Am. J. Hum. Genet. 66:629–650, 2000

Estimation of Variance Components of Quantitative Traits in Inbred Populations

Mark Abney,[1,2] Mary Sara McPeek,[2] and Carole Ober[1]

Departments of [1]Human Genetics and [2]Statistics, University of Chicago, Chicago

Summary

Use of variance-component estimation for mapping of quantitative-trait loci in humans is a subject of great current interest. When only trait values, not genotypic information, are considered, variance-component estimation can also be used to estimate heritability of a quantitative trait. Inbred pedigrees present special challenges for variance-component estimation. First, there are more variance components to be estimated in the inbred case, even for a relatively simple model including additive, dominance, and environmental effects. Second, more identity coefficients need to be calculated from an inbred pedigree in order to perform the estimation, and these are computationally more difficult to obtain in the inbred than in the outbred case. As a result, inbreeding effects have generally been ignored in practice. We describe here the calculation of identity coefficients and estimation of variance components of quantitative traits in large inbred pedigrees, using the example of HDL in the Hutterites. We use a multivariate normal model for the genetic effects, extending the central-limit theorem of Lange to allow for both inbreeding and dominance under the assumptions of our variance-component model. We use simulated examples to give an indication of under what conditions one has the power to detect the additional variance components and to examine their impact on variance-component estimation. We discuss the implications for mapping and heritability estimation by use of variance components in inbred populations.

Received July 30, 1999; accepted for publication November 4, 1999; electronically published February 18, 2000.

Address for correspondence and reprints: Dr. Mark Abney, Department of Human Genetics, University of Chicago, 924 East 57th Street, R-102, Chicago, IL 60637. E-mail: abney@genetics.uchicago.edu

0002-9297/2000/6602-0030$02.00

Introduction

The use of variance-component analysis to study quantitative traits began early in the 20th century (Weinberg 1909; Fisher 1918). Fisher (1918) described a partition of the total variance (V_t) of a quantitative trait in an outbred population into variance due to environment (V_e), additive genetic effects (V_a), dominance (V_d), and epistasis (V_I). The sum of all those components, aside from environmental variance, is generally called the "genetic variance" (V_g). In principle, one can allow for gene-environment interactions as well. Such a variance decomposition can be used to assess heritability of a trait. Heritabilities in the broad and narrow sense provide two measures of the importance of genetic factors to a trait, with the broad-sense heritability, V_g/V_t, also called the "coefficient of genetic determination," expressing the extent to which the phenotype is explained by genotype in a particular population. The narrow-sense heritability, given in an outbred population by V_a/V_t and typically referred to simply as the "heritability," measures the degree to which, in the given population, the offspring phenotype is explained by the parental phenotypes. Variance-component estimation may be done in conjunction with segregation analysis, in an effort to elucidate the genetic model (Elston and Stewart 1971; Morton and MacLean 1974). More recently, variance-component analysis has been used for the mapping of quantitative traits, by including in the model one or more components of variance due to a major gene linked to a particular locus (Goldgar 1990; Schork 1993; Amos 1994; Almasy and Blangero 1998). In practice, many mapping and heritability studies in outbred populations consider only environmental, additive, and dominance variance or only environmental and additive variance. With appropriate breeding designs in animal or plant models, it is, in principle, possible to estimate epistatic variance, but this is not feasible in humans.

Commonly used methods for estimation of components of variance of quantitative traits include parent-offspring regression, analysis of variance (ANOVA) applied to sib and/or half-sib families, and ANOVA applied to MZ and DZ twins (for a detailed survey, see, e.g., the work of Lynch and Walsh [1998]). ANOVA has the

advantage of providing unbiased estimators even when the data are not normally distributed, although the assumption of normality is generally used to assess uncertainty in the estimate. Such approaches are particularly suited to controlled breeding situations but are not ideal for populations for which one has information on many relationship types and unbalanced family sizes. The situation is particularly extreme in isolated populations in which most individuals are fairly closely related and there is detectable inbreeding. In such a population, it may be impossible to obtain many true full or half-sibs, because, for example, apparent "half-sibs" will usually have a relationship that, because of additional relatedness of their parents, is closer than half-sib. In fact, in some populations, such as the Hutterites, each pair of individuals has such complicated interconnecting lines of relationship that it is only in very circumscribed situations that two different relative pairs have exactly the same relationship (see Appendix A).

A more flexible alternative to ANOVA methods for estimation of variance components is maximum likelihood (ML) (or restricted maximum likelihood [REML]) variance-component estimation (Hartley and Rao 1967; Patterson and Thompson 1971). Recently this methodology has gained interest for purposes of mapping of quantitative traits (Goldgar 1990; Schork 1993; Amos 1994; Almasy and Blangero 1998). For the price of assuming a particular distribution, generally multivariate normal, for the phenotype, the method allows one to partition the variance into its basic genetic and nongenetic components, using a sample of individuals of known relationship. Because the analysis can use the information from all types of relative pairs in the data, without concern for balanced numbers of families of restricted relationship types, the information inherently available in the data is used more efficiently than in ANOVA methods.

The computational burden of ML/REML variance-component estimation in an inbred population can be great. In part, this is due to there being more dominance-variance components that must be estimated in the inbred than in the noninbred case (Harris 1964; Jacquard 1974; Cockerham and Weir 1984). The resulting computational difficulties have limited such studies in the past (de Boer and Hoeschele 1993). Recently, Shaw and Woolliams (1999) have estimated these additional dominance-variance components, using REML, apparently for the first time in a livestock species. Shaw et al. (1998) performed such an estimation in the flowering annual *Nemophila menziesii*. Shaw and Woolliams's (1999) study of sheep found little evidence of either non-0 inbreeding depression or inbreeding dominance components. Shaw et al. (1998), on the other hand, detected significance in several traits of both one of the three inbreeding dominance components and the inbreeding depression. To our knowledge, no studies estimating inbreeding dominance components have been done in humans. There have been a number of studies attempting to estimate inbreeding depression of various traits in humans—for example, the studies by Barrai et al. (1964), Mange (1964), Bittles and Neel (1994). In the studies by Barrai et al. (1964) and Bittles and Neal (1994), association, in many populations, between marital consanguinity and social factors cause confounding. In the Hutterite study by Mange (1964), relationships among individuals in the population are not taken into account in the assessment of uncertainty, potentially inflating the significance of the inbreeding depression.

Although most human genetic studies include outbred families, there has been a recent growth of interest in inbred isolates, because of their apparent advantages in genetic mapping studies. Although simplifying the full variance-component model for inbred individuals to that for the noninbred case may be justifiable at times, it may, in general, introduce bias in the estimates, depending on the nature of the trait and the population. Here we present a variance-component method that, in its estimates of the additive, dominance, and environmental variance components and of inbreeding depression, fully takes into account the effects of inbreeding in the population. We apply the method to a Hutterite sample of 806 individuals who are related by a 13-generation, 1,623-member pedigree. First we describe the identity coefficients for pairs of inbred individuals and describe how they are used in the general variance-component framework. We then apply the method to the sample of Hutterites with HDL as the phenotype.

Methods

General Identity States

In order to estimate variance components by use of pairs of relatives in an inbred pedigree, it is necessary to specify the probabilities of identity-by-descent (IBD) sharing for pairs of individuals, on the basis of their relationship. First, we will explain more carefully what we mean by (pairwise) relationship and how we determine it in practice. Then we will describe the possible identity states for single-locus genotypes of pairs of individuals in an inbred population and will define the associated condensed coefficients of identity (Gillois 1964; Harris 1964).

Examples of pairwise relationships include sib, half-sib, avuncular, and first-cousin. These can be thought of as equivalence classes in pedigrees, in which the pedigree for a pair of individuals is a directed graph with nodes for the two individuals in the pair, with a node for each individual who is ancestral to at least one of the two

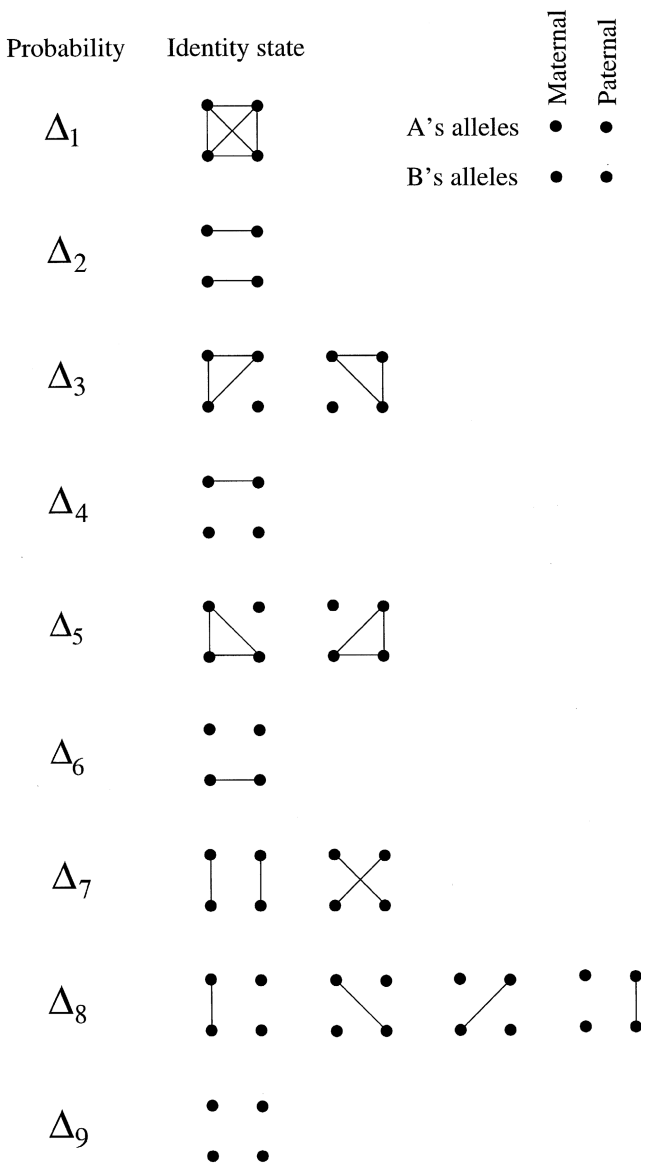

Figure 1 15 Possible identity states for individuals A and B, grouped according to their nine condensed states. Lines indicate alleles that are IBD (adapted from the work of Lynch and Walsh [1998]).

individuals in the pair and with directed edges connecting parents to offspring. In practice, pedigree information will include a limited number of generations, and founders of the pedigree will be assumed to be unrelated, although in fact they are likely to be at least very distantly related. In "outbred" populations, two individuals who have a sib relationship as inferred on the basis of a two-generation pedigree of the individuals and their parents should also have a sib relationship, as inferred on the basis of an *n*-generation pedigree of the individuals and their ancestors, where *n* is some large positive integer; that is, one must go back many generations before any of the sibs' ancestors are found to be related. However, in an inbred population, if two individuals have a sib relationship as inferred on the basis of a two-generation pedigree, then examination of, for instance, a five-generation pedigree for those individuals and their ancestors may reveal that their parents are second cousins. In this case their relationship is in fact closer than sib. Consideration of a six- or seven-generation pedigree for this pair of individuals and their ancestors may reveal additional shared ancestry, allowing a more precise determination of the pair's relationship. We can define a partial ordering on the set of pairwise relationships and choose a precise definition of the *n*-generation pedigree for two individuals (see Appendix A). We then have that the *n*-generation pedigree for a pair of individuals gives a lower bound on the true relationship for the pair, with the accuracy of the approximation increasing with *n*. In the Hutterite data set, we have a 13-generation pedigree containing nodes for all the individuals in the study. In our calculations, we assume that the founders of this pedigree are unrelated.

For a pair of individuals in an inbred population, there are 15 possible ways in which the four alleles of the pair at a particular locus can be shared IBD (fig. 1) (Jacquard 1974). One often restricts attention to genetic models for which the distribution of the phenotype depends on the alleles inherited from the parents, without regard for which allele is maternal and which is paternal. In this case, the 15 possible IBD-sharing configurations among the four alleles for a pair of individuals may be collapsed

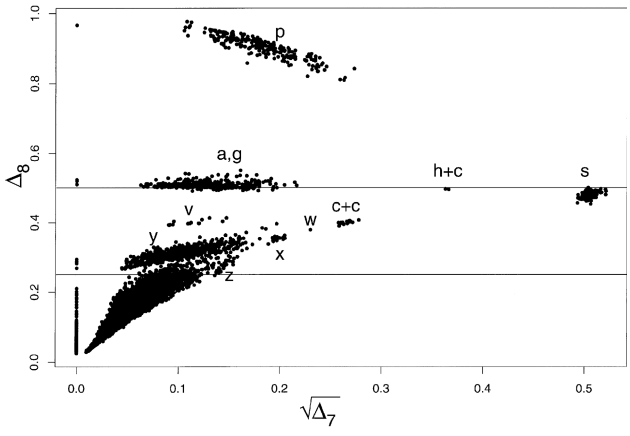

Figure 2 Plot of Δ_8 vs. $\sqrt{\Delta_7}$, for the 324,415 pairs of individuals in the Hutterite sample, where the labeled clouds of points are described in table 1. The horizontal lines are at values $\Delta_8 = .25$ and $\Delta_8 = .5$.

