Genomic Measures of Relationship and Inbreeding

P.M. VanRaden
Animal Improvement Programs Laboratory, Agricultural Research Service,
United States Department of Agriculture, Beltsville, MD, USA 20705-2350

Abstract

Models that include genomic relationships can predict genetic effects more accurately than those that use expected relationships from pedigrees. Relationship matrices can estimate the expected fraction of genes identical by descent, the actual fraction of DNA shared, or the fraction of alleles shared for loci that affect a particular trait. Each may be a valid answer to the question "Are two individuals related?" Several options are available for including genomic relationships in genetic evaluations.

Introduction

Genomic relationship matrix **G** uses genotypic data to estimate the fraction of total DNA that two individuals share. Measures of genomic similarity are useful in selection and parentage testing (Dodds *et al.,* 2005) and have been used to manage genetic diversity (Caballero and Toro, 2002) because of advantages over measures of genetic distance. Estimators of genomic relationships were compared (Wang, 2002) and provided similar results unless the population was small or the markers had many alleles.

Additive genetic relationship matrix **A** uses only pedigree data to calculate probabilities that gene pairs are identical by descent (Wright, 1922). Malécot (1948) derived the same probabilities without crediting Wright and showed how to correct the probabilities for mutation, which "is insignificant for close relatives" but could become important "when very distant ancestors are involved." Expected relationships are widely used in statistical analyses and in animal breeding because the matrix inverse is sparse and simple to obtain (Henderson, 1975), even for millions of individuals.

Relationship matrix **T** estimates fractions of alleles of quantitative trait loci (QTL) that two individuals share only for loci that affect a specific trait. The term QTL often refers to loci with the largest effects but includes all loci that affect the trait in this paper. Matrix **T** requires both phenotypic and genotypic data to estimate QTL locations and allele effects, which in most cases cannot all be known. However, marker genotypes may be weighted across loci by size of allele effects to estimate their total genomic effect on a trait.

How related are relatives?

The proportion of alleles that is identical by descent is a function of the number of loci that influence the trait. If only one locus is considered, full sibs have a 0.25 chance of sharing two alleles, 0.5 chance of sharing one allele, and 0.25 chance of sharing neither allele. With two loci, the probabilities are 0.0625, 0.25, 0.375, 0.25, and 0.0625 of sharing zero, one, two, three, or four alleles, respectively. The general formula for k alleles in common with n independent loci (and with ! denoting factorial) is $0.5^n n!/[k!(n-k)!]$. For larger numbers of loci, the distribution of full-sib alleles in common approximates normal with $50\% \pm 50\%/(2n)^{0.5}$. Results for full sibs and for half sibs, who have one rather than both parents in common and share half as many genes, are in Table 1.

Table 1. Mean and standard deviation of alleles shared by full sibs and by half sibs.

Independent loci	Percentage of alleles shared	
	Full-sib mean (SD)	Half-sib mean (SD)
1	50 (35.4)	25 (17.7)
5	50 (15.8)	25 (7.9)
10	50 (11.2)	25 (5.6)
50	50 (5.0)	25 (2.5)
100	50 (3.5)	25 (1.8)
Infinite	50 (0.0)	25 (0.0)

Standard deviation for percentage of alleles shared by full sibs does not decline below about 3.5% as number of loci becomes large because the loci are actually linked rather than independent. Alleles on the same chromosome are inherited together unless a crossover occurs between them, which causes closely linked genes on a chromosome segment to act as a single allele. Simulation showed that 100 unlinked multiallelic loci or 300 unlinked biallelic loci would provide the same relationship pattern as ≥10,000 linked loci that were distributed randomly across 30 pairs of chromosomes. Table 1 is based on multiallelic loci and shows that actual covariances among relatives are sufficiently different from expected values used in Wright's (1922) relationship matrix to increase reliability of quantitative predictions, especially as the numbers of relatives increase.

Relationships of parents to progeny are 50% ± 0% because each progeny receives exactly half of the two parents' autosomal DNA from the two gametes. Male or female progeny may be more related to their father or mother, respectively, if inheritance of mitochondria, genes on the X and Y sex chromosomes, or gametic imprinting are considered. Of course, genotyping mistakes, pedigree errors, and mutations may also occur. Relationships of grandparents to grandprogeny are the same as those in Table 1 for half sibs.

Unrelated individuals?

Pedigrees may include many generations but must end eventually. Traditional models assume that the base or founding individuals are unrelated and share no genes in common, but genomic analyses reveal that individuals in the earliest recorded generation always share genes from more remote ancestors. Alleles shared through a known ancestor are said to be identical by descent, and additional alleles may be alike in state if inherited from a common unknown ancestor or if mutation resulted in the same gene sequence or function. In the traditional **A**, unrelated individuals have 0 for off-diagonal elements and 1 for diagonal elements, but they may have more or fewer alleles in common and more or less heterozygosity than average when actual genotypes are examined.

