Genomic Selection in Plants: Empirical Results and Implications for Crop Improvement

Mark E. Sorrells



Cornell University
Department of Plant Breeding & Genetics

Celebrating 100 Years of Crop Improvement

Presentation Overview

Genomic Selection

- Marker-assisted Breeding
- Characterizing GxE using marker effects
- Integration in a breeding program
- Biparental vs Multi-Family predictions
- Recurrent Genomic Selection



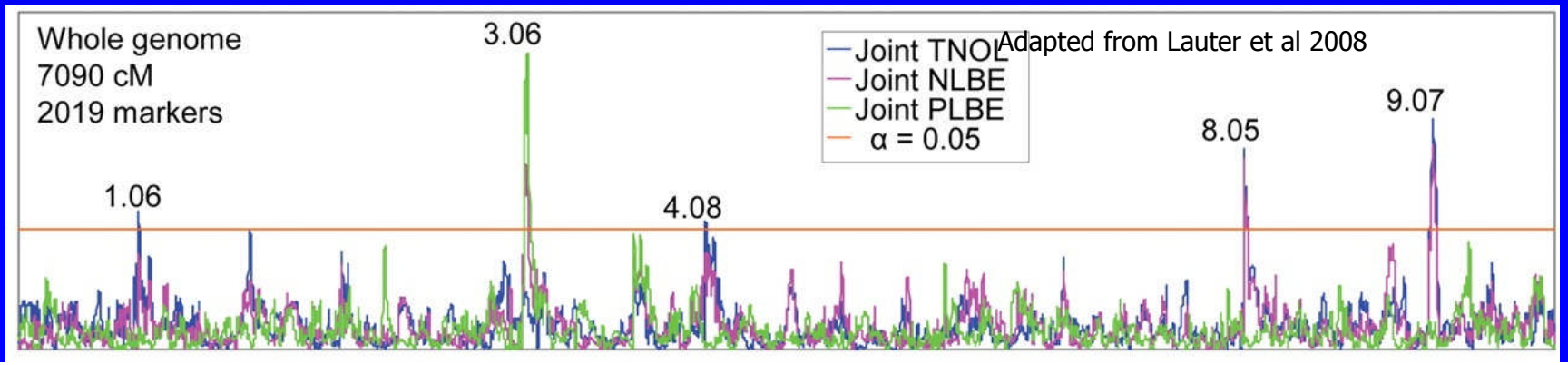
Plant Breeding is a Predictive Science

Novel Breeding Strategies are driven by Technology & New Knowledge

Prediction accuracy is important at every step in the breeding program & breeding methods are designed to improve accuracy.

- Family Selection Methods
- Progeny Testing
- Marker Assisted Selection
- Genomic Selection





Adapted from Lauter et al 2008

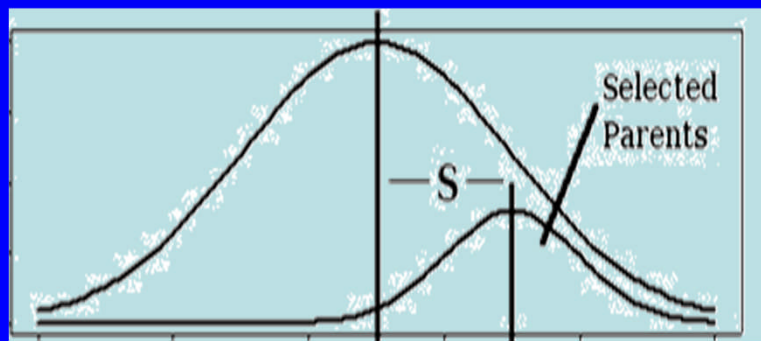
Molecular Breeding Goals

- Allele discovery
- Allele characterization & validation
- Parental & progeny selection for superior alleles at multiple loci to generate transgressive segregation

Increase response (R_t) from selection

1) Increase selection intensity (i)
2) More accurate selection (r)
3) Increase the genetic variation (σ_A)
4) Reduce selection cycle time (t)

$$R_t = \frac{ir\sigma_A}{t}$$



Molecular Breeding

- Marker-assisted selection (MAS) or marker-assisted recurrent selection (MARS) use only significant markers.
- Genomic selection (GS) uses genome-wide genotypic data and phenotypic data to predict any trait with an accuracy sufficient to realize genetic gain based only on that prediction.
- GS estimates all marker effects simultaneously, thus capturing markers with small effects.

Why Genomic Selection?

- Decreasing cost of genotyping
- Large cost of phenotyping
- Most traits are controlled by QTL with small effects
- Simulation & empirical studies have shown better gain from GS than from traditional selection

Genomic Selection Methodology

Haley & Visscher 1998; Meuwissen , Hayes & Goddard 2001

A Training Population is genotyped with a large number of markers and phenotyped for important traits (number of predictors > number of observations)

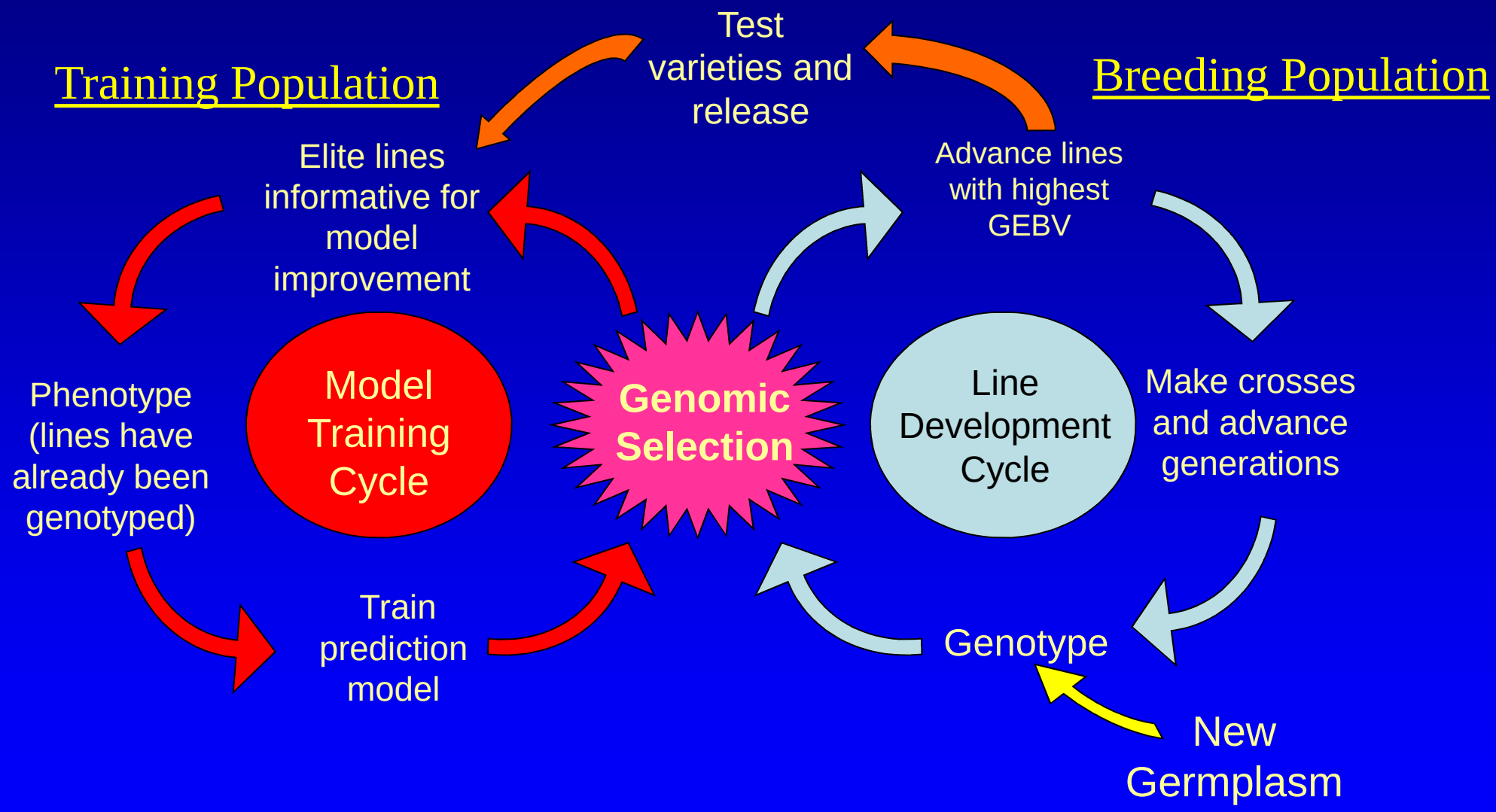
- Genome-wide markers are considered to be random effects and all marker effects on the phenotype are estimated simultaneously in a single model
- One or more markers are assumed to be in LD with each QTL affecting trait
- Prediction model attempts to captures the total additive genetic variance to estimate breeding value of individuals based on sum of all marker effects

In a Breeding Population individuals are genotyped but not phenotyped

- A genomic estimated breeding value (GEBV) for each individual is obtained by summing the marker effects for that genotype
- Prediction model can be used to impose multiple generations of selection

GS in a Plant Breeding Program

Heffner, Sorrells & Jannink. Crop Science 49:1-12



Genomic selection reduces cycle time & cost by reducing frequency of phenotyping

Factors Affecting the Accuracy of GEBVs

- Level and distribution of LD between markers and QTL
  - Meuwissen 2009: Minimum number of markers for across family= N*e**L where N*e* is the effective population size and L is the genome size in Morgans E.g. Wheat: 50x30=1500
  - Marker number must scale with effective population size.
- Size of training population
  - Larger is better and over time re-training models will be required
  - Meuwissen 2009: Minimum number of records for across family prediction = 2*N*e**L E.g. Wheat: 2X50x30=3000
- Distribution of QTL effects
  - Many small effect QTL or low LD favor BLUP for capturing small effect QTL that may not be in LD with a marker
- Heritability of the trait
  - Like phenotypic selection, more records are required for low heritability traits





Adapting Genomic Selection to Plant Breeding



Mating schemes:

- Plant breeders can use many diverse mating schemes
- Animal Parental values are mainly based on half-sib families

Population size:

- For most crop species large populations can be generated
- For animals, many daughters are tested for each bull & data is accumulated over generations and families

Replication:

- Inbred lines, testcross hybrids and clonally propagated crops can be replicated in time and space
- Each animal is a unique genotype and heterozygous
- GxE is a major issue in plant breeding

Choosing a Statistical Model for GS

Problem: Must estimate many QTL effects from a limited number of phenotypes or records (large ‘p’ small ‘n’ problem)

Previous Approach: Least Squares regression for variable selection

- Markers are fixed effects and an arbitrary threshold for significance is used to fit the markers
- Results in overestimation of significant effects and loss of small effects
- Options: Variable selection or shrinkage estimation can be used to deal with oversaturated regression models

Current methods: Linear mixed models, Bayesian estimation, Machine Learning

- No Testing for significance
- Many QTL effects, set as random effects, can be estimated simultaneously

Choosing a Statistical Model for GS

Current methods differ based on their assumptions and treatment of marker effects:

Ridge Regression Best Linear Unbiased Prediction (BLUP):

- Assumes equal variance of marker effects giving a Gaussian distribution of marker effects and equally shrinks all QTL effects towards zero

Bayesian Regression:

- Bayesian Shrinkage Regression - a.k.a. Bayes A (Meuwissen et al), uses an inverse chi-square prior on marker variances resulting in a scaled t distribution for marker effects. Uses all markers.
- Bayesian Variable Selection - e.g. Bayes B , uses an inverse chi-square prior on marker variances resulting in a scaled t distribution for marker effects. Uses all markers. Some marker values =0.
- Bayesian LASSO -has an exponential prior on the marker variances giving a double exponential distribution for the marker effects resulting in more shrinkage for small marker effects and less shrinkage for large effects.

Non Parametric Methods:

- Reproducing Kernel Hilbert Spaces regression - based on genetic distance and a kernel function (Gaussian) with a smoothing parameter to control the distribution of the QTL effects (Gianola et al)
- Random Forest – A machine learning algorithm based on a set of regression trees grown from samples of observations using a random set of predictors. Predictions are averaged over the different trees.

Choosing a Statistical Model for GS

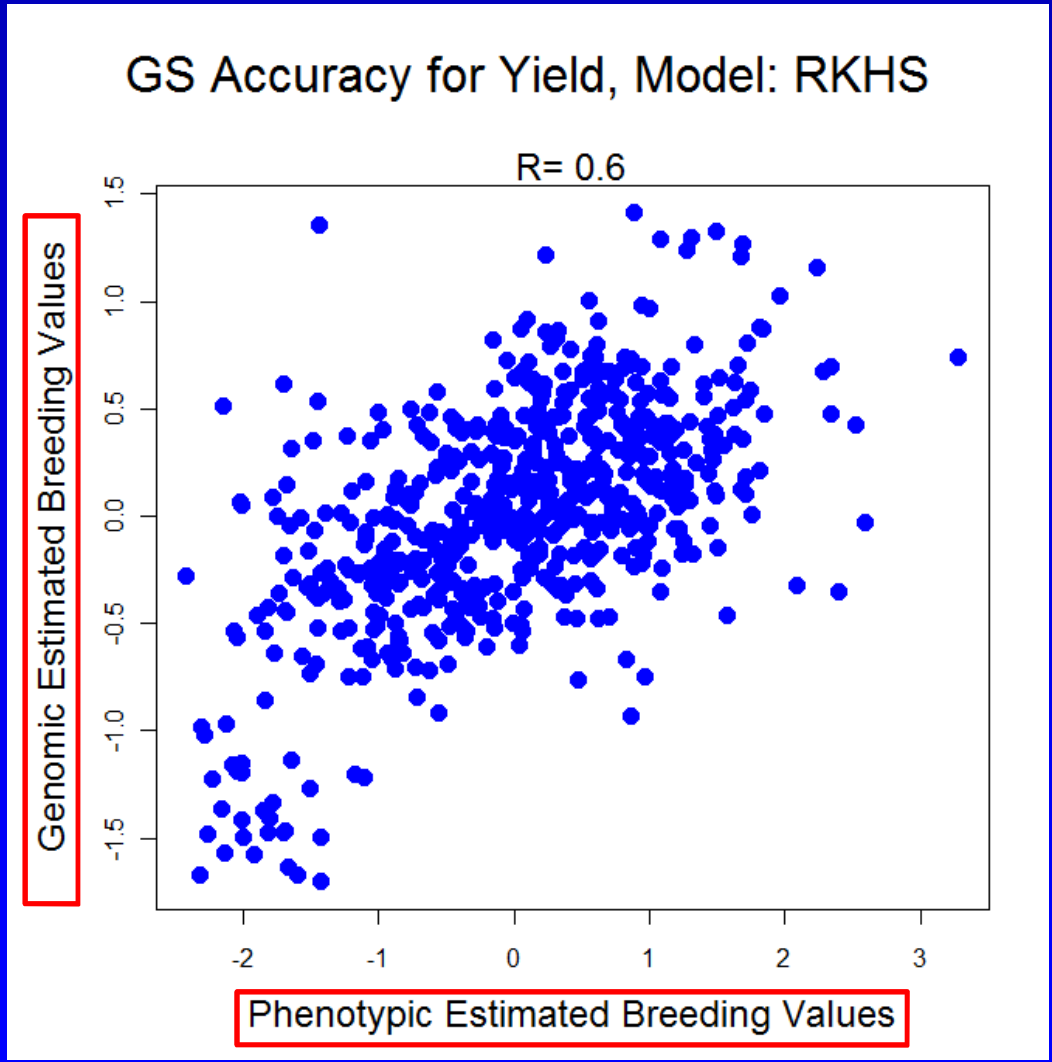
Model performance is based on correlation between GEBV and True Breeding Value (TBV)

Accuracy (*r*) = Pearson's correlation coefficient using:

- Breeding values based on phenotype
- Breeding values estimated based on GS prediction models

r is analogous to the square root of the narrow sense heritability

Must use different individuals for creating the prediction model and evaluating the accuracy



Optimizing Prediction using Marker Effects

Nicolas Heslot

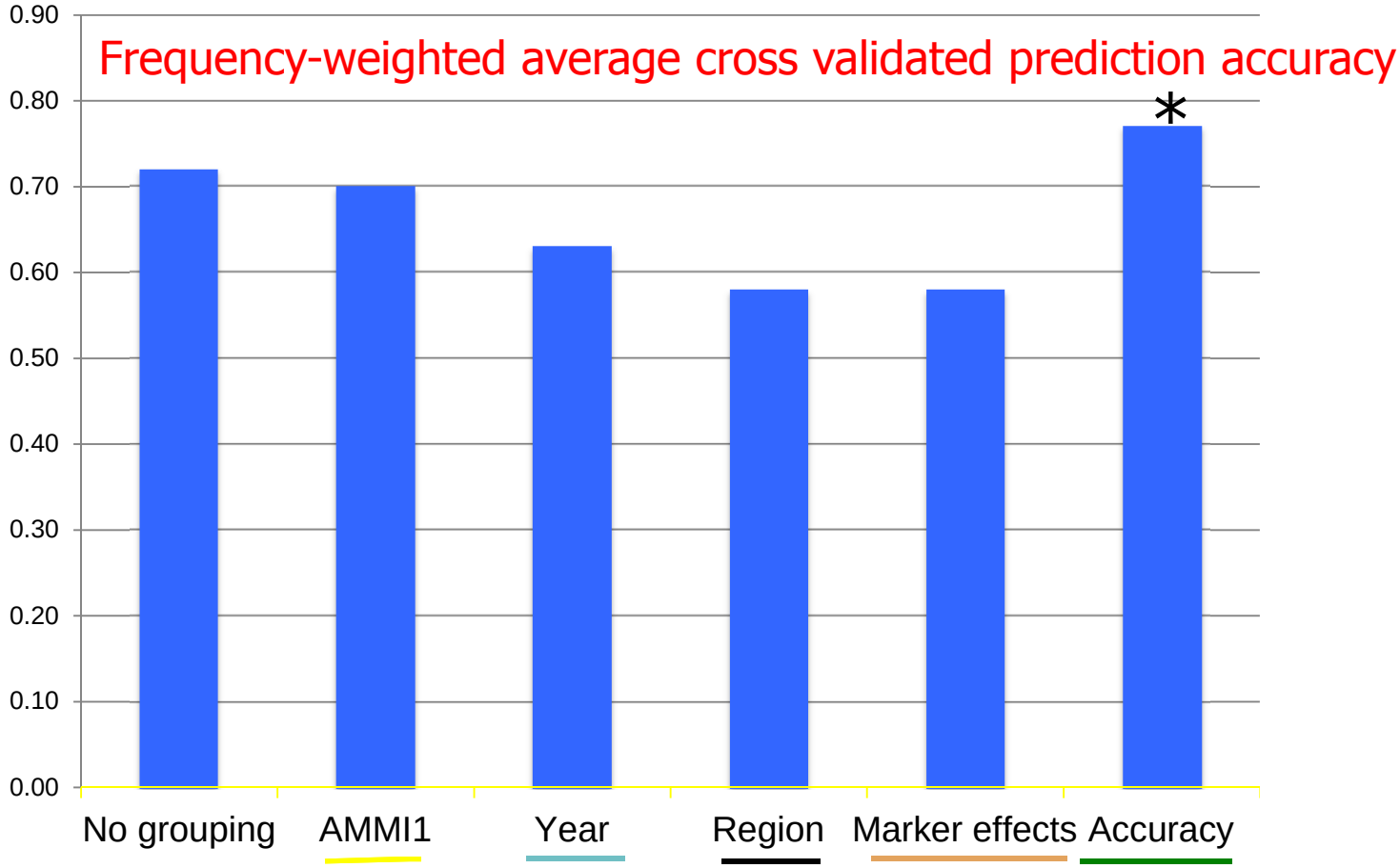
- Limagrain Europe Barley
  - 996 F6 &F7 lines
  - 58 environments
  - 13,682 plots
  - Unbalanced – only 18 genotypes present in >50% of the environments
- Bayesian LASSO GS model
- Characterize allele-effect estimates for each test location
  - Identify outlier environments
  - Group relevant mega-environments



Identifying Relevant Environment Subgroups for Genomic Selection

Environments were grouped based on Additive Main Effects & Multiplicative Interaction (AMMI), year, region, marker effects, and pairwise prediction accuracy between environments.

