Copyright © 2008 by the Genetics Society of America
DOI: 10.1534/genetics.108.091058

Mixed Effects Models for Quantitative Trait Loci Mapping With Inbred Strains

Lara E. Bauman,*,†,[1] Janet S. Sinsheimer,†,‡,§ Eric M. Sobel§ and Kenneth Lange†,§,**

**Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Texas 78245-0549, †Department of Biomathematics, University of California, Los Angeles, California 90095-1766, ‡Department of Biostatistics, University of California, Los Angeles, California 90095-1772, §Department of Human Genetics, University of California, Los Angeles, California 90095-7088 and **Department of Statistics, University of California, Los Angeles, California 90095-1554*

Manuscript received May 14, 2008
Accepted for publication September 5, 2008

ABSTRACT

Fixed effects models have dominated the statistical analysis of genetic crosses between inbred strains. In spite of their popularity, the traditional models ignore polygenic background and must be tailored to each specific cross. We reexamine the role of random effect models in gene mapping with inbred strains. The biggest difficulty in implementing random effect models is the lack of a coherent way of calculating trait covariances between relatives. The standard model for outbred populations is based on premises of genetic equilibrium that simply do not apply to crosses between inbred strains since every animal in a strain is genetically identical and completely homozygous. We fill this theoretical gap by introducing novel combinatorial entities called strain coefficients. With an appropriate theory, it is possible to reformulate QTL mapping and QTL association analysis as an application of mixed models involving both fixed and random effects. After developing this theory, our first example compares the mixed effects model to a standard fixed effects model using simulated advanced intercross line (AIL) data. Our second example deals with hormone data. Here multivariate traits and parameter identifiability questions arise. Our final example involves random mating among eight strains and vividly demonstrates the versatility of our models.

IN analyzing gene mapping data from inbred strains, there is always the temptation to borrow models more pertinent to outbred populations. The vast majority of statisticians are wise enough to resist this temptation and turn to analysis methods tailored to specific breeding designs. Fortunately, the typical backcross or F_2 design has sufficient symmetry to permit analysis of variance by standard statistical packages. As mammalian geneticists explore more complicated designs involving multiple strains and multiple generations, this analysis paradigm has begun to fracture. It is therefore hardly surprising that the last decade and a half have seen a revival of interest in statistical models for gene mapping with inbred strains. Although we briefly review some of the important contributions to this literature in the next section, it is fair to say that most modern models rely heavily on fixed effects. In contrast, the most successful models for mapping quantitative trait loci (QTL) in outbred populations invoke random effects (Hopper and Mathews 1982; Goldgar 1990; Schork 1993; Amos 1994; Blangero and Almasy 1997).

The premise of this article is that, properly formulated, random effects models hold equal promise for more complicated inbred strain data. If a QTL is segregating between two strains, backcross and F_2 designs reliably detect it (Valdar *et al.* 2006). Models based on fixed allelic effects play a critical role in this process. Traditional designs have two drawbacks. First, the scarcity of recombination events often gives long mapped intervals. Second, when two founder strains of related ancestry are chosen, there may be no segregating QTL. To increase the number of recombination events and the number of segregating QTL, geneticists are turning to more complex designs involving multiple strains. Although the rationale for more complex designs is compelling, they bring in their wake problems of overparameterization. Random effects models neatly circumvent some of the parameterization issues encountered with fixed effects models. Unfortunately, the standard outbred QTL model does not make sense for inbred strains. All individuals of a particular strain are genetically identical and completely homozygous. These cardinal characteristics have subtle consequences when we calculate trait covariances for the descendants of matings between different strains. A logically correct theory for specifying covariances between pairs of individuals is the key to making random effects models respectable for inbred strains.

In this article, we take two approaches to QTL mapping; both capture polygenic background as a source of random variation. The two approaches differ in how they handle variation caused by the QTL. In

[1]*Corresponding author:* Department of Genetics, Southwest Foundation for Biomedical Research, P.O. Box 760549, San Antonio, TX 78245-0549. E-mail: lbauman@sfbrgenetics.org

association mapping, markers are treated one by one as candidate genes, and observed genotypes or allele counts at a marker serve as fixed predictors of trait means. In linkage mapping, markers in the vicinity of the QTL provide prior information on gene sharing, and the QTL contribution is modeled as a random effect. The greatest defect of our models is the blanket assumption of additivity. The greatest strength of our models is their generality in other regards. Thus, there is no limit to the number of founding strains, the depth and complexity of pedigrees, or the number of traits in a multivariate analysis.

To avoid breaking the flow of our discussion, much of the mathematical detail is relegated to the APPENDIXES. The following sections summarize previous contributions, lay out the model with full attention to computation of strain coefficients and relative covariances, resolve the thorny issue of identifiability, apply the models to real and simulated data, and discuss the broader implications and limitations of the models.

METHODS

A brief survey of previous methods: Inbred mammalian strains have unique advantages in genetics. All members of a strain are genetically identical and completely homozygous. Simple crosses between strains involve no phase ambiguities, and any genes mapped can be quickly located in humans and other species by synteny. With mice and other small mammals, breeding is reasonably straightforward, generation times are fairly short, and the environment can be exquisitely controlled.

For decades, QTL mapping in inbred strains was considered an exercise in fixed effects modeling. Testing for association between marker genotypes and trait values is readily carried out using several available statistical packages. In the interval method introduced by LANDER and BOTSTEIN (1989), the QTL is allowed to take any position along a chromosome. This makes QTL genotypes unobservable and requires computation of posterior distributions given observed genotypes at the flanking markers. Although the EM algorithm is applicable in this context, it is often slow to converge, and the regression method of HALEY and KNOTT (1992) provides a quick approximation. The permutation test of CHURCHILL and DOERGE (1994) handles multiple testing problems gracefully. The recent program R/qtl (BROMAN *et al.* 2003), which capitalizes on the R software environment, combines several of these methods with hidden Markov modeling of missing genotypes. Despite these admirable advances, interval mapping is still limited to simple crosses where polygenic background is confounded with random environment. As the field embraces more complex crosses, geneticists no longer have the luxury of ignoring polygenic background, and it seems self-evident that explicitly modeling it will improve statistical inference.

The composite interval mapping method of ZENG (1993, 1994) implemented in QTL Cartographer generalizes interval mapping by including the direct effects of one or more markers unlinked to the QTL. Hence, composite interval mapping can be viewed as an attempt to incorporate polygenic background through fixed effects. If the number of typed markers is large, then it becomes hopeless to include all of them, and some automatic selection of background markers is desirable (MANLY and OLSON 1999).

Although XIE *et al.* (1989) take important first steps toward including polygenic background as a random effect, they do not derive general covariance expressions. This failure makes it difficult to deal with nonstandard crosses and awkward to combine data from different crosses. In the meantime, the pressure to increase the number of strains per cross has been growing (REBAI and GOFFINET 1993). Of 21 cloned mouse genes listed in Tables 1 and 2 of the review by FLINT *et al.* (2005), 7 rely on cloning strategies involving multiple strains or outbred mice. These practical concerns are stimulating intense efforts to revamp experimental design and statistical analysis of inbred cross data (LIU and ZENG 2000; HITZEMANN *et al.* 2002; PLETCHER *et al.* 2004; LI *et al.* 2005; CERVINO *et al.* 2007). Other recent models that delve into multiple QTL models and epistasis are both frequentist (KAO *et al.* 1999; JANNINKA and JANSENA 2001; SEATON *et al.* 2002; BROMAN *et al.* 2003) and Bayesian oriented (SILLANPÄÄ and ARJAS 1998; SEN and CHURCHILL 2001; BROMAN *et al.* 2003).

Trait means, variances, and covariances: We begin our theory development with a basic model applicable to any inbred strain design, including F_2, advanced intercross lines, and random mating. Suppose that i and j are two animals generated by a complex cross involving s inbred strains. At t traits of interest, i and j exhibit random vectors X_i and X_j of trait values. For the sake of simplicity, assume further that X_i and X_j reflect the contributions of a single gene whose alleles have additive effects. Our immediate goal is to calculate the expected vectors $E(X_i)$ and $E(X_j)$ and the covariance matrix $\mathrm{Cov}(X_i, X_j)$. When $i = j$, we recover variances as well as covariances. Because of our assumption of additivity, X_i decomposes as the sum $Y_i + Z_i$ of a maternal contribution Y_i plus a paternal contribution Z_i. To calculate $E(Y_i)$, let M_i denote the originating strain of the maternal gene of i. Although M_i is unobserved, we can calculate the probability $\Pr(M_i = a)$ for any given strain a. In terms of these probabilities and the $t \times 1$ mean vector $\mu(a)$ of allelic effects on each trait for strain a, we have

$$E(Y_i) = \sum_{a=1}^{s} \Pr(M_i = a)\mu(a).$$

Invoking a similar expression for $E(Z_i)$, it follows that

$$E(X_i) = 2\sum_{a=1}^{s} \gamma_i(a)\mu(a), \tag{1}$$

where $\gamma_i(a)$ is the probability that a randomly sampled gene from i originates from strain a. We refer to γ_i as the strain fraction vector for animal i; γ_i has dimension $s \times 1$.

Covariances are derived by the same kind of reasoning. Decompose X_j into the sum $V_j + W_j$ of a maternal contribution V_j plus a paternal contribution W_j. In view of the bilinearity of the covariance operator and the symmetry of maternal and paternal alleles, it suffices to find the covariance $\mathrm{Cov}(Y_i, V_j)$. Let N_j denote the originating strain of the maternal gene of j. Conditioning on the joint value of M_i and N_j then yields

$$\begin{aligned}\mathrm{Cov}(Y_i, V_j) &= E(Y_i V_j^*) - E(Y_i)E(V_j)^* \\ &= \sum_a \sum_b \Pr(M_i = a, N_j = b)\mu(a)\mu(b)^* \\ &\quad - \left[\sum_a \Pr(M_i = a)\mu(a)\right] \\ &\quad \times \left[\sum_b \Pr(N_j = b)\mu(b)\right]^*,\end{aligned}$$

where the superscript * indicates a vector or matrix transpose. By analogy with kinship coefficients, we define the strain coefficient $\psi_{ij}(a, b)$ to be the joint probability that a randomly drawn gene from animal i originates from strain a and a randomly drawn gene from the same locus of animal j originates from strain b. If i and j coincide, then sampling is done with replacement. The $t \times t$ covariance matrix between the trait values of i and j becomes

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{Cov}(Y_i, V_j) + \mathrm{Cov}(Y_i, W_j) + \mathrm{Cov}(Z_i, V_j) \\ &\quad + \mathrm{Cov}(Z_i, W_j) \\ &= 4\sum_a \sum_b \psi_{ij}(a, b)\mu(a)\mu(b)^* \\ &\quad - 4\sum_a \gamma_i(a)\mu(a) \sum_b \gamma_j(b)\mu(b)^* \\ &= 4\sum_a \sum_b C_{ij}(a, b)\mu(a)\mu(b)^*, \end{aligned} \tag{2}$$

where $C_{ij}(a, b) = \psi_{ij}(a, b) - \gamma_i(a)\gamma_j(b)$, which we collect into an $s \times s$ matrix, denoted C_{ij}.

