CFANS
COLLEGE OF FOOD, AGRICULTURAL AND NATURAL RESOURCE SCIENCES

PLANT BREEDING CENTER

Data Bootcamp for Genomic Prediction in Plant Breeding

Aaron Lorenz

University of Minnesota

Goals

- **My goal**: Provide to you basic (and some advanced) knowledge and skills so that you can begin to perform your own genomic prediction analyses on your own data.

- **Student goal**: Understand each line of each script well enough so you can begin to write your own scripts for studying genomic prediction using your own data.

Pedagogical methods

- Familiarize students with some basic and advanced topics in genomic prediction for plant breeding.
- Demonstrate some techniques for handling, formatting, and manipulating genome-wide marker datasets for use in genomic prediction.
  - Balance use of base R code with use of R packages.
- Provide examples and hands-on use of some R packages useful for various applications of genomic prediction.
- Provide opportunity for extensive hands-on experience performing genomic prediction analyses combined with line-by-line instruction.

Pre-requisites

Basic knowledge of molecular markers

What they are, various types, scoring, usage, etc.

Basic knowledge of statistics

Linear regression

ANOVA

Basic knowledge of plant breeding and genetics

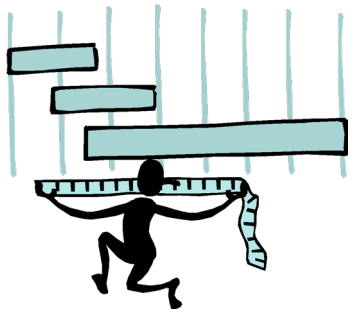
Basic knowledge of the R environment and programming language

Plant breeding in the 21st century

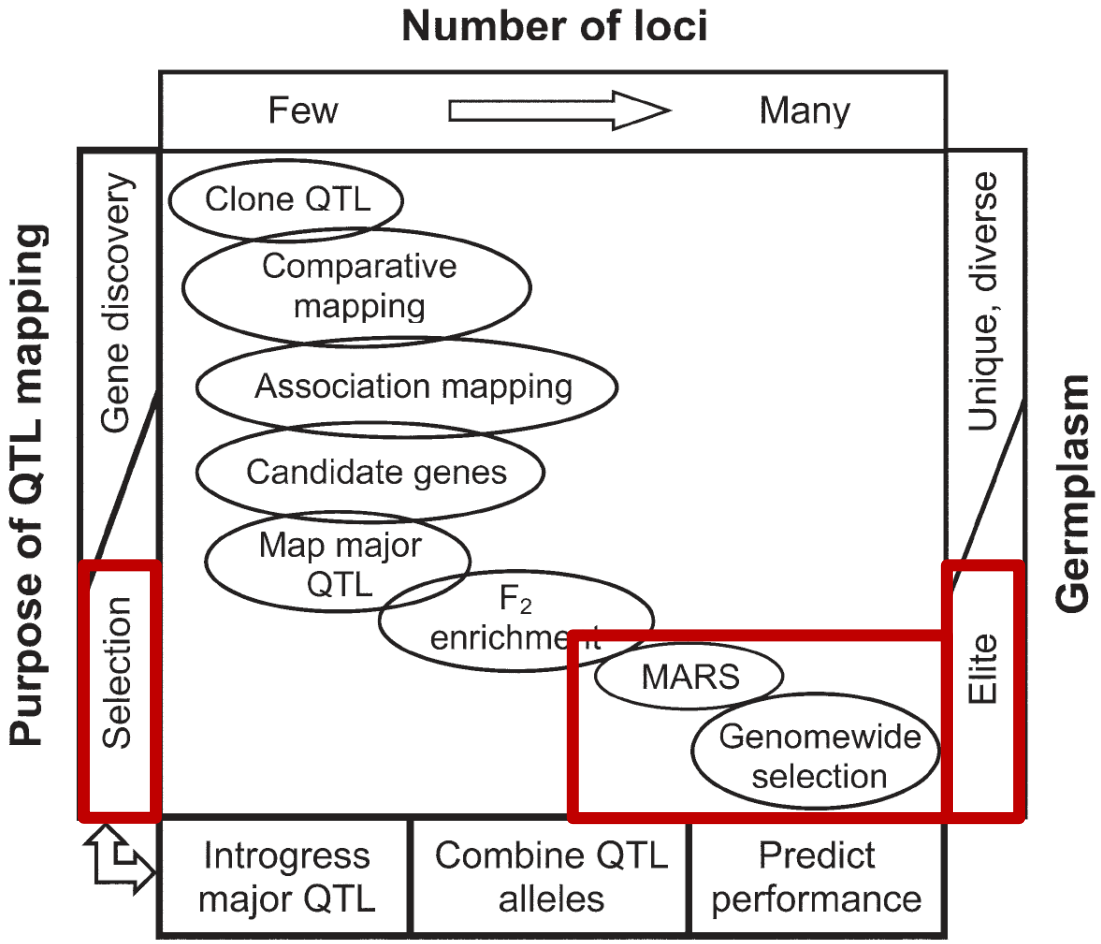
Two important trends

Genotypic data

Phenotypic data

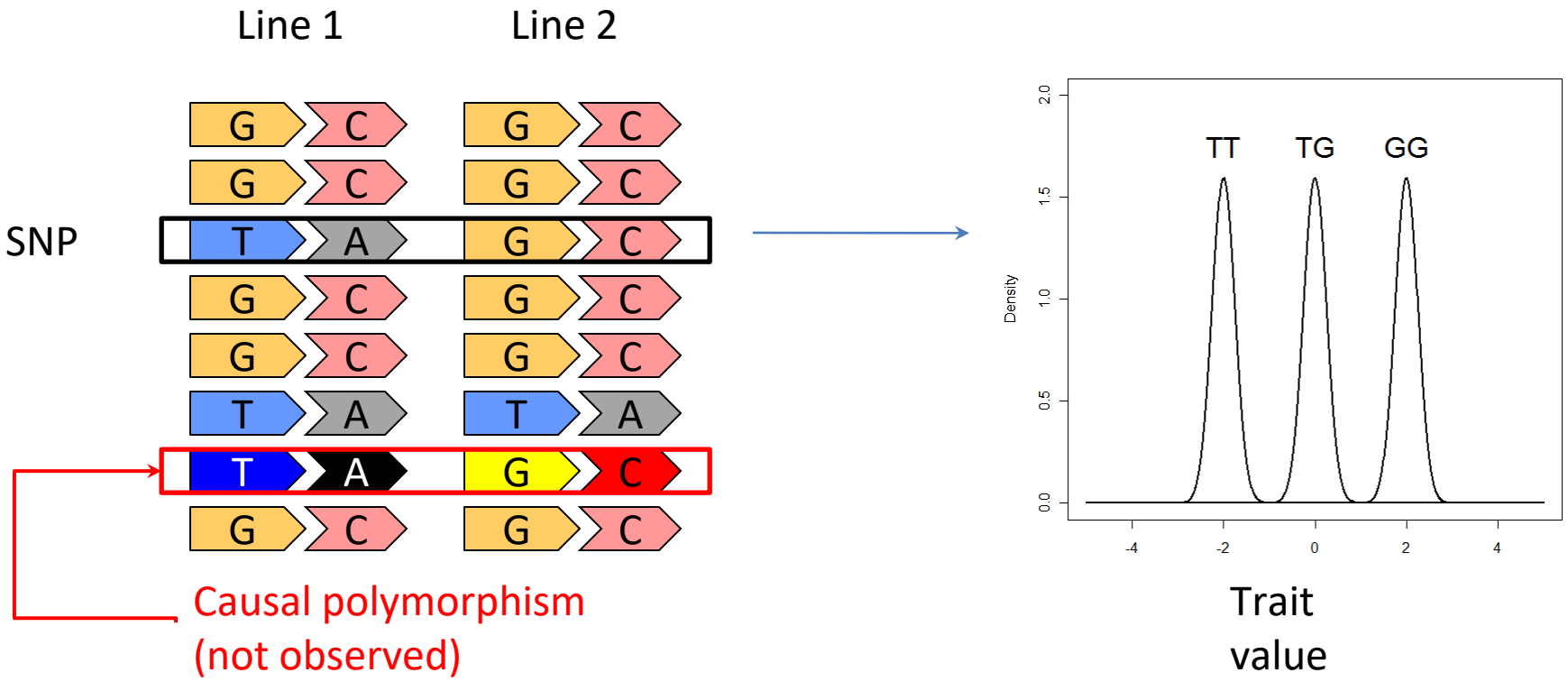


Focus for this course will be on using markers for prediction



Bernardo (2008)

Genotype Phenotype for genetic value



Reasons to incorporate molecular markers into selection decisions

1) **Marker information is potentially less expensive**

2) **Selections on marker information can be made during the off season**

3) **It's possible marker information provides more accurate predictions of genetic value**

Reasons to incorporate molecular markers into selection decisions

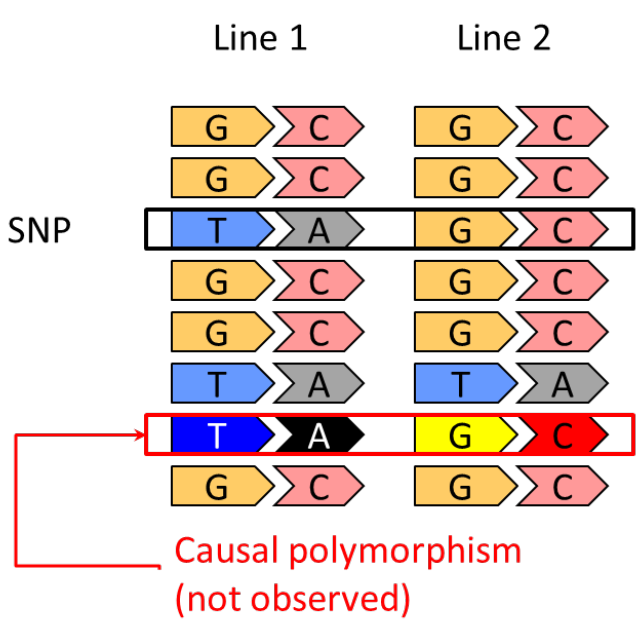
1) **Marker information is potentially less expensive** if phenotyping is laborious or requires specialized procedures, like detailed chemical analyses. Note that some phenotyping is relatively inexpensive, such as plant height and flowering time. Currently markers are not always cheaper than phenotypes, but they are cheaper in many cases.

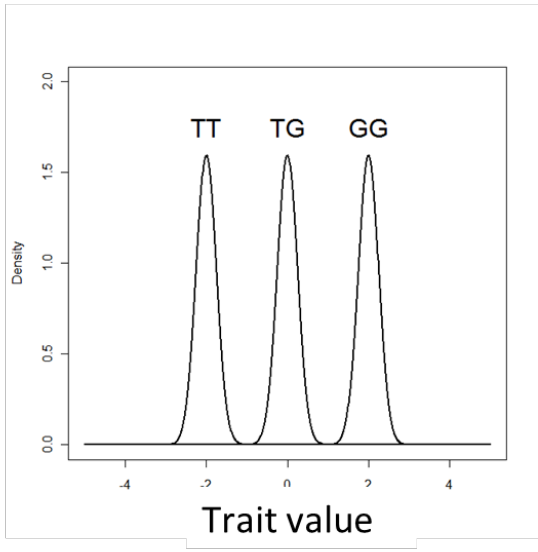
2) **Selections on marker information can be made during the off season**, thereby increasing the number of selection cycles per year compared to phenotypic selection when phenotypes can only be measured once per year in temperate environments.

3) For phenotypes of low heritability, it's **possible marker information provides more accurate predictions of genetic value** compared to phenotypes. This is especially true if large-effect QTL have been tagged by tightly linked markers, or very large training datasets are available for genomic prediction.

Linkage disequilibrium (LD) between markers and QTL

- LD refers to the non-random association of alleles between loci.
- Affected by:
  - Physical linkage, effective population size (drift), mating pattern, admixture, selection
- The rate of decay of LD over physical genomic distance determines the density of markers needed to effectively use markers for selection.





Linkage disequilibrium

$$D = p_{AB} - p_A p_B$$

p_{AB} = Frequency of gamete AB

p_A = Frequency of allele A

p_B = Frequency of allele B

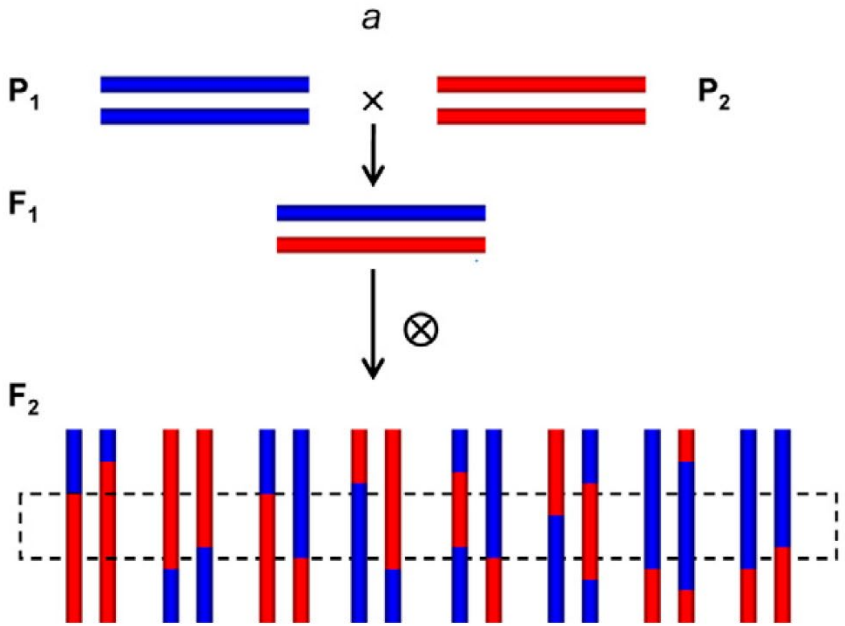
Note that *D* may be positive or negative depending on whether the alleles are in coupling or repulsion phase linkage

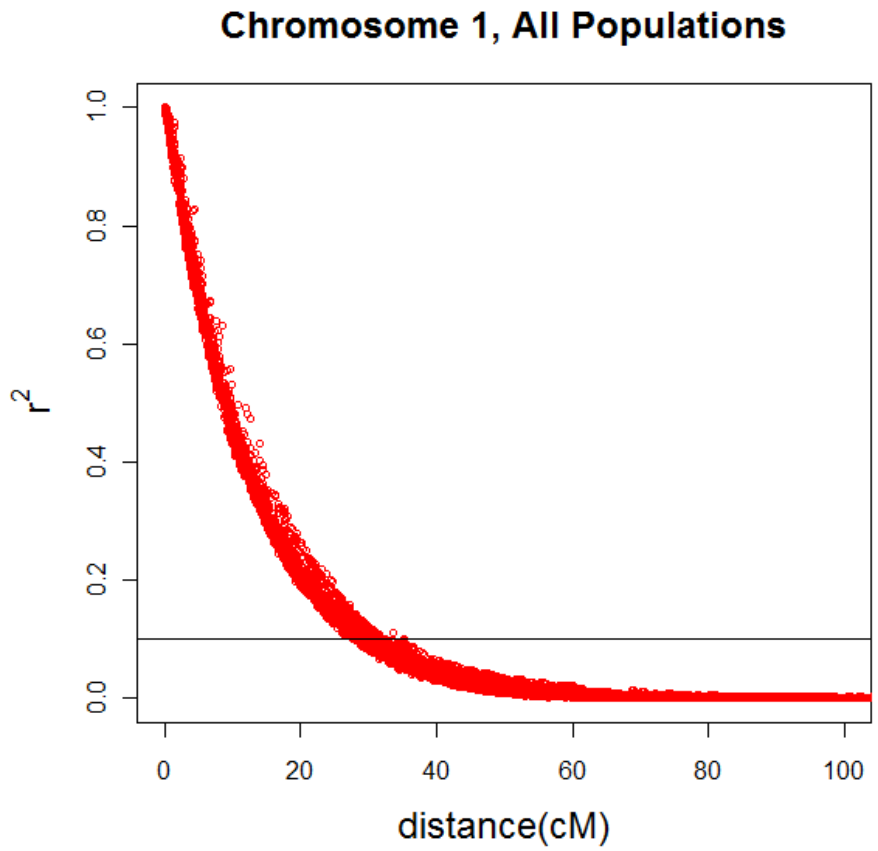
Common statistic to quantify LD.
Normalized value of D.

