Statistical Modeling of Epistasis and Linkage Decay using Logic Regression

T. B. Parker[1], P. Sz cs[1], W. F. Mahaffee[2], J.-L. Jannink[3] and J. A. Henning[4]*

[1]Department of Crop and Soil Science, Oregon State University, Corvallis, OR 97331, USA.

[2]USDA-ARS Horticultural Crops Research Laboratory, Corvallis, OR 97331, USA.

3USDA-ARS, R.W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY 14853, USA.

[4]USDA-ARS National Forage Seed Processing Research Center, Corvallis OR, 97331, USA.

*Corresponding author: E-mail: John.Henning@oregonstate.edu.

Abstract Logic regression has been recognized as a tool that can identify and model non-additive genetic interactions using Boolean logic groups. Logic regression, TASSEL-GLM and SAS-GLM were compared for analytical precision using a previously characterized model system to identify the best genetic model explaining epistatic interaction of vernalization-sensitivity in barley. A genetic model containing two molecular markers identified in vernalization response in barley was selected using logic regression while both TASSEL-GLM and SAS-GLM included spurious associations in their models. The results also suggest the logic regression can be used to identify dominant/recessive relationships between epistatic alleles through its use of conjugate operators.

Introduction

Recent concerns about potential loss of genetic variation in our crop plants (Yu and Bernardo 2004) make it important to understand genetic modeling in an attempt to correctly measure levels of variation within elite breeding germplasm**.** Unfortunately traditional techniques in genetic modeling are thought to underestimate many forms of epistasis (Solomon et al. 2007), which in outcrossing species, is thought to play a significant role in the maintenance of genetic diversity under potential bottleneck conditions as those encountered during advanced stages in the breeding cycle (Yu and Bernardo 2004). In addition, epistatic interactions have been long thought to play a vital

role in the evolutionary diversification of species (Wright 1931; Orr 1995). Epistatic interaction of the *Arabidopsis FRI* and *FLC* flowering time genes is indicated to determine the generation of a latitude cline in the species (Caicedo et al. 2004). Furthermore, quantitative trait loci (QTL) analysis identified epistatic interactions that resulted in natural phenotypic variation in *Arabidopsis*, *Drosophila* and *Caenorhabditis elegans* (Shook and Johnson 1999; Dilda and Mackay 2002; Ungerer et al. 2002; Weinig et al. 2003). Unfortunately, modern statistical analyses to identify higher-order epistatic interactions are not trivial (Hahn et al. 2003; Blanc et al. 2006; Millstein et al. 2006) and there are numerous fundamental methodological issues that need to be addressed before significant gains in genetic modeling of epistasis are realized (Solomon et al. 2007).

Detection of gene x gene interactions can be accomplished using parametric or non-parametric methods. Parametric methods use a genetic model to describe the genetic effects of each marker on the measured phenotype which provides information on the mean and variance components of the trait used in statistical inference of the genetic model parameters. Importantly, when the assumptions of linear modeling are violated (as is often the case in real world data), careful analysis is warranted to determine whether these violations compromise the validity of the parametric test and their results.

Non-parametric methods utilize data mining approaches in the detection of disease susceptibility genes involved in epistatic interactions (Solomon et al. 2007). Data mining has been defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Bradley 1968). In non-parametric methods, the genetic model is not specified a priori as it is not known (Solomon et al. 2007), instead the model is determined from the data. Therefore, non-parametric methods are more powerful if the genetic model is unknown and they are less burdened by assumptions about the parameter in the probability density function (Solomon et al. 2007). Recent interest in non-parametric methods of data analysis, for use in discovering gene x gene

interactions, has generated much interest, due in part, to the increased flexibility and their ability to handle high-dimensional data such as SNP (Solomon et al. 2007). By investigating novel modeling paradigms, it is hoped that we can better understand and model epistasis within a quantitative genetics framework.

Recent advances adaptive regression methodologies have been developed to explore high-order interactions in genomic data. One such technique, logic regression (Kooperberg et al. 2001; Ruczinski et al. 2003; Ruczinski et al. 2004), utilizes a simulated annealing algorithm in a data reduction framework to identify statistical models for binary data sets. Logic regression constructs models consisting of Boolean combinations of binary covariates (Ruczinski et al. 2004). With $X_1 \ldots X_k$ as binary predictors and Y as the response, logic regression will fit regression models in the form $g(E[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j$, where L_j is a Boolean expression of the predictors X_i (Ruczinski et al. 2004). These are collectively called logic models. In evaluating models of varying sizes, logic regression identifies signal vs. noise in the data set. In statistical modeling, *signal* is identified by asking whether the slope (b) is equal to zero or not equal to zero (Ruczinski 2007). Signal is when X is associated with Y. When additional covariates not associated with Y are added to the model, this is considered *noise* (Ruczinski 2007). By evaluating models of various sizes for signal vs. noise, researchers can determine the level of over-fitting (noise) in each model-size class. In addition, potentially troublesome data sets where there are unacceptable levels of noise are quickly identified so that no further time is wasted in the analysis of these problematic data.

