

|  |
| --- |
| Strategies to improve adaptation in plant breeding with complex genotype by environment interactions  Bioinformatics Research Proposal |

|  |
| --- |
| Student Exam Number: B244770  I **have** used AI-based methods (e.g. ChatGPT) in the preparation of this Research Proposal |

# Abstract

Plant and animal breeding are inherently complex due to the presence of genotype by environment (GxE) interaction, which manifests as differential responses of individuals (genotypes) to different environments. In plant breeding, the assessment of GxE interaction is critical as they significantly influence the adaptability and yield of crop genotypes under varying environmental conditions. This project aims to evaluate different experimental designs within and across a range of environments to optimize the detection and measurement of GxE interaction. The experimental designs will be evaluated using stochastic simulations, which enable us to modify a large number of parameters and compare numerous scenarios in real time. By maintaining a fixed number of experimental plots, our research focuses on identifying optimal configurations that improve the accuracy of genotype performance within and across environments. Through a strategic variation of genotype replication (classified as replicated twice, once, or not at all), this project seeks to determine the most effective method to capture essential data on environmental responsiveness, thus enhancing the precision of selection for better-adapted crop genotypes. The outcomes of this research are expected to provide valuable insights into the design of breeding programs, facilitating the development of crop genotypes with greater stability and productivity across diverse farming conditions.

**Keywords**: **Plant Breeding, Genotype by Environment (GxE) Interaction, Statistical Modelling, Experimental Design, Stochastic Simulation**

# BACKGROUND

Genotype by environment (GxE) interaction significantly complicate plant breeding by affecting phenotypic stability across diverse environments, challenging the predictability of breeding outcomes1,2. GxE interaction represents the differential response of individuals (genotypes) to different environments3. GxE interaction can be broadly classified as either non-crossover or crossover interaction, which represent changes in scale performance between environments or changes in rank4. Crossover interaction is of particular importance to plant breeders because their selection decisions are more complicated by changes in rank compared to changes in scale5.

Multi-environment trials (METs) are fundamental in assessing genotype by environment (GxE) interactions, which are key to understanding the adaptability and yield stability of crop genotypes across varied conditions. These trials are integral to plant breeding strategies, particularly when employing innovative experimental designs like sparse testing and partially replicated ("p-rep") designs. Sparse testing strategically evaluates genotypes in a limited subset of potential environments to efficiently gather data without extensive resource investment, thereby broadening environmental sampling and capturing critical GxE interactions6–8. Partially replicated designs enhance trial efficiency by focusing resources on replicating a select group of genotypes while assessing others in single-plot trials, improving data reliability and the accuracy of performance assessments9. These approaches are scrutinized through simulations and statistical analyses to verify their effectiveness in enhancing the precision of genotype selection (e.g. Figure 1), aiming to develop more stable and productive crop varieties suited to diverse agricultural settings.

Stochastic simulation is an essential tool for evaluating various experimental designs in plant breeding programs, particularly when addressing the complexities of genotype by environment interactions and the effectiveness of sparse testing. This method offers a practical and cost-effective way to assess and compare multiple breeding scenarios over extended periods, which might be unfeasible in real-world settings due to resource and time constraints. By simulating breeding programs under diverse design and management strategies, researchers can predict long-term outcomes, optimize breeding processes, and refine decision-making without conducting extensive field trials. Software tools like AlphaSimR and FieldSimR are instrumental in these simulations, allowing breeders to model complex scenarios across multiple generations10,11. This is crucial for developing strategies that enhance the adaptability and yield stability of crops under varied environmental conditions. Specifically, FieldSimR generates realistic plot data for multi-environment trials, capturing spatial variations and environmental effects that significantly impact phenotype expressions. This functionality is vital for testing and optimizing sparse testing designs, where genotypes are not uniformly tested across all environments. Through these simulations, breeders can anticipate the long-term effects of specific experimental designs, thereby enhancing the precision and adaptability of crop genotypes to diverse agricultural conditions.

The aim of this project is to evaluate different experimental designs within and across multi-environment trials (METs), crucial for evaluating the GxE interactions effectively. The objective is to optimise replication configurations to ensure that genotypic performance assessments are robust and reflective of a genotype's potential across different environmental scenarios. By doing so, this research aims to contribute to the development of plant genotypes that are not only high-yielding but also resilient and well-adapted to a range of environmental conditions, thereby enhancing agricultural sustainability and productivity in the face of global climatic challenges1,2.

