On Marker-Assisted Prediction of Genetic Value: Beyond the Ridge

Daniel Gianola,*,1 Miguel Perez-Enciso† and Miguel A. Toro‡

*Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53706, †Station d'Amelioration Génétique des Animaux, Institut National de la Recherche Agronomique, 31326 Castanet-Tolosan, France and ‡Departamento de Mejora Genética Animal, Instituto Nacional de Investigaciones Agrarias, 28040-Madrid, Spain

ABSTRACT

Marked-assisted genetic improvement of agricultural species exploits statistical dependencies in the joint distribution of marker genotypes and quantitative traits. An issue is how molecular (*e.g.*, dense marker maps) and phenotypic information (*e.g.*, some measure of yield in plants) is to be used for predicting the genetic value of candidates for selection. Multiple regression, selection index techniques, best linear unbiased prediction, and ridge regression of phenotypes on marker genotypes have been suggested, as well as more elaborate methods. Here, phenotype-marker associations are modeled hierarchically via multilevel models including chromosomal effects, a spatial covariance of marked effects within chromosomes, background genetic variability, and family heterogeneity. Lorenz curves and Gini coefficients are suggested for assessing the inequality of the contribution of different marked effects to genetic variability. Classical and Bayesian methods are presented. The Bayesian approach includes a Markov chain Monte Carlo implementation. The generality and flexibility of the Bayesian method is illustrated when a Lorenz curve is to be inferred.

THE availability of a plethora of markers has led to consideration of the issue of the extent to which molecular information can be used to advantage in genetic improvement programs of agricultural species, such as maize or dairy cattle. There is a large body of literature on this matter (*e.g.*, SOLLER and BECKMANN 1983; SMITH and SIMPSON 1986; LANDE and THOMPSON 1990); WHITTAKER (2001) gives a review. Marker-assisted selection can be effective as long as part of the genetic variance can be associated with segregating marker loci (LANDE and THOMPSON 1990).

The basic idea in marker-assisted selection is to exploit statistical dependencies (linkage disequilibrium) existing in the joint distribution of marker and quantitative trait loci (QTL) genotypes. For example, when two inbred lines are crossed, the disequilibrium is manifest in the F_2 generation. On the other hand, when there is linkage equilibrium at the population level, only the joint distribution of marker and QTL genotypes within a family is nontrivial (OLLIVIER 1998). GEORGES *et al.* (1995) have exploited high levels of within-family disequilibria in dairy cattle populations for QTL mapping. Most livestock populations have some disequilibrium due at least to chance (small effective size), as shown by FARNIR *et al.* (2000).

Linkage disequilibrium between markers and QTL can be used for two main purposes: (1) to infer genomic location and effects of QTL affecting a trait and (2) to arrive at improved (in some statistical sense) predictions of genetic merit of candidates for selection in a breeding program. These two objectives may not be disjoint (*e.g.*, FERNANDO and GROSSMAN 1989). However, genetic cartography of QTL is not a requirement for prediction of genetic merit or marker-assisted selection (LANGE and WHITTAKER 2001). In fact, remarkable advances have been made in prediction of breeding values in livestock since the introduction of the best linear unbiased predictor (BLUP; HENDERSON *et al.* 1959; HENDERSON 1973). This method uses only phenotypic and pedigree information, with the QTL viewed only in an abstract manner. On the other hand, it has been argued and shown, at least in simulations, that molecular information may enhance the accuracy of selection (*e.g.*, WHITTAKER *et al.* 2000).

Our concern is with statistical models and methods for inferring genetic merit using molecular and phenotypic information. The objective is to describe phenotype-marker associations using multilevel hierarchical linear models. The setting is mainly as in WHITTAKER *et al.* (2000) and LANGE and WHITTAKER (2001), *i.e.*, situations where conditional independence of genetic sampling can be assumed, such as in an F_2 population derived from a cross between inbred lines. Model features include chromosome-specific effects, spatial associations of markers within chromosomes, existence of background genetic variability, and heterogeneity among families, if some such clustering exists. Classical and Bayesian methods are described. First, we present a mixed-effects model formulation and a BLUP implementation, with the dispersion components estimated

[1] *Corresponding author:* Department of Animal Sciences, 1675 Observatory Dr., Madison, WI 53706. E-mail: gianola@calshp.cals.wisc.edu

by some likelihood-based procedure. An extension of the mixed-effects model is given subsequently. Finally, a Bayesian formulation is presented, including a Markov chain Monte Carlo procedure for drawing samples from target posterior distributions. Possible applications to outbred populations are discussed.

A MIXED-EFFECTS MODEL FORMULATION

Hierarchical representation: Let the phenotypic value of individual i for a quantitative trait in an F_2 from a cross between inbred lines be described by the model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + a_i + e_i; \quad i = 1, 2, \ldots, n. \tag{1}$$

Here, the $p \times 1$ vector $\boldsymbol{\beta}$ contains some systematic effects representing, *e.g.*, year of harvest, level of fertilization, or plant density; $\mathbf{x}_i'$ is a known incidence vector relating $\boldsymbol{\beta}$ to y_i; a_i is an unobserved genetic value; and e_i is an independently distributed random residual reflecting environmental variability or inadequacy of the model. It is assumed that the genetic value a_i results from an unknown number (K, say) of QTL acting additively, so that

$$a_i = \sum_{k=1}^{K} \alpha_k Q_{ik},$$

where Q_{ik} is the genotype at biallelic (assumed for simplicity) locus k for individual i and α_k is the per-allele effect. If a_i is random, (1) is a special case of the well-known mixed linear model (*e.g.*, HENDERSON 1973). When K goes to infinity this becomes the classical infinitesimal model of quantitative genetics.

It is conceptually convenient to develop model (1) hierarchically. The first level of a Gaussian hierarchy is given by

$$y_i|\boldsymbol{\beta}, a_i, \sigma_e^2 \sim N(\mathbf{x}_i'\boldsymbol{\beta} + a_i, \sigma_e^2), \tag{2}$$

where $N(\cdot)$ indicates a normal distribution and σ_e^2 is the environmental variance. If the n environmental deviates are independently and identically distributed, this leads to the matrix representation

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{a}, \mathbf{I}\sigma_e^2), \tag{3}$$

where $\mathbf{X}$ is an $n \times p$ incidence matrix assumed (without loss of generality) to have full-column rank, $\mathbf{a} = \{a_i\}$, and $\mathbf{I}$ denotes an identity matrix, in this case $n \times n$. Suppose next that individual i has been typed for marker genotypes at each of l loci; this is represented by the vector

$$\mathbf{m}_i' = [m_{i1}, m_{i2}, \ldots, m_{il}],$$

with

$$m_{ij} = \begin{cases} -1 & \text{for } aa \\ 0 & \text{for } Aa \\ 1 & \text{for } aa. \end{cases}$$

Assume that all individuals have been typed for all markers (this is not realistic, but see the DISCUSSION). Then, the unobserved genetic value can be modeled as

$$a_i = \sum_{k=1}^{l} m_{ik}\gamma_k + \varepsilon_i = \mathbf{m}_i'\boldsymbol{\gamma} + \varepsilon_i, \tag{4}$$

where $\boldsymbol{\gamma} = \{\gamma_k\}$ is of order $l \times 1$. We refer to γ_k as the "marked effect" of marker locus k on genetic value, while ε_i can be interpreted as some residual or "background" genetic effect not involved in the association between genetic value and the markers but, yet, having an effect on phenotype. The vector $\boldsymbol{\gamma}$ is the gradient or regression of the unobservable additive genetic value on the observable marker genotype, that is, $\boldsymbol{\gamma} = \partial a_i/\partial \mathbf{m}_i$. As noted by LANDE and THOMPSON (1990), dominance can be introduced by expanding (4) as, *e.g.*,

$$a_i = \mathbf{m}_i'\boldsymbol{\gamma}_A + \mathbf{m}_i^{2'}\boldsymbol{\gamma}_D + \varepsilon_i,$$

where $\mathbf{m}_i^{2'}$ is a row vector with elements consisting of the squares of the corresponding entries of $\mathbf{m}_i$. Interactions between marked effects for different marker loci can be modeled via cross-products between appropriate elements of $\mathbf{m}_i$. For simplicity, additivity of marked effects is assumed throughout.

The second level of the hierarchy is represented by a distribution describing the uncertainty about genetic values, given the marked effects, that is the background genetic variability. We adopt the Gaussian model,

$$a_i|\mathbf{m}_i, \boldsymbol{\gamma}, \sigma_\varepsilon^2 \sim N(\mathbf{m}_i'\boldsymbol{\gamma}, \sigma_\varepsilon^2),$$

where σ_ε^2 is the background additive genetic variance. It is assumed (rightly or wrongly, depending on the context), given the marker genotypes and $\boldsymbol{\gamma}$, that the "background" genetic effects ε_i of different individuals are mutually independent. This implies that either there is no family structure or that, conditionally on the marked effects $\boldsymbol{\gamma}$, the family structure is not relevant. Family clustering is taken up in a later section. In matrix notation, and consistently with (4), the assumption of independence leads to

$$\mathbf{a}|\mathbf{M}, \boldsymbol{\gamma}, \sigma_\varepsilon^2 \sim N(\mathbf{M}\boldsymbol{\gamma}, \mathbf{I}\sigma_\varepsilon^2), \tag{5}$$

where $\mathbf{M}$ is the $n \times l$ matrix of known marker genotypes. Unless there is some prior knowledge about σ_e^2 and σ_ε^2 (or some clustering of individuals, such as a family structure), the background effect ε_i must be lumped together with e_i, because of nonidentifiability. On the other hand, if the variances σ_ε^2 and σ_e^2 are known *a priori*, it is possible to "predict" ε_i distinctly from e_i, in the same way that one can predict additive genetic and environmental effects via BLUP when dispersion parameters are known.

WHITTAKER *et al.* (2000) and WHITTAKER (2001) treat $\boldsymbol{\gamma}$ as a fixed parameter and employ ridge regression for

estimation of this vector. From a Bayesian perspective (*e.g.*, LINDLEY and SMITH 1972; ZELLNER and VANDAELE 1975), this is equivalent to regarding $\boldsymbol{\gamma}$ as having the distribution

$$\boldsymbol{\gamma}|\sigma_\gamma^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_\gamma^2), \tag{6}$$

with σ_γ^2 elicited in some manner. Assumption (6) is adopted as the third level of the hierarchy. In a frequentist setting, the assumption states that the marked effects $\boldsymbol{\gamma}$ are drawn at random from the multivariate normal distribution (6) in each conceptual repetition of a crossing experiment. In a Bayesian setting, this would be part of the prior ensemble of the model; this is discussed later. Regression coefficients for markers that are not adjacent to QTL are expected to be null under the assumption of no interference (ZENG 1993). However, without knowing the location of the QTL in relation to the markers (the usual situation), it is not obvious how such a prior consideration can be incorporated into the model. Since (6) is a prior distribution in the Bayesian sense, its influence on inferences can be tempered by measuring enough individuals. At this point, it suffices to say that $\boldsymbol{\gamma}$ can be treated as a random effect merely as a device for obtaining possibly improved (in some sense) predictions of genetic value. HAYES and GODDARD (2001) assumed that unobservable gene or chromosome effects followed a Gamma distribution, so these effects would be strictly positive even though estimated values can be negative. MEUWISSEN *et al.* (2001) used Gamma deviates for simulating gene effects, but then "tossed a coin" to determine their sign. Therefore, it is unclear what would be gained from assuming a Gamma distribution for the elements of $\boldsymbol{\gamma}$. Note that (6) implies that the "marked effects" are independent and identically distributed, but this assumption is relaxed later on. The normality assumption in (6) is probably adequate enough and facilitates computation significantly.

The three-stage hierarchy can be condensed by inserting (5) into (3), so that the model describing the phenotypic values can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \mathbf{e} \tag{7}$$

$$= \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} + \mathbf{e}, \tag{8}$$

where $\boldsymbol{\varepsilon} = \{\varepsilon_i\}$ and $\mathbf{e} = \{e_i\}$. This can be viewed as a frequentist mixed-effects model, where $\boldsymbol{\beta}$ is a fixed location parameter and $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are random terms. Unless additional assumptions are made or some knowledge about the partition of variance in the population is available, ε_i and e_i are "confounded." However, it is conceptually useful to maintain these two vectors as distinct. The marginal (frequentist) distribution of the phenotypes induced by model (8) is the normal process

$$\mathbf{y}|\boldsymbol{\beta}, \sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{M} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{MM}'\sigma_\gamma^2 + \mathbf{I}(\sigma_\varepsilon^2 + \sigma_e^2)), \tag{9}$$

where $\mathbf{MM}'\sigma_\gamma^2$ is the variance-covariance matrix of marked genotypic values, conditionally on the marker genotypes $\mathbf{M}$ observed in the experiment. In scalar notation, the "total variance" of a single observation is

$$\mathrm{Var}(y_i|\boldsymbol{\beta}, \sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{M}) = \sum_{k=1}^l m_{ik}^2\sigma_\gamma^2 + \sigma_\varepsilon^2 + \sigma_e^2.$$

Hence,

$$\frac{\sum_{k=1}^l m_{ik}^2\sigma_\gamma^2}{\mathrm{Var}(a_i)} = \frac{\sum_{k=1}^l m_{ik}^2\sigma_\gamma^2}{\sum_{k=1}^l m_{ik}^2\sigma_\gamma^2 + \sigma_\varepsilon^2} \tag{10}$$

is interpretable as the fraction of genetic variance attributable to marked effects, in an experiment repeated over and over with the marker genotypes fixed across replications. The variance "due to" the association with the markers is $\sum_{k=1}^l m_{ik}^2\sigma_\gamma^2$, which depends nontrivially on $\mathbf{m}_i$, the specific marker genotype of individual i. On the other hand, since the marker genotypes vary at random over replications,

$$\mathrm{Var}(\mathbf{M}\boldsymbol{\gamma}) = E[\mathrm{Var}(\mathbf{M}\boldsymbol{\gamma}|\mathbf{M})] + \mathrm{Var}[E(\mathbf{M}\boldsymbol{\gamma}|\mathbf{M})]$$

$$= \sigma_\gamma^2 E(\mathbf{MM}'), \tag{11}$$

where the expectation and covariance matrix are taken over the distribution of marker genotypes in the population. Knowledge of the distribution of marker genotypes is needed for evaluation of (11).