Table 1

Approximate Relationships Corresponding to Labeled Point Clouds of Figure 2, with Number of Occurrences in the Hutterite Data Set and with Corresponding IBD Probabilities for Outbred Individuals

Label	Approximate Relationship(s)	$(\sqrt{\Delta_7},\Delta_8)$ in Outbred Individuals	No. of Pairs
s	Sib	(.5,.5)	1,601
p	Parent-offspring	(0,1)	1,114
h+c	Half-sib plus first cousin[a]	(.354,.5)	68
h	Other half-sib	(0,.5)	0
a	Avuncular	(0,.5)	4,354
g	Grandparent-grandchild	(0,.5)	748
c+c	Double first cousin	(.25,.375)	557
v	(Half-sib+cousin) once removed[b]	(0,.375)	425
w	Cousin plus second (half-sib+cousin)[c]	(.2165,.344)	12
x	Cousin plus (cousin once removed)[d]	(.1768,.3125)	377
y	Other first cousin	(0,.25)	7,897
	Grand-avuncular	(0,.25)	2,314
	Great-grandparent/child	(0,.25)	98
	Half-avuncular	(0,.25)	33
	(Double cousin) once removed	(0,.25)	834
z	Double (cousin once removed)	(.125,.21875)	424

[a] Individuals 1 and 2 in figure 3*A*.
[b] Individuals 1 and 4 in figure 3*A*.
[c] Individuals 3 and 4 in figure 3*A*.
[d] Individuals 1 and 2 in figure 3*B*.

into nine equivalence classes, which we refer to as "identity states" (fig. 1). Following Jacquard (1974) (also see the work of Gillois [1964] and Harris [1964]; Lange [1997] provides a detailed exposition), for a pair of individuals we let Δ_i, $i = 1,...,9$ denote the conditional probability of identity state i, given the relationship between the individuals. These Δ_i's are called the "condensed coefficients of identity." With these identity coefficients, one can define other, more commonly used measures of relatedness, such as the kinship and inbreeding coefficients. The kinship coefficient between individuals A and B, which can be interpreted as the probability that, at a given locus, a randomly chosen allele from individual A is IBD to a randomly chosen allele from individual B, conditional on the relationship between A and B, is $\Phi_{AB} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8$, and the inbreeding coefficients of A and B, (i.e., the kinship coefficient of A's parents and the kinship coefficient of B's parents), are $f_A = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$ and $f_B = \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6$. The k coefficients of Cotterman (1940)—k_0, k_1, and k_2, which are equivalent to Δ_9, Δ_8, and Δ_7 (with $k_0 = \Delta_9$, $k_1 = \Delta_8/2$, and $k_2 = \Delta_7$)—are sufficient to fully specify the relationship between any pair of individuals only when neither of the individuals is inbred.

A recursive algorithm for computing the identity coefficients for any pedigree has been given by Karigl (1981) (also see the work of Harris [1964] and Lange [1997]). Although this algorithm works well for small- to moderate-size pedigrees, large inbred pedigrees can prove very problematic computationally. De Boer and Hoeschele (1993) have developed a similar algorithm, which uses a tabular method essentially based on the recursion relations given by Harris (1964), and were able to calculate the needed relationship coefficients in a simulated population of 200 inbred individuals constituting five generations. However, they concluded that implementation of their method along with subsequent variance-component analysis for larger populations was not yet feasible, owing to the size of the resulting relationship matrices. We chose to use the basic algorithms of Karigl (1981) and a computational strategy that increases their efficiency to the point where they are feasible even in large inbred pedigrees (see Appendix B).

Quantitative-Trait Model

We consider the following model for a quantitative trait y, $y_k = \mathbf{x}_k^T\boldsymbol{\beta} + g_k + e_k$, where y_k is the phenotypic (trait) value for the kth individual, $\mathbf{x}_k$ is a vector of covariate values for the kth individual, $\boldsymbol{\beta}$ is a vector of fixed effects, g_k is the random genetic effect for individual k, and the environmental effects e_k are assumed to be independent and identically distributed normal$(0,V_e)$ random variables, with $\mathbf{e}$ and $\mathbf{g}$ independent and $(\mathbf{e},\mathbf{g})$ not depending on $\mathbf{X}$, the matrix of covariate values. Furthermore, $g_k = \Sigma_{l=1}^{L} g_l[(i,j)_{l,k}]$ and $g_l[(i,j)_{l,k}] = a_{li} + a_{lj} + d_{lij}$,

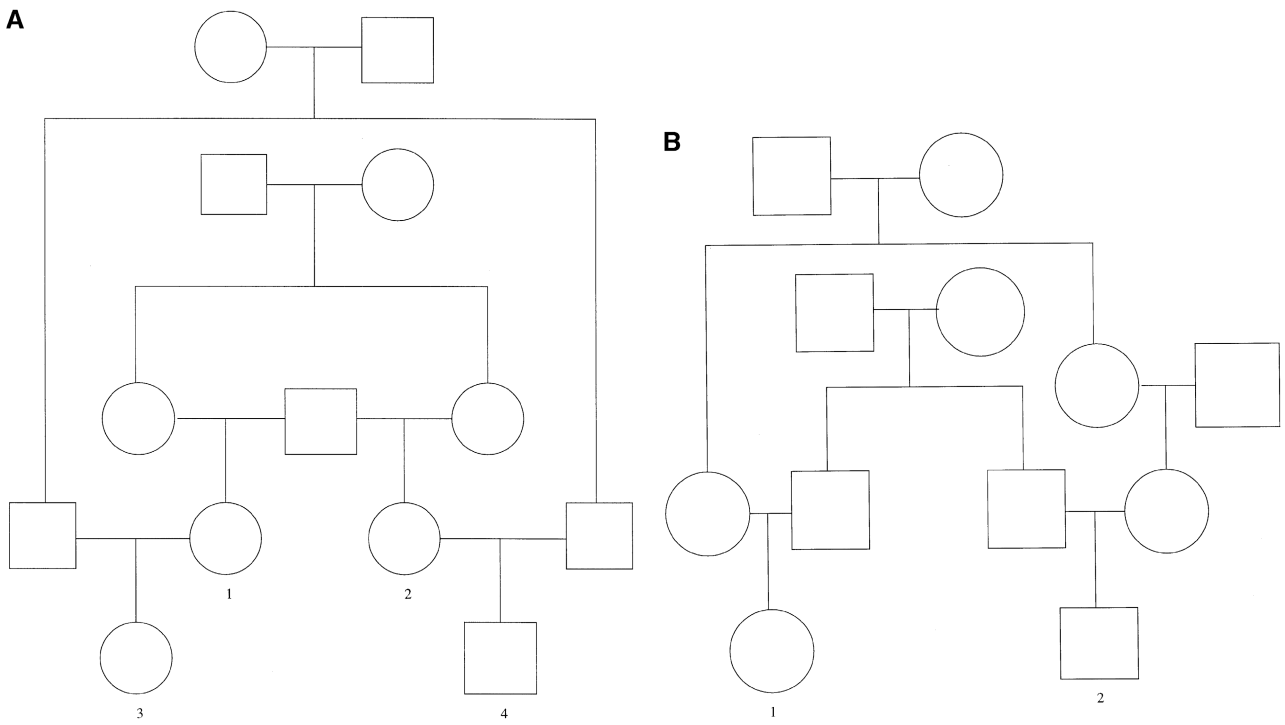

Figure 3 Example pedigrees 1 (*A*) and 2 (*B*) from the Hutterite sample

where l indexes the genetic loci influencing the trait, $(i,j)_{l,k}$ is the genotype of the kth individual at the lth locus, a_{li} and a_{lj} are the additive effects, and d_{lij} is the dominance (i.e., interaction) effect between the two alleles at the lth locus, with a_{li}, a_{lj}, and $d_{lij} = d_{lji}$ all fixed. As described in more detail below, we will generally take $\mathbf{g}$ to be mean 0 in outbred individuals, by including a constant term in $\mathbf{x}_k^T\boldsymbol{\beta}$, and we will define a and d to minimize the mean squared deviation of $g_l[(i,j)_{l,k}]$ from $a_{li} + a_{lj}$ in an outbred random-mating population.

Note that the genotype $(i,j)_{l,k}$ of the kth individual at the lth locus is random and dependence among relatives is induced by the inheritance process. In principle, one can think of the relationships among individuals in the population as being known and can consider the above model conditional on them, or one can think of the relationships among individuals in the population as following some probability model. We will usually condition on the relationships among individuals. This approach allows estimation of the variance components described below and is applicable to the Hutterite data set that we are analyzing, for which extensive pedigree information is available.

The case of a small number of loci with large genetic effects, representing one or more major genes, would result in the phenotype y_k being a mixture of normal random variables, with a complicated dependence structure for $\mathbf{y}$, owing to the inheritance process. Fitting this model to data by maximum likelihood presents serious computational challenges even for pedigrees much simpler than that of the Hutterites. Here we consider instead the polygenic case in which $\mathbf{g}$—and, hence, $\mathbf{y}$—is assumed to be multivariate normal. Fisher (1918) suggested this as an approximation to the case of a large number of loci with small genetic effects and additivity across loci, assuming that conditions of a central-limit theorem are met. The polygenic model is widely used even for mapping, where, presumably, the locus being mapped is actually a major gene (Goldgar 1990; Schork 1993; Amos 1994; Almasy and Blangero 1998). Some sufficient conditions for a central-limit theorem have been discussed by Lange (1978); however, these conditions exclude the case in which there is both inbreeding and non-0 dominance variance. The trait models that we consider in the present paper have both inbreeding and non-0 dominance variance, with an assumption of either (1) unlinked loci or (2) linked loci with the constraint that, on each chromosome, at most one locus has non-0 inbreeding depression. In Appendix C, we give an extension of the central-limit theorem to this case.

Under the assumption of multivariate normality, we are concerned with the first and second moments of $\mathbf{y}$

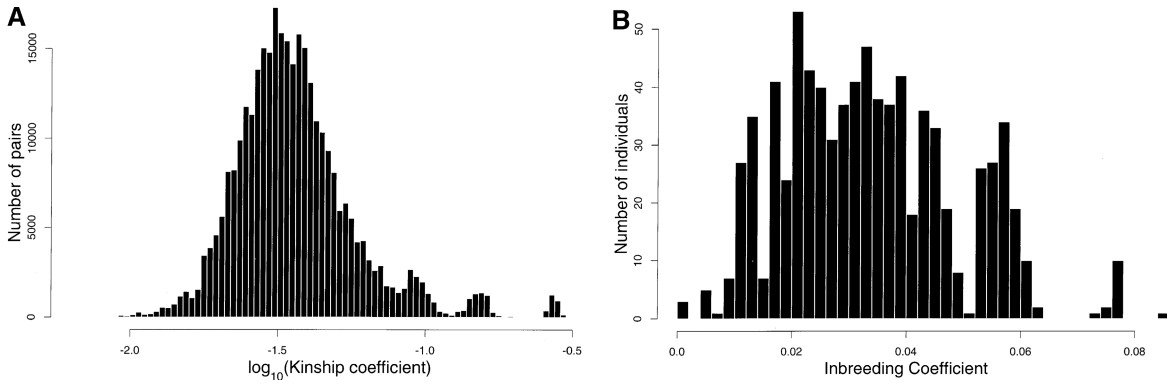

Figure 4 Histograms of (*A*) $\log_{10}$(kinship coefficient) for the 324,415 pairs of individuals in the Hutterite sample and (*B*) inbreeding coefficient for the 806 individuals in the Hutterite sample.

induced by the inheritance process, conditional on the pedigree, ignoring higher moments. Initially, we drop the subscript indexing the locus and focus on first and second moments for the single-locus genetic model, which is then extended to multiple loci.

Variance Components in an Outbred Population

Consider an infinitely large, random mating population—hence, one with no inbreeding. Let p_{ij} be the probability that a randomly chosen individual has genotype (i,j), $1 \leq i \leq n$, $1 \leq j \leq n$, where n is the number of alleles at the locus. Let p_i be the probability that a randomly chosen allele is i. We assume Hardy-Weinberg equilibrium—that is, $p_{ij} = 2p_ip_j$ for $i \neq j$ and $p_{ii} = p_i^2$. For a single locus, we now write the genetic effect of genotype (i,j) as $g_{ij} = a_i + a_j + d_{ij}$. In this population, all individuals will have the same mean genetic effect. We can assume that this mean is 0, because a non-0 mean can be absorbed into the constant term in $\mathbf{x}_k^T\boldsymbol{\beta}$. Let $E(a) = \Sigma_{i=1}^n a_ip_i$, $E(d_j) = \Sigma_{i=1}^n d_{ij}p_i$, and $\mathrm{Cov}(a,d) = \Sigma_{i=1}^n\Sigma_{j=1}^n a_id_{ij}p_ip_j$. Here, $E(a)$ and $E(d_j)$ represent the mean additive effect of a randomly chosen allele at a locus and the mean dominance effect of allele j at a locus, respectively, where the mean is taken with respect to the distribution of alleles in the population. $\mathrm{Cov}(a,d)$ is the covariance of additive and dominance effects of an allele in the population. Then we can define a_i and d_{ij} so that $\mathrm{Cov}(a,d) = E(a) = 0$, $E(d_j) = 0$ for all j. This is equivalent to choosing the a_i and d_{ij} to minimize the expected squared deviation of g_{ij} from $a_i + a_j$, where (i,j) is a genotype chosen at random from the population.

