Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v41i1.45361

GenomicLand: Software for genome-wide association studies and genomic prediction

Camila Ferreira Azevedo[1*], Moysés Nascimento[1], Vitor Cunha Fontes[2], Fabyano Fonseca e Silva[3], Marcos Deon Vilela de Resende[4] and Cosme Damião Cruz[5]

[1]Departamento de Estatística, Laboratório de Inteligência Computacional, Universidade Federal de Viçosa, Av. P H Rolfs, s/n, 36570-900, Campus Universitário, Viçosa, Minas Gerais, Brazil. [2]Departamento de Engenharia Agrícola, Grupo de Pesquisa em Interação Atmosfera-Biosfera, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. [3]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. [4]Embrapa Florestas/Departamento de Engenharia Florestal, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. [5]Departamento de Biologia Geral, Laboratório de Bioinformática, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. *Author for correspondence. E-mail: camila.azevedo@ufv.br

ABSTRACT. GenomicLand is free software intended for prediction and genomic association studies based on the R software. This computational tool has an intuitive interface and supports large genomic databases, without requiring the user to use the command line. GenomicLand is available in English, can be downloaded from the Internet (https://licaeufv.wordpress.com/), and requires the Windows or Linux operating system. The software includes statistical procedures based on mixed models, Bayesian inference, dimensionality reduction and artificial intelligence. Examples of data files that can be processed by GenomicLand are available. The examples are useful to learn about the operation of the modules and statistical procedures.

Keywords: statistical analysis; genomic analysis; molecular markers; biometrics.

Received on November 14, 2018.
Accepted on November 28, 2018.

Introduction

Genome-wide selection (GWS) and genome-wide association studies (GWAS) consist of analyzing a large number of single nucleotide polymorphism (SNP) markers widely distributed in the genome, capturing quantitative trait loci (QTLs) that affect a quantitative trait. The use of GWS in breeding programs allows for an increase in the efficiency in predicting genetic values (Wellmann & Bennewitz, 2012; Goddard, 2012; VanRaden, 2008, Meuwissen, Hayes, & Goddard, 2001), speed in the identification of genetically superior individuals (Vandenplas, Calus, & Gorjanc, 2018; Meuwissen et al., 2001; VanRaden, 2008), increase in the rate of genetic gain and reduction of the generation interval (Polejaeva et al., 2013; Resende et al., 2012a). The GWAS allows for identifying possible associations between genetic regions and traits. Both these analyses have had an important impact on breeding programs due to the study of the molecular controls of complex traits and biological processes.

The study of statistical methodologies and computational tools applied to genomic prediction and association studies is considered an important line of research. Therefore, several different approaches are available in the literature, for example, RR-BLUP, G-BLUP, Bayesian alphabet, partial least squares (PLS), and kernel-based regressions, among others. These methodologies, focusing on solutions for multicollinearity, dimensionality and nonnormal trait distributions, are implemented in different software, which, in general, require knowledge of programming languages, making them difficult for lay users. Additionally, some of these require a license for their use and have complicated interfaces.

In this context, developing a free and intuitive software will provide practitioners (academic or not) easy access to this high-level technology and methodology. For breeding programs interested in using the benefits of genomics, this development will lead to advances in the area of computational genomics.

For this purpose, GenomicLand was developed as a computational tool for genomic prediction and association studies based on the free R software, which has an intuitive interface and supports large genomic databases. The software is freely available to the scientific community at https://licaeufv.wordpress.com/.

Description

The software GenomicLand can be used in the Windows or Linux operating system. Some configuration settings are indispensable, such as a screen resolution of 1024 x 768 (large fonts) and the use of a decimal symbol expressed by points. The software is available in English. Python was the chosen language for the interface due to its free, easy operation and versatility.

Integration with R software

R is a language and an integrated development environment for statistical procedures and graphing. The R software has been increasingly accepted by universities and companies around the world. Currently, the acquisition costs of statistical software that are similar to R or even poorer in terms of analytical capacity are very high, especially for the predominantly small and medium businesses in our country. The development of GenomicLand under the R interface made it possible to use the main packages of this software without the use of command lines by the user. The user can use all statistical procedures necessary for a complete genomic approach without the need to use other software. In addition, the scheduling of computational routines is prepared for the dimensionality of the genomic data of the plant area and is optimized in order to reduce computational time and effort.

