University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Faculty Papers and Publications in Animal Science | Animal Science Department

3-2-2007

Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood

Z. Zhang
Cornell University, Ithaca, NY

E. S. Buckler
Cornell University, Ithaca, NY

L. Dale Van Vleck
University of Nebraska - Lincoln, dvan-vleck1@unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/animalscifacpub

Part of the Animal Sciences Commons

Zhang, Z.; Buckler, E. S.; and Van Vleck, L. Dale, "Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood" (2007). *Faculty Papers and Publications in Animal Science.* Paper 124.
http://digitalcommons.unl.edu/animalscifacpub/124

This Article is brought to you for free and open access by the Animal Science Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Papers and Publications in Animal Science by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood

Z. Zhang,* R. J. Todhunter,† E. S. Buckler,*‡ and L. D. Van Vleck§[1]

*Institute for Genetic Diversity, Cornell University, Ithaca, NY 14853; †Department of Clinical Sciences, Cornell University, Ithaca, NY 14853; ‡USDA-ARS Cornell University, 159 Biotechnology Bldg, Ithaca, NY 14853; and §USDA-ARS A218 Animal Sciences, University of Nebraska, Lincoln 68583-0908

ABSTRACT: The widespread use of the set of multiple-trait derivative-free REML programs for prediction of breeding values and estimation of variance components has led to significant improvement in traits of economic importance. The initial version of this software package, however, was generally limited to pedigree-based relationships. With continued advances in genomic research and the increased availability of genotyping, relationships based on molecular markers are obtainable and desirable. The addition of a new program to the set of multiple-trait derivative-free REML programs is described that allows users the flexibility to calculate relationships using standard pedigree files or an arbitrary relationship matrix based on genetic marker information. The strategy behind this modification and its design is described. An application is illustrated in a QTL association study for canine hip dysplasia.

Key words: breeding value estimation, multiple-trait derivative-free restricted maximal likelihood, quantitative trait locus association study, variance component estimation

©2007 American Society of Animal Science. All rights reserved.

J. Anim. Sci. 2007. 85:881–885
doi:10.2527/jas.2006-656

INTRODUCTION

As advances in genomic research lead to increased availability and affordability of genotyping, relationships based on molecular markers are becoming more obtainable in animal and plant breeding programs. Zhang et al. (2006b) showed that these marker-based relationships can be superior to conventional pedigree-based relationships in controlling for false positives in QTL association studies. In lieu of probabilities based on identity by descent, marker-based kinships may provide a better alternative in situations where pedigrees are not deep, are not complete (e.g., the father or mother is unknown), or are unknown.

To exploit these sources of genetic information, many algorithms have been developed to estimate kinships based on genetic markers (Loiselle et al., 1995; Ritland, 1996; Hardy, 2003). Several software packages have also been developed, including SPAGeDi (Hardy and Vekemans, 2002) and TASSEL (Zhang et al., 2006a).

Although widely used to estimate variance components, to predict breeding values, and to study associations between markers and QTL, the initial version of the set of multiple-trait derivative-free REML (**MTDFREML**) programs was generally limited to the use of relationships based on pedigree only (e.g., Moody et al., 1999; Riley et al., 2002; Blott et al., 2003; Ge et al., 2003; Kim et al., 2004). One notable exception is Bromley et al. (2000), who successfully employed the MTDFREML programs to compare models with line effects treated as random or fixed factors by using coefficients of coancestry (one-half additive relationships) among 58 corn inbred lines to calculate the inverse of the coancestry matrix.

Modification of MTDFREML to include an alternate version of the first program of the set of programs will now allow users to calculate relationships based on standard pedigree files or on relationships derived directly from molecular markers.

This technical note describes the necessary modifications and includes an empirical example of their implementation.

MATERIALS AND METHODS

The MTDFREML software package consists of 3 executable programs: MTDFNRM, MTDFPREP, and MTDFRUN (Boldman et al., 1995). The first program (1) calculates the inverse of the relationship matrix to be used in the mixed model equations and makes use of the Henderson (1975, 1976) and Quaas (1976) rules to

[1]Corresponding author: lvanvleck@unlnotes.unl.edu
Received September 26, 2006.
Accepted November 1, 2006.

calculate the inverse of the relationship matrix directly from a list of animals and their parents, (2) provides individual identification for matching phenotypic records to individuals, (3) calculates inbreeding coefficients, and (4) calculates the logarithm of the determinant of the relationship matrix needed to calculate the logarithm of the likelihood function. The second program of the set prepares coefficients for the mixed model equations based on the statistical model (fixed and random factors) for single and multiple trait analyses. The third program solves the mixed model equations and finds variance component estimates that maximize the restricted likelihood given the phenotypic data.

