University of São Paulo
"Luiz de Queiroz" College of Agriculture

snpReady and BGGE: R packages to prepare genomic datasets and perform genome-enabled predictions

Italo Stefanine Correia Granato

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2018

Italo Stefanine Correia Granato
Agronomist

snpReady and BGGE: R packages to prepare datasets and perform genome-enabled predictions

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **ROBERTO FRITSCHE NETO**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2018

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Granato, Italo Stefanine Correia

snpReady and BGGE: R packages to prepare datasets and perform genome-enabled predictions / Italo Stefanine Correia Granato. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2018.

40 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1 R 2 Seleção genômica 3 Interação genótipo x ambiente 4 Controle de qualidade 5 Modelos mistos Bayesianos I Título

ACKNOWLEDGMENTS

To my parents Luzinete and Geraldo for all the caring, dedication and commitment to my personal and professional formation.

To my family for the affection and support to reach my goals.

To my advisor Roberto Fritsche Neto, for the guidance, encouragement, and trust through years.

To ESALQ, especially to the Graduate Program in Genetics and Plant Breeding, for the opportunity and the CAPES and CNPq for the granting of the scholarship.

To my colleagues from Allogamous Breeding Laboratory. All of you have an important part in this work.

To the CIMMYT staff, especially Dr. José Crossa and Juan Burgueño, and my Mexican friends for hosting and given me the support needed to complete my job.

To the old friends who have helped me to get here, in especial, Andrea Lanna, Lorena Batista and João Paulo for backing me up and the lifetime friendship.

To all who directly or indirectly contributed to the accomplishment of this work.

CONTENTS

RESUMO

snpReady e BGGE: pacotes do R para preparar dados genômicos e realizar predições genômicas

O uso de marcadores moleculares permite um aumento na eficiência da seleção, bem como uma melhor compreensão dos recursos genéticos em programas de melhoramento. No entanto, com o aumento do número de marcadores, é necessário o processamento deste antes de deixa-lo disponível para uso. Além disso, para explorar a interação genótipo x ambiente (GA) no contexto da predição genômica, algumas matrizes de covariância precisam ser obtidas antes da etapa de predição. Assim, com o objetivo de facilitar a introdução de práticas genômicas nos programa de melhoramento, dois pacotes em R foram desenvolvidos. O primeiro, *snpReady*, foi criado para preparar conjuntos de dados para realizar estudos genômicos. Este pacote oferece três funções para atingir esse objetivo, organizando e aplicando o controle de qualidade, construindo a matriz de parentesco genômico e com estimativas de parâmetros genéticos populacionais. Além disso, apresentamos um novo método de imputação para marcas perdidas. O segundo pacote é o *BGGE,* criado para gerar *kernels* para alguns modelos genômicos de interação GA e realizar predições genômicas. Consiste em duas funções (*getK* e *BGGE*). A primeira é utilizada para criar *kernels* para os modelos GA, e a última realiza predições genômicas, com alguns recursos especifico para os *kernels* GA que diminuem o tempo computacional. Os recursos abordados nos dois pacotes apresentam uma opção rápida e direta para ajudar a introdução e uso de análises genômicas nas diversas etapas do programa de melhoramento.

Palavras-chave: R; Seleção genômica; Interação genótipo x ambiente; Controle de qualidade; Modelos mistos Bayesianos

ABSTRACT

snpReady and BGGE: R packages to prepare genomic datasets and perform genome-enabled predictions

The use of molecular markers allows an increase in efficiency of the selection as well as better understanding of genetic resources in breeding programs. However, with the increase in the number of markers, it is necessary to process it before it can be ready to use. Also, to explore Genotype x Environment (GE) in the context of genomic prediction some covariance matrices needs to be set up before the prediction step. Thus, aiming to facilitate the introduction of genomic practices in the breeding program pipelines, we developed two R-packages. The former is called *snpReady*, which is set to prepare data sets to perform genomic studies. This package offers three functions to reach this objective, from organizing and apply the quality control, build the genomic relationship matrix and a summary of a population genetics. Furthermore, we present a new imputation method for missing markers. The latter is the *BGGE* package that was built to generate kernels for some GE genomic models and perform predictions. It consists of two functions (*getK* and *BGGE*). The former is helpful to create kernels for the GE genomic models, and the latter performs genomic predictions with some features for GE kernels that decreases the computational time. The features covered in the two packages presents a fast and straightforward option to help the introduction and usage of genome analysis in the breeding program pipeline.

Keywords: R; Genome selection; Genotype x environment; Quality control; Bayesian mixed models

1. INTRODUCTION

In recent years, advances in genomics through next-generation sequencing have allowed the discovery of millions of new markers, which have continuously been used in studies of important agronomic traits (Edwards and Batley 2010). Thus, markers enable studies that allow association of markers with the phenotypic expression of traits of interest or, additionally, in research that involves population studies at loci level allowing an efficient management of genetic resources (Peiffer et al. 2014; Swarts et al. 2017). Unfortunately, the data obtained from genotyping platforms are not readily usable. Hence, there is the need for quality control and appropriate transformations of these data such as removal markers with some missing rate (call rate), low minor allele frequency (MAF), and imputation. After a proper quality control, markers can be used to obtain population genetic parameters for several purposes, for instance, to estimate the genetic variability available, identify and separate the germplasm into heterotic groups, and rare and exclusive alleles.

In the context of genomic prediction, several models have been proposed, and among them, the linear mixed model Genomic BLUP (G-BLUP) has stood out due to its simplicity and low computational requirements (Gianola et al. 2014). Given its popularity, there are many algorithms to carry it out and build the relationship matrices as well. In the proposal of the genome selection by Meuwissen et al. (2001) Bayesian models have been introduced in the context of whole-genome regression and since then have become common in genomic prediction (Gianola 2013). Within this framework, the appropriate prior distributions and simulations via Markov chain Monte Carlo (MCMC) allow convergence for predictive posterior distributions that do not have solution analytically. Using molecular markers in classical parametric regression can lead to the problem of $n \ll p$ setting, which can be suppressed by using semi-parametric regression as RKHS or the linear mixed models (Gianola and Van Kaam 2008; de los Campos et al. 2010)

Usually, genomic predictions do not take into account Genotype x Environment (GE) interaction. Nevertheless, in the breeding program pipelines, genotypes are evaluated in multi-environment trials leading to GE interaction. Recently, it has been proved the advantage of genomic models that take into account information from multi-environment trials simultaneously (Burgueño et al. 2012). Hence, a family of genomic models was developed to account the GE interaction besides allows incorporating fixed effects of environments and several genetic effects in a variety of linear mixed models (Jarquín et al. 2014; Lopez-Cruz et al. 2015; e Souza et al.

2017). These models are complex and need particular covariance matrices to be built before genome prediction.

In this context, we propose two R packages to facilitate the adoption of genome selection in breeding pipelines. The snpReady package was built to prepare datasets to run genomic analyses and has three functions, the raw.data that allows quality control and conversion data from genotyping platform into different formats. Also, we proposed and implemented a fast and straightforward method of imputation; the G.matrix, allows to create genomic relationship matrices; popgen estimates the genetics of population parameters. The BGGE package was built to generate kernels for some GE genomic models and perform genomic prediction. It consists of two functions, getK, a to create kernels for the GE genomic models; BGGE, the function to perform genomic prediction.

References

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52:707–719. doi: 10.2135/cropsci2011.06.0299

de los Campos G, Gianola D, Rosa GJ, et al (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb) 92:295–308. doi: 10.1017/S0016672310000285

e Souza MB, Cuevas J, Couto EG de O, et al (2017) Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype × Environment Interaction. G3-Genes Genom Genet g3.117.042341. doi: 10.1534/g3.117.042341

Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8:2–9. doi: 10.1111/j.1467-7652.2009.00459.x

Gianola D (2013) Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 194:573–596. doi: 10.1534/genetics.113.151753

Gianola D, Van Kaam JBCHM (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303. doi: 10.1534/genetics.107.084285

Gianola D, Weigel K, Krämer N, et al (2014) Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One. doi: 10.1371/journal.pone.0091693

Jarquín D, Crossa J, Lacaze X, et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607. doi: 10.1007/s00122-013-2243-1

Lopez-Cruz M, Crossa J, Bonnett D, et al (2015) Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker × Environment Interaction Genomic Selection Model. G3-Genes Genom Genet 5:569–82. doi: 10.1534/g3.114.016097

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Peiffer JA, Romay MC, Gore MA, et al (2014) The genetic architecture of maize height. Genetics 196:1337–56. doi: 10.1534/genetics.113.159152

Swarts K, Gutaker RM, Benz B, et al (2017) Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science 357:512–515. doi: 10.1126/science.aam9425

2. *SNPREADY*: A TOOL TO ASSIST BREEDERS IN GENOMIC ANALYSIS

ABSTRACT

The snpReady R-package is a new instrument developed to help breeders in genomic projects such as genomic prediction and association studies. This package offers three different methods to build the genomic relationship matrix, a new imputation method for missing markers based on Wright`s theory, and a population genetics overview. Therefore, we implemented three functions (raw.data, G.matrix, and popgen). Hence, this tool allows the raw data to be transformed from different genotyping platforms to numeric matrices and performs quality control (missing data and allele frequency). Moreover, the package generates and exports four different relationship matrices depending on the purpose and software to be used in further analysis. Finally, based on the genotypic matrix, the package estimates the genetic variability, effective population size, and endogamy, among other population genetic parameters. Empirical comparisons between the method of imputation proposed and other well-known approaches have shown a lower accuracy of imputation, however, with no significant impact on the genome prediction accuracies when a lower amount of missing data is allowed. The functions and arguments were designed to carry out the preparation of genomic datasets in a straightforward, fast, and more computationally efficient way. The package and its details are available at CRAN.