Application of the program

An application of the software GenomicLand usually includes the following steps.

a. *Examples of data files*: examples of data files that can be processed by GenomicLand are available. The examples are useful to learn about the operation of the modules and statistical procedures. Each procedure is represented by an icon that accesses the file containing an illustrative example of a particular procedure, with the files having the advantage of a complete description of all the parameters for immediate data analysis.

b. *Supplying data for processing*: statistical procedures usually have a common sequence of steps. Essentially, the user provides the folder path and selects the data file to be processed. The software prints and saves the results in the same folder. After these steps, the software requires the definition of the parameters (number of variables, number of fixed effects, number of random effects, order of the variable analyzed, etc.). It is recommended that these data files have headers. Regardless of the file format (i.e., .txt or .csv), the results will be saved in .csv format.

c. *Help with the parameters of statistical procedures*: a file containing a summary and the description of the required parameters to perform the procedure.

d. *Parameter Description*: for each procedure, the user must provide specific information about the data file that will be used for the processing. For example, to perform the genomic BLUP method, the user must provide the number of variables contained in the data file, the order of the variables to be analyzed, the number of subsamples defined to assess the accuracy and predictability of the model ($k - fold$), and the number of fixed and random effects that should be included in the template. In addition, the user must provide information about which genetic effects will be included in the model: additive, due to dominance and epistatic. The control buttons are common to all the available procedures. The button functions are described below:

Run: after the inclusion of the values of the desired parameters for each analysis, the user can perform the analyses through this button.

Stop: the user can abort the analysis if necessary.

Clear: the user can clear the description of the parameters if necessary.

Help: a file containing a summary of the statistical procedure used and a description of the parameters necessary to perform the analysis.

Example: for each procedure, the user will have an example file available, which is preloaded as part of the interface.

Print: the first six rows of the data file loaded in the software will be printed on the screen. An error in reading the data would definitely lead to errors in the data processing; thus, to ensure the correct reading of the data and analysis, the user must apply the necessary corrections, according to the specifications of each procedure.

e. *Result output:* the main results are printed on the software screen. However, all available results are saved in the .csv format in the folder chosen by the user. The results can also be exported to Excel, Libreoffice or other related programs.

Modules

The GenomicLand software system contains analysis modules that involve several procedures to perform genomic prediction analysis and association studies. These procedures are described below.

Initial analysis

This module contains initial statistical analyses. The initial procedures serve to verify the consistency and accuracy of the data, describing and exploring the study sample, and preparing the data for further analysis. It is crucial that this is done before undertaking complex analyses. The procedures are described below:

a. Convert SNP genotype data: the procedure converts a file with nitrogenous base pairs into 0, 1, and 2. The procedure consists of checking which nitrogenous base is most frequent, assigning it the weight 1, whereas the less frequent one receives a weight of 0.

b. Phenotype correction: procedures for the correction of phenotypes for fixed and random environmental effects (factors) and for the correction of population structure by principal components or by eigenvectors, as reported by Azevedo et al. (2017).

c. Quality control: the quality control of the marker files consists of the elimination of nonpolymorphic loci and/or loci with a low call rate and/or loci that are not frequent according to the Hardy-Weinberg equilibrium.

d. Relationship matrix (G): this matrix calculates the genomic relationship matrix (G) according to VanRaden (2008).

e. Heatmap of G: heatmaps are used to visualize the relationships among individuals through the genomic relationship matrix.

f. Principal Components of G: plots of the first and second principal components of the genomic relationship matrix.

Items e) and f) are used to study the genetic diversity among the individuals.

