Software engineering the mixed model for genome-wide association studies on large samples

Zhiwu Zhang, Edward S. Buckler, Terry M. Casstevens and Peter J. Bradbury

Submitted: 6th May 2009; Received (in revised form): 25th September 2009

Abstract

Mixed models improve the ability to detect phenotype-genotype associations in the presence of population stratification and multiple levels of relatedness in genome-wide association studies (GWAS), but for large data sets the resource consumption becomes impractical. At the same time, the sample size and number of markers used for GWAS is increasing dramatically, resulting in greater statistical power to detect those associations. The use of mixed models with increasingly large data sets depends on the availability of software for analyzing those models. While multiple software packages implement the mixed model method, no single package provides the best combination of fast computation, ability to handle large samples, flexible modeling and ease of use. Key elements of association analysis with mixed models are reviewed, including modeling phenotype-genotype associations using mixed models, population stratification, kinship and its estimation, variance component estimation, use of best linear unbiased predictors or residuals in place of raw phenotype, improving efficiency and software–user interaction. The available software packages are evaluated, and suggestions made for future software development.

Keywords: *mixed model; association study; quantitative trait loci; genome-wide; kinship; population structure*

INTRODUCTION

Using association mapping to identify quantitative trait loci (QTL) in structured or stratified populations presents clear challenges. Detecting associations between genotypes and phenotypes and using them to locate QTL in unstructured populations is relatively straight-forward [1]. Unfortunately, unstructured populations or populations in which all individuals are equally related do not exist in nature. Almost by definition, association populations do not meet this criterion.

Correlations between unlinked markers and QTL arise from population stratification [2]. Population stratification or population structure means that subgroups within a population are reproductively isolated, at least partially. As a result, over time, the allele frequencies of the subgroups can diverge. In the most extreme case, a single locus may become fixed for different alleles within each subgroup. If the subgroup means differ for a trait of interest, all loci which differ in allele frequencies between those subgroups will be associated with the phenotype.

The causes and extent of reproductive isolation can vary. In natural populations, reproductive isolation is often the result of physical isolation. Researchers might bring those locally isolated groups together in a germplasm collection then perform genetic analysis on the collection as a whole.

Corresponding author. Zhiwu Zhang, Institute for Genomic Diversity, Cornell University, Ithaca, New York, USA. Tel: +1-607-255-3270; Fax: +1-607-255-6249; E-mail: zz19@cornell.edu

Zhiwu Zhang is a statistical geneticist in the Institute for Genomic Diversity at Cornell University. His research interests in computational genetics include algorithm and software development.

Edward S. Buckler is a research geneticist employed by USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, developing genomic, statistical, bioinformatic and germplasm resources to scan plant genomes for functional polymorphisms using association mapping. He is also an adjunct professor in the Department of Plant Breeding and Genetics at Cornell University.

Terry Casstevens is a software engineer in the Institute for Genomic Diversity at Cornell University working on TASSEL software architecture and the Gramene Diversity project.

Peter J. Bradbury is a computational biologist employed by USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, working on QTL mapping and software development.

Plant breeders may restrict sharing of germplasm between different breeding programs, create distinct populations for discrete markets, or establish heterotic groups and not make breeding crosses between those groups. Even though the underlying population might not be stratified, a sample in which groups of individuals are closely related can mimic population structure. In a sense, structure can arise from selection or biased sampling as well as population admixture.

Whatever the cause, successful association analysis must be able to remove spurious associations arising from structure or unequal relatedness within populations. As a result methods have been developed that correct for structure. These methods and the software that implements them have been reviewed previously [3]. One of the most successful approaches incorporates both major population structure and the relatedness from all pairs of individuals within those populations in a linear mixed model to remove spurious associations [4, 5]. This method has been shown to perform better than alternatives [6].

This statistical advance brings challenges for software development especially when the method is used for large genome wide association studies (GWAS). Here, QTL detection is complicated by the fact that traits may be controlled by many genes, each with a small effect. In addition, many of the QTL have rare alleles, which association studies have low power to detect. Maintaining a large enough sample size is critical in order to have the power to detect such QTL, especially when high resolution is needed for fine mapping and cloning [7]. Large sample sizes with many markers distributed across the entire genome bring challenges for software engineering in terms of data management, model complexity, the time required for the analysis, and the desire to make the method accessible to a large number of users.

Interest in genome-wide association mapping of QTL in plants is being driven by the development of relatively inexpensive methods for high throughput genotyping [8, 9]. Mixed models have been shown to be effective for association analysis, but the challenge will be to apply them to large datasets potentially involving thousands of individuals and hundreds of thousands of markers. While the interest here is in models that will be applied to plant populations, the methods are heavily influenced by the work of animal breeders. This review discusses mixed model approaches and related software packages as they apply to association analysis in plants although some of the non-mixed model approaches (packages) are still commonly used [10–14]. It focuses on methods to reduce spurious associations and improve computational efficiency. Related topics which deserve attention but will not be covered here, include threshold models, generalized linear models and multivariate analysis.

