Brazilian Journal of Biometrics, **43**, 1–18 (2025)
doi:10.28951/bjb.v43i4.865

ARTICLE

A framework for building enviromics matrices in mixed models

Bruno Achcar Trevisan,*[,1] Vinícius Silva Junqueira,[2] Bruna de Mello Florêncio,[3] Alexandre Siqueira Guedes Coelho,[1] Gustavo Eduardo Marcatti,[3] and Rafael Tassinari Resende*[,3]

[1]School of Agronomy, Federal University of Goiás, GO, Brazil
[2]Bayer Crop Science, Uberlândia, MG, Brazil
[3]TheCROP, Operating Virtually, Goiânia (GO) and Sete Lagoas (MG), Brazil
*Corresponding author. Email: trevisan23@discente.ufg.br; rafael.tassinari@ufg.br

(Received: March 25, 2025; Revised: June 18, 2025; Accepted: July 25, 2025; Published: November 26, 2025)

Abstract

This study unravels a framework for constructing enviromics matrices within mixed models to integrate genetic and envirotypic data, enhancing phenotypic predictions in plant breeding. Enviromics leverages diverse data sources, such as climate and soil, to characterize genotype-by-environment (G×E) interactions. The approach uses block-diagonal structures in the design matrix **Z** to incorporate random effects from genetic and envirotypic covariates across trials. The covariance structure is modeled through the Kronecker product of the genetic relationship matrix **A** and an identity matrix **I** representing envirotypic effects, effectively capturing both genetic and environmental variability. This dual representation facilitates more accurate predictions of crop performance (**y**) across environments, enabling improved selection strategies in breeding programs. The framework is compatible with widely used mixed model software, including rrBLUP and BGLR, and is adaptable to account for more complex interactions. By integrating genetic relationships (**A**) and environmental influences (**Z**), this approach provides a robust tool for advancing G×E studies and accelerating the development of superior crop varieties.

Keywords: Genotype-by-environment interaction; Envirotypic covariates; Plant breeding; Predictive modeling; Linear algebra.

1. Introduction

The omics alphabet has grown rapidly with advancements in high-throughput typing technologies (i.e. tools capable of efficiently measuring and analyzing large-scale data with high precision), encompassing fields such as genomics, transcriptomics, proteomics, metabolomics, phenomics, and

© Brazilian Journal of Biometrics. This is an open access article distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/)

so on. Recently, enviromics has emerged as a valuable addition, providing a framework for characterizing the "envirome" - the complete set of environmental factors affecting biological processes. Initially applied in fields such as psychiatry (Anthony *et al.,* 1995), it can be expanded to areas such as agriculture or plant biology (Teixeira *et al.,* 2011). Integrating diverse sources of environmental data, including climate, soil properties, air quality, and socioeconomic factors, captures micro- and macroenvironmental influences on complex traits. This approach offers a more comprehensive understanding of how environmental components interact with biological systems, making it a powerful tool across disciplines.

The genetic basis for producing a phenotype is complex and involves the activation and interaction of several genes and metabolic pathways. The most significant benefit of enviromics in the context of breeding is the possibility of modeling additional features beyond phenotypes and genetic markers when predicting future crop performance. In plant and animal breeding, the concept was first introduced in a 2019 bioRxiv preprint (Resende *et al.,* 2021). It is particularly valuable for breeding studies, as it helps breeders and scientists assess the various factors that influence crop performance under different environmental conditions. In this context, "environmental conditions" encompasses more than just geographical factors, such as specific plots, fields, and locations where crops are planted. Instead, enviromics considers any information derived from Earth's surface, atmospheric, remote sensing, and socioeconomic factors that may contribute to changes in genotypic performance across multiple locations. It is especially useful for studying genotype-by-environment-by-management (G×E×M) interactions, helping to decouple diverse envirotypic factors affecting crop performance (Costa-Neto & Fritsche-Neto, 2021; Crossa *et al.,* 2021; Resende *et al.,* 2021). This capability optimizes breeding strategies and decisions, enabling breeders and scientists to make informed site-specific recommendations.

Many advancements in modeling genotype-by-environment (G×E) interactions highlight the benefits of integrating environmental matrix structures with genetic data to improve predictive accuracy in breeding. Crossa *et al.* (2006) initially focused on genetic covariances for G×E modeling. Jarquín *et al.* (2014) employed reaction norm models that combined high-dimensional genomic and environmental data, and Cuevas *et al.* (2017) used Bayesian kernel methods to refine predictions by accounting for genotype-specific responses to environments. Hadamard and Kronecker products have been shown to model covariance structures in interactions effectively, capturing complex relationships in crop trials (Martini *et al.,* 2020). Costa-Neto *et al.* (2021) incorporated such data into nonlinear kernel-based models, broadening the scope of environmental variability considered. These developments underscore the shift from purely genetic models to approaches that integrate comprehensive environmental descriptors, aligning with the principles of enviromics.

The use of this approach in breeding is vast, as it can be applied to develop new cultivars within breeding programs. Still, it is also beneficial for activities that follow after the new cultivars are identified (Resende *et al.,* 2024). For instance, identifying the most suitable environment for multiplying these cultivars is transformative, as it can enhance the profitability of seed production while minimizing seed multiplication costs. Production costs are a key factor influencing seed prices for customers, and since genetics vary in their adaptability to environments, multiplying seeds in less optimal conditions can increase costs (Gevartosky *et al.,* 2023). While our intention is not to exhaust all possible applications, we emphasize that this is a powerful tool extending beyond performance prediction. Its benefits span from developing new cultivars to optimizing seed multiplication, defining agronomic recommendations, and guiding in-farm practices aimed at maximizing the yield potential of the cultivars.

This document outlines the process of building enviromic matrices for mixed models, aiming to integrate these tools into widely used software packages such as BGLR (Pérez & de los Campos, 2014) and rrBLUP (Endelman, 2011). The preparation of this material was inspired by the work of Bates *et al.* (2015), specifically adapted here to assist users in applying linear algebra to handle

random multiple regressions. Please, note that a solid understanding of matrix algebra is required for accurate statistical modeling (see for instance Searle & Khuri (2017)). These matrices are pivotal in mixed models that analyze G×E interactions and enable the integration of enviromic data into genomic selection frameworks. Properly constructed matrices facilitate the modeling of intricate environmental and genetic relationships, improving the prediction accuracy of phenotypic performance and ultimately accelerating the development of improved plant varieties.

2. Enviromics Modeling Documentation

Enviromics provides a robust framework for modeling genotype-by-environment interactions by integrating multiple sources of variation. It allows breeders to account for complex environmental influences and genetic relationships simultaneously, enabling more accurate prediction of phenotypic responses under varying conditions. The approach uses a block-diagonal structure for the matrix $\mathbf{Z}$ to include random effects associated with different genotypic and envirotypic covariates throughout the trials, offering flexibility in capturing site-specific and genotype-specific responses. The covariance structure $\Sigma \otimes \mathbf{A}$, where Σ represents variance-covariance among environmental effects, and A denotes the additive relationship matrix. This dual representation allows us to predict crop performance across diverse environments, optimizing selection strategies in breeding programs.