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

r^2 here is equivalent to square of Pearson's correlation coefficient between allelic states of loci

LD decay in bi-parental linkage mapping populations

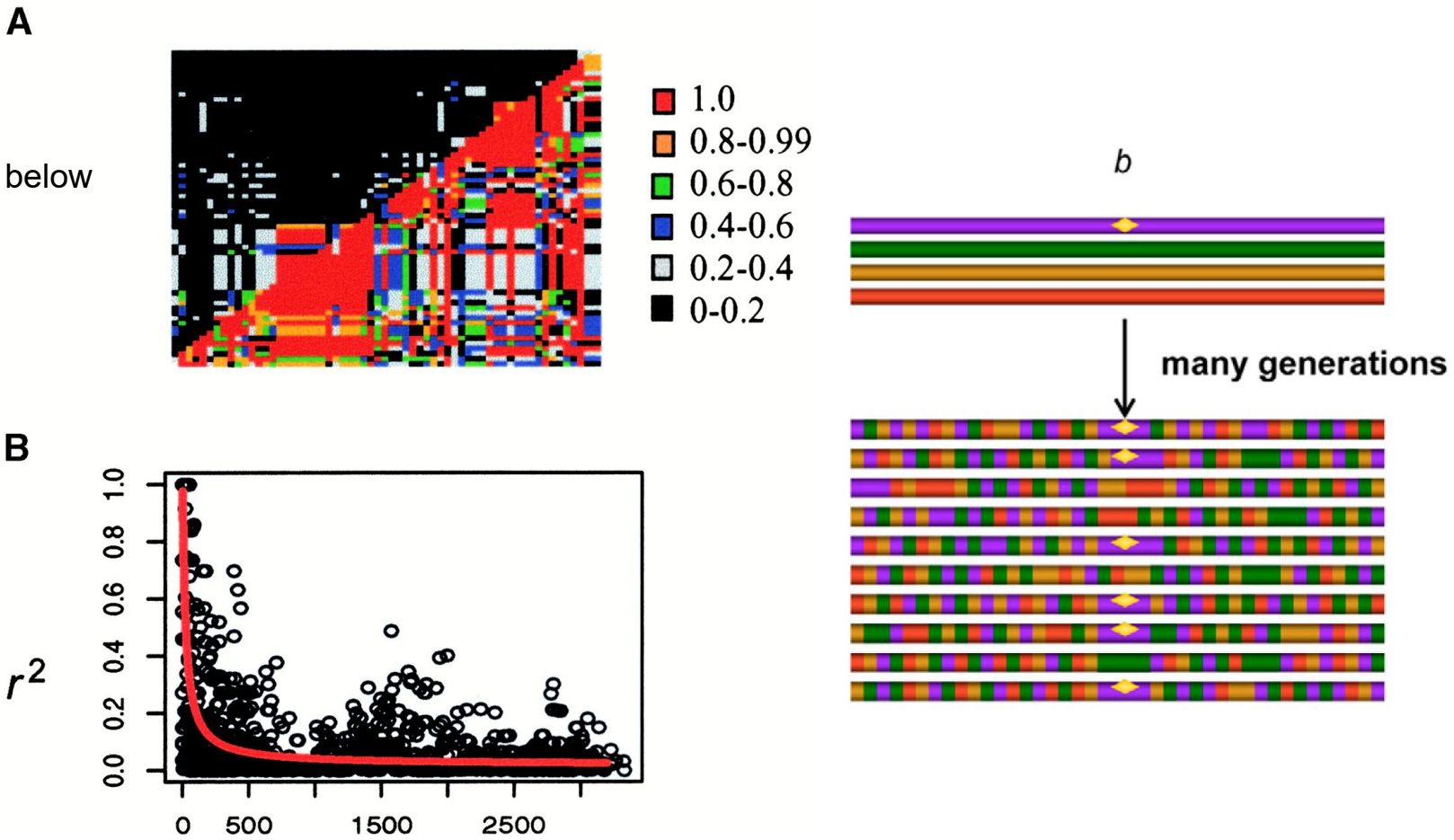




Peter Bradbury

Plots of LD across the Maize d3 Gene in a diverse panel of maize lines (Remington et al., 2001).

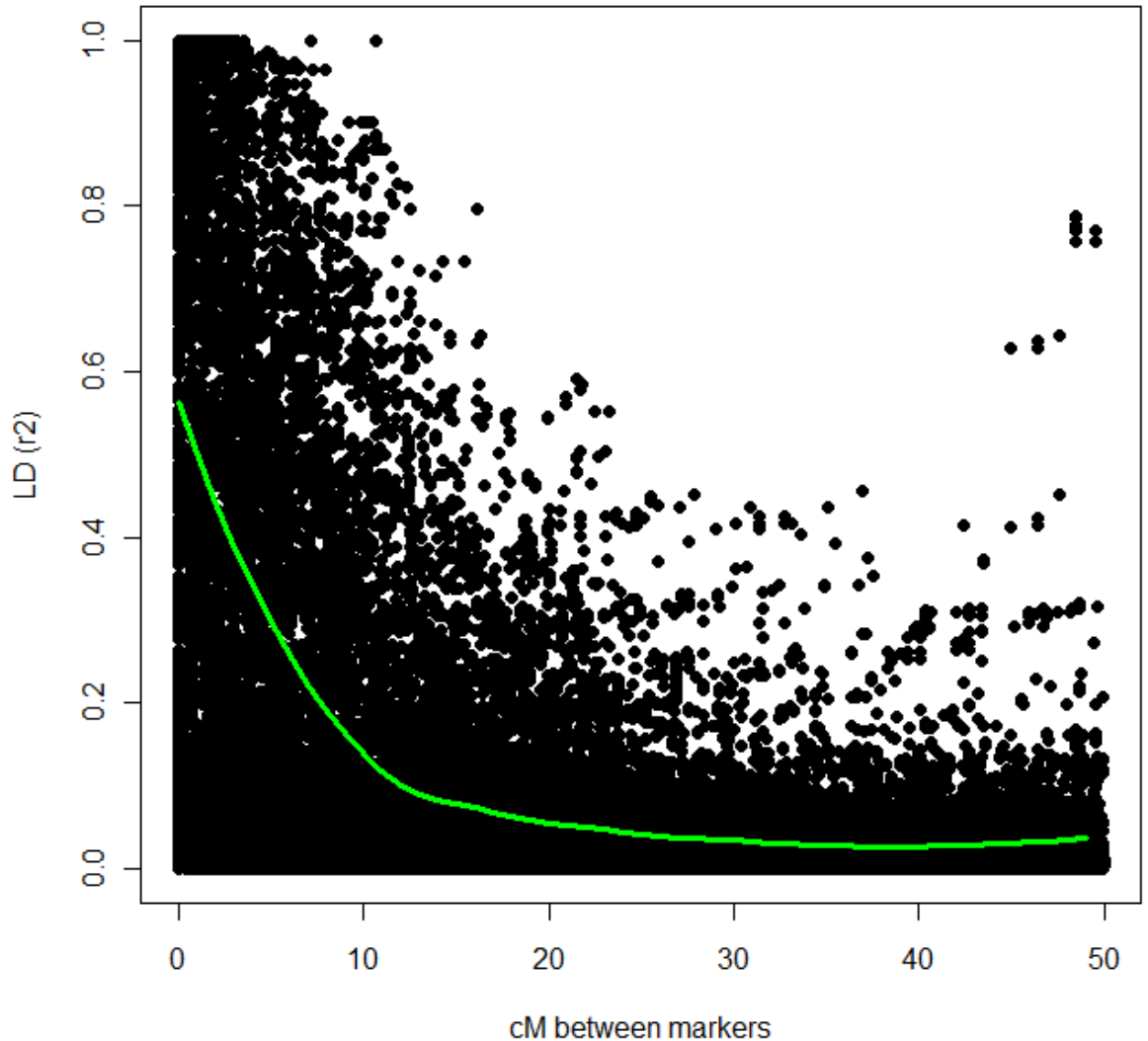
r^2 above diagonal, D' below diagonal



Note that LD drops to nearly 0 within 500 base pairs

Gaut B. S., Long A. D. Plant Cell 2010:15:1502-1506

Extensive LD in Barley of the Upper Midwest



Marker-based or marker-assisted selection

Marker score: The sum of effects associated with the marker alleles carried by the individual or family

Marker-assisted selection: Selection on the basis of the marker score, or weighted combination of the marker score and phenotype together into an index.

Note: Sometimes “marker-based selection” is used to describe the selection on marker score alone, while “marker-assisted selection” is used to describe the selection on marker score + phenotype.

Marker score

Marker score (*w*): The sum of effects associated with the marker alleles carried by the individual or family

$$w = \sum_{i=1}^{n} b_i X_i$$

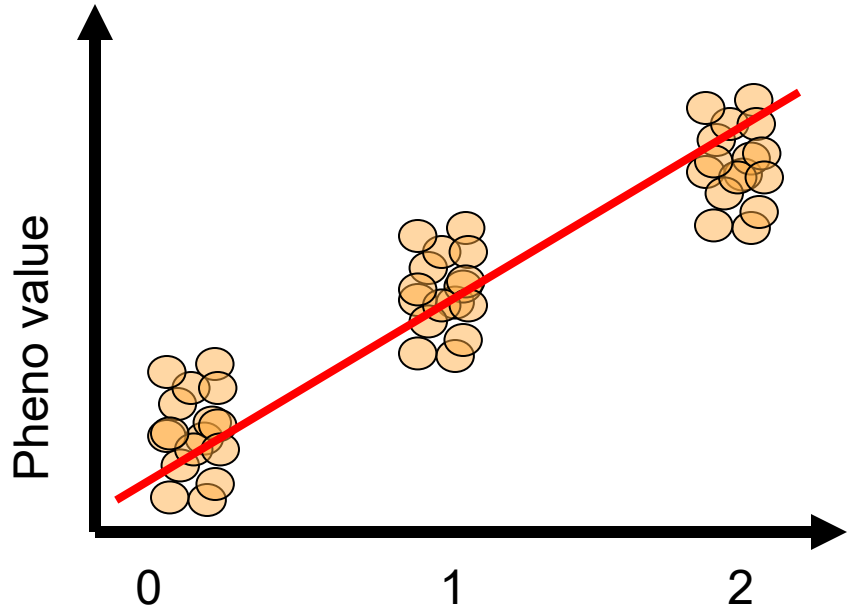
- b_i is half the difference between the fitted means of the M_iM_i and m_im_i marker genotypes at the i^{th} marker locus
- X_i is an indicator variable indicating the genotype
- n is the number of marker loci.

Coded genotypes

Geno	*X*
MM	2
Mm	1
mm	0

Toy example

- 500 random individuals from a population phenotyped and genotyped
  - Genotypes were scored for one marker linked to a candidate gene
  - Individuals scored as mm = 0, Mm = 1, MM = 2.



$$y = \mu + bx + \varepsilon$$
$$H_0: b = 0$$
$$H_A: b \neq 0$$

Example. Consider three loci

	Mean values		
Geno	Locus 1	Locus 2	Locus 3
M_iM_i	0.40	0.60	0.80
m_im_i	0.20	0.30	0.20
b_i	0.10	0.15	0.30

Two different individuals with different genotypes

$M_1M_1\ M_2m_2\ m_3m_3$: $w = (0.10)(2) + (0.15)(1) + (0.30)(0) = 0.35$

$m_1m_1\ M_2M_2\ M_3M_3$: $w = (0.10)(0) + (0.15)(2) + (0.30)(2) = 0.90$

Estimation of marker effects

- There are two goals of linear models: estimation of model parameters and estimation of any appropriate variances.

$$y = \mu + \beta x + e$$

In this simple linear regression model, we consider μ and β to be **fixed** constants we are trying to estimate (**fixed effects**).

However, we consider *e* to be drawn from some probability distribution and thus consider these effects to be **random**.

Source: Bruce Walsh

Fixed vs. Random Effects

- If we consider a parameter to be **“fixed”**, we are trying to estimate its fixed effect that is unchanging and free of any distributional assumptions.

- If a parameter is considered to be **“random”**, we are trying to make inferences about its underlying probability distribution. If we assume the probability distribution to be a Normal distribution, we try to estimate the variance of that distribution.

- By convention, we **estimate fixed effects** and **predict random effects**:

    - BLUE: Best linear unbiased estimates of fixed effects

    - BLUP: Best linear unbiased predictions of random effects

Source: Bruce Walsh

Fixed vs. Random Effects (cont.)

- Assume we have *k* fixed effects. If we treat these as fixed, we lose *k* degrees of freedom.

- However, if we assume these *k* effects are drawn from an underlying probability distribution with mean 0 and an unknown variance, **only one degree of freedom is lost** for estimating the variance of the distribution.

    - We then predict the values of the *k* realizations.

Source: Bruce Walsh

Least-squares estimation

$$y = X\beta + e$$

Assume $e \sim (0, I_n\sigma^2)$ where e is a vector of residuals. This says that the residuals are independent and identically distributed (iid).

Least squares aims to estimate β to minimize the residual sums of squares (RSS)

$$RSS = \sum_{i=1}^{n} (\hat{e}_i)^2 = \hat{e}^T \hat{e} = (y - X\hat{\beta})^T (y - X\hat{\beta})$$

Taking the derivative and setting to zero it can be shown that RSS is minimized by

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

Estimating marker effects for MAS and MARS

- Usually, *p* (the number of markers) greatly exceeds *n* (the number of individuals in a population).

