A GENERALIZED APPROACH AND COMPUTER TOOL FOR QUANTITATIVE GENETICS STUDY

Jixiang Wu

Johnie N. Jenkins

Jack C. McCarty

A Generalized Approach and Computer Tool for Quantitative Genetics Study

Jixiang Wu*, Johnie N. Jenkins, and Jack C. McCarty

J. Wu: Plant Science Department, South Dakota State University, Box 2140C, Brookings, SD 57007; * Corresponding author; J. N. Jenkins and J. C. McCarty: Crop Science Research Laboratory, USDA-ARS, Mississippi State, Box 5367, MS 39762

Abstract

Quantitative genetics is one of the most important components to provide valuable genetic information for improving production and quality of plants and animals. The research history of quantitative genetics study could be traced back more than one hundred years. Since the Analysis of Variance (ANOVA) methods were proposed by Fisher in 1925, several useful genetic models have been proposed and have been widely applied in both plant and animal quantitative genetics studies. Useful examples included various North Carolina (NC) and diallel cross mating designs. However, many genetic models derived from these mating designs are ANOVA method based, so there are several major limitations. For example, ANOVA based methods are constricted to simple genetic models and specific mating designs and require balanced data structures. Though mixed linear model approaches were proposed in the 1960s, their applications in quantitative genetics study were limited until the early 1990s. The advantages of the mixed linear model approaches include the flexibility for unbalanced genetic data structures and complex genetic model systems. In the past years the mixed linear models have been applied to analyze various useful genetic models and a number of computer programs have been developed. In addition, researchers are not only interested in finding appropriate data structures needed for specific genetic models but also want to identify appropriate genetic models suitable for a specific data structure. Therefore, a generalized computer tool has been developed for both model evaluations and actual data analyses. In this paper, various genetic models will be detailed and generalized by mixed linear model approaches and the features of the new computer tool GenMod will be described.

1. Introduction

Since an analysis of variance (ANOVA) approach was proposed (Fisher, 1925), geneticists have been extensively using this approach for quantitative genetic data analyses because of its convenience and simplicity (i.e. Garder and Eberhart, 1966; Borges, 1987;

Haullauer and Miranda, 1988; Das and Griffley, 1994; Lynch and Walsh, 1998). However, ANOVA based approaches are often associated with several major limitations. For example, these methods are often challenged by various irregular genetic mating designs and unbalanced data structures. It is often difficult to follow specific genetic mating designs (Comstock and Robison, 1948; Griffing, 1956) when a large number of parents are used for crossing either due to flowering differences or resource constraints (i.e. Cheatham et al., 2003; Saha et al., 2006). In addition, insect damage and/or environmental conditions could contribute to data missing or data structures being unbalanced. It is also very common that F_2 (second generation) populations are used to replace F_1 (first generation) populations because of F_1 seed supply (i.e. Meredith, 1990; Tang 1996; Jenkins et al., 2006, 2007, 2009); however, the genetic structures for F_2 are different for F_1. Furthermore, genetic model structures can be very complex. For example, some seed traits may be controlled by gene systems in seeds and their maternal plants because maternal plants provide nutrition to seed growth and development (i.e. Zhu and Weir, 1994a,b; Wang et al., 1996a,b; Wu et al., 2010). Thus, genetic data containing important genetic information can be underscored and underutilized if inappropriate statistical methods or genetic models are used.

Since the 1960s, mixed linear model approaches have been proposed and can be used for unbalanced data structures and complex models (i.e. Hartley and Rao, 1967; Patterson and Thompson, 1971; Rao, 1971; Searle et al., 1992; Little et al., 1996; Zhu, 1998). These approaches are matrix- and vector- based approaches, which offer flexibility to analyze complex genetic models and/or data structures. For example, procedure mixed in recent SAS versions can be used not only for missing data but also for various repeated measurements. Typically, there are three types of mixed linear model approaches: maximum likelihood (ML), restricted maximum likelihood (REML), and minimum norm quadratic unbiased estimation (MINQUE) (Hartley and Rao, 1967; Rao, 1971; Searle et al., 1992). Although these mixed linear model approaches were proposed and employed for many years, the applications to quantitative genetic data analyses have not been widely received until the late 1980s due to their mathematical complexity and computational constraints. Some valuable genetic models were proposed and can be analyzed by mixed linear model approaches. Various crop systems were investigated, including cotton, rice, barley, and canola (i.e. McCarty et al., 2004a,b; Shi et al., 1997; Yan et al., 1998) and covering agronomic traits (i.e. McCarty et al., 2004a,b; Jenkins et al., 2006, 2007), seed traits (Wu et al., 1995, 2010; Wang et al., 1996a,b; Shi et al., 1997), developmental traits

(Ye et al., 2003; McCarty et al., 2006; Wu et al., 2009), and traits for chromosome substitution lines (Saha et al., 2006; Jenkins et al., 2006; 2007; McCarty et al., 2008; Wu et al., 2006a).

