HOS 6236 Molecular Marker Assisted Plant Breeding
Fall 2017

Last Class:

Methods in Genomic Selection

Today's Class

The G-Matrix, GBLUP, and beyond the additive model

Different Methods to Predict Marker Effect

- **BLUP-Based**: G-BLUP, RR-BLUP, RR-BLUP_B
- **Bayes-Based**: BayesA, BayesB, BayesCπ, BayesR
- **LASSO-Based**: Bayesian Lasso Regression, Improved Lasso
- **Semi-Parametric Regression**: RKHS
- **Non-Parametrics:** Suport Vector Machine, Neural-Networks
- **Others...**

Meuwissen et al 2001; Habier et al 2011; De los Campos et al 2009; Legarra et al 2011; Gianola et al 2006; Long et al 2011; Gianola et al 2011

Basic Linear Mixed Model

Yield = Environment Effect+ Genetic Effect + Residual

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j sum of additive (g_a), dominance (g_d), and epistasis (g_i)

e_{ij} random error of the ij^{th} observation

g_j → average additive effect of genes an individual receive from both parents (breeding value)

Each parent contributes a sample half of its genes to its progeny. The average effect of of this sample is the general combining ability (GCA) of the parent → half of the breeding value

Basic Linear Mixed Model

Yield = Environment Effect+ Genetic Effect + Residual

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j sum of additive (g_a), dominance (g_d), and epistasis (g_i)

e_{ij} random error of the ij^{th} observation

Breeding value (BV) of progeny then is the sum of the GCA of both parents.

Since the GCA is a function of the genes transmitted from parents to progeny, it is the only components that can be selected for.

Dominance and epistasis assumed to be insignificants.

Basic Linear Mixed Model

Yield = Environment Effect+ Genetic Effect + Residual

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j sum of additive (g_a), dominance (g_d), and epistasis (g_i)

e_{ij} random error of the ij^{th} observation

It is assumed that y follows a multivariate normal distribution (MVN), implying that traits are determined by infinitely many additive genes of infinitesimal effect, infinitesimal model (Fisher, 1918).

Thus BV for individual i → $a_i = g_a = 1/2a_f + 1/2a_m + m_i$

m_i = *Mendelian sampling*

I can estimate the BV of an individuals if I know the parents and their BV!!!

The Mixed Model Equation

Consider a model with block as fixed and variety as random effects.

yield = μ + block + variety + error

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j random effect of the j^{th} variety, $E(g_j) = 0$, $V(g_j) = \mathbf{A}\sigma_g^2 = \mathbf{G}$

e_{ij} random error of the ij^{th} observation, $E(e_{ij}) = 0$, $V(e_{ij}) = \mathbf{I}\sigma^2 = \mathbf{R}$

$Cov(g_j, e_{ij}) = 0$

Variances need to be estimated first with one method: OLS, method of the moments, ML, REML, or Bayesian

Note: G in this slide is not the Genomic Relationship Matrix

Variance Component Estimation

- Variance components need to be estimated before obtaining estimates of fixed/random effects and performing any type of inference.

$$\hat{\mathbf{G}} = \mathbf{G}(\hat{\theta})$$
$$\hat{\mathbf{R}} = \mathbf{R}(\hat{\theta})$$
$$\Rightarrow \hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}'+\hat{\mathbf{R}}$$

- Restricted/residual maximum likelihood (REML) is a likelihood-based method used to estimate these variance components and is based assuming that both **g** and **e** follow a multivariate normal distribution.
- The REML variance component estimates are later used to estimate the solutions of fixed and random effects.
- Henderson (1950) derived the Mixed Model Equations (MME) to obtain the solutions of all effects:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \qquad \text{BLUE} \rightarrow \text{EBLUE}$$
$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad \text{BLUP} \rightarrow \text{EBLUP}$$

Breeding Value Prediction and BLUP

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$\hat{\mathbf{g}}$	vector of random effect predictions.
$\hat{\mathbf{G}}\mathbf{Z}' = \mathbf{C}'$	covariance matrix between observations and random (genetic) effects to be predicted.
$\hat{\mathbf{V}}$	variance-covariance matrix for the observations.
$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$	individual observations 'corrected' by fixed effects.

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$\hat{g}_i = [\sigma_a^2 / \sigma_p^2] \times (y_i - \bar{y})$$

$\rightarrow$ ΔGain

$$\hat{g}_i = h^2 \times (y_i - \bar{y})$$

Note: the expression changes depending of what model is being evaluated (**y**).

Pedigree and Relationship Matrices

- Why worry about the pedigree in genetic analyses?

  Statistically, random genetic effects (i.e. BLUPs) are not independent and their matrix of correlations or co-variances (**G** or **A**) needs to be specified.

  Genetically, it is important to consider information about relatives as they will share some alleles, and therefore their response is correlated.

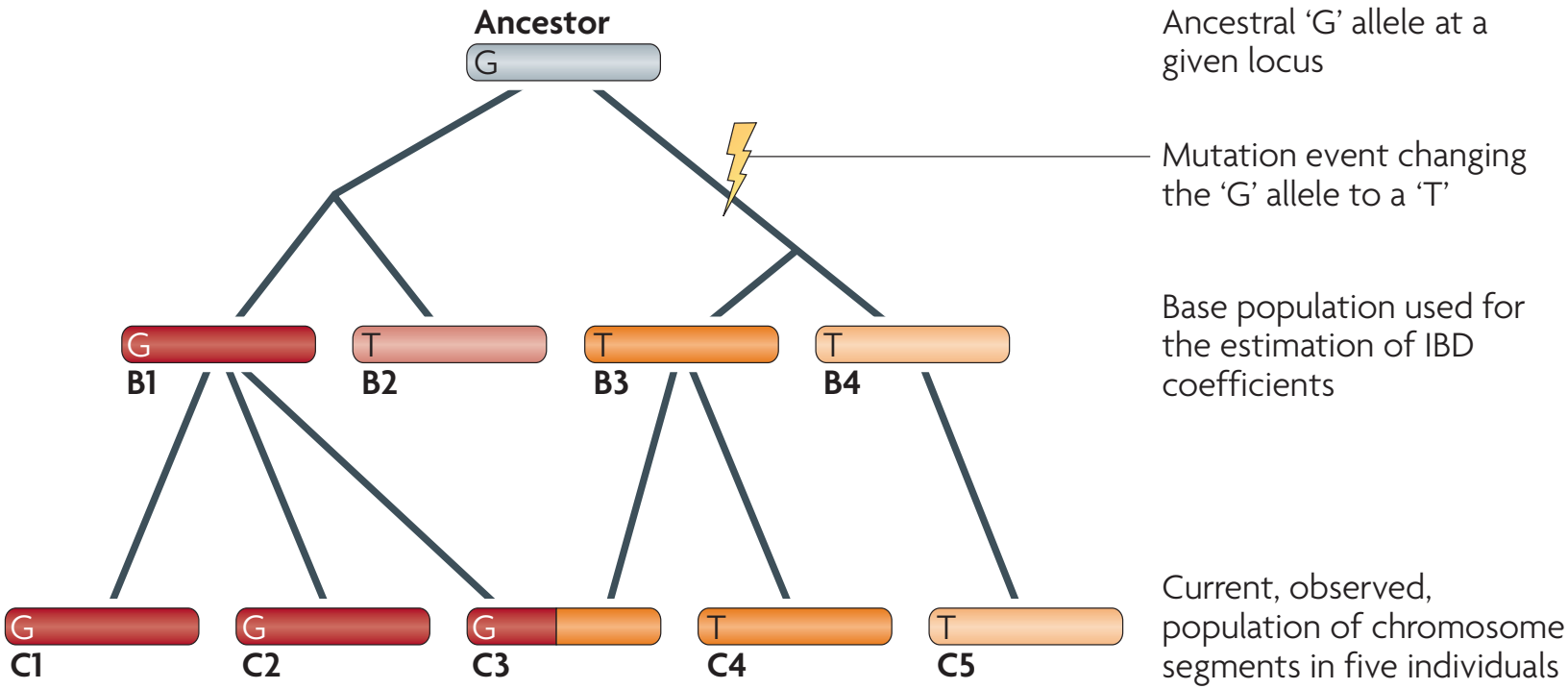
- How to incorporate this information?

