Genomic Selection Theory

11 July 2012

Jean-Luc Jannink

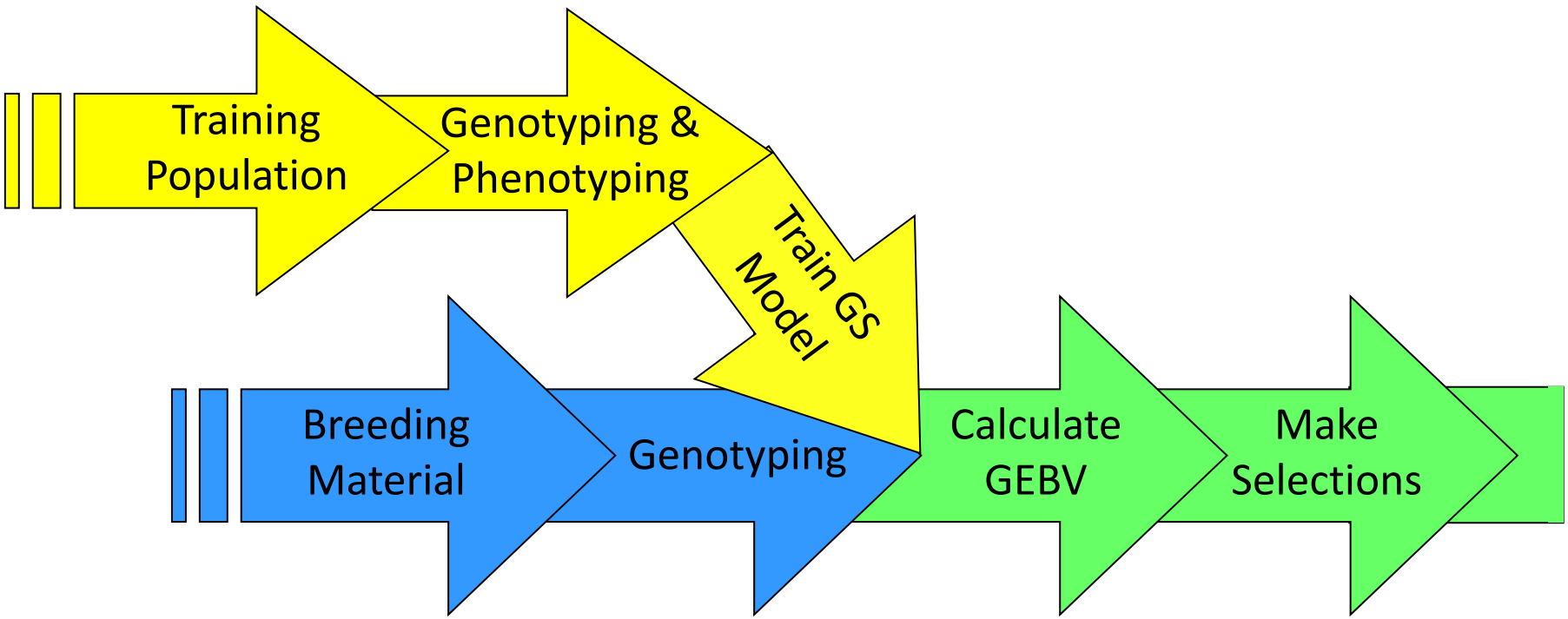
Acknowledgments

- Martha T. Hamblin, Nicolas Heslot, Jeff Endelman, Jessica Rutkoski, Delphine Ly, Julie Dawson, Mark Sorrells
- Vanessa Weber, Gary Atlin, Albrecht Melchinger
- Kevin P. Smith, Vikas Vikram, Ahmad H. Sallam, Aaron J. Lorenz
- Moshood A. Bakare, Melaku Gedil, Ismail Rabbi, Peter Kulakow

Outline

- What is genomic selection?
  - Minimal model
  - Prediction accuracy
- Number of markers
- Size of the training population
  - Should you replicate lines in the TP?
  - Population structure
- Relationship between the training population and the selection candidates