Best prediction and best linear prediction: Under standard assumptions, with $\boldsymbol{\beta}$ and the variance components σ_γ^2, σ_ε^2, and σ_e^2 known, the joint distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$, given the phenotypes and the marker genotypes, is the multivariate normal process:

$$\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} \Big| \boldsymbol{\beta}, \sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{y}, \mathbf{M} \sim N\left(\begin{bmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\varepsilon}} \end{bmatrix}, \begin{bmatrix} \mathbf{V}_{\gamma\gamma} & \mathbf{V}_{\gamma\varepsilon} \\ \mathbf{V}_{\varepsilon\gamma} & \mathbf{V}_{\varepsilon\varepsilon} \end{bmatrix} \right). \tag{12}$$

Here,

$$\begin{bmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\varepsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}'\mathbf{M} + \mathbf{I}\tau_\gamma & \mathbf{M}' \\ \mathbf{M} & \mathbf{I} + \mathbf{I}\tau_\varepsilon \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{M}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \tag{13}$$

where $\tau_\gamma = \sigma_e^2/\sigma_\gamma^2$, $\tau_\varepsilon = \sigma_e^2/\sigma_\varepsilon^2$, and

$$\begin{bmatrix} \mathbf{V}_{\gamma\gamma} & \mathbf{V}_{\gamma\varepsilon} \\ \mathbf{V}_{\gamma\varepsilon} & \mathbf{V}_{\varepsilon\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{M}'\mathbf{M} + \mathbf{I}\tau_\gamma & \mathbf{M}' \\ \mathbf{M} & \mathbf{I} + \mathbf{I}\tau_\varepsilon \end{bmatrix}^{-1} \sigma_e^2. \tag{14}$$

The best linear predictor [BLP; also the best predictor (BP) under normality; HENDERSON 1973] of the unobserved total genetic values $\mathbf{a} = \mathbf{M}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ is

$$E(\mathbf{M}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}|\boldsymbol{\beta}, \sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{y}, \mathbf{M}) = \mathbf{M}\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\varepsilon}} = \widehat{\mathbf{a}}, \tag{15}$$

and the variance-covariance matrix of the prediction

error (under normality this is also the covariance matrix of the conditional distribution of $\mathbf{a}$) is

$$
\mathrm{Var}[\mathbf{M\gamma} + \mathbf{\varepsilon} - (\widehat{\mathbf{M\gamma}} + \widehat{\mathbf{\varepsilon}})|\mathbf{\beta}, \sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{M}]
$$

$$
= \mathrm{Var}(\mathbf{M\gamma} + \mathbf{\varepsilon}|\mathbf{\beta}, \sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{y}, \mathbf{M})
$$

$$
= [\mathbf{M}\ \mathbf{I}]\begin{bmatrix} \mathbf{M'M} + \mathbf{I}\tau_\gamma & \mathbf{M'} \\ \mathbf{M} & \mathbf{I} + \mathbf{I}\tau_\varepsilon \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{M'} \\ \mathbf{I} \end{bmatrix}\sigma_e^2. \quad (16)
$$

The best predictor has the smallest possible mean-squared error of prediction. Hence, it would be difficult to improve upon this, provided that the model is reasonable, normality holds, and parameters are known. Thus, (13–16) provide an alternative to a ridge regression approach to prediction. Generalization to multiple traits measured in different individuals is straightforward, but this is not dealt with here. Since, given $\mathbf{M}$, the BLP or BP is unbiased (*e.g.*, HENDERSON 1973), it follows automatically that it is also unbiased unconditionally. However, unconditionally,

$$
\mathrm{Var}[\mathbf{M\gamma} + \mathbf{\varepsilon} - (\widehat{\mathbf{M\gamma}} + \widehat{\mathbf{\varepsilon}})|\mathbf{\beta}, \sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2]
$$

$$
= \sigma_e^2 E\left\{ [\mathbf{M}\ \mathbf{I}]\begin{bmatrix} \mathbf{M'M} + \mathbf{I}\tau_\gamma & \mathbf{M'} \\ \mathbf{M} & \mathbf{I} + \mathbf{I}\tau_\varepsilon \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{M'} \\ \mathbf{I} \end{bmatrix}\right\},
$$

with the expectation taken with respect to the joint distribution of marker genotypes in the entire population. A drawback of BP or BLP is that it is unrealistic to assume that $\mathbf{\beta}$, σ_γ^2, σ_ε^2, and σ_e^2 are known without error.

Best linear unbiased prediction: An obvious improvement is to use BLUP. BLUP takes into account uncertainty about $\mathbf{\beta}$, which is not the case of BP or BLP above, where $\mathbf{\beta}$ is treated as known. Under normality, BLUP($\mathbf{a}$) can be interpreted as the mean of the conditional distribution of the predictand $\mathbf{a} = \mathbf{M\gamma} + \mathbf{\varepsilon}$, given a vector of "error contrasts," denoted as $\mathbf{w}$. For example, take $\mathbf{w} = \mathbf{y} - \mathbf{X}\widehat{\mathbf{\beta}}$, where $\widehat{\mathbf{\beta}}$ is either the ordinary least-squares or the generalized least-squares estimator of $\mathbf{\beta}$. In such a setting, BLUP is the best predictor under normality, but only in the class of linear translation invariant predictors (SEARLE 1974; GIANOLA and GOFFINET 1982). It is well known that for (8) and (9),

$$
\begin{bmatrix} \mathbf{\gamma} \\ \mathbf{\varepsilon} \end{bmatrix}\bigg|\sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{w}, \mathbf{M} \sim N\left(\begin{bmatrix} \widehat{\widehat{\mathbf{\gamma}}} \\ \widehat{\widehat{\mathbf{\varepsilon}}} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{\gamma\gamma} & \mathbf{C}_{\gamma\varepsilon} \\ \mathbf{C}_{\varepsilon\gamma} & \mathbf{C}_{\varepsilon\varepsilon} \end{bmatrix}\right), \quad (17)
$$

where, for $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$,

$$
\begin{bmatrix} \widehat{\widehat{\mathbf{\gamma}}} \\ \widehat{\widehat{\mathbf{\varepsilon}}} \end{bmatrix} = \begin{bmatrix} \mathbf{M'PM} + \mathbf{I}\tau_\gamma & \mathbf{M'P} \\ \mathbf{PM} & \mathbf{P} + \mathbf{I}\tau_\varepsilon \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{M'Py} \\ \mathbf{Py} \end{bmatrix} \quad (18)
$$

gives $\widehat{\widehat{\mathbf{\gamma}}} = \mathrm{BLUP}(\mathbf{\gamma})$ and $\widehat{\widehat{\mathbf{\varepsilon}}} = \mathrm{BLUP}(\mathbf{\varepsilon})$. Further,

$$
\begin{bmatrix} \mathbf{C}_{\gamma\gamma} & \mathbf{C}_{\gamma\varepsilon} \\ \mathbf{C}_{\varepsilon\gamma} & \mathbf{C}_{\varepsilon\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{M'PM} + \mathbf{I}\tau_\gamma & \mathbf{M'P} \\ \mathbf{PM} & \mathbf{P} + \mathbf{I}\tau_\varepsilon \end{bmatrix}^{-1}\sigma_e^2. \quad (19)
$$

The BLUP of $\mathbf{a} = \mathbf{M\gamma} + \mathbf{\varepsilon}$ is

$$
\widehat{\widehat{\mathbf{a}}} = E(\mathbf{M\gamma} + \mathbf{\varepsilon}|\sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{w}, \mathbf{M}) = \mathbf{M}\widehat{\widehat{\mathbf{\gamma}}} + \widehat{\widehat{\mathbf{\varepsilon}}},
$$

and the variance-covariance matrix of the prediction errors is

$$
\mathrm{Var}[\mathbf{M\gamma} + \mathbf{\varepsilon} - (\mathbf{M}\widehat{\widehat{\mathbf{\gamma}}} + \widehat{\widehat{\mathbf{\varepsilon}}})|\sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{M}]
$$

$$
= \mathrm{Var}(\mathbf{M\gamma} + \mathbf{\varepsilon}|\sigma_\gamma^2, \sigma_\varepsilon^2, \sigma_e^2, \mathbf{w}, \mathbf{M})
$$

$$
= [\mathbf{M}\ \mathbf{I}]\begin{bmatrix} \mathbf{M'PM} + \mathbf{I}\tau_\gamma & \mathbf{M'P} \\ \mathbf{PM} & \mathbf{P} + \mathbf{I}\tau_\varepsilon \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{M'} \\ \mathbf{I} \end{bmatrix}\sigma_e^2.
$$

As noted, BLUP of the unobservable genetic values (marked or background genetic effects or any linear combination thereof) takes into account the uncertainty about $\mathbf{\beta}$, although given τ_γ, τ_ε, and σ_e^2. Since BLUP is conditionally unbiased, it follows that it is also so unconditionally (averaged over marker genotypes). The unconditional covariance matrix of the prediction errors is the average of (19) taken over the distribution of marker genotypes.

The needed variance components: Sensible values of the dispersion parameters must be specified for implementing BLUP. As stated, unless there is prior or external knowledge, it is not possible to separate the variance of the background genetic effects (σ_ε^2) from that of the environmental influences (σ_e^2). The model can be rewritten as

$$
\mathbf{y} = \mathbf{X\beta} + \mathbf{M\gamma} + \mathbf{e}^*, \quad (20)
$$

where $\mathbf{e}^* \sim N(\mathbf{0}, \mathbf{I}\sigma_e^{2*})$. Methods for obtaining maximum-likelihood estimates of σ_γ^2 and of $(\sigma_\varepsilon^2 + \sigma_e^2)$ for (20) are well known, *e.g.*, SEARLE *et al.* (1992). In this setting only the marked part of the genetic value can be predicted. Here, the conditional distribution used for inferring $\mathbf{\gamma}$ would be

$$
\mathbf{\gamma}|\sigma_\gamma^2, \sigma_e^{2*}, \mathbf{w}, \mathbf{M} \sim N(\tilde{\mathbf{\gamma}}, \mathbf{C}_{\gamma\gamma}^*), \quad (21)
$$

where

$$
\tilde{\mathbf{\gamma}} = \left[\mathbf{M'PM} + \mathbf{I}\frac{\sigma_e^{2*}}{\sigma_\gamma^2}\right]^{-1}\mathbf{M'Py}
$$

is the best predictor (given the error contrasts) or BLUP($\mathbf{\gamma}$). The covariance matrix of the conditional distribution (21) is

$$
\mathbf{C}_{\gamma\gamma}^* = \left[\mathbf{M'PM} + \mathbf{I}\frac{\sigma_e^{2*}}{\sigma_\gamma^2}\right]^{-1}\sigma_e^{2*},
$$

with this being the same as the covariance matrix of the prediction errors $\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}$, given the marker genotypes $\mathbf{M}$. The BLUP of the marked quantitative trait genotype is $\mathbf{M}\tilde{\boldsymbol{\gamma}}$, and the prediction error dispersion matrix is $\mathbf{M}\mathbf{C}^{*}_{\gamma\gamma}\mathbf{M}'\sigma^{2*}_{e}$. Hence, $\mathbf{M}\tilde{\boldsymbol{\gamma}}$ could be used as a criterion for genetic evaluation in marker-assisted selection ignoring background effects. If reasonable maximum-likelihood (ML) or restricted maximum-likelihood (REML) estimates of σ^{2}_{γ} and of σ^{2*}_{e} are available, these can be treated as "true" values in (21).

Suppose that estimates of the additive genetic $\left(\widehat{\sigma^{2}_{a}}\right)$ and of the environmental variance $\left(\widehat{\sigma^{2}_{e}}\right)$ are available (or of heritability $\widehat{h^{2}}$ and the phenotypic variance $\widehat{\sigma^{2}_{y}}$) from an analysis ignoring marker information. Then, the variance of the background genetic effects (assuming additivity) could be estimated as $\widehat{\sigma^{2}_{\varepsilon}} = \widehat{\sigma^{2*}_{e}} - \widehat{\sigma^{2}_{e}} = \widehat{\sigma^{2*}_{e}} - (1 - \widehat{h^{2}})\widehat{\sigma^{2}_{y}}$ (hoping that this value will be positive). Then, form

$$\widehat{\tau}_{\gamma} = \frac{(1 - \widehat{h^{2}})\widehat{\sigma^{2}_{y}}}{\widehat{\sigma^{2}_{\gamma}}}; \quad \widehat{\tau}_{\varepsilon} = \frac{(1 - \widehat{h^{2}})\widehat{\sigma^{2}_{y}}}{\widehat{\sigma^{2}_{\varepsilon}}},$$

where $\widehat{\sigma^{2}_{\gamma}}$ is the ML or REML estimate of σ^{2}_{γ}, and proceed with the BLUP implementation in (17). The resulting predictor is an empirical BLUP having properties that depend largely on the accuracy of the variance component estimates and on the adequacy of model in (7) and (9).

Differences with ridge regression: A main difference with the ridge regression approach of WHITTAKER *et al.* (2000) and WHITTAKER (2001) is how the value of τ_{γ} (or of $\tau^{*}_{\gamma} = \sigma^{2*}_{e}/\sigma^{2}_{\gamma}$) is obtained. Here, the formalism of the random-effects model (with a justification from a Bayesian viewpoint) is favored over an *ad hoc* procedure for doing shrinkage and tempering colinearity in a fixed-effects model, such as ridge regression. LINDLEY and SMITH (1972), ZELLNER and VANDAELE (1975), and Box (1980) discuss ridge regression from a Bayesian perspective. Further, a random-effects treatment of a fixed-effects model can lead to estimators and predictors with a superior frequentist performance (*e.g.*, ZELLNER and VANDAELE 1975; GIANOLA 1990; WEIGEL *et al.* 1991). An advantage of a random-effects treatment is the flexibility to accommodate additional model features. For example, the ridge regression estimator of WHITTAKER (2001) implies that all marked effects in $\boldsymbol{\gamma}$ are independent (*a priori*). Evidence of coexpression of genes in at least the same chromosome (CARON *et al.* 2001) indicates that the assumption that marked sections or segments within a chromosome have independent effects is not always tenable. Adjacent QTL may coexpress more (that is, be less independent) than QTL further apart. Thus, some spatial covariance structure along the chromosome might be in order. Ridge regression overstates the prior precision of the model, by virtue of implicitly assuming independence. A possible consequence is that "more genotype is marked" than warranted. Also, there

may be chromosomes that are "hotbeds" of QTL, whereas other chromosomes may be arid, leading to between-chromosome variation. This heterogeneity can be accommodated by introducing chromosome effects in the model, with markers in different chromosomes having (possibly) distinct distributions. These extensions are dealt with in the following section.

EXTENDING THE HIERARCHY

Model without background genetic effects: For simplicity, assume first that the state of prior knowledge does not allow disentangling background genetic effects from environmental deviates. The first tier is then as in (20), so that

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{M}, \boldsymbol{\gamma}, \sigma^{2*}_{e} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\gamma}, \mathbf{I}\sigma^{2*}_{e}). \quad (22)$$

The second tier is given by the distribution of the marked genetic effects. Here, instead of assuming that $\boldsymbol{\gamma}|\sigma^{2}_{\gamma} \sim N(\mathbf{0}, \mathbf{I}\sigma^{2}_{\gamma})$, a "one-way layout" is adopted, to partition the variability of marked effects into between- and within-chromosome components. Arrange markers sequentially according to their order within each chromosome and assume a dense marker map. The model for the second tier is

$$\boldsymbol{\gamma} = \mathbf{T}\mathbf{c} + \mathbf{v}, \quad (23)$$

where $\mathbf{c} = \{c_i\}$ is an $n_c \times 1$ vector of "marked chromosomal effects" (n_c is the number of pairs of chromosomes), $\mathbf{T}$ is a known incidence matrix of order $l \times n_c$ relating marked effects to chromosomes, and $\mathbf{v} = \{v_{ij}\}$ is an $l \times 1$ vector of marked "within-chromosome" deviates. In scalar notation, $\gamma_{ij} = c_i + v_{ij}$ is the "marked quantitative effect of the jth marker at the ith chromosome." It is assumed that

$$\boldsymbol{\gamma}|\mathbf{c}, \mathbf{V} \sim N(\mathbf{T}\mathbf{c}, \mathbf{V}). \quad (24)$$

Here, $\mathbf{V} = \text{Diag}(\mathbf{V}_i)$ is taken to be a block-diagonal matrix with n_c blocks, where $\mathbf{V}_i$ is the $l_i \times l_i$ variance-covariance matrix of marked within-chromosome deviates in chromosome i, and l_i is the number of markers in chromosome i. The block diagonality of $\mathbf{V}$ in (24) implies that marked within-chromosome effects are independent across chromosomes. However, a within-chromosome dependence will be introduced. The matrix $\mathbf{V}_i$ can be structured in several different manners, and some forms of modeling such dependence are discussed below. In (23) and (24) the $n_c \times 1$ vector of chromosome effects is assumed to possess the distribution

$$\mathbf{c}|\sigma^{2}_{c} \sim N(\mathbf{0}, \mathbf{I}\sigma^{2}_{c}), \quad (25)$$

where σ^{2}_{c} is the variance between chromosome effects. Some of the possible forms of $\mathbf{V}_i$ (among many possible ones) are considered next:

1. Within-chromosome deviates are correlated according

to a first-order autoregressive process. Partition $\mathbf{v} = [\mathbf{v}_1', \mathbf{v}_2', \ldots, \mathbf{v}_{n_c}']'$ with

$$\mathbf{v}_i = [v_{i1}, v_{i2}, \ldots, v_{il_i}]'.$$

Then, if markers are equally spaced,

$$\mathbf{V}_i = \mathrm{Var}(\mathbf{v}_i|\rho, \sigma_{v_i}^2) = \begin{bmatrix} 1 & \rho & \rho^2 & . & \rho^{l_i-1} \\ \rho & 1 & \rho & . & \rho^{l_i-2} \\ . & . & . & . & . \\ . & . & . & 1 & \rho \\ \rho^{l_i-1} & \rho^{l_i-2} & . & \rho & 1 \end{bmatrix} \sigma_{v_i}^2,$$

where $\sigma_{v_i}^2$ is the variance between deviates in chromosome i ($i = 1, 2, \ldots, n_c$) and ρ is a parameter taking values between -1 and 1. This covariance structure satisfies the assumption that adjacent within-chromosome deviates are more strongly correlated than those further apart. A form of relaxing the requirement that markers be equally spaced is immediately below.