  Genetic relationships (pedigree) can be calculated using genetic theory (expected values) or molecular information (e.g. SNPs), and included into the linear mixed model by specifying a Relationship matrix

- Are there other benefits?

  Many. It is a more efficient use of the information about individuals, but also genetic values of individual not tested, but with relatives tested, can be *predicted* and selected.

Pedigree and Relationship Matrices



Ancestor is the point of coalescent for current alleles C1-C5.
Identity by descent (IBD) of current alleles can be defined respect to B1-B4, thus
G allele in C1-C3 are IBD
T allele in C4 and C5 are identity by state (IBS)
The chromosome segment C2 and C3 are IBS as well. *Powell et al 2010*

Pedigree and Relationship Matrices

- The additive genetic relationship between two individuals is twice their IBD (aka coancestry or kinship coefficient)

- The matrix that indicates the additive genetic relationship among individuals is the numerator relationship matrix (**A**).
- Properties of **A:**

  - Is symmetric

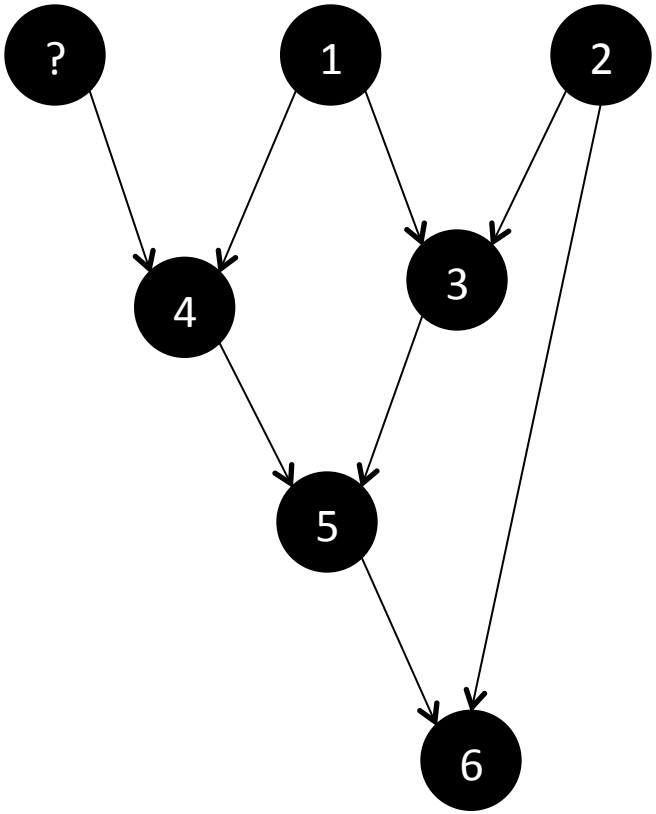
  - Diagonal elements is equal to $1+F_i$ (F_i is the inbreeding coefficient on individual i)

  - Can be computed by different methods, the recursive method (Henderson 1976) is simpler.

Pedigree and Relationship Matrices

Example

Pedigree of a group of individuals:



Individual	Male	Female
3	1	2
4	1	Unknown
5	4	3
6	5	2

Pedigree and Relationship Matrices

Numerator relationship matrix (A)

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1.00 & 0.00 & 0.50 & 0.50 & 0.50 & 0.25 \\ & 1.00 & 0.50 & 0.00 & 0.25 & 0.625 \\ & & 1.00 & 0.25 & 0.625 & 0.563 \\ & & & 1.00 & 0.625 & 0.313 \\ & & & & 1.125 & 0.688 \\ & & & & & 1.125 \end{bmatrix} \end{array}$$

- Linked to the concept of **identity by descent**.
- **Diagonal** $a_{ii} = 1 + F_i$ (inbreeding coefficient on individual *i*)

  Twice the probability that two gametes taken at random from animal *i* will carry identical alleles by descent.
- **Off-diagonal** a_{ij} numerator of the coefficient of relationship between animal *i* and *j*.
- Several algorithms are available in ASReml to obtain this matrix.

Pedigree and Relationship Matrices

CALCULATING THE A MATRIX

- Let $\mathbf{A} = \{a_{ij}\}$ be the relationship matrix.
- Let $a_{i,-j}$ the the *i-th* row of **A** except for the *j-th* element.
- Assume the relationship matrix for the base individual is known (e.g. unrelated, non inbred). This will for a base matrix (e.g. identity)
- The row of the relationship matrix for the progeny of two parents is generates as the average of the relationship matrix rows for the parents:

$$a_{i,-j} = (a_{s,-i} + a_{d,-i})/2$$

- The diagonal element, $a_{i,i}$ of this new individual is:

$$a_{i,i} = 1 + a_{s,d}/2 = 1 + F_i$$

where F_i is the inbreeding coefficient.

Pedigree and Relationship Matrices

Construction / Check

- Pedigree information is associated with proper management and validation/check of data.
- Individuals need to be ordered by generation (e.g. parents need to be defined before progeny).
- All parents need to be defined in pedigree file (the inclusion of founder parents is optional).
- All individuals present in dataset (i.e. levels associated with pedigree file) need to be defined in pedigree file.
- Individuals can be defined as male or female parents (but this should be checked if is not biologically possible).

Genome Selection: Animal Model/ GBLUP

$$y_i = a_i + \varepsilon_i \quad (i = 1,\ldots,n)$$

$$a \sim MVN\left[0, G\sigma_a^2\right] \qquad \varepsilon \sim MVN\left[0, I\sigma_\varepsilon^2\right]$$

Additive Relationship Matrix $G_A \propto coancestry$

Genomic Relationship Matrix $G_M \propto XX'$

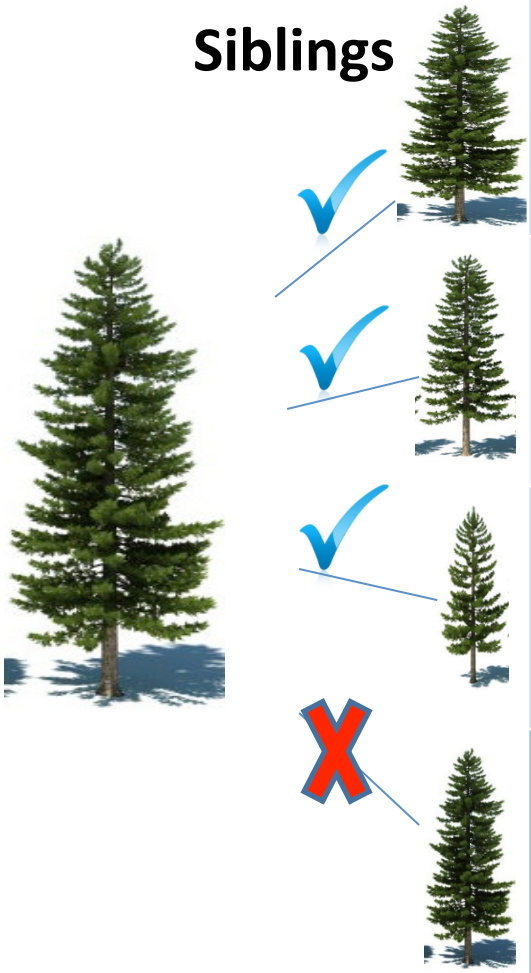
$$\text{TRN} \Rightarrow \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

$$\text{TST} \Rightarrow$$

$$\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim MVN\left[\mathbf{0}, \begin{pmatrix} \mathbf{G}_1 & \mathbf{G}_{12}\sigma_a^2 & \mathbf{G}_{22}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{G}_{21}\sigma_a^2 & \mathbf{G}_{22}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_1\sigma_\varepsilon^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_2\sigma_\varepsilon^2 \end{pmatrix}\right]$$