For s strains and t traits, it is convenient to stack the allelic effects into a column vector μ of length st with transpose

$$\mu^* = [\mu_1(1), \ldots, \mu_1(s), \ldots, \mu_t(1), \ldots, \mu_t(s)].$$

The positive semidefinite matrix $\Omega = \mu\mu^*$ can then be split into t^2 blocks Ω_{kl} each of size $s \times s$. Restricting our attention to the block corresponding to traits k and l, the covariance matrix (2) has entries given by the trace formula

$$\mathrm{Cov}(X_{ik}, X_{jl}) = 4\,\mathrm{tr}(C_{ij}\Omega_{kl}). \tag{3}$$

In polygenic inheritance, many independent loci contribute in an additive manner to the traits under consideration. Since trait means and covariances add in this setting, the mean expression (1) and the covariance expressions (2) and (3) remain valid provided we replace μ by $\sum_l \mu^l$ and Ω by $\sum_l \mu^l(\mu^l)^*$. Here μ^l denotes the vector contribution corresponding to locus l rather than the lth component of μ. APPENDIX A shows that every pair (μ, Ω) consisting of a vector μ and a positive semidefinite matrix Ω can be represented as two such coordinated linear combinations. Hence, to capture polygenic background, it suffices to estimate arbitrary μ and Ω. We see later that there is an identifiability issue that must be surmounted in estimating Ω.

Computation of strain coefficients: Because the combinatorial coefficients $\gamma_i(a)$ and $\psi_{ij}(a, b)$ are essential in calculating trait means and variances, we need good algorithms to compute these coefficients. Fortunately, we can mimic the logic used in calculating kinship coefficients for outbred populations. Since a pedigree founder i is assumed to be strain pure, one entry of the vector $\gamma_i = 1$, and the remaining entries $= 0$. Likewise for two founders i and j, one entry of the matrix $\psi_{ij} = 1$, and the remaining entries $= 0$. All other strain fraction vectors γ_i and strain coefficient matrices ψ_{ij} are defined recursively starting with the founders.

To avoid circular reasoning, pedigree members are numbered so that parents always precede their children. If animal i is not a founder, then it has parents k and l. Assuming that k and l have already been visited in filling in the strain fractions, we set

$$\gamma_i = \frac{1}{2}(\gamma_k + \gamma_l). \tag{4}$$

If $j \neq i$, then without loss of generality we can assume j has been visited already, and we can set

$$\psi_{ij} = \frac{1}{2}(\psi_{kj} + \psi_{lj}) \tag{5}$$

$$\psi_{ji} = \frac{1}{2}(\psi_{jk} + \psi_{jl}). \tag{6}$$

This leaves only the case $j = i$. There are four equally likely possibilities when we sample two genes of i: (a) both genes coincide with the gene passed by k, (b) both genes coincide with the gene passed by l, (c) the first gene comes from k and the second from l, and (d) the first gene comes from l and the second from k. These considerations produce the matrix recurrence

$$\psi_{ii} = \frac{1}{4}[\mathrm{diag}(\gamma_k) + \mathrm{diag}(\gamma_l) + \gamma_k\gamma_l^* + \gamma_l\gamma_k^*], \tag{7}$$

where $\mathrm{diag}(\gamma)$ denotes a diagonal matrix whose diagonal entries coincide with the entries of the vector γ.

The initial conditions on founders and the recurrences (4)–(7) completely determine γ_i and ψ_{ij}. These in turn determine the C_{ij} matrices, which have a richer mathematical structure than the strain coefficient matrices ψ_{ij}. APPENDIX B describes several fascinating properties of the C_{ij} matrices. One such property is $C_{ij} = \mathbf{0}$ between most members of simple crosses, for example, for all F_2 animals when $i \neq j$ or whenever i is a founder or F_1.

Variance component models for QTL mapping with outbred populations require conditional kinship coefficients in addition to theoretical kinship coefficients. For exactly the same reasons, we also need conditional strain fractions and coefficient matrices. These depend on observed marker genotypes in the vicinity of a putative QTL. On small pedigrees, it is possible to compute conditional strain coefficient matrices exactly by considering all descent graphs (gene flow patterns) at the QTL and neighboring markers (KRUGLYAK *et al.* 1996). In practice, inbred strain pedigrees are so large that the number of possible descent graphs is astronomical. Stochastic sampling provides a workable substitute for exhaustive enumeration of descent graphs (SOBEL and LANGE 1996). The Markov chain Monte Carlo (MCMC) method incorporated in the computer program SimWalk samples relevant descent graphs with the appropriate conditional probabilities. Given a descent graph at the QTL, it is trivial to compute strain fractions for all animals and strain coefficient matrices for all pairs of animals in a pedigree. The averages of these quantities over all sampled descent graphs serve as approximations to the conditional strain fractions and strain coefficient matrices.

Strain coefficients convey more information than strain fractions. For instance, it is obvious that

$$\gamma_i(a) = \sum_b \psi_{ii}(a, b).$$

We can put this extra information to good use in predicting QTL genotypes. At a given genomic location, imagine a marker with a different allele for each strain. Let $\hat{\beta}_i(a/b)$ be the conditional probability that animal i has unordered genotype a/b at the hypothetical marker given the observed data at the ordinary markers. The relations

$$\hat{\psi}_{ii}(a, a) = \frac{1}{2}\hat{\gamma}_i(a) + \frac{1}{2}\hat{\beta}_i(a/a),$$
$$\hat{\psi}_{ii}(a, b) = \frac{1}{4}\hat{\beta}_i(a/b), b \neq a$$

connect the conditional genotype probabilities to the conditional strain fractions and coefficients. These relations in turn imply that

$$\hat{\beta}_i(a/a) = 2\hat{\psi}_{ii}(a, a) - \hat{\gamma}_i(a),$$
$$\hat{\beta}_i(a/b) = 4\hat{\psi}_{ii}(a, b), b \neq a. \quad (8)$$

Thus, we can impute strain genotypes as well as strain fractions.

Variance component models: Variance component models revolve around the multivariate normal distribution or related distributions such as the multivariate *t*. Every multivariate normal distribution is uniquely determined by its mean vector ν and variance matrix Σ. If we decompose trait values into independent, additive contributions, then ν and Σ can be expressed as sums over the various contributions. As long as we are willing to take the leap of faith that all random contributions are Gaussian, then trait vectors will be Gaussian as well. For each random contribution, variance matrices are constructed from a constant part and a parametric part. The genetic covariance formula (3) is typical in this regard. The constant parts C_{ij} are forced on us by the nature of the pedigree. The parametric part Ω with blocks Ω_{kl} requires estimation.

The environmental contribution to the mean is usually modeled as the sum of a grand mean η plus covariate effects such as age or sex. Random environment and cage effects can be modeled by Kronecker products of variance matrices, provided we order trait values so that all values corresponding to a given trait are contained in a single block, and animals are consistently enumerated across blocks. Given these conventions, the variance matrix under random environment reduces to the Kronecker product $\Upsilon \otimes I$ of the trait variance matrix Υ and the identity matrix I. Obviously, Υ is the parametric part; it describes the environmental covariation of the traits in a single animal. The matrix I reflects the independence of the random environments for the various animals. For a random cage effect, we replace the identity matrix by a cage matrix $H = (h_{ij})$, where $h_{ij} = 1$ if animals i and j belong to the same cage and 0 otherwise. The matrix replacing Υ describes the environmental covariation of the traits for animals in a single cage (LANGE 2002). As an example, heritability analyses generally specify two random effects, additive polygenes and random error/environment,

$$E(X_{ik}) = 2\sum_{a=1}^{s} \gamma_i(a)\mu_k(a) + \sum_{c=l}^{C} \alpha_{ic}\beta_{kc} + \eta \quad (9)$$

$$\mathrm{Cov}(X_{ik}, X_{jl}) = 4\,\mathrm{tr}(C_{ij}\Omega_{kl}) + \Upsilon_{kl}, \quad (10)$$

where α_{ic} is the cth of C covariates measured on animal i and β_{kc} is the corresponding regression coefficient for trait k.

Once we specify the mean and variance components, the loglikelihood of a pedigree can be written as

$$\mathcal{L} = -\frac{1}{2}\ln\det\Sigma - \frac{1}{2}(x - \nu)^{*}\Sigma^{-1}(x - \nu),$$

using the observed trait values x, the mean vector ν such as that of Equation 9, and the variance matrix Σ such as that of Equation 10. Assuming pedigrees behave

independently, their loglikelihoods add. Given the overall loglikelihood, parameters can be estimated by maximum likelihood, and statistical inference conducted by standard likelihood ratio tests comparing alternative hypotheses to null hypotheses. Lange (2002) develops this frequentist approach to estimation and inference in detail. Our computer program Mendel relies on a quasi-Newton algorithm for maximum likelihood estimation. Bauman *et al.* (2005) discusses an alternative EM algorithm as well as factor-analytic parameterizations of variance matrices. Given the presence of covariates and heterogenous pedigree structures, permutation testing is rarely possible. To aid the user in judging significance and model fitting, Mendel reports standard errors of parameters, pedigree deviances, outlier individuals, and various goodness-of-fit statistics.

Two QTL mapping strategies: There are two specific strategies, association and linkage, for QTL mapping. Variance component models are pertinent to both. Although the two strategies differ in how they portray QTL effects, each captures polygenic background as a random effect. In addition to the strain effects appearing in Equation 1, most models include a grand mean η and fixed effects tied to plausible predictors. If we specify η, then we must impose the vector constraint $\sum_a \mu(a) = \mathbf{0}$ on the polygenic mean vector μ. Here the index a ranges over all strains. Random effects include the polygenic effect summarized by Equation 3, random environment plus measurement error, and possibly correlated environment such as cage effects. As described in the next section, the polygenic variance matrix Ω is not identifiable, and complicated constraints must be imposed on it to compensate for this fact. Regardless of the nature of these constraints, we must compute theoretical strain fractions and strain coefficients to estimate μ and Ω under the null hypothesis of no QTL effect.

In linkage mapping, markers serve to tag chromosome segments and keep track of recombination events. The genotypes of the causative QTL are unobserved, and the QTL is allowed to assume any position along the genome. Under the alternative hypothesis in linkage mapping, we model the QTL as a random effect in the same way that we modeled the contribution of a single gene with additive effects. The only difference is that we use strain fractions and coefficients calculated conditional on the observed marker data. From here on, we refer to these as conditional strain fractions and coefficients; those calculated unconditionally we call theoretical strain fractions and coefficients. Motivated by Equations 1 and 3, we let $\varepsilon(a)$ denote the additive effect of the QTL in strain a. Then our earlier reasoning shows that the QTL contribution has mean

$$2\sum_a \hat{\gamma}_i(a)\,\varepsilon(a)$$

for animal i and covariance

$$4\sum_a\sum_b \widehat{C_{ij}}(a,\,b)\varepsilon(a)\varepsilon(b)^*$$

for animals i and j. Here the circumflexes indicate conditional versions of the strain fractions and coefficients estimated from the marker data. Under the alternative hypothesis, we estimate the entries of ε.