In quantitative genetic data analysis, two important questions are often asked. The first question is which genetic models are appropriate for a given genetic data set. The second question is that given a biologically meaningful genetic model being employed, what types of genetic data structures are required. Since genetic data structures to be analyzed or genetic models to be employed for a given data structure are case-specific, a computer program that can specifically evaluate such appropriateness of data structures or genetic models is needed.

In this paper, various commonly used genetic models will be addressed and generalized in terms of vectors and matrices, so that quantitative genetic data analyses can be conducted in a more generalized way. Section 2 will detail various genetic models and their generalization. Mixed linear model approaches for generalized genetic models will be addressed in Section 3. In Section 4, a computer program will be briefly introduced and results from an actual cotton genetic data set in cotton will be summarized as an example. The major objective of this study was to provide a generalized way to analyze various genetic data structures so that useful genetic information can be used for crop and animal improvement.

2. A Generalized Genetic Model

2.1. A simple genetic model

For a number of genotypes grown in multiple environments with repeated plots under a random complete block (RCB) design, a linear genetic model can be expressed as in equation (1):

$$y = \mu + E + G + GE + B(E) + e \quad (1)$$

Where y is an observed value, μ is population mean, E is an environmental effect, G is a genotypic effect, GE is a genotype-by-environment (GE) interaction effect, B(E) is a block effect within environment, and e is a random error. In equation (1), three components are partitionable. For example, E could be year, location, treatment, and their interaction effects. A genotypic effect could include additive, dominance, and epistatic effects (Cockerham, 1980). A GE interaction effect can include various GE interaction effects corresponding to the partitioning of a genotypic effect.

For the purpose of consistency, all of the following genetic models are expressed with genetic effects and their corresponding genotype-by-environment (GE) interaction effects with possible block effects (within environment). If an experiment follows a completely randomized (CR) design, the block effects can be deleted from each genetic model. In addition, if an experiment is only conducted in one environment, the environmental effect and all GE interaction effects should be deleted from the model. The difference in considering fixed and random effects is debatable; however, based on our experience in data analyses we observed there is not much difference in environmental and genetic effects being obtained when they are considered fixed or random. For this reason and for convenience, we may treat all effects as random effects except population mean and environmental effects. Several genetic models are detailed as follows.

2.2. Genotype and genotype-by-environment interaction (GE) model

The observation y_{hij} for ith genotype grown in jth block in hth environment can be expressed as the following linear model in equation (2):

$$y_{hij} = \mu + E_h + G_i + GE_{hi} + B_{j(h)} + e_{hij} \quad (2)$$

where μ is the population mean, E_h is the environmental effect, G_i is genotypic effect, GE_{hi} is the genotype-by-environment interaction effect, $B_{j(h)}$ is the block effect, and e_{hij} is the random error.

2.3.Nested model

In breeding programs, a number of lines are often derived from each of multiple crosses (families) and are evaluated in different environments with repeated plots. In this or similar cases, a nested genetic model can be applied. This model can be applied for evaluation of germplasm lines collected from different regions. The observation y_{hijk} for jth line within ith family grown in kth block within hth environment can be expressed in equation (3):

$$y_{hijk} = \mu + E_h + F_i + L_{j(i)} + FE_{hi} + LE_{hj(i)} + B_{k(h)} + e_{hijk} \quad (3)$$

where F_i is the family effect; $L_{j(i)}$ is the within family line effect; FE_{hi} is the family-by-environment interaction effect, $LE_{hj(i)}$ is the within family line-by-environment interaction effect; $B_{k(h)}$ is the block effect; and e_{hijk} is the random error.

2.4. Additive-dominance (AD) model

The AD model is one of the most popular genetic models. A large number of applications of AD model in quantitative genetic study can be found in the literature (i.e. Tang 1996; Jenkins 2006, 2007, 2009; McCarty et al., 2007). Given a number of parents and their F_1 or F_2 progenies evaluated in multiple environments, the AD genetic model can be expressed in linear form as follows regarding parent i or a cross between parents i and j at different generations.

For parent

$$y_{hiik\,(P)} = \mu + E_h + 2A_i + D_{ii} + 2AE_{hi} + DE_{hii} + B_{k(h)} + e_{hiik} \quad (4)$$

For F_1:

$$y_{hijk\,(F_1)} = \mu + E_h + A_i + A_j + D_{ij} + AE_{hi} + AE_{hj} + DE_{hij} + B_{k(h)} + e_{hijk\,(F_1)} \quad (5)$$

For F_2:

$$y_{hijk\,(F_2)} = \mu + E_h + A_i + A_j + \frac{1}{4}D_{ii} + \frac{1}{4}D_{jj} + \frac{1}{2}D_{ij} + AE_{hi} + AE_{hj} + \frac{1}{4}DE_{hij} + \frac{1}{4}DE_{hij} + \frac{1}{2}DE_{hij} + B_{k(h)} + e_{hijk\,(F_2)} \quad (6)$$

For F3:

$$y_{hijk\,(F_3)} = \mu + E_h + A_i + A_j + \frac{3}{8}D_{ii} + \frac{3}{8}D_{jj} + \frac{1}{4}D_{ij} + AE_{hi} + AE_{hj} + \frac{3}{8}DE_{hii} + \frac{3}{8}DE_{hjj} + \frac{1}{4}DE_{hij} + B_{k(h)} + e_{hijk\,(F_1)} \quad (7)$$

Where μ is the population mean, a fixed effect; E_h is the environment effect, either random or fixed (fixed in this study); A_i (or A_j) is additive effect from parent *i* or *j*; D_{ii}, D_{jj} or D_{ij} is the dominance effect; AE_{hi} (or AE_{hj}) is additive by environment interaction effect; DE_{hii}, DE_{hjj}, or DE_{hij} is the dominance by environment interaction effect; $B_{k(h)}$ is the block effect; and $e_{hijk\,(.)}$ is the random error.

