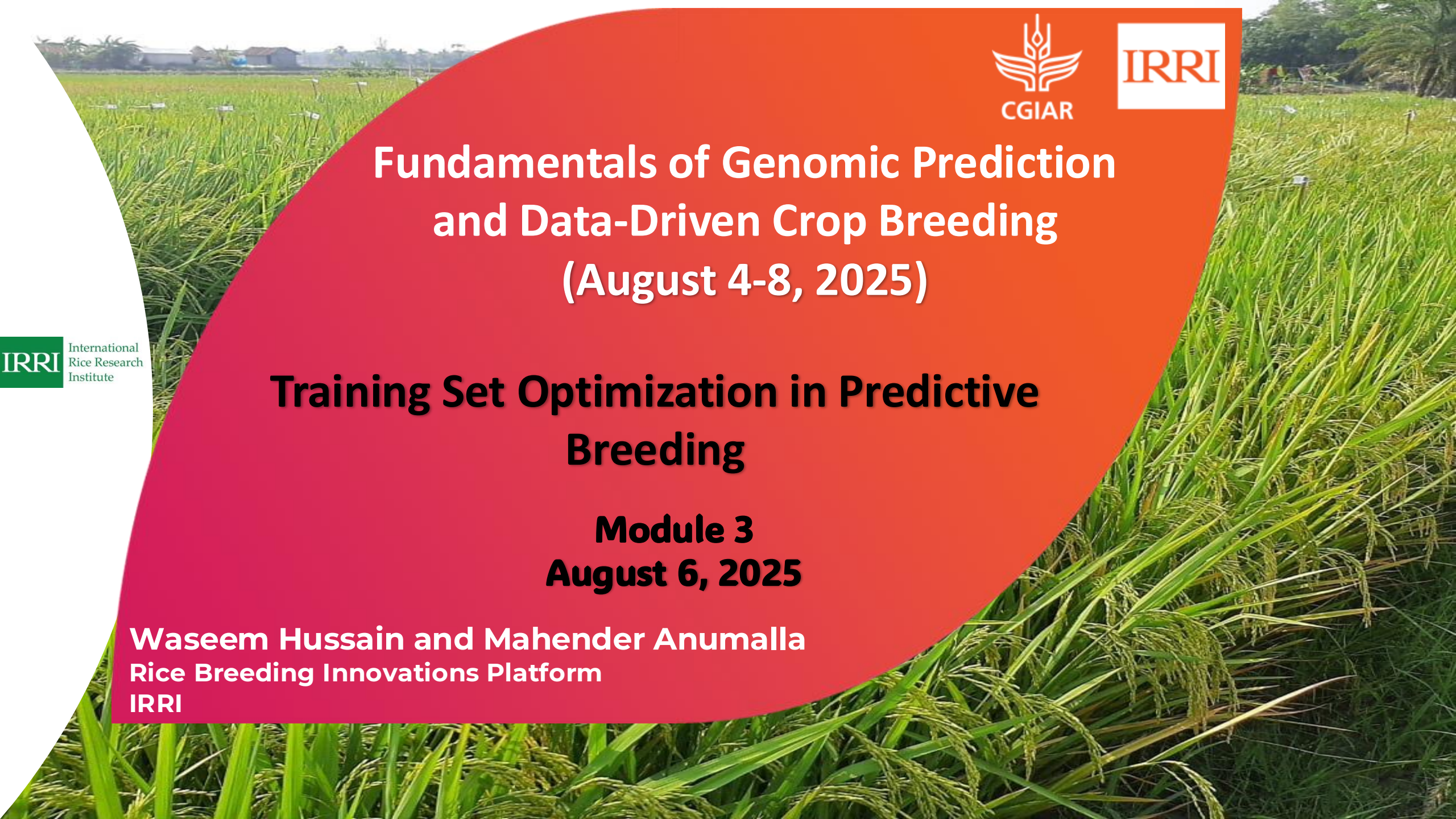


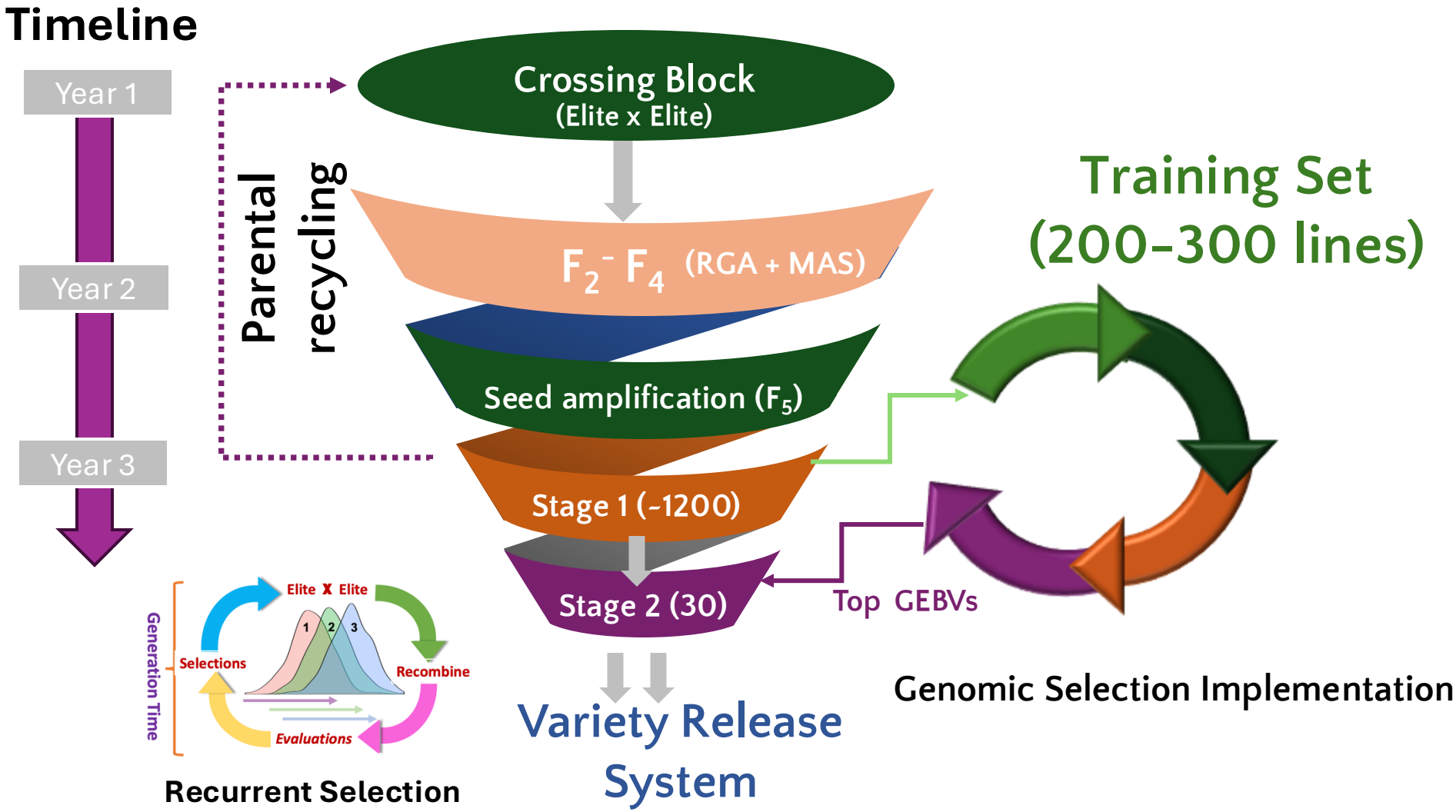
Fundamentals of Genomic Prediction and Data-Driven Crop Breeding (August 4-8, 2025)

