2024 Soybean Breeders Workshop

Sparse Testing Designs at the Industrial Level: An Application in Soybeans

Diego Jarquin

February 12, 2024

OUTLINE

Predicting for CABBI and other related topics

Multi-Omics Integration and Allocation

Sparse Testing Designs at the Industrial Level

- Intro
- Resource allocation
- Training set sizes and composition
- Results and Conclusions

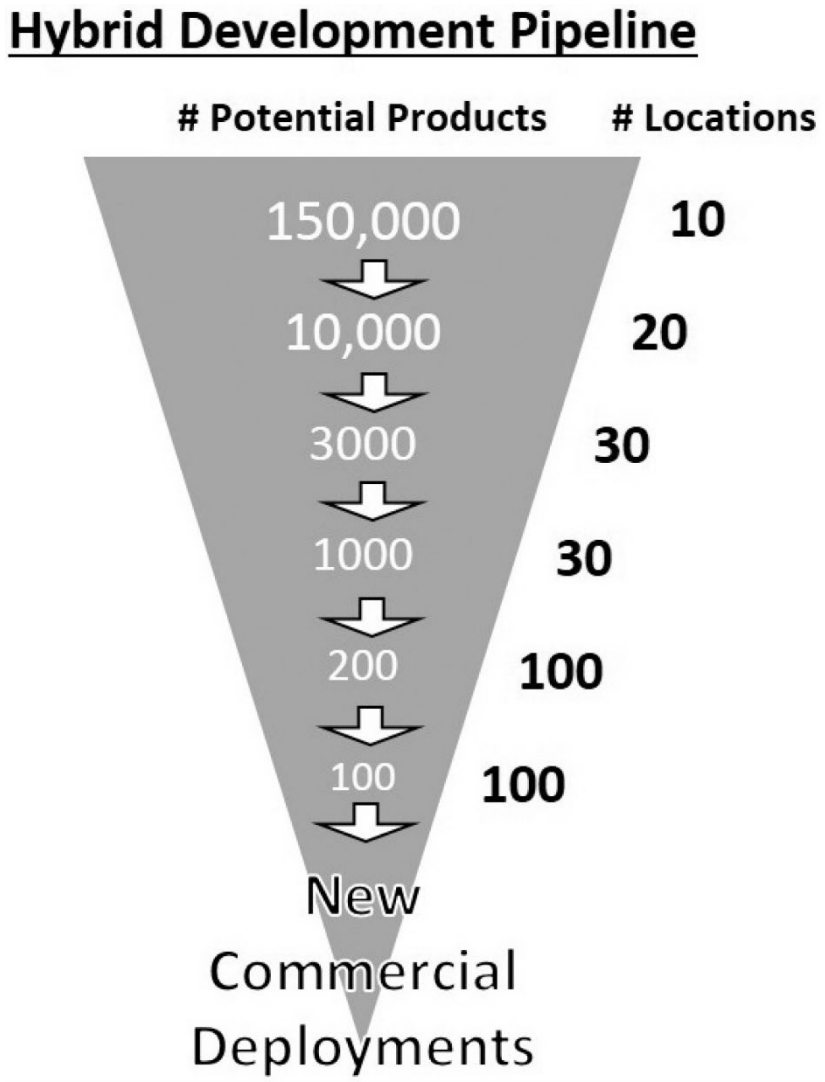
Sparse testing designs at the industrial level [Intro]

- Breeders are interested in the release (development) of stable genotypes that outperform current elite materials in a broad set of environments (**G×E matters**).

- Necessary to conduct multi-environmental trials.

- Budget constrains do not allow testing all genotypes in all environments.
    - Just a reduced number of the combinations (genotypes-environments).

Sparse testing designs at the industrial level [Intro]

