Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1679-9275
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v35i3.21251

GENES - a software package for analysis in experimental statistics and quantitative genetics

Cosme Damião Cruz

Laboratório de Bioinformática, Universidade Federal de Viçosa, Av. P H Rolfs, s/n, 36570-000, Campus Universitário, Viçosa, Minas Gerais, Brazil. E-mail: cdcruz@ufv.br

ABSTRACT. GENES is a software package used for data analysis and processing with different biometric models and is essential in genetic studies applied to plant and animal breeding. It allows parameter estimation to analyze biological phenomena and is fundamental for the decision-making process and predictions of success and viability of selection strategies. The program can be downloaded from the Internet (http://www.ufv.br/dbg/genes/genes.htm or http://www.ufv.br/dbg/biodata.htm) and is available in Portuguese, English and Spanish. Specific literature (http://www.livraria.ufv.br/) and a set of sample files are also provided, making GENES easy to use. The software is integrated into the programs MS Word, MS Excel and Paint, ensuring simplicity and effectiveness in data import and export of results, figures and data. It is also compatible with the free software R and Matlab, through the supply of useful scripts available for complementary analyses in different areas, including genome wide selection, prediction of breeding values and use of neural networks in genetic improvement.

Keywords: software, statistical analysis, genetic analysis, quantitative genetics, biometry.

GENES - software para análise de dados em estatística experimental e em genética quantitativa

RESUMO. O programa GENES é um software destinado à análise e processamento de dados por meio de diferentes modelos biométricos. Seu uso é de grande importância em estudos genéticos aplicados ao melhoramento vegetal e animal, por permitir estimativa de parâmetros para entendimento de fenômenos biológicos e fundamentais em processo de tomada de decisão e na predição do sucesso e viabilidade da estratégia de seleção. O programa é obtido pela rede Internet (http://www.ufv.br/dbg/genes/genes.htm ou http://www.ufv.br/dbg/biodata.htm) e está disponível nos idiomas português, inglês e espanhol. Conta com literatura específica (http://www.livraria.ufv.br/) e um conjunto de arquivos de exemplos, tornando-o de fácil utilização. O GENES está integrado aos aplicativos MS Word, MS Excel e Paint permitindo importar dados e exportar resultados, dados e figuras de forma simples e eficiente. Está também integrado ao software livre R e ao aplicativo Matlab, por meio da disponibilização de scripts úteis para análises complementares em áreas diversas, incluindo seleção genômica ampla, predição de valores genéticos e uso de redes neurais no melhoramento genético.

Palavras-chave: software, análise estatística, análise genética, genética quantitativa, biometria.

Introduction

To breed genetically superior plants, the selected individuals must simultaneously unite a series of properties to produce a comparatively higher yield and to meet consumer demands. A way to increase the chances of success of a breeding program is to perform reliable experiments, generating a great volume of experimental data. Based on an adequate processing of these data, genetic parameters can be estimated and biological phenomena interpreted. In this phase of result analysis and interpretation, appropriate software systems and computer resources are of utmost importance.

The development of software in the field of plant Genetics and Breeding is crucial due to the scarcity of such resources available to the scientific community. The availability of such tools would supply the increasing demand of users in numerous research institutions who deal with an enormous volume of data, requiring adequate ways of processing to accurately estimate statistical and biological parameters.

Particularly in the case of plant genetics, it is noted that the intensive breeding of many species and the complexity of the most important traits require the use of increasingly accurate selection criteria. In all breeding stages, breeders must use information that is expressed in parameters of the biometric models, which are usually available in the output of most scientifically oriented software systems.

For this purpose, GENES (CRUZ, 2006a, 2006b, 2006c and 2008) was developed to meet the specific needs in the areas of Genetics and Experimental Statistics. The software is freely available to the scientific community at www.ufv.br/dbg/genes/genes.htm, www.ufv.br/dbg/genes.htm or http://www.ufv.br/dbg/biodata.htm.

Description

The software GENES is compatible with IBM PCs and requires the Windows 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 package comprises 257 executable projects, 131 text documents in rtf format, occupying about 285Mbytes, available in English and Portuguese.

