Ciencia e Investigación AGRARIA

RESEARCH PAPER

Cien. Inv. Agr. 37(3):83-89. 2010
www.rcia.uc.cl

Generalized composite interval mapping offers improved efficiency in the analysis of *loci* influencing non-normal continuous traits

Freddy Mora[1], Carlos Alberto Scapim[2], Adam Baharum[3], and Antonio Teixeira do Amaral Júnior[4]

[1]Facultad de Ciencias Forestales, Universidad de Concepción. Victoria 631, Barrio Universitario, Concepción, Chile.

[2]Departamento de Agronomia, Centro de Ciências Agrárias, Universidade Estadual de Maringá. Avenida Colombo, 5790, Bloco 5-PGM, Maringá-PR, Brasil.

[3]School of Mathematical Sciences, Universiti Sains Malaysia. 11800 Penang, Malaysia.

[4]Laboratório de Melhoramento Genético Vegetal, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Parque California, 28013-602, Campos dos Goytacazes, RJ, Brasil.

Abstract

F. Mora, C.A. Scapim, A. Baharum, and A.T. Amaral-Junior. 2010. Generalized composite interval mapping offers improved efficiency in the analysis of *loci* influencing non-normal continuous traits. Cien. Inv. Agr. 37(3):83-89. In genetic studies, most Quantitative Trait *Loci* (QTL) mapping methods presuppose that the continuous trait of interest follows a normal (Gaussian) distribution. However, many economically important traits of agricultural crops have a non-normal distribution. Composite interval mapping (CIM) has been successfully applied to the detection of QTL in animal and plant breeding. In this study we report a generalized CIM (GCIM) method that permits QTL analysis of non-normally distributed variables. GCIM was based on the classic Generalized Linear Model method. We applied the GCIM method to a F_2 population with co-dominant molecular markers and the existence of a QTL controlling a trait with Gamma distribution. Computer simulations indicated that the GCIM method has superior performance in its ability to map QTL, compared with CIM. QTL position differed by 5 cM and was located at different marker intervals. The Likelihood Ratio Test values ranged from 52 (GCIM) to 76 (CIM). Thus, wrongly assuming CIM may overestimate the effect of the QTL by about 47%. The usage of GCIM methodology can offer improved efficiency in the analysis of QTLs controlling continuous traits of non-Gaussian distribution.

Key words: bioinformatics, Generalized Linear Model, molecular markers, Quantitative Trait *Loci* (QTL).

Introduction

The great majority of traits of agricultural crops are a result of the joint action of several genes with generally continuous phenotypic variation. The expression of these quantitative traits is controlled by *loci* termed QTL (Quantitative Trait *Loci*). There are two basic categories of molecular markers (Ovesná *et al*., 2002): markers that segregate and determine the presence of a single, dominant or recessive gene and QTL-associated markers. It is much easier and economic to develop markers for a trait inherited by a single gene than markers based on QTL.

Received August 28, 2009. Accepted November 18, 2009.
Corresponding author: morapoblete@gmail.com, freddymora@udec.cl

The accuracy of QTL mapping (i.e., detection or analysis) must therefore be as high as possible. Statistical procedures have been extensively studied because they are essential for improving the accuracy of genetic analyses (Mora *et al*., 2008a). In a simulation study on the accuracy of position and effect estimates of linked QTL, Mayer *et al*. (2004) found that the reduction of the marker interval size from 10 cM (centiMorgan) to 5 cM led to a higher power in QTL detection and to an improvement of the QTL position as well as the QTL effect estimates.

The statistical association between a marker locus and the genomic region of a QTL (QTL mapping within the chromosome) is an important tool for genetic improvement programs. The genotypes of known QTLs can be added to the information about plant performance and together with genealogical information, can be used to increase the prediction accuracy of genetic values in traditional breeding methods (Gonçalves-Vidigal *et al*., 2008): as in the case of BLUP (Best Linear Unbiased Prediction). With the advancement of genomic studies, molecular markers linked to QTL are becoming increasingly available as additional information for genetic evaluation and could be used to increase selection efficiency via marker-assisted selection (Liu and Zeng, 2005).