Because the 3 programs are sequentially executed, a program was added that allows incorporation of marker-based relationships (Yu et al., 2006). This new program is run before the second program and is independent of the original first program, except that the output of both must be in the same form. Thus, the change in the architecture of MTDFREML allows users to start from the original first program or the alternative first program before continuing on with the second and the third programs.

The new program, called MTDFARM, receives input in the form of coefficients of a relationship matrix that can be a marker-based relationship matrix or any arbitrary relationship matrix (**ARM**), such as the numerical relationship matrix calculated from a pedigree. The input file contains coefficients of the upper half of the arbitrary relationship matrix, with the location of each relationship coefficient identified by row and column numbers. The 3 fields (row, column, and relationship coefficient) are delimited by spaces, and any set having a relationship coefficient of zero can be omitted from the file. The row and column numbers of the relationship matrix also serve as identification for individuals. Therefore, the individual identification in the file of phenotypic records should correspond to the row and column numbers in the relationship matrix beginning with 1, 2, ...

After the upper half of the relationship matrix is read, the matrix is inverted using a Gaussian algorithm, and the lower half stored elements of the inverse are then saved in a binary file (MTDF44) to be used as input for the third program (MTDFRUN). The determinant of the relationship matrix is also calculated and written at the beginning of the binary file for use in calculating the log likelihood in MTDFRUN.

The new program, MTDFARM, also writes a numerical file (MTDF11) required by the second program (MTDFPREP), with the first record containing the number of animals (**n**) in the relationship matrix. The records that follow have 2 integer fields, with the first containing recoded identification (1, , n) and the second containing the original identification, which in this case is identical to the first field. This file is used by MTDFPREP to create equation numbers for first (and second) animal genetic effects for all animals included in the relationship matrix, regardless of whether or not they have a record.

The tests of MTDFARM were in conjunction with the 3 original programs using a variety of empirical and simulated data. Application to a QTL association study for canine hip dysplasia is described.

Animals and Pedigree

All procedures involving animals were approved by the Cornell Institutional Animal Care and Use Committee.

A research colony of Labrador Retrievers has been maintained at the Baker Institute at Cornell University for more than 30 yr. A total of 116 Labrador Retrievers were sampled, including 12 founders with unknown parents. Of these 12 founders, 8 (4 males and 4 females) were crossed with 7 Greyhound founders (2 males and 5 females) selected from racing stock for development of a pedigree of crossbred animals for use in mapping QTL for hip dysplasia. The crossbred pedigree consisted of 143 progeny over 4 generations ($F_1 \times$ Greyhound and Labrador Retriever founders, F_2, and $\frac{3}{4} \times \frac{3}{4}$ Labrador Retriever). A total of 266 dogs comprised the entire Labrador and crossbreed pedigree.

Marker-Based Relationships

The marker-based relationship matrix was calculated as a kinship matrix using the method of Loiselle et al. (1995), as implemented by the software package SPAGeDi (Hardy and Vekemans, 2002) using 471 microsatellite markers. The markers were described by Mateescu et al. (2005).

Candidate Markers

A total of 171 SNP markers, covering the midportion of canine chromosome 29 over a 25 cM region, were genotyped by the Biotechnology Research Center at Cornell University. The DNA was isolated from whole anticoagulated blood using standard protocols and genotyped using the SNPlex genotyping system (Applied Biosystems, Foster City, CA). Results for one of these markers are presented below.

Phenotypes

Hip radiographs were taken at 8 mo of age to obtain the distraction (laxity) index, dorsolateral subluxation score, and Norberg angle as described by Mateescu et al. (2005). The Norberg angle measured on the left side was used for this example.