Our basic linkage model therefore specifies the trait means and covariances

$$E(X_{ik}) = 2\sum_{a=1}^{s}\gamma_i(a)\mu_k(a) + 2\sum_{a=1}^{s}\hat{\gamma}_i(a)\varepsilon_k(a) + \sum_c \alpha_{ic}\beta_{kc} + \eta \tag{11}$$

$$\mathrm{Cov}(X_{ik}, X_{jl}) = 4\,\mathrm{tr}(C_{ij}\Omega_{kl}) + \mathrm{tr}(\hat{C}_{ij}\varepsilon_k\varepsilon_l^*) + 1_{\{i=j\}}\Upsilon_{kl} \tag{12}$$

for two animals i and j. Here k and l index two traits, α_{ic} is covariate c of animal i, and β_{kc} is the corresponding regression coefficient for trait k. If we let $\bar{\varepsilon}$ denote the average $(1/s)\sum_a \varepsilon(a)$, then all QTL models that include a grand mean require the constraint $\bar{\varepsilon} = \mathbf{0}$. In the presence of this constraint, the likelihood ratio test of linkage follows asymptotically a χ^2 distribution with $st - t$ degrees of freedom.

In association mapping, QTL fixed effects are tied to the current marker. The marker is viewed as a candidate gene whose genotypes or alleles directly influence trait means (Lange *et al.* 2005); random QTL effects are omitted. Hence, in Equation 12 we drop the random effect $\mathrm{tr}(\widehat{C_{ij}}\varepsilon_k\varepsilon_l^*)$, and in Equation 11 we amend the fixed effect $2\sum_{a=1}^{s}\hat{\gamma}_i(a)\varepsilon_k(a)$ to represent regression on observed allele counts at the current marker. If the additive model for allelic effects is viewed as too restrictive, then we can regress on observed genotypes. Association testing is again conducted by likelihood ratio statistics.

In the presence of missing genotypes in association testing, we fall back on imputed allele counts or imputed genotype counts. Because genotypes at markers are usually directly observed, little is lost in imputation by ignoring genotypes at flanking markers. In this simpler setting, a fast deterministic algorithm is available for imputation (Lange *et al.* 2005). Flanking marker genotypes occasionally resolve phase ambiguities caused by combining closely spaced single nucleotide polymorphisms (SNPs) into supermarkers. Accordingly, the current version of Mendel also accepts MCMC estimates of conditional strain fractions from SimWalk. When each strain carries a different allele at the marker, the allele counts delivered by SimWalk are computed by doubling the conditional strain fractions at the marker. When two strains share a common allele at the marker, the corresponding strain fractions are added before doubling.

Identifiability: We have seen that the polygenic covariance expression (3) between trait k of animal i and

trait l of animal j involves the $s \times s$ trait block Ω_{kl} of an $st \times st$ variance matrix Ω. Unfortunately, estimation of Ω collides with an identifiability issue. The crux of the problem is the existence of nontrivial matrices Λ with

$$\operatorname{tr}[C_{ij}(\Omega_{kl} + \Lambda_{kl})] = \operatorname{tr}(C_{ij}\Omega_{kl})$$

for every legitimate choice of C_{ij} and every trait pair (k, l). Proposition 2 of APPENDIX B explains this phenomenon by representing C_{ij} as a convex combination of the matrix $\mathbf{0}$ and $\binom{s}{2}$ matrices E_{mn} indexed by unordered strain pairs $\{m, n\}$. Here all entries of E_{mn} are 0 except for the diagonal entries $e_{mm} = e_{nn} = 1$ and the off-diagonal entries $e_{mn} = e_{nm} = -1$. It follows that

$$\operatorname{tr}(C_{ij}\Lambda_{kl}) = \sum_{\{m,n\}} a_{ij,mn}\operatorname{tr}(E_{mn}\Lambda_{kl}) = 0$$

provided

$$\operatorname{tr}(E_{mn}\Lambda_{kl}) = \lambda_{kl,mm} + \lambda_{kl,nn} - \lambda_{kl,mn} - \lambda_{kl,nm} = 0 \quad (13)$$

for every strain pair $\{m, n\}$ and every $s \times s$ trait block $\Lambda_{kl} = (\lambda_{kl,mn})$ of Λ.

We can solve the identifiability problem by subtracting the nonidentifiable part of Ω from Ω. To achieve this end, we view the positive semidefinite matrix Ω as a vector in the Euclidean space $\mathbb{R}^{st \times st}$. In this setting the trace function $\langle A,B\rangle = \operatorname{tr}(AB^*)$ and Frobenius norm $\|A\|_F = \operatorname{tr}(AA^*)^{1/2}$ reduce to the standard inner product and Euclidean norm. To find the nonidentifiable part of Ω, one projects Ω onto the vector subspace $\mathcal{S}$ of symmetric matrices satisfying Equation 13 for every strain pair $\{m, n\}$ and every trait block Ω_{kl}. Formally, the projection $P(\Omega)$ is defined to be the matrix X giving the minimum of $\|\Omega - X\|_F^2$ for $X \in \mathcal{S}$.

Fortunately, minimization of $\|\Omega - X\|_F^2$ separates into subproblems corresponding to different trait blocks. First, consider a diagonal block Ω_{kk} of Ω. To simplify notation, denote its entries by $y_{mn} = \Omega_{kk,mn}$ and the entries of the corresponding block of the projection by $x_{mn} = P(\Omega)_{kk,mn}$. To find $P(\Omega)_{kk}$ we must minimize the sum of squares

$$\frac{1}{2}\sum_m\sum_n (y_{mn} - x_{mn})^2$$

subject to the constraints $x_{mm} + x_{nn} = x_{mn} + x_{nm}$ for every pair $\{m, n\}$. Now consider off-diagonal blocks $\Omega_{kl} = \Omega_{lk}^*$. These come in pairs that must be handled together, so we let

$$y_{mn} = \Omega_{kl,mn} = \Omega_{lk,nm}$$

and

$$x_{mn} = P(\Omega)_{kl,mn} = P(\Omega)_{lk,nm}$$

and minimize the sum of squares

$$\frac{1}{2}\sum_m\sum_n (y_{mn} - x_{mn})^2 + \frac{1}{2}\sum_m\sum_n (y_{nm} - x_{nm})^2 = \sum_m\sum_n (y_{mn} - x_{mn})^2$$

subject to the constraints $x_{mm} + x_{nn} = x_{mn} + x_{nm}$ for every pair $\{m, n\}$. It follows that diagonal blocks and off-diagonal blocks lead to the same constrained minimization problem.

APPENDIX C shows that each of these least-squares problems has solution $X = (x_{mn})$ with residual

$$Y - X = \frac{1}{2}(U + U^*),$$

where $Y = (y_{mn})$, $U = QYQ$, and Q is the $s \times s$ projection matrix $I - (1/s)\mathbf{1}\mathbf{1}^*$. In calculating a covariance, we can ignore symmetrization and replace the matrix $\frac{1}{2}(U + U^*)$ by U. Indeed, the symmetry of C_{ij} implies that

$$\frac{1}{2}\operatorname{tr}[C_{ij}(U + U^*)] = \frac{1}{2}\operatorname{tr}(C_{ij}U) + \frac{1}{2}\operatorname{tr}(C_{ij}U^*) = \operatorname{tr}(C_{ij}U).$$

Thus, $\operatorname{tr}(C_{ij}Q\Omega_{kl}Q)$ faithfully represents the covariance between trait k of animal i and trait l of animal j. By the same reasoning, we can replace the entire residual matrix $\Omega - P(\Omega)$ by the matrix

$$R = \operatorname{diag}(Q)\Omega\operatorname{diag}(Q). \quad (14)$$

Here $\operatorname{diag}(Q)$ is a diagonal block matrix with all t diagonal blocks equal to Q. One can easily check that $\operatorname{diag}(Q)$ is a projection matrix and that R inherits the properties of symmetry and positive semidefiniteness from Ω.

In reparameterizing Ω, it is convenient to define an orthogonal matrix O mapping the vector $(1/\sqrt{s})\mathbf{1}$ to the standard basis vector e_1. (See APPENDIX D for one version of O.) It follows that

$$OQO^* = O(I - \frac{1}{s}\mathbf{1}\mathbf{1}^*)O^* = I - e_1e_1^*.$$

Observe that pre- and postmultiplying any square matrix by $I - e_1e_1^*$ zeros out the first row and first column of the matrix. To take advantage of this fact, we express the residual matrix (14) as

$$R = \operatorname{diag}(O^*)\operatorname{diag}(OQO^*)\operatorname{diag}(O)\Omega\operatorname{diag}(O^*)\operatorname{diag}(OQO^*)\operatorname{diag}(O) = \operatorname{diag}(O^*)\Upsilon\operatorname{diag}(O). \quad (15)$$

The matrix

$$\Upsilon = \operatorname{diag}(OQO^*)\operatorname{diag}(O)\Omega\operatorname{diag}(O^*)\operatorname{diag}(OQO^*)$$

is a positive semidefinite replacement for $\operatorname{diag}(O)\Omega\operatorname{diag}(O^*)$. By our earlier remark, a block Υ_{kl} of Υ equals the corresponding block of $\operatorname{diag}(O)\Omega\operatorname{diag}(O^*)$ with its first row and column zeroed out.

We are now close to the desired goal of reparameterizing the residual. The matrix Υ has entire rows and

columns consisting of zeros. Permuting its rows and columns appropriately will move its nontrivial part to an upper-left block, which will be positive definite whenever Ω is positive definite. The Cholesky decomposition of this upper-left block then serves as a good parameterization of R. To compute the number of parameters for s strains and t traits, observe that the matrix Υ is $st \times st$. A total of t rows and columns are lost in the zeroing-out process. This leaves an $(st - t) \times (st - t)$ upper-left block with $(st - t)(st - t + 1)/2$ diagonal or subdiagonal entries. For example, with three strains and two traits, there are 10 parameters.

For the sake of clarity, let us summarize how our proposed parameterization leads to trait covariances. It begins with a Cholesky decomposition Δ of an $(st - t) \times (st - t)$ positive definite matrix. The matrix $\Delta\Delta^*$ is then subdivided into $(s - 1) \times (s - 1)$ trait blocks $(\Delta\Delta^*)_{kl}$, and each block is promoted to an $s \times s$ trait block Υ_{kl} by adding a top row and left column of zeros. In matrix notation, $\Upsilon_{kl} = Z(\Delta\Delta^*)_{kl}Z^*$ with Z the $s \times (s - 1)$ matrix

$$Z = \begin{pmatrix} \mathbf{0}^* \\ I \end{pmatrix}.$$

Finally, we construct the residual matrix R via Equation 15, using the orthogonal matrix O.

With these conventions, the covariance between trait k of animal i and trait l of animal j amounts to

$$\begin{aligned} \operatorname{Cov}(X_{ik}, X_{jl}) &= 4 \operatorname{tr}(C_{ij}R_{kl}) \\ &= 4 \operatorname{tr}(C_{ij}O^*\Upsilon_{kl}O) \\ &= 4 \operatorname{tr}[C_{ij}O^*Z(\Delta\Delta^*)_{kl}Z^*O] \\ &= 4 \operatorname{tr}[Z^*OC_{ij}O^*Z(\Delta\Delta^*)_{kl}]. \end{aligned} \tag{16}$$

In computing covariances over large pedigrees, it saves time and storage to precompute and store the $(s - 1) \times (s - 1)$ matrices $4Z^*OC_{ij}O^*Z$ and discard the $s \times s$ matrices C_{ij}. Note that the action $A \mapsto Z^*AZ$ on an $s \times s$ matrix A deletes the first row and first column of A.

