How Should I Select the Individuals of my Training Population to Make Selections in Genomic Selection?

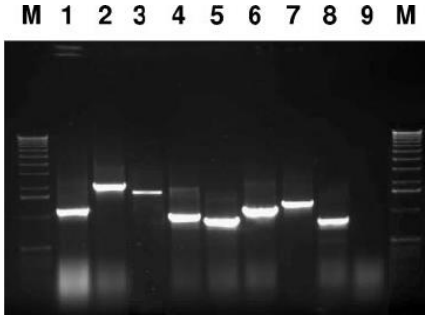
Julio Isidro Sánchez

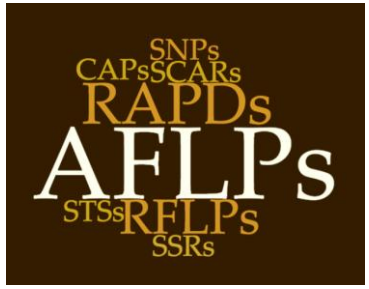
10th International Oat Conference

11-15 July 2016 Saint Petersburg-Rusia

Tools

Phenotypic Selection

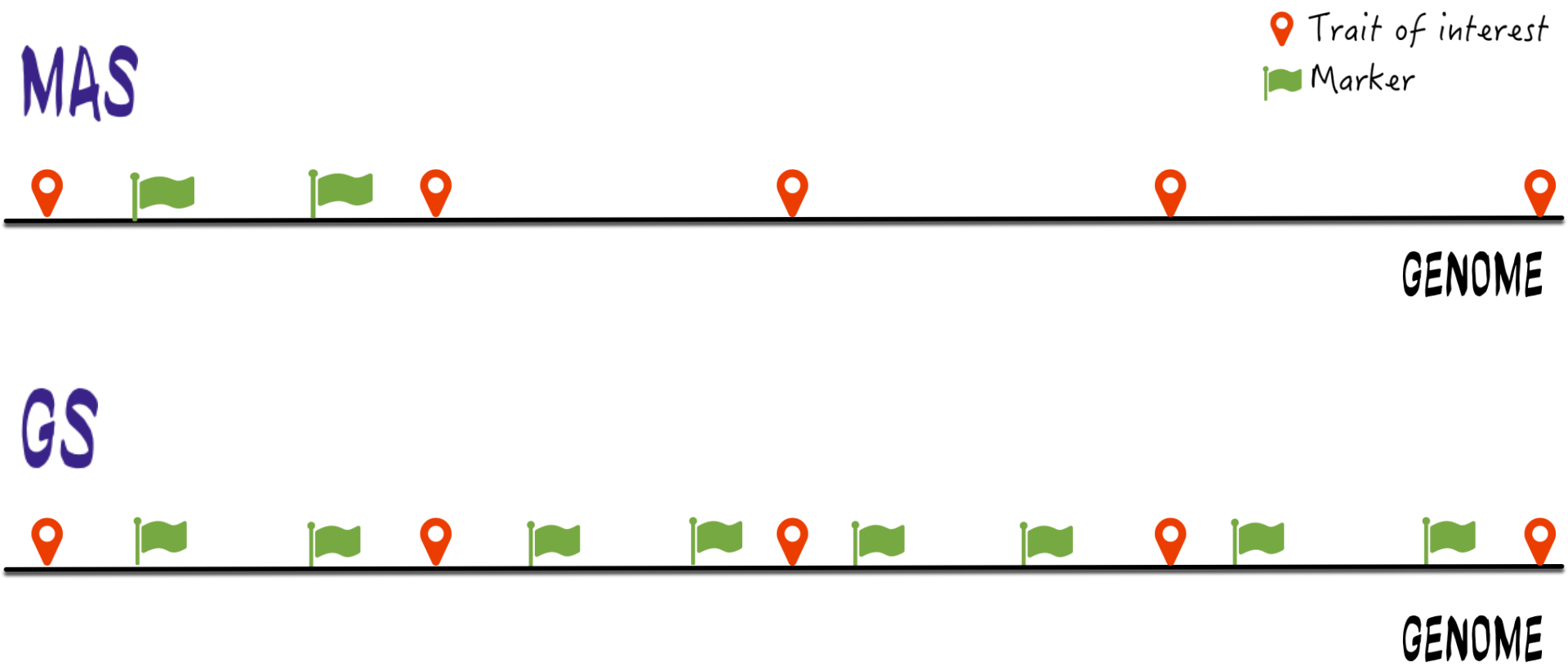




MAS

Without the rigth tools,
We are just playing.

