frontiers in Genetics

REVIEW
published: 21 August 2018
doi: 10.3389/fgene.2018.00341

Genome-Wide Expression Quantitative Trait Loci Analysis Using Mixed Models

*Chaeyoung Lee**

Department of Bioinformatics and Life Science, Soongsil University, Seoul, South Korea

Expression quantitative trait loci (eQTLs) are important for understanding the genetic basis of cellular activities and complex phenotypes. Genome-wide eQTL analyses can be effectively conducted by employing a mixed model. The mixed model includes random polygenic effects with variability, which can be estimated by the covariance structure of pairwise genomic similarity among individuals based on genotype information for nucleotide sequence variants. This increases the accuracy of identifying eQTLs by avoiding population stratification. Its extensive use will accelerate our understanding of the genetics of gene expression and complex phenotypes. An overview of genome-wide eQTL analyses using mixed model methodology is provided, including discussions of both theoretical and practical issues. The advantages of employing mixed models are also discussed in this review.

Keywords: expression quantitative trait locus, genetic association, genetic variance, heritability, mixed model

OPEN ACCESS

Edited by:
Mariza De Andrade,
Mayo Clinic, United States

Reviewed by:
Paola Sebastiani,
Boston University, United States
Wan-Yu Lin,
National Taiwan University, Taiwan

***Correspondence:**
Chaeyoung Lee
clee@ssu.ac.kr

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *09 May 2018*
Accepted: *09 August 2018*
Published: *21 August 2018*

Citation:
Lee C (2018) Genome-Wide Expression Quantitative Trait Loci Analysis Using Mixed Models. Front. Genet. 9:341. doi: 10.3389/fgene.2018.00341

INTRODUCTION

Gene expression is the frontier process linking genotypes to phenotypes, and thus the genetics of gene expression is critical for dissecting the genetic basis of complex phenotypes. Currently, the genetics of gene expression largely depends on identifying an expression quantitative trait locus (eQTL), i.e., an association between gene expression and the genotype at a locus. Genome-wide eQTL studies have shown that the eQTLs explain a substantial proportion of variation in gene expression (Spielman et al., 2007); about 90% of the variation in the expression of many genes has been attributed to nucleotide variants (Yang S. et al., 2014). Genome-wide eQTL analyses have enabled us to obtain a profile of regulatory signals for each gene and to compare multiple profiles for cells with different functions. Furthermore, eQTL analyses for a variety of molecular traits can provide evidence for the specific regulatory stages and functions of gene expression. Data production for such eQTL analyses is increasing dramatically with the continuous development of technology. The Geuvadis consortium generated RNA sequencing data on lymphoblastoid cell lines of 462 individuals from the 1000 Genome Project (Lappalainen et al., 2013), and the Genotype-Tissue Expression consortium reported RNA sequencing data on 1641 samples across 43 tissues from 175 individuals (GTEx Consortium, 2015). The choice of statistical method for analyzing these data is increasingly important to draw better inferences.

Mixed model methodology is an emerging method for genome-wide association studies (GWASs); it was originally applied to the genome-wide identification of loci associated with a phenotypic trait but can be extended to analyses of associations between loci and intermediate molecular traits, such as RNA and protein expression levels. The mixed model methodology has been employed for nearly a half century for genetic analyses because it can explain polygenic

effects while this is the intractable problem using fixed models. The polygenic effects can be assessed as random effects which are the special feature of mixed models, using pedigree-based genetic relationships. Currently, GWAS data are evaluated by mixed models with genomic similarity among unrelated individuals modified from pedigree information to nucleotide variant information. The direct approach of the genetic difference among individuals is an efficient way of avoiding population stratification that is one of the critical problems producing spurious genetic associations in GWAS. An overview of eQTL analyses by the mixed model methodology is provided, with an emphasis on important issues. Only essential mathematical notation for analytical models and relevant estimation methods are concisely presented in this review to ensure a clear presentation of mixed models and to avoid the intricacies of specific conditions.

HISTORICAL LOOK AT MIXED MODELS

Henderson (1950, 1953) developed the mixed model and the corresponding parameter estimation method for applications in genetics. Prior to its development, Fisher (1925) estimated variance components using mean squares of analysis of variance (ANOVA) and their expected values, but this estimation is limited to balanced data. Henderson's mixed model methodology has been utilized extensively for the genetic improvement of animals. Since the mixed model is hierarchical, the estimation of variance components for fixed and random effects is stressed as a priority. Various estimation methods for variance components have been applied, and they can be categorized into four general types: ANOVA-based estimation, distribution-free quadratic estimation, likelihood-based estimation, and Bayesian estimation (**Table 1**). An example of distribution-free quadratic estimation is the minimum variance quadratic unbiased estimation (MIVQUE), in which a local best unbiased estimate is obtained with minimum variance of the quadratic form of a random variable (Rao, 1971). Empirical bias was observed in the application of the MIVQUE to genetic and residual variance components (Van Tassell et al., 1995). Likelihood-based estimation has attractive statistical properties, such as asymptotic unbiasedness and asymptotic efficiency (Casella and Berger, 1990). Nevertheless, restricted maximum likelihood (REML; Patterson and Thompson, 1971) estimation has been dominantly preferred to maximum likelihood (ML; Hartley and Rao, 1967) estimation. This is because the degrees of freedom in estimating fixed effects are explained for REML, but not for ML. Empirical unbiasedness of REML estimates has been verified using simulated data, even for artificial selection in animals (Jensen and Mao, 1991; Lee and Pollak, 1997b). REML has been utilized as the standard method for estimating variance components in mixed model analyses. Representative algorithms for obtaining REML estimates include the quasi-Newton method (Kennedy and Gentle, 1980), average information method (Johnson and Thompson, 1995), expectation maximization method (Laird et al., 1987), and derivative-free method (Boldman and Van Vleck, 1991). The Bayesian estimation of variance components is feasible by Markov chain Monte Carlo (MCMC), a numerical procedure for sampling from a desired probability distribution at equilibrium in a Markov chain. Bayesian estimation is increasingly used for variance component estimation. The advantages of Bayesian estimation are briefly discussed in the section on parameter estimation.