This ends our theoretical overview of the model. APPENDIX E shows how to differentiate covariances with respect to parameters, and APPENDIX F supplies a counterexample connecting identifiability and symmetry. We now move on to data analysis.

APPLICATIONS

A simulated advanced intercross line: An AIL starts with F_1 offspring from an intercross of two inbred strains. The F_1 animals are randomly bred to produce the F_2 animals, the F_2 animals are randomly bred to produce the F_3 animals, and so on for a total of n generations. An AIL differs from repeated brother–sister mating, because it involves enough animals to preserve genetic diversity. It draws its strength from the steady accumulation of recombination events over many generations (DARVASI and SOLLER 1995). Simulating data according to an AIL design permits us to compare our mixed effects results with the fixed effects results of the benchmark program QTL Cartographer. This exercise is not meant to be a substitute for an exhaustive study of power and experimental design. Also, the comparison is not entirely fair because QTL Cartographer analyzes the F_n data at the last generation ignoring the previous generations. To reconstruct missing marker information, QTL Cartographer applies an inflated recombination fraction scaled to reflect n.

To create our simulated AIL data, we mated two inbred founder animals and subjected their descendants in each generation to virtual random mating. Generation 10 contained 175 animals in 140 sibships with 492 animals overall. Placing the QTL locus at the midpoint of markers 5 and 6 of 11 equally spaced marker loci, we simulated genotypes by gene dropping and assigned QTL effects on the basis of the genotypes at the QTL. QTL genotypes were then discarded from further analysis. We modeled a univariate trait with a grand mean $\eta = 4$, an environmental variance $\sigma^2_{\text{env}} = 1$, and a 2×2 polygenic variance matrix

$$\Omega = \begin{pmatrix} 1.0 & 0.20 \\ 0.20 & 0.29 \end{pmatrix}.$$

For this simulated trait, strain one has a genetic variance comparable to the environmental variance and larger than the genetic variance of strain two. The two strains share a modest genetic correlation. For reasons explained in the next section, a single generation of data in a symmetric cross of this sort does not sustain estimation of strain-specific polygenic means. To circumvent this problem in our comparisons, we set the strain-specific polygenic means equal to 0. We chose small strain-specific QTL effects $\varepsilon_1 = 0.2$ and $\varepsilon_2 = -0.2$ centered around 0. In view of our discussion of identifiability, we can estimate only a single parameter p_1 characterizing Ω. The projection technique discussed yields the value $p_1 = 0.667$. The discussion of the C_{ij} matrices in APPENDIX B explains why genotype data on a single generation also prevent estimation of p_1.

To provide the most informative comparisons, we ran three analyses: (1) Mendel on the full pedigree with complete genotype and phenotype data (Mendel Full), (2) Mendel on the full pedigree but with phenotype data on only the final F_{10} generation (Mendel F_{10}), and (3) QTL Cartographer on the final F_{10} generation with complete genotype and phenotype data (Cartographer). Simply comparing cases Mendel Full and Cartographer is hardly fair; the full pedigree contains more than twice the number of animals in the final generation. Mendel F_{10} takes advantage of the full genealogy and all genotype data in computing theoretical and conditional strain coefficients. It limits itself to the phenotype sample in the last generation to enable a better comparison to QTL Cartographer.

Before turning to QTL mapping in the Mendel analyses, we fit a baseline model including the grand mean, the polygenic variance, and the environmental variance. We then estimated conditional strain coefficients at each of the 11 marker loci. This put us in a position to estimate the global parameters and the QTL-specific parameters simultaneously at each locus. The evidence in favor of the QTL is summarized by a likelihood ratio test (LRT) statistic following a χ^2_1 distribution; a nonlinear false discovery rate (FDR) correction (Benjamini *et al.* 2001) corrects for multiple testing for all three analyses. Table 1 summarizes the type I error rate, power, and coverage as well as the generating parameters, their estimates, and the standard errors of the estimates at the loci adjacent to the QTL. Successful coverage occurs when the equivalent one-LOD drop interval (4.6 LRT units) includes the QTL. We reject the null hypothesis of no QTL effect when the LRT is significant at the 0.05 level.

The results in Table 1 reflect 100 simulations for a QTL-effect size that yields power >90% for Mendel Full; type I error rates are given as confidence intervals based on 500 simulations under the null hypothesis of no QTL effect. Clearly, the power to detect linkage is drastically reduced when only the F_{10} generation is available for analysis. This absence of data also makes it difficult for Mendel F_{10} to estimate the polygenic parameter p_1 accurately. For Mendel Full all estimates are within one standard error of their true values, and standard errors are small. QTL Cartographer exhibits slightly better power and coverage than Mendel F_{10}, but with a largely inflated type I error rate. Both methods are easily bested by Mendel Full. These trends continue over a range of smaller QTL effects (data not shown). We are pleased with these results. In our view they demonstrate that application of the mixed effects model sacrifices little in simple settings while generalizing readily to complex pedigrees.

A multivariate four-way cross: To illustrate the analysis of multivariate traits, we next consider the hormone data of Burke and colleagues (Harper *et al.* 2003) on aging UM-HET3 mice. Figure 1 shows how the UM-HET3 mice were created from four founder strains: BALB/cJ (C), C57BL/6J (B6), C3H/HeJ (C3), and DBA/2J (D2). $CB6F_1$ females crossed with $C3D2F_1$ males provided 967 F_2 full siblings. At markers with four different alleles, all F_2 mice were heterozygous. Thus compared to a two-way cross, the four-way cross doubles the number of founder strains without sacrificing phase certainty. Hormone levels of insulin-like growth factor I (IGF), leptin (Lep), and thyroxine (T4) were measured at 4 and 15 months on each of the F_2 mice. Testing maternal and paternal effects separately, Harper *et al.* found several linked markers in these data via ANOVA, including a maternal allele at D3Mit25 linked to IGF at 15 months, a paternal allele at D3Mit127 linked to Lep at 4 months, and both maternal and paternal alleles linked to Lep at 15 months. It is worth pointing out that ANOVA or MANOVA must be carried out at marker loci. Only here do marker genotypes or allele counts unambiguously define factor levels. With complete genotyping, our model collapses in this setting to the classical models.

This multistrain cross highlights identifiability pitfalls inherent in the structure of some crosses and the data collected on them. For example, all F_2 mice share the strain fraction vector $\frac{1}{4}\mathbf{1}$. Hence, the polygenic mean is confounded with the grand mean. Using strain trait averages or phenotyping members of the original strains would allow us to estimate the polygenic means, but this is not an option for the current data.

Although the rigid structure of the four-way cross preserves phase certainty, it reduces uncertainty to the point where the polygenic covariance matrix cannot be estimated. Polygenic covariances depend on the combinatorial matrices C_{ij}. We have already noted that $C_{ij} = \mathbf{0}$ whenever i is a founder or i and j are F_1 mice. Straightforward calculations for F_2 mice i and j with $i \neq j$ yield

$$C_{ij} = \mathbf{0}, \quad C_{ii} = \frac{1}{16}\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

Inspection of Equation 3 therefore shows that the polygenic covariance matrix Ω is confounded with the matrix describing the environmental covariances.

Finally, there are identifiability problems with the QTL allelic effects. At the covariance level, the conditional coefficient matrix $\widehat{C_{ij}}$ is identically 0 when typing is full and different alleles are present in each strain. At the mean level, imposition of the constraint $\varepsilon_4 = -\varepsilon_1 - \varepsilon_2 - \varepsilon_3$ shows that the genotype-specific means in a purely allelic model can be expressed as the vector

$$\begin{pmatrix} \varepsilon_1 + \varepsilon_3 + \eta \\ \varepsilon_1 + \varepsilon_4 + \eta \\ \varepsilon_2 + \varepsilon_3 + \eta \\ \varepsilon_2 + \varepsilon_4 + \eta \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & -1 & -1 & 1 \\ 0 & 1 & 1 & 1 \\ -1 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \eta \end{pmatrix}.$$

Because the matrix on the right of this equation has less than full rank, some mean vectors are not representable. As a substitute for the additive QTL contributions, we assign a different mean effect to each of the four F_2 genotypes.

We analyze these data in the same manner as the simulated AIL except for graphing the $-\log_{10}(P\text{-value})$ instead of LRTs and analyzing multiple map points in the intervals between marker loci. We enjoy two advantages over ANOVA or MANOVA; namely, we can use phenotyped individuals with wholly or partially missing genotypes, and we can estimate both QTL location and effect size.

To carry out a multivariate analysis, one must decide which univariate traits to analyze together. This is not a trivial matter because combining traits exacerbates the

TABLE 1

AIL: type I error, power, coverage, and average estimates

	Mendel Full	Mendel F_{10}	Cartographer	True value
Type I error	0.83–3.39	4.92–9.60	25.6–33.8	NA
Power	96	46	51	NA
Coverage	92	39	48	NA
Point 5				
η	3.9913 (0.2058)	3.9661 (0.2355)	NA	4.000
p_1	0.6486 (0.1183)	0.3259 (0.0878)	NA	0.667
ε_1	0.1910 (0.0487)	0.1759 (0.0764)	0.1869	0.200
ε_2	−0.1910 (0.0487)	−0.1759 (0.0764)	−0.1869	−0.200
σ^2_{env}	1.0112 (0.1002)	1.1494 (0.1721)	0.9581	1.000
Point 6				
η	3.9998 (0.2059)	3.9465 (0.2344)	NA	4.000
p_1	0.6492 (0.1176)	0.3370 (0.0967)	NA	0.667
ε_1	0.1932 (0.0477)	0.1764 (0.0758)	0.1781	0.200
ε_2	−0.1932 (0.0477)	−0.1764 (0.0758)	−0.1781	−0.200
σ^2_{env}	1.0092 (0.0998)	1.1458 (0.1746)	0.9606	1.000

Type I error rates, power, and coverage are percentages; estimates and standard errors are averages. Type I error rates are confidence intervals based on 500 simulations under the null hypothesis; other table entries are based on 100 simulations under the alternative hypothesis. We count successful coverage when the equivalent one-LOD drop (4.6 LRT units) interval around a significant map point includes the QTL. The QTL is located at the midpoint of points 5 and 6. Parameter p_1 is the reduced-dimension polygenic background parameter; η, the grand mean; ε_i, the QTL effect on strain i; and σ^2_{env}, the residual covariance. The Mendel estimates have average standard errors in parentheses. NA, not applicable.

multiple testing problem and may add noise and degrade power (Amos *et al.* 2001; Bauman *et al.* 2005). With outbred populations it is intertwined with the issue of ascertainment (Dawson and Elston 1984); it may also be a problem with inbred populations since strains are often chosen for a particular experiment on the basis of their average phenotype. We present here the results of two multivariate analyses making these points. The most interesting results from this example data set are on chromosome 3, and we focus on three traits, leptin measurements at both 4 and 15 months and insulin-like growth factor I at 15 months in this region. In univariate analysis, both IGF-15 and Lep-4 show significant linkage to markers on chromosome 3, while Lep-15 shows suggestive linkage. Multivariate analyses are indicated biologically, spatially, and temporally.