MOST GENETIC CHARACTERS ARE QUANTITATIVES

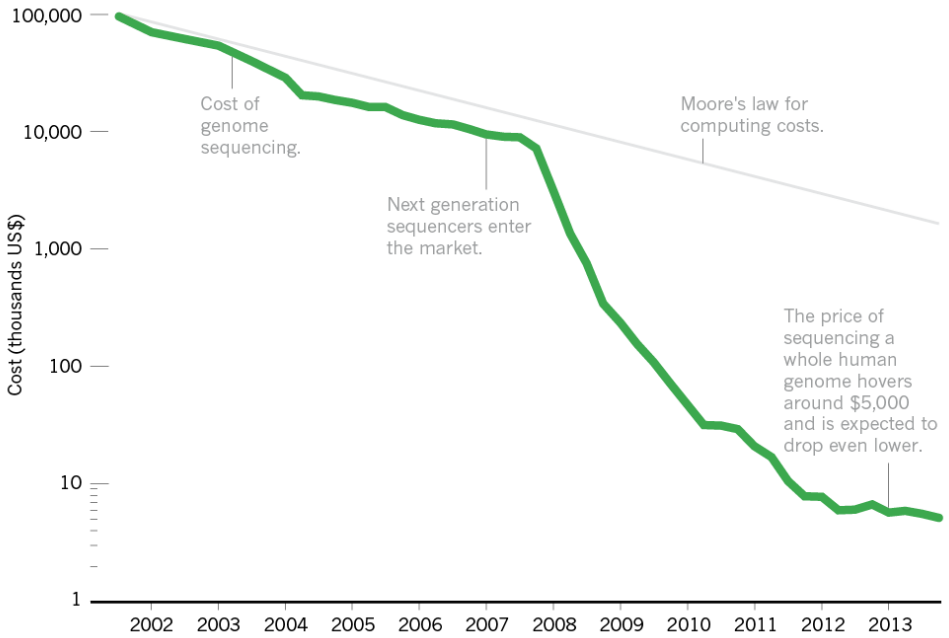


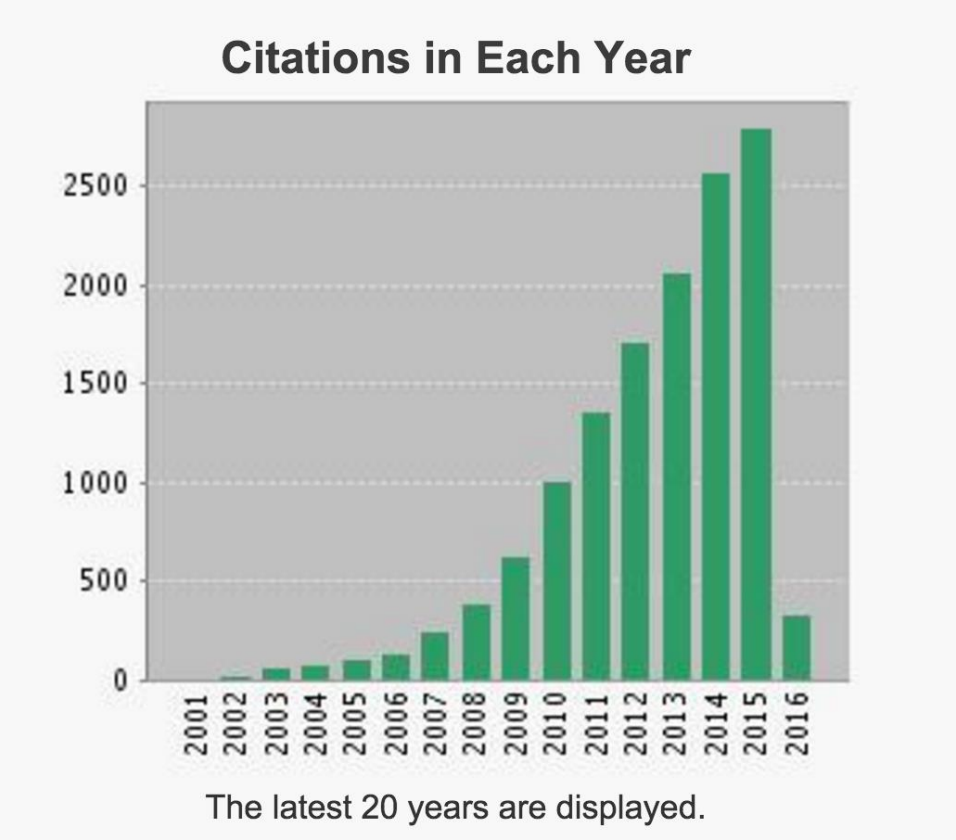
Tools

GENOMIC SELECTION

Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.





Number of Citations of GS per year. Web of science

WHAT IS GENOMIC SELECTION?

Genomic Selection is a "new" tool in plant and animal breeding that use statistical modeling to predict how a plant/animal will perform (Breeding Value) before it is phenotyped.

Big Picture in Plant Breeding

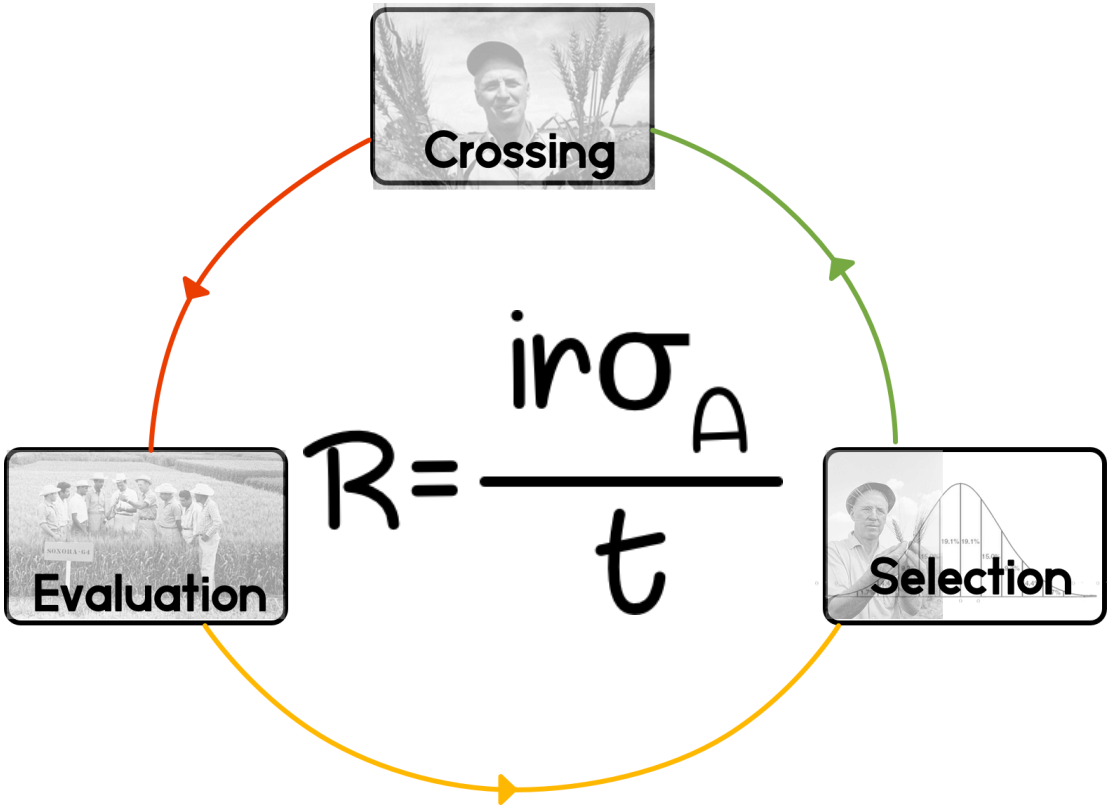
$$R = \frac{ir\sigma_A}{t}$$

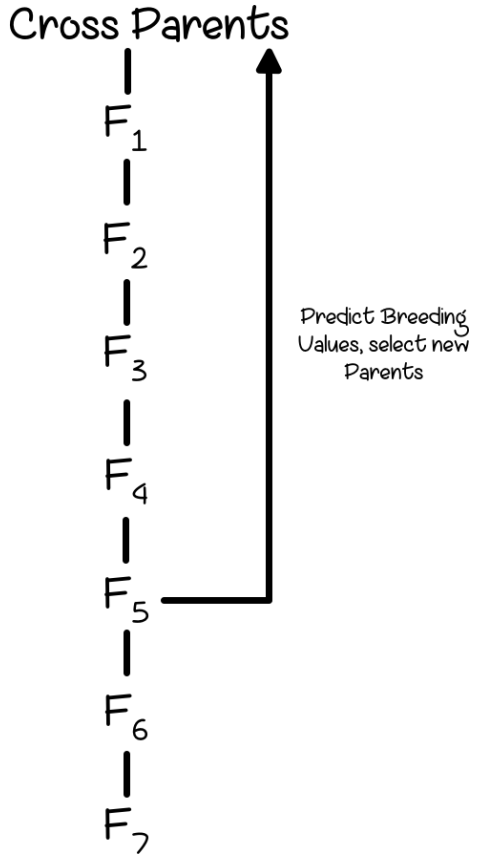
i=Intensity of selection

r=Accuracy of selection

σ_A=Additive genetic variance (standard deviation)