- Thus, some subset selection is needed to reduce *p*.
    - Because the focus is on prediction, precise localization of QTL is not necessary and thus QTL mapping techniques are not necessarily used.

- Rather, some stepwise model selection procedure is performed, such as:

Example of a stepwise model selection procedure for MAS

1. Perform multiple regression of phenotypic value on marker state for all marker loci on single chromosome.
2. Eliminate non-significant markers through backward elimination
    - Relaxed significance level, $P = 0.20$ to 0.30
3. Repeat for each chromosome
4. Combine all remaining significant markers on each chromosome and build multiple regression model.
5. Eliminate non-significant markers.
6. Estimate effects for markers in final marker model using multiple linear regression.

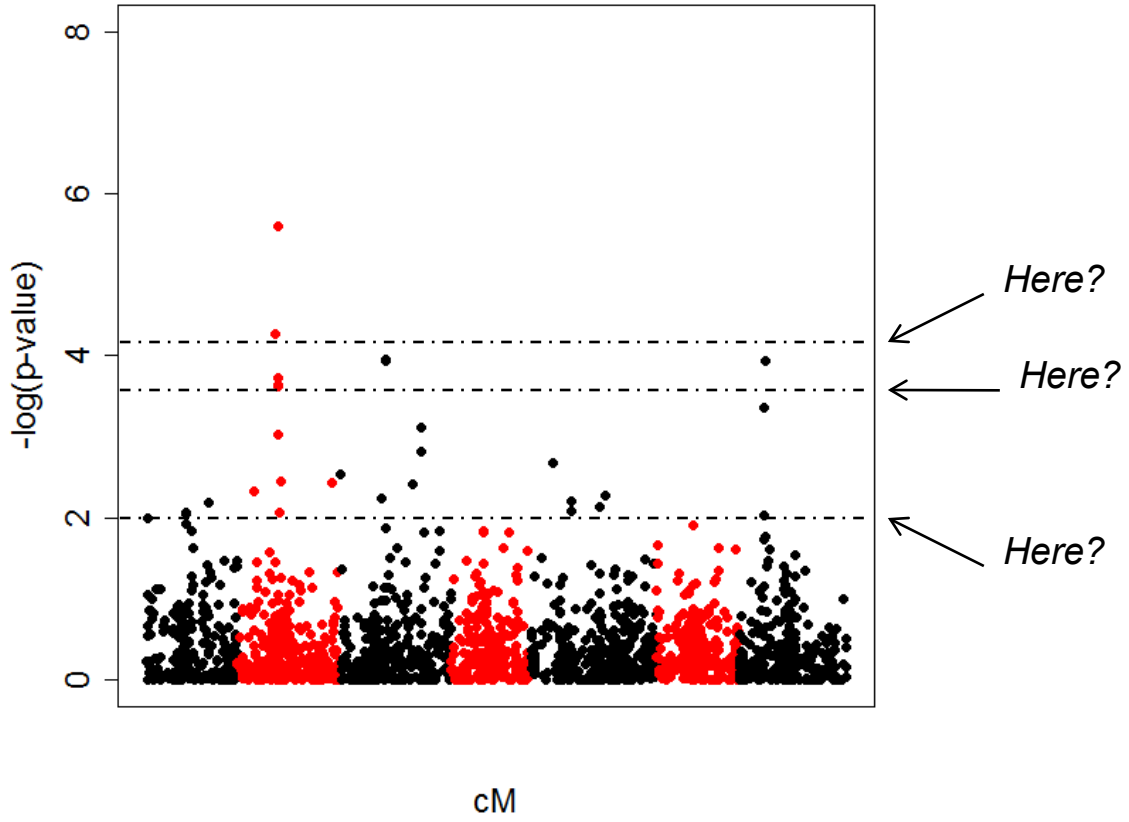
Problems associated with fixed effect estimation and model selection for marker-based selection

- The QTL number, and hence selected marker number, is often grossly underestimated for quantitative traits.

- Marker effects are often overestimated, especially in low-powered situations (i.e., “Beavis effect”).

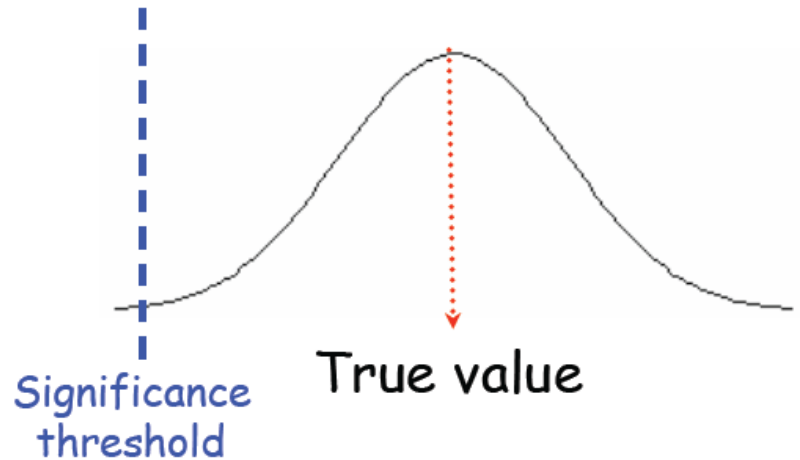
- Arbitrary significance testing → markers just meeting level of significance included in prediction model, while those just missing are left entirely out of prediction model.

Choice of markers in traditional marker-based selection is essentially arbitrary



Beavis Effect

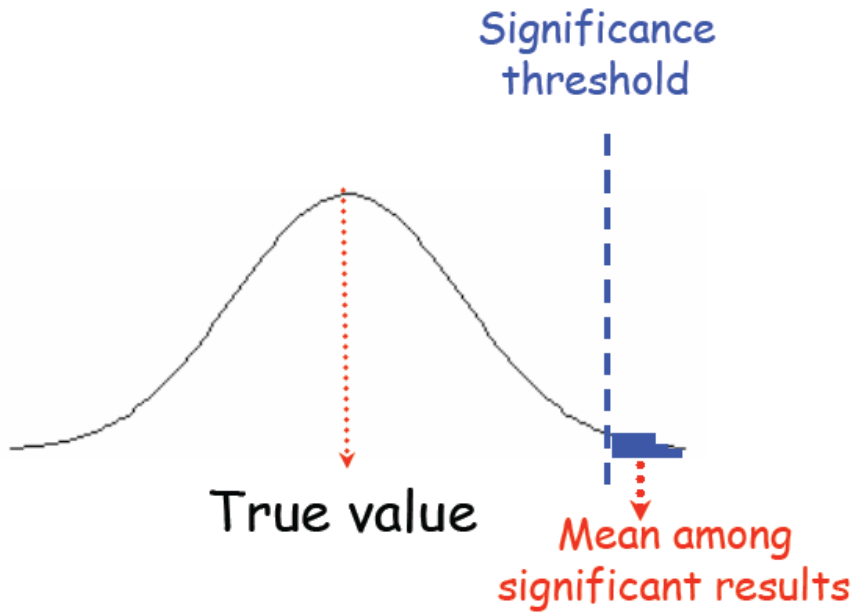
Also called the "winner's curse" in the GWAS literature



Distribution of the realized value of an effect in a sample

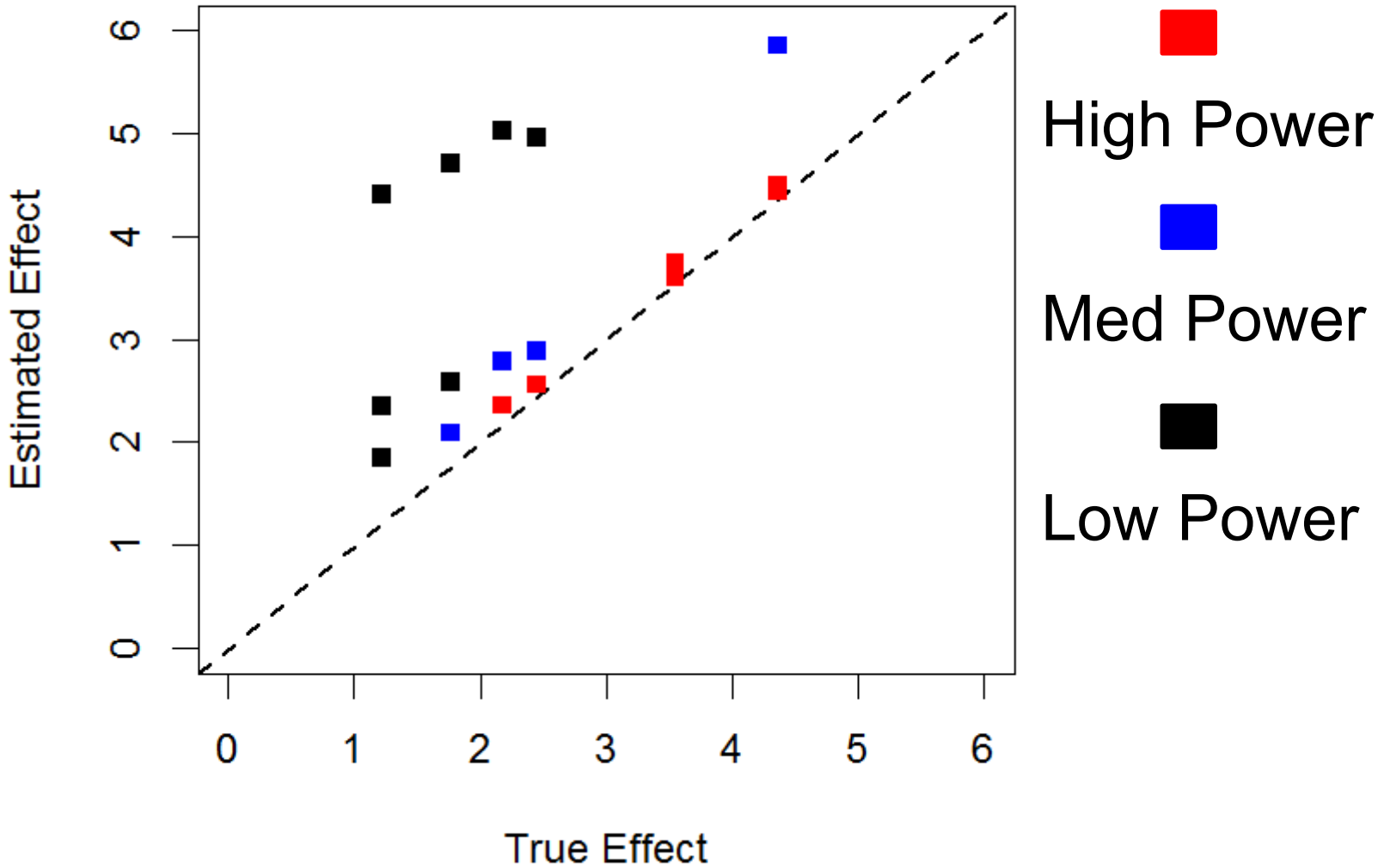
High power setting: Most realizations are to the right of the significance threshold, and the average value of these approaches the true value

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



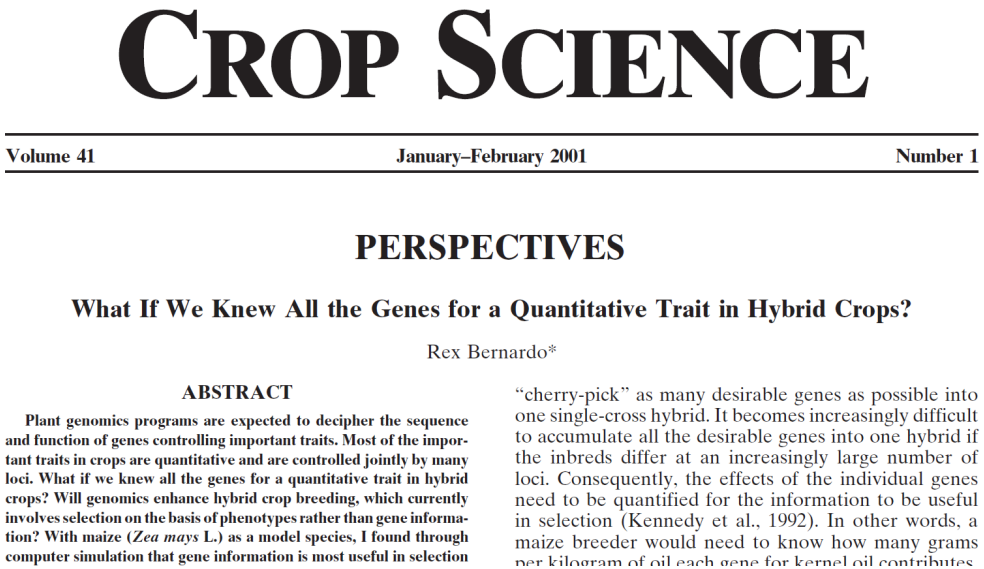
However, the mean of those declared significant is much larger than the true mean

Beavis (1998)


Estimated Effect
0
1
2
3
4
5
6
True Effect
0
1
2
3
4
5
6
High Power
Med Power
Low Power

Bernardo (2001) – All the Genes

- What if we knew the exact location of all causal polymorphisms underlying genetic variation for a quantitative trait?
- Does inclusion of this marker information in a selection index enhance response relative to phenotypic selection?
- Simulation
  - Varied genetic architecture and % of loci known
  - Used multiple linear regression to estimate QTL effects as fixed



CROP SCIENCE

Volume 41 January–February 2001 Number 1

PERSPECTIVES

What If We Knew All the Genes for a Quantitative Trait in Hybrid Crops?