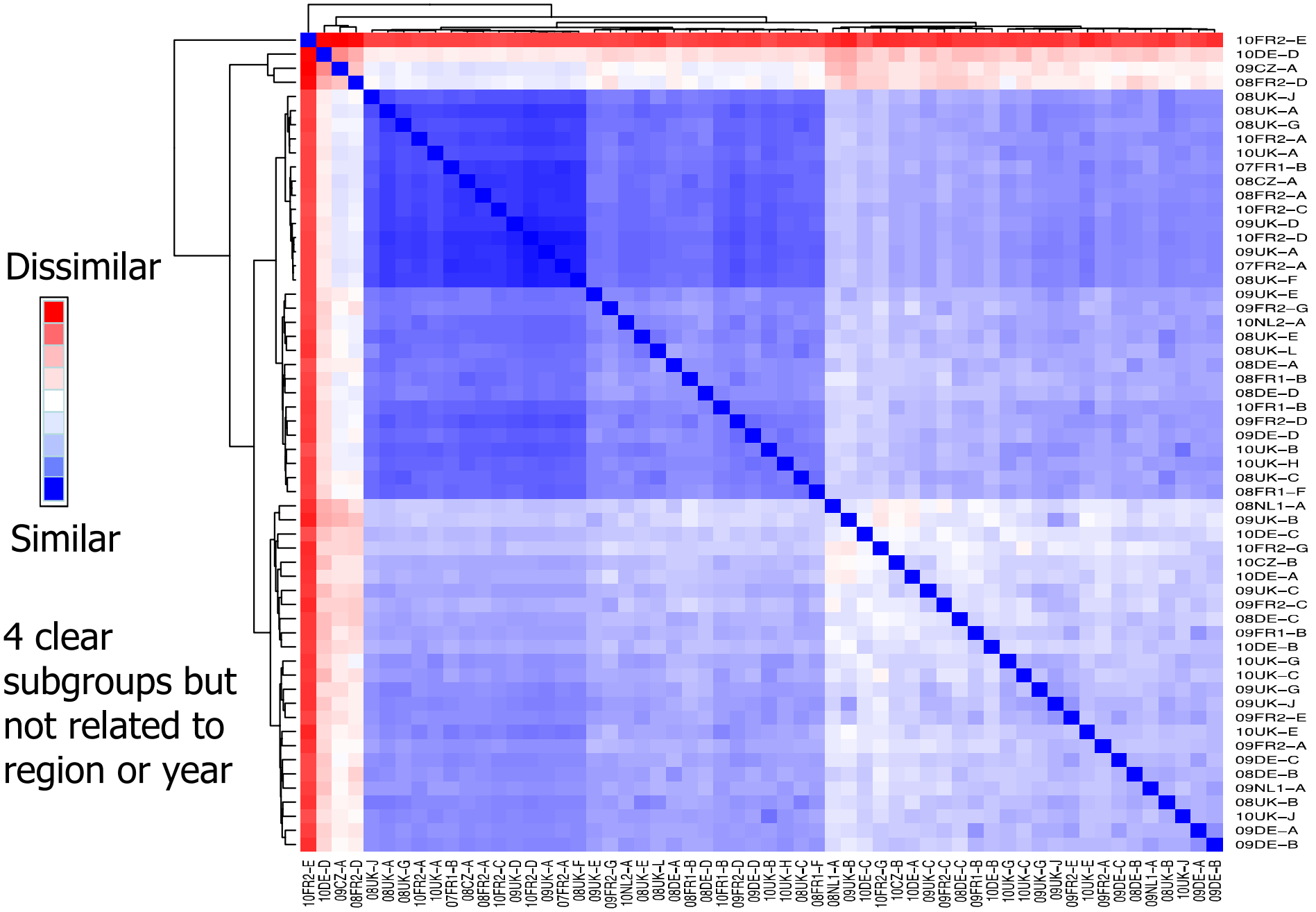
GxE accounted for 29% of the variance in the mixed model and was 2X the genotypic variance.



Subdivision of the environments was advantageous only for the partition made using the **average reciprocal prediction accuracies** to cluster the environments.

Use of Marker Effects to Cluster Environments

Marker effects for all lines in each environment formed a balanced dataset for computing Euclidean distances



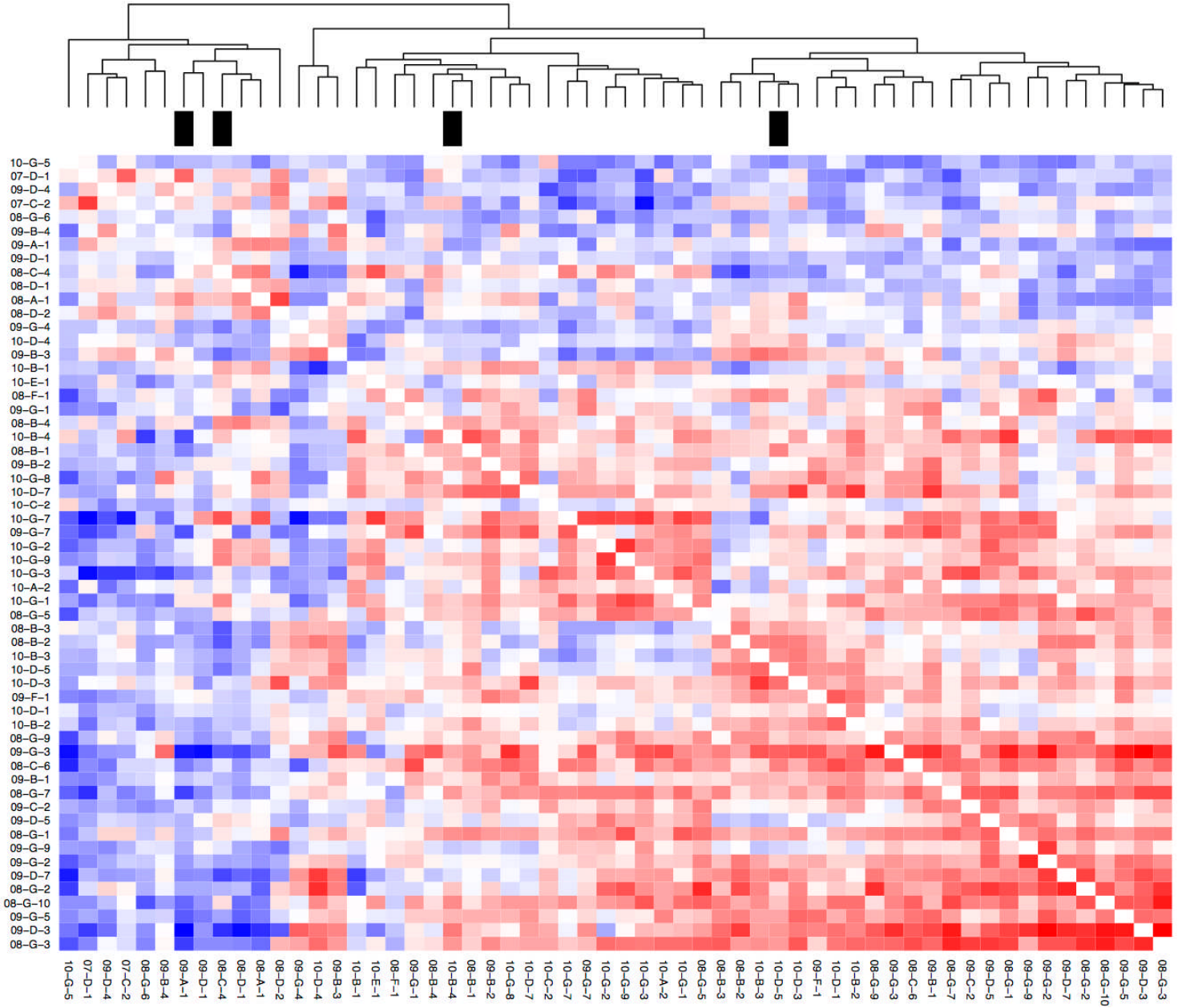
4 clear subgroups but not related to region or year

Outlier environments were identified but there was no appreciable gain in accuracy.

Use of Pairwise Prediction Accuracy to Cluster Environments

(Heslot et al Crop Science In press)

Black bars represent outlier environments



Grouping based on average reciprocal accuracies increased prediction accuracy

The patterns are not as clear as grouping based on marker effects.

Optimizing the Composition of the Training Population
(Heslot et al Crop Sci. In Press)

Train GS model in each environment

↓

Compute mean accuracy of each training environment for predicting line performance in the other environments then rank them

↓

Remove environment one at a time starting from the least predictive

↓

Train and cross validate a GS model on the remaining training population: Predictive set

↓

Predict the removed data with the GS model: Unpredictive set

↓

Use both accuracy measurements to decide the cut-off point

Uses the predictive ability of an environment to optimize the composition of the training population derived from the complete dataset



Optimizing the Composition of the Training Population
(Heslot et al Crop Sci. In press)

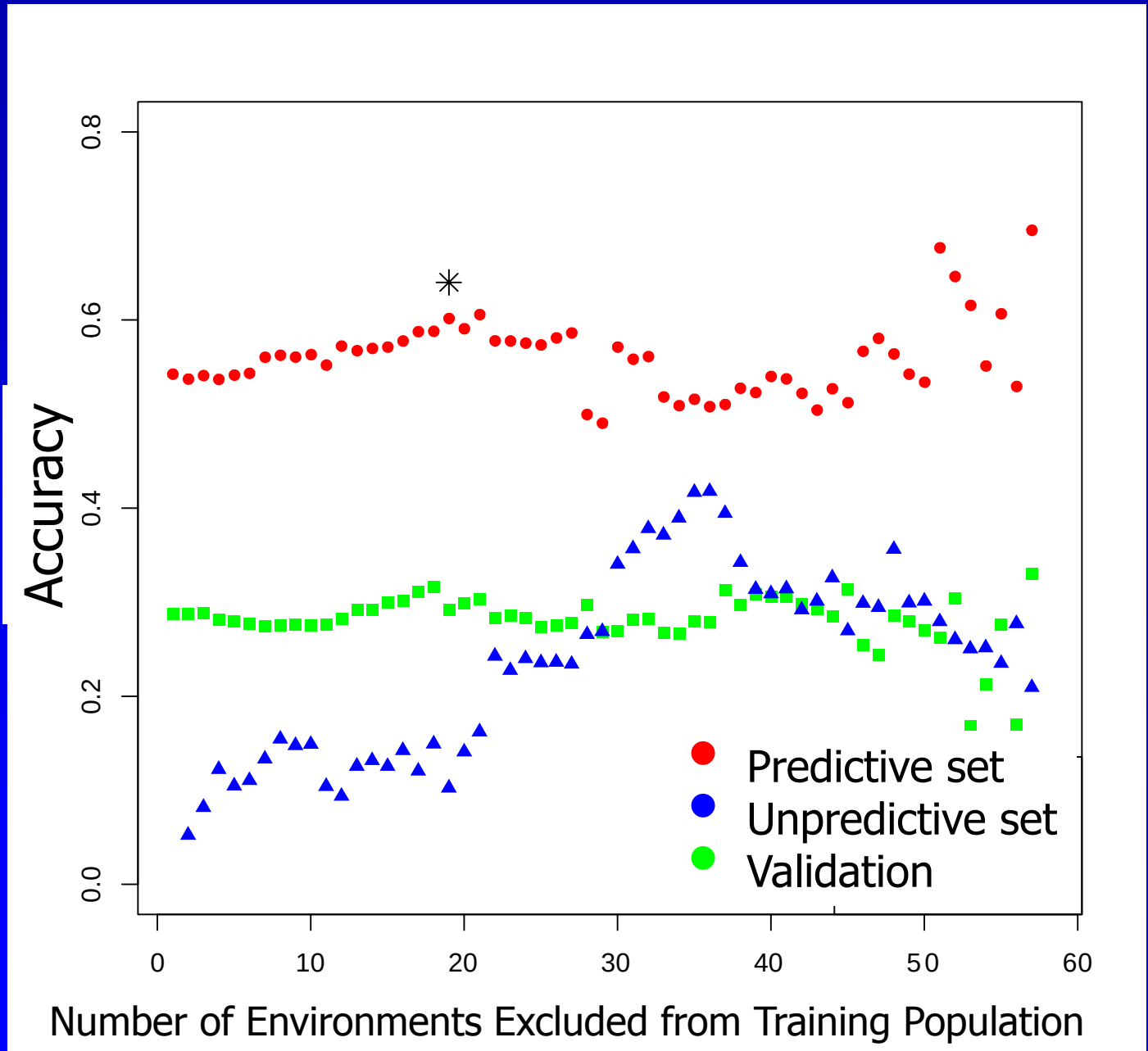
Remove least predictive environments one at a time, then cross validate

Prediction accuracy rose from 0.54 to 0.61 and no change in heritability

Some outlier environments were included in the optimal model

Accuracy in the validation set increased from 0.279 to 0.292

Only 1 of the 996 barley lines were excluded in the optimal model training set



Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E

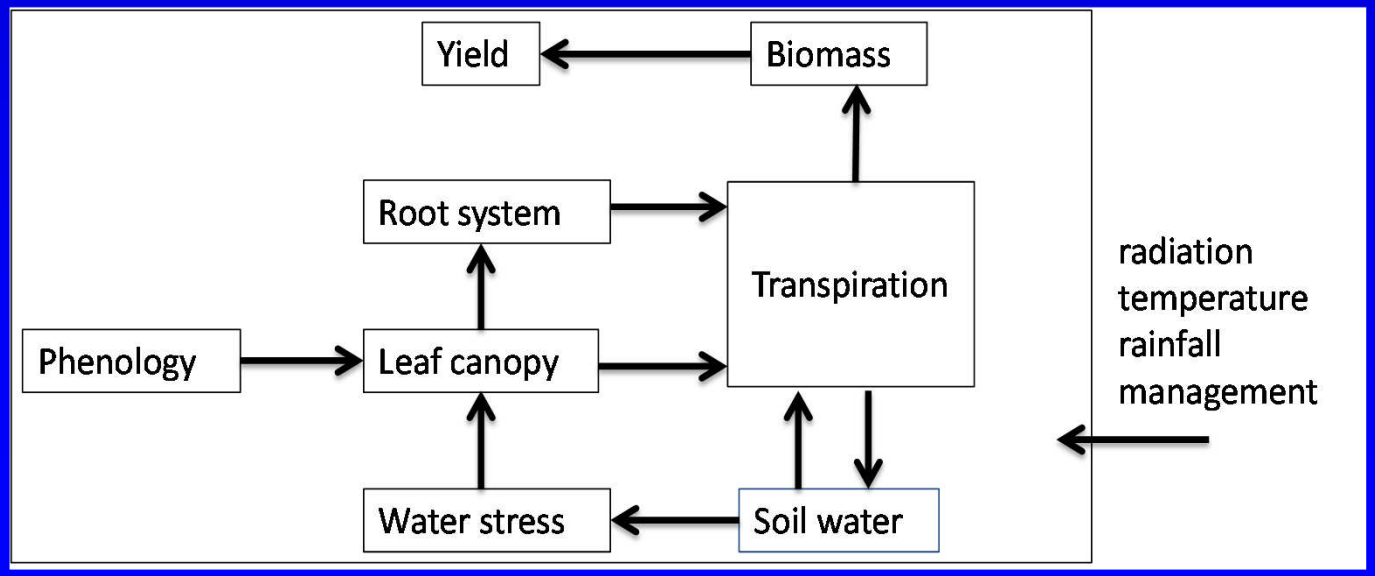
Heslot et al. 2013: *TAG 127:463-480*

Objectives:

- Integrate environmental data & crop modeling into the genomic selection framework to predict G*E and increase prediction accuracy.
- Model genome-wide markers and their differential response to the environment to better understand the genetic architecture of G*E.

