Biostatistics (2008), **9**, 4, *pp.* 635–657
doi:10.1093/biostatistics/kxm055
Advance Access publication on March 14, 2008

A likelihood-based approach to mixed modeling with ambiguity in cluster identifiers

ANDREA S. FOULKES*, RECAI YUCEL, XIAOHONG LI

Division of Biostatistics, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA, USA
foulkes@schoolph.umass.edu

SUMMARY

This manuscript describes a novel, linear mixed-effects model–fitting technique for the setting in which correlated data indicators are not completely observed. Mixed modeling is a useful analytical tool for characterizing genotype–phenotype associations among multiple potentially informative genetic loci. This approach involves grouping individuals into genetic clusters, where individuals in the same cluster have similar or identical multilocus genotypes. In haplotype-based investigations of unrelated individuals, corresponding cluster assignments are unobservable since the alignment of alleles within chromosomal copies is not generally observed. We derive an expectation conditional maximization approach to estimation in the mixed modeling setting, where cluster assignments are ambiguous. The approach has broad relevance to the analysis of data with missing correlated data identifiers. An example is provided based on data arising from a cohort of human immunodeficiency virus type-1–infected individuals at risk for antiretroviral therapy–associated dyslipidemia.

Keywords: Expectation conditional maximization; Genotype; Haplotype; HIV-1; Lipids; Missing identifiers; Mixed-effects models; Phenotype; Population-based genetic association studies.

1. INTRODUCTION

Mixed-effects modeling is a well-established method for the analysis of correlated data where correlation among observations can arise from repeated measures or clustering. Since the landmark paper of Laird and Ware (1982), an extensive literature has developed that spans a range of model-fitting techniques and applications, including Diggle *and others* (1994), Vonesh and Chinchilli (1997), Pinheiro and Bates (2000), Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Demidenko (2004), and Fitzmaurice *and others* (2004), among others. Together, these provide a clear and comprehensive discussion of state-of-the-art methods for estimation, testing, and prediction in the context of linear, generalized linear, and nonlinear mixed-effects modeling. In addition, a broad array of applications are presented with complete discussion of available software tools for implementation of existing methods. To our knowledge, a fully likelihood–based method that specifically addresses unobservable correlated data indicators, that is, missing individual or cluster identifiers, has not been described.

*To whom correspondence should be addressed.

The data settings motivating our research are population-based genetic association studies of unrelated individuals for whom haplotypic phase, that is, the alignment of alleles on a single chromosome, is unobservable. In a recent manuscript, we describe a multistage approach for this setting that involves (1) estimating haplotype frequencies using only the available genetic information, (2) multiply imputing cluster membership identifiers, (3) for each of these imputations, fitting a mixed-effects model for the outcome of interest, such as a measure of disease progression, using existing analytical tools, and (4) combining the results across imputations to make inference (Foulkes *and others*, 2007). While this approach is straightforward to implement, it does not provide knowledge about the outcome to inform the haplotype frequency estimation. That is, estimation of haplotype frequencies (step 1 above) is done independently of the mixed-effects model–fitting procedure (step 3). In the present manuscript, we derive a novel, likelihood-based approach that incorporates the haplotype estimation component into the model-fitting procedure. Our approach has the marked advantage of drawing strength from a clinical measure (outcome in the model framework) to update the haplotype frequency estimates.

Specifically, we derive an expectation conditional maximization (ECM) algorithm for this missing data setting. Expectation–maximization (EM)-type algorithms have been described for fitting mixed-effects models (Laird and Ware, 1982; Jennrich and Schluchter, 1986; Laird *and others*, 1987; Jamshidian and Jennrich, 1993). In its original formulation, model random effects are treated as missing data (McCulloch and Searle, 2001, p. 264). We extend this for our setting by letting both the random effects and the correlated data indicators together constitute the missing component. We also distinguish our setting from the more common missing data settings in which covariate or response data are missing and/or there is imbalance in the design, that is, unevenly spaced measurements over time. Methods for these settings are well described as noted in Fitzmaurice *and others* (2004, p. 375) and McCulloch and Searle (2001, p. 94).

The ECM approach originally proposed by Meng and Rubin (1993) extends the EM algorithm of Dempster *and others* (1977) to reduce complexities in the maximization step by partitioning the set of parameters into disjoint and exhaustive subsets with likelihood functions that are easier to maximize. Two alternative maximization algorithms are well described in the context of fitting mixed models, Newton–Raphson and Fisher scoring (FS) (Lindstrom and Bates, 1988; Wolfinger *and others*, 1994; Pinheiro and Bates, 2000; Demidenko, 2004), and combinations of each with EM-type algorithms provide both efficiency and stability. A combination of FS and the EM was recently proposed for missing covariate and response data by Schafer and Yucel (2002). While further extensions for missing cluster identifiers are tenable, the ECM algorithm is efficient, provides simple interpretable solutions at each step, and converges reliably to maximum likelihood (ML) estimates by guaranteeing an increase in the likelihood function at each iteration (Little and Rubin, 1987).

Unobservable cluster identifiers can arise in a variety of settings. For example, hospital records may have incomplete information on patients' local area identifiers such as ZIP codes, which may be desirable in modeling treatment patterns (Chiu *and others*, 2005). Alternatively, clusters may define underlying biological constructs that are not observable. In general, investigators can identify a subset of clusters that are consistent with the observed data. For example, additional information available from either census records or hospital records may identify a set of possible ZIP codes. In the context of characterizing biological states, genetic indicators can inform us about the set of possible groupings of individuals (Foulkes and DeGruttola, 2002).

The data motivating our research arise from a cohort of human immunodeficiency virus type-1 (HIV-1)–infected individuals on highly active antiretroviral therapy (HAART). Long-term exposure to HAART has be associated with an array of lipid abnormalities that can lead to early onset of cardiovascular disease in this population. Our investigation aims to characterize the associations among genetic polymorphisms and lipids, controlling for the effects of drug exposures and other relevant clinical and demographic factors. Ultimately, understanding the pharmacogenomic underpinnings to complex diseases, such as

cardiovascular disease, will have broad implications for tailoring therapy decisions to patient-specific characteristics.

In general, the pair of single nucleotide polymorphisms (SNPs) at each locus within a gene is observed; however, the alignment of these nucleotides across loci for a given chromosomal copy is unobservable. This unobservable information, commonly referred to as haplotypic phase, can be biologically and clinically informative and ignoring it may lead to a loss of power to detect associations. In this manuscript, we describe how the ECM approach accounting for uncertainty in cluster identifiers can be applied to the setting of ambiguous-phase haplotype data to discover clinically relevant biological associations. This approach represents a contribution to existing methodology since it addresses simultaneously the need to consider multiple genetic indicators and the unobservable aspect of haplotypic phase using a fully likelihood–based approach.

Recently, Foulkes *and others* (2005) proposed applying mixed-effects models to data arising from genetic association studies of unrelated individuals. A primary strength of this approach is that it allows for assessing overall variability across combinations of multiple genetic polymorphisms using a single, omnibus test while controlling for potential confounding by environmental and clinical characteristics. Empirical Bayes estimates of multilocus genotype effects and corresponding prediction intervals lend additional insight into the specific polymorphisms contributing to measures of disease progression. The method proposed herein extends this approach to handle the setting in which genetic information is unobservable.

The proposed method also extends the generalized linear modeling approach of Lake *and others* (2003) and Lin and Zeng (2006) that both describe implementation of an EM algorithm for unobservable haplotype data. Notably, both the mixed modeling approach and the methods given in Lin and Zeng (2006) can accommodate specific departures from Hardy–Weinberg equilibrium (HWE). The primary difference between the approaches is that the mixed modeling approach we present assumes that haplotype effects are random, arising from an underlying probability distribution. This provides a flexible analytic framework for characterizing a large number of genetic indicators and may offer a solution to the degrees-of-freedom problem inherent in tests of haplotype–trait associations as described by Tzeng *and others* (2006).

Finally, mixed-effects models have been described as a special case of structural equation models (SEMs) or latent class models (Sanchez *and others*, 2005). Here, we introduce a doubly latent class structure since there are latent cluster random effects as well as unobservable cluster identifiers. Notably, the inclusion of both latent class indicators and latent random effects has been described in the SEM literature. For example in Muthen and Shedden (1999), alcohol dependency classes have latent indicators while person-specific random effects are included to account for repeated measures over time. In our setting, clusters similarly have latent indicators but it is the same clusters (and not individuals) that are assumed to have random effects. This renders our setting distinct. The heterogeneity model of Verbeke and Lesaffre (1996), on the other hand, assumes an unobservable mixture distribution on the random effects, that is, that the random cluster effects are themselves clustered. This is again different from our setting since we assume that the cluster effects arise from a single distribution while the membership to these clusters is potentially unobservable. These are subtle distinctions but important ones requiring novel associated methods.