2. Within-chromosome deviates are correlated according to a Gaussian decay model (*e.g.*, VERBEKE and MOLENBERGHS 1997). The covariance structure here is

$$\mathrm{Var}(\mathbf{v}_i|\rho_i, \sigma_{v_i}^2) = \begin{bmatrix} 1 & \lambda_{i,12} & \lambda_{i,13} & . & \lambda_{i,1(l_i-1)} \\ \lambda_{i,12} & 1 & \lambda_{23} & . & \lambda_{i,2(l_i-2)} \\ . & . & . & . & . \\ . & . & . & 1 & \lambda_{i,(l_i-1)l_i} \\ \lambda_{i,1(l_i-1)} & \lambda_{i,2(l_i-2)} & . & \lambda_{i,(l_i-1)l_i} & 1 \end{bmatrix} \sigma_{v_i}^2,$$

where

$$\lambda_{i,kt} = \exp\left(-\frac{d_{i,kt}^2}{\rho_i^2}\right)$$

is the correlation between marked within-chromosome deviates k and t in chromosome i. Here, $d_{i,kt}$ is the distance (*e.g.*, in physical units such as kilobases) between markers k and t in chromosome i, and $\rho_i > 0$ is a chromosome-specific parameter. When the distance between markers $\to 0$, $\lambda_{i,kt} \to 1$; on the other hand, when $d_{i,kt} \to \infty$, the correlation between within-chromosome deviates is null. The parameter ρ_i governs the rate at which the correlation decreases. When ρ_i is close to 0, such correlation falls rapidly; when ρ_i increases, the drop is more gentle. It may be convenient to assume that $\rho_i = \rho$ for all chromosomes.

3. Suppose that the number of markers is constant across chromosomes (with l markers and n_c chromosomes there would be $z = l/n_c$ markers per chromosome). If the markers are equally spaced and if the variance of within-chromosome deviates is constant across chromosomes, one may pose the Toeplitz covariance specification

$$\mathrm{Var}(\mathbf{v}_i|\boldsymbol{\lambda}, \sigma_v^2) = \begin{bmatrix} 1 & \lambda_{12} & \lambda_{13} & . & \lambda_{1z} \\ \lambda_{12} & 1 & \lambda_{23} & . & \lambda_{2z} \\ . & . & . & . & . \\ . & . & . & 1 & \lambda_{(z-1)z} \\ \lambda_{1z} & \lambda_{2z} & . & \lambda_{(z-1)z} & 1 \end{bmatrix} \sigma_v^2$$

(VERBEKE and MOLENBERGHS 1997) for $i = 1, 2, \ldots, n_c$. Here, the correlation between within-chromosome deviates k and t is

$$\mathrm{Corr}(v_{ik}, v_{il}) = \lambda_{kt}.$$

Note that λ_{kt} here, even though a correlation coefficient, is not the same as $\lambda_{i,kt}$ in the preceding section. Using (23) in (22) the model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}(\mathbf{Tc} + \mathbf{v}) + \mathbf{e}^*$$
$$= \mathbf{X}\boldsymbol{\beta} + \mathbf{MTc} + \mathbf{Mv} + \mathbf{e}^*. \qquad (26)$$

Under Gaussian assumptions, the joint distribution of $\mathbf{c}$ and $\mathbf{v}$, given a vector or linearly independent error contrasts $\mathbf{w}$, is

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{v} \end{bmatrix} \middle| \sigma_c^2, \mathbf{V}, \sigma_e^{2*}, \mathbf{w}, \mathbf{M}, \mathbf{T} \sim N\left(\begin{bmatrix} \widehat{\widehat{\mathbf{c}}} \\ \widehat{\widehat{\mathbf{v}}} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{cc} & \mathbf{C}_{cv} \\ \mathbf{C}_{vc} & \mathbf{C}_{vv} \end{bmatrix} \right). \quad (27)$$

Here, $\widehat{\widehat{\mathbf{c}}} = \mathrm{BLUP}(\mathbf{c})$ and $\widehat{\widehat{\mathbf{v}}} = \mathrm{BLUP}(\mathbf{v})$ are computed as

$$\begin{bmatrix} \widehat{\widehat{\mathbf{c}}} \\ \widehat{\widehat{\mathbf{v}}} \end{bmatrix} = \begin{bmatrix} \mathbf{T'M'PMT} + \mathbf{I}(\sigma_e^{2*}/\sigma_c^2) & \mathbf{T'M'PM} \\ \mathbf{M'PMT} & \mathbf{M'PM} + \mathrm{Diag}(\mathbf{V}_i^{-1})\sigma_e^{2*} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{T'M'Py} \\ \mathbf{M'Py} \end{bmatrix}, \quad (28)$$

and

$$\begin{bmatrix} \mathbf{C}_{cc} & \mathbf{C}_{cv} \\ \mathbf{C}_{vc} & \mathbf{C}_{vv} \end{bmatrix} = \begin{bmatrix} \mathbf{T'M'PMT} + \mathbf{I}(\sigma_e^{2*}/\sigma_c^2) & \mathbf{T'M'PM} \\ \mathbf{M'PMT} & \mathbf{M'PM} + \mathrm{Diag}(\mathbf{V}_i^{-1})\sigma_e^{2*} \end{bmatrix}^{-1} \sigma_e^{2*}. \quad (29)$$

The BLUP of the $\mathbf{M}\boldsymbol{\gamma}$ is $\widehat{\widehat{\mathbf{Mtc}}} + \widehat{\widehat{\mathbf{Mv}}}$. Since the predictor is conditionally unbiased (given $\mathbf{M}$), it is also unbiased unconditionally. The unconditional covariance matrix of the prediction errors is the average of (29) taken with respect to the distribution of marker genotypes.

Here, it is possible to predict marked effects on a chromosomal basis. Consider model (26) and recall that $\boldsymbol{\gamma} = \mathbf{Tc} + \mathbf{v}$. Suppose that a hypothetical species has four pairs of chromosomes. For the n individuals assayed, the incidence matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \begin{bmatrix} \mathbf{m}_{11}' & \mathbf{m}_{12}' & \mathbf{m}_{13}' & \mathbf{m}_{14}' \\ \mathbf{m}_{21}' & \mathbf{m}_{22}' & \mathbf{m}_{23}' & \mathbf{m}_{24}' \\ . & . & . & . \\ . & . & . & . \\ \mathbf{m}_{n1}' & \mathbf{m}_{n2}' & \mathbf{m}_{n3}' & \mathbf{m}_{n4}' \end{bmatrix},$$

where $\mathbf{m}'_{ij}$ are the codes for the marker genotypes of individual i at chromosome j. Likewise, the "true" marked effects on a chromosome-by-chromosome basis can be written as

$$\begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \boldsymbol{\gamma}_3 \\ \boldsymbol{\gamma}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{l_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{l_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{l_3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{l_4} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \end{bmatrix},$$

where c_j is an effect peculiar to all markers in chromosome pair j and $\mathbf{1}_{l_j}$ is a vector of 1's of order l_j. Then, after solving (28), one has the following relationships: (a) $\text{BLUP}(c_j) = \widehat{\widehat{c}}_j$; (b) $\text{BLUP}(\boldsymbol{\gamma}_j) = \mathbf{1}_{l_j}\widehat{\widehat{c}}_j + \widehat{\widehat{\mathbf{v}}}_j$; and (c) $\text{BLUP}(\mathbf{m}'_i\boldsymbol{\gamma}) = \Sigma_{j=1}^4 \mathbf{m}'_{ij} \text{ BLUP}(\boldsymbol{\gamma}_j) = \Sigma_{j=1}^4 \mathbf{m}'_{ij}\mathbf{1}_{l_j}\widehat{\widehat{c}}_j + \Sigma_{j=1}^4 \mathbf{m}'_{ij}\widehat{\widehat{\mathbf{v}}}_j$.

For many reasonable covariance structures, all dispersion parameters are identifiable in this model and can be estimated by tailoring some suitable algorithm for maximum likelihood. VERBEKE and MOLENBERGHS (1997, 2000) present examples of applications of standard statistical software for restricted maximum-likelihood estimation.

Model with background genetic effects: The model is now an expanded version of (26):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{MTc} + \mathbf{Mv} + \boldsymbol{\varepsilon} + \mathbf{e}. \tag{30}$$

The best linear unbiased estimator of $\boldsymbol{\beta}$ and the BLUP of $\mathbf{c}$, $\mathbf{v}$, and $\boldsymbol{\varepsilon}$ can be obtained by solving the system:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{MT} & \mathbf{X}'\mathbf{M} & \mathbf{X}' \\ \mathbf{T}'\mathbf{M}'\mathbf{X} & \mathbf{T}'\mathbf{M}'\mathbf{MT} + \mathbf{I}(\sigma_e^2/\sigma_c^2) & \mathbf{T}'\mathbf{M}'\mathbf{M} & \mathbf{T}'\mathbf{M}' \\ \mathbf{M}'\mathbf{X} & \mathbf{M}'\mathbf{MT} & \mathbf{M}'\mathbf{M} + \text{Diag}(\mathbf{V}_i^{-1})\sigma_e^2 & \mathbf{M}' \\ \mathbf{X} & \mathbf{MT} & \mathbf{M} & \mathbf{I} + \mathbf{I}(\sigma_e^2/\sigma_\varepsilon^2) \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\widehat{\mathbf{c}}} \\ \widehat{\widehat{\mathbf{v}}} \\ \widehat{\widehat{\boldsymbol{\varepsilon}}} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{T}'\mathbf{M}'\mathbf{y} \\ \mathbf{M}'\mathbf{y} \\ \mathbf{y} \end{bmatrix}. \tag{31}$$

As before, if an estimate of the total genetic variance is available, one can set σ_ε^2 as $\widehat{\sigma}_\varepsilon^2 = \widehat{\sigma}_\varepsilon^{2*} - (1 - \widehat{h}^2)\widehat{\sigma}_y^2$. Some estimates (*e.g.*, REML) of σ_c^2 and of the parameters defining the structure of $\mathbf{V}_i$ can be obtained when estimating $\widehat{\sigma}_e^{2*}$ from the model where background genetic effects are ignored or from the full model but fixing the value of σ_ε^2. The BLUP of the "total" genetic value of the entire collection of individuals in the sample is then

$$\text{BLUP}(\mathbf{MTc} + \mathbf{Mv} + \boldsymbol{\varepsilon}) = \mathbf{MT}\widehat{\widehat{\mathbf{c}}} + \mathbf{M}\widehat{\widehat{\mathbf{v}}} + \widehat{\widehat{\boldsymbol{\varepsilon}}}, \tag{32}$$

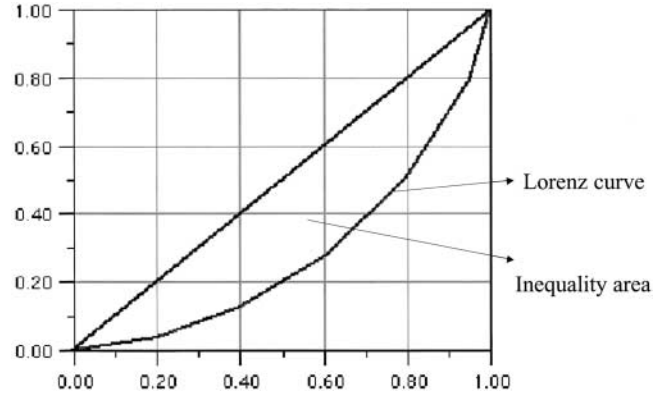and the covariance matrix of the prediction errors is

FIGURE 1.—Lorenz curve depicting the relative contribution of marked effects to marked genetic variability. Ordinate, cumulative proportion of marked genetic variance explained by marked effects; abscissa, cumulative proportion of markers.

$$\text{Var}(\mathbf{MTc} + \mathbf{Mv} + \boldsymbol{\varepsilon} - \mathbf{MT}\widehat{\widehat{\mathbf{c}}} - \mathbf{M}\widehat{\widehat{\mathbf{v}}} - \widehat{\widehat{\boldsymbol{\varepsilon}}})$$

$$= [\mathbf{MT} \quad \mathbf{M} \quad \mathbf{I}]$$

$$\times \begin{bmatrix} \mathbf{T}'\mathbf{M}'\mathbf{PMT} + \mathbf{I}(\sigma_e^2/\sigma_c^2) & \mathbf{T}'\mathbf{M}'\mathbf{PM} & \mathbf{T}'\mathbf{M}'\mathbf{P} \\ \mathbf{M}'\mathbf{PMT} & \mathbf{M}'\mathbf{PM} + \text{Diag}(\mathbf{V}_i^{-1})\sigma_e^2 & \mathbf{M}'\mathbf{P} \\ \mathbf{PMT} & \mathbf{PM} & \mathbf{P} + \mathbf{I}(\sigma_e^2/\sigma_\varepsilon^2) \end{bmatrix}$$

$$\times \begin{bmatrix} \mathbf{T}'\mathbf{M}' \\ \mathbf{M}' \\ \mathbf{I} \end{bmatrix}. \tag{33}$$

Differential contribution of marked effects to genetic variability: A representation of the degree of inequality of a frequency distribution is given by the Lorenz curve (*e.g.*, GASTWIRTH 1971, 1972). This is used for measuring income inequality in economics and was applied first in quantitative genetics by URIOSTE *et al.* (2001) in an assessment of heterogeneity of variance of milk yield in cattle among herds. We place the cumulative proportion of the markers $1/l, 2/l, \ldots, i/l, \ldots, (l-1)/l$ in the abscissa and plot this against the cumulative proportion of a measure of the marked "genetic variability." A hypothetical example is in Figure 1. A straight line with a slope of $45°$ indicates perfect equality. For example, if 20% of the markers explain 20% of the variability, 40% of the markers account for 40%, and so on, a straight line results. On the other hand, if a few markers account for a large portion of the variability, the curve bends downward because the distribution of effects is unequal. A measure of the inequality of contributions to variability is the Gini coefficient or twice the "inequality area," that is, the area between the Lorenz curve and the straight line. Since the square in Figure 1 has an area equal to 1, the maximum value of the inequality area is 0.5. Hence, the Gini coefficient takes values between 0 (all marked effects contributing equally to variability) and 1 (a single marked effect explains all variability).

Let $\mathbf{M}_{\text{unique}}$ be a set of rows of $\mathbf{M}$ defining u unique marker genotypes (u may be much smaller than n), $\mathbf{m}'_{\text{unique},j}$ be the jth row of $\mathbf{M}_{\text{unique}}$, and $m_{\text{unique},jk}$ be the entry for marker locus k in $\mathbf{m}'_{\text{unique},j}$. Define the corresponding "unique marked effect" as

$$\mu_j = \sum_{k=1}^{l} m_{\text{unique},jk}\gamma_k$$
$$= \mathbf{m}'_{\text{unique},j}\gamma$$
$$= \mathbf{m}'_{\text{unique},j}(\mathbf{Tc} + \mathbf{v}); \quad j = 1, 2, \ldots, u. \quad (34)$$

Arbitrarily, define "total marked genetic variability" as the sum of the squared unique marked genetic effects, so

$$S(\mathbf{c}, \mathbf{v}) = \sum_{j=1}^{u} \mu_j^2. \quad (35)$$

This definition does not take into account the frequency of the unique marker genotypes in the population, but this can be accounted for easily by weighting each μ_j appropriately. Let now $\mu_{[j]}^2$ be the jth ordered value of μ_j^2, sorted in an increasing order. The ordinate values in the Lorenz curve, $i.e.$, the cumulative proportion of the "total marked genetic variation," are calculated as

$$L\left(\frac{i}{u}\right) = \frac{\sum_{j=1}^{i}\mu_{[j]}^2}{S(\mathbf{c}, \mathbf{v})} = \frac{\sum_{j=1}^{i}\mu_{[j]}^2}{\sum_{j=1}^{u}\mu_j^2}, \quad (36)$$

where i/u is the cumulative proportion of unique marker genotypes. The Lorenz curve results from plotting $L(i/u)$ against i/u; observe that $L(0) = 0$ and $L(1) = 1$. It can be shown (AGUILAR-GUTIERREZ 2000; URIOSTE $et~al.$ 2001) that the Gini coefficient is approximately equal to

$$G = \frac{2}{(u\sum_{j=1}^{u}\mu_j^2)}\sum_{j=1}^{u}\left(j - \frac{1}{2}\right)\mu_{[j]}^2. \quad (37)$$

A main difficulty in estimating the Lorenz curve and the Gini coefficient is that nonlinear functions of unknown quantities are involved. Simple method of moment estimators $\widehat{\widehat{L}}(i/u)$ and $\widehat{\widehat{G}}$ are obtained by replacing μ_j^2 in (36) and (37) by

$$\widehat{\widehat{\mu}}_j^2 = \mathbf{m}'_{\text{unique},j}(\widehat{\widehat{\mathbf{Tc}}} + \widehat{\widehat{\mathbf{v}}})(\widehat{\widehat{\mathbf{Tc}}} + \widehat{\widehat{\mathbf{v}}})'\mathbf{m}_{\text{unique},j},$$

with $\widehat{\widehat{\mathbf{c}}}, \widehat{\widehat{\mathbf{r}}}$ obtained from (31). This statistic does not take into account the uncertainty associated with $\widehat{\widehat{\mathbf{Tc}}} + \widehat{\widehat{\mathbf{v}}}$.