- **Model Complexity**: The enviromics approach integrates multiple Z_i submatrices (e.g., $Z_1, Z_2, \ldots, Z_n$) into a block-diagonal matrix $\mathbf{Z}$ (Bates *et al.,* 2015), each corresponding to a different trial or environmental condition. This structure provides a flexible framework for modeling random effects across multiple genotypes, accommodating diverse environmental scenarios and trial conditions.

- **Block-Diagonal Z**: The matrix $\mathbf{Z}$ combines various random effects, capturing the structure of the data in the model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_n \end{bmatrix}$$

  where $\mathbf{y}$ is the response vector, $\mathbf{X}$ is the design matrix for fixed effects, $\boldsymbol{\beta}$ represents the fixed effects, $\mathbf{Z}$ is the design matrix for random effects, $\mathbf{u}$ is the vector of random effects, and $\mathbf{e}$ is the residual error. Each block $\mathbf{Z}_i$ corresponds to the design matrix for a specific genotype and combination of environmental conditions.
- **Random Effects Vector**: The random effects vector u encapsulates the responses of genotypes under different envirotypic covariates, where each $u_{i,j}$ represents the effect of genotype i and envirotypic covariate j:

$$\mathbf{u}^\top = [u_{1,1}, u_{1,2}, \ldots, u_{n,m}], \quad \text{with} \quad \mathbf{u} \sim N(0, \mathbf{A} \otimes \Sigma)$$

  This formulation allows for modeling the joint distribution of genetic and environmental factors, enhancing the accuracy of predictions.

- **Covariance Structure**: The envirotypic covariance matrix Σ can be represented as the identity matrix I_{m+1}, where m is the number of envirotypic covariates plus one random intercept (of the genotype), assuming no additional variance within each environment. The matrix **K** is defined as the Kronecker product of the kinship matrix **A** and the identity matrix $\mathbf{I}_{m+1}$, expressed as $\mathbf{K} = \mathbf{A} \otimes \mathbf{I}_{m+1}$. This structure allows for modeling genetic relationships among genotypes while accounting for environmental effects, leading to a comprehensive framework for predicting phenotypic outcomes in mixed models.

The matrix **Z** must have dimensions matching the data structure in mixed models. The number of rows of this matrix should match the dimension of the number of observations, and the number of columns should align with the number of parameters to estimate. Then, the number of columns must also align with the number of rows (or columns) of the kernel matrix (e.g., kinship or similarity matrix). Such alignment ensures proper matrix multiplication and integration of random effects into the model.

2.1 Enviromics Prediction

The enviromics prediction model aims to accurately estimate phenotypic responses by regressing random effects associated with genotypes and multiple envirotypic covariates. The general prediction equation for multiple genotypes can be expressed as:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\mathbf{u}}$$

where $\hat{\mathbf{y}}$ is the vector of predicted phenotypic responses for all genotypes; **X** is the design matrix for fixed effects; $\hat{\beta}$ is the vector of estimated fixed effects (i.e., BLUE in the frequentist paradigm); **Z** is the design matrix for random effects, with dimensions $n \times m$, where n is the number of observations and m is the number of random effects; and $\hat{\mathbf{u}}$ is the vector of estimated random effects (i.e., BLUP in frequentist paradigm), containing the random effects for all genotype-environment combinations. And the specific prediction for each genotype i can then be written as:

$$\hat{y}_i = \beta_0 + \sum_{j=0}^{k} \hat{u}_{ij} \cdot \mathrm{EC}_j$$

where $\hat{y}_i$ is the predicted phenotypic response for genotype i; β_0 is the fixed intercept; $\hat{u}_{ij}$ represents the estimated random effect for the j-th envirotypic covariate (EC) and genotype i; EC_j is the value of the j-th envirotypic covariate, with $\mathrm{EC}_0 = 1$ to account for the random intercept; and k is the total number of envirotypic covariates considered in the model.

The predicted phenotypic responses for each genotype i (where i varies from 1 to n) are summarized in Table 1. Each equation incorporates random effects associated with envirotypic covariates j (where j varies from 1 to m). This allows for a comprehensive understanding of how genetic and environmental factors contribute to phenotypic variation. In a related context, Barros *et al.* (2024) demonstrated the predictive value of Landsat-based indices for classifying sugarcane areas using principal component analysis (PCA) and machine learning algorithms, supporting the integration of such descriptors into crop modeling frameworks.

Envirotypic covariates (ECs) are measurable environmental factors that influence phenotypic performance, while enviromic markers represent these covariates in predictive models, analogous to genomic markers (Resende *et al.,* 2024). According to these authors, an enviromic marker should

have the following properties: (i) it should have a linear relationship with the trait of interest; (ii) it should indicate that a location with a high phenotypic mean is also a favorable or potential site; (iii) a generally favorable location is not necessarily favorable for all genotypes; and (iv) it consists of different sets of environmental covariates, where a single environmental covariate is not always the best predictor for all genotypes.

However, in this work, for didactic purposes, we assumed ECs to be enviromic markers, acknowledging the risk of potentially missing important nonlinear effects. To address such effects, see Costa-Neto *et al.* (2021). To improve the diagnostic evaluation of enviromic predictions, residuals that are minimally confounded with random effects may be used in place of marginal or conditional residuals. Queiroz de Oliveira *et al.* (2024) showed that such residuals enhance the detection of model misspecification in mixed models with hierarchical random structures.

Table 1. Predicted phenotypic responses for genotypes incorporating random effects from envirotypic covariates

Genotype	Equation
1	$\hat{y}_1 = \beta_0 + \hat{u}_{1,0} + \hat{u}_{1,1} \cdot \mathrm{EC}_1 + \hat{u}_{1,2} \cdot \mathrm{EC}_2 + \ldots + \hat{u}_{1,m} \cdot \mathrm{EC}_m$
2	$\hat{y}_2 = \beta_0 + \hat{u}_{2,0} + \hat{u}_{2,1} \cdot \mathrm{EC}_1 + \hat{u}_{2,2} \cdot \mathrm{EC}_2 + \ldots + \hat{u}_{2,m} \cdot \mathrm{EC}_m$
⋮	⋮
n	$\hat{y}_n = \beta_0 + \hat{u}_{n,0} + \hat{u}_{n,1} \cdot \mathrm{EC}_1 + \hat{u}_{n,2} \cdot \mathrm{EC}_2 + \ldots + \hat{u}_{n,m} \cdot \mathrm{EC}_m$

3. Data Description (a Toy Example)

To exemplify the enviromics modeling, we will use genotypic observations across various locations with associated envirotypic covariates (e.g., temperature, soil properties, and precipitation). This information forms the basis for building design matrices and covariance matrices in Enviromics. Each location represents a different trial, with varying numbers of genotypes. The plotted points correspond to the geographic coordinates (Longitude and Latitude), while the labels (G_A, G_B, G_C, or G_D) indicate the genotypes ('`gen`') available at each trial site.