Logic regression was designed as a tree-based Boolean expression search and reduction algorithm, which constructs logic models where the binary predictors are modeled using a subset of permissible rules (moves) with logical operators (Ruczinski et al. 2004). Terminal *knots* (locations of elements in logic trees), which are not further

subdivided, are called leaves (Ruczinski et al. 2004). Logic regression offers numerous scoring functions for linear regression (residual sums of square), logistic regression (deviance), classification (misclassification) and proportional hazards models (partial likelihood) (Ruczinski et al. 2004). In addition, the statistical procedure allows for inclusion of binary or non-binary additive predictors in the model. Furthermore, by creating statistical models consisting of Boolean combinations of binary covariates, this methodology shows promise in the identification of dominant forms of epistatic interactions between molecular markers. Although logic regression has been identified as a parametric method (Solomon et al. 2007), one can argue that it is a combined approach, which utilizes non-parametric, logic trees in a linear regression framework. This small, but important distinction may offer the 'best of both worlds' approach for identifying gene x gene interaction, which will be discussed later in this paper.

Epistasis among the alleles at the *VRN-H1*, *VRN-H2* and *VRN-H3* loci is the hypothesized determinant for vernalization-sensitivity in cultivated barley (*Hordeum vulgare* subsp. *vulgare*) (Takahashi and Yasuda 1971). There is no allelic variation at *VRN-H3* in most cultivated barley genotypes, reducing the genetic model to a two-locus epistatic model (Takahashi and Yasuda 1971). *VRN-H2* encodes a dominant flowering repressor (ZCCT-H) down-regulated by vernalization (Yan et al. 2004). *VRN-H1* is a MADS-box floral meristem identity gene (*HvBM5A*) (Danyluk et al. 2003; Yan et al. 2003) and large deletions within the first intron result in a dominant *VRN-H1* allele and spring growth habit (Fu et al. 2005; von Zitzewitz et al. 2005). A molecular model has been recently proposed to explain the *VRN-H2/VRN-H1* epistatic interaction where dominant *VRN-H2* inhibits the expression of recessive *VRN-H1* alleles (Yan et al. 2004). Based on this model, genotypes with *VRN-H2_/vrn-H1vrn-H1*/*vrn-H3vrn-H3* allelic architecture flower late in the absence of vernalization (vernalization-sensitive) and all other allelic configurations lead to a lack of significant vernalization-sensitivity. This

well-validated epistatic interaction (Sz cs et al. 2007) was used as a model system to test the ability of logic regression in identifying epistasis in binary molecular data.

The objective of this work was to determine if logic regression can be used to identify the interaction between molecular markers associated with the days to flowering phenotype in barley with little or no spurious associations by comparing logic regression with traditional linear-modeling techniques. In addition, we wanted to determine logic regression's capabilities at identifying spurious associations using a linkage decay series.

Materials and methods

Plant material, phenotype and data set

'Dicktoo' (*vrn-H2vrn-H2*/*vrn-H1vrn-H1*), 'Calicuchima'-sib (*Vrn-H2Vrn-H2*/*Vrn-H1Vrn-H1*) and the 'Oregon Wolf Barley Dominant' genetic stock (hereafter referred to as 'OWB-D') (*Vrn-H2Vrn-H2*/*Vrn-H1Vrn-H1*) are vernalization-insensitive barley genotypes (Sz cs et al. 2007). 'Dicktoo' was crossed with 'Calicuchima'-sib and 'OWB-D' and two F_2 populations were established (Sz cs et al. 2007). Flowering time was measured for all unvernalized F_2 plants grown under long-day greenhouse conditions with supplemental lighting and constant temperature (Sz cs et al. 2007). Previously reported gene-specific primers were used to assign *VRN-H2* and *VRN-H1* allele-types for each F_2 individual (Sz cs et al. 2007). We sequenced the recently cloned *VRN-H3* gene (Yan et al. 2006) from the three parents and confirmed that 'Calicuchima'-sib (EU007825), 'Dicktoo' (EU007827), and 'OWB-D' (EU007829) contains the recessive allele.

VRN-H1 and *VRN-H2* molecular markers were coded as binary. Heterozygotes were bulked with the homozygous dominants and scored as 1 while the homozygous

recessives were scored as 0. In addition to the actual molecular markers, *VRN-H1* and *VRN-H2*, we created 100 randomized binary markers (simulated data) for a total of 102 binary markers.