Family heterogeneity: Suppose that the phenotypic and molecular marker information can be partitioned into F nonoverlapping or "independent" clusters representing some familial aggregation. For example, XU (1998) discussed the use of "families" of line crosses for QTL mapping. This author pointed out that single-family methods present the risk that if two lines involved in the cross are not segregating at a QTL, then the latter would not be detected, irrespective of the number of individuals scored in a backcross or F_2 population. He

allowed for multiple families in the model, but treated their effects as fixed. Irrespective of whether these effects are viewed as fixed or random, they constitute statistical nuisances from the point of view of QTL mapping (XU 1998). Here, a different point of view is taken. Allowance is made for heterogeneity of family effects on the marker-phenotype relationships. The stylized situation described in standard treatments of quantitative genetics is adopted; for example, individuals may be clustered into full-sib or half-sib families. We emphasize that our approach does not include recombination rates between markers and QTL as model parameters. A rigorous treatment for outbred populations requires extremely involved algebra ($e.g.$, WANG $et~al.$ 1998). The purpose here is to bypass these complex procedures via straightforward linear models having a much simpler covariance structure.

Let the vector of phenotypic records be partitioned as $\mathbf{y} = [\mathbf{y}'_1, \mathbf{y}'_2, \ldots, \mathbf{y}'_F]'$, where $\mathbf{y}_i$ contains f_i observations pertaining to family i ($i = 1, 2, \ldots, F$). Assume now that heterogeneity in the marker-phenotype association exists. In the presence of family-specific marked effects the entire vector γ can be written as

$$\gamma_{Fl\times 1} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \cdot \\ \cdot \\ \gamma_F \end{bmatrix},$$

where γ_i is an $l \times 1$ set of effects peculiar to family i. The model for phenotypic values, allowing now for family-specific background effects, ε_F, becomes

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \cdot \\ \cdot \\ \mathbf{y}_F \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \cdot \\ \cdot \\ \mathbf{X}_F \end{bmatrix}\beta + \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_2 & . & . & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & . & . & \mathbf{M}_F \end{bmatrix}\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \cdot \\ \cdot \\ \gamma_F \end{bmatrix}$$

$$+ \begin{bmatrix} \mathbf{1}_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_2 & . & . & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbf{0} & \mathbf{0} & . & . & \mathbf{1}_F \end{bmatrix}\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_F \end{bmatrix} + \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \cdot \\ \cdot \\ \mathbf{w}_F \end{bmatrix}$$

$$= \mathbf{X}\beta + \mathbf{M}\gamma + \mathbf{F}\varepsilon_f + \mathbf{w}, \quad (38)$$

where $\mathbf{X} = [\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_F]'$ and $\mathbf{X}_i$ has as many rows as there are observations for family i (f_i) and p columns. Further, $\mathbf{M} = \text{Diag}(\mathbf{M}_i)$. The incidence matrix $\mathbf{M}_i$ has f_i rows and l columns, the number of markers. The $F \times 1$ vector $\varepsilon_f = \{\varepsilon_i\}$ contains F family background effects

related to the phenotypic values via the known incidence matrix $\mathbf{F}$ (consisting of vectors of ones, $\mathbf{1}_i$, in appropriate locations and of zeroes elsewhere); $\mathbf{w}$ is an $n \times 1$ residual vector of independently distributed within-family deviations, with common variance σ_w^2 and where $n = \Sigma_{i=1}^F f_i$.

In the absence of marker information, the phenotypic variance can be partitioned into between- and within-family components. For example, with half-sib families, the between-family variance contains one-quarter of the additive genetic variance (supposing additive inheritance), and the within-family component includes the environmental variance plus three-quarters of the additive genetic variance. In the setting of (38) it is assumed that ε_f and $\mathbf{w}$ are independently distributed vectors, with distributions $\varepsilon_f | \sigma_{\varepsilon_f}^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_{\varepsilon_f}^2)$ and $\mathbf{w} | \sigma_w^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_w^2)$, respectively. This partitioning reassigns the unmarked additive genetic variance into between ($\sigma_{\varepsilon_f}^2$) and within-family (σ_w^2) components. Now, the γ_i marked effects can be treated as random regressions following the distribution

$$\gamma_i | \gamma_0, \mathbf{B} \sim N(\gamma_0, \mathbf{B}); \quad i = 1, 2, \ldots, F. \quad (39)$$

Here, γ_0 is an $l \times 1$ parameter that is common to all families (the "fixed" part of the regression) and $\mathbf{B}$ is the $l \times l$ variance-covariance matrix between the regression coefficients. This implies that $\gamma_i = \gamma_0 + \delta_i$, where $\delta_i \sim N(\mathbf{0}, \mathbf{B})$ is a vector of deviations from the common regression γ_0 that is specific to family i. In a simpler hierarchical model it could be postulated, for example, that $\gamma_i | \sigma_\gamma^2 \sim N(\mathbf{0}, \mathbf{I}_i \sigma_\gamma^2)$ for $i = 1, 2, \ldots, F$. The model for the more general specification becomes

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ . \\ . \\ \mathbf{y}_F \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{M}_1 \\ \mathbf{X}_2 & \mathbf{M}_2 \\ . & . \\ . & . \\ \mathbf{X}_F & \mathbf{M}_F \end{bmatrix} \begin{bmatrix} \beta \\ \gamma_0 \end{bmatrix} + \mathbf{M}\delta + \mathbf{F}\varepsilon_f + \mathbf{w}$$

$$= \mathbf{X}^*\beta^* + \mathbf{M}\delta + \mathbf{F}\varepsilon_f + \mathbf{w}, \quad (40)$$

where $\beta^{*\prime} = [\beta', \gamma_0']$, $\mathbf{X}^* = [\mathbf{X} \ \mathbf{M}_0]$, $\mathbf{M}_0 = [\mathbf{M}_1' \ \mathbf{M}_2' \cdot \mathbf{M}_F']'$, and $\delta' = [\delta_1, \delta_2', \ldots, \delta_F]$ is a vector of family-specific deviations from the overall regression γ in (39).

Given the dispersion parameters $\mathbf{B}$, $\sigma_{\varepsilon_f}^2$, and σ_w^2, the best linear unbiased estimator of β^* and the BLUP of each of the family-specific deviations δ_i are found by solving the system

$$\begin{bmatrix} \mathbf{X}^{*\prime}\mathbf{X}^* & \mathbf{X}^{*\prime}\mathbf{M} & \mathbf{X}^{*\prime}\mathbf{F} \\ \mathbf{M}'\mathbf{X}^* & \oplus_{i=1}^F(\mathbf{M}_i'\mathbf{M}_i + \mathbf{B}^{-1}\sigma_w^2) & \mathbf{M}'\mathbf{F} \\ \mathbf{F}'\mathbf{X}^* & \mathbf{F}'\mathbf{M} & \mathbf{F}'\mathbf{F} + \mathbf{I}\frac{\sigma_w^2}{\sigma_{\varepsilon_f}^2} \end{bmatrix} \begin{bmatrix} \widehat{\widehat{\beta}}^* \\ \widehat{\widehat{\delta}} \\ \widehat{\widehat{\varepsilon}}_f \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{*\prime}\mathbf{y} \\ \mathbf{M}'\mathbf{y} \\ \mathbf{F}'\mathbf{y} \end{bmatrix}, \quad (41)$$

where $\oplus_{i=1}^F$ is the "direct sum" of matrices notation (*e.g.*, SEARLE 1982), denoting a block-diagonal concatenation

of the F matrices $\mathbf{M}_i'\mathbf{M}_i + \mathbf{B}^{-1}\sigma_w^2$, with all these matrices being $l \times l$. Note that

$$\mathbf{X}^{*\prime}\mathbf{X}^* = \sum_{i=1}^F \begin{bmatrix} \mathbf{X}_i'\mathbf{X}_i & \mathbf{X}_i'\mathbf{M}_i \\ \mathbf{M}_i'\mathbf{X}_i & \mathbf{M}_i'\mathbf{M}_i \end{bmatrix} = \sum_{i=1}^F \mathbf{X}_i^{*\prime}\mathbf{X}_i^*,$$

$$\mathbf{X}^{*\prime}\mathbf{M} = [\mathbf{X}_1^{*\prime}\mathbf{M}_1 \quad \mathbf{X}_2^{*\prime}\mathbf{M}_2 \quad \cdot \quad \mathbf{X}_F^{*\prime}\mathbf{M}_F],$$

$$\mathbf{X}^{*\prime}\mathbf{F} = [\mathbf{X}_1^{*\prime}\mathbf{1}_1 \quad \mathbf{X}_2^{*\prime}\mathbf{1}_2 \quad \cdot \quad \mathbf{X}_F'\mathbf{1}_F],$$

$$\mathbf{M}'\mathbf{F} = \oplus_{i=1}^F \mathbf{M}_i'\mathbf{1}_i,$$

$$\mathbf{F}'\mathbf{F} = \mathrm{Diag}(f_i),$$

$$\mathbf{X}^{*\prime}\mathbf{y} = \sum_{i=1}^F \mathbf{X}_i^{*\prime}\mathbf{y}_i,$$

$$\mathbf{M}'\mathbf{y} = [\mathbf{M}_1'\mathbf{y}_1 \quad \mathbf{M}_2'\mathbf{y}_2 \ldots \mathbf{M}_F'\mathbf{y}_F]',$$

and

$$\mathbf{F}'\mathbf{y} = [\mathbf{1}_1'\mathbf{y}_1 \quad \mathbf{1}_2'\mathbf{y}_2 \quad \cdot \quad \mathbf{1}_F'\mathbf{y}_F]'.$$

The variance-covariance matrix of the prediction errors is given by the inverse of the coefficient matrix in (41) multiplied by σ_w^2. Unless there are many families (*e.g.*, independent half-sib groups in dairy cattle) it will be difficult to obtain reliable maximum-likelihood estimates of $\mathbf{B}$.

The hierarchy can be extended as in (23) and (24) to include chromosome and within-chromosome effects that would now be family specific. Put

$$\begin{bmatrix} \delta_1 \\ \delta_2 \\ . \\ \delta_F \end{bmatrix} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 & . & . & \mathbf{0} \\ . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & . & \mathbf{T}_F \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ . \\ \mathbf{c}_F \end{bmatrix} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ . \\ \mathbf{v}_F \end{bmatrix}, \quad (42)$$

where δ_i is $l \times 1$, $\mathbf{c}_i$ is an $n_c \times 1$ vector of chromosome effects peculiar to family i, $\mathbf{T}_i$ is an $l \times n_c$ incidence matrix, and $\mathbf{v}_i$ is an $l \times 1$ vector of within-chromosome effects for family i. It is convenient to assume that the within-chromosome marked effects are independent across families and chromosomes, but possibly dependent within chromosomes. Here, the variance-covariance matrix $\mathbf{V}$ in (24) would be $Fl \times Fl$ and would take the form

$$\mathbf{V} = \mathrm{Diag}(\mathbf{V}_i), \quad i = 1, 2, \ldots, F,$$

where $\mathbf{V}_i = \mathrm{Var}(\mathbf{v}_i)$ is an $l \times l$ variance-covariance matrix that is specific to family i. Now, partition

$$\mathbf{v}_i = [\mathbf{v}_{i1}', \mathbf{v}_{i2}', \ldots, \mathbf{v}_{in_c}']',$$

where $\mathbf{v}_{ij}$ is an l_j (number of markers in chromosome j) $\times 1$ vector of within-chromosome deviates for family i in chromosome j. Then, under the assumption of independence of deviates across chromosomes,

$$\mathbf{V}_i = \mathrm{Var}(\mathbf{v}_i) = \mathrm{Diag}[\mathrm{Var}(\mathbf{v}_{ij})] = \mathrm{Diag}(\mathbf{V}_{ij}).$$

With this,

$$\mathbf{V} = \oplus_{i=1, j=1}^{F, n_c} \mathbf{V}_{ij},$$

and $\mathbf{V}_{ij}$ can be assigned any of the spatial structures discussed earlier. Then, the distribution of all family-specific deviations from regression would be

$$\begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ . \\ . \\ \boldsymbol{\delta}_F \end{bmatrix} \Big| \mathbf{T}, \mathbf{c}, \mathbf{V} \sim N \left(\begin{bmatrix} \mathbf{T}_1 \mathbf{c}_1 \\ \mathbf{T}_2 \mathbf{c}_2 \\ . \\ . \\ \mathbf{T}_F \mathbf{c}_F \end{bmatrix}, \oplus_{i,j} \mathbf{V}_{ij} \right). \quad (43)$$

The hierarchy is completed by assigning a distribution to the family-specific chromosome effects $\mathbf{c}_i$. An extension of assumption (25) is

$$\begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ . \\ . \\ \mathbf{c}_F \end{bmatrix} \Big| \mathbf{K} \sim \mathbf{N} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{0} & . & \mathbf{0} \\ \mathbf{0} & \mathbf{K} & . & \mathbf{0} \\ . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & \mathbf{K} \end{bmatrix} \right), \quad (44)$$

where $\mathbf{K} = \{k_{ij}\}$ is an $n_c \times n_c$ matrix of covariances between chromosome effects of the same family, with this dispersion structure assumed homogeneous across families. For example, element k_{45} of $\mathbf{K}$ would be the covariance between the effects of chromosomes 4 and 5 within the same family. As a side note, observe that unconditionally to $\mathbf{c}_i$, (43) and (44) imply that $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \mathbf{T}_i \mathbf{K} \mathbf{T}_i' + \mathbf{V}_i)$. Hence, assumption (39) is equivalent to (43) and (44) if and only if $\mathbf{T}_i \mathbf{K} \mathbf{T}_i' + \mathbf{V}_i = \mathbf{B}$ for all i.

Using (42) in (40) leads to the following model for phenotypes,

$$\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{M} \mathbf{T} \mathbf{c} + \mathbf{M} \mathbf{v} + \mathbf{F} \boldsymbol{\varepsilon}_f + \mathbf{w}, \quad (45)$$

where $\mathbf{M} = \mathrm{Diag}(\mathbf{M}_i)$ and $\mathbf{T} = \mathrm{Diag}(\mathbf{T}_i)$. The BLUP of all family-specific chromosome and within-chromosome effects is obtained by solving

$$\begin{bmatrix} \mathbf{X}^{*\prime}\mathbf{X}^* & \mathbf{X}^{*\prime}\mathbf{MT} & \mathbf{X}^{*\prime}\mathbf{M} & \mathbf{X}^{*\prime}\mathbf{F} \\ \mathbf{T}'\mathbf{M}'\mathbf{X}^* & \mathbf{T}'\mathbf{M}'\mathbf{MT} + (\mathbf{I}_F \otimes \mathbf{K}^{-1})\sigma_w^2 & \mathbf{T}'\mathbf{M}'\mathbf{M} & \mathbf{T}'\mathbf{M}'\mathbf{F} \\ \mathbf{M}'\mathbf{X}^* & \mathbf{M}'\mathbf{MT} & \mathbf{M}'\mathbf{M} + \oplus_{i,j}\mathbf{V}_{ij}^{-1}\sigma_w^2 & \mathbf{M}'\mathbf{F} \\ \mathbf{F}'\mathbf{X}^* & \mathbf{F}'\mathbf{MT} & \mathbf{F}'\mathbf{M} & \mathbf{F}'\mathbf{F} + \mathbf{I}(\sigma_w^2/\sigma_\varepsilon^2) \end{bmatrix}$$

$$\times \begin{bmatrix} \widehat{\boldsymbol{\beta}^*} \\ \widehat{\widehat{\mathbf{c}}} \\ \widehat{\widehat{\mathbf{v}}} \\ \widehat{\widehat{\boldsymbol{\varepsilon}_f}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{*\prime}\mathbf{y} \\ \mathbf{T}'\mathbf{M}'\mathbf{y} \\ \mathbf{M}'\mathbf{y} \\ \mathbf{F}'\mathbf{y} \end{bmatrix}. \quad (46)$$

The variance-covariance matrix of the prediction errors is the inverse of the coefficient matrix in (46) times σ_w^2.