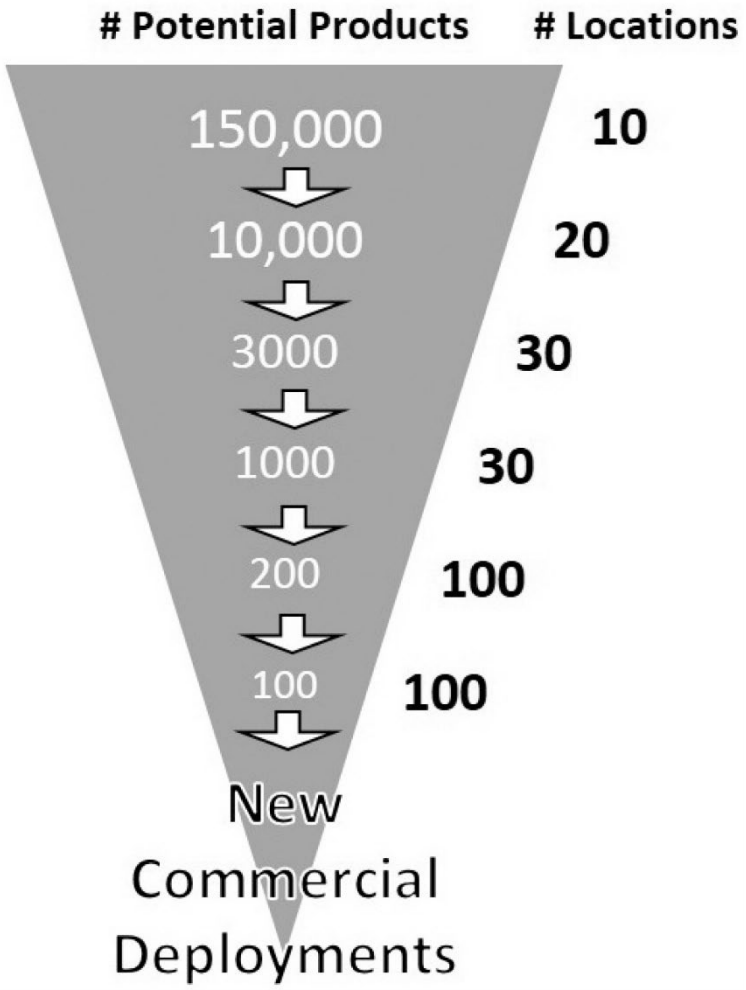


Glenn et al. 2017

- Initially, screening a large number of genotypes in few environments.
- As the more promising lines are advanced, these need to be tested in more environments.
- Only very few genotypes make it to the end of the breeding program (~3-5).
- These are released as commercial varieties.
- The genetics of those genotypes at the bottom is the same than at the top.
- Hence, a method for identifying these promising genotypes at early stages would help to speed up the breeding cycle.

Sparse testing designs at the industrial level [Intro]

Hybrid Development Pipeline

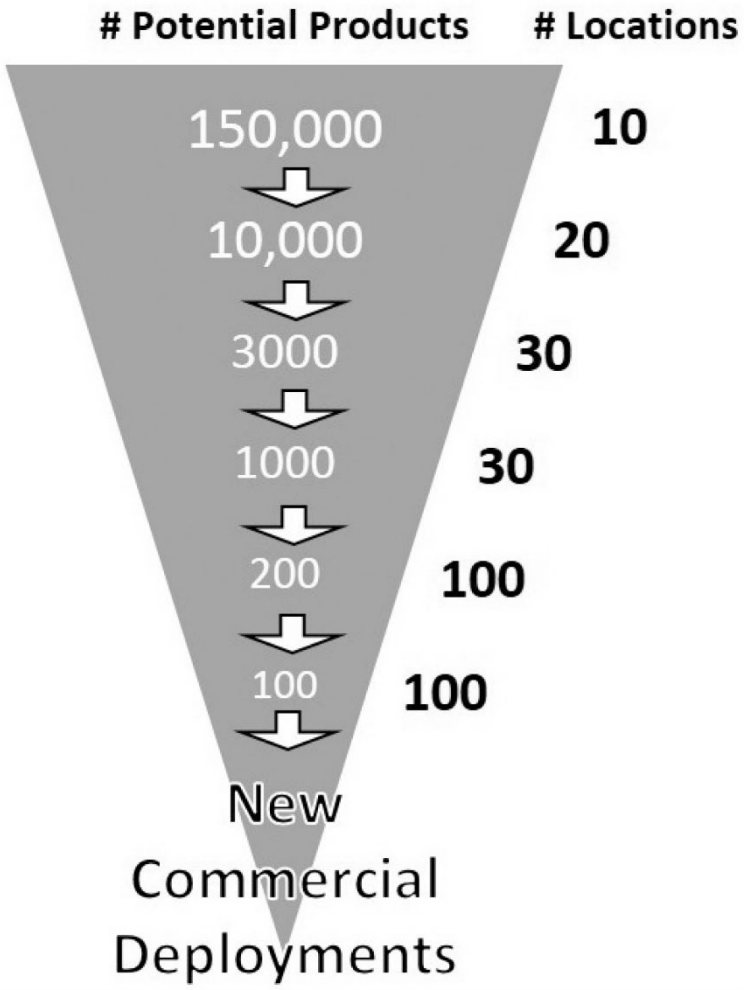


- Ideal scenario (nonrealistic): Testing all genotypes in all environments.
  - 150,000 genotypes in 100 environments/locations = 15,000,000 combinations.
- Total number of "phenotypes" along the pipeline ~ 1,760,000
  - 150,000 × 10 = 1,500,000
  - 10,000 × 20 = 200,000
  - 1,000 × 30 = 30,000
  - 200 × 100 = 20,000
  - 100 × 100 = 10,000

Glenn et al. 2017

Sparse testing designs at the industrial level [Intro]

