\sum_{j=1}^k (1 - 2p_j(1 - p_j))$ from p_j
- ▶ Compute U from W and S
- ▶ Compute G

In practice:

- * The computed matrix G is often singular \implies cannot be inverted.
- * G can be approximated by $G^* = G + 0.01 * I$ or $G^* = G + 0.01 * A$
- * G^* can be inverted and is used in the Mixed model equations.

BVM is a linear mixed effect model using genomic relationship matrix G .
 G is computed as shown.