Rex Bernardo*

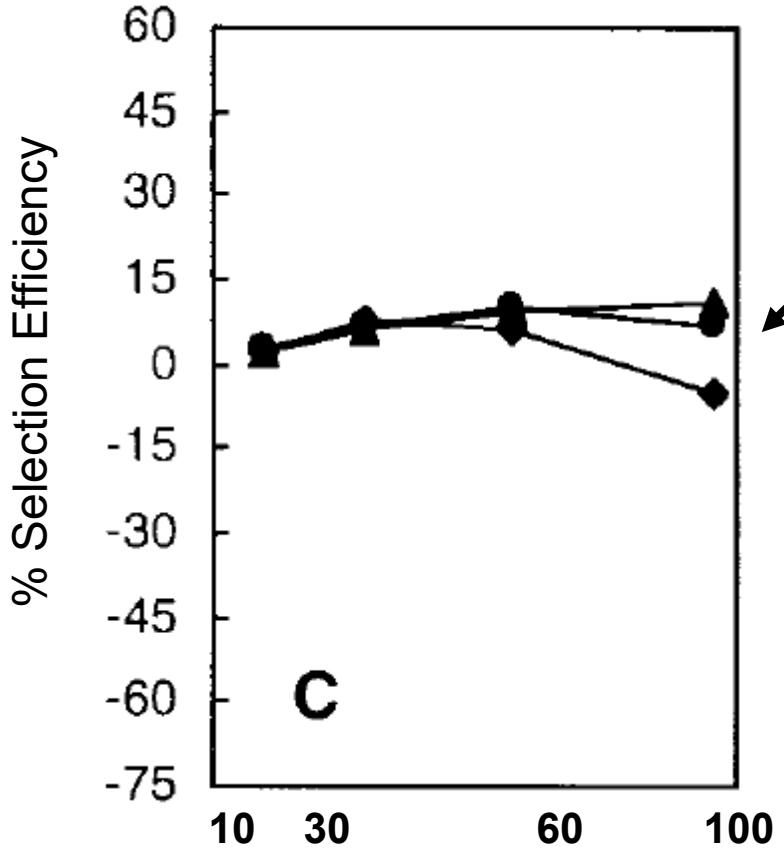
ABSTRACT

Plant genomics programs are expected to decipher the sequence and function of genes controlling important traits. Most of the important traits in crops are quantitative and are controlled jointly by many loci. What if we knew all the genes for a quantitative trait in hybrid crops? Will genomics enhance hybrid crop breeding, which currently involves selection on the basis of phenotypes rather than gene information? With maize (*Zea mays* L.) as a model species, I found through computer simulation that gene information is most useful in selection

"cherry-pick" as many desirable genes as possible into one single-cross hybrid. It becomes increasingly difficult to accumulate all the desirable genes into one hybrid if the inbreds differ at an increasingly large number of loci. Consequently, the effects of the individual genes need to be quantified for the information to be useful in selection (Kennedy et al., 1992). In other words, a maize breeder would need to know how many grams per kilogram of oil each gene for kernel oil contributes

◆ $h^2 = 0.20$, ● $h^2 = 0.50$, ▲ $h^2 = 0.80$

#QTL = 50, n = 500



Why doesn't selection efficiency get better as we model more genes?

It's an estimation problem!

% Loci known

Bernardo (2001)

Genet. Res., Camb. (2000), **75**, *pp.* 249–252. Printed in the United Kingdom

Marker-assisted selection using ridge regression

JOHN C. WHITTAKER[1]*, ROBIN THOMPSON[2] AND MIKE C. DENHAM[1]

[1] *Department of Applied Statistics, University of Reading, PO Box 240, Earley Gate, Reading RG6 2FN, UK*

[2] *IACR Rothamsted, Harpenden, Herts AL5 2JQ, UK, and Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK*

(*Received 2 March 1999 and in revised form 29 July 1999*)

Whittaker et al. (2000)

- When doing MAS, cannot include all the markers, so must select subset of markers to fit.
- No entirely satisfactory way of doing this exists.
- Objective was to evaluate ridge regression.
  - Superior to subset selection when objective is to make predictions.

Whittaker et al. (2000)

- Find subset of markers Q.
- Interested in $\hat{a}_i = \sum_{k \in Q} \hat{\beta}_i x_i$

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

- Cannot include all markers in Q
  - Increases variance of β
  - If number of markers really large, not enough d.f.

Whittaker et al. (2000)

- Ridge regression – include all variables, but replace normal least-squares estimators with

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

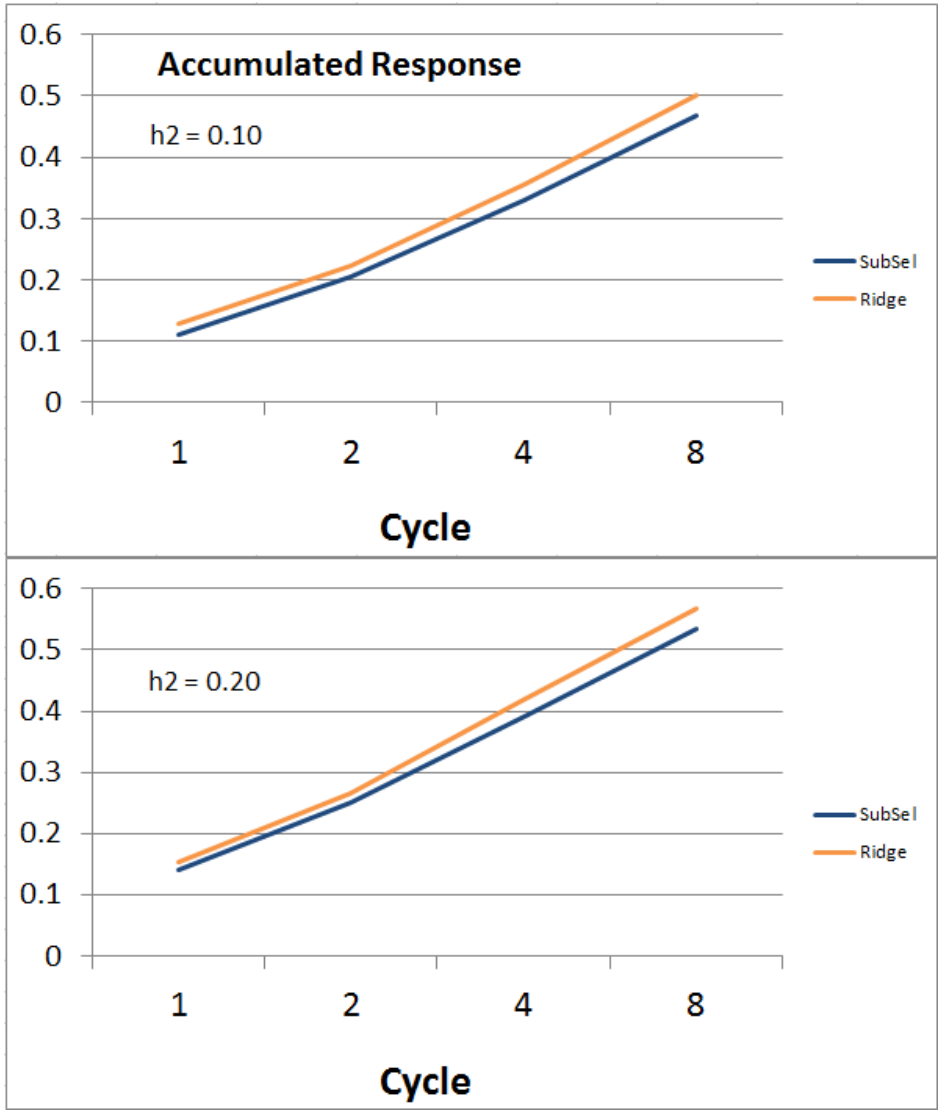
- Normal estimates shrunk toward 0
  - Degree of shrinkage determined by lambda

- Evaluated range of lambda and chose lambda to minimize model error (ME)

$$\hat{ME} = \mathrm{RSS} - \mathrm{n}\sigma_e^2 + 2\ \sigma_e^2 \mathrm{tr}[\mathbf{X}^\mathbf{T}\mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}]$$

- Addition of $\lambda\mathbf{I}$ term reduces collinearity and prevents the matrix $\mathbf{X}^\mathbf{T}\mathbf{X}$ from becoming singular.

Whittaker et al. (2000)



Genomic prediction and genomic selection

- Genomic prediction: The prediction of an individual's genetic value using genome-wide markers.
- Genomic selection: Selection of individuals based on their genomic prediction value(s).

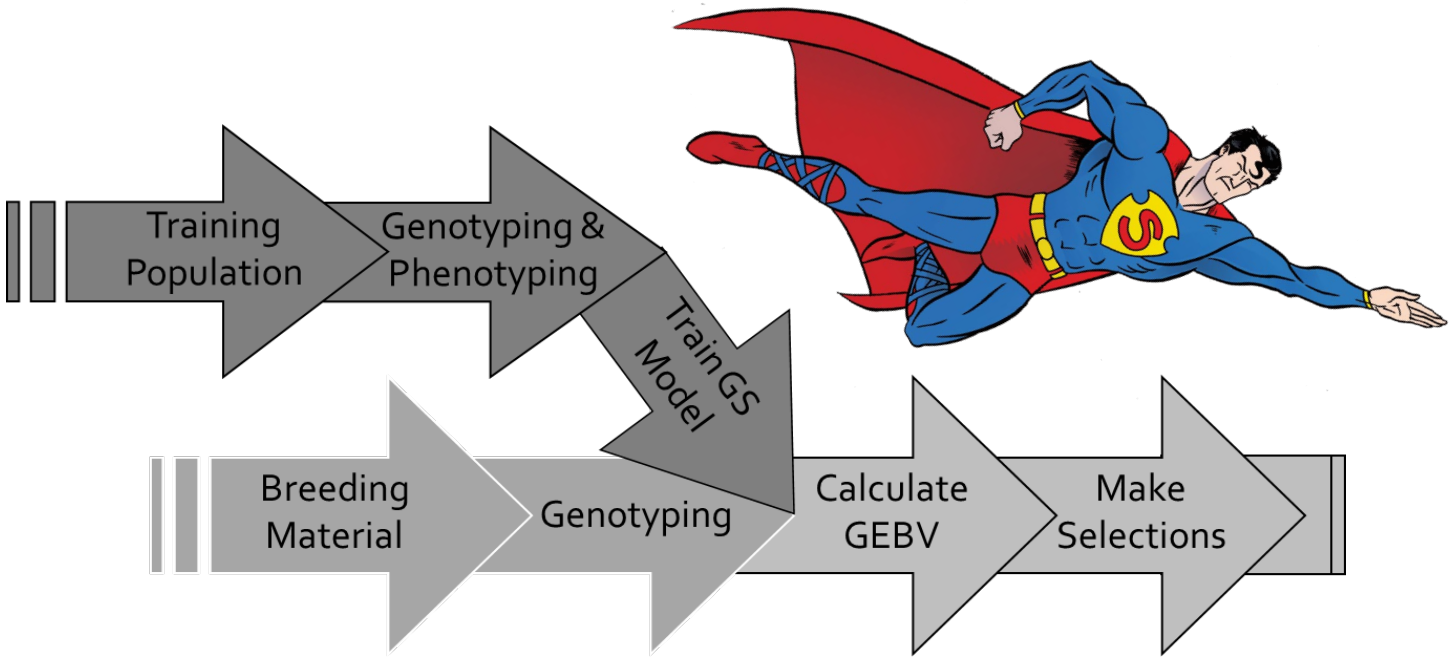
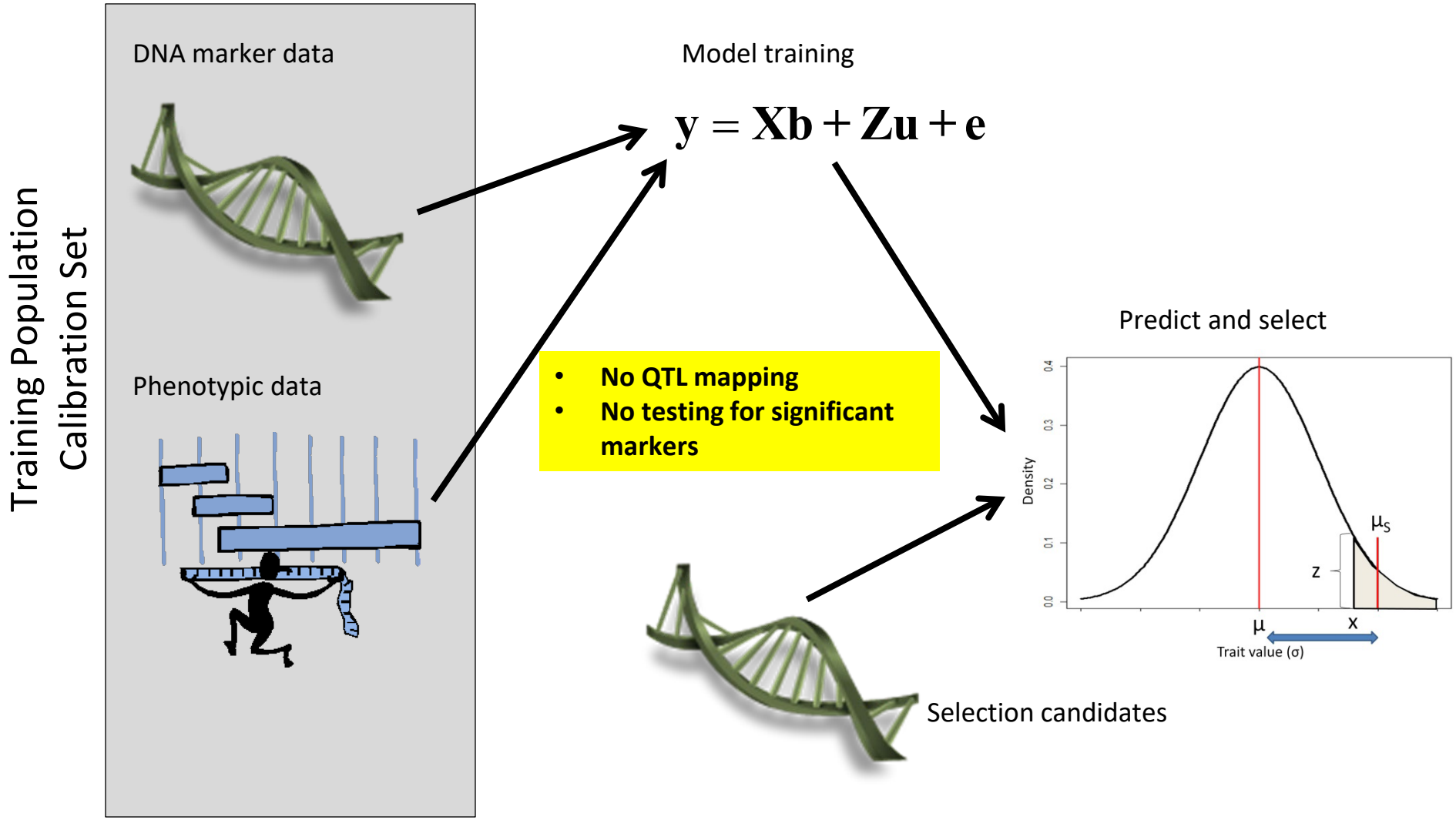


Figure from Heffner et al. (2009)

Genomic Prediction and Selection



Genomic selection: The essentials

- Genomic selection = Genome-wide selection
  - A form of marker-based selection
    - Avoids QTL mapping
    - Includes all markers in model
- Made possible by:
  - High-throughput marker technologies
  - *Relatively* new and adopted statistical models for high-dimensional data
  - High-performance computing
- Advantages over traditional marker-based selection
  - No arbitrary statistical threshold
  - Able to capture small allelic effects

Commonly used terminology

- **Training population**: A population of individuals that have been both phenotyped and genotyped and which are used to develop a statistical model relating markers to phenotypes.
    - Also referred to as calibration set, training set

- **Model training**: The process of using a training population to develop a prediction model.

- **Target population**: The population of individuals on which predictions are to be made.
    - Also referred to validation population (set), test population (set), especially in the context of cross-validation.