Consider $\mathrm{Cov}(\mathbf{y})$, the covariance matrix of $\mathbf{y}$ conditional on the pairwise relationships of the sampled individuals. $\mathrm{Cov}(\mathbf{y})$ has constant diagonal $\mathrm{Var}(y) = V_a + V_d + V_e$, where V_a is the additive variance $V_a = 2\Sigma_{i=1}^n p_ia_i^2$, V_d is the dominance variance $V_d = \Sigma_{i=1}^n\Sigma_{j=1}^n p_ip_jd_{ij}^2$, and V_e is the residual environmental variance. Note that V_a, which represents the individual's phenotype variance due to additive genetic effects, has a factor of two, to account for the two alleles in a genotype. For A ≠ B, the ABth element of $\mathrm{Cov}(\mathbf{y})$—that is, the covariance between the genetic values of individuals A and B, conditional on their relationship, is $\mathrm{Cov}(y_A,y_B) = \mathrm{Cov}(g_A,g_B) = 2\Phi_{AB}V_a + \Delta_7V_d$ where Φ_{AB} is the kinship coefficient for A and B.

Variance Components in an Inbred Population

Inbred populations violate the above assumptions. Nevertheless, following Harris (1964), we can still define a_i and d_{ij} as above—that is, in an infinitely large population undergoing random mating with allele frequencies equal to the frequencies in the population in question. The population model that we assume here is one of finite size undergoing random mating so that the probability p_{ij} of individual k being genotype (i,j), conditional on the relationship between the individual's parents, is $p_{ii} = f_kp_i + (1 - f_k)p_i^2$ and $p_{ij} = (1 - f_k)2p_ip_j$ when $i \neq j$, where f_k is individual k's inbreeding coefficient. In this case, the mean genetic effect for individual k is

$$E(g_k) = f_k\mu_h \ , \qquad (1)$$

where $\mu_h = \Sigma_{i=1}^n p_id_{ii}$ and μ_h is the mean in the homozygous population and is known as the "inbreeding depression" (for a review of inbreeding depression, see the work of Lynch and Walsh [1998]). (Recall that $E(g_k)$ for an outbred individual was set to 0 by absorbing any non-0 value into the constant term of the trait model.) In ad-

dition, the homozygosity increase that results from inbreeding alters the matrix Cov(**y**). Note that the additive genetic component of covariance is not affected by the increase in homozygosity, because the allele frequencies are unchanged. In contrast, the dominance variance now takes on three additional components. According to Jacquard (1974), these components are as follows:

1. V_h, the dominance variance in the homozygous population, where $V_h = \Sigma p_i d_{ii}^2 - \mu_h^2$;
2. $\mathrm{Cov}_h(a,d)$, the covariance of additive and dominance effects in the homozygous population, where $\mathrm{Cov}_h(a,d) = \Sigma p_i a_i d_{ii}$;
3. μ_h^2, the square of the inbreeding depression.

The matrix Cov(**y**) has as its kth diagonal element

$$\mathrm{Var}(y_k) = (1 + f_k)V_a + (1 - f_k)V_d + f_k V_h + 4f_k \mathrm{Cov}_h(a,d) + f_k(1 - f_k)\mu_h^2 + V_e ,$$

where f_k is the inbreeding coefficient of the kth individual. The off-diagonal elements become

$$\mathrm{Cov}(y_A,y_B) = \mathrm{Cov}(g_A,g_B) = 2\Phi_{AB}V_a + \Delta_7 V_d + \Psi_4 V_h + 2(\Psi_3 + \Psi_4) \times \mathrm{Cov}_h(a,d) + (\Delta_1 + \Delta_2 - f_A f_B)\mu_h^2 ,$$

where $\Psi_4 = \Delta_1$ is the probability that all four alleles of A and B are IBD, and $\Psi_3 = \Delta_1 + (\Delta_3 + \Delta_5)/2$ is the probability that three alleles taken at random from A and B are IBD.

For the polygenic model with additivity of genetic effects across loci, no linkage disequilibrium among loci, and no more than one locus per chromosome having non-0 inbreeding depression, we get the expressions (see Appendix C)

$$\mathrm{Var}(y_k) = (1 + f_k)V_a + (1 - f_k)V_d + f_k V_h + 4f_k \mathrm{Cov}_h(a,d) + f_k(1 - f_k)SS_{\mu_h} + V_e \qquad (2)$$

and

$$\mathrm{Cov}(y_A,y_B) = \mathrm{Cov}(g_A,g_B) = 2\Phi_{AB}V_a + \Delta_7 V_d + \Psi_4 V_h + 2(\Psi_3 + \Psi_4) \times \mathrm{Cov}_h(a,d) + (\Delta_1 + \Delta_2 - f_A f_B)SS_{\mu_h} , \qquad (3)$$

where V_a, V_d, V_h, and Cov(a,d) are now redefined to be the sums, over all loci, of their single-locus values. Here, $SS_{\mu_h} = \Sigma_l \mu_{hl}^2$, where the sum is over all loci, and μ_{hl} is the inbreeding depression of the lth locus. In equation (1), the expression for the mean vector, μ_h is redefined to be the sum, over all loci, of its single-locus value. If the number of loci can be arbitrary, we have the following constraints:

1. V_a,V_d,V_e,V_h and SS_{μ_h} are all nonnegative (and, we assume, not all 0);
2. If $SS_{\mu_h} = 0$, then $\mu_h = 0$;
3. $|\mathrm{Cov}_h(a,d)| \leq \sqrt{V_h V_a/2}$.
4. If $V_d = 0$, then $SS_{\mu_h} = 0$ and $V_h = 0$ (and, as a consequence of 2 and 3, $\mu_h = 0$ and $\mathrm{Cov}_h(a,d) = 0$ as well).

The first constraint is due to the fact that V_a, V_d, V_e, and V_h are all variances and that SS_{μ_h} is a sum of squares. The second constraint simply reflects the fact that a sum of squares equals 0 only if all the components are 0, implying that their sum is 0. The third constraint is a version of the well-known inequality $|\mathrm{Cov}(x,y)| \leq \sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}$, which is a special case of the Cauchy-Schwartz inequality and which is equivalent, when the variances are non-0, to the constraint that the correlation must lie between -1 and 1. Here, we have $|\mathrm{Cov}_h(a,d)| \leq \sqrt{\mathrm{Var}_h(a)\mathrm{Var}_h(d)}$, where the h subscript indicates that the quantities are calculated in the homozygous population,

$$\mathrm{Var}_h(d) = \sum_i p_i d_{ii}^2 - \left(\sum_i p_i d_{ii}\right)^2 = V_h ,$$

$$\mathrm{Var}_h(a) = \sum_i p_i a_i^2 - \left(\sum_i p_i a_i\right)^2 = \frac{1}{2} V_a .$$

Thus, $|\mathrm{Cov}_h(a,d)| \leq \sqrt{V_h V_a/2}$. To obtain the fourth constraint, we have $V_d = 0$ only if $d_{ij} = 0$ for all i and j and for all loci, because V_d is a weighted sum of squares of the d_{ij}'s. When every d_{ij} is 0, we have $\mu_{hl} = 0$, $V_h = 0$, and $SS_{\mu_h} = 0$.

Using this model, we can now partition the total var-

Figure 5 Plot of inbreeding coefficient versus year of birth for the 806 individuals in the Hutterite study sample

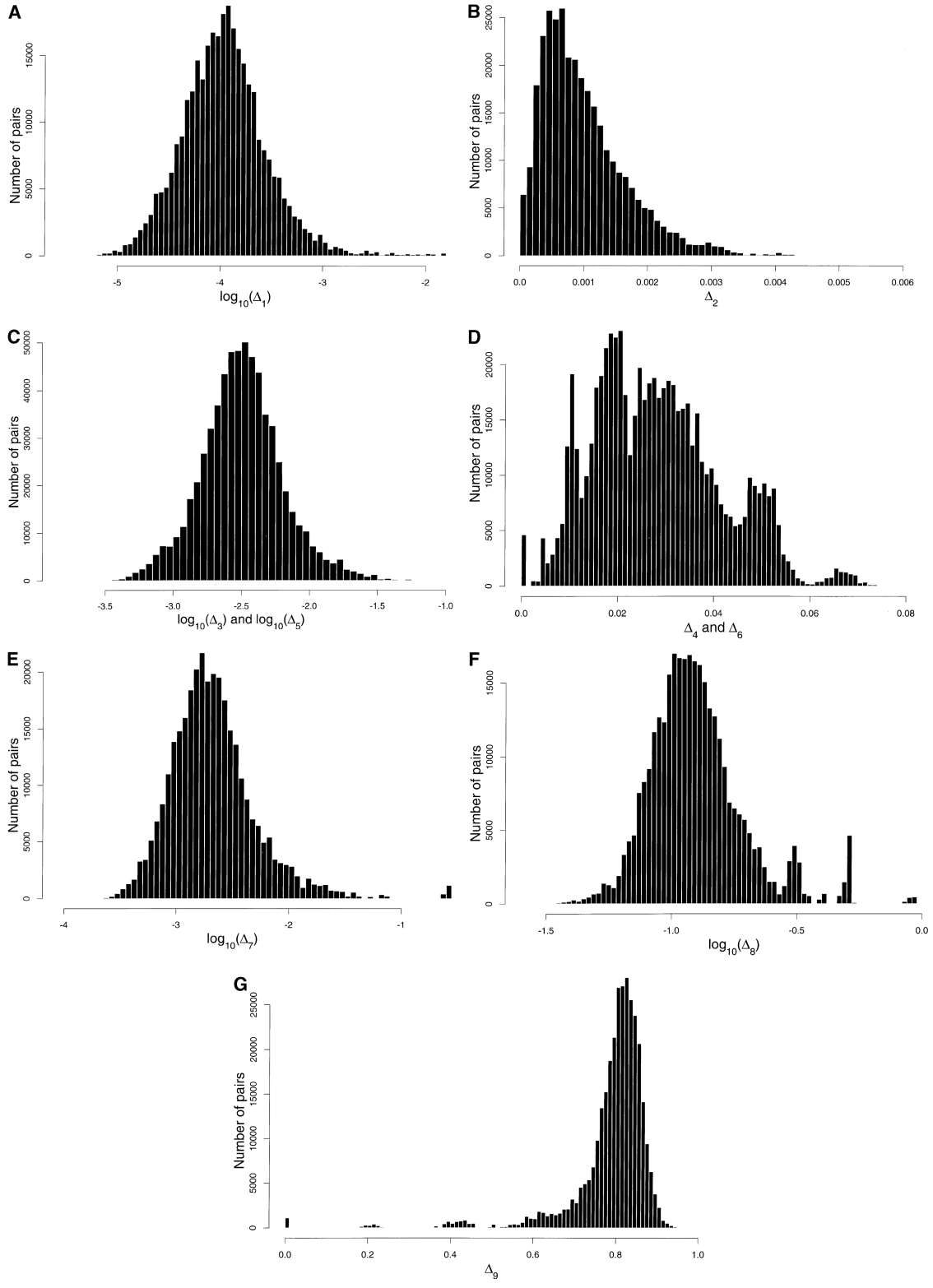

Figure 6 Histograms of (*A*) $\log_{10}(\Delta_1)$, (*B*) Δ_2, (*C*) $\log_{10}(\Delta_3)$ combined with $\log_{10}(\Delta_5)$, (D) Δ_4 combined with Δ_6, (*E*) $\log_{10}(\Delta_7)$, (*F*) $\log_{10}(\Delta_8)$, and (*G*) Δ_9, for the 324,415 pairs of individuals in the Hutterite sample.

Table 2

Mean and SD of Identity Coefficients, Taken over All Pairs of Hutterite Individuals

	Mean (SD)		
Coefficient	Untransformed	$\log_{10}$ Transformation[a]	Geometric Mean[a]
Δ_1	.000217 (.000747)	−3.96 (.42)	.000111
Δ_2	.000993 (.000678)	...	...
Δ_3,Δ_5	.00411 (.00382)	−2.49 (.30)	.00321
Δ_4,Δ_6	.0283 (.0134)	...	...
Δ_7	.00444 (.01833)	−2.67 (.39)	.00212
Δ_8	.141 (.089)	−.899 (.190)	.126
Δ_9	.788 (.101)	...	...

[a] Three individuals are removed from the sample before the $\log_{10}$ transformation or geometric mean is taken (see text). With these three individuals removed, Δ_1, Δ_3, Δ_5, Δ_7, and Δ_8 are non-0 for all pairs; note that Δ_2, Δ_4, Δ_6, and Δ_9 will always be 0 for parent-offspring pairs.

iance of the quantitative trait in the population. Consider choosing an individual uniformly at random from the population, and let the random variable Z denote that individual's trait value. Let F be the random variable representing the inbreeding coefficient of the individual chosen. Then, using a standard identity of conditional probability, we have

$$\begin{aligned} V_t = \mathrm{Var}(Z) &= E[\mathrm{Var}(Z|F)] + \mathrm{Var}[E(Z|F)] \\ &= (1+\mu_f)V_a + (1-\mu_f)V_d + \mu_f V_b + 4\mu_f \mathrm{Cov}_b(a,d) \\ &\quad + [\mu_f(1-\mu_f) - \sigma_f^2]SS_{\mu_b} + V_e + \sigma_f^2\mu_b^2 , \end{aligned}$$

where μ_f and σ_f^2 are the mean and variance, respectively, of the inbreeding coefficient in the population. Then the narrow and broad sense heritabilities in this population are $h^2 = (1+\mu_f)V_a/V_t$ and $H^2 = 1 - V_e/V_t$, respectively.