2.5. Additive-dominance with additive-by-additive interaction (ADAA)

The ADAA model is one of the important extended AD models, investigating additive-by-additive interaction (epistatic) effects (Cockerham, 1980; Zhu 1998). Several applications are available in the literature (Xu and Zhu, 1999; McCarty et al., 2004a,b, 2005, 2008, Saha et al., 2010). This model was also evaluated when data structures are unbalanced (Wu et al., 2006a). Additive-by-additive interaction effects can be used for both inbred line and hybrid development (Xu and Zhu, 1999; McCarty et al., 2004a, b). Given a number of parents and their F_1 or F_2 planted in multiple environments, this genetic model can be expressed in linear form as follows regarding parent i or a cross between parents i and j at different generations.

For parent:

$$y_{hiik\,(P)} = \mu + E_h + 2A_i + D_{ii} + 4AA_{ii} + 2AE_{hi} + DE_{hii} + 4AAE_{hii} + B_{k(h)} + e_{hiik} \quad (8)$$

For F_1:

$$y_{hijk\,(F_1)} = \mu + E_h + A_i + A_j + D_{ij} + AA_{ii} + AA_{jj} + 2AA_{ij} + AE_{hi} + AE_{hj} + DE_{hij} + AAE_{hii} + AAE_{hjj} + 2AAE_{hij} + B_{k(h)} + e_{hijk\,(F_1)} \quad (9)$$

For F_2:

$$y_{hijk\,(F_2)} = \mu + E_h + A_i + A_j + \frac{1}{4}D_{ii} + \frac{1}{4}D_{jj} + \frac{1}{2}D_{ij} + AA_{ii} + AA_{jj} + 2AA_{ij} + AE_{hi} + AE_{hj} + \frac{1}{4}DE_{hii} + \frac{1}{4}DE_{hjj} + \frac{1}{2}DE_{hij} + AAE_{hii} + AAE_{hjj} + 2AAE_{hij} + B_{k(h)} + e_{hijk\,(F_2)} \quad (10)$$

For F_3:

$$y_{hijk\,(F_3)} = \mu + E_h + A_i + A_j + \frac{3}{8}D_{ii} + \frac{3}{8}D_{jj} + \frac{1}{4}D_{ij} + AA_{ii} + AA_{jj} + 2AA_{ij} + AE_{hi} + AE_{hj} + \frac{3}{8}DE_{hii} + \frac{3}{8}DE_{hjj} + \frac{1}{4}DE_{hij} + AAE_{hii} + AAE_{hjj} + 2AAE_{hij} + B_{k(h)} + e_{hijk\,(F_3)} \quad (11)$$

Where A_i (or A_j) is the additive effect from parent i (or j); D_{ii}, D_{jj} or D_{ij} is the dominance effect; AA_{ii}, AA_{jj}, or AA_{ij} is the additive-by-additive (AA) epistatic effect; AE_{hi} (or AE_{hj}) is additive-by-environment interaction effect; DE_{hii}, DE_{hjj} or DE_{hij} is the dominance by environment interaction effect; AAE_{hii}, AAE_{hjj}, or AAE_{hij} is the AA-by-environment interaction effect; $B_{k(h)}$ is the block effect; and $e_{hijk\,(.)}$ is the random error.

2.6. Other extended AD models

Genetic modeling is case or data structure specific. For example, genetic systems for agronomic traits could be different from seed traits. Thus, genetic modeling needs to maximally reflect its biological meaning for a trait to be investigated. In addition to AD and ADAA models, other different genetic models have been reported in the literature. Examples include AD model with cytoplasmic effects (ADC model: Wu et al., 2010), AD model with maternal effects (ADM model: Zhu, 1994), seed models (Zhu and Weir 1994a, b; Wu et al., 1995; Wang et al., 1996a, b); AD model with single marker effects (Wu et al., 2000), and a chromosome model (Wu et al., 2006a).