Lande-Thompson regression model revisited: LANDE and THOMPSON (1990) suggested splitting the molecular scores into between- and within-family components, so that the components could be combined "optimally" for the purpose of genetic improvement. In our context and for simplicity, consider the following variant of their model, using (38). Write the marker codes for family i as $\mathbf{M}_i = \overline{\mathbf{M}}_i + (\mathbf{M}_i - \overline{\mathbf{M}}_i)$, where $\overline{\mathbf{M}}_i$ is an $f_i \times l$ matrix with columns equal to the mean values of the corresponding columns of $\mathbf{M}_i$, with all its rows being equal. A model of interest might be

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ . \\ . \\ \mathbf{y}_F \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ . \\ . \\ \mathbf{X}_F \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \overline{\mathbf{M}}_1 \\ \overline{\mathbf{M}}_2 \\ . \\ . \\ \overline{\mathbf{M}}_F \end{bmatrix} \boldsymbol{\alpha}_B + \begin{bmatrix} \mathbf{M}_1 - \overline{\mathbf{M}}_1 \\ \mathbf{M}_2 - \overline{\mathbf{M}}_2 \\ . \\ . \\ \mathbf{M}_F - \overline{\mathbf{M}}_F \end{bmatrix} \boldsymbol{\alpha}_W$$

$$+ \begin{bmatrix} \mathbf{1}_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_2 & . & . & \mathbf{0} \\ . & . & . & . & . \\ . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & . & \mathbf{1}_F \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ . \\ . \\ \boldsymbol{\varepsilon}_F \end{bmatrix} + \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ . \\ . \\ \mathbf{w}_F \end{bmatrix}$$

$$= \mathbf{X}\boldsymbol{\beta} + \overline{\mathbf{M}}\boldsymbol{\alpha}_B + \mathbf{M}_W \boldsymbol{\alpha}_W + \mathbf{F}\boldsymbol{\varepsilon} + \mathbf{w}, \quad (47)$$

where $\boldsymbol{\alpha}_B$ is the regression of phenotypes on mean family molecular scores, $\mathbf{M}_W$ is a matrix of within-family deviations from the mean scores for the appropriate families, and $\boldsymbol{\alpha}_W$ is the vector of regressions of phenotypes on the within-family deviations. Assume that $\boldsymbol{\alpha}_B \sim N(0, \mathbf{I}_F \sigma_{\alpha F}^2)$ and $\boldsymbol{\alpha}_W \sim N(0, \mathbf{I}_F \sigma_{\alpha W}^2)$ where $\sigma_{\alpha F}^2$ and $\sigma_{\alpha W}^2$ are components of variance, and suppose that the two random vectors are independently distributed. The best linear unbiased predictor of the between- and within-family regressions can be calculated by solving

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\overline{\mathbf{M}} & \mathbf{X}'\mathbf{M}_W & \mathbf{X}'\mathbf{F} \\ \overline{\mathbf{M}}'\mathbf{X} & \overline{\mathbf{M}}'\overline{\mathbf{M}} + \mathbf{I}(\sigma_w^2/\sigma_{\alpha F}^2) & \overline{\mathbf{M}}'\mathbf{M}_W & \overline{\mathbf{M}}'\mathbf{F} \\ \mathbf{M}_W'\mathbf{X} & \mathbf{M}_W'\overline{\mathbf{M}} & \mathbf{M}_W'\mathbf{M}_W + \mathbf{I}(\sigma_c^2/\sigma_{\alpha W}^2) & \mathbf{M}_W'\mathbf{F} \\ \mathbf{F}'\mathbf{X} & \mathbf{F}'\overline{\mathbf{M}} & \mathbf{F}'\mathbf{M}_W & \mathbf{F}'\mathbf{F} + \mathbf{I}(\sigma_w^2/\sigma_\varepsilon^2) \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\widehat{\boldsymbol{\alpha}_b}} \\ \widehat{\widehat{\boldsymbol{\alpha}_W}} \\ \widehat{\widehat{\boldsymbol{\varepsilon}}} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \overline{\mathbf{M}}'\mathbf{y} \\ \mathbf{M}_W'\mathbf{y} \\ \mathbf{F}'\mathbf{y} \end{bmatrix}. \quad (48)$$

The variance-covariance matrix of the prediction errors can be obtained from the inverse of the coefficient matrix in (48), times σ_w^2. Model (47) can be extended by allowing the within-family regressions to be family-specific. In this extended model, in addition to $\boldsymbol{\alpha}_B$, there would be $\boldsymbol{\alpha}_{Wi}$ ($i = 1, 2, \ldots, F$) within-family regressions.

A BAYESIAN FORMULATION

We adopt now a Bayesian point of view and describe a Markov chain Monte Carlo (MCMC) implementation. The focus is on model (40) and its hierarchical expansion in (42–44), leading to (45) and (46) in the mixed model and BLUP treatment. This is the most general and richly parameterized specification among those discussed so far. The developments follow the typical hierarchical structure of Bayesian multilevel models (LINDLEY and SMITH 1972).

Joint posterior distribution: Start with (40) as sampling model (data-generating process) or first level of the Bayesian hierarchy. Conditionally on $\boldsymbol{\beta}^*$, $\boldsymbol{\delta}$, and $\boldsymbol{\varepsilon}_f$, it is assumed that the phenotypic values are drawn independently from the distribution

$$\mathbf{y}|\boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\varepsilon}_f, \sigma_{\mathrm{w}}^2 \sim N(\mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{M}\boldsymbol{\delta} + \mathbf{F}\boldsymbol{\varepsilon}_f, \mathbf{I}\sigma_{\mathrm{w}}^2). \qquad (49)$$

It is convenient to express the joint density of all phenotypes, given $\boldsymbol{\beta}^*$, $\boldsymbol{\delta}$, $\boldsymbol{\varepsilon}_f$, and σ_{w}^2 as the product of F densities corresponding to the contributions to information made by the phenotypic values in each of the families. Thus,

$$p(\mathbf{y}|\boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\varepsilon}_f, \sigma_{\mathrm{w}}^2) = \prod_{i=1}^{F} N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i, \mathbf{I}\sigma_{\mathrm{w}}^2). \qquad (50)$$

The notation $N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i, \mathbf{I}\sigma_{\mathrm{w}}^2)$ indicates a normal density or distribution with $\mathbf{y}_i$ as random variable, $\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i$ as mean vector, and $\mathbf{I}\sigma_{\mathrm{w}}^2$ as covariance matrix; a similar notation is used if the distribution involves a scalar.

The second level assigns a prior distribution to all unknown parameters of the first tier. It is assumed that the joint prior density can be written as

$$p(\boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\varepsilon}_f, \sigma_{\mathrm{w}}^2|\mathbf{c}, \mathbf{V}(\boldsymbol{\phi}), \sigma_{\varepsilon_f}^2) = N(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0^*, \mathbf{I}s_{\boldsymbol{\beta}}^2*)N(\boldsymbol{\delta}|\mathbf{Tc}, \mathbf{V}(\boldsymbol{\phi}))$$
$$\times N(\boldsymbol{\varepsilon}_f|\mathbf{0}, \mathbf{I}\sigma_{\varepsilon_f}^2)p(\sigma_{\mathrm{w}}^2|s_{\mathrm{w}}^2, \nu_{\mathrm{w}}). \qquad (51)$$

Above, $\boldsymbol{\beta}_0^*$ is the known mean of the prior distribution of $\boldsymbol{\beta}^*$ and $s_{\boldsymbol{\beta}}^2*$ is a known scalar that tunes the degree of vagueness of the prior for this parameter. Different degrees of vagueness may be assigned to the two components of $\boldsymbol{\beta}^*$; thus, there would be two distinct tuning parameters, $s_{\boldsymbol{\beta}}^2$ and $s_{\boldsymbol{\gamma}_0}^2$. The notation $\mathbf{V}(\boldsymbol{\phi})$ means that the variance-covariance matrix of $\mathbf{v}$ (the within-chromosome deviates) depends on some parameter vector $\boldsymbol{\phi}$. For example, if within-chromosome deviates are correlated according to an autoregressive process having ρ as parameter and with chromosome-specific variances $\sigma_{v_1}^2, \sigma_{v_2}^2, \ldots, \sigma_{v_{n_c}}^2$, then $\boldsymbol{\phi} = [\rho, \sigma_{v_1}^2, \sigma_{v_2}^2, \ldots, \sigma_{v_{n_c}}^2]'$. The autoregressive process is assumed hereafter, to illustrate one of the possible specifications. In (51), s_{w}^2 and ν_{w} are known parameter values (hyperparameters) of a scaled inverted chi-square distribution (*e.g.*, GELMAN *et al.* 1995) assigned to σ_{w}^2. Now, as stated in the developments following (42) and leading to (43), given the chromosome effects $\mathbf{c}$, the marked effects $\boldsymbol{\delta}$ of different families are taken to be mutually independent. Hence, (51) can be put as

$$p(\boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\varepsilon}_f, \sigma_{\mathrm{w}}^2|\mathbf{c}, \mathbf{V}(\boldsymbol{\phi}), \sigma_{\varepsilon_f}^2) = N(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0^*, \mathbf{I}s_{\boldsymbol{\beta}}^2*)$$
$$\times \prod_{i=1}^{F}[N(\boldsymbol{\delta}_i|\mathbf{T}_i\mathbf{c}_i, \mathbf{V}_i(\boldsymbol{\phi}))\times N(\varepsilon_i|0, \sigma_{\varepsilon_f}^2)]p(\sigma_{\mathrm{w}}^2|s_{\mathrm{w}}^2, \nu_{\mathrm{w}}). \qquad (52)$$

The third level of the hierarchy consists of the prior distribution assigned to $\mathbf{c}$, $\boldsymbol{\phi}$, and $\sigma_{\varepsilon_f}^2$, the parameters of the second tier. It is assumed that the corresponding density takes the form

$$p(\mathbf{c}, \boldsymbol{\phi}, \sigma_{\varepsilon_f}^2|\mathbf{K}, \boldsymbol{\eta}, s_v^2, \nu_v, s_f^2, \nu_f) = \left[\prod_{i=1}^{F} N(\mathbf{c}_i|\mathbf{0}, \mathbf{K})\right]$$
$$\times p(\rho, \sigma_{v_1}^2, \sigma_{v_2}^2, \ldots, \sigma_{v_{n_c}}^2|\boldsymbol{\eta}, s_v^2, \nu_v)p(\sigma_{\varepsilon_f}^2|s_f^2, \nu_f)$$
$$= \left[\prod_{i=1}^{F} N(\mathbf{c}_i|\mathbf{0}, \mathbf{K})\right]$$
$$\times p(\rho|\boldsymbol{\eta})[\prod_{i=1}^{n_c} p(\sigma_{v_i}^2|s_v^2, \nu_v)]p(\sigma_{\varepsilon_f}^2|s_f^2, \nu_f). \qquad (53)$$

In the preceding, it is assumed that (a) the chromosome effects of different families are mutually independent, although correlated within a family with an $n_c \times n_c$ covariance matrix $\mathbf{K}$, as in (44); (b) the parameter ρ has a prior distribution indexed by some known parameter vector $\boldsymbol{\eta}$ and is independently distributed (*a priori*) of all within-chromosome variances $\sigma_{v_i}^2$; (c) the n_c variances $\sigma_{v_i}^2$ are independently and identically distributed as scaled inverted chi square with known hyperparameters s_v^2 and ν_v; and (d) the variance between families $\sigma_{\varepsilon_f}^2$ follows a scaled inverted chi-square distribution with known parameters s_f^2 and ν_f.

The fourth and final level of the hierarchical model is the prior distribution assigned to $\mathbf{K}$, the covariance matrix between effects of the same family on different chromosomes. It is assumed that $\mathbf{K}$ follows an inverted Wishart process of order n_c (the order of $\mathbf{K}$) with density $p(\mathbf{K}|\nu_c\mathbf{S}_c, \nu_c)$, where $\nu_c\mathbf{S}_c$ is a known scale matrix and ν_c is a known positive parameter usually referred to as the "degrees of freedom" of the distribution (*e.g.*, GELMAN *et al.* 1995). Our parameterization is such that, *a priori*, $E(\mathbf{K}|\nu_c\mathbf{S}_c, \nu_c) = \nu_c\mathbf{S}_c/(\nu_c - n_c - 1)$. For a very large value of ν_c, $\mathbf{S}_c$ then approximates the mean value of the prior distribution. In less structured models, *e.g.*, independence of chromosome effects within a family, $\mathbf{K}$ would take a simpler form, in which case the prior distribution would be modified appropriately.

Some of the prior distributions given above are difficult or impossible to elicit. In such a situation, one may consider using some of the standard default improper priors, carry out a technically involved reference prior analysis (BERNARDO 1979; BERNARDO and SMITH 1994), or resort to maximum entropy fits; see CANTET *et al.* (1992) and SORENSEN and GIANOLA (2002) for an elementary discussion of the latter in a genetics context. Collecting (50), (52), (53), $p(\mathbf{K}|\nu_c\mathbf{S}_c, \nu_c)$ and letting $\mathbf{H}$ denote the set of all known hyperparameters, the joint posterior density of the uncertain parameters is, after rearrangement,

$$p(\boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\varepsilon}_f, \mathbf{c}, \sigma_{\mathrm{w}}^2, \rho, \sigma_{v_1}^2, \sigma_{v_2}^2, \ldots, \sigma_{v_{n_c}}^2, \sigma_{\varepsilon_f}^2, \mathbf{K}|\mathbf{y}, \mathbf{H})$$
$$\propto \left[\prod_{i=1}^{F} N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i, \mathbf{I}\sigma_{\mathrm{w}}^2)N(\boldsymbol{\delta}_i|\mathbf{T}_i\mathbf{c}_i, \mathbf{V}_i(\boldsymbol{\phi}))N(\varepsilon_i|0, \sigma_{\varepsilon_f}^2)\right]$$
$$\times N(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0^*, \mathbf{I}s_{\boldsymbol{\beta}}^2*)\left[\prod_{i=1}^{F} N(\mathbf{c}_i|\mathbf{0}, \mathbf{K})\right]\left[\prod_{i=1}^{n_c} p(\sigma_{v_i}^2|s_v^2, \nu_v)\right]$$
$$\times p(\sigma_{\mathrm{w}}^2|s_{\mathrm{w}}^2, \nu_{\mathrm{w}})p(\sigma_{\varepsilon_f}^2|s_f^2, \nu_f)p(\rho|\boldsymbol{\eta})p(\mathbf{K}|\mathbf{S}_c, \nu_c). \qquad (54)$$

Markov chain Monte Carlo scheme: At least in theory, all marginal or joint distributions of sets of parameters of interest can be derived from (54) by effecting the necessary integrations, to take uncertainty about nuisance parameters into account properly. Since the joint posterior distribution is not in a form amenable to analytical treatment, MCMC (*e.g.*, Gilks *et al.* 1996) can be used for simulating draws from the posterior. Features of any target posterior distribution can be estimated from the simulated samples. Monte Carlo error can be made negligible by sheer computing force; however, suitable reparameterizations may enhance the efficiency of the simulation.