Hybrid Development Pipeline

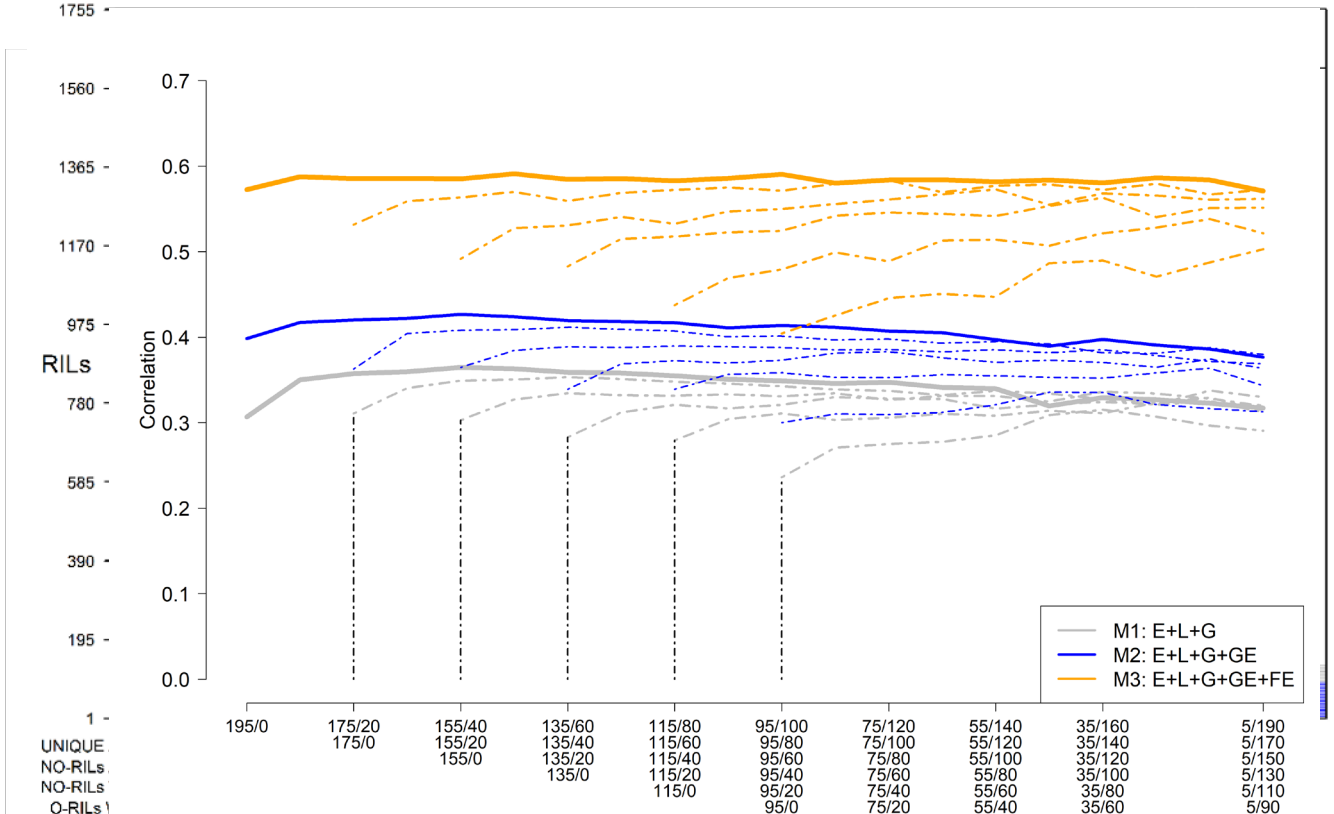


Glenn et al. 2017

- A more convenient allocation would help us to "evaluate" all genotypes (150,000) in all environments/locations (100) using the same resources.
- This would help us to find the most "promising" genotypes in less time.
- For example, sparsely observe the 150,000 genotypes in the 100 locations (~1,760,000 phenotypes) and predict the remaining combinations (13,240,000).
- We are already dedicating recourses to phenotype 11.73% of all these combinations along pipeline.
  - ~1,760,000
- Using the same budget, the use of sparse testing designs could help us to assess the convenience of testing a fraction of these combinations and predicting the remaining ones for selection of superior cultivars saving time.

Sparse testing designs at the Industrial level [Resource Allocation]

- **Objectives**
  - Assess the convenience of implementing sparse testing designs to reduce the number of years for selecting superior cultivars at a given budget.
  - Evaluate different methods to select training sets (composition and sizes) at the industrial level.
- **Preliminary results**
  - Previously, we implemented sparse testing designs in maize (Jarquin et al., 2020), wheat (Crespo-Herrera et al., 2021), and soybean (Persa et al., 2023) considering a large number of genotypes (851-1,755) in a reduced number of environments (3-9).