Genomic selection: Prediction using many markers

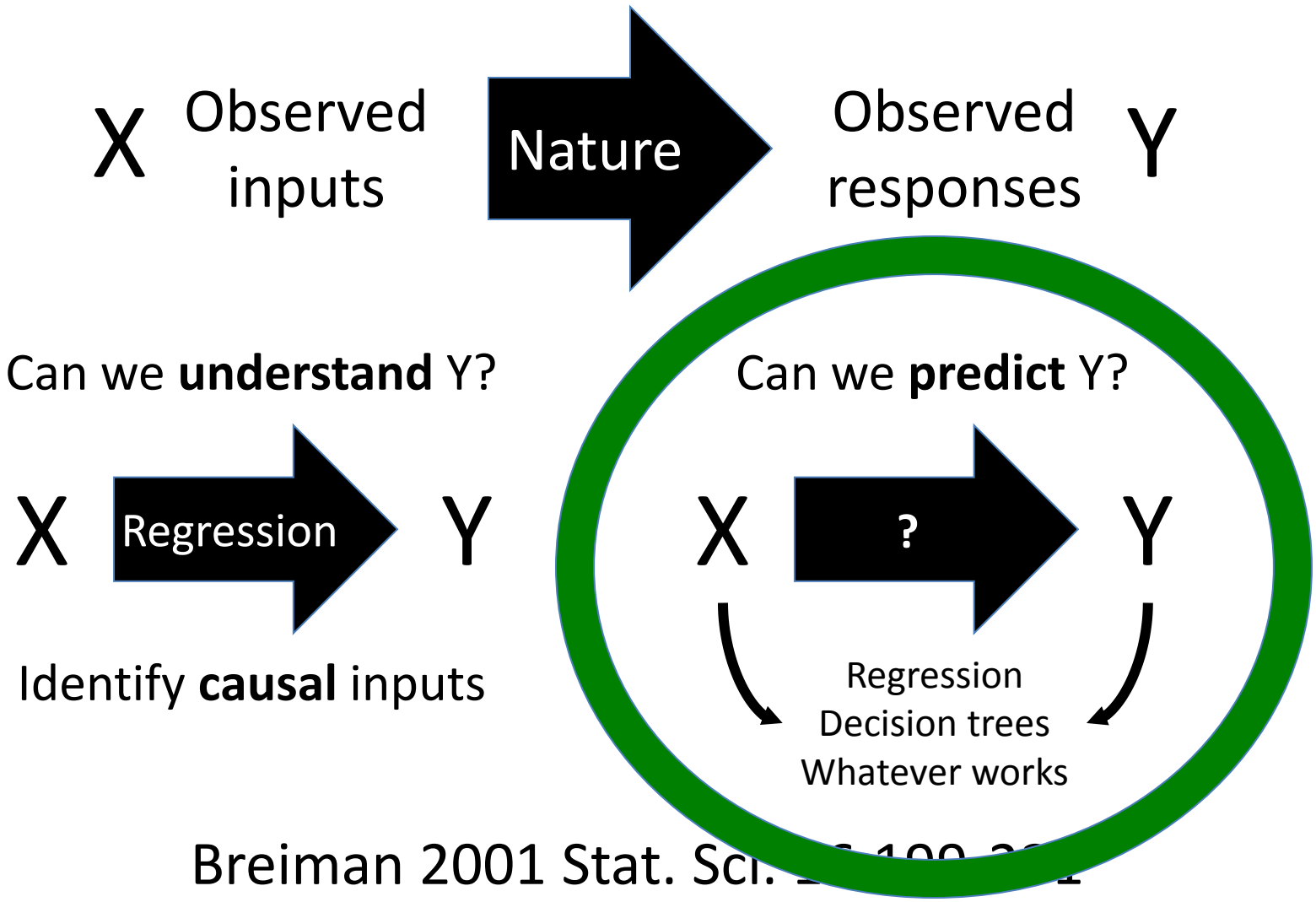


Meuwissen et al. 2001 Genetics 157:1819-1829

Genomic selection principles

- Meuwissen et al. 2001 Genetics 157:1819-1829
- No distinction between “significant” and “non-significant”: all markers contribute to prediction
- (–) More markers than there are phenotypes
- (+) Estimated effects are unbiased
- (+) Capture small effects

Statistical modeling: The two cultures



Breiman 2001 Stat. Sci.

Baseline model

$$\mathbf{y} = \mu + \sum_k x_k \beta_k + \mathbf{e}$$

This is an allele dosage!

$$\beta_k \sim N(0, \sigma_\beta^2)$$

Marker effects –> additive relationship

$$\mathbf{y} = \mu + \sum_k x_k \beta_k + \mathbf{e} \qquad \beta_k \sim N(0, \sigma_\beta^2)$$

$$\hat{a}_i = \sum_k x_k \hat{\beta}_k = \mathbf{X}\hat{\beta}$$

$$var(\hat{\mathbf{a}}) = \hat{\mathbf{A}}\hat{\sigma}_a^2$$

$$\hat{\mathbf{A}} \propto \mathbf{X}\mathbf{X}^T$$

Prediction accuracy = Correlation(predicted, true)

$$R = i r_A \sigma_A$$

r_A = corr(selection criterion, breeding value)

- On simulated data corr(Â, A) is easy
- On real data: $corr(\hat{A}, P) = \frac{cov(\hat{A}, A+E)}{\sqrt{\sigma^2_{\hat{A}}(\sigma^2_A + \sigma^2_E)}}$

$$= \frac{cov(\hat{A}, A)}{\sqrt{\sigma^2_{\hat{A}}(\sigma^2_A + \sigma^2_E)}}$$

$$\left(h^2 = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_E}\right)$$

$$= \frac{cov(\hat{A}, A)}{\sqrt{\sigma^2_{\hat{A}}(\sigma^2_A / h^2)}}$$

$$= corr(\hat{A}, A) \times h$$

Prediction accuracy = Correlation(predicted, true)

$$R = i r_A \sigma_A$$

r_A = corr(selection criterion, breeding value)

- On simulated data corr(Â, A) is easy
- On real data: $corr(\hat{A}, P)$

$$\left(h^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}\right)$$

$$= \quad corr(\hat{A}, A) \times h$$

Effective population size

- Idealized populations randomly mate (real populations don't)
- Increasing **N** weakens **drift** and strengthens **selection** force to shift allele frequencies
- Those forces have a certain strength in a real population
- $\mathbf{N_e}$ summarizes that with reference to an idealized population

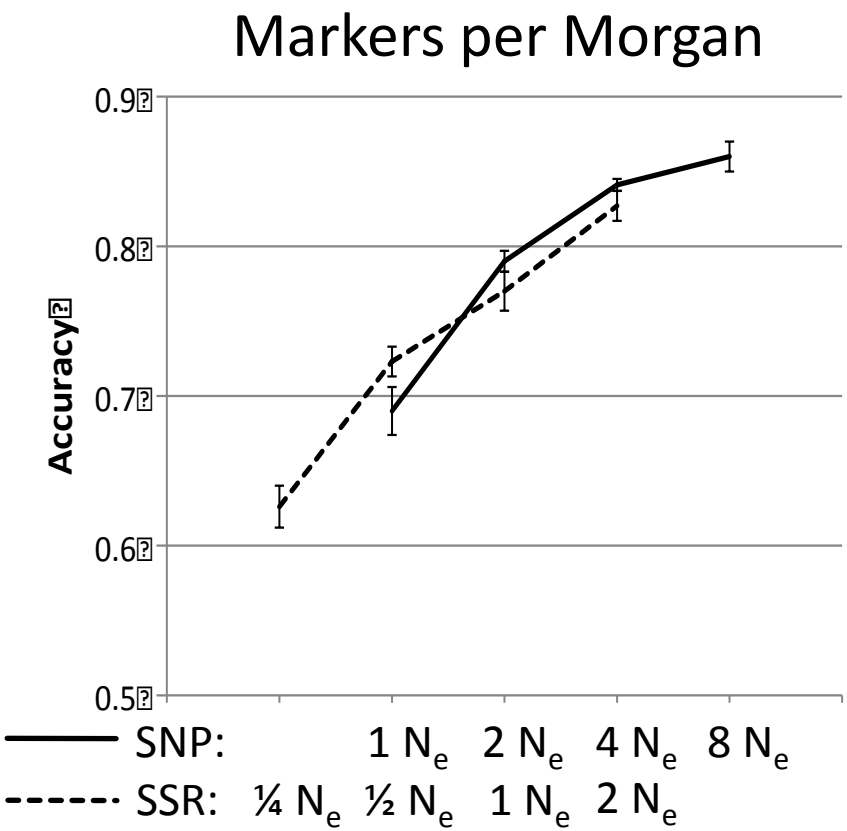
Why N_e matters

- As N_e increases "historical recombination" increases
  - Segments stay in the population longer before being eliminated or fixed
  - Recombination generates segments that segregate independently
  - Expectation: $2N_e$ effective segments per Morgan

$$E(r^2) = 1/(1 + 4N_e c)$$
$$E(r^2) = (5 + 2N_e c)/(11 + 26N_e c + 8[N_e c]^2)$$