The first step involves loading the dataset with observations across multiple locations and genotypes. The dataset includes envirotypic covariates (EC_1, EC_2, EC_3, EC_4, EC_5) that vary by location and may influence the response of different genotypes. Each row represents a unique combination of location and genotype, providing the necessary structure for building the design matrices and the covariance matrix. The column named 'y' represents the phenotypic variable (trait), which can be associated with crop production or productivity (whatever). The following code snippet demonstrates how to read the dataset into the R environment and display its contents for initial inspection (see the R chunk below):

```
# Read the data from a file
dat <- read.table("dat.txt", header = TRUE)

# Display the data
print(dat)
   loc lon lat gen   EC1    EC2    EC3    EC4    EC5     y
1   L1   1   4  GA -0.230 -1.265  0.111 -0.050 -0.270 12.88
2   L5   2   5  GA  0.129  1.224  0.498  0.867  0.968 17.83
3   L6   3   6  GA  1.715  0.360 -1.967  1.061 -0.132 18.41
4   L2   2   2  GB -0.560  0.461  0.401  0.391  0.472 12.08
```

5	L1	1	4	GB	−0.230	−1.265	0.111	0.022	−0.046	11.84
6	L4	4	4	GB	0.071	−0.446	1.787	0.580	0.710	15.82
7	L6	3	6	GB	1.715	0.360	−1.967	0.832	−0.486	18.16
8	L2	2	2	GC	−0.560	0.461	0.401	0.588	0.322	12.34
9	L3	3	3	GC	1.559	−0.687	−0.556	0.522	0.166	16.11
10	L2	2	2	GD	−0.560	0.461	0.401	0.770	0.692	15.25
11	L1	1	4	GD	−0.230	−1.265	0.111	0.108	−0.142	14.91
12	L4	4	4	GD	0.071	−0.446	1.787	0.607	1.262	19.24
13	L5	2	5	GD	0.129	1.224	0.498	0.683	0.706	18.16
14	L6	3	6	GD	1.715	0.360	−1.967	1.052	−0.299	16.18

Figure 1 shows the pedigree of the four individuals and their relationship (kinship) matrix. On the left, the pedigree presents the relationships between founders (P_1 to P_5) and their offspring (genotypes G_A, G_B, G_C, and G_D). On the right, the relationship matrix quantifies the degree of relatedness between the genotypes, with colors representing the levels of relatedness. The relationship matrix includes the coefficient of relationship of Wright (1921) and can be easily implemented following Henderson's rules to derive its inverse directly (Henderson, 1976). Since we are dealing with a small number of individuals (genotypes) here, the kinship coefficients were provided directly; in cases involving a large number of individuals with pedigree data, we recommend using packages such as AGHMatrix to compute them efficiently (Amadeu *et al.,* 2023). Although this study considers a diploid species, the methodology remains generally applicable to polyploid plants, provided that the breeder correctly computes the kinship coefficients.

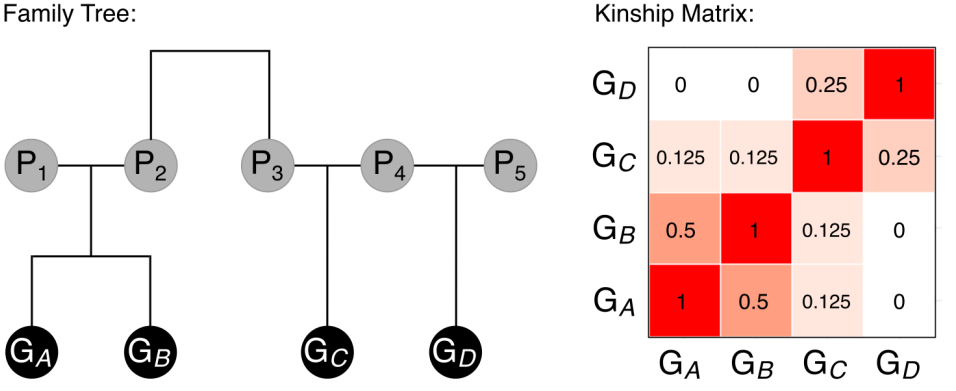

Figure 1. Pedigree and kinship matrix. The left side shows the genealogical relationships between the founders (P_1 to P_5) and their offspring (G_A, G_B, G_C, and G_D). The right side displays a heatmap of the kinship matrix, with colors representing the kinship coefficients between the genotypes.

4. Preliminary Steps

The initial steps for matrix preparation include the following:

- The data is first sorted by genotype and then by location.
- Constructing block-diagonal design matrices to represent the structure of the observations.
- Build the relationship matrix based on specified genetic relationships, which will later be used to build kernel matrices.

5. Enviromics Matrices Preparation

Building enviromics matrices involves organizing data to capture genetic and environmental components in statistical models. The process begins with constructing a design matrix (Z) to represent random effects for each genotype and envirotypic covariate. This block-diagonal matrix accommodates multi-environment trials, allowing the model to separate effects across genotypes. A kernel matrix (K) also models genetic relationships (which is, in fact, merely an expanded kinship matrix), integrating genetic information with random effects through the Kronecker product for a comprehensive covariance structure. Together, these matrices form the foundation for mixed model analyses, enabling the prediction of phenotypic responses under various conditions. The process involves the following components:

- **Design Matrix (Z)**: A block-diagonal matrix where each block represents the design structure for a specific genotype, incorporating envirotypic covariates. The matrix captures the design structure for random effects with the following structure:

```
# Split the data by genotype
split_data <- split(dat, dat$gen)
# Create the block-diagonal matrix Z
Z <- bdiag(lapply(split_data, function(df) {
  # First column of 1's and columns for covariates:
  # EC1, EC2, EC3, EC4, EC5
  Z_i <- cbind(1, as.matrix(df[, c("EC1", "EC2", "EC3",
  "EC4", "EC5")]))
  Matrix(Z_i, sparse = TRUE)
}))
# Names the columns of matrix Z
colnames(Z) <- unlist(lapply(1:4, function(i) {
    paste0("u", 0:5, "_G", LETTERS[i])
    }))
# Display the matrix Z
print(Z)
```

Table 2 shows the design matrix Z used in the mixed model. This matrix is constructed in a block-diagonal format, where each block represents the design effects for different genotypes (G_A, G_B, G_C, G_D). The first column of each block represents an intercept (u_0), while the subsequent columns represent envirotypic covariates ($u_0, u_1, u_2, u_3, u_4, u_5$). The block-diagonal structure reflects the separation of effects for different genotypes, allowing efficient modeling of specific random effects.