2.7. Genetic model generalization

As we have seen, genetic models can be trait or case dependent. In addition, genetic structures vary at different generations for the same model and data can be missing or unbalanced, which often cause data analyses to be performed on a case by case basis. Thus, it will be helpful to generalize different genetic models in a simple yet practical way: not only for model extension but for data analyses as well. With the use of mixed linear model approaches, these genetic models can be expressed in forms of vectors and matrices described as follows.

$$\boldsymbol{y} = \sum_{i=1}^{f} \boldsymbol{X_i b_i} + \sum_{u=1}^{r} \boldsymbol{U_u e_u} = \boldsymbol{Xb} + \sum_{u=1}^{r} \boldsymbol{U_u e_u} \quad (12)$$

where $\boldsymbol{y}$ is an observed vector with size of $n \times 1$; $\boldsymbol{b}_i$ is an unknown fixed effect vector to be estimated with dimension of $t_i \times 1$ and $\boldsymbol{X}_i$ is the known information for vector $\boldsymbol{b}_i$; $\boldsymbol{e}_u$ is an unknown random effect vector to be calculated with dimension of $s_u \times 1$ and $\boldsymbol{U}_u$ is the known information for vector $\boldsymbol{e}_u$. Note that the last item $\boldsymbol{e}_r$ in equation (12) is random error. When $\boldsymbol{e}_u$ is independently and identically distributed, then $\boldsymbol{U}_r$ is an identical matrix. Since the values of *f* and *r* in equation (12) can be any numbers, it can generalize various genetic models for various data structures and it can generalize computer programming.

3. Statistical Approaches

3.1. Variance component estimation

ANOVA based approaches are challenged by missing data points, irregular genetic mating designs, and/or complex genetic models. On the other hand, mixed linear model approaches offer flexibility for analyzing complex genetic models and various unbalanced data structures. There are three general types of mixed linear model approaches, which can be used for analyzing mixed linear models: maximum likelihood (ML), restricted maximum likelihood (REML), and minimum norm quadratic unbiased estimation (MINQUE) approaches (Hartley and Rao, 1967; Patterson and Thompson, 1971; Rao, 1971; Searle et al., 1992; Zhu, 1998). Both ML and REML approaches require iteration process and assuming data being normally distributed. MINQUE approaches require no iteration process and can be applied to different data distribution (Rao, 1971). Given a reasonable large data set with normal distribution, each variance component can be tested by asymptotic chi-square distribution. However, chi-square test has some limitations: (1) a reasonable size of a data structure for a specific model is difficult to determine; (2) data structures may not follow normal distributions; (3) it may be difficult to test parameters like genetic correlations, genetic covariances, and proportions. An alternative

method is using resampling approaches including jackknife, permutation, and bootstrap tests (Miller, 1974; Efron, 1982; Davison and Hinkley, 1997; Wu et al., 2008). Jackknife methods have been widely used for testing significance of each parameter of interest (Miller, 1974; Wu et al., 2008). The results are equivalent through permuting and bootstrapping residuals; however, bootstrapping observations often result in abnormal variance component estimation and effect estimation or prediction. We observed that group-based jackknife methods are stable for the data set with replications; however, it might be more appropriate to use permutation test for data sets with only single replication or very irregular data sets.

3.2. Genetic effects, heterosis, and genotypic values

Breeders are not only interested in estimated genetic variance components, but genetic effects as well. Predicted genetic effects give information about which parents should be used for crossing or which crosses should be used for selection. If genetic variance components are known, a best linear unbiased prediction (BLUP) for genetic effects can be obtained. However, genetic variance components normally are unknown, estimated variance components are normally used for predicting genetic effects using the BLUP approach. These predicted genetic effects cannot be guaranteed to be linear or best since estimated variance components are quadratic functions of observations. Two other prediction methods, which can result in linear and unbiased predictions, are linear unbiased prediction (LUP) (Zhu and Weir, 1994a) and adjusted unbiased prediction (AUP) (Zhu, 1993) methods. When the genetic effects were predicted subject to different genetic models, heterosis and genotypic values of each cross either over environments or in a specific environment can be calculated as well (i.e. McCarty et al., 2004b, 2005).

4. GenMod: A Generalized Computer Program

4.1.The features of GenMod

Using C++ language, we developed a new computer program GenMod specifically for implementing previously described genetic models. When using the MINQUE approach, prior values for each variance components are required. Our analysis based on simulated and actual data showed that different prior values generate almost identical results. The methods used in this computer program are detailed as follows: MINQUE approach with all prior values being 1,

MINQUE1 approach (Zhu, 1989) for variance component estimation; adjusted unbiased prediction (AUP) method for genetic effect prediction; and two jackknife methods for calculating standard error of each parameter. Though many genetic models can be added to this computer program, only several commonly used genetic models are available. They include genotype model with GE interaction model, nested model, AD model, ADAA model, ADC model, and ADM model. However, other biologically meaningful genetic models and more functions can be easily added to this computer program.

The computer program has the following advantages: (1) it simultaneously conducts analysis for an actual data set or model evaluation for a data set with only experimental design information but no traits included; (2) it can analyze a data set with missing data points, crosses, irregular genetic mating designs; (3) it can analyze data sets where genotypes vary across environments; (4) it is able to analyze data sets for different generations; and (5) it provides a significance test for each parameter with jackknife methods.