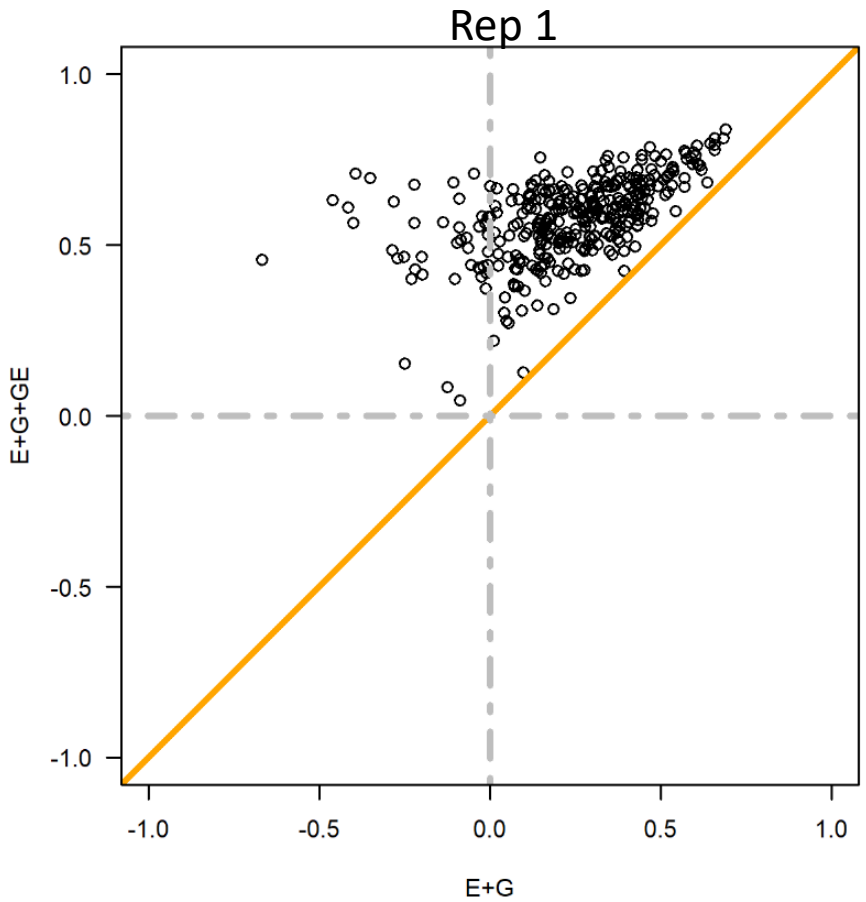


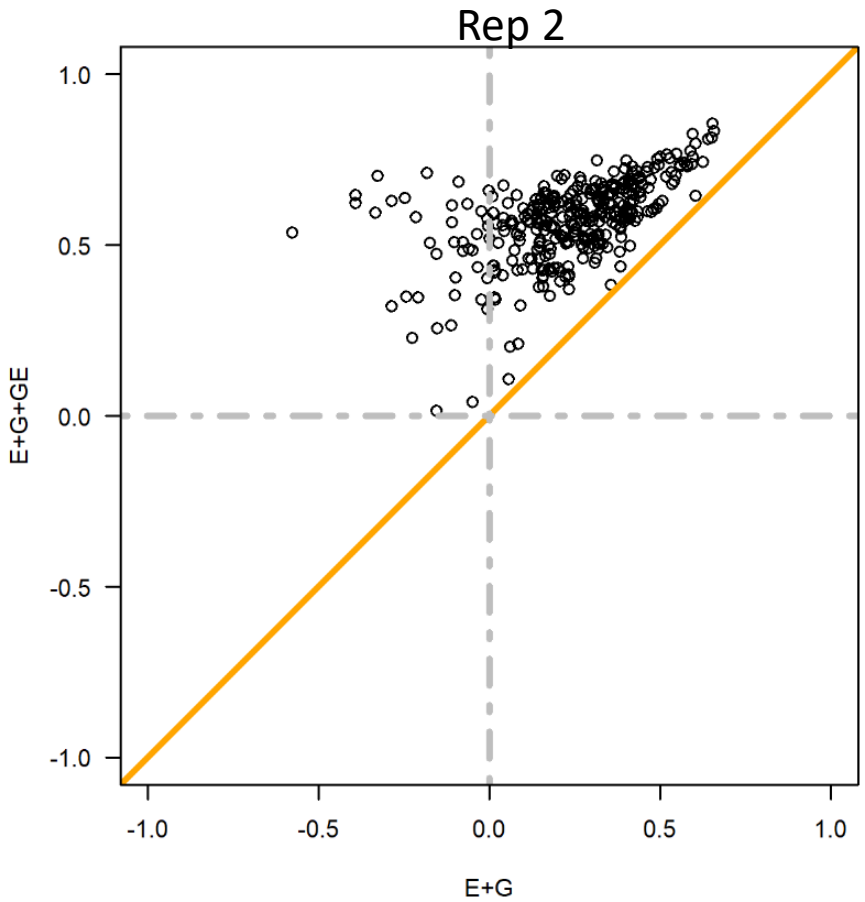
Sparse testing designs at the industrial level [Training set sizes and composition]

- **Data description**
  - Large soybean dataset (“The Company”) comprised of 2,500 genotypes tested in 340 environments.
  - All genotypes tested in all environments. Total number of datapoints: 850,000 yield records).
  - Genomic data on 2,300 marker SNPs was also available.

Sparse testing designs at the industrial level [Results and Conclusions]

- Considering 2% of the combinations for model training.
  - 17,000 of the 850,000 combinations to predict the remaining 833,00 [98%] (or 2,450 genotypes per environment).
  - 50 genotypes observed within environment (7 unique and 43 overlapped with the adjacent environment).





Scatterplot of the correlations of the 340 environments considering two models.