Application of the program

An application of the program Genes usually includes the following steps:

a. *Examples of data files:* Examples of data files to be processed by Genes are available, which are particularly useful in initial studies, for providing a double learning effect about the operation of the application itself and of the statistical and biometrical techniques used. Each procedure is represented by an icon that accesses the file containing an illustrative example of a particular procedure, with the advantage of the complete description of all its parameters for immediate data analysis.

b. *Supplying data for processing:* The procedures generally have a common sequence of data analysis. Basically, the user provides the name of the file containing the data to be processed, information about the parameters (number of variables, treatments, blocks etc.), the names of the variables (optional), and then prints or saves the results. It is recommended that these data files should be in .txt format, but they are importable from excel spreadsheets.

The data are supplied in a file containing a data spreadsheet, in which each column represents a certain characteristic to be analyzed and each row the experimental observation. Sometimes, the first columns are reserved to describe classificatory variables or descriptor effects, e.g., treatments, blocks, years, locations etc.

c. *Parameter Description:* For each procedure, the user must provide specific information on the data file that will be used in processing. For each procedure, specific information is requested. Thus, for example, to perform variance analysis, the user should provide the number of variables to be analyzed, the number of treatments and the number of blocks or replications. In other procedures, other information will be solicited, but in the different procedures, the control buttons on the right side of the screen are common. These buttons represent:

Return: ends operations on the screen of parameter identification.

Read Data: reads the file data, considering all rows and columns. This option is useful to identify gaps in the structure of the data file. Through statistics indicating average, maximum and minimum, possible typos can be detected. An error in data reading would definitely lead to errors in data processing, so that the user must apply the necessary corrections, according to the specification of each procedure, to ensure correct data reading and analyses.

d. *Definition of names of variables:* After providing the information in the procedure of parameter identification, the user can name the variables analyzed. If the variables are not named, the program will apply the description: $X_1, X_2, \dots , X_n$.

e. *Result output:* Results are provided by a proper editor of the program Genes. However, the output file can be exported to Word, allowing the use of all features of this powerful editor. In this case, we present the results in a file with font Courier New 8, with customized heading and page numbering. Results can also be exported to Excel or Wordpad and diagrams and figures to Excel or Mspaint.

Modules

The Genes software system contains analysis modules that involve several procedures of biometric analysis, as described below.

Biometrics

Biometrics is the application of statistics to the biological field, being essential for planning, assessment and interpretation of all data obtained in research in the biological area. A growing user demand is noted in various research institutions in the biological area, especially with a view to genetic studies, which deal with large data volumes. This requires an adequate processing, to ensure an accurate estimation and interpretation of the statistical and biological parameters. But there is a market gap for software that would supply this demand. In this context, the program GENES was developed to cover mainly the area of biometrics, with numerous procedures for an adequate data processing. The following procedures are available in this module:

a. Genotype x Environment Interaction: stratification analysis, dissimilarity and correlations between environments.

b. Stability and Adaptability: analysis by methods based on ANOVA (traditional, PLAISTED; PETERSON, 1959; WRICKE, 1965; ANNICCHIARICO, 1992), regression (EBERHART; RUSSELL, 1966; FINLAY; WILKINSON, 1963; TAI, 1971), bi-segmented regression (VERMA et al., 1978; CRUZ et al., 1989) nonparametric analysis (HUEHN, 1990, visual analysis and LIN; BINNS, 1988), analysis of factors and main components or centroids.

c. Genetic gains from selection – Indices: calculation of gains by selection between families (univariate and indices), considering direct and indirect selection, the classic index of Smith (1936) and Hazel (1943), based on the sum of ranks of Mulamba and Mock (1978), base index of Williams (1962), multiplicative index of Subandi et al. (1973), weight-free index of Elston (1963), based on the desired gains of Pesek and Baker (1969) and on the genotype-ideotype distance index. Calculation of gains by selection between families by univariate methods or by the following restricted indices: classic index of Smith (1936) and Hazel (1943), of Kempthorne and Nordskog (1959), of Tallis (1962), of James (1968) and of Cunningham et al. (1970). Calculation of gain by selection among families considering collinearity indices, indices of gains by selection among and within families, in balanced and unbalanced experiments, by massal and stratified selection among and within families. Visual selection analysis, multi-environment selection and prediction of gains by selection within, without information from plants within a plot.