MODELING ASSOCIATION: THE MIXED LINEAR MODEL

Genetic markers and other co-factors that explain the phenotype can be simultaneously fit in a mixed linear model to reduce spurious associations and increase statistical power. Mixed linear model refers to a linear model containing both fixed and random effects [15]. In general, fixed effects take on only a few values and all the values that interest the investigator are included in the study. On the other hand, random effects have values that are taken from a larger population, and the investigator is interested in estimating the mean and variance of that population. In the context of association analysis, the most important distinction is that the covariance structure of random effects can be included in the model, whereas fixed effects have no covariance structure. Why that is important is explained below.

Some of the earliest work to analyze mixed models in a quantitative genetics framework can be traced to Dr C.R. Henderson's efforts at Cornell University. His abstract for the 1949 annual meeting of the American Dairy Science Association marks the birth of Henderson's widely used mixed model equations [16–18]. However, the properties of estimators of fixed effects and predictors of random effects were not proved until ten years later when Shayle Searle joined him as his graduate student. The joint analysis of fixed and random effects results in best linear unbiased estimators (BLUE) of fixed effects and best linear unbiased predictors (BLUP) of random effects such as breeding values [19–21]. The use of the mixed models to predict breeding values was later extended to include genetic markers as random effects [22, 23].

The most common use of a mixed model to test the association between a genetic marker and a phenotype is to fit the marker as a fixed effect and a polygenic component modeled as a random effect. The random effect is the individual taxon (strains, inbred lines or varieties). A likelihood ratio test against the chi-square distribution (when using maximum likelihood), or the Wald test against either the chi-square or normal distribution when using restricted maximum likelihood (REML), is performed to assess the significance of the effect of a polymorphic marker [24].

The covariance matrix of the random effect that Henderson used for predicting animal breeding values is a constant, the additive genetic variance, multiplied by an additive numerator relationship matrix. The diagonal of the numerator relationship matrix equals the inbreeding coefficient plus one. The off-diagonals are Wright's coefficient of correlation [25] multiplied by the square root of the product of the diagonal elements for the two parents. The additive numerator relationship matrix is known as the A matrix in animal breeding. Traditionally, the A matrix is calculated from pedigrees in quadratic time proportional to the number of individuals in the pedigree.

The mixed linear model method has been extensively used in animal breeding and in other fields [26–28]. In addition to the random additive effect, a mixed model can also include other random effects such as a dominance effect, which varies widely among species and traits [29–31]. In genetic marker and phenotype association analysis, the mixed model has been effectively used to adjust for population structure and unequal relatedness among individuals. With the availability of large sets of genetic markers that provide good coverage of whole genomes, a marker-based relationship matrix became a reasonable substitute for pedigree-based relationship matrices, especially where pedigrees were not available or were incomplete [6], especially in plants. In fact when a sufficient number of markers are used, marker-based matrices more precisely describe relatedness between individuals because pedigree-based kinship is based on expected values [32]. Details of marker-based relationship estimates will be discussed later.

Phenotype-marker association can be tested in a variety of ways using mixed models. Markers can either be genotypes of single loci or haplotypes composed of multiple loci on the same chromosome. The direct measurement of haplotypes in heterozygous individuals is difficult. Fortunately, haplotypes can be inferred from genotype [33]. Inferred haplotypes are less informative because of uncertainty about phasing, but the information loss that arises from phasing is small when linkage disequilibrium is strong [34]. Additionally, when the genotype is fit as a fixed effect class variable, the overall test of marker effects can be partitioned into additive and dominance components.

POPULATION STRATIFICATION

One method of association analysis that attempts to correct for population stratification is structured association (SA) analysis [35]. This method assumes that individuals in an existing population trace back to a certain number of discrete populations. Existing individuals could either belong to a single population or be derived from a mix of populations. Once the number of populations is identified and the fractional membership determined for individuals being studied, various methods can be used to correct estimates of phenotype-marker associations for population structure.

Structured analysis (SA) was proposed [35] to use a maximum likelihood ratio to test the association between a segregating locus and a phenotype. The numerator is the likelihood of the observed allele frequencies given the phenotypes and population structure while the denominator is the likelihood of the allele frequencies assuming no association with phenotypes. Pritchard described the use of this method to test binary phenotypes such as the presence and absence of a disease. The method was later modified for quantitative traits [36]. Thornsberry used logistic regression to estimate the likelihoods of observing the allele frequencies under the alternate hypotheses of association and no association with the phenotypes. Alternatively, the association can be tested in a fixed effects linear model [6], which is essentially linear regression.

For SA, linear and logistic regression are closely related and yield similar results. Logistic regression models the genotype as the dependent variable with population structure and phenotype as independent variables. The linear model treats phenotype as the dependent variable and population structure and genotype as independent variables. Since the dependent variable is generally considered to be measured with error, while the independent variables are not, the linear model formulation seems more intuitive. In addition, the linear model can be extended to the mixed model analysis and can easily incorporate other fixed and random effects.

One widely used method developed for calculating a population structure matrix (Q-matrix) models a population of individuals as a metapopulation composed of *k* populations [35]. Each of those *k* populations is assumed to be in Hardy–Weinberg equilibrium. Furthermore, the markers used to determine population membership are assumed to be unlinked and in linkage equilibrium within the *k* populations. This method has been implemented in the software STRUCTURE [35]. The method was shown to correctly identify underlying population structure in specific cases. Advantages include the fact that the output is in a form that can be directly used in SA and the wide acceptance of this software for identifying population membership. Disadvantages include long run times, the fact that the exact number of underlying populations is often not clearly identified, that many populations, self-pollinated plant species in particular, violate the underlying assumption of Hardy–Weinberg equilibrium in the underlying populations, and that the discrete population model is not always appropriate for describing relationships among individuals. To address one of these issues, a method was introduced [37] to determine the number of populations.