Logic regression analysis

To test the null hypothesis that logic regression cannot identify epistasis amongst vernalization in barley, the datasets from the two F_2 populations were modeled with logic regression in the statistical software R (R Foundation for Statistical Computing, Vienna, Austria) using the linear regression scoring function (Kooperberg and Ruczinski 2005). Day's to flowering was used as the continuous response variable and the molecular marker data were used as binary predictors. Initially, logic regression was allowed to choose the high and low temperatures for the simulated annealing algorithm using a single-fit selection with one tree. Once the program chose the annealing algorithm parameters, the high and low temperatures were optimized according to the author's instructions (Kooperberg and Ruczinski 2005) for selection of a single-fit model. After analyzing the single-fit model data, multiple-fit model selection was performed for use in model selection. When the results warranted further investigation, we performed null model tests to test for statistical signal vs. noise in the data. Upon verification of a strong statistical signal with little noise, we ran a cross-validation test to identify the logic trees with the best predictive capability. In the final step, permutation tests were run to confirm the results of the search algorithm so that we could positively identify the best model that describes the association between predictors and response.

TASSEL analysis

The two F_2 datasets were analyzed using the association mapping software TASSEL-GLM (Trait Analysis by Association Evolution and Linkage) (Zhang et al. 2006). The binary-coded, two vernalization markers and the 100 randomly generated markers were imported into TASSEL along with the phenotypic data. A population structure Q-matrix was designed with all 1's to suggest a single population for our data effectively removing that predictor from the model. The general linear model function was selected for analysis.

SAS-GLM

Analysis of variance was performed on both F_2 datasets using the general linear model (GLM) of SAS Version 9.1 (SAS Institute, Cary, NC). The individual markers which were identified as being significantly associated with the phenotype in TASSEL were analyzed in SAS using a type III fixed effects model analysis to confirm the single marker association results in TASSEL. A type III fixed effects full model containing all the significantly associated markers was performed in SAS to identify marker interactions.

Linkage decay data

Spurious associations between trait and randomly generated markers were tested with logic regression using a linkage decay series to determine the point at which logic regression could no longer make valid associations between truly linked markers and random noise. Two randomly generated sets of linkage decay markers were created each set based upon one of the F_2 populations in our study. Both sets of linkage decay markers had 90%, 80%, 70%... 0% similarity to *VRN-H1*. Our goal was to create randomly

generated markers that would decay in a predictable pattern as the signal in the data became progressively weaker as the similarity to the original vernalization marker decreased (Fig. 1). The decay series data was created by randomly changing 10% of the 1's to zeros and, thereby, using this *new* linkage decay marker as the basis for creating the next marker in the decay series. Original *VRN-H1* and *VRN-H2* markers were removed from the analysis as they interfered with the analysis of the decay series due to their strength of association with the phenotype. This procedure created a linkage decay series where the model association became progressively weaker as the linkage to the phenotype decayed resulting in a smooth logarithmic response (Fig. 1).

Results

VRN-H1/VRN-H2 model selection

Logic regression correctly identified the genetic model explaining the epistatic interaction of the vernalization alleles in data sets for both crosses. The search resulted in a model with a score (residual sums of square) of 12.47 and the equation [+74.9 * (*VRN-H2* and (not *VRN-H1*))] for the in the 'Dicktoo' x 'Calicuchima'-sib data and a model score 8.83 and the equations [+85.4 * (*VRNH2* and (not *VRNH1*))] for the 'Dicktoo' x 'OWB-D' data. The single-fit model searches were repeated 100 times and it was found that the scores (Fig. 1) and the coefficients of the selected models never changed. The conjugate form of the model for each data set, e.g. [-85.4 * (*VRNH1* or (not *VRNH2*))], were also chosen during the search, however the conjugate forms of the models are equal.

In addition, the null model tests suggested that there was a strong signal in the data with very little noise because 0% of the model scores were better than the best score (Supplementary Table 1 online). The cross-validation and the 1000-randomization

permutation tests on the multiple-fit model analyses for the two data sets confirmed the results of the single-fit model search. Tests indicated the optimum model to be model two with one tree and two leaves as it had the lowest cross-validation test average (Fig. 2**)**. Further, the permutation tests on the two data sets identified the same model with one tree and two leaves as being the optimum sized and correct model for the data set as that was the point where the mean of the randomization scores stopped decreasing as the model size increased (Supplementary Table 2 online).

TASSEL analysis

TASSEL-GLM results showed *VRN-H1* and *VRN-H2* as being associated with the days to flowering phenotype in the 'Dicktoo' x 'Calicuchima'-sib and 'Dicktoo' x 'OWB-D' data (Table 1). TASSEL-GLM also identified the randomly generated marker RANDOM 70 as being associated with the days to flowering phenotype in the 'Dicktoo' x 'Calicuchima'-sib data and randomly generated markers RANDOM 46 and RANDOM 58 as being associated with the phenotype in the 'Dicktoo' x 'OWB-D' data (Table 1).