- **Kernel Matrix (K)**: Calculated using the Kronecker product, combining a genetic relationship matrix with an identity matrix representing different effects. This matrix is used to model the covariance structure between genotypes:

```
# Create the relationship matrix based on pedigree
A <- matrix(c(
  1.000, 0.500, 0.125, 0.000,
  0.500, 1.000, 0.125, 0.000,
```

```
    0.125, 0.125, 1.000, 0.250,
    0.000, 0.000, 0.250, 1.000
), nrow = 4, byrow = TRUE)

# Define the genotype names
rownames(A) <- colnames(A) <- unlist(lapply(1:4, function(i)
{paste0("G", LETTERS[i])    }))

# Create the 6x6 identity matrix for the effects
(u0, u1, u2, u3, u4, u5)
I6 <- diag(6)
rownames(I6) <- colnames(I6) <- c("u0", "u1", "u2",
"u3", "u4", "u5")

# Calculate the Kernel matrix using the Kronecker product
K <- kronecker(A, I6)
# The resulting matrix will have dimensions 24x24

# Rename the rows and columns of the K matrix to reflect
#genotypes and effects
rownames(K) <- colnames(K) <- unlist(lapply(1:4, function(i)
{paste0("u", 0:5, "_G", LETTERS[i])    }))
```

Table 2. Design matrix Z for the mixed model. Each block represents the design structure for a specific genotype (G_A, G_B, ..., G_D), with the first column (u_0) as an intercept, the last column (u_m) as a covariate, and ellipses (...) indicating omitted intermediate envirotypic covariates. In our case, $m = 5$, corresponding to ($u_0, u_1, ..., u_5$)

G_A				G_B				G_C				G_D			
u_0	u_1	...	u_m	u_0	u_1	...	u_m	u_0	u_1	...	u_m	u_0	u_1	...	u_m
1	-0.230	...	-0.270	0	0	...	0	0	0	...	0	0	0	...	0
1	0.129	...	0.968	0	0	...	0	0	0	...	0	0	0	...	0
1	1.715	...	-0.132	0	0	...	0	0	0	...	0	0	0	...	0
0	0	...	0	1	-0.560	...	0.472	0	0	...	0	0	0	...	0
0	0	...	0	1	-0.230	...	-0.046	0	0	...	0	0	0	...	0
0	0	...	0	1	0.071	...	0.710	0	0	...	0	0	0	...	0
0	0	...	0	1	1.715	...	-0.486	0	0	...	0	0	0	...	0
0	0	...	0	0	0	...	0	1	-0.560	...	0.322	0	0	...	0
0	0	...	0	0	0	...	0	1	1.559	...	0.166	0	0	...	0
0	0	...	0	0	0	...	0	0	0	...	0	1	-0.560	...	0.692
0	0	...	0	0	0	...	0	0	0	...	0	1	-0.230	...	-0.142
0	0	...	0	0	0	...	0	0	0	...	0	1	0.071	...	1.262
0	0	...	0	0	0	...	0	0	0	...	0	1	0.129	...	0.706
0	0	...	0	0	0	...	0	0	0	...	0	1	1.715	...	-0.299

Figure 2 illustrates the expanded kinship matrix K, which combines the genetic relationship matrix with the Identity (I_{m+1}) matrix for different effects. The matrix K is constructed using the Kronecker product, capturing the covariance structure between genotypes (G_A, G_B, G_C, and G_D) across various random effects ($u_0, u_1, u_2, u_3, u_4, u_5$). The heatmap visualizes the kinship values, with a color gradient indicating the degree of relatedness.

In the context of the kernel matrix, it is assumed that there is no relationship between the random effects $u_0, u_1, u_2, u_3, u_4, u_5$ within each genotype. This follows Fisher's infinitesimal model, which posits that numerous small, independent effects contribute to the overall genetic variation. In practical applications of enviromics, the model can accommodate thousands of such effects $(u_1, u_2, \ldots, u_m)$, representing various envirotypic covariates. However, when considering specific effects across genotypes, such as $u_0 \times u_0$, $u_1 \times u_1$, $u_2 \times u_2$, $u_3 \times u_3$, $u_4 \times u_4$, and $u_5 \times u_5$, there is a structured relationship that reflects the genetic relatedness specified by the kinship matrix. This kinship can be based on pedigree information or derived from molecular markers like SNPs (VanRaden, 2008), allowing the model to incorporate genetic relationships across different random effects.

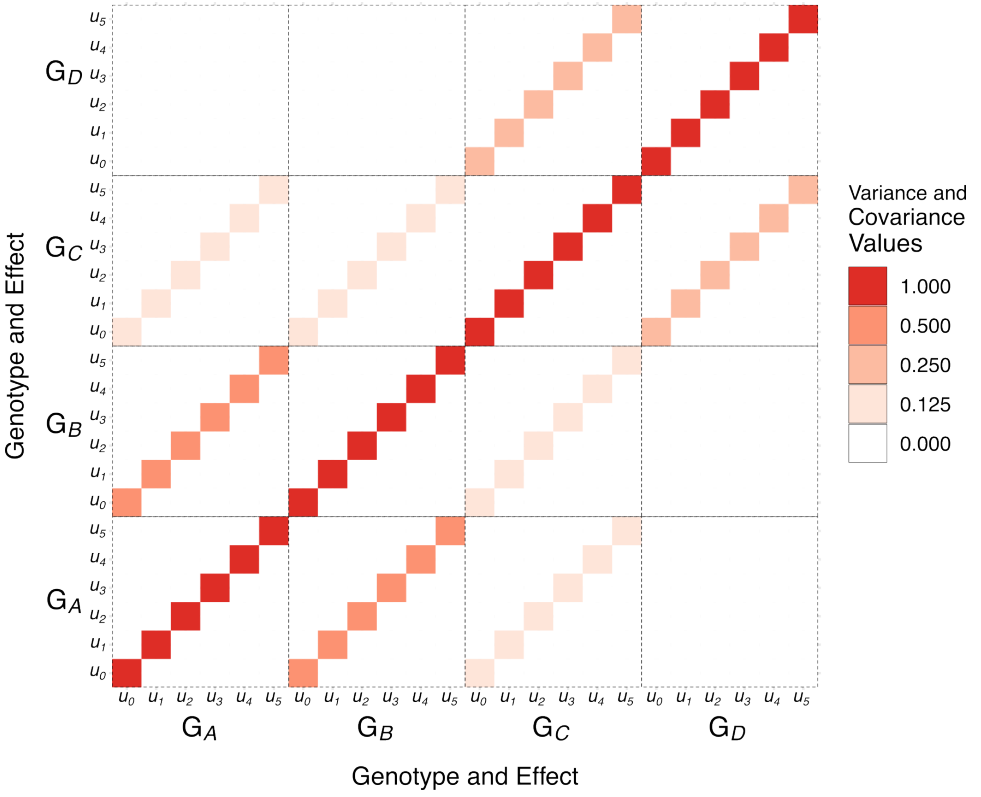

Figure 2. Heatmap visualization of the expanded kinship matrix K, here also called as "Kernel" matrix, representing the genetic relationships across genotypes and effects. The matrix was generated using the Kronecker product to expand the genetic relationship matrix by different random effects $(u_0, u_1, u_2, u_3, u_4, u_5)$. This matrix is defined as the Kronecker product of the kinship matrix **A** and the a identity matrix **I**.

The matrix methodology presented here is similar to that described in Resende *et al.* (2025), which achieved predictive ability approximately 20–25% higher than traditional G×E-based selection methods. This model was applied using the `rrBLUP` package (Endelman, 2011) for a frequentist approach and the `BGLR` package (Pérez & de los Campos, 2014) for a Bayesian approach, as both allow flexible manipulation of matrices. However, many other R packages can be used for this purpose, depending on the user's preferences and familiarity. To understand the predictive performances of the models described here, please refer to the studies by Resende *et al.* (2021), and Resende *et al.* (2025).

