"SNP Snappy": A Strategy for Fast Genome-Wide Association Studies Fitting a Full Mixed Model

Karin Meyer[1] and Bruce Tier
Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351, Australia

ABSTRACT A strategy to reduce computational demands of genome-wide association studies fitting a mixed model is presented. Improvements are achieved by utilizing a large proportion of calculations that remain constant across the multiple analyses for individual markers involved, with estimates obtained without inverting large matrices.

GENOME-WIDE association studies (GWAS) have become a routine task for geneticists in a range of areas. Analyses employing a mixed model are widely used as this provides a flexible framework to account for systematic differences and covariances due to other sources, such as population stratification and a family structure among genotyped individuals (Kang *et al.* 2010; Price *et al.* 2010; Zhang *et al.* 2010). A common type of investigation involves solving a system of mixed model equations (MME) fitting one or a few single nucleotide polymorphism markers (SNP) at a time, treating SNP effects as covariables, with variance components fixed at their estimates from an analysis omitting SNP. Typically this is done by inverting the coefficient matrix in the MME for each analysis. While individual analyses take only seconds, analyzing all markers for a high-density chip imposes a considerable computational burden. Hence, estimation of SNP effects by first fitting the mixed model excluding any SNP effects and then applying the SNP-wise analysis to the resulting residuals has been suggested (Aulchenko *et al.* 2007). However, this may lead to biased results if genotypes are not randomized across the effects in the model or if SNP effects and population strata are partially confounded. A typical example is an analysis comprising animals of different breeds with different allele frequencies (Johnston and Graser 2010).

When fitting the full model, we can partition the pertaining MME into a small part due to SNP effects and a part due to the other effects in the model. For complete genotype information only the former changes as different SNPs are considered. This structure can be exploited to reduce computational requirements. We present the strategy to do so, describe its implementation in freely available mixed model software, and show an example application.

doi: 10.1534/genetics.111.134841
Manuscript received September 21, 2011; accepted for publication October 19, 2011
[1]Corresponding author: Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351, Australia. E-mail: kmeyer.agbu@gmail.com

Computing Strategy

Consider a mixed model

$$\mathbf{y} = \mathbf{X}\mathbf{b}^k + \mathbf{Z}\mathbf{u}^k + \mathbf{W}^k\mathbf{s}^k + \mathbf{e}^k, \tag{1}$$

with $\mathbf{y}$, $\mathbf{b}^k$, $\mathbf{u}^k$, $\mathbf{s}^k$, and $\mathbf{e}^k$ denoting the vector of observations (phenotypes), fixed effects other than SNP effects, random effects, SNP effects and residuals, and $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}^k$ the incidence matrices pertaining to $\mathbf{b}^k$, $\mathbf{u}^k$, and $\mathbf{s}^k$. As emphasized by the superscript k, only $\mathbf{W}^k$ differs between analyses for different SNPs, with the elements of $\mathbf{W}^k$ equal to the number of copies of the reference allele—0, 1, or 2 in a biallelic model—for the SNP(s) in the kth analysis. To estimate $\hat{\mathbf{s}}^k$ we need to solve the MME

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W}^k \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{W}^k \\ \mathbf{W}^{k'}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}^{k'}\mathbf{R}^{-1}\mathbf{Z} & \mathbf{W}^{k'}\mathbf{R}^{-1}\mathbf{W}^k \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}}^k \\ \hat{\mathbf{u}}^k \\ \hat{\mathbf{s}}^k \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}^{k'}\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}, \tag{2}$$

with $\mathbf{R} = \mathrm{Var}(\mathbf{e})$ and $\mathbf{G} = \mathrm{Var}(\mathbf{u})$ the covariance matrices among residuals and random effects, respectively. Rewrite Equation 2 as

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12}^k \\ \mathbf{C}_{21}^k & \mathbf{C}_{22}^k \end{pmatrix} \begin{pmatrix} \hat{\mathbf{v}}^k \\ \hat{\mathbf{s}}^k \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2^k \end{pmatrix}, \tag{3}$$

with $\mathbf{C}_{11}$ of size $n_1 \times n_1$ denoting the part of the coefficient matrix that is constant and $\mathbf{r}_1$ the pertaining vector of right-hand sides, $\mathbf{C}_{22}^k$, of size $n_2 \times n_2$, $\mathbf{r}_2^k$ the corresponding terms

for the effects changing with each analysis, and $\mathbf{C}_{12}^k$ and $\mathbf{C}_{21}^k$ the off-diagonal blocks in the coefficient matrix. With $\hat{\mathbf{v}}^k$ generally not of interest, we can estimate $\hat{\mathbf{s}}^k$ as a solution to

$$(\mathbf{C}_{22}^k - \mathbf{C}_{21}^k\mathbf{C}_{11}^{-1}\mathbf{C}_{12}^k)\hat{\mathbf{s}}^k = \mathbf{r}_2^k - \mathbf{C}_{21}^k\mathbf{C}_{11}^{-1}\mathbf{r}_1. \quad (4)$$