Winter Wheat Trials from Limagrain Europe

- 2,437 elite inbred lines
- 44 environments over 6 years in France
- 9024 mean grain yields
- 1,287 SNP markers
- Daily weather data (AGRI4CAST)
- Wheat Crop Model (SirusQuality)

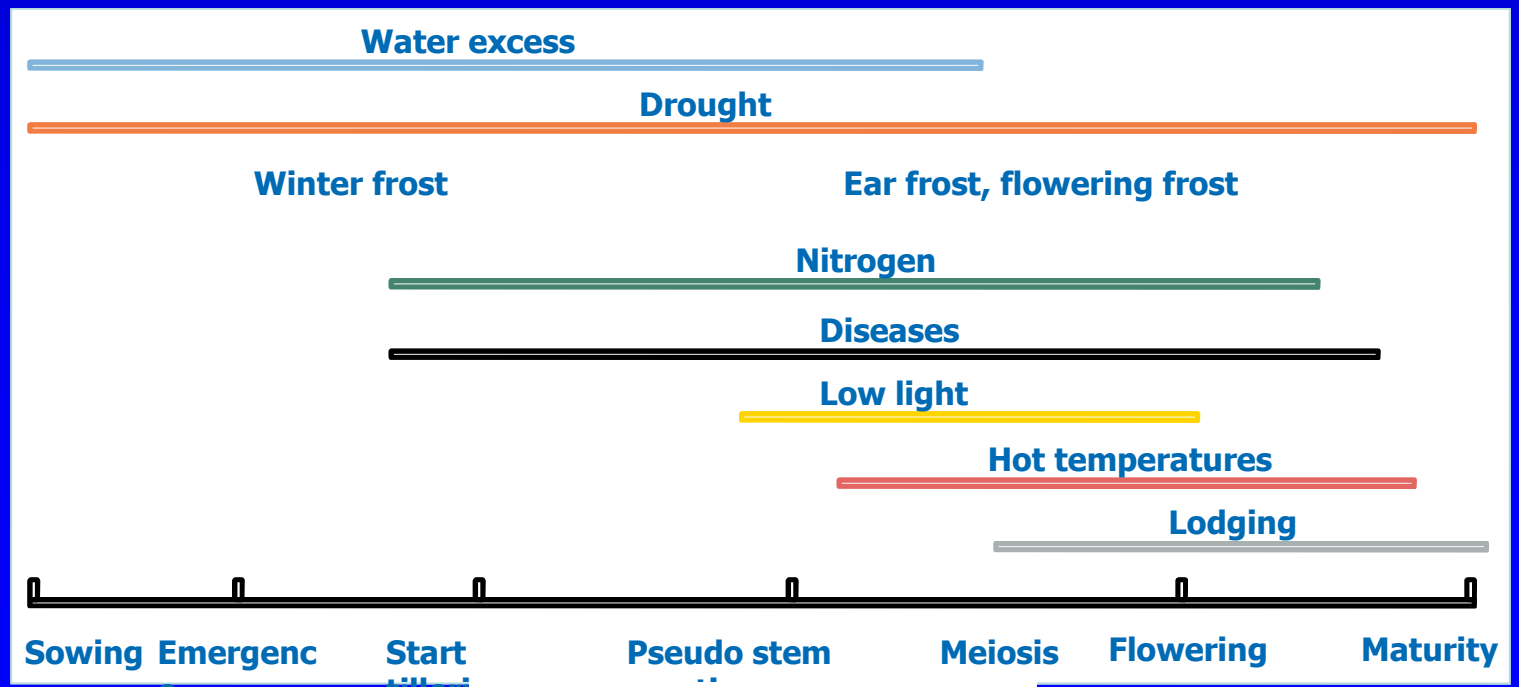


Typical structure of a crop model

Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E

- Physiological integration of the environment data:
  - Use a crop model (SiriusQuality) to compute the phenology and sync early, medium, and late maturing genotypes to the weather data
  - Derive stress covariates by developmental stage by using knowledge about sensitivity of specific growth stages to abiotic stresses (101 Stress Covariates)
  - Stress covariates were used as independent variables in statistical genetic models for effect estimation and prediction

Stresses



Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E

(Heslot et al. 2013: *TAG 127:463-480*)

- Integration in a GS prediction model
  - The factorial regression model was extended to the genomic selection context
  - For each marker, fit a main effect and a sensitivity to each of the stress covariates
  - Machine learning "soft rule fit" was used to capture the interaction between markers and stress covariates and non-linearity
  - Performance is predicted as main effect + G*E deviation

$$y_{ij} = \mu + \beta_j + u_i + \gamma_i f\left(x_j\right) + \varepsilon_{ij}$$

Intercept

Environmental effect

Genotype main effect

Sensitivity of each genotype to a stress covariate x, that can be transformed by a function f()

Residuals

Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E

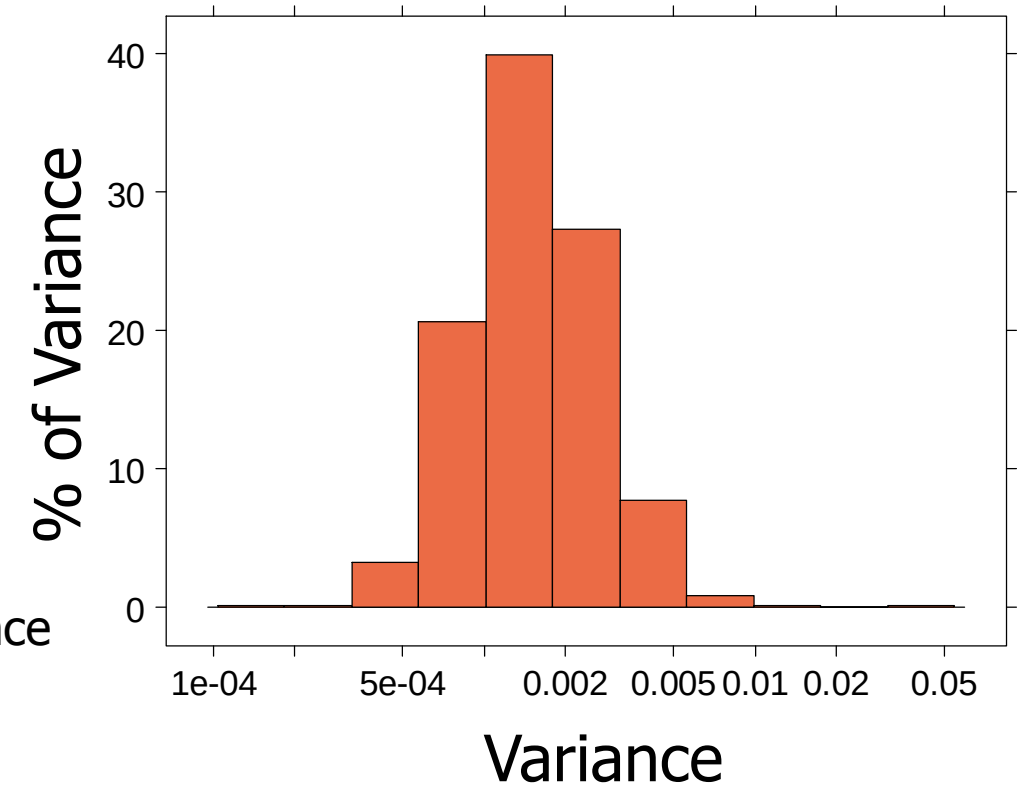
Dimensionality reduction

Nmarkers * Ncovariate predictors in the simplest case

Not all markers are highly variable across environments

Preselection of markers based on their variance

Ppd-D1 had the highest variance but alone did not capture a significant part of the G*E variance



Optimal model based on cross-validation used 250 markers + soft rule fit level 4

Significant Stress Covariates

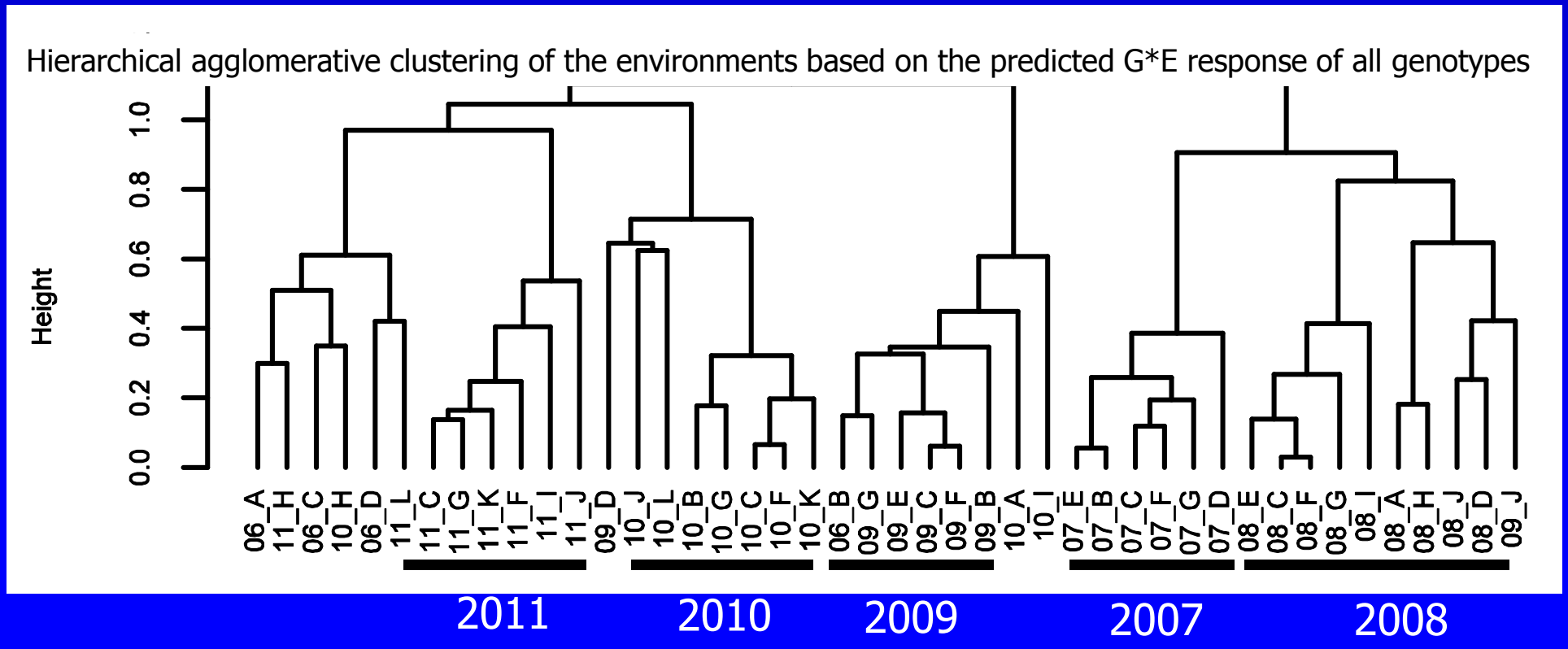
Stress Covariate	Stress Type	Importance
stmpmf	Sum average daily temperature meiosis to flowering	100
ntddr	Drought early spring–Total no. dry days to 350dd	89
spetpe1	Drought early spring - Sum precipitation, ETP	85
st25ef	Heat stress before flowering-Sum max temp>25°	80
nd25ef	Heat stress before flowering-No. days max>25°	78
st25fh	Heat stress early grain fill - Sum max temp>25°	76
nd25fh	Heat stress early grain filling-No. days max>25°	72
latitude	North/South trend	30

Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E

Environment clustering based on predicted G*E

- G*E can be predicted for any genotype in any environment
- We can use the balanced data set of predicted G*E's to cluster environments and investigate the structure of the TPE.

Environments were mainly grouped by year but also showed a North/South trend.



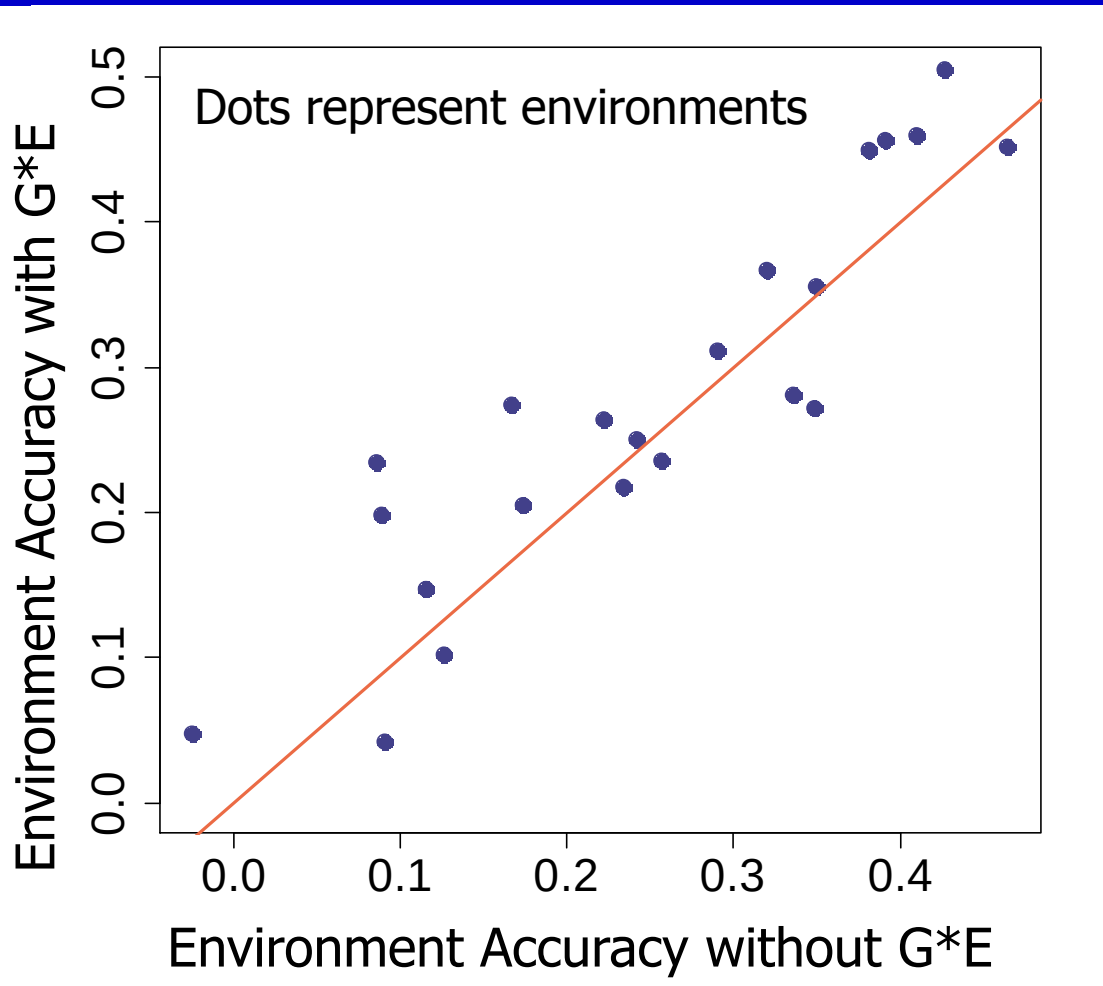
Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E

(Heslot et al. 2013: *TAG 127:463-480*)

- A dot above the line indicates higher accuracy with the model including G*E
- 11.1% increase in mean accuracy (P-value 0.02)
- 10.8% decrease in accuracy CV

Benefits of using crop modeling:

- Leverage agronomy knowledge
- Reduce dimensionality & non linearity
- Enables use of existing breeding data
- Interpretable model that can identify specific environmental stresses



Optimizing Training Population in the Presence of Population Structure

(Isidro et al. 2014 TAG. DOI: 10.1007/s00122-014-2418-4)

Five methods for choosing genotypes in sub-populations to optimize the training population:

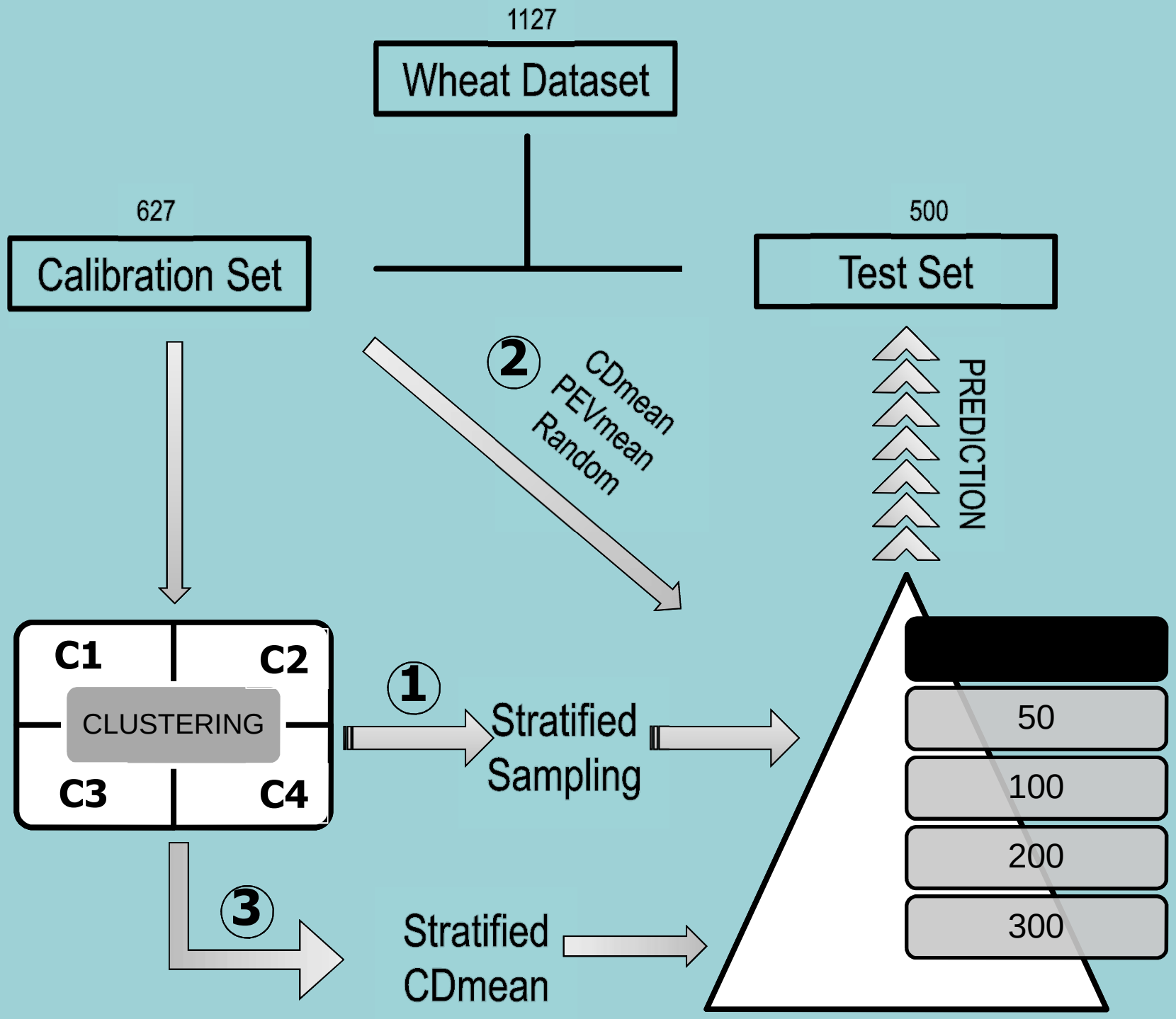
- Random Sampling
- Stratified Sampling
- Coefficient of determination mean (CDmean) (Rincent et al. 2012)
- Prediction Error Variance mean (PEVmean) (Rincent et al. 2012)
- Stratified Coefficient of Determination (StratCDmean)