6. Enviromics Matrices: Frequentist and Bayesian Approaches

The study of genetic and non-genetic components of populations can be approached through a variety of methodologies. From a statistical perspective, these methods are broadly classified into supervised and unsupervised approaches. Supervised methods, in particular, can be further divided into parametric and non-parametric categories, each with distinct assumptions and analytical frameworks. In this work, we focus on supervised parametric approaches within the frequentist and Bayesian paradigms. These methods have been extensively applied over decades to investigate the genetic architecture of plants, animals, microorganisms, and human populations, contributing to significant advancements in quantitative genetics and breeding. In the following sections, we delve into the technical foundations of these parametric approaches, with an emphasis on genotype-by-environment interaction modeling. Additionally, we illustrate their application through a simple example, demonstrating the construction of design matrices in the emerging field of enviromics.

For a large number of envirotypic covariates (ECs), convergence issues may arise both in Frequentist and Bayesian approaches. To address this, 'ensemble models' can be applied by grouping ECs and running separate models for each group in enviromics models (Resende *et al.,* 2025). Going further, in practical / empirical applications, it is also essential to cross-validate these models by implementing training and validation sets. For this purpose, we recommend exploring strategies for multi-environment and/or multi-year validations across regions, as discussed in Rogers & Holland (2022).

6.1 Frequentist approach

Using supervised parametric methods in genetics study involves estimating parameters that describe the source variations of a given population. Over the years, several techniques, such as Methods I, II, and III of Henderson (1953), Methods MINQUE and MIVQUE of Rao (1970), Rao (1971a), and Rao (1971b), ANOVA, maximum likelihood (ML), and restricted (or residual) maximum likelihood (REML) (Patterson & Thompson, 1971), were developed to help disentangle the source of variations into different components. Each one of these methods aims to optimize quadratic forms under different assumptions related to the data structure (i.e., balanced and unbalanced), as well as the bias and variance of estimates. Currently, the most widely used method is REML due to its reliability and flexibility in managing complex data structures and its ability to produce unbiased estimates for fixed and random effects.

As long as the variance components are available, estimating the genetic values becomes a (relatively) trivial task under Henderson's mixed model equations. However, in practice, the estimation of variance components and genetic value prediction occurs in a two-step iterative process. The first step estimates the variance components through maximum likelihood or restricted maximum likelihood methods. Then, the fixed and random effects are derived in the second step, assuming that the variance components estimated in the first step are true. The model converges to a stable solution after a few rounds of iterative process. Currently, several computationally efficient algorithms can execute this task.

In the Genomic Era, mixed models have been used successfully to predict genomic values in many species. Nowadays, multiple models are used, and they can differ if the goal is to estimate the genetic values or marker (e.g., SNP) effects. Among the methods, the ridge regression BLUP (RRBLUP) is a popular one. RRBLUP employs ridge regression to handle multicollinearity issues common in high-dimensional datasets, such as those encountered in genomics, where usually there are way more markers than observations. RRBLUP works by partitioning phenotypic variation into genetic and non-genetic components using mixed models, where fixed effects account for general trends and random effects capture genotype-specific variations. Fixed effects capture consistent influences shared across all genotypes, while random effects model genotype-specific deviations from the population mean. This framework integrates a design matrix $\mathbf{Z}$, a kernel matrix $\mathbf{K}$, and

phenotypic values **y** to effectively model genotype-by-environment interactions. The estimated fixed and random effects are organized into structured data frames, providing a systematic approach for downstream analysis and interpretation (Burgueño *et al.,* 2007).

The following code demonstrates the application of the `rrBLUP` R package to partition phenotypic variation into genetic and environmental components. The phenotypic data (*y*) is stored, and the model is fitted using the design matrix (*Z*) and the kernel matrix (*K*). Fixed effects are extracted and stored for further interpretation, while random effects are arranged and labeled in a structured data frame for downstream analysis.

The observed effects shown in Figure 3 were predicted using the mixed model equation $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\mathbf{u}}$, where $\mathbf{X}\hat{\beta}$ represents the fixed effects and $\mathbf{Z}\hat{\mathbf{u}}$ accounts for the random effects. These components were extracted and arranged using the `rrBLUP` package, as shown in the provided code. The fixed and random effects for each genotype and envirotypic index were then combined to generate the predicted values plotted in the figure.

Using the fitted models, phenotypic responses are predicted for each genotype across a gradient of envirotypic indices. These predictions combine fixed effects, representing baseline responses, with random effects, accounting for genotype-specific deviations. This dual-component structure enables a nuanced understanding of how genotypes respond to varying environmental conditions, as in Table 1. Such modeling captures the complexities of genotype-by-environment interactions, facilitating the identification of genotypes with optimal performance under specific conditions (Cuevas *et al.,* 2018).

```
# Storing phenotypic data
y <- as.matrix(dat$y)

# Fitting the rrBLUP model
rrblupmodel <-  mixed.solve (y = y, Z = Z, K = K)

# Storing fixed effect
fixef_rrblup <- as.data.frame(rrblupmodel$beta)[, 1]

#Storing , arranging and naming random effects
ranef_rrblup <- as.data.frame(rrblupmodel$u)[, 1]
ranef_matrix_rrblup <- matrix(ranef_rrblup, nrow = 4,
ncol = 6, byrow = TRUE)
ranef_df_rrblup <- as.data.frame(ranef_matrix_rrblup)
rownames(ranef_df_rrblup) <- c("GA","GB","GC","GD")
colnames(ranef_df_rrblup) <- c("u_0","u_1","u_2",
                               "u_3","u_4","u_5")
```

6.2 Bayesian approach

Bayesian framework is another method commonly used to estimate variance components and derive genetic values. Bayesian methods date back to Thomas Bayes, who was credited with developing Bayes' Theorem, a fundamental probability theory and statistics principle. However, the broader Bayesian paradigm, as we know it today, was shaped and expanded by many contributors over time. Such continued contributions by other scientists enabled the utilization of the Bayesian approach in practical conditions. The central principle of the Bayesian approach is that the posterior probability of the parameters of interest—such as fixed and random effects and variance components—is determined by the joint distribution of the data, which is represented by the likelihood

function alongside additional terms related to the prior probability of the model's parameters. The prior probability in the Bayesian framework reflects our a priori knowledge about a specific parameter. In the context of genomics, this prior information typically relates to the probability function that describes the distribution of genetic markers. For instance, if we assume that several genes of small effect influence a particular trait, a normal distribution may be a suitable model for that trait. Conversely, if a few genes of large effect control the trait under investigation, an exponential or t-distribution may better capture its characteristics than a normal distribution. Additionally, there is the option to adjust the functions that model the likelihood of a specific marker's inclusion—or exclusion—in influencing phenotypic expression. These models are referred to as Bayesian variable selection methods.