Statistical Analyses

Genotypes for the SNP marker were fitted as fixed effects. The model also included sex and breed group as fixed effects. A random variable for animal was included to capture any remaining polygenic effects. The

Table 1. Estimates of parameters and twice the negative of the logarithm of the likelihood from the 3 analyses[1]

Parameter	PED	MAR	IBD
σ_a^2	0.00327	0.85198	0.00327
σ_e^2	37.91204	37.03153	37.91204
h^2	0.00009	0.02249	0.00009
–2logL	1,110.95331	1,110.87255	1,110.95331

[1]PED uses the usual pedigree file to compute the elements of inverse of additive relationship matrix with MTDFNRM; MAR uses marker alleles to compute a mean relationship matrix to be inverted to obtain a marker-based inverse of the relationship matrix with MTDFARM; and IBD uses an additive relationship matrix computed for identity by descent from the pedigree to be inverted to obtain the inverse of the relationship matrix with MTDFARM.

Table 2. Estimates of fixed effects and statistics for testing significance from the 3 analyses[1]

Factor	PED	MAR	IBD
Sex			
Male	0.35185	–0.22598	0.35185
Female	1.16336	1.05684	1.16336
t-statistic[2]	1.014	1.040	1.014
Breed group[3]			
G	–0.40277	–0.26154	–0.40277
L	–3.06634	–2.84668	–3.06634
F_1	–3.73444	–3.63222	–3.73444
F_2	1.60419	1.74645	1.60419
BG	0.00000	0.00000	0.00000
BL	2.93819	3.01070	2.93819
¾L × ¾L	–3.19287	–3.12475	–3.19287
F-statistic[4]	1.893	1.811	1.893
Genotype			
CC	–1.56526	–1.55532	–1.56526
CT	–2.57350	–2.57088	–2.57350
TT	0.00000	0.00000	0.00000
F-statistic[5]	3.857	3.799	3.857

[1]PED uses usual pedigree file to compute elements of inverse of additive relationship matrix through MTDFNRM; MAR uses marker alleles to compute a mean relationship matrix to be inverted to obtain a marker-based inverse of the relationship matrix through MTDFARM; and IBD uses an additive relationship matrix computed for identity by descent from the pedigree to be inverted to obtain the inverse of the relationship matrix with MTDFARM.
[2]To test the difference due to sex.
[3]G = Greyhound; L = Labrador Retriever; BG = backcross to the Greyhound; and BL = backcross to the Labrador Retriever.
[4]To test the hypothesis of no differences among breed groups, 6 df.
[5]To test the hypothesis of no differences among genotypes, 2 df.

covariance matrix for the random animal genetic effects was defined as the additive relationship matrix multiplied by the additive genetic variance, which is an unknown scalar estimated by MTDFRUN. The residuals were assumed to be identically and independently distributed with unknown variance, which is also estimated by MTDFRUN. Three analyses were conducted. Analysis I (PED) used the original first program of the set of MTDFREML programs (MTDFNRM) to build the inverse of the relationship matrix from the standard pedigree file of animal, sire of animal (if known), and dam of animal (if known). Analysis II (MAR) used the new program (MTDFARM) to incorporate marker-based relationships. For analysis III (IBD), the additive relationship matrix from identity by descent was first calculated by using the tabular method (Cruden, 1949; Emik and Terrill, 1949), with founder animals as well as those with records, and then MTDFARM was utilized to obtain the inverse of the additive relationship matrix. Analyses I and III must yield identical results (Henderson, 1975; Quaas, 1976) because they are equivalent models.

RESULTS AND DISCUSSION

As expected, results obtained from using the new program (MTDFARM) with relationships calculated from the pedigree (analysis III) were identical to those obtained from using MTDFNRM, which also relies on the standard pedigree (analysis I). Differences arising between these 2 analyses and the analysis using MTDFARM with marker-based relationships (analysis II) are most likely attributable to incomplete pedigree information because the analysis contained 19 founder dogs (12 Labrador Retrievers and 9 Greyhounds with unknown parents). As a result, the pedigree-based relationship matrix may have been less informative than the marker-based matrix and did seem to capture less genetic variance (see Tables 1 and 2).

The new alternative first program also provides MTDFREML users with the flexibility to use information from pedigrees that do not adhere to conventional format of sire and dam or of sire and maternal grandsire. A common example where this would be desirable would include development of a synthetic line from more than 2 parent lines, which frequently occurs in poultry.