Keywords: R; Software; Quality control; Genomic data; Relationship matrix

2.1. INTRODUCTION

In recent years, advances in genomics and next-generation sequencing have allowed the discovery of millions of new markers, which have continuously been used in studies of important agronomic traits (Edwards and Batley 2010). Due to new high-throughput sequencing and genotyping technologies, the cost per datapoint has declined, and the amount of information has increased considerably (Buermans and den Dunnen 2014). With this information, breeders have focused on studies that allow the association of markers with the phenotypic expression of characteristics of interest or, additionally, in research that involves population studies and diversity analyses.

Unfortunately, the data obtained from genotyping platforms are not readily usable. Hence, there is the need for quality control and appropriate transformations of these data, such as call rate, minor allele frequency and imputation. Regarding the latter example, many imputation methods have been developed, though most are computationally intense. For instance, there are methods based on hidden Markov models, which are implemented in some

software, such as BEAGLE (Browning and Browning 2016) and IMPUTE v2 (Howie et al. 2009). However, these approaches require the construction of the haplotype phase. Therefore, we proposed a simple and fast method of imputation, which is implemented by the raw.data function. This method is based on the Wright equilibrium theory (Wright 1922).

In genomic prediction, the preparation and computation of the genomic kinship matrix are important in the prediction models of breeding programs (Forni et al. 2011; Da et al. 2014). Among the several prediction methods, linear mixed model genomic prediction (G-BLUP) has stood out due to its simplicity and low computational requirements (Gianola et al. 2014). Given its popularity, there are many algorithms to carry this function out and build the relationship matrices as well. However, there is a different package to each one of those algorithms. Therefore, the G.matrix function was developed to provide a tool that allows researchers to create different types of relationship matrices in the proper format for many software packages.

Population genetics parameters have been used in breeding programs for several purposes, including estimating the genetic variability available, identifying and separating the germplasm into heterotic groups, and rare and exclusive alleles in stratified populations. In this context, we built the popgen function, which provides some fundamental parameters and measures of the populations under selection that reflect the importance and usefulness of those markers.

Hence, in the following sections, we present this new tool, which is a straightforward approach to assist genomic analysis in preparation of genomic data and estimation of population parameters. Moreover, for quality control of genomic data, we also present a comparison between the imputation method proposed in our package to other well-known methods.

2.2. MATERIAL AND METHODS

2.2.1. Data sets

To demonstrate the package functions, we used a set of 452 tropical maize single-crosses provided by Helix Sementes®, Brazil. Hybrids were obtained from crosses between 128 inbred lines and were evaluated for grain yield (GY) and plant height (PH). Field trials were carried out using a randomized complete block design with two replicates each, allocated across five sites for GY and three for PH during the growing season of 2015.

Inbred lines were genotyped via the Affymetrix® Axiom® Maize Genotyping Array (Unterseer et al. 2014) with 660K SNP markers. Quality control for SNPs was made based on Call Rate (CR), in which all markers with any missing data point were excluded, and Minor Allele

Frequency (MAF) procedures, in which markers with a low level of polymorphism (MAF < 0.05) were removed. Hybrid genotypes were scored by an allelic combination of homozygous markers of parental lines. After quality control, 37,625 SNP were used to compare the imputation methods.

2.2.2. Imputation simulation

We simulated ten replicates of missing data. In each replicate, 60% of the markers were allowed to have up to 5% missing data, i.e., the amount of missing data points could not exceed the MAF and call rate thresholds, which were 0.05 and 0.95, respectively. The following three methods of imputation were compared: a map-dependent method and two map-independent methods.

The map-dependent method BEAGLE uses a hidden markov model (HMM) to reconstruct missing genotypes by the flanking markers. Hence, this methods needs the localized LD structure and cluster haplotypes at each marker to improve prediction of alleles at one group of markers given alleles at downstream markers on a haplotype (Browning and Browning 2016). The random method assumes Hardy-Weinberg equilibrium for all loci. Thus, missing values for a marker j are sampled with the following probabilities:

$$P(x_{ij}) = \begin{cases} p(x=0) = (1-p_j)^2 \\ p(x=1) = 2p_j(1-p_j) \\ p(x=2) = p_j^2 \end{cases} \quad i = 1, 2, \ldots, n \tag{1}$$

where p_j is the minor allele frequency of marker j.

For the k-nearest neighbor imputation (kNNI) (Troyanskaya et al. 2001), missing data points were imputed by replacing them with the weighted average of the data points at the k closest markers. The marker distance was constructed using the Euclidean distance.

2.2.3. Imputation accuracy

The imputation accuracy was assessed by the rate of correct imputation (C):

$$C = \frac{match}{total\ of\ missing}$$

where $match$ is the number of right imputation.

For the kNNI method, data points are imputed as continuous values; thus, to assess the imputation accuracy, we used the R^2 equation proposed by Rutkoski et al. (2013). The R^2 was defined as:

$$R^2 = 1 - \frac{\sum_j \left(x_j\ true - x_j\ imputed\right)^2}{\sum_j \left[x_j\ true - mean(x)\right]^2}$$

where x is a vector of length j (amount of missing data) for a given marker. The simulations to convert the R^2 to the rate of success were carried out as presented by Rutkoski et al. 92013). Briefly, for each marker with missing data, we simulated 1001 vectors, each one with a length of 1,000, based on their own MAF. The first marker was used as the reference, and for each percentage of incorrect values (ranging from 0.01 to 100 with an interval of 0.1), the proportion of the miss-assignment was simulated, and the R^2 was estimated. The best rate of success was the simulation with the R^2-value closest to the original.

The imputation of the map-dependent method via HMM and the random method were conducted using default options in the R package Synbreed (Wimmer et al. 2012). In the HMM, the package uses the BEAGLE algorithm (Browning and Browning 2016). The kNNI imputation was made using all default options of the R package impute (Hastie et al. 2017). For each replicate of simulated missing datasets, the time (in seconds) required for the whole imputation process was recorded. All analyses were run in Windows with 8 GB of RAM and 4 CPUs core.

2.2.4. Assessing prediction accuracies

Predictions were carried out considering an additive genetic model. The following two kernels were used to compare imputed datasets: GBLUP (VanRaden 2008) and the Gaussian kernel (GK). The former is a linear kernel related to additive quantitative genetics theory, and the latter is a non-linear kernel that can capture small complex interactions and non-additive variation (de los Campos et al. 2010). Thus, the following general equation was used:

$$y = \mathbf{1}\mu + \boldsymbol{Z}u + \varepsilon \quad (2)$$

where y is the vector of genotypic values of hybrids; μ is the general mean; u is the random genetic effects, assuming that its distribution is multivariate normal and follows $u \sim N(0, \sigma_u^2 G)$; and ε is the residual random effects, in which it is assumed that the errors are independent with homogeneous variance, σ_ε^2, distributed as $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$. $\mathbf{Z}$ is the incidence matrix related to the vector u. Statistical analysis was performed by rrBLUP (Endelman 2011) for G-BLUP kernel and BGLR (Pérez and de los Campos 2014) for GK kernel.

Individuals were randomly divided into two populations: a training set (*TS*), consisting of 70% of the individuals, and a validation set (*VS*), with the remaining 30% of the individuals. The phenotypes of individuals from *VS* were assigned as missing and were then predicted based on the *TS*. To estimate the accuracies of each imputation simulation, we carried out 20 random

sampling between *TS* and *VS*, and for each one, we performed the correlation between predicted and observed genetic values from VS.

2.3. RESULTS AND DISCUSSION

2.3.1. Quality control

The *raw.data* function was developed to transform raw data derived from different SNP marker genotyping platforms and prepare them for use in genomic analyses. The arguments of *raw.data* are presented below:

```
raw.data(data, frame = c("long", "wide"), hapmap = NULL, base =
TRUE, sweep.sample = 1, call.rate = 0.95, maf = 0.05, imput = TRUE,
imput.type = c("wright", "mean", "knni"), outfile = c("012", "-
101","structure"), plot = FALSE)
```

The *data* argument is the marker matrix, which will be passed by the quality control. The dataset can be inputted in two *frames*, the wide, consisting of one row per sample and markers in the columns, and the long, consisting of four columns, for samples, markers and one for each allele. *Base* is a logical argument determining if the matrix is coded as the nitrogenous bases or numerical. The *sweep.sample*, *call.rate* and *maf* are thresholds for quality control based on missingness and frequency allele for samples and markers. Imputation is optional and case chosen three methods are currently supported. The new matrix after quality control is exported in three formats by *outfile*, in especial, a proper format to be used in the software STRUCTURE (Pritchard et al. 2000). Also, graphical output of quality control is generated when *plot* is true.