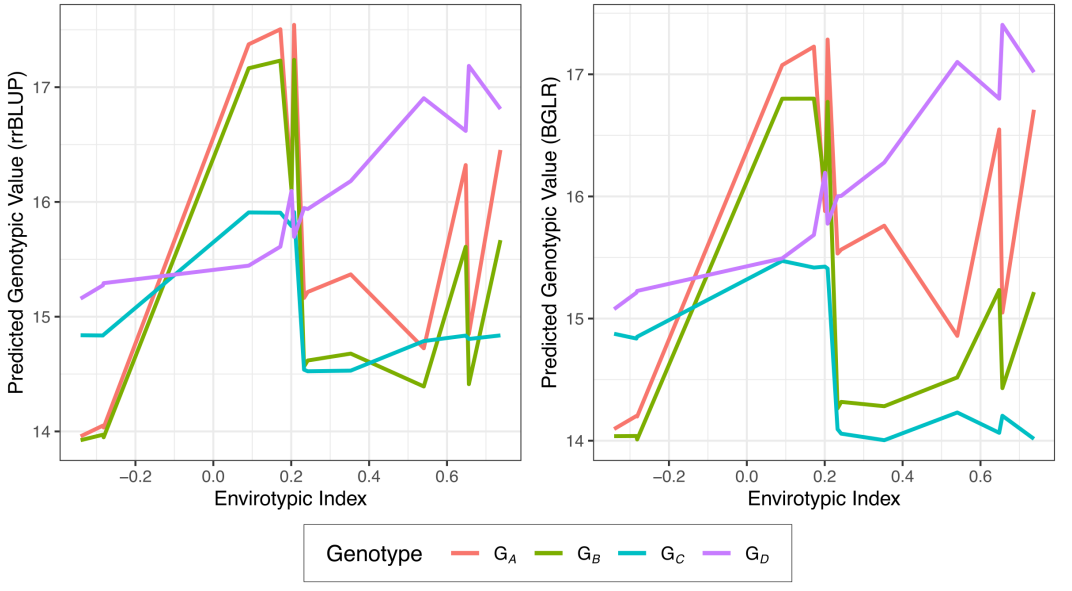

Figure 3. Comparison of Predicted Genotypic Values from the rrBLUP and BGLR models across an Envirotypic Index for different genotypes (G_A, G_B, G_C, and G_D). The "Envirotypic Index" was assumed to be simply the average of the five envirotypic covariates (EC_1–EC_5).

A key concept in the Bayesian approach is that all effects are treated as random. Each effect included in the model is represented using a specific probability distribution function that describes its behavior. In Bayesian analysis, when we want to indicate that a frequentist fixed effect is included explicitly, we refer to it as a systematic effect. Typically, these systematic effects are incorporated into the model using either a uniform or normal distribution with a very large variance. This approach allows us to assume that the effect remains relatively consistent across all levels of the other parameters in the model. In genetics, we need to explicitly define the probability function that will be used to describe either marker effects or genetic effects. The Bayesian framework is particularly advantageous for exploring the various hypotheses related to the genetic control of traits and their impact on predictions. There are many options available within this framework. One common option is to use a normal prior with equal variance for all markers, which is implemented in the RRBLUP method. In the Bayes A method, each marker is assigned its normal prior with a specific variance. Bayes B is similar to Bayes A but requires a predefined proportion (π) of markers to have non-zero effects, facilitating variable selection. Methods like Bayes Cπ and Dπ incorporate a probability function related to the proportion of markers with non-zero effects (Gianola, 2013). These methods are improved versions of Bayes B, allowing the model to determine the optimal

π based on available data (likelihood function) and the assumed prior distributions. This aspect of Bayesian analysis is particularly beneficial as it enables a learning process during the evaluation.

Several computational packages are available for analyzing genomic data using Bayesian methods, such as `bWGR` (Xavier *et al.,* 2020). In this demonstration, we illustrate the application of Bayesian Ridge Regression (BRR) using the `BGLR` package. This versatile package allows users to choose from various probabilistic models and employs Markov Chain Monte Carlo (MCMC) methods via the Gibbs sampling algorithm for parameter estimation. In our model, we utilize a normal prior for the kernel matrix **K** and specify both the kernel matrix and the design matrix **Z** within the `ETA` list. The analysis is executed over 50,000 iterations, with a burn-in period set at 10% of the total iterations. Systematic effects, treated as fixed effects in frequentist approaches, are estimated and recorded, while the remaining random effects are structured and labeled in a data frame for downstream analysis. This structured output is particularly useful for interpreting genotype-by-environment interactions. The BGLR framework enhances genomic analysis through its flexibility and ability to incorporate prior knowledge. Unlike deterministic estimators like rrBLUP, BGLR uses iterative sampling to estimate parameters, providing a quantification of uncertainty in predictions (Montesinos-López *et al.,* 2016). Our implementation adopts a normal prior distribution, consistent with frequentist assumptions, to enable direct comparisons between Bayesian and frequentist results. Following the completion of the MCMC process, the first step is to discard the burn-in samples to mitigate the influence of initial conditions. To further address autocorrelation often present in MCMC samples, we retain only one sample for every n-th iteration (thinning). After this initial processing, it is critical to evaluate the convergence of the retained samples. Convergence diagnostics typically involve graphical inspections (e.g., trace plots) and statistical tests, such as the Geweke diagnostic, to ensure the reliability of parameter estimates.

```
# Fitting the BGLR model
ETA <- list(A = list(X = Z, K = K, model = 'BRR'))

niter <- 50000

BGLRmodel <- BGLR(y = y,
        ETA=ETA,
        nIter = niter,
        burnIn = niter*.10,
        thin = 20,
        saveAt = 'BGLRoutput')


# Storing fixed effect
fixef_bglr <- as.data.frame(BGLRmodel$mu)[, 1]

#Storing, arranging and naming random effects
ranef_bglr <- as.data.frame(BGLRmodel$ETA$A$b)[, 1]
ranef_matrix_bglr <- matrix(ranef_bglr, nrow = 4,
ncol = 6, byrow = TRUE)
ranef_df_bglr <- as.data.frame(ranef_matrix_bglr)
rownames(ranef_df_bglr) <- c("GA","GB","GC","GD")
colnames(ranef_df_bglr) <- c("u_0","u_1","u_2","u_3","u_4",
"u_5")
```

Model performance is assessed by comparing predictions from rrBLUP and BGLR side-by-side in Figure 3. The high correlation (0.973) between predictions highlights their consistency in estimating genotypic values. This agreement underscores the robustness of these methods in capturing additive and additive-by-additive genetic interactions across environments (Solaymani *et al.,* 2020). However, the Bayesian approach's ability to incorporate prior knowledge and quantify uncertainty offers additional flexibility, particularly for complex multi-environment trials involving multiple traits.

Bayesian analysis allows the integration of different sources of genetic information, such as the relationship matrix (**A**) and the genomic relationship matrix (**G**) for genomic-enviromic predictions (Cuevas *et al.,* 2017). The possibility of using different priors can be a key feature that enhances the model's ability to derive reliable predictions. Informative priors are valuable in real-world breeding studies when prior knowledge exists about genetic relationships, environmental effects, or trait heritability. For instance, priors based on pedigree information can guide the model when genomic data is sparse, while priors derived from historical yield data can improve predictions in environments with limited phenotypic observations. Additionally, shrinkage priors, such as those used in Bayesian Ridge Regression, are effective for controlling overfitting when working with a large number of predictors. By incorporating these priors, `BGLR` allows breeders and scientists to leverage prior knowledge effectively, resulting in more accurate and reliable predictions for complex breeding scenarios.