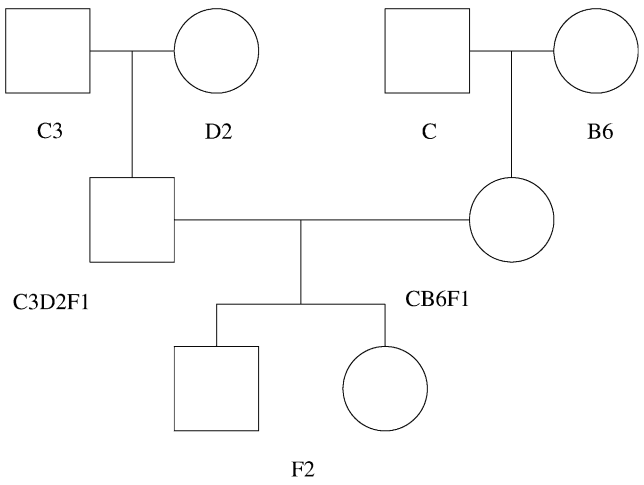

Figure 1.—Four-way cross for UM-HET3 Mice. UM-HET3 mice are created from four founder strains: BALB/cJ (C), C57BL/6J (B6), C3H/HeJ (C3), and DBA/2J (D2); the F_2 generation results from $CB6F_1$ females crossed with $C3D2F_1$ males.

We carried out a number of multivariate analyses; some of the results are summarized in Figures 2 and 3 and Table 2. The graphs of $-\log_{10}(P\text{-value})$ along chromosome 3 in Figure 2 correspond to the univariate analyses of IGF-15, Lep-4, and Lep-15 and the bivariate analysis of Lep-4 and Lep-15. The univariate graph of IGF-15 peaks over marker D3Mit5. Subjecting the *P*-values for IGF-15 to the nonlinear FDR correction (Benjamini *et al.* 2001) suggests a single location for IGF-15. Both of the univariate leptin graphs as well as the bivariate graph peak over D3Mit127. After FDR correction, at least two significant map points are suggested over D3Mit127 for the bivariate leptin analysis. Table 2 reports estimates and standard errors for the bivariate leptin mean parameters at marker D3Mit127. These estimates are very similar at the two time points. Although likelihood ratios improve over univariate analysis, *P*-values do not because the degrees of freedom of the χ^2 test double. The estimated environmental covariance matrix

$$\hat{\Sigma}_e = \begin{pmatrix} 0.96(0.05) & 0.50(0.04) \\ 0.50(0.04) & 0.97(0.05) \end{pmatrix} \qquad (17)$$

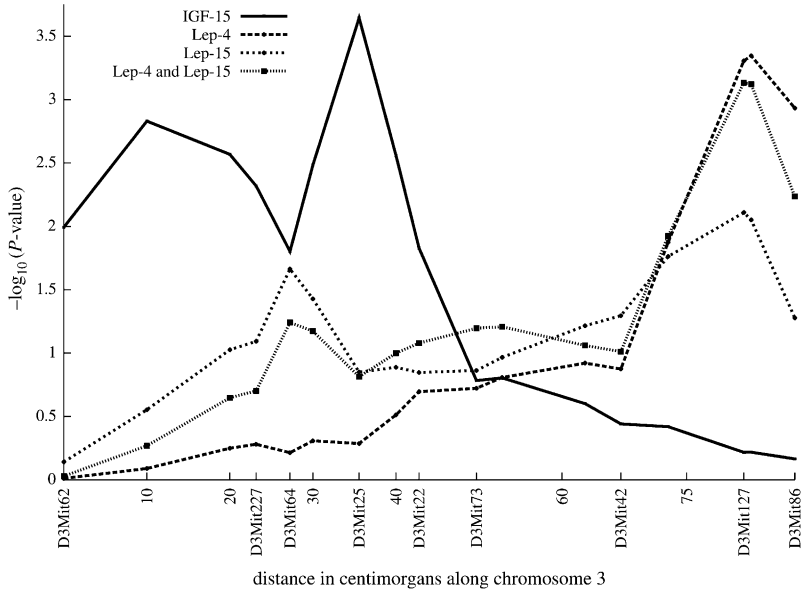

FIGURE 2.—Univariate and bivariate results, four-way cross on chromosome 3, univariate results for IGF-15 peak over marker D3Mit25. Univariate and bivariate results for Lep-4 and Lep-15 peak over marker D3Mit127.

is consistent with the raw correlation of the two traits. In the matrix (17), the standard error of each estimate appears in parentheses.

A trivariate analysis of IGF-15, Lep-4, and Lep-15 clearly illustrates that in the case of multivariate traits, more is not always better. Comparing Figure 2 to Figure 3 shows two large peaks: one at marker D3Mit25 and one over marker D3Mit127. After FDR adjustment only the first peak survives, and the evidence for it is compromised. Thus, the trivariate analysis provides no additional linkage information and actually degrades the power to detect linkage. While leptin and IGF share numerous biological interactions, there is no evidence in these data for a common genetic determinant on chromosome 3.

An eight-strain simulated cross: Our first two examples demonstrated the equivalence of the random effects model to the fixed effects model for standard cross designs and hint at the flexibility of our approach. To demonstrate this flexibility, we now present an eight-strain simulated example that (a) documents how correctly accounting for polygenic background can be beneficial and (b) demonstrates how it is possible to test hypotheses with the kind of unbalanced pedigree data encountered in human studies. As with the simulated AIL example, this exercise is not meant to be a substitute for an exhaustive study of power and experimental design.

Simulation specifics: Our simulated cross involves a univariate trait, eight inbred strains, and seven pedigrees of nine generations each. We are motivated in part by the heterogeneous stock (MOTT *et al.* 2000) and the collaborative-cross designs (WILLIAMS *et al.* 2002). Starting with strain-pure founders, we constructed each pedigree by random mating with a decreasing number of progeny per animal per generation. The average number of animals per pedigree is 366. Random mating ensures substantial diversity in theoretical and conditional strain fractions and coefficients. On the basis of the marker map for chromosome 2 in the UM-HET example of the previous section, we simulated genotypes at six loci using the gene-dropping option of Mendel. Locus 3 serves as the QTL and the remaining loci as markers. Genotypes at the QTL are omitted during linkage analysis.

We generated univariate trait values independently for each pedigree by sampling from a multivariate normal distribution with prescribed means and covariances. If animal i has QTL genotype a/b and trait value X_i, then

$$E(X_i) = \eta + 2\gamma_i^*\mu + \varepsilon(a) + \varepsilon(b),$$

where η is the grand mean, μ is the vector of polygenic deviations from the mean, and ε is the vector of QTL deviations from the mean. For animals i and j, the polygenic and random environment contributions entail the covariance

$$\mathrm{Cov}(X_i,\ X_j) = 4\,\mathrm{tr}(C_{ij}\Omega) + 1_{\{i=j\}}\sigma_e^2.$$

Note the absence here of a QTL variance contribution. Although the data are analyzed conditionally given observed marker genotypes, they are generated unconditionally. Table 3 displays the values of the parameters used for the simulations. These values were chosen randomly subject to constraints such as $\sum_a \mu(a) = 0$.

Our simulation choices present both opportunities and challenges. For example, the fact that each strain is assigned a unique QTL allele suggests that even a simple F_2 cross between two strains would be adequate to map the QTL. This advantage is tempered by the long genetic distances separating the QTL from the flanking markers,

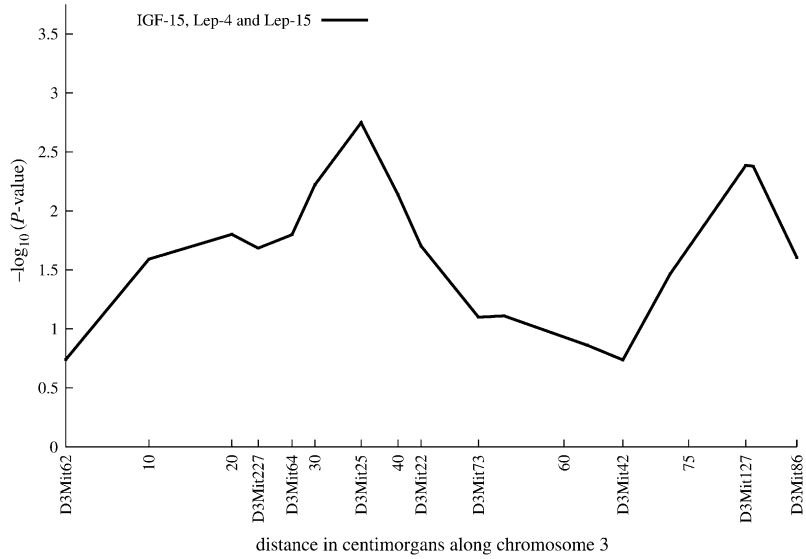

FIGURE 3.—Trivariate analysis, four-way cross on chromosome 3, trivariate results peak over marker D3Mit25 and D3Mit86. These peaks are lower than those obtained with univariate and bivariate analyses.

by the smallness of the QTL effects, by the similarity of these effects in some strains, and by the discordance of the QTL effects and the polygenic means effects.

In using random effect models for QTL mapping, inclusion of polygenic background is usually a good idea. If polygenic background is present but ignored, then the only way of accounting for relative correlations is through the QTL component. When we analyze the current data omitting polygenic background, every single chromosome location in the linkage analysis achieves a *P*-value <0.00001. Adding polygenic background causes *P*-values to reach more reasonable levels, ranging from 0.0019 to 0.3835. Subjecting the *P*-values to the (FDR) procedure highlights the QTL and one neighboring point as significant (BENJAMINI *et al.* 2001). Figure 4 plots the function $-\log_{10}(P\text{-value})$ along the chromosome; as earlier, the *P*-values reflect the likelihood ratio tests of the QTL component. The QTL is located at 30 cM from the origin between marker D2Mit323 at 23 cM and marker D2Mit37 at 42 cM.

We also used these data to illustrate the application of the QTL association model. As in our linkage analysis, omitting polygenic background leads to unrealistically small *P*-values. Figure 4 plots the $-\log_{10}(P\text{-value})$ for the association analysis with the polygenic background. The association results are similar to the linkage results. The marker with the most significant result is D2Mit323, which is the marker nearest to the QTL. The FDR procedure singles out D2Mit323 as the only significant association.

Comparison of computation times between the two models illustrates the speed of the association analysis. The linkage model requires ~4 hr for calculation of the coefficient matrices for each pedigree and ~20 hr to estimate the parameters for each of the 17 points. The association model requires ~1.5 hr for all calculations at each of the five markers.

DISCUSSION

In the hope of mapping QTL with small effects, geneticists are undertaking more ambitious crosses with multiple strains, multivariate traits, and dense marker sets. The random effects models developed here will enable a smooth transition to more sophisticated statistical analysis. The greatest strength of the models

TABLE 2

Four-way cross: mean estimates for bivariate leptin analysis at D3Mit127

	Lep-4		Lep-15	
Mean effect	Estimate	SE	Estimate	SE
B6/C3	0.1273	0.0616	0.1258	0.0638
B6/D2	−0.1686	0.0599	−0.1675	0.0632
C/C3	0.1812	0.0593	0.1198	0.0615
C/D2	−0.1399	0.0615	−0.0781	0.0646
Grand	−0.0157	0.0346	−0.0243	0.0362

SE, standard error.