SAS general linear model analysis of variance

The type III fixed effects full model for the 'Dicktoo' x 'Calicuchima'-sib data revealed a significant interaction between *VRN-H1* and *VRN-H2*, but there were no significant singular effects or interactions with the randomly generated marker RANDOM 70 (Table 2). The type III fixed effects full model for the 'Dicktoo' x 'OWB-D' data revealed a significant interaction between *VRN-H1* and *VRN-H2*, but there were no significant singular effects with either marker RANDOM 46 or marker RANDOM 58 (Table 3). A spurious interaction between *VRN-H1* and RANDOM 58 was identified using the Proc

GLM procedure in SAS (Table 3).

Linkage decay

Linkage decay results for the two data sets showed they were quite different in how they responded in a controlled decay simulation. The 'Dicktoo' x 'Calicuchima'-sib data showed less overall variation in single-fit model scores when compared with the 'Dicktoo' x 'OWB-D' data (Fig. 1 and 3). Closer examination of the 'Dicktoo' x 'Calicuchima'-sib data revealed a large increase in variation (CV) within the single-fit model selection scores when linkage decay reached 40% similar to *VRN-H1* (Fig. 3), which corresponded where logic regression could no longer distinguish between linkage decay markers and the simulated markers. Also, there were large variations in the single-fit model scores for *VRN-H2* over multiple runs, which resulted in extremely large CVs (Fig. 3).

Stable single-fit regression model was *only* identified when both the markers appeared in the data set (Fig. 1). Furthermore, large increases in the CV were observed when *VRN-H1* was modeled in the linkage decay series (Fig. 3). The 'Dicktoo' x 'Calicuchima'-sib data up to 40% similar to *VRN-H1* (the point where logic regression could no longer distinguish between decay and dummy markers) had CVs of less than 6% (Fig. 3).

Discussion

Logic regression correctly identified the associated vernalization markers and what's more remarkable is that the model explained the epistasis as a dominant/recessive interaction. Vernalization response in barley is hypothesized to be an interaction where

dominant *VRN-H2* inhibits the expression of recessive *VRN-H1* alleles (Yan et al. 2004). Because a dominant/suppression form of epistasis has been hypothesized to govern vernalization response in barley, we suggest that logic regression correctly identified this proposed genetic model. Numerous QTL and genic studies have identified multiple loci involved in the expression of a single trait (Calborg and Haley 2004). In addition, interval mapping and composite interval mapping have been used successfully to identify QTL associated with specific phenotypes that led to the identification of statistically significant interactions (Shook and Johnson 1999; Lefebvre et al 2003; Ma et al. 2006). However, in all these cases, additional studies were required to ascertain the actual genetic model defining the interaction of the various loci. The use of logic regression appears to address both problems simultaneously.

Identification of epistatic interaction using linear modeling refers to a deviation from additive effects of alleles at different loci as a contribution to the quantitative phenotype (Fischer 1918). This definition leads one to conclude whatever is not additive is therefore epistatic. In addition, it has been reported that linear regression methods are ineffective in pure epistatic models, which have simple main effects and antagonistic epistasis with zero marginal main effects (Solomon et al. 2007). Therefore, traditional linear modeling has limitations in its ability to identify some forms of genetic interaction. Furthermore, there are often difficulties separating out the epistatic variances from additive and dominance (Yu and Bernardo 2004) making estimates of epistatic variances difficult to the point that it was suggested that a non-significant epistatic variance does not suggest an absence of epistasis (Yu and Bernardo 2004). This can be especially true when trying to identify potential epistasis in a large outcrossing population.

Epistasis was originally described as a masking effect where a ‘variant’ (allele) at one locus prevents the variant at the other locus from manifesting an effect (Bateson 1909). Cordell (2002) suggests epistasis has been confused by the fact that we have two

uniquely different definitions for epistasis and unfortunately, there is no correspondence between biological models and those that are statistically motivated (Cordell 2002), or is there? Although logic regression has been reported to be a parametric method of modeling (Solomon et al. 2007), it uses tree-based logic methods in its analysis. These tree-based methods are ideally suited for identification of dominant forms of epistasis. Indeed, our results clearly show that logic regression outperformed traditional linear methods in the simple main effect/synergistic epistasis (Solomon et al. 2007) found in barley vernalization response. In addition, to our surprise, logic regression also identified the precise genetic model explaining the vernalization response in barley. Therefore, although Cordell (2002) suggested there is no correspondence between biological models and those that are statistically motivated, we provide evidence showing logic regression can indeed bridge that gap. Remember, in a diallelic system, there are four ways in which the alleles can interact: dominance x dominance, additive x additive, dominance x additive and additive x dominance. This suggests that in certain instances, linear modeling falls short in the identification of epistasis while in other instances non-parametric methods fall short. In fact, we are told that there is currently no single method that is recognized as the 'best' for detecting, characterizing and interpreting gene x gene interactions and suggest that real breakthroughs will be realized when combined methodologies are used (Solomon et al. 2007). Our analysis suggests logic regression to be a combined system, which bridges the gap between parametric and non-parametric methods.