# method

## Description of Experimental Designs

In this plant breeding project, we plan to utilize a sparse testing approach that involves the strategic distribution of a selected number of genotypes across multiple environments. This approach is specifically designed to efficiently explore genotype by environment (GxE) interactions and to maximize the use of limited resources.

手机屏幕的截图

中度可信度描述已自动生成

Figure 1 The visualizations depict three replicates each of sparse and partially replicated experimental designs on a 10x10 grid. In the sparse design (left column), half the plots are randomly filled, simulating a scenario where genotypes are tested across a limited number of environments. In the partially replicated design (right column), most plots are filled, representing a strategic approach where certain genotypes are tested multiple times for enhanced data reliability. Each replicate illustrates a potential layout variation within a field trial.

The experimental design will start with an initial set of genotypes, for instance, 1,000, which will be evaluated under a range of environmental conditions. The allocation of genotypes to environments will be based on a strategy that aims to reveal the most information about each genotype's adaptability and performance across a spectrum of conditions:

### Initial Assessment:

All genotypes will initially be assessed in a limited number of key environments, typically one or two, to establish a baseline understanding of their performance.

### Scaled Expansion:

As the study progresses, the number of environments will be incrementally increased. This expansion is not fixed but will be adapted based on preliminary results and resource availability. For example, from testing in two environments initially, we might expand to five or more, depending on the diversity of environments available and their relevance to the breeding objectives.

### Dynamic Allocation:

Genotypes will be dynamically allocated to environments based on their performance and stability in initial tests. This approach allows us to focus resources on genotypes showing promise under specific conditions, thus enhancing the efficiency of the breeding program.

### Integration of Partial Replications:

To further optimize the design, partial replications will be incorporated where only a subset of promising genotypes will be selected for additional testing across an expanded set of environments. This method will balance the need for comprehensive data collection with resource constraints, focusing on potential high-performers under varied environmental stresses.

This adaptive experimental design will enable us to systematically explore a wide range of GxE interactions while managing the logistical and financial constraints typical of large-scale breeding programs. By employing an iterative approach to environment selection and genotype testing, we aim to refine our understanding of genotype adaptability and improve the precision of our selection processes.

## Overview of Simulation and Data Generation

In this project, advanced simulation tools FieldSimR and AlphaSimR will be utilized to study the effects of genotype by environment interactions on grain yield in inbred wheat genotypes. These tools are key for representing the detailed genetic architecture of wheat and the complex dynamics typical of agricultural environments.

### Generating Plot Errors and Genetic Values:

FieldSimR will simulate plot-specific errors for agricultural field trials, accurately capturing the different types of variability that impact experimental outcomes (Figure 1):

图表, 树状图

描述已自动生成

Figure 2 Example of plot\_effects(), showing the total plot errors for environment 1

* Spatial Errors: These will reflect structured variability due to environmental gradients within the field, such as soil fertility and moisture levels, which affect plant growth and development systematically across the trial landscape.
* Random Errors: Representing unstructured, stochastic noise that occurs naturally in field data, these errors account for unexplained variations that are not due to observable environmental factors, often arising from measurement inaccuracies or other random disturbances.
* Extraneous Errors: These will model non-environmental systematic variations that can influence trial outcomes, such as plot position effects or artifacts introduced by specific agricultural practices like the layout of irrigation systems or the pattern of planting.

Simultaneously, AlphaSimR will be used to generate genetic values employing an additive genetic model. This model integrates essential genetic parameters, including allele frequencies and linkage disequilibrium, to simulate breeding values that represent the potential yield under varying genetic and environmental scenarios.

### Constructing Phenotypes:

The process will integrate the simulated plot errors with genetic values to create realistic phenotypes. This integration will use both FieldSimR’s capabilities and custom R scripts to accurately merge genetic data with environmental effects. This step ensures that our simulation output not only reflects the natural variability seen in field trials but also incorporates the complex interactions between genotype and environment. The precision in merging these data layers is crucial for simulating realistic outcomes that can subsequently inform breeding decisions.