Forming a covariance structure of random effects is a critical step for a mixed model analysis. The structure could be generated as a matrix with elements of pairwise genetic relationships among individuals based on pedigree information. It was first called a numerator relationship matrix, and efficient algorithms for building and inverting the matrix enabled geneticists to handle large matrices (Quaas, 1976, 1988). This matrix should be modified according to genetic model, and only a portion of the genetic variance can be explained by the analytical model with the matrix. For example, while an animal model explains all of the genetic variance (Quaas and Pollak, 1980), a sire model explains only a quarter (Wang et al., 1993), and a sire-maternal grand sire model explains 3/8 (Lee and Pollak, 2002).

The mixed model has been applied to GWAS (Kang et al., 2010; Zhang et al., 2010; Yang et al., 2011). The only difference is that genetic covariance between individuals is assessed by genotype information, instead of pedigree information. Genotype information for a large number of nucleotide sequence variants is available owing to dramatic improvements in sequencing technologies. The mixed model can be employed to analyze gene expression, instead of phenotypes.

In summary, mixed models and the corresponding estimation methodology have been improved since their development to explain the genetics of complex phenotypes. They were, of course, adapted for applications to particular disciplines, such as evolution (Wilson et al., 2006), ecology (Melbourne and Hastings, 2008), and social sciences (Bartholomew et al., 2008). Irrespective of their application, they have a covariance structure of random effects as a common characteristic and this common structure is the reason why mixed models are used. The covariance structure for the genetics of complex phenotypes is represented as the numerator relationship matrix constructed by various ways. In genome-wide eQTL analyses by mixed models, the numerator relationship matrix can be constructed using genome-widely available nucleotide variant data, explaining polygenic effects. Since polygenic effects reflect different genetic backgrounds among individuals, the mixed model analysis is a powerful method for identifying accurate eQTLs. Of course, it may avoid spurious eQTLs produced by confounding effects of population stratification and kinship. Note that "genomic similarity matrix" is used as the variant-based genetic relationship matrix hereafter in this review.

GENOME-WIDE eQTL ANALYSIS USING MIXED MODELS

A mixed model for genome-wide eQTL analyses is presented in a generalized form with matrices as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon}$$

TABLE 1 | Major methods for variance component estimation in a mixed model framework.

Category	Method (abbreviation)	Property
ANOVA-based estimation	Henderson's method 3	Unbiasedness Possibility of negative estimate (e.g., out of parameter space) Unknown distribution Lack of uniqueness
Distribution-free quadratic estimation	Minimum norm quadratic unbiased estimation (MINQUE)	No normality assumption Possibility of negative estimate
	Iterative MINQUE (I-MINQUE)	No normality assumption Asymptotic normality Possibility of negative estimate
	Minimum variance quadratic unbiased estimation (MIVQUE)	Equivalent to MINQUE with null priors (MINQUE0) Properties shared with MINQUE
Likelihood-based estimation	Maximum likelihood (ML)	Normality assumption Non-negative estimate by maximization within parameter space Asymptotic unbiasedness Asymptotic efficiency No closed form solution
	Restricted maximum likelihood (REML)	Explaining degrees of freedom involved in fixed effects Relatively free from normality assumption[1] Non-negative estimate Asymptotic unbiasedness Asymptotic efficiency No closed form solution Various numerical solutions are available The most popular method
Bayesian estimation	Gibbs sampling	Direct inference from posterior distribution[2]
	Metropolis and Hastings	Direct inference from posterior distribution[2] Data augmentation

[1]REML was originally derived under normality assumption as ML, but REML estimates can be corresponding to I-MINQUE, which does not require the normality assumption (Brown, 1976).
[2]Point (e.g., posterior mean) and interval (e.g., 95% credible interval) estimates are directly obtained using the samples of posterior distribution generated by a Markov chain Monte Carlo.