ML and REML Estimation

Given a random sample of N individuals from the population, with the relationships among them known, the multivariate normal log-likelihood is

$$\begin{aligned} l(\boldsymbol{\beta},\mu_b,\mathbf{V};\mathbf{y}) &= -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\Omega| \\ &\quad -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}\mu_b)^T\Omega^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f}\mu_b) , \end{aligned} \qquad (4)$$

where $\boldsymbol{\beta}$ is the vector of fixed effects; μ_b is the inbreeding depression; $\mathbf{V} = [V_a,V_d,V_b,\mathrm{Cov}_b(\mathrm{a,d}),SS_{\mu_b},V_e]$, the vector of variance-component parameters; Ω is the covariance matrix Cov(**y**), which is a function of **V** and the identity coefficients and is given in equations (2) and (3); $|\Omega|$ is the determinant of Ω; **y** is the vector of phenotypic values; **X** is the covariate matrix; and **f** is the vector of inbreeding coefficients. We maximize the log likelihood by implementing a search over **V**, using the simplex algorithm of Nelder and Mead (1965). For each choice of **V**, the maximum-likelihood estimates of $\boldsymbol{\beta}$ and μ_b can easily be found by generalized regression. Standard errors are obtained from the observed Fisher information.

We consider the full model of equations (1)– (3) and, also, various submodels obtained by setting some of the variance components to 0. In addition to the parameter estimates, standard errors (SEs), and maximized log likelihoods, we report the values of two model-selection criteria, the Akaike information criterion (AIC [Akaike 1974, pp. 267–281) and the Bayesian information criterion (BIC [Schwartz 1978]). AIC is defined to be $-2\hat{l} + 2k$, and BIC is defined to be $-2\hat{l} + k\log(n)$, where $\hat{l}$ is the maximized log likelihood, k is the number of freely varying parameters, and n is the sample size. Minimization of these criteria has been proposed for use in selecting from a number of models of different dimension. The term $2k$ or $k\log(n)$ is an attempt to enforce parsimony by penalizing the log likelihood for estimation of additional parameters.

Maximum-likelihood estimation of variance components does not, in general, take into account the loss in degrees of freedom that results from estimation of the fixed effects, and, as a result, ML estimators tend to be biased. In particular, estimates of the variance components are generally downwardly biased (Corbeil and Searle 1976), with the bias increasing as the number of fixed effects increases. If the sample size is small, this bias can become quite substantial. An alternative to ML estimation is REML estimation (Searle et al. 1992), which essentially maximizes only that portion of the likelihood that depends on the variance components and not on the fixed effects. Hence, bias of this type is removed by REML in a manner analogous to the removal of bias in a variance estimator by dividing by the degrees of freedom rather than by dividing by the sample size. The question of whether ML or REML is the preferred method in any particular situation, however, is unclear, because each has advantages and disadvantages (Harville 1977; Searle et al. 1992). We chose to implement REML in addition to ML, to see how they compare in this situation.

REML, instead of using the data vector **y** directly, is based on a linear transformation of the data, where the transformation is chosen in such a way that the fixed effects are eliminated from the model. Given the mixed model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \mathbf{e}$, consider a matrix **K** such that $\mathbf{KX} = \mathbf{0}$. Applying this transformation to our model equation results in $\mathbf{Ky} = \mathbf{Kg} + \mathbf{Ke}$. If **y** is normal, with mean $\mathbf{X}\boldsymbol{\beta}$ and variance **V**, then **Ky** is also normal, with

Table 3

Maximum-Likelihood Estimates of Variance Components and Inbreeding Depression for HDL in the Hutterites, with Resulting Narrow- and Broad-Sense Heritabilities

Model	μ_b (SE)	V_e (SE)	V_a (SE)	V_d (SE)	V_b (SE)	h^2 (SE)	H^2 (SE)	AIC	BIC	Log Likelihood
1		.89 (.05)						467	484	−229.43
2		.33 (.06)	.60 (.11)			.65 (.06)	.65 (.06)	381	402	−185.56
3	4.84 (2.73)	.33 (.06)	.60 (.11)			.65 (.06)	.65 (.06)	381	407	−184.54
4		.23 (.14)	.59 (.12)	.13 (.16)		.65 (.14)	.76 (.12)	382	408	−185.13
5	4.55 (2.83)	.26 (.13)	.59 (.11)	.089 (.15)		.66 (.14)	.73 (.11)	383	412	−184.33
6		.22 (.09)	.55 (.11)		5.03 (2.88)	.59 (.10)	.77 (.08)	379	405	−183.67
7	4.13 (2.96)	.23 (.09)	.55 (.11)		4.57 (2.86)	.60 (.10)	.76 (.08)	380	410	−183.07
8		.21 (.14)	.54 (.11)	.023 (.17)	4.87 (3.21)	.62 (.16)	.79 (.12)	381	411	−183.66
9	4.12 (2.96)	.26 (.13)	.55 (.11)	.006 (.17)	4.53 (3.18)	.63 (.16)	.76 (.12)	382	416	−183.07

mean **0** and variance $\mathbf{KVK}^T$. REML then proceeds as ML, but with the transformed data vector and variance matrix. Although REML apparently requires one to compute the matrix **K**, it is possible to formulate the REML equations only in terms of **y**, **X**, and **V** (e.g., see the work of Searle et al. [1992] and Hofer [1998]).

Comparison of Models With and Without Additional Inbreeding Dominance Components and Inbreeding Depression

In studying a quantitative trait in an inbred population, one might choose to consider a model simpler than that given in equations (1)–(3), by setting μ_b, V_b, $\mathrm{Cov}_b(a,d)$, and SS_{μ_b} to 0. In an inbred population, this model is equivalent to setting $d_{ii} = 0$ for all *i;* that is, at each locus, dominance effects occur only between distinct alleles, and the genetic effect of having two copies of a particular allele is always twice the additive effect of a single copy. Although it is not particularly natural or reasonable to assume this, it makes the analysis considerably easier. In addition to a parameter space of smaller dimension, we have the advantage that the only information needed from the pedigree is the matrix of kinship coefficients Φ (here we refer to standard kinship coefficients defined for pairs of individuals, not the generalizations defined in Appendix B) and the matrix of Δ_7, which are vastly simpler to obtain than are the additional identity coefficients needed in the more general case (see Appendix B). The model suggested by Almasy and Blangero (1998) for mapping of a major gene in arbitrary pedigrees would correspond, in inbred pedigrees, to this simplified model just described. It seems sensible to examine whether, in practice, the more complicated and more accurate model is worth the considerable extra effort to fit. There can, of course, be no definitive answer to this question, but, by considering various examples, we may get an idea of those situations in which the difference between these models is likely to have a practical impact. We examine this question in three ways: (1) by a comparison of variance-components analyses of HDL in the Hutterites, under the models with and without the additional inbreeding dominance components and inbreeding depression; (2) by a small-scale simulation study in which trait values are simulated for various combinations of the Hutterite sample and additional inbred individuals and in which results of inference under the two models are compared; and (3) by comparing, in some examples, the results that one would obtain asymptotically under the two different models. By looking at such asymptotic results, we are able to gain insight into the comparison of the models when the number of replicates is large, a situation for which simulation would be infeasible. We now describe the last of these three approaches; the first and second are presented in the Results section.

Suppose that the true model for **y** is that described in equations (1)–(3). At the moment, we are not concerned with fixed effects, so we assume that there are none, with the exception of a constant term μ. Assume that the parameter vector $\boldsymbol{\theta} = (\mu,\mu_b,\mathbf{V})$ lies in parameter space $\Theta_1 \subset \mathbb{R}^8$ defined by the constraints described previously. Let $\Theta_2 \subset \Theta_1$ be the convex subspace with $\mu_b = V_b = \mathrm{Cov}_b(a,d) = SS_{\mu_b} = 0$. Let the true set of parameter values be represented by $\boldsymbol{\theta}^* \in \Theta_1$, where $\boldsymbol{\theta}^*$ may or may not lie within Θ_2. Assume that the data consist of a random sample of individuals from a population, where the relationships among the individuals in the sample are known and conditioned on. We consider an asymptotic scenario in which independent, identically distributed replicates of the sample are drawn, where, in each replicate, the same relationships hold among the individuals. Suppose that we have n such replicates generated under the model with true parameter $\boldsymbol{\theta}^*$ and that we maximize the log-likelihood over all $\boldsymbol{\theta} \in \Theta_2$. Let $\hat{\boldsymbol{\theta}}_n$ represent the resulting estimator. We consider the (almost sure) limit of $\hat{\boldsymbol{\theta}}_n$ as n approaches infinity; that is, we consider the result that would be obtained asymptotically if the variance components were estimated under

Table 4

Scenarios Considered in Projections and Simulations Comparing Models With and Without Additional Inbreeding Dominance Components

Scenario	Population	Average Inbreeding	Model
A	806 Hutterites	.034	Three-quarters dominant
B	806 Hutterites	.034	Fully dominant
C	2 × 806 Hutterites	.034	Fully dominant
D	806 Hutterites + 500 avuncular- pair offspring	.069	Three-quarters dominant
E	806 Hutterites + 800 cousin-pair offspring	.048	Fully dominant
F	806 Hutterites + 500 avuncular- pair offspring	.069	Fully dominant
G	2 × 806 Hutterites + 1,000 avuncular- pair offspring	.069	Fully dominant

a model assuming $\mu_h = V_h = Cov_h(a,d) = SS_{\mu_h} = 0$, when, in fact, these components may be non-0.

Let $\tilde{\boldsymbol{\theta}}$ maximize over all $\boldsymbol{\theta} \in \Theta_2$ the expected log likelihood

$$E_{\theta^*} \log L(\boldsymbol{\theta}; y) = \int L(\boldsymbol{\theta}^*; z) \log L(\boldsymbol{\theta}; z) dz \ ,$$

where the expected value is under the true parameter $\boldsymbol{\theta}^*$, and L is the likelihood function whose log is given in formula (4). Note that this is equivalent to choosing $\boldsymbol{\theta} \in \Theta_2$ to minimize the Kullback-Leibler divergence from the distribution represented by $\boldsymbol{\theta}^*$ to the distribution represented by $\boldsymbol{\theta}$ (Kullback and Leibler 1951; Kullback 1959). Because, in our case, $E_{\theta^*} \log L(\boldsymbol{\theta}; y)$ is a strictly convex function of $\boldsymbol{\theta}$, a unique maximum $\boldsymbol{\theta}$ exists. Under the asymptotic scenario described, with probability 1, $\tilde{\boldsymbol{\theta}}$ is the limiting value of $\hat{\boldsymbol{\theta}}_n$ as n approaches infinity (Huber 1967; Akaike 1973, pp. 267–281; reviewed, for the exponential family case, by McCulloch [1988]).

For a given choice of pedigree and of $\boldsymbol{\theta}^*$, we can compare $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ to see what is the effect asymptotically of considering the simpler model, and we can then try to evaluate the practical implications. For a given pedigree structure and given $\boldsymbol{\theta}^*$, we can calculate $\boldsymbol{\theta}$ by maximizing a likelihood over $\boldsymbol{\theta} \in \Theta_2$, where, in the likelihood, we plug in the expected values of the sufficient statistics under $\boldsymbol{\theta}^*$, in place of the sufficient statistics calculated from the data. This calculation follows from the proof of Result 1 of McCulloch (1988), for the exponential family. Writing $\boldsymbol{\theta}^* = (\mu^*, \mu_h^*, \mathbf{V}^*)$, with $\mu^* = 0$, $\mathbf{V}^* = (V_a^*, V_d^*, V_h^*, \mathrm{Cov}_h(a,d)^*, SS_{\mu h}^*, V_e^*)$, and, defining Ω^* to be the covariance matrix given in equations (2) and (3), evaluated at $\mathbf{V}^*$, we obtain $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}, 0, \tilde{\mathbf{V}})$, with $\tilde{\mathbf{V}} = (\tilde{V}_a, \tilde{V}_d, 0, 0, 0, \tilde{V}_e)$, where $\tilde{\mu} = \mu_h^* (1^T \tilde{\Omega}^{-1} \mathbf{f}) / (1^T \tilde{\Omega}^{-1} 1)$, $\tilde{\Omega}$ is the covariance matrix given in equations (2) and (3) and evaluated at $\tilde{\mathbf{V}}$, and $\tilde{\mathbf{V}}$ is the maximizer, over Θ_2, of

$$-\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln |\tilde{\Omega}| - \frac{1}{2} \left[\mathrm{tr}(\tilde{\Omega}^{-1} \Omega^*) + \mu_h^{*2} \mathbf{f}^T \tilde{\Omega}^{-1} \mathbf{f} - \mu_h^{*2} \frac{(1^T \tilde{\Omega}^{-1} \mathbf{f})^2}{1^T \tilde{\Omega}^{-1} 1} \right] .$$

If we allow $\mu^* > 0$, then $\tilde{\mathbf{V}}$ is unchanged, and μ^* is added to $\tilde{\mu}$. Thus, using the likelihood-maximizing routine developed to analyze the data, we can calculate the asymptotic value $\boldsymbol{\theta}$ of the maximum-likelihood estimate of the parameters when the true model has the inbreeding dominance components and inbreeding depression but the fitted model does not. We apply this method in the Results section, to evaluate the effect asymptotically of ignoring the inbreeding dominance components and inbreeding depression. This gives us an idea of the effect of ignoring these components in large samples, for which simulations are infeasible.