Marker density requirements

Solberg et al. 2008



Calus & Veerkamp 2007

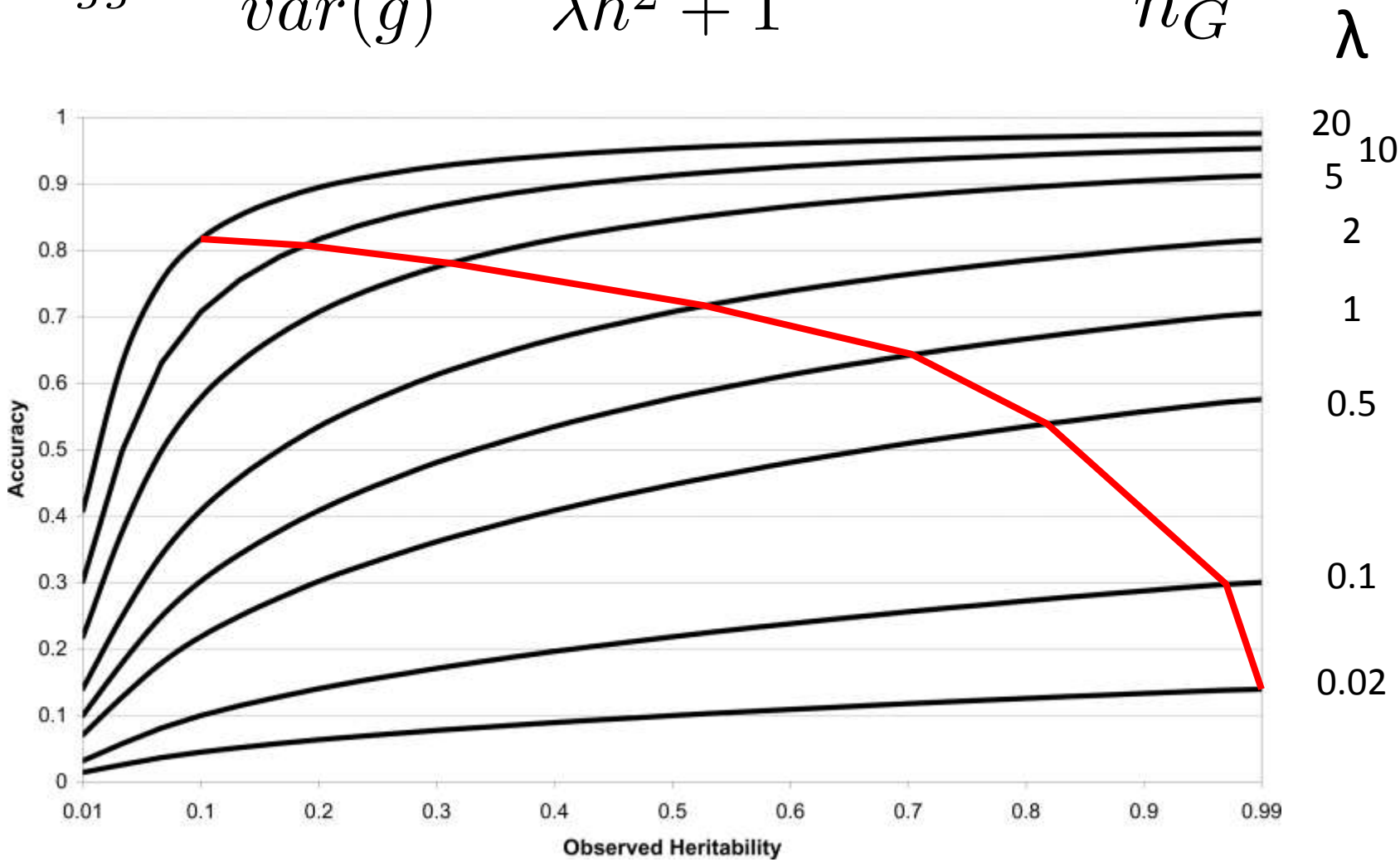
- Average adjacent marker LD should be $r^2 = 0.20$
- Implies a density of $4N_e$ markers per Morgan

Training population size requirements

- Daetwyler, H.D. et al. 2008. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3:e3395

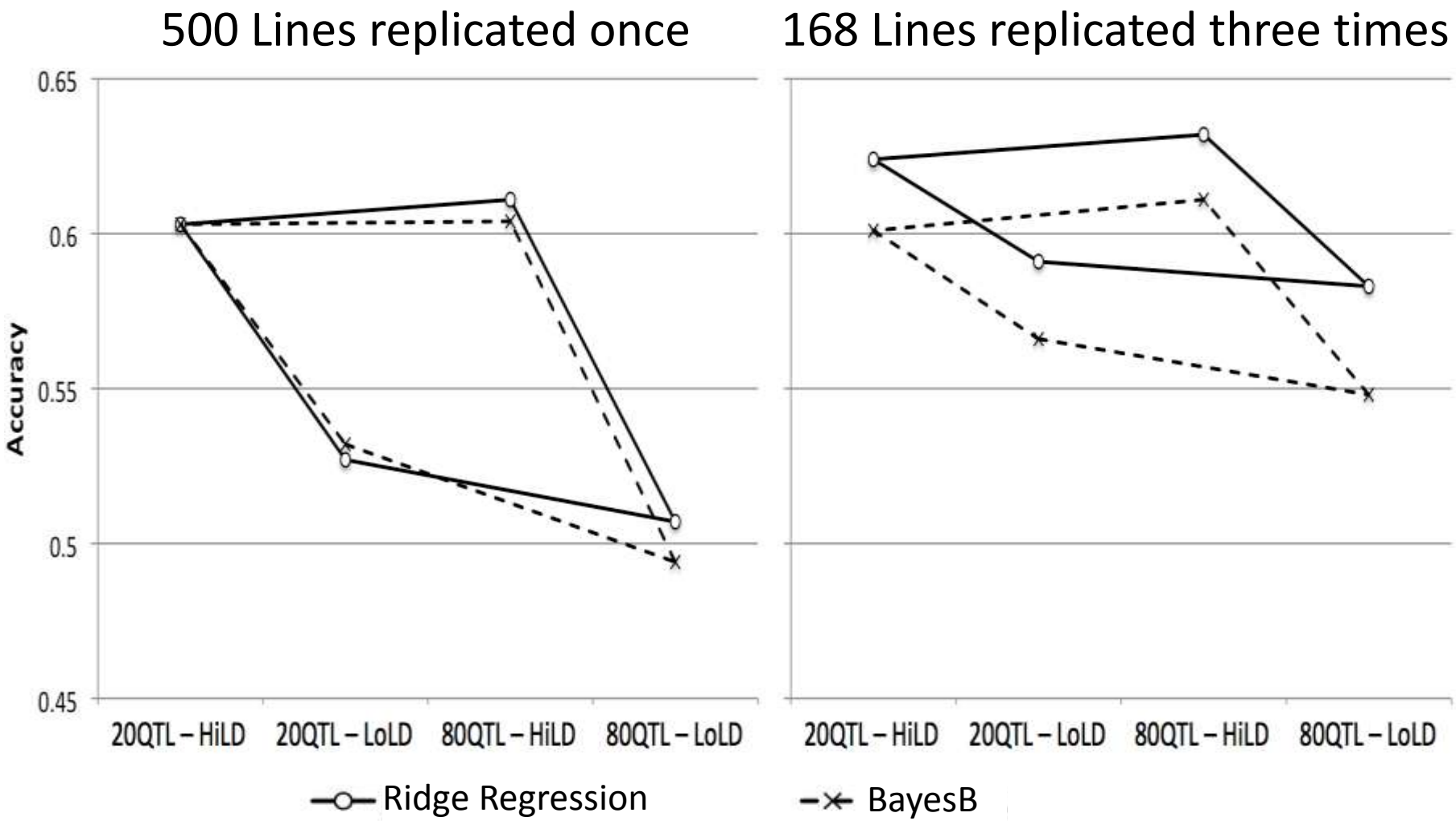
- Assume all loci affecting the trait are known and are independent
- Assume marker effects are **fixed**

$$r^2_{g\hat{g}} = \frac{var(g)}{var(\hat{g})} = \frac{\lambda h^2}{\lambda h^2 + 1} \qquad \lambda = \frac{n_P}{n_G}$$