TABLE 3

Eight-strain cross: simulation generating parameters

Inbred strain a	$\mu(a)$	$\varepsilon(a)$
BALB	−3.94	0.45
C57	5.62	0.24
C3H	−1.95	−0.24
DBA	2.13	−0.53
CAST	−3.68	−0.10
RIII	2.22	−0.68
I	−4.88	0.29
AKR	4.48	0.57
Grand	6.31	NA

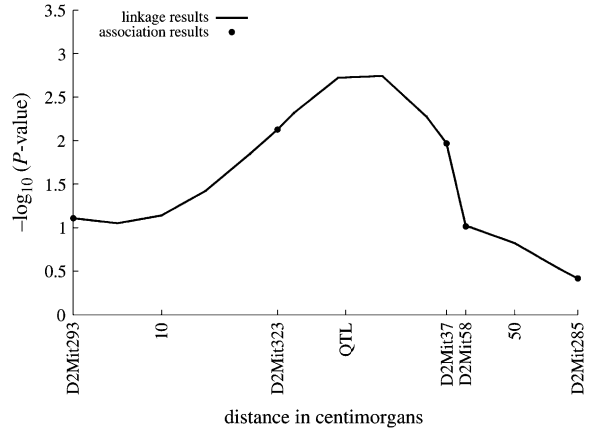

FIGURE 4.—Eight-strain cross example—linkage and association results for the simulated eight-strain random-mating example. Association results mirror linkage results at the markers. The linkage results peak over the QTL, located between markers D2Mit323 and D2Mit58.

is their ability to capture polygenic background parsimoniously. A second strength is their versatility in handling large pedigrees, large numbers of contributing strains, and multivariate traits. While we have warned against importing ideas wholesale from the rest of statistical genetics, judicious adaptations are fully warranted. For example, since environment can be exquisitely controlled for inbred strain experiments, models of gene-by-environment interaction can be put to good use on the mean level (BLANGERO 1993) and on the variance level (LANGE 1986; ITOH and YAMADA 1990). These techniques apply both to continuous traits (PLETCHER 1999; PLETCHER and GEYER 1999; JAFFRÉIC and PLETCHER 2000; PLETCHER and JAFFRÉIC 2002; PURCELL 2002; PURCELL and SHAM 2002; MEYER and KIRKPATRICK 2005) and to categorical traits (TOWNE *et al.* 1997; VIEL *et al.* 2005). It is also straightforward to model multiple QTL acting additively (LANGE 2002).

Balanced against these strengths is the need for better-conceived study designs. Unless crosses are carefully structured, some parameters will be unidentifiable. One antidote is to scale back the complexity of a model and reparameterize. Our first two examples illustrate this tactic. Another antidote is to avoid monolithic designs and opt for a mixture of designs that individually reveal different features of a model. Our third example does this.

In random effects models, trait values for most animals are correlated. Logically, one should treat all animals as members of a single large pedigree. At some point this requirement becomes unwieldy. The computational demands of the random effects models are fairly high, so tactics such as pedigree splitting, marker thinning, and marker amalgamation should not be dismissed. It will probably take a combination of these tactics to cope with the large-scale mapping projects now under way (PLETCHER *et al.* 2004). Fortunately, our experiences with simulated data suggest that a moderate amount of pedigree splitting sacrifices little information.

We have omitted a detailed discussion of how the program SimWalk delivers conditional strain fractions and coefficients. In our experience, SimWalk's MCMC algorithm adequately samples descent graph space. In association analysis, this lengthy process can be dispensed with if information at neighboring markers is ignored. Deterministic algorithms that produce approximate kinship and strain coefficients may ultimately be a better choice than stochastic sampling (GAO *et al.* 2004; GAO and HOESCHELE 2005). In maximizing loglikelihoods, it is also worth mentioning that Mendel allows the user to set initial parameter values and bounds. This flexibility is valuable in exploring multimodal likelihood surfaces.

Our QTL parameters enter the model at both the mean and the variance level and are not subject to nonnegativity constraints. Thus, the asymptotic distribution of a likelihood ratio test follows a chi-square distribution with degrees of freedom equal to the difference in the number of independent parameters between the underlying nested models. Model selection can be accomplished by likelihood ratio tests or modified criteria such as the Akaike information criteria (AIC) or the Bayesian information criteria (BIC). Multiple testing is certainly an issue. The FDR correction of Benjamini and Hochberg (BENJAMINI *et al.* 2001) for dependent tests is often a useful cure and provided us with correct inferences in our simulated examples. Extensions such as Storey's optimal discovery procedure (STOREY 2007; STOREY *et al.* 2007) can lead to more accurate *P*-values and should be kept in mind.

The assumption of multivariate normality is helpful in maximum likelihood estimation. For univariate traits with excess kurtosis, the multivariate *t* distribution is a workable substitute for the multivariate normal distribution and is an implemented option in Mendel. It is reasonable to conjecture that some version of the central limit theorem should hold for a polygenic trait over a pedigree (LANGE 1978; LANGE and BOEHNKE 1983). For simple pedigrees generated *en masse* in a cross, one can check the normality assumption empirically. The impact of departures from normality has been considered by several researchers (BEATY *et al.* 1985; ALLISON *et al.* 1999; PRATT *et al.* 2000). BLANGERO *et al.* (2000) and SHAM *et al.* (2000) suggest solutions to gross violations. One can object that QTL effects by their discrete nature cannot be normal. Three responses are possible. First, this objection has never stopped ordinary QTL mapping with outbred populations. Second, under the null hypothesis, the discrete effects disappear. Third, in all but the simplest crosses, application of a rigorous model incorporating both polygenes and major genes is very computationally demanding.

The web site (http://www.genetics.ucla.edu/software) offers the current versions of Mendel and SimWalk for

several computing platforms. Ample documentation and sample problems are provided. The experimental versions of Mendel and SimWalk featured in this article will be released publicly as soon as it is practical.

The authors are grateful to David Burke for access to the UM-HET3 data, to Karl Broman for his editorial interest and guidance, and to the anonymous reviewers for their helpful comments. This investigation was supported by U.S. Public Health Service grants MH59490, GM53275, T32-HG02536, and HL28481.

LITERATURE CITED

Allison, D. B., M. C. Neale, R. Zannolli, N. J. Schork, C. I. Amos *et al.*, 1999 Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am. J. Hum. Genet. **65:** 531–544.

Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. Am. J. Hum. Genet. **54:** 535–543.

Amos, C. I., M. de Andrade and D. K. Zhu, 2001 Comparison of multivariate tests for genetic linkage. Hum. Hered. **51:** 133–144.

Bauman, L. E., L. Almasy, J. Blangero, R. Duggirala, J. Sinsheimer *et al.*, 2005 Fishing for pleiotropic QTLs in a polygenic sea. Ann. Hum. Genet. **69:** 590–611.

Beaty, T. H., S. G. Self, K.-Y. Liang, M. A. Connolly, G. A. Chase *et al.*, 1985 Use of robust covariance components models to analyse triglyceride data in families. Ann. Hum. Genet. **49:** 315–328.

Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi and I. Golani, 2001 Controlling the false discovery rate in behavior genetics research. Behav. Brain Res. **125:** 279–284.

Blangero, J., 1993 Statistical approaches to human adaptability. Hum. Biol. **65:** 941–966.

Blangero, J., and L. Almasy, 1997 Multipoint oligogenic linkage analysis of quantitative traits. Genet. Epidemiol. **14:** 959–964.

Blangero, J., J. T. Williams and L. Almasy, 2000 Robust lod scores for variance component-based linkage analysis. Genet. Epidemiol. **19:** S8–S14.

Broman, K. W., H. Wu, Ś. Sen and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19:** 889–890.

Cervino, A. C. L., A. Darvasi, M. Fallahi, C. C. Mader and N. F. Tsinoremas, 2007 An integrated *in silico* gene mapping strategy in inbred mice. Genetics **175:** 321–333.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Darvasi, A., and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141:** 1199–1207.

Dawson, D. V., and R. C. Elston, 1984 A bivariate problem in human-genetics—ascertainment of families through a correlated trait. Am. J. Med. Genet. **18:** 435–448.

Flint, J., W. Valdar, S. Shifman and R. Mott, 2005 Strategies for mapping and cloning quantitative trait genes in rodents. Nat. Rev. Genet. **6:** 271–286.

Gao, G., and I. Hoeschele, 2005 Approximating identity-by-descent matrices using multiple haplotype configurations on pedigrees. Genetics **171:** 365–376.

Gao, G., I. Hoeschele, P. Sorensen and F. Du, 2004 Conditional probability methods for haplotyping in pedigrees. Genetics **167:** 2055–2065.

Goldgar, D. E., 1990 Multipoint analysis of human quantitative genetic-variation. Am. J. Hum. Genet. **47:** 957–967.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

Harper, J. M., A. T. Galecki, D. T. Burke, S. L. Pinkosky and R. A. Miller, 2003 Quantitative trait loci for insulin-like growth factor I, leptin, thyroxine, and corticosterone in genetically heterogeneous mice. Physiol. Genomics **15:** 44–51.

Hitzemann, R. W., B. Malmanger, S. Cooper, S. Coulombe, C. Reed *et al.*, 2002 Multiple cross mapping (MCM) markedly improves the localization of a QTL for ethanol-induced activation. Genes Brain Behav. **1:** 214–222.

Hopper, J. L., and J. D. Mathews, 1982 Extensions to multivariate normal models for pedigree analysis. Ann. Hum. Genet. **46:** 373–383.

Itoh, Y., and Y. Yamada, 1990 Relationships between genotype x environment interaction and genetic correlation of the same trait measured in different environments. Theor. Appl. Genet. **80:** 11–16.

Jaffréic, F., and S. D. Pletcher, 2000 Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. Genetics **156:** 913–922.

Janninka, J.-L., and R. Jansena, 2001 Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **157:** 445–454.

Kao, C.-H., Z.-B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. S. Lander, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. **58:** 1347–1363.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

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

Lange, K., 1986 Cohabitation, convergence and environmental covariances. Am. J. Med. Genet. **24:** 483–491.

Lange, K., 2002 *Mathematical and Statistical Methods for Genetic Analysis*, Ed. 2. Springer-Verlag, New York.

Lange, K., and M. Boehnke, 1983 Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. Am. J. Med. Genet. **14:** 513–524.

Lange, K., J. S. Sinsheimer and E. M. Sobel, 2005 Association testing with Mendel. Genet. Epidemiol. **29:** 36–50.

Li, R., M. A. Lyons, H. Wittenburg, B. Piagen and G. A. Churchill, 2005 Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. Genetics **169:** 1699–1709.

Liu, Y., and Z.-B. Zeng, 2000 A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. Genet. Res. **75:** 345–355.

Manly, K. F., and J. M. Olson, 1999 Overview of QTL mapping software and introduction to Map Manager QT. Mamm. Genome **10:** 327–334.

Meyer, K., and M. Kirkpatrick, 2005 Up hill, down dale: quantitative genetics of curvaceous traits. Philos. Trans. R. Soc. **360:** 1443–1455.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins and J. Flint, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. USA **97:** 12649–12654.

Pletcher, M. T., P. McClurg, S. Batalov, A. I. Su, S. W. Barnes *et al.*, 2004 Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. PLoS Biol. **2:** 2159–2169.

Pletcher, S. D., 1999 Model fitting and hypothesis testing for age-specific mortality data. J. Evol. Biol. **12:** 430–439.

Pletcher, S. D., and C. J. Geyer, 1999 The genetic analysis of age-dependent traits: modeling the character process. Genetics **153:** 825–835.