Courtesy: Gustavo de los Campos

Use of relationships to estimate genetic effects



Pedigree Relationships	Marker Relationships	Benefit
0.5	0.38	Improve relationship estimation; By estimating Mendelian segregation
0.5	0.52	
0.5	0.63	
0.0	0.40	Estimate relationships inexistent by pedigree

From markers to G

Example:

$$\hat{a} = (X - P)\hat{g} \qquad P = E(X) = 2p; \cdots X - P = W$$

$$W = \begin{bmatrix} 1-p_i & 0-p_i & 1-p_i & 2-p_i \\ 2-p_i & 2-p_i & 0-p_i & 2-p_i \\ 2-p_i & 1-p_i & 1-p_i & 0-p_i \\ 0-p_i & 2-p_i & 2-p_i & 1-p_i \end{bmatrix} \quad \hat{g} = \begin{bmatrix} 0.24 \\ 0.02 \\ -0.08 \\ 0.14 \end{bmatrix} \quad \hat{a} = \begin{bmatrix} 0.44 \\ 0.80 \\ 0.42 \\ 0.02 \end{bmatrix}$$

- If the markers are capturing all genetic variation, then we can assume that (Van Raden, 2008): $a = Wg$
- If we also assume: $V(g) = I\sigma_g^2$
- Then we get: $V(a) = WW'\sigma_g^2$

which is a covariance matrix for the individual breeding values **"a"**

From markers to G

- Ideally, we want to model this covariance using the same classical Linear Mixed Model framework, therefore, it would be desirable to have this matrix in terms of σ^2_a

$$\sigma_a^2 = \sum_{i=1}^{ALL\_SNPs} 2 p_i q_i \sigma_g^2 \longrightarrow \sigma_g^2 = \frac{\sigma_a^2}{\sum_{i=1}^{ALL\_SNPs} 2 p_i q_i}$$

- Gianola *et al.* (2009) showed that (under Hardy-Weinberg equilibrium):

$$V(a) = WW' \sigma_g^2$$

  - If we recall then:

$$V(a) = \frac{WW' \sigma_a^2}{\sum_i 2 p_i q_i} = G_A \sigma_a^2$$

by replacing σ^2_m.

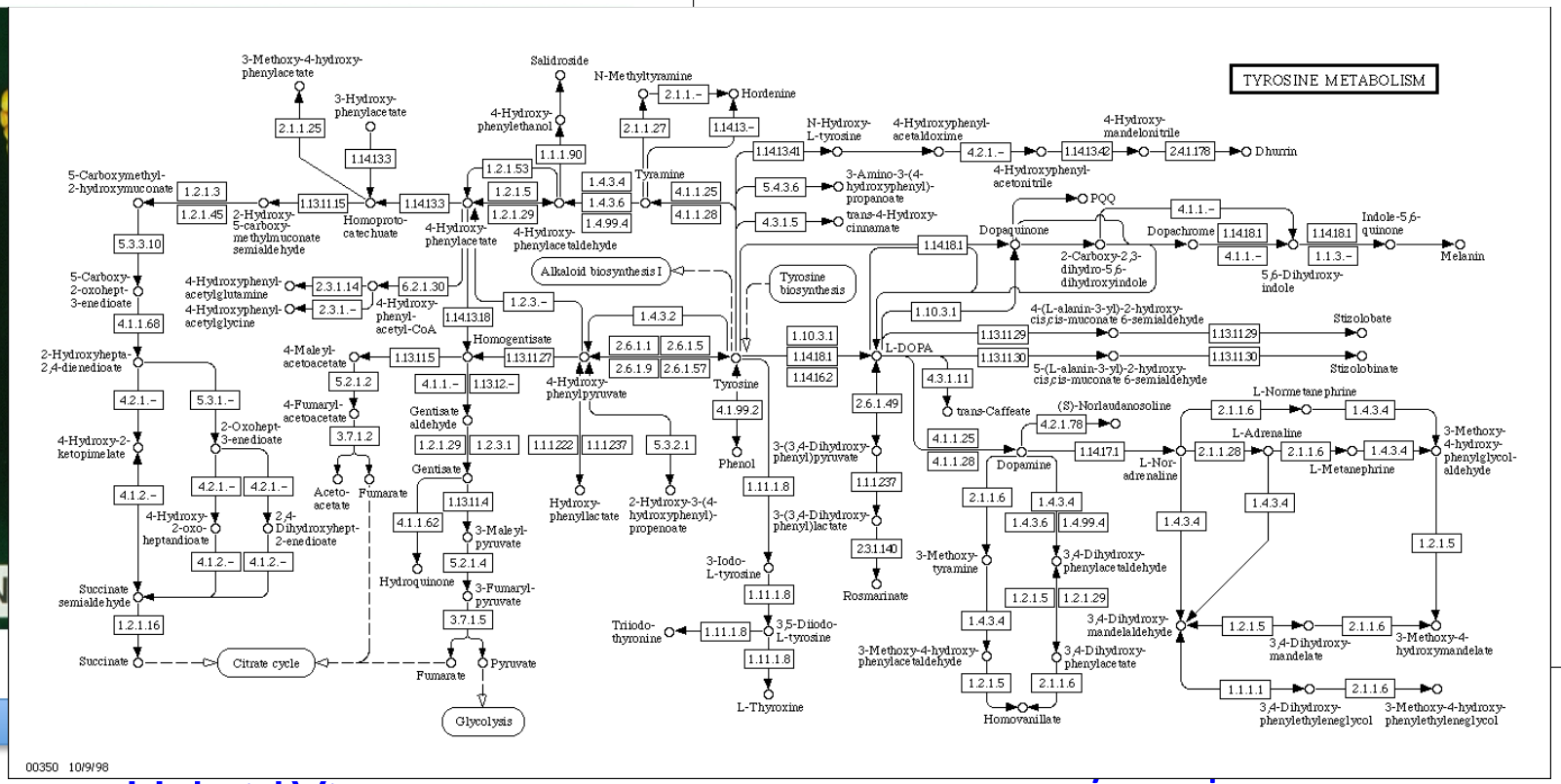
Non-Additive Effects Exist! - Evidence

Molecular evidence

- Breeding at gene level
- Interaction of different genes

Inbreeding Depression

DO WE REALLY THINK EVERYTHING IS ADDITIVE!!