Commonly used terminology (cont)

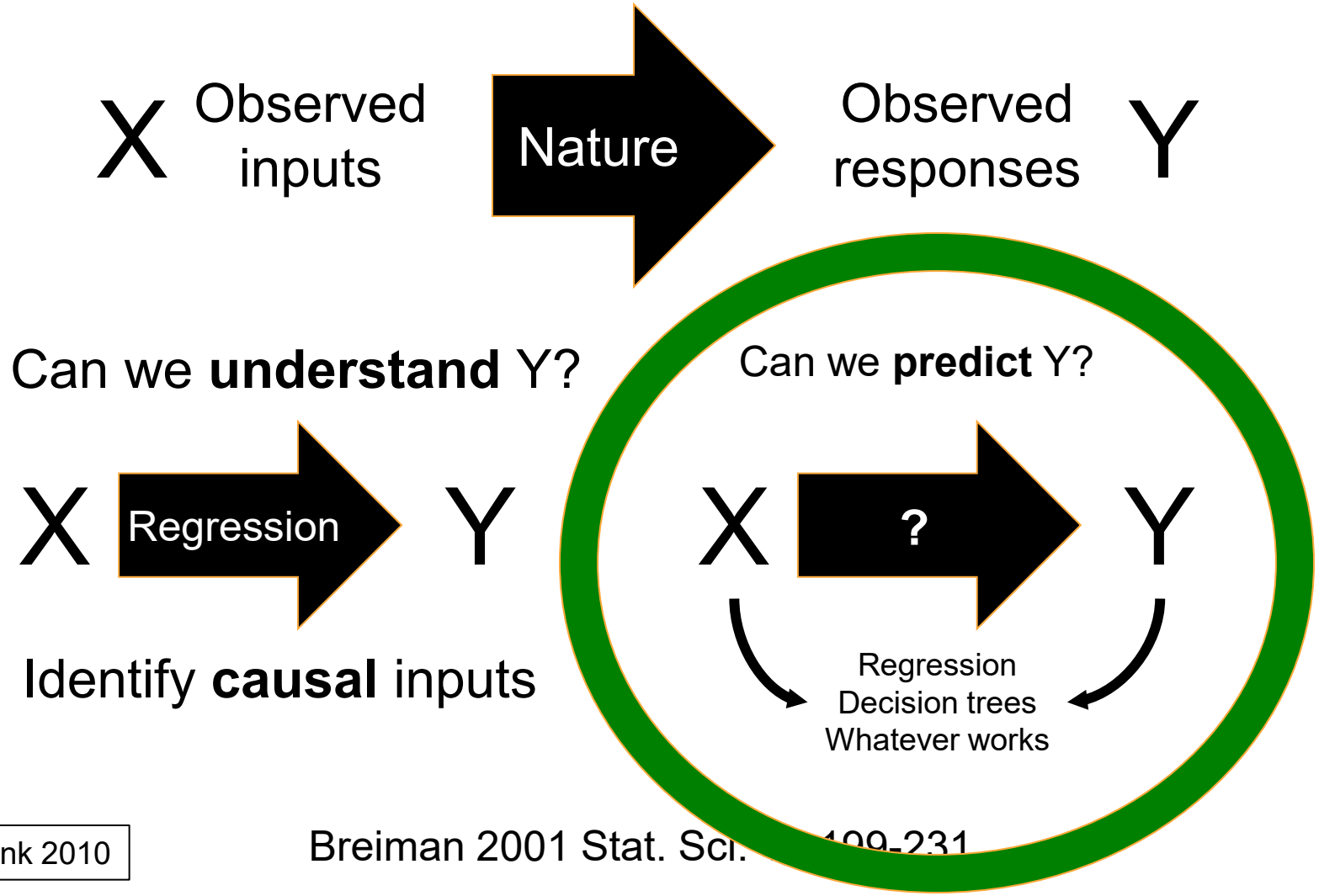
- **Prediction accuracy**: Measured as the correlation between the estimated breeding value and true breeding value. In genomic selection, we use the genomic estimated breeding value (GEBV).

- **Predictive ability:** Measured as the correlation between the phenotype and estimated genotypic value. Heritability of the phenotypes used for validation will set an upper bound to “predictive ability”.

Genomic selection

- Has been described as a “black box” approach to marker-based selection.
- It is a “black box” approach to the extent that phenotypic selection is a “black box” approach

Statistical modeling: The two cultures



JL Jannink 2010

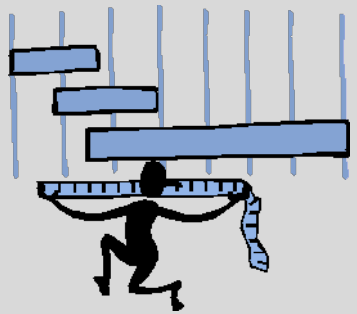
Breiman 2001 Stat. Sci. 199-231

Genomic Prediction and Selection

Training Population
Calibration Set

DNA marker data

Phenotypic data



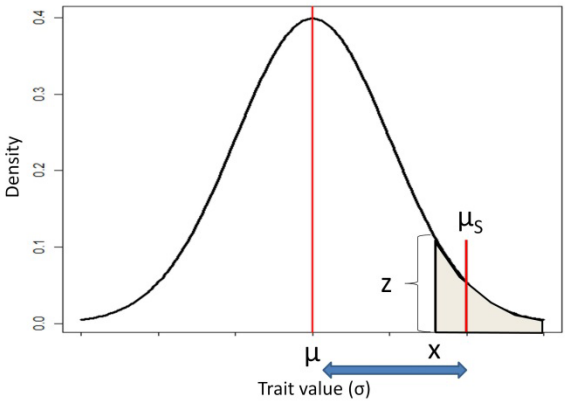
Model training

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

Why?

No QTL mapping!!
Use all markers!

Predict and select



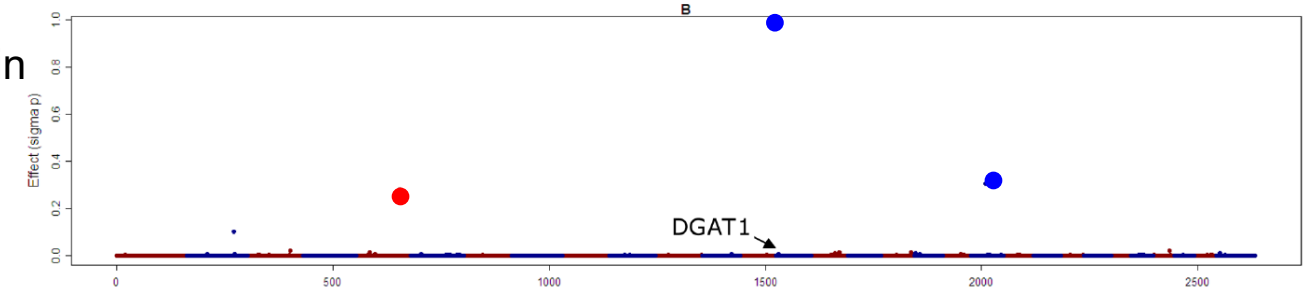
Selection candidates

Complex traits are controlled by many small-effect alleles

Hayes et al. (2010)

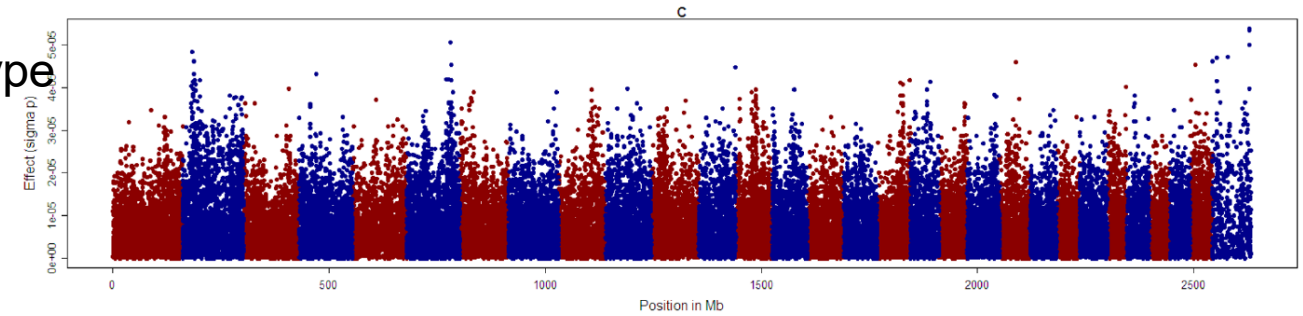
Simpler

Holstein
Fat %

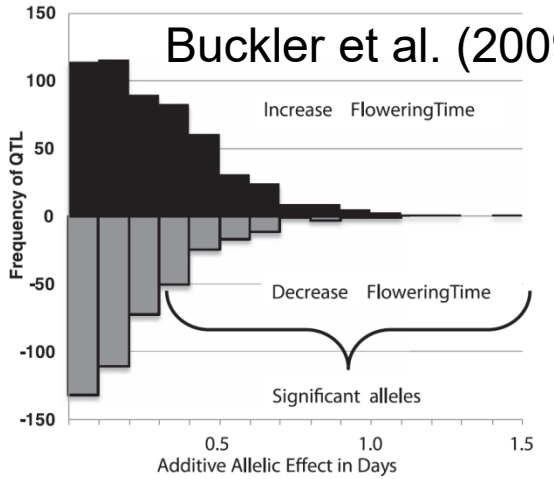


Complex

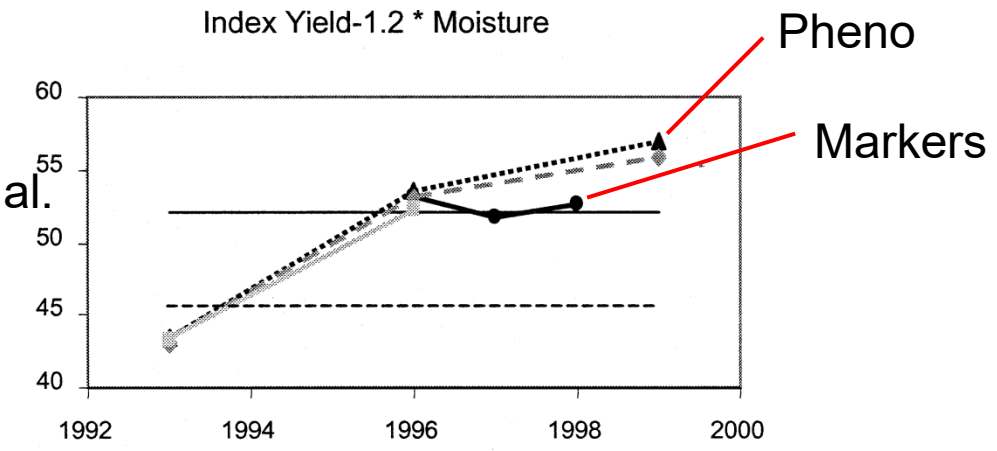
Holstein
Overall type



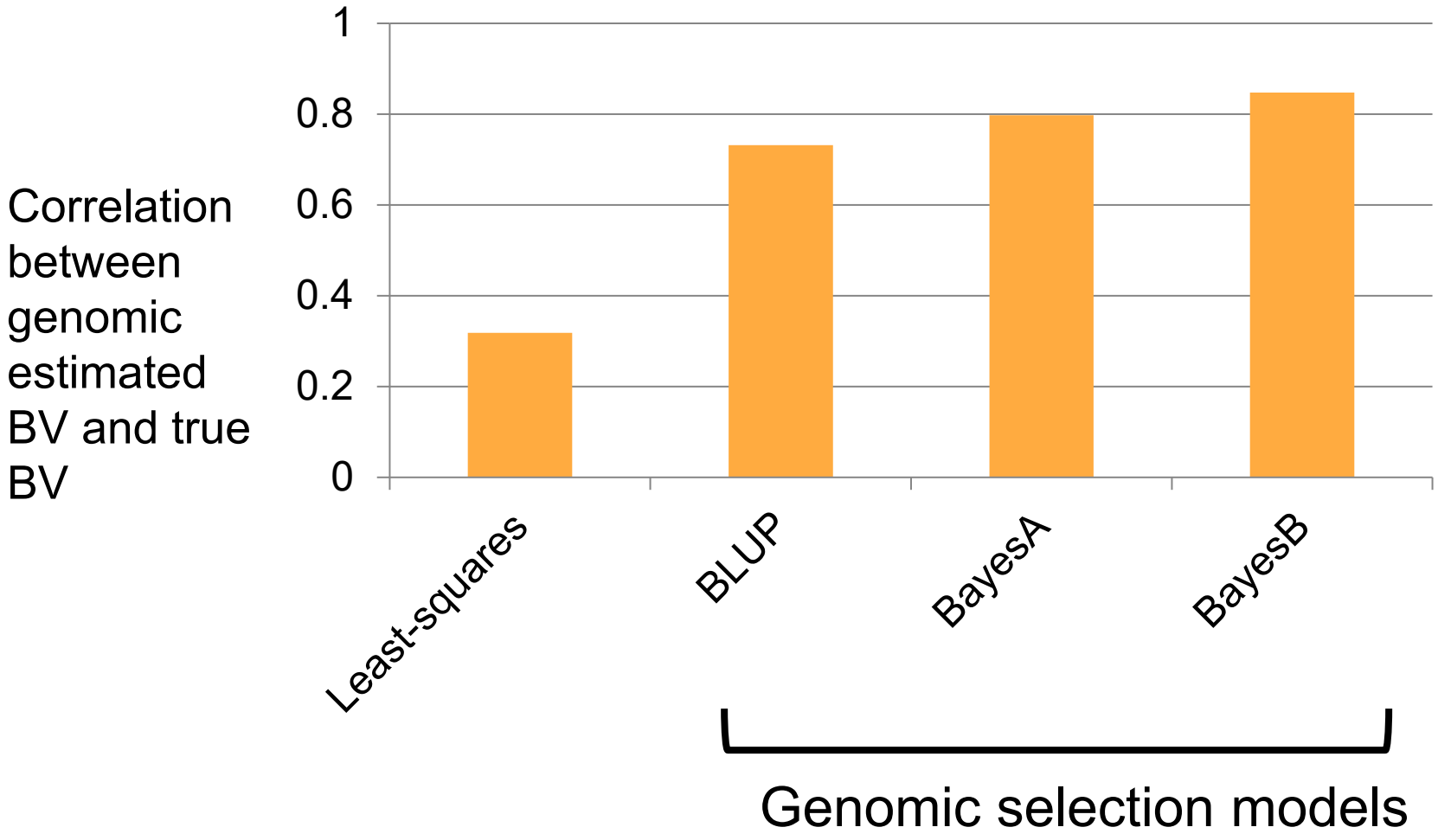
Buckler et al. (2009)



