DOI: 10.1534/genetics.107.071977

Note

Increasing the Efficiency of Variance Component Quantitative Trait Loci Analysis by Using Reduced-Rank Identity-by-Descent Matrices

Lars Rönnegård,*,[1] Kateryna Mischenko,†,‡ Sverker Holmgren† and Örjan Carlborg*

**Linnaeus Centre for Bioinformatics, Uppsala University, SE-75124 Uppsala, Sweden, †Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-75124 Uppsala, Sweden and ‡Department of Mathematics and Physics, Mälardalen University, SE-72123 Västerås, Sweden*

Manuscript received February 12, 2007
Accepted for publication April 21, 2007

ABSTRACT

Recent technological development in genetics has made large-scale marker genotyping fast and practicable, facilitating studies for detection of QTL in large general pedigrees. We developed a method that speeds up restricted maximum-likelihood (REML) algorithms for QTL analysis by simplifying the inversion of the variance–covariance matrix of the trait vector. The method was tested in an experimental chicken pedigree including 767 phenotyped individuals and 14 genotyped markers on chicken chromosome 1. The computation time in a chromosome scan covering 475 cM was reduced by 43% when the analysis was based on linkage only and by 72% when linkage disequilibrium information was included. The relative advantage of using our method increases with pedigree size, marker density, and linkage disequilibrium, indicating even greater improvements in the future.

THE use of variance component models is rapidly increasing in the field of QTL analysis (Lynch and Walsh 1998). As the cost for genotyping decreases, the sizes of the analyzed pedigrees are likely to increase, making full genome scans computationally slow or even infeasible. However, current algorithms commonly used for variance component estimation were not specifically developed for QTL analysis and there is a need to re-evaluate the computational efficiency and robustness of these algorithms.

Variance component estimation has been included in general statistical software, such as Proc Mixed in SAS (Littell *et al.* 1996), where an arbitrary covariance structure of the random effect can be given by the user. These programs have in common that they use iterative procedures, Fisher's scoring or Newton–Raphson, to maximize the likelihood or the restricted likelihood (Pawitan 2001). More specific programs for applications in animal breeding have also been developed over the last two decades such as ASReml, DMU, and VCE (see Druet and Ducrocq 2006 and references therein). These programs use a mixture of Fisher's scoring and Newton–Raphson to maximize the restricted likelihood, called the "average information restricted maximum-likelihood (AI-REML) algorithm." The most computationally demanding part of AI-REML is the inversion of the variance–covariance matrix (**V**) of the response vector (**y**). This inversion has to be performed on each iteration. We study a model with a random QTL effect and a residual effect, with variance–covariance matrix of the form $\mathbf{V} = \Pi\sigma_v^2 + \mathbf{I}\sigma_e^2$, where Π is the symmetric identity-by-descent (IBD) matrix, σ_v^2 is the QTL variance, **I** is the identity matrix, and σ_e^2 is the residual variance. If Π is positive definite, then the inversion of **V** can be simplified by inverting **V** in parts. This has been implemented in the software package ASReml by setting up the mixed-model equations (MME) (Gilmour *et al.* 1995; Johnson and Thompson 1995; Jensen *et al.* 1997).

The AI-REML algorithm can be implemented by combining the MME with sparse matrix techniques (as done in ASReml), which gives fast solutions when the covariance structure of the random effect is sparse and positive definite. In traditional animal breeding applications, the covariance structure is given by the average relationship between individuals (Lynch and Walsh 1998). This covariance structure is usually sparse and always positive definite. The IBD matrix is not necessarily sparse or positive definite, however, and the advantage of using sparse matrix techniques together with inversion of **V** by means of MME in AI-REML may be questioned. Lee and van der Werf (2006) found that the AI-REML algorithm could be faster and more robust in QTL analysis if direct inversion of **V** is used, especially in linkage-disequilibrium linkage (LDL) mapping

[1] *Corresponding author:* Linnaeus Centre for Bioinformatics, Uppsala University, SE-75124 Uppsala, Sweden. E-mail: lars.ronnegard@lcb.uu.se

(Meuwissen and Goddard 2000) since the IBD matrices used in LDL are usually dense and positive semidefinite. The reason for this is that a covariance structure is added to the base generation alleles (Meuwissen and Goddard 2000; Hernandez-Sanchez *et al.* 2006), which increases the number of nonzero elements in Π.