Pletcher, S. D., and F. Jaffréic, 2002 Generalized character process models: estimating the genetic basis of traits that cannot be observed and that change with age or environmental conditions. Biometrics **58:** 157–162.

Pratt, S. C., M. J. Daly and L. Kruglyak, 2000 Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. Am. J. Hum. Genet. **6:** 1153–1157.

Purcell, S., 2002 Variance components models for gene-environment interaction in twin analysis. Twin Res. **5:** 554–571.

Purcell, S., and P. Sham, 2002 Variance components models for gene-environment interaction in quantitative trait locus linkage analysis. Twin Res. **5:** 572–576.

Rebai, A., and B. Goffinet, 1993 Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor. Appl. Genet. **86:** 1014–1022.

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

Seaton, G., C. S. Haley, S. A. Knott, M. Kearsey and P. M. Visscher, 2002 QTL Express: mapping quantitative trait loci in simple and complex pedigrees. Bioinformatics **18:** 339–340.

Sen, Ś., and G. A. Churchill, 2001 A statistical framework for quantitative trait mapping. Genetics **159:** 371–387.

Sham, P. C., J. H. Zhao, S. S. Cherny and J. K. Hewitt, 2000 Variance-components QTL linkage analysis of selected and non-normal samples: conditioning on trait values. Genet. Epidemiol. **19:** S22–S28.

Sillanpää, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics **148:** 1373–1388.

Sobel, E. M., and K. Lange, 1996 Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. **58:** 1323–1337.

Storey, J. D., 2007 The optimal discovery procedure: a new approach to simultaneous significance testing. J. R. Stat. Soc. B **69:** 347–368.

Storey, J. D., J. Y. Dai and J. T. Leek, 2007 The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. Biostatistics **8:** 414–432.

Towne, B., R. M. Siervogel and J. Blangero, 1997 Effects of genotype-by-sex interaction on quantitative trait linkage analysis. Genet. Epidemiol. **14:** 1053–1058.

Valdar, W., J. Flint and R. Mott, 2006 Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. Genetics **172:** 1783–1797.

Viel, K. R., D. M. Warren, A. Buil, T. D. Dyer, T. E. Howard *et al.*, 2005 A comparison of discrete versus continuous environment in a variance components-based linkage analysis of the COGA data. BMC Genet. **6**(Suppl. 1): S57.

Williams, R. W., K. W. Broman, J. M. Cheverud, G. A. Churchill, R. W. Hitzemann *et al.*, 2002 A collaborative cross for high-precision complex trait analysis. First workshop report of the complex trait consortium. Technical report.

Xie, C., D. D. G. Gessler and S. Xu, 1989 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics **149:** 1139–1146.

Zeng, Z.-B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: K. W. Broman

APPENDIX A: REPRESENTATION OF POSITIVE DEFINITE MATRICES

Given a $k \times k$ positive definite matrix Ω and a $k \times 1$ vector μ, we now prove that vectors $\mu_1, \ldots, \mu_n$ exist such that $\sum_i \mu_i = \mu$ and $\sum_i \mu_i \mu_i^* = \Omega$. To simplify the proof, we pass to the spectral decomposition $\Omega = O^* D O$ of Ω. Here O is an orthogonal matrix, and D is a diagonal matrix whose jth diagonal entry d_j is an eigenvalue of Ω. If n vectors $\nu_1, \ldots, \nu_n$ exist such that $\sum_i \nu_i = O\mu = \nu$ and $\sum_i \nu_i \nu_i^* = D$, then taking $\mu_i = O^* \nu_i$ for each i completes the proof.

With the transformed problem, we can work on each dimension j separately. Suppose we can find scalars $a_1, \ldots, a_m$ such that $a_1 + \ldots + a_m = \nu_j$ and $a_1^2 + \ldots + a_m^2 = d_j$. Then we construct m vectors $w_1, \ldots, w_m$ whose entries are 0 except for their jth entries $w_{ij} = a_i$. These m vectors compose part of the solution set $\nu_1, \ldots, \nu_n$ and do not impinge on the parts contributed by other dimensions. To show that appropriate scalars $a_1, \ldots, a_m$ exist, we consider optimizing the function $f(a) = a_1^2 + \ldots + a_m^2$ subject to the affine constraint $a_1 + \ldots + a_m = \nu_j$. By introducing a Lagrange multiplier, we can prove that $f(a)$ attains its minimum ν_j^2 / m when all $a_i = \nu_j / m$. The maximum of $f(a)$ is infinite in all but the trivial case $m = 1$. For instance, we can take $a_1 = \frac{1}{2}(\nu_j + p)$, $a_2 = \frac{1}{2}(\nu_j - p)$, and all other $a_i = 0$ and send p to ∞. Since d_j must be positive, some positive integer m exists with $\nu_j^2 / m < d_j$. This choice of m puts us in a position to invoke the intermediate value theorem. The set of vectors $a = (a_1, \ldots, a_m)$ satisfying the constraint is convex and therefore connected. A continuous function on a connected set attains every value between its minimum and maximum values. Hence, there is some a with $f(a) = d_j$.

APPENDIX B: PROPERTIES OF THE C_{ij} MATRICES

The role of the matrix C_{ij} in formula (3) suggests its importance. Mathematically C_{ij} is better behaved than the strain coefficient matrix ψ_{ij}. Recall that the founder initial conditions and the recurrences (4)–(7) completely determine the strain fraction vectors γ_i and the strain coefficient matrices ψ_{ij}. If we retain the conventions that i has parents k and l and j is an animal previously considered, then the last three recurrences translate into the similar recurrences

$$\begin{aligned} C_{ij} &= \psi_{ij} - \gamma_i \gamma_j^* \\ &= \frac{1}{2}(\psi_{kj} + \psi_{lj}) - \frac{1}{2}(\gamma_k + \gamma_l)\gamma_j^* \\ &= \frac{1}{2}(C_{kj} + C_{lj}) \end{aligned} \tag{B1}$$

$$C_{ji} = \frac{1}{2}(C_{jk} + C_{jl}) \tag{B2}$$

and

$$\begin{aligned} C_{ii} &= \psi_{ii} - \gamma_i \gamma_i^* \\ &= \frac{1}{4}[\text{diag}(\gamma_k) + \text{diag}(\gamma_l) + \gamma_k \gamma_l^* + \gamma_l \gamma_k^*] \\ &\quad - \frac{1}{4}(\gamma_k + \gamma_l)(\gamma_k + \gamma_l)^* \\ &= \frac{1}{4}[\text{diag}(\gamma_k) + \text{diag}(\gamma_l) - \gamma_k \gamma_k^* - \gamma_l \gamma_l^*] \end{aligned} \tag{B3}$$

on the C_{ij} matrices. The next proposition collects some relevant facts.

PROPOSITION 1. *In addition to satisfying the recurrences* (*B*1), (*B*2), *and* (*B*3), *the matrix* C_{ij}

a. *has all entries* 0 *when either* i *or* j *is a founder,*
b. *is symmetric,*
c. *is positive semidefinite,*
d. *has the vector* $\mathbf{1}$ *in its null space,*
e. *has entries* $C_{ij}(m, n)$ *confined to the interval* $[-\frac{1}{8}, 0]$ *for* $n \neq m$ *and to the interval* $[0, \frac{1}{8}]$ *for* $n = m$.

Proof.

a. If i is a founder belonging to strain q and j is a founder belonging to strain r, then by definition $\gamma_i(m) = 1_{\{m=q\}}$, $\gamma_j(n) = 1_{\{n=r\}}$, and $\psi_{ij}(m, n) = 1_{\{m=q\}}1_{\{n=r\}}$. Thus, all entries of C_{ij} vanish. If i or j is a founder but the other is not, then induction and the recurrences (B1) and (B2) show that all entries of C_{ij} vanish.
b. Formula (B3) forces C_{ii} to be symmetric, and the recurrences (B1) and (B2) preserve symmetry.
c. Because the recurrences (B1) and (B2) preserve positive semidefiniteness, it suffices to prove that C_{ii} is positive semidefinite. Inspection of formula (B3) further demonstrates that it suffices to prove that $\text{diag}(\gamma_k) - \gamma_k \gamma_k^*$ is positive semidefinite for all k. Accordingly, let v be an arbitrary vector. The quadratic form

$$v^*[\text{diag}(\gamma_k) - \gamma_k \gamma_k^*]v = \sum_m \gamma_k(m) v_m^2 - \left[\sum_m \gamma_k(m) v_m\right]^2$$

is nonnegative owing to Cauchy's inequality

$$\left[\sum_m \gamma_k(m) v_m\right]^2 \leq \left[\sum_m \gamma_k(m)\right]\left[\sum_m \gamma_k(m) v_m^2\right]$$

and the fact that $\sum_m \gamma_k(m) = 1$.
d. Again this is a consequence of the recurrences (B1) and (B2) and the validity of the assertion for C_{ii}. In the latter case, the equality

$$[\text{diag}(\gamma_k) - \gamma_k \gamma_k^*]\mathbf{1} = \gamma_k - \gamma_k \gamma_k^* \mathbf{1} = \mathbf{0}$$

is obvious.
e. Because the stated bounds are preserved by recurrences (B1) and (B2), it suffices to consider C_{ii}. The contribution $\frac{1}{4}\gamma_k(m)[1 - \gamma_k(m)]$ to a diagonal term in Equation B3 is bounded below by 0 and above by $\frac{1}{16}$. The contribution $-\frac{1}{4}\gamma_k(m)\gamma_k(n)$ to an off-diagonal term is bounded below by $-\frac{1}{16}$ and above by 0. ■

The collection $\mathcal{C}$ of all C_{ij} matrices over a pedigree has considerable structure. For example, the symmetry of C_{ij} entails $C_{ij} = C_{ji}$. With just $s = 2$ strains, parts b, d, and e of Proposition 1 imply that every C_{ij} is representable as

$$C_{ij} = a_{ij}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

for some constant $a_{ij} \in [0, \frac{1}{8}]$. Furthermore, since $a_{ij} = 0$ whenever i or j is a founder or an F_1 individual, straightforward recursive arguments show that within any strictly linear mating designs like F_n, $a_{ij} = 0$ for all $i \neq j$. Two-strain systems also produce uninteresting *conditional* coefficients; straightforward calculations show that $\hat{C}_{ij} = \mathbf{0}$ for all i and j at markers that differentiate between the strains with complete genotyping.

To generalize this representation to more than two strains, it is helpful to introduce the $s \times s$ matrix E_{mn} where all entries of E_{mn} are 0 except for $e_{mm} = e_{nn} = 1$ and $e_{mn} = e_{nm} = -1$. There are $\binom{s}{2}$ such matrices.