We begin in Section 2 by outlining our notation, the assumed underlying model, and a brief summary of estimation via the EM algorithm in the usual linear mixed modeling setting in which cluster assignments are fully observed. We then describe a novel estimation approach in the general context of cluster ambiguity in Section 3. Extensions for investigation of genetic associations are provided in Section 4. Finally, in Section 5 we present a summary of results from a simulation study and from applying this approach to a study of HAART-associated dyslipidemia in HIV-1-infected individuals. A comprehensive discussion of the simulation approach and corresponding findings is provided in Appendix 3.

2. BACKGROUND

2.1 *Notation and model*

Consider the linear mixed-effects model given in (2.1) for $i = 1, \ldots, M$, where $\mathbf{Y}_i$ is an $n_i \times 1$ vector with jth element equal to the response for the jth observation in cluster i, n_i is the number of observations in cluster i, $\mathbf{X}_i$ is a corresponding matrix of covariates, and $\mathbf{Z}_i$ is the design matrix for random cluster effects. We assume $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$, $b_i \overset{\text{iid}}{\sim} N(0, D)$, and $\epsilon_{ij} \perp b_i$. In the general mixed modeling setting, $\mathbf{Z}_i$ is observed and an EM approach is used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, where $\boldsymbol{\beta}$ is the vector of mean parameters and $\boldsymbol{\theta} = (D, \sigma_\epsilon^2)$ is a vector of variance components. This approach is described in Laird and Ware (1982) and summarized in Section 2.2 below:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i b_i + \epsilon_i. \tag{2.1}$$

Now, suppose $\mathbf{Z}_{N\times M} = \text{blkdiag}[\mathbf{Z}_i]$, where $N = \sum_{i=1}^M n_i$. In the ambiguous cluster setting, both $\mathbf{Z}$ (the indicator for cluster membership) and n_i are potentially unobserved. In addition, the elements of $\mathbf{Y}_i$ and $\mathbf{X}_i$ will vary depending on cluster assignments. Let the observed data relevant to cluster assignments be G. We define $\mathcal{S}$ to be the set of all design matrices $\mathbf{Z}$ that are consistent with these observed data. For simplicity of notation in subsequent sections, we let $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_M^T)^T$, $\mathbf{X} = [\mathbf{X}_1^T, \mathbf{X}_2^T, \ldots, \mathbf{X}_M^T]^T$, $\mathcal{D} = \text{blkdiag}[D]$, $b = (b_1^T, \ldots, b_M^T)^T$, and $\epsilon = (\epsilon_1^T, \ldots, \epsilon_N^T)^T$. The model in (2.1) can be rewritten in complete matrix notation as described in (2.2). The variance of $\mathbf{Y}$ is given by $W = \mathbf{Z}\mathcal{D}\mathbf{Z}^T + \sigma^2\mathbf{I}_{N\times N}$:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}b + \epsilon. \tag{2.2}$$

2.2 *Estimation in the fully observed cluster setting*

First, consider the usual linear mixed modeling setting in which cluster assignments are fully observed and the traditional EM approach to estimation of Laird and Ware (1982). This approach proceeds by first calculating the ML estimate of $\boldsymbol{\beta}$ assuming the current estimate of $\boldsymbol{\theta}$. This calculation is straightforward since a closed-form solution exists. Second, we update the estimate of $\boldsymbol{\theta}$ assuming the current estimates of $\boldsymbol{\beta}$. Estimation at this step proceeds using an EM algorithm, which involves first determining the conditional expectation of the complete-data log-likelihood (E-step) and then maximizing this to arrive at new parameter estimates (M-step). This process is then repeated iteratively until a convergence criterion is met.

If the variance parameters are known, the ML estimate of $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}}$ in (2.3). In general, $\boldsymbol{\theta}$ is not known and we replace W in this equation with its ML estimate given by $\widehat{W} = \mathbf{Z}\widehat{\mathcal{D}}\mathbf{Z}^T + \widehat{\sigma}^2\mathbf{I}$. Based on the complete-data likelihood, where the complete data consist of $\mathbf{Y}$, b, and ϵ, the sufficient statistics for $\boldsymbol{\theta} = [\sigma^2, D]$ are given by $t_1 = \sum_{i=1}^M \epsilon_i^T\epsilon_i$ and $t_2 = \sum_{i=1}^M b_i b_i^T$. The M-step of the EM algorithm is composed of arriving at ML estimates $\widehat{\sigma}^{2(k+1)} = t_1^{(k)}/N$ and $\widehat{D}^{(k+1)} = t_2^{(k)}/M$ assuming the current estimate of $\Omega = (\boldsymbol{\beta}, \boldsymbol{\theta})$:

$$\widehat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}^T W^{-1}\mathbf{X})^{-1}\mathbf{X}^T W^{-1}\mathbf{Y}. \tag{2.3}$$

The E-step involves setting the sufficient statistics $t_1^{(k)}$ and $t_2^{(k)}$ equal to their expectation conditional on the observed data $\mathbf{Y}$, as summarized in (2.4). It is straightforward to show $E(\epsilon_i^T\epsilon_i|\mathbf{Y}_i, \widehat{\boldsymbol{\theta}}^{(p)}) = \text{tr}\{\widehat{\sigma}^4\widehat{W}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})^T\widehat{W}_i^{-1} + \widehat{\sigma}^2[\mathbf{I} - \widehat{\sigma}^2\widehat{W}_i^{-1}]\}$ and $E(b_i b_i^T|\mathbf{Y}, \widehat{\Omega}) = \widehat{D}\mathbf{Z}_i^T\widehat{W}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})^T\widehat{W}_i^{-1}\mathbf{Z}_i\widehat{D} + \widehat{D} - \widehat{D}\mathbf{Z}_i^T\widehat{W}_i^{-1}\mathbf{Z}_i\widehat{D}$. Restricted maximum likelihood (REML) estimates are obtained by adding $\text{Var}(E(\epsilon_i|\mathbf{Y}_i, \widehat{\Omega})) = \widehat{\sigma}_\epsilon^4\widehat{W}_i^{-1}\mathbf{X}_i\text{Var}(\widehat{\boldsymbol{\beta}})\mathbf{X}_i^T\widehat{W}_i^{-1}$ and $\text{Var}(E(b_i|\mathbf{Y}_i, \widehat{\Omega})) = \widehat{D}\mathbf{Z}_i^T\widehat{W}_i^{-1}\mathbf{X}_i\text{Var}(\widehat{\boldsymbol{\beta}})\mathbf{X}_i^T\widehat{W}_i^{-1}\mathbf{Z}_i\widehat{D}$ to each equation, respectively, where $\text{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}_i^T\widehat{W}_i^{-1}\mathbf{X}_i)^{-1}$.

These additional terms account for estimation of the mean parameter $\boldsymbol{\beta}$:

$$t_1^{(k)} = E(t_1|\mathbf{Y}, \widehat{\Omega}^{(k)}) = \sum_{i=1}^{M} E(\epsilon_i^T \epsilon_i|\mathbf{Y}_i, \widehat{\Omega}^{(k)}),$$
$$t_2^{(k)} = E(t_2|\mathbf{Y}, \widehat{\Omega}^{(k)}) = \sum_{i=1}^{M} E(b_i b_i^T|\mathbf{Y}_i, \widehat{\Omega}^{(k)}). \tag{2.4}$$

3. Methods

In this section, we extend the methods described in Section 2.2 to handle ambiguity in the correlated data indicators. That is, we assume that the i of (2.1) is not observed. Since estimation of $\boldsymbol{\beta}$ requires knowledge of the unobserved cluster assignments, an additional implementation of the EM algorithm is required at the first step. Specifically, estimation of $\boldsymbol{\beta}$ will depend on weights equal to the estimated posterior probabilities of each potential cluster assignment given the observed data. This requires first assuming a distribution for the number of observations in each cluster as described in Section 3.1 below. We let this distribution be a function of the parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$, where α_i is the population frequency of cluster i. We then proceed similarly to the unambiguous setting by first estimating the mean components $\Phi = [\boldsymbol{\beta}, \boldsymbol{\alpha}]$ assuming $\boldsymbol{\theta}$ is known and, second, estimating the variance parameters $\boldsymbol{\theta}$ given the current estimate of Φ.