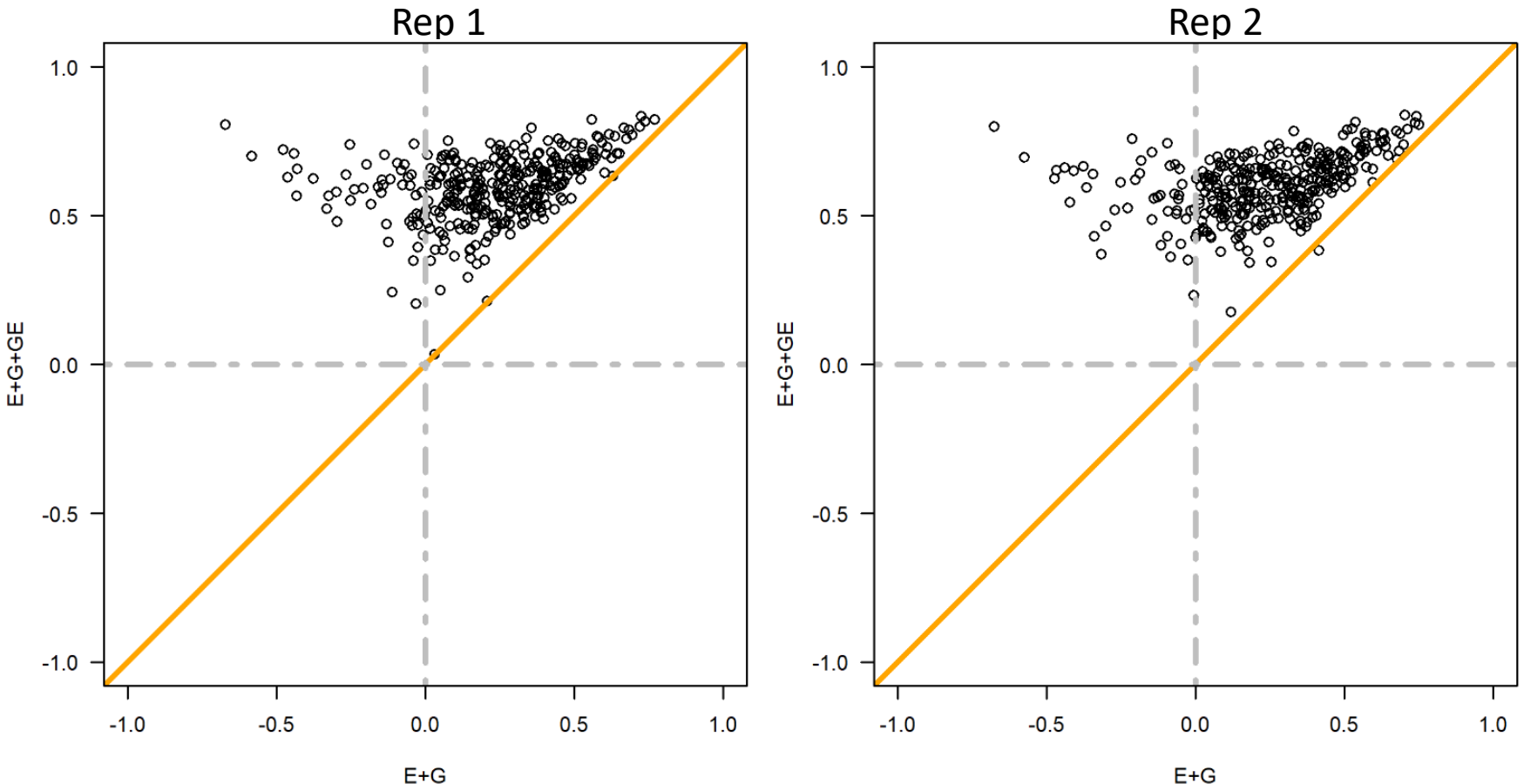
GBLUP [$\rho = 0.236$]

vs.

Reaction Norm [$\rho = 0.579$]

Sparse testing designs at the industrial level [Results and Conclusions]

- Considering 2% of the combinations for model training
  - 17,000 of the 850,000 combinations to predict the remaining 833,00 [98%] (or 2,450 genotypes per environment).
  - 50 genotypes randomly selected and observed across all the 340 environments (full overlap).



Scatterplot of the correlations of the 340 environments considering two models.

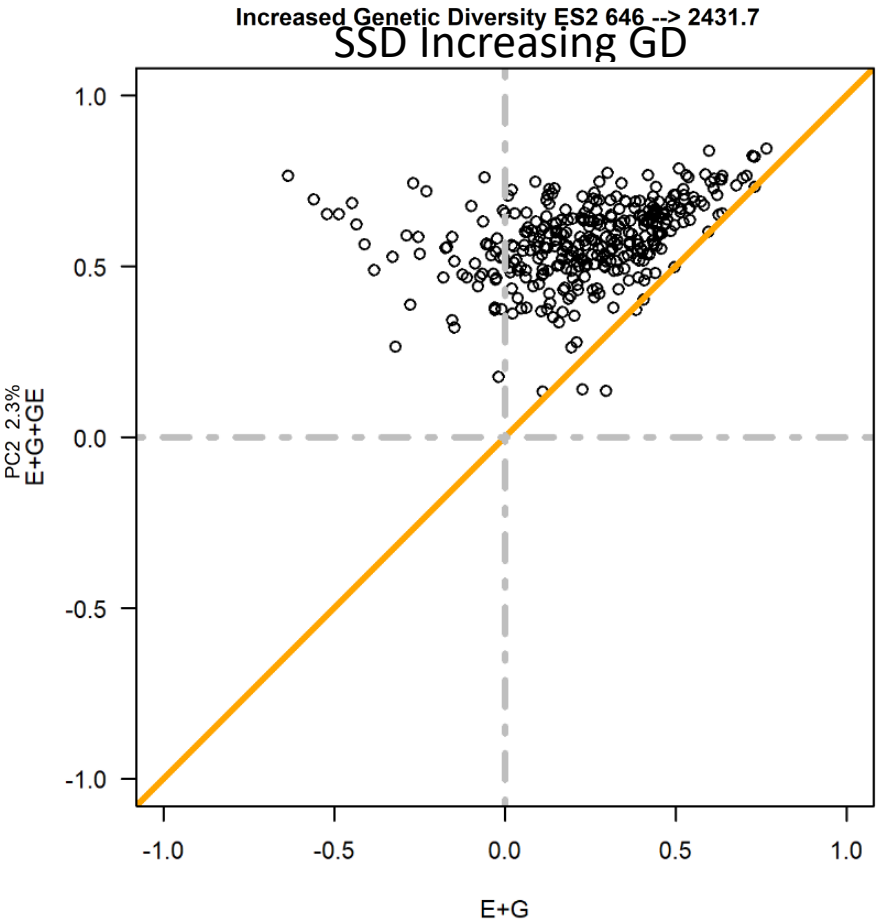
GBLUP [$\rho = 0.294$]