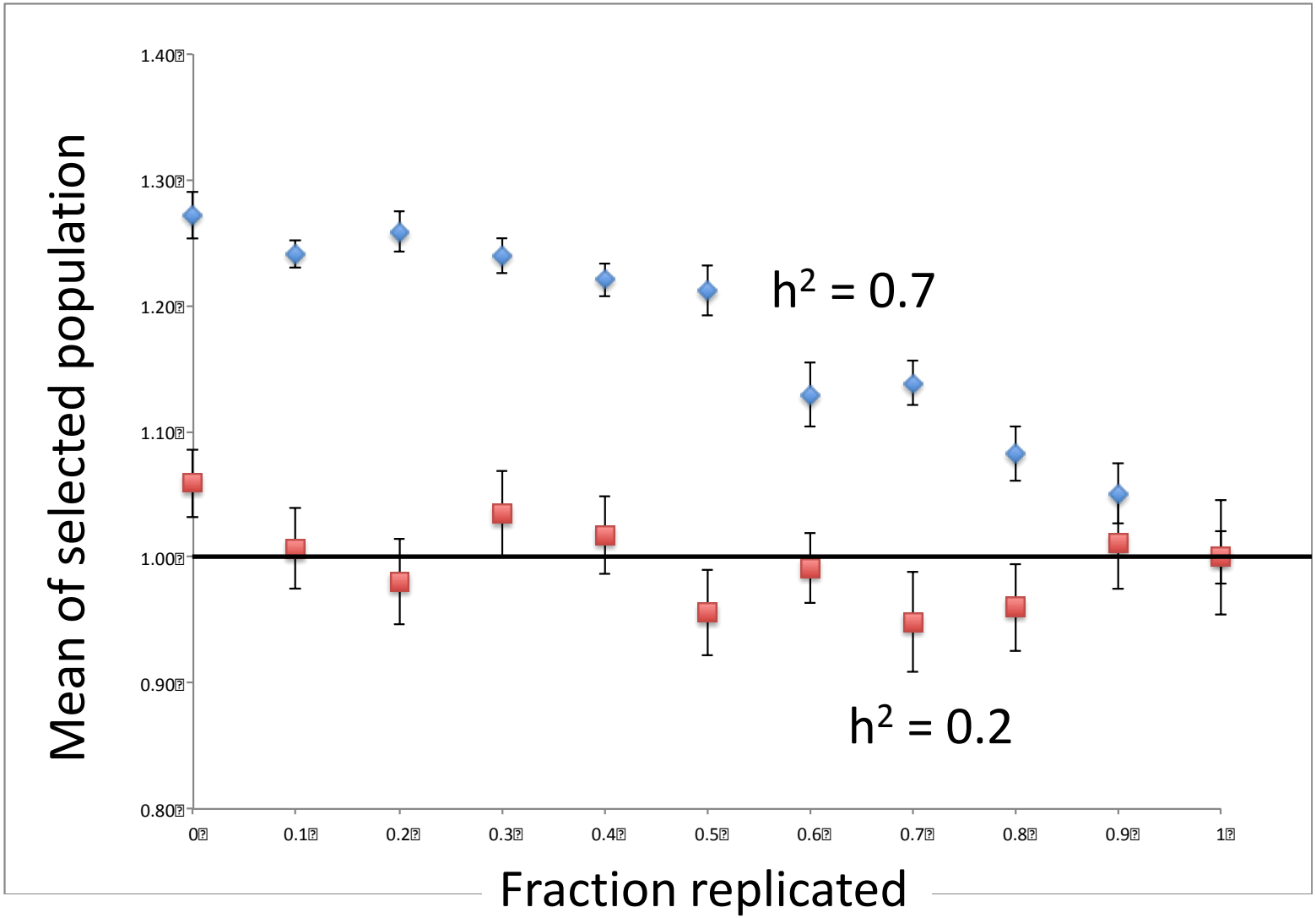


Replicating hurts: 2000 with 1 plot is better than 1000 with 2 plots

To replicate or not to replicate



More lines –> higher pressure



The importance of structure

- Correlation between prediction and phenotype for Grain Yield, Anthesis Date, and Anthesis–Silking Interval in maize
- CIMMYT diversity panel
- 50K SNPs; TP size: 200; VP size 50

GY: 0.44	**AD: 0.45**	**ASI: 0.36**

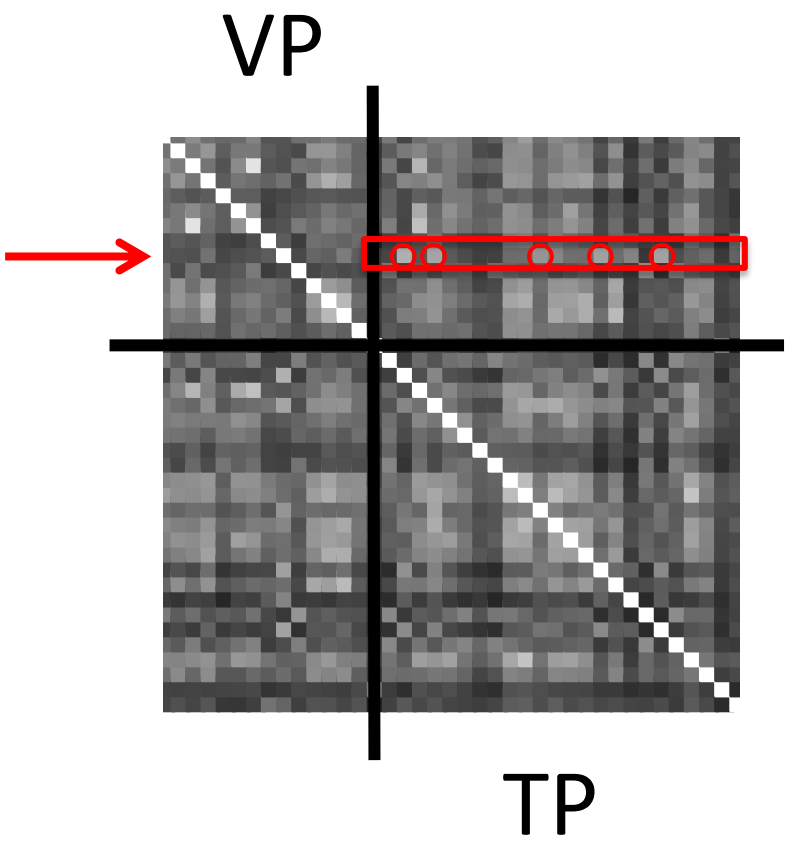
- Structure: 8 Subpopulations
  - Use 200 to estimate subpopulation mean then prediction is mean for subpopulation of origin