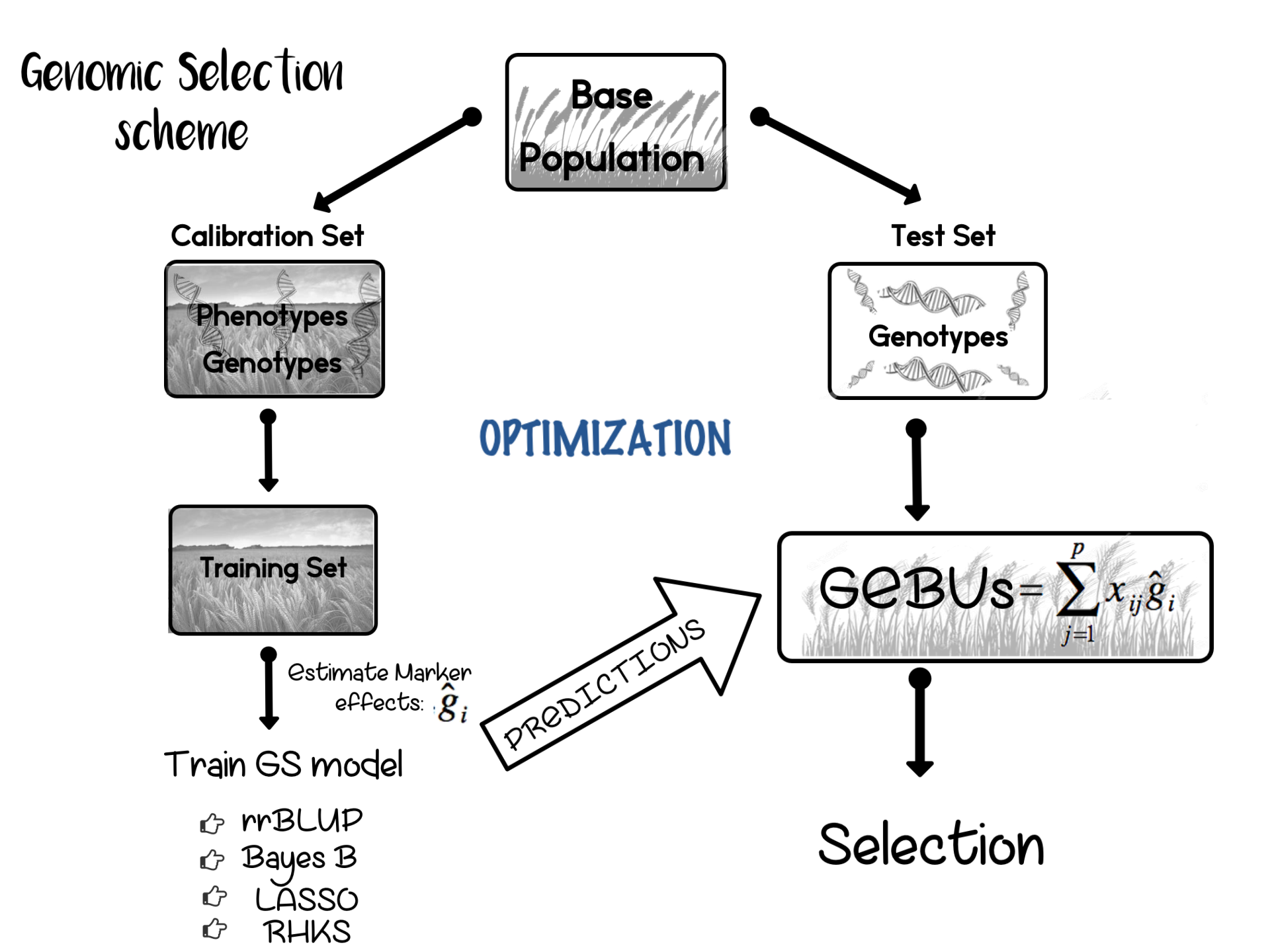
t= Time





Increase Genetic Gain by:

- Increase Accuracy of Selection
- Decrease Generation interval


Genomic Selection scheme
Base Population
Calibration Set
Test Set
Phenotypes
Genotypes
Genotypes
OPTIMIZATION
Training Set
GEBUs = $\sum_{j=1}^{p} x_{ij}\hat{g}_i$
Estimate Marker effects: $\hat{g}_i$
PREDICTIONS
Train GS model
rrBLUP
Bayes B
LASSO
RHKS
Selection

Questi on

Theor Appl Genet (2015) 128:145–158
DOI 10.1007/s00122-014-2418-4

ORIGINAL PAPER

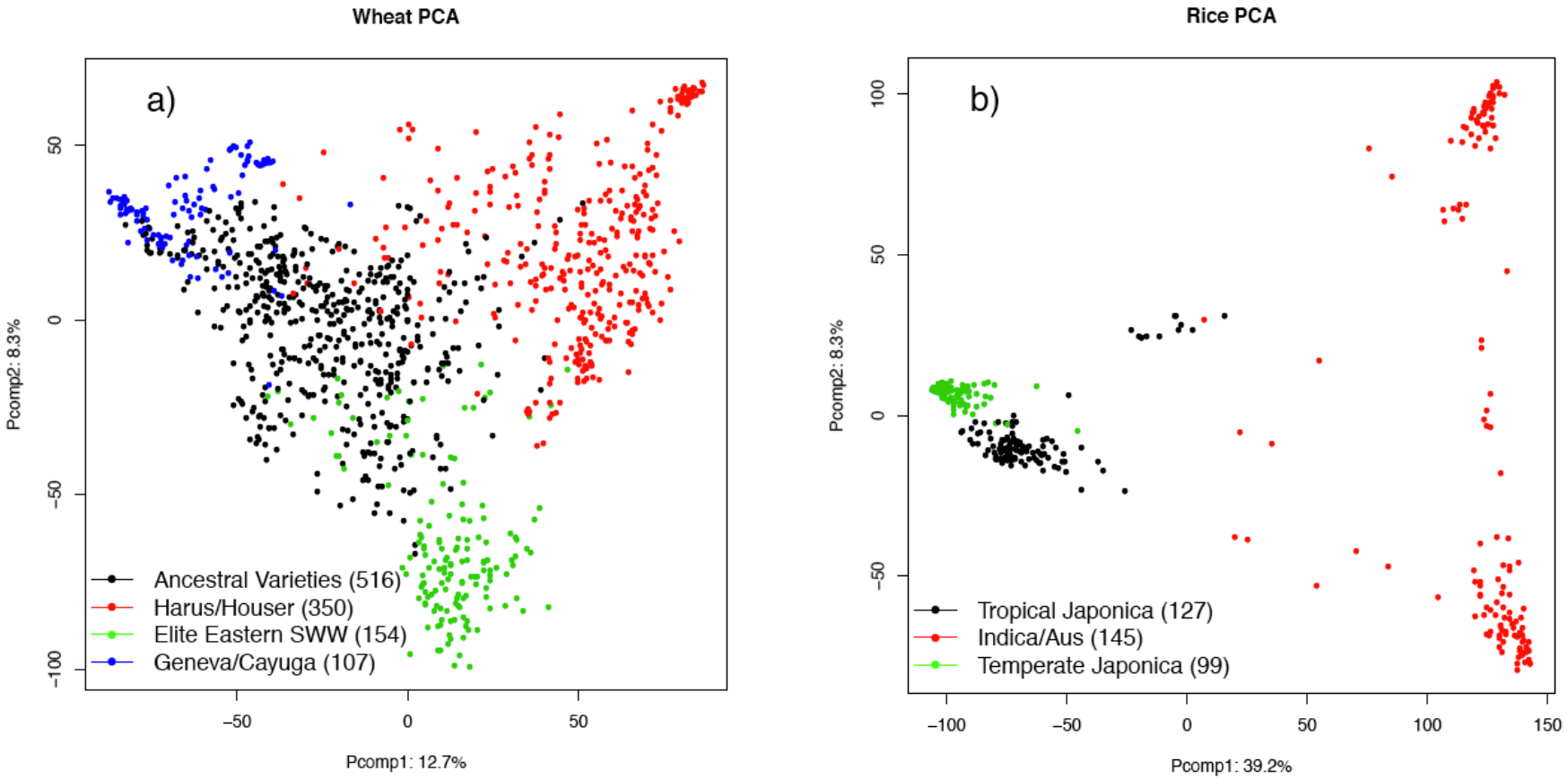
Training set optimization under population structure in genomic selection

Julio Isidro · Jean-Luc Jannink · Deniz Akdemir · Jesse Poland · Nicolas Heslot · Mark E. Sorrells

Does population structure has an impact on the optimization of the training population?.