d. Diallel Analysis: Analysis of balanced diallels (Methodologies of GRIFFING, 1956; GARDNER; EBERHART, 1966; HAYMAN, 1954; COCKHERHAN; WEIR, 1977, tests among hybrids and reciprocals crosses, prediction of compounds and hybrids and of family indices) joint diallel analysis (of balanced diallels of GRIFFING, 1956; GARDNER; EBERHART, 1966, and of partial and circulating diallels), Partial diallels (by the methodologies of GERALDI; MIRANDA FILHO, 1988, of MIRANDA FILHO; GERALDI, 1984, of KEMPTHORNE, 1966, of VIANA et al., 1999, and prediction of triple and double hybrids). Analysis of circulating, circulating partial and unbalanced diallels.

e. Analysis of Segregating and non-segregating generations: scale joint test (P1, P2, F1, F2 with optional inclusion of BC1 and BC2), analysis of experiments of segregating lines and parents in alternating rows and analysis of plants in generation Ft and the derived Ft+1 lines.

f. Repeatability: Analysis of original or classified data.

g. Combined selection: analysis of experiments of families with balanced and unbalanced data. Analysis of genetic design proposed by Comstock and Robinson (1948).

h. Genetic and Environmental Progress.

i. Nuclear Collection.

Multivariate Analysis

The designation multivariate analysis represents a large number of methods and techniques that simultaneously use all variables in the analysis, interpretation and processing of the data set from a biological phenomenon under study. The mathematical complexity, typical of multivariate methods, has inhibited the transfer of the underlying stochastic fundamentals and principles to the researchers. However, the key part, which is the statistical inference, has been stimulated through the use of well-constructed software with a user-friendly interface for researchers. In the program Genes, the scientist will find the following:

a) Analysis of structural simplification: Principal Components and Canonical Variable Analysis.

b) Association Analysis: Path analysis, Canonical Correlations and Factor analysis

c) Analysis of diversity: Discriminant Analysis (by the method proposed by Anderson or based on principal components). Measures of Dissimilarity: based on continuous, multicateegoric or binary phenotipic quantitative variables. Analysis of molecular data from dominant or codominant markers; cluster analysis: Tocher optimization method, hierarchical, graphic dispersion and 2D and 3D projection. Identification of more and less similar accessions. Importance of traits: by main components or the distance by Mahalanobis' Generalized distance and canonical variable analysis.

Simulation

One the major contributions of computing is that phenomena can be studied by simulating a complex situation in which parameters and constraints are established, so that the effect of certain controllable factors can be conveniently studied. Simulation is defined as a way of imitating the behavior of a real system by computational resources to study its functioning under alternative conditions, involving certain types of logic models to describe, as best as possible, the natural system .

Simulations are highly useful in genetic studies in various contexts, including studies of populations, the individual or of the proper genome. They require the development of appropriate biological models to represent the phenomena of interest as ideally as possible by researchers and suitable procedures of processing by programmers, according to the parameters and constraints, so that the influence of certain factors can be assessed.

Genes contained the procedures: Simulation of experiments, Simulation of Samples (*p* populations and *v* variables), Optimal Number of Families, Optimal Number of Plants (Random or predifined Sampling) and Optimal Number of Replications or Optimal Sample Size

Genetic Diversity

Studies on diversity can be directed to plant breeding, evolutionary associations, conservation and management of plant material, among other purposes. In each case, an adequate methodology and appropriate information are required. The data of measured units, plants, accessions or taxa can be phenotypic or genotypic. Phenotypic information is derived from the evaluation of characteristics with continuous or discrete distributions, of which the latter can be multicategoric or binary. Genotypic data are obtained from molecular markers or DNA sequencing. In the case of markers, there are dominant or co-dominant and diallelic or multiallelic types. All these situations are addressed in the application Genes, by the approach:

a. Diversity between accessions: based on continuous, multi-categoric, binary phenotypic variables, and analysis of data of dominant and codominant (multi-allelic) markers.