Training Set Optimization in Predictive Breeding

Module 3
August 6, 2025

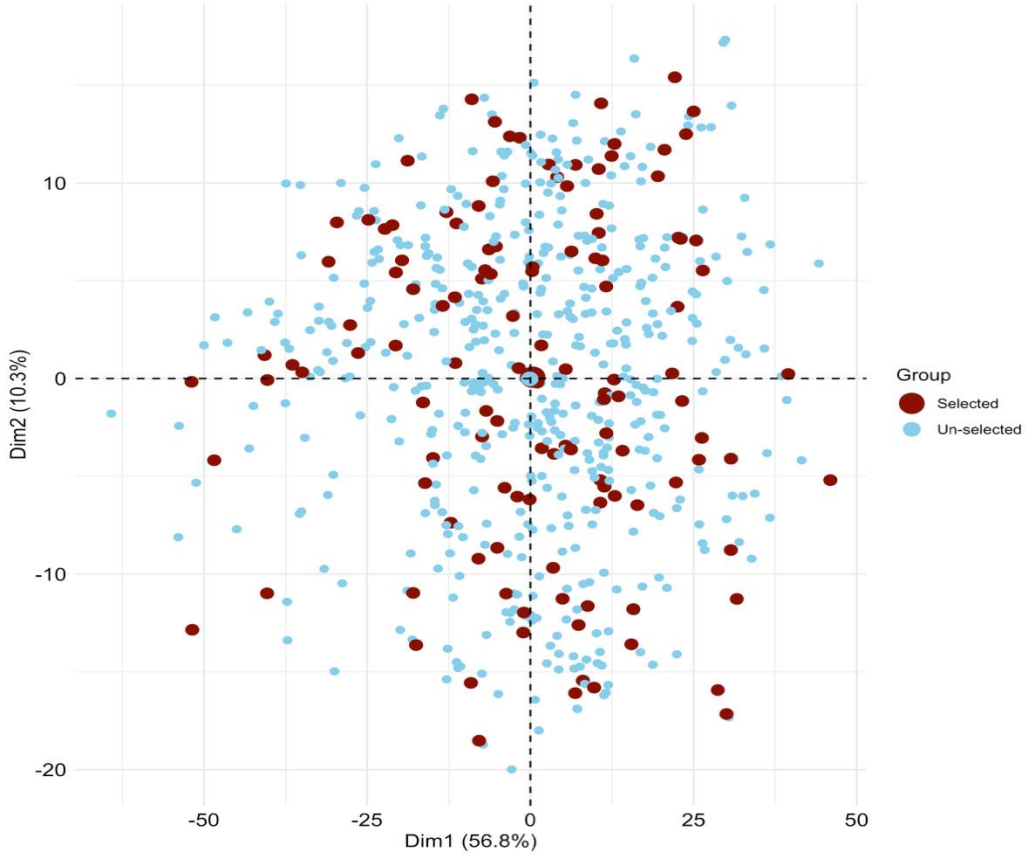
Waseem Hussain and Mahender Anumalla
Rice Breeding Innovations Platform
IRRI

General Breeding Pipeline



How to Design The Training Set

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



Example of IRRI's drought breeding program
Dark red represents the training set

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*

Rincent et al. 2012; Rincent et al. 2012

Prediction Error Variance (PEV)

(How Accurate we are!)

Prediction error variance (PEV): fraction of additive variance not accounted for by the prediction

$$PEV = var\,(\hat{a}_i - a_i) = (1 - r_i^2)\sigma_a^2$$

- The closer the PEV to true values, the closer the reliability is to 1.
- PEV depends upon the n, the individuals with more information have small PEV

$$\begin{matrix} X'X & X'Z \\ Z'X & Z'X + I\lambda \end{matrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

$$\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}^{-1} = \begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix}$$

Issue: Individuals are assumed to be unrelated, thus does not consider the decrease in genetic variance when close relatives are sampled

Now, $\mathbf{PEV} = \rho(\hat{a}_i, a_i) = C^{22}\sigma_e^2$

For ach level of random effect or individual breeding value $PEV_i = (d_i\sigma_e^2)$,

Where, d_i is the diagonal element of C^{22}

Critical Difference Mean (CD)