To address the application of the method to self-pollinated plant species, the software InStruct was developed [38]. The method uses the same approach as STRUCTURE but relaxes the assumption of Hardy–Weinberg equilibrium in the underlying populations. In addition to assigning individuals to populations, it estimates inbreeding coefficients. Both InStruct and STRUCTURE uses Markov Chain Monte Carlo (MCMC) to fit a model to the data. As with STRUCTURE, the output includes Q-matrix values that can be used in SA. Additional software that uses Bayesian clustering to determine population structure at low levels of population differentiation or that incorporates spatial information has been reviewed [39, 40].

Principal components analysis (PCA) provides a faster alternative to the MCMC model-based methods to identify population structure [41]. A PCA analysis of large data sets, hundreds of thousands of markers and thousands of samples for example, may take a few hours whereas the model based methods would simply not be practical. Not only does PCA run much faster, but it also gives results similar to STRUCTURE. Patterson and co-authors [41] describe the theoretical basis for using PCA. They show why PCA can be expected to identify the same populations as STRUCTURE and demonstrate, using simulations, that PCA is at least as accurate in many situations. In addition to greatly improved speed, a test of significance based on the Tracy–Widom distribution can be applied to the PCA axes to determine the number of populations present. PCA axes can be readily calculated using almost any general statistical software and can be used as the Q-matrices in the previously described models. Alternatively, the axes can be used to adjust genotype and phenotype scores prior to testing for associations [42].

MARKER-BASED RELATIONSHIP

In addition to using estimates of population membership to correct for structure, mixed models can use estimates of relatedness between individuals. The additive numerator relationship matrix (A matrix) described earlier, also called the covariate coefficient matrix, provides these estimates. In diploid species, the covariate coefficient matrix is twice the coancestry coefficient matrix or kinship matrix. Malecot's coancestry coefficient refers to the probability that any two alleles, sampled at random (one from each individual), are identical copies of an ancestral allele. In other words, the two alleles are identical by descent (IBD). The inbreeding coefficient of an individual, a related value, equals the coancestry coefficient of its parents [43].

As an alternative to calculating this matrix from pedigrees, both coancestry and inbreeding coefficients can be estimated from similarity matrices based on marker identity by state (IBS). Different methods have been used to estimate IBD from IBS calculated from markers. The methods in the program SPAGeDi [44] start with *P*(IBS), the probability that a pair of alleles are IBS. That probability is adjusted based on the frequency of that allele in the population as a whole, reasoning that when two alleles are IBS that they are more to likely to be IBD if the overall frequency of that allele is low.

While the pedigree approach assumes that ancestors with unknown parents are unrelated, the marker approach, as implemented in SPAGeDi, requires an arbitrary decision about which taxa are unrelated, with the default being the population average. The resulting matrices (i) contain negative values, (ii) may contain relationship coefficients greater than 1 and (iii) may fail to be non-negative definite (n.n.d.) [45]. Items (i) and (ii) violate the classic definition of IBD but do not present any particular difficulty in solving the resulting mixed model equations. Item (iii) presents computational problems because, under the assumption of normality, the likelihood equation derived from the mixed model will be undefined for some values of genetic variance and residual variance. As a result, some computational methods may have difficulty finding a solution for the resulting mixed model.

A different method of deriving an IBD kinship matrix from an IBS kinship matrix is based on the relationship $P(\text{IBD}) = [P(\text{IBS}) - T]/(1 - T)$, where $T = P(\text{IBS} \mid \text{not IBD})$ [46]. Unfortunately, the value of T is generally unknown. It was suggested [47] that T could be treated as a parameter in the mixed model and a maximum likelihood estimate of T derived. Furthermore, if the adjustment results in any negative values, those values are set to zero. Application of this method to data from several plant species showed that a T-adjusted kinship matrix could provide improved power for QTL detection compared to a kinship matrix calculated by SPAGeDi with negative values set to zero [48].

Alternatively, calculating an IBS kinship matrix based on percent shared alleles is a simple but effective method. It was shown that with proper handling of missing values, a kinship matrix based on allele sharing is guaranteed to be n.n.d. [45]. Furthermore, it was shown that an IBS kinship matrix is as effective as an IBD matrix derived from SPAGeDi. Kinship matrices based on allele sharing were shown to be preferable to using the relationship coefficients calculated using SPAGeDi [45, 49]. While kinship matrices with elements equal to the fraction of alleles shared is effective for testing markers, in order to obtain estimates of additive genetic variance and heritability, these kinship matrices must be appropriately scaled [50].

Weighted alikeness in state (WAIS) is another method of calculating kinship matrices that always produces a matrix that is n.n.d. [51]. Using simulation, it was demonstrated that this method estimates IBD as accurately as several other marker-based methods including those already discussed. Calculating the weights used to adjust the IBS terms requires defining sets of unrelated lines, e.g. maize lines belonging to different heterotic groups.