where $\mathbf{y}$ is the $n \times 1$ vector of gene expression levels, n is the number of the gene expression levels, β is the $n_f \times 1$ vector of fixed effects (e.g., gender, age, and nucleotide variant effects), n_f is the number of the fixed effects, and $\mathbf{X}$ is the $n \times n_f$ design matrix for the fixed effects. The fixed effects include the minor allele effect of the candidate nucleotide variant, and the corresponding column of $\mathbf{X}$ includes elements of 0, 1, and 2 for the homozygous major allele, heterozygous genotype, and homozygous minor allele under the assumption of an additive model with a biallelic nucleotide variant. $\mathbf{g}$ is the $n \times 1$ vector of random polygenic effects ($\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$) where $\mathbf{G}$ is the $n \times n$ genomic similarity matrix with elements of pairwise genomic similarity coefficients based on genotypes of nucleotide variants, and σ_g^2 is the polygenic variance component. The genomic similarity coefficient between individuals j and k can be calculated as follows:

$$g_{jk} = \frac{1}{n_v} \sum_{i=1}^{n_v} \frac{(\tau_{ij} - 2f_i)(\tau_{ik} - 2f_i)}{2f_i(1 - f_i)}$$

where n_v is the number of nucleotide variants that contribute to the genomic similarity, τ_{ij} and τ_{ik} represent the number (0, 1, or 2) of minor alleles for the nucleotide variant i, and f_i is the frequency of the minor allele. ε is the $n \times 1$ vector of random environmental effects ($\boldsymbol{\varepsilon} \sim \boldsymbol{N(0, \mathrm{I}\sigma_\varepsilon^2)}$), where $\mathbf{I}$ is the $n \times n$ identity matrix, and $\boldsymbol{\sigma_\varepsilon^2}$ is the environmental variance component. Variance in gene expression is thus defined as $\boldsymbol{var(\mathbf{y}) = \mathbf{V} = \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2}$.

To avoid underestimating the association of gene expression with the candidate nucleotide variant by proximal contamination, genomic similarity coefficients can be estimated by excluding nucleotide variants that are in linkage disequilibrium with the candidate. One strategy is to exclude all variants located on the same chromosome as the candidate variant (Lippert et al., 2011). Theoretically, different genomic similarity coefficients are required for evaluating associations with every nucleotide variant, but this strategy reduces the burden by estimating the genomic similarity matrix for the same number of chromosomes. This efficiency should be stressed because the computing and memory costs for the genomic similarity matrix based on nucleotide variant information are expensive, unlike dealing with the sparse genetic relationship matrix based on pedigree information.

Considering only independent variants with more than a certain effect size is another efficient way to substantially reduce the cost. Only representative variants based on linkage disequilibrium can be used to explain polygenic effects (Lippert et al., 2011). The excessive exclusion of variants, however, may lead to insufficient correction for stratification (Yang J. et al., 2014). An example for selecting representative variants is to maximize polygenic variance by a stepwise selection of variants in linkage equilibrium with $r^2 < 0.8$. The selection process should

be conducted for every gene and thus requires an expensive computing cost. Thus, it might be convenient to select variants ($r^2 < 0.8$) with an arbitrary significance threshold ($P < 0.05$).

The variance components for polygenic and environmental effects are usually estimated by employing REML prior to estimating fixed and random effects. For example, variance components can be estimated by maximizing the log restricted likelihood (Harville, 1977; Searle, 1979) as follows:

$$l_r \propto -\frac{1}{2}(\log|\mathbf{G}\sigma_g^2| + \sigma_\varepsilon^{2n} + |\mathbf{C}| + \mathbf{y}'\mathbf{P}\mathbf{y})$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$ and $\mathbf{C} = \begin{pmatrix} \mathbf{X'X} & \mathbf{X'} \\ \mathbf{X} & \mathbf{I} + \frac{\sigma_\varepsilon^2}{\sigma_g^2}\mathbf{G}^{-1} \end{pmatrix}$, which is the coefficient matrix of Henderson's mixed model equation (MME; Henderson et al., 1959):

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'} \\ \mathbf{X} & \mathbf{I} + \frac{\sigma_\varepsilon^2}{\sigma_g^2}\mathbf{G}^{-1} \end{pmatrix} \begin{bmatrix} \beta \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{y} \end{bmatrix}.$$

Any closed form solutions for variance components are unavailable because the likelihood is non-linear, and various computing algorithms for obtaining REML estimates of variance components have been suggested as a non-trivial task. The log likelihood function presented above is efficient for obtaining REML estimates as one of the simplified forms, especially for the derivative-free algorithm incorporating the Choleski decomposition and simplex method (Boldman and Van Vleck, 1991). For detailed descriptions of a variety of variance component estimation methods (see Searle et al., 2009).

The fixed and random effects are then solved with the estimated variance components under the MME. The heavy computational burden on inverse matrices for solving MME can be avoided by Choleski decomposition or by iteration methods, such as the Jacobi and Gauss–Seidel algorithms (Lee, 2016b).

The identification of eQTL is performed by a *t*-test with 1 degree of freedom for candidate variant effects. Since a large number of tests for associations with genome-wide nucleotide variants are usually conducted, adjustments for multiple testing are employed to avoid spurious eQTLs. The most common method is the Bonferroni correction under the assumption of independence among individual tests. Researchers often use a less conservative method (e.g., false discovery rate), especially for a huge number of tests. Regardless of the number of tests, significance threshold value of $P = 5 \times 10^{-8}$ is acceptable for GWAS (Dudbridge and Gusnanto, 2008; Jannot et al., 2015). If only *cis*-regulatory eQTLs are considered, a smaller significance threshold value can be used. For example, significance threshold values used for the *cis*-eQTLs within 1 Mb from the transcription start site were $P = 2.82 \times 10^{-5}$ (Koopmann et al., 2014) and $P = 9.22 \times 10^{-5}$ (Gong et al., 2017). The selection of eQTLs for polygenic random effects might be distinguished from the eQTL identification addressed above. While the identification of eQTLs focuses on avoiding spurious eQTLs, the selection of eQTLs focuses on the appropriate reflection of polygenic effects. Thus, eQTLs might be selected without any correction for multiple testing.