The strategy used to detect whether a marker and a QTL are linked is the statistical analysis of association between the phenotypic variation of the trait and markers. The methods of simple regression, maximum likelihood and those based on Monte Carlo Markov Chains have been frequently used to detect QTLs, which are focused mainly on a continuous distribution (Normal or Gaussian) of the trait of interest (Thomson, 2003). However, several studies related to plant breeding have shown that several traits of economic and scientific interest have non-Gaussian distribution (Spyrides-Cunha *et al*., 2000; Ribeiro *et al*., 2005; Mora *et al*., 2007; Mora *et al*., 2009).

In this sense, the methodology of the Generalized Linear Models (GLM) was developed in the 70's. It is based on distributions of the exponential type (termed exponential family distributions), and uses methods similar to traditional linear approaches for normal data distribution (Myers *et al*., 2002). Distributions such as: Gamma, Poisson, Binomial, Multinomial and Normal are some examples of distributions that belong to the exponential family, which are frequently found in agronomic experiments (Mora *et al*., 2008b; Mora *et al*., 2008c). In the GLM approach, the assumptions of normality required for conventional analysis can be relaxed if the trait of interest follows any exponential family distribution.

The current study has been motivated by the existence of traits of interest with other than Gaussian distributions (Mora *et al*., 2007; Mora *et al*., 2008b; Rodovalho *et al*., 2008; Mora *et al*., 2008d). Therefore, this study aimed to map a quantitative trait *locus* by using a generalized linear regression modeling approach where the agronomical trait is non-normally distributed. Knowledge about the statistical method used to map QTL will enhance the accuracy of genetic-quantitative analyses and improve our understanding on the genomic regions that control agronomic traits of interest.

Material and methods

The methodology of the generalized linear regression was used to map a QTL that controls a trait of non-Gaussian distribution. For this purpose, data of 11 codominant molecular markers, distributed within a chromosome of 105.6 centiMorgan (cM, a distance of Haldane), were simulated using the program GQMOL (Schuster and Cruz, 2004). The supposed breeding population consisted of a F_2 population with 252 individuals.

It was assumed that the trait under selection is genetically controlled by an infinite number of additive *loci*, each one with an infinitesimal effect (poly-genes), plus a single two-allele QTL (alleles Q and q). The agronomical data were simulated in R program (Ihaka and Gentleman, 1996) using statistical parameters from Mora *et al*. (2008b). Therefore, it was assumed a trait of interest with Gamma distribution, which was subsequently adjusted to data of molecular markers. The population was assumed to be a breeding F_2 population.

Let $y_1, ..., y_n$ denote n independent observations on a response; a realization of a random variable Y_i. According to previous assumptions, the quantitative trait simulated here has a distribution that belongs to the exponential family (Dobson, 2001). The density and probability function for the observed response y can be expressed as:

$$f(y_i;\theta_i,\phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\} \quad [1]$$

where $a_i(\phi)$, $b(\theta_i)$ and $c(y_i,\phi)$ are called specific functions. The parameter θ_i is related to the mean of the distribution. ϕ, called the dispersion parameter, typically is known and is usually related to the variance of the distribution (Dobson, 2001; Myers *et al*., 2002).

If Y_i has a distribution in the exponential family then it has mean and variance:

$$E(Y_i) = \mu_i = b(\theta_i) \quad [2]$$

$$\text{var}(Y_i) = \sigma_i^2 = b``(\theta_i)a_i(\phi) \quad [3]$$

The generalized regression model, considering the composite interval mapping (CIM) of QTL (Haley and Knott, 1992), includes additive and dominant effects of the QTL.

Assuming a Generalized Composite Interval Mapping (GCIM) for QTL analysis with an exponential family distribution (in this case Gamma), the model is constructed around the linear predictor:

$$\eta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \quad [4]$$

where β_0 and β_i are the model parameters (intercept, additive and dominance effects of the QTL and the markers considered as co-factors). The model is found through the use of a link function: $\eta_i = g(\mu_i)$. In this study, the logarithmic link function was used according to Myers *et al*. (2002).