## Modeling and Evaluation

In our study, we will fit various statistical and machine learning models to evaluate and predict genotype effects under diverse environmental conditions. Our objective is to understand how these models can accurately forecast genotype performance both within specific environments and across different settings, which is crucial for effective plant breeding.

### Model Implementation:

* Best Linear Unbiased Prediction (BLUP): This traditional statistical method, fitted within a linear mixed model framework, will predict random effects leveraging available trial data. BLUP is foundational for estimating genetic values based on phenotypic and pedigree data.
* Pedigree Best Linear Unbiased Prediction (PBLUP): To enhance BLUP's predictions, PBLUP incorporates pedigree information. This model improves prediction accuracy by using known genetic relationships, which is beneficial in environments with documented lineage information.
* Genomic Best Linear Unbiased Prediction (GBLUP): Advancing beyond PBLUP, GBLUP uses genomic similarity matrices derived from SNP data. This model offers a more precise estimation of genetic relationships, improving predictions in studies with extensive genomic data.
* Factor Analytic Genomic Best Linear Unbiased Prediction (FA-GBLUP): Specializing in multi-environment trials, FA-GBLUP adds a factor analytic layer to GBLUP to address genotype-by-environment interactions. This method is particularly useful for parsing complex interactions that influence phenotypic variance across diverse conditions.
* Machine Learning Methods: We will also deploy advanced machine learning techniques such as support vector machines, neural networks, and ensemble methods like random forests and gradient boosting machines. These methods are selected for their capacity to model non-linear relationships and interactions more effectively than traditional models, providing nuanced insights into genetic data.

### Evaluation of Model Performance:

Each model will be evaluated for its accuracy in predicting genotype effects. We will measure the models' ability to predict phenotypic outcomes based on genetic inputs within individual environments and assess their performance across multiple environments. The accuracy of these predictions is vital for understanding which models are most effective in different agricultural contexts, thus guiding breeders in selecting optimal genotypes.

## Data Visualisation

In this study, we extensively used visualization tools to enhance our understanding of genotype by environment (GxE) interactions and the complex dynamics present in field trial simulations. One of the key techniques involved using FieldSimR's plot\_matrix() function to visually represent these interactions, which are critical for discerning how different genotypes respond under varying environmental conditions. This function allowed us to visualize the performance of each genotype across a range of environmental setups, clearly showing where certain genotypes performed better or worse. The matrix format of the visualization helped in quickly identifying patterns and anomalies in genotype responses, facilitating a deeper analysis of environmental impacts on phenotypic traits.

The visual outputs from the plot\_matrix() function provided a clear, intuitive understanding of the GxE interactions by mapping the performance of genotypes across different environmental conditions in a grid-like structure. Each cell in the matrix represented a specific environment-genotype combination, colored or shaded according to the phenotype observed. This method of visualization not only highlighted the overall trends and outliers but also pinpointed specific genotype-environment combinations where unique or unexpected responses occurred.

# Working Plan

* **Weeks 1-2 (May 1-14): Initial Simulation Setup**
  + **Simulation Software Setup**: Install and configure FieldSimR and AlphaSimR, ensuring all packages are updated and operational.
  + **Parameter Definition**: Define the simulation parameters for the plant breeding trial, including genotype numbers, plot dimensions, and environmental settings.
* **Weeks 3-4 (May 15-28): Data Generation and Error Simulation**
  + **Sequence Generation**: Start generating sequences using AlphaSimR for the base genetic data.
  + **Plot Error Simulation**: Begin simulating plot errors with FieldSimR to include spatial, random, and extraneous variability.
* **Weeks 5-6 (June 1-14): Genetic Data Simulation and Model Preparation**
  + **Genetic Value Simulation**: Complete the genetic value simulation in AlphaSimR and prepare datasets for model integration.
  + **Model Parameterization**: Start parameterizing the model to fit genetic data against simulated plot errors.
* **Weeks 7-8 (June 15-28): Integration and Preliminary Analysis**
  + **Data Integration**: Combine genetic values with plot errors to generate phenotypic data.
  + **Initial Data Analysis**: Conduct preliminary analysis on the integrated data to check for consistency and expected trends.
* **Week 9 (June 29 - July 5): Heritability Estimation**
  + **Heritability Analysis**: Calculate plot-level and line-mean heritability from the simulated phenotypic data.
  + **Model Refinement**: Refine model parameters based on heritability outcomes.
* **Week 10 (July 6-12): Advanced Data Analysis and Visualization**
  + **Advanced Statistical Analysis**: Perform in-depth statistical analysis on the simulated data to identify patterns of genotype performance across different environmental conditions.
  + **Data Visualization**: Create visualizations of the simulated trial data to facilitate better understanding and interpretation of the results.
* **Weeks 11-11.5 (July 13-15): Dissertation Drafting**
  + **Writing the Methodology Section**: Begin drafting the Methods section of the dissertation based on the simulated results and analysis conducted thus far.