The rank of Π at a marker depends on the size of the base generation and how polymorphic the marker is (Rönnegård and Carlborg 2007). In a QTL linkage analysis, the rank of Π is twice the size of the base generation when the marker is fully informative (*i.e.*, all marker alleles are unique) and the rank does not depend on the total pedigree size, whereas the number of rows (and columns) in Π equals the total number of individuals in the pedigree, n. Hence, at marker locations, Π will have many eigenvalues equal to zero, and the number of zero-valued eigenvalues increases with the difference between the total number of individuals and the number of base individuals. In nonmarker locations, the number of eigenvalues in Π that approaches zero, when the distance to the marker decreases, is equal to twice the difference between the total number of individuals and the number of base individuals. Thus, for a dense marker map most eigenvalues in all IBD matrices will be either equal to zero (in marker positions) or close to zero (in nonmarker positions).

In this article we develop a fast genome scan method for variance component QTL analysis using AI-REML. The method utilizes an efficient inversion of $\mathbf{V}$ that takes advantage of the fact that Π has many eigenvalues close to zero.

An efficient inversion of V using the Sherman–Morrison–Woodbury formula: A simple mixed linear model for QTL detection is given by $\mathbf{y} = \mathbf{X}\beta + \mathbf{v} + \mathbf{e}$, where $\mathbf{X}$ and β are the design matrix and parameter vector, respectively, for the fixed effects, $\mathbf{v}$ is the vector of QTL genotype effects (length n), $\mathbf{v} \sim \text{MVN}\ (0, \Pi\sigma_v^2)$, and $\mathbf{e} \sim \text{MVN}(0, \mathbf{I}\sigma_e^2)$. Thus the variance–covariance matrix of $\mathbf{y}$ is $\mathbf{V} = \Pi\sigma_v^2 + \mathbf{I}\sigma_e^2$. Now let the spectral decomposition of Π be $\Gamma\mathbf{D}\Gamma^{\text{T}}$, where Γ is the matrix of eigenvectors and $\mathbf{D}$ is the diagonal eigenvalue matrix (superscript T denotes matrix transpose). Using this decomposition, we approximate Π. Let $\mathbf{D}_{\text{red}}$ be a submatrix of $\mathbf{D}$, where the eigenvalues larger than a threshold value τ are included, and Γ_{red} be the matrix of eigenvectors corresponding to these eigenvalues. Then an approximate IBD matrix with reduced rank is given by $\Pi_{\text{red}} = \Gamma_{\text{red}}\mathbf{D}_{\text{red}}\Gamma_{\text{red}}^{\text{T}}$. In our applications τ was set equal to λ_i for which the cumulative sum of λ_1 to λ_i divided by the total sum of eigenvalues was 0.8, where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the ordered eigenvalues from largest to smallest. Using the Sherman–Morrison–Woodbury formula (Golub and van Loan 1996) we then get an efficient equation for inverting $\mathbf{V}$: $\mathbf{V}^{-1} \approx \mathbf{I}(1/\sigma_e^2) - (\sigma_v^2/(\sigma_e^2)^2)\Gamma_{\text{red}}\mathbf{D}_{\text{red}}\left(\mathbf{I}_{\text{red}} + (\sigma_v^2/\sigma_e^2)\mathbf{D}_{\text{red}}\right)^{-1}\Gamma_{\text{red}}^{\text{T}}$.

Here $\mathbf{I}_{\text{red}}$ is the identity matrix of the same size as $\mathbf{D}_{\text{red}}$. This approximation dramatically decreases the number of mathematical operations needed to invert $\mathbf{V}$. Let k be the rank of Π_{red}; then the number of floating point arithmetic operations (flops) in calculating the approximated inverse is $n^2k + 3nk^2$, whereas the number of flops to invert $\mathbf{V}$ directly is $n^3/2$. The spectral decomposition of Π requires on the order of n^3 flops, but this computation is performed only once in each REML estimation, whereas the inversion of $\mathbf{V}$ is performed in each iteration. As an example suppose $n = 500$, $k = 20$, and that AI-REML converges in 10 iterations; then the total number of flops for inverting $\mathbf{V}$ directly is 6.25×10^8 with the direct method and 1.81×10^8 with the reduced method, giving a 3.5-fold speedup.