3.1 *Defining a distribution for cluster counts*

We assume that the probability of a particular configuration of cluster assignments (represented by the design matrix $\mathbf{Z}$) follows a multinomial distribution. This probability density is given explicitly in (3.1), where $n_{Z,i}$ is the number of observations in cluster i for the given $\mathbf{Z}$, $\alpha = (\alpha_1, \ldots, \alpha_M)$, α_i is the population frequency of cluster i, and $\sum_{i=1}^{M} \alpha_i = 1$ or, equivalently, $\alpha_M = 1 - \sum_{i=1}^{M-1} \alpha_i$. Note that the usual constant term $(N!/n_1! \cdots n_M!)$ is not included in this formula since the probability is for a single configuration $\mathbf{Z}$:

$$\Pr(\mathbf{Z}|\alpha) = \prod_{i=1}^{M} \alpha_i^{n_{Z,i}}. \tag{3.1}$$

The number of clusters, given by M, is assumed to be known. This is a reasonable assumption in most data settings. For example, in Section 4 clusters are formulated based on pairs of haplotypes; the number of possible pairs is a fixed number that depends on the number of SNPs under investigation within a gene. Alternatively, clusters may represent hospitals or schools and the number of such units is generally fixed at the onset of a study.

3.2 *Estimating mean parameters, conditional on $\boldsymbol{\theta}$*

The complete data consist of $\mathbf{Y}$, $\mathbf{Z}$, b, and ϵ and are denoted $\mathbf{Y}_{\text{complete}}$. In estimating the mean parameters, we treat b and ϵ as known and write the complete-data likelihood for $\Phi = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ given $\boldsymbol{\theta}$ as $L_c(\Phi|\mathbf{Y}_{\text{complete}}, \boldsymbol{\theta})$ in (3.2). Here, $\Pr(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the marginal conditional density for the observed data $\mathbf{Y}$ and is given by (3.3). Note that the particular configuration of cluster assignments will contribute to W

and we therefore include an additional Z subscript in our notation:

$$L_{\mathrm{c}}(\Phi|\mathbf{Y}_{\mathrm{complete}}, \boldsymbol{\theta}) = \Pr(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta})\Pr(\mathbf{Z}|\alpha), \tag{3.2}$$

$$\begin{aligned}\Pr(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta}) &= L(\beta|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}) \\ &= \frac{1}{|W_Z|^{1/2}} \exp\{-1/2(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T W_Z^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\}.\end{aligned} \tag{3.3}$$

The E-step involves calculating the conditional expectation of the complete-data log-likelihood. This conditional expectation is given in (3.4), where $p_Z(\Omega)$ is the posterior probability of the combination of cluster assignments (again denoted by the design matrix $\mathbf{Z}$) given $\mathbf{Y}$ and $\Omega = (\Phi, \boldsymbol{\theta})$. Recall that $\mathcal{S}$ is the set of all design matrices $\mathbf{Z}$ that are consistent with the observed data. A formulation of this posterior probability is given in (3.5). At this step, we update our estimate of $p_Z(\Omega)$ assuming the current estimate of Ω. That is, we calculate $p_Z(\widehat{\Omega}^{(k)})$, where $\widehat{\Omega}^{(k)} = [\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\sigma}_\epsilon^{(k)}, \widehat{D}^{(k)}]$ is the vector of current parameter estimates:

$$E[\log L_{\mathrm{c}}(\Phi)|\mathbf{Y}, \boldsymbol{\theta}] = \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\Omega)[\log \Pr(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta}) + \log \Pr(\mathbf{Z}|\alpha)], \tag{3.4}$$

$$p_Z(\Omega) = p_\Omega(\mathbf{Z}|\mathbf{Y}) = \frac{\Pr(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta})\Pr(\mathbf{Z}|\alpha)}{\sum_{\mathbf{Z}\in\mathcal{S}} \Pr(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta})\Pr(\mathbf{Z}|\alpha)}. \tag{3.5}$$

The M-step involves maximizing the conditional expectation of the complete-data log-likelihood conditional on the current estimate of the posterior probabilities $p_Z(\widehat{\Omega}^{(k)})$ calculated in the E-step. Maxima can be obtained by maximizing the system of equations given in (3.6) and (3.7). Here, we use the relationship that the derivative of the conditional expectation is equal to the conditional expectation of the score function. Resulting closed-form solutions for $\widehat{\alpha}_i$ and $\widehat{\boldsymbol{\beta}}$ are given in (3.8) and (3.9):

$$\begin{aligned}\frac{\partial E[\log L_{\mathrm{c}}(\Phi)|\mathbf{Y}_{\mathrm{obs}}, \boldsymbol{\theta}]}{\partial \boldsymbol{\beta}} &= E\left[\frac{\partial \log L_{\mathrm{c}}(\Phi)}{\partial \boldsymbol{\beta}}\middle|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}\right] \\ &= \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}^{(k)})\frac{\partial \log L(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}^{(k)})\mathbf{X}^T W_Z^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),\end{aligned} \tag{3.6}$$

$$\begin{aligned}\frac{\partial E[\log L_{\mathrm{c}}(\Phi)|\mathbf{Y}_{\mathrm{obs}}, \boldsymbol{\theta}]}{\partial \alpha_i} &= E\left[\frac{\partial \log L_{\mathrm{c}}(\Phi)}{\partial \alpha_i}\middle|\boldsymbol{Y}, \mathbf{Z}, \boldsymbol{\theta}\right] \\ &= \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}^{(k)})\frac{\partial \log \Pr(\mathbf{Z}|\alpha)}{\partial \alpha_i} \\ &= \frac{\sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}^{(k)})n_{Z,i}}{\alpha_i} - \frac{\sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}^{(k)})n_{Z,M}}{\alpha_M},\end{aligned} \tag{3.7}$$

$$\widehat{\beta}^{(k+1)} = \left(\sum_{\mathbf{Z} \in \mathcal{S}} p_Z(\widehat{\Omega}^{(k)}) \mathbf{X}^T W_Z^{-1} \mathbf{X} \right)^{-1} \sum_{\mathbf{Z} \in \mathcal{S}} p_Z(\widehat{\Omega}^{(k)}) \mathbf{X}^T W_Z^{-1} \mathbf{Y}, \tag{3.8}$$

$$\widehat{\alpha}_i^{(k+1)} = \frac{\sum_{\mathbf{Z} \in \mathcal{S}} p_Z(\widehat{\Omega}^{(k)}) n_{Z,i}}{N}. \tag{3.9}$$

In the case that the variance parameters are known, we iterate between updating our estimates of $p_Z(\Omega)$ and updating our estimates of Φ. The EM algorithm ensures that we will increase the likelihood at each iteration. In general, the variance components $\boldsymbol{\theta}$ are not known; however, we can obtain ML estimates of $\boldsymbol{\theta}$ and condition on these estimates. This amounts to substituting these ML estimates into the above equations. In the following paragraphs, we describe a modification of the EM algorithm for estimation of $\boldsymbol{\theta}$ that additionally incorporates posterior probabilities associated with each combination of cluster assignments.

3.3 *Estimating variance components, conditional on* Φ

For the purpose of estimating variance parameters, we define the complete-data log-likelihood by $L_c(\boldsymbol{\theta}|\mathbf{Y}_{\text{complete}}, \Phi)$ in (3.10). Note $f(\mathbf{Y}|b, \epsilon, \boldsymbol{\beta})$ is a dirac function (=1 under model) and only depends on $\boldsymbol{\beta}$ so is ignored in estimation of $\boldsymbol{\theta}$. Using the same approach as described in Section 2.2, we set the sufficient statistics for $\boldsymbol{\theta}$ equal to their expectation. Here, we sum additionally over the set of all design matrices $\mathbf{Z}$ that are consistent with the observed data and weight by corresponding posterior probabilities $p_Z(\widehat{\Omega}^{(k)})$. Again the maximization step involves setting $\widehat{\sigma}^{2(k+1)} = t_1^{(k)}/N$ and $\widehat{D}^{(k+1)} = t_2^{(k)}/\tilde{M}$, where $\tilde{M} = \sum_{Z \in \mathcal{S}} p_Z(\widehat{\Omega}^{(k)}) M_Z$ and M_Z is the number of clusters corresponding to the specific configuration given by Z. Adjustments to these equations to arrive at REML estimates proceed as described in Section 2.2:

$$\begin{aligned} L_c(\theta|\mathbf{Y}_{\text{complete}}, \Phi) &= \Pr(\mathbf{Z}|\alpha) \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) \\ &= \Pr(\mathbf{Z}|\alpha) f(\mathbf{Y}|b, \epsilon, \boldsymbol{\beta}) f(b|D) f(\epsilon|\sigma^2) \\ &\propto \left(\frac{\mathrm{e}^{-1/2 \sum_{i=1}^M b_i^T D b_i}}{|D|^{M/2}} \right) \left(\frac{\mathrm{e}^{-1/2 \sum_{i=1}^M \epsilon_i^T \epsilon_i / \sigma^2}}{\sigma^N} \right), \end{aligned} \tag{3.10}$$

$$\begin{aligned} t_1^{(k)} &= E(t_1|\mathbf{Y}, \widehat{\Omega}^{(k)}) = \sum_{Z \in \mathcal{S}} p_Z(\widehat{\Omega}^{(k)}) \sum_{i=1}^M E(\epsilon_i^T \epsilon_i|\mathbf{Y}_i, \widehat{\Omega}^{(k)}), \\ t_2^{(k)} &= E(t_2|\mathbf{Y}, \widehat{\Omega}^{(k)}) = \sum_{Z \in \mathcal{S}} p_Z(\widehat{\Omega}^{(k)}) \sum_{i=1}^M E(b_i b_i^T|\mathbf{Y}_i, \widehat{\Omega}^{(k)}). \end{aligned} \tag{3.11}$$