Two-Allele Models with Full and Three-Quarters Dominance

In addition to the general model described in equations (1)–(3), we consider two special cases. These two special cases are used, in the Results section, to simulate phenotypes with some degree of dominance, in order to assess the impact, on variance-component estimation, of fitting a model with all the inbreeding dominance components, as opposed to a simpler model in which these components are fixed at 0. We now describe the two polygenic models used to simulate the phenotypes. In the first, every locus is assumed to be biallelic with complete dominance; that is, the genetic effect of the heterozygote is the same as that of one of the two homozygotes. In the second, every locus is assumed to be biallelic with three-quarters dominance; that is, the genetic effect of the heterozygote is three-quarters of the way from the midpoint of the homozygote effects to one of the two homozygote effects. For simplicity, we further assume that the allele frequencies are the same at all loci.

Table 5

Asymptotic Values of Estimated Variance Components and Heritabilities When a Model without Inbreeding Dominance Components Is Fit, with the Dominant Allele Having Frequency .8 and with Scenarios as given in Table 4

							TRUE (CALCULATED)	
	V_e	V_a	V_d	V_b	$\mathrm{Cov}_b(a,d)$	SS_{μ_b}	h^2	H^2
True model is fully dominant:								
True value	1.00	1.00	2.00	4.50	1.50	2.00		
Scenario:								
B	1.02	1.21	2.19				.23 (.29)	.77 (.77)
E	.95	1.44	2.25				.23 (.33)	.78 (.79)
F	.97	1.93	2.03				.22 (.42)	.80 (.80)
True model is three-quarters dominant:								
True value	1.00	3.36	2.00	4.50	1.50	2.00		
Scenario:								
A	1.02	3.59	2.18				.51 (.54)	.85 (.85)
D	.80	4.34	2.18				.47 (.62)	.87 (.89)

Let p be the frequency of the recessive allele (call this "allele 1"), and let $q = 1 - p$ be the frequency of the dominant allele (call this "allele 2"). In each case, let $m_l = [g_l(2,2) + g_l(1,1)]/2$ be the midpoint of the homozygote effects for locus l, let $\delta_l = g_l(1,2) - m_l$ be the excess of the heterozygote effect above the midpoint of the homozygote effects for locus l, let $\gamma_1 = \Sigma_{l=1}^{L}\delta_l$, and let $\gamma_2 = \Sigma_{l=1}^{L}\delta_l^2$, with L being the total number of loci. In the fully dominant model, we have $\delta_l = [g_l(2,2) - g_l(1,1)]/2$, and we can explicitly compute the variance components to be (Jacquard 1974) $\mu_b = -2pq\gamma_1$, $V_a = 8p^3q\gamma_2$, $V_d = 4p^2q^2\gamma_2$, $V_b = 4pq(p^3 + q^3)\gamma_2$, $\mathrm{Cov}_b(a,d) = -4p^2q(p - q)\gamma_2$, and $SS_{\mu_b} = V_d$.

In the three-quarter–dominant model, we have $\delta_l = \frac{3}{4}\{[g_l(2,2) - g_l(1,1)]/2\}$, so that γ_1 is three-quarters of its value and γ_2 is $(\frac{3}{4})^2$ of its value in the fully dominant case. In the three-quarter–dominant case, the expression for the additive variance becomes

$$V_a = 2pq\left(\frac{7}{3}p + \frac{1}{3}q\right)^2\gamma_2 .$$

The expressions for the other genetic-variance components are unchanged (only the values of γ_1 and γ_2 have changed).

For the case when $L = 1$, we would have $\mu_b = -\sqrt{V_d}$. As a consequence, for the model with all loci biallelic, our assumption that no more than one locus per chromosome has non-0 inbreeding depression is equivalent to an assumption that no more than one locus per chromosome has non-0 dominance variance. This is not generally true for nonbiallelic loci. In particular, μ_b can be 0 when V_d is non-0, for a locus with three or more alleles. This implies that, in general, it is possible to have a number of loci on a chromosome that exhibit non-0 dominance variance, with only one of them having non-0 inbreeding depression.

Hutterite Population

We here apply the methods described above to a sample of individuals taken from a large, complex Hutterite pedigree. The Hutterites are a religious sect that originated in the Tyrolean Alps during the 1500s. Between the mid 1700s and the mid 1800s, while in Russia, the population grew in size from ~120 to >1,000 members (Hostetler 1974). During the 1870s, ~900 of these members migrated to what is now South Dakota, and roughly half settled on three communal farms. The population has since expanded dramatically, with >35,000 Hutterites now living in >350 communal farms (called "colonies") in the northern United States and in western Canada. The Hutterites' communal life-style ensures that all members are exposed to a relatively uniform environment. Genealogical records trace all extant Hutterites to <90 ancestors who lived from the early 1700s to the early 1800s (Martin 1970). The relationships among these ancestors are unknown, but some of them may have been related. The three original South Dakota colonies have given rise to the three major subdivisions of the modern Hutterite population: the Schmiedeleut (S-leut), the Dariusleut (D-leut), and the Leherleut (L-leut). Members of each leut have remained reproductively isolated from each other since 1910 (Bleibtreu 1964).

The subjects of our study, the S-leut Hutterites of South Dakota, are descendants of 64 Hutterite ancestors and consist of four main lineages (Mange 1964). Information on the relationships among members of our sample are in the form of a 13-generation, 12,903-member

Table 6

Number of Simulations Having Various *P* Values, and Average Variance-Component Estimates, with Heritabilities, for Simulated Data Sets

A. Significance of Inbreeding Dominance-Variance Components, for Different Scenarios Given in Table 4

Scenario[a]	No. of Simulations Having P =			
	> .10	.10–.05	.05–.01	< .01
A	3	2	0	0
B	9	1	0	0
C	2	1	2	0
D	1	1	0	3
E	2	0	1	2
F	1	0	1	3
G	0	0	1	4

B. Average Variance-Component Estimates, with Heritabilities, for Simulated Data Sets and Scenarios as Given in Table 4[a]

Model and Scenario[b]	V_e	V_a	V_d	V_b	$\mathrm{Cov}_b(a,d)$	SS_{μ_b}	h^2	H^2
Three-quarters dominant model:								
Scenario A:								
True value	1.00	3.36	2.00	4.50	1.50	2.00	.51	.85
Mean (SE)	1.7 (.7)	2.6 (.4)	2.9 (1.0)				.38 (.06)	.76 (.09)
√MSE	1.5	1.1	2.2				.17	.20
Mean (SE)	1.8 (.5)	2.9 (.4)	2.3 (.9)	9.8 (4.4)	−2.3 (1.1)	.002 (.002)	.43 (.07)	.75 (.08)
√MSE	1.3	.9	1.8	10.3	4.4	2.0	.15	.19
Scenario D:								
True value	1.00	3.36	2.00	4.50	1.50	2.00	.47	.87
Mean (SE)	1.5 (.6)	3.5 (.5)	2.6 (.9)				.49 (.07)	.80 (.08)
√MSE	1.3	1.0	1.9				.13	.17
Mean (SE)	1.7 (.6)	2.9 (.5)	2.2 (.9)	15.0 (6.4)	−1.3 (1.6)	1e-5 (7e-5)	.39 (.06)	.78 (.07)
√MSE	1.3	1.1	1.8	16.5	4.2	2.0	.15	.17
Fully dominant model:								
Scenario B:								
True value	1.00	1.00	2.00	4.50	1.50	2.00	.23	.77
Mean (SE)	1.2 (.2)	1.0 (.1)	2.3 (.3)				.23 (.03)	.72 (.04)
√MSE	.5	.4	1.0				.10	.13
Mean (SE)	1.3 (.1)	1.3 (.2)	1.8 (.2)	7.0 (2.6)	−1.4 (.6)	2.3 (1.6)	.29 (.04)	.71 (.03)
√MSE	.5	.5	.7	8.2	3.4	4.7	.12	.12
Scenario C:								
True value	1.00	1.00	2.00	4.50	1.50	2.00	.23	.77
Mean (SE)	1.4 (.2)	1.2 (.1)	2.1 (.4)				.23 (.03)	.70 (.06)
√MSE	.6	.3	.7				.06	.14
Mean (SE)	1.4 (.2)	1.4 (.1)	1.7 (.4)	9.0 (6.3)	−1.4 (1.1)	.02 (.02)	.31 (.02)	.69 (.06)
√MSE	.6	.4	.9	13.4	3.6	2.0	.09	.14
Scenario E:								
True value	1.00	1.00	2.00	4.50	1.50	2.00	.23	.78
Mean (SE)	1.1 (.2)	1.5 (.3)	2.2 (.4)				.32 (.06)	.76 (.04)
√MSE	.4	.7	.8				.15	.07
Mean (SE)	1.2 (.2)	1.1 (.2)	1.9 (.4)	6.1 (.9)	.4 (.9)	3.3 (3.3)	.24 (.05)	.75 (.04)
√MSE	.4	.5	.9	2.5	2.1	6.7	.10	.08
Scenario F:								
True value	1.00	1.00	2.00	4.50	1.50	2.00	.22	.80
Mean (SE)	1.3 (.3)	1.9 (.2)	1.8 (.3)				.40 (.03)	.73 (.05)
√MSE	.6	.9	.7				.19	.12
Mean (SE)	1.3 (.2)	1.3 (.3)	1.6 (.4)	14.1 (5.4)	−1.4 (1.2)	1e-4 (2e-5)	.28 (.05)	.73 (.04)
√MSE	.5	.6	.8	14.5	3.8	2.0	.12	.11
Scenario G:								
True value	1.00	1.00	2.00	4.50	1.50	2.00	.22	.80
Mean (SE)	1.3 (.2)	2.0 (.1)	1.9 (.4)				.41 (.03)	.75 (.05)
√MSE	.6	1.0	.8				.20	.12
Mean (SE)	1.4 (.2)	1.4 (.1)	1.6 (.4)	10.2 (4.2)	−1.3 (1.1)	3.4 (3.4)	.29 (.02)	.73 (.05)
√MSE	.6	.4cx	.8	10.2	3.6	6.9	.08	.13

[a] The number of replicates for scenarios A and C–G is 5; that for scenario B is 10.

[b] MSE = average squared error of the estimate.

genealogy. We focus on a subset of these individuals, consisting of everyone of age >5 years (736 individuals) in nine colonies drawn from three of the four lineages of the S-leut, with an additional 70 S-leut individuals from other colonies. For these 806 individuals, we have a 13-generation pedigree consisting of 1,623 individuals from which we calculate all nine identity coefficients for every pair of individuals among the 806. Extensive information has been collected on a number of traits in these colonies. We focus here on HDL level, which is available for 521 of the individuals.

Results

Identity Coefficients and Pairwise Relationships in the Hutterites

For the 806 Hutterite individuals in our study, we calculated all nine identity coefficients for every pair of individuals, as well as the two possible identity coefficients for each individual with him/herself. The computational difficulty of this operation should not be underestimated, since (*a*) very few pairs share the same relationship (see Appendix A) and (*b*) there are a large number of pairs that need to be considered (325,221 total pairs among the 806 individuals, when individuals are included with themselves). Furthermore, the recursive algorithms of Karigl (1981) require that one evaluate kinship coefficients of not only pairs but also of trios and quartets of individuals (see Appendix B)—and not only among the 806 study individuals themselves but also among their ancestors in the 13-generation 1,623-member pedigree. This results in many billions of combinations. Initial estimates indicated that calculating all the identity coefficients could take years with a simple execution of the algorithms, but, with some computational speed-ups (see Appendix B), we were able to do the entire calculation in <1 wk of computer time. Note that these calculations are based only on the pedigree, not on the phenotype. Thus, these identity coefficients need only be calculated once for a given set of individuals, with the same results used for variance-components analyses of any phenotype, for any study, including mapping and heritability studies, based on any subset of those individuals.

Calculation of the identity coefficients for all pairs of individuals among 806 S-leut Hutterites provides a rare opportunity to consider in detail the structure of relationships among individuals in an isolated population. In figure 2, we plot Δ_8 (i.e., the probability that a pair of individuals shares two alleles IBD and that neither individual's own alleles are IBD) against $\sqrt{\Delta_7}$ (i.e., the square root of the probability that a pair of individuals shares exactly one allele IBD and that neither individual's own alleles are IBD) for the 324,415 pairs that consist of two distinct individuals from the 806 Hutterites in the sample. The square-root transformation for Δ_7 is used simply to reduce the amount of empty space in the plot and to make the distinct features more visible. Various point clouds are labeled, representing certain close relationship types. These points have been classified into groups based on their approximation by various outbred relationships. Table 1 lists the relationships corresponding to the different labels, giving also the number of pairs having that approximate relationship, and the location of the single point on the plot where the relationship would lie if it occurred for an outbred pair. To aid in comparison of the point clouds with these reference points, horizontal lines corresponding to $\Delta_8 = .25$ and $\Delta_8 = .5$ are added to the plot.