ADVANTAGES OF MIXED MODELS FOR eQTL ANALYSES

The mixed model framework not only enables the identification of eQTLs by determining the statistical significance of associations with gene expression, but also shows polygenetic variance explained by nucleotide variants. Thus, genome-wide eQTL analyses using the mixed model may substantially reduce "missing heritability," which is usually attributed to inherent difference between GWAS and pedigree-based genetic analyses. Furthermore, the element of vector g indicates the relative genetic ability of each individual for gene expression.

The use of a genomic similarity matrix would help to control for population stratification, to explain polygenic effects, and thus to reduce false positive and negative genetic associations. The mixed model analysis for data simulated with a variety of designs performed better than the fixed model analysis incorporating genomic control or principal component analysis in respect to empirical type 1 error rate and statistical power (Widmer et al., 2014; Shin and Lee, 2015a). The improvement by the mixed models increased more with a highly admixed population, a large narrow-sense heritability, a small number of causal variant, or a large number of related individuals (Widmer et al., 2014; Shin and Lee, 2015a,b).

The assumption of an infinitesimal model is not required for the identical-by-state (IBS) genetic relationship (i.e., genomic similarity matrix) based on genotype information, unlike for the identical-by-descent genetic relationship based on pedigree information. That is, the IBS genetic relationship matrix can be flexibly constructed using genotype information for a customized set of selected nucleotide variants. This is useful for eQTL mapping where the cell-specific genomic similarity matrix should be constructed with different loci. Gene expression is regulated by the cell environment, and the cell environment is produced by gene expression regulation. Thus, trans-regulators as well as *cis*-regulators should be stressed to construct the genomic similarity matrix, and different genomic similarity matrices among cells are largely attributed to cell-specific trans-regulators. Cell-specific genomic similarity increases the accuracy of eQTL identification and heritability estimates explained by eQTLs. Gene-specific similarity is also required because the loci used to estimate the genomic similarity matrix vary widely in kind and size depending on gene functions. The selection of loci is determined by statistical significance for associations with gene expression using a specific significance threshold. A subjective significance threshold is employed, or the value is determined by maximizing polygenic variance estimated with the selected loci. In conclusion, a cell- and gene-specific genomic similarity matrix should be constructed for eQTL analyses, without the unjustifiable assumption of an infinitesimal model.

Using mixed models, it is feasible to extend the additive genetic analysis presented in this review to non-additive genetic analyses. For example, an analytical model may include random

dominance polygenic effects with corresponding variances, e.g., a dominance genomic similarity matrix multiplied by dominance genetic variance (Da et al., 2014). Similarly, additive-by-additive, additive-by-dominance, dominance-by-dominance, and/or higher order epistatic terms can be added, each with their own variance, e.g., $G^2\sigma^2_{g\times g}$ (Martini et al., 2016). However, careful modeling is required for epistatic analyses because unreliable results are more likely as the degree of interaction increases. Filtering according to biological relevance-, gene module-, and marginal effect-based strategies may avoid exhaustive searches for epistasis (Huang et al., 2013). Filtering also helps to overcome another challenge arising from a large number of weak signals.

Mixed models enable the partitioning of the polygenic variance by its subsets as sell as the estimation of polygenic variance with any customized set of nucleotide variants. For example, polygenic variance may be partitioned by nucleotide variants proximal and distal to the gene of interest in order to infer that they are *cis*- and *trans*- eQTLs, respectively. These *cis*- and *trans*- eQTLs might be interpreted further as potential global and cell-specific regulators (Thalayasingam et al., 2018).

ISSUES WITH ANALYTICAL MODELS

Simulation studies have shown that mixed models perform well, regardless of the use of IBD or IBS genetic relationships, showing empirical unbiasedness (Lee and Pollak, 1997a; Ryoo and Lee, 2014). The fitness of the analytical model employed for analyzing real data must be confirmed prior to the analysis. If the normality assumption is violated for gene expression, data transformations should first be considered, such as normalization and log-transformation. Alternatively, more flexible distributions might be assumed for analytical models, such as generalized linear mixed models (Breslow and Clayton, 1993) and hierarchical generalized linear models (Lee and Nelder, 1996).

The analytical model presented in this review assumes a consistent effect for every eQTL in assessing the genetic covariance structure among individuals. This unrealistic assumption might produce bias in the genetic variance component (Ryoo and Lee, 2014). Thus, heterogeneous effects can be incorporated into the model. For example, the genomic similarity may reflect a penalty based on functions for each eQTL effect size (Yi and Xu, 2008), and a Bayesian approach with priors on the number of major eQTLs is also plausible (Lee et al., 2008).

When analyzing gene expression, it is possible to use gene expression at a previous stage as a covariate in order to determine the regulatory stage. For example, an eQTL for protein level might result from the regulation of transcription or translation. This is tested by employing an analytical model with the protein level as a dependent variable and RNA level as a covariate. The test can be used to confirm a candidate gene for a phenotype by employing an analytical model with phenotype as a dependent variable and its expression level as a covariate, especially when the candidate gene (i.e., the nearest gene to the association signal) obtained by GWAS differs from the gene identified by the eQTL analysis. Paired *t*-tests with expression data obtained at two stages for every individual can avoid inflating the sampling variance for statistics estimated from the separate analysis of independent two-stage expression data for unrelated individuals. Such joint modeling also provides the proportion of phenotypic variance explained by all eQTLs identified for the expression of a specific candidate gene (Huang et al., 2014).