For actual data analysis, the program provides estimated variance components, estimated proportional variance components to the phenotypic variances, and predicted genetic effects if a variance component estimate is numerically greater than zero. For simulation studies, the computer program provides the parameter values σ_u^2 (true or preset values of variance components), estimated values $\bar{\sigma}_u^2$, and the respective bias calculated by bias=$\bar{\sigma}_u^2$ - σ_u^2. The statistical testing power is defined in this program as power=1-β, where β is the probability level for type II error at different levels. The mean square error (MSE) for each parameter is calculated by MSE= $bias^2 + \text{var}(\bar{\hat{\sigma}}_u^2)$ and the coefficient of efficiency, $CE = \sqrt{\frac{MSE}{|\sigma_u^2|+|bias|}}$ (Zhu and Weir, 1994a, Wu et al., 2006a, b, 2010). If a preset value is zero, then the power is actually the Type I error at a specific nominal value α. Thus, this computer program can be used to test both Type I error rate and testing power.

4.2. The use of GenMod

Since this computer program is able to analyze different genetic models with different functions, a full user manual will be developed separately. However, the use of this computer program is straightforward and general procedures to run this computer program are briefly described as follows.

Step 1: Prepare a data file.

Different models require different data format. For purpose of demonstration in this paper, a cotton data set (*realf2.txt*) including 12 F_2 populations and their eight parents (two years and six replications for each year) will be analyzed subject to an AD model. Given this genetic model, the first five columns in the example data file (Table 1) are required and represent environment (e.g, year or location), female, male, generation, and block (replication). The data identifiers should be consecutive positive integers, each beginning with 1 for columns 1, 5, and 2 or 3. The generation codes for column 4 are 0 for parent, 1 for F_1, 2 for F_2, and 3 for F_3. Enter observed data in columns 6 to p if they are available. In addition, data need to be sorted by environment followed by sorting by replication, which can be done easily in Excel.

Step 2: Prepare an information file.

Given the data set in Table 1, an information file (i.e. *adinf.txt*) can be developed in second column with comments in third column of Table 2.

Step 3: Conduct data analysis

Given the above data sample in Table 1 and the information file in Table 2, we can conduct an actual data analysis for the data set mentioned in Step 1. After clicking the computer program GenMod and entering in the information file (adinf.txt) the results will be saved in realf2advar.csv. For each trait, the results include estimated variance components (excluding block), proportional variance components, population means in each environment, predicted genetic effects (if the corresponding variance component is numerically greater than zero). In addition, standard error (SE), probability value (P value), and significance (Sign.) for each parameter are provided. NS means non-significant while S+, S*, and S** mean significance at probability levels of 0.10, 0.05, and 0.01, respectively. The following results included additive effects, dominance effects, additive × environment interaction effects, and dominance × environment interaction effects calculated for lint percentage (LP). Estimated variance components and proportional variance components are listed in Tables 3 and 4, respectively. Predicted additive effects and dominance effects are listed in Tables 5 and 6, respectively.

Estimated variances for additive effects was 1.545, dominance effects, 3.287, additive × environment effects, 0.114, dominance × environment effects, 0.003, residuals, 0.641, and total, 5.589, which were all significantly different from zero for lint percentage (Table 3). Next are the estimated proportional variance components to the phenotypic variance that measure the narrow sense heritability (1.545/5.589*100=28%) and broad sense heritability (27.6%+58.8%=86.4%)

(Table 4). Additive effects for eight parents were provided, showing that all were different from zero (P=0.05) (Table 5). Parents 1 and 2 were two good general combiners that can be used as parents to increase lint percentage (Table 5). The other parents except 7 were associated with negative additive effects, indicating that these lines will reduce lint percentage if they are used as parents. Dominance effects, including homozygous and heterozygous are summarized in Table 6. Among eight homozygous dominance effects, six had significantly positive effects for lint percentage while heterozygous dominance effects either were significantly negative or not different from zero (Table 6). Results suggest that most crosses showed reduced lint percentage (negative heterosis) at their early generations. Then following the dominance effects were the predicted additive × environment and dominance × environment interaction effects (not listed in this paper due to limited space).

The above application is an example of demonstration of using GenMod for actual data analysis. Random data points (some lines in Table 1) can be deleted and new data sets can be generated for additional data analyses. Interested readers may compare results from the complete data set and reduced data sets. By deleting the values of lint percentage various simulations can be conducted as well. New data structures can be generated by deleting lines either randomly or on purpose (for example, delete last replication in second environment). Other genetic models can be applied for other data analyses by using this computer program. For detailed information, please contact the contact author of this paper (Jixiang.wu@sdstate.edu).

Summary

Quantitative genetics is one of the most important components to provide valuable genetic information for improving production and quality of plants and animals. ANOVA based methods are very common statistical methods for quantitative genetics study but are constricted to simple genetic models and specific mating designs and require balanced data structures (i.e. Griffings, 1956; Garder and Eberhart, 1966; Borges, 1987; Haullauer and Miranda, 1988; Das and Griffley, 1994; Lynch and Walsh, 1998). Mixed linear model approaches that were proposed

in the 1960s and 1970s (Hartley and Rao, 1967; Patterson and Thompson, 1971; Rao, 1971) offer the flexibility to analyze unbalanced data structures and complex model systems. However, since the 1980s, these approaches have been introduced into the quantitative genetics study and various useful genetic models and a number of computer programs have been developed (i.e. Zhu, 1998). This paper gives an overview of several useful genetic models that can be generalized by mixed linear models suitable for various data structures.