3.4 *Summary of approach*

In summary, ML estimation proceeds by iterating between 2 EM algorithms: (1) estimation of Φ and (2) estimation of $\boldsymbol{\theta}$. For computational efficiency, we implement one iteration of the first EM algorithm conditional on the current estimate of $\boldsymbol{\theta}$ and then one iteration of the second EM algorithm conditional on

the current estimate of Φ. This is then repeated until a convergence criterion is met. Initial values for the parameter estimates are arrived at by randomly assigning clusters in the case of ambiguity and fitting the usual mixed-effects model. This approach is summarized in the following step-by-step procedure:

1. Initialize $k = 0$. Determine initial values for $\widehat{\Omega}^{(k)}$ by randomly assigning cluster membership in the case of ambiguity and fitting usual mixed effects model.
2. Calculate $p_Z(\widehat{\Omega}^{(k)})$ based on Equation 3.5 using the current estimate of $\Omega = \widehat{\Omega}^{(k)} = (\widehat{\Phi}^{(k)}, \widehat{\theta}^{(k)})$.
3. Update $\widehat{\beta}$ and $\widehat{\alpha}$ using Equations 3.8 and 3.9, assuming the current estimates of $p_Z(\Omega) = p_Z(\widehat{\Omega}^{(k)})$ and $\theta = \widehat{\theta}^{(k)}$. Denote the new estimate for Φ by $\widehat{\Phi}^{(k+1)} = (\widehat{\beta}^{(k+1)}, \widehat{\alpha}^{(k+1)})$.
4. Update $\widehat{\sigma}_\epsilon$ and $\widehat{D}$ using Equation 3.11, assuming the current estimates of $\Phi = \Phi^{(k+1)}$, $p_Z(\Omega) = p_Z(\widehat{\Omega}^{(k)})$ and $\theta = \widehat{\theta}^{(k)}$. Denote the new estimate for θ by $\widehat{\theta}^{(k+1)} = (\widehat{\sigma}_\epsilon^{(k+1)}, \widehat{D}^{(k+1)})$.
5. Let $k = k + 1$ and repeat steps (2)–(4) until a convergence criterion is met.

4. CONSIDERATIONS FOR GENETIC ASSOCIATION STUDIES

Data arising from genetic association studies of unrelated individuals are generally composed of 3 components: (1) one or more outcomes (commonly referred to as phenotypes, these can be continuous or a binary indicator for case–control status), (2) covariates and potential confounders, including clinical and demographic factors, and (3) genotypes, consisting of the pair of nucleotides present at each locus within and across the candidate genes under consideration. In general, the alignment of nucleotides on a single chromosomal copy, commonly referred to as haplotypic phase, is not observable. For example, if an individual is heterozygous at 2 loci within a gene so that their observed genotype is (Aa, Bb), then the corresponding possible haplotype pairs for this individual are (AB, ab) or (Ab, aB).

4.1 *Defining clusters*

The mixed modeling approach to the analysis of genetic association studies begins by grouping individuals into clusters so that individuals within the same cluster have similar or identical underlying genetic compositions. For example, in Foulkes *and others* (2005) individuals with identical multilocus genotypes (i.e. the same pattern of SNPs across multiple loci within or across a gene) are deterministically assigned to a corresponding cluster. Once these clusters are defined, analysis proceeds using the mixed-effects modeling framework just as it would in a typical clustered data setting. In the context of studying haplotypic variations, such a grouping based on genetic compositions is generally unobservable. That is, just as haplotypic phase is unobservable, cluster assignments based on this information must also be unobservable.

More formally, suppose $\mathcal{H} = (h_1, \ldots, h_K)$ represents the set of all haplotype pairs (diplotypes) consistent with the observed genotypes for a given gene. We define clusters $C_1, \ldots, C_M$ such that an individual with haplotype combination contained in $\mathcal{H}_i$ belongs to cluster C_i, where $\mathcal{H}_i \subset \mathcal{H}$. In the most general case, we assume that the number of clusters equals the number of haplotype pairs, that is, $K = M$ and $\mathcal{H}_i = h_i$ so that all individuals within the same cluster have identical diplotypes. For example, in the case of 2 SNPs within a gene, there are 4 possible haplotypes and $K = 10$ possible diplotypes. In the general case, we define 10 corresponding clusters.

Several alternative formulations of the clusters are tenable, and these can reflect the biological hypothesis under investigation. For example, returning to the simple case above, we can group all diplotypes that contain at least one copy of the rare haplotype into a single cluster. This would result in 7 clusters and is consistent with a dominant genetic model in which one copy of the disease allele results in an altered phenotype. A visual representation of these 2 approaches to defining clusters is given in Figure 1.

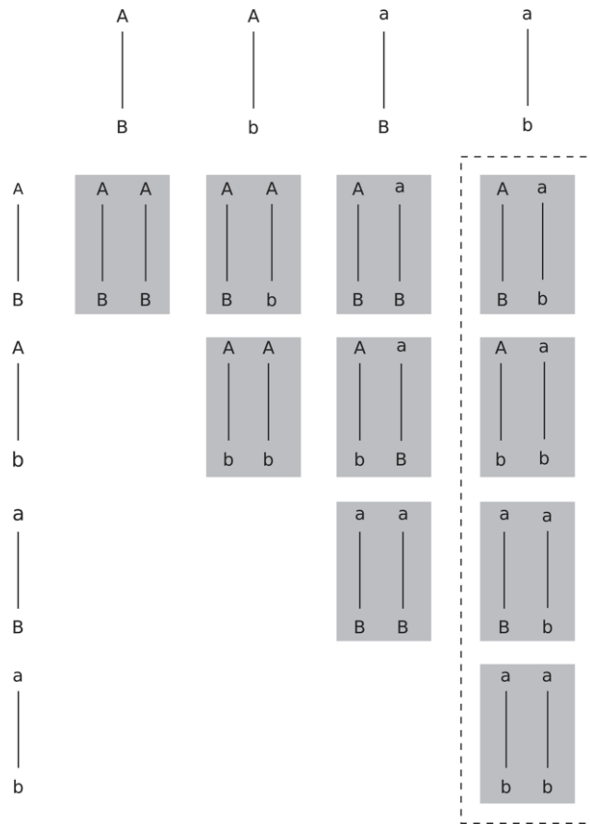

Fig. 1. Sample approaches to defining clusters. For the 2 SNP example in which the observed genotypes are (AA, Aa, or aa) and (BB, Bb, or bb), there are 4 possible haplotypes, AB, Ab, aB, and ab, and 10 possible diplotypes. The most general approach to defining clusters results in 10 clusters consisting of all these possible combinations of 2 haplotypes. These are indicated by shaded rectangles. An alternative approach groups all diplotypes with at least one copy for the rare ab haplotype into a single cluster. This is indicated by the dashed rectangle that combines 4 of the previously defined clusters into a single cluster. In this case, there are a total of 7 clusters.

4.2 *Estimation*

Returning to the model in (2.1), $\mathbf{Z}_i$ again indicates membership to cluster i (or equivalently for the general case, presence of haplotype pair h_i) and is potentially unobservable. The observed data G that inform the cluster memberships are the observed genotypes. Recall that the population frequency of each cluster is given by α_i for $i = 1, \ldots, M$. Under the assumption of HWE, the probability of a pair of haplotypes is equal to the product of the corresponding marginal frequencies.

In our setting, the HWE assumption is not required since we estimate cluster-level (diplotype) probabilities. This results from the fact that we allow for 2 components of the data to inform our estimation of cluster frequencies. The first is those individuals whose cluster membership is unambiguous and the second is the phenotype ($\mathbf{Y}$). In the special case that we are estimating the frequencies of clusters that are completely ambiguous, that is, all individuals within the clusters are ambiguous, then we rely solely on $\mathbf{Y}$ for this purpose, unless we make an additional assumption such as HWE. In this extreme case, while we are able to estimate cluster frequencies and calculate corresponding empirical Bayes estimates of random

effects, it is not possible to distinguish which values correspond to which clusters without additional assumptions. Notably, the omnibus test for overall variability in the random cluster effects is still valid.