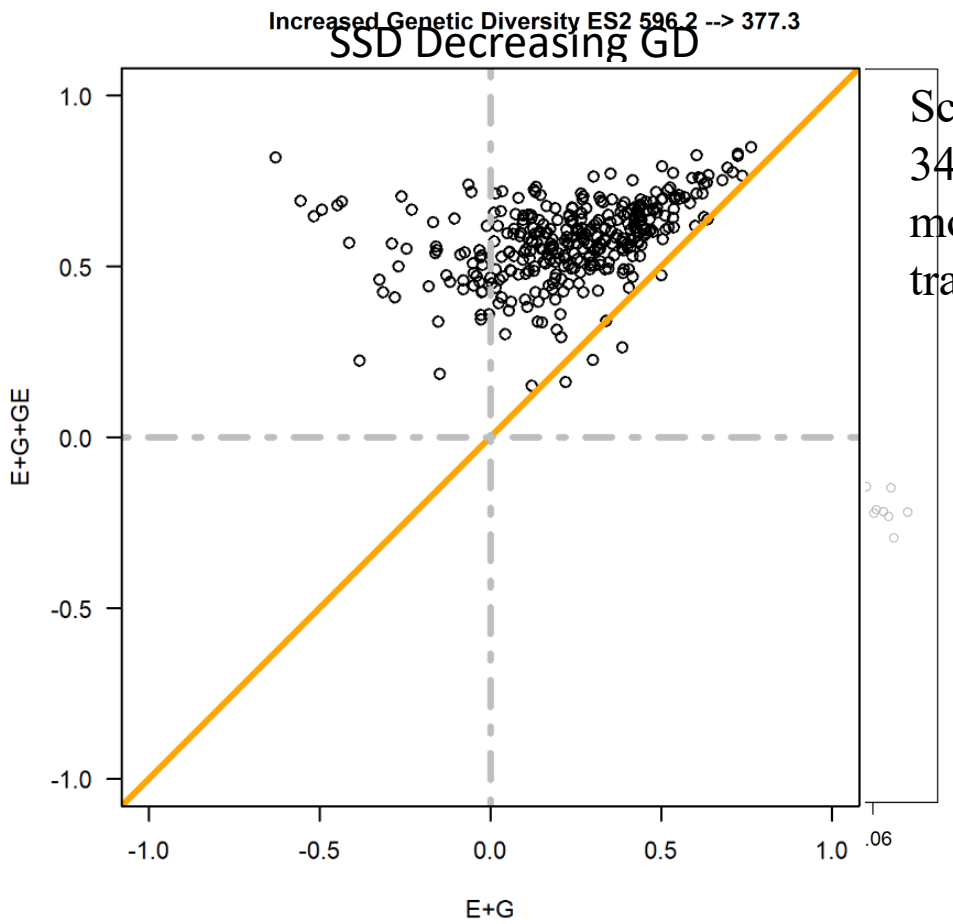
vs.

Reaction Norm [$\rho = 0.599$]

Sparse testing designs at the industrial level [Results and Conclusions]

- Considering 2% of the combinations for model training
  - 17,000 of the 850,000 combinations to predict the remaining 833,00 [98%] or (2,450 genotypes per environment).
  - 50 genotypes selected using the super saturated design for increasing/decreasing genomic diversity (full overlap) of the training set.