For the input of a marker-based relationship matrix into MTDFARM, animals must be coded 1, , n. Not all animals in the relationship matrix, however, need to have a record in the data file. The animal identification in the phenotypic data file 1) must match the 1, , n coding, or 2) the second field of the MTDF11 output file from MTDFARM must be modified to contain the uncoded (integer) animal identification in the data file to match the 1, , n coding in the first column. The only field in the first record of the MTDF11 file is the integer corresponding to the number of animals in the relationship matrix.

The only field in the first record of the binary file, MTDF44, required by MTDFRUN, generated by MTDFNRM or the new MTDFARM, is one-half the logarithm of the determinant of the relationship matrix, which is used in MTDFRUN to calculate $-2\log L \mid y$. The records that follow in MTDF44 contain the lower half stored elements of the inverse of the relationship matrix in the form of 2 integer fields and 1 decimal field: row (i), column (j), and coefficient (i,j). These are used in forming the lower half stored coefficients of the mixed model equations. With rules for A-inverse as used by MTDFNRM, more than 1 coefficient for a row and col-

umn may be created that will be summed by MTDFRUN (e.g., the sire, dam coefficients for animals with the same sire and dam).

Some Questions

Although the MTDFARM program allows the MTDFREML software package to use molecular markers to establish additive genetic relationships among animals for random additive genetic effects, several questions require further investigation.

One advantage to using an ARM based on mean relationships from many markers is that knowledge of parent-progeny relationships is not required, which might be the case for wild populations or populations lacking birth date or mating information. In the process of segregation even in the absence of selection, there are wide deviations in relatedness. For example, an offspring could be twice as related to one grandparent as to another. Selection toward 1 parental phenotype during breeding only increases this variability in segregation. The use of a large number of markers is the only reliable way to track the vagaries of segregation.

What is clear, however, which can be illustrated by the canine study, is that the arbitrary relationship matrix can greatly increase the density (fraction of nonzero elements) of the coefficient matrix. The method of Henderson (1975, 1976) and Quaas (1976), which uses an animal, sire of animal, dam of animal file to directly compute, without calculation of the relationship matrix, the inverse elements of the relationship matrix needed for the coefficient matrix of the mixed model equations, creates remarkably few nonzero coefficients. In the canine example with 266 animals with potentially 35,511 half-stored elements, only 700 were not zero. Perhaps even more remarkable, however, is that all except 2 coefficients were not zero for the inverse of the marker-based relationship matrix. This creates a dense coefficient matrix that leads to a filled Choleski factor, which in turn negates benefits derived from the sparse matrix methods used in MTDFRUN. Such density is not critical for a relatively small data set, but with thousands of individuals (as compared with the 266 used in the example), the number of computational steps required to achieve convergence for estimates of variance components or to find solutions for fixed genotypic effects may become overwhelming. It may be worthwhile to explore setting some of the smallest marker-based estimates of relatedness or inverse elements to zero in order to make the analysis more tractable. If possible, it could be even more worthwhile to find an efficient algorithm to calculate the inverse of the marker-based relationship matrix directly from markers.

Some research by Yu et al. (2006) suggests that marker-based relationships can be used to reduce type I error and increase the power of the test for association mapping. This question, however, needs to be investigated over larger and more diverse samples than illustrated here. In addition, the effectiveness of marker-based relationships for predicting additive genetic values needs to be investigated.

An obvious difficulty with ARM based on markers would occur if, for example, 2 genetic lines derived from the same sources were identical for a set of markers. That would be the same situation as for identical twins and would lead to a singular relationship matrix. A singular relationship matrix could also result from use of a small number of markers. Thus, a few hundred markers randomly distributed through the genome would be desirable for solving the singularity problem and more importantly would provide good estimates of relationships.

Use of markers to validate or invalidate identification of parents (or progeny) may be important for use of pedigree-based methods such as Henderson-Quaas (Henderson, 1976 and Quaas, 1976). Such use would allow for identification of progeny of sires in multiple sire pastures or of dams when many are calving (lambing, etc.) at the same time. In such cases, the original MTDFNRM program could be used with the corrected pedigree.

In conclusion, the options of 1) use of an arbitrary relationship matrix derived from markers in common or from portions of a full relationship matrix or 2) use of traditional use of rules for the inverse of the augmented relationship matrix allow for flexibility in accounting for polygenic effects for analyses of genomic data or for estimating polygenic breeding values and components of variance due to polygenic effects.

