International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 7, Issue 5 [May 2018] PP: 10-13

Exploring GAPIT Package Features with Simulated Genomic Data

Michele Barbosa[1], Leila Maria Ferreira[2], Fernando Ribeiro Cassiano[3], Laís Mesquita Silva[4], Thelma Sáfadi[5]

[1]*PhD student in Agricultural Statistics and Experimentation, Federal University of Lavras, Brazil; Theater of the Federal University of Alfenas, Campus Varginha, Brazil,*
[2]*PhD student in Agricultural Statistics and Experimentation, Federal University of Lavras, Brazil,*
[3]*PhD student in Agricultural Statistics and Experimentation, Federal University of Lavras, Brazil,*
[4]*PhD student in Agricultural Statistics and Experimentation, Federal University of Lavras, Brazil,*
[5]*Teacher of the Department of Agricultural Statistics and Experimentation, Federal University of Lavras, Brazil*
Corresponding Author: Michele Barbosa

Abstract:*The agronomic information process in the genetics prospect currently presents voluminous data, which necessitates the continuous development of innovative and quality statistical methods that meet the Big Data era in which much is being reestablished in the form of generating, analyzing and absorbing data. The R language is an efficient tool for these types of analyzes and allows some structures and functionalities, such as the GAPIT package, which conducts genomic association (GWAS) and genomic prediction (GS). For this reason, this work aims at the study and exploration of characteristics of the GAPIT package, around a simulated data set. We demonstrate some output files that include graphical resources and tables ready for publication of results, identifying the GAPIT package as a viable alternative when working with large volumes of genotype data and obtaining quick and efficient responses.*

Keywords:*Genetics, GAPIT package, Genomic association*

Date of Submission: 05-05-2018 Date of acceptance: 21-05-2018

I INTRODUCTION

Nowadays genetic information from agronomic point of view has the need to perform the processing and analysis of large amount of data. For this reason, the use of computational tools and adequate statistical methods becomes indispensable for the accomplishment of these activities.

In the R language, numerous packages have been developed that allow the use of data and present high performance, packages aimed at the most diverse genetic treatments, such as genetics of populations that implement methods represented by genotypes and haplotype data with functions to estimate and Hardy-Weinberg tests and linkage disequilibrium (LD). LDheatmap package creates and plots a heat map for pairwise LD. Packages such as adegenet, hierfstat, gap for population analysis. Details about these packages can be obtained in Jombart (2008), Goudet (2005) and Zhao (2007). In addition to these we can mention the packageGenABELwhichis a package developed for manipulation of GWAS data. A current list of these packages can be found on the CRAN page https://cran.r-project.org/, which is the page of the entity that makes available the R and its official packages.

One tool that presents statistical methods for association studies and prediction genomic data is the GAPIT (Genome Association and Prediction Integrated Tool). The GAPIT package for R software, implements advanced statistical methods, manipulates large datasets with more than 10,000 individuals and 1 million single nucleotide polymorphisms (SNPs) (Lipka et al., 2012). An alternative to obtain better performance with large data analyzes is the use of functions of this package.

In addition, the GAPIT package is an alternative to writing quick functions, with simple data structure, where you can interpret, analyze, access easily and discover information about the data. The visualization approach is enhanced by methods implemented in GAPIT (produces comprehensive charts, statistical tests, quality tables and a variety of outputs) and becomes important to harness the power of statistical evaluations and accuracy of predictions and genomic results.

However, we highlight some structures and functionalities of the GAPITpackage for the software R, which conducts genomic association (GWAS) and genomic prediction (GS) studies. We consider a simulated data set and present how GAPIT receives a large volume of genotypic data,evidencing the graphical features ready for publication of results.

II MATERIAL AND METHODS

To obtain the genotypic data that were used for analysis in the GAPIT, three diallelpopulations of the GENES software were simulated (CRUZ, 2013).

A matrix$\boldsymbol{X} = \left[x_{ij}\right]_{n\times k}$ generated with the simulated data consisted of n=900 individuals (300 individuals by population) containing$k = 6000$co-dominant markers SNPs distributed on 6 chromosomes (1000 marks per chromosome).

The elements of the matrix $\boldsymbol{X} = \left[x_{ij}\right]_{n\times k}$ follow a Binomial distribution with the genotypic frequencies: $p(22) = p^2$,$p(12) = 2$pq,$p(11) = q^2$, where $p(x = 2) =$ p and$q = 1 - p = p(x = 1)$, by the Hardy-Weinberg equilibrium (HW). The markers (SNPs) were coded as 0, 1 and 2, considering the genetic values a, d and -a, for alleles 11, 12 and 22 respectively. We assume

$$x_{ij} = \begin{cases} 0, with\ probability\ p^2 \\ 1, with\ probability\ 2pq \\ 2, with\ probability\ q^2, \end{cases}$$

thepopulation mean for a gene as $E\left(x_{ij}\right) = (p - q)a + 2pqd$

and the genetic variance given by gene as $var\left(x_{ij}\right) = \sigma_g^2 = 2pq\alpha^2 + (2pqd)^2$

where$\alpha$is the effect of allelic substitution, given by: $\alpha = [a - (p - q)d]$.

