



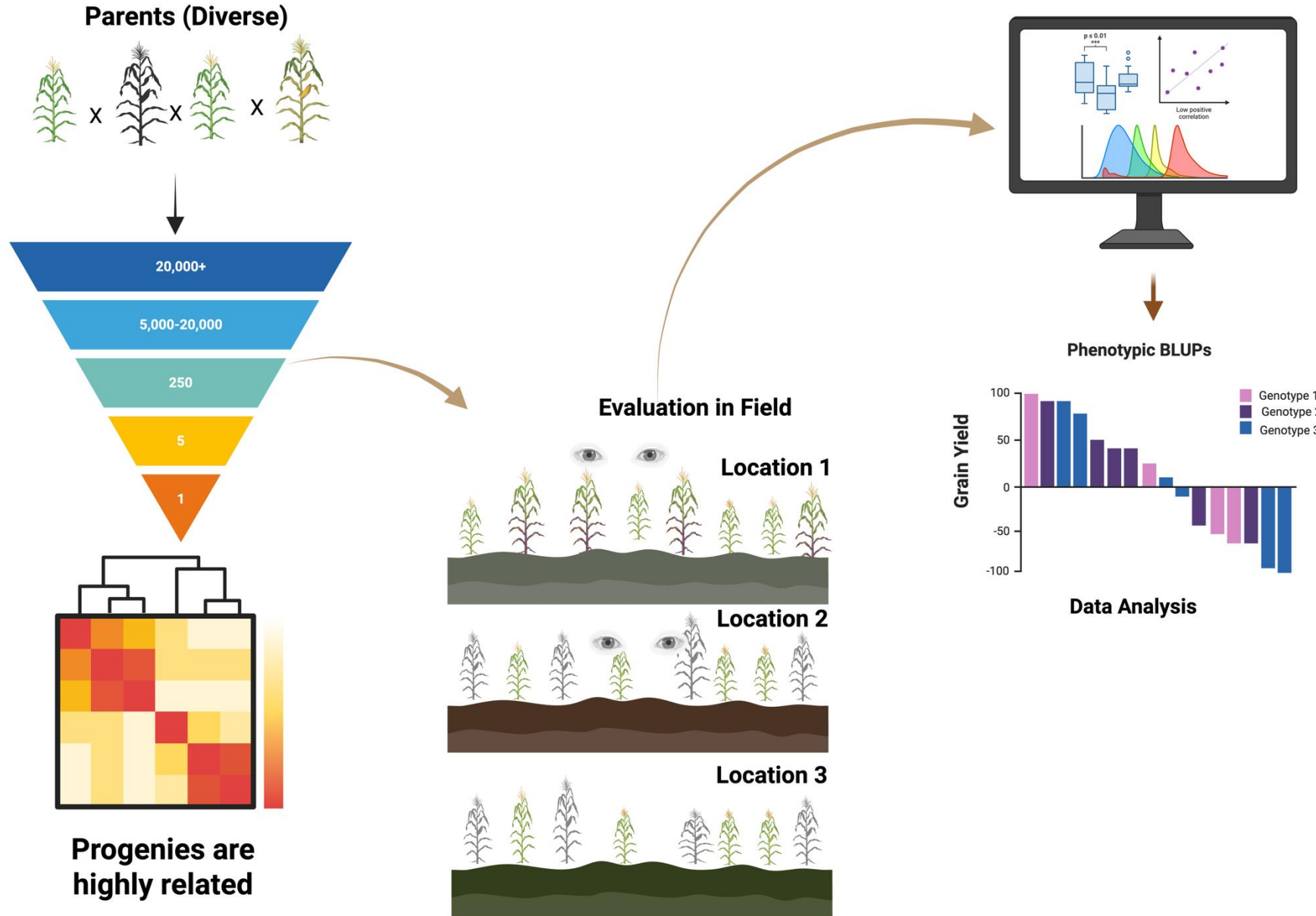
# **Fundamentals of Genomic Prediction and Data-Driven Crop Breeding (November 24-28, 2025)**

## **Fundamentals of Genomic Predictions**

**Module 2  
November 25, 2025**

**Waseem Hussain, Mahender Anumalla and Margaret  
Catolos**  
Rice Breeding Innovations Platform  
IRRI

# Traditional Breeding Phenotypic Selection



Assumption that all  
Genotypes are  
independent  
(No Relationship)