WHEAT		RICE	
Population Size	1127	405	
Markers	38,893 GBS	36901 SNPs	
Subpopulation	4	3	
Environments	3	2	
Years	6	2	
Trait	h^2	Trait	h^2
YLD	0.79	FP	0.78
TWT	0.92	FT	0.85
LODG	0.78	PH	0.89
HD	0.94	PC	0.70
HT	0.95		

Statistical Model

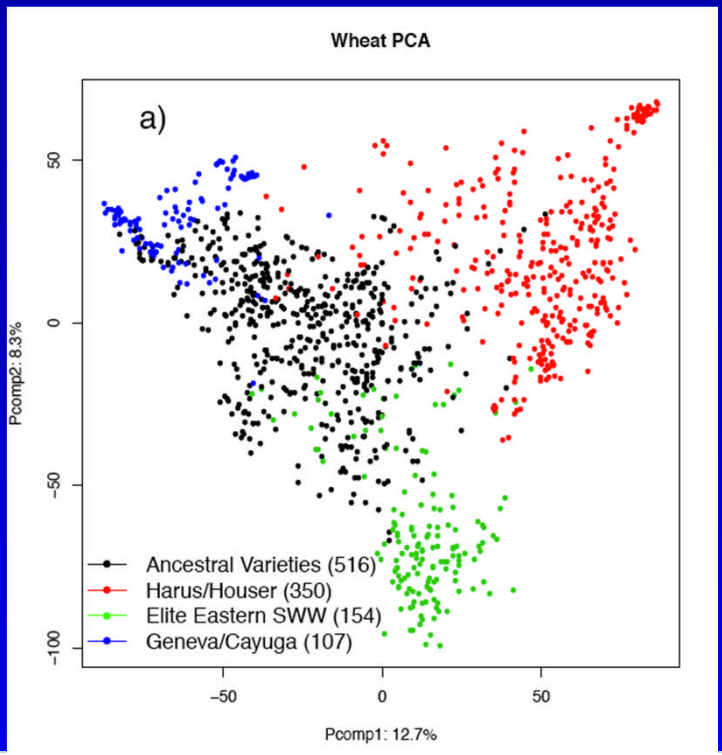
GBLUP(Genomic Best Linear Unbiased Predictor)
Same variance for all loci
Traits determined by a large number of gene

Ward's hierarchical clustering criterion


1127
Wheat Dataset
627
Calibration Set
500
Test Set
2
CDmean
PEVmean
Random
PREDICTION
C1
C2
CLUSTERING
C3
C4
1
Stratified
Sampling
3
Stratified
CDmean
50
100
200
300

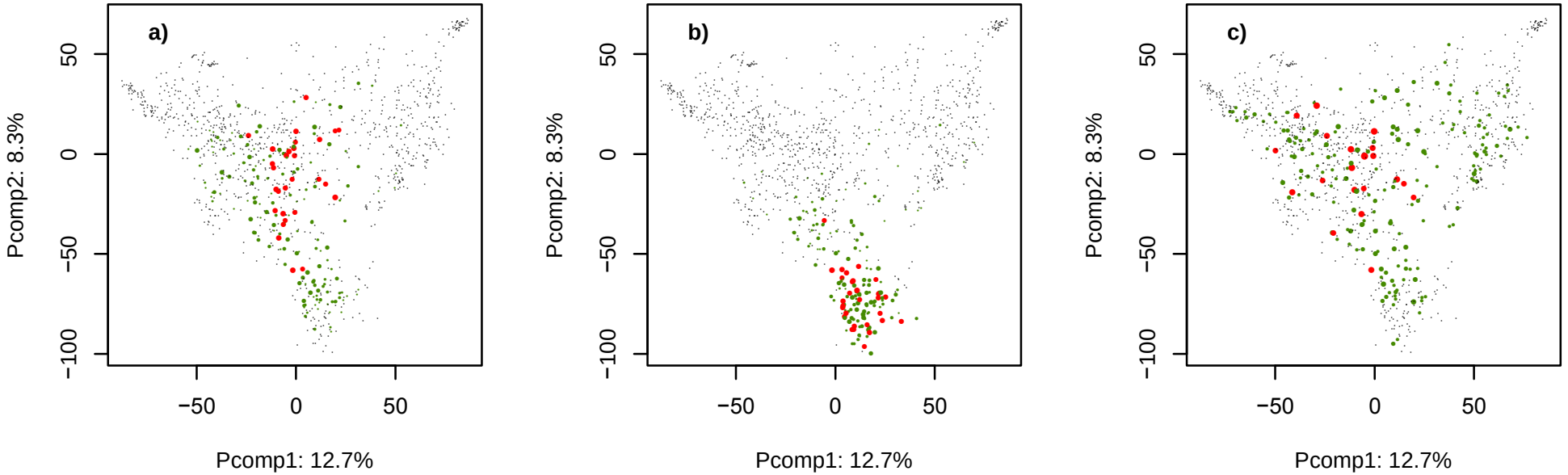
Principle Components Analysis for a Wheat Breeding Population

Mild Population Structure



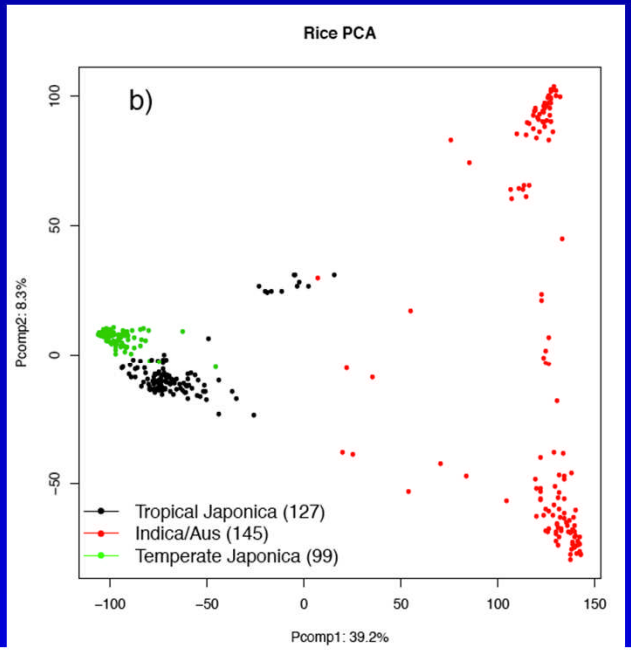
Wheat CDmean | **Wheat PEVmean** | **Wheat StratCDmean**

Green dots - genotypes selected by the algorithm over 50 runs. *Red dots* - genotypes that were selected more than 15 times.

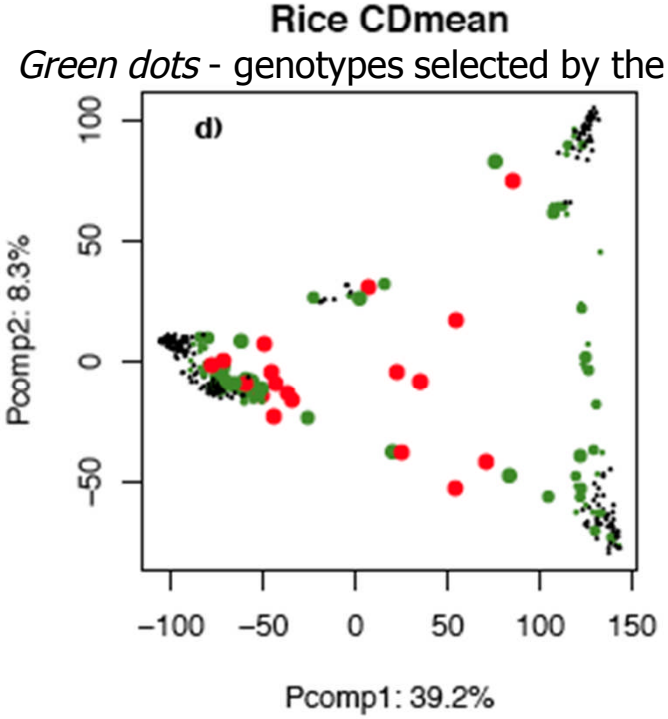


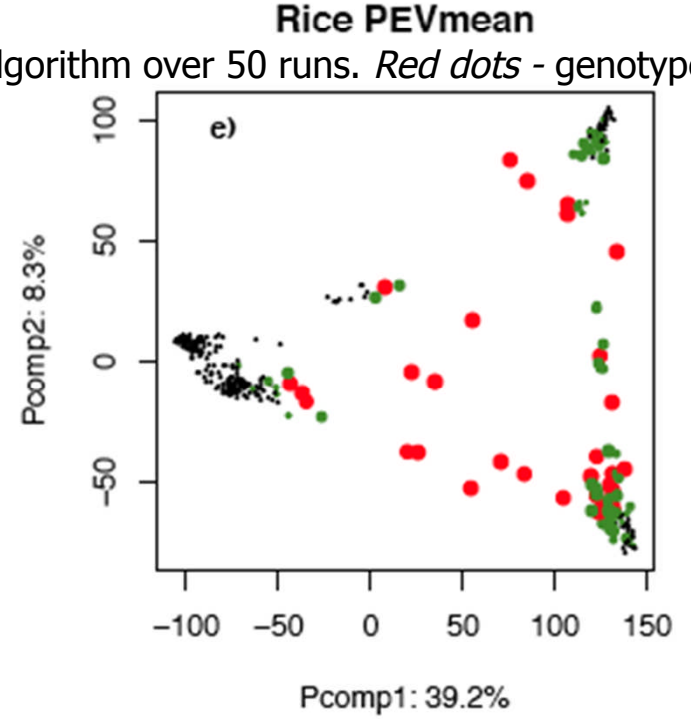
Principle Components Analysis for a Rice Global Diversity Panel

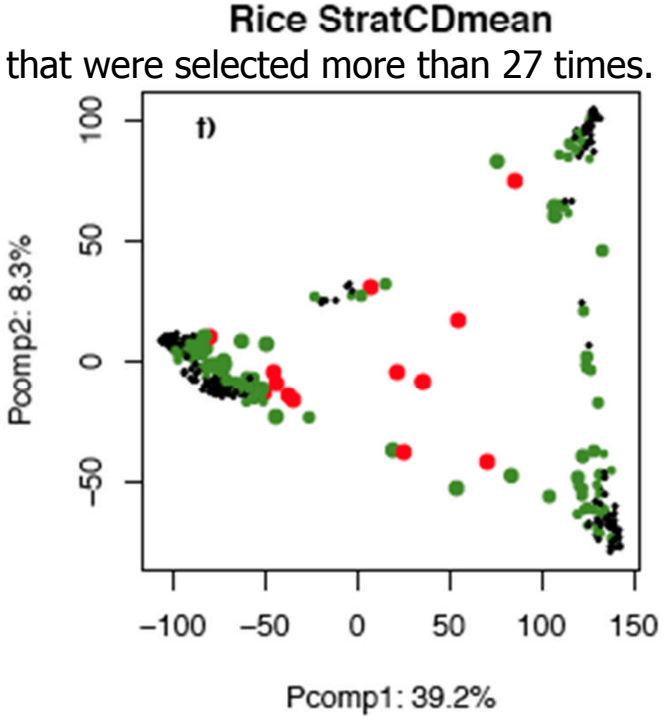
Strong Population Structure



Green dots - genotypes selected by the algorithm over 50 runs. *Red dots* - genotypes that were selected more than 27 times.



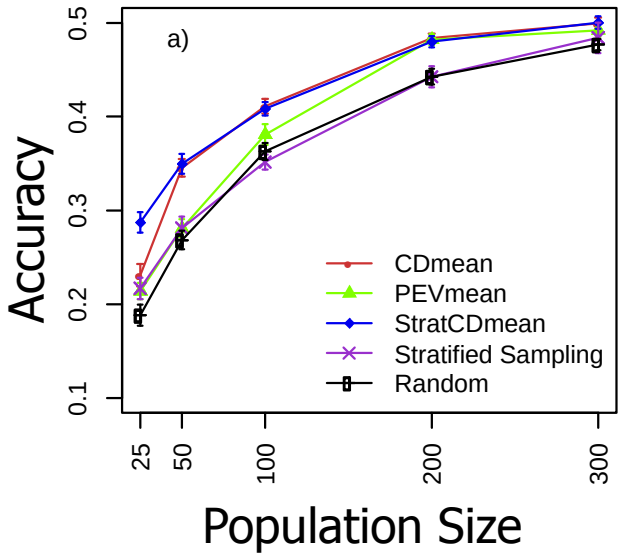




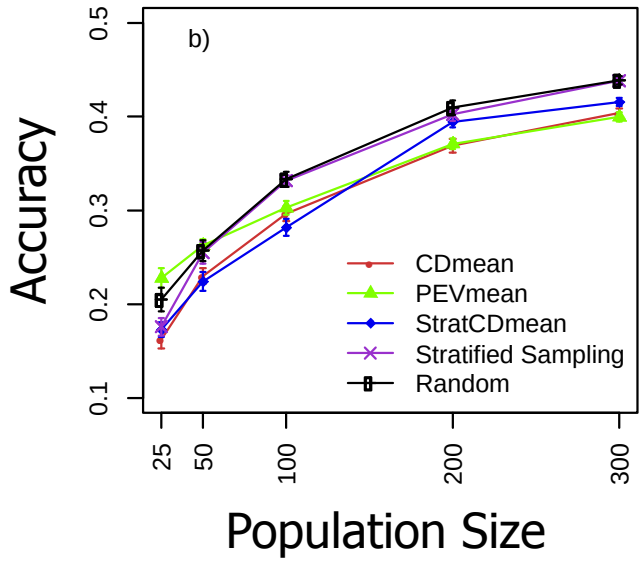
Prediction Accuracies for Training Populations Selected Using Different Criteria