Two important questions are then: How are the maximum log-likelihood values and the variance component estimates affected by the approximation? How much is the computational time reduced in a genome scan in practice? To answer these questions we tested the method on chicken chromosome 1 (475 cM) in a Jungle Fowl–White Leghorn F_2 cross, where the measured trait was body weight at 200 days of age. Kerje *et al.* (2003) reported two QTL for this trait on chromosome 1 at 68 and 420 cM. There were 4 F_0, 41 F_1, and 767 F_2 individuals in the pedigree. In our analysis, population mean and sex were included as fixed effects. There were 14 genotyped markers located at 0, 27.7, 35.3, 91.3, 124.3, 154.2, 189.7, 209.3, 233.0, 258.8, 337.4, 407.9, 425.9, and 475.4 cM. The IBD matrices were estimated at every 1 cM using the Markov chain Monte Carlo-based program package Loki (Heath 1997; Heath *et al.* 1997). The estimated IBD matrices were dense with >80% nonzero elements (elements were defined as nonzero for values $\geq 10^{-3}$). The likelihood-ratio (LR) statistic in AI-REML was calculated as twice the difference between the maximized $\log(L)$ and $\log(L)$ with $\sigma_v^2 = 0$. The convergence criterion used was change in log-likelihood $<10^{-4}$, with $\log(L) = -\frac{1}{2}(\log|\mathbf{V}| + \log|\mathbf{X}^{\text{T}}\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}^{\text{T}}\mathbf{P}\mathbf{y})$ and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\text{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}\mathbf{V}^{-1}$, and the starting value of σ_v^2 was 0.01 times the residual variance. The computer code to calculate Π_{red} and the AI-REML algorithm was implemented in R (R Development Core Team 2004).

Accuracy and efficiency of the method: At individual chromosomal locations, our method was up to five times faster than the direct inversion of $\mathbf{V}$ and taken over the whole chromosome it reduced the computation time by 43% (Figure 1). It was faster at all tested locations, except at positions >30 cM from the closest marker. The rank of Π_{red} was smallest close to marker positions, where also the greatest speedups were achieved.

The correlation between the LR values obtained with Π and Π_{red} was 0.9999. Both methods resulted in maximum LR values at 54 and 426 cM (Figure 2). The relative difference in LR between the two methods was 0.5 and 4.3% at these locations, respectively. In our experience, the shape of the likelihood-ratio curve is not substantially affected by the inclusion of polygenic

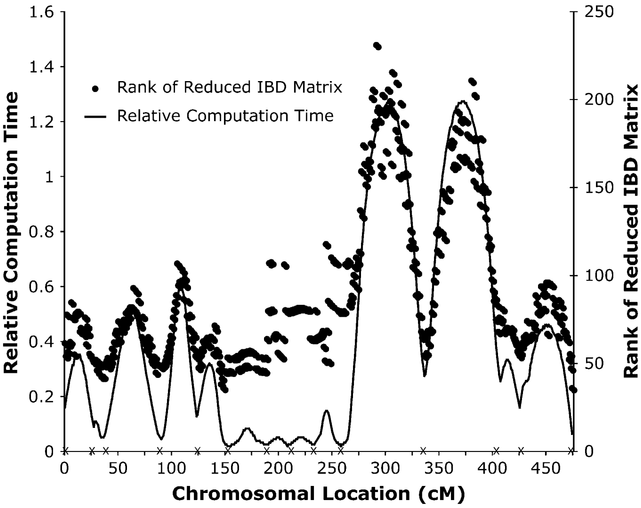

FIGURE 1.—The rank of Π_{red} (right axis) and the relative computation time (left axis) along chicken chromosome 1. The relative computation time is the ratio of the total number of flops used in AI-REML for the inversion of **V** by the reduced method and direct inversion. This is $[n^3 + N_{\mathrm{red}}(n^2k + 3nk^2)]/[N_{\mathrm{full}}n^3/2]$, where N_{red} and N_{full} are the number of iterations until convergence using Π_{red} and Π, respectively, n is the number of observations, and k is the rank of Π_{red}. A marker position is given by an "X" along the x-axis.