If the HWE assumption is reasonable, the proposed method can be refined further to define cluster probabilities as the product of the corresponding 2 haplotype probabilities. That is, (3.1) can be reexpressed as described in (4.1), where H_j represents a single haplotype, M^* is the number of unique haplotypes, and $\tilde{n}_{Z,j} = \sum_i [n_{Z,i}(1 + \mathbf{I}[C_i = \{H_j, H_j\}])]^{H_j \in C_i}$ is the number of copies of H_j observed across all clusters for the configuration given in $\mathbf{Z}$. The ML estimate of δ_j is as given in (3.9), where $n_{Z,i}$ is replaced with $\tilde{n}_{Z,j}$. For the simple example described in Figure 1, under the HWE assumption the number of frequency parameters reduces from $M = 10$ to $M^* = 4$. Note, however, that in both cases, the number of random effects is 10 since there are 10 clusters. Sensitivity of this approach to violations of HWE is described in Section 5.1 and Appendix 3:

$$\Pr(\mathbf{Z}|\alpha) = \prod_{i=1}^{M} \alpha_i^{n_{Z,i}} = \prod_{j=1}^{M^*} \delta_j^{\tilde{n}_{Z,j}}. \tag{4.1}$$

4.3 *Testing and prediction*

For the case in which we are interested in overall genetic effects and not interactions between genes and other covariates, b_i reduces to a scalar and $\mathbf{Z}_i = \mathbf{1}_{n_i}$ is an $n_i \times 1$ vector of 1s. In the application of mixed modeling to genetic data, interest may lie in testing for significant variability across random effects (e.g. $H_0 : \sigma_b^2 = 0$). A likelihood ratio test comparing the expected complete-data log-likelihood for the full model (with random cluster effects) to the reduced model (without random cluster effects) can be applied. Finally, empirical Bayes estimates of the random effects inform us about the cluster-specific effects on the phenotype under consideration. These are calculated in the usual way with additional weights equal to the posterior probabilities of cluster assignments and given in (4.2):

$$\widehat{b_i} = E(b_i|\mathbf{Y}_i) = \sum_{\mathbf{Z} \in \mathcal{S}} p_Z(\widehat{\Omega}^{(k)}) \mathcal{D} \mathbf{Z}^T W_Z^{-1} (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{4.2}$$

4.4 *Computational and identifiability considerations*

Suppose heterozygosity is observed in our sample at exactly r sites for the gene under investigation. In this case, the number of possible haplotypes is 2^r and the number of haplotype pairs (clusters) is $R = \binom{2^r}{2} + \binom{2^r}{1}$. Notably, some clusters consist only of individuals whose haplotypic phase is fully determined. For example, consider the clusters illustrated in Figure 1. The cluster consisting of the haplotype pair (AB, Ab) corresponds uniquely to individuals with the genotype (AA, Bb). Since this genotype has heterozygosity at only a single site, the corresponding haplotypic phase is known. That is, the haplotypic phase of individuals within the (AB, Ab) cluster is completely observed. On the other hand, for the example provided, the true cluster assignment for individuals with uncertainty in phase will be either (AB, ab) or (Ab, aB). In general, $R^* = \binom{2^r}{2} - r2^{r-1}$ of the R clusters consists of individuals for whom haplotypic phase is ambiguous.

If there are K individuals with ambiguous phase, then the number of possible cluster assignment configurations is $|\mathcal{S}| = (R^*)^K$. This is the number of elements $\mathbf{Z}$ of the set $\mathcal{S}$ in (3.4)–(4.2). For the simple case in which $r = 2$, we have $R^* = 2$ and $|\mathcal{S}| = 2^K$. Thus, the computational burden of the proposed modeling approach is clearly quite large; however, a few matrix identities help to reduce the computational intensity. Specifically, suppose each individual has a single random effect so that $\mathbf{Z} = \mathbf{1}_{n_{Z,i}}$

and $\mathcal{D} = \sigma_b^2\mathbf{I}$. In this case, we can write W_Z^{-1} as in (4.3), where $J_{n_{Z,i}}^{n_{Z,i}} = \mathbf{1}_{n_{Z,i}}\mathbf{1}_{n_{Z,i}}^T$. A formal derivation of this identity is given in Appendix 1(a). Note that W_Z^{-1} depends only on the number of individuals within each cluster and not the specific configuration of the individuals. Since multiple elements of $\mathcal{S}$ yield the same numbers of observations per cluster, use of (4.3) reduces the number of calculations of W^{-1} from 2^K to $K+1$:

$$W_Z^{-1} = \sigma_\epsilon^{-2}\left(\mathbf{I} - \text{blkdiag}\left[\frac{\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2} J_{n_{Z,i}}^{n_{Z,i}}\right]\right). \tag{4.3}$$

If we further assume $X = \mathbf{1}_N$ so that the model given in (2.1) consists of an intercept and no covariates, then $\widehat{\boldsymbol{\beta}}$ reduces to (4.4). A detailed derivation is provided in Appendix 1(b). The sum over $\widetilde{N} \in S$ now represents a sum over all combinations of cluster sizes and $\dot{p}_{\widetilde{N}}(\widehat{\Omega})$ is the sum of $p_Z(\widehat{\Omega})$ for all Z consistent with the configuration $\widetilde{N}$. Note that $\sum_{j=1}^{n_{Z,i}} \mathbf{Y}_{ij}$ does depend on the particular configuration of cluster assignments and thus must be determined for each $Z \in \mathcal{S}$. Again calculation of the inverse (the first term in the product in (4.4)) depends only on the number of individuals per cluster, thus reducing the number of computations substantially:

$$\widehat{\beta} = \left\{ N - \sum_{\widetilde{N}\in\mathcal{S}} \dot{p}_{\widetilde{N}}(\widehat{\Omega}) \sum_{i=1}^{M} n_{\widetilde{N},i}^2 \left[\frac{\sigma_b^2}{n_{\widetilde{N},i}\sigma_b^2 + \sigma_\epsilon^2}\right]\right\}^{-1}$$

$$\times \left\{ N\bar{Y} - \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} \left[\frac{n_{Z,i}\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2} \sum_{j=1}^{n_{Z,i}} \mathbf{Y}_{ij}\right]\right\}. \tag{4.4}$$

Gains in computational efficiency can also be achieved by partitioning $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ into their ambiguous and unambiguous components. Let $\mathbf{Y}^T = [\mathbf{Y}_\text{a}^T|\mathbf{Y}_\text{u}^T]$, $\mathbf{X}^T = [\mathbf{X}_\text{a}^T|\mathbf{X}_\text{u}^T]$, and $\mathbf{Z} = \begin{pmatrix} Z_\text{a} & 0 \\ 0 & Z_\text{u} \end{pmatrix}$, where the subscripts a and u indicate subsets of the observed data corresponding to individuals whose cluster assignments are ambiguous and unambiguous, respectively. Since W is block diagonal (clusters are assumed independent), W^{-1} can be partitioned similarly into $W^{-1} = \begin{pmatrix} W_\text{a}^{-1} & 0 \\ 0 & W_\text{u}^{-1} \end{pmatrix}$, where $W_\text{a}^{-1} = (Z_\text{a}\mathcal{D}Z_\text{a}^T + \sigma_\epsilon^2\mathbf{I})^{-1}$ and $W_\text{u}^{-1} = (Z_\text{u}\mathcal{D}Z_\text{u}^T + \sigma_\epsilon^2\mathbf{I})^{-1}$. We can now write (3.8), (3.9), and (3.11) in terms of sums of ambiguous and unambiguous components, as described in Appendix 2(a)–(c). The unambiguous components need only be calculated once while the ambiguous components depend on $Z \in \mathcal{S}$.

The most general approach to defining clusters in the haplotype setting, as described in Figure 1, results in all individuals with ambiguity in phase belonging to a subset of clusters while no fully observed individuals belong to clusters in this subset. As noted in Section 4.2, estimation of cluster-level frequencies in this setting relies solely on the response variable and is not identifiable in the sense that we cannot distinguish which frequencies and empirical Bayes estimates correspond to which particular clusters. While the omnibus test for overall variability is valid in this setting, estimation of haplotype frequencies under the HWE assumption may be more relevant. In the general setting of missing correlated data identifiers, this would represent an extreme case.

5. Data results

5.1 *Simulation study*

In order to evaluate the mixed modeling approach for characterizing haplotype–trait associations, we conducted a simulation study that includes the following components: (1) a sensitivity analysis of the mixed

modeling approach to the number of clusters (haplotypes) and model misspecification. Both founder effect models (assuming dominant and recessive traits) and departures from HWE are considered. (2) A comparison to alternative methods, including a traditional analysis of variance (ANOVA) approach and a 2-stage multiple imputation (MI) approach (Foulkes *and others*, 2007). (3) Detailed simulation findings, including power, coverage rates (CRs), and false-discovery rates (FDRs) for varying effect sizes (ratios of standard deviations) and degrees of ambiguity in cluster assignments. Details of the simulation study and corresponding results are provided in Appendix 3.