ESTIMATION OF VARIANCE COMPONENTS

The literature on variance component estimation using mixed models is extensive. The development of mixed models to estimate genetic variance components was recently reviewed [52]. Solving this model in its simplest form involves treating the genetic relationship matrix as a known constant and deriving maximum likelihood estimates of the additive genetic and residual variances. Methods for deriving these estimates are well developed [53–55] and typically use either Newton–Raphson or Expectation-Maximization (EM) algorithms to iteratively search for values that maximize the likelihood equation. Often, the search is reduced to a single dimension by maximizing with respect to the ratio of the variance components and subsequently solving for the variance components.

Because almost all of the time required to compute a solution for the mixed model is spent estimating these variance components, an important area of mixed model research is finding ways to reduce the computational burden of variance component estimation. Each iteration in the solution algorithm requires inverting a matrix that is approximately the size of the kinship matrix. That inversion is an $O(n^3)$ process, where $n =$ the number of taxa. Many of the larger genetic studies being conducted will be too large to be analyzed with this method using standard mixed model software available in computing packages. Methods for reducing the computational burden have included average information REML [56], implemented in the ASREML software among others, and, more recently, a simplification of the likelihood equations using eigenvalue decomposition [45], implemented in the EMMA R-package and in the TASSEL software. Additional strategies for improving speed are discussed below.

ASSOCIATION WITH BLUPS OR RESIDUALS

Instead of directly analyzing raw phenotypes, BLUPs from a mixed model may be substituted as the dependent variable. One reason for using BLUPs is that phenotypes and genotypes are not evaluated for all individuals. For example, milk yield can be scored only on cows and not on bulls, but genotypes are more commonly evaluated on bulls with many progeny. Another reason is to reduce computation time. When the reliability of BLUPs is high, their direct use has statistical power for testing association similar to analysis with raw phenotypes. For example, the growth potential of a bull can be more accurately predicted by using its own record plus the records from parents, siblings and progeny. Progeny testing is particularly accurate for popular bulls where artificial insemination has produced thousands of progeny each. The average reliability of milk yield is above 90% for all US bulls born by 2003 [57]. In such cases, the association analysis using BLUPs can be performed with many fewer observations and require much less time [58].

More recently, researchers have evaluated the use of residuals instead of raw data. The rationale is that after removing all the effects except the marker, including the polygenic genetic variance captured by the BLUPs, the signal due to marker association is still contained in the residuals. Signal from the marker will be removed only to the extent that it is correlated with the other effects. The residual approach performs as well as the approach using raw phenotype directly for low heritability traits [24, 59, 60]. Because the association test using residuals is performed without including the polygenic random effect, tests of individual markers run quickly. The mixed model equations with thousands of individuals only need to be solved once for any particular phenotype. After that, the millions of association tests for individual markers can then be performed using simple *t*-tests or *F*-tests of the marker classes.

PROGRAMING FOR GENOME-WIDE ASSOCIATION

Various strategies exist for testing associations between markers and traits. The most common methods of association analysis involve fitting one marker at a time. An alternative, stepwise regression, proceeds by fitting the marker with the strongest association first, then retesting the remaining markers for significance after. Additional markers are added in a similar fashion until a stopping criterion is met. A different strategy is to fit all the markers simultaneously as random effects. The distribution of the markers can then be modeled according to a Bayesian framework [61].

Some of the statistical methods and algorithms discussed above are relatively simple and easily implemented or can be performed with existing general statistical software. These include PCA for population structure and association with residuals [24, 59, 60]. However, solving mixed model equations (MME) with variance component estimation is more complex and requires the development of software with a number of functions, such as convenient data processing, flexibility of modeling, fast computation and the ability to handle large datasets.

While a variety of methods for solving MME have been implemented in software packages (Table 1) used by animal breeders, work needs to be done to determine which of these will be most useful for GWAS in plants. Because performing association analysis in plants is a somewhat different problem than predicting breeding values in animals, the animal breeding software is not directly useable for most plant studies. Key differences include (i) the existing software focuses on estimating breeding values of individual animals not on testing trait-marker associations, (ii) QTL mapping functions have used pedigree-derived IBD to perform linkage analysis, (iii) the existing software generally solves large, complex models with a number of random effects one time rather than solving a simpler model millions of times and (iv) animal software targets outbred populations with extensive pedigree information while plant populations are often inbred with limited pedigree information.

Among the programming approaches for mixed models, the following three strategies are commonly used. One strategy for handling large datasets is to derive the left-hand side (LHS) of the MME directly instead of building the design matrix first and deriving the LHS from that. This method can result in large gains in efficiency when individuals are measured in multiple environments. For a complete dataset without missing values, using the average across the environments is equivalent to using each measurement to perform association tests on genetic markers. This approach allows an analysis to be done much faster compared to software packages which build the design matrix first. For example, dramatic differences can be seen when SAS Proc Mixed is

Table I: Software packages useful for mixed model for association mapping[a]