The basic workflow consists of recodification of the bases from counting the number of copies of the reference allele. For each marker locus, the reference allele is defined by alphabetical order. Thus, when one of the alleles present in the locus is the adenine (A), for each sample, the number of copies of this allele is counted. Quality control for SNP markers is mainly based on allele frequency and percentage of missing data. Markers with the minor frequency of the secondary allele (MAF) should be removed, as it is not very informative. In other words, the MAF is used to remove markers that do not exhibit relevant polymorphism among individuals. The proportion of missing data to be removed, i.e., markers with failures in genotyping call rate, is optimized for the call rate (CR) (Cooper et al. 2013). Thus, the function performs the MAF and CR and the threshold is determined by the user.

After elimination of markers through quality control, the function performs imputation, if chosen, of the remaining missing data through two methodologies. One method is based on Wright's theory and it is very similar to the probabilities presented in (1). However, for a missing

data point, the function will assume that the probability of a genotype is dependent on two factors: the allele frequency of the marker and the inbreeding coefficient of the individual. Thus, for any missing data point x_{ij}, in the marker j and individual i, the probabilities are:

$$P(x_{ij}) = \begin{cases} p(x=0) = (1-p_j)^2 + p_j(1-p_j)F_i \\ p(x=1) = 2p_j(1-p_j) - 2p_j(1-p_j)F_i \\ p(x=2) = p_j^2 + p_j(1-p_j)F_i \end{cases} \quad (3)$$

where p_j is the frequency of allele j, and F_i is the observed homozygosity of the individual i. The second approach is the mean, where the missing positions for an individual i are replaced by the mean of the marker. The third method is the kNN, which impute missing data points in markers by the mean of the *k*-nearest neighbor. The nearest markers are found by using a Euclidian distance. This method uses the function impute.knn from the package impute (Hastie et al. 2017).

2.3.1.1. Comparing imputation methods

The four methods of imputation tested can be divided into two classes: those that use information from neighboring SNPs and those that use only information from the markers alone. It was observed that methods considering neighboring SNPs had, on average, a high accuracy, achieving 99% of correct data point imputations for the BEAGLE algorithm and 96% for kNNI (Table 1). He et al. (2015) compared three map-dependents methods with one map-independent and found that the map-dependent outperforms the latter. These approaches take advantage of the physical linkage between markers and thus can be a more reliable estimator (Rutkoski et al. 2013). In our study, the methods based on Hardy-Weinberg equilibrium and Wright equilibrium were less accurate than the other two methods. However, we expect that the Wright equilibrium approach (proposed here) is more realistic than imputation at random or assuming Hardy-Weinberg equilibrium because in populations under selection, a large part of the loci of interest for a trait trend toward non-equilibrium.

Despite the difference in imputation accuracy, it was noticed that different imputation methods did not influence the accuracy of prediction (Table 2), regardless of the kernel used. In our comparison, the missing input rate was based on a standard procedure of filtering SNPs with less than 5% of data missing. Nonetheless, He et al. (2015) and Rutkoski et al. (2013) compared the GS accuracy in several methods of imputation with different levels of missingness and observed that differences between the methods increase as long as missing data increase. Furthermore, at low rates of missingness, the GS accuracy is similar between these methods. This dynamic exists because after filtering the SNPs by the quality control process, the amount of missing data tends to decrease, reducing the influence of missing data imputation on GS

methods. In this context, to perform the genome prediction, a simple imputation method as proposed or based on the mean (Rutkoski et al. 2013) could be enough.

2.3.2. Genomic relationship matrix

Additive and dominant genomic relationship matrices (GRM) can be created through the *G.matrix* function. These matrices are suitable for use in studies of genomic prediction through the GBLUP model and GWAS analyses.

```
G.matrix(M, method = c("VanRaden", "UAR", "UARadj", "GK"), format =
c("wide", "long"), plot = FALSE)
```

The *M* argument is the marker matrix obtained after the quality control and imputation process. Based on that, this function allows building three forms of GRM and the non-linear Gaussian kernel (GK) through *method* argument. Moreover, results can be exported in two *formats*, the wide, which presents a pairwise relationship in a square matrix, and the long, which presents only the lower diagonal of GRM inverse in three columns. The use of *plot* produces graphical outputs.

The first method is based on that proposed by VanRaden (2008), which creates additive and dominant relationship matrices. Thus, the marker matrix is reparametrized in the *W* matrix with m markers, in which, for any marker locus *j*:

$$w_j = \begin{cases} A_1A_1 = 0 - 2p_j \\ A_1A_2 = 1 - 2p_j \\ A_2A_2 = 2 - 2p_j \end{cases}$$

where $j = 1,2,\dots,m$ and p_j is the frequency of reference allele. Thus, the construction of the additive relationship matrix follows:

$$G_A = \frac{WW'}{trace(WW')/n}$$

in which *n* represents the number of individuals.

Along with the additive GRM and following the same proposal, there is the formation of the dominance genome matrix. Consequently, there is appropriate parameterization of the marker matrix S for dominance deviations with m markers, in which, for each marker locus *j*:

$$s_j = \begin{cases} A_1A_1 = -2\left(1 - p_j\right)^2 \\ A_1A_2 = 2p_j\left(1 - p_j\right) \\ A_2A_2 = 2p_j^2 \end{cases}$$

Thus, the dominance genomic relationship matrix is estimated by:

$$G_D = \frac{SS'}{trace(SS')/n}$$

These matrices have a different denominator from the one originally proposed. As reported by Forni et al. (2011), non-normalized matrices can inflate accuracy estimates, as well as provide the mean of the diagonal elements, and the inbreeding coefficient of the individuals, other than 1.

Yang et al. (2010) proposed the other two methods for estimation of the GRM and later was named by Powell et al. (2010) as *Unified Additive Relationship* (UAR) and adjusted UAR. These methodologies proposed a correction in the calculation of the individuals' inbreeding coefficient individual. Thus, the relationship between two genotypes is obtained by:

$$G_{UAR} = \frac{1}{N}\sum\nolimits_i A_{ijk} = \begin{cases} \frac{1}{N}\sum\limits_i \frac{(x_{ji} - 2p_j)(x_{jk} - 2p_j)}{2p_j(1-p_j)} & j \neq k \\ 1 + \frac{1}{N}\sum\limits_i \frac{x_{ji}^2 - (1 + 2p_j)x_{jk} + 2p_j^2}{2p_j(1-p_j)} & j = k \end{cases}$$

where N is the number of SNPs, and p_j is the frequency of reference allele. Yang et al. (2010) proposed an adjustment in the original UAR matrix to reduce the bias in estimation of variance in the relationship in causal loci. Thus, the adjusted UAR matrix is described as:

$$G_{UARa} = \begin{cases} \beta A_{jk} & j \neq k \\ 1 + \beta(A_{jk} - 1) & j = k \end{cases}$$

where β is a regression coefficient empirically established:

$$\beta = 1 - \frac{c + 1/N}{var(A_{jk})}$$

where c is a constant dependent on MAF of causal variants. Here, we assume $c = 0$ for causal loci and SNPs on the same spectrum of allele frequency.

Another important available GRM is the Gaussian kernel, which is a reproducing kernel (RK) matrix used especially in semi-parametric methods such as Reproducing Kernel Hilbert Space (RKHS) (Pérez-Elizalde et al. 2015) and is estimated as:

$$K(x_i, x_{i'}) = \exp\left(\frac{-d_{ii}^2}{q}\right)$$

where x_i and $x_{i'}$ are the vectors of markers for individuals i and i' respectively, $d_{ii'}^2 = \sum_k (x_{ik} - x_{i'k})^2$ is the square of the Euclidean distance, and q is its fifth quantile.

Regardless the GRM chosen, the package can generate and export these data in different formats (full pairwise matrix or columns) depending on the purpose and software to be used in further analysis.

2.3.3. Summary of population genetics

The *popgen* function presents an overview of the population and markers parameters based on genomic information.

```
popgen(M, subgroups = NULL, plot = FALSE)
```

The *M* argument is the marker matrix to estimate of populational genetic parameters. The use of *subgroups* allows assigning individuals to subpopulations and *plot* produces proper graphical outputs. Therefore, for each marker locus, the following are estimated: minor allele frequency (MAF), observed heterozygosity (Ho) and expected heterozygosity (He), the *Nei's genetic diversity index* (GD), and polymorphic information content (PIC), the χ^2 estimates for Hardy-Weinberg equilibrium test and its p-values. In addition, the function presents estimates of observed heterozygosity (Ho) and inbreeding (*Fi*), at the individual level. The latter is measured based on the amount of the observed number of homozygous genotypes within an individual relate to what would be expected under random mating (Keller et al. 2011).

The function has an option to assign subpopulations to the individuals, thus allowing estimates of the same population genetic parameters described previously at the subpopulation level. Furthermore, it estimates exclusive and rare alleles present in each subpopulation. Moreover, in the presence of population subdivisions, the effective population size and the additive and dominance variance components for each one will also be estimated. In addition, F-statistics (F_{IT}, F_{IS}, and F_{ST}) are estimated from subpopulations that must be previously assigned.