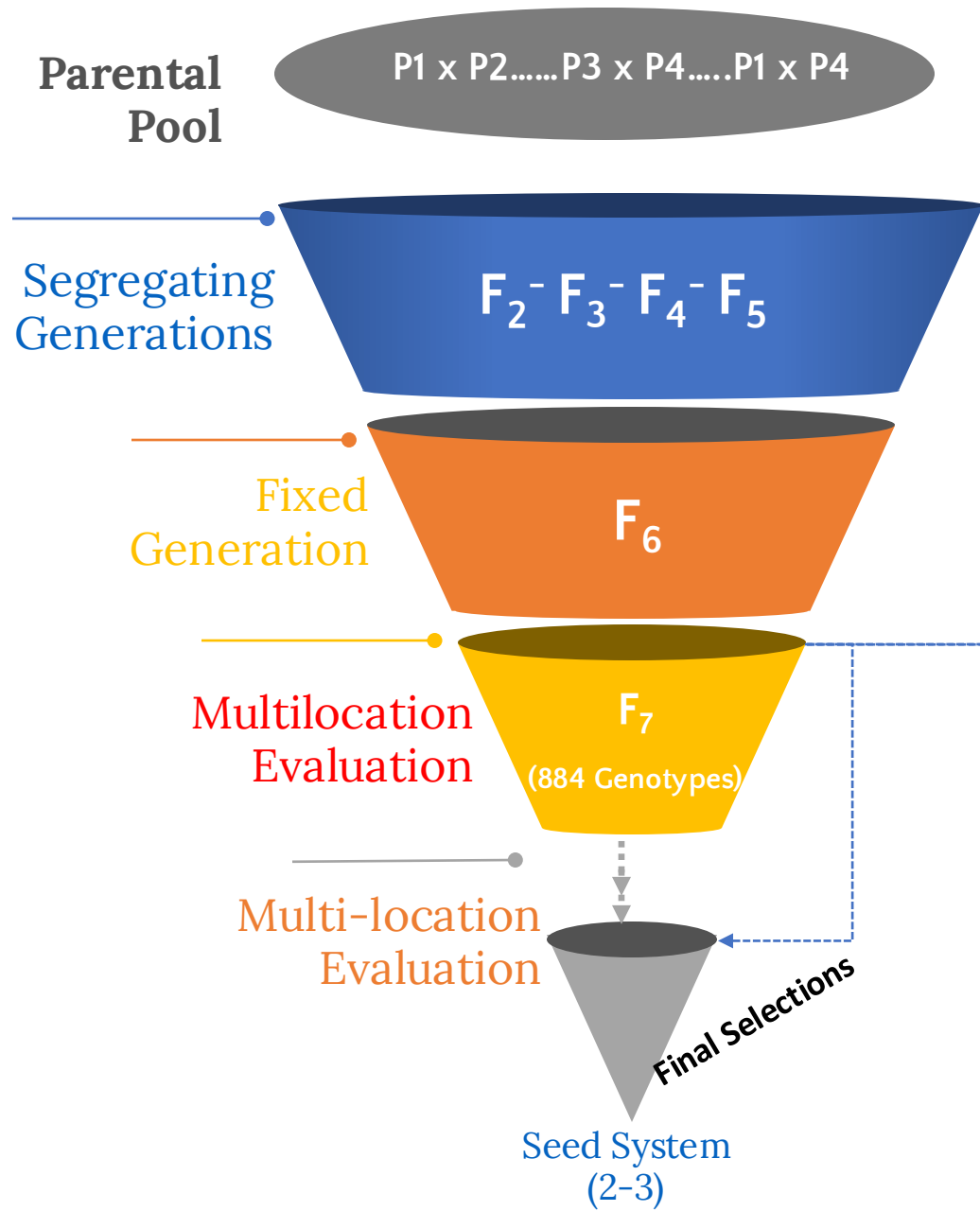
***But that is not  
True!***

(Genotypes have a half-sib or full-sib relationship).