Wheat

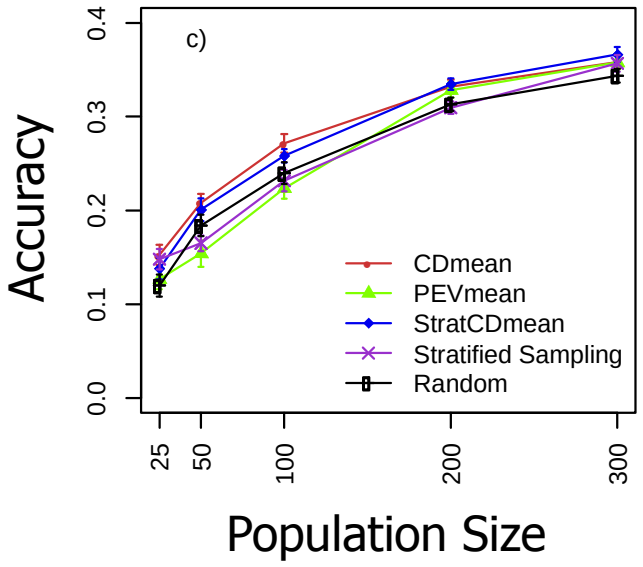
Yield



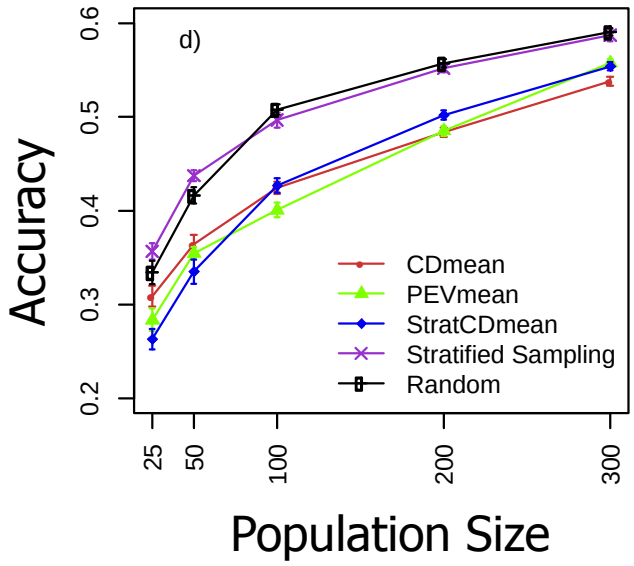
Test Weight



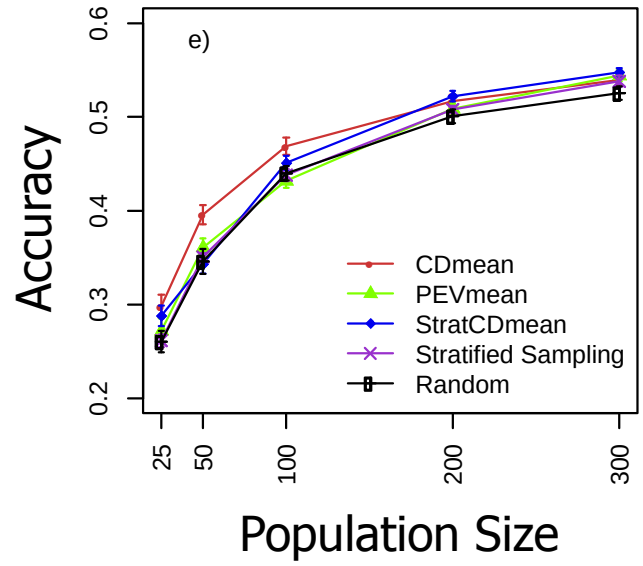
Lodging



Heading Date

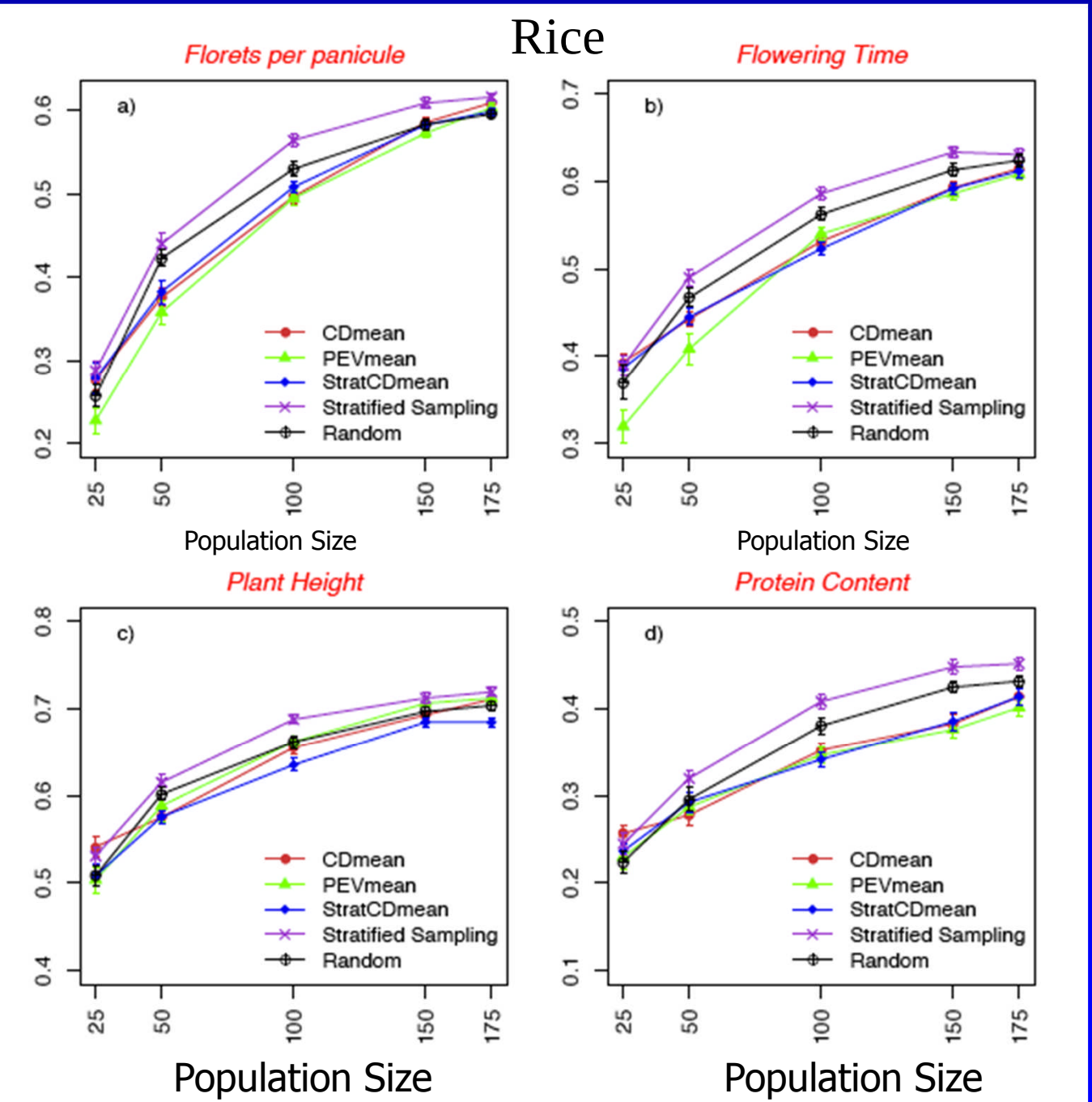


Plant Height



- CDmean and StratCDmean gave the highest accuracies except for TW and HD
- For TW and HD, Stratified and Random sampling produced the highest accuracies
- **Best criteria with mild population structure were CDmean and StratCDmean**

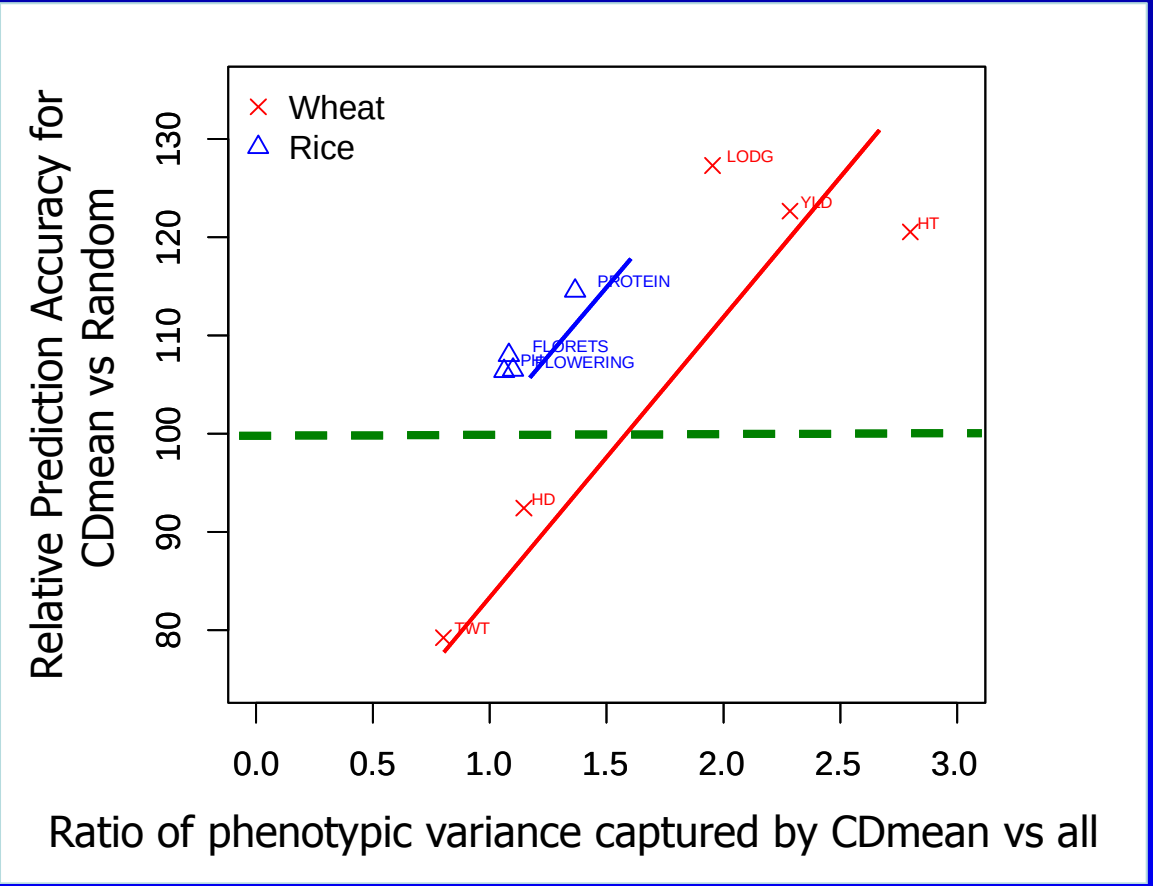
Prediction Accuracies for Training Populations Selected Using Different Criteria



- Stratified Sampling showed the highest accuracy
- At the smallest TRS size CD and StratCDmean showed the highest accuracies
- **With strong population structure Stratified Sampling was best**

Relative Prediction Accuracy vs Phenotypic Variation Captured using Cdmean

(Isidro et al. 2014 TAG. DOI: 10.1007/s00122-014-2418-4)



- The large genotypic variance obtained by using CDmean for genotypes does not always sample the largest phenotypic variance
- Highest accuracies were obtained with methods that captured more phenotypic variation

Biparental vs. Multi-Family GS

Biparental

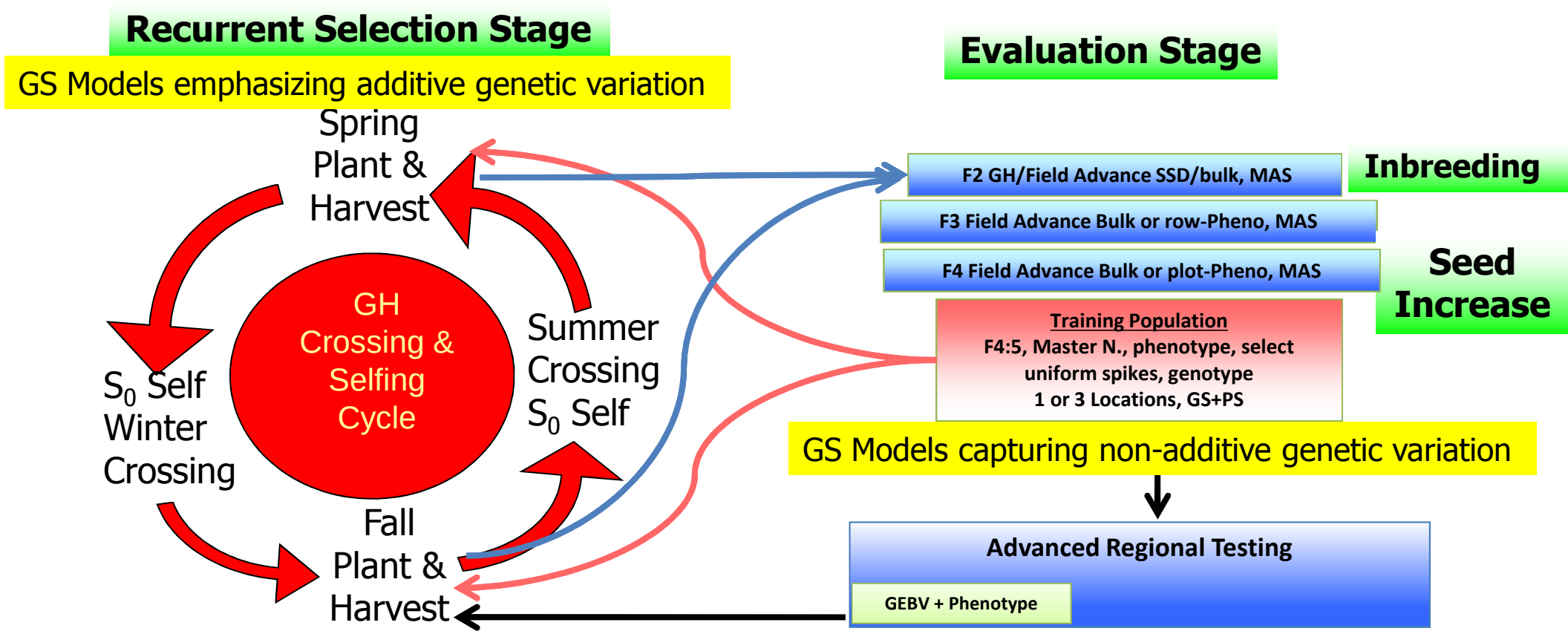
1) Population specific
2) Reduced epistasis
3) Reduced number of markers required
4) Smaller training populations required
5) Balanced allele frequencies
6) Best for introgression of exotic germplasm

Multi-Family

1) Allows prediction across a broader range of adapted germplasm
2) Allows sampling of more environments
3) Cycle duration is reduced because retraining model is on-going
4) Allows larger training populations
5) Greater genetic diversity



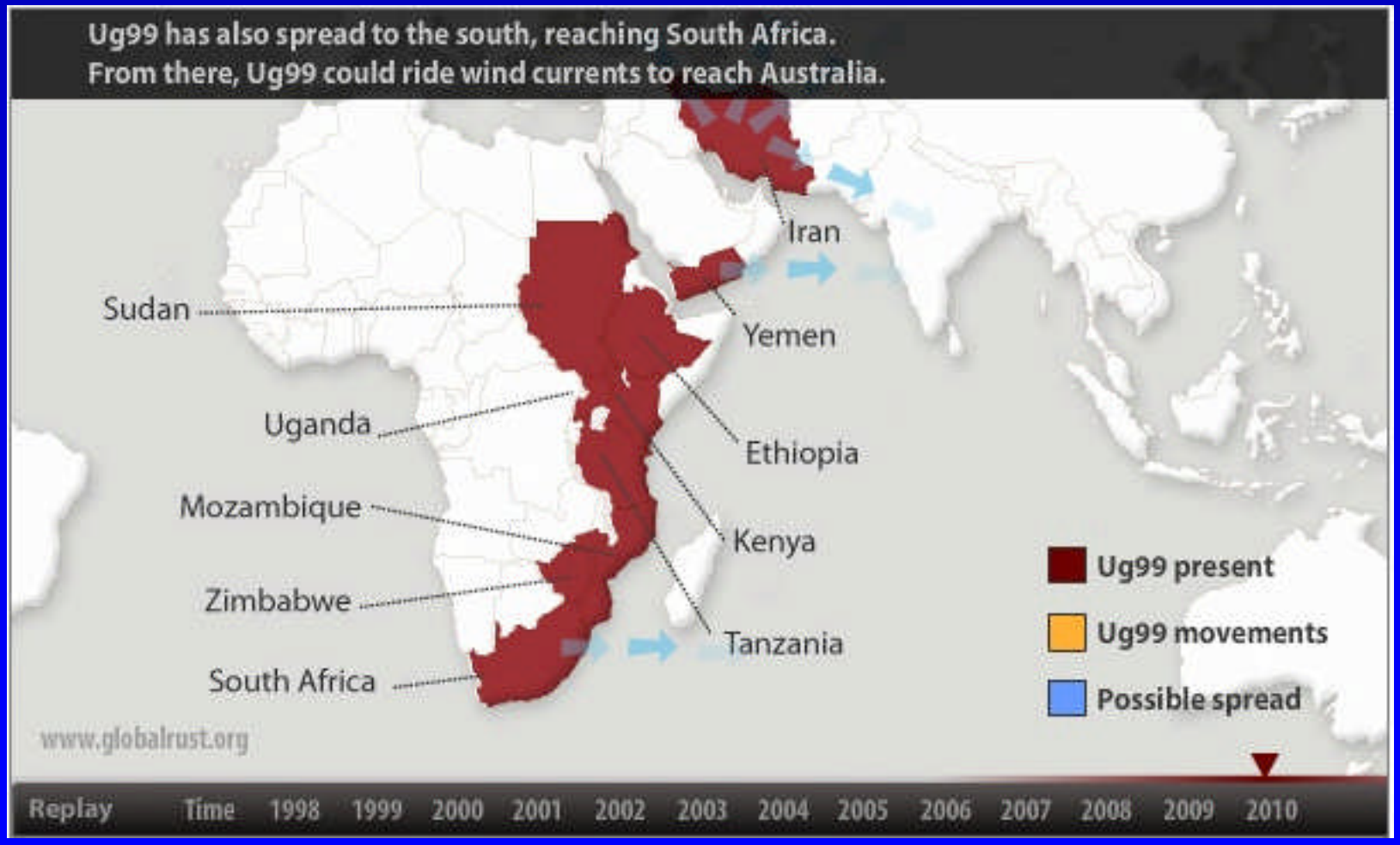
Integrating Genomic Selection in a Breeding Program



Planting GS-selected individuals (S_0), intermating, self-pollination and S_1 harvesting seed occur twice a year
F_2s can be field or GH planted depending on space available. MAS &/or GS can be applied
F_3s will begin seed increases of selected individuals in single rows
F_4s will be phenotyped and 50 F_5 spikes selected for uniformity and for GBS genotyping
Selected lines enter the Master Training Nursery
Each year selected lines are entered in the regional trials and/or recycled in the crossing program

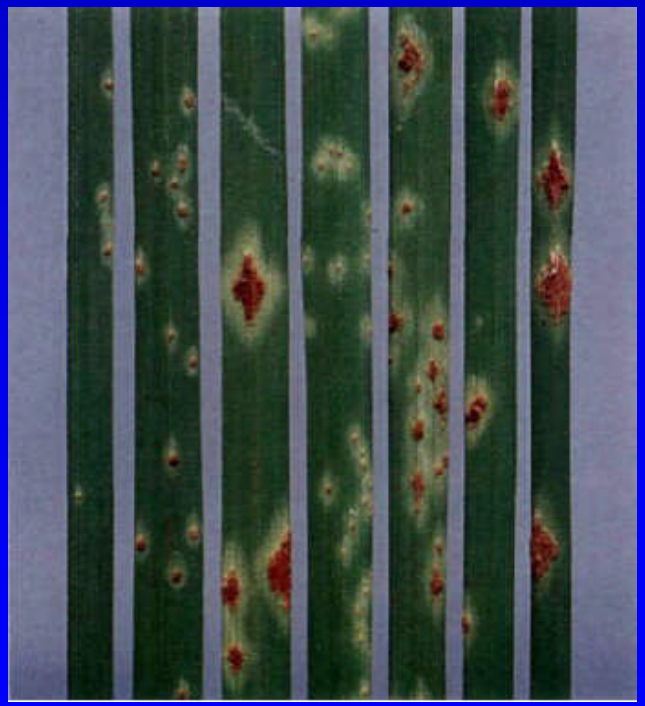
Durable Rust Resistance in Wheat
Funded by the Bill & Melinda Gates Foundation

Since the 1999 discovery of virulence on *Sr31* in Uganda, Ug99 races have overcome *Sr36* and *Sr24* and spread north to Iran and south to South Africa



Stem Rust Resistance- Two Types:

All-stage resistance- conferred by major R-genes that elicit a hypersensitive response



VS.

Adult plant resistance- conferred by additive loci that alone do not provide complete resistance



Validation of GS Models For Stem Rust APR Using Genotyping by Sequencing (GBS)

Jessica Rutkoski

Objectives:

1) Compare prediction accuracies of 4 different GS models:
   - Ridge Regression (RR)
   - Bayesian Lasso (BL)
   - RKHS
   - Random Forest (RF)

2) Evaluate different % missing data thresholds for filtering markers

3) Evaluate different methods for Imputation (inferring missing genotype data)

Materials and Methods

Phenotypic data:

370 CIMMYT breeding lines evaluated for stem rust APR at the Ug99 screening nursery in Njoro, Kenya and Ethiopia for 8 environments.