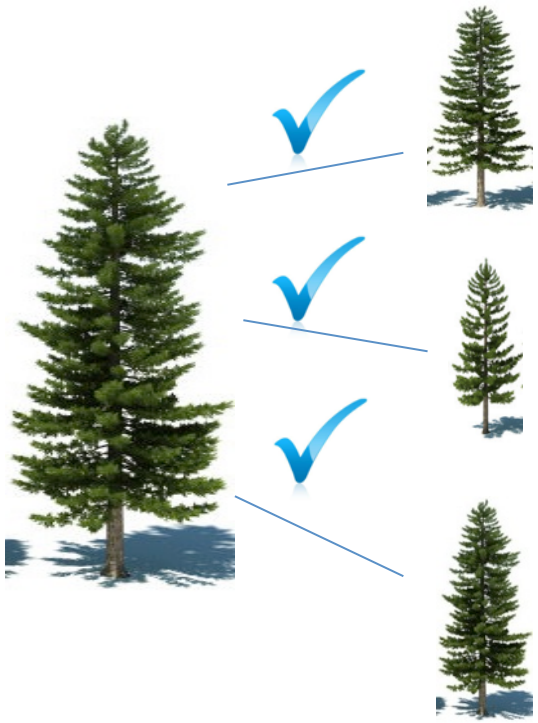
Hybrid Vigor

http://www.genome.jp/kegg/pathway.html

McKeand et al 2007

Markers-derived Relationships

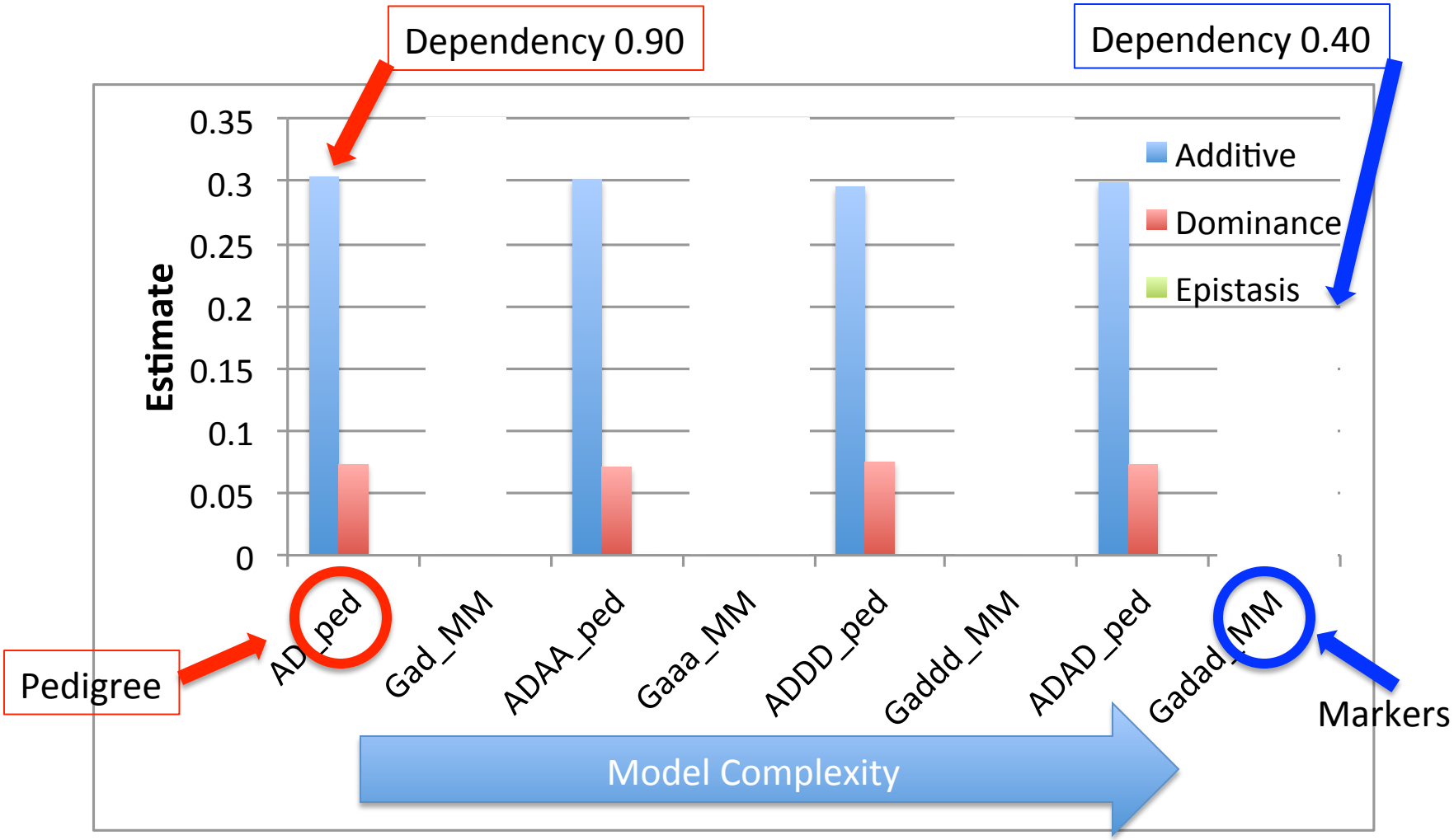
Siblings



Additive Pedigree Relationships	Dominance Pedigree Relationships
0.5	0.25
0.5	0.25
0.5	0.25

Van Raden (2008) Yang *et al.* (2010)

Estimates of Variance Components



Even when complexity of the model increases partition is the same with pedigree.

A different architecture is observed with marker-derived matrices

Prediction in 10-fold Cross Validation

Model	Cor(RBV,PBV)	MSE(RBV,PBV)	Top10%RankCor
A_ped	0.640	1335.800	0.17
Ga_MM	0.670	1291.800	0.34
ADAD_ped	0.727	657.258	0.16
Gadad_MM	0.872	108.240	0.37
ADDD_ped	0.732	638.464	0.18
Gaddd_MM	0.873	151.199	0.32

RBV → Breeding Value using all data within model
PBV → Breeding Value in cross validation

Top10%RankCor → correlation between R-ranking and P-ranking for top 10% performance

Munoz et al 2014 - GENETICS

Estimates of Variance Components

GENOMIC SELECTION

Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices

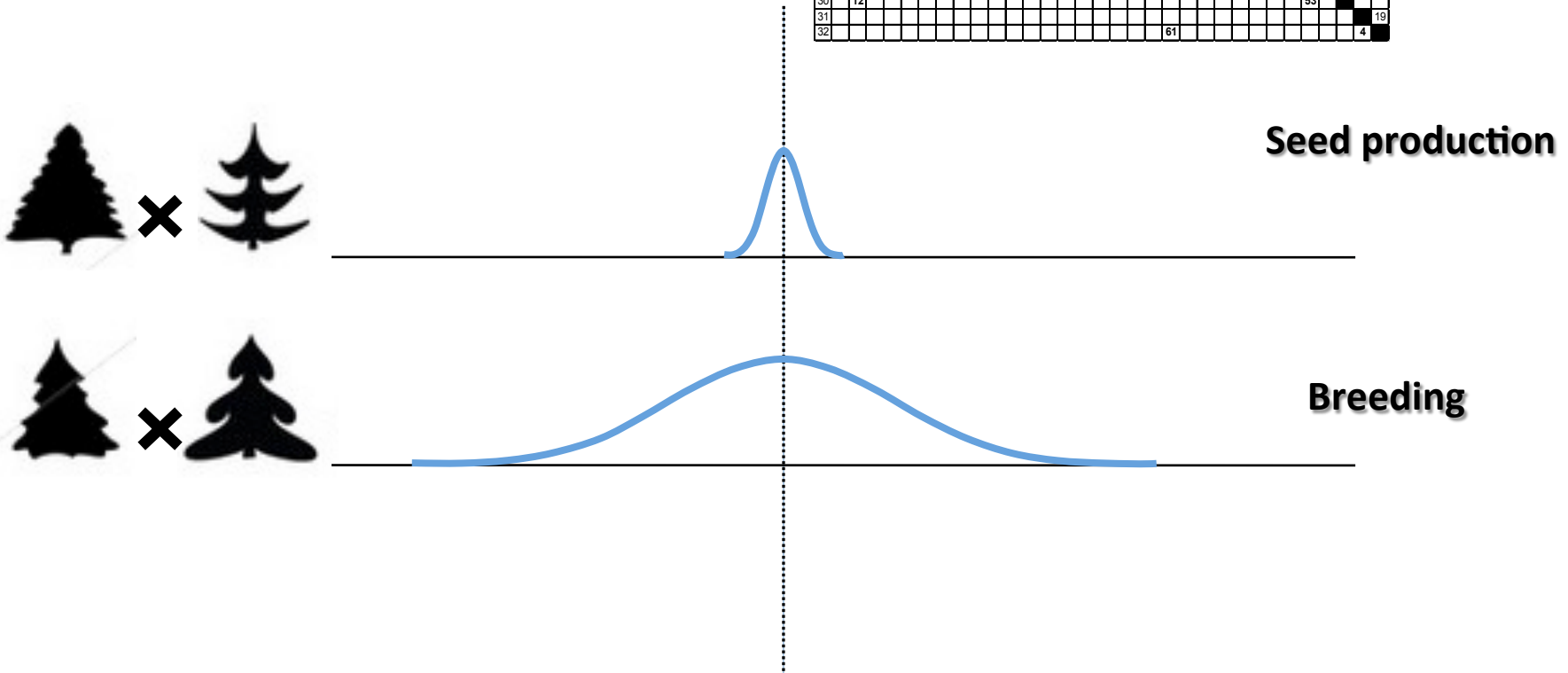
Patricio R. Muñoz,*,†,1,2 Marcio F. R. Resende, Jr.,†,‡,1 Salvador A. Gezan,† Marcos Deon Vilela Resende,§, Gustavo de los Campos,†† Matias Kirst,†,‡‡ Dudley Huber,† and Gary F. Peter†,‡‡,2**