The R program (Ihaka and Gentleman, 1996) was used for QTL mapping, according to Myers *et al*. (2002), to adjust the model of generalized linear regression. The Likelihood Ratio Test (LRT) was used to compare the models of each interval mapping. These results were compared with an analysis where a normal distribution of the agronomic trait of interest was assumed (CIM).

Results and discussion

The linkage group (chromosome) and the respective distances between the co-dominant molecular markers are shown in Figure 1-A. The Haldane distance ranged from 5.5 cM (estimated between the markers SSR6 and SSR7) to 19.1 cM (between SSR2 and SSR3). The observed and expected distribution of the trait is shown in Figure 1-B (obtained using R program). The relatively small asymmetry in relation to the normal distribution was confirmed by the statistical tests Shapiro-Wilk and Lilliefors ($P < 0.01$). The graphic analysis confirmed the Gamma distribution of the response variable, according to Freund (1992).

The results of the markers associated with the trait of interest are shown in Table 1. This is a key procedure to identify co-factors that will be used in the composite interval mapping. The generalized linear regression identified the markers

Figure 1. (A) Diagram of the linkage group constructed with 11 molecular markers (MM), in the lower part of the linkage group, the Haldane distances between MM are shown in cM; (B) Diagram of expected (Gaussian: dotted line) and observed (Gamma: continuous line) distribution of the quantitative trait.

SSR3 and SSR11 as significant ($P < 0.05$). A similar analysis using stepwise regression (considering Normal error distribution) confirmed these molecular markers, but also indicated statistical significance for SSR9 ($P < 0.05$). Although the results of composite interval mapping showed no differences to the procedure with simple intervals (data not shown) due to the simplicity of the genomic data, it is important to emphasize that the differences found in the selection of cofactors can significantly affect QTL detection according to Schuster and Cruz (2004).

Figure 2 shows a diagram with the QTL location depending on the analysis method used: 1) ignoring the real trait distribution that is, erroneously assuming normal distribution (QTL1) and 2) assuming the real distribution of the response variable (Gamma) and therefore using the principle of Generalized Linear Models for QTL mapping (QTL2). The Likelihood Ratio Test (LRT) values ranged from 52 (QTL2) to 76 (QTL1). Thus, wrongly assuming the normal distribution model, the real effect of the QTL was overestimated by about 47%.

In this study, it may be noted that the QTL analysis or mapping, where the normality rejection is disregarded, was relatively robust in the sense that both approaches detected significant evidence for a QTL ($P < 0.01$) within this genetic linkage group. However, the QTL location, between one methodology and the other, differed by 5 cM, despite the small deviation in

Table 1. Generalized Linear Regression method used for the selection of molecular markers as co-factors for the composite interval mapping of QTL. In bold, the significant molecular markers ($P < 0.05$).

Marker	Estimate	Standard error	Wald Confidence limits (95%)		χ^2	$P > \chi^2$
SSR1	-0.0210	0.1595	-0.3336	0.2916	0.02	0.895
SSR2	0.1213	0.1845	-0.2403	0.4828	0.43	0.511
SSR3	0.4454	0.1553	0.1409	0.7498	8.22	0.004
SSR4	0.1952	0.1954	-0.1878	0.5782	1.00	0.318
SSR5	0.0605	0.2715	-0.4717	0.5927	0.05	0.824
SSR6	-0.1520	0.2813	-0.7034	0.3994	0.29	0.589
SSR7	0.2954	0.255	-0.2044	0.7952	1.34	0.247
SSR8	-0.2948	0.2304	-0.7464	0.1567	1.64	0.201
SSR9	-0.0398	0.1703	-0.3736	0.2941	0.05	0.816
SSR10	-0.1234	0.1796	-0.4754	0.2286	0.47	0.492
SSR11	0.3671	0.1586	0.0562	0.6780	5.36	0.021