1. Random Sampling
2. Stratified Sampling
3. Coefficient of determination (CD)
4. Prediction Error Variance (PEV)
5. Stratified Coefficient of

OPTIMIZATION



Mild Population Structure in Wheat

Improving genetic diversity using kinship matrix

$$PEV = diag\left[\frac{c'(Z'MZ + \lambda G^{-1})^{-1}c}{c'c}\right] \times \sigma^2_{\varepsilon}$$

$$CD = diag\left[\frac{c'(G - \lambda(Z'MZ + \lambda G^{-1})^{-1}c}{c'Gc}\right]$$

Coefficient of determination use Kinship matrix in its calculation

OPTIMIZATION

Method 1- STRATIFIED SAMPLING

100 pop size

1127

Wheat Dataset

Random

627

Calibration Set

500

Test Set

PREDICTION

50X

C1-188

C2-175

CLUSTERING

C3-153

C4-111

Stratified Sampling

34 C1

29 C2

21 C3

16 C4

100 TRS

Optimization Training Population Size

25

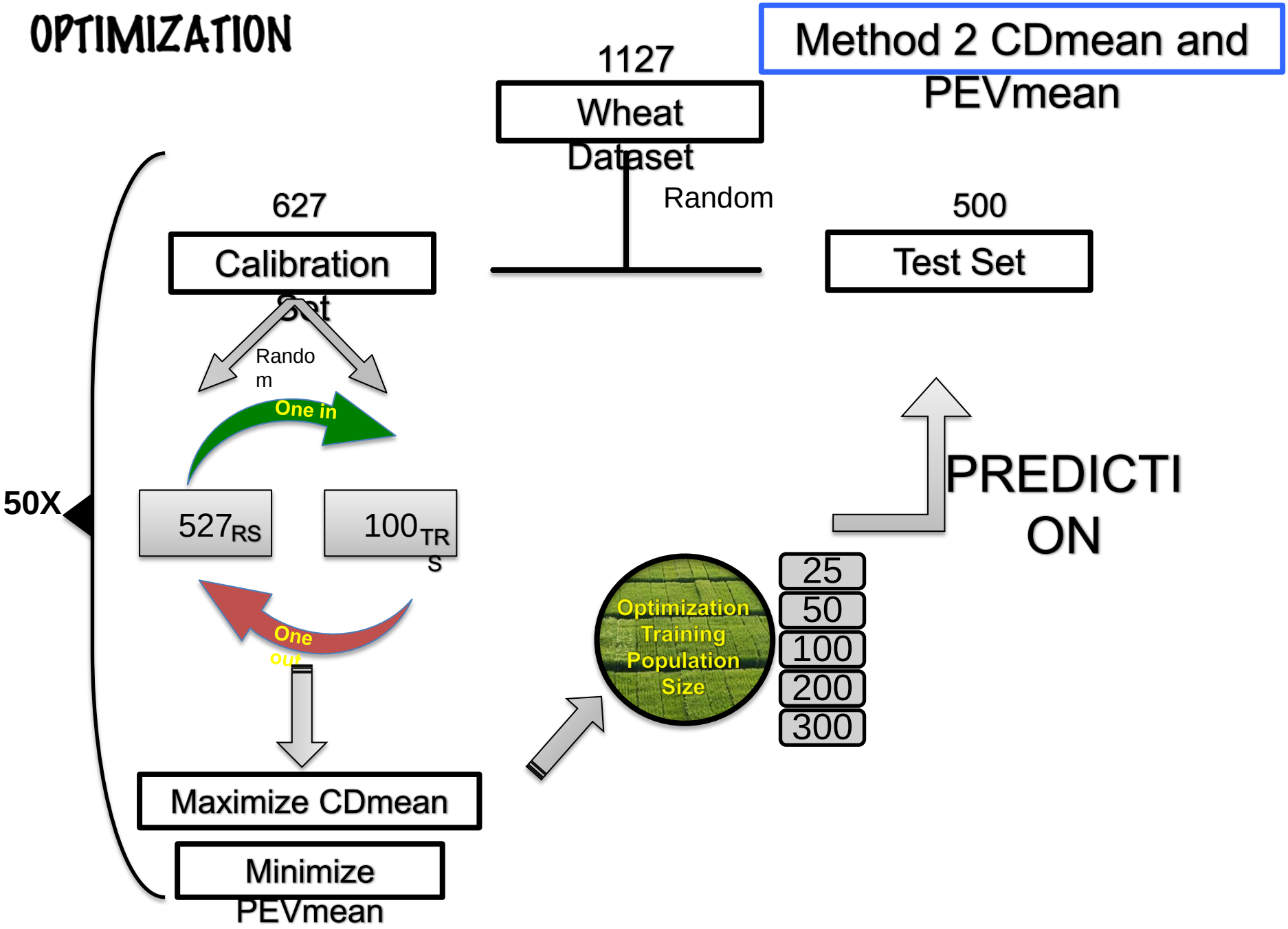
50

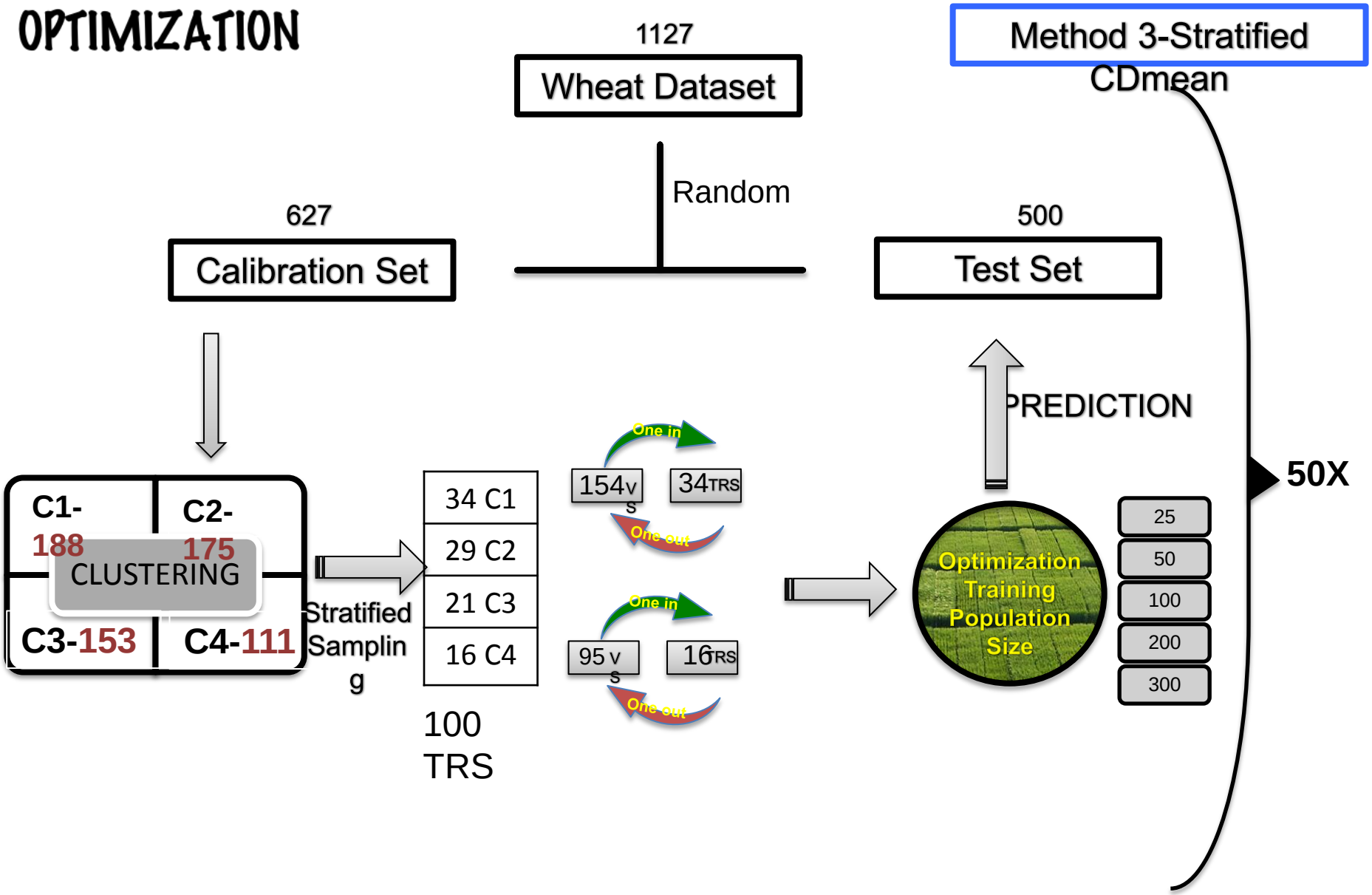
100

200

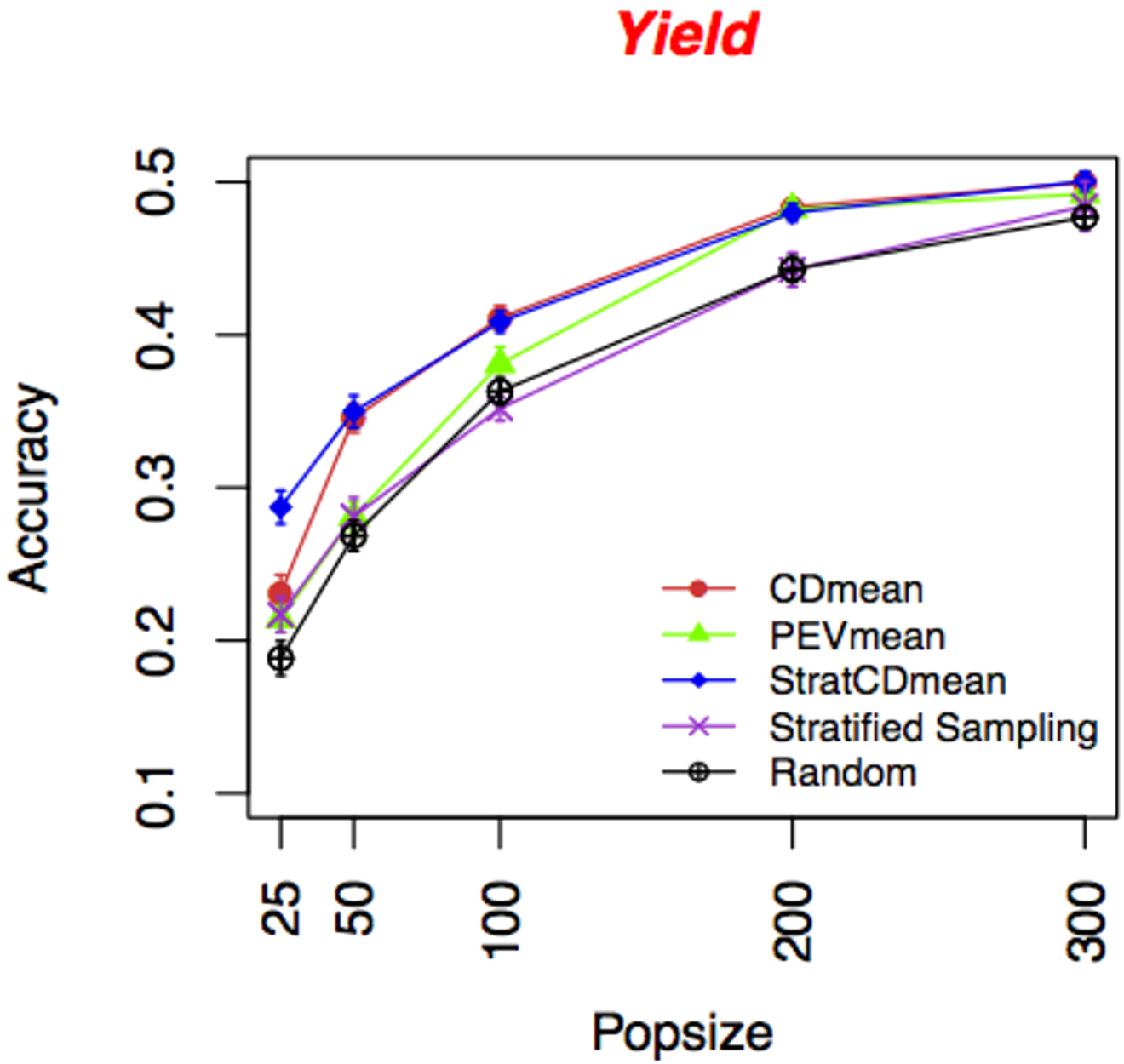
300

OPTIMIZATION




OPTIMIZATION
1127
Wheat Dataset
Method 3-Stratified
CDmean
Random
627
Calibration Set
500
Test Set
PREDICTION
50X
C1-188
C2-175
CLUSTERING
C3-153
C4-111
Stratified Sampling
34 C1
29 C2
21 C3
16 C4
100 TRS
One in
154 VS
34 TRS
One out
One in
95 VS
16 TRS
One out
Optimization Training Population Size
25
50
100
200
300

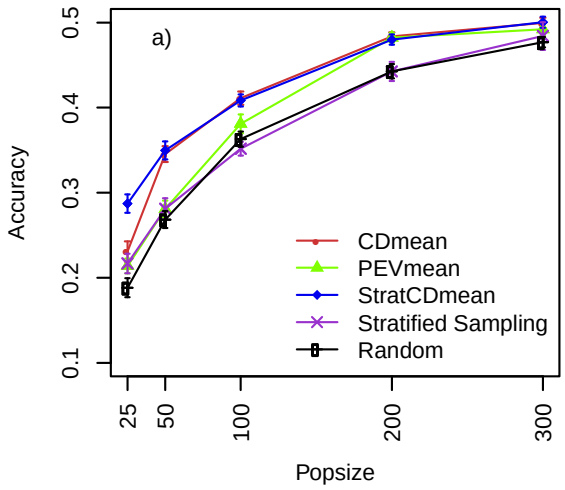
Improving genetic diversity using kinship matrix