Genotypic data:

Genotyping by Sequencing (GBS) polymorphisms

Filtered using :

A) 20% missing data threshold -> 2,014 markers

B) 50% missing data threshold -> 10,461 markers

C) 70% missing data threshold -> 136,188 markers

Imputation Methods for **unordered** markers:

A) Substituting missing values with the mean allele value at a particular locus

B) K Nearest Neighbors (Troyanskya et al 2001)

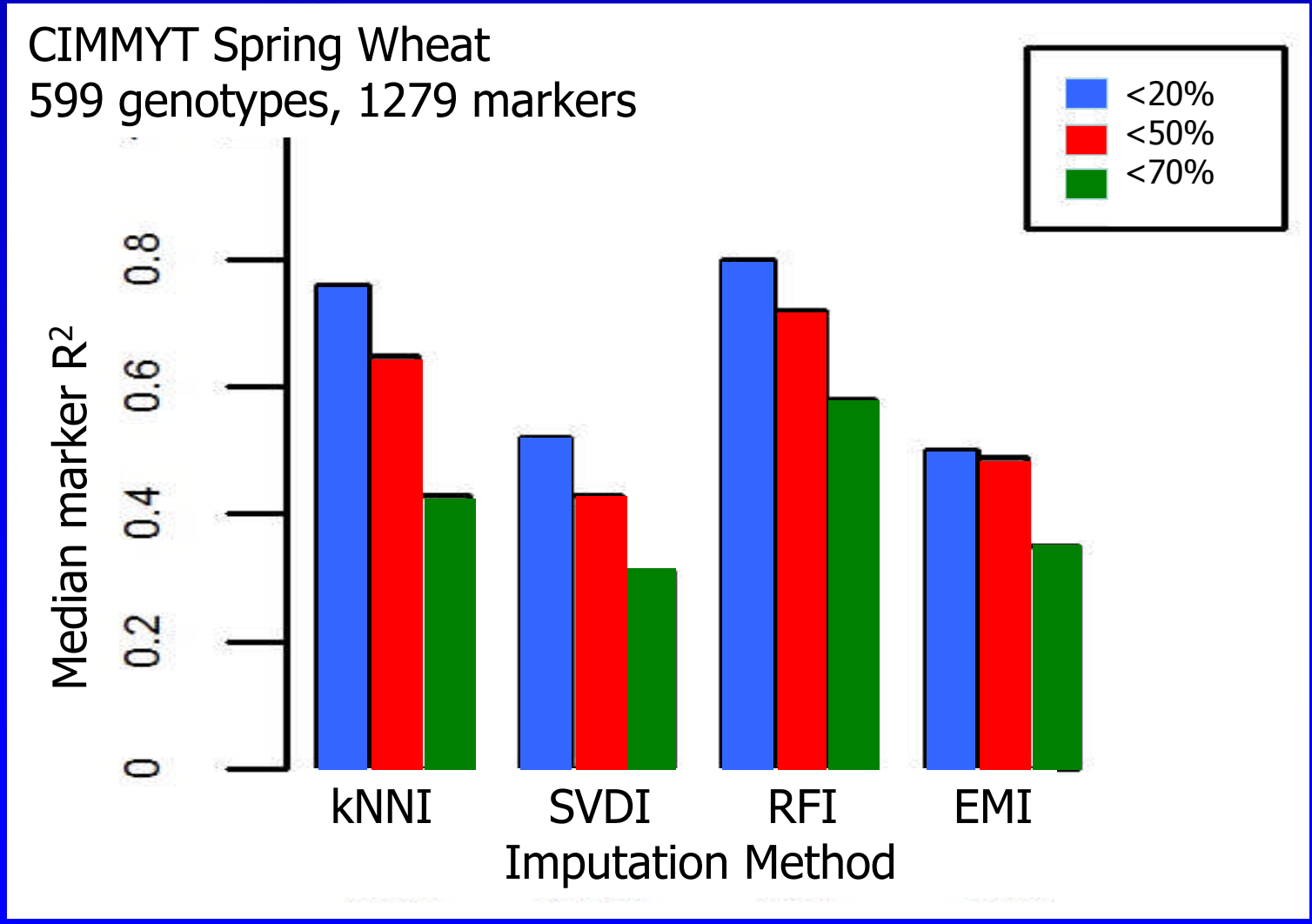
C) Singular value decomposition (Troyanskya et al 2001)

D) Expectation maximization (Dempster et al 1977)

E) Machine learning algorithm, Random Forest (RF)

Imputation Accuracy Using Four Different Imputation Methods (Rutkoski et al. 2014. G3)

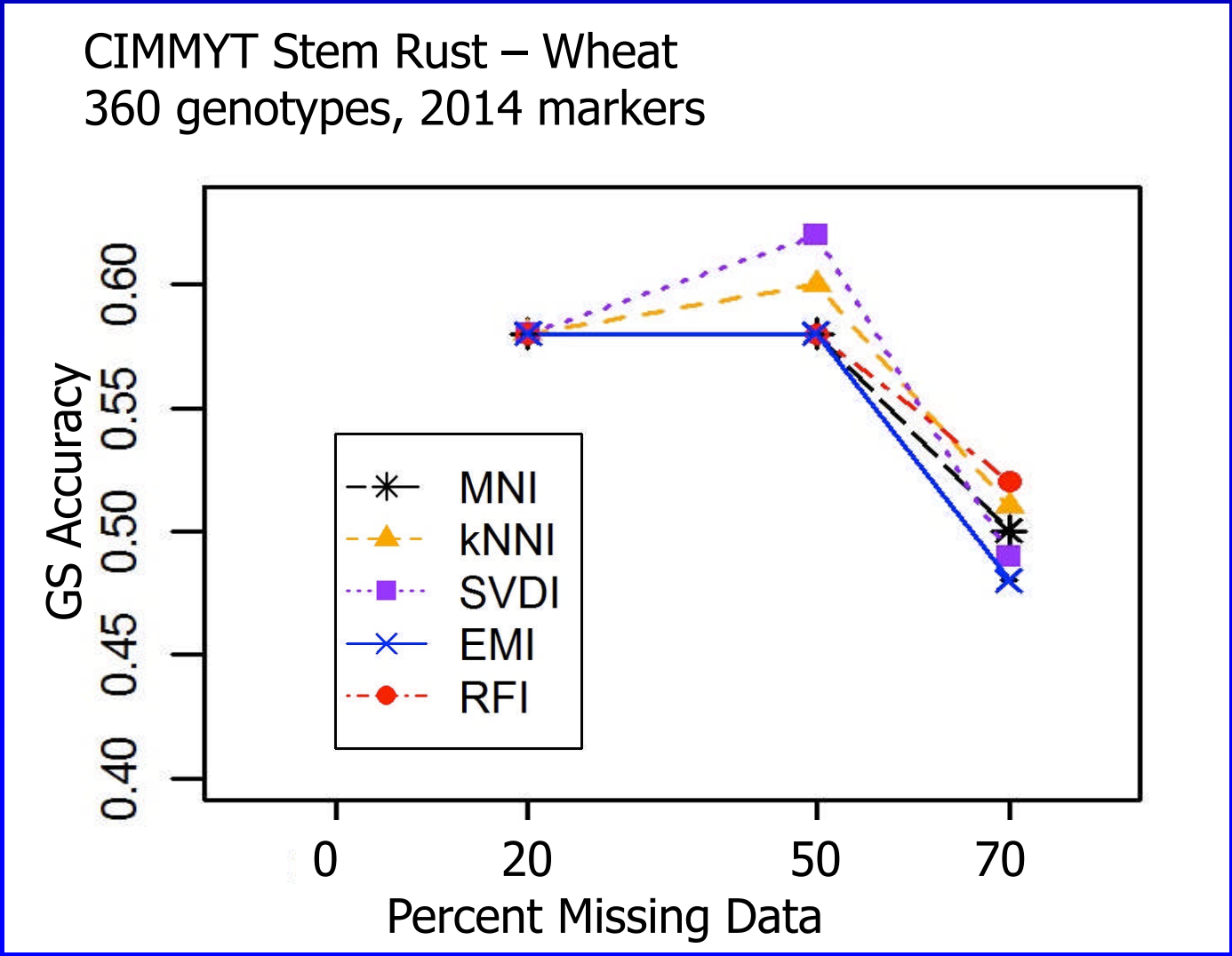
Each imputation method was compared using the same number of markers with <20, 50, or 70% missing data. Marker imputation accuracy was based on the R^2 between the predicted and original data points.



GS Accuracy Using GBS for Five Different Imputation Methods

(Rutkoski et al. Accepted G3)

Each imputation method was compared using the same number of markers with <20, 50, or 70% missing data.



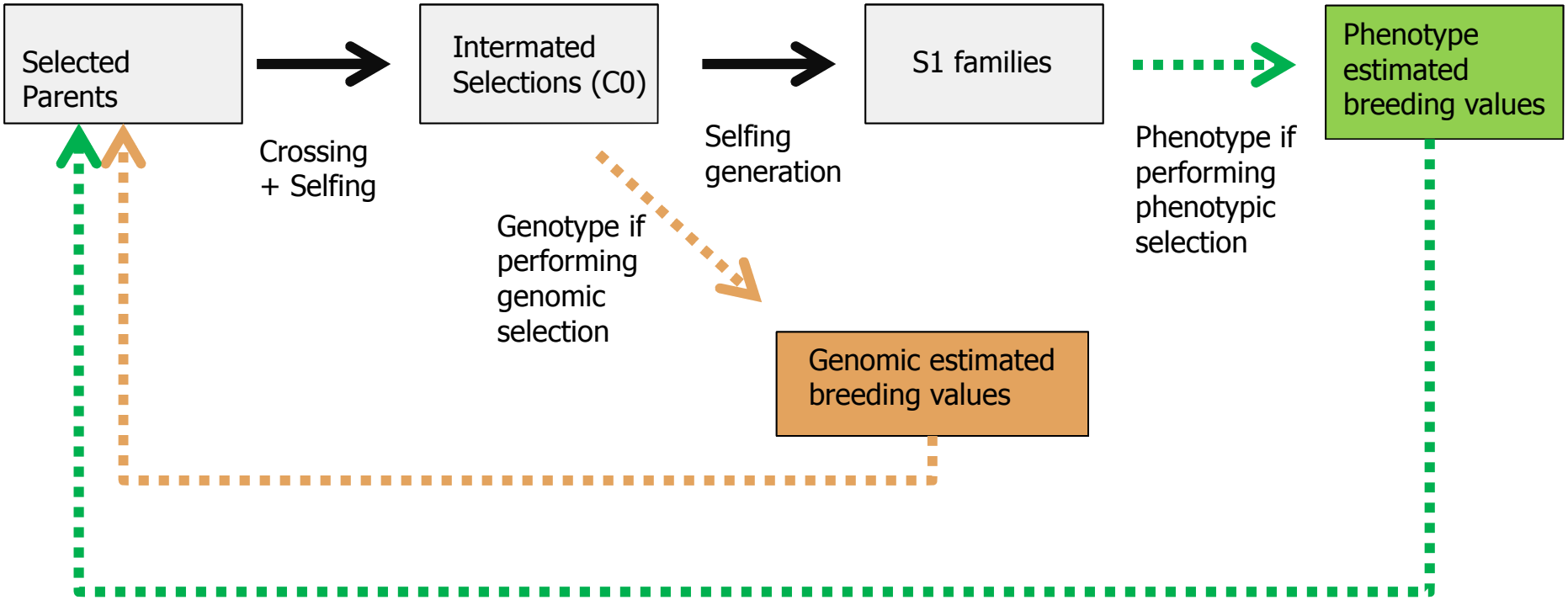
Methods produced similar GS accuracies

Up to 50% missing data could be imputed with no loss in GS accuracy

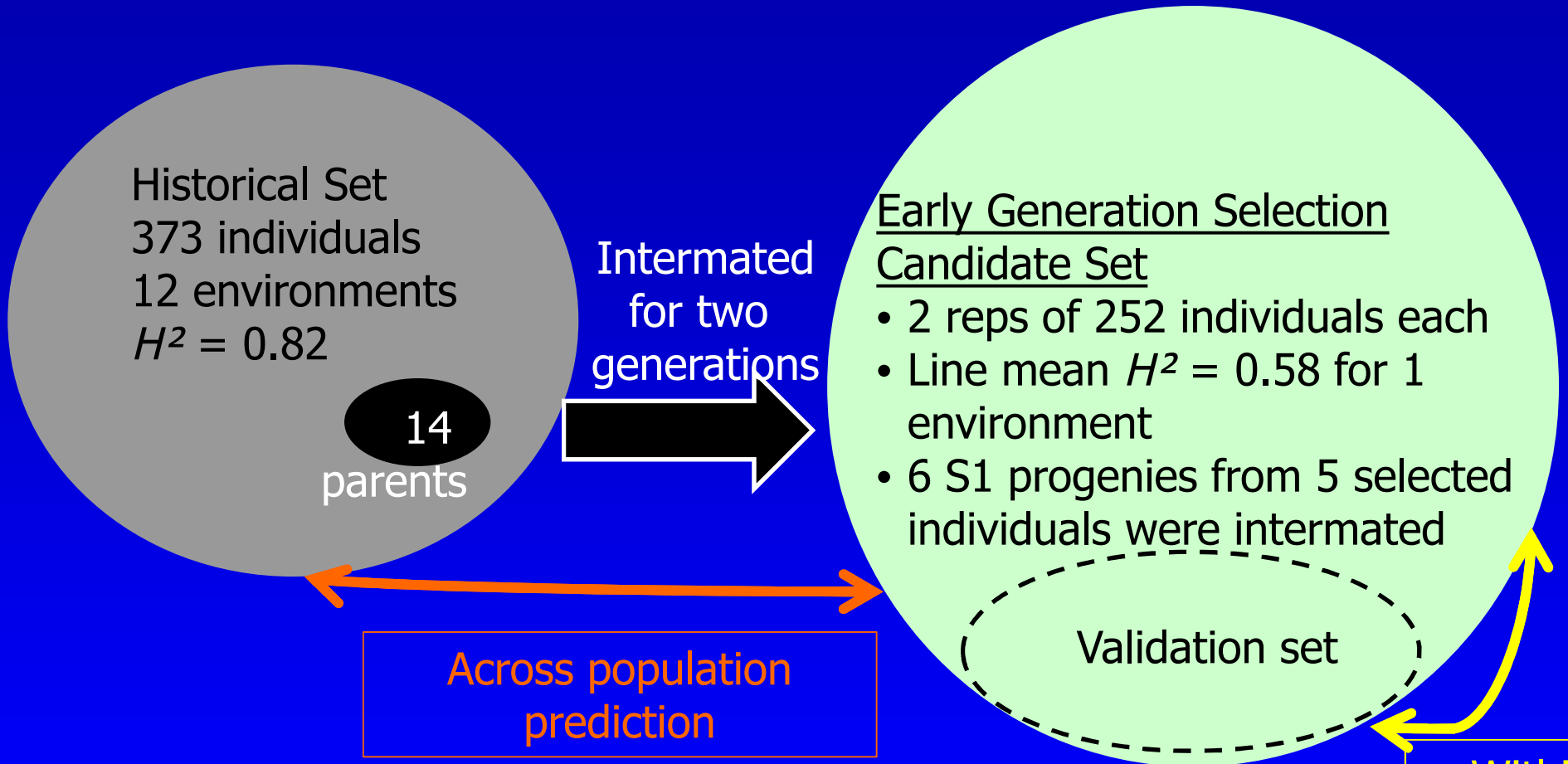
Implementation Strategy to Evaluate Realized Gain from GS for Adult Plant Stem Rust Resistance

(Rutkoski et al. 2015. The Plant Genome In press)

Recurrent Selection, Genomic vs. Phenotypic

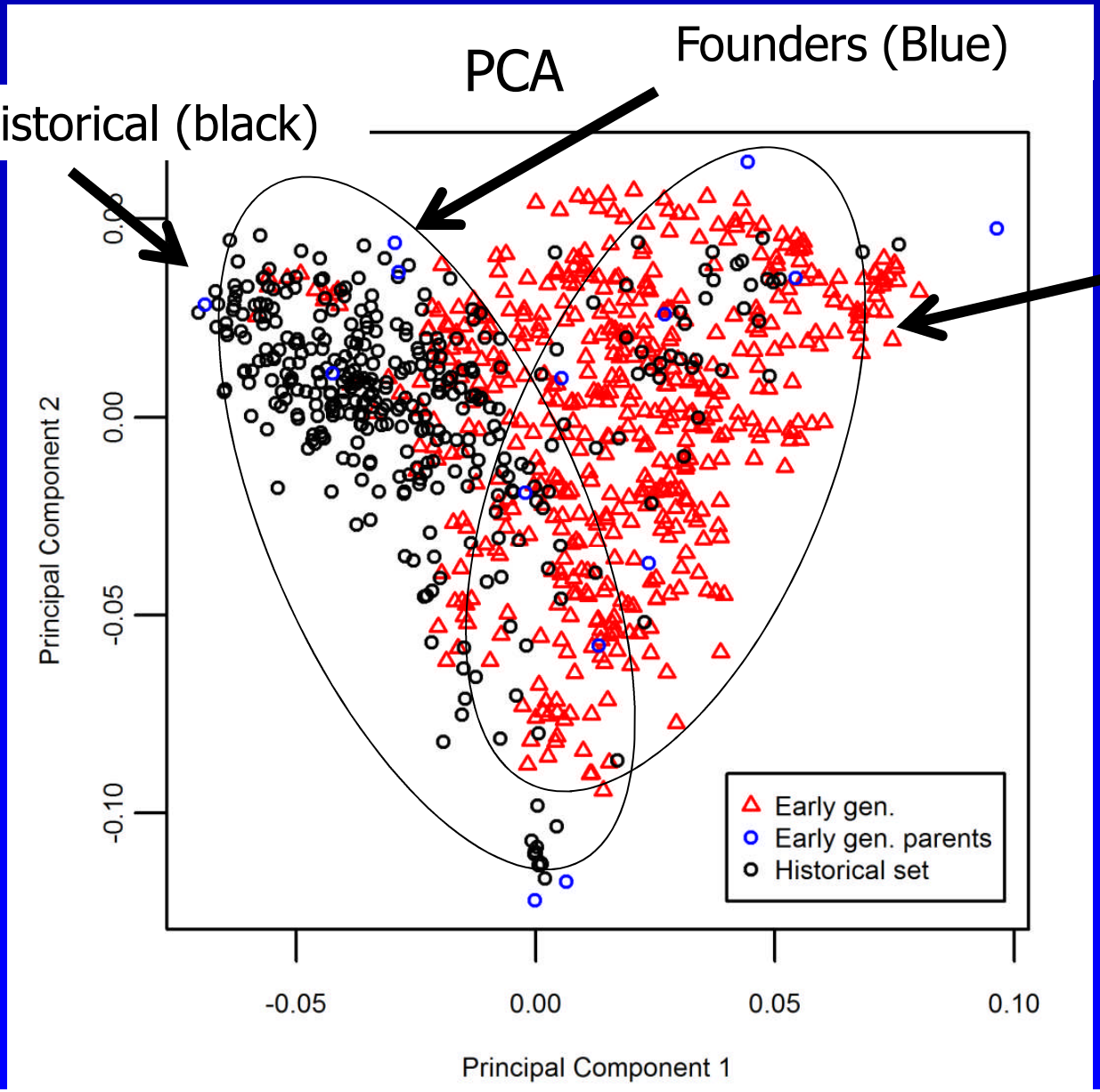


Populations Used for Recurrent Genomic Selection for Stem Rust Adult Plant Resistance (Rutkoski et al. 2015)



27,434 GBS markers filtered to 18,653 with <20% missing data

Relationship Between the Historical and the New Early Generation (EG) Selection Candidate Populations (Rutkoski et al.)



