CROP BREEDING AND APPLIED BIOTECHNOLOGY

NOTE

Be-Breeder – an application for analysis of genomic data in plant breeding

Crop Breeding and Applied Biotechnology
17: 54-58, 2017
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332017v17n1n8

Filipe Inácio Matias[1*], Italo Stefanine Correa Granato[1], Gabriel Dequigiovanni[1] and Roberto Fritsche-Neto[1]

Abstract: *Be-Breeder is an application directed toward genetic breeding of plants, developed through the Shiny package of the R software, which allows different phenotype and molecular (marker) analysis to be undertaken. The section for analysis of molecular data of the Be-Breeder application makes it possible to achieve quality control of genotyping data, to obtain genomic kinship matrices, and to analyze genomic selection, genome association, and genetic diversity in a simple manner on line. This application is available for use in a network through the site of the Allogamous Plant Breeding Laboratory of ESALQ-USP (http://www.genetica.esalq.usp.br/alogamas/R.html).*

Key words*: Quality control, kinship matrix, genomic selection, genetic diversity, SNP.*

INTRODUCTION

The application of molecular information in genetic breeding of plants can lead to savings in time and money and increase precision and intensity of selection for programs that use this information (Moose and Mumm 2008, Varshney et al. 2009). In addition, this information has allowed advances in studies in population genetics and species evolution through facilitating observation and comparison of existing genetic diversity measured by molecular markers (Govindaraj et al. 2015).

Due to the complexity of molecular markers, as well as the size of data sets generated, different statistical and computational approaches are necessary for researchers to extract all the information available for analysis (Cole et al. 2012). For that reason, software programs with different computational approaches have been developed to undertake molecular analysis in specific areas of genetics and plant breeding. In this context, the aim of this study was to develop an easily accessible application through scripts and packages of the R software, the molecular section of app Be-Breeder (Fritsche-Neto and Matias 2016) which simultaneously includes different molecular analysis as tools for hanalysisandling and transformation of datasets and matrices, methodologies of genomic selection and association, and mechanisms for studies of genetic diversity.

MATERIAL AND METHODS

The Be-Breeder application was constructed through the Shiny package of the R software (Chang et al. 2015) with a click point interface that allows the

***Corresponding author:**
E-mail: filipeinacio23@hotmail.com

Received: 09 September 2016
Accepted: 31 November 2016

[1] Universidade de São Paulo (USP), Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ), Departamento de Genética, Av. Pádua Dias, 11, CP 83, 13.418-900, Piracicaba, SP, Brazil

user to perform different analysis with molecular (marker) data. For that reason, scripts in R language were developed and implemented in the "molecular breeding" section by the team of the Allogamous Plant Breeding Laboratory of the Genetics Department of the Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ) of the Universidade de São Paulo (USP). This is available for use at http://www.genetica.esalq.usp.br/alogamas/R.html.

In each one of the tabs to be presented below, there are practical examples and explanatory tutorials in the application itself regarding data entry and details of analysis (*Help*).

Table 1. Example of data entry in the column format for the "*Quality Control*" section

Sample	Marker	Allele.1	Allele.2
A01	PHM4468-13	G	G
A01	PHM2770-19	G	G
A01	PHM523-21	<NA>	<NA>
A01	PZA00485-2	A	A
A01	PZA00522-7	A	A
A01	PZA00627.1	G	G
A01	PZA00473.5	G	G
A01	PHM5232-11	C	C
...	...	...	...
H12	PZA00516-3	G	G

Genotyping data

Modifying and constructing matrices that are in agreement with the presupposed statistics or even those that are suitable for entry in analytical software is one of the main challenges encountered by researchers. In this respect, this section of the application provides the user with the tabs "*Quality Control*" and "*Kinship Matrix*".

The first tab uses the "*raw.data*" function of the snpReady package of R (under construction). This allows the user to control the quality of the genomic data. Data can be uploaded as matrix or columns, codified in nitrogen bases, as a function of the structure of the initial dataset received from genotyping (Table 1).

The present parameters of quality control are:

- MAF (Minor Allele Frequency): consists of removal of markers with minor allele frequency, that is, lower than the threshold defined by the user.

- Call Rate: call rate parameter, which refers to the quality of genotyping in relation to lost data. Thus, markers with a Call Rate lower than the threshold defined by the user will be eliminated.