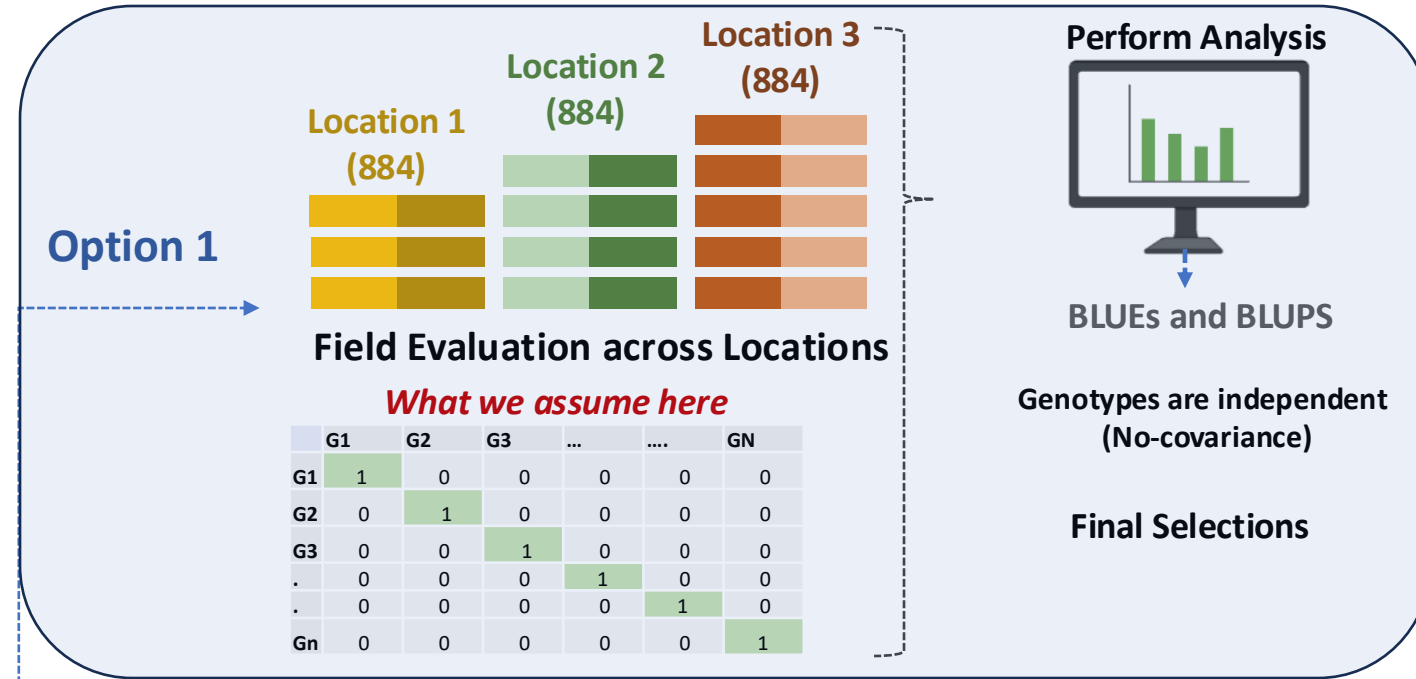
There is variance between the families, within families, But Also covariance (genetic similarity within families and across families)