Sex effects might be simply treated as a covariate in the analytical model. However, heterogeneous effects of sex are also important as an interaction between eQTLs and sex, especially for genes with hormone-dependent functions. Heterogeneous eQTLs by sex can be explained by separate analyses with partitioned data or by sex-stratified bivariate mixed models (Lee, 2016a). A straightforward sex-stratified model is the two-phenotype model in which male expression is treated as one phenotype and female expression as the other (Lee et al., 1997, 2012). The heterogeneous eQTLs by sex clarify the heterogeneous genetic architecture with respect to sex for various complex phenotypes.

Joint modeling in the mixed model framework can be extended to analyze various kinds of expression simultaneously. For example, expression data for multiple tissues can be treated as different phenotypes, and these analyses provide genetic covariance components and genetic correlations between tissues. A mixed model meta-analysis can be applied to the identification of eQTLs with heterogeneous effects across multiple tissues (Sul et al., 2013).

Spurious eQTLs easily result from microarray expression data because confounding effects are induced by various measurement errors (Churchill, 2002; Akey et al., 2007). Mixed model analyses, such as the intersample correlation emended method (Kang et al., 2008), probabilistic analysis of genomic data (Fusi et al., 2012), and confounding factor estimation through independent component analysis (Ju et al., 2017), have been suggested to correct for the confounding effects. These analytical models incorporate random effects with an intersample covariance structure that might explain unknown confounding factors produced by measurement errors.

ISSUES RELATED TO PARAMETER ESTIMATION

The mixed model is usually applied with the statistical property of best linear unbiased estimator (BLUE, a solution of fixed effects) and best linear unbiased predictor (BLUP, a solution of random effects) derived from Henderson's mixed model equations. The BLUE and BLUP are, however, only valid under the assumption of known variance components. In reality, variance components are unknown for specific data, and should be estimated with the data used for BLUE and BLUP. This contradictory assumption has not limited its practical application. This might be because employing REML works well for the practical estimation of variance components. In order to actively address with the problem, BLUE can be estimated simultaneously with the variance and covariance components by a Bayesian approach implemented with MCMC methods, such as Gibbs sampling

TABLE 2 | Useful software for genome-wide eQTL analysis using mixed models.

Program	Method and algorithm	Website (http)	MA[1]	Source code	Reference
GCTA	Average information restricted maximum likelihood (AIREML)	cnsgenomics.com/software/gcta	Δ	C++	Yang et al., 2011
GEMMA	Newton–Raphson restricted maximum likelihood (NRREML) Bayesian using Metropolis and Hastings	www.xzlab.org/software.html	O	C++	Zhou and Stephens, 2014 Zhou et al., 2013
TASSEL	NRREML	www.maizegenetics.net/tassel	X	Java	Zhang et al., 2010
MTG2	AIREML	sites.google.com/site/honglee0707/mtg2	O	FORTRAN	Lee and van der Werf, 2016
GENSEL	Bayesian using Gibbs sampling	archive.is/bigs.ansci.iastate.edu	X	C++	Kizilkaya et al., 2010
MMAP	AIREML, NRREML, Expectation-maximization restricted maximum likelihood, Fisher information restricted maximum likelihood	mmap.github.io	X	Undisclosed	O'Connell, 2014
FaST-LMM	Maximum likelihood[2], Restricted maximum likelihood[2]	www.microsoft.com/en-us/research/project/fastlmm	X	Python	Lippert et al., 2011

[1] *Multivariate analysis; Δ indicates that bivariate analysis is available.*
[2] *Spectral decomposition-based algorithm is employed.*