2.3.4. Graphical outputs

Each of the available functions produces graphical output. For the *raw.data()*, when the map is included, a plot with the proportion of removed markers by call rate and MAF and proportion of imputed data, for each chromosome. For the *G.matrix()*, a heatmap of the genomic relationship matrix and a 3D principal component plot is produced. Finally, the *popgen()* function produces a histogram for the estimates of minor allele frequency, Nei's genetic diversity, polymorphic information content and expected heterozygosity considering the whole population and subpopulations, when it is available. In addition, when information of subpopulations is available, a heat map of the pairwise F_{ST} between populations is produced.

More information about *snpReady* and its functions can be obtained from the package documentation (see the Comprehensive R Archive Network website: https://cran.r-project.org/web/packages/snpReady/index.html) and the package vignette.

REFERENCES

Browning BL, Browning SR (2016) Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98:116–126. doi: 10.1016/j.ajhg.2015.11.020

Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: Advances and applications. Biochim Biophys Acta - Mol Basis Dis 1842:1932–1941. doi: 10.1016/j.bbadis.2014.06.015

Cooper TA, Wiggans GR, VanRaden PM (2013) Short communication: Relationship of call rate and accuracy of single nucleotide polymorphism genotypes in dairy cattle1. J Dairy Sci 96:3336–3339. doi: 10.3168/jds.2012-6208

Da Y, Wang C, Wang S, Hu G (2014) Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS One. doi: 10.1371/journal.pone.0087666

de los Campos G, Gianola D, Rosa GJ, et al (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb) 92:295–308. doi: 10.1017/S0016672310000285

Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8:2–9. doi: 10.1111/j.1467-7652.2009.00459.x

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4:250–255. doi: 10.3835/plantgenome2011.08.0024

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43:1–7. doi: 10.1186/1297-9686-43-1

Gianola D, Weigel K, Krämer N, et al (2014) Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One. doi: 10.1371/journal.pone.0091693

Hastie T, Tibshirani R, Narasimhan B, Chu G (2016) impute: Imputation for microarray data. R package version 1.52.0

He S, Zhao Y, Mette M, et al (2015) Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC Genomics 16:168. doi: 10.1186/s12864-015-1366-y

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. doi: 10.1371/journal.pgen.1000529

Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics 189:237–249. doi: 10.1534/genetics.111.130922

Pérez-Elizalde S, Cuevas J, Pérez-Rodríguez P, Crossa J (2015) Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. J Agric Biol Environ Stat 20:512–532. doi: 10.1007/s13253-015-0229-y

Pérez P, de los Campos G (2014) Genome-wide regression & prediction with the BGLR statistical package. Genetics 198:483–495. doi: 10.1534/genetics.114.164442

Powell JE, Visscher PM, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet 11:800–5. doi: 10.1038/nrg2865

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945 LP-959. doi: 10.1111/j.1471-8286.2007.01758.x

Rutkoski JE, Poland J, Jannink J-L, Sorrells ME (2013) Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. G3-Genes Genom Genet 3:427–439. doi: 10.1534/g3.112.005363

Troyanskaya O, Cantor M, Sherlock G, et al (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–5.

Unterseer S, Bauer E, Haberer G, et al (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15:823. doi: 10.1186/1471-2164-15-823

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23. doi: 10.3168/jds.2007-0980

Wimmer V, Albrecht T, Auinger H-J, Schön C-C (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–2087. doi: 10.1093/bioinformatics/bts335

Wright S (1922) Coefficients of Inbreeding and Relationship. Am Nat 56:330–338. doi: 10.1086/279872

Yang J, Benyamin B, Mcevoy BP, et al (2010) Common SNPs explain a large proportion of heritability for human height. Nature 569:565–569. doi: 10.1038/ng.608

TABLES

Table 1 – Mean of time and accuracy measured by the rate of correct imputation (C) for ten runs of stochastic missing data simulation in 452 hybrids of maize.

Method	**Time***	**RS**
rawdata	18.82	0.609
random	46.43	0.606
BEAGLE	867.30	0.999
kNNI	27.61	0.96

* Time measured in seconds

Table 2 - Mean of prediction accuracy of genomic prediction for two statistical methods and four methods of imputation, for plant height (PH) and grain yield (GY) in 452 maize single-crosses (SD in parenthesis)

Method	PH		GY	
	GBLUP	GK	GBLUP	GK
Rawdata	0.771	0.784	0.528	0.656
	(0.028)	(0.026)	(0.059)	(0.044)
Random	0.772	0.784	0.529	0.656
	(0.028)	(0.026)	(0.059)	(0.044)
Beagle	0.772	0.784	0.529	0.656
	(0.028)	(0.027)	(0.059)	(0.044)
kNNI	0.772	0.784	0.528	0.656
	(0.028)	(0.027)	(0.06)	(0.044)

3. *BGGE*: A NEW PACKAGE FOR GENOMIC PREDICTION DEALING WITH GENOTYPE BY ENVIRONMENTS MODELS

ABSTRACT

One of the major issues in plant breeding is the occurrence of genotype by environment (GE) interaction. Several models have been created to understand this phenomenon and explore it. In the genomic era, several models were employed to simultaneously improve selection by using markers and account for GE interaction. Some of these models use special genetic covariance matrices. In addition, multi-environment trials scales are getting larger, and this increases the computational challenges. In this context, we propose an R package that, in general, allows building GE genomic covariance matrices and fitting linear mixed models, in particular, to a few genome GE models. Here we propose a function to create the genomic kernels needed to fit these models. This function makes genome predictions through a Bayesian linear mixed model approach. A particular treatment is given for sparse covariance matrices; in particular, to block diagonal matrices that are present in some GE models in order to decrease the computational demand. In empirical comparisons with Bayesian Genomic Linear Regression (BGLR), accuracies and the mean squared error were similar; however, the computational time was up to five times lower than when using the classic approach. Bayesian Genomic Genotype × Environment Interaction (BGGE) is a fast, efficient option to create genome GE kernels and make genomic predictions.

Keywords: R; Genotype × Environment interaction (G×E); Genomic selection; Bayesian linear mixed models

3.1. INTRODUCTION

Genomic selection has the advantage of saving time and resources when selecting genotypes by adopting predictive methods for complex traits, along with information on pedigree, molecular markers or even environmental covariates (Crossa *et al.* 2017). In the genomic selection proposed by Meuwissen *et al.* (2001), Bayesian models were introduced in the context of whole-genome regression and since then have become common in genomic prediction (Gianola 2013). Within this framework, appropriate prior distributions and simulations via Markov Chain Monte Carlo (MCMC) allow convergence for predictive posterior distributions that cannot be solved analytically. However, these methods require thousands of iterations to ensure convergence, so that if the model is complex, the sampling process can increase the computational time. In this context, attempts were made to reduce the computational time of Bayesian models with approaches that do not use MCMC, such as variational Bayesian methods

(Montesinos-López *et al.* 2017) and Integrated Nested Laplace Approximation (INLA) (Holand *et al.* 2013). These methods are faster; however, they have constraints that may lead to lower prediction accuracy, which is undesired.

Using molecular markers in classic parametric regression can lead to the problem of $n \ll p$ setting, which can be reduced by using such semi-parametric regression as RKHS (Reproducing Kernel Hilbert Spaces) or linear mixed models (Gianola and Van Kaam 2008; de los Campos *et al.* 2010). These approaches assume the contribution of molecular markers as a random variable in some distributions with a covariance matrix that consists of a scalar variance component and a known covariance kernel obtained by markers. This covariance kernel can model genetic effects as additive, dominance, and epistasis, as a mixture of these effects or even as genetic and non-genetic remaining effects (Crossa *et al.* 2010; Technow *et al.* 2012; Azevedo *et al.* 2015). Genomic predictions are usually made using models that do not take into account genotype by environment interaction (GE). Nevertheless, the advantage of genomic models that take into account information from multi-environment trials simultaneously has been proved (Burgueño *et al.* 2012). Hence, a family of genomic models was developed to account for GE interaction; these models also allow incorporating fixed effects of environments and several genetic effects into a variety of linear mixed models (Jarquín *et al.* 2014; Lopez-Cruz *et al.* 2015; e Souza *et al.* 2017).

In this paper, we propose the R package BGGE (Bayesian Genomic Genotype x Environment) (Granato *et al.* 2017) models that makes it possible to quickly adjust linear mixed models, especially to a family of genomic GE models, such as those proposed by Jarquín *et al.* (2014) and Lopez-Cruz *et al.* (2015). The increase in speed is achieved by reparameterization through orthogonal rotation of the random vectors, allowing the use of univariate distributions in the sampling process (Cavalier 2008; Cuevas *et al.* 2014). Also, some special treatments are given for structured dispersed covariance matrices, in particular, those structured as a block diagonal, prevalent in some GE models (e Souza *et al.* 2017). In section two, we present statistical models and algorithms with a generic linear mixed model and its Bayesian counterpart, which is the base of the Bayesian genomic GE prediction (BGGE) package, as well as the most representative parts of the prediction process and kernel construction for genomic GE models. In section three, the *getK* and *BGGE* functions are presented along with some examples of their use; we also compare them to other packages that use Bayesian approaches.