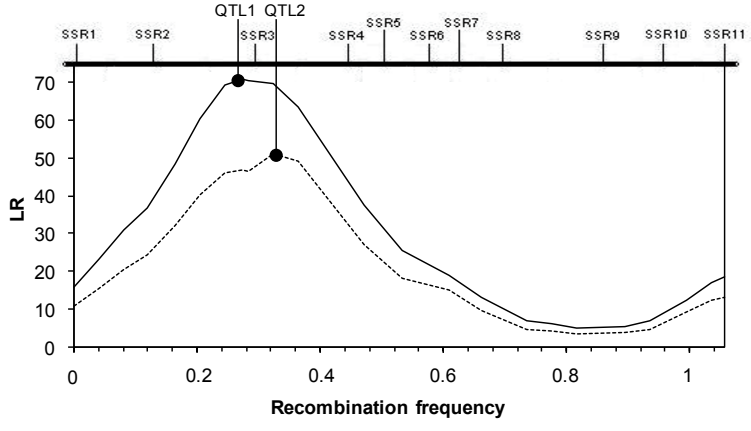

Figure 2. QTL location, which was dependent on the analysis method: erroneously assuming normal distribution (QTL1), and assuming the true distribution (Gamma) of the response variable (QTL2), which was analyzed by using a generalized linear regression model. LR is the likelihood ratio test; SSR1, SSR2... and SSR11 are the co-dominant markers.

the trait distribution. Another important result is that the QTL was detected at different intervals. The differences may seem relatively small, but we emphasize the fact that the results presented here are based on an extremely simple molecular data (only one linkage group and one QTL). The accuracy of QTL mapping must be as high as possible (Mora *et al*., 2008a). Mayer *et al*. (2004) for example, found that the reduction of the marker interval size from 10 cM to 5 cM led to a higher power in QTL detection and to a remarkable improvement of the QTL position as well as the QTL effect estimates.

The generalized linear model fitted to the data of this study was based on the logarithmic link function. Technically, this link is not the canonical (natural) function of the Gamma distribution, but it is often used (Myers *et al*., 2002) since mathematical problems can be overcome with the use of the reciprocal function, which is the canonical link of the Gamma distribution.

Application of molecular marker techniques has helped to better understand characters controlled by multiple genes. In principle, QTL can be used for Marker Assisted Selection (MAS). Moreover, with the help of markers it is possible to begin with specific selection in earlier generations. For these reasons, the procedures for mapping quantitative trait loci must be meticulously studied because they are essential for determining statistical association between molecular data and the agronomical trait of interest (Mora *et al*., 2008a). In the current study, the results by the approach of generalized linear modeling for QTL mapping were promising. If we presume that 1 cM is equivalent to approximately 1 million base pairs in this study, the differences between one procedure and the other can therefore be of the order of 5 million base pairs. Furthermore, the fact that the QTLs were found at different intervals can mean the loss of important resources, when the QTL localization is needed.

Originally, the theory of the GLMs is an alternative approach to data analysis where the normality assumption is unrealistic, as reported here. The use of the GLM method can be interesting and effective in QTL mapping in situations where the distribution of the response variable belongs to the exponential family (Dobson, 2001). Although there are several methods of detecting QTLs, from the simple regression to the Bayesian methods (Thomson, 2003), most procedures assume a normal trait distribution.

As in our breeding programs, Thomson (2003) argued that many traits of both scientific and economic interests have non-normal distribution. For example, binary data are often found when the aim is to improve traits such as disease status (Setiawan *et al*., 2000; Park *et al*., 2001), mortality or survival (Mora *et al*., 2008d), flowering (Missiaggia *et al*., 2005; Mora *et al*., 2007; Mora *et al*., 2009), among other traits. Yang *et al*. (2009) stated that deviations from this assumption may affect the accuracy of QTL detection and lead to detection of spurious QTLs. The current study confirmed that the mapping of QTLs on control traits of non-Gaussian distribution could be improved by the theory of Generalized Linear Models.