Genomic prediction methods

Based on Mixed Model

The linear mixed model (LMM) has been widely used in genetics and is an extension of the linear regression model, in which the variables are divided into two groups: fixed effects and random effects. The equations of mixed models (EMM) proposed by Henderson (1975) are used in the estimation of the best linear unbiased estimator (BLUE) for fixed effects and its functions of interest, as well as in the estimation of the best linear unbiased prediction (BLUP) for random effects. This method is done by using the covariance matrices of the random effects of the model and the estimates of the variance components obtained through the restricted maximum likelihood (REML) method. The statistical methods available in this module are described below:

a. G-BLUP: in genomic BLUP (genomic best linear unbiased predictor – VanRaden, 2008), the main random effects of the model are the genetic values of the individuals and their covariance matrix is given by the genomic relationship matrix between the individuals. This model can be considered a model with the inclusion of additive genetic effects, dominant and epistatic effects (additive × additive). The genomic relationship matrix associated with the additive effects and due to dominance are obtained as per Vitezica, Varona and Legarra (2013), and that associated with the additive epistatic effects is obtained as per Su, Christensen, Ostersen, Henryon and Lund (2012). The G-BLUP considers the homogeneous shrinkage for all markers' effects.

b. G-BLUP heterogeneous: this model is a mixed model, based on the linear model, for which the genomic relationship matrix considers the heterogeneous shrinkage for the markers' effects (Resende, Silva, Lopes & Azevedo, 2012b). The variance of each marker is estimated via the Bayesian least absolute shrinkage and selection operator (BLASSO) method.

Based on dimensionality reduction

Dimensionality reduction methods consist of constructing linear combinations of the explanatory variables, called components, with the purpose of reducing the dimensionality of the studied problem. In the context of genomic prediction, the explanatory variables are molecular markers. These methodologies guarantee the absence of multicollinearity between the components and a solution for the problems of high dimensionality. In this module, only the inclusion of additive effects is possible, as described by Azevedo, Resende, Silva, Lopes, and Guimarães (2013), Azevedo et al. (2014) and Azevedo et al. (2015a). The statistical methods and procedures available in this module are described below:

a. Number of components: this procedure assists in the choice of the number of components inserted in the model using principal component regression (PCR) and partial least squares (PLS). Three kinds of information are available to assist in the choice: percentage of explanation of the variance of X (molecular markers), percentage of explanation of the variance of Y (phenotype) and the number of components associated with a greater predictive capacity in the prediction of genomic values.

b. Principal Component Regression: in PCR, the components are actual orthogonal linear combinations that maximize the total variance.

c. Partial Least Squares: in PLS, the components are real orthogonal linear combinations that maximize the covariance between the components and the variable Y (phenotype).

d. Independent Component Regression: in ICR, the components are linear combinations that maximize the independence between them.

The prediction of the genomic values of the individuals is made based on a specific number of components determined by the user. The number of components must be less than or equal to $\min\left(\frac{N}{k-fold}, n\right) - 1$, where: N is the number of individuals contained in the database and n is the number of markers.

Based on Bayesian inference

Bayesian inference treats the vector of unknown parameters of the model as random quantities, and any initial information about them can be represented by means of probabilistic models. Thus, probability distributions, called *prior* distributions, are assumed for all unknown quantities. Bayesian methods are associated with systems of nonlinear equations and nonlinear predictions, and consequently, they may be better when the effects of QTLs are not normally distributed due to the presence of large-effects genes. In this module, it is possible to include additive effects and those due to dominance as per Azevedo et al. (2015b). The difference between the Bayesian approach regression methods applied to GWS is mainly due to the assumed *prior* distribution for the effects of the markers.

a. Bayesian Ridge Regression: as *prior* distribution for the effects of markers (additives and due to dominance), this regression assumes a normal distribution with a common variance term that leads to a homogeneous shrinkage through the effects of the markers.

b. BayesA: as *prior* distribution for the effects of markers (additives and due to dominance), this regression assumes a t distribution and a specific variance for each marker (Meuwissen et al., 2001).

c. BayesB: for a fraction π of the markers, this regression assumes the same *prior* distribution as BayesA and that the fraction 1 - π of the markers has no effect, π being adopted subjectively by the user (Meuwissen et al., 2001).

d. BayesCπ: as *prior* distribution for the effects of markers (additives and due to dominance), this regression assumes a normal distribution with a common variance for a fraction π of the markers and that the fraction 1 - π of the markers has no effect. It does so more effectively than BayesB (Gianola, de Los Campos, Hill, Manfredi, & Fernando, 2009). The π fraction is estimated by means of a Beta probability distribution.

e. BLASSO: as *prior* distribution for the effects of markers (additives and due to dominance), this method assumes a double exponential distribution and a specific variance for each marker (de los Campos et al., 2009).