EGs have begun to diverge from the Historical set

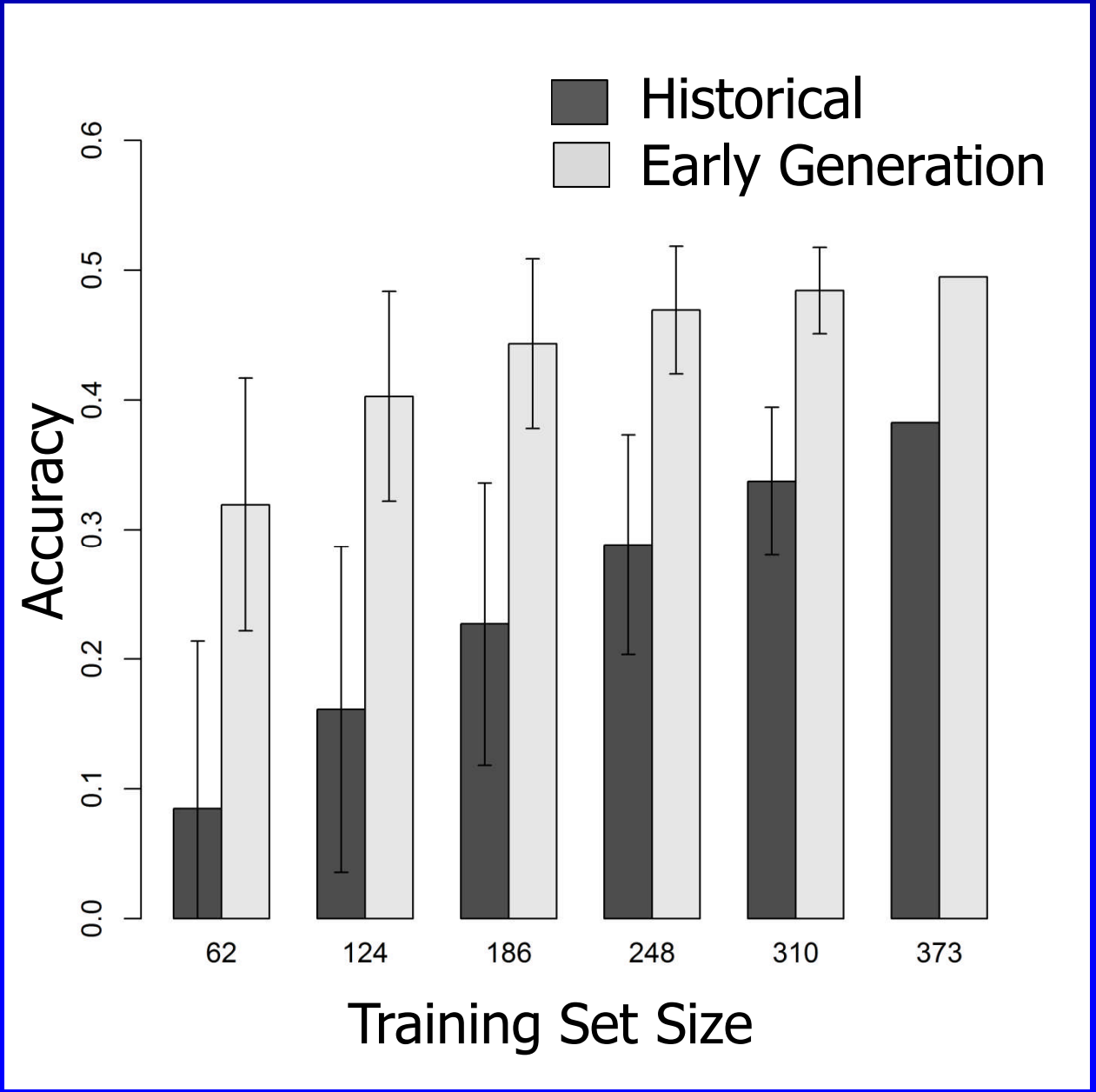
PC1 explains 7% of variance
PC2 explains 5% of variance

Accuracy using Training Sets of Historical Lines or EGs and Effect of Training Set Size

(Rutkoski et al. 2015 TAG In press.)

EGs had higher prediction accuracy because they are more closely related to the selection candidates

Accuracy for the EGs plateaus but not for the Historical Lines

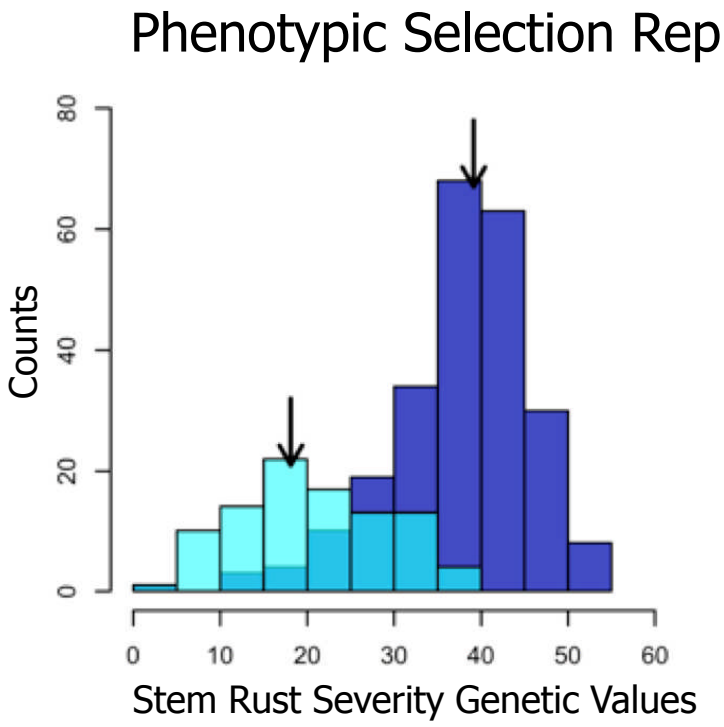


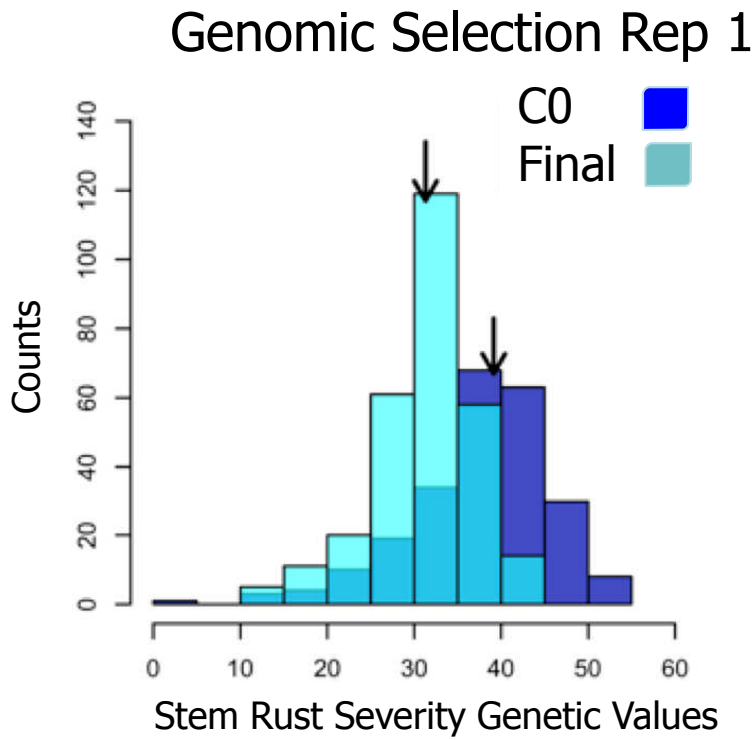
Genetic Values for Adult Plant Stem Rust Resistance Comparing C0 with the final populations

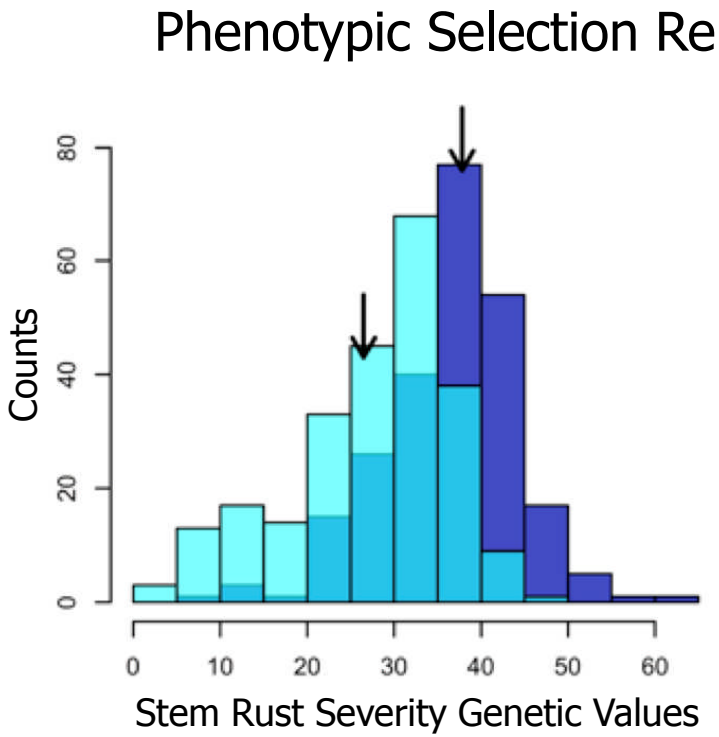
(Rutkoski et al. 2015. The Plant Genome)

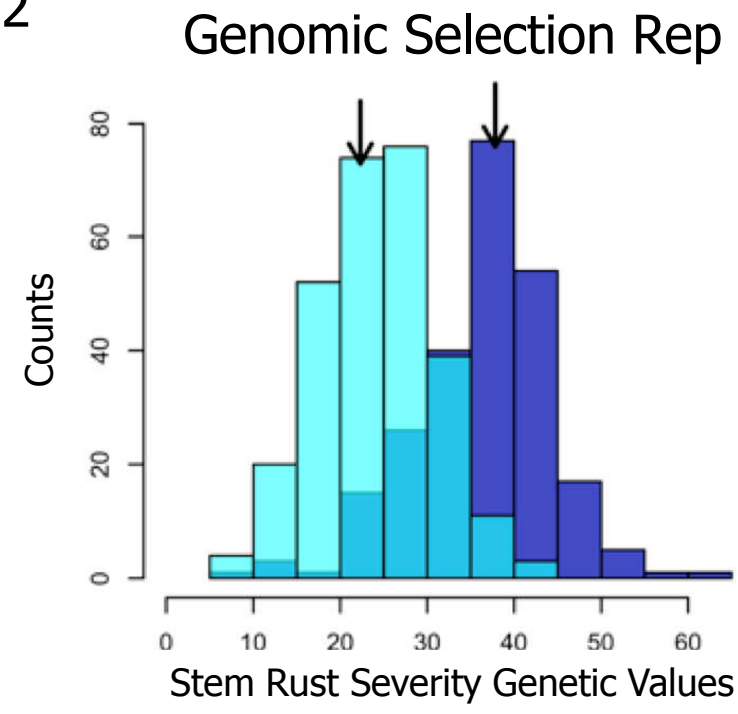
Rep 1 - PS realized gain was > GS realized gain (21 vs. 8%).

Rep 2 – GS realized gain was > PS realized gain (16 vs. 11%) .

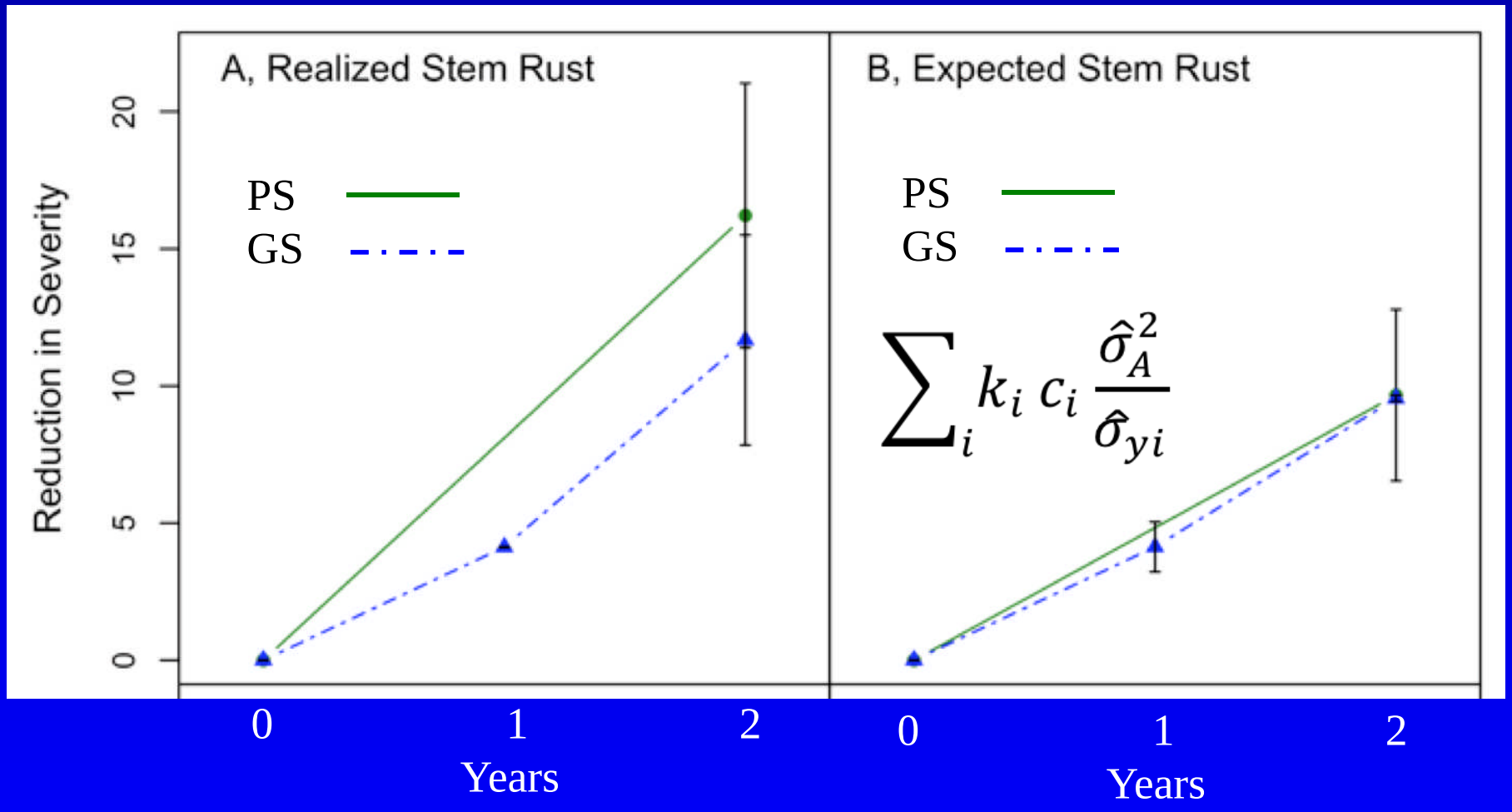






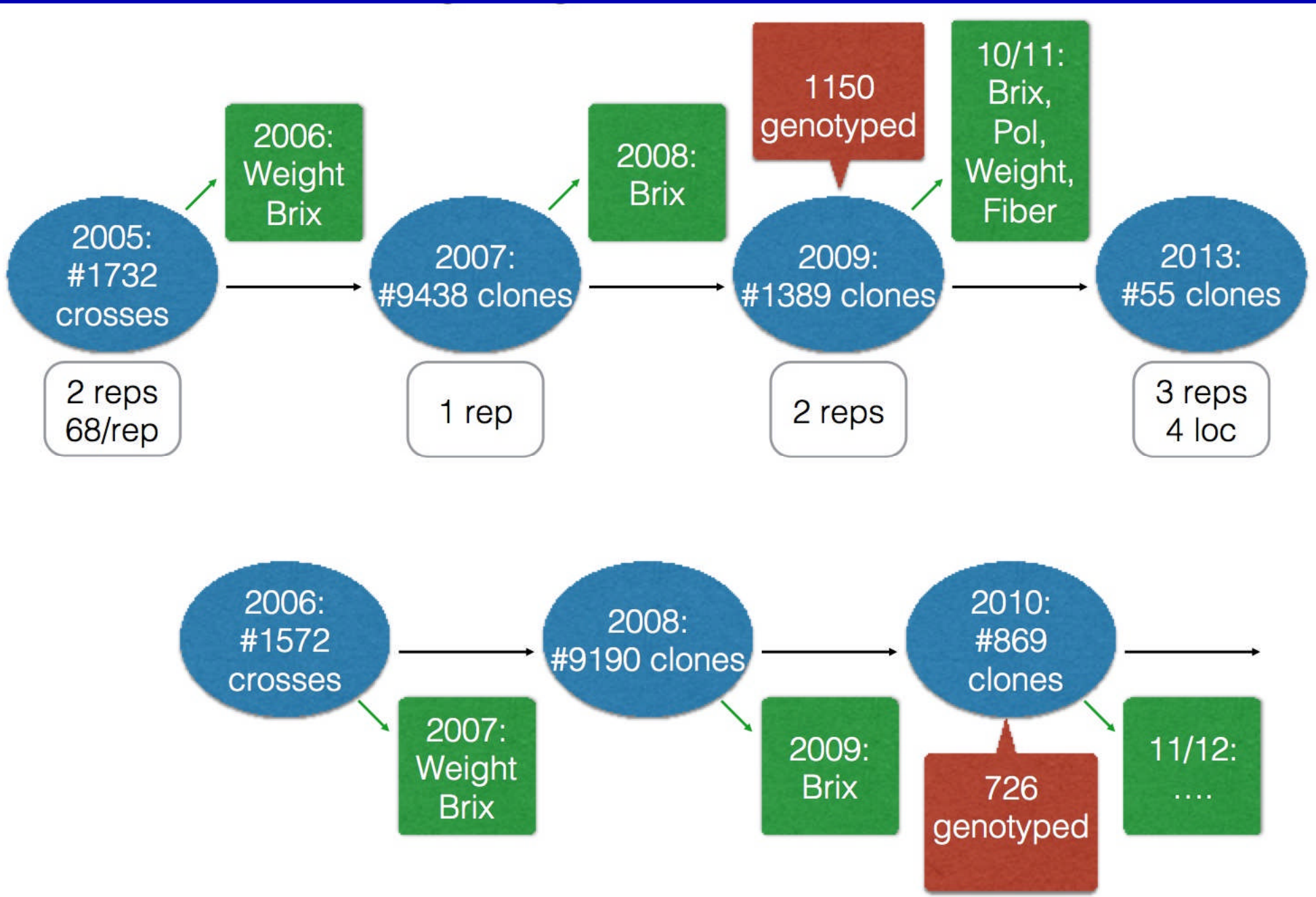


Realized and Expected Gain from Selection for Adult Plant Stem Rust Resistance (Rutkoski et al. 2015: In Press)