3.2. STATISTICAL MODELS AND ALGORITHMS

3.2.1. Classical linear mixed model

Consider the following basic linear mixed models that covers the diversity of models that can be applied to uni or multi-enviroments trials. Assume l vectors of random effects:

$$\boldsymbol{y} = \boldsymbol{\mu}1 + \boldsymbol{X\beta} + \sum_{r=1}^{l} \boldsymbol{u}_r + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{y}$ is the vector combining phenotypic observations. The scalar $\boldsymbol{\mu}$ is the common intercept or the mean. The matrix $\boldsymbol{X}$ represents the design matrix associated with the vector of fixed effects $\boldsymbol{\beta}$. The random vectors $\boldsymbol{u}_r$ $(r = 1, 2, \ldots, l)$ are assumed to be independent of other random effects. We expected that $\boldsymbol{u}_r$ follows a normal distribution with zero mean and a covariance matrix of the form $\Lambda_r = \varphi_r \boldsymbol{K}_r$, where φ_r is a scalar representing the unknown variance parameter to be estimated from $\boldsymbol{u}_r$ and $\boldsymbol{K}_r$ is a known symmetric positive semi-definite covariance matrix.

Finally, the random error vector $\boldsymbol{\varepsilon}$, of same length of $\boldsymbol{y}$, follows a normal distribution with zero-mean and form $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}\sigma_\varepsilon^2$, and $\boldsymbol{I}$ is an identity matrix.

3.2.2. Linear mixed model reparametrization

Singular value decomposition (SVD) is commonly used in parametric regression models applied to genomic prediction (Meuwissen *et al.* 2017). In linear mixed model, the covariance matrix can be approached by using the eigen-decomposition of $\boldsymbol{K}$ (de los Campos *et al.* 2010). Hence, $\boldsymbol{K} = \boldsymbol{USU'}$, where $\boldsymbol{S}$ is a diagonal matrix with the n positive or zero eigenvalues and $\boldsymbol{U}$ is a orthogonal matrix with the eigenvectors associated to the n eigenvalues. In order to facilitate reading, we use a single kernel model and considering that $\boldsymbol{y}^* = \boldsymbol{y} - \mu\boldsymbol{1} - \boldsymbol{X\beta}$. Cuevas *et al.* (2014) propose an orthogonal transformation by multiplying both sides of (1) by $'$:

$$\boldsymbol{U'y}^* = \boldsymbol{U'u} + \boldsymbol{U'\varepsilon} \quad (2)$$

The model (2) becomes:

$$\boldsymbol{d} = \boldsymbol{b} + \boldsymbol{e} \quad (3)$$

where $\boldsymbol{d} = \boldsymbol{U'y}^*$ $\boldsymbol{b} = \boldsymbol{U'u}$ and $\boldsymbol{e} = \boldsymbol{U'\varepsilon}$. The $\boldsymbol{b}$ follows that $\boldsymbol{U'u} \sim N(0, \boldsymbol{U'KU}\varphi) = N(0, \boldsymbol{U'USU'U}\varphi) = N(0, \boldsymbol{S}\varphi)$, considering $\boldsymbol{U'U} = \boldsymbol{I}$. Similarly, $\boldsymbol{e}$ follows that $\boldsymbol{U'\varepsilon} \sim N(0, \boldsymbol{U'U}\sigma_\varepsilon^2) = N(0, \boldsymbol{I}\sigma_\varepsilon^2)$. The rotation causes the elements of $\boldsymbol{b}$ to be independent with univariate normal distributions. Also, it is worth noting that the eigenvalues very close to zero

(less than $1x10^{-10}$) reflects the noise (and numerical errors) and their associated eigenvectors can be eliminated and therefore reducing the dimension of the matrices $\boldsymbol{U}$ and $\boldsymbol{S}$.

3.2.3. Bayesian linear mixed models

The BGGE solves the linear mixed models through a Bayesian hierarchical modelling. The distribution of the transformed data $\boldsymbol{d}$, given $\boldsymbol{b}$ and $\sigma_{\boldsymbol{\varepsilon}}^2$ is:

$$f(\boldsymbol{d}|\, \boldsymbol{b}, \sigma_{\varepsilon}^2\,) = \prod_{i=1}^{n} N(\boldsymbol{d}_i|\boldsymbol{b}_i, \sigma_{\varepsilon}^2) \quad (4)$$

The Bayesian regression assumes unknown variance parameters. As pointed out by de los Campos *et al.* (2010) there is a relationship between the singular values and the extent of the shrinkage. Thus, if $p(\boldsymbol{u}|\varphi) = N(\boldsymbol{u}|0, \boldsymbol{K}\varphi)$, then the conditional prior distribution of $\boldsymbol{b}_i$ is as it follows:

$$p(\boldsymbol{b}_i|\sigma_{\boldsymbol{u}}^2) = N(\boldsymbol{b}_i|0, \varphi s_i) \quad (5)$$

where s_i are the singular values and φ is the unknown scale. A scaled inverse χ^2 distribution is assigned to φ and $\sigma_{\boldsymbol{\varepsilon}}^2$. Hence, the joint posterior distribution of $(\boldsymbol{b}, \varphi, \sigma_{\boldsymbol{\varepsilon}}^2, Sc_{\varphi}, Sc_{\varepsilon})$ given $\boldsymbol{d}$ and $\boldsymbol{S}$ is:

$$\begin{aligned} p\big(\boldsymbol{b}, \varphi, \sigma_{\boldsymbol{\varepsilon}}^2, Sc_{\varphi}, Sc_{\varepsilon}\big)\, \big|\boldsymbol{d}, \boldsymbol{S}\big) & \\ \propto \left\{\prod_{i=1}^{n} N(d_i|b_i, \sigma_{\varepsilon}^2) N(b_i|0, \varphi s_i)\right\} & \\ \times \chi^{-2}\big(\varphi\big|\nu_{\varphi}, \nu_{\varphi} Sc_{\varphi}\big) p\big(Sc_{\varphi}\big) \chi^{-2}(\sigma_{\varepsilon}^2|\nu_{\varepsilon}, \nu_{\varepsilon}\, Sc_{\varepsilon}) & \end{aligned}$$

Flat prior distributions are assumed for μ and $\boldsymbol{\beta}$. For variance parameters priors, distributions are assumed scaled inverse χ^2, arbitrarily assigned to ν_{φ} and ν_{ε} a value of 3 in order to not generate infinite values, and a priori flat distribution for the scale (Sc_{φ}) hyperparameter (see Cuevas *et al.* 2014). From equations (5) and (6), conditional distributions can be constructed to generate the MCMCs through a Gibbs sampler. Details are presented in Appendix A.

3.2.4. Sparse matrices

In an attempt to speed up the prediction algorithm, some special treatments are given for sparse matrices. In some GE models (Jarquín *et al.* 2014; Lopez-Cruz *et al.* 2015), some random $\boldsymbol{u}$ effects have an associated covariance matrix $\boldsymbol{K}$ that can be considered sparse with submatrices in a known structure. Thus, instead of applying the eigen-decomposition in the $\boldsymbol{K}$,

we identify, individualize and apply the eigen-decomposition in these submatrices that compose the block diagonal.

3.2.5. Obtaining multi-environment kernels

Different multi-environment models are defined based on the construction of the kernel matrices, using information available on genotypes, molecular markers and the environment (Jarquín *et al.* 2014; e Souza *et al.* 2017). The construction of multi-environment kernels depends on two primary processes: the choice of covariance function and the multi-environment model.

3.2.5.1. Choice of covariance function

Two covariance or kernel functions are generated internally. GBLUP (GB) is the standard linear kernel from the properties of a multivariate normal distribution in linear mixed models and is usually referenced as the genomic relationship matrix. Thus, GB is obtained as follows:

$$\boldsymbol{GB} = \frac{XX'}{p}$$

where $\boldsymbol{X}$ is the marker matrix, and $\boldsymbol{p}$ is the number of markers. This matrix was proposed by VanRaden (2008) and since then has been successfully used in genomic prediction (de los Campos et al. 2009).

Another covariance function offered is the Gaussian kernel (GK). The GK appeared as a reproducing kernel (RK) in the semi-parametric model Reproducing kernel Hilbert spaces (RKHS) (González-Camacho *et al.* 2012) and is defined as follows:

$$\boldsymbol{GK} = K(x_i, x_j) = \exp\left(\frac{-hd_{ij}^2}{q}\right)$$

where h is the bandwidth parameter that controls the decay rate of covariance between genotypes, q is the percentile of the square of the Euclidean distance $d_{ij} = \sum_k (x_{ik} - x_{jk})^2$, a measure of genetic distance between individuals. Some results have shown better performance of GK over GB (Cuevas *et al.* 2016a; e Souza *et al.* 2017).