b. Diversity between populations: Nei's Genetic identity Calculation (1972) and the following distances: Euclidean, of Rogers, Angular, of Goldstein et al. (1985) and of Hedrick.

c. Diversity within populations: calculation of the coefficient of endogamy and heterozygosis, Shannon-Wiener index and heterozygosis from binary data.

d. Diversity among and within populations: descriptive analysis, Nei's diversity index (1973), Wright's fixation index (Two alleles or Multiple alleles), analysis of heterozygosity of Weir (1996). Analysis of Contingency Table, ANOVA of allelic frequency (F, f and θ), AMOVA of Excoffier et al. (1992) and analysis of binary data.

e. Discriminant Analysis: discriminant analysis of Anderson, analysis based on main components or in k-nearest neighbors. Discriminant analyses from the dissimilarity matrices.

f. Grouping analysis: using the following methods: Tocher optimization and hierarchical methods, by graphic dispersion, 2D and 3D projection and analysis of more and less similar accessions. Matrices of Dissimilarities: calculation of the correlation and sum between elements of matrices of dissimilarity. Importance of traits: considering phenotypic quantitative characters or molecular information, by means of MANOVA.

g. Optimization: Analysis of the optimal number of binary or multi-allelic markers for the study of genetic variance. Simulation: simulation of populations, crossings and population samples, under the effect of divergent selection or genetic drift.

h. Relationship coefficient and Hardy-Weinberg Equilibrium: Population analysis based on the information of codominant diallelic or multi-allelic markers. Analysis of Gametic Disequilibrium.

Experimental Statistics

This module contains procedures based on statistical models with wide application in various areas of research and undergraduate and graduate teaching. The importance of statistical analysis is the probabilistic proof of the truth of a particular hypothesis formulated based on extensive studies and on analyses of the research results. In statistics, parameters estimates related to the data are presented and interpreted per se, or hypothesis are tested and results are associated with probability values by means of statistical tests. Usually, the use of a particular inferential statistics is directed by the study question. The software Genes offers the following procedures for statistical analyses:

Descriptive Statistics, Normality Test and Stand Correction Methods

a. Variance Analysis: analysis of completely randomized designs and schemes, of experiments with regular and non-regular treatments, in randomized blocks, factorial and subdivided plots. Analysis of origin/progeny/plant, simple and triple lattices and hierarchical models.

b. Regressions: simple linear, non-linear, multiple and polynomial, response surface and 3D graphical analysis.

c. Correlations: calculation of genetic correlations, partial and canonical Pearson and Spearman correlations. Path analysis (involving 1 or 2 chains) and path analysis under collinearity.

d. Comparison of Means: Tests of Tukey, Duncan, Scheffé and Scott and Knott, Tukey test with variable number of replications, Dunnett, t test, Tocher, and chi-square test to evaluate hypotheses, heterogeneity and factorial linkage.

Matrices

The study of matrices is considered fundamental because it is an important tool in this area of mathematics related to calculations and parameter estimation. It is widely applied in estimation methods and model adjustments, such as least squares and maximum likelihood and different matrix analyses. The following procedures are available in Genes:

a. Diagnosis of multicollinearity

b. Algebra of matrices

c. Solution of the system $Y = X\beta + \varepsilon$

d. Solution of the system $X'X\hat{\beta} = X'Y$

Integration with other software

Currently, the software GENES has 205 executable projects involving the modules of experimental statistics, biometrics, multivariate analysis, genetic diversity, and simulation matrices. Thus, each procedure has a particular data set for which an appropriate biometric template is prepared that will allow the user to process data and generate and properly interpret results of the studied phenomenon. However, additional analyses may be required or even some differentiated form of carrying out the same kind of study may be evaluated. In this case, the user would surely be willing to try a new analysis option provided no effort is required to understand the particular access to an alternative program or supplement. The user of software GENES has direct access to other applications such as:

Microsoft Word: designed to receive output results and emit reports

Microsoft Excel: designed to receive outputs or results of complementation analysis, in particular graphical analysis.

Microsoft Paint: designed to receive figures, images, and diagrams resulting from the analysis to which, as the researcher sees fit, graphical resources can be applied to improve the aesthetics of the result.