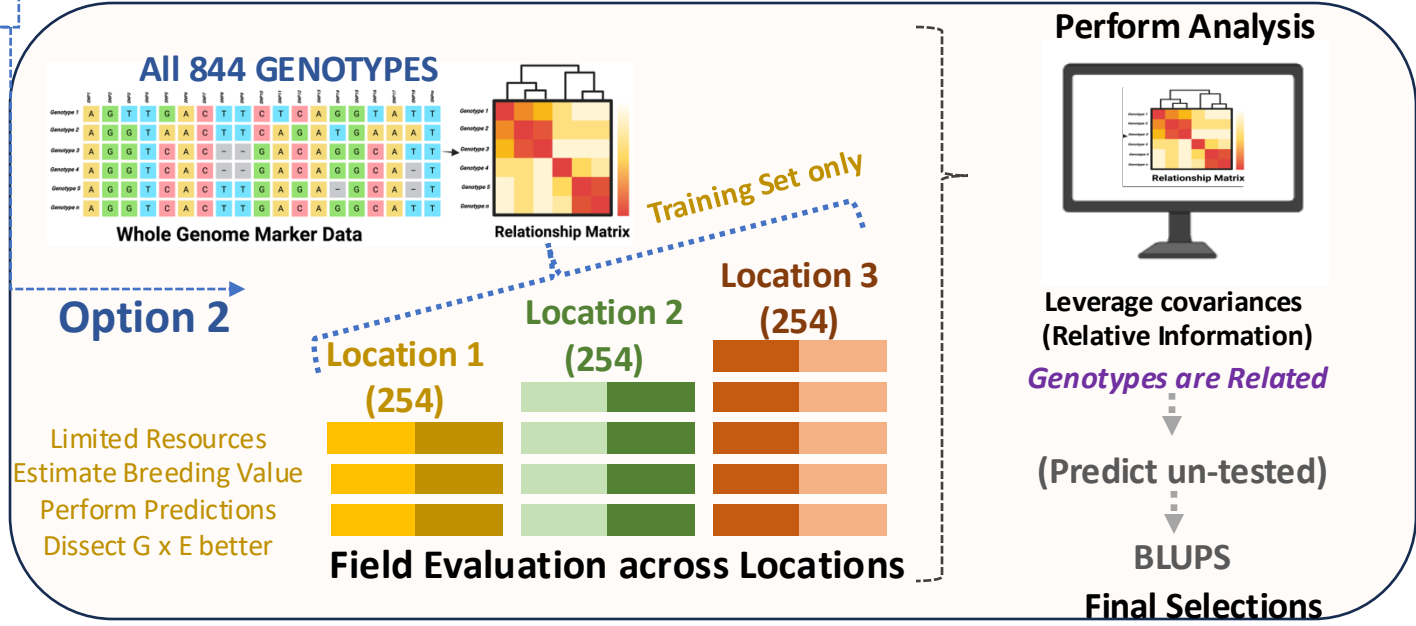
# General Breeding Pipeline



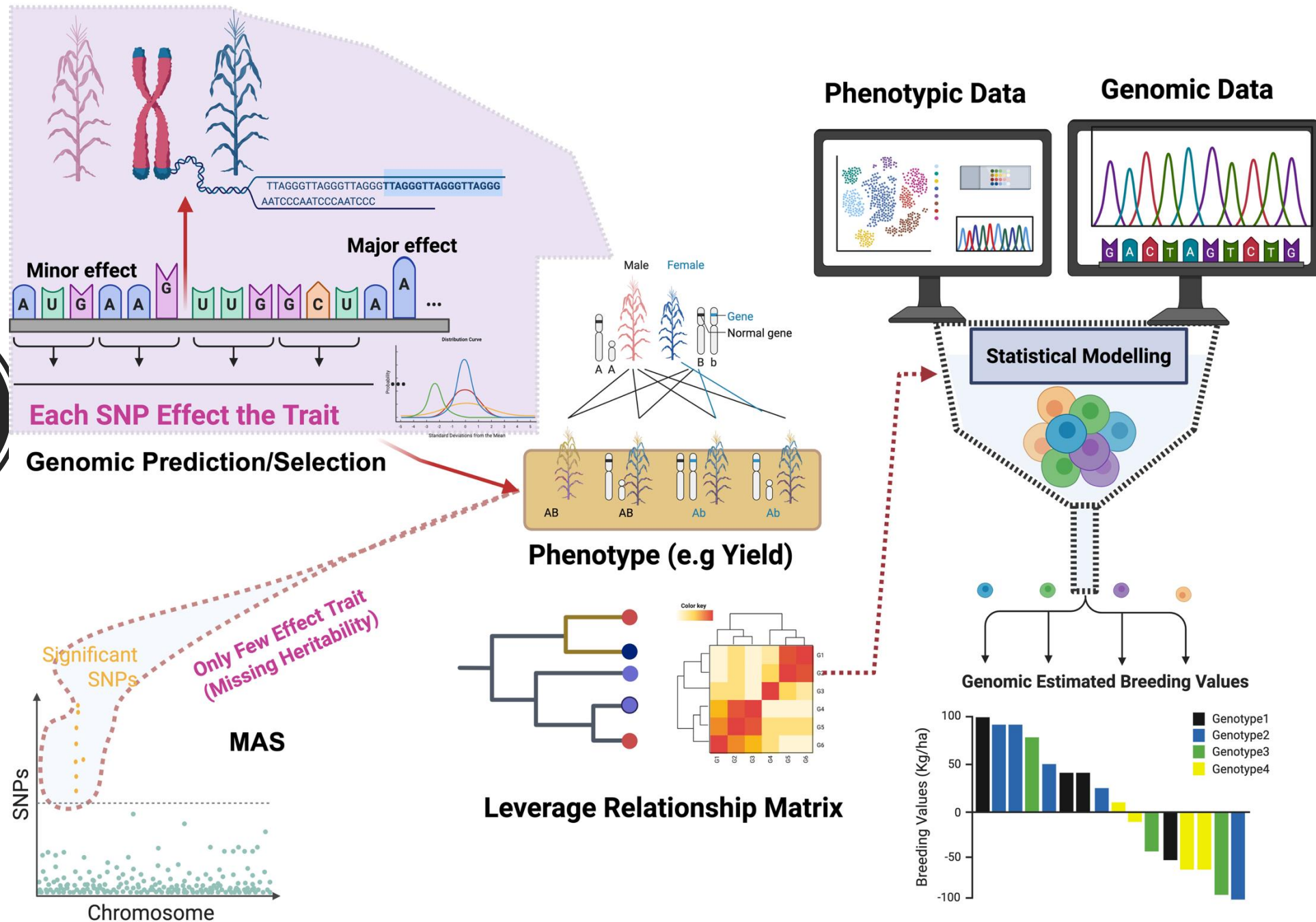
## Phenotypic Selection



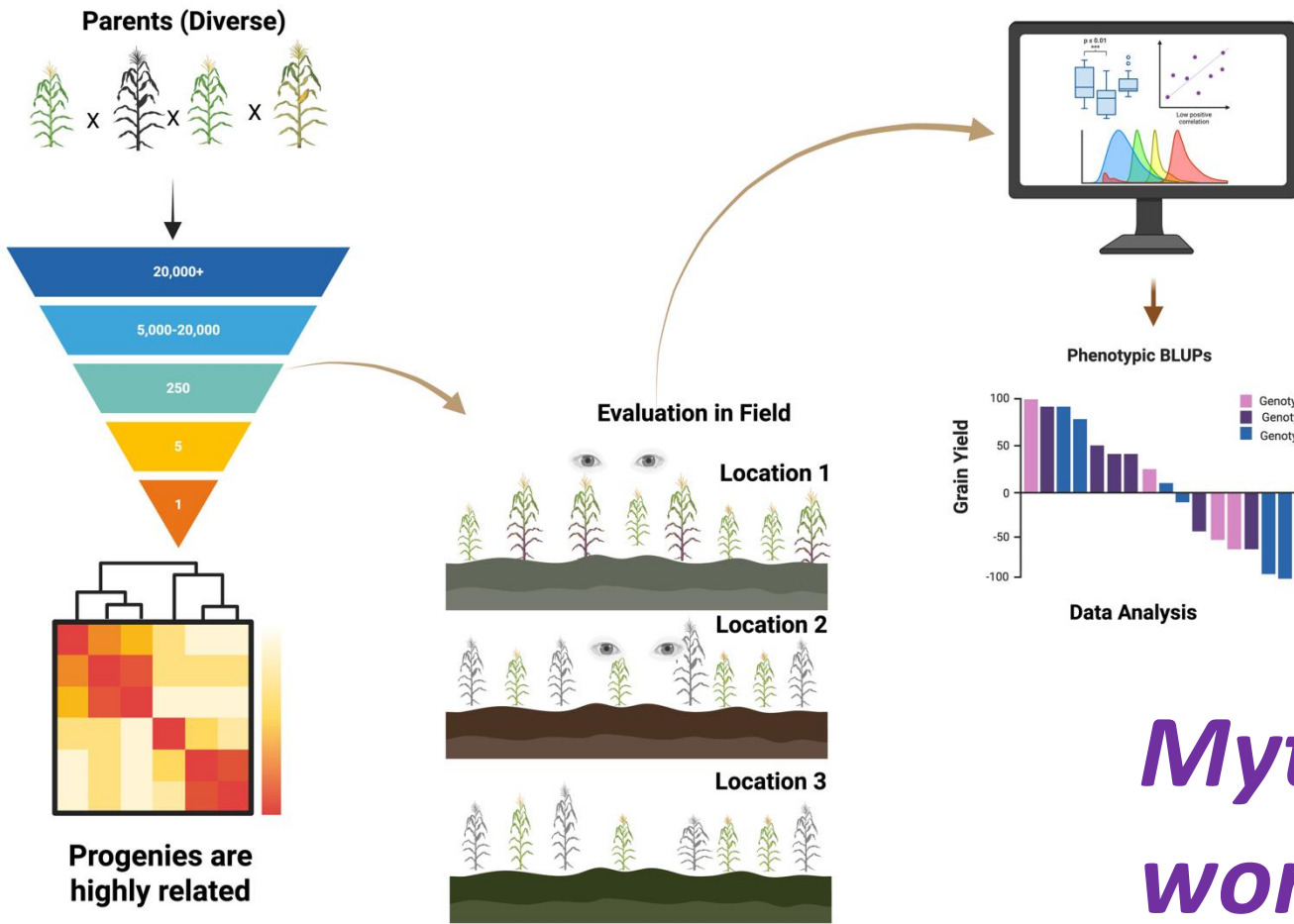
## Genomic Selection



# Genomic Selection (Leveraging Relationship)



# Phenotypic Selection vs MAS vs Genomic Selection



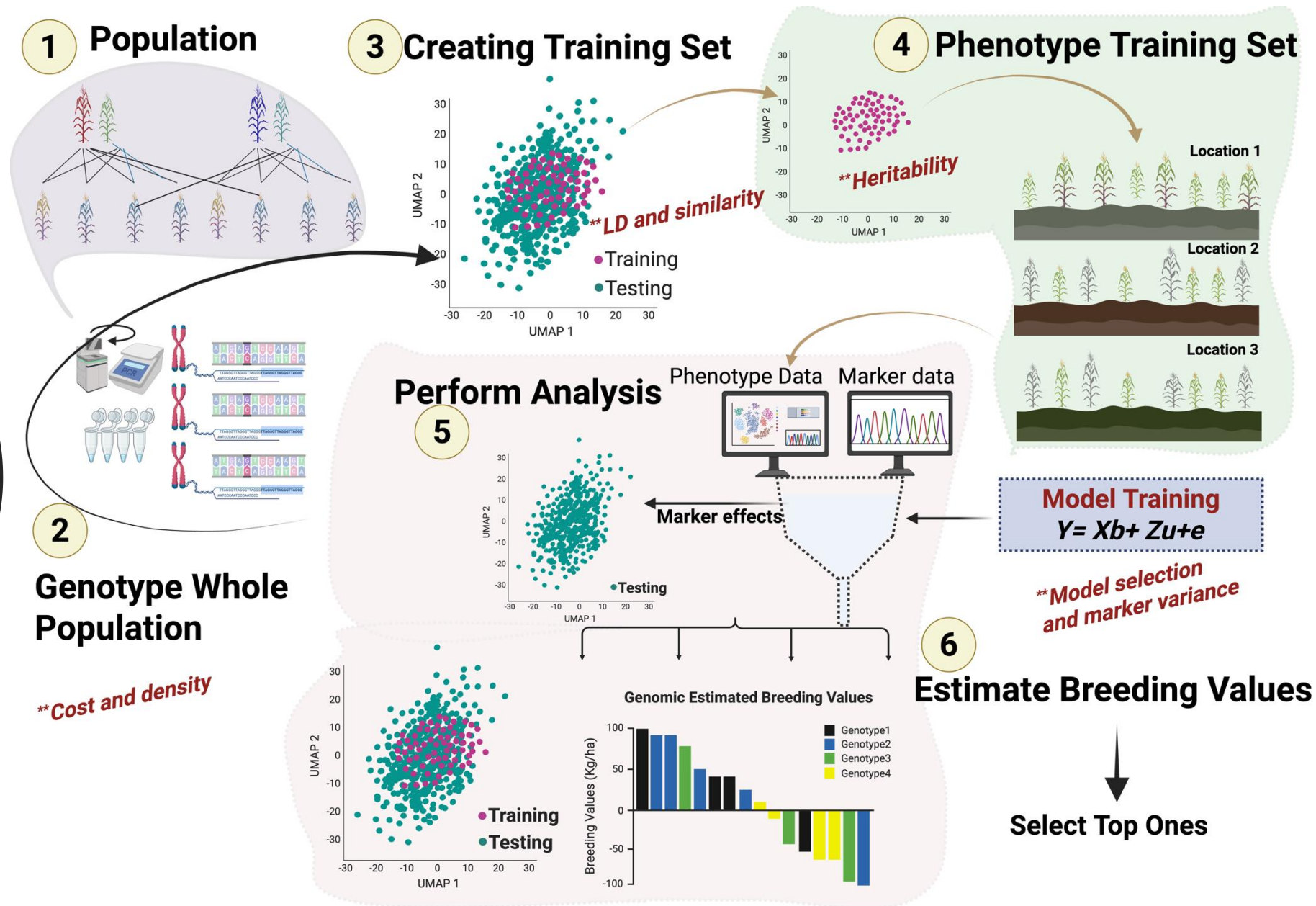
*Phenotypic selection: No Relationship*

*MAS: No relationship*