ACKNOWLEDGMENTS

We thank D. White and N. Stevens for technical editing of this manuscript. The Fortran code for the MTDFREML programs is available from the USDA by e-mail from the corresponding author (lvanvleck@unlnotes.unl.edu). The Fortran code for MTDFARM is available from the senior author (zz19@cornell.edu) or the corresponding author. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

LITERATURE CITED

Blott, S., J. J. Kim, S. Moisio, A. Schmidt-Kntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, M. Georges, F. Farnir, and W. Coppieters. 2003. Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163:253–266.

Boldman, K. G., L. A. Kriese, L. D. VanVleck, C. P. Van Tassell, and S. D. Kachman. 1995. A manual for use of MTDFREML. A set of programs to obtain estimates of variance and covariance. USDA, Agriculture Research Service, Clay Center, NE. (DRAFT)

Bromley, C. M., L. D. Van Vleck, B. E. Johnson, and O. S. Smith. 2000. Estimation of variance components due to line effects from F_1 performance with and without pedigree relationships among lines. Crop Sci. 40:651–655.

Cruden, D. 1949. The computation of inbreeding coefficients for closed populations. J. Hered. 40:248–251.

Emik, L. O., and C. E. Terrill. 1949. Systematic procedures for calculating inbreeding coefficients. J. Hered. 40:51–55.

Ge, W., M. E. Davis, H. C. Hines, K. M. Irvin, and R. C. M. Simmen. 2003. Association of single nucleotide polymorphisms in the growth hormone and growth hormone receptor genes with blood serum insulin-like growth factor I concentration and growth traits in Angus cattle. J. Anim. Sci. 81:641–648.

Hardy, O. J. 2003. Estimation of pair wise relatedness between individuals and characterization of isolation by distance processes using dominant genetic markers. Mol. Ecol. 12:1577–1588.

Hardy, O. J., and X. Vekemans. 2002. SPAGeDi: A versatile computer program to analyze spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618–620.

Henderson, C. R. 1975. A rapid method for computing inverse of a relationship matrix. J. Dairy Sci. 58:1727–1730.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83.

Kim, N. K., Y. W. Seo, G. H. Kim, J. H. Joh, O. H. Kim, E. R. Chung, and C. S. Lee. 2004. A previously unreported DraI polymorphism within the regulatory region of the bovine growth hormone gene and its association with growth traits in Korean Hanwoo cattle. Anim. Genet. 35:152–154.

Loiselle, B., V. Sork, J. Nason, and C. Graham. 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am. J. Bot. 82:1420–1425.

Mateescu, R. G., Z. Zhang, K. L. Tsai, J. Phavaphutanon, N. I. Burton-Wurster, G. Lust, R. L. Quaas, K. E. Murphy, G. M. Acland, and R. J. Todhunter. 2005. Analysis of allele fidelity, polymorphic information content, and density of microsatellites in a genome wide screen for hip dysplasia in a cross-breed pedigree. J. Hered. 96:847–853.

Moody, D. E., D. Pomp, M. K. Nielsen, and L. D. Van Vleck. 1999. Identification of quantitative trait loci influencing traits related to energy balance in selection and inbred lines of mice. Genetics 152:699–711.

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32:949–953.

Riley, D. G., C. C. Chase, Jr., A. C. Hammond, R. L. West, D. D. Johnson, T. A. Olson, and S. W. Coleman. 2002. Estimated genetic parameters for carcass traits of Brahman cattle. J. Anim. Sci. 80:955–962.

Ritland, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. Camb. 67:175–185.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208.

Zhang, Z., P. J. Bradbury, D. E. Kroon, T. M. Casstevens, and E. S. Buckler. 2006a. TASSEL 2.0: A software package for association and diversity analyses in plants and animals. Plant and Anim. Genomes XIV Conf. January 14-18, 2006, San Diego, CA. http://www.intl-pag.org/14/abstracts/PAG14_C012.html Accessed Jan. 12, 2007.

Zhang, Z., R. G. Mateescu, G. Lust, J. Phavaphutanon, K. Tsai, K. Murphy, R. J. Todhunter, and E. S. Buckler. 2006b. Association mapping accounting for background QTLs: Relationship based on pedigree vs. molecular markers. Proc. 8th World Conf. Genetics Appl. Livest. Prod., Belo Horizonte, Brazil, Aug. 13–18, 2006. CD-ROM communication No. 20-08.