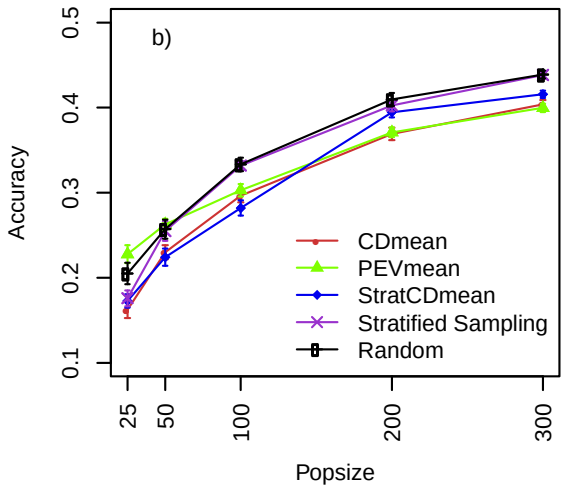
$$R = \frac{ir\sigma_A}{t}$$

OPTIMIZATION

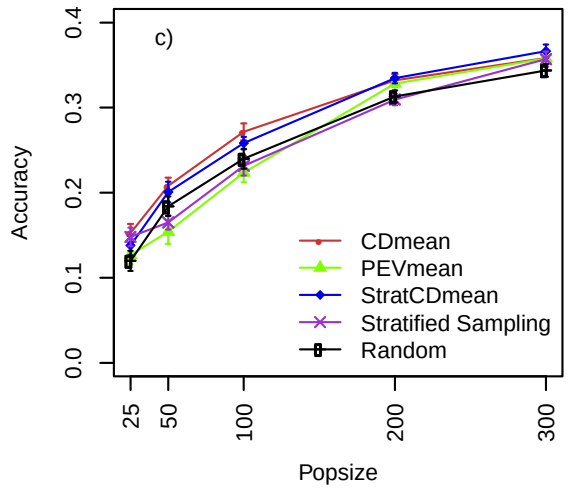
Yield



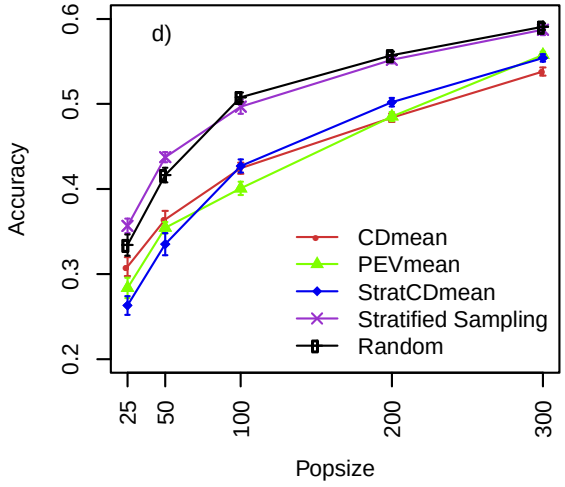
Test Weight



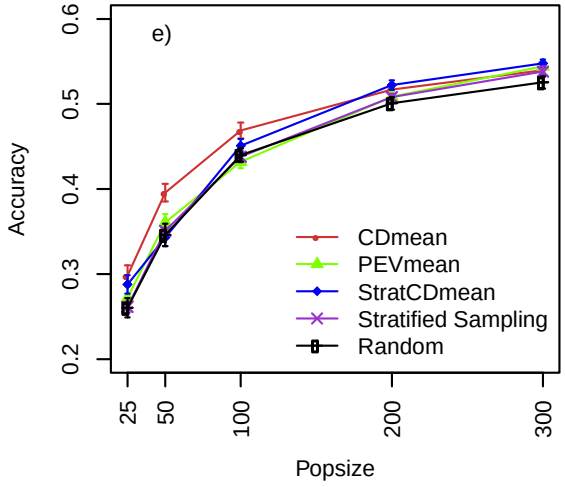
Lodging



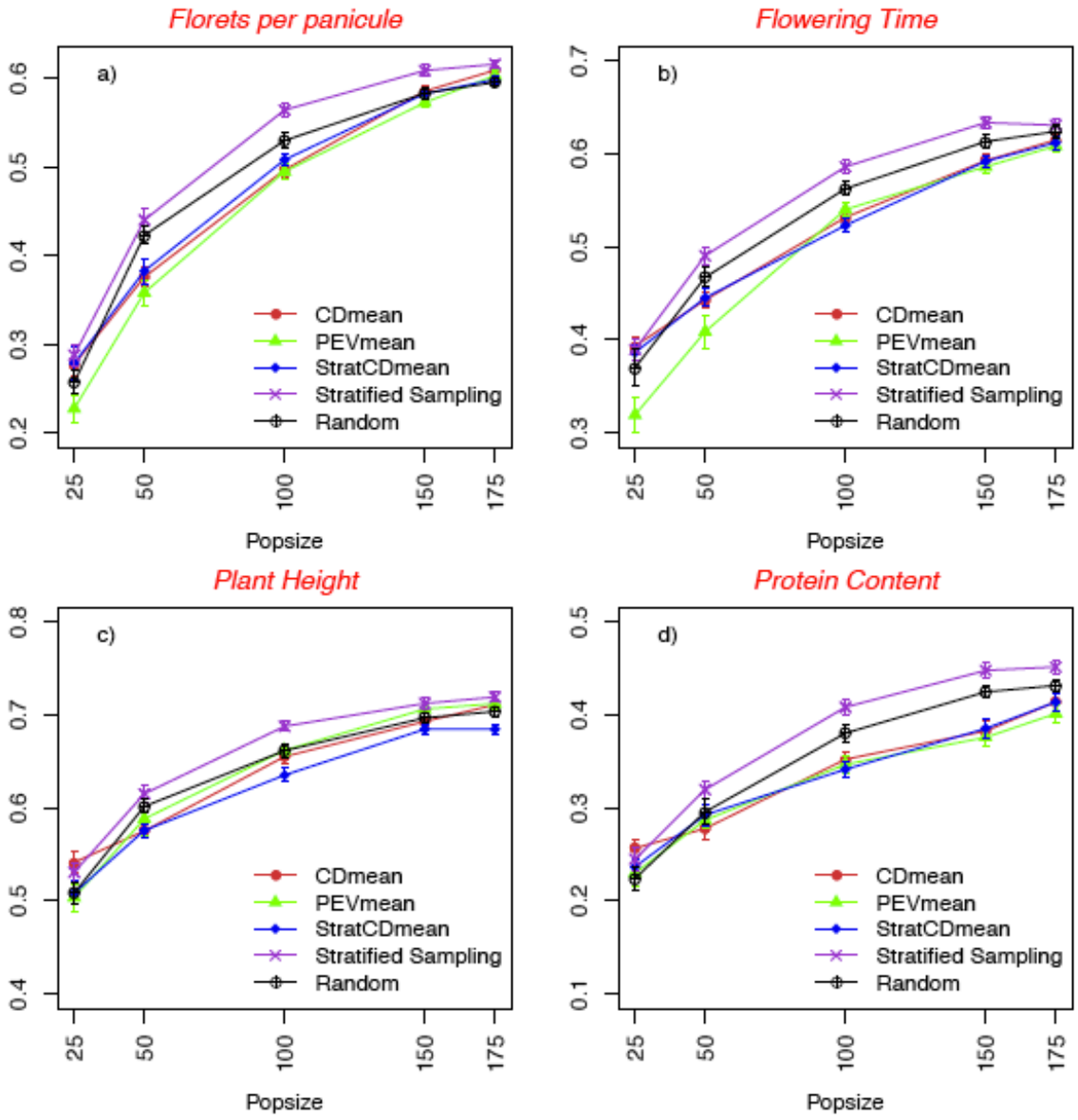
Heading Date



Plant Height



OPTIMIZATION

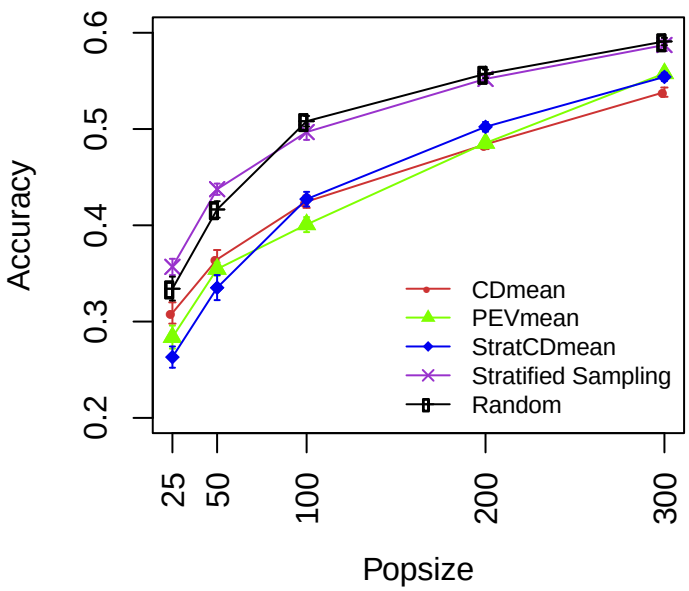


OPTIMIZATION

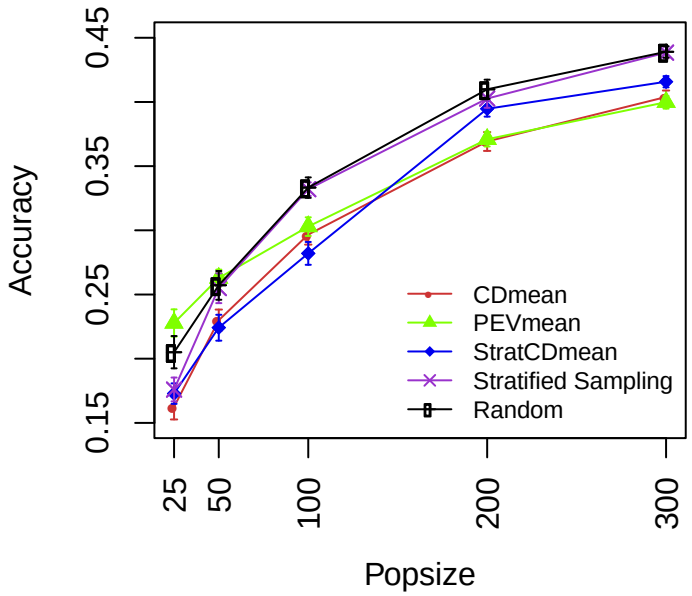