Free software environment R: For each procedure available within software GENES, the user finds a set of instructions for the appropriate settings so that the data can be accessed and processed by program R, according to the researcher's demand. The program R has been increasingly accepted by universities and companies around the world. Nowadays, the acquisition costs of statistical software packages that are similar or even poorer in terms of analysis capacity, are very high, especially for the predominantly small and medium businesses in our country. Thus, the inclusion of this facility in Genes is yet a another contribution to the use of R, intended to break barriers and facilitate the construction of diagrams and data analyses of quality data, at no cost and with the same reliability as of other software.

Matlab®: Software GENES generates established scripts by a set of sentences or commands to perform or solve problems of a particular type of study based on a set of data or information, within the Matlab program. Matlab is an interactive system whose basic data element is an array that does not require dimensioning. This system allows the resolution of many numerical problems in a fraction of the time one would spend writing a similar program in Fortran, Basic or C. Moreover, solutions to problems are expressed almost exactly as they are written mathematically. Each script consists of set of methods organized and documented by one or more parts of a process allowing, if necessary, the identification and correction of errors by means of debugging the script to obtain a solution with no errors.

Conclusion

GENES is a software package very important for data analysis and processing with different biometric models and is essential in genetic studies applied to plant and animal breeding. The software should supply the increasing demand of users in numerous research institutions who deal with an enormous volume of data, requiring adequate ways of processing to accurately estimate statistical and biological parameters.

Acknowledgements

The author gratefully acknowledges FAPEMIG, CAPES and CNPq for their financial support.

References

ANNICCHIARICO, P. Cultivar adaptation and recomendation from alfafa trials in Northern Italy. **Journal of Genetics and Plant Breeding**, v. 4, n. 1, p. 269-278, 1992.

COCKHERHAN, C. C.; WEIR, B. S. Quadratic analyses of reciprocal crosses. **Biometrics**, v. 33, n. 3, p. 187-203, 1977.

COMSTOCK, R. E.; ROBINSON, H. F. The components of genetic variance in populations of biparental progenies and their use in estimating the average degree of dominance. **Biometrics**, v. 4, n. 1, p. 254-266,1948.

CRUZ, C. D. **Programa Genes** - Estatística Experimental e Matrizes. 1st ed. Viçosa: UFV, 2006a.

CRUZ, C. D. **Programa Genes** - Biometria. 1st ed. Viçosa: UFV, 2006b.

CRUZ, C. D. **Programa Genes** - Análise multivariada e simulação. 1st ed. Viçosa: UFV, 2006c.

CRUZ, C. D. **Programa Genes** - Diversidade Genética. 1st ed. Viçosa: UFV, 2008.

CRUZ, C. D.; TORRES, R. A.; VENCOVSKY, R. An alternative approach to the stability analysis proposed by Silva e Barreto. **Revista Brasileira de Genética**, v. 12, n. 3, p. 567-580, 1989.

CUNNINGHAM, E. P.; MOEN, R. A.; GJEDREM, T. Restriction of selection indexes. **Biometrics**, v. 26, n. 4, p. 67-74, 1970.

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties. **Crop Science**, v. 6, n. 1, p. 36-40, 1966.

ELSTON, R. C. A weight-free index for the purpose of ranking or selection with respect to several traits at a time. **Biometrics**, v. 19, n. 5, p. 85-97, 1963.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. **Genetics**, v. 131, n. 5, p. 479-491, 1992.

FINLAY, K. W.; WILKINSON, G. N. The analysis of adaptation in a plant breeding programme. **Australian Journal of Agricultural Research**, v. 14, n. 3, p. 742-754, 1963.

GARDNER, C. O.; EBERHART, S. A. Analysis and interpretation of the variety cross diallel and related populations. **Biometrics**, v. 22, n. 5, p. 439-452, 1966.

GERALDI, I. O.; MIRANDA FILHO, J. B. Adapted models for the analysis of combining ability of varieties in partial diallel crosses. **Revista Brasileira de Genética**, v. 11, n. 1, p. 419-430, 1988.