*Genomic Selection: **Relationship is Foundation***

*Myth 1: Genomic Prediction works in all Populations like MAS*

# How Genomic Prediction Works



# Evaluate Prediction Performance

Predictive correlation ( $r$ ) =  $\text{cor}(\text{True value}, \text{Predicted value})$

Divide the data into training and testing set

Training set  $\in (y_{\text{training}}, X_{\text{training}})$

Testing set  $\in (y_{\text{testing}}, X_{\text{testing}})$

$$y_{\text{training}} = X_{\text{training}} \beta_{\text{training}} + \varepsilon_{\text{training}}$$

Perform cross-validation

$$\hat{y}_{\text{testing}} = X_{\text{testing}} \beta_{\text{testing}}$$

$$= \text{cor}(y_{\text{testing}}, \hat{y}_{\text{testing}})$$

$$= \text{cor}(y_{\text{testing}}, X_{\text{testing}} \beta_{\text{testing}})$$

## Breeding Pipeline

$P_1 \times P_2$



$F_2$



$F_3-F_4$



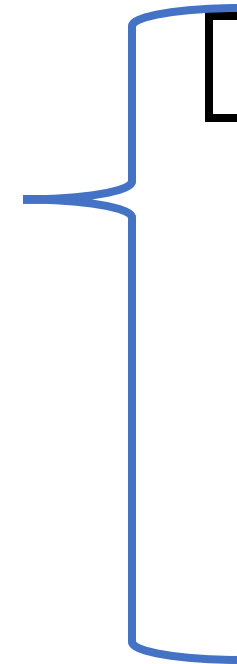
$F_5$



$F_6$



$F_7$





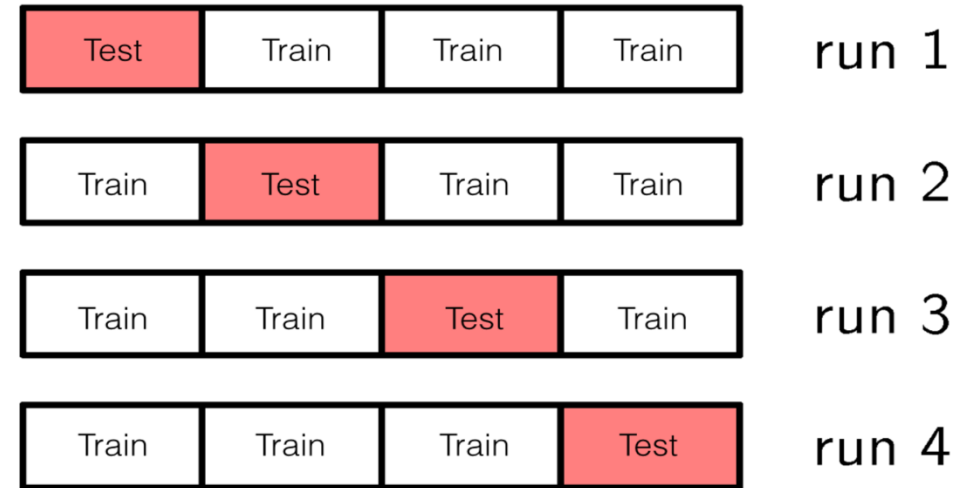
# Cross-validation

(evaluate prediction performance)



- ❖ Take model uncertainty into account
- ❖ Divide data into training and testing sets
- ❖ Train the model in the training set
- ❖ Evaluate predictive performance in the testing set
- ❖ Predictive correlation:  $r = \text{cor}(y, y_{\text{predicted}})$