References

(1) Hajjarpoor, A.; Nelson, W. C. D.; Vadez, V. How Process-Based Modeling Can Help Plant Breeding Deal with G x E x M Interactions. *Field Crops Research* **2022**, *283*, 108554. https://doi.org/10.1016/j.fcr.2022.108554.

(2) Egea-Gilabert, C.; Pagnotta, M. A.; Tripodi, P. Genotype × Environment Interactions in Crop Breeding. *Agronomy* **2021**, *11* (8), 1644. https://doi.org/10.3390/agronomy11081644.

(3) Allard, R. W.; Bradshaw, A. D. Implications of Genotype‐Environmental Interactions in Applied Plant Breeding 1. *Crop Science* **1964**, *4* (5), 503–508. https://doi.org/10.2135/cropsci1964.0011183X000400050021x.

(4) Gail, M.; Simon, R. Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. *Biometrics* **1985**, *41* (2), 361. https://doi.org/10.2307/2530862.

(5) Yang, R. ‐C.; Baker, R. J. Genotype‐Environment Interactions in Two Wheat Crosses. *Crop Science* **1991**, *31* (1), 83–87. https://doi.org/10.2135/cropsci1991.0011183X003100010021x.

(6) Jarquin, D.; Howard, R.; Crossa, J.; Beyene, Y.; Gowda, M.; Martini, J. W. R.; Covarrubias Pazaran, G.; Burgueño, J.; Pacheco, A.; Grondona, M.; Wimmer, V.; Prasanna, B. M. Genomic Prediction Enhanced Sparse Testing for Multi-Environment Trials. *G3 Genes|Genomes|Genetics* **2020**, *10* (8), 2725–2739. https://doi.org/10.1534/g3.120.401349.

(7) Atanda, S. A.; Govindan, V.; Singh, R.; Robbins, K. R.; Crossa, J.; Bentley, A. R. Sparse Testing Using Genomic Prediction Improves Selection for Breeding Targets in Elite Spring Wheat. *Theor Appl Genet* **2022**, *135* (6), 1939–1950. https://doi.org/10.1007/s00122-022-04085-0.

(8) Montesinos-López, O. A.; Saint Pierre, C.; Gezan, S. A.; Bentley, A. R.; Mosqueda-González, B. A.; Montesinos-López, A.; Van Eeuwijk, F.; Beyene, Y.; Gowda, M.; Gardner, K.; Gerard, G. S.; Crespo-Herrera, L.; Crossa, J. Optimizing Sparse Testing for Genomic Prediction of Plant Breeding Crops. *Genes* **2023**, *14* (4), 927. https://doi.org/10.3390/genes14040927.

(9) Cullis, B. R.; Smith, A. B.; Coombes, N. E. On the Design of Early Generation Variety Trials with Correlated Data. *JABES* **2006**, *11* (4), 381–393. https://doi.org/10.1198/108571106X154443.

(10) Gaynor, R. C.; Gorjanc, G.; Hickey, J. M. AlphaSimR: An R Package for Breeding Program Simulations. *G3 Genes|Genomes|Genetics* **2021**, *11* (2), jkaa017. https://doi.org/10.1093/g3journal/jkaa017.

(11) Werner, C. R.; Gemenet, D. C.; Tolhurst, D. J. FieldSimR: An R Package for Simulating Plot Data in Multi-Environment Field Trials. *Front. Plant Sci.* **2024**, *15*, 1330574. https://doi.org/10.3389/fpls.2024.1330574.