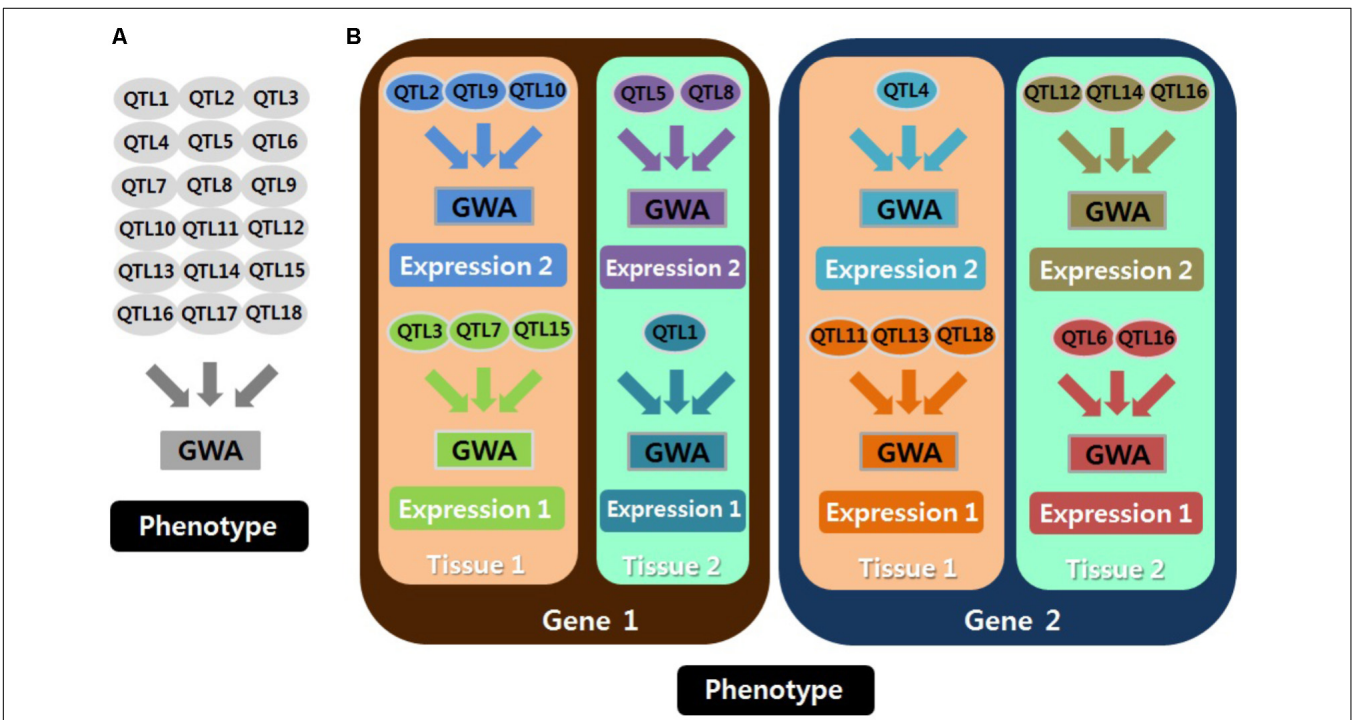

FIGURE 1 | Schematic concepts for analyzing quantitative trait locus (QTL) associated with phenotypes **(A)** and with gene expression **(B)**. Each genome-wide association analysis (GWAS) is presented in a different color. Many GWAS should be conducted for QTL associated with gene expression. Thus, employing a mixed model is important because every GWA requires a different genomic similarity matrix, and the cellular environment for gene expression is largely the result of genetic effects. Expression 1 and expression 2 are distinguished by different kinds or stages of expression (e.g., RNA expression and protein expression).

and the Metropolis and Hastings algorithm (Gilks et al., 1995). Another advantage of Bayesian methods is that they reflect uncertainty in unknown parameters, such as variance components for the analytical model, by treating the parameters as random variables. As a result, the Bayesian approaches provide a probability distribution called the posterior for each parameter. This enables us to make straightforward inferences for the parameters. For example, specific credible intervals for every parameter considered in an analytical model can be directly obtained using the samples of posterior distribution generated by the MCMC. The posterior avoids doubts about undesirable local ML estimates produced from a frequentist approach.

In practice, a search for the maximum of a likelihood function is mathematically and computationally challenging.

SOFTWARE

Mixed model methods for GWAS have been implemented with a variety of software. Most of them provide REML and Bayesian estimates of parameters, and some useful software for genome-wide eQTL analysis are presented in **Table 2**. In particular, GEMMA and TASSEL employed the Newton–Raphson algorithm using observed Fisher information matrix (i.e., Hessian matrix) as the second derivative of likelihood for REML (Zhang et al., 2010; Zhou and Stephens, 2014), and GCTA and MTG2 employed the average information algorithm using both of the Hessian matrix and Fisher information matrix (Yang et al., 2011; Lee and van der Werf, 2016). The algorithms were both used in MMAP (O'Connell, 2014).

CLOSING REMARKS

Mixed models are important for GWAS to explain polygenic effects and thus to avoid population stratification. In particular, polygenic effects for gene expression might be more sensitive than those for phenotypes. This is because the cellular environment for gene expression results largely from genetic effects, and noise produced by a long process from genotype to phenotype is decreased in gene expression analyses. Furthermore, accurate analyses are essential to identify specific regulatory stages and functions of eQTLs for gene expression, and eQTLs can be specified with the corresponding technique, i.e., chromatin modification eQTL (Degner et al., 2012; Grubert et al., 2015), including DNase I sensitivity QTL (dsQTL), methylation QTL (meQTL), and histone QTL (hQTL); transcriptional eQTL (Lappalainen et al., 2013; Li et al., 2016), including narrow-sense eQTL, splicing QTL (sQTL), transcript ratio QTL (trQTL), miRNA QTL (mirQTL), allele specific expression QTL (aseQTL), RNA synthesis rate QTL (rsQTL), and RNA decay QTL (rdQTL); chromatin interaction eQTL (Tang et al., 2015), including chromatin interaction QTL (cQTL) and promoter enhancer interaction QTL (peQTL); and translational eQTL (Battle et al., 2015), including ribosome occupancy QTL (rQTL) and protein abundance QTL (pQTL). The accuracy of eQTL identification and parameter estimation depends on the customized genomic similarity matrix for each genome-wide analysis by expressed molecules and tissues as well as by genes (**Figure 1**). Of course, the specified analyses can be extended to the identification of any other potential heterogeneity in eQTLs. An example is age-dependent eQTLs, which may explain the heterogeneous heritability of complex phenotypes by age (Lee and Lee, 2015). Thus, employing a mixed model should be emphasized to reduce spurious eQTLs in genome-wide eQTL analyses.