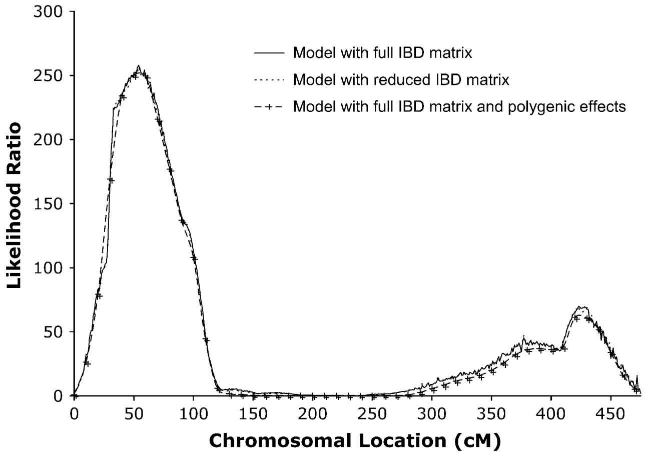

FIGURE 2.—QTL likelihood-ratio curves for body weight at 200 days of age along chicken chromosome 1 for three different variance component models.

effects in the variance component QTL model, which was confirmed in our analysis (Figure 2). Both the full model including polygenic effects and our method identified the same two peaks, with a relative difference in LR of 2.7% at 54 cM and 8.1% at 426 cM.

In LDL mapping, a correlation structure is added to the base generation individuals in a pedigree with a varying degree of positive correlation between alleles. In a second analysis of our chicken data, we computed Π assuming fixation within lines. This is an extreme case of LDL mapping, where the base generation alleles are assumed fully correlated within lines. Assuming fixation of QTL alleles within lines is equivalent to fitting a line effect, and for a fully informative marker the rank of Π will therefore be 2. The rank of Π_{red} was consequently reduced further in this example, where the rank was 2 in all positions except at positions >30 cM from the closest marker. In this case our method was up to five times faster than direct inversion in specific positions and reduced the computations over the whole chromosome by 72%. Both models with either Π or Π_{red} gave similar LR values (correlation 0.9999) with QTL at 60 and 430 cM. The relative differences in LR were 1.2 and 0.7%, respectively, for the two QTL.

Conclusions: Our efficient AI-REML method approximates the likelihood-ratio statistic very well. It is, therefore, a useful method to locate the regions in the genome with the strongest support for a QTL. The efficiency of the method increases when a covariance structure is added to the base generation alleles, which is the case in LDL mapping. This increase in efficacy was illustrated with an extreme covariance structure where the QTL alleles were assumed fixed within lines.

For accurate testing and estimation of the variance components at the QTL identified using this fast method, we suggest that a full model including polygenic effects should be used. Our results indicate that inclusion of polygenic effects is not important in the variance component model when the aim is to detect the location of potential QTL. Other studies have shown that the risk of getting false positives increases in QTL studies if polygenic effects are not included (KENNEDY *et al.* 1992), but this will not be a problem in our suggested application.

MEUWISSEN and GODDARD (2000) did not include polygenic effects in their presentation of their LDL-mapping model, but accounted for different kinds of population substructures by modeling a general residual variance–covariance matrix **R**, where $\mathbf{V} = \Pi\sigma_v^2 + \mathbf{R}\sigma_e^2$. This is also possible in our model, and **V** may then be inverted using the more general version of the Sherman–Morrison–Woodbury formula (GOLUB and VAN LOAN 1996) $\mathbf{V}^{-1} = \mathbf{R}^{-1}(1/\sigma_e^2) - (\sigma_v^2/(\sigma_e^2)^2)\mathbf{R}^{-1}\Gamma_{\mathrm{red}}(\mathbf{I}_{\mathrm{red}} + (\sigma_v^2/\sigma_e^2)\Gamma_{\mathrm{red}}^{\mathrm{T}}\mathbf{R}^{-1}\Gamma_{\mathrm{red}}\mathbf{D}_{\mathrm{red}})^{-1}\Gamma_{\mathrm{red}}^{\mathrm{T}}\mathbf{R}^{-1}$. The inversion of **R** has to be performed only once for the whole genome scan, since it does not change between positions, and the loss in computation time merely depends on how sparse $\mathbf{R}^{-1}$ is.

We have not developed the method for several QTL (see LEE and VAN DER WERF 2006). In principle, the method should be extendable to several QTL, either by developing a more general version of the Sherman–Morrison–Woodbury formula or by developing an orthogonal model where the variance components can be tested one at a time (following THOMSEN 1975).