In addition to actual quantitative genetic data analyses, researchers are not only interested in finding appropriate data structures needed for specific genetic models but also want to determine appropriate genetic models suitable for a specific data structure. Using C++ language, we developed a new computer program GenMod specifically for implementing genetic models being described in this paper. This computer program has the following advantages: (1) it simultaneously conducts analysis for an actual data set or model evaluation for a data set with only experimental design information but no traits included; (2) it is suitable for various genetic data structures; and (5) it provides a significance test by jackknife resampling approaches. Additional genetic models can be added to this computer program. Interested readers can contact the authors of this paper.

References

Borges, O. L.F. 1987. Diallel analysis of maize resistance to sorghum downy mildew. Crop Sci. 27:178-180.

Cheatham, C. L., J. N. Jenkins, J. C. McCarty Jr., C. E. Watson, and J. Wu. 2003. Genetic variance and combining ability of crosses of American cultivars, Australian culitivars, and wild cottons. J. Cotton Sci. 7: 16-22.

Cockerham, C.C. 1980. Random and fixed effects in plant genetics. Theor. Appl. Genet. 56:119–131.

Comstock, R.E., and H.F. Robinson. 1948. The components of genetic variance in populations of biparental progenies and their use in estimating the average degree of dominance. Biometrics 4:254–266.

Das, M. K., and C. A. Griffley. 1994. Diallel analysis of adult-plant resistance to powdery mildew in wheat. Crop Sci. 34:948-952.

Davison, A.C., and D.V. Hinkley. 1997. Bootstrap methods and their application. Cambridge Univ. Press, Cambridge, UK.

Efron, B. 1982. The Jackknife, the bootstrap and other resampling plans. Capital City Press, Montpelier, VT.

Fisher, R.A. 1925. Statistical methods for research workers. 1st ed. Oliver & Boyd, Edinburgh and London, UK.

Garder, C. O., and S. A. Eberhart. 1966. Analysis and interpretation of the variety cross diallel and related populations. Biometrics 22:439-452.

Griffing, B. 1956. Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. 9:463-493.

Hartley, H. O., J. N.K. Rao. 1967. Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika 54:93-108.

Haullauer, A. R. and J. B. Miranda. 1988. Quantitative genetics in maize breeding. Iowa State Univ. Press, Ames, IA.

Jenkins, J. N., J. C. McCarty, J. Wu, and O. A. Gutierrez. 2009. Genetic variance components and genetic effects among eleven diverse upland cotton lines and their F_2 hybrids. Euphytica 167: 397-408.

Jenkins, J. N., J. C. McCarty, J. Wu, S. Saha, O. Guitierrez, R. Hayes, and D. Stelly. 2007. Genetic effects of thirteen Gossypium barbadense L. chromosome substitution lines in topcrosses with upland cotton cultivars: II. Fiber quality traits. Crop Sci. 47: 561-572.

Jenkins, J. N., J. Wu, J. C. McCarty, S. Saha, O. Gutierrez, R. Hayes, and D. M. Stelly. 2006. Genetic evaluation for thirteen chromosome substitution lines crossed with five commercial cultivars: I. yield traits. Crop Sci. 46: 1169-1178.

Littell, R. C., G. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. SAS system for mixed models. SAS institute Inc., Cary, NC.

Lynch, M. and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc. Sunderland, MA.

McCarty, J. C. Jr., J. N. Jenkins, and J. Wu. 2004a. Primitive accession germplasm by cultivar crosses as sources for cotton improvement I: Phenotypic values and variance components. Crop Sci. 44: 1226-1230.

McCarty, J. C. Jr., J. N. Jenkins, and J. Wu. 2004b. Primitive accession germplasm by cultivar crosses as sources for cotton improvement II: Genetic effects and genotype values. Crop Sci. 44: 1231-1235.

McCarty, J. C., J. Wu, and J. N. Jenkins. 2007. Use of primitive derived cotton accessions for agronomic and fiber traits improvement: variance components and genetic effects. Crop Sci. 47: 100-110.

McCarty, J. C., J. Wu, J. N. Jenkins, X. Mo. 2005. Evaluating American and China cotton cultivars and their crosses for improvement. J. Cotton Sci (China) 17(1): 47-55.

McCarty, J. C., J. Wu, J. N. Jenkins. 2008. Genetic associations of cotton yield with its component traits in derived primitive accessions crossed by elite Upland cultivars using the conditional ADAA genetic model. Euphytica 161: 337–352.

McCarty, J. C., J. Wu, S. Saha, J. N. Jenkins, and R. Hayes, 2006. Effects of chromosome 5sh from Gossypium barbadense L. on flower production in G. hirsutum L. Euphytica 152:99-107.

Miller, R.G. 1974. The jackknife: a review. Biometrika 61: 1-15.

Patterson, H.D., and R. Thompson. 1971. Recovery of inter-block information when block size are unequal. Biometrika 58:545-554.

Rao, C.R. 1971. Estimation of variance and covariance components MINQUE theory. J. Multivar. Anal. 1:257–275.