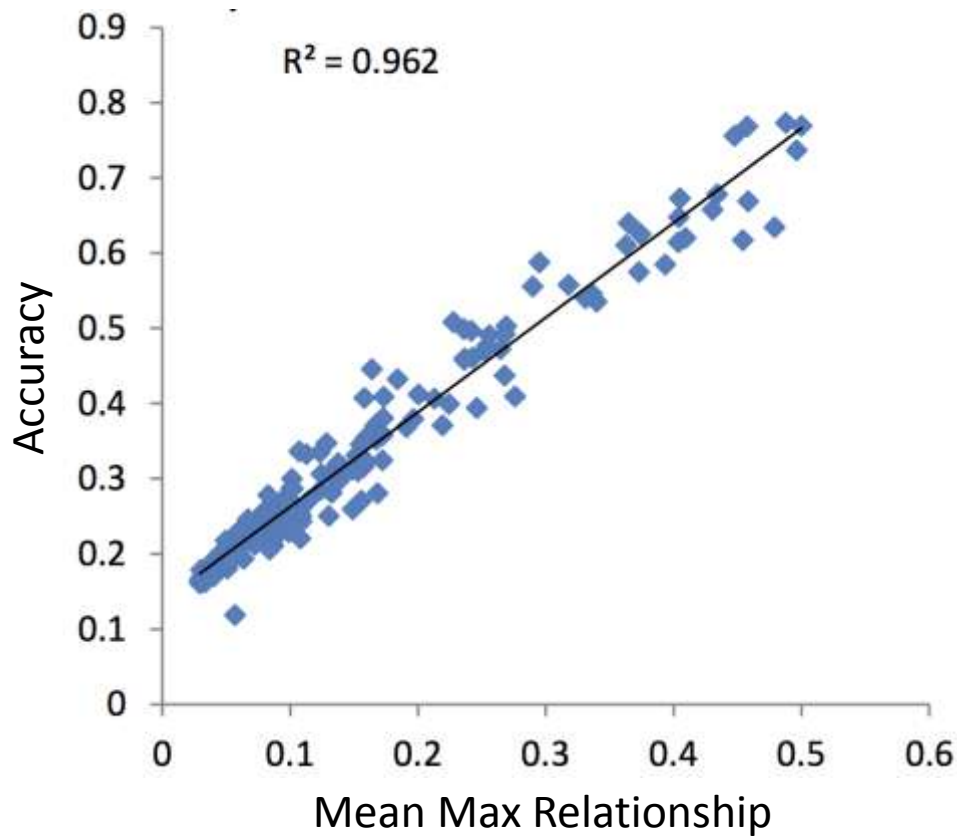
GY: 0.50	**AD: 0.44**	**ASI: 0.46**

- Fraction of the variation due to subpopulation

GY: 0.26	**AD: 0.16**	**ASI: 0.27**

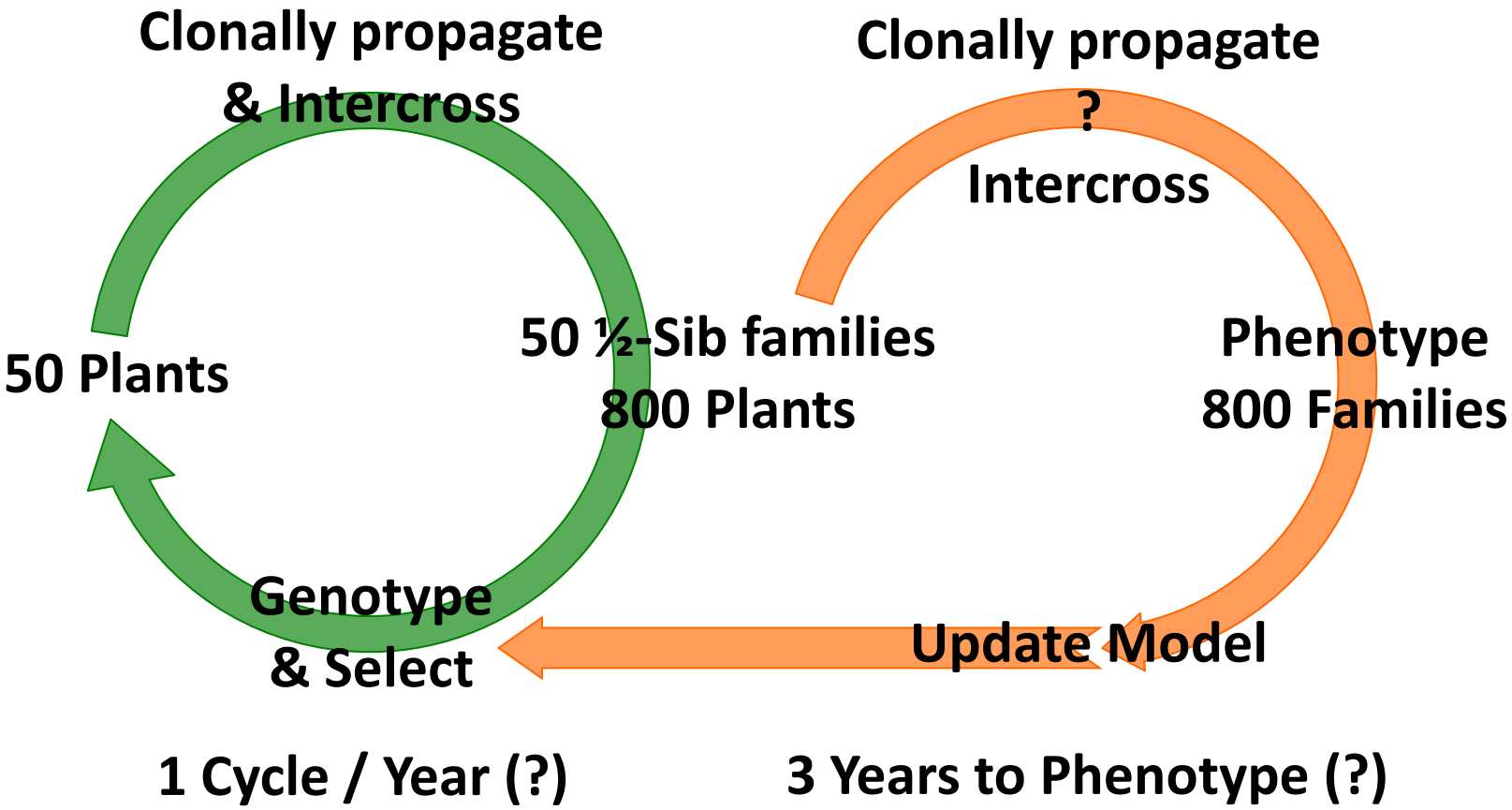
The importance of being in relationship





Clark et al. *GSE* 2012 44:4

As it might play out in alfalfa



Take home messages

- This approach is totally feasible now
  - Statistical and marker technologies are easily powerful enough
  - Logistics and informatics are a challenge
- Planning the training population is most critical
- Expect some outcomes to be non-intuitive
- Don’t trust a theoretician: go out and do it

Questions?