Briefly, we found that the mixed modeling approach has reasonable power for a sample size of $n = 200$ and between 10 and 36 clusters (4–8 haplotypes). Performance is relatively poor, however, under misspecification of the random-effects distribution. The reduction in power is especially pronounced for the recessive founder model in which only a single cluster effect arises from a normal distribution with >0 variability and the remaining cluster effects have 0 variability. On the other hand, power appears stable under moderate deviations from HWE when this is assumed. For a ratio of standard deviations (σ_b/σ_ϵ) of 0.4 and sample size of $n = 200$, power for the ANOVA and mixed modeling approaches is comparable while the number of clusters is less than 20. For more than 20 clusters, power of the mixed modeling approach (based on the single degree of freedom test) is greater. Power for the 2-stage MI approach is comparable to the fully likelihood–based approach described herein for one data example. Modest reductions in power are observed for increasing ambiguity in cluster identifiers while corresponding CRs for variance parameters are lower. Finally, FDRs increase from 5% to 7–9% with an increase in cluster ambiguity up to 20%.

5.2 *Example*

Recent studies indicate that long-term exposure to certain combination antiretroviral therapies (ARTs) may lead to a host of lipid abnormalities including increases in triglycerides and total cholesterol and a reduction in high-density lipoprotein cholesterol (HDL-C). In turn, this can lead to accelerated onset of cardiovascular disease and death, presenting a grave concern for HIV-1-infected individuals receiving continuous long-term therapy. However, the large number of available ARTs provides a great potential to tailor treatments to individual-level characteristics. Furthermore, understanding the characteristics of individuals at greatest risk for cardiovascular complications will provide clinicians the opportunity to target interventions, such as administration of lipid-lowering therapies.

The data motivating our research arise from a cohort of $N = 626$ HIV-1-infected individuals at risk for ART-associated dyslipidemia. These data were collected as part of multiple AIDS Clinical Trials Group studies and combined under New Work Concept Sheet 224. The primary aim of this study is to identify genetic factors that predict lipid abnormalities after controlling for traditional risk factors and other clinical parameters, including age, sex, use of lipid-lowering therapy, and current ART exposure. First-stage analysis results and general descriptive information on the cohort are provided in Foulkes *and others* (2006). This analysis revealed potential effect modification by race/ethnicity, and so for the purpose of illustration, we describe here application of the above method within Hispanics ($N = 109$).

The effects of haplotypic variation in endothelial lipase (EL) on HDL-C are considered. The SNPs chosen for analysis are rs12970066, Asn396Ser, and rs3829632 (-1309A/G) and were determined based on prior knowledge of association with plasma lipoproteins and for capturing genetic variability within this gene. A haplotype-based analysis can be advantageous if the observed SNPs are in linkage disequilibrium with the disease-causing variant and are not themselves functional; in general and in this setting, the functionality of specific SNPs is not fully characterized, and thus, a haplotype-based analysis can provide new insight. We assume HWE within the single race/ethnicity group and apply the ECM approach described in Section 4.2. A summary of genotype frequencies is given in Table 1. In this sample, $N = 13$ individuals have uncertainty in haplotypic phase due to heterozygosity at rs12970066 (AG)

Table 1. *EL genotype within Hispanics. Genotype counts for combination of* 3 *SNPs in EL. Although variability in rs3829632 is not observed within the subset of Hispanics, this SNP is included in the presentation for completeness*

	EL genotypes			Count (%)
	rs12970066	Asn396Ser	rs3829632 (-1309A/G)	
1	AA	CC	AA	23 (0.21)
2	AA	CG	AA	24 (0.22)
3	AA	GG	AA	4 (0.04)
4	AG	CC	AA	31 (0.28)
5	AG	CG	AA	13 (0.12)
6	GG	CC	AA	14 (0.13)
				Total: 109

Table 2. *Estimated haplotype frequencies within Hispanics. Estimated haplotype frequencies based on application of mixed model–fitting procedure assuming HWE*

	EL haplotypes			Estimated frequency
	rs12970066	Asn396Ser	rs3829632(-1309A/G)	
1	A	C	A	0.470
2	A	G	A	0.205
3	G	C	A	0.325
4	G	G	A	<0.001

and Asn396Ser (CG). Notably, variability is not observed in the third SNP (rs3829632) within Hispanics; however, we include this SNP in our presentation for completeness. Covariates included in model fitting are age, gender, CD4 count, current ART exposure, use of lipid-lowering therapy, and study. $N = 100$ individuals with complete data are included in the analysis.

A convergence criterion of a maximum absolute percentage change in parameter estimates from one iteration to the next of $<1 \times 10^{-5}$ is used. Convergence is met after 20 iterations. Resulting haplotype frequency estimates are given in Table 2. The estimated variance of the random haplotype effects is $\widehat{\sigma}_b^2 = 0.013$ with a corresponding likelihood ratio test statistic of $\chi^2_{1,0} = 6.69$ ($p < 0.05$). The estimated error variance is $\widehat{\sigma}_\epsilon^2 = 0.057$. Empirical Bayes estimates of the random haplotype pair effects are given in Figure 2. These results suggest overall variability in the haplotypic effects of EL on HDL-C. The cluster (ACA, AGA) has the largest absolute estimated effect, suggesting that individuals with this pair of haplotypes will have a lower predicted HDL-C level. Since HDL-C is considered as the good cholesterol, these individuals may be at greatest risk for ART associated lipid complications and candidates for targeted intervention.

6. Discussion

In this manuscript, we describe an ECM approach to finding ML parameter estimates in the linear mixed-effects model setting when the correlated data indicators are ambiguous. This research was motivated by interest in characterizing genetic effects on a phenotype when haplotypic phase is unobservable. The proposed approach, however, has broader relevance to other settings in which cluster identifiers are not known with certainty. Notably, a similar approach can be applied to missing genotype data, where multilocus genotype group identifiers are treated as ambiguous.

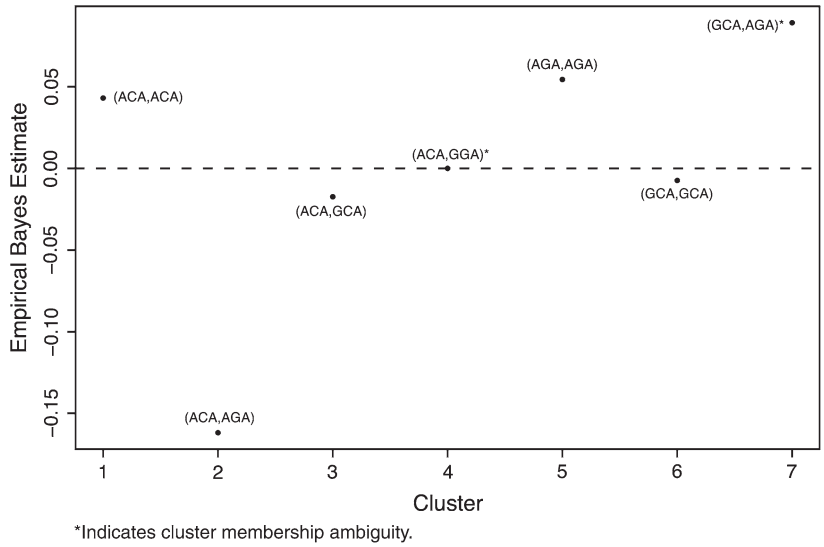

Fig. 2. Empirical Bayes predictions of random EL cluster effects. Asterisk indicates cluster membership ambiguity.

We focused on the simple linear mixed-effects model with a single random cluster effect. Alternative formulation of the design matrix for the random effects allows for assessing interactions between patient-specific characteristics and haplotypes. For example, inclusion of an ART drug indicator in the $\mathbf{Z}$ matrix of (2.1) would allow us to investigate drug-by-gene effects in a pharmacogenomic study. Extensions to settings with alternative, noncontinuous outcomes and semiparametric mixed models that relax the normality assumption on the random effects require additional consideration. As expected, our simulation study suggests relatively poor performance in the context of a recessive, founder model in which the normality assumption of the random effects is severely violated. Further investigations of performance under alternative model formulations, as well as explorations of the utility of semiparametric procedures and model diagnostics in these settings, would be interesting.