*Agronomy Department, †School of Forest Resources and Conservation, ‡Genetics and Genomics Graduate Program, and ‡‡University of Florida Genetics Institute, University of Florida, Gainesville, Florida 32611, §EMBRAPA Forestry, Estrada da Ribeira, Colombo, PR 83411-000 Brazil, **Department of Forest Engineering, Universidade Federal de Viçosa, Viçosa, MG 36571-000 Brazil, and ††Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama 35294

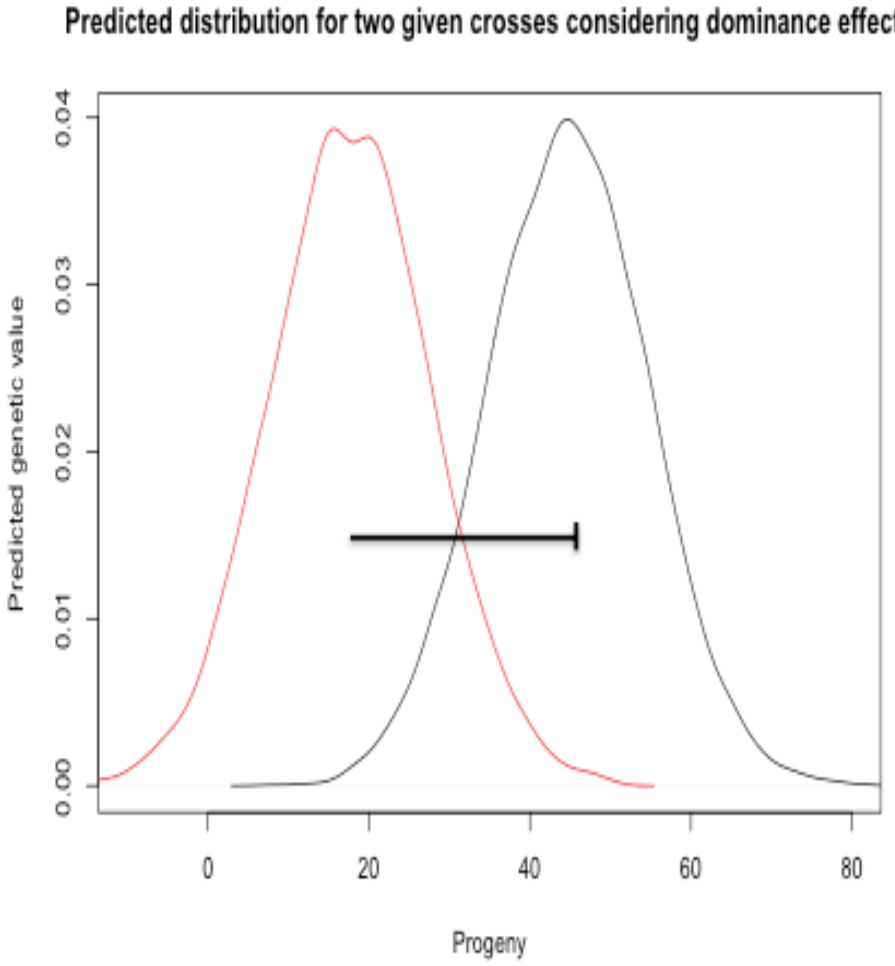
What should you cross?

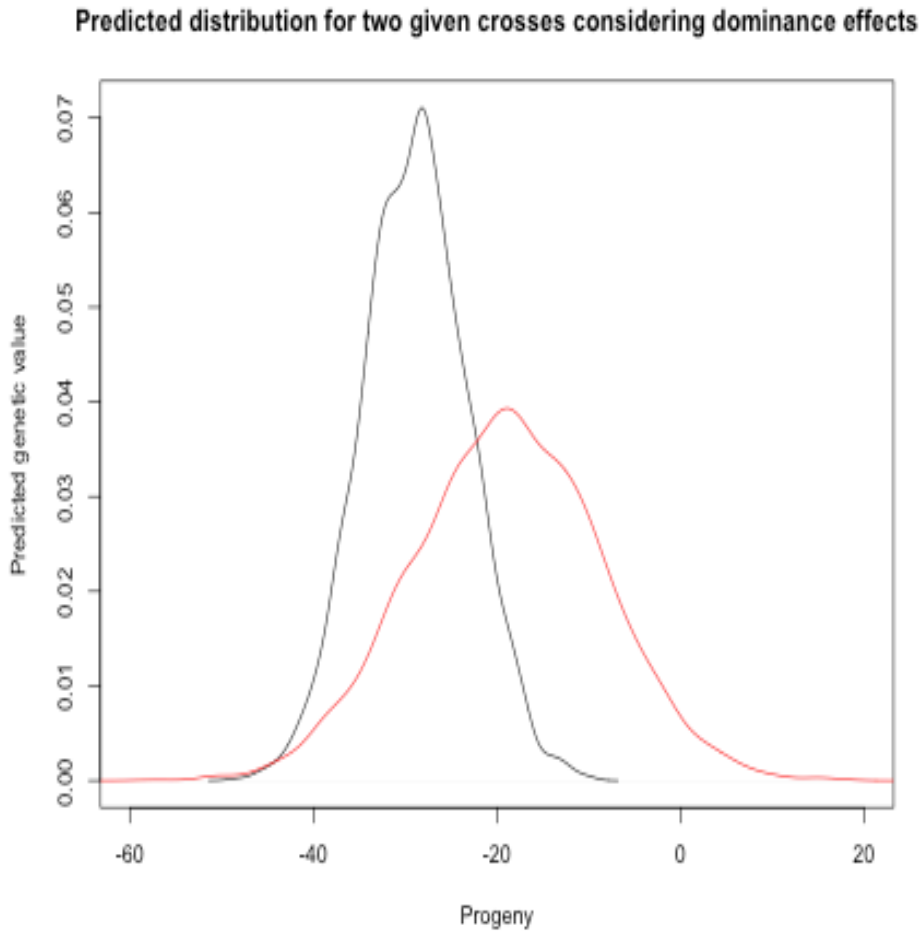
- ~950 individuals
- ~450,000 possible crosses