In these methodologies, the user must provide the number of iterations for the Markov chain Monte Carlo (MCMC) algorithms and the number of initial iterations that will be discarded in the Markov chain;

thus, the effect of the initial values on the later inference is minimized (burn-in). The user must also specify the thinning that corresponds to the interval at which iterations are recorded.

Based on alternative methods

In this module, methods that can be used for genomic prediction, which do not fit into the previous modules, are described.

a. Least Absolute Shrinkage and Selection Operator: the LASSO regression proposed by Tibshirani (1996) combines the selection of covariates and the regularization by shrinkage of the regression coefficients. LASSO assumes the effects of markers as fixed effects; thus, the effects of markers that move away from 0 suffer a penalty. The LASSO solution allows for up to N - 1 coefficients different from zero, where N is the number of individuals. The shrinkage is dictated by the λ penalty parameter.

b. Lambda λ: this function helps the user define the best penalty parameter for the dataset. It calculates the *k - fold* cross-validated mean squared prediction error for the LASSO.

c. Machine learning: models based on regression trees and their improvements, such as bootstrap aggregation (Bagging) and Random Forest. A value is assigned for each region formed in the tree, which is used to predict the value of the variable response of a new individual. This value is the average of all individuals belonging to the region used in the construction of the respective tree. For a genomic prediction application of these methodologies, see González-Camacho et al. (2018).

Association studies

Due to the study of the molecular controls of the biological processes of complex traits, the study of the genome-wide association studies between QTLs and the genetic values of individuals is of extreme interest to breeding programs. However, in practice, the study of these associations is performed between molecular markers and phenotypes, and this is possible by means of linkage disequilibrium (LD) between the marker and the QTLs that control the trait of interest. The methods available for association studies are as follows:

a. Single-marker models: the traditional GWAS approach, by fitting one marker at a time in the phenotype and a hypothesis test, it is possible to detect the significance of this effect. In this function, it is possible to reduce the false positive rate by including the vector of polygenic effects in the model along with the genomic kinship matrix (Hayes et al., 2007; Macleod et al., 2010) or by principal components (Zhang et al., 2010; Azevedo et al., 2017).

b. Quantile regression: a quantile regression-based GWAS; different to the traditional GWAS approach, it allows for the fitting of models to all portions of a probability distribution of the trait. In other words, QR enables measuring the impact of an SNP on specific quantiles of the trait. For an application in association studies of this methodology, see Barroso et al. (2017) and Nascimento et al. (2018).

c. Manhattan: GWAS Manhattan plots, where the genomic coordinates are displayed along the X-axis, with the negative logarithm of the association P-value for each SNP displayed on the Y-axis, such that each dot on the Manhattan plot signifies an SNP.

Conclusion

GenomicLand is a free and intuitive tool for the analysis and processing of molecular datasets. This software includes some of the main models used in genome-wide association studies and genomic prediction. The software responds to the growing demand of users (academic or not) with different academic backgrounds, some of which lack knowledge of programming languages.

Acknowledgements

The authors would like to thank FAPEMIG, CAPES, and CNPq for the financial support.

References

Azevedo, C. F., Resende, M. D. V., Silva, F. F., Lopes, P. S., & Guimarães, S. E. F. (2013). Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. *Pesquisa Agropecuária Brasileira*, *48*(6), 619-626. DOI: 10.1590/S0100-204X2013000600007

Azevedo, C. F., Silva, F. F., Resende, M. D. V., Lopes, M. S., Duijvesteijn, N., Guimarães, S.E.F., … Knol, E. F. (2014). Supervised independent component analysis as an alternative method for genomic selection in pigs. *Journal of Animal Breeding and Genetics*, *131*(6), 452-461. DOI: 10.1111/jbg.12104

Azevedo, C. F., Nascimento, M., Silva, F. F., Resende, M. D. V., Lopes, P. S., Guimarães, S. E. F., & Glória, L. S. (2015a). Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. *Genetics and Molecular Research*, *14*(4), 12217-12227. DOI: 10.4238/2015.October.9.10

