SOFTWARE RELEASE

Be-Breeder: an R/Shiny application for phenotypic data analyses in plant breeding

Filipe Inácio Matias[1*], Italo Granato[1] and Roberto Fritsche-Neto[1]

Crop Breeding and Applied Biotechnology
18: 241-243, 2018
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332018v18n2s36

Abstract: *In order to successfully achieve the final goal of genotype selection in plant breeding programs, many aspects must be considered and carefully thought regarding cost, time, and efficiency. Thus, we have developed the Be-Breeder application to perform main biometric and statistical analyses using mixed and multivariate models. Implemented using the Shiny R package, this is one the first online platforms proposed in this context. Be-Breeder is available at http://www.genetica.esalq.usp.br/alogamas/R.html.*

Key words*: Online tool, mixed models, selection index, genetics.*

INTRODUCTION

Plant breeding consists in the fundamental base for the provision of food and raw material to society, being a major driving force for science-based enhancements in crops (Bhat et al. 2016). Currently, many aspects are arising from advances in science, technology and more profound statistical studies that enable conventional breeding to be more accurate, rapid, and precise.

In order to successfully obtain genetic gains, it is essential to choose models and strategies for the species, the goals of the breeding program, the experimental designs, and the population structure. Once these aspects are correctly defined, it is possible to increase selective accuracy and thus more rapidly achieve the planned objective (Lynch and Walsh 1998). Therefore, development of computational tools that support breeders and researchers in the decision-making processes throughout a breeding program is of great interest.

In this study, we present Be-Breeder, a Shiny application for interactive analysis considering phenotypic data and mixed models under multiple experimental conditions. This tool provides a user-friendly interface to aid researchers and breeders in essential steps regarding the achievement of the final goal of a breeding program with a high degree of interconnectivity of academic plant research and plant breeding accurately.

DESCRIPTION

Phenotypic Breeding section of Be-Breeder was constructed using the Shiny R package (Chang et al. 2015) in R language, HTML, CSS, JavaScript, PHP, and C#. The scripts in R language (S4) were developed and adapted to be easily implemented in the Be-Breeder web-based application, where users can quickly obtain information on their breeding populations based on phenotypic data.

***Corresponding author:**
E-mail: filipematias23@usp.br

Received: 20 August 2016
Accepted: 15 March 2017

[1] Universidade de São Paulo, Escola Superior de Agricultura 'Luiz de Queiroz', Av. Pádua Dias, 11, 13.418-900, Piracicaba, São Paulo, Brazil

Be-Breeder is freely available at http://www.genetica.esalq.usp.br/alogamas/R.html.

The application has an intuitive approach. Regardless of the type of analysis, users can always input data in *.txt* or *.csv* format under different types of settings concerning heading, column separation, and quotes to correctly fit their structure. All the outputs can be exported in a *.txt* format according to the needs of the user. To do so, one simply types the desired name of the file and then click on the option Download. For each analysis, there is also a Help checkbox option with explanations on data entry and details regarding the statistical procedures.

FEATURES

Experimental design analysis

On the topic Statistical Model from Experimental Design Analysis tab, the dataset is analyzed using mixed models by the lme4 package in R (Bates et al. 2014). This package requires at least one random effect parameter to estimate fixed effects and predict the random effects. For illustration, a mixed linear model can be written as $y = X\beta + Z\upsilon + \varepsilon$, where y is a vector of phenotypic values, β is the fixed effects vector, υ is the random effects vector, and ε is the experimental error. X and Z are the incidence matrices for the vectors β and υ. Analyses are performed using mixed model equations of Restricted Maximum Likelihood / Best Linear Unbiased Predictor (REML/BLUP) method, in which the variance-covariance matrix of genotype effect is an identity matrix since they are considered as non-related (Resende 2002).

The user must define the effect of each factor in the model as fixed or random. Fixed effects should be declared using the name of their respective columns in the dataset. Random effects must be declared between parentheses, with "1" followed by the name of the effect, separated by a vertical bar. For example, considering an experiment with two factors, in which factor *A* is declared as fixed and factor *B* as random, the model should be typed as $y \sim A + (1|B)$. Interaction between factors are declared with their names separated by a colon, i.e. (1|A:*B*), and the nature of its effect should be typed as described for individual factors.