- Population structure has an impact on the optimization of the training population.
  - Mild PS---CDmean and StratCDmean
  - Strong PS---Stratified Samplin

OPTIMIZATION

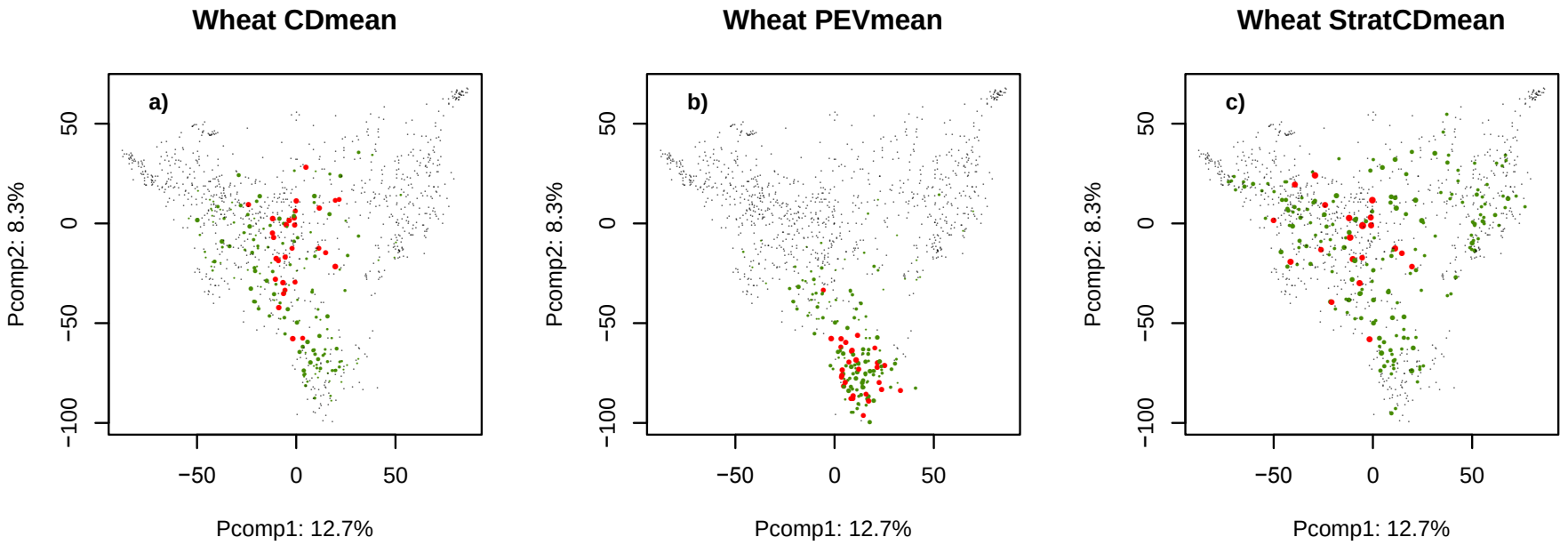
Heading Date



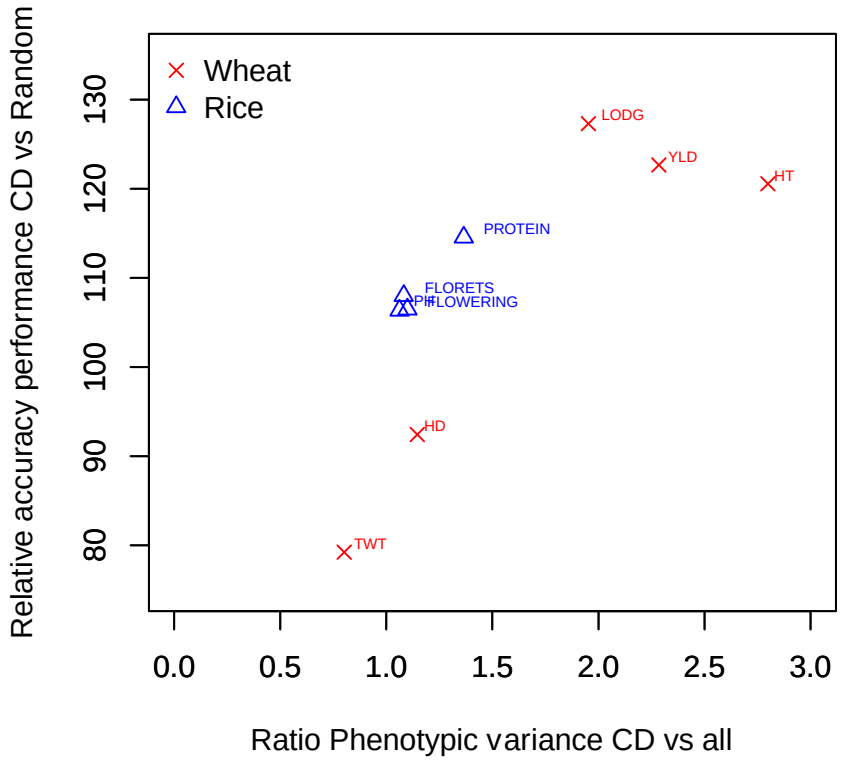
Test Weight



Improving genetic diversity using kinship matrix



OPTIMIZATION



Large genotypic variance obtained by CDmean doesn´t translate into large phenotypic variance