In another recent manuscript, we describe an MI approach for this setting in which the haplotype frequencies are estimated independently of the outcome (Foulkes *and others*, 2007). An EM algorithm as described by Excoffier and Slatkin (1995) can be applied for haplotype reconstruction and then multiple imputed data sets derived by repeated weighted sampling based on the estimated posterior probabilities. The primary advantage of the joint approach we describe herein is that it incorporates information about the phenotype in the estimation procedure. For example, if ambiguity rests between 2 clusters with effects b_i and $b_{i'}$, where $b_i > b_{i'}$, then this approach will tend to assign individuals with higher observed phenotypes to C_i and individuals with lower phenotypes to $C_{i'}$. Notably, in one simple simulation study, the 2 approaches yielded similar results while the computational burden associated with the joint approach is much greater. In light of the theoretical advantages, however, further and extensive consideration of alternative settings and the extent to which incorporating this additional layer of information indeed results in greater efficiency is warranted.

A primary limitation of the mixed modeling approach for haplotype–trait association studies is that as the number of SNPs increases, the number of haplotypes (and therefore clusters) can quickly approach the number of individuals under investigation. In genome-wide association studies, taking a random sample of SNPs within a known disease pathway or genetic region may be tenable and appropriate for the random-effects modeling framework. Paradoxically, increasing the number of variables (SNPs) can also lead to

greater phase ambiguity in the data, suggesting an important trade-off between information gained from more accurate haplotype reconstruction and potential power loss associated with increasing ambiguity.

As mentioned in Section 1, the proposed approach represents an extension of SEMs with a doubly latent class structure defined by both latent random effects and unobservable class identifiers. Further extensions that draw on the literature of SEMs may provide additional tools for incorporating known biological function, such as gene-specific pathways to disease. For example, multiple random effects based on sets of genes with similar, known functionality may provide additional insight into the determinants of complex diseases. The framework we describe allows for this multivariable investigation while accounting for the unobservable nature of haplotypic phase in association studies.

Acknowledgments

We thank the AIDS Clinical Trials Group New Works Concept Sheet 224 study team for helpful discussions and providing access to data. *Conflict of Interest:* None declared.

Funding

National Institute of Allergy and Infection Diseases (NIAID) (AI056983); National Institute of Diabetes and Digestive and Kidney Diseases (DK021224); Adult AIDS Clinical Trials Group funded by the NIAID (AI38858); CRI: Computational Biology Facility for Western Massachusetts (CNS 0551500) to computing cluster.

APPENDIX 1

(A)

$$W_Z^{-1} = (\sigma_\epsilon^2\mathbf{I} + \mathbf{Z}\mathcal{D}\mathbf{Z}^T)^{-1} = (\sigma_\epsilon^2\mathbf{I} + \sigma_b^2\mathbf{Z}\mathbf{Z}^T)^{-1} = \sigma_\epsilon^{-2}\mathbf{I} - \sigma_\epsilon^{-4}\mathbf{Z}(\sigma_\epsilon^{-2}\mathbf{Z}^T\mathbf{Z} + \sigma_b^{-2}\mathbf{I})^{-1}\mathbf{Z}^T$$

$$= \sigma_\epsilon^{-2}\mathbf{I} - \sigma_\epsilon^{-4}\mathbf{Z}(\sigma_\epsilon^{-2}\mathrm{diag}[n_i] + \sigma_b^{-2}\mathbf{I})^{-1}\mathbf{Z}^T = \sigma_\epsilon^{-2}\mathbf{I} - \sigma_\epsilon^{-4}\mathbf{Z}\left(\mathrm{diag}\left[\frac{n_{Z,i}}{\sigma_\epsilon^2} + \frac{1}{\sigma_b^2}\right]\right)^{-1}\mathbf{Z}^T$$

$$= \sigma_\epsilon^{-2}\mathbf{I} - \sigma_\epsilon^{-4}\mathbf{Z}\mathrm{diag}\left[\frac{\sigma_\epsilon^2\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2}\right]\mathbf{Z}^T = \sigma_\epsilon^{-2}\left(\mathbf{I} - \mathbf{Z}\mathrm{diag}\left[\frac{\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2}\right]\mathbf{Z}^T\right)$$

$$= \sigma_\epsilon^{-2}\left(\mathbf{I} - \mathrm{blkdiag}\left[\frac{\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2}J_{n_{Z,i}}^{n_{Z,i}}\right]\right).$$

(B)

$$\widehat{\beta} = \left\{\sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega})(\mathbf{X}^T W_Z^{-1}\mathbf{X})\right\}^{-1} \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega})\mathbf{X}^T W_Z^{-1}\mathbf{Y}$$

$$= \left\{\sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega})\left(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathrm{blkdiag}\left[\frac{\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2}J_{n_{Z,i}}^{n_{Z,i}}\right]\mathbf{X}\right)\right\}^{-1}$$

$$\times\left\{\sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega})\left(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathrm{blkdiag}\left[\frac{\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2}J_{n_{Z,i}}^{n_{Z,i}}\right]\mathbf{Y}\right)\right\}$$

$$= \left\{ \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \left(N - \sum_{i=1}^{M} n_{Z,i}^2 \left[\frac{\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2} \right] \right) \right\}^{-1}$$

$$\times \left\{ \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \left(N\bar{Y} - \sum_{i=1}^{M} \left[\frac{n_{Z,i}\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2} \sum_{j=1}^{n_{Z,i}} \mathbf{Y}_{ij} \right] \right) \right\}$$

$$= \left\{ N - \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} n_{Z,i}^2 \left[\frac{\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2} \right] \right\}^{-1}$$

$$\times \left\{ N\bar{Y} - \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} \left[\frac{n_{Z,i}\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2} \sum_{j=1}^{n_{Z,i}} \mathbf{Y}_{ij} \right] \right\}$$

$$= \left\{ N - \sum_{\widetilde{N}\in\mathcal{S}} \dot{p}_{\widetilde{N}}(\widehat{\Omega}) \sum_{i=1}^{M} n_{\widetilde{N},i}^2 \left[\frac{\sigma_b^2}{n_{\widetilde{N},i}\sigma_b^2 + \sigma_\epsilon^2} \right] \right\}^{-1}$$

$$\times \left\{ N\bar{Y} - \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} \left[\frac{n_{Z,i}\sigma_b^2}{n_{Z,i}\sigma_b^2 + \sigma_\epsilon^2} \sum_{j=1}^{n_{Z,i}} \mathbf{Y}_{ij} \right] \right\}.$$

APPENDIX 2

(A)

$$\widehat{\beta} = \left(\sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \mathbf{X}^T W_Z^{-1} \mathbf{X} \right)^{-1} \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \mathbf{X}^T W_Z^{-1} \mathbf{Y}$$

$$= \left(\sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} \mathbf{X}_i^T W_{Z,i}^{-1} \mathbf{X}_i \right)^{-1} \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} \mathbf{X}_i^T W_{Z,i}^{-1} \mathbf{Y}_i$$

$$= \left\{ \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \left(\sum_{i=1}^{M_a} \mathbf{X}_{a,i}^T W_{a,Z,i}^{-1} \mathbf{X}_{a,i} + \sum_{i=1}^{M_u} \mathbf{X}_{u,i}^T W_{u,Z,i}^{-1} \mathbf{X}_{u,i} \right) \right\}^{-1}$$

$$\times \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \left(\sum_{i=1}^{M_a} \mathbf{X}_{a,i}^T W_{a,Z,i}^{-1} \mathbf{Y}_{a,i} + \sum_{i=1}^{M_u} \mathbf{X}_{u,i}^T W_{u,Z,i}^{-1} \mathbf{Y}_{u,i} \right)$$

$$= \left\{ \sum_{i=1}^{M_{\mathrm{u}}} \mathbf{X}_{\mathrm{u},i}^T W_{\mathrm{u},Z,i}^{-1} \mathbf{X}_{\mathrm{u},i} + \sum_{\mathbf{Z}\in S} p_Z(\widehat{\Omega}) \sum_{i=1}^{M_{\mathrm{a}}} \mathbf{X}_{\mathrm{a},i}^T W_{\mathrm{a},Z,i}^{-1} \mathbf{X}_{\mathrm{a},i} \right\}^{-1}$$

$$\times \left\{ \sum_{i=1}^{M_{\mathrm{u}}} \mathbf{X}_{\mathrm{u},i}^T W_{\mathrm{u},Z,i}^{-1} \mathbf{Y}_{\mathrm{u},i} + \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M_{\mathrm{a}}} \mathbf{X}_{\mathrm{a},i}^T W_{\mathrm{a},Z,i}^{-1} \mathbf{Y}_{\mathrm{a},i} \right\}.$$

(B)

$$\widehat{\alpha}_j = \frac{1}{N} \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) n_j = \begin{cases} n_j/N, & \text{membership to cluster } j \text{ is fully observed,} \\ \sum_{\mathbf{Z}\in\mathcal{S}} p_Z(\widehat{\Omega}) n_j/N, & \text{membership to cluster } j \text{ is ambiguous.} \end{cases}$$