Saha S., J. N. Jenkins, J. Wu, J. C. McCarty, R.G. Percy, R. G. Cantrell, and D. M. Stelly. 2006. Effect of chromosome specific introgression in Upland cotton on fiber and agronomic traits. Genetics 172: 1927-1938.

Saha, S., J. Wu, J. N. Jenkins, J. C. McCarty, R. Hayes, and D. M.Stelly. 2010. Genetic dissection of chromosome substitution lines of cotton to discover novel gossypium barbadense L. alleles for improvement of agronomic traits. Theor. Appl.Genet. 120: 1193-1205.

Searle, S. R., G. Casella, and C.E. McCulloch. 1992. Variance components. Wiley, New York, NY.

Shi, C., J. Zhu, R. Zeng, and G. Chen. 1997. Genetic and heterosis analysis for cooking quality traits of indica rice in different environments. Theor. Appl. Genet., 95:294-300

Tang, B, J.N. Jenkins, C. E. Watson, J. C. McCarty, and R. G. Creech. 1996. Evaluation of Genetic variances, heritabilities, and correlations for yield and fiber traits among cotton F_2 hybrid populations. Euphytica 91:315-322.

Wang, G., J. Zhu, R. Zang, F. Xu, and D. Ji. 1996a. Analysis of covariance components between seed and agronomy traits in upland cotton. Acta Gossypii Sinica 8(6): 395-300.

Wang, G., J. Zhu, R. Zang, F. Xu, and D. Ji. 1996b. Analysis of genetic correlation among seed nutrient quality traits and seed physical traits in upland cotton. J Zhejiang Agri. Univ. 22:585-590.

Wu, J., J. C. McCarty, J.N. Jenkins. 2010. Cotton chromosome substitution lines crossed with cultivars: Genetic model evaluation and seed trait analyses. Theor. Appl. Genet. 120: 1473-1483

Wu, J., J. C. McCarty, S. Saha, J. N. Jenkins, and R. Hayes. 2009. Genetic changes in plant growth and their associations with chromosomes from *Gossypium barbadence* L. in G. *hirsutum* L. Genetica 137: 57-66.

Wu**,** J., J. N. Jenkins, and J. C. McCarty. 2008. Testing variance components by two jackknife techniques. Proc. Appl.Stat. Agri.1-17.

Wu, J., J. N. Jenkins, J. C. McCarty, S. Saha, and D. M. Stelly. 2006a. An additive-dominance model to determine chromosomal effects in chromosome substitution lines and other germplasms. Theor. App. Genet. 112:391-399.

Wu, J., J. N. Jenkins, Jack C. McCarty, and D. Wu. 2006b. Variance component estimation using the ADAA model when genotypes vary across environments. Crop Science 46: 174-179.

Wu, J., J. Zhu**,** D. Ji, and F. Xu. 1995. Genetic analysis of direct and maternal effects of seed traits in upland cotton. (Chinese). Acta Agronomica Sinica. 21(6): 659-664.

Wu, J. J.N. Jenkins, J. McCarty Jr., C. Cheatham. 2000. Separation of single gene effects from additive-dominance genetic models. Proceedings of Applied Statistics in Agriculture, Kansas State University, KS

Xu, Z.C., and J. Zhu. 1999. A new approach for predicting heterosis based on an additive, dominance and additive × additive model with environment interaction. J. Hered. 82:510–517.

Yan, X., S. Xu, Y. Xu, and J. Zhu. 1998. Genetic investigation of contributions of embryo and endosperm genes to malt kolbach index, alpha-amylase activity and wort nitrogen content in barley. Theor. Appl. Genet. 96:709-715.

Ye, Z., Z. Lu, and J. Zhu. 2003. Genetic analysis for developmental behavior of some seed quality traits in Upland cotton (*Gossypum hirsutum* L.). Euphytica. 129: 183-191.

Zhu, J. 1989. Estimation of genetic variance components in the general mixed model. Ph.D. Dissertation, North Carolina State University, Raleigh, NC

Zhu, J. 1993. Methods of predicting genotype value and heterosis for offspring of hybrids. (Chinese). J.Biomath. 8(1): 32-44.

Zhu, J. 1994. General genetic models and new analysis methods for quantitative traits (Chinese). J. Zhejiang Agric. Univ. 20:551–559.

Zhu, J. 1998. Analytical methods for genetic models. Press of China Agriculture, Bejing, China.

Zhu, J., and B.S. Weir. 1994a. Analysis of cytoplasmic and maternal effects: I. A genetic model for diploid plant seeds and animals. Theor. Appl. Genet. 89:153–159.

Zhu, J., and B.S. Weir. 1994b. Analysis of cytoplasmic and maternal effects: I. Genetic models for triploid endosperms. Theor. Appl. Genet. 89:160–166.

Table 1. A cotton data set including 12 F_2 and their eight parents with two years and six replications.