The use of the mixed model for genome-wide eQTL analyses provides more reliable results than conventional fixed model analyses. Of course, accuracy will be further improved by decreasing errors produced from current RNA-seq techniques and costs (e.g., sequencing death). Understanding the genetic architecture of complex phenotypes will be accelerated by genome-wide eQTL analyses using mixed models with a profile of transcriptome-wide gene expression for the activity of a single cell and further with multiple profiles across cells with different functions.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (Grant No. NRF-2018R1A2B6004867).

REFERENCES

Akey, J. M., Shameek, B., Leek, J. T., and Storey, J. D. (2007). On the design and analysis of gene expression studies in human populations. *Nat. Genet.* 39, 807–808. doi: 10.1038/ng0707-807

Bartholomew, D. J., Steele, F., Galbraith, J., and Moustaki, I. (2008). *Analysis of Multivariate Social Science Data*. Boca Raton, FL: Chapman and Hall.

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., et al. (2015). Impact of regulatory variation from RNA to protein. *Science* 347, 664–667. doi: 10.1126/science.1260793

Boldman, K. G., and Van Vleck, L. D. (1991). Derivative-free restricted maximum likelihood estimation in animal models with a sparse matrix solver. *J. Dairy Sci.* 74, 4337–4343. doi: 10.3168/jds.S0022-0302(91)78629-3

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88, 9–25.

Brown, K. G. (1976). Asymptotic behavior of MINQUE-like estimators of variance components. *Ann. Stat.* 73, 141–146. doi: 10.1093/biostatistics/kxs024

Casella, G., and Berger, R. L. (1990). *Statistical Inference*. Pacific Grove, CA: Wadsworth & Brooks.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32, 490–495. doi: 10.1038/ng1031

Da, Y., Wang, C., Wang, S., and Hu, G. (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* 9:e87666. doi: 10.1371/journal.pone.0087666

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394. doi: 10.1038/nature10808

Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* 32, 227–234. doi: 10.1002/gepi.20297

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fusi, N., Stegle, O., and Lawrence, N. D. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.* 8:e1002330. doi: 10.1371/journal.pcbi.1002330

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., et al. (2017). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res.* 46, D971–D976. doi: 10.1093/nar/gkx861

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., et al. (2015). Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* 162, 1051–1065. doi: 10.1016/j.cell.2015.07.048

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Hartley, H. O., and Rao, J. N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* 54, 93–108. doi: 10.1093/biomet/54.1-2.93

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.10480998

Henderson, C. R. (1950). Estimation of genetic parameters. *Ann. Math. Stat.* 21, 309–310.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* 9, 226–252. doi: 10.2307/3001853

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15, 192–218. doi: 10.2307/2527669

Huang, Y., Wuchty, S., and Przytycka, T. M. (2013). eQTL epistasis–challenges and computational approaches. *Front. Genet.* 4:51. doi: 10.3389/fgene.2013.00051

Huang, Y. T., VanderWeele, T. J., and Lin, X. (2014). Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann. Appl. Stat.* 8, 352–376. doi: 10.1214/13-AOAS690

Jannot, A. S., Ehret, G., and Perneger, T. (2015). P $<$ 5 $\times$ 10-8 has emerged as a standard of statistical significance for genome-wide association studies. *J. Clin. Epidemiol.* 68, 460–465. doi: 10.1016/j.jclinepi.2015.01.001

Jensen, J., and Mao, I. L. (1991). Estimation of genetic parameters using sampled data from populations undergoing selection. *J. Dairy Sci.* 74, 3544–3551. doi: 10.3168/jds.S0022-0302(91)78546-9

Johnson, D. L., and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* 78, 449–456. doi: 10.3168/jds.S0022-0302(95)76654-1

Ju, J. H., Shenoy, S. A., Crystal, R. G., and Mezey, J. G. (2017). An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci. *PLoS Comput. Biol.* 13:e1005537. doi: 10.1371/journal.pcbi.1005537

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kang, H. M., Ye, C., and Eskin, E. (2008). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180, 1909–1925. doi: 10.1534/genetics.108.094201

Kennedy, W. J., and Gentle, J. E. (1980). *Statistical Computing*. New York, NY: Marcel Dekker.

Kizilkaya, K., Fernando, R. L., and Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88, 544–551. doi: 10.2527/jas.2009-2064

Koopmann, T. T., Adriaens, M. E., Moerland, P. D., Marsman, R. F., Westerveld, M. L., Lal, S., et al. (2014). Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. *PLoS One* 9:e97380. doi: 10.1371/journal.pone.0097380

Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *J. Am. Stat. Assoc.* 82, 97–105. doi: 10.1080/01621459.1987.10478395

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531

Lee, C. (2016a). Analytical models for genetics of human traits influenced by sex. *Curr. Genom.* 17, 439–443. doi: 10.2174/1389202917666160420142601

Lee, C. (2016b). Best linear unbiased prediction of individual polygenic susceptibility to sporadic vascular dementia. *J. Alzheimers Dis.* 53, 1115–1119. doi: 10.3233/JAD-160391

Lee, C., and Pollak, E. J. (1997a). Influence of partitioning data by sex on genetic variance and covariance components for weaning weight in beef cattle. *J. Anim. Sci.* 75, 61–67.

Lee, C., and Pollak, E. J. (1997b). Relationship between sire $\times$ year interactions and direct-maternal genetic correlation for weaning weight of Simmental cattle. *J. Anim. Sci.* 75, 68–75.