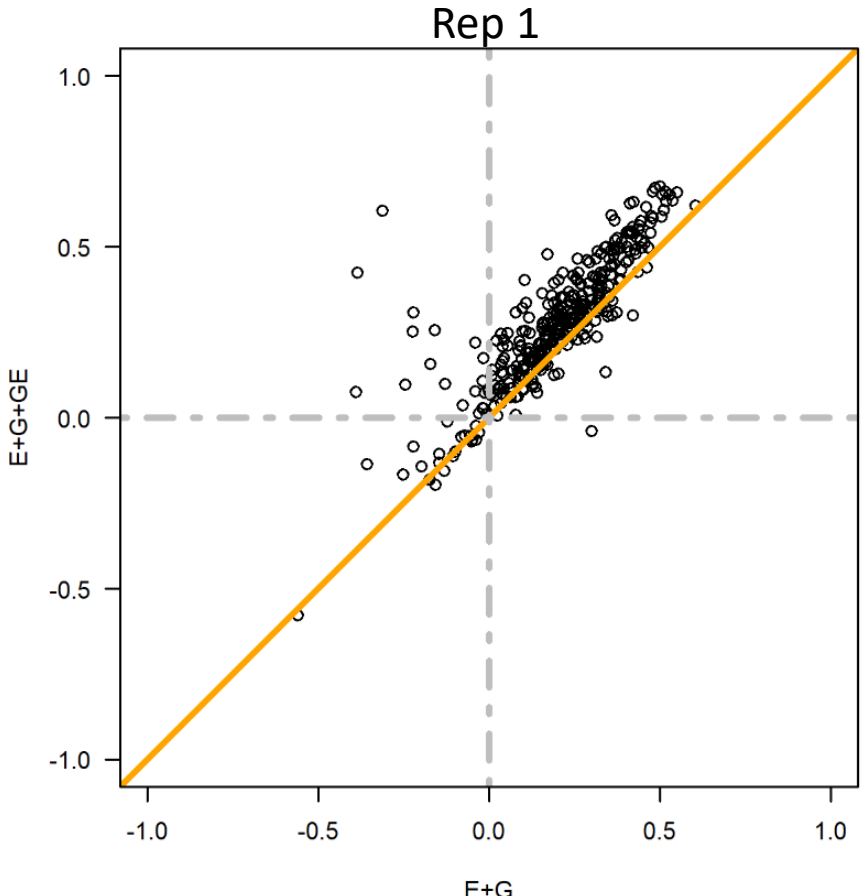
Scatterplot of the correlation of the 340 environments considering two models and two manners to select training sets.

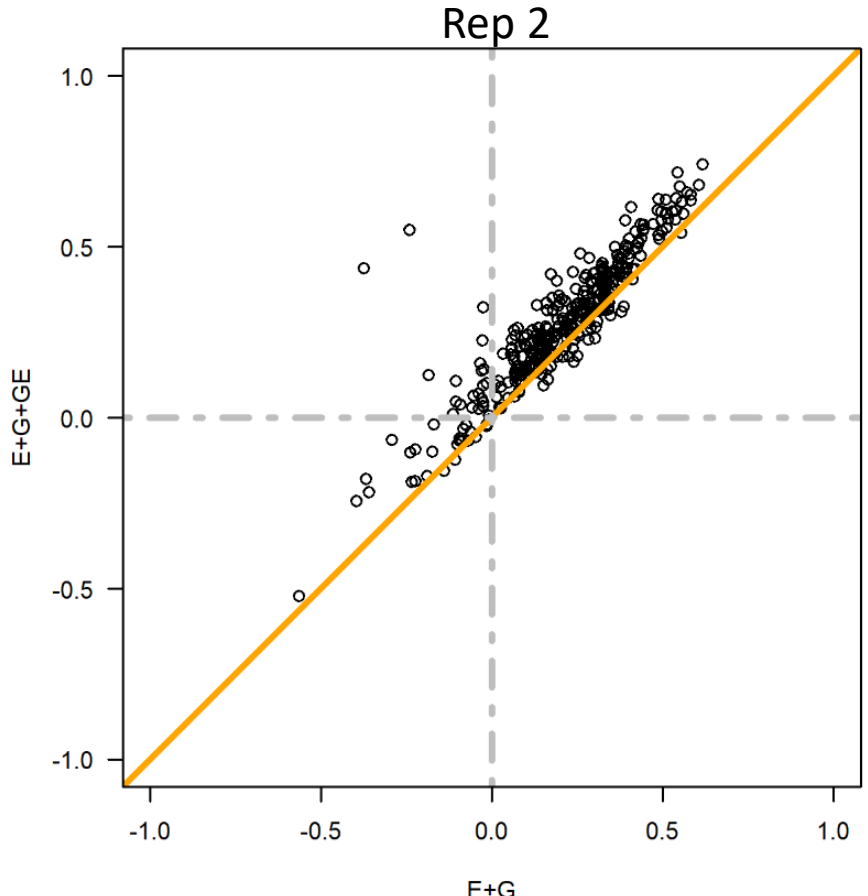
Increasing GD
GBLUP [$\rho = 0.232$]
vs.
Reaction Norm [$\rho = 0.576$]

Decreasing GD
GBLUP [$\rho = 0.233$]
vs.
Reaction Norm [$\rho = 0.575$]

Sparse testing designs at the industrial level [Results and Conclusions]

- Considering 0.3% of the combinations for model training
  - 2,500 of the 850,000 combinations to predict the remaining 847,500 [99.7%] (or 2,493 genotypes per environment).
  - Between 7 and 8 genotypes (randomly selected) were observed at each environment (each genotype was observed only once across environments).





Scatterplot of the correlation for each one of the 340 environments considering two models.

GBLUP [$\rho = 0.200$]

vs.