In this section, it is also possible to obtain outputs such as the overall phenotypic mean, estimation of fixed effects, prediction of random effects, test of statistical significance for both of them, and predicted means for each genotype and trait. These latter are displayed taking into account the Selection Intensity (scrollbar) set by the user, being possible to simulate different values to visualize the selected individuals and then choose the best fit for their conditions.

A distinctive feature of this section is the flexibility to define a statistical model by the experimental conditions from users. For more detailed descriptions and options, we recommend checking the Help option.

Selection index

In this tab from the Phenotypic Breeding section, it is possible to perform the selection of genotypes using additive selection index (SI), which combine several quantitative traits simultaneously. The general model for SI considering n traits is $SI = X_1b_1 + X_2b_2 + ... + X_nb_n$, where SI is the value of the index, X_i is the mean of trait i, and b_i is the weight attributed to each trait i. There is no restriction regarding the number of traits considered.

User's dataset must contain the genotype identification in the first column, named *Genotype*, followed by the information of traits, each one in a different column. The weights for each trait and their respective definition of favorable (+) or unfavorable (-) effect must be written appropriately in the Index Analysis window, according to the order the traits are mentioned in the data file. The value zero should be attributed as the weight to a trait for removing its effect from the index.

Here, we recommend the use of predicted means or BLUP values, which can be obtained from the topic Statistical Model from Experimental Design Analysis tab, as previously described. As output, genotypes are sorted by the index value, from highest to lowest. It is also possible to simulate different levels of selection intensity and verify their effect on the outcome.

Path analysis

Following the workflow, there is the Path Analysis tab, where one can carry out path analysis among traits (Wright

1923). This analysis is performed using the *agricolae* R package (Mendiburu 2016). Just as it occurs for Selection Index, input dataset must contain the genotype identification in the first column, named Genotype, followed by information of as many traits as desired, each one in a different column.

In this tab, phenotypic mean must be considered as trait information. In addition, it is necessary to indicate the main trait, so the estimates of direct and indirect effects between that and the other ones are supplied. Finally, the user can choose whether perform path analysis or traits correlation (Pearson's) in the option Choose Results.

Biplot analysis: G × E interaction

Evaluation of the genotype × environment interaction is contemplated in the tab Biplot Analysis, carried out using the functions *princomp* and *biplot* in R. In the entry dataset, genotypes should be arranged in rows and environments in columns.

Two possibilities of analysis are available. The former is GE Biplot, which performs the principal components analysis, releasing the summary, the scores for each component calculated either for genotypes (G) or for environments (E), and the biplot graph, where the first two principal components are considered, being the genotypes represented by points and the environments by arrows. The latter is GE Cluster, that executes cluster analysis in the same context (in here, the user may inform the desired number of clusters), issuing cluster information, the group genotypes belong to, mean of clusters, and a dendrogram built using the R function *hclust*.

CONCLUSION

Be-Breeder is a web-application available to the public that bridges a wide range of phenotypic data with a network of experimental analysis towards plant breeding programs. Integrated with several R packages, Be-Breeder does not require advanced mathematical or programming skills and provides a user-friendly interface with flexible query options allowing the exploration, visualization, and export of outputs.

ACKNOWLEDGMENTS

We thank the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) for financial support.

REFERENCES

Bates D, Mächler M, Bolker BM and Walker SC (2015) Fitting linear mixed-effects models using lme4. **Journal of Statistical Software 67**: (10.18637/jss.v067.i01).

Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A, Mushtaq M, Jain N, Singh PK, Singh GP and Prabhu KV (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. **Frontiers in Genetics 7**: (https://doi.org/10.3389/fgene.2016.00221).

Chang W, Cheng J, Allaire J, Xie Y and McPherson J (2015) Shiny: web application framework for R. **R package version 0.11 1**.

Griffing B (1956) Concept of general and specific combining ability in relation to diallel crossing systems. **Australian Journal of Biological Sciences 9**: 463-493.

Lynch M and Walsh B (1998) **Genetics and analysis of quantitative traits**. Sinauer Associates, Sunderland, 980p.

Mendiburu Fd (2016) Agricolae: statistical procedures for agricultural research. **R package version 1.2-4**: 1-6.

Resende MDV (2002) **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informação Tecnológica, Brasília, 975p.

Wright S (1923) The theory of path coefficients a reply to Niles's criticism. **Genetics 8**: 239-255.

Wright S (1951) The genetical structure of populations. **Annals of Eugenics 15**: 323-354.