Category	Program	Web address (http)	Availability	User interface[b]	Flexible modeling	Automatic GWAS[c]	Sample size[d]	Population structure	Build Kinship from pedigree	Build Kinship from marker	Number of Random Effects	Reference
Multi-purpose	TASSEL	www.maizegenetics.net	Free	G/C	No	Yes	S	Yes	Yes	Yes	1	[71]
	SAS	www.sas.com	Licensed	C	Yes	Yes	S	Yes	Yes	Yes	≥1	[73]
	JMP Genomics	www.jmp.com/software/genomics	Licensed	G	Yes	Yes	NA	Yes	NA	Yes	≥1	[74]
Mixed model	ASREML[d]	www.vsni.co.uk/software/asreml	Licensed	C	Yes	Yes	NA	Yes	Yes	No	≥1	[56]
	MTDFREML	aipl.arsusda.gov/curtvt/mtdfreml.html	Free	C	Yes	No	L	Yes	Yes	No	≥1	[75, 76]
	DMU	www.dmu.agrsci.dk	Free	C	Yes	No	L	Yes	Yes	No	≥1	[77]
	QxPak	nce.ads.uga.edu/~ignacy/newprograms.html	Free	C	Yes	Yes	L	Yes	Yes	No	≥1	[78]
	WOMBAT	agbu.une.edu.au/~kmeyer/wombat	Free	C	Yes	NA	L	Yes	Yes	No	≥1	[79]
	EMMA(R)	mouse.cs.ucla.edu/emma	Free	C	No	Yes	M	No	No	Yes	1	[45]
Structure	InStruct	cbsuapps.tc.cornell.edu/InStruct.aspx	Free	C			S	Yes				[38]
	Eigensoft	genepath.med.harvard.edu/~reich/Software.htm	Free	C			M	Yes				[41, 42]
	STRUCTURE	pritch.bsd.uchicago.edu/structure	Free	G			S	Yes				[35]
Kinship	PowerMarker	statgen.ncsu.edu/powermarker	Free	G			S	No	No	Yes		[80]
	SPAGeDi	www.ulb.ac.be/sciences/ecoevol/spagedi.html	Free	C			S	No	No	Yes		[44]

NA: not available.

[a] Software packages are sorted roughly by number of functions and desirable features. The evaluation of functions and features were based solely on authors' judgements, which may be biased. While the software Genstat was included in the text, it is not included in this table because of the authors' lack of familiarity with it.

[b] ASREML is available as standalone and as S language and R (ASREML-S and ASREML-R) add-ons.

[c] A software package was considered to have a graphical (G) user interface when all analyses could be performed by mouse clicking and guided keyboard input; otherwise, it was classified as command (C) line interface.

[d] Automatic GWAS refers to whether a single analysis automatically tests all markers across the genome as opposed to manually testing one marker at a time.

[e] Sample size for which software can perform an association test in minutes per marker or estimate structure or kinship in hours: Small (S): less than 1000; Medium (M): between 1000 and 5000 and Large (L): larger than 5000. Estimates of software capacity are approximate and based on authors' experience rather than exhaustive testing.

used to analyze individuals averaged across environments compared to using individuals measured in each environment.

A second strategy is to avoid inverting large matrices since the computing time for matrix inversion is proportional to the cube of the number of rows (or columns) in the matrix [54]. Matrix inversion occurs in two places when solving the MME. First, the kinship matrix is inverted before adding it to the LHS. Instead of inverting the original kinship matrix, its inverse can be directly derived from pedigrees with the time required proportional to the square of the number of taxa rather than its cube [62–65]. This shortcut is not available when kinship is calculated from markers. The second instance is inversion of the LHS to solve the MME, which is an iterative process. Methods for solving the MME vary considerably in time required and numerical accuracy, especially when the LHS is singular or nearly so. Methods developed for genomic selection [66] or used in software for estimating breeding values in animals may help, but most have not been evaluated in the context of association mapping.

A third strategy is to take advantage of the sparse nature of the LHS that often arises when it is derived from pedigrees. Although the kinship matrix itself may have many non-zero elements, the inverse of the kinship is sparse, that is, it contains many zeros. As indicated by the direct inverse algorithm [62, 63], each individual only contributes three elements: two are pairings of the progeny with each parent, and the third is the element between the two parents. Because the matrix is symmetric, there are only six non-zero off-diagonal elements for each progeny. As a result, sparse matrix libraries or algorithms can be used to reduce the time needed to solve the MME [67].

Legarra and Misztal [66] reviewed methods for solving MME in the context of genomic selection. They found matrix free GSRU (Gauss-Siedel with residual updating) and PCG (preconditioned conjugate gradient) methods were much faster than Cholesky decomposition, a common method used in solving MME. These techniques may hold promise for reducing the time required to analyze large association studies. To estimate variance components, Misztal [68] recommends using AI REML (average information restricted maximum likelihood) for problems of moderate size and complexity and Bayesian analysis for large datasets and complex models. While promising, Bayesian analysis is an area that needs additional research before it can be used for routine association analysis. Misztal points out that Bayesian analysis does not always succeed and that each run needs to be inspected to make sure problems did not occur. That makes the method infeasible for fitting large numbers of markers individually. In addition, association studies in plants that use Bayesian analysis to fit all markers simultaneously have shown that method to be very slow even for relatively small data sets [69, 70].