In our analysis we used Loki to estimate Π. If, however, an approximate decomposition of Π could be rapidly estimated directly from the marker data, then the spectral decomposition would be redundant and the speed

TABLE 1

Pedigree structure of a commercial chicken pedigree and an outbred pig cross used for QTL mapping

	Pedigree structure		Computational reduction	
	Size of base	Total size	% highly informative markers[a]	% fully informative markers[b]
Commercial chicken pedigree[c]	146	2.708	43	71
Duroc–Landrace/Yorkshire cross[d]	710	11.000	36	64

The reduction in the number of flops to invert **V** with our method compared to direct inversion is given for two cases: (1) highly informative and moderately dense markers (<10 cM apart) and (2) a fully informative marker at each tested position.

[a] Rank of Π_{red} is twice the size of the base generation, the number of AI-REML iterations is assumed to be seven, and spectral decomposition of Π is included in calculations.

[b] Rank of Π is twice the size of the base generation, and decomposition is estimated directly from the data.

[c] Pedigree structure is from Rowe *et al.* (2006).

[d] Pedigree structure is from M. S. Lund (personal communication; Danish Institute of Agricultural Sciences).

of AI-REML would increase significantly. This decomposition can be made in LDL mapping if the number of possible haplotypes is limited, as shown by Meuwissen and Goddard (2000). Furthermore, Rönnegård and Carlborg (2007) have recently developed a general method for estimating a decomposition of Π directly from marker information. Their method is based on single-marker information, but is in principle extendable to a multiple-marker framework. In our analysis of chicken chromosome 1 the number of flops to invert **V** would have been decreased by 98% if a decomposition of Π had been estimated directly. In Table 1, we have also compared the potential of our method when applied to an existing commercial chicken pedigree (Rowe *et al.* 2006) and an outbred pig cross (M. S. Lund, personal communication; Danish Institute of Agricultural Sciences).

In conclusion, the efficiency of our method will increase when the ratio between the total pedigree size and base generation size increases, the density and informativeness of markers increases, and the correlation between base generation alleles increases. Hence, the relative efficacy of the method can be expected to increase in the future as deeper pedigrees and more markers become available.

LITERATURE CITED

Druet, T., and V. Ducrocq, 2006 Innovations in software packages in quantitative genetics. Paper no. 27-10. World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil.

Gilmour, A. R., R. Thompson and B. R. Cullis, 1995 Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics **51:** 1440–1450.

Golub, G. H., and C. Van Loan, 1996 *Matrix Computations*, Ed. 3. Johns Hopkins University Press, Baltimore.

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Heath, S. C., G. L. Snow, E. A. Thompson, C. Tseng and E. M. Wijsman, 1997 MCMC segregation and linkage analysis. Genet. Epidemiol. **14:** 1011–1015.

Hernandez-Sanchez, J., C. S. Haley and J. A. Woolliams, 2006 Prediction of IBD based on population history for fine gene mapping. Genet. Sel. Evol. **38:** 231–252.

Jensen, J., E. A. Mantysaari, P. Madsen and R. Thompson, 1997 Residual maximum likelihood estimation of (co)variance components in multivariate mixed linear models using average information. J. Indian Soc. Agric. Stat. **49:** 215–236.

Johnson, D. L., and R. Thompson, 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.

Kennedy, B. W., M. Quinton and J. A. Van Arendonk, 1992 Estimation of effects of single genes on quantitative traits. J. Anim. Sci. **70:** 2000–2012.

Kerje, S., Ö. Carlborg, L. Jacobsson, K. Schütz, C. Hartmann *et al.*, 2003 The twofold difference in adult size between the red junglefowl and White Leghorn chickens is largely explained by a limited number of QTLs. Anim. Genet. **34:** 264–274.

Lee, S. H., and J. H. J. van der Werf, 2006 An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genet. Sel. Evol. **38:** 25–43.

Littell, R. C., G. A. Milliken, W. W. Stroup and R. D. Wolfinger, 1996 *SAS System for Mixed Models.* SAS Institute, Cary, NC.

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

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics **155:** 421–430.

Pawitan, Y., 2001 *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press, Oxford.

R Development Core Team, 2004 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna (http://www.R-project.org).

Rönnegård, L., and Ö. Carlborg, 2007 Separation of base allele and sampling term effects gives new insights in variance component QTL analysis. BMC Genet. **8:** 1.

Rowe, S., R. Pong-Wong, C. S. Haley, S. Knott and D. J. de Koning, 2006 Variance component estimation of additive and dominance QTL effects in outbred populations applied to commercial broilers. Paper no. 20-09. World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil.

Thomsen, I., 1975 Testing hypotheses in unbalanced variance components models for two-way layouts. Ann. Stat. **3:** 257–265.

Communicating editor: E. Arjas