Although logic regression appears to have many strong points, there are some limitations inherent in the program. First, the program handles binary data for markers. This means identified interactions are limited to those with dominant interaction. Another limitation is that logic regression is a fixed-effects model and therefore is limited in its usefulness in estimating QTL effect. Until recently, there was no method for deriving

additive genetic effects from dominant markers (Hardy 2003). Therefore, there was no way to reliably estimate the effect of a QTL identified using dominant markers like AFLP. With the recent advances in *F*-statistic theory (Hardy 2003) and Bayesian inference (Holsinger et al. 2002) for use with dominant markers, it is now possible to use dominant-scored DNA data sets such as AFLP in a mixed-model analysis.

New modeling strategies utilizing Bayesian model choice (Yi et al. 2003; Xu 2007) and connected designs (Blanc et al. 2006) search for interactive quantitative trait loci have been proposed for mapping epistasis. These new methods show promise in analyzing QTL data and interactions by BLUP (Best Linear Unbiased Prediction) in random effects models to estimate epistatic effects. Unfortunately, Bayesian model choice was limited to just the estimation of the epistatic effects and was not used for variable selection (Xu 2007). However, the model was used to derive an accurate BLUP for the epistatic effect, which can be used, ultimately, to derive the proportion of phenotypic variance explained by an effect of a QTL (h^2). The connected design, although promising, is computationally complex involving numerous 'connected' models with the final results going through a FDA (false discovery rate; Storey and Tibshirani 2003) analysis for significance testing (Blanc et al. 2006). Although these methods are extensible for handling complex models containing additive-by-dominance, dominance-by-additive and dominance-by-dominance interactions, they cannot yet identify the precise genetic model governing epistasis, which includes the dominant/recessive relationships among alleles. In addition, it is unclear how many potentially useful interactions might be missed because of Type I and Type II errors resulting from the limitations of linear modeling. In contrast, logic regression has shown promise in elucidating the precise genetic model through the use of Boolean logic groups with conjugate (recessive) forms of the markers. Recently, logic regression was shown to outperform mixed and other forms of modeling in simulated data trials (Ruczinski et al.

2003). However, it remains unclear as to whether Bayesian model choice and connected designs outperform the combined use of logic trees and data reduction via the simulated annealing algorithm.

Recently, two new Bayesian logic regression packages have been developed for use with high-dimensional SNP data (Clark et al. 2007; Fritsch and Ickstadt 2007). These methods show promise in their ability to use prior information in the logic regression framework. The method by Clark et al. (2007) allows for the inclusion of population genetics information in the model, unfortunately, this method allows for only binary response data and the search algorithm is constrained to logic trees consistent with perfect phylogeny. The full Bayesian method developed by Fritsch and Ickstadt (2007) also has some limitations. Their version allows for only the inclusion of binary response and binary predictors with model selection being limited to those models where SNP interactions fall within known biological pathways. However, it should be possible to construct a full Bayesian logic regression which would work for continuous response variables with the inclusion of non-binary predictors such as population structure and kinship coefficient (Q-matrix and K-matrix) data. This would extend the use of this powerful technique to traits shown to be continuous and would also take population stratification and relatedness into account. It is widely accepted that population structure/stratification and relatedness among individuals can lead to spurious associations (associations without linkage) between a candidate marker and phenotype (Lander and Schork 1994; Bacanu et al. 2000; Pritchard et al. 2000a,b; Devlin et al. 2001; Yu et al. 2006) and that this type of structure needs to be accounted for in the data analysis.

Modeling linkage decay helped demonstrate the power of logic regression to accurately model data sets where linkage between markers may be incomplete or spurious. The coefficient of variation (CV) is a dimensionless value used to quantify

uncontrolled experimental error. The coefficient of variation results suggest there were measurable differences between the two crosses, however, this data alone does not provide any diagnostic information about acceptable levels of variation within the data sets. However, when the CV data was compared with TASSEL-GLM (Table 1) and SAS-GLM (Table 3) output, we discovered that the CV might be diagnostic in the identification of potentially troublesome data sets. The substantial increase the CV at 40% similar to *VRN-H1* (Fig. 3) was the precise point in the decay series where logic regression could no longer differentiate between the decay-marker and randomly generated markers. This jump in CV within the decay series suggests there may be a limit in predictive capability. Our results suggest the 'limit of predictability' threshold may be where there this substantial increase in CV was observed (Fig. 3). Comparing CV values with the results from the SAS-GLM suggest CVs above 6% may result in the modeling of spurious associations.