SOFTWARE–USER INTERACTION

An important component in software design is the choice of user interface. There are two main ways to design software interfaces. One involves using a front-end graphical user interface (GUI) client. The other executes software using a Command Line Interface (CLI). In the context of solving mixed model analyses, a GUI client can guide users through the data input process by providing dialogs for users to select data, input parameters, run analyses, and view results. With a CLI, data sources and parameter values are specified before program execution as command line options. Once the program has started, the analysis runs non-interactively to completion.

A CLI has several advantages. For instance, users can setup executions and let them run unattended for as long as necessary. Also, multiple executions can be setup to iterate over varying data sets or model parameters without continued user involvement. Using a CLI can save time by avoiding repetitive mouse clicks and parameter entries when running a series of similar analyses. Not only that, batch jobs can be organized to run on multiple machines and/or multi-processor machines such as computing clusters. Spreading these jobs over multiple processors greatly reduces the time required to complete long running mixed models. Many of the software packages for solving mixed models and estimation of variance components were implemented with CLI.

The main advantage of a GUI is that it can guide users through an analysis and significantly reduce the time it takes for new users to become productive or be more likely to be used by those intimidated by CLI software. In addition, it relieves users of the need to learn command line syntax. Software designs with GUIs also allow users to view intermediate results and make decisions about analyses based on those intermediate results.

For some software, both GUI and CLI versions exist. For example, TASSEL's [71] architecture is organized in modules, called plug-ins, which perform various functions. These plug-ins are implemented both with a GUI and a CLI. As a result, consistent results are achieved independent of the interface. In the GUI, the plug-ins are invoked by clicking buttons on the interface. With the CLI, the plug-ins are used in a predetermined pipeline that passes the output from one step to the input of another. Depending on the needs of the user, unlimited pipeline setups can be designed. There are significant advantages to creating a software architecture that easily accommodates both GUI and CLI versions. A number of sophisticated software packages offer a combination of GUI and CLI. SAS, ASREML and R-Package software are examples of the 'hybrid' approach that provides a GUI for submitting commands interactively.

OVERVIEW OF SOFTWARE PACKAGES

Features of software useful for solving mixed models of the type considered here are summarized in Table 1. Because the mixed model method was developed initially for predicting breeding values for animals and breeding populations are usually very large (over millions), most of the resulting software packages can solve mixed models for large samples reasonably quickly. These software packages are used routinely on data that have a relatively stable format but have limited flexibility for handling other types of data or performing non-standard analyses. Reasons why these packages may not be well suited to plant association analysis were discussed above. Of the packages designed for animal breeding, only QxPak and ASREML, will automatically run association analysis for a series of markers. In addition, MTDFREML, ASREML, WOMBAT and DMU will accept a user-supplied additive relationship matrix (or kinship matrix). All can calculate that matrix based on pedigrees.

Public, freely available software suitable for association analysis using mixed models in plants include TASSEL and EMMA/R. Both analyze moderately large datasets in a reasonable amount of time but only allow a single effect (samples or taxa) to be fit as a random effect. All other effects are treated as fixed. EMMA relies on the R [72] for data management and visualization whereas TASSEL handles those functions itself. Several commercial software packages can be used for association. ASREML and JMP Genomics are specifically engineered for genetic analysis and can handle more complex models while general purpose packages such as SAS Proc Mixed and Genstat can perform association analysis but require more expertise and programming on the part of the user. Few timed comparisons of software suitable for association analysis have been published. One of those software [45] compared EMMA/R, TASSEL, ASREML and SAS Proc Mixed. Using a data set with 553 SNPs, 277 lines and 3 phenotypes, the study reported that EMMA/R was faster than ASREML by a factor of 10, which in turn was faster than SAS Proc Mixed and TASSEL by a factor of six. The method used by EMMA/R has since been implemented in TASSEL.

From the user's perspective, none of the available software packages provide an optimal combination of advanced statistical methods, efficient algorithms, and ease of use. When a less powerful method is used, it means less value is realized from data that is often expensive to collect. When a less efficient algorithm is used, it can mean waiting hours, days or even weeks with less opportunity to investigate alternative models. Complicated software can require a long period of training and provide results that can be difficult to interpret.

Clearly, freely available software plays an important role in scientific investigation. Without the developers' hard work embodied in the computer code, the analyses of many empirical studies would be much harder or even impossible [66]. However, public funding for software development is limited with the result that free software often lacks flexibility, ease of use or user support.

CONCLUSION

The theory of mixed models has been well developed, and many factors impacting association analysis have been investigated. However, software does not exist that combines all the analytical tools that users would like, including ease of use, modeling flexibility, computing efficiency, capacity for large samples, permutation testing and marker selection. In large part, this results from the rapid pace of development of new sequencing and data analysis methods. Software development necessarily lags behind method development. For existing software packages that analyze mixed models, improvements for handling large data sets and marker derived kinship matrices would contribute greatly to the analysis of the increasingly large GWAS being conducted.

Key Points

- Mixed models provide a powerful method for detecting phenotype–genotype associations but are resource intensive especially for large data sets.
- Good software design can improve the efficiency of mixed model methods and make them accessible to a broader group of users.
- Alternatives exist for calculating kinship, estimating population structure and solving mixed models.
- The choice among those alternatives has a significant impact on computation time and results.