Lee, C., and Pollak, E. J. (2002). Genetic antagonism between body weight and milk production in beef cattle. *J. Anim. Sci.* 80, 316–321. doi: 10.2527/2002.802316x

Lee, C., Van Tassell, C. P., and Pollak, E. J. (1997). Estimation of genetic variance and covariance components for weaning weight in Simmental cattle. *J. Anim. Sci.* 75, 325–330. doi: 10.2527/1997.752325x

Lee, D., and Lee, C. (2015). Age- and gender-dependent heterogeneous proportion of variation explained by SNPs in quantitative traits reflecting human health. *Age* 37:19. doi: 10.1007/s11357-015-9756-2

Lee, S. H., and van der Werf, J. H. (2016). MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* 32, 1420–1422. doi: 10.1093/bioinformatics/btw012

Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* 4:e1000231. doi: 10.1371/journal.pgen.1000231

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28, 2540–2542. doi: 10.1093/bioinformatics/bts474

Lee, Y., and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. R. Stat. Soc. B* 58, 619–678.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604. doi: 10.1126/science.aad9417

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835. doi: 10.1038/nmeth.1681

Martini, J. W., Wimmer, V., Erbe, M., and Simianer, H. (2016). Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor. Appl. Genet.* 129, 963–976. doi: 10.1007/s00122-016-2675-5

Melbourne, B. A., and Hastings, A. (2008). Extinction risk depends strongly on factors contributing to stochasticity. *Nature* 454, 100–103. doi: 10.1038/nature06922

O'Connell, J. R. (2014). *MMAP User Guide*. Baltimore, MD: University of Maryland.

Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554. doi: 10.1093/biomet/58.3.545

Quaas, R. L. (1976). Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* 32, 949–953. doi: 10.2307/2529279

Quaas, R. L. (1988). Additive genetic model with groups and relationships. *J. Dairy Sci.* 71, 1338–1345. doi: 10.3168/jds.S0022-0302(88)79691-5

Quaas, R. L., and Pollak, E. J. (1980). Mixed model methodology for farm and ranch beef cattle testing programs. *J. Anim. Sci.* 51, 1277–1287. doi: 10.2527/jas1981.5161277x

Rao, C. R. (1971). Minimum variance quadratic unbiased estimation of variance components. *J. Multivar. Anal.* 1, 445–456. doi: 10.1016/0047-259X(71)90019-4

Ryoo, H., and Lee, C. (2014). Underestimation of heritability using a mixed model with a polygenic covariance structure in a genome-wide association study for complex traits. *Eur. J. Hum. Genet.* 22, 851–854. doi: 10.1038/ejhg.2013.236

Searle, S. R. (1979). *Notes on Variance Component Estimation: A Detailed Account of Maximum Likelihood and Kindred Methodology*. Technical Report BU-673-M. Ithaca, NY: Cornell Univ.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance Components*. New York, NY: John Wiley and Sons.

Shin, J., and Lee, C. (2015a). A mixed model reduces spurious genetic associations produced by population stratification in genome-wide association studies. *Genomics* 105, 191–196. doi: 10.1016/j.ygeno.2015.01.006

Shin, J., and Lee, C. (2015b). Statistical power for identifying nucleotide markers associated with quantitative traits in genome-wide association analysis using a mixed model. *Genomics* 105, 1–4. doi: 10.1016/j.ygeno.2014.11.001

Spielman, R. S., Bastone, L. A., Burdick, J. T., Morley, M., Ewens, W. J., and Cheung, V. G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* 39, 226–231. doi: 10.1038/ng1955

Sul, J. H., Han, B., Ye, C., Choi, T., and Eskin, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* 9:e1003491. doi: 10.1371/journal.pgen.1003491

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163, 1611–1627. doi: 10.1016/j.cell.2015.11.024

Thalayasingam, N., Nair, N., Skelton, A. J., Massey, J., Anderson, A. E., Clark, A. D., et al. (2018). CD4+ and B lymphocyte expression quantitative traits at rheumatoid arthritis risk loci in patients with untreated early arthritis. *Arthritis Rheumatol.* 70, 361–370. doi: 10.1002/art.40393

Van Tassell, C. P., Casella, G., and Pollak, E. J. (1995). Effects of selection on estimates of variance components using Gibbs sampling and restricted maximum likelihood. *J. Dairy Sci.* 78, 678–692. doi: 10.3168/jds.S0022-0302(95)76680-2

Wang, C. S., Rutledge, J. J., and Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genet. Sel. Evol.* 25, 41–62. doi: 10.1186/1297-9686-25-1-41

Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., et al. (2014). Further improvements to linear mixed models for genome-wide association studies. *Sci. Rep.* 4:6874. doi: 10.1038/srep06874

Wilson, A. J., Pemberton, J. M., Pilkington, J. G., Coltman, D. W., Mifsud, D. V., Clutton-Brock, T. H., et al. (2006). Environmental coupling of selection and heritability limits evolution. *PLoS Biol.* 4:e216. doi: 10.1371/journal.pbio.0040216

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* 46, 100–106. doi: 10.1038/ng.2876

Yang, S., Liu, Y., Jiang, N., Chen, J., Leach, L., Luo, Z., et al. (2014). Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics* 15:13. doi: 10.1186/1471-2164-15-13

Yi, N. J., and Xu, S. H. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179, 1045–1055. doi: 10.1534/genetics.107.085589

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genomewide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* 9:e1003264. doi: 10.1371/journal.pgen.1003264

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* 11, 407–409. doi: 10.1038/nmeth.2848

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.