In most cases, a point on the plot represents a pair of sibships. The reason is that, if A and B are sibs and if neither is an ancestor to C, then the relationship between A and C is the same as that between B and C. Similarly, if there are two sibships of size m and n, then all $m \times n$ pairs, in which one individual is taken from each sibship, have the same relationship (see Appendix A). Thus, the points representing these relationships are coincident. For the sib pairs in the "s" cloud, each sibship is represented by a single point. Points for different sibships in the "s" cloud may occasionally coincide—for example, in the case of two sibships that are double first cousins to one another. The points along the left side of figure 2, with $\sqrt{\Delta_7} = 0$, are due to three individuals—a sib pair whose mother is not known to have any ancestors in the pedigree and another individual whose father is not known to us. For the remaining 803 individuals, identity coefficients Δ_1,Δ_3, Δ_5, Δ_7, and Δ_8 are non-0 for every pair of individuals. Below, when we consider the logarithms of these five identity coefficients, we first eliminate these three individuals. Note that Δ_2,Δ_4, Δ_6, and Δ_9 will be 0 for every parent-offspring pair.

All the half-sib pairs occurring in the sample are of a special type, given by the relationship between individuals 1 and 2 in figure 3*A*, which we call "half-sib plus first cousin"; that is, the two individuals share a father and their mothers are sisters. (It is not uncommon in the Hutterites for a widower to marry his deceased spouse's sister.) There are 68 such half-sib–plus–first-cousin pairs in the sample, but they are represented in figure 2 by only two points, labeled "h+c", because they all arise from two occurrences of a widower marrying his deceased wife's sister. In one case, a man had 6 children with one woman and 10 with her sister; in another case, a man had 8 children with one woman and 1 with her sister. Other interesting relationships arise from such families. These relationships are complex, so, in describing them, we use parentheses to convey the order in

which relationship modifiers such as "once removed" are to be applied. The relationships arising in figure 3*A* include what we call "(half-sib plus first cousin) once removed"—that is, the relationship between individuals 1 and 4 and that between individuals 2 and 3, in figure 3*A*—and "cousin plus second (half-sib plus first cousin)"—that is, the relationship between individuals 3 and 4 in figure 3*A*. Both of these relationships occur in the sample, and the resulting sets of points are labeled "*v*" and "*w*," respectively, in figure 2. The cloud of points *x* in figure 2 represents relationships of the type between individuals 1 and 2 in figure 3*B*, which we call "cousin plus (cousin once removed)," because the two individuals are first cousins through their fathers and are first cousins once removed through their mothers.

In figure 2, the large cloud of points *y* represents the relationships first cousin, great-grandparent/grandchild, grand-avuncular, half-avuncular, and (double cousin) once removed. Note that (double cousin) once removed is distinct from the relationship double (cousin once removed), which is represented by point cloud *z*. The first of these two occurs, for instance, when individual A is a double first cousin to a parent of individual B. The second of the two occurs, for instance, when individual A is a first cousin to both the mother and father of individual B, but through separate lines of descent, so that the mother and father of individual B need not be related. Alternatively, individual A could be a first cousin to one parent of B, and individual B could be a first cousin to one parent of A, through separate lines of descent. All of these situations appear in the Hutterite sample.

Panels *A* and *B* of figure 4 show the histograms of $\log_{10}(\Phi)$ and $\mathbf{f}$, respectively, for the 806 individuals in the sample. Their mean kinship coefficient is .042 with SD .031, although, on the basis of the rough symmetry on the log scale in figure 4*A*, it is perhaps more natural to take the geometric mean, which is .036. These values, .042 and .036 are between first cousins (.0625) and first cousins once removed (.03125). The mean inbreeding coefficient in the sample is .034, which is slightly more than that for the offspring of first cousins once removed. In figure 5 we plot the inbreeding coefficient as a function of year of birth for the 806 individuals in our sample. It is evident that, over time, there has been an increase in the level of inbreeding of the Hutterites, as would be expected, given their reproductive isolation, and as has been noted previously by Ober et al. (1999). We note that this effect is not explained by any decreased longevity of individuals with higher inbreeding. The phenomenon persists when deceased individuals are included (not shown). Histograms of the identity coefficients, some of them on the log scale, are given in figure 6, where Δ_3 and Δ_5 are combined and Δ_4 and Δ_6 are combined. Note that the combined distribution of Δ_3 and Δ_5 is almost exactly log normal and that Δ_1 is also close to log normal. Means and SDs are given in table 2.

Variance-Components Analysis of HDL in the Hutterites

Using the methods described above, we estimated the inbreeding depression and components of variance of HDL in the Hutterites. Because of skewness in the HDL distribution, we normalized by taking the square root of the phenotype and applied ML to the transformed data. Fixed effects of age and sex and a constant term were included in all models. We fit several models, setting some of the parameters to 0; the results are shown in table 3. In addition, we fit models including $\mathrm{Cov}_h(a,d)$ and SS_{μ_h}. However, estimation of $\mathrm{Cov}_h(a,d)$ and SS_{μ_h} made very little change in the log likelihood, and the estimates are consistently close to 0. Models including these parameters are not included in table 3. When all models were refit with REML instead of ML, the results were virtually identical.

Note that the constraints of our model dictate that, if $SS_{\mu_h} = 0$, then $\mu_h = 0$, and that, if $V_d = 0$, then all other inbreeding dominance components and the inbreeding depression are 0. However, this constraint has little practical impact if we allow the number of loci to be arbitrary. In that case, if the value of V_d is infinitesimal but still positive, then the other dominance components are unconstrained. Thus, we argue that in any real data set, a model with, say V_e, V_a, and V_h estimated with V_d set to 0 can be interpreted as an approximation to the case in which V_d is set to some arbitrarily small positive value.

When the log likelihood of model 1 is compared with that for the other models, it is clear that HDL has a strong genetic component. Model 2, with both V_e and V_a, was most favored according to BIC, giving heritabilities of 65%. Model 6, which includes the inbreeding dominance component V_h, is most favored according to AIC. A likelihood-ratio test comparing model 6 to model 2 is at the borderline of significance, giving a *P* value of .052. In all cases, the inbreeding depression μ_h was not significant. Furthermore, the estimate of V_a was largely unaffected by which other variance components were included in the analysis. This is also evident in the fact that the narrow-sense heritability, h^2, is little changed from model to model. The broad-sense heritability, H^2, has a slightly different estimate, depending on whether any dominance components are included, but the magnitude of these variations is close to the level of the sampling variability.

A lack of detectable inbreeding depression and only suggestive evidence of any non-0 inbreeding dominance components in the data may be influenced by a number

of factors, particularly the genetic model for the trait, the level of inbreeding, and the sample size. Through numerical examples and simulations, we examined the roles of these factors.

Comparison of Models With and Without Additional Inbreeding Dominance Components and Inbreeding Depression

For a given pedigree and genetic model, we compared the results that would be obtained asymptotically—that is, without sampling variability—under the models with and without the additional inbreeding dominance components and inbreeding depression. We did this by projecting the full model onto the subspace with the inbreeding dominance components and inbreeding depression set to 0, where the projection is obtained by minimizing the Kullback-Leibler divergence as described in the Methods section. The first genetic model that we considered is a biallelic fully dominant model with all loci having frequency $q = .8$ for the dominant allele. This frequency of .8 was chosen to be in a range in which the ratios of the inbreeding dominance components to the additive variance are relatively high. We set the relative magnitudes of the genetic effect and environmental variance to give narrow- and broad-sense heritabilities in the ranges of .22–.23 and .77–.80, respectively, for the populations that we considered. We also considered a biallelic three-quarters-dominant model with all loci having frequency $q = .8$ for the dominant allele. We set the magnitudes of the genetic effect and environmental variance to give the same values of dominance-variance components and V_e as are in the fully dominant model that we chose. We examined the models, conditional on each of three different populations: (i) the 806-member Hutterite sample, (ii) the Hutterite sample with an additional 800 independent offspring of first-cousin marriages added to the sample, and (iii) the Hutterite sample with an additional 500 independent offspring of avuncular marriages added to the sample, where (ii) and (iii) are used to increase the level of inbreeding. The cases in which projections are examined are scenarios A, B, and D–F in table 4; results are given in table 5. The estimate of the total genetic component of variance was essentially unaffected by use of the wrong model, although the additive variance could be considerably inflated when the level of inbreeding was .048 or .069.

Simulation Studies

We performed small-scale simulation studies to see in which situations we could detect that the extra dominance components and inbreeding depression are non-0—and whether they could be estimated with reasonable accuracy. The time-consuming nature of the computations led us to consider only a few realizations, so these should be considered simulated examples, rather than a full-scale simulation study of the sampling distribution of the estimates. Section A of table 6 gives the ranges of P values for the likelihood-ratio test of the full model versus the model with all inbreeding dominance components set to 0. For the fully dominant models with average level of inbreeding .069, there appears to be reasonable power to detect the inbreeding dominance components, (four of five cases and five of five cases, respectively, for scenarios F and G). When the Hutterite sample is doubled and the model is fully dominant (scenario C), in two of five cases the inbreeding dominance components were significant. Section B of table 6 gives the averages of the estimates when the full and restricted models are used under the different simulated scenarios, with square roots of their average squared errors. From these results, it appears that the sample sizes considered here are not adequate to obtain accurate estimates of the inbreeding dominance components, even though there may be some power to detect that they are different from 0.

Discussion

Both calculation of the nine condensed coefficients of identity and estimation of dominance-variance components in the Hutterite sample are computationally challenging because of the size and complexity of the sample's genealogy (13 generations, 1,623 individuals). We succeeded in calculating the identity coefficients, and we used these to examine pairwise relationships in the Hutterite population. We fitted variance-component models including the extra inbreeding dominance components to HDL in the Hutterite population, apparently the first time that such components have been estimated in a human population. In previous studies of inbreeding depression in humans, socioeconomic factors may have been confounded with consanguinity of parents (Barrai et al. 1964; Bittles and Neel 1994). In the Hutterite population that we studied, however, the belief system and the custom of communal living create a very homogeneous environment and social structure, and no such association is expected. We note that there is structure in mate selection with respect to Hutterite-colony lineages, with preferences for marriages involving two individuals from the same lineage (Bleibtreu 1964; O'Brien 1987; Ober et al. 1997). One might expect that inbreeding may be associated with the type of marriage (i.e., whether the individuals are from different lineages or, if not, which lineage they are from). There is, however, no evidence for any association of this type (authors' unpublished data).

Variance-components analysis of HDL in the Hutterites does not indicate significance of any of the inbreeding dominance components or inbreeding depression. Both ML and REML are used to estimate the variance components, with little difference between the results. A similar study of the annual plant *Nemophila menziesii* (Shaw et al. 1998) found significance of inbreeding dominance components and inbreeding depression in several traits, whereas a study of sheep (Shaw and Woolliams 1999) did not. Both studies had much higher inbreeding levels than were seen in the Hutterites.

The results of the data analysis, numerical examples, and simulations suggest that, even with levels of inbreeding that are high for human populations, and even with a model having large inbreeding dominance-variance components and inbreeding depression, a quite substantial sample size would be necessary in order to obtain reasonable estimates of the components. In our simulated samples of 1,306 individuals with average inbreeding .069 in a fully dominant model, there is apparently some power to detect non-0 inbreeding dominance components (detected in four of five cases). However, even when that sample size is doubled, the estimates of these components are poor (although, with only a few simulations, the conclusions that we can draw about the sampling distributions of these estimators are necessarily limited). It is worth noting that the models that we simulated have biallelic loci. Models with multiallelic loci differ in some respects and potentially may produce dominance effects that are more pronounced than those which can be achieved with biallelic loci. Estimates of heritability did not seem to be very sensitive to estimation of the additional dominance components. The sample size would probably need to be much larger in order for the impact of the inbreeding dominance components on heritability to be above the level of the sampling variability. For mapping, this suggests that, at least for modeling of background polygenic effects, consideration of inbreeding dominance components will not have a great effect in studies of humans.

In estimating variance components, we have assumed a multivariate normal model for the genetic effects. Lange (1978) has given sufficient conditions for a central-limit theorem on pedigrees but explicitly excludes the case with both inbreeding and non-0 dominance variance. We have extended the work of Lange (1978) to a central-limit theorem on pedigrees that have both inbreeding and non-0 dominance variance, provided that no more than one locus per chromosome has non-0 inbreeding depression, which is the situation that we have assumed in our models. For mapping, the basis of the assumption of multivariate normality is questionable. Assuming a major-gene model, one might expect the distribution of the genetic effects to be a mixture of normal distributions or of some other distribution, but fitting such a model to the Hutterite population is computationally impractical, especially if it must be done repeatedly in a search for genes. Amos et al. (1996) considered the robustness of maximum-likelihood estimators of genetic parameters obtained under an assumption of normality. In simulations of a major-gene effect modeled by a mixture of normal distributions and having no dominance variance, Amos et al. (1996) found that estimators obtained under the assumption of normality performed well. For nonnormal distributions, Beaty et al. (1985), Amos (1994), and Amos et al. (1996) have considered quasi-likelihood estimation.