Using the software R (TEAM, 2016), a phenotypic value $\boldsymbol{y}_{n\times 1}$was initially simulated from a linear genetic model

$\boldsymbol{y} = (\boldsymbol{X} - 2p)\alpha$.

For the simulation of the phenotype an error of a Gaussian distribution was added with $\sigma_e^2 = \sigma_a^2\left(\frac{1-h^2}{h^2}\right)$, considering heritability$h^2 = 0.2$ to an oligogenic scenario, whereσ_a^2 and$\sigma_e^2$are components of the genetic and residual variance respectively.

When entering phenotype data, simulated genotype matrix, the results will be saved in the working directory in output files that are saved in formats (.csv) or (.pdf) include summaries of both GWAS and GS.

Figure 1 shows a program in which command lines use phenotypic (myY) and genotypic (myGD) data files. The myGM file contains information for SNPs, with identitycolumns, chromosome, and position.

```
#Pass (1)
myGD<- read.table("GD.txt" , head = TRUE)
myGM<- read.table("infsnp.txt", head = TRUE)  # information SNPs, chromoss, position
myY<- read.table("yo2.txt", head = TRUE)
```

Figure 1: Software R with information on phenotypic and genotypic data.

The second step shown in the Figure 2 is to execute the GAPIT function and the files with results.Graphics will be saved in your directory.

```
myGAPIT<- GAPIT(
  Y=myY,
  GD=myGD,
  GM=myGM,
group.from=nrow(myY),
  group.to=nrow(myY),
PCA.total=3,
Model.selection=TRUE)
```

Figure 2: Software R with GAPIT package functions.

The calculation of the kinship matrix can be done by methods like VanRaden (VanRaden et al., 2008). The GAPIT package also offers the option of calculating main components from the genotypic data.

Outputs are displayed as files, include summaries with results in GWAS and GS. The GWAS results are summarized by Manhattan graphs, Q-Q plot graphs and tables. Similarly, GS results are presented in a Heatmap and tables. Graphs of estimates of heritability, likelihood function,among others.

The analyzes were performed in the software R, version 3.3.2 (R Development Core Team 2016).

GAPIT is a package run by software R and can be downloaded from http://www.r-project.org or http://www.rstudio.com. The user can provide genotypic data, phenotypic, kinship matrices, population structure and store them in HapMap (format commonly used to store sequences of SNPs, chromosomes and position) or in a data.frame.

To facilitate the use of the GAPIT package, sample data, results, manual and source code are available at: http: //zzlab.net/GAPIT.

III RESULTS AND DISCUSSION

Figure 3 shows the dispersion diagram, histogram, box-plot graph and cumulative distribution, and can visualize and validate the phenotypic distribution, verifying the corresponding graphs. Attention can be given to the requirement of normality over residual effects. These graphs also help to detect atypical values, which are important sources of error and may need correction to avoid false positives.

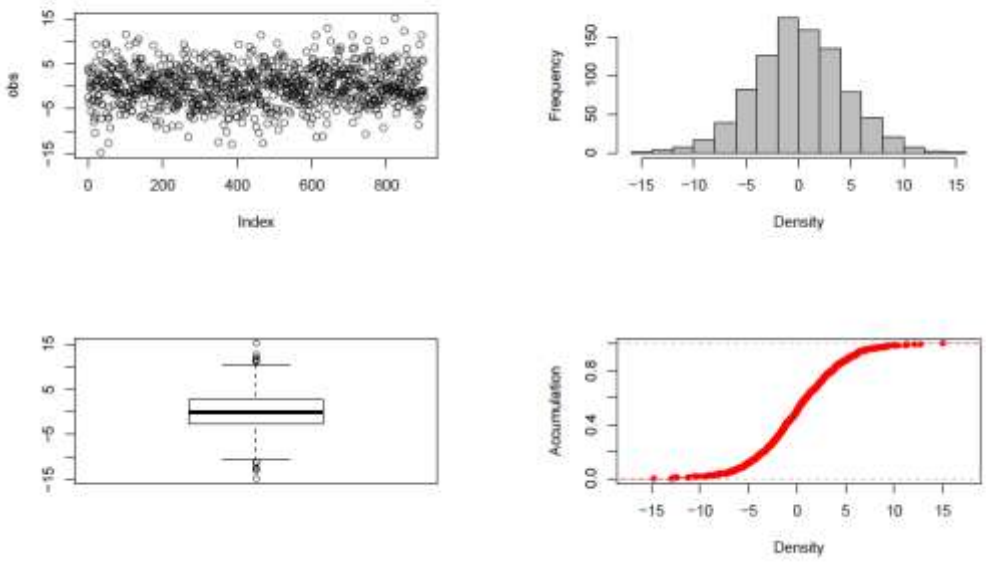


Figure 3: Scatter diagram, histogram, boxplot and cumulative distribution for phenotypic data.

Figure 4 shows main components (PCs) in two-dimensional and three-dimensional graphics. The kinship matrix is presented as Heatmap where red indicates the highest correlation between pairs of individuals and yellow indicates the lowest correlation. A tree of hierarchy among individuals is displayed based on their kinship.