Genomic relationships

Linear model predictions of total genetic effects can be obtained from either mixed model or selection index equations. Let vector **u** contain the additive genetic effects for each allele or each marker, and let **M** be the incidence matrix that specifies which alleles each individual inherited. Elements of **M** are 0 or 1 in a gametic incidence matrix, and 0, 1, or 2 in a genotypic matrix. Off-diagonals of $\mathbf{MM}'$ show the number of alleles shared by relatives, and diagonals show the individual's relationship to itself (inbreeding). In contrast, off-diagonals of $\mathbf{M}'\mathbf{M}$ show how many times two different alleles were inherited by the same individual, and diagonals show how many individuals inherited each allele. Matrix $\mathbf{M}'\mathbf{M}$ has a larger dimension than $\mathbf{MM}'$ when the total number of alleles or haplotypes is larger than the number of individuals (e.g., with dense genotyping).

Let **P** contain frequencies p_i of the second allele at each locus such that column i of **P** is $\mathbf{1}p_i$ for gametic models or $\mathbf{2}p_i$ for genotypic models. Subtraction of **P** from **M** gives **Z**, which is needed to set the expected value of **u** to 0. Subtraction of **P** gives more credit to rare alleles than to common alleles when calculating genomic relationships. Also, the genomic inbreeding coefficient is higher if the individual is homozygous for rare alleles than for common alleles.

Allele frequencies in **P** should be from the unselected base population rather than those that occur after selection or inbreeding. Gengler *et al.* (2007) presented a simple strategy to obtain such frequencies. An earlier or later base population can lead to greater or fewer relationships and to more or less inbreeding. For populations with more than one subpopulation, base relationships and inbreeding can be set to 0 for the two least related subpopulations (VanRaden, 1992).

Relationship matrix **G** is $\mathbf{ZZ}'/[2\Sigma p_i(1-p_i)]$. Division by $2\Sigma p_i(1-p_i)$ makes **G** analogous to **A**. Matrix **G** is positive semi-definite but can be singular. Two individuals can have identical genotypes if limited numbers of loci are considered, and identical twins cause singularity

even in **A**. Matrix **G** is also singular if the total number of alleles is less than the number of individuals genotyped. The rank of $\mathbf{ZZ}'$ cannot exceed the columns in **Z** if **Z** has fewer columns than rows. An improved, non-singular matrix, $\mathbf{G}_w$, can be obtained as a weighted (w) average, $w\mathbf{G} + (1-w)\mathbf{A}$, if numbers of markers are limited.

Genomic models

If each individual is measured for a trait and the inheritance of all alleles is known, then data vector **y** can be modeled as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e},$$

where **Xb** is the mean and **e** is a random error vector with variance equal to $\mathbf{R}\sigma_e^2$. The sum **Zu** over all marker loci then is assumed to equal the vector of breeding values (**a**). With many markers, that should provide a good approximation to the true, unobservable biological model $\mathbf{a} = \mathbf{Qq}$, where **Q** and **q** are the incidence matrix and effects of only loci that affect the trait.

Mixed model estimates of **u** ($\hat{\mathbf{u}}$) are obtained by using matrix $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$, vector $\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})$, and a scalar k defined as the ratio σ_e^2/σ_u^2, which equals $2\Sigma p_i(1-p_i)$ times the ratio σ_e^2/σ_a^2. Estimated breeding values $\hat{\mathbf{a}}$ then are obtained as $\mathbf{Z}\hat{\mathbf{u}}$, and the resulting equations are:

$$\hat{\mathbf{a}} = \mathbf{Z}[\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{I}k]^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}).$$

Selection index equations predict $\hat{\mathbf{a}}$ directly using **G**. Selection index equations are constructed as the covariance of **y** and **a** multiplied by the inverse of the variance of **y** multiplied by the deviation of **y** from $\mathbf{X}\hat{\mathbf{b}}$ or:

$$\hat{\mathbf{a}} = \mathbf{G}(\mathbf{G} + \mathbf{R}\sigma_e^2/\sigma_a^2)^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}).$$

The selection index and mixed model equations provide the same estimates of $\hat{\mathbf{a}}$ if the same estimates of $\mathbf{X}\hat{\mathbf{b}}$ are used (Henderson, 1963). Thus, selection index and mixed model methods should be identical in many genomic analyses because daughter yield deviations or de-regressed proofs are the data source, and the means already have been removed. However, more numerical problems may result from using mixed model equations (Lee and van der Werf, 2006). Estimates of $\hat{\mathbf{u}}$ could also be obtained if needed using the selection index equations by substituting $\mathbf{Z}'$ for the left-most **G** in the expression above, which shows again that $\hat{\mathbf{a}}$ is the sum $\mathbf{Z}\hat{\mathbf{u}}$ over all alleles that the individual inherited.