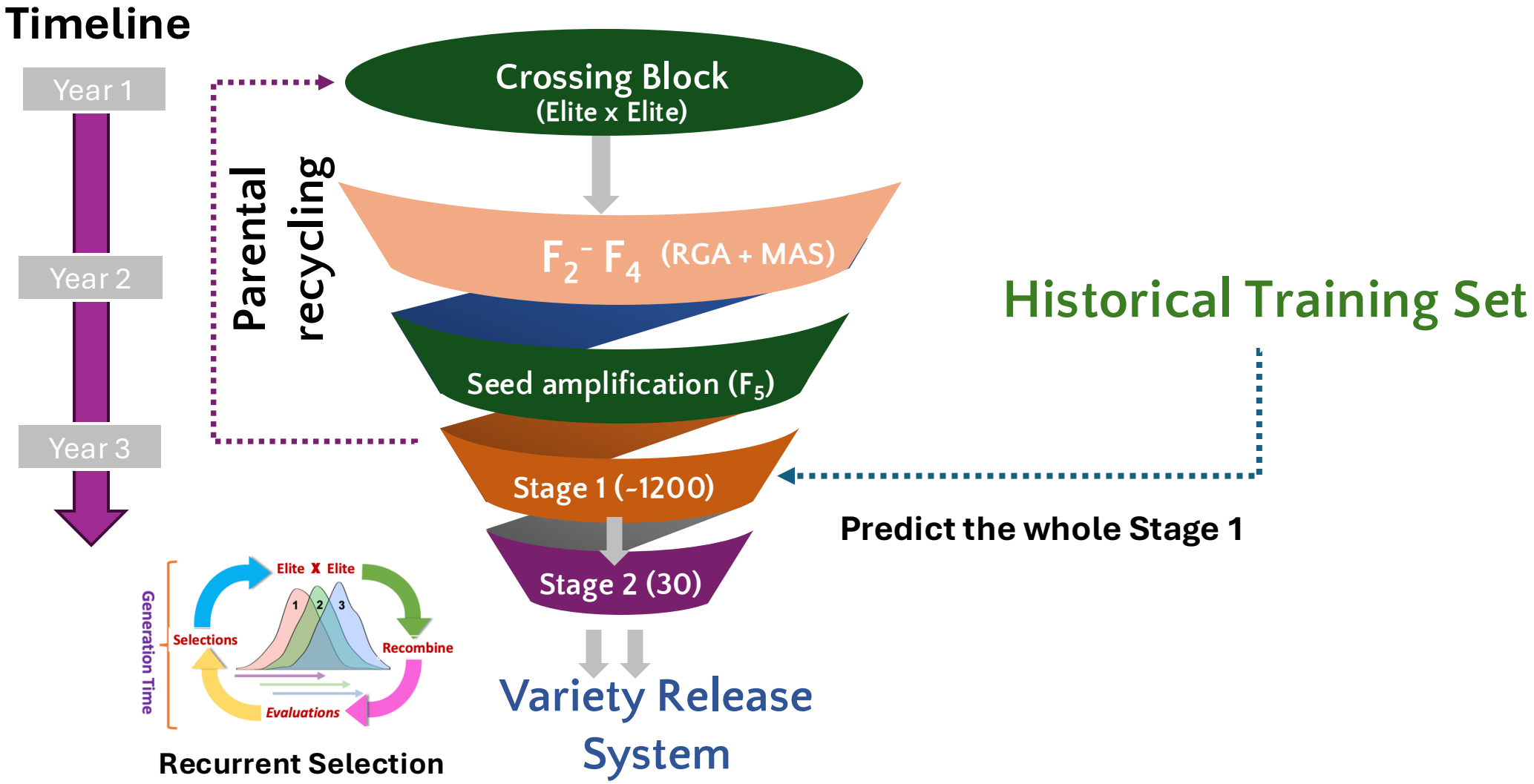
- Complementary Approach is the expected reliability of the prediction of contrasts
- CD is the expected reliability of the contrast (true and predicted)
- CD is a balance between PEV and genetic variance, which **takes into account the relationship**.
- Implemented with generalized Coefficient of Determination

$$\mathbf{CD}\,(\mathbf{c}) = \mathrm{diag}\left[\frac{\mathbf{c}'\left(\mathbf{A} - \lambda\left(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \lambda\mathbf{A}^{-1}\right)^{-1}\right)\mathbf{c}}{\mathbf{c}'\mathbf{A}\mathbf{c}}\right]$$

takes into account covariances between the candidate individuals

Note: CD Mean is superior over PEV Mean because it takes into account the reduction in variance due to the relatedness between individuals

General Breeding Pipeline



Training Population Update

- Over generations, recombination between markers and QTL will cause LD to decay, while selection and drift will potentially act to generate new LD or tighten the LD between closely linked loci
- Shifts in the pattern of QTL-marker LD, if not captured, will result in decreased prediction accuracy.
- Training populations must be updated during recurrent selection to maintain prediction accuracy

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5427505/

Additional Read

- https://link.springBer.com/article/10.1007/s00122-021-03916-w
- https://doi.org/10.3835/plantgenome2019.04.0028
- https://doi.org/10.1534/genetics.112.141473
- https://link.springer.com/article/10.1007/s00122-014-2418
- https://link.springer.com/article/10.1186/s12711-015-0116
- https://link.springer.com/article/10.1007/s00122-021-03916-w