Acknowledgements

The authors thank Dale Van Vleck for his helpful comments. The authors also thank Linda Rigamer Lirette and Sean Myles for editing of the manuscript.

FUNDING

NSF-Plant Genome Program (DBI-0321467, 0703908 and 0820619), NSF-Plant Genome Comparative Sequencing Program (DBI-06638566) and USDA-ARS.

References

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;**273**:1516–17.
2. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999;**65**:220–8.
3. Forabosco P, Falchi M, Devoto M. Statistical tools for linkage analysis and genetic association studies. *Expert Rev Mol Diagn* 2005;**5**:781–96.
4. Gwaze DP, Zhou Y, Reyes-Valdes MH, *et al*. Haplotypic QTL mapping in an outbred pedigree. *Genet Res* 2003;**81**: 43–50.
5. Schenkel FS, Miller SP, Ye X, *et al*. Association of single nucleotide polymorphisms in the leptin gene with carcass and meat quality traits of beef cattle. *J Anim Sci* 2005;**83**: 2009–20.
6. Yu JM, Pressoir G, Briggs WH, *et al*. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 2006;**38**:203–8.
7. Shalom A, Darvasi A. Experimental designs for QTL fine mapping in rodents. *Methods Mol Biol* 2002;**195**:199–223.
8. Ersoz ES, Yu J, Buckler ES. Applications of linkage disequilibrium and association mapping in maize. In: Kriz A, Larkins B (eds). *Molecular Genetic Approaches to Maize Improvement*. Berlin: Springer, 2008;173–95.
9. Zhu C, Gore M, Buckler ES, *et al*. Status and prospects of association mapping in plants. *Plant Genome* 2008;**1**:5–20.
10. Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
11. Excoffier L, Heckel G. Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet* 2006; **7**:745–58.
12. Lange C, DeMeo DL, Laird NM. Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet* 2002;**71**:1330–41.
13. Horvath S, Xu X, Lake SL, *et al*. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol* 2004;**26**: 61–9.
14. Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 2000;**66**:279–92.
15. Gelman A. Analysis of variance: why it is more important than ever. *Ann Stat* 2005;**33**:1–53.
16. Schaeffer LR. C. R. Henderson: contributions to predicting genetic merit. *J Dairy Sci* 1991;**74**:4052–66.
17. Freeman AE. C. R. Henderson: contributions to the dairy industry. *J Dairy Sci* 1991;**74**:4045–51.
18. Van Vleck LD. Charles Roy Henderson, National Academy of Sciences (Biographical Memoirs) 1998;73:182–207.
19. Searle SR. C. R. Henderson, the statistician; and his contributions to variance components estimation. *J Dairy Sci* 1991;**74**:4035–44.
20. Henderson CR. Estimation of variance and covariance components. *Biometrics* 1953;**9**:226–52.
21. Henderson CR. Selection index and expected genetic advance. In: Hanson WD, Robinson HF (eds) *Statistical Genetics and Plant Breeding*. NAS-NRC. Pub. No: 982, Washington, DC, 1963;141–63.
22. Fernando RL, Grossman M. Marker assisted selection using best linear unbiased prediction. *Genet Sel Evol* 1989; **21**:467–77.
23. Thallman RM, Hanford KJ, Kachman SD, *et al*. Sparse inverse of covariance matrix of QTL effects with incomplete marker data. *Stat Appl Genet Mol Biol* 2004;**3**: Article 30.
24. Lam AC, Schouten M, Aulchenko YS, *et al*. Rapid and robust association mapping of expression quantitative trait loci. *BMC Proc* 2007;**1**(Suppl. 1):S144.
25. Wright SI. Coefficient of inbreeding and relationship. *Am Naturalist* 1922;**56**:330–8.
26. Zhang Z, Zhu L, Sandler J, *et al*. Estimation of heritabilities, genetic correlations, and breeding values of four traits that collectively define hip dysplasia in dogs. *Am J Vet Res* 2009;**70**:483–92.
27. Hannan PJ, Murray DM. Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed-model and the logistic mixed-model analyses in simulated community trials with a dichotomous outcome variable at the individual level. *Eval Rev* 1996;**20**:338–52.
28. Holditch-Davis D, Edwards LJ, Helms RW. Modeling development of sleep-wake behaviors: I. Using the mixed general linear model. *Physiol Behav* 1998;**63**:311–18.

29. Wei M, van der Werf JH. Animal model estimation of additive and dominance variances in egg production traits of poultry. *J Anim Sci* 1993;**71**:57–65.

30. Johansson K, Kennedy BW, Quinton M. Prediction of breeding values and dominance effects from mixed models with approximations of the dominance relationship matrix Livestock Production. *Science* 1993;**34**:213–23.

31. Misztal I. Estimation of variance components with large-scale dominance models. *J Dairy Sci* 1997;**80**:965–74.

32. Zhu L, Zhang Z, Friedenberg S, *et al*. The long (and winding) road to gene discovery for canine hip dysplasia. *Vet J* 2009;**181**:97–110.

33. Marchini J, Cutler D, Patterson N, *et al*. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 2006;**78**:437–50.

34. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;**7**:781–91.

35. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**:945–59.

36. Thornsberry JM, Goodman MM, Doebley J, *et al*. Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 2001;**28**:286–9.

37. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol Notes* 2005;**14**:2611–20.

38. Gao H, Williamson S, Bustamante CD. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* 2007;**176**:1635–51.

39. Latch E, Dharmarajan G, Glaubitz J, *et al*. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conserv Genet* 2006;**7**: 295–302.

40. Chen C, Durand E, Forbes F. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Mol Ecol Notes* 2007;**7**:747–56.

41. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190.

42. Price AL, Patterson NJ, Plenge RM, *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**:904–9.

43. Falconer DS, Mackay T. *Introduction to Quantitative Genetics*. Fourth edition, Addison Wesley Longman, Harlow, Essex, UK, 1996

44. Hardy OJ, Vekemans X. spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2002;**2**:618–20.

45. Kang HM, Zaitlen NA, Wade CM, *et al*. Efficient control of population structure in model organism association mapping. *Genetics* 2008;**178**:1709–23.

46. Bernardo R. Estimation of coefficient of coancestry using molecular markers in maize. *Theor Appl Genet* 1993; **85**:1055–62.

47. Stich B, Mohring J, Piepho H-P, *et al*. Comparison of mixed-model approaches for association mapping. *Genetics* 2008;**178**:1745–54.

48. Stich B, Melchinger AE. Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. *BMC Genomics* 2009; **10**:94.

49. Zhao K, Aranzana MJ, Kim S, *et al*. An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 2007;**3**:4.

50. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci* 2008;**91**:4414–23.

51. Maenhout S, De Baets B, Haesaert G. Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. *Theor Appl Genet* 2009;**118**:1181–92.

52. Thompson R. Estimation of quantitative genetic parameters. *Proc Biol Sci* 2008;**275**:679–86.

53. Searle SR, Casella G, McCulloch CE. *Variance Components,* John Wiley and Sons, Hoboken, NJ, 1992.

54. Henderson CR. *Applications of Linear Models in Animal Breeding,* University of Guelph, Guelph, Ontario, Canada, 1984.

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

56. Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 1995;**51**:1440–50.

57. Powell RL, Sanders AH, Norman HD. Accuracy of foreign dairy bull evaluations in predicting United States evaluations for yield. *J Dairy Sci* 2004;**87**:2621–6.

58. Calvo JH, Marcos S, Jurado JJ, *et al*. Association of the heart fatty acid-binding protein (FABP3) gene with milk traits in Manchega breed sheep. *Anim Genet* 2004;**35**: 347–9.

59. Aulchenko YS, de Koning D-J, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 2007; **177**:577–85.

60. Amin N, van Duijn CM, Aulchenko YS. A genomic background based method for association analysis in related individuals. *PLoS ONE* 2007;**2**:e1274.

61. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 2001;**157**:1819–29.

62. Henderson CR. Rapid method for computing the inverse of a relationship matrix. *J Dairy Sci* 1975;**58**: 1727–30.

63. Henderson CR. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 1976;**32**:69–83.

64. ter Heijden E, Chesnais JP, Hickman CG. An efficient method of computing the numerator relationship matrix and its inverse matrix with inbreeding for large sets of animals. *Theoret Appl Genet* 1977;**45**:237–41.

65. Quaas RL. Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* 1976;**32**: 949–53.

66. Legarra A, Misztal I. Technical note: computing strategies in genome-wide selection. *J Dairy Sci* 2008;**91**:360–6.

67. Boldman KG, Van Vleck LD. Derivative-free restricted maximum likelihood estimation in animal models with a sparse matrix solver. *J Dairy Sci* 1991;**74**:4337–43.

68. Misztal I. Reliable computing in estimation of variance components. *J Anim Breed Genet* 2008;**125**:363–70.

69. Sillanpaa MJ, Bhattacharjee M. Bayesian association-based fine mapping in small chromosomal segments. *Genetics* 2005;**169**:427–39.

70. Iwata H, Uga Y, Yoshioka Y, *et al*. Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among Oryza sativa L. germplasms. *Theor Appl Genet* 2007;**114**:1437–49.

71. Bradbury PJ, Zhang Z, Kroon DE, *et al*. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007;**23**:2633–5.

72. R Development Core Team. *R: A Language and Environment for Statistical Computing*, http://www.R-project.org 2009.

73. SAS Institute Inc. *Statistical Analysis Software for Windows,* Cary, NC, USA, 2002

74. SAS Institute Inc. *JMP Genomics*, Cary, NC, USA, 2007.

75. Boldman KG, Kriese LA, Van Vleck LD *et al*. A manual for use of MTDFREML. USDA-ARS, Clay Center, Nebraska, 1993.

76. Zhang Z, Todhunter RJ, Buckler ES, *et al*. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *J Anim Sci* 2007;**85**:881–5.

77. Madsen P, Sørensen P, Su G *et al*. DMU – a package for analyzing multivairate mixed models. In *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, August 13–18, Belo Horizonte, MG, Brasil, 2006.

78. Perez-Enciso M, Misztal I. Qxpak: a versatile mixed model application for genetical genomics and QTL analyses. *Bioinformatics* 2004;**20**:2792–8.

79. Meyer K. WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *J Zhejiang Univ Sci B* 2007;**8**:815–21.

80. Liu K, Muse SV. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 2005;**21**:2128–9.