With n_2 small, inversion of the coefficient matrix and direct solution of Equation 4 is undemanding. While $\mathbf{C}_{11}^{-1}$ remains constant and thus needs only to be determined once, computations for the inversion are proportional to n_1^3 and thus can be nontrivial for large n_1. Fortunately, we can obtain $\hat{\mathbf{s}}^k$ without inverting $\mathbf{C}_{11}$. Let

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{21}^k & \mathbf{L}_{22}^k \end{pmatrix} \quad (5)$$

denote the Cholesky factor of the coefficient matrix in Equation 3 with $\mathbf{C}_{11} = \mathbf{L}_{11}\mathbf{L}_{11}'$, $\mathbf{C}_{21}^k = \mathbf{L}_{21}^k\mathbf{L}_{11}'$ and $\mathbf{C}_{22}^k = \mathbf{L}_{21}^k\mathbf{L}_{21}^{k\prime} + \mathbf{L}_{22}^k\mathbf{L}_{22}^{k\prime}$. Substituting these terms in Equation 4 yields

$$\mathbf{L}_{22}^k\mathbf{L}_{22}^{k'}\,\hat{\mathbf{s}}^k = \mathbf{r}_2^k - \mathbf{L}_{21}^k\hat{\mathbf{t}}_1 \qquad \text{with } \hat{\mathbf{t}}_1 = \mathbf{L}_{11}^{-1}\mathbf{r}_1. \quad (6)$$

This suggests that estimates $\hat{\mathbf{s}}^k$ for $k = 1, \ldots, K$ can be obtained efficiently by splitting computations as follows.

To be performed once

- Set up $\mathbf{C}_{11}$ and $\mathbf{r}_1$, *i.e.*, the MME omitting SNP effects.
- Perform the Cholesky factorization of $\mathbf{C}_{11}$ to obtain $\mathbf{L}_{11}$.
- Determine $\hat{\mathbf{t}}_1$ as a solution to $\mathbf{L}_{11}\hat{\mathbf{t}}_1 = \mathbf{r}_1$. With $\mathbf{L}_{11}$ triangular, this involves forward substitution steps

$$\hat{t}_1 = r_1/\ell_{11} \qquad \text{and}$$

$$\hat{t}_i = \left(r_i - \sum_{j=1}^{i-1}\ell_{ij}\hat{t}_j\right)/\ell_{ii} \text{ for } i = 2, n_1$$

for r_i and $\hat{t}_i$, the ith element of $\mathbf{r}_1$ and $\hat{t}_1$, and ℓ_{ij} the ijth element of $\mathbf{L}_{11}$.

To be performed for each set of SNPs

- Determine parts of the MME specific to the kth analysis, $\mathbf{C}_{21}^k$, $\mathbf{C}_{22}^k$, and $\mathbf{r}_2^k$.
- Set up $\mathbf{L}^*$, representing the intermediate matrix arising in factorizing the coefficient matrix in Equation 3 after rows 1 to n_1 have been processed:

$$\mathbf{L}^* = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{C}_{21}^k & \mathbf{C}_{21}^k \end{pmatrix}$$

- "Complete" the factorization steps for rows $n_1 + 1$ to $n_1 + n_2$ using

$$\ell_{ij}^* = \left(\ell_{ij}^* - \sum_{k=1}^{j-1}\ell_{ik}^*\ell_{jk}^*\right)/\ell_{jj}^* \text{ for } j<i \quad \text{and}$$

$$\ell_{ii}^* = \sqrt{\ell_{ii}^* - \sum_{k=1}^{i-1}(\ell_{ik}^*)^2}$$

(for ℓ_{ij}^* the ijth element of $\mathbf{L}^*$). Processing columns 1 to n_1 column-wise replaces $\mathbf{C}_{21}^k$ in $\mathbf{L}^*$ with $\mathbf{L}_{21}^k$. The remaining elements (in columns $n_1 + 1$ to $n_1 + n_2$) are then adjusted row-wise, overwriting $\mathbf{C}_{22}^k$ with $\mathbf{L}_{22}^k$.