3.2.5.2. Choice of multi-environment model

Several models focused on genotype by environment genomic prediction have been developed (Jarquín *et al.* 2014; Lopez-Cruz *et al.* 2015; Cuevas *et al.* 2018). These models use different approaches to account for GE interaction in the context of genomic selection. Three

models are covered. First multi-environment model is the main genotypic effect model (MM), proposed by Jarquin et al. (2014) and assumes that genetic effects across environments are constant. Another model is the multi-environment, single variance genotype × environment deviation model (MDs), which is an extension of the main genetic effect model (MM), but incorporating a random deviation effect of genotype by environment interaction. Finally, the third model is the multi-environment, environment-specific variance genotype × environment deviation model (MDe). This model was proposed by Lopez-Cruz et al. (2015) to allow the genetic effects to change across the environments. Hence, it splits the genetic effect into two components: a common effect for all environments (across-environment effect) and a random deviation effect for each environment (environment-specific effect).

3.3. APPLICATION DESCRIPTION

DESCRIBING FUNCTIONS

In this section, we present the usage, and the main aspects of the two functions offered the *getK* and *BGGE*. The *getK* was proposed to create multi-environment kernels for the MM, MDs, and MDe models.

Box 1

```
getK(Y, X, kernel = c("GK", "GB"), setKernel = NULL, bandwidth = 1,
model = c("SM", "MM", "MDs", "MDe"), quantil = 0.5)
```

The box one contains the main arguments of the getK function. The data frame **`Y`** is the phenotypic file and must include information that connects genotypes and environments and the trait of interest. **`X`** is the marker matrix; the vector **`kernel`** is the covariance functions used to construct the ME kernels. In case the Gaussian kernel (GK), **`bandwidth`** and **`quantil`** arguments are equivalent to bandwidth parameter and the quantile as defined previously. In case of choosing other covariance functions than GB and GK, these kernels are passed by the argument **`setKernel`**. The present GE models (SM, MM, MDe, and MDs) must be defined in the **`model`**.

The *BGGE* function has the goal of performing genomic prediction through linear regression for continuous variables.

Box 2

```
BGGE(y, K, XF = NULL, ne, ite = 1000, burn = 200, thin = 3)
```

In the box 2 are the arguments for the BGGE function. The **`y`** argument is the response variable. **`K`** is a two level-list defining the kernel associated with each random effect to

be fitted. The **XF** is the design matrix for the fixed effects to be fitted. **ne** is a vector defining the number of genotypes in each environment and **ite**, **burn**, and **thin** defines the number of iterations of the sampler, the number of samples to be discarded, and the thinning used to compute posterior means, respectively. Further details about K and ne are given in the examples below.

In the following examples, the wheat data set will be used to demonstrate the application of the BGGE. This dataset comprehends 599 wheat lines derived from 25 Field trials grouped into four mega-environments, with the phenotypic trait grain yield (GY). The 599 wheat lines were genotyped using 1447 Diversity Array Technology (DArT) markers generated by Triticarte Pty. Ltd.. This dataset is available at the BGLR package (Pérez and de los Campos 2014).

EXAMPLE 1: FITTING MM MODEL

In this example, we show how to fit the main genotypic model (MM) (Jarquín *et al.* 2014) along with the linear kernel GBLUP. First, we obtain the kernel through *getK*.

Box 3

```
library(BGLR)
data(wheat)
env <- ncol(X)
gen <- nrow(X)
rownames(X) <- 1:gen

pheno_geno <- data.frame(env = gl(n = env, k = gen),
                         GID = gl(n=gen, k=1, length = gen*env),
                         Y = as.vector(wheat.Y))
K1 <- getK(Y = pheno_geno, X = X, kernel = "GB", model = "MM")
```

The phenotypic file must be provided as a data frame with three columns indentifying the environments, the individuals or genotypes and the trait of interest. When in the presence of marker matrix, is necessary to choose the covariance function to create the kernel. The *getK* returns a two-level list with the kernels for the respective model and a definition of the type of matrix. The MM model produces only one kernel considered as dense.

Box 4

```
ne <- as.vector(table(pheno_geno$env))
fit <- BGGE(y = pheno_geno$Y, K = K1, ne = ne)
```

In the box four is presented the basic syntax for *BGGE* function. The input for K is the two-level list returned by the *getK* function. The call above for the *BGGE* fits a multi-environment main genotypic model (MM), with a total of 1000 cycles of a Gibbs sampler (the default value for the number of iterations), and the first 200 samples are discarded (the default value for burn-in). Also, samples are collected at an interval of three (the default for thinning

interval). The *BGGE* function returns a list with estimated posterior means for each random term in the linear model and the genetic values predicted. To assess convergence and to estimate Monte Carlo error, samples of the intercept and random effects variances are stored and returned in the same output list.

EXAMPLE 2: FITTING MDE MODEL

In this example, we show how to fit the environment-specific variance genotype × environment deviation model (MDe) (Lopez-Cruz *et al.* 2015; Cuevas *et al.* 2016b) along with the non-linear Gaussian kernel (GK).

Box 5

```
K2 <- getK(Y = pheno_geno, X = X, kernel = "GK", bandwidth = 1,
model = "MDe")
fit <- BGGE(y = pheno_geno$Y, K = K2, ne = ne)
```

In the call from *getK*, using the Gaussian kernel, multi-environment kernels are obtained by default using a bandwidth parameter of one. However, this can be modified by `h` argument. For the MDe model, the function returns the kernels for the main genotypic effect and kernels for each environment. This model is characterized by structured matrices for specific environments.

For the MDe model, in the prediction step by calling *BGGE* the properties of structured matrices are used. The `ne` argument is used to extract the sub-matrices for each environment instead of decomposing in singular values the big sparse matrix. The BGGE returns the estimated posterior mean of genetic (main effect + specific-environment effects) and residual variances.

EXAMPLE 3: FITTING MULTI-KERNEL MULTI-ENVIRONMENT MODELS

When using the Gaussian kernel (GK), the problem of selecting the best bandwidth parameter arises. As pointed by de los Campos *et al.* (2010), with extreme bandwidth values the information of the markers is practically lost, making it necessary optimizing the best parameter. Endelman (2011) and Pérez-Elizalde *et al.* (2015) proposed two different approaches to optimizing this parameter via REML and Bayesian framework, respectively. Nevertheless, de los Campos et al. (2010) address the problem by proposing a multi-kernel approach in which a sequence of kernels are obtained from a grid of bandwidth parameters, named kernel averaging (KA).

Box 6

```
K3 <- getK(Y = pheno_geno, X = X, kernel = "GK", bandwidth =
c(0.25,1,2.5), model = "MDs")
fit <- BGGE(y = pheno_geno$Y, K = K3, ne = ne)
```

We will use the MDs model as an example. As the `bandwidth` argument accepts a vector as input, it can be used as a solution to create the multi-kernels using a range of bandwidth values. For the present models, getK will create $n \times v$ kernels, in which n is the number of basic kernels for each model and v is the number of bandwidth parameters.

EXAMPLE 4: FITTING ADDITIVE+DOMINANCE MODELS

Several kernels were proposed as *t* (Tusell *et al.* 2014) and exponential (Endelman 2011) as well as other estimators of the genomic relationship between subjects (Astle and Balding 2009; Yang *et al.* 2010; Wang and Da 2014) in attempt to improve predictions. Hence, it is possible to use others kernels than GB and GK to create the multi-environment kernels. In this example, we will illustrate how to apply external kernels to fit genome prediction to the model MDs (Jarquín *et al.* 2014). For instance, we use the additive and dominance matrices to fit the multi-environment MDs model. Using SNP matrix is it possible to compute the additive and dominance relationship matrices and it can be combined and used to build the multi-environment kernels.

Box 7

```
# GBa - Kernel GBLUP for additive
# GBd - Kernel GBLUP for dominance
Ker <- list(GBa, GBd)
K5 <- getK(Y = pheno_geno, setKernel = Ker, model = "MDs")
fit <- BGGE(y = pheno_geno$Y, K = K5, ne = ne)
```

In the initial call for *getK*, we introduce the `setKernel` argument that allows passing a list of other kernels than those computed internally. Thus, it will create $n \times k$ kernels, in which n is the number of basic kernels for each model and k is the number of kernels introduced by user.

EMPIRICAL COMPARISONS

The approach used in the BGGE using different features was compared against the standard Bayesian kernel regression proposed by de los Campos *et al.* (2010). The two approaches were compared empirically to the different genotype by environment models through prediction accuracy, predictive mean squared error (PMSE), variance components and time. A maize dataset (Helix Seeds Company) was used in the comparisons. It consists of 452 maize hybrids evaluated in five sites. For more details see e Souza *et al.* (2017). For this comparative study, we only used Grain Yield (GY). Credits in phenotyping to Helix Sementes Ltda, Brazil.

In order to assess accuracy model prediction, we perform 50 random partitions into training and test set. For each partition, 80% of the observations were set as training, and 20% as a test. For multi-environment models, the separation between training and test set was made according to two cross-validation designs, the cross-validation 1 (CV1) and cross-validation 2 (CV2) (Burgueño *et al.* 2012). The PMSE was assessed by computing the mean of the squares of the errors. The posterior of variance components were estimated using full data. Also, computational time was used in the comparison. The genomic GE models were fitted using the method RKHS in the package BGLR (Pérez and de los Campos 2014) and BGGE, using the Gibbs sampler with 30,000 iterations, a burn-in of 5000 and a thinning interval of 5. Kernels for GE models were built in the *getK* function.