PROPOSITION 2. *Every matrix C_{ij} from the collection $\mathcal{C}$ can be represented as a linear combination*

$$C_{ij} = \sum_{\{m,n\}} a_{ij,mn} E_{mn}. \tag{B4}$$

Furthermore, the coefficients $a_{ij,mn}$ are nonnegative dyadic rationals satisfying

$$\sum_{\{m,n\}} a_{ij,mn} \leq \frac{1}{4}\left(1 - \frac{1}{s}\right).$$

Proof. Each matrix $C_{ii} = \mathbf{0}$ corresponding to a founder i clearly qualifies. The representation (B4) is preserved by the averaging process of the recurrences (B1) and (B2), so it suffices to prove the representation for a matrix C_{ii} generated by a nonfounder. Again the averaging nature of recurrence (B3) allows us to verify the representation (B4) for a matrix of the form $\frac{1}{2}[\mathrm{diag}(\gamma_k) - \gamma_k \gamma_k^*]$. Because the set of dyadic rationals constitutes an algebraic field, it is clear by induction that all entries of γ_k are dyadic rationals. We now claim that

$$\frac{1}{2}[\mathrm{diag}(\gamma_k) - \gamma_k \gamma_k^*] = \frac{1}{2} \sum_{\{m,n\}} \gamma_k(m)\gamma_k(n) E_{mn}. \tag{B5}$$

Equality (B5) is certainly true for the off-diagonal entries of the matrices on both sides. For the diagonal entries, it is a consequence of the identity

$$\sum_{n \neq m} \gamma_k(m)\gamma_k(n) = \gamma_k(m)[1 - \gamma_k(m)].$$

Because the coefficients $\frac{1}{2}\gamma_k(m)\gamma_k(n)$ are dyadic rationals, all that remains is to check that the sum of the coefficients is properly bounded. This follows from

$$\begin{aligned} \frac{1}{2} \sum_{\{m,n\}} \gamma_k(m)\gamma_k(n) &= \frac{1}{4} \sum_m \sum_{n \neq m} \gamma_k(m)\gamma_k(n) \\ &= \frac{1}{4} \sum_m \gamma_k(m)[1 - \gamma_k(m)] \\ &\leq \frac{1}{4}\left(1 - \frac{1}{s}\right). \end{aligned} \tag{B6}$$

The upper bound (B6) can be proved by introducing a Lagrange multiplier corresponding to the constraint $\sum_m \gamma_k(m) = 1$. Equality is achieved only when all $\gamma_k(m) = 1/s$. ∎

APPENDIX C: COMPUTATION OF THE PROJECTION

Consider minimizing the function

$$f(x) = \frac{1}{2} \sum_m \sum_n (y_{mn} - x_{mn})^2$$

subject to the constraints $x_{mm} + x_{nn} = x_{mn} + x_{nm}$ for every unordered pair $\{m, n\}$. We proceed by seeking a stationary point of the Lagrangian

$$\begin{aligned} \mathcal{L}(x, \mu) = &\frac{1}{2} \sum_m \sum_n (y_{mn} - x_{mn})^2 \\ &+ \sum_{\{m,n\}} \mu_{\{m,n\}}(x_{mm} + x_{nn} - x_{mn} - x_{nm}). \end{aligned}$$

This point is characterized by the equations

$$\frac{\partial}{\partial x_{mn}} \mathcal{L}(x, \mu) = -(y_{mn} - x_{mn}) - \mu_{\{m,n\}} = 0 \tag{C1}$$

$$\frac{\partial}{\partial x_{mm}}\mathcal{L}(x,\,\mu) = -(y_{mm} - x_{mm}) + \sum_{n\neq m}\mu_{\{m,n\}} = 0$$

$$\frac{\partial}{\partial \mu_{\{m,n\}}}\mathcal{L}(x,\,\mu) = x_{mm} + x_{nn} - x_{mn} - x_{nm} = 0 \tag{C2}$$

with the convention that $m \neq n$. Rearrangement of Equation (C1) gives

$$x_{mn} = y_{mn} + \mu_{\{m,n\}}. \tag{C3}$$

If we interchange m and n in Equation C3, add the result to Equation C3, and invoke the constraint (C2), then we get the equation

$$x_{mm} + x_{nn} = x_{mn} + x_{nm} = y_{mn} + y_{nm} + 2\mu_{\{m,n\}},$$

determining $\mu_{\{m,n\}}$ as

$$\mu_{\{m,n\}} = \frac{1}{2}(x_{mm} + x_{nn} - y_{mn} - y_{nm}).$$

From Equation C3 it follows that

$$x_{mn} = \frac{1}{2}(y_{mn} - y_{nm}) + \frac{1}{2}(x_{mm} + x_{nn}). \tag{C4}$$

It is easy to check that the constraint (C2) is implicit in this solution. Furthermore, the solution entails the residual

$$y_{mn} - x_{mn} = \frac{1}{2}(y_{mn} + y_{nm}) - \frac{1}{2}x_{mm} - \frac{1}{2}x_{nn}.$$

If we set $a_{mn} = \frac{1}{2}(y_{mn} + y_{nm})$, our objective function can now be expressed as

$$f(x) = \frac{1}{2}\sum_m\sum_n\left(a_{mn} - \frac{1}{2}x_{mm} - \frac{1}{2}x_{nn}\right)^2.$$

Neither the off-diagonal entries x_{mn} nor the constraints now appear. To solve this unconstrained problem, we center the a_{mn} by subtracting their average value $\bar{a} = \bar{y}$. This allows us to reparameterize $f(x)$ as

$$\begin{aligned} f(x) &= \frac{1}{2}\sum_m\sum_n\left[a_{mn} - \bar{y} - \frac{1}{2}(x_{mm} - \bar{y}) - \frac{1}{2}(x_{nn} - \bar{y})\right]^2 \\ &= \frac{1}{2}\sum_m\sum_n(b_{mn} - u_m - u_n)^2 \end{aligned}$$

in more or less obvious notation.

Minimizing the objective function in this form coincides with a classical problem in population genetics. If we assume that m and n represent two possible alleles from s equally frequent alleles and b_{mn} represents a trait value determined by the genotype m/n, then minimizing $f(x)$ corresponds to the problem of determining the additive genetic variance of a centered trait. The solution to this problem is known to be

$$u_m = \frac{1}{s}\sum_n b_{mn}.$$

It follows that

$$x_{mm} = \frac{2}{s}\sum_n b_{mn} + \bar{y} = \frac{2}{s}\sum_n a_{mn} - \bar{y}.$$

A final substitution for a_{mn} gives

$$x_{mm} = \frac{1}{s}\sum_n y_{mn} + \frac{1}{s}\sum_n y_{nm} - \bar{y}$$

and the general formula

$$x_{mn} = \frac{1}{2}(y_{mn} - y_{nm}) + \frac{1}{2s}\sum_k y_{mk} + \frac{1}{2s}\sum_k y_{km} + \frac{1}{2s}\sum_k y_{nk} + \frac{1}{2s}\sum_k y_{kn} - \bar{y} \tag{C5}$$

based on Equation (C4) and valid for both $m \neq n$ and $m = n$.

The projection solution (C5) reduces the residual $r_{mn} = y_{mn} - x_{mn}$ to

$$r_{mn} = \frac{1}{2}(y_{mn} + y_{nm}) - \frac{1}{2s}\sum_k y_{mk} - \frac{1}{2s}\sum_k y_{km} - \frac{1}{2s}\sum_k y_{nk} - \frac{1}{2s}\sum_k y_{kn} + \bar{y}.$$

If we define a matrix R with entries r_{mn} and a matrix U with entries

$$u_{mn} = y_{mn} - \frac{1}{s}\sum_k y_{kn} - \frac{1}{s}\sum_k y_{mk} + \bar{y},$$

then it is clear that

$$r_{mn} = \frac{1}{2}(u_{mn} + u_{nm}).$$

In other words, the residual matrix $R = \frac{1}{2}(U + U^*)$ is a symmetrized version of U. Fortunately, we can represent U as the matrix product

$$U = \left(I - \frac{1}{s}\mathbf{1}\mathbf{1}^*\right) Y \left(I - \frac{1}{s}\mathbf{1}\mathbf{1}^*\right)$$

of $Y = (y_{mn})$ sandwiched between two copies of the orthogonal projection $Q = I - \frac{1}{s}\mathbf{1}\mathbf{1}^*$.

APPENDIX D: CONSTRUCTION OF AN ORTHOGONAL MATRIX

An orthogonal matrix O mapping the vector $(1/\sqrt{s})\mathbf{1}$ to the standard basis vector e_1 can be explicitly constructed by the Gramm–Schmidt process applied to the basis $\{(1/\sqrt{s})\mathbf{1},\ e_1, \ldots,\ e_{n-1}\}$, where e_k is the standard basis vector with 1 in position k and zeros elsewhere. The first row of O is just $o_1^* = (1/\sqrt{s})\mathbf{1}^*$; the subsequent rows take the form

$$o_k^* = \frac{1}{\sqrt{(s-k+1)(s-k+2)}}(0, \ldots, 0,\ s-k+1,\ -1, \ldots,\ -1),$$

where $k - 2$ zeros precede the entry $s - k + 1$. The reader can easily check that the row vectors o_k^* provide an orthonormal basis.

APPENDIX E: DIFFERENTIATION OF VARIANCES AND COVARIANCES

Because the fastest maximum likelihood algorithms rely on exact derivatives, there is an obvious need to calculate the partial derivatives of each covariance $\mathrm{Cov}(X_{ik},\ X_{jl})$ with respect to the entries of $\Delta = (\delta_{mn})$. If we let ∂_{mn} denote partial differentiation with respect to δ_{mn}, then formula (16) immediately leads to

$$\partial_{mn}\,\mathrm{Cov}(X_{ik},\ X_{jl}) = 4\,\mathrm{tr}[Z^*OC_{ij}O^*Z\partial_{mn}(\Delta\Delta^*)_{kl}],$$

so it suffices to compute the partial derivatives of $\Delta\Delta^* = (d_{uv})$. Since

$$d_{uv} = \sum_{w=1}^{\min\{u,v\}} \delta_{uw}\delta_{vw},$$

the product rule of differentiation yields

$$\partial_{mn} d_{uv} = 1_{\{m=u\}} 1_{\{n \le v\}} \delta_{vn} + 1_{\{m=v\}} 1_{\{n \le u\}} \delta_{un}.$$

Thus, $\partial_{mn}(\Delta\Delta^*)$ consists entirely of zeros except for row m and column n. This fact considerably simplifies computation of derivatives.

APPENDIX F: A COUNTEREXAMPLE ON IDENTIFIABILITY

Finally, we consider a counterexample that illustrates some of the subtleties of identifiability. We noted that projection replaces each trait block $Y = \Omega_{kl}$ with a symmetrized block residual

$$Y - X = \frac{1}{2}(U + U^*).$$

For purposes of computing covariances, we argued that symmetrization is unnecessary and avoiding it simultaneously yields correct covariances and reduces the number of parameters. We have not actually demonstrated that no further reduction is possible. Furthermore, exploiting the symmetrized version may lead to a residual $\Omega - P(\Omega)$ that fails to be positive semidefinite. Consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & 0 & 0 & 2 \end{pmatrix}.$$

Straightforward algebra leads to the positive semidefinite matrix

$$B = AA^* = \begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 16 & 4 & 0 \\ 1 & 4 & 3 & -1 \\ -1 & 0 & -1 & 5 \end{pmatrix}.$$

If we symmetrize each 2×2 block of B, then we get

$$C = \begin{pmatrix} 1 & 0 & 1 & \frac{3}{2} \\ 0 & 16 & \frac{3}{2} & 0 \\ 1 & \frac{3}{2} & 3 & -1 \\ \frac{3}{2} & 0 & -1 & 5 \end{pmatrix}.$$

A tedious computation shows that

$$\det C = -\frac{291}{16},$$

and C cannot be positive semidefinite.