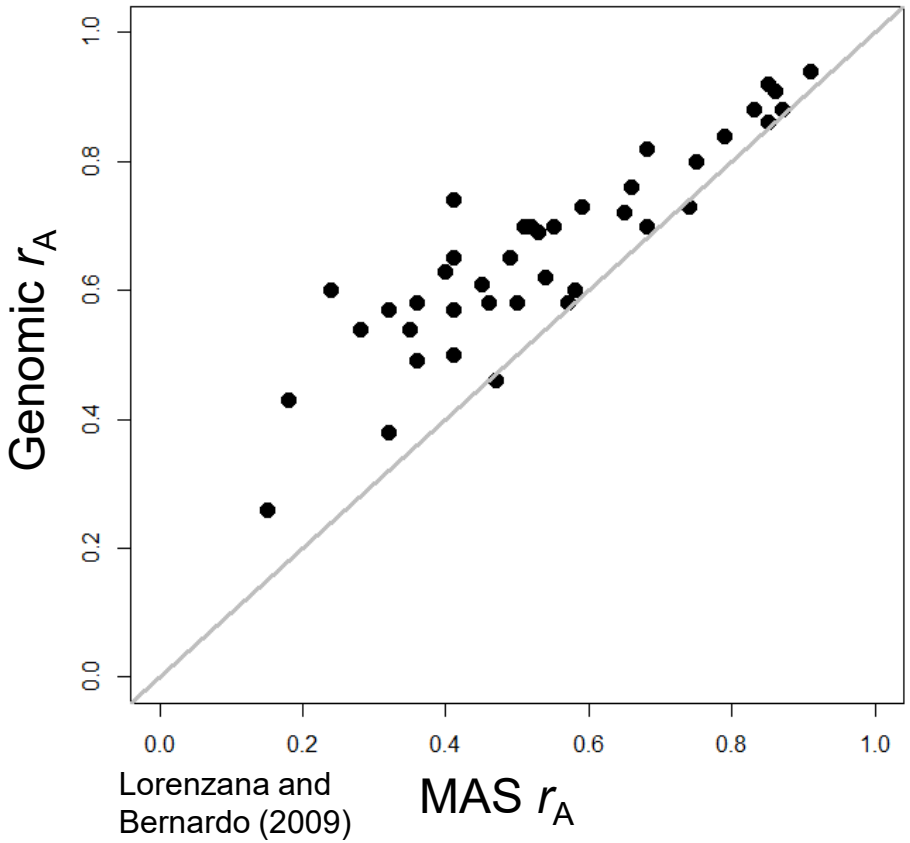
Moreau et al. (2004), Maize



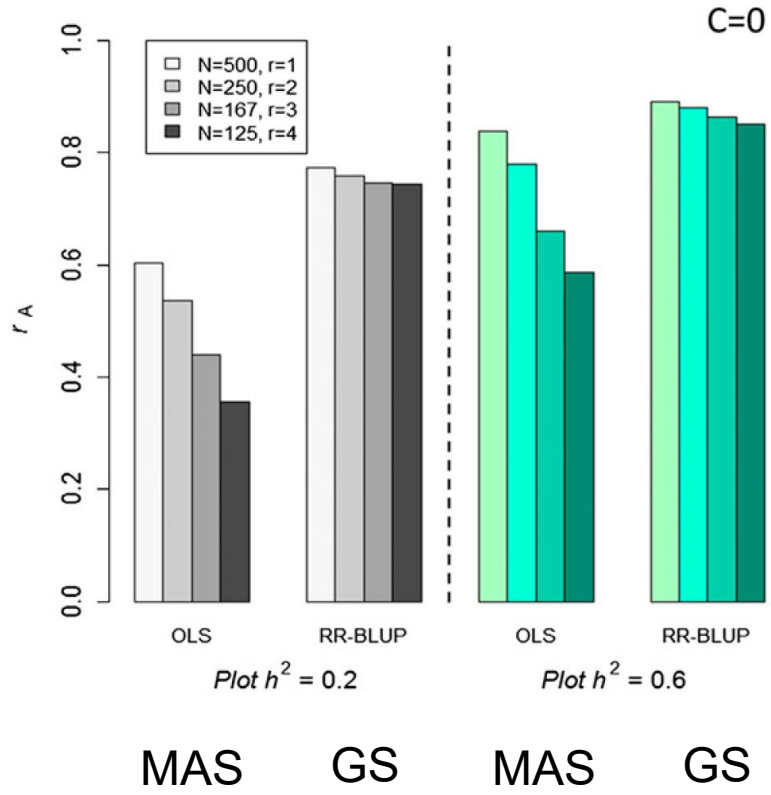
MHG 2001


Correlation between genomic estimated BV and true BV
1
0.8
0.6
0.4
0.2
0
Least-squares
BLUP
BayesA
BayesB
Genomic selection models

A genome-wide approach typically provides better predictions



Lorenzana and Bernardo (2009)



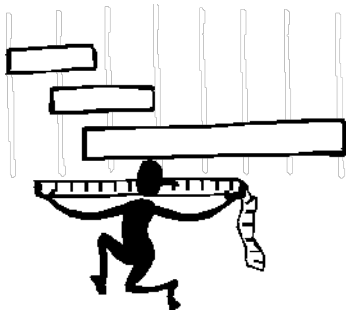
Lorenz (2013)

Genomic Prediction

Training Population
Calibration Set

DNA marker data

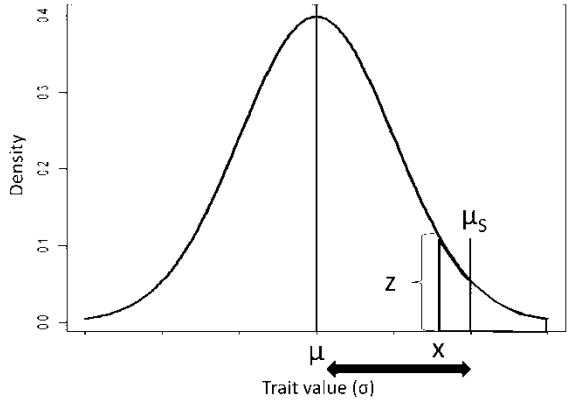
Phenotypic data



Model training

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

Predict and select



Selection candidates

Genomic prediction models

LARGE *p* !!

Training population

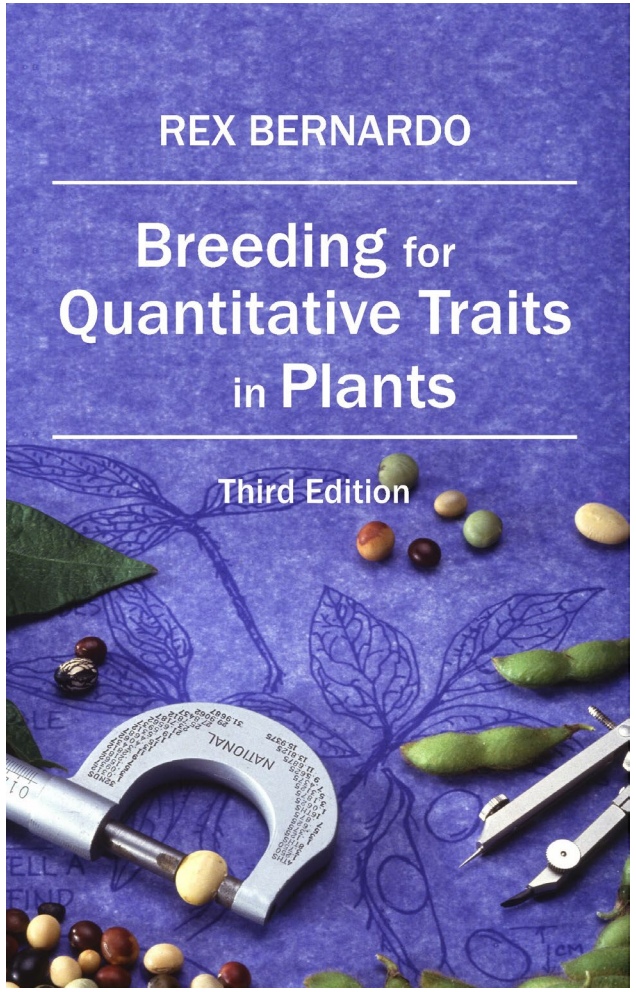
Line	Yield	Mrk 1	Mrk 2	…	Mrk *p*
Line 1	76	1	1		1
Line 2	56	1	1		1
Line 3	45	1	1		1
Line 4	67	0	1		0
⋮					
Line *n*	22	1	1		1

smaller *n* !!

Applications and Optimization

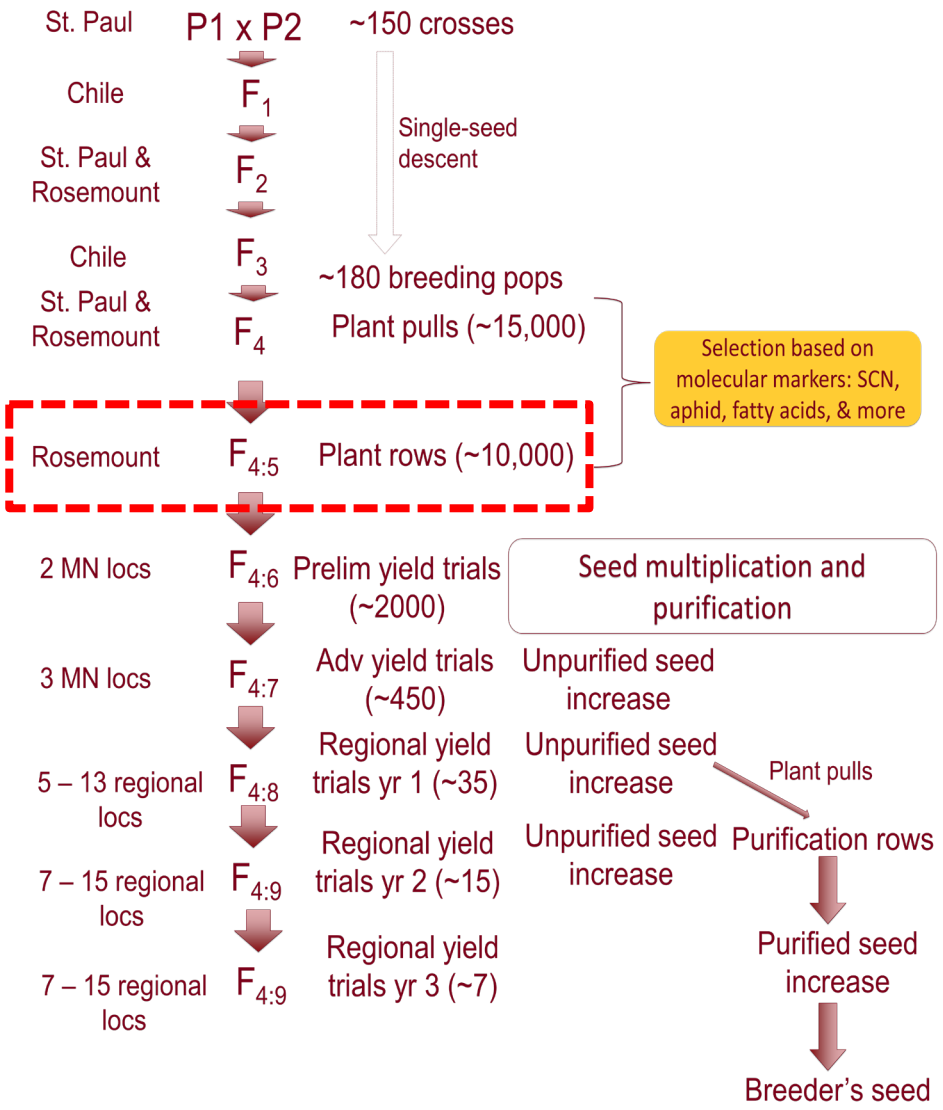
When to use genomic predictions and selection

1. When phenotypic selection is ineffective
2. To increase gain per unit time
3. For traits that are difficult to measure
4. For other target populations of environments
5. To reduce phenotyping
6. When there are too many candidates to phenotype
7. When seed amounts are insufficient



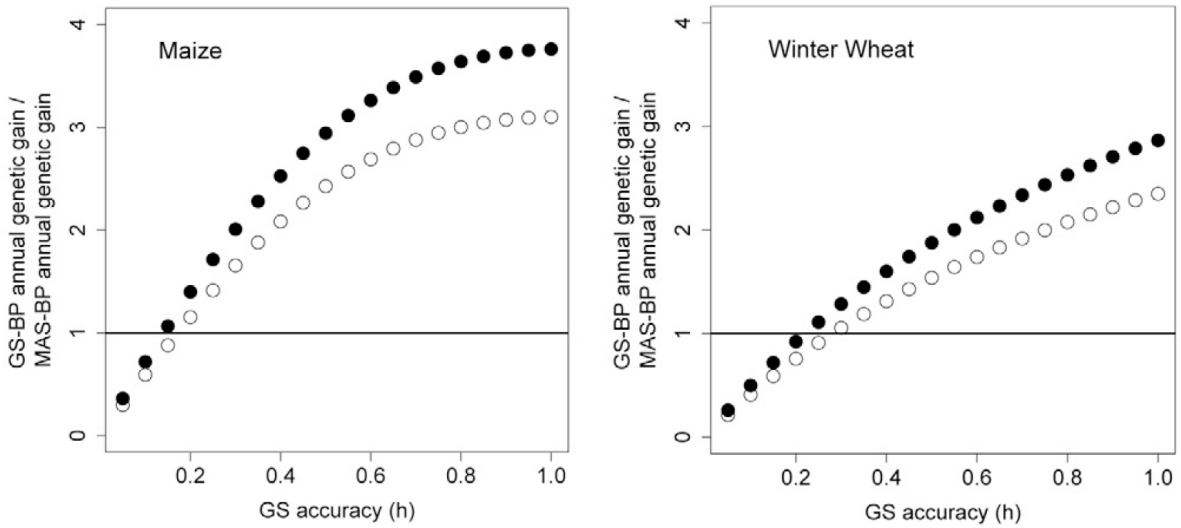
1. When phenotypic selection in ineffective

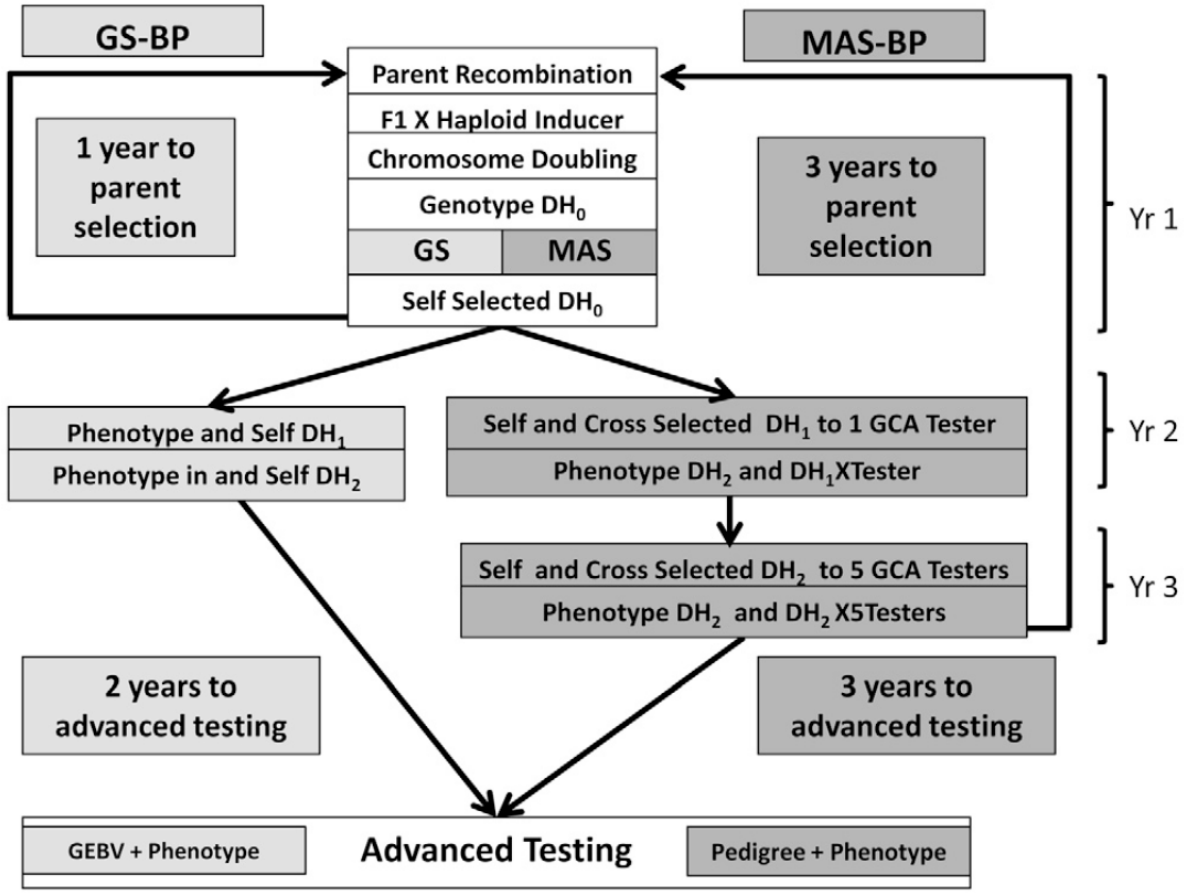
- Examples in soybean breeding:
  -- Single plants, plant rows.
  -- Heritability of yield at the plant row stage is near 0.