Empirical vignettes

- Genomic selection can work for a difficult crop: Cassava
- We have a field-validated success in barley

Is cassava a little like a forage?

- Cassava is a “young” crop (still a bit wild)
- Cassava is outcrossed and clonally propagated

Phenotypic data

- DNA from 623 Clones important at IITA
- ~ 50,000 plots of ***historic*** data
- Very high differences in replication
- 17 traits (disease, morphology, yield)

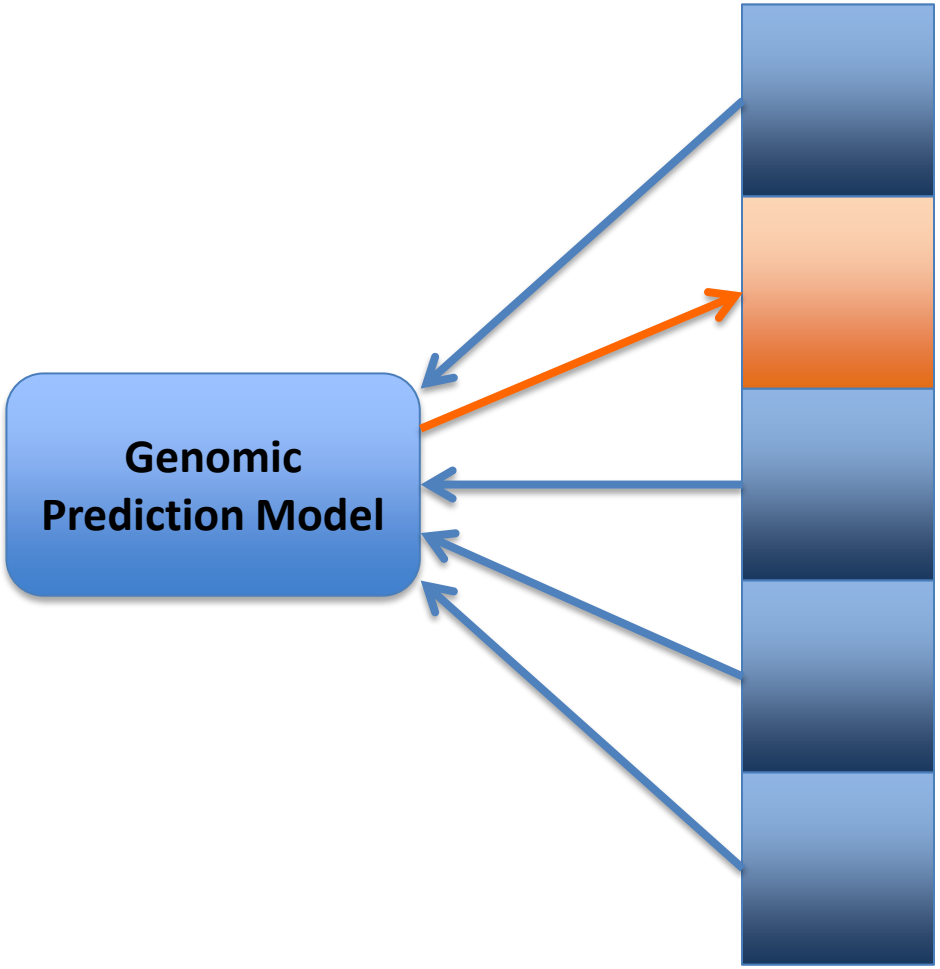
GBS markers

- 4984 SNPs
- Average of 25.5% missing data per marker
- Calling heterozygotes is a challenge