- Evaluated all populations in Kenya and Ethiopia in 2014 (r=0.66)
- Gain from one cycle of PS (42%) & two cycles of GS (31%) were not significantly different
- Second GS cycle increased gain because model was retrained on EG population
- Realized gain similar to expected gain for both PS and G

CTC Breeding Program
Itaraju Brum



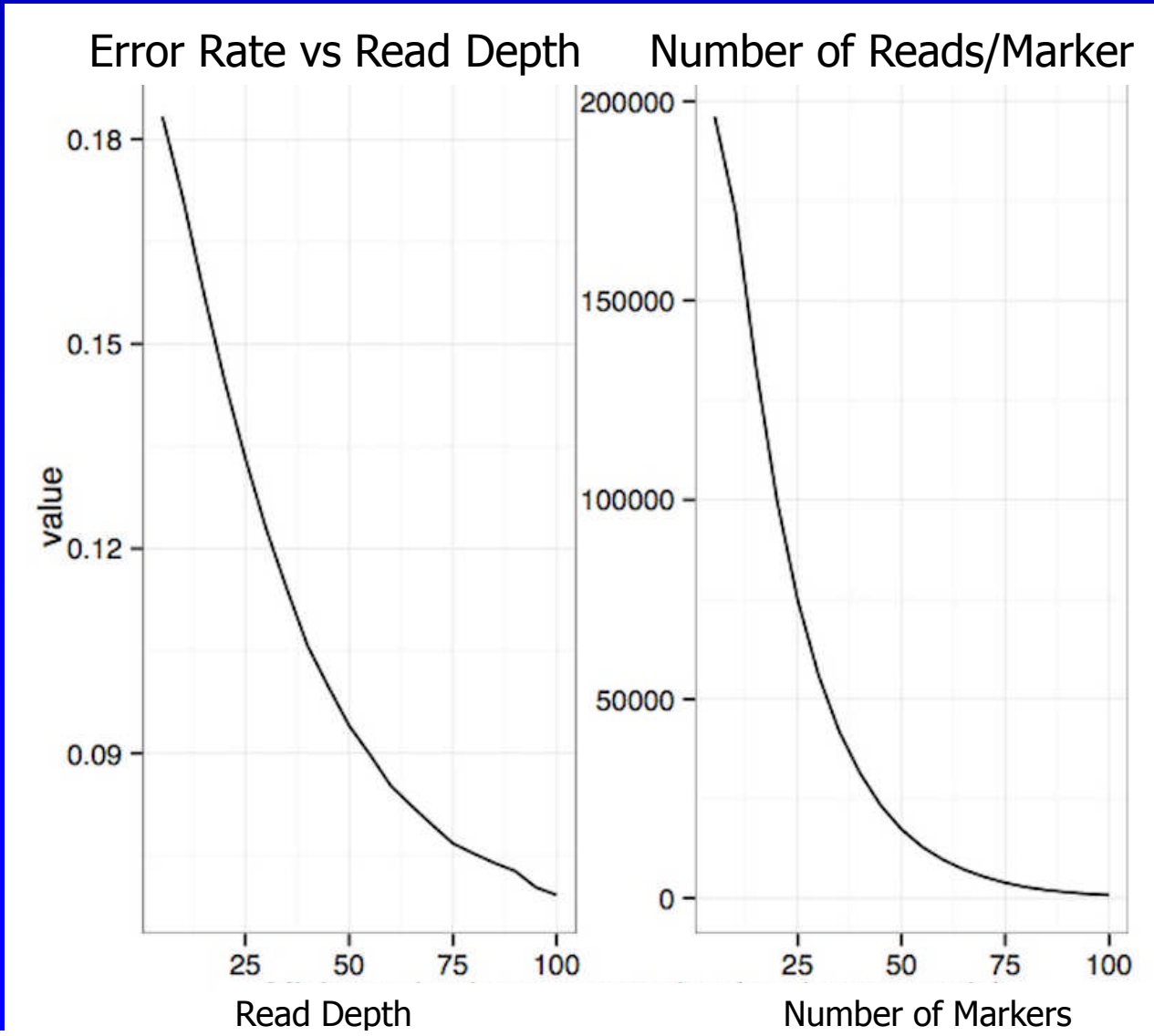
GBS Genotyping

Chance of not observing a single dose allele in a dodecaploid

- Raw data: 245,923 SNPs x 1,919 samples
- Raw-2 criteria: MAF<0.05, het < 0.05, markers missing < 75%
  196,302 SNPs x 1,850 Samples

Final set criteria: Read depth >= 40, markers missing < 25%

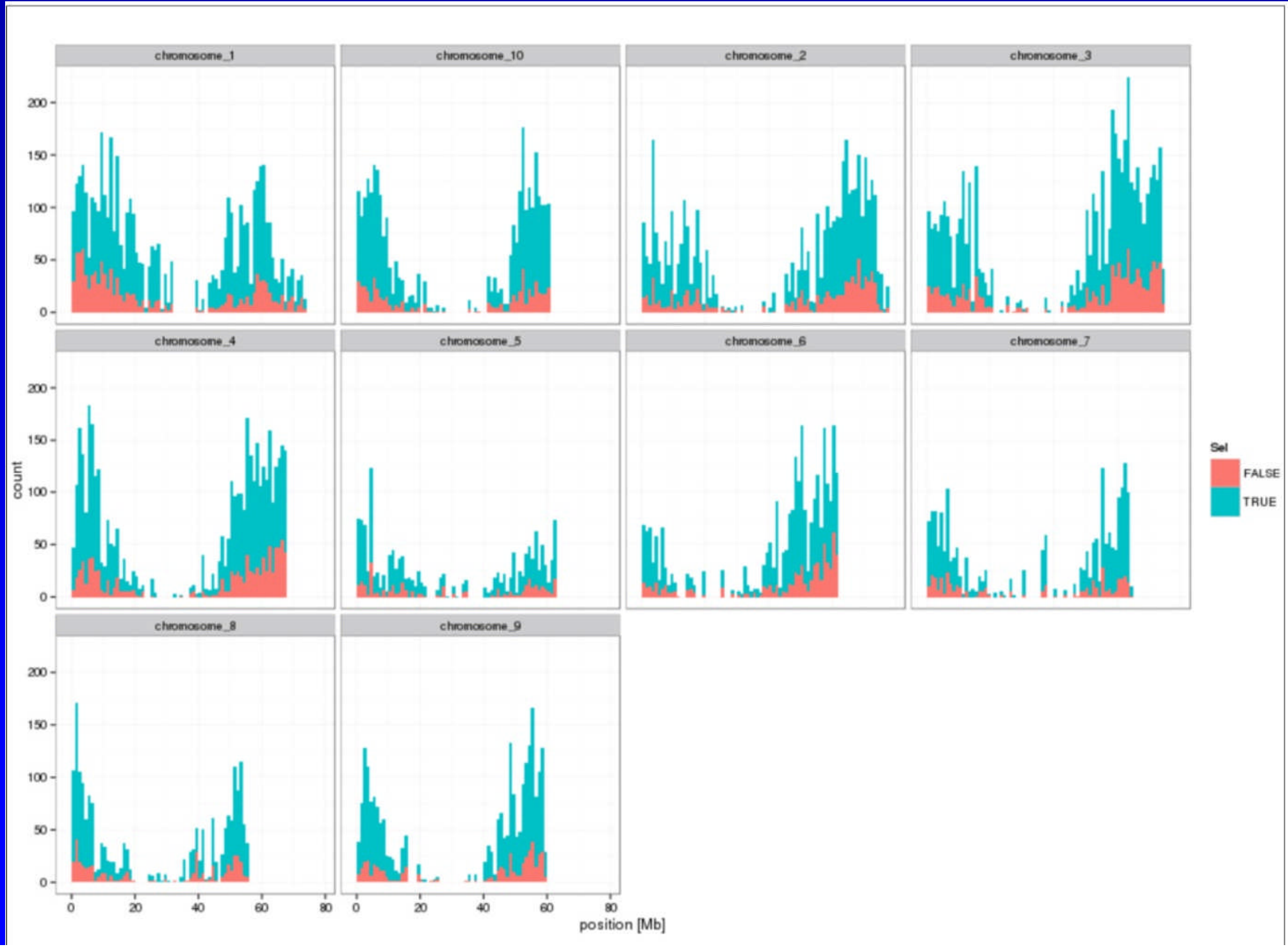
30,667 SNPs x 1,785 samples



GBS Marker Redundancy

R^2	#grouped markers	#markers to remove	% to remove	#final
0.9	1198	634	2.1	30033
0.8	5351	2972	9.7	27695
0.7	10966	6754	22.0	23913
0.6	15123	10049	32.8	20618
0.5	18451	12943	42.2	17724
0.4	21105	15474	50.5	15193
0.3	23493	17851	58.2	12816

Sorghum Genome Location of Filtered Markers



Challenges for GS

- Genotyping is an added expense
- Need uniform benchmarks for statistical methods, design of training populations and reporting of genomic prediction results.
- Genomic selection has not been implemented for most species and integration will vary for different crops.
- Methods for optimization in breeding programs are needed.
- Moving from the “black box” approach to discovery mode.





Summary: Genomic Selection

- GS is complementary to MAS and Association Breeding.
- GS does not require knowledge of the underlying genetic control and/or biological function, but GS can facilitate their investigation and understanding.
- Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E increases prediction accuracy and provides insight into the genetic architecture controlling G*E.
- GS was effective for increasing quantitative stem rust resistance.
- Most important advantages are reductions in the length of the selection cycle and associated phenotyping cost resulting in greater genetic gain per year.

Acknowledgements

- Bill & Melinda Gates Foundation grant to Cornell University for Borlaug Global Rust Initiative Durable Rust Resistance in Wheat.
- Bill & Melinda Gates Foundation grant to Cornell University & CIMMYT to implement a pilot project on genomic selection in wheat & maize breeding programs
- USDA Cooperative State Research, Education and Extension Service, Coordinated Agricultural Project 2005-05130-Wheat Applied Genomics
- USDA National Institute of Food and Agriculture, NRI Triticeae Coordinated Agricultural Project 2011-68002-30029 Improving Barley and Wheat Germplasm for Changing Environments
- USDA National Needs Fellowship Grant 2005-38420-15785: Provided Fellowship for Elliot Heffner
- USDA National Needs Fellowship Grant 2008-38420-04755: Provided Fellowship for Jessica Rutkoski
- Limagrain Europe (France) provided support for Nicolas Heslot
- US Wheat & Barley Scab Initiative: Provided funding for the fusarium head blight research

USDA United States Department of Agriculture

KARI

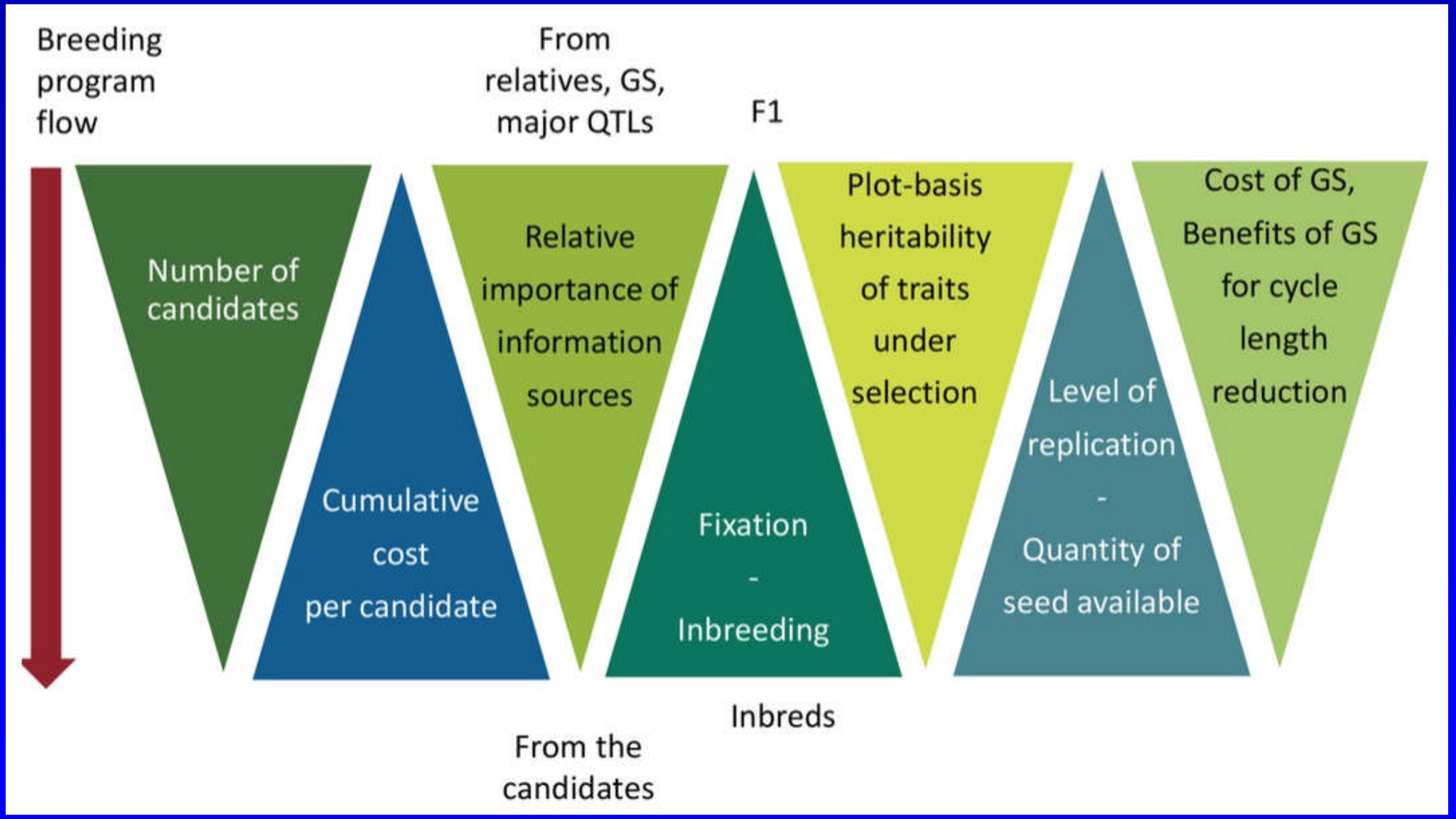
Limagrain Europe

CIMMYT

Questions?

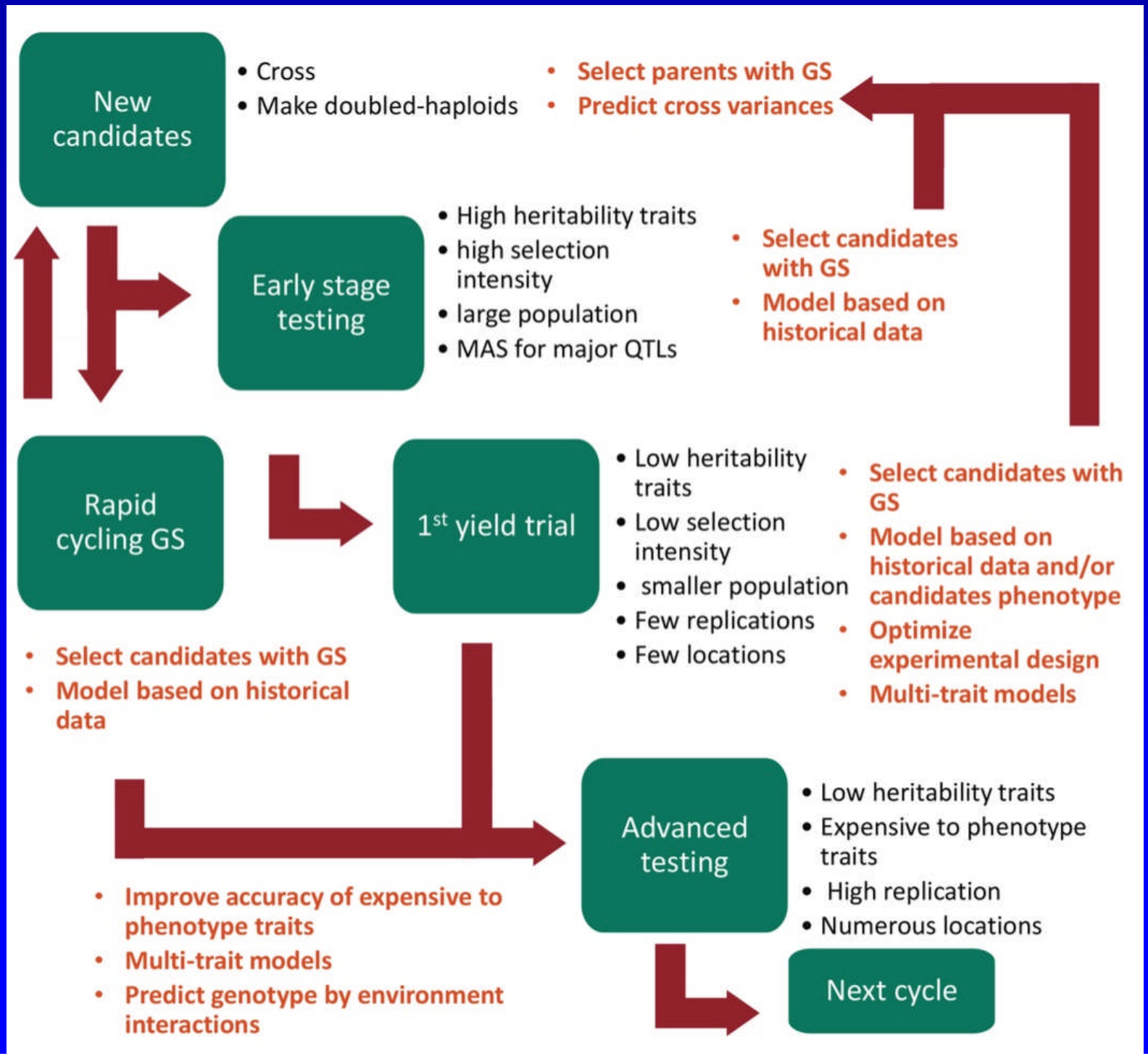


Perspectives for Genomic Selection
Heslot et al. 2015. Crop Sci. 55:1-12



Changes in key parameters during a breeding cycle

Perspectives for Genomic Selection



Questions?