It is important to emphasize that various approaches have been studied to deal with non-normal phenotypes in QTL mapping. The Generalized Estimating Equation procedure, for example, is a natural extension of generalized linear models, which has shown to be useful for mapping QTLs affecting longitudinal non-normal traits (Lange and Whittaker, 2001; Mora *et al*., 2008a). Bayesian methods are also considered to be able to study QTL affecting non-normal traits. Recently, Yang *et al*. (2009), for example, studied the genetic architecture of quantitative trait by combining the flexibility of Bayesian approach in modeling multiple QTL and their interactions and the better phenotypic fitting of symmetric and long-tailed distributions in characterizing non-normal traits.

In the current study, the accuracy in QTL mapping depended on the methodological approach. Information about the distribution of agronomic data should be studied and included in breeding programs to improve the reliability of the QTL mapping. In this sense, the generalized composite interval mapping method can offer improved efficiency in the analysis of QTLs controlling continuous traits of non-Gaussian distribution.

Resumen

F. Mora, C.A. Scapim, A. Baharum y A.T. Amaral-Junior. 2010. Mapeo por intervalos compuestos generalizados entrega una mejorada eficiencia en el análisis de loci que afectan características continuas no-normales. Cien. Inv. Agr. 37(3):83-89. En estudios genéticos, la mayoría de los métodos de mapeo de Loci de Característica Cuantitativa (LCC) presupone que la característica de interés (continua) sigue una distribución Normal (Gaussiana). Sin embargo, muchas de las características económicamente importantes de cultivos agrícolas tienen una distribución no-normal. Mapeo por intervalos compuestos (MIC) ha sido aplicado exitosamente para la detección de LCC en el mejoramiento de plantas y animales. En este estudio se investigó el método generalizado de MIC (MICG) que permite el análisis de LCC de variables distribuidas no normalmente. MICG fue basado en el método clásico de los Modelos Lineales Generalizados. Se aplicó el MICG a una población F_2 con marcadores moleculares co-dominantes y la existencia de un LCC que controla una característica con distribución Gamma. Simulaciones de computador indicaron que el método MICG tiene un poder superior en su habilidad para mapear LCC, en comparación con MIC. La posición del LCC difirió en 5 cM, y fue localizado en diferentes intervalos de marcadores. Los valores de la prueba de la Razón de Verosimilitud variaron de 52 (MICG) a 76 (MIC). Por lo tanto, asumiendo erróneamente MIC, se podría sobreestimar el efecto del LCC en alrededor de 47%. El uso de MICG puede ofrecer una mejorada eficiencia en el análisis de LCC que controlan características de distribución diferente a la Gaussiana.

Palabras clave: bioinformática, *Loci* de Característica Cuantitativa (LCC), marcador molecular, Modelo Linear Generalizado.

References

Dobson, A.J. 2001. An introduction to generalized linear models. London, England. Chapman and Hall, 225 pp.

Freund, JE. 1992. Mathematical statistics. New Jersey, USA. Prentice-Hall. 658 pp.

Gonçalves-Vidigal, M.C., F. Mora, T.S. Bignotto, R.E.F. Munhoz, and L.D. Souza. 2008. Heritability of quantitative traits in segregating common bean families using a Bayesian approach. Euphytica 164:551-560.

Haley, C.S., and S.A. Knott. 1992. A simple regression method for mapping quantitative trait loci in linecrosses using flanking markers. Heredity 69:315-324.

Ihaka, R., and R. Gentleman. 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5:299-314.

Lange, C., and J.C. Whittaker. 2001. Mapping quantitative trait loci using Generalized Estimating Equations. Genetics 159:1325-1337.

Liu, Y., and Z.B. Zeng. 2005. Mixture model equations for marker-assisted genetic evaluation. J. Anim. Breed. Genet. 122:229-239.