The approach used for prediction has an orthogonal transformation of the observed data (y), as well as different prior variance assigned to the regression parameters. Despite the theoretically expected differences between these two approaches, the difference is not significant. For the two data sets, the residual variance was slightly lower when using the BGGE approach (Table 1). In contrast, the genetic variance components were high for BGGE. For accuracy assessed by the two cross-validation designs (CV1 and CV2), there is no clear advantage of using one method instead of other and differences are no longer than 0.02 (Table 2; Table 3). In the computational time, differences are up to five times slower using the BGGE approach (Table 4).

The primary difference between the two approaches relies on the reparametrization of observations and a prior that changes the shrinkage. In the context regression on markers, Cuevas *et al.* (2014) have found that new parameterization allows reducing the dimensionality by decreasing the number of parameters, which gives a computational advantage due to simulating a smaller number of parameters with univariate distributions. The prior is the main force driving various Bayesian models, however, if the variance matches between them, the differences between models would rely on the extent of theirs shrinkage effects (Gianola 2013). The prior assigned has to influence on the weighting factor, however, despite the different priors assumed and the features between the two approaches, their behavior was very similar. The extra features of sparse structure matrix assumed in the BGGE algorithm reduce dimensionality decreasing the computational time.

3.4. CLOSING REMARKS

The proposal package was built to perform genomic prediction for continuous variable focused on the genomic GE models. Using information from multi-environment trials can

improve predictions, and several models have been created (e Souza *et al.* 2017). However, each GE model has their properties, and thus specific kernels must be created.

The *getK* function has the purpose of efficiently generate the kernel for three genomic GE models. Some features were included like using covariance function created internally or another external kernel, which opens possibilities for various combinations of covariance matrices for GE models, as additive, dominant, and epistasis. Also, for the Gaussian kernel, different values of bandwidth parameters can be introduced to create the several kernels as defined in the kernel averaging (de los Campos *et al.* 2010). The output is in the proper format to be used in the BGGE prediction function.

The BGGE function uses another reparametrization and different priors for the hyperparameters (Cuevas *et al.* 2014) in the linear mixed model regression in the Bayesian context. These features reduce the dimensionality of the vector of regression parameters and change the regularization parameter. Also, we explore the properties of structured sparsity in some GE kernels to decrease the computational time. Therefore, the package arises as fast and efficient option to perform predictions of genetic values. The BGGE is programmed entirely in R, and has no dependence of other packages. Despite their speed the authors continue to work on this issue, hoping that future versions to reduce the computational time.

REFERENCES

Astle, W., and D. Balding, 2009 Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat. Sci. 24: 451–471.

Azevedo, C. F., M. D. V. de Resende, F. F. E Silva, J. M. S. Viana, M. S. F. Valente *et al.*, 2015 Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genet. 16: 105.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa, 2012 Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52: 707–719.

Cavalier, L., 2008 Nonparametric statistical inverse problems. Inverse Probl. 24: 34004.

Crossa, J., G. De Los Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín *et al.*, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci.

Cuevas, J., J. Crossa, O. Montesinos-Lopez, J. Burgueno, P. Perez-Rodriguez *et al.*, 2016 Bayesian Genomic Prediction with Genotype × Environment Interaction Kernel Models. G3-Genes Genom Genet 7: 41–53.

Cuevas, J., J. Crossa, V. Soberanis, S. Pérez-Elizalde, P. Pérez-Rodríguez *et al.*, 2016 Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models. Plant Genome 9: 0.

Cuevas, J., S. Pérez-Elizalde, V. Soberanis, P. Pérez-Rodríguez, D. Gianola *et al.*, 2014 Bayesian Genomic-Enabled Prediction as an Inverse Problem. G3-Genes Genom Genet 4: 1991–2001.

e Souza, M. B., J. Cuevas, E. G. de O. Couto, P. Pérez-Rodríguez, D. Jarquín *et al.*, 2017 Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype × Environment Interaction. G3-Genes Genom Genet g3.117.042341.

Endelman, J. B., 2011 Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4: 250–255.

Gianola, D., 2013 Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 194: 573–596.

Gianola, D., and J. B. C. H. M. Van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289–2303.

González-Camacho, J. M., G. de los Campos, P. Pérez, D. Gianola, J. E. Cairns *et al.*, 2012 Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125: 759–771.

Granato, I. S. C., F. J. Luna-Vázquez, and J. Cuevas, 2017 BGGE - R package.

Holand, A. M., I. Steinsland, S. Martino, and H. Jensen, 2013 Animal Models and Integrated Nested Laplace Approximations. G3: Genes|Genomes|Genetics 3: 1241–1251.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt *et al.*, 2014 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127: 595–607.

Legarra, A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92: 4656–4663.

Lopez-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland *et al.*, 2015 Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker × Environment Interaction Genomic Selection Model. G3-Genes Genom Genet 5: 569–82.

de los Campos, G., D. Gianola, G. J. Rosa, K. Weigel, and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. (Camb). 92: 295–308.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–85.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Meuwissen, T. H. E., U. G. Indahl, and J. Ødegård, 2017 Variable selection models for genomic selection using whole-genome sequence data and singular value decomposition. Genet. Sel. Evol. 49: 94.

Montesinos-López, O. A., A. Montesinos-López, J. Crossa, J. C. Montesinos-López, F. J. Luna-Vázquez *et al.*, 2017 A Variational Bayes Genomic-Enabled Prediction Model with Genotype × Environment Interaction. G3-Genes Genom Genet 7: g3.117.041202.

Pérez-Elizalde, S., J. Cuevas, P. Pérez-Rodríguez, and J. Crossa, 2015 Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. J. Agric. Biol. Environ. Stat. 20: 512–532.

Pérez, P., and G. de los Campos, 2014 Genome-wide regression & prediction with the BGLR statistical package. Genetics 198: 483–495.

Technow, F., C. Riedelsheimer, T. A. Schrag, and A. E. Melchinger, 2012 Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor. Appl. Genet. 125: 1181–94.

Tusell, L., P. Pérez-Rodríguez, S. Forni, and D. Gianola, 2014 Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: A case study with pig litter size and wheat yield. J. Anim. Breed. Genet. 131: 105–115.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–23.

Wang, C., and Y. Da, 2014 Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. PLoS One 9: e114484.

Yang, J., B. Benyamin, B. P. Mcevoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of heritability for human height. Nature 569: 565–569.

APPENDIX A

The conditional posterior distribution of $\boldsymbol{b}$:

$$p(\boldsymbol{b}|\boldsymbol{d}, S, \varphi, \tau) = \prod_{i=1}^{n} N\left(b_i \middle| \frac{\tau\, \varphi s_i d_i}{1 + \tau\, \varphi s_i}, \frac{s_i \varphi}{1 + \tau\, \varphi s_i}\right)$$

where $\tau = 1/\sigma_\varepsilon^2$. Assuming that $\boldsymbol{b} = \boldsymbol{U}'\boldsymbol{u}$, the genetic effects $\boldsymbol{u}$ can be recovered by $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{b}$.

The conditional distribution of σ_ε^2, φ, Sc_φ and Sc_ε are:

$$p(\sigma_\varepsilon^2|\boldsymbol{y}^*, \boldsymbol{u}, \nu_\varepsilon, Sc_\varepsilon) = \chi^{-2}(\sigma_\varepsilon^2|n + \nu_\varepsilon, (\boldsymbol{y}^* - \boldsymbol{u})'(\boldsymbol{y}^* - \boldsymbol{u}) + Sc_\varepsilon)$$

$$p(\varphi|\boldsymbol{b}, \nu_\varphi, Sc_\varphi) = IG\left(\varphi \middle| \frac{\nu_\varphi + n}{2}, \frac{\boldsymbol{b}'\boldsymbol{S}^{-1}\boldsymbol{b} + \nu_\varphi Sc_\varphi}{2}\right)$$

$$p(Sc_\varphi|\boldsymbol{d}, \boldsymbol{b}, \varphi, \nu_\varphi) = Ga\left(Sc \middle| \frac{\nu_\varphi}{2} + 1, \frac{\nu_\varphi}{2\varphi}\right)$$

$$p(Sc_\varepsilon|\sigma_\varepsilon^2, \nu_\varepsilon) = Ga\left(Sc \middle| \frac{\nu_\varepsilon}{2} + 1, \frac{\nu_\varepsilon}{2\sigma_\varepsilon^2}\right)$$

TABLES

Table 1 - HEL data set. Estimate of variance components obtained by BGGE and BGLR functions, for the multi-environment models, main genotypic effect (MM), single variance G×E deviation model (MDs) and environment-specific variance G×E deviation model (MDe) with two kernels, GBLUP (GB) and Gaussian (GK).