The hybrid algorithm proposed uses the conditional posterior distributions as candidate-generating densities, whenever possible. Save for the presence of $p(\rho|\boldsymbol{\eta})$ (whose structure has not been specified yet), the joint posterior process with density (54) is in a normal-inverse Gamma or normal-inverse Wishart form. All fully conditional distributions (except that of ρ) can be identified and sampled with relative ease. This defines a Gibbs sampler for drawing from most posterior distributions of interest. Since the conditional process is not recognizable for ρ, a Metropolis procedure is tailored. The most relevant expressions needed for implementing this hybrid algorithm are presented below. The notation [parameter or parameters|ELSE] is used throughout to denote a fully conditional posterior distribution, that is, the posterior distribution of a scalar or vector parameter given all other parameters, the data, and **H**. For details, see Wang *et al.* (1993, 1994), Sorensen *et al.* (1994), and Sorensen and Gianola (2002).

Conditional distribution of $\boldsymbol{\beta}^*$: This is arrived at by retaining in (54) only the terms involving $\boldsymbol{\beta}^*$. The resulting conditional density is

$$p(\boldsymbol{\beta}^*|\text{ELSE}) \propto \prod_{i=1}^{F} N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i, \mathbf{I}\sigma_w^2) N(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0^*, \mathbf{I}s_{\boldsymbol{\beta}}^2{}^*).$$

This can be identified as the density of the multivariate Gaussian distribution,

$$\boldsymbol{\beta}^*|\text{ELSE} \sim N(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{C}}_{\boldsymbol{\beta\beta}}), \qquad (55)$$

where

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{I}\frac{\sigma_w^2}{s_{\boldsymbol{\beta}}^2*}\right]^{-1}\left[\mathbf{X}^*(\mathbf{y} - \mathbf{M}\boldsymbol{\gamma} - \mathbf{F}\varepsilon_f) + \frac{\sigma_w^2}{s_{\boldsymbol{\beta}}^2*}\boldsymbol{\beta}_0^*\right],$$

and

$$\tilde{\mathbf{C}}_{\boldsymbol{\beta\beta}} = \left[\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{I}\frac{\sigma_w^2}{s_{\boldsymbol{\beta}}^2}\right]^{-1}\sigma_w^2.$$

Obtaining samples from (55) is straightforward, especially if the order of $\boldsymbol{\beta}^*$ is not too large. When the prior distribution of $\boldsymbol{\beta}^*$ is diffuse ($s_{\boldsymbol{\beta}}^2* \to \infty$), this step involves essentially least-squares type computations, after making an offset of the data vector.

Conditional distributions of family-specific marked effects $\boldsymbol{\delta}$: Inspection of (54) reveals that all family-specific marked

regressions $\boldsymbol{\delta}_i$ are mutually independent, given all other parameters. The density of the conditional posterior distribution of $\boldsymbol{\delta}_i$ ($i = 1, 2, \ldots, F$) is

$$p(\boldsymbol{\delta}_i|\text{ELSE}) \propto N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i, \mathbf{I}\sigma_w^2) N(\boldsymbol{\delta}_i|\mathbf{T}_i\mathbf{c}_i, \mathbf{V}_i(\phi)).$$

The distribution can be shown to be the *l*-variate normal process

$$\boldsymbol{\delta}_i|\text{ELSE} \sim N(\tilde{\boldsymbol{\delta}}_i, \tilde{\mathbf{C}}_{\boldsymbol{\delta}i\boldsymbol{\delta}i}), \quad i = 1, 2, \ldots, F, \quad (56)$$

where

$$\tilde{\boldsymbol{\delta}}_i = [\mathbf{M}_i'\mathbf{M}_i + \mathbf{V}_i^{-1}(\phi)\sigma_w^2]^{-1}[\mathbf{M}_i'(\mathbf{y}_i - \mathbf{X}_i^*\boldsymbol{\beta}^* - \mathbf{1}_i\varepsilon_i)$$
$$+ \mathbf{V}_i^{-1}(\phi)\sigma_w^2\mathbf{T}_i\mathbf{c}_i],$$

and

$$\tilde{\mathbf{C}}_{\boldsymbol{\delta}i\boldsymbol{\delta}i} = [\mathbf{M}_i'\mathbf{M}_i + \mathbf{V}_i^{-1}(\phi)\sigma_w^2]^{-1}\sigma_w^2.$$

All family-specific marked effects can be drawn on a family-by-family basis by sampling from the multivariate normal distribution given in (56).

Conditional distributions of family-specific background effects ε: From (54), it follows that

$$p(\varepsilon_f|\text{ELSE}) \propto \left[\prod_{i=1}^{F} N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i, \mathbf{I}\sigma_w^2) N\left(\varepsilon_i|0, \sigma_{\varepsilon_f}^2\right)\right]$$
$$= \prod_{i=1}^{F} p(\varepsilon_i|\text{ELSE}),$$

so that family-specific background effects are conditionally independent, given all other parameters. In particular, for the ith family one has

$$p(\varepsilon_i|\text{ELSE}) \propto N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\varepsilon_i, \mathbf{I}\sigma_w^2) N(\varepsilon_i|0, \sigma_{\varepsilon_f}^2).$$

This can be shown to be the kernel of the density of the univariate normal distribution,

$$\varepsilon_i|\text{ELSE} \sim N(\tilde{\varepsilon}_i, \tilde{c}_{\varepsilon_f\varepsilon_i}), \quad i = 1, 2, \ldots, F, \quad (57)$$

where

$$\tilde{\varepsilon}_i = \left(\mathbf{1}_i'\mathbf{1}_i + \frac{\sigma_w^2}{\sigma_{\varepsilon_f}^2}\right)^{-1}\mathbf{1}_i'(\mathbf{y}_i - \mathbf{X}_i^*\boldsymbol{\beta}^* - \mathbf{M}_i\boldsymbol{\delta}_i)$$
$$= \frac{\sigma_{\varepsilon_f}^2}{\sigma_{\varepsilon_f}^2 + (\sigma_w^2/f_i)}\mathbf{1}_i'(\mathbf{y}_i - \mathbf{X}_i^*\boldsymbol{\beta}^* - \mathbf{M}_i\boldsymbol{\delta}_i)$$

and

$$\tilde{c}_{\varepsilon_i\varepsilon_i} = \left(\mathbf{1}_i'\mathbf{1}_i + \frac{\sigma_w^2}{\sigma_{\varepsilon_f}^2}\right)^{-1}\sigma_w^2$$
$$= \sigma_{\varepsilon_f}^2\left(1 - \frac{\sigma_{\varepsilon_f}^2}{\sigma_{\varepsilon_f}^2 + \sigma_w^2/f_i}\right).$$

Note that $\mathbf{1}_i'\mathbf{1}_i = f_i$ is the number of individuals in family i and that $\sigma_{\varepsilon_f}^2/(\sigma_{\varepsilon_f}^2 + \sigma_w^2/f_i)$ is proportional to the "heritability of a family mean" (whenever the variance of background effects between families is proportional to the additive genetic variance), in the terminology of stan-

dard quantitative genetics. Sampling from distribution (57) is straightforward.

Conditional distributions of chromosome effects: The density of the conditional distribution of the vector of chromosome effects $\mathbf{c}$ is, from (54),

$$p(\mathbf{c}|\text{ELSE}) \propto \left[\prod_{i=1}^{F} N(\boldsymbol{\delta}_i|\mathbf{T}_i\mathbf{c}_i, \mathbf{V}_i(\phi)) N(\mathbf{c}_i|\mathbf{0}, \mathbf{K}) \right]$$

$$= \prod_{i=1}^{F} p(\mathbf{c}_i|\text{ELSE}).$$

Thus, family-specific chromosome effect vectors $\mathbf{c}_i$ ($i = 1, 2, \ldots, F$) are conditionally independent of each other. After algebra, as in GIANOLA and FERNANDO (1986) or GIANOLA *et al.* (1990), it can be shown that the *i*th conditional distribution is the n_c-variate normal process,

$$\mathbf{c}_i|\text{ELSE} \sim N(\tilde{\mathbf{c}}_i, \tilde{\mathbf{C}}_{c_i c_i}), \quad i = 1, 2, \ldots, F, \quad (58)$$

where

$$\tilde{\mathbf{c}} = \{\mathbf{T}_i' \mathbf{V}_i^{-1}(\phi) \mathbf{T}_i + \mathbf{K}^{-1}\}^{-1} \mathbf{T}_i' \mathbf{V}_i^{-1}(\phi) \boldsymbol{\delta}_i,$$

and

$$\tilde{\mathbf{C}}_{c_i c_i} = \{\mathbf{T}_i' \mathbf{V}_i^{-1}(\phi) \mathbf{T}_i + \mathbf{K}^{-1}\}^{-1}.$$

For instance, in a species with 30 pairs of chromosomes such as cattle, a draw would need to be made from a 30-variate normal distribution for each of the *F* families represented in the data set.

Between- and within-family variances of background effects: The density of the conditional posterior distribution of the between-family variance of background effects, $\sigma_{\epsilon_f}^2$, is

$$p(\sigma_{\epsilon_f}^2|\mathbf{y}, \mathbf{H}) \propto \left[\prod_{i=1}^{F} N(\epsilon_i|0, \sigma_{\epsilon_f}^2) \right] p(\sigma_{\epsilon_f}^2|s_f^2, \nu_f).$$

Writing the *F* Gaussian densities and the density of the scaled inverted chi-square prior process of $\sigma_{\epsilon_f}^2$ explicitly, one arrives at

$$p(\sigma_{\epsilon_f}^2|\mathbf{y}, \mathbf{H}) \propto (\sigma_{\epsilon_f}^2)^{-((F+\nu_f)/2 + 1)} \exp\left[-\frac{\sum_{i=1}^{F}\epsilon_i^2 + s_f^2\nu_f}{2\sigma_{\epsilon_f}^2} \right].$$

This indicates that the conditional posterior distribution of $\sigma_{\epsilon_f}^2$ is scaled inverted chi square (*e.g.*, WANG *et al.* 1993; GELMAN *et al.* 1995), that is,

$$\sigma_{\epsilon_f}^2|\mathbf{y}, \mathbf{H} \sim \left(\sum_{i=1}^{F}\epsilon_i^2 + s_f^2\nu_f \right) \chi_{F+\nu_f}^{-2}. \quad (59)$$

To draw samples from (59), one extracts a random deviate from a central chi-square distribution on $F + \nu_f$ degrees of freedom, takes its reciprocal, and multiplies the inverted deviate by $(\sum_{i=1}^{F}\epsilon_i^2 + s_f^2\nu_f)$.

Likewise, the density of the conditional posterior distribution of the within-family variance is

$$p(\sigma_w^2|\text{ELSE}) \propto \left[\prod_{i=1}^{F} N(\mathbf{y}_i|\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{M}_i\boldsymbol{\delta}_i + \mathbf{1}_i\epsilon_i, \mathbf{I}\sigma_w^2) \right] p(\sigma_w^2|s_w^2, \nu_w).$$

Writing the densities explicitly and rearranging one arrives at

$$p(\sigma_w^2|\text{ELSE}) \propto (\sigma_w^2)^{-((n+\nu_w)/2+1)}$$

$$\times \exp\left[-\frac{(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^* - \mathbf{M}\boldsymbol{\delta} - \mathbf{F}\epsilon)'(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^* - \mathbf{M}\boldsymbol{\delta} - \mathbf{F}\epsilon) + s_w^2\nu_w}{2\sigma_w^2} \right].$$

This is the distribution of the scaled inverted chi-square random variable

$$\sigma_w^2|\text{ELSE} \sim [(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^* - \mathbf{M}\boldsymbol{\delta} - \mathbf{F}\epsilon)'(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^* - \mathbf{M}\boldsymbol{\delta} - \mathbf{F}\epsilon) + s_w^2\nu_w]\chi_{n+\nu_w}^{-2}.$$
$$(60)$$

The sampling process is similar to that described in connection with (59).

Autoregressive parameter ρ*:* Note in (54) that the only terms involving ρ are the prior densities of the family-specific marked effects $\boldsymbol{\delta}_i$ and the prior density $p(\rho|\boldsymbol{\eta})$. Hence,

$$p(\rho|\text{ELSE}) \propto \prod_{i=1}^{F} N(\boldsymbol{\delta}_i|\mathbf{T}_i\mathbf{c}_i, \mathbf{V}_i(\phi)) p(\rho|\boldsymbol{\eta}).$$

Now, recall that $\mathbf{V}_i(\phi)$ is an $l \times l$ covariance matrix such that

$$\mathbf{V}_i(\phi) = \bigoplus_{j=1}^{n_c} \mathbf{V}_{ij}(\phi),$$

where $\mathbf{V}_{ij}(\phi)$ is the $l_j \times l_j$ covariance matrix of within-chromosome deviates for family *i* in chromosome *j*. Specifically,

$$\mathbf{V}_{ij}(\phi) = \begin{bmatrix} 1 & \rho & \rho^2 & . & \rho^{l_j-1} \\ \rho & 1 & \rho & . & \rho^{l_j-2} \\ . & . & . & . & . \\ . & . & . & 1 & \rho \\ \rho^{l_j-1} & \rho^{l_j-2} & . & \rho & 1 \end{bmatrix} \sigma_{v_j}^2$$

$$= \mathbf{W}_j(\rho)\sigma_{v_j}^2. \quad (61)$$

Note that this covariance matrix is chromosome specific. Using this, the conditional posterior density above has the form

$$p(\rho|\text{ELSE}) \propto \prod_{i=1}^{F} |\oplus_{j=1}^{n_c}\mathbf{W}_j(\rho)\sigma_{v_j}^2|^{-1/2} \exp\left\{ -\frac{(\boldsymbol{\delta}_i - \mathbf{T}_i\mathbf{c}_i)'\left[\oplus_{j=1}^{n_c}\mathbf{W}_j^{-1}(\rho)/\sigma_{v_j}^2\right](\boldsymbol{\delta}_i - \mathbf{T}_i\mathbf{c}_i)}{2} \right\}$$

$$\times p(\rho|\boldsymbol{\eta}). \quad (62)$$

When viewed as a function of ρ, (62) is not in a recognizable form, irrespective of the form of the prior density $p(\rho|\boldsymbol{\eta})$ adopted for the autoregressive parameter. Effecting a draw from this distribution requires a more involved procedure, such as a single-site Metropolis or Metropolis-Hastings algorithm (METROPOLIS *et al.* 1953; TANNER 1993; GELMAN *et al.* 1995; GILKS *et al.* 1996; SORENSEN and GIANOLA 2002). Perhaps the simplest is to adopt a Metropolis step. Here, a proposal distribution is needed for generating candidate values of ρ. As in FISHER (1921), and for the purpose of facilitat-

ing the generation of proposals, consider transforming the correlation coefficient ρ into ζ as

$$\zeta = \frac{1}{2}\log\left(\frac{1 + \rho}{1 - \rho}\right),$$

with the Jacobian of the transformation being

$$J = \left[\frac{2\exp(\zeta)}{1 + \exp(2\zeta)}\right]^2.$$

Then, the conditional posterior density of ζ is

$$p(\zeta|\text{ELSE}) \propto \prod_{i=1}^{F}|\oplus_{j=1}^{n_c}\mathbf{W}_j(\zeta)\sigma_{v_j}^2|^{-1/2}\exp\left\{-\frac{(\boldsymbol{\delta}_i - \mathbf{T}_i\mathbf{c}_i)'\left[\oplus_{j=1}^{n_c}\mathbf{W}_j^{-1}(\zeta)/\sigma_{v_j}^2\right](\boldsymbol{\delta}_i - \mathbf{T}_i\mathbf{c}_i)}{2}\right\}$$

$$\times p(\zeta|\boldsymbol{\eta})\left[\frac{\exp(\zeta)}{1 + \exp(2\zeta)}\right]^2, \tag{63}$$

where $\boldsymbol{\zeta} = [\zeta, \sigma_{v_1}^2, \sigma_{v_2}^2, \ldots, \sigma_{v_{n_c}}^2]'$. Suppose now that the state of ρ at iteration t of the algorithm is $\rho^{[t]}$ and that all other parameters in its conditional posterior distribution have been updated. This implies that the state of ζ is

$$\zeta^{[t]} = \frac{1}{2}\log\left(\frac{1 + \rho^{[t]}}{1 - \rho^{[t]}}\right).$$

We update ζ via a Metropolis jump. Here, a candidate value ζ* must be sampled from a symmetric candidate generating distribution with density $Q(\zeta^*|\zeta^{[t]})$; symmetry means that $Q(\zeta^*|\zeta^{[t]}) = Q(\zeta^{[t]}|\zeta^*)$ for all pairs $(\zeta^*, \zeta^{[t]})$ and for every t (GELMAN *et al.* 1995). A distribution meeting this requirement is a *t*-process; the *t*-distribution has thicker tails than the normal, so "extreme" values appear at higher density, specially when the degrees of freedom are low (4–10, say). A possible Metropolis jump is as follows.