2. To increase genetic gain per unit time

Even modest prediction accuracies can result in more genetic gain per unit time because parent selection can happen earlier.





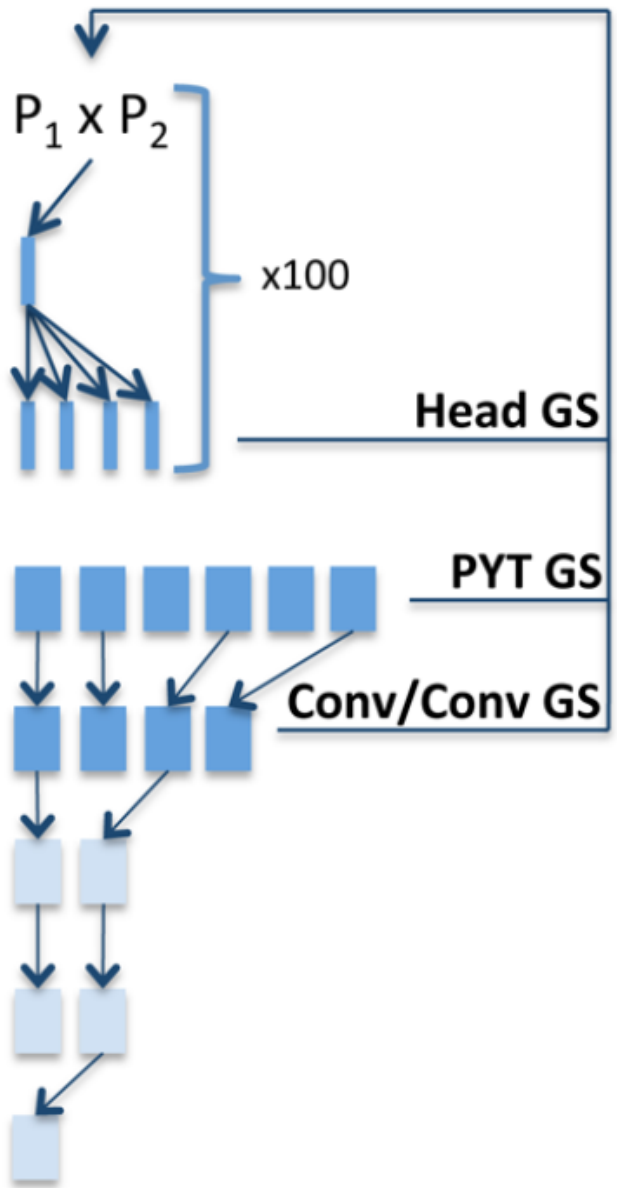
Heffner et al. (2010)

RESEARCH

A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines

R. Chris Gaynor, Gregor Gorjanc, Alison R. Bentley, Eric S. Ober,
Phil Howell, Robert Jackson, Ian J. Mackay, John M. Hickey*

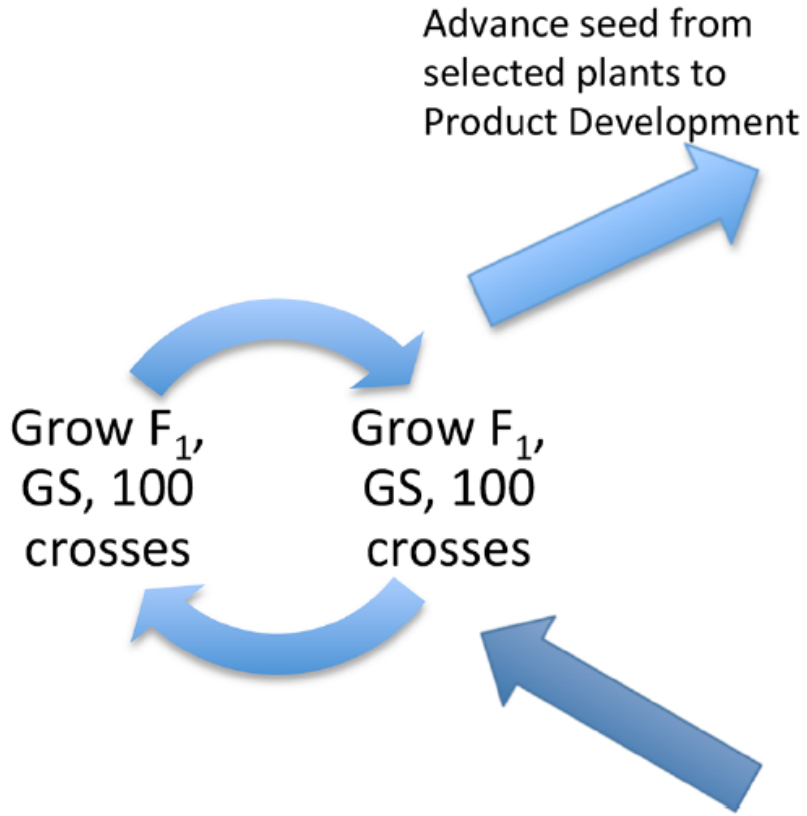
Year	Stage	Number of Plants	Action
1	Crossing	100 crosses	Make bi-parental crosses
1-2	F_1/DH	100 full-sib families	Produce DH lines
3	Headrows	100 x $N^{\dagger}$ DH lines	Advance 500 lines, genotype/cross (**Head GS**)
4	PYT	500 DH lines	Yield trial, genotype/cross (**PYT GS**)
5	AYT	50 DH lines	Yield trial, cross (**Conv**), genotype/cross (**Conv GS**)
6	EYT	10 DH lines	Yield trial
7	EYT	10 DH lines	Yield trial
8	Variety	1 DH line	Release variety



Gaynor et al. (2017)

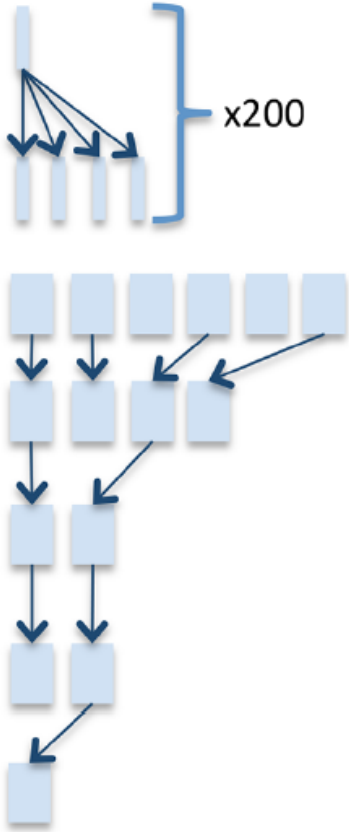
Gaynor et al. (2017)

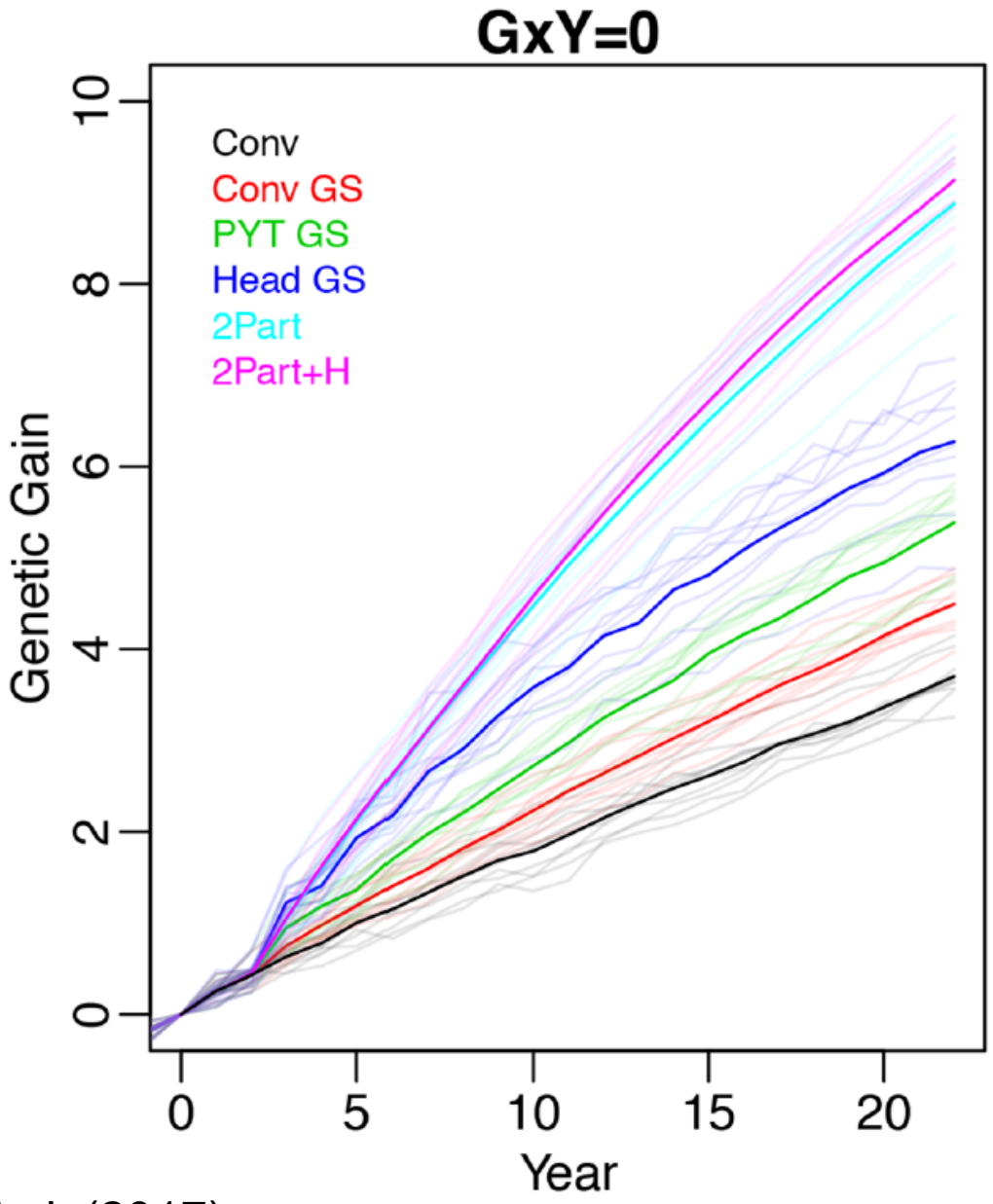
Population Improvement

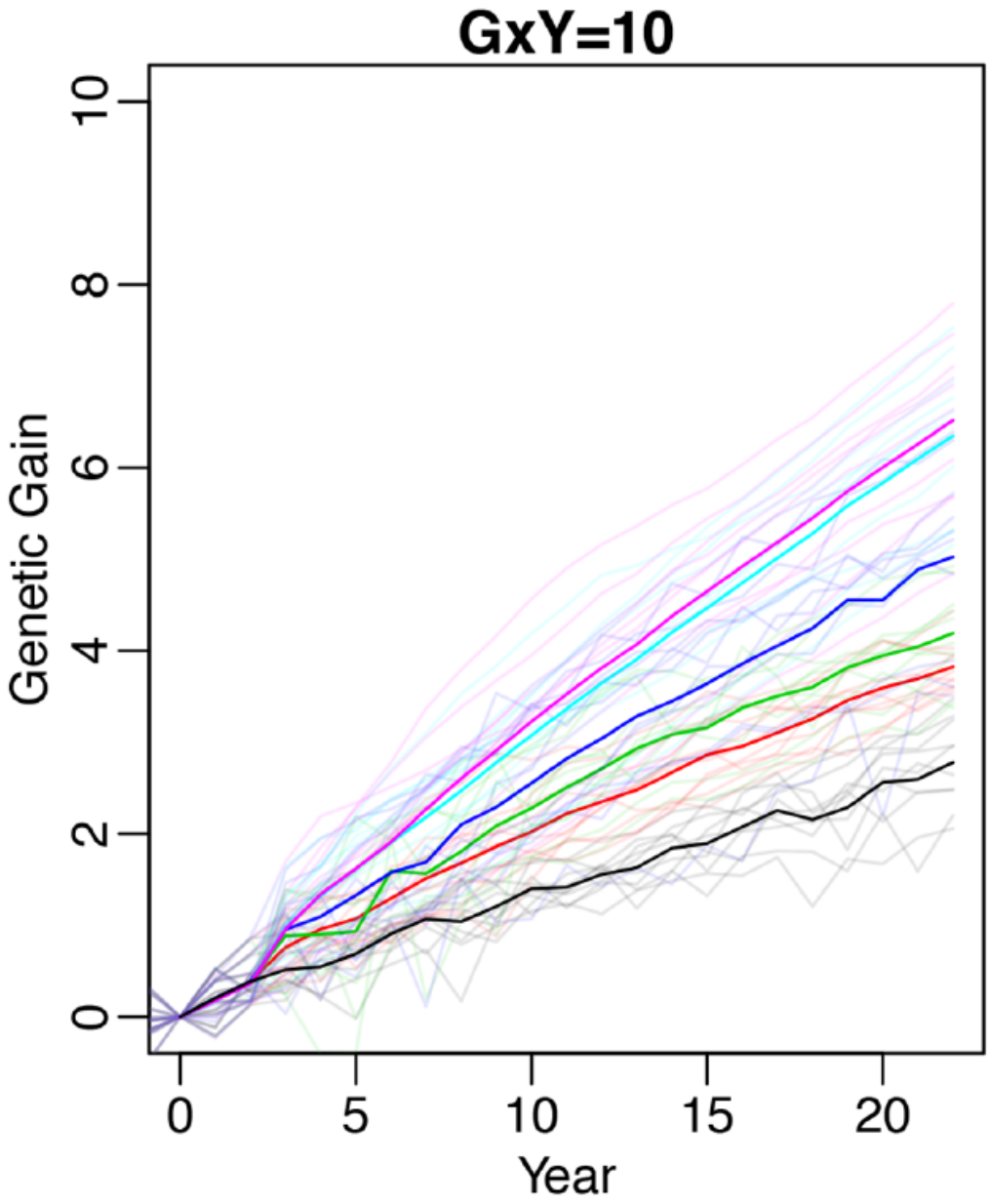


Product Development

Year	Stage	Number of Plants	Action
1-2	F_1/DH	200 half-sib families	Produce DH lines
3	Headrow	200 x N^+ DH lines	Advance 500 lines, genotype (**2Part+H**)
4	PYT	500 DH lines	Yield trial, genotype (**2Part**)
5	AYT	50 DH lines	Yield trial
6	EYT	10 DH lines	Yield trial
7	EYT	10 DH lines	Yield trial
8	Variety	1 DH line	Release variety