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1		16	17				18										19															
2	6		18	17									20															18	19		22	
3	34	23		19	19																											
4		69	37			15																										
5			58			19	18				20		21																			
6				15	10		19	19							18																	
7	45				11	64		21	19																							
8						50	56		18	19																						
9							40	26		20																						
10								25	49		21																					
11					39					44		18							21													
12											28		21	19												21						
13		35			41							19		19	21																	
14												46	65			18																
15						7							68			20	22															
16														70	38		22	21														
17	16														36	42		3	31													
18											27					60	0		21	22												
19																	1	22		22	15											
20																		30	57		21	20										21
21																			13	29			19									
22																			8		19	18							20			
23								48												67	20		18	20								
24																					51	55		17								
25												66										43	62		21	19						
26																								9		18	19					
27		54																						31	32		20					
28		33																							52	21		14	22			
29																						3					2					
30		12																									53					
31																															19	
32																				61									4			



GWS to Guide Crosses

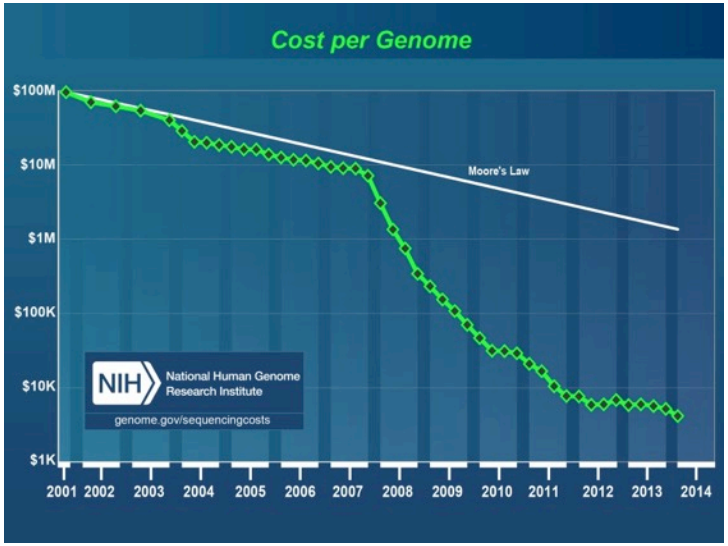




Family-Bulk Genome-wide Prediction

Goal: perform predictions using bulk genotyping and phenotyping

In many species “cultivars” corresponded to a population of genotypes (e.g. forages).



Even when genotyping cost has decreased significantly using NGS, it is still too high for most breeding programs.

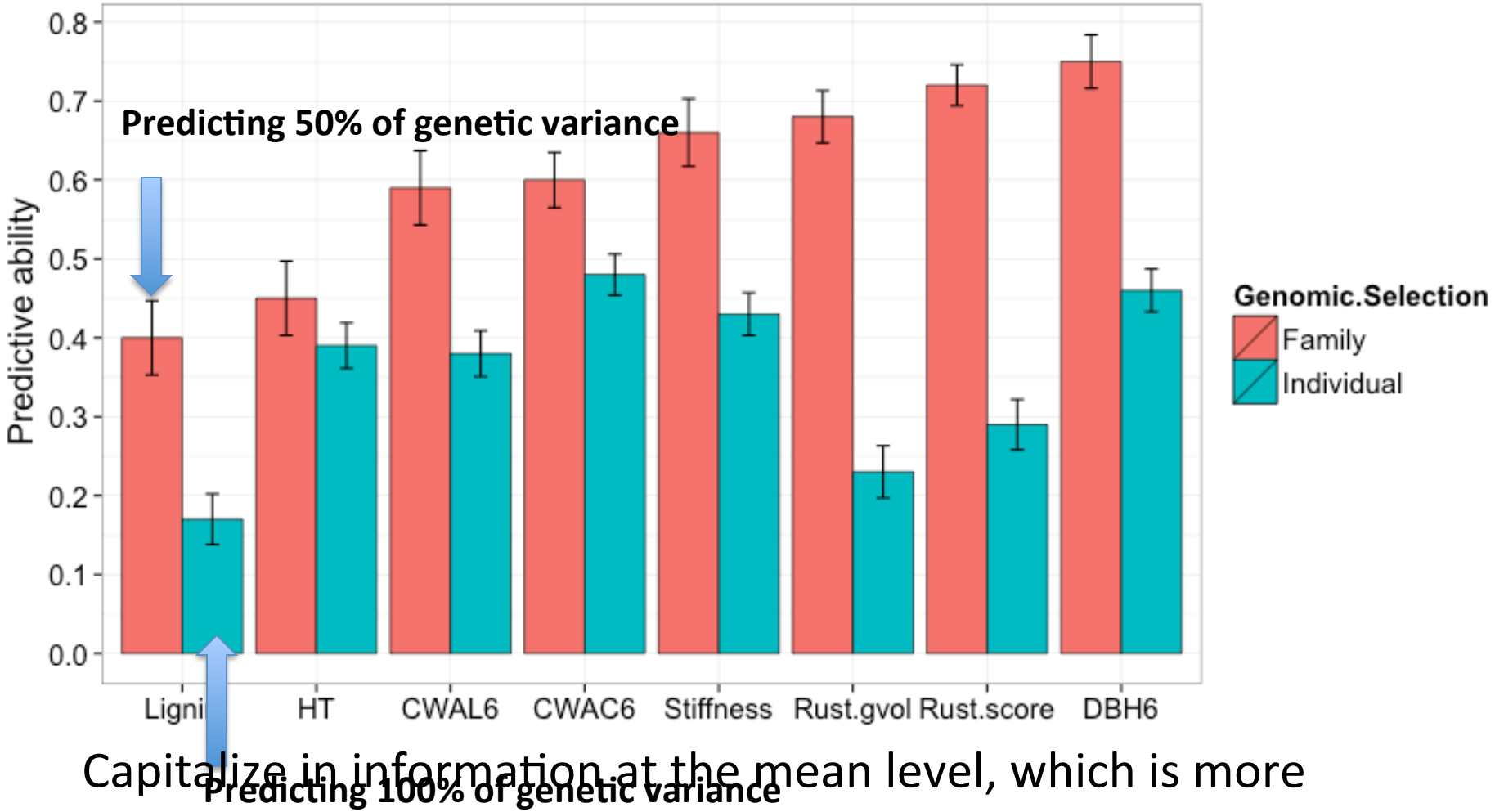
Model Species and Method

Pine breeding (CCLONES) population with 71 families and genotyped with Chip for ~4,700 SNPs markers