(C)

$$E(t_1|\mathbf{Y},\widehat{\Omega}) = \sum_{Z\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} E(\epsilon_i^T \epsilon_i|\mathbf{Y}_i,\widehat{\Omega}) = \sum_{i=1}^{M_{\mathrm{u}}} E(\epsilon_i^T \epsilon_i|\mathbf{Y}_i,\widehat{\Omega}) + \sum_{Z\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M_{\mathrm{a}}} E(\epsilon_i^T \epsilon_i|\mathbf{Y}_i,\widehat{\Omega}),$$

$$E(t_2|\mathbf{Y},\widehat{\Omega}) = \sum_{Z\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M} E(b_i b_i^T|\mathbf{Y}_i,\widehat{\Omega}) = \sum_{i=1}^{M_{\mathrm{u}}} E(b_i b_i^T|\mathbf{Y}_i,\widehat{\Omega}) + \sum_{Z\in\mathcal{S}} p_Z(\widehat{\Omega}) \sum_{i=1}^{M_{\mathrm{a}}} E(b_i b_i^T|\mathbf{Y}_i,\widehat{\Omega}).$$

APPENDIX 3

We begin by describing the results of a simulation study aimed at assessing the sensitivity of the mixed modeling approach to both the number of clusters and the model misspecification. Founder effect models as well as departures from HWE are considered. A comparison of the mixed modeling to a more traditional ANOVA approach for identifying haplotype associations as well as a more recently described 2-stage MI approach (Foulkes *and others*, 2007) is also provided. We then summarize precision and power for a range of percentages of individuals with ambiguous cluster membership. Finally, estimates of the FDRs associated with varying degrees of missingness are presented.

Performance and sensitivity results for varying numbers of clusters and models are provided in Figures 3(a) and (b). These results are based on 400 iterations per condition, samples of size $n = 200$, and fully observed haplotypes. Cluster assignments are resampled within each iteration based on assumed frequencies. The ratio of standard deviations is defined as σ_b/σ_ϵ, and for simplicity, we set $\sigma_\epsilon = 1$ across all simulations. Power is defined as the proportion of simulations for which the likelihood ratio test comparing the mixed model (with a random cluster effect) to a fixed-effects model (intercept only) is significant at the 0.05 level. That is, power is the proportion of times we reject the omnibus null hypothesis H_0: $\sigma_b^2 = 0$. A significance cutoff is chosen based on a 50–50 mixture of a χ_1^2 and χ_0^2 distribution since we are testing a variance parameter at a boundary.

Figure 3(a) illustrates power for a range of variance ratios where the number of clusters ranges from 3 to 36. For the most general case in which clusters are defined by unique haplotype pairs, this corresponds to 2–8 observed haplotypes. Assumed cluster frequencies are determined based on population haplotype frequencies. For the case of 36 clusters (8 haplotypes), haplotype frequencies are set equal to (0.20, 0.15, 0.15, 0.12, 0.10, 0.10, 0.10, 0.08) and corresponding cluster probabilities are calculated assuming independence (HWE). For 21, 10, and 3 clusters, the corresponding haplotype probabilities are set equal to (0.20, 0.20, 0.20, 0.15, 0.15, 0.10), (0.40, 0.20, 0.20, 0.20), and (0.60, 0.40), respectively, and

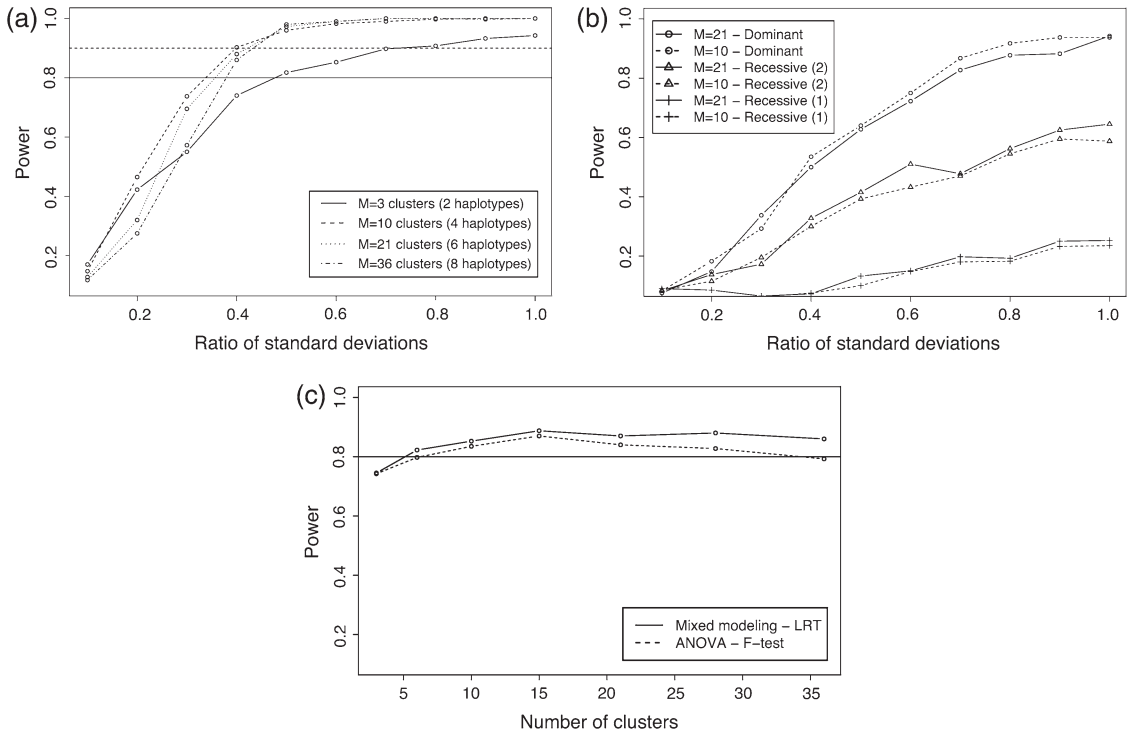

Fig. 3. Performance and sensitivity of the mixed modeling approach. (a) Power for detecting haplotype effect variability by number of clusters ($n = 200$). (b) Power under dominant and recessive founder models ($n = 200$). For recessive model, population frequencies of founder haplotype (H_{d}) equal to (1) 0.20 and (2) 0.40 are considered. For dominant model, a frequency of 0.20 is illustrated. (c) Power for mixed model and ANOVA approaches ($n = 200$, $\sigma_b/\sigma_\epsilon = 0.40$).

again HWE is assumed to determine cluster frequencies. For example, in the case of 2 haplotypes, the cluster frequencies are $(0.60^2, 2 \times 0.60 \times 0.40, 0.40^2) = (0.36, 0.48, 0.16)$.

These results suggest that for $\sigma_b/\sigma_\epsilon > 0.4$, the omnibus test has reasonable power (>80%) for detecting variability in the cluster effects in the case of at least 10 clusters (4 haplotypes). Power increases to greater than 90% for a ratio of standard deviations of more than 0.5. The difference in power between differing numbers of clusters is more marked for smaller effect sizes. For example, $M = 10$ and $M = 21$ clusters yield comparable power for $\sigma_b/\sigma_\epsilon \geqslant 0.4$, while the power differential is 15–20% for $\sigma_b/\sigma_\epsilon = 0.3$ and 0.2. Interestingly, the power gains for a larger number of clusters diminish entirely for effect sizes of greater than 0.5. This may be due in part to the decreasing number of observations per cluster for a fixed sample size of $n = 200$ across simulations.

In the second case, illustrated in Figure 3(b), data are generated according to a founder effect model in which a single haplotype (H_{d}) is associated with the disease phenotype. Both recessive and dominant genetic models are considered. In the case of the recessive model, the presence of 2 copies of H_{d} is required to effect the phenotype, while in the dominant model, a single copy of H_{d} effects the phenotype. More specifically, for the recessive model $b_i \sim N(0, \sigma_b^2)$ for the unique i such that $C_i = (H_{\mathrm{d}}, H_{\mathrm{d}})$ and $b_i = 0$ otherwise. For the dominant model, on the other hand, $b_i \sim N(0, \sigma_b^2)$ for all i such that $H_{\mathrm{d}} \in C_i$ and $b_i = 0$ otherwise. In the case of the dominant model, the frequency of H_{d} is set equal to 0.20. In the recessive model examples, founder haplotype frequencies of 0.20 and 0.40 are considered, corresponding

to single cluster frequencies of 0.04 and 0.16, respectively. Model fit is based on the assumption of normality of all the random effects, and thus, we indicate that the model has been misspecified.

As expected, power is dramatically lower under model misspecification. Under the recessive founder model, only a single cluster effects the phenotype while the variability in the effect of the other clusters is assumed to