We have neglected epistasis and assortative mating in our models. If present, assortative mating would have the effect of inflating the additive variance (Crow and Kimura 1970). Although it is not possible, on the basis of phenotype data alone, to estimate epistatic variance components in natural populations, epistasis can greatly inflate the additive and/or dominance components, and it may prove to have great practical importance for mapping.

Acknowledgments

We thank Nancy Cox for her valuable discussions and thoughtful input. The comments and suggestions of the anonymous reviewers were helpful in clarifying various points throughout this article. This work was supported by Sloan Foundation Fellowship 97-7-1 CMB (to M.A.), by National Institutes of Health grants HG01645 (to M.S.M.), HL49596 (to C.O.), and HL56399 (to C.O.), and by a grant from Hoffman-La Roche, Inc. (to C.O.).

Appendix A

Notes on Relationships

First, we define pairwise relationships and a partial ordering on the set of pairwise relationships. Next, we define the n-generation pedigree for two individuals. With these definitions, it is clear that the n-generation pedigree for two individuals gives a lower bound on the true relationship for the pair, with the accuracy of the approximation increasing with n. Finally, we describe some conditions that are sufficient for two different relative pairs to have the same relationship. These sufficient conditions are particularly relevant in an inbred isolate such as the Hutterites, where most instances in which two pairs have the same relationship, on the basis of the observed 13-generation pedigree, can be attributed to one of these conditions.

Relationships can be thought of as equivalence classes on pedigrees. Here we restrict attention to pairwise re-

lationships. The pedigrees that we consider include nodes only for the two individuals, call them "*a*" and "*b,*" and for some finite nontrivial subset of ancestors of *a* and/or *b* (with no more than one node for each individual), with directed edges from parents to offspring and with every node connected by a path to either *a* or *b*. Furthermore, we restrict attention to pedigrees in which every node *c* in the pedigree is connected to *a* or *b* in such a way that a *directed* path can be taken from *c* to either *a* or *b* (i.e., each step is from a parent to an offspring). (An example of a pedigree ruled out by the last condition would be that resulting from the family depicted in fig. 3*B,* if individual 2's maternal grandmother had a node in the pedigree but individual 2's mother did not.) Let this set of pedigrees be denoted by **P**. We define P_1 and P_2 in **P** to be equivalent if there exists $P_3 \in \mathbf{P}$ such that there are injective maps f_1 and f_2 mapping the nodes of P_1 and P_2, respectively, into the nodes of P_3, where the maps preserve *a* and *b* (which we take to be uniquely defined in each element of **P**) and where, for each directed edge in P_3, say from *c* to *d,* at least one of the following two conditions must hold: (i) there is no other directed edge starting from *c,* or (ii) there exist nodes c' and d' in P_1 and c'' and d'' in P_2 such that $c = f_1(c') = f_2(c'')$, $d = f_1(d') = f_2(d'')$, there is a directed edge from c' to d' in P_1, and there is a directed edge from c'' to d'' in P_2. It can be shown that this is an equivalence relation on **P**. We define the set of pairwise relationships to be the resulting set of equivalence classes **C**. We now define a partial ordering on this set. For relationships $C_1, C_2 \in \mathbf{C}$, we define $C_1 \prec C_2$ (i.e., C_2 represents a closer relationship than C_1), if there are pedigrees $P_1 \in C_1$ and $P_2 \in C_2$ and a surjection f_1 mapping the nodes of P_1 onto the nodes of P_2 such that *a* and *b* are preserved and such that, if there is a directed edge from *c* to *d* in P_1, then there is a directed edge from $f_1(c)$ to $f_1(d)$ in P_2. It can be shown that this satisfies the conditions of a partial ordering.

Note that members of a pedigree cannot necessarily be divided into discrete generations (e.g., see fig. 3*B*). However, we can, somewhat arbitrarily, define the *n*-generation pedigree for individuals *a* and *b* to be the element $P \in \mathbf{P}$ having the largest number of nodes among those *P* satisfying the following conditions: there is an injective map *f* from the nodes of *P* to the set of individuals ancestral to *a* and/or *b,* such that there is a directed edge from *c* to *d* in *P* if and only if $f(c)$ is a parent of $f(d)$ and such that there is at least one directed path of length $\leq n$ from any element to either *a* or *b*. When P_n is taken to be the *n*th-generation pedigree of *a* and *b* and C_n is taken to be the equivalence class with P_n as a member for $n \geq 1$, it can be shown that $C_n \prec C_{n+1}$ for all *n*. From this we conclude that the *n*th-generation pedigree for two individuals gives a lower bound on the true relationship for the pair, with the accuracy of the approximation increasing with *n*.

Finally, we describe conditions sufficient for two relative pairs to have the same relationship when all four individuals are embedded in a large pedigree. Define two individuals to be sibs if and only if they have the same parents. Note that different sib pairs need not have the same relationship. Define two individuals *a* and *b* to be double first cousins if and only if (i) parent 1 of *a* is a sib to parent 1 of *b* and parent 2 of *a* is a sib to parent 2 of *b,* for some choice of labeling of *a*'s and of *b*'s parents, and (ii) *a*'s parents are not sibs. Again, two different double–first-cousin pairs need not have the same relationship. Then conditions sufficient for two relative pairs to have the same relationship include the following: (i) if individuals A and B are sibs and neither is an ancestor to C, then the relationship between A and C is the same as that between B and C; (ii) as a consequence of (i), we conclude that, if A and B are sibs and C and D are sibs and if none is ancestral to another, then A and C have the same relationship as B and C, A and D, and B and D; (iii) if individuals A and B are double first cousins and if they are neither ancestors nor offspring of C and if C is not a descendant of a parent of A or B, then the relationship between A and C is the same as that between B and C. In the Hutterites, these conditions account for the vast majority cases in which two pairs of individuals are found to have the same relationship based on the observed 13-generation pedigree.

Appendix B

Computational Speed-Ups in the Calculation of Identity Coefficients

To calculate the identity coefficients, we follow the method of Karigl (1981). This method requires that one calculate a set of generalized kinship coefficients, from which one can obtain the identity coefficients via a linear transformation. The generalized kinship coefficients are Φ_{ab}, the standard kinship coefficient for two individuals, together with kinship coefficients for three individuals, four individuals, and two pairs of individuals: Φ_{abc}, Φ_{abcd} and $\Phi_{ab,cd}$, respectively. The definition of Φ_{ab} is the probability that a randomly chosen allele from *a* is IBD with a randomly chosen allele from *b*. Similarly, Φ_{abc} (or Φ_{abcd}) is the probability that three (or four) randomly chosen alleles, one from each individual, are IBD. $\Phi_{ab,cd}$ is the probability that a random allele from *a* is IBD with a random allele from *b* and that a random allele from *c* is IBD with a random allele from *d*. Using the notation $a \not\geq b$ to mean that *a* is not an ancestor of *b* and using "*f*" and "*m*" to denote, respectively, the

father and mother of individual *a*, we can write recursion formulas for the generalized kinship coefficients:

$$\begin{aligned}
\Phi_{ab} &= \frac{1}{2}(\Phi_{fb} + \Phi_{mb}) && \text{for} \quad a \ngeq b \\
\Phi_{aa} &= \frac{1}{2}(1 + \Phi_{fm}) \\
\Phi_{abc} &= \frac{1}{2}(\Phi_{fbc} + \Phi_{mbc}) && \text{for} \quad a \ngeq b,c \\
\Phi_{aab} &= \frac{1}{2}(\Phi_{ab} + \Phi_{fmb}) && \text{for} \quad a \ngeq b \\
\Phi_{aaa} &= \frac{1}{4}(1 + 3\Phi_{fm}) \\
\Phi_{abcd} &= \frac{1}{2}(\Phi_{fbcd} + \Phi_{mbcd}) && \text{for} \quad a \ngeq b,c,d \\
\Phi_{aabc} &= \frac{1}{2}(\Phi_{abc} + \Phi_{fmbc}) && \text{for} \quad a \ngeq b,c \\
\Phi_{aaab} &= \frac{1}{4}(\Phi_{ab} + 3\Phi_{fmb}) && \text{for} \quad a \ngeq b \\
\Phi_{aaaa} &= \frac{1}{8}(1 + 7\Phi_{fm}) \\
\Phi_{ab,cd} &= \frac{1}{2}(\Phi_{fb,cd} + \Phi_{mb,cd}) && \text{for} \quad a \ngeq b,c,d \\
\Phi_{aa,bc} &= \frac{1}{2}(\Phi_{bc} + \Phi_{fm,bc}) && \text{for} \quad a \ngeq b,c \\
\Phi_{ab,ac} &= \frac{1}{4}(2\Phi_{abc} + \Phi_{fb,mc} + \Phi_{mb,fc}) && \text{for} \quad a \ngeq b,c \\
\Phi_{aa,ab} &= \frac{1}{2}(\Phi_{ab} + \Phi_{fmb}) && \text{for} \quad a \ngeq b \\
\Phi_{aa,aa} &= \frac{1}{4}(1 + 3\Phi_{fm}).
\end{aligned}$$

Furthermore, $\Phi_{ab} = \Phi_{abc} = \Phi_{abcd} = 0$ when there is no common ancestor to the four individuals *a*–*d* and $\Phi_{ab,cd} = 0$ unless there are two common ancestors, one for *a* and *b* and one for *c* and *d*. From these rules, all of the generalized kinship coefficients may be calculated for a given pedigree.

The identity coefficients may be found from the kinship coefficients, on the basis of the following linear transformation:

$$\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
2 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 \\
2 & 2 & 1 & 1 & 2 & 2 & 1 & 1 & 1 \\
4 & 0 & 2 & 0 & 2 & 0 & 2 & 1 & 0 \\
8 & 0 & 4 & 0 & 2 & 0 & 2 & 1 & 0 \\
8 & 0 & 2 & 0 & 4 & 0 & 2 & 1 & 0 \\
16 & 0 & 4 & 0 & 4 & 0 & 2 & 1 & 0 \\
4 & 4 & 2 & 2 & 2 & 2 & 1 & 1 & 1 \\
16 & 0 & 4 & 0 & 4 & 0 & 4 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
\Delta_1 \\ \Delta_2 \\ \Delta_3 \\ \Delta_4 \\ \Delta_5 \\ \Delta_6 \\ \Delta_7 \\ \Delta_8 \\ \Delta_9
\end{pmatrix}
=
\begin{pmatrix}
1 \\ 2\Phi_{aa} \\ 2\Phi_{bb} \\ 4\Phi_{ab} \\ 8\Phi_{aab} \\ 8\Phi_{abb} \\ 16\Phi_{aabb} \\ 4\Phi_{aa,bb} \\ 16\Phi_{ab,ab}
\end{pmatrix}.$$

The recursive nature of the kinship-coefficient equations suggests two basic alternative strategies for their calculation. The first method is to start with the founders of the pedigree and descend through the pedigree one generation at a time, calculating the kinship coefficients for all pairs within a generation. This method proves efficient in that coefficients for particular combinations of individuals are calculated only once. Also, in the case in which there are many generations, the information necessary in order to calculate the kinship coefficients for the next generation is encompassed entirely within the current generation and only those members of previous generations who have mated with members in the current generation. If there are no cross-generational matings, only the coefficients for the current generation need to be retained. Although this method is a standard way to calculate the two-person kinship coefficient, it proves problematic when one is working with a large, complex pedigree such as the Hutterites. The difficulty is a consequence of the very large number of three-individual, four-individual, and two-pair kinship coefficients. As a result, the memory demands required in order to consider any generation in the Hutterites, aside from the first few, quickly exceed the capacity of available computers. It is conceivable that the memory demands could be reduced by a consideration of only those groupings of individuals that are needed to calculate the identity coefficients for the study sample, rather than all possible groupings within a generation; however, there is no *a priori* method to identify these combinations. For these reasons, this approach is impractical, given the size and complexity of the Hutterite pedigree.

Another strategy may be characterized as a bottom-up approach. Here, one starts with a particular pair for which the identity coefficients are desired and recursively applies the equations discussed above, calculating only those kinship coefficients that are needed for that particular pair. The disadvantage here is that, even though we calculate only those kinship coefficients that are necessary for the pair, when we select the next pair from our study sample we may end up recalculating many of the same values, wasting large amounts of time. If, instead, we are able to store these values and recall them as needed, there is potential for greatly speeding up the process. The problem then becomes one of efficiently storing and recalling the various kinship coefficients that have been calculated, given the large number of combinations and the need to minimize memory search time. Also, because it is unclear how many of each type of kinship coefficient will be calculated, the memory scheme must be flexible.

To satisfy these requirements, we have implemented a hash table. This allows us to set aside a large amount of memory that could be used to store any of the coefficients, as necessary, but provides a fairly fast method for recall of previously calculated values. Efficiency is improved by ordering the pairs such that one may take

advantage of relationship equivalency classes. For instance, once person A's coefficients of identity with a set of people B_i have been calculated, the coefficients of identity of that set of people with A's siblings are immediately known, as long as none of the B_i are descendants of either A or a sibling of A.