Mayer, M., L. Yuefu, and G. Freyer. 2004. A simulation study on the accuracy of position and effect estimates of linked QTL and their asymptotic standard deviations using multiple interval mapping in an F2 scheme. Genet. Sel. Evol. 36(4):455-479.

Missiaggia, A.A., A.L. Piacezzi, and D. Grattapaglia. 2005. Genetic mapping of Eef1, a major effect QTL for early flowering in *Eucalyptus grandis*. Tree Genetics & Genomes 1:79-84.

Mora, F., S. Perret, C.A. Scapim, E.N. Martins, and M.P. Molina. 2007. Source-dependent blooming variability of *Eucalyptus cladocalyx* in the Region of Coquimbo, Chile. Cien. Inv. Agr. 34:131-139.

Mora, F., F. Tapia, C.A. Scapim, E.N. Martins, R.J.B. Pinto, and A. Ibacache. 2008a. Early performance of *Olea europaea* cv. Arbequina, Picual and Frantoio in southern Atacama Desert. Crop Breeding and Applied Biotechnology 8:30-38.

Mora, F., A.I. Santos, and C.A. Scapim. 2008b. Mapping quantitative trait loci (QTLs) using a multi-

variate approach. Cien. Inv. Agr. 35(2):137-145.

Mora, F., L.M. Gonçalves, C.A. Scapim, M.F.P.S. Machado, and E.N. Martins. 2008c. Generalized lineal models for the analysis of binary data from propagation experiments of Brazilian orchids. Brazilian Archives of Biology and Technology 51(5):963-970.

Mora, F., M.C. Gonçalves-Vidigal, and A.I. Santos. 2008d. Bayesian analysis of the genetic control of survival in F_3 families of common bean. Chilean Journal of Agricultural Research 68:334-341.

Mora, F., R. Gleadow, S. Perret, and C.A. Scapim. 2009. Genetic variation for early flowering, survival and growth in sugar gum (*Eucalyptus cladocalyx* F. Muell) in southern Atacama Desert. Euphytica 169:335-344.

Myers, R.H., D.C. Montgomery, and G.G. Vining. 2002. Generalized linear models, with applications in engineering and the sciences. New York, USA. John Wiley and Sons Press, 342 pp.

Ovesná, J., K. Poláková, and L. Leišová. 2002. DNA analyses and their applications in plant breeding. Czech J. Genet. Plant Breed. 38:29-40.

Park, S.O., D.P. Coyne, J.R. Steadman, and P.W. Skroch. 2001. Mapping of QTL for resistance to white mold disease in common bean. Crop Science 41:1253-1262.

Ribeiro, A.O., E. Bearzoti, and T. Sáfadi. 2005. QTL mapping of Poisson traits: a simulation study. Crop Breeding and Applied Biotechnology 5:310-317.

Rodovalho, M.A., F. Mora, E.M. Santos, C.A. Scapim, and E. Arnhold. 2008. Survival heritability in 169 families of white grain popcorn: a Bayesian approach. Cien. Inv. Agr. 35:255-260.

Setiawan, A., G. Koch, S.R. Barnes, and C. Jung. 2000. Mapping quantitative trait loci (QTLs) for resistance to Cercospora leaf spot disease (*Cercospora beticola* Sacc.) in sugar beet (*Beta vulgaris* L.). Theoretical and Applied Genetics 100:1176-1182.

Schuster, I., and C.D. Cruz. 2004. Estatística genômica aplicada a populações derivadas de cruzamentos controlados. Viçosa, Brasil. Universidade Federal de Viçosa. 568 pp.

Spyrides-Cunha, M.H., C.G.B. Demetrio, and L.E.A. Camargo. 2000. Proportional odds model applied to mapping of disease resistance genes in plants. Genetics and Molecular Biology 23:223-227.

Thomson, P.C. 2003. A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. Genet. Sel. Evol. 35:257-280.

Yang, R., X. Wang, J. Li, and H. Deng. 2009. Bayesian robust analysis for genetic architecture of quantitative traits. Bioinformatics 25(8):1033–1039.