7. Next Steps

The next phase of implementing enviromic models involve:

- Validating the matrix construction with real datasets.
- Extending the approach to account for more complex genetic and environmental interactions.
- Prediction of parents, for instance, when dealing with lines, to obtain specific hybrids.
- Interpolate the genotypic prediction across an entire framed area based on the "Target Population of Environments" (TPE) (Cruz *et al.,* 2025) for the crop.
- Evaluating the impact of different covariance structures on the predictive performance of the mixed models.

8. Conclusions

Developing enviromics matrices for mixed models using frequentist or Bayesian methods is a robust framework for integrating genetic and environmental data in plant breeding. The construction of design and kernel matrices enables the accurate modeling of genotype-by-environment interactions, allowing for improved prediction of phenotypic performance under diverse conditions. While the current approach uses a block-diagonal structure to separate random effects across genotypes and assumes no relationships between different **u**'s within the same genotype, future advancements could extend these methods to more complex scenarios.

Mathematical developments in matrix expansions could involve incorporating multi-trait models, where multiple phenotypic traits are analyzed simultaneously, accounting for correlations among them. Another promising direction is including diverse experimental types within a unified framework, allowing for the simultaneous analysis of trials with different designs, such as replicated trials, unreplicated trials, or even observational data. Additionally, introducing covariances between random effects **u**'s across genotypes could further enhance the modeling of shared environmental influences or interactions among traits.

Such expansions would require sophisticated matrix derivations and computational techniques, potentially involving sparse matrix algebra operations and high-dimensional statistical methods. By

pursuing these advancements, enviromics could provide even more comprehensive insights into genetic and environmental interactions, driving more effective breeding strategies and accelerating the development of improved crop varieties. In addition, future research should focus on developing more computationally efficient strategies to enable large-scale analyses. As environmental data becomes widely available and included in the genetic evaluation of breeding programs, new software will be required to manage the amount of data, especially when integrated with additional omics information such as genomic, transcriptomic, proteomics, metabolomics, epigenomics.

Acknowledgments

We express our gratitude to the PPGGMP/UFG ("Programa de Genética e Melhoramento de Plantas" at UFG–"Universidade Federal de Goiás") and the LAMP ("Laboratório de Melhoramento de Precisão") for their support during the development of this work. We declare no conflicts of interest. Also, we would like to thank reviewers and editors for their comments and suggestions.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: RESENDE, R.T. **Data curation**: - **Formal analysis**: TREVISAN, B.A.; RESENDE, R.T.; JUNQUEIRA, V.S.; FLORÊNCIO, B.M. **Funding acquisition**: - **Investigation**: TREVISAN, B.A.; RESENDE, R.T. **Methodology**: RESENDE, R.T.; COELHO, A.S.G **Project administration**: RESENDE, R.T. **Software**: - **Resources**: - **Supervision**: RESENDE, R.T.; MARCATTI, G.E. **Validation**: TREVISAN, B.A.; JUNQUEIRA, V.S.; COELHO, A.S.G.; MARCATTI, G.E. **Visualization**: TREVISAN, B.A.; JUNQUEIRA, V.S.; COELHO, A.S.G.; MARCATTI, G.E. **Writing - original draft**: RESENDE, R.T. **Writing - review and editing.**: TREVISAN, B.A.; JUNQUEIRA, V.S.; FLORÊNCIO, B.M.; COELHO, A.S.G.; MARCATTI, G.E.

References

1. Amadeu, R. R., Garcia, A. A. F., Munoz, P. R. & Ferrão, L. F. V. AGHmatrix: genetic relationship matrices in R. *Bioinformatics* **39** (ed Schwartz, R.) ISSN: 1367-4811. doi:10.1093/bioinformatics/btad445. http://dx.doi.org/10.1093/bioinformatics/btad445 (2023).
2. Anthony, J. C., Eaton, W. W. & Henderson, A. S. Looking to the Future in Psychiatric Epidemiology. *Epidemiologic Reviews* **17,** 240–242. ISSN: 0193-936X. doi:10.1093/oxfordjournals.epirev.a036182. http://dx.doi.org/10.1093/oxfordjournals.epirev.a036182 (1995).
3. Barros, A. C. A. V. B. d., Silva, M. A., Marques, R. D., Villegas, C. & Luciano, A. C. d. S. Classification of sugarcane areas in Landsat images using machine learning algorithms. *Brazilian Journal of Biometrics* **42,** 147–157. ISSN: 2764-5290. doi:10.28951/bjb.v42i2.693. http://dx.doi.org/10.28951/bjb.v42i2.693 (2024).
4. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software* **67.** ISSN: 1548-7660. doi:10.18637/jss.v067.i01. http://dx.doi.org/10.18637/jss.v067.i01 (2015).
5. Burgueño, J., Crossa, J., Cornelius, P. L., Trethowan, R., McLaren, G. & Krishnamachari, A. Modeling Additive × Environment and Additive × Additive × Environment Using Genetic Covariances of Relatives of Wheat Genotypes. *Crop Science* **47,** 311–320. ISSN: 1435-0653. doi:10.2135/cropsci2006.09.0564. http://dx.doi.org/10.2135/cropsci2006.09.0564 (2007).

6. Costa-Neto, G. & Fritsche-Neto, R. Enviromics: bridging different sources of data, building one framework. *Crop Breeding and Applied Biotechnology* **21.** ISSN: 1518-7853. doi:10.1590/1984-70332021v21sa25. http://dx.doi.org/10.1590/1984-70332021v21Sa25 (2021).
7. Costa-Neto, G., Fritsche-Neto, R. & Crossa, J. Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* **126,** 92–106. ISSN: 1365-2540. doi:10.1038/s41437-020-00353-1. http://dx.doi.org/10.1038/s41437-020-00353-1 (2021).
8. Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R. & Krishnamachari, A. Modeling Genotype × Environment Interaction Using Additive Genetic Covariances of Relatives for Predicting Breeding Values of Wheat Genotypes. *Crop Science* **46,** 1722–1733. ISSN: 1435-0653. doi:10.2135/cropsci2005.11-0427. http://dx.doi.org/10.2135/cropsci2005.11-0427 (2006).
9. Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A. & Bentley, A. R. The Modern Plant Breeding Triangle: Optimizing the Use of Genomics, Phenomics, and Enviromics Data. *Frontiers in Plant Science* **12.** ISSN: 1664-462X. doi:10.3389/fpls.2021.651480. http://dx.doi.org/10.3389/fpls.2021.651480 (2021).
10. Cruz, D. D. M., Heinemann, A. B., Marcatti, G. E. & Resende, R. T. Defining the target population of environments (TPE) for enviromics studies using R-based GIS tools. *Crop Breeding and Applied Biotechnology* **25.** ISSN: 1518-7853. doi:10.1590/1984-70332024v25n1s09. http://dx.doi.org/10.1590/1984-70332024v25n1s09 (2025).
11. Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P. & de los Campos, G. Bayesian Genomic Prediction with Genotype × Environment Interaction Kernel Models. *G3 Genes|Genomes|Genetics* **7,** 41–53. ISSN: 2160-1836. doi:10.1534/g3.116.035584. http://dx.doi.org/10.1534/g3.116.035584 (2017).
12. Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., Bandeira e Sousa, M. & Crossa, J. Genomic-Enabled Prediction Kernel Models with Random Intercepts for Multi-environment Trials. *G3 Genes|Genomes|Genetics* **8,** 1347–1365. ISSN: 2160-1836. doi:10.1534/g3.117.300454. http://dx.doi.org/10.1534/g3.117.300454 (2018).
13. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* **4,** 250–255. ISSN: 1940-3372. doi:10.3835/plantgenome2011.08.0024. http://dx.doi.org/10.3835/plantgenome2011.08.0024 (2011).
14. Gevartosky, R., Carvalho, H. F., Costa-Neto, G., Montesinos-López, O. A., Crossa, J. & Fritsche-Neto, R. Enviromic-based kernels may optimize resource allocation with multi-trait multi-environment genomic prediction for tropical Maize. *BMC Plant Biology* **23.** ISSN: 1471-2229. doi:10.1186/s12870-022-03975-1. http://dx.doi.org/10.1186/s12870-022-03975-1 (2023).
15. Gianola, D. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. *Genetics* **194,** 573–596. ISSN: 1943-2631. doi:10.1534/genetics.113.151753. http://dx.doi.org/10.1534/genetics.113.151753 (2013).
16. Henderson, C. R. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics* **32,** 69. ISSN: 0006-341X. doi:10.2307/2529339. http://dx.doi.org/10.2307/2529339 (1976).
17. Henderson, C. R. Estimation of Variance and Covariance Components. *Biometrics* **9,** 226. ISSN: 0006-341X. doi:10.2307/3001853. http://dx.doi.org/10.2307/3001853 (1953).