- Determine a general inverse of $\mathbf{L}_{22}^k$, $\mathbf{L}_{22}^{k-}$, to obtain $\hat{\mathbf{s}}^k = \mathbf{L}_{22}^{k-}(\mathbf{L}_{22}^{k-})'(\mathbf{r}_2^k - \mathbf{L}_{21}^k\hat{\mathbf{t}}_1)$. Sampling variances of $\hat{\mathbf{s}}^k$ are given by the diagonal elements of $\mathbf{L}_{22}^{k-}(\mathbf{L}_{22}^{k-})'$.

Note that $\mathbf{L}_{22}^k$ can have diagonal elements of zero if a SNP is monomorphic or if SNPs with proportional allele counts are considered simultaneously. This is accounted for in the generalized inverse. If n_2 is not small or if sampling variances are not required, an alternative to solve for $\hat{\mathbf{s}}^k$ is a series of forward and backward substitution steps.

Implementation

The strategy described above has been implemented in the mixed model package WOMBAT (Meyer 2007), utilizing the existing capabilities to set up the MME for an arbitrary model and sparse matrix calculations, including Cholesky factorization of the coefficient matrix. Estimation of SNP effects is invoked through a run-time option. In addition to the data, pedigree, and parameter files as required for standard analyses, allele counts for each SNP analysis are expected to be read sequentially from a separate file. The software and user manual together with a worked example illustrating its use for GWAS analyses are available for download from http://didgeridoo.une.edu.au/km/wmbdownloads.php.

Application

Our strategy was applied to estimate effects for 4541 SNPs on age at first *corpus luteum* in beef cattle. Any missing allele counts were imputed so that marker information was complete. Records were a subset of data analyzed previously (Hawken *et al.* 2011). The model of analysis fitted five fixed effects and two linear covariables as well as a linear regression on a single SNP effect. Animals' additive genetic effects were fitted as random effects with the relationship matrix determined from pedigree information. There were 941 animals with genotypes and phenotypes. Including additive genetic effects for parents without records yielded a total of 3858 animals in the model and 3909 equations in total.

Calculations were carried out on a desktop computer with an Intel I7 processor rated at 3.2 GHz. Performing single SNP analyses one by one, inverting the complete

coefficient matrix in the MME each time required a total of 1784 sec CPU time. Estimation using our new strategy reduced this to 16 sec. Repeating SNP information 150 times to mimic a high-density chip with 681,150 SNPs, analysis was completed in 2295 sec. With a set-up time required of ~1 sec, this gave an average of 297 SNPs analyzed per second.

Conclusions

Computational demands for GWAS analyses fitting a full mixed model can be reduced by orders of magnitude utilizing that a large part of the MME and computations involved remain constant. The computing strategy described to exploit this is straightforward and is readily implemented in existing mixed model software. Savings that can be achieved increase with the number of effects in the mixed model and are proportional to the number of SNP effects considered.

Acknowledgment

This work was supported by Meat and Livestock Australia under grant B.BFG.0050. We are indebted to the CRC for Beef Genetic Technologies for the data used in the example. AGBU is a joint venture of New South Wales Department of Primary Industries and the University of New England.

Literature Cited

Aulchenko, Y. S., D.-J. de Koning, and C. Haley, 2007 Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177: 577–585.

Hawken, R. J., Y. Zhang, M. R. S. Fortes, E. Collis, A. Reverter *et al.*, 2011 Dissecting the genetics underlying reproduction rate in tropically adapted beef cattle. In *Applied Genomics for Sustainable Livestock Breeding*. Sir Mark Oliphant Conferences, Melbourne.

Johnston, D. J., and H.-U. Graser, 2010 Estimated gene frequencies of GeneSTAR markers and their size of effects on meat tenderness, marbling, and feed efficiency in temperate and tropical beef cattle breeds across a range of production systems. J. Anim. Sci. 88: 1917–1935.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42: 348–354.

Meyer, K., 2007 WOMBAT – a tool for mixed model analyses in quantitative genetics by REML. J. Zhejiang Uni. SCIENCE B 8: 815–821.

Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson, 2010 New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11: 459–463.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42: 355–360.

Communicating editor: G. A. Churchill