Factor	BGGE						BGLR					
	MM		MDs		MDe		MM		MDs		MDe	
	GK	GB	GK	GB	GK	GB	GK	GB	GK	GB	GK	GB
σ^2	0.582 (0.02)	0.749 (0.03)	0.278 (0.02)	0.591 (0.02)	0.247 (0.02)	0.592 (0.02)	0.585 (0.02)	0.751 (0.03)	0.288 (0.02)	0.595 (0.02)	0.279 (0.02)	0.605 (0.02)
σ_g^2	0.821 (0.1)	0.356 (0.08)	0.938 (0.1)	0.370 (0.08)	0.931 (0.1)	0.390 (0.09)	0.783 (0.1)	0.307 (0.07)	0.882 (0.1)	0.311 (0.07)	0.899 (0.1)	0.33 (0.07)
σ_{gE}^2			0.525 (0.06)	0.188 (0.03)					0.504 (0.06)	0.181 (0.03)		
σ_1^2					0.372 (0.1)	0.237 (0.08)					0.318 (0.08)	0.178 (0.07)
σ_2^2					0.769 (0.2)	0.076 (0.06)					0.526 (0.16)	0.05 (0.03)
σ_3^2					0.370 (0.09)	0.259 (0.08)					0.312 (0.07)	0.209 (0.06)
σ_4^2					1.143 (0.21)	0.255 (0.09)					0.922 (0.18)	0.167 (0.07)
σ_5^2					0.688 (0.16)	0.158 (0.07)					0.507 (0.14)	0.101 (0.05)

Table 2 - HEL data set. Mean correlation (50 partitions) and predictive mean squared error (PMSE) for CV1 design, obtained by BGGE and BGLR packages. For the multi-environment models, main genotypic effect (MM), single variance G×E deviation model (MDs) and environment-specific variance G×E deviation model (MDe) with two kernels, GBLUP (GB) and Gaussian (GK) for Grain Yield (standard deviation in parentheses).

	Mean Correlation						PMSE					
	MM		MDs		MDe		MM		MDs		MDe	
	GK	GB	GK	GB	GK	GB	GK	GB	GK	GB	GK	GB
						BGGE						
1	0.575 (0.09)	0.426 (0.11)	0.752 (0.06)	0.607 (0.08)	0.755 (0.05)	0.618 (0.09)	0.684 (0.15)	0.824 (0.14)	0.436 (0.08)	0.652 (0.15)	0.435 (0.08)	0.64 (0.14)
2	0.506 (0.09)	0.361 (0.07)	0.54 (0.09)	0.394 (0.08)	0.538 (0.09)	0.394 (0.08)	0.747 (0.18)	0.879 (0.17)	0.715 (0.16)	0.858 (0.16)	0.712 (0.13)	0.855 (0.15)
3	0.662 (0.06)	0.533 (0.07)	0.758 (0.05)	0.662 (0.07)	0.754 (0.05)	0.671 (0.04)	0.591 (0.11)	0.723 (0.13)	0.428 (0.06)	0.565 (0.1)	0.432 (0.07)	0.559 (0.08)
4	0.346 (0.1)	0.219 (0.1)	0.527 (0.06)	0.321 (0.1)	0.524 (0.08)	0.339 (0.09)	0.901 (0.17)	1.031 (0.23)	0.724 (0.13)	0.938 (0.15)	0.73 (0.13)	0.916 (0.19)
5	0.455 (0.1)	0.293 (0.11)	0.576 (0.07)	0.376 (0.1)	0.555 (0.09)	0.383 (0.11)	0.794 (0.18)	0.939 (0.19)	0.677 (0.14)	0.88 (0.19)	0.693 (0.14)	0.872 (0.19)
						BGLR						
2	0.573 (0.07)	0.427 (0.11)	0.753 (0.05)	0.619 (0.09)	0.751 (0.07)	0.614 (0.09)	0.682 (0.12)	0.817 (0.15)	0.444 (0.09)	0.628 (0.13)	0.439 (0.09)	0.642 (0.15)
3	0.509 (0.11)	0.361 (0.1)	0.555 (0.07)	0.402 (0.12)	0.548 (0.09)	0.382 (0.1)	0.741 (0.16)	0.866 (0.17)	0.692 (0.13)	0.848 (0.17)	0.70 (0.13)	0.858 (0.18)
4	0.664 (0.06)	0.544 (0.07)	0.762 (0.04)	0.663 (0.06)	0.759 (0.05)	0.662 (0.05)	0.59 (0.08)	0.714 (0.12)	0.424 (0.06)	0.565 (0.09)	0.427 (0.07)	0.569 (0.09)
5	0.349 (0.1)	0.223 (0.1)	0.515 (0.09)	0.349 (0.1)	0.518 (0.08)	0.328 (0.09)	0.9 (0.19)	1.028 (0.15)	0.734 (0.14)	0.901 (0.15)	0.732 (0.12)	0.918 (0.2)
6	0.459 (0.08)	0.296 (0.1)	0.561 (0.07)	0.401 (0.1)	0.551 (0.08)	0.38 (0.11)	0.791 (0.19)	0.927 (0.16)	0.689 (0.13)	0.852 (0.19)	0.698 (0.15)	0.864 (0.18)

Table 3 - HEL data set. Mean correlation (50 partitions) and predictive mean squared error (PMSE) for CV2 design, obtained by BGGE and BGLR packages. For the multi-environment models, main genotypic effect (MM), single variance G×E deviation model (MDs) and env variance G×E deviation model (MDe) with two kernels, GBLUP (GB) and Gaussian (GK) for Grain Yield (standard deviation in parentheses).

	Mean Correlation						PMSE					
	MM		MDs		MDe		MM		MDs		MDe	
	GK	GB	GK	GB	GK	GB	GK	GB	GK	GB	GK	GB
						BGGE						
1	0.595 (0.08)	0.51 (0.11)	0.807 (0.04)	0.683 (0.08)	0.804 (0.05)	0.678 (0.07)	0.641 (0.09)	0.745 (0.14)	0.354 (0.06)	0.56 (0.13)	0.351 (0.06)	0.564 (0.12)
2	0.601 (0.08)	0.469 (0.1)	0.627 (0.08)	0.472 (0.08)	0.616 (0.09)	0.473 (0.11)	0.647 (0.13)	0.778 (0.15)	0.608 (0.1)	0.78 (0.17)	0.619 (0.11)	0.784 (0.17)
3	0.645 (0.06)	0.584 (0.08)	0.776 (0.04)	0.697 (0.05)	0.778 (0.04)	0.693 (0.05)	0.592 (0.09)	0.674 (0.1)	0.4 (0.05)	0.519 (0.07)	0.395 (0.06)	0.528 (0.08)
4	0.427 (0.1)	0.296 (0.1)	0.591 (0.07)	0.39 (0.08)	0.592 (0.08)	0.395 (0.1)	0.854 (0.14)	0.956 (0.2)	0.648 (0.09)	0.863 (0.19)	0.645 (0.1)	0.863 (0.15)
5	0.558 (0.07)	0.396 (0.11)	0.666 (0.06)	0.466 (0.08)	0.662 (0.06)	0.468 (0.09)	0.694 (0.16)	0.844 (0.19)	0.556 (0.1)	0.794 (0.17)	0.565 (0.11)	0.79 (0.18)
						BGLR						
1	0.598 (0.09)	0.51 (0.12)	0.804 (0.05)	0.67 (0.09)	0.8 (0.05)	0.669 (0.07)	0.636 (0.1)	0.745 (0.15)	0.358 (0.08)	0.566 (0.13)	0.365 (0.07)	0.578 (0.1)
2	0.591 (0.08)	0.475 (0.1)	0.626 (0.07)	0.46 (0.09)	0.625 (0.07)	0.482 (0.07)	0.652 (0.14)	0.782 (0.17)	0.611 (0.11)	0.791 (0.14)	0.609 (0.1)	0.773 (0.19)
3	0.651 (0.05)	0.588 (0.08)	0.774 (0.05)	0.694 (0.04)	0.774 (0.04)	0.688 (0.06)	0.592 (0.08)	0.674 (0.11)	0.398 (0.05)	0.526 (0.07)	0.404 (0.07)	0.529 (0.08)
4	0.421 (0.1)	0.289 (0.12)	0.592 (0.07)	0.407 (0.09)	0.583 (0.07)	0.389 (0.08)	0.852 (0.14)	0.954 (0.19)	0.651 (0.1)	0.848 (0.2)	0.667 (0.11)	0.853 (0.15)
5	0.559 (0.07)	0.397 (0.09)	0.667 (0.07)	0.462 (0.11)	0.667 (0.07)	0.466 (0.08)	0.692 (0.16)	0.847 (0.21)	0.556 (0.12)	0.79 (0.17)	0.561 (0.12)	0.786 (0.15)

Table 4 - Total time (in seconds) to execution of BGGE and BGLR function (mean of 50 partitions). For the multi-environment models, main genotypic effect (MM), single variance G×E deviation model (MDs) and environment-specific variance G×E deviation model (MDe) with two kernels, GBLUP (GB) and Gaussian (GK) (standard deviation in parentheses).

		HEL	
Model	Kernel	BGGE	BGLR
MM	GK	155.42 (7.81)	316.15 (56.57)
	GB	65.46 (5.69)	136.47 (36.06)
MDs	GK	290.47 (23.2)	1253.7 (182.47)
	GB	118.69 (9.54)	502.05 (113.54)
MDe	GK	318.88 (27.13)	1694.62 (332.46)
	GB	152.85 (11.95)	908.53 (221.58)