18. Jarquín, D. *et al.* A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* **127,** 595–607. ISSN: 1432-2242. doi:10.1007/s00122-013-2243-1. http://dx.doi.org/10.1007/s00122-013-2243-1 (2014).

19. Martini, J. W. R., Crossa, J., Toledo, F. H. & Cuevas, J. On Hadamard and Kronecker products in covariance structures for genotype × environment interaction. *The Plant Genome* **13.** ISSN: 1940-3372. doi:10.1002/tpg2.20033. http://dx.doi.org/10.1002/tpg2.20033 (2020).

20. Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M. & Rutkoski, J. A Genomic Bayesian Multi-trait and Multi-environment Model. *G3 Genes|Genomes|Genetics* **6,** 2725–2744. ISSN: 2160-1836. doi:10.1534/g3.116.032359. http://dx.doi.org/10.1534/g3.116.032359 (2016).

21. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58,** 545–554. ISSN: 1464-3510. doi:10.1093/biomet/58.3.545. http://dx.doi.org/10.1093/biomet/58.3.545 (1971).

22. Pérez, P. & de los Campos, G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* **198,** 483–495. ISSN: 1943-2631. doi:10.1534/genetics.114.164442. http://dx.doi.org/10.1534/genetics.114.164442 (2014).

23. Queiroz de Oliveira, A., Poczynek, M., Maris Machado Bittar, C. & Gonçalves de Lima, C. Linear mixed-effects models and least confounded residuals in the modelling of Holstein calves' performance. *Brazilian Journal of Biometrics* **42,** 103–118. ISSN: 2764-5290. doi:10.28951/bjb.v42i2.681. http://dx.doi.org/10.28951/bjb.v42i2.681 (2024).

24. Rao, C. R. Estimation of Heteroscedastic Variances in Linear Models. *Journal of the American Statistical Association* **65,** 161–172. ISSN: 1537-274X. doi:10.1080/01621459.1970.10481070. http://dx.doi.org/10.1080/01621459.1970.10481070 (1970).

25. Rao, C. Estimation of variance and covariance components—MINQUE theory. *Journal of Multivariate Analysis* **1,** 257–275. ISSN: 0047-259X. doi:10.1016/0047-259x(71)90001-7. http://dx.doi.org/10.1016/0047-259X(71)90001-7 (1971).

26. Rao, C. Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis* **1,** 445–456. ISSN: 0047-259X. doi:10.1016/0047-259x(71)90019-4. http://dx.doi.org/10.1016/0047-259X(71)90019-4 (1971).

27. Resende, R. T., Hickey, L., Amaral, C. H., Peixoto, L. L., Marcatti, G. E. & Xu, Y. Satellite-enabled enviromics to enhance crop improvement. *Molecular Plant* **17,** 848–866. ISSN: 1674-2052. doi:10.1016/j.molp.2024.04.005. http://dx.doi.org/10.1016/j.molp.2024.04.005 (2024).

28. Resende, R. T., Piepho, H.-P., Rosa, G. J. M., Silva-Junior, O. B., e Silva, F. F., de Resende, M. D. V. & Grattapaglia, D. Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theoretical and Applied Genetics* **134,** 95–112. ISSN: 1432-2242. doi:10.1007/s00122-020-03684-z. http://dx.doi.org/10.1007/s00122-020-03684-z (2021).

29. Resende, R. T., Xavier, A., Silva, P. I. T., Resende, M. P. M., Jarquin, D. & Marcatti, G. E. GIS-based G×E modeling of maize hybrids through enviromic markers engineering. *New Phytologist* **245,** 102–116. ISSN: 1469-8137. doi:10.1111/nph.19951. http://dx.doi.org/10.1111/nph.19951 (2025).

30. Rogers, A. R. & Holland, J. B. Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. *G3 Genes|Genomes|Genetics* (ed Lipka, A) ISSN: 2160-1836. doi:10.1093/g3journal/jkab440. http://dx.doi.org/10.1093/g3journal/jkab440 (2022).

31. Searle, S. R. & Khuri, A. I. *Matrix algebra useful for statistics* ISBN: 978-1-118-93514-9. https://www.wiley.com/en-us/Matrix+Algebra+Useful+for+Statistics%2C+2nd+Edition-p-9781118935149 (John Wiley & Sons, 2017).

32. Solaymani, S., Ayatollahi Mehrgardi, A., Esmailizadeh, A., Tusell, L. & Momen, M. Performance of pedigree and various forms of marker-derived relationship coefficients in genomic prediction and their correlations. *Journal of Animal Breeding and Genetics* **137,** 423–437. ISSN: 1439-0388. doi:10.1111/jbg.12467. http://dx.doi.org/10.1111/jbg.12467 (2020).

33. Teixeira, A. P. *et al.* Cell functional enviromics: Unravelling the function of environmental factors. *BMC Systems Biology* **5.** ISSN: 1752-0509. doi:10.1186/1752-0509-5-92. http://dx.doi.org/10.1186/1752-0509-5-92 (2011).

34. VanRaden, P. Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91,** 4414–4423. ISSN: 0022-0302. doi:10.3168/jds.2007-0980. http://dx.doi.org/10.3168/jds.2007-0980 (2008).

35. Wright, S. Systems of mating. I. The biometric relations between parent and offspring. *Genetics* **6,** 111–123. ISSN: 1943-2631. doi:10.1093/genetics/6.2.111. http://dx.doi.org/10.1093/genetics/6.2.111 (1921).

36. Xavier, A., Muir, W. M. & Rainey, K. M. bWGR: Bayesian whole-genome regression. *Bioinformatics* **36** (ed Schwartz, R.) 1957–1959. ISSN: 1367-4811. doi:10.1093/bioinformatics/btz794. http://dx.doi.org/10.1093/bioinformatics/btz794 (2020).