Phenotype: family phenotype mean

Genotype: family allele frequency

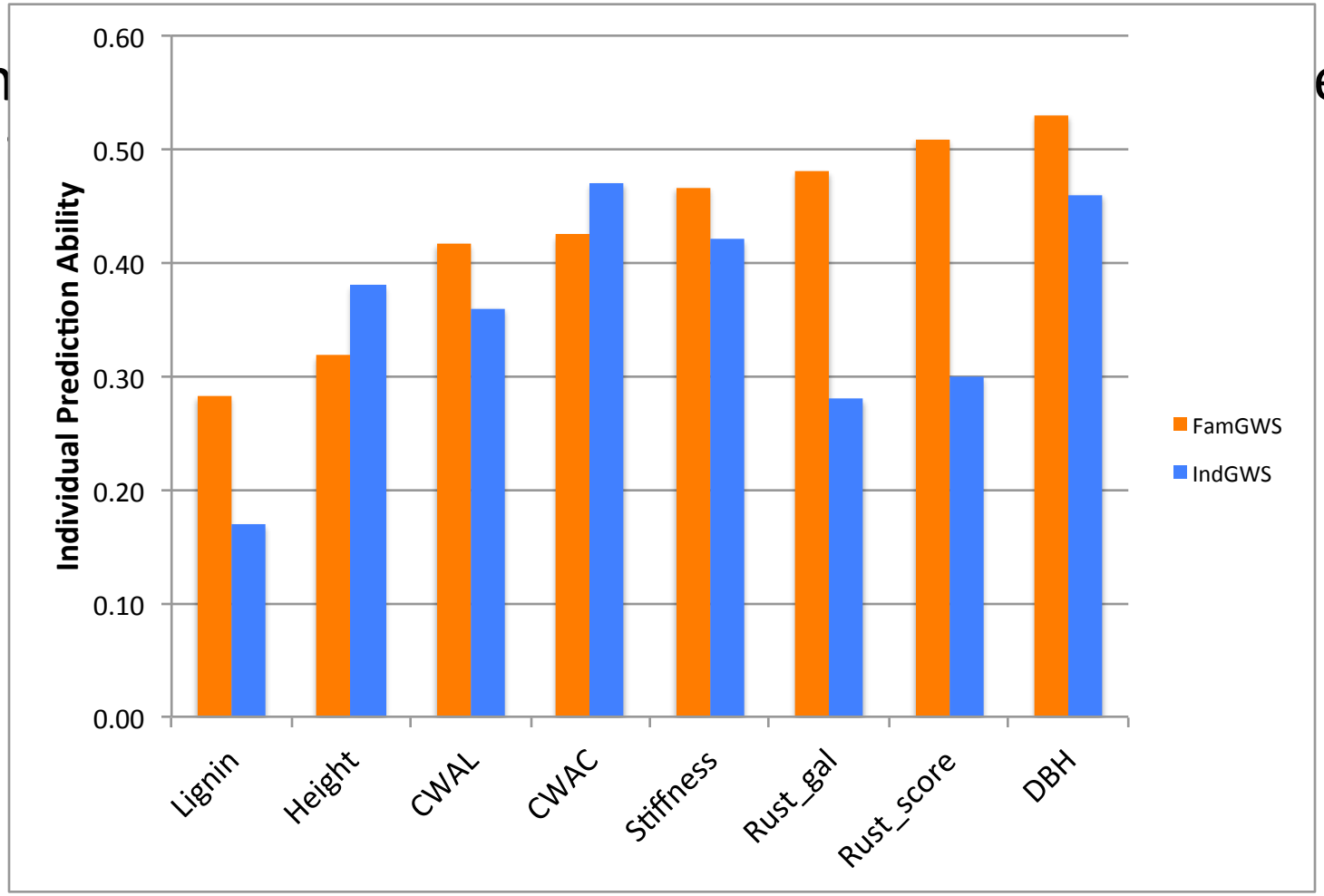
Family-Bulk Genotyping vs Traditional GWS Pine



Capitalize in information at the mean level, which is more precise than at the individual level. But…

Family-Bulk Genotyping vs Traditional GWS Pine

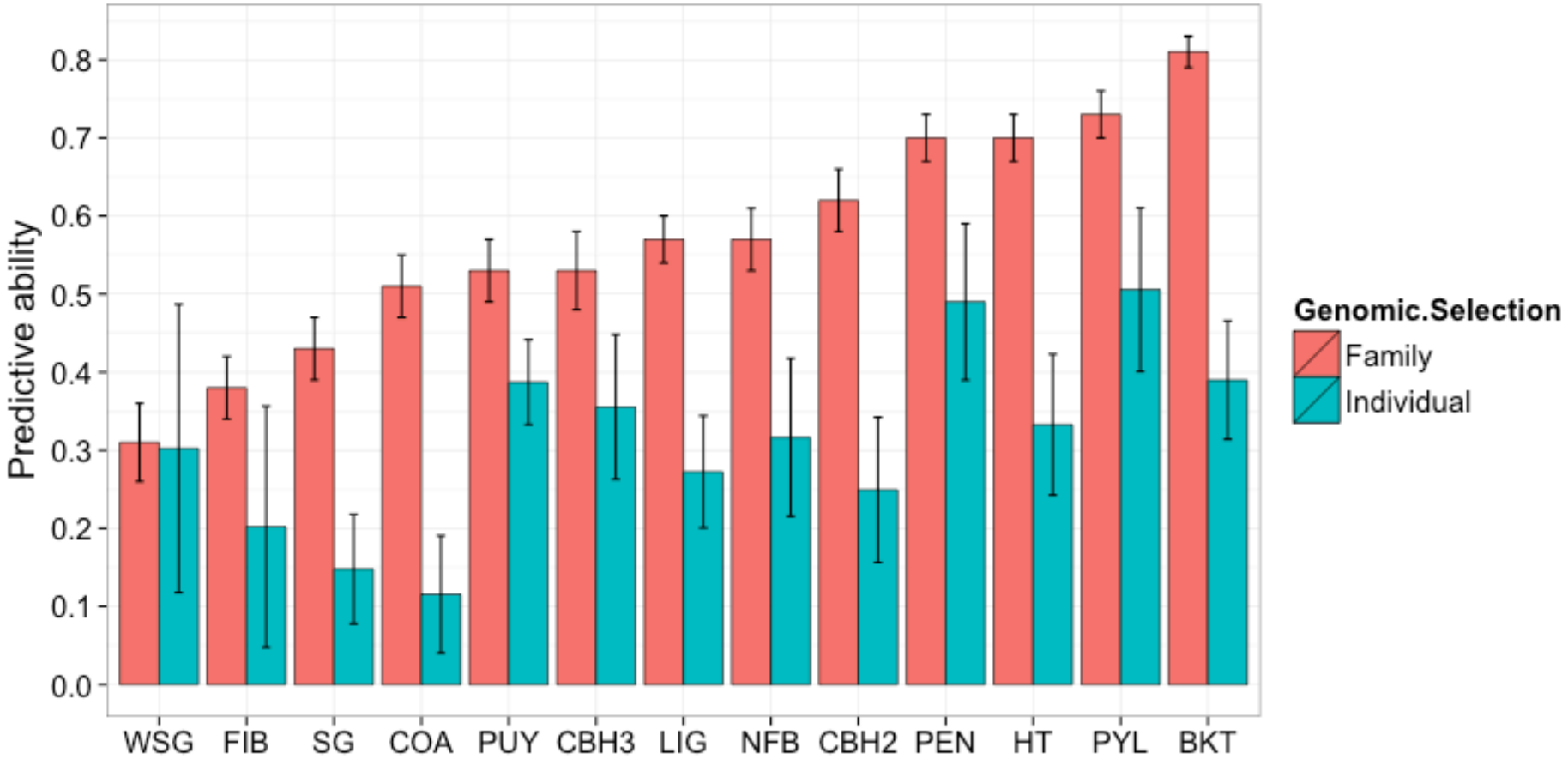
To n e
fam



Still better in 6 out of 8 traits

Family-Bulk Genotyping vs Traditional GWS Eucalyptus

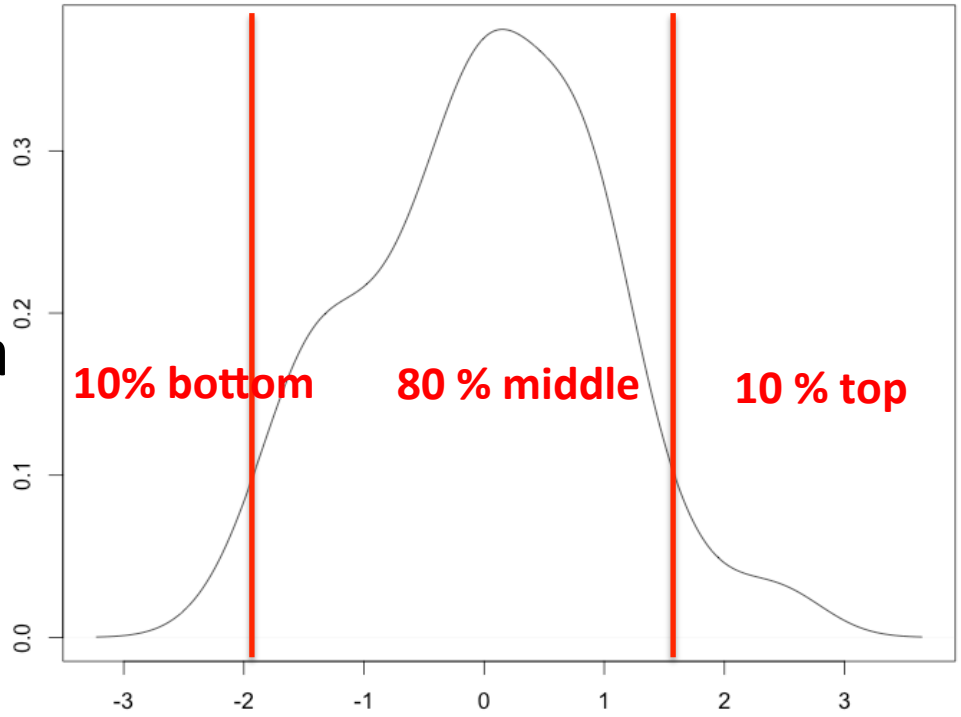
Eucalyptus breeding population with 68 families and genotyped with sequence capture for 500,000 SNPs markers.