Another equivalent model presented by Garrick (2007) could be more efficient than selection index because **G** can be inverted just once and then additional traits can be processed using iteration:

$$\hat{\mathbf{a}} = (\mathbf{R}^{-1} + \mathbf{G}^{-1}\sigma_e^2/\sigma_a^2)^{-1}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}).$$

Gains in reliability from using **G** instead of **A** in the mixed model depend on how large the differences are between traditional average relationships and the more exact fractions of genes in common available from genomic studies. Non-linear models increase reliability further (Meuwissen *et al.,* 2001) by using prior information about the expected distribution of QTL effects ($\mathbf{V}_q$). In the linear model, marker effects are assumed normally distributed. In the non-linear model, large effects are regressed less and small effects more so that **T,** which equals $\mathbf{QV}_q\mathbf{Q}'$ divided by the total genetic variance, can be approximated better. Estimation of haplotype effects instead of single-marker regression can also improve accuracy.

Reliability of predictions

Average reliability and accuracy was determined from simulated data using 50,000 markers and varying numbers of full sibs. The predictions were for an individual with genotype known but no data, whereas breeding values of the sibs were assumed to be measured almost without error (reliability = 0.99) to provide an upper limit regarding the reliability that they provide for this individual. Squared correlations of breeding value with estimated breeding value for 100 replicates are in Table 2 for comparison with theoretical reliability. Linear models that include **G** or non-linear models that better approximate **T** can obtain much higher reliability than tradi-

Table 2. Average reliability using traditional (**A**) or genomic (**G**) relationships with 100 loci assumed to affect the trait.

Full sibs	Reliability	
	A	**G**
1	0.250	0.261
10	0.454	0.502
100	0.495	0.773
1000	0.499	0.970

tional models that include **A**. Villanueva *et al.* (2005) also reported that genomic relationships constructed using large numbers of markers increase reliability even if no major genes affect a trait.

Conclusions

Genetic similarity can be defined in several ways using pedigree data, genotypic data, phenotypic data, or combinations of those data. Use of exact fractions of shared genes in **G** can provide more accurate predictions than use of the expected fractions in **A**. Non-linear models can change the weights on individual markers to match actual fractions of shared QTL alleles in **T** more closely. Full sibs may actually share 45 or 55% of their DNA rather than the expected 50%. Accounting for those small differences in the relationship matrix and tracing individual genes can greatly increase reliability, especially if the number of genotyped individuals is large.

References

Caballero, A. & Toro, M.A. 2002. Analysis of genetic diversity for the management of conserved subdivided populations. *Conservation Genetics 3*, 289–299.

Dodds, K.G., Tate, M.L. & Sise, J.A. 2005. Genetic evaluation using parentage information from genetic markers. *J. Anim. Sci. 83*, 2271–2279.

Garrick, D.J. 2007. Equivalent mixed model equations for genomic selection. *J. Anim. Sci. 85* (Suppl. 1), 376 (Abstr. 418).

Gengler, N., Mayeres, P. & Szydlowski., M. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal 1*, 21–28.

Henderson, C.R. 1963. Selection index and expected genetic advance. In: Hanson, W.D., Robinson, H.F. (Eds.), *Statistical Genetics and Plant Breeding*, Natl. Res. Counc. Publ. 982. National Academy of Science–National Research Council, Washington, DC.

Henderson, C.R. 1975. Rapid method for computing the inverse of a relationship matrix. *J. Dairy Sci. 58*, 1727–1730.

Lee, S.H. & van der Werf, J.H.J. 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.

Malécot, G. 1948. *Les mathématiques de l'hérédité*. Masson et Cie, Paris. [*The mathematics of heredity* (1968 English translation by D.M. Yermanos). W.H. Freeman and Co., San Francisco.]

Meuwissen, T.H., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157*, 1819–1829.

VanRaden, P.M. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. *J. Dairy Sci. 75*, 3136–3144.

Villanueva, B., Pong-Wong, R., Fernández, J. & Toro, M.A. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci. 83*, 1747–1752.

Wang, J. 2002. An estimator for pairwise relatedness using molecular markers. *Genetics 160*, 1203–1215.

Wright, S. 1922. Coefficients of inbreeding and relationship. *Amer. Nat. 56*, 330–338. (Available online at http://aipl.arsusda.gov/publish/other/wright1922.pdf).