- Sweep sample: refers to the quality of genotyping in relation to the data lost in the individuals. Thus, individuals with a Sweep sample lower than the threshold defined by the user will be eliminated.

This tab also allows data imputation, which is carried out from the combined probability of the alleles of a determined SNP$_{(i)}$ (frequencies of p_i and q_i) and the level of homozygosity of the individual that has a marker to be imputed. The output generated by the function is a "clean" matrix, in which the output format can also be defined by the user, with the options of counting a reference allele for each locus (0,1,2), this matrix centered in zero (-1,0,1), or even the matrix in the appropriate format for entry in the Structure software. Thus, the resulting matrix is in the adequate format to be used in other software or packages or, moreover, to proceed with other analysis in the Be-Breeder application itself.

The "*Kinship Matrix*" tab uses the "*G.matrix*" function of the snpReady package and allows construction of different kinship matrices, as indicated by the formulas below:

- UAR: unified additive kinship matrix (Yang et al. 2010)

$$G_{UAR} = \frac{1}{N}\sum_i A_{ijk} = \begin{cases} \frac{1}{N}\Sigma_i \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-2p_i)}, j \neq k \\ 1+\frac{1}{N}\Sigma_i \frac{x_{ij}^2-(1+2p_i)x_{ij}+2p_i^2}{2p_i(1-p_i)}, j = k \end{cases}$$

- UARadj: adjusted unified additive kinship matrix (Yang et al. 2010)

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

- WW: kinship matrix of VanRaden (VanRaden 2008)

$$G_a = \frac{(W^{*}W^{*`})}{\sum_i^n [2p_i(1-p_i)]}$$

Thus, in this tab, in addition to the possibility of generating these three types of matrices, it is possible to change the format of the generated output. The formats available are the traditional kinship matrix (complete matrix) or an alternative format, in which the inverse of the kinship matrix is generated and then organized in columns for use in the package ASREML-R (Butler et al. 2009):

Genomic selection (GS)

Genomic selection is a potent tool in genetic plant breeding, the basis of which is the use of powerful statistical models to carry out prediction using thousands of markers. These models estimate the effects of molecular markers to predict characteristics of the individuals of this population without the need for direct phenotype observation (Meuwissen et al. 2001). This results in time and resources saved in the selection process (Heffner et al. 2010). For that reason, the statistical model must first be trained and validated and then use the effects of the markers as a predictive tool. The Be-Breeder application uses the RR-BLUP method, implemented in the rrBLUP package (Endelman 2011). Thus, it uses a mixed models approach (1), adopting the effects of the markers as random (2) through the REML (Restricted Maximum-Likelihood) approach.

$$y = X\beta + Zu + \varepsilon$$

$$BLUP(u) = \overset{*}{u} = \sigma_u^2\, KZ'V^{-1}(y - X\beta^{*})$$

In the "*Prediction and Selection*" tab, the user must enter with the marker matrix, in the format centered in zero (-1,0,1) to carry out analysis, in which the individuals are in the lines and the markers in the columns (Table 2). The phenotype dataset is also necessary (Table 3). In this tab, there is also the possibility of using the effects of the markers generated and the matrix of markers of non-phenotyped individuals to predict the genetic values of the individuals.

***Table 2**. Example of genotype data entry for the "Genomic Selection" (GS)* section

Genotype	M1	M2	M3	M4	M5	M6	...	Mm
G1	1	1	1	1	-1	-1	1	-1
G2	1	-1	1	-1	-1	1	-1	-1
G3	1	-1	-1	1	1	-1	-1	1
G4	1	-1	1	1	-1	1	1	1
...	1	1	-1	1	-1	-1	-1	1
Gn	1	1	-1	-1	-1	-1	1	1

***Table 3**. Example of phenotype data entry for the "Genomic Selection" (GS)* section

Genotype	Y
G1	-14.67
G2	-49.55
G3	29.87
G4	-3.14
...	...
Gn	-3.24

Genome association (GWAS)

Determination of regions in the gene related to a trait (QTL) through molecular markers is an initial step and of extreme importance for understanding the regulation and genetic structure of a phenotype of interest (Korte and Farlow 2013). This information allows the user to direct other bioinformation approaches, such as verification of genes associated with this QTL, to try to correlate the different transcripts with the phenotypes (Kamatani 2016). To do so, an important tool in breeding is genome association (GWAS), which uses mixed models associated with multiple regression, and by means of a threshold (LOD), determined by the user, which allows molecular markers to be found that exhibit greater correlation with the phenotypic variability of the trait under study. For that reason, in the "*GWAS*" section of the Be-Breeder, the user supplies the phenotype (Table 3) and genome (Table 4) data. From these data and through the "*GWAS*" function of the rrBLUP package (Endelman 2011), analysis of association

***Table 4**. Example of genotype data entry for the "Genomic Association" (GWAS)* section

Marker	Chrom	Pos	G1	G2	G3	G4	G5	...	G*n*
1	1	1	-1	-1	-1	-1	1	1	-1
2	1	2	-1	-1	1	1	-1	1	1
3	1	3	1	1	1	-1	-1	1	1
4	1	4	-1	1	-1	1	1	1	1
5	1	5	1	-1	-1	1	-1	-1	-1
6	1	6	1	-1	1	-1	-1	1	-1
7	1	7	-1	-1	1	1	-1	-1	1
m	1	8	1	1	-1	-1	1	-1	-1

between the markers and the QTL for the trait under study is carried out, considering the following model:

$$y = X\beta + Zg + S\tau + \varepsilon$$

in which y is the phenotype vector, β is the fixed effects vector, g is the genetic effects vector given as random, τ is the additive effects vector of the SNP being considered as fixed, and ε is the residue vector. X, Z, and S are incidence matrices of the model.

In addition to the estimate of the marker effect, the application creates a Manhattan plot graph with all the chromosomes, markers, and their respective effects.

Analysis of genetic diversity

To study genetic diversity among individuals and populations, molecular markers are quite useful and highly used currently (Govindaraj et al. 2015). In this respect, Be-Breeder provides the user with two options.

In the first tab, "*Genetic Diversity*", the user can enter files in text format (.txt) generated by the *Genpop* (Raymond and Rousset 1995) (Table 5) or *Structure* (Pritchard et al. 2000) software (Table 6), as indicated in the *Help* section and in the example files. These analysis were constructed through scripts that associate functions of the "ape" (Paradis et al. 2004), "poppr" (Kamvar et al. 2014), and "adegenet" (Jombart 2008) packages of the R software. As output, the user has a general summary with information on the populations and markers. From these estimates, observed heterozygosity (*Ho*) is obtained through "*Ho* = number of heterozygotes in each locus / number of individuals" and expected heterozygosity (*He*) by $He = 1 - \sum_{i=1}^{\neq alleles} p_i^2$, for in reference to the frequency of the *i*-th allele (Nei 1978). The F of Wright, for its part, is estimated by (Wright 1965). Graphic outputs that involve multivariate principal component analysis, population structure, and dendrograms are obtained based on the Nei distance (Nei 1972).

The second tab "*Population genetics*" uses the "popgen" function of the snpReady package and allows information to be obtained in reference to allele frequency per marker (p and q), minor allele frequency (MAF), expected heterozygosity (He) estimated through the formula $He = 2*p*q$, observed heterozygosity (Ho), genetic diversity (DG), obtained by $DG = -1 -p^2 -q^2$, and polymorphic information content (PIC) estimated by $PIC = -1 -(p^2 + q^2) - (2*p^2*q^2)$ (Hartl and Clark 2010). The function furthermore allows an argument for attributions of individuals in subpopulations, if there are any and they are known *a priori*. To carry out analysis, the user must provide the marker matrix parametrized for allele count (0,1,2) (Table 7) and, as an option, a vector with the information of subgroups within the population (subgroups) (Table 8). Thus, it will be possible to identify each one of these parameters for each subpopulation, as well as the existence of exclusive,

Table 5. Example of the *.txt* file of genome data generated by the *Genpop* software to be used in the "*Genetic Diversity*" section

Stacks version 1.29; Genepop version 4.1.3						
SSR	61_82	106_11	135_52	255_50	255_51	385_23
pop						
Bo1	101	102	303	103	103	404
Bo10	101	101	303	303	101	404
Bo11	101	101	303	303	101	404
pop						
Bo2	101	101	303	303	101	404
Bo20	101	101	303	303	101	404
Bo21	101	202	303	303	101	304

Table 6. Example of the *.txt* file of genomic data generated by the *Structure* software to be used in the "*Genetic Diversity*" tab

Genotype	Population	M1	M2	M3	M4	M5	M6	M7	M8
I_1	Pop1	249	253	232	236	249	249	-9	-9
I_2	Pop1	249	253	232	236	249	249	241	241
I_3	Pop1	249	253	236	236	249	249	241	241
I_4	Pop1	249	253	236	236	249	249	241	241
I_5	Pop1	249	253	236	236	249	249	241	241
I_6	Pop1	249	253	232	244	249	263	225	225
I_7	Pop2	249	249	244	244	249	249	241	241
I_8	Pop2	249	249	244	244	249	249	213	213
I_9	Pop2	247	247	232	232	263	263	245	245
I_10	Pop2	247	247	232	232	249	249	213	213
I_11	Pop2	245	245	232	232	249	249	245	245
I_12	Pop2	249	249	232	232	249	249	213	213

Table 7. Example of genotype data entry for the "*Population Genetics*" section

Genotype	PHM4468.13	PHM2770.19	PZA00485.2	...	PZA00473.5
30A37PW	2	2	1	2	1
DKB.340.PRO	2	2	1	2	1
2B688PW	1	2	2	1	1
BM820	2	2	2	2	0
Truck.TL	2	1	2	1	1
2B587PW	2	2	1	1	2
2B710PW	1	2	2	2	2
DKB.310.PRO	2	2	2	2	1

absent, and fixed alleles in the subpopulations. If this last piece of information is not supplied, the application will consider only a single population.

RESULTS AND DISCUSSION

The Be-Breeder application allows the use of important tools of molecular breeding (GS and GWAS), as well as studies related to genetic diversity, in an easy way on line. It is expected that the practicality of the application allows the use of more complex analysis, superseding the need for mastery of R software programming. This application will assist researchers in providing access to quality results in a dynamic and interactive manner, integrating genome data to the routine of plant breeding programs.

***Table 8**.* Example of information entry of subgroups within the population, for use in the "*Population Genetics*" section

Genotype	Subgroup
30A37PW	1
DKB.340.PRO	2
2B688PW	1
BM820	3
Truck.TL	4
2B587PW	1
2B710PW	1
DKB.310.PRO	2

REFERENCES

Butler DG, Cullis BR, Gilmour AR and Gogel BJ (2009) **ASReml-R reference manual**. Department of Primary Industries and Fisheries, The State of Queensland, 149p.

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

Cole J, Newman S, Foertter F, Aguilar I and Coffey M (2012) Breeding and genetics symposium: Really big data: Processing and analysis of very large data sets. **Journal of Animal Science 90**: 723-733.

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. **The Plant Genome 4**: 250-255.

Fritsche-Neto R and Matias FI (2016) Be-Breeder-Learning: a new tool for teaching and learning plant breeding principles. **Crop Breeding and Applied Biotechnology 16**: 240-245.

Govindaraj M, Vetriventhan M and Srinivasan M (2015) Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. **Genetics Research International 2015**: 1-14.

Hartl DL and Clark A (2010) **Princípios de genética de populações**. Artmed, Porto Alegre, 660p.

Heffner EL, Lorenz AJ, Jannink J-L and Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. **Crop Science 50**: 1681.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. **Bioinformatics 24**: 1403-1405.

Kamatani Y (2016) Genome wide association study: its theory and methodological review. **Clinical Calcium 26**: 525.

Kamvar ZN, Tabima JF and Grünwald NJ (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. **PeerJ 2**: 281.

Korte A and Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. **Plant Methods 9**: 1.

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

Moose SP and Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. **Plant Physiology 147**: 969-977.

Nei M (1972) Genetic distance between populations. **American Naturalist**: 283-292.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. **Genetics 89**: 583-590.

Paradis E, Claude J and Strimmer K (2004) APE: analysis of phylogenetics and evolution in R language. **Bioinformatics 20**: 289-290.

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. **Genetics 155**: 945-959.

Raymond M and Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. **Journal of Heredity 86**: 248-249.

VanRaden P (2008) Efficient methods to compute genomic predictions. **Journal of Dairy Science 91**: 4414-4423.

Varshney RK, Nayak SN, May GD and Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. **Trends in Biotechnology 27**: 522-530.

Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. **Evolution**: 395-420.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG and Montgomery GW (2010) Common SNPs explain a large proportion of the heritability for human height. **Nature Genetics 42**: 565-569.