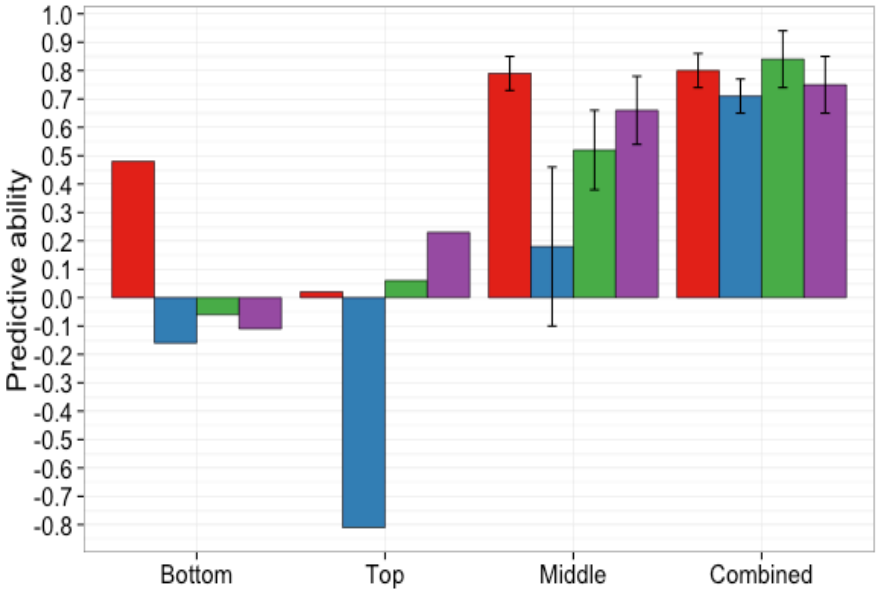


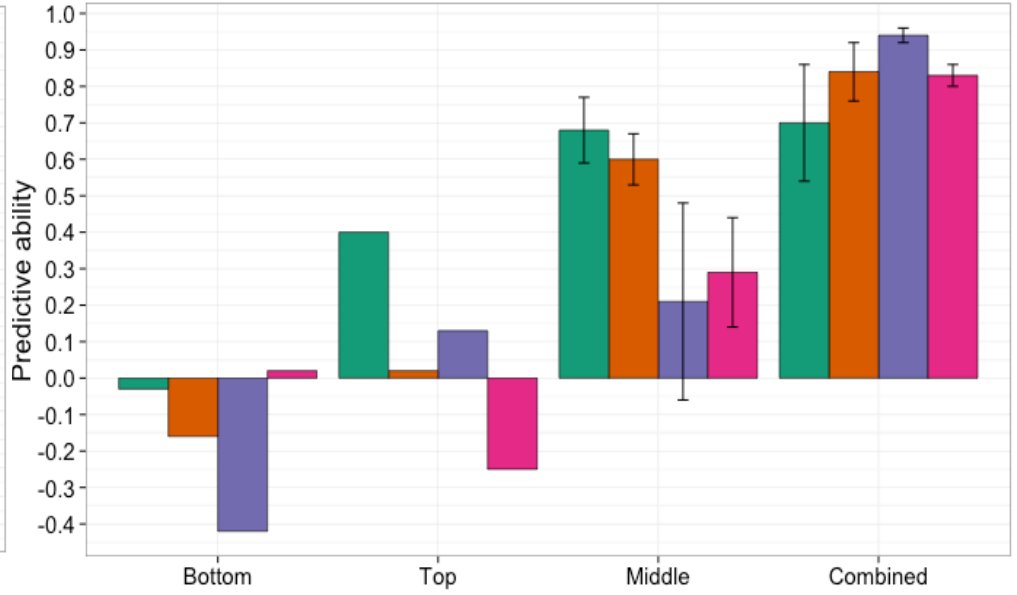
Still better in 8 out of 13 traits at the <u>individual level</u>

Challenges:

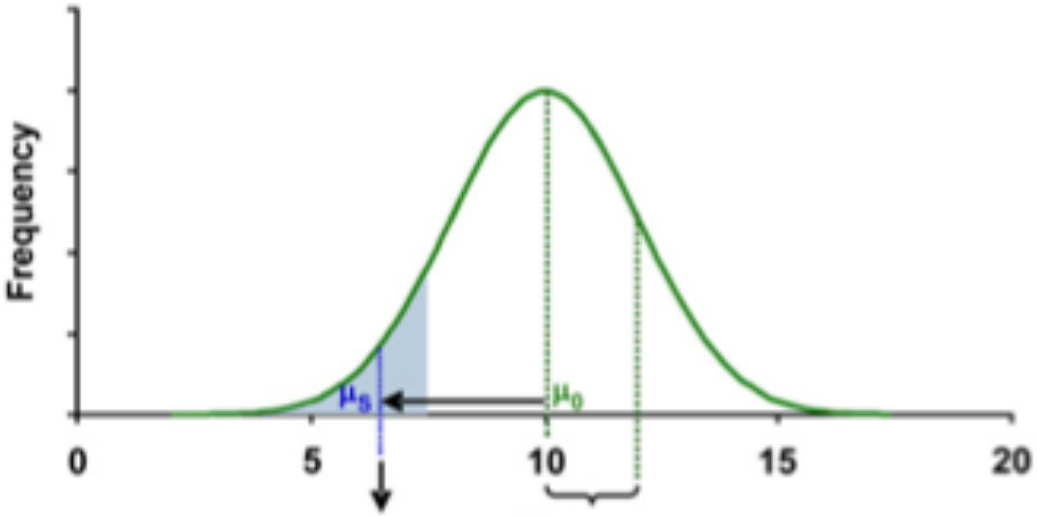
Very contra ation folds







The Goal!!



Genetic Gain = $\Delta G = h^2 \sigma_P i / L$ → length of cycle interval (usually 1 generation)

heritability

Selection Intensity
proportion of population selected to produce the next generation

$$\sigma^2_{Genotype} - \sigma^2_{Dominance} - \sigma^2_{Epistasis} = \sigma^2_{Additive}$$

$$\sigma^2_{Genotype} + \sigma^2_{Environment} + \sigma^2_{GxE} + \sigma^2_{error} = \sigma^2_P$$

phenotypic variability in population

Moose and Mumm 2008

Breeder's Equation

$$Genetic\_Gain = \Delta G = \frac{h^2 * \sigma_p^2 * i}{L}$$

h^2 = Narrow-sense heritability = is the portion of phenotypic variance due to additive genetic variation.

σ_p = phenotypic standard deviation = phenotypic variability in the population

i = intensity of selection

L = length of breeding cycle interval

Why this is important???