First, sample a deviate r from a Gamma(d.f./2, d.f./2) where d.f. denotes the degrees of freedom of the *t*-distribution.

Second, following Box and TIAO (1973), sample a candidate ζ* from

$$\zeta^* \sim N\left(\frac{1}{2}\log\left(\frac{1 + \rho^{[t]}}{1 - \rho^{[t]}}\right) - \frac{5\rho^{[t]}}{2l}, \frac{1}{l\sqrt{r}}\right).$$

Third, using (63), compute the Metropolis ratio:

$$\delta = \frac{p(\zeta^*|\text{ELSE})}{p(\zeta^{[t]}|\text{ELSE})}.$$

The integration constant cancels out in the numerator and denominator, so δ is calculated by direct evaluation of ratios involving (63).

Fourth, draw a deviate U from a Uniform(0, 1) distribution, and set

$$\zeta^{[t+1]} = \begin{cases} \zeta^* & \text{if } U \leq \delta \\ \zeta^{[t]} & \text{otherwise}. \end{cases}$$

Invert the candidate as

$$\rho^* = \frac{\exp(2\zeta^*) - 1}{1 + \exp(2\zeta^*)},$$

and set

$$\rho^{[t+1]}\begin{cases} \rho^* & \text{if } U \leq \delta \\ \rho^{[t]} & \text{otherwise}. \end{cases}$$

The preceding completes the updating of the autoregressive parameter.

Variances of within-chromosome deviates $\sigma_{v_i}^2$: From (54),

$$p(\sigma_{v_1}^2, \sigma_{v_2}^2, \ldots, \sigma_{v_{n_c}}^2|\text{ELSE}) \propto \left[\prod_{i=1}^{F}N(\boldsymbol{\delta}_i|\mathbf{T}_i\mathbf{c}_i, \mathbf{V}_i(\boldsymbol{\phi}))\right]\left[\prod_{j=1}^{n_c}p\left(\sigma_{v_j}^2|s_v^2, \nu_v\right)\right].$$

It is convenient to rewrite the model for marked effects presented in (42) into a chromosome-specific basis. Write

$$\begin{bmatrix} \boldsymbol{\delta}_1^* \\ \boldsymbol{\delta}_2^* \\ \cdot \\ \boldsymbol{\delta}_{n_c}^* \end{bmatrix} = \begin{bmatrix} \mathbf{T}_1^*\mathbf{c}_1^* \\ \mathbf{T}_2^*\mathbf{c}_2^* \\ \cdot \\ \mathbf{T}_{n_c}^*\mathbf{c}_{n_c}^* \end{bmatrix} + \begin{bmatrix} \mathbf{v}_1^* \\ \mathbf{v}_2^* \\ \cdot \\ \mathbf{v}_{n_c}^* \end{bmatrix},$$

where $\boldsymbol{\delta}_j^*$ is a vector of order $Fl_j \times 1$ containing the marked effects of the l_j markers in chromosome j for the F families, and $\mathbf{v}_j^*$ is a vector of within-chromosome deviates. An assumption made earlier is that within-chromosome deviates are conditionally mutually independent across chromosomes and families, but correlated within chromosomes according to the covariance matrix in (61). Since the preceding is merely a rearrangement of (42), it follows that

$$\prod_{i=1}^{F}N(\boldsymbol{\delta}_i|\mathbf{T}_i\mathbf{c}_i, \mathbf{V}_i(\boldsymbol{\phi})) = \prod_{j=1}^{n_c}N(\boldsymbol{\delta}_j^*|\mathbf{T}_j^*\mathbf{c}_j^*, \mathbf{I} \otimes \mathbf{V}_{ij}(\boldsymbol{\phi})),$$

where the identity matrix is $F \times F$. Hence,

$$p(\sigma_{v_1}^2, \sigma_{v_2}^2, \ldots, \sigma_{v_{n_c}}^2|\text{ELSE}) \propto \left[\prod_{j=1}^{n_c}N(\boldsymbol{\delta}_j^*|\mathbf{T}_j^*\mathbf{c}_j^*, \mathbf{I} \otimes \mathbf{V}_{ij}(\boldsymbol{\phi}))p\left(\sigma_{v_j}^2|s_v^2, \nu_v\right)\right]$$

$$\propto \prod_{j=1}^{n_c}p(\sigma_{v_j}^2|\text{ELSE}); \quad j = 1, 2, \ldots, n_c$$

indicating that the within-chromosome variances $\sigma_{v_j}^2$ have independent fully conditional posterior distributions. In particular, and writing the densities explicitly, one has for $\sigma_{v_j}^2$,

$$p(\sigma_{v_j}^2|\text{ELSE}) \propto |\mathbf{I} \otimes \mathbf{V}_{ij}(\boldsymbol{\phi})|^{-1/2}$$

$$\times \exp\left\{-\frac{(\boldsymbol{\delta}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*)'[\mathbf{I} \otimes \mathbf{V}_{ij}^{-1}(\boldsymbol{\phi})](\boldsymbol{\delta}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*)}{2}\right\}$$

$$\times p(\sigma_{v_j}^2|s_v^2, \nu_v)$$

$$\propto \left| \mathbf{I} \otimes \frac{\mathbf{W}_j^{-1}}{\sigma_{v_j}^2} \right|^{-1/2}$$

$$\times \exp\left\{ -\frac{(\boldsymbol{\gamma}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*)'\left[\mathbf{I} \otimes \mathbf{w}_j^{-1}/\sigma_{v_j}^2\right](\boldsymbol{\gamma}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*)}{2} \right\}$$

$$\times \, p(\sigma_{v_j}^2|s_v^2,\nu_v) \propto (\sigma_{v_j}^2)^{-((F_j + \nu_0)/2+1)}$$

$$\times \exp\left\{ -\frac{(\boldsymbol{\delta}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*)'[\mathbf{I} \otimes \mathbf{W}_j^{-1}](\boldsymbol{\delta}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*) + s_v^2\nu_v}{2} \right\}. \quad (64)$$

This is the density of the scale-inverted chi-square variate:

$$\sigma_{v_j}^2|\text{ELSE} \sim \left[(\boldsymbol{\delta}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*)'[\mathbf{I} \otimes \mathbf{W}_j^{-1}](\boldsymbol{\delta}_j^* - \mathbf{T}_j^*\mathbf{c}_j^*) + s_v^2\nu_v\right]\chi_{F_j}^{-2} + \nu_v;$$

$$j = 1, 2, \ldots, n_c. \quad (65)$$

Thus, the within-chromosome variance can be sampled independently, chromosome by chromosome.

Covariance matrix of chromosome effects $\mathbf{K}$: Retaining in (54) only those terms involving $\mathbf{K}$ leads to

$$p(\mathbf{K}|\text{ELSE}) \propto \left[\prod_{i=1}^{F} N(\mathbf{c}_i|\mathbf{0}, \mathbf{K})\right] p(\mathbf{K}|\mathbf{S}_c, \nu_c).$$

Now, we write explicitly the F normal densities and the inverse Wishart density $p(\mathbf{K}|\mathbf{S}_c, \nu_c)$, to obtain

$$p(\mathbf{K}|\text{ELSE}) \propto \left[\prod_{i=1}^{F} |\mathbf{K}|^{-1/2}\exp\left(-\frac{\mathbf{c}_i'\mathbf{K}^{-1}\mathbf{c}_i}{2}\right)\right]|\mathbf{K}|^{-(n_c + \nu_c + 1)/2}$$

$$\times \exp\left[-\frac{tr(\mathbf{S}_c\nu_c\mathbf{K}^{-1})}{2}\right]$$

$$\propto |\mathbf{K}|^{-(n_c + F + \nu_c + 1)/2}\exp\left\{-\frac{1}{2}tr\left[\mathbf{K}^{-1}\left(\sum_{i=1}^{F}\mathbf{c}_i\mathbf{c}_i' + \mathbf{S}_c\nu_c\right)\right]\right\}. \quad (66)$$

This indicates that the fully conditional posterior distribution is inverse Wishart of order n_c, degrees of freedom equal to $F + \nu_c$, scale matrix $\sum_{i=1}^{F}\mathbf{c}_i\mathbf{c}_i' + \mathbf{S}_c\nu_c$, and posterior expectation

$$E(\mathbf{K}|\text{ELSE}) = \frac{\sum_{i=1}^{F}\mathbf{c}_i\mathbf{c}_i' + \mathbf{S}_c\nu_c}{F + \nu_c - n_c - 1}$$

(given all other parameters). Procedures for sampling from inverse Wishart distributions in a quantitative genetics context are in, *e.g.*, JENSEN *et al.* (1994), KORSGAARD *et al.* (1999), and SORENSEN and GIANOLA (2002).

The MCMC scheme in a nutshell: A discussion of how Markov chain Monte Carlo schemes can be tuned, run, and monitored for convergence can be found in GELMAN *et al.* (1995), GILKS *et al.* (1996), COWLES and CARLIN (1996), ROBERT and CASELLA (1999), and SORENSEN and GIANOLA (2002). A possible implementation follows.

1. Set starting values that are hopefully "inside" of the target joint posterior distribution.
2. Sample systematically from each of the following con-ditional posterior distributions, perhaps changing the order of visitation at each round: (55) for $\boldsymbol{\beta}^*$; (56), (57), and (58) for the family-specific marked effects, background effects, and chromosome effects, respectively; (59) and (60) for the between- and within-family variances of the background effects; (62) for the autoregressive parameter, after incorporating the Metropolis step; (65) for the n_c variances of within-chromosome deviates; and (66) for $\mathbf{K}$, the covariance matrix of chromosome effects. This constitutes a single completed iteration of the scheme. At each subsequent loop, update the values of the appropriate conditioning values with the new draws.
3. Repeat the iteration as many times as needed to ensure (a) that draws can be reasonably claimed as belonging to the posterior distribution and (b) that posterior features are estimated with a sufficiently small Monte Carlo error.
4. If necessary, discard early iterations, in what is called the "burn-in" period.
5. Using the remaining samples, estimate any feature of the posterior distribution, *e.g.*, a posterior mean or variance, a posterior density, or the distribution of an order statistic.

To illustrate the flexibility of the approach, consider inferring the Lorenz curve and the Gini coefficient in (36) and (37), respectively, and contrast this with the crude methods discussed in *Differential contribution of marked effects to genetic variability*. Let $\boldsymbol{\gamma}^{(k)}$ be a draw from the posterior distribution of the marked effects in the simplest model. Then, as in (34), $\mu_j^{(k)} = \mathbf{m}_{\text{unique},j}'\boldsymbol{\gamma}^{(k)}$ is a draw from the posterior distribution of μ_j and $S(\boldsymbol{\gamma}^{(k)})$ is a sample from the posterior distribution of $S(\boldsymbol{\gamma})$ in (35). Next, order all the $\mu_j^{2(k)}$, so that

$$\left[\mu_{[1]}^{2(k)}, \mu_{[2]}^{2(k)}, \ldots \mu_{[u]}^{2(k)}\right]$$

is a sample from posterior distribution of the ordered squared "unique" marked effects. Then

$$L^{(k)}\left(\frac{i}{u}\right) = \frac{\sum_{j=1}^{i}\mu_{[j]}^{2(k)}}{S(\boldsymbol{\gamma}^{(k)})} = \frac{\sum_{j=1}^{i}\mu_{[j]}^{2(k)}}{\sum_{j=1}^{u}\mu_j^{2(k)}}$$

is a sample from the posterior distribution of the Lorenz curve in (36). Likewise,

$$G^{(k)} = \frac{2}{(u\sum_{j=1}^{u}\mu_j^{2(k)})}\sum_{j=1}^{u}\left(j - \frac{1}{2}\right)\mu_{[j]}^{2(k)}$$

is a sample from the posterior distribution of the Gini coefficient in (37), measuring unequal contribution of marked effects to genetic variability. The Bayesian analysis produces an entire description of the uncertainty about the Lorenz curve and the Gini coefficient.

Draws from the posterior distribution of the "total" genetic value (in the simplest model) would be obtained

as $\mathbf{M}\gamma^{(k)} + \varepsilon^{(k)}$, and posterior means, variances, and co-variances can be computed directly from such samples.

DISCUSSION

Several methods for incorporating molecular information into predictors of genetic merit of candidates for selection in improvement programs have been proposed in recent years. For example, Lande and Thompson (1990) suggested using multiple regression (in a least-squares sense) of the phenotypic value of an individual on the number of copies of alleles at polymorphic loci; in situations where the linkage map is known, separate regressions could be calculated for each chromosome. They argued that, given a sufficient number of markers, nearly all of the additive genetic variance could be accounted for by the regression model. These authors calculated a "molecular score" (the fitted value of the least-squares regression) and combined the molecular score with the phenotype for the quantitative trait via standard selection index techniques. Lande and Thompson (1990) also introduced a familial structure and discussed how selection could be optimized (in some sense) by combining family means and within-family deviations for both the molecular score and the phenotypic values via selection index theory (Smith 1936; Hazel 1943). There are some problems with this approach. First, the method breaks down when formulated in a vectorial manner. Even in the trivial case where candidates are independent and identically distributed (essentially their setting), the covariance matrix of the molecular scores is singular since the distribution of fitted values in regression is defined only in a p-dimensional space, rather than in n dimensions (the number of individuals with molecular scores). Hence, a selection index approach coined in the more general framework of best linear prediction (Henderson 1973) leads to an infinite number of solutions. Hence, care must be exercised to ensure that the functions of genetic merit to be predicted are indeed predictable (Harville 1976).

Another difficulty with the least-squares method arises when the number of markers is almost as large as the number of individuals typed. For example, >1.5 million single-nucleotide polymorphisms (the current desideratum of genetic marker) have been identified in the human genome and their positions located at an average spacing of 2×10^{-3} cM (Hartl and Jones 2002). If advantage is to be taken of this map density, this leads to a situation where $l \gg n$. Hence, colinearity of predictor variables creates an insurmountable parameter identification problem unless some dimension-reduction technique, such as singular value decomposition, or some *ad hoc* marker model selection procedure is employed. Also, it is well known in statistics that least-squares is an inadmissible estimation procedure. Stein (1955) showed for the orthonormal linear regression model that whenever $l > 2$, the least-squares estimator is dominated (in the mean-squared error sense) throughout the parameter space by a nonlinear, shrinkage type, estimator. This result, which torpedoes best linear unbiased estimation and maximum-likelihood estimation under normality, has been ignored in quantitative genetics (Gianola 1990), although it is widely acknowledged in econometrics (Theil 1971; Judge *et al.* 1985). On these grounds, not much should be expected from least-squares derivatives in marker-assisted prediction of genetic value, as found by Meuwissen *et al.* (2001).

Whittaker *et al.* (2000) and Whittaker (2001) noted that having "too many" markers in the regression model produces serious colinearity, causing unstable least-squares estimates and a poor prediction of the molecular score. The method leads to apparent paradoxes, as follows. At least in theory, the regression coefficients of phenotype on markers flanking a QTL must have the same sign (Whittaker *et al.* 1996). However, the estimated regression coefficients can differ in sign even when the true parameters agree in sign (Hwang and Nettleton 2002). The latter authors point out that sign inconsistency is greatest when flanking markers are close together and, thus, highly correlated. As an alternative to selection of subsets of markers to include in the model, Whittaker *et al.* (2000) and Whittaker (2001) suggested employing ridge regression (Hoerl and Kennard 1970) for estimation. This *ad hoc* procedure shrinks the least-squares estimates toward zero, improves the condition of the coefficient matrix in least-squares, and often produces a reduction in mean-squared error of estimation. Here, one encounters the typical dilemma: the shrinkage estimator may be more precise but less accurate, since it is biased. Whether or not this reduces mean-squared error depends on a bias *vs.* variance trade-off, and the performance of ridge relative to least-squares regression is a function of the true value of the regression vector, which is obviously unknown. One never knows how much shrinkage should be effected, although there are some informal recipes for tuning the shrinkage parameter. On the other hand, ridge regression makes more sense when viewed from a Bayesian perspective (Lindley and Smith 1972; Zellner and Vandaele 1975; Gianola and Fernando 1986). Ridge regression is equivalent to adopting a normal prior for the regression vector centered at zero and with a prior covariance structure equal to an identity matrix times a scalar, the variance of the prior distribution. A small variance means large prior precision and more shrinkage (a stronger belief that the regression is null); a large variance implies diffuse prior knowledge of the regression vector. The scalar prior covariance matrix implies that the individual regression coefficients of phenotypes on markers are viewed as independently and identically distributed *a priori*. Shrinking all marked effects toward a common prior mean may be questionable. Also, assuming that

different chromosome sections have independent expression levels is contradicted by evidence from gene expression trials (CARON *et al.* 2001).