OPTIMIZATION

- There isn´t a best selection criterion to optimize the TRS under population structure. PS plays an important role in optimization of TRS in GS
- Before optimization, population structure must be evaluated
- Highest accuracies with methods that capture more phenotypic variance.
- CDmean is an optimal criterion for long-

Akdemir *et al. Genetics Selection Evolution* (2015) 47:38
DOI 10.1186/s12711-015-0116-6

GSE Genetics Selection Evolution

RESEARCH Open Access

Optimization of genomic selection training populations with a genetic algorithm

Deniz Akdemir[1*], Julio I Sanchez[1] and Jean-Luc Jannink[2]

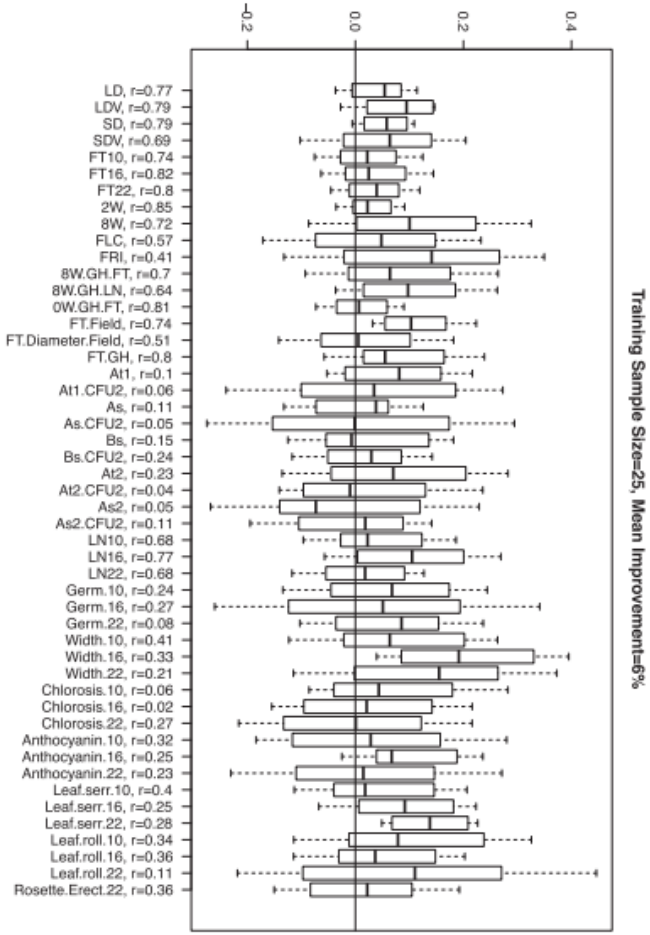
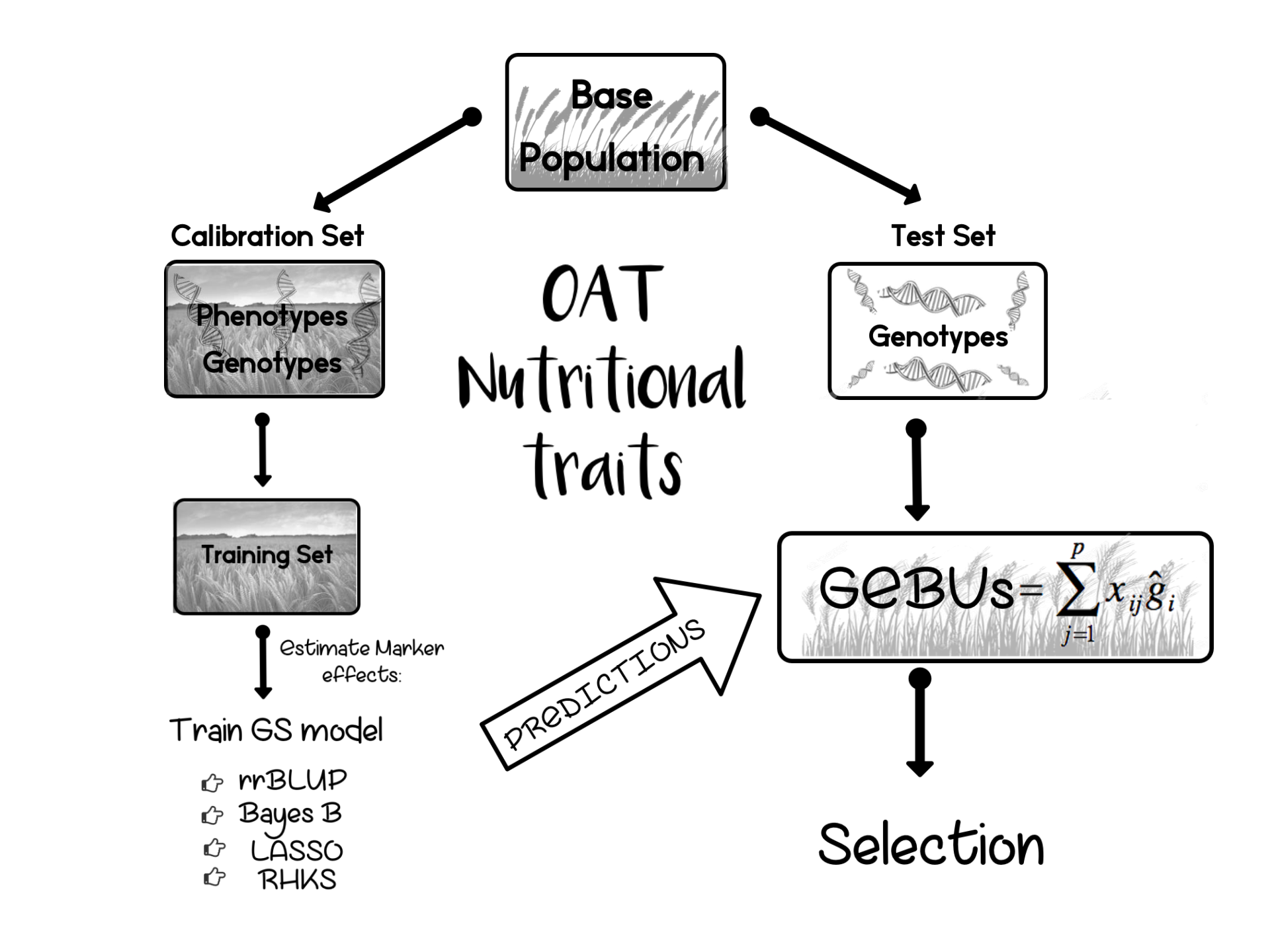


Figure 1 Arabidopsis data. The difference between the accuracies of the models trained on optimized populations versus random samples. Positive values indicate the cases for which the optimized population performed better as compared to a random sample.


Base Population
Calibration Set
Test Set
OAT Nutritional traits
Phenotypes
Genotypes
Genotypes
Training Set
Estimate Marker effects:
Train GS model
rrBLUP
Bayes B
LASSO
RHKS
PREDICTIONS
$GEBVs = \sum_{j=1}^{p} x_{ij}\hat{g}_i$
Selection

UCD Centre for Plant Science

THANKS

About | People | Research Areas | Education | Facilities | News and Events

Crop Improvement for Food Security

Trait selection for yield stability and disease resistance | Learn more ...

Plant Signalling and Development

Understanding fundamental cellular processes | Learn more ...

Plant Bioprocessing

Producing fuel and value-added products | Learn more ...

Adaptation to Future Climates

Predicting and evaluating crop productivity and performance under future climatic conditions | Learn more ...