GOLDSTEIN D. B.; LINARES A. R.; CAVALLI-SFORZA, L. L.; FELDMAN M. W. Genetic absolute dating based on microsatellites and the origin of modern humans. **Proceedings of the National Academy of Sciences**, v. 92, n. 3, p. 6723-6727, 1995.

GRIFFING, B. Concept of general and specific combining ability in relation to diallel crossing systems. **Australian Journal of Biological Sciences**, v. 9, n. 1, p. 463-493, 1956.

HAYMAN, B. I. The theory and analysis of diallel crosses. **Genetics**, v. 39, n. 2, p. 789-809, 1954.

HAZEL, L. N. The genetic basis for constructing selection indexes. **Genetics**, v. 28, n. 1, p. 476-490, 1943.

HUENH, M. Nonparametric measures of phenotypic stability. Part 1: Theory. **Euphytica**, v. 47, n. 3, p. 189-194, 1990.

JAMES, J. W. Index selection with restrictions. **Biometrics**, v. 24, n. 1, p. 1015-1018, 1968.

KEMPTHORNE, O. **An introduction to genetic statistics**. New York: John Wiley and Sons, 1966.

KEMPTHORNE, O.; NORDSKOG, A. W. Restricted selection indices. **Biometrics**, v. 15, n. 1, p. 10-19, 1959.

LIN, C. S.; BINNS, M. R. A superiority measure of cultivar performance for cultivar x location data. **Canadian Journal of Plant Science**, v. 68, n. 3, p. 193-198, 1988.

MIRANDA FILHO, J. B.; GERALDI, I. O. An adapted model for the analysis of partial diallel crosses. **Revista Brasileira de Genética**, v. 7, n. 1, p. 667-688, 1984.

MULAMBA, N. N.; MOCK, J. J. Improvement of yield potential of the Eto Blanco maize (*Zea mays* L.) population by breeding for plant traits. **Egyptian Journal of Genetics and Cytology**, v. 7, n. 1, p. 40-51, 1978.

NEI, M. Analysis of gene diversity in subdivided population. **Proceedings of the National Academy of Sciences**, v. 70, n. 12, p. 3321-3323, 1973.

NEI, M. Genetic distance between populations. **American Naturalist**, v. 106, n. 1, p. 283-292, 1972.

PESEK, J.; BAKER, R. J. Desired improvement in relation to selected indices. **Canadian Journal of Plant Science**, v. 49, n. 1, p. 803-804, 1969.

PLAISTED, R. L.; PETERSON, L. C. A technique for evaluating the ability of selections to yield consistently in different locations and seasons. **American Potato Journal**, v. 36, n. 2, p. 381-385, 1959.

SMITH, H. F. A discriminant function for plant selection. **Annals of Eugenics**, v. 7, n. 1, p. 240-250, 1936.

SUBANDI, W.; COMPTON, A.; EMPIG, L. T. Comparison of the efficiencies of selection indices for three traits in two variety crosses of corn. **Crop Science**, v. 13, n. 1, p. 184-186, 1973.

TAI, G. C. C. Genotypic stability analysis and its application to potato regional trials. **Crop Science**, v. 11, n. 1, p. 184-190, 1971.

TALLIS, G. M. A selection index for optimum genotype. **Biometrics**, v. 18, n. 2, p. 120-122, 1962.

VERMA, M. M.; CHAHAL, G. S.; MURTY, B. R. Limitations of conventional regression analysis: a proposed modification. **Theoretical and Applied Genetics**, v. 53, n. 3, p. 89-91, 1978.

VIANA, J. M. S.; CRUZ, C. D.; CARDOSO, A. A. Theory and analysis of partial diallel crosses. **Genetics and Molecular Biology**, v. 22, n. 4, p. 591-599, 1999.

WEIR B. S. **Genetic Data Analysis II**. Sunderland: Sinauer Associates Inc., 1996.

WILLIANS, J. S. The evaluation of a selection index. **Biometrics**, v. 18, n. 1, p. 375-393, 1962.

WRICKE, G. Zur berechning der okovalenz bei sommerweizen und hafer. **Pflanzenzuchtung**, v. 52, n. 1, p. 127-138, 1965.

Received on June 10, 2012.
Accepted on June 21, 2012.