In practice, all individuals cannot be typed for all markers. Further, if typing costs continue to be abated, the number of predictor variables (markers) may grow faster than the number of observations. This situation is encountered routinely in animal breeding, and it can be solved by positing random-effect models. The BLUP methods used in animal breeding can be viewed as based on conditional or posterior distributions (GIANOLA and FERNANDO 1986), illustrating why the dimensionality and identification problems arising in least-squares are circumvented. MEUWISSEN *et al.* (2001) noted this and compared least-squares with BLUP and with two Bayesian methods for marker-assisted prediction. In their simulation, effects of 50,000 marker haplotypes needed to be inferred from 2200 observations. Since this is not possible via least-squares, they calculated regressions on markers found in chromosome segments of 1 cM each and evaluated a log-likelihood (assuming normality) for every segment. Then, "significant" segments were selected using an arbitrary procedure on the basis of said likelihoods. Eventually, the authors combined all such segments into a single multiple regression model. In the BLUP analysis, all chromosome effects were regarded as independent and identically distributed random effects with known variance. Their first Bayesian model was as in BLUP, but assigning independent scaled inverted chi-square distributions to the variances of the chromosome segments. The second Bayesian model used a mixture prior distribution for the variance of the chromosome segment variances. Their main conclusions were: (1) least-squares had a low accuracy of prediction of genetic value, (2) the BLUP and Bayesian methods differed by little, and (3) the Bayesian methods assigning prior distributions to the variance associated with chromosome segments gave the most accurate predictions.

The classical and Bayesian methods outlined here do not accommodate complex genealogical structures such as those arising in animal breeding. To illustrate, consider the hypothetical situation in which the markers and the QTL are identical, so that there are no background effects if gene action is additive. Hence, in the notation of A MIXED-EFFECTS MODEL FORMULATION, $a_i = \sum_{k=1}^{l} m_{ik}\gamma_k$, and, therefore, one would expect *a priori* that the covariance between the genetic value of any two relatives i and i' is equal to

$$\text{Cov}\left(\sum_{k=1}^{l} m_{ik}\gamma_k, \sum_{k=1}^{l} m_{i'k}\gamma_k\right) = \sigma_{\gamma}^2 E(\mathbf{m}_i'\mathbf{m}_i').$$

In the absence of knowledge of genotype frequencies, the term $E(\mathbf{m}_i'\mathbf{m}_i')$ cannot be assessed. In a fully Bayesian method one needs to assign a prior to the genotypic frequencies and take the MCMC method even further

(*e.g.*, HOESCHELE 2001). For this reason, all analyses presented in our article treat the matrix $\mathbf{M}$ as nonstochastic. Concerning the background genetic effects, one can introduce a full relationship matrix under additive inheritance.

Model (38) introduces a family structure, and (47) expands LANDE and THOMPSON (1990) to a matrix domain, but with a random-effects treatment. To illustrate a Bayesian analysis of the latter, consider the following terms: (1) $\overline{\mathbf{M}}_i\boldsymbol{\alpha}_B$, the $f_i \times 1$ vector of unobserved (contrary to least-squares fitted values) mean molecular scores for all members of family i (all its elements are equal); (2) $(\mathbf{M}_i - \overline{\mathbf{M}}_i)\boldsymbol{\alpha}_W$, the within-family contributions to the molecular scores; (3) ε_i or background genetic effect of family i (*e.g.*, the additive genetic contribution, net of the marked part of the variability); and (4) $\mathbf{w}_i$, a vector of within-family deviations (also net of the marked part of the variability). The mean of the posterior distribution of the total genetic merit of the individuals in family i would be obtained by averaging $\overline{\mathbf{M}}_i\boldsymbol{\alpha}_B^{(k)} + (\mathbf{M}_i - \overline{\mathbf{M}}_i)\boldsymbol{\alpha}_W^{(k)} + \mathbf{1}_i\varepsilon_i^{(k)}$ over the MCMC samples. Similarly, the posterior mean of the within-family deviations can be estimated by averaging $\mathbf{w}_i^{(k)} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}^{(k)} - \mathbf{M}_i\boldsymbol{\gamma}_i^{(k)} - \mathbf{1}_i\varepsilon_i^{(k)}$. LANDE and THOMPSON (1990) consider "selection optimization" by defining total merit of a candidate for selection as a linear combination of family and within-family information. For instance, one could write the merit vector

$$\mathbf{T}_i = h_{F\alpha}\overline{\mathbf{M}}_i\boldsymbol{\alpha}_B + h_{F\varepsilon}\mathbf{1}_i\varepsilon_i + h_{w\alpha}(\mathbf{M}_i - \overline{\mathbf{M}}_i)\boldsymbol{\alpha}_W + h_{w\varepsilon}\mathbf{w}_i,$$

where $h_{F\alpha}$ and $h_{F\varepsilon}$ are some nonstochastic relative weights assigned to the unobserved molecular and quantitative trait family means, respectively, and $h_{w\alpha}$, $h_{w\varepsilon}$ are corresponding weights assigned to the unobserved within-family deviations. The means, variances, covariances, and any feature of a posterior distribution involving the unobserved total merits $\mathbf{T}_i$ can be estimated from the MCMC samples. The procedure takes the ideas of LANDE and THOMPSON (1990) to a higher level of generality, with the uncertainty about nuisance parameters taken into account fully in the usual Bayesian manner (BOX and TIAO 1973).

We have assumed throughout that all individuals have been typed for all markers, but this is seldom the case, as noted earlier. In the Bayesian context missing markers can be dealt with automatically via an augmentation of the posterior distribution. Subsequently, MCMC is used to make imputations for the unobserved part of the molecular information, at least if missingness is at random in the sense of RUBIN (1976). See HOESCHELE (2001) for technical details.

It remains to be seen to what extent the proposed procedures hold in practice or in simulations what they seem to promise on theoretical grounds. In a simulation, WHITTAKER *et al.* (2000) found that a simple ridge regression procedure enhanced response to selection over plain phenotypic selection or selection based on

least-squares fits of the molecular scores. The results of Meuwissen *et al.* (2001) point in the same direction. This is encouraging. Another issue is the computer intensiveness of the required calculations. The BLUP-REML analysis can be carried out with standard software for large data sets. On the other hand, the MCMC procedures require a higher level of computation; however, they are more flexible and lead to richer inference (*e.g.*, estimation of the Lorenz curve). At any rate, computational burden is unimportant relative to the cost of genotyping individuals. Also, the Bayesian analysis requires elicitation of some complex priors distributions and this is difficult. In practice, a possibility is to use some of the standard default noninformative priors (*e.g.*, Sorensen and Gianola 2002) although at the risk of causing impropriety. Another option is to carry out the Bayesian analysis with different priors. If inferences are invariant with respect to the set of priors used, this would suggest that the data are "informative enough" and that the prior is overwhelmed by the information contributed by the data.

LITERATURE CITED

Aguilar-Gutierrez, X., 2000 *Desigualdad y Pobreza en México, son inevitables?* Instituto de Investigaciones Económicas, Universidad Nacional Autónoma de México, Mexico City.

Bernardo, J. M., 1979 Reference posterior distributions for Bayesian inference. J. R. Stat. Soc. B **41:** 113–147.

Bernardo, J. M., and A. F. M. Smith, 1994 *Bayesian Theory.* John Wiley & Sons, New York.

Box, G. E. P., 1980 Sampling and Bayes' inference in scientific modelling and robustness. J. R. Stat. Soc. A **143:** 383–430.

Box, G. E. P., and G. C. Tiao, 1973 *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, UK.

Cantet, R. J. C., R. L. Fernando and D. Gianola, 1992 Bayesian inference about dispersion parameters of univariate mixed models with maternal effects: theoretical considerations. Genet. Sel. Evol. **24:** 107–135.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins *et al.*, 2001 The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291:** 1289–1292.

Cowles, M. K., and B. P. Carlin, 1996 Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Am. Stat. Assoc. **91:** 883–904.

Farnir, F., W. Coppieters, J. J. Arranz, P. Berzi, N. Cambisano *et al.*, 2000 Extensive genome-wide linkage disequilibrium in dairy cattle. Genome Res. **10:** 220–227.

Fernando, R. L., and M. Grossman, 1989 Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **33:** 209–229.

Fisher, R. A., 1921 On the probable error of a coefficient of correlation deduced from a small sample. Metron **4:** 3–32.

Gastwirth, J. L., 1971 A general definition of Lorenz curve. Econometrica **39:** 1037–1039.

Gastwirth, J. L., 1972 The estimation of the Lorenz curve. Rev. of Econ. Stat. **54:** 306–316.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis.* Chapman & Hall, London.

Georges, M. D., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto *et al.*, 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics **139:** 907–920.

Gianola, D., 1990 Can BLUP and REML be improved upon?, pp. 445-449 in *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production,* Vol. XIII. Joyce Darling, Penicuik, UK.

Gianola, D., and R. L. Fernando, 1986 Bayesian methods in animal breeding theory. J. Anim. Sci. **63:** 217–244.

Gianola, D., and B. Goffinet, 1982 Sire evaluation with best linear unbiased predictors. Biometrics **38:** 1085–1088.

Gianola, D., S. Im and F. W. Macedo, 1990 A framework for prediction of breeding value, pp. 210–238 in *Advances in Statistical Methods for Genetic Improvement of Livestock,* edited by D. Gianola and K. Hammond. Springer-Verlag, Heidelberg.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

Hartl, D. L., and E. W. Jones, 2002 *Essential Genetics: A Genomics Perspective.* Jones & Bartlett, Sudbury, MA.

Harville, D. A., 1976 Extension of the Gauss-Markov theorem to include the estimation of random effects. Ann. Stat. **4:** 384–395.

Hayes, B., and M. E. Goddard, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. **33:** 209–229.

Hazel, L. N., 1943 The genetic basis for constructing selection indexes. Genetics **28:** 476–490.

Henderson, C. R., 1973 Sire evaluation and genetic trends, pp. 10–41 in *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush.* American Society of Animal Science and American Dairy Science Association, Champaign, IL.

Henderson, C. R., O. Kempthorne, S. R. Searle and C. M. von Krosigk, 1959 The estimation of environmental and genetic trends from records subject to culling. Biometrics **15:** 192–218.

Hoerl, A. E., and R. W. Kennard, 1970 Ridge regression: biased estimation and applications for non-orthogonal problems. Technometrics **12:** 55–82.

Hoeschele, I., 2001 Mapping quantitative trait loci in outbred pedigrees, pp. 599–644 in *Handbook of Statistical Genetics,* edited by D. J. Balding M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Hwang, J. T., and D. Nettleton, 2002 Investigating the probability of sign inconsistency in the regression coefficients of markers flanking quantitative trait loci. Genetics **160:** 1697–1705.

Jensen, J., C. S. Wang, D. A. Sorensen and D. Gianola, 1994 Bayesian inference on variance and covariance components for traits influenced by maternal and direct genetic effects using the Gibbs sampler. Acta Agric. Scand. **44:** 193–201.

Judge, G. G., W. E. Griffiths, R. Carter, H. Lutkepohl and T. C. Lee, 1985 *The Theory and Practice of Econometrics.* John Wiley & Sons, New York.

Korsgaard, I. R., A. H. Andersen and D. Sorensen, 1999 A useful reparameterisation to obtain samples from conditional inverse Wishart distributions. Genet. Sel. Evol. **31:** 177–181.

Lande, R., and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124:** 743–756.

Lange, C., and J. C. Whittaker, 2001 On prediction of genetic values in marker-assisted selection. Genetics **159:** 1375–1381.

Lindley, D. V., and A. F. M. Smith, 1972 Bayes estimates for the linear model. J. R. Stat. Soc. B **34:** 1–41.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculation by fast computing machines. J. Chem. Phys. **21:** 1087–1091.

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

Ollivier, L., 1998 The accuracy of marker-assisted selection for quantitative traits in populations in linkage equilibrium. Genetics **148:** 1367–1372.

Robert, C. P., and G. Casella, 1999 *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

Rubin, D. B., 1976 Inference and missing data. Biometrika **63:** 581–592.

Searle, S. R., 1974 Prediction, mixed models and variance components, pp. 229–266 in *Reliability and Biometry,* edited by F. Proschan and R. J. Serfling. Society for Industrial and Applied Mathematics, Philadelphia.

Searle, S. R., 1982   *Matrix Algebra Useful for Statistics.* John Wiley & Sons, New York.

Searle, S. R., G. Casella and C. E. McCulloch, 1992   *Variance Components.* John Wiley & Sons, New York.

Smith, C., and S. P. Simpson, 1986   The use of genetic polymorphisms in livestock improvement. J. Anim. Breed. Genet. **103:** 205–217.

Smith, H. F., 1936   A discriminant function for plant selection. Ann. Eugen. **7:** 240–250.

Soller, M., and J. S. Beckmann, 1983   Genetic polymorphisms in varietal identification and genetic improvement. Theor. Appl. Genet. **67:** 25–33.

Sorensen, D., and D. Gianola, 2002   *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.* Springer-Verlag, New York.

Sorensen, D., C. S. Wang, J. Jensen and D. Gianola, 1994   Bayesian analysis of genetic change due to selection using Gibbs sampling. Genet. Sel. Evol. **26:** 333–360.

Stein, C., 1955   Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, pp. 192-206 in *Proceedings of the Third Berkeley Symposium on Mathematics, Statistics, and Probability.* University of California Press, Berkeley.

Tanner, M. A., 1993   *Tools for Statistical Inference.* Springer-Verlag, New York.

Theil, H., 1971   *Principles of Econometrics.* John Wiley & Sons, New York.

Urioste, J. I., D. Gianola, R. Rekaya, W. F. Fikse and K. A. Weigel, 2001   Evaluation of extent and amount of heterogeneous variance for milk yield in Uruguayan Holsteins. Anim. Sci. **72:** 259–268.

Verbeke, G., and G. Molenberghs, 1997   *Linear Mixed Models in Practice: A SAS Oriented Approach.* Springer-Verlag, New York.

Verbeke, G., and G. Molenberghs, 2000   *Linear Mixed Models for Longitudinal Data.* Springer, New York.

Wang, C. S., J. J. Rutledge and D. Gianola, 1993   Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. **25:** 41–62.

Wang, C. S., J. J. Rutledge and D. Gianola, 1994   Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet. Sel. Evol. **26:** 91–115.

Wang, T., R. L. Fernando and M. Grossman, 1998   Genetic evaluation by best linear unbiased prediction using marker and trait information in a multibreed population. Genetics **148:** 507–515.

Weigel, K. A., D. Gianola, R. J. Tempelman, C. A. Matos, I. H. C. Chen *et al.*, 1991   Improving estimates of fixed effects in a mixed linear model. J. Dairy Sci. **74:** 3174–3182.

Whittaker, J. C., 2001   Marker-assisted selection and introgression, pp. 673–693 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Whittaker, J. C., R. Thompson and P. M. Visscher, 1996   On the mapping of QTL by regression of phenotype and marker-type. Heredity **77:** 23–32.

Whittaker, J. C., R. Thompson and M. C. Denham, 2000   Marker-assisted selection using ridge regression. Genet. Res. **75:** 249–252.

Xu, S., 1998   Mapping quantitative trait loci using multiple families of line crosses. Genetics **148:** 517–524.

Zellner, A., and W. Vandaele, 1975   Bayes-Stein estimators for k-means, regression and simultaneous equation models, pp. 627–653 in *Studies in Bayesian Econometrics and Statistics,* edited by S. Fienberg and A. Zellner. North-Holland, Amsterdam.

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.

Communicating editor: J. B. Walsh