Reaction Norm [$\rho = 0.278$]

Sparse testing designs at the industrial level [Results and Conclusions]

- The identification of superior cultivars can be accelerated by implementing sparse testing designs.
- No need to increase budget for already large-scale programs.
- For a given budget, the screening/testing capacity can be easily increased by 10 folds (e.g., current budget 5,000 phenotypes).
  - For a target number of 10 environments: “screening/evaluate” 5,000 genotypes instead of considering 500 (500 genotypes per environment).
  - For a target number of 500 genotypes: “screening/evaluate” 100 environments instead of 10 (50 genotypes per environment).
  - Or combinations of these (e.g., 250 genotypes and 20 environments).
- For a target population of genotypes and environments the phenotyping costs can be reduced up to 90% (500 genotypes & 10 Environments).
  - Screening 50 genotypes per environment (total 500 across 10 environments = 5,000 phenotypes) and predict the remaining 450 (4,500 across environments).
  - 500 phenotypes instead of 5,000.

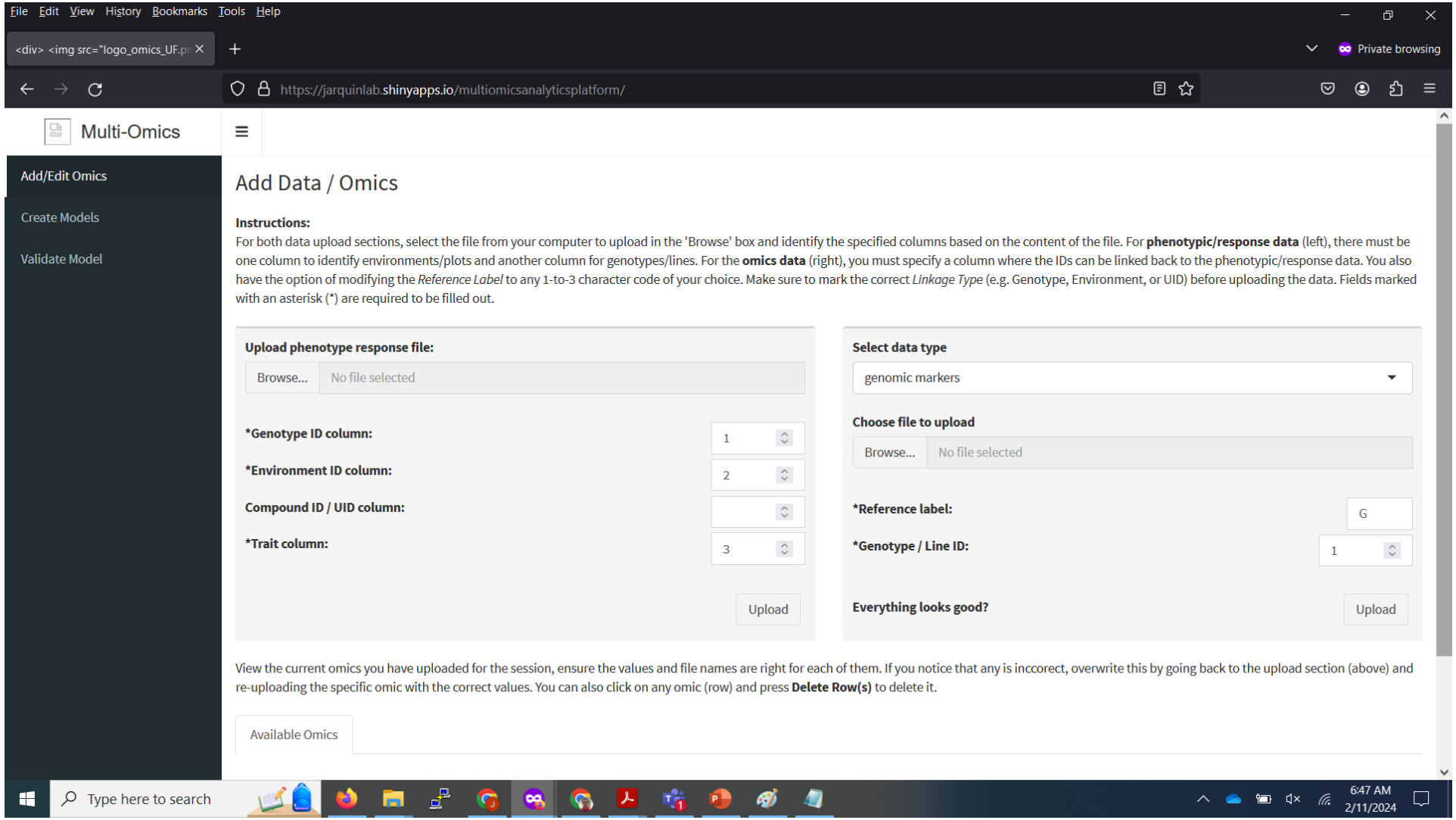
ChIDO: Characterization & Integration of Driven Omics

A no-code solution to build models with interaction matrices

Francisco Gonzalez
MS student
Modern Apps Lead - Google

Julian Garcia-Abadillo
PhD student
Intern at Google

https://jarquinlab.shinyapps.io/multiomicsanalyticsplatform/



Visit Gainesville this summer

90 attendees – 12 Countries 5 continents
– 26 States USA

This year we are partnering with Google, NCSU, UG, and UARK

THANK YOU

GO GATORS!!!

Questions?

diego.jarquin@gmail.com
jhernandezjarqui@ufl.edu