Azevedo, C. F., Resende, M. D. V., Silva, F. F., Viana, J. M. S., Valente, M. S. F., Resende Jr, M. F. R., & Muñoz, P. (2015b). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genetics*, *16*(105), 1-13. DOI: 10.1186/s12863-015-0264-2

Azevedo, C. F., Resende, M. D. V., Silva, F. F., Nascimento, M., Viana, J. M. S., & Valente, M. S. F. (2017). Population structure correction for genomic selection through eigenvector covariates. *Crop Breeding and Applied Biotechnology*, *17*(4), 350-358. DOI: 10.1590/1984-70332017v17n4a53

Barroso, L. M. A., Nascimento, M., Nascimento, A. C. C., Silva, F. F., Serão, N. V. L., Cruz, C. D., … Guimarães, S. E. F. (2017). Regularized quantile regression for SNP marker estimation of pig growth curves. *Journal of Animal Science and Biotechnology*, *8*(59), 1-9. DOI: 10.1186/s40104-017-0187-z

de Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., … Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375-385. DOI: 10.1534/genetics.109.101501

Gianola, D., de Los Campos, G., Hill, W. G., Manfredi, E., & Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, *183*(1), 347-363. DOI: 10.1534/genetics.109.103952

Goddard, M. E. (2012). Uses of genomics in livestock agriculture. *Animal Production Science*, *52*(3), 73-77. DOI: 10.1071/AN11180

González-Camacho, J. M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., & Crossa, J. (2018). Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance. *The Plant Genome*, *11*(2), 1-15. DOI: 10.3835/plantgenome2017.11.0104

Hayes, B. J., Chamberlain, A. J., Mcpartlan, H., Macleod, I., Sethuraman, L., & Goddard, M. E. (2007). Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. *Genetical Research*, *89*(4), 215-220. DOI: 10.1017/S0016672307008865

Henderson, C. R. (1975). Best linear estimation and prediction under a selection model. *Biometrics*, *31*(2), 423-447. DOI: 10.2307/2529430

Macleod, I. M., Hayes, B. J., Savin, K., Chamberlain, A. J., Cpartlan, H., & Goddard, M. E. (2010). Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. *Journal of Animal Breeding and Genetics*, *127*(2), 133-142. DOI: 10.1111/j.1439-0388.2009.00831.x

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829.

Nascimento, M., Nascimento, A. C. C., Silva, F. F., Barili, L. D., Vale, N. M., Carneiro, J. E., … Serão, N. V. L. (2018). Quantile regression for genome-wide association study of flowering time-related traits in common bean. *PLoS ONE*, *13*(1), e0190303. DOI: 10.1371/journal.pone.0190303

Polejaeva, I. A., Broek, D. M., Walker, S. C., Zhou, W., Walton, M., Benninghoff, A. D., & Faber, D. C. (2013). Longitudinal Study of Reproductive Performance of Female Cattle Produced by Somatic Cell Nuclear Transfer. *PLoS ONE*, *8*(12), e84283. DOI: 10.1371/journal.pone.0084283

Resende, M. D. V., Resende Jr, M. F. R., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., … Grattapaglia, D. (2012a). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist*, *194*(1), 116-128. DOI: 10.1111/j.1469-8137.2011.04038.x

Resende, M. D. V., Silva, F. F., Lopes, P. S., & Azevedo, C. F. (2012b). *Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial*. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. Retrieved on September 15, 2018 from http://www.ppestbio.ufv.br/?page_id=448.

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., & Lund, M. S. (2012). Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. *PLoS ONE*, *7*(9), e45293. DOI: 10.1371/journal.pone.0045293

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, *58*(1), 267-288.

Vandenplas, J., Calus, M. P. L., & Gorjanc, G. (2018). Genomic Prediction Using Individual-Level Data and Summary Statistics from Multiple Populations. *Genetics*, *210*(1), 53-69. DOI: 10.1534/genetics.118.301109

Vanraden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, *91*(11), 4414-4423. DOI: 10.3168/jds.2007-0980

Vitezica, Z. G., Varona, L., & Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, *195*(4), 1223-1230. DOI: 10.1534/genetics.113.155176

Wellmann, R., & Bennewitz, J. (2012). Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genetics Research*, *94*(1), 21-37. DOI: 10.1017/S0016672312000018

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., … Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355-360. DOI: 10.1038/ng.546