Env	Female	Male	Gen	Rep	LP
1	1	3	2	1	37.15
1	2	3	2	1	36.44
1	1	4	2	1	37.03
1	2	4	2	1	36.28
1	1	5	2	1	37.76
1	2	5	2	1	36.26
1	1	6	2	1	38.14
1	2	6	2	1	37.09
1	1	7	2	1	37.88
1	2	7	2	1	37.63
1	1	8	2	1	36.3
1	2	8	2	1	35.08
1	3	3	0	1	34.22
1	4	4	0	1	36.07
1	5	5	0	1	34.69
1	6	6	0	1	33.97
1	7	7	0	1	35.43
1	8	8	0	1	32.99
1	1	1	0	1	40.95
1	2	2	0	1	41.19
.	.	.	.	.	.
2	1	3	2	6	37.6
2	2	3	2	6	37.07
2	1	4	2	6	36.2
2	6	6	0	6	34.58
2	7	7	0	6	37.26
2	8	8	0	6	36.6

Table 2. An information file with comments for an AD model analysis

Line		Comments
1	2	Code for AD model
2	0	Code for actual data analysis
3	Realf2.txt	Input data file name
4	Realf2advar.csv	Output file name
5	1	Code for block (1 for yes and 0 for no)
6	1	Code for block jackknife
7	1	Number of blocks to be jackknifed
8	1	Pseudo value based jackknife (0 for non-pseudo value based)
9	1	Negative variance components are adjusted to zero (0 for no adjustification)

Table 3. Estimated variance components for lint percentage

Parameter†	Estimate	SE.	Pvalue	Sign.
Var<Add.>	1.545	0.338	<0.001	S**
Var<Dom.>	3.287	0.691	<0.001	S**
Var<Add.*Env.>	0.114	0.109	0.316	NS
Var<Dom.*Env.>	0.003	0.274	0.991	NS
Var<Resi.>	0.641	0.092	<0.001	S**
Var<Pheno.>	5.589	0.388	<0.001	S**

†: Var<Add.> = additive variance, Var<Dom.> =dominance variance, Var<Add.*Env.> = variance for additive-by-environment interaction, Var<Dom.*Env.> = variance for dominance-by-environment interaction, Var<Resi.> = variance for residual, and Var<Pheno.> = phenotypic variance

Table 4. Estimated variance components expressed as proportions to the phenotypic variance for lint percentage

Parameter†	Estimate	SE.	Pvalue	Sign.
V<Add.>/V<P>	0.276	0.080	0.002	S**
V<Dom.>/V<P>	0.588	0.096	<0.001	S**
V<Add.*Env.>/V<P>	0.020	0.022	0.338	NS
V<Dom.*Env.>/V<P>	0.001	0.047	0.979	NS
V<Resi.>/V<P>	0.115	0.022	<0.001	S**

†: V<Add.>/V<P>, V<Dom.>/V<P>, V<Add.*Env.>/V<P>, V<Dom.*Env.>/V<P>, and V<Resi.>/V<P> are the proportions to the phenotypic variance for additive, dominance, additive-by-environment interaction, dominance-by-environment interaction, and residual.

Table 5. Predicted additive effects of eight parents for lint percentage

Parameter†	Estimate	SE.	Pvalue	Sign.
Add.<1>	1.428	0.183	<0.001	S**
Add.<2>	1.274	0.136	<0.001	S**
Add.<3>	-0.717	0.083	<0.001	S**
Add.<4>	-0.485	0.092	<0.001	S**
Add.<5>	-0.481	0.090	<0.001	S**
Add.<6>	-0.365	0.083	<0.001	S**
Add.<7>	0.185	0.053	0.005	S**
Add.<8>	-0.834	0.092	<0.001	S**

†: Additive effects for eight parents.

Table 6. Homozygous and heterozygous dominance effects for lint percentage

Parameter†	Estimate	SE.	Pvalue	Sign.
Dom.<1*1>	3.909	0.551	<0.001	S**
Dom.<2*2>	4.043	0.277	<0.001	S**
Dom.<3*3>	1.426	0.398	0.004	S**
Dom.<4*4>	0.949	0.275	0.005	S**
Dom.<5*5>	0.751	0.289	0.025	S*
Dom.<6*6>	-0.163	0.300	0.599	NS
Dom.<7*7>	-0.260	0.287	0.385	NS
Dom.<8*8>	1.503	0.298	<0.001	S**
Dom.<1*3>	-1.533	0.441	0.005	S**
Dom.<1*4>	-2.241	0.347	<0.001	S**
Dom.<1*5>	-2.006	0.644	0.010	S**
Dom.<1*6>	0.525	0.658	0.442	NS
Dom.<1*7>	0.398	0.372	0.308	NS
Dom.<1*8>	-0.975	0.579	0.120	NS
Dom.<2*3>	-2.320	0.715	0.008	S**
Dom.<2*4>	-0.333	0.429	0.453	NS
Dom.<2*5>	-0.162	0.441	0.720	NS
Dom.<2*6>	-0.699	0.230	0.011	S*
Dom.<2*7>	0.389	0.425	0.380	NS
Dom.<2*8>	-3.196	0.555	<0.001	S**

†: Rows 1 to 8 are homozygous dominance effects while rows 9-20 are heterozygous dominance effects