Because, SAS-GLM identified a spurious interaction in the 'Dicktoo' x 'OWB-D' data set with both vernalization markers present (Table 3), suggested the data set may be problematic right from the start due to noise. The ANOVA suggested noise (by identifying a spurious interaction) and the CV analysis on the single-fit model scores suggest variation above 6% may lead to spurious association. In support of the hypothesized modeling limit, it was reported when there are large variations in single-fit model scores during initial model identification, there may be problems with the data set (Kooperberg et al. 2001). Unfortunately, it's unclear where that cutoff might be. This was a concern for us and it became one of the major reasons for performing the linkage decay series. Based on our results, we suggest any data set that has a single-fit model selection CV of 6% or less should prove reliable and identify real associations.

Our results suggests logic regression works in accurate identification of epistatic interaction and that the model building algorithm appears to be more robust and accurate

when compared with traditional general linear modeling in QTL analysis. From a theoretical point of view, logic regression may use the more appropriate approach for modeling epistasis by forming logic groups as a means of identifying marker interaction independent of linear model assumptions and rules. This suggests logic regression to be a combined non-parametric/parametric approach to modeling epistasis, which has been suggested before real gains in epistasis research are realized. Advances in QTL analysis related to logic regression may prove cost effective by creating a useful random effects models for dominant DNA data which would identify many forms of epistasis. By investigating Boolean logic's utility in high-level mixed and/or Bayesian models, definitive statements on the usefulness of logic groups in quantitative genetic modeling can be made. Regardless of these *potential* future applications, it does appear that logic regression is a useful tool in data mining applications[28] and provides researchers with a complement to traditional QTL identification.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Patrick M. Hayes for his critical suggestions on the data analysis and manuscript preparation. The authors would also like to thank Johns Hopkins University statisticians Dr. Ingo Rusczinski, Dr. Charles Kooperburg and Dr. Mel LeBlanc for making logic regression freely available for use in R statistics.

REFERENCES

Bacanu S-A, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum

Genet 66: 1933-1944

Bateson W (1909) Mendel's principles of heredity. Cambridge University Press, Cambridge

Blanc G, Charcosset A, Mangin B, Gallais A, Moreau L (2006) Connecting populations For detecting quantitative trait loci and testing for epistasis: an application in maize. Theor Appl Genet 113: 206-224

Bradley JV (1968) Distribution free statistical tests. Engelwood Cliffs, NJ, p 15

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitude cline in a life history trait. Proc Natl Acad Sci USA 101: 15670-15675

Carlborg Ö, Haley CS (2004) Epistasis: too often neglected in complex trait studies. Nature 5: 618-625

Clark TG, De Iorio M, Griffiths RC (2007) Bayesian logistic regression using a perfect phylogeny. Biostatistics 8: 32-52

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11: 2463-2468

Danyluk J, Kane ND, Breton G, Limin AE, Fowler DB, Sarhan F (2003) *TaVRT-1*, a putative transcription factor associated with vegetative to reproductive transition in cereals. Pl Physiol 132: 1849–1860

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Pop Biol 60: 155-166

Dilda CL, Mackay TFC (2002) The genetic architecture of *Drosophila* sensory bristle number. Genetics 162: 1655-1674

Fischer RA (1918) The correlation between relatives on the supposition of Mendelian Inheritance. Trans R Soc Edin 52: 399-433

Fritsch A, Ickstadt K (2007) Comparing logic regression based methods for identifying SNP interactions. In: Hochreiter S, Wagner R (eds) LNBI 4414. Springer-Verlag, Berlin Heidelberg, pp 90-103

Fu D, Sz cs P, Yan L, Helguera M, Skinner JS, von Zitzewitz J, Hayes PM, Dubcovsky J (2005) Large deletions within the first intron in *VRN-1* are associated with spring growth habit in barley and wheat. Mol Genet Genom 273:54–65

Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene x gene and gene x environment interactions. Bioinformatics 19: 376-382

Hardy OJ (2003) Estimation between individuals and characterization of isolation-by-distance processes using dominant genetic markers. Mol Ecology 12: 1577-1588

Holsinger KE, Lewis PO, Dey DK (2002) A Bayesian approach to inferring population structure for dominant markers. Mol Ecology 11:1157-1164

Kooperberg C, Bis JC, Marciante KD, Heckbert SR, Lumley T, Psaty BM (2007) Logic Regression for analysis of the association between genetic variation in the rennin-angiotensin system and myocardial infarction or stroke. Am J Epidemiol 165: 334-343

Kooperberg C, Ruczinski I (2005) The logic regression package. In: Contributed Packages. R Project. http://cran.rproject.org/src/contrib/PACKAGES.html

Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L (2001) Sequence analysis using logic regression. Genetic Epidemiology 21 (Suppl. 1): S626-S631

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037-2048

Lefebvre V, Daubèze A–M, Rouppe van der Voort J, Peleman J, Bardin M, Palloix A (2003) QTLs for resistance to powdery mildew in pepper under natural and artificial infections. Theor Appl Genet 107: 661-666

Ma H–X, Bai G–H, Zhang X, Lu W–Z (2006) Main effects, epistasis, and environmental interactions of quantitative trait loci for Fusarium head blight resistance in a recombinant inbred population. Phytopath 96: 534-541

Millstein J, Conti DV, Gililand FD, Gauderman WJ (2006) A testing framework for identifying susceptibility genes in the presence of epistasis. Am J of Hum Genet 78: 15-27

Orr HA (1995) The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139: 1805-1813

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

Pritchard JK, Stevens M, Donelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67: 170-181

Ruczinski I, Kooperberg C, LeBlanc M (2003) Logic regression. J Comput Graph Stat 12:

475–511

Ruczinski I, Kooperburg C, LeBlanc ML (2004) Exploring interactions in high-dimensional genomic data: an overview of logic regression with applications. J Multi Analysis 90: 178-195

Ruczinski, I (2007) Personal communication

Shook DR, Johnson TE (1999) Quantitative trait loci affecting survival and fertility related traits in *Caenorhabditis elegans* show genotype-environment interactions, pleiotropy and epistasis. Genetics 153: 1233-1243

Solomon KM, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari H, Allison D (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered 63: 67-84

Storey JD, Tibshirani R (2003) Statistical significance for genome wide studies. Proc

Natl Acad Sci USA 100: 9440-9445

Sz cs P, Skinner JS, Karsai I, Cuesta-Marcos A, Haggard KG, Corey AE, Chen THH, Hayes PM (2007) Validation of the *VRN-H2/VRN-H1* epistatic model in barley reveals that intron length variation in *VRN-H1* may account for the continuum of vernalization sensitivity. Mol Genet Genom 277: 249-261

Takahashi R, Yasuda S (1971) Genetics of earliness and growth habit in barley. In: Nilan RA (ed) Barley Genetics II. Proceedings of the Second International Barley Genetics Symposium. Washington State University Press Pullman, pp. 388–408

Ungerer MC, Halldorsdottir SS, Modliszewski JL, Mackay TFC and Purugganan MD (2002) Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*. Genetics 160: 1133-1151

von Zitzewitz J, Sz cs P, Dubcovsky J, Yan L, Pecchioni N, Francia E, Casas A, Chen THH, Hayes PM, Skinner JS (2005) Molecular and structural characterization of barley vernalization genes. Plant Mol Biol 59:449–467

Weinig C, Dorn LA, Kane NC, German ZM, Halldorsdottir SS, Ungerer M, Toyonaga Y, Mackay TFC, Purugganan MD, Schmitt J (2003) Heterogenous selection at specific loci in natural environments in *Arabidopsis thaliana*. Genetics 165: 321-329

Wright S (1931) Evolution in Mendelian populations. Genetics 16: 97-159

Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63: 513-521

Yan L, Fu D, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. Proc Natl Acad Sci USA 103: 19581-19586

Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, SanMiguel P, Bennetzen

JL, Echenique V, Dubcovsky J (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science 303: 1640–1644

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci USA 100: 6263–6268

Yi N, Xu S, Allison DB (2003) Bayesian model choice and search strategies for mapping interacting quantitative trait loci. Genetics 165: 867-883

Yu J, Bernardo R (2004) Changes in Genetic variance during advanced cycle breeding in maize. Crop Sci 44: 405-410

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203-208

Zhang Z, Bradbury PJ, Kroon DE, Casstevens TM, Buckler ES (2006) TASSEL 2.0: a software package for association and diversity analyses in plants and animals, Plant & Animal Genomes XIV Conference

Figure 1 The relationship between model score and linkage decay for the ‘Dicktoo’ x ‘Calicuchima’-sib and ‘Dicktoo’ x ‘OWB-D’ data. The figure shows data for when both vernalization markers are present and modeled, *VRN-H2* alone and modeled and the linkage decay data for *VRN-H1* starting at 100% *VRN-H1* and regressing to 0% similar to *VRN-H1*. RSS is the residual sums of square with error bars indicating the standard deviation from the mean for 100 replicated single-fit model searches. VH1/VH2=Both vernalization markers present, (%)VH1=(%)*VRN-H1* alone and (%)VH2=(%)*VRN-H2* alone. Percentages in front of VH1 indicate percent similarity to the original decay marker indicating the level of introduced randomness at any given point in the decay series.