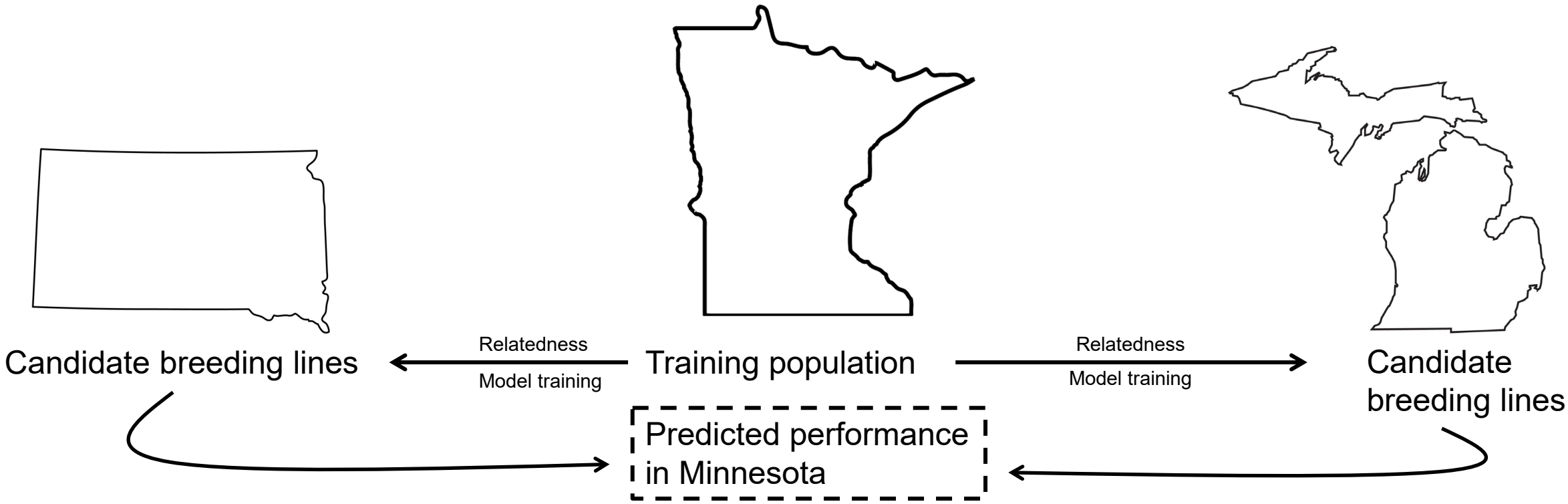




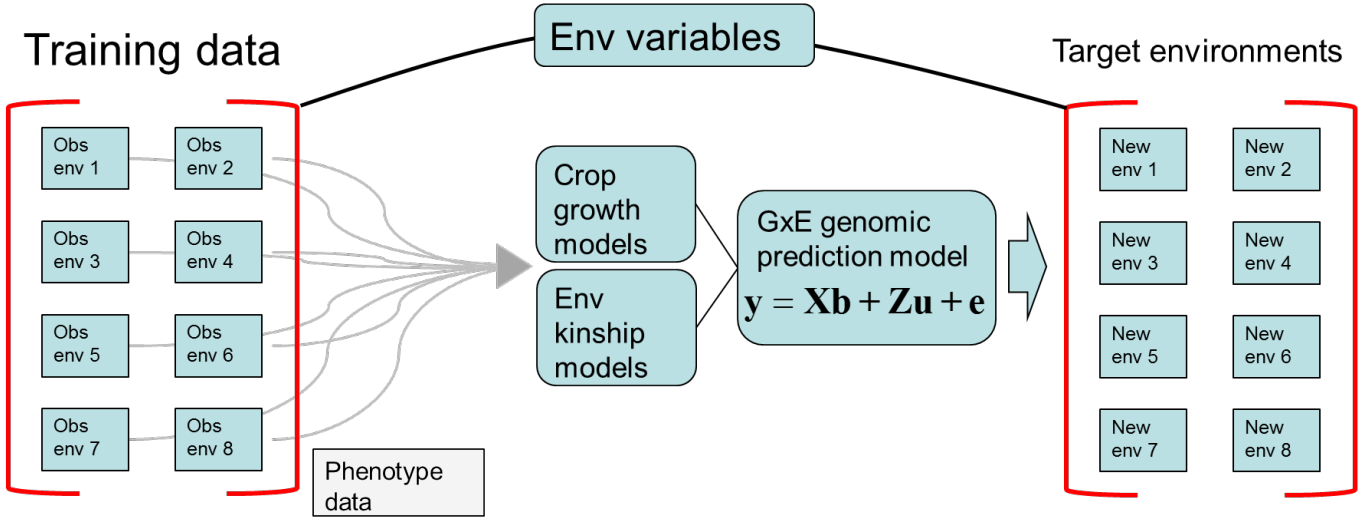


Gaynor et al. (2017)

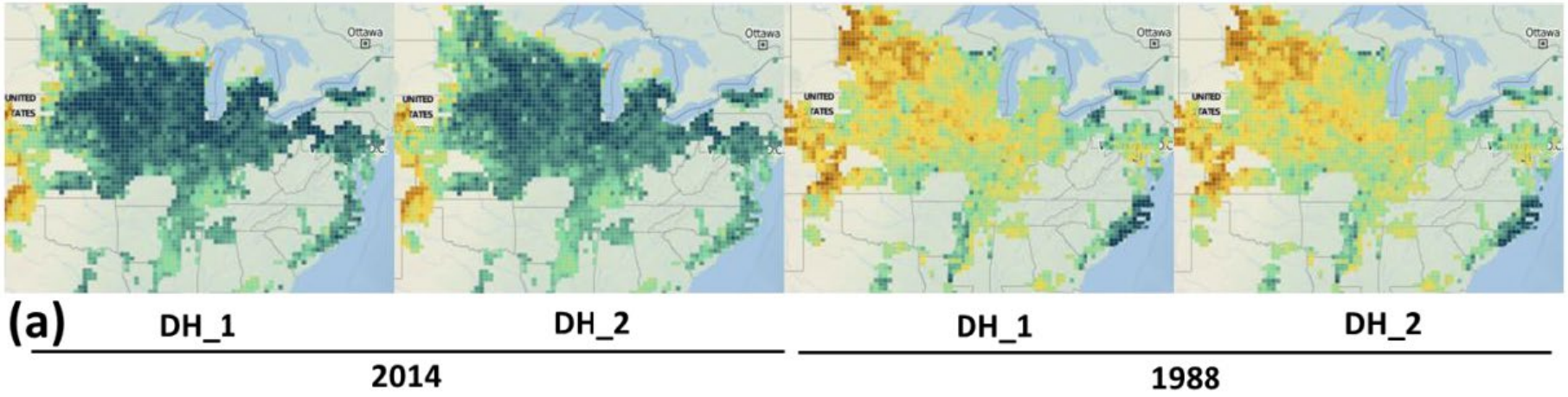
4. For other target populations of environments



4. For other target populations of environments



Knowing expected weather conditions allows characterization of genetics at all candidate growing sites!

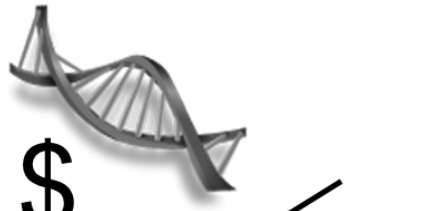


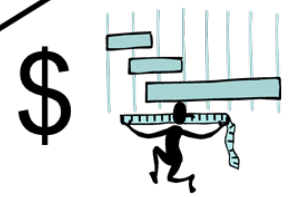
Messina et al. (2017)

5. Reduce phenotyping

Relative cost of genotyping to phenotyping dictates how genotyping/phenotyping resources should be allocated

$$Rc = \frac{\$}{\$}$$





= cost of genotyping one entry relative to cost of phenotyping one field plot unit

Reduced Phenotyping

Year	Heterotic group 1	
1	A x B	
1	F1 x haploid inducer	
2	Haploid doubling	
2	DH seed increase + genotyping	
3	***50% DH for topcrossing***	***50% in cold storage***
3	***Yield trial 50%***	***Predict 50%***
4	Cross selected DHs to 3 elite testers	
4	Yield trial testcrosses	
	Selected DHs	
5		
6		
7		

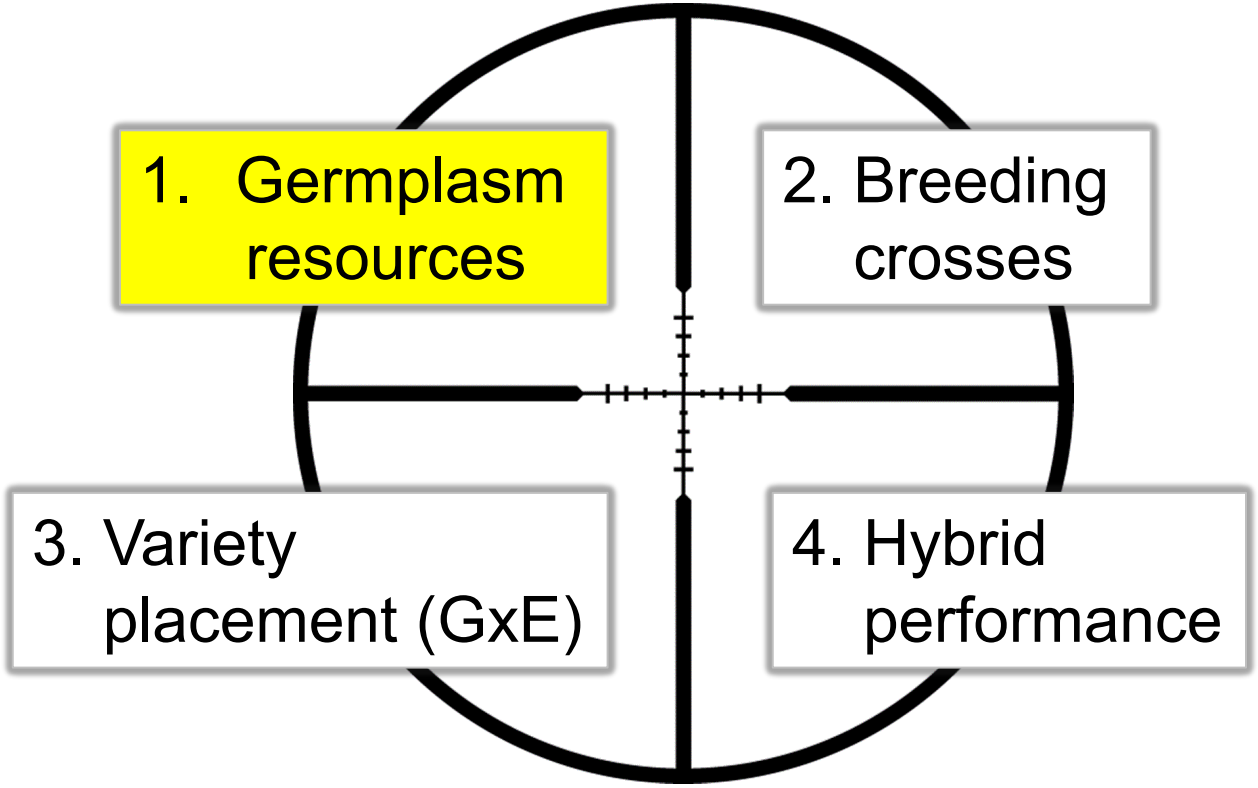
How many reps do we need for phenotyping to maintain acceptable accuracy?

Single crosses

Advanced yield testing

Pre-commercial testing

6. When there are too many candidates to evaluate



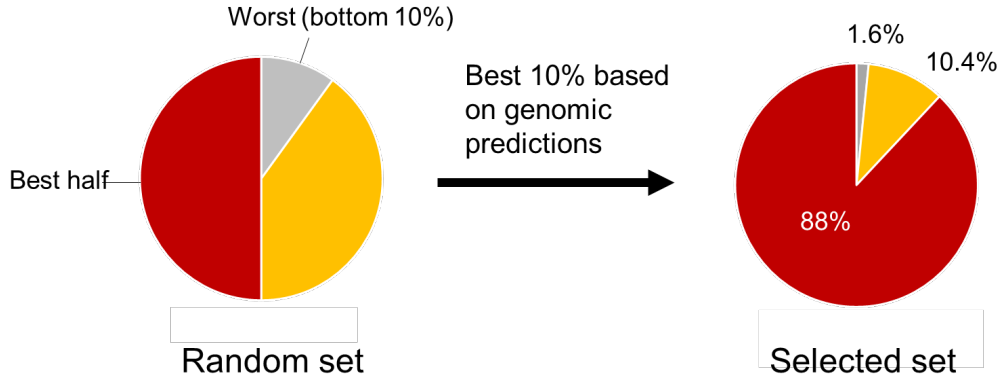
6. When there are too many candidates to evaluate

USDA Soybean Germplasm Collection

~22,000 accessions
Genotyped with 50K SNPs

Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions

Diego Jarquin,* James Specht,* and Aaron Lorenz†,1
*Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Nebraska 68583-0915, and †Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, Minnesota 55108



8. When weather and just bad luck prevent you from harvesting your trials