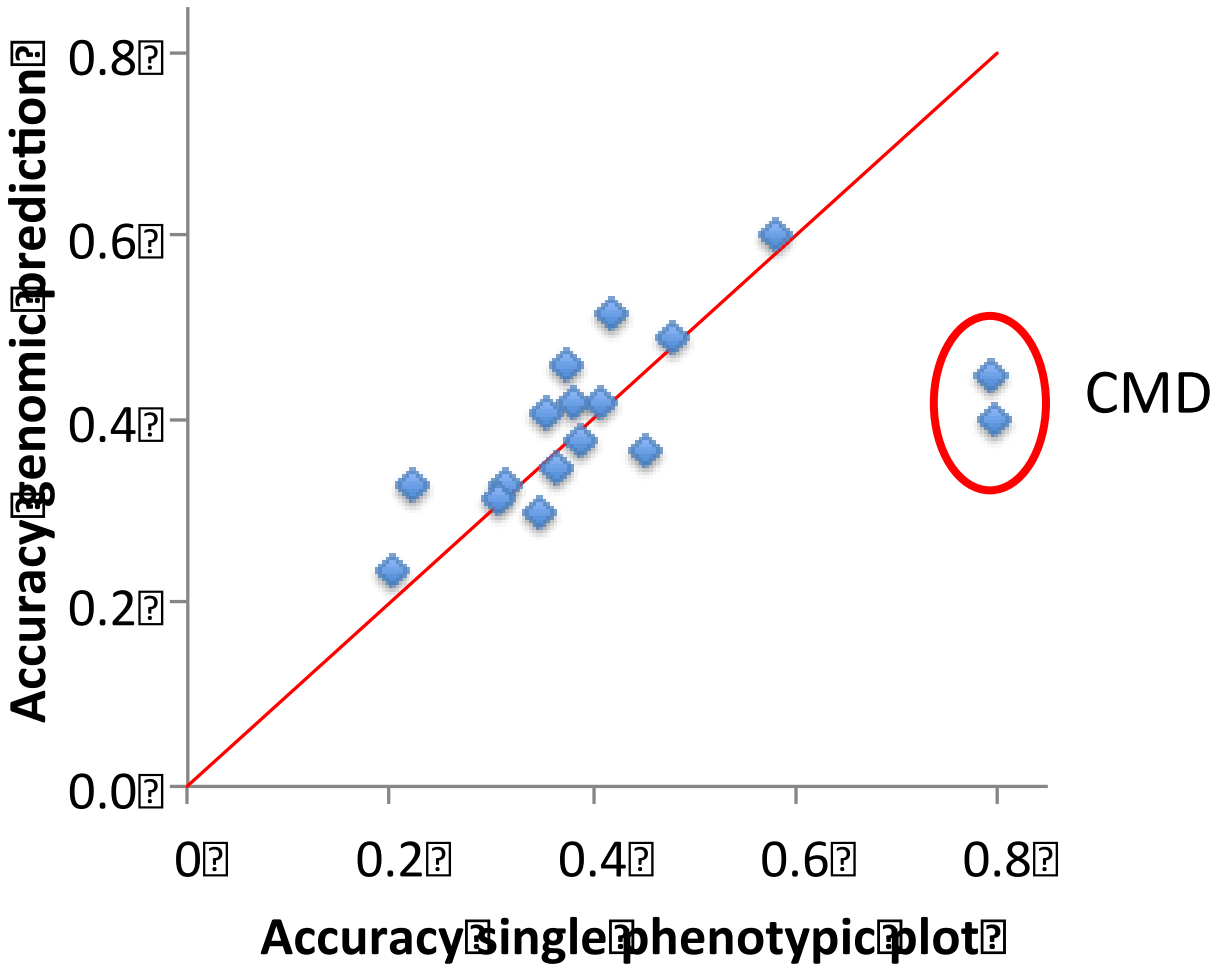
Validation method

- Cross validation:

  Predict subsets of the data that did *not* contribute to building the model



Comparison to phenotypic accuracy

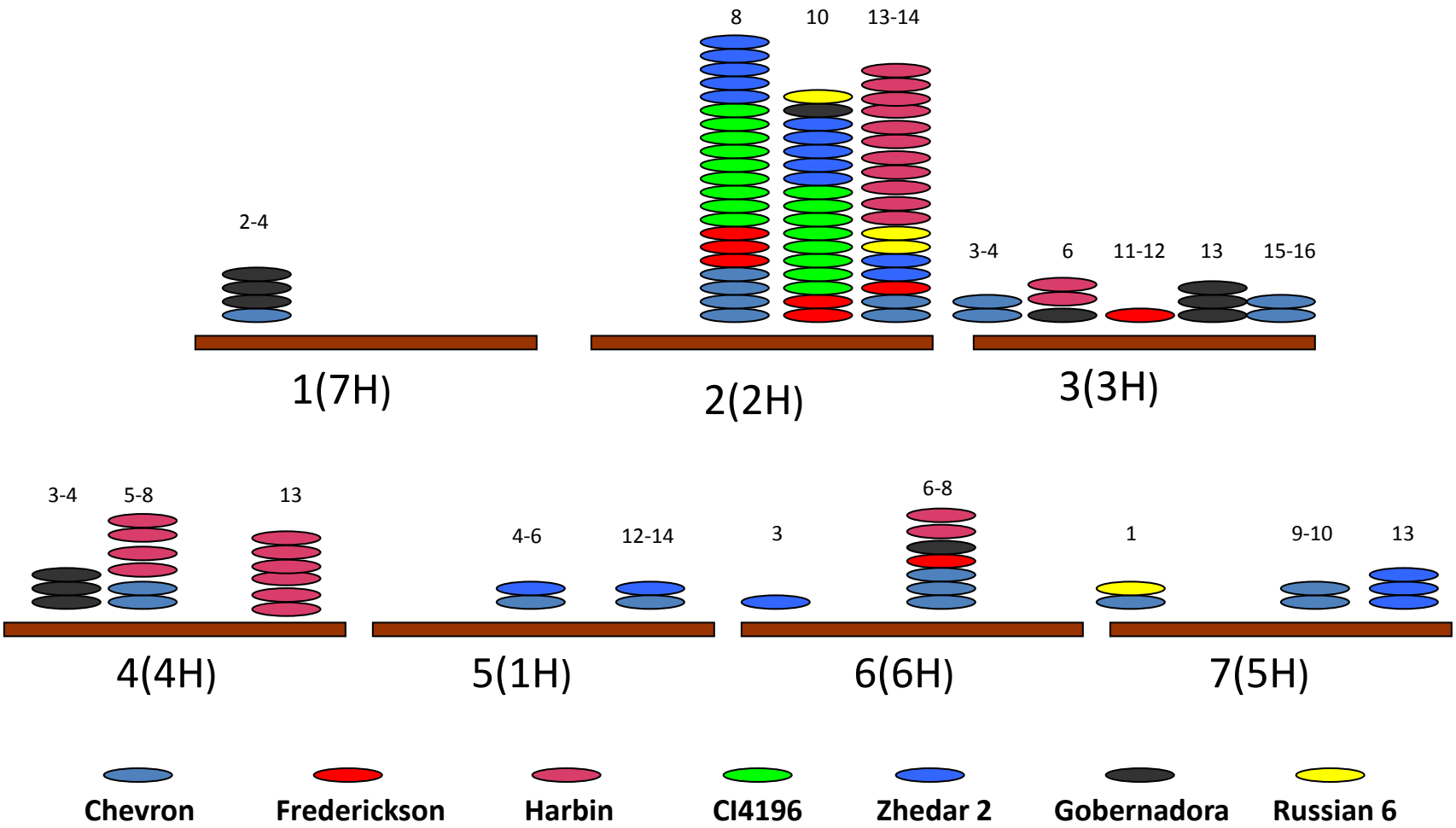


Higher accuracy would be good but

- GS will reduce the breeding cycle time from 5 to 2 years [*2.5 × Faster*]
- **Any accuracy above 0.4 *cannot* be beat by phenotypic selection**
- Biases make phenotypic look better and genomic look worse than reality