Figure 2 Cross-validation plots showing the test scores for models with one logic tree and with one to five leaves (model size) on the x-axis. The 'Dicktoo' x 'Calicuchima'-sib are data represented by black circles and the 'Dicktoo' x 'OWB-D' data represented by open circles. Models with the smallest test set RSS have the best predictive performance[35] and when there are ties in the score, we condition on the smallest model within the group. The cross-validation test is used to determine the logic tree with the best predictive capability by assessing how well the best model of size *k* performs in comparison to other size models[15]. The data set is divided into *m* (approximately) equal sized groups of cases[15]. For each of the *m* groups of cases, the ith groups are removed[15]. Then, the best scoring model of size *k* is found using only (*m*-1) groups[15]. The cases within group *i* are all scored under this model which yields a score ε_{ki} [15]. The cross-validated test score for model size *k* equals $\varepsilon_k = (1/m)\sum_i \varepsilon_{ki}$. The cross-validated scores for the model-size classes are then compared to determine the model with the best predictive capability.

Figure 3 The coefficients of variation (CV=s/μ *100, where s = standard deviation and μ = mean) for 100 single-fit model scores shown in **Figure 1**. VH1/VH2=Both vernalization markers present, (%)VH1=(%)*VRN-H1* alone and (%)VH2=(%)*VRN-H2* alone. Percentages in front of VH1 indicate percent similarity to the original decay marker indicating the level of introduced randomness at any given point in the decay series.

Table 1 Markers associated with the days to flowering phenotype, using TASSEL-GLM.

Source of Variation	DF	Mean Square	*F*-value	*p*-value	R^2
'Dicktoo' x 'Calicuchima'-sib$^{\psi}$					
VRN-H1	1	43312.83	87.11	6.33E-15***	0.49
VRN-H2	1	10188.23	11.83	8.80E-04***	0.12
RANDOM 70	1	6834.69	7.61	0.0070**	0.08
Error	91				
'Dicktoo' x 'OWB-D'$^{\psi}$					
VRN-H1	1	87543.90	209.33	2.53E-25***	0.70
VRN-H2	1	16319.55	13.59	3.86E-04***	0.13
RANDOM 58	1	16155.45	12.24	7.27E-04***	0.12
RANDOM 46	1	5061.18	4.16	0.04*	0.04
Error	91				

ψ data obtained from Sz cs, P. *et al.*[24]

p-values and R^2 values for molecular markers identified by TASSEL-GLM.
*Significant at the 0.05 level, ** Significant at the 0.01 level,
*** Significant at the 0.001 level

Table 2 Analysis of variance results for the 'Dicktoo' x 'Calicuchima'-sib data full model for markers found to be associated with the days to flowering phenotype in TASSEL. Results are from a type III fixed effects model showing corresponding *p*-values.

Source of Variation	DF	Type III SS	Mean Square	*F*-value	*p*-value
VRN-H1	1	14374.4	14374.4	101.4	<0.001**
VRN-H2	1	14615.0	14615.0	103.1	<0.001**
RANDOM 70	1	83.3	83.3	0.6	0.445(NS)
*VRN-H1***VRN-H2*	1	6367.3	6367.3	44.9	<0.001**
*VRN-H1**RANDOM 70	1	200.1	200.1	1.4	0.238(NS)
*VRN-H2**RANDOM 70	1	1.4	1.4	0.01	0.921(NS)
*VRN-H1***VRN-H2**RANDOM 70	1	26.3	26.3	0.2	0.668(NS)
Error	85				

** Significant at the 0.001 level

Table 3 Analysis of variance results for the 'Dicktoo' x 'OWB-D' data full model for markers found to be associated with the days to flowering phenotype in TASSEL. Results are from a type III fixed effects model showing corresponding *p*-values.

Source of Variation	DF	Type III SS	Mean Square	*F*-value	*p*-value
VRN-H1	1	9236.0	9236.0	136.42	<0.001**
VRN-H2	1	8188.2	8188.2	120.94	<0.001**
RANDOM 46	1	53.0	53.0	0.78	0.3799(NS)
RANDOM 58	1	169.4	169.4	2.50	0.118(NS)
*VRN-H1***VRN-H2*	1	5705.0	5705.0	84.26	<0.001**
*VRN-H1**RANDOM 46	1	23.8	23.8	0.35	0.555(NS)
*VRN-H1**RANDOM 58	1	301.1	301.1	4.45	0.038*
*VRN-H2**RANDOM 46	1	10.6	10.6	0.16	0.694(NS)

*VRN-H2**RANDOM 58	1	208.9	208.9	3.09	0.083(NS)
*VRN-H1*VRN-H2**RANDOM 46	1	17.9	17.9	0.26	0.609(NS)
*VRN-H1*VRN-H2**RANDOM 58	1	185.6	185.6	2.74	0.1017(NS)
Error	81				

*Significant at the 0.05 level, ** Significant at the 0.001 level