Another problem arises because, even though we are now restricting the calculations to only those kinship coefficients that are necessary for the pairs in question, the available memory can become quickly exhausted (we used an R10000 CPU on an SGI Power Challenge XL with 1 GB of RAM dedicated to the computation). In such a circumstance, we decided to erase the existing calculations and to refill the memory as needed. Although this procedure entails recomputation of many kinship coefficients, it proves to be quicker than it would be to replace existing calculations one at a time, since this latter procedure requires the overhead of searching through the entire memory space before the replacement. Implementation of the strategy outlined above allowed us to calculate the nine identity coefficients for all 325,221 pairs in our study sample, in ~5 d of computer time.

Appendix C

Sufficient Conditions for a Central-Limit Theorem with Both Inbreeding and Non-0 Dominance Variance

Lange's (1978) conditions for a central-limit theorem for polygenic-trait values in a pedigree or collection of pedigrees require either no inbreeding or no dominance variance. In the present report, we consider models in which both inbreeding and dominance variance are allowed, with the restriction that no more than one locus per chromosome has non-0 inbreeding depression. We extend Lange's (1978) second central-limit theorem to this case. Following Lange (1978), we assume Hardy-Weinberg equilibrium and that all loci are in linkage equilibrium, that there is no assortative mating or epistasis, that the number of chromosomes goes to infinity, that there are a fixed number of individuals m, and that there is an upper bound q on the number of loci per chromosome.

Let X_k^i be the random contribution of locus k to the trait value of individual i. Let H_k^i denote the event that individual i's two alleles at locus k are IBD. Let μ_{hk} be the inbreeding depression at locus k. Following Lange (1978), we have $\mathrm{Cov}(X_k^i,X_l^j) = [Pr(H_k^i \cap H_l^j) - Pr(H_k^i)Pr(H_l^j)]\mu_{hk}\mu_{hl} = [Pr(H_k^i \cap H_l^j) - f_i f_j]\mu_{hk}\mu_{hl}$. If loci k and l are linked, the relationship between individuals i and j can be chosen so that the first factor is either positive or negative. Thus, we have the following result.

LEMMA: Assume that $k \neq l$. Then $\mathrm{Cov}(X_k^i,X_l^j) = 0$ for all possible relationships between individuals i and j if and only if at least one of the following holds: (i) loci k and l are unlinked or (ii) either $\mu_{hk} = 0$ or $\mu_{hl} = 0$.

As noted in the Methods section, if k is a biallelic locus, $\mu_{hk} = 0$ implies $V_{dk} = 0$. However, for a locus with more than two alleles, it is possible to have large dominance variance with 0 inbreeding depression. Thus, condition (ii) is weaker than Lange's (1978) condition (c) (absence of dominance variance).

Let X_k be the random column vector with ith entry X_k^i. Let $S_L = \sum_{k=1}^L X_k$. Here, L is the number of loci, and we will let $L \to \infty$. Let $m_L = E(S_L) = f\sum_{k=1}^L \mu_{hk}$, where f is the vector of inbreeding coefficients. Suppose for the moment that $\Omega_L \equiv \mathrm{Var}(S_L)$, calculated by means of equations (2) and (3), with V_e set to 0, is positive definite. For any column vector u with m components, we have $\mathrm{Var}[u^T(S_L - m_L)] = u^T\Omega_L u$. Let $a_L^2 = \sum_{k=1}^L (V_{ak} + V_{dk} + V_{hk} + \mathrm{Cov}(a,d)_k + \mu_{hk}^2)$, where V_{ak} is the additive variance, V_{dk} is the dominance variance, V_{hk} is the homozygous dominance variance, and $\mathrm{Cov}(a,d)_k$ is the homozygous additive by dominance covariance due to the kth locus. Assume that $V_{ak} + V_{dk} > 0$ for each k. This implies that $a_L^2 > 0$ for each L. We give sufficient conditions for $(S_L - m_L)/a_L \Rightarrow \mathrm{MVN}(0,\Sigma)$. We apply Orey's (1958) univariate central-limit theorem for q-dependent random variables to the sequence $u^T(X_1 - f\mu_{h1}),\ldots,u^T(X_L - f\mu_{hL})$. If we assume that we have ordered the loci by chromosome and that there are $\leqslant q$ loci per chromosome, then this sequence is q dependent. Following Lange (1978), under the conditions of the Lemma, we will have $u^T(S_L - m_L)/\sqrt{u^T\Omega_L u} \Rightarrow \mathrm{N}(0,1)$, the standard normal distribution, provided that the Lindeberg condition $1/b_L^2 \sum_{k=1}^L \int_{\{|W_k| > \epsilon b_L\}} W_k^2 dP \to 0$ holds for every $\epsilon > 0$, where $W_k = u^T(X_k - f_{\mu_{hk}})$ and $b_L^2 = u^T\Omega_L u$. The following is an extension of Lange's (1978) second central-limit theorem to the case of non-0 dominance variance with inbreeding, with no more than one locus per chromosome having non-0 inbreeding depression:

THEOREM: Suppose that $V_{ak} + V_{dk} > 0$ for all k, $(1/a_L^2)\sum_{k=1}^L V_{ak} \to \sigma_a^2 < \infty$, $(1/a_L^2)\sum_{k=1}^L V_{dk} \to \sigma_d^2 < \infty$, $(1/a_L^2)\sum_{k=1}^L V_{hk} \to \sigma_h^2 < \infty$, and $(1/a_L^2)\sum_{k=1}^L \mathrm{Cov}_h(a,d)_k \to \sigma_{ad}$ with $|\sigma_{ad}| < \infty$ (so $(1/a_L^2)\sum_{k=1}^L \mu_{hk}^2 \to s = 1 - \sigma_a^2 - \sigma_d^2 - \sigma_h^2 - \sigma_{ad}$). Also assume that $\lim_{L\to\infty}(1/a_L^{2+\delta})\sum_{k=1}^L \max_{j\in\{1,\ldots,m\}} E|X_k^j - f_j\mu_{hk}|^{2+\delta} = 0$ for some $\delta > 0$. Then $(S_L - m_L)/a_L \Rightarrow \mathrm{MVN}(0,\Omega)$, where Ω is as given in equations (2) and (3), with $V_e = 0$, V_i replaced by σ_i^2, $i = a,d,h$, $\mathrm{Cov}_h(a,d)$ replaced by σ_{ad}, and SS_{μ_h} replaced by s.

The proof closely follows that of Lange (1978). It suffices to prove that $u^t(S_L - m_L)/a_L \Rightarrow N(0,u^T\Omega u)$ for each column vector u. If $u^T\Omega u = 0$, then $\mathrm{Var}(u^T(S_L - m_L)/a_L) \Rightarrow 0$, and the result is immediate. Otherwise, for each random variable R, define $||R||_{2+\delta} = (E|R|^{2+\delta})^{1/(2+\delta)}$. Applying Minkowski's inequality and letting $Y_k^j = X_k^j - f_j\mu_{hk}$, we get

$$\begin{aligned}\sum_{k=1}^{L} E|u^T Y_k|^{2+\delta} &= \sum_{k=1}^{L} ||u^T Y_k||_{2+\delta}^{2+\delta} \\ &\leq \sum_{k=1}^{L} \left(\sum_{j=1}^{m} ||u_j Y_k^j||_{2+\delta}\right)^{2+\delta} \\ &= \sum_{k=1}^{L} \left(\sum_{j=1}^{m} |u_j| \ \ ||Y_k^j||_{2+\delta}\right)^{2+\delta} \\ &\leq \left(\sum_{j=1}^{m} |u_j|\right)^{2+\delta} \sum_{k=1}^{L} \max_{j\in\{1,\ldots,m\}} E|Y_k^j|^{2+\delta} \ .\end{aligned}$$

(Note that $E|Y_k^j|$ depends on j only through f_j.) Let $c_L^2 = u^T \Omega_L u$. Then, for $\epsilon > 0$,

$$\begin{aligned}&\left(\frac{1}{c_L^2}\right)\sum_{k=1}^{L} \int_{\{|u^t Y_k| > \epsilon c_L\}} |u^T Y_k|^2 dP \\ &\leq (\epsilon^\delta c_L^{2+\delta})^{-1} \sum_{k=1}^{L} \int |u^T Y_k|^{2+\delta} dP \\ &\leq \left(\sum_{j=1}^{m} |u_j|\right)^{2+\delta} / (\epsilon^\delta c_L^{2+\delta}) \sum_{k=1}^{L} \max_{j\in\{1,\ldots,m\}} E|Y_k^j|^{2+\delta} \\ &= \left(\sum_{j=1}^{m} |u_j|\right)^{2+\delta} / [\epsilon^\delta (c_L/a_L)^{2+\delta}] a_L^{-2-\delta} \\ &\quad \times \sum_{k=1}^{L} \max_{j\in\{1,\ldots,m\}} E|Y_k^j|^{2+\delta} \to 0 \ ,\end{aligned}$$

since $c_L/a_L \to u^T \Omega u > 0$.□

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle: Second International Symposium on Information Theory. Akadémiai Kiadó, Budapest

——— (1974) A new look at the statistical model identification. IEEE Trans Automatic Control 19:716–723

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62: 1198–1211

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535–543

Amos CI, Zhu DK, Boerwinkle E (1996) Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 60:143–160

Barrai I, Cavalli-Sforza LL, Mainardi M (1964) Testing a model of dominant inheritance for metric traits in man. Heredity 19:651–668

Beaty TH, Self SG, Liang KY, Connolly MA, Chase GA, Kwiterovich PO (1985) Use of robust variance components models to analyse triglyceride data in families. Ann Hum Genet 49:315–328

Bittles AH, Neel JV (1994) The costs of human inbreeding and their implications for variations at the DNA level. Nat Genet 8:117–121

Bleibtreu HK (1964) Marriage and residence patterns in a genetic isolate. PhD thesis, Harvard University, Cambridge, MA

Cockerham CC, Weir BS (1984) Covariances of relatives stemming from a population undergoing mixed self and random mating. Biometrics 40:157–164

Cotterman CW (1940) A calculus for statistico-genetics. PhD thesis, Ohio State University, Columbus

Corbeil RR, Searle SR (1976) A comparison of variance component estimators. Biometrics 32:779–791

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York

de Boer IJM, Hoeschele I (1993) Genetic evaluation methods for populations with dominance and inbreeding. Theor Appl Genet 86:245–258

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52:399–433

Gillois M (1964) La relation d'identité en génétique. Ann Inst Henri Poincaré B 2:1–94

Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. Am J Hum Genet 47:957–967

Harris DL (1964) Genotypic covariances between inbred relatives. Genetics 50:1319–1348

Hartley HO, Rao JNK (1967) Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika 54: 93–108

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–338

Hofer A (1998) Variance component estimation in animal breeding: a review. J Anim Breeding Genet 115:247–265

Hostetler JA (1974) Hutterite society. Johns Hopkins University Press, Baltimore

Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. Fifth Berkeley Symp 1:221–223

Jacquard A (1974) The genetic structure of populations. Springer-Verlag, New York

Karigl G (1981) A recursive algorithm for the calculation of identity coefficients. Ann Hum Genet 45:299–305

Kullback S (1959) Information theory and statistics. John Wiley, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Lange K (1978) Central limit theorems for pedigrees. J Math Biol 6:59–66

——— (1997) Mathematical and statistical methods for genetic analysis. Springer-Verlag, New York

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA

Mange AP (1964) Growth and inbreeding of a human isolate. Hum Biol 36:104–133

Martin AO (1970) The founder effect in a human isolate: evolutionary implications. Am J Phys Anthropol 32: 351–368

McCulloch RE (1988) Information and the likelihood function in exponential families. Am Stat 42:73–75

Morton NE, MacLean CJ (1974) Analysis of family resemblance. III. Complex segregation of quantitative traits. Am J Hum Genet 26:489–502

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:3038–313

Ober C, Hyslop T, Hauck WW (1999) Inbreeding effects on fertility in humans: evidence for reproductive compensation. Am J Hum Genet 64:225–231

Ober C, Weitkamp LR, Cox N, Dytch H, Kostyu D, Elias S (1997) HLA and mate choice in humans. Am J Hum Genet 61:497–504

O'Brien E (1987) The correlation between population structure and genetic structure in the Hutterites. In: Chepko-Sade BD, Halpin ZT (eds) Mammalian dispersal patterns: the effects of social structure on population genetics. University of Chicago Press, Chicago, pp 193–210

Orey S (1958) A central limit theorem for *m*-dependent random variables. Duke Math J 25:543–546

Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. Biometrika 58: 545–554

Schork NJ (1993) Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am J Hum Genet 53:1306–1319

Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Searle SR, Casella G, McCulloch CE (1992) Variance components. John Wiley & Sons, New York

Shaw FH, Woolliams JA (1999) Variance component analysis of skin and weight data for sheep subjected to rapid inbreeding. Genet Sel Evol 31:43–59

Shaw RG, Byers DL, Shaw FH (1998) Genetic components of variation in Nemophila menziesii undergoing inbreeding: morphology and flowering time. Genetics 150:1649–1661

Weinberg W (1909) Über Vererbungsgesetze beim Menschen. Z Induktive Abstammungs-Vererbungslehre 1:377–392