Barley Fusarium head blight

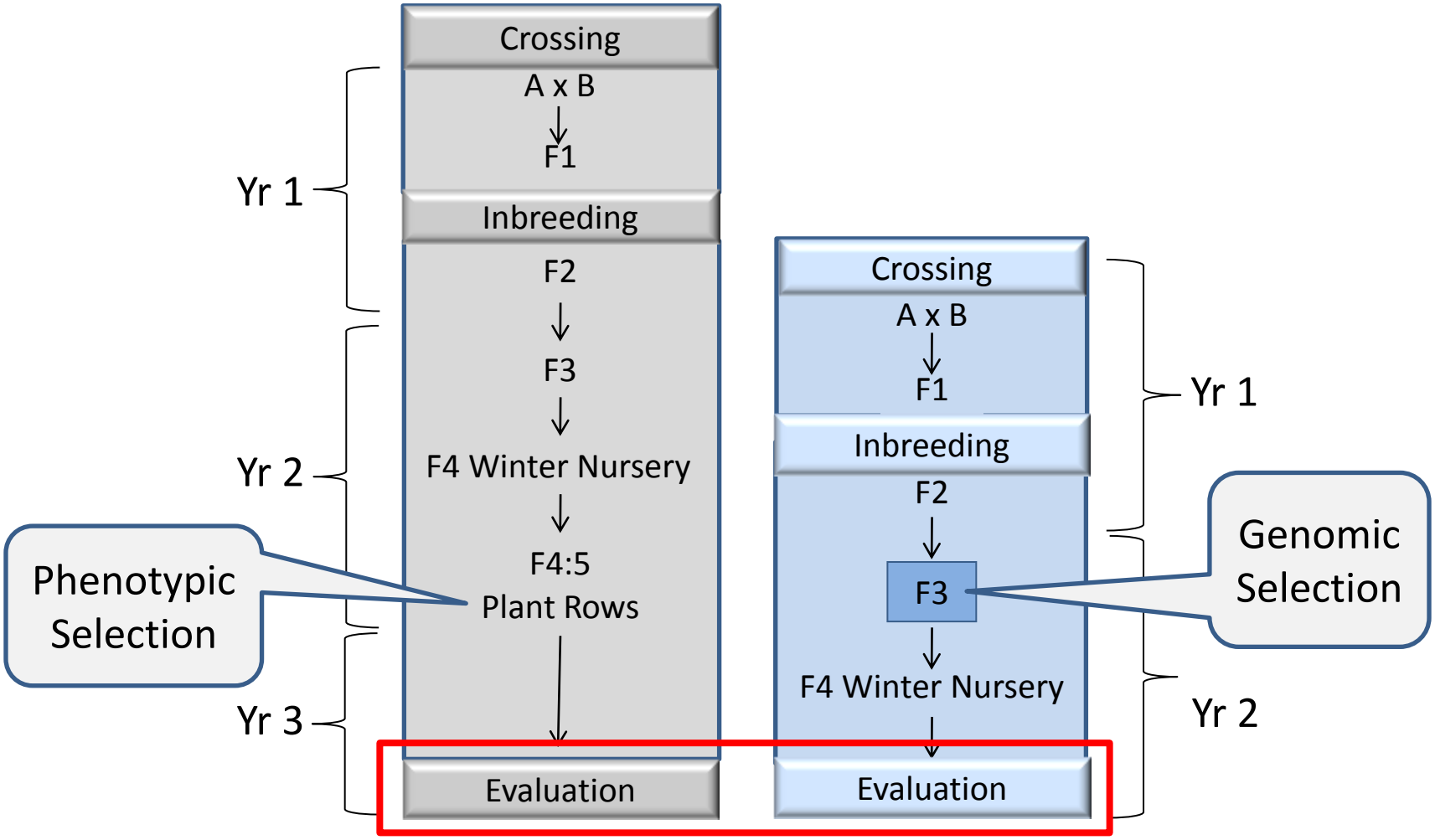
FHB resistance is polygenic



Genomic selection set up

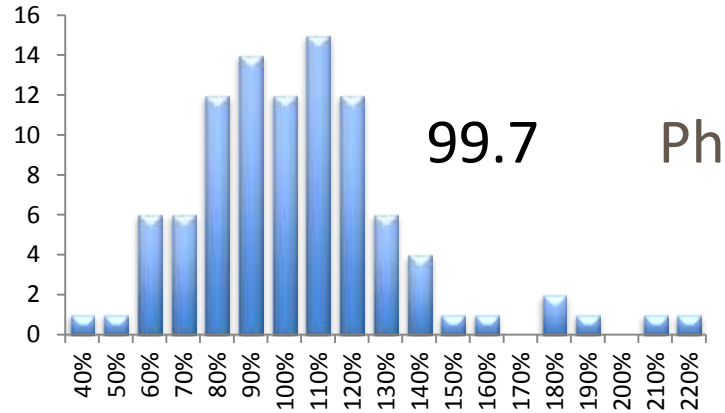
- Three breeding programs
- 685 barley (6-row) lines
- Three years, 2 locations
- 1500 SNPs
- Choose 384 optimized for PIC & distribution
- 1440 progeny from 60 crosses
- “Project” parental SNP onto progeny

Head-to-head comparison



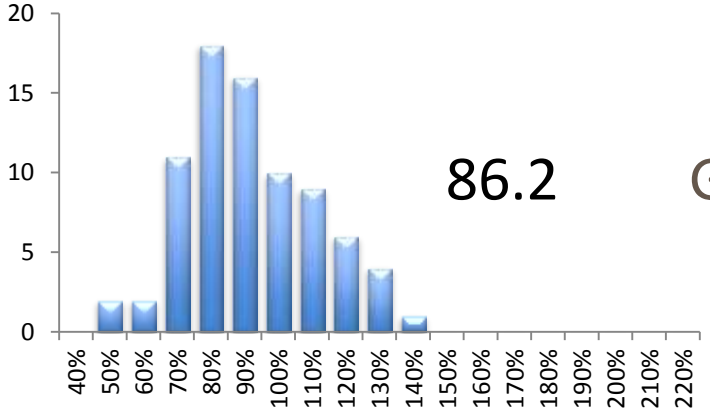
Phenotypic vs. Genomic Selection

FHB



99.7

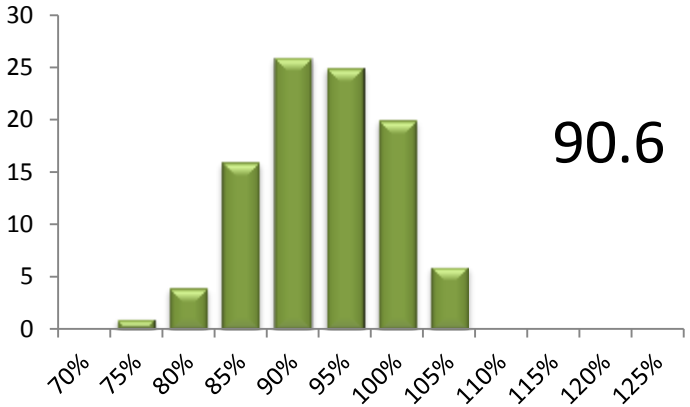
Phenotypic



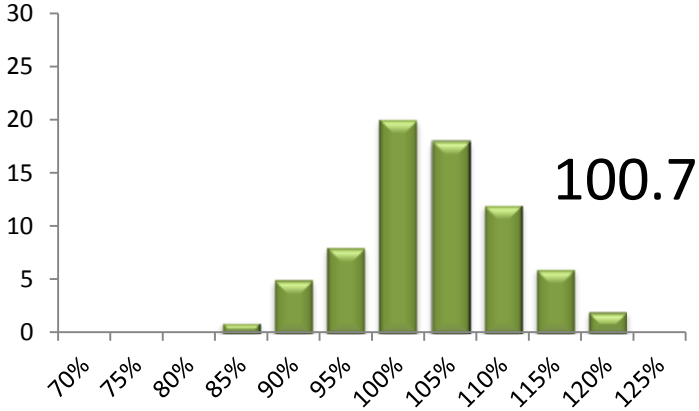
86.2

Genomic

Yield



90.6



100.7