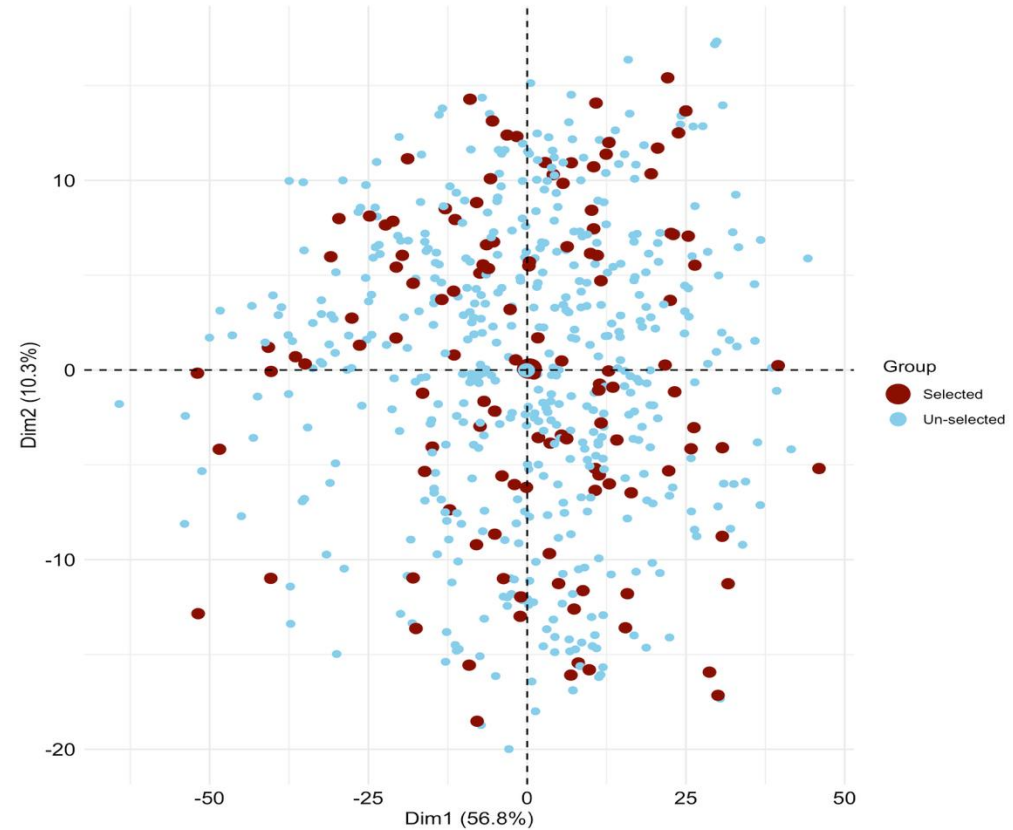
## K-fold Cross-validation





# How to Design The Training Set

- Maximizes the relationship between training and testing set
- Key for success of genomic selection



## Various Approaches to Design Training Set

- ❖ Prediction error variance (PEV) means: *Minimize the error variance*
- ❖ Critical Difference (CD) means: *Minimize error variance and Relationship*
- ❖ K-means algorithm: *Based on G matrix and similarity*



# Literature to Read on Matrix

- <https://courses.lumenlearning.com/wmopen-collegealgebra/chapter/introduction-matrices-and-matrix-operations/>
- <https://courses.lumenlearning.com/boundless-algebra/chapter/introduction-to-matrices/>
- [https://math.libretexts.org/Bookshelves/Applied Mathematics/Book%3A Applied Finite Mathematics \(Sekhon and Bloom\)/02%3A Matrices/2.01%3A Introduction to Matrices](https://math.libretexts.org/Bookshelves/Applied_Mathematics/Book%3A_Applied_Finite_Mathematics_(Sekhon_and_Bloom)/02%3A_Matrices/2.01%3A_Introduction_to_Matrices)
- <https://byjus.com/maths/singular-matrix/>
- <https://www.onlinemathlearning.com/singular-matrix.html>



**Genomic Selection is  
Simply a tool to  
Supplement the  
Breeding Pipeline**

**Thank You**

**Questions**