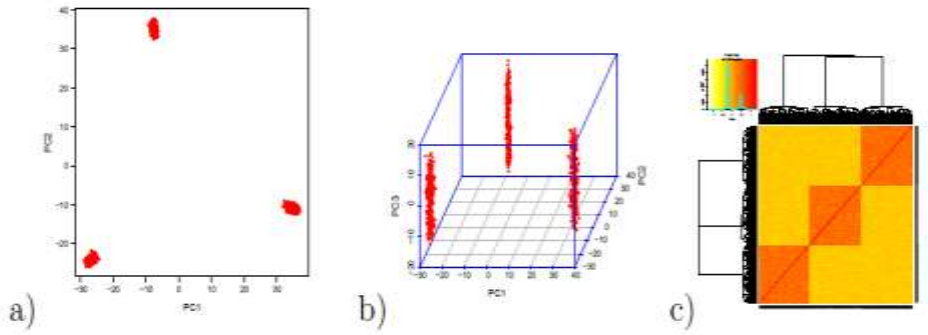
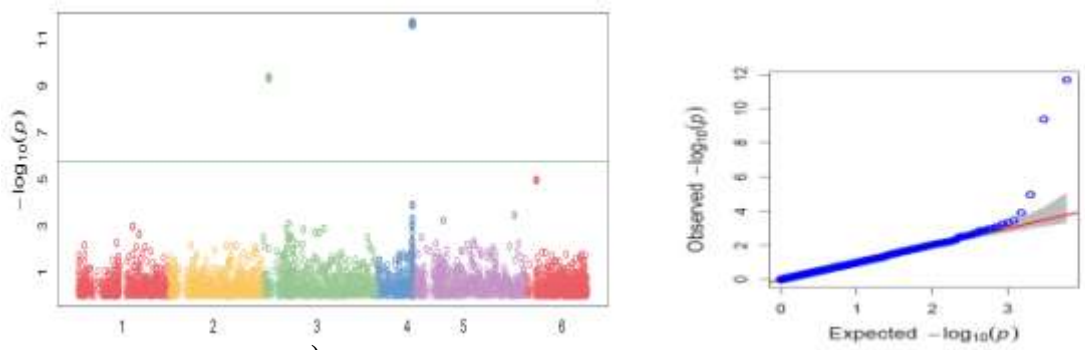


Figure 4: Population and Kinship Structures. Main components in two dimensions (a) and three dimensions (b). Kinship matrix displayed through the heat map (c).

The associations between phenotypes and genetic markers are displayed in text files (.csv), presented in the Table 1.

Table 1: Results in GWAS of all analyzed SNPs

SNP	Chromosome	Position	p-value	maf	nobs	R^2 without SNP	R^2 with SNP	FDR ajusted p-value
3963	4	24789723	2.12e-12	0.126	900	0.045746056	0.0996477	1.27e-08
2069	3	4386122	4.48e-10	0.475	900	0.045746056	0.0879567	1.35e-06
5044	6	80278699	1.13e-5	0.277	900	0.045746056	0.0664350	0.02267



a) b)

Figure 5: Manhattan Simulated data (a). Q-Q plot (b).

The Manhattan graph indicates the association between the markers by -log10 (p), the higher the height, the stronger the association (Figure 5 (a)). The Q-Q plot graphic (Figure 5 (b)) allows interpreting GWAS results, indicating the markers that are or are not associated with the phenotype. The red line indicates hope. The gray area is the 95% confidence interval.

IV CONCLUSION

It is observed that the processing and analysis of data in genetics is currently something essentialfor organizations that have a database with a large amount of information. The analysis runs using GAPIT are shown to be simplicate and it ensures that it is a solution for the analysis of bulky data in association and genomic prediction. In addition, enabling fast data analysis with comprehensive content and results.

ACKNOWLEDGEMENTS

The authors would like to thank the financial support and scholarships granted by FAPEMIG, CAPESandCNPq.

REFERENCES

[1]. Cruz, C. D. GENES - a software package for analysis in experimental statistics and quantitative genetics. Second edition. London: MIT press, 2010. 1064 p.
[2]. Goudet, J. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes, 2005, 5: 184-186.
[3]. Jombart, T. Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 2008,24: 1403-1405.
[4]. Lipka, A. E.; Tian F.; Wang Q.; Peiffer J.; Li M.; Bradbury P. J.; Gore M. A.; Buckler E. S.; Zhang Z. GAPIT: genome association and prediction integrated tool. Bioinformatics, 2012, 28: 2397-2399.
[5]. Zhao, J. H. Gap: Genetic Analysis Package. Journal of Statistical Software, 2007,23:1-18.
[6]. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[7]. VanRaden, P.M.;Van TassellC.P.;WiggansG.R.;SonstegardT.S.;SchnabelR.D.;Taylor J.F.;Schenkel F. S.. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci.,2009, 92:16-24.

Michele Barbosa." Exploring GAPIT Package Features with Simulated Genomic Data" International Journal of Engineering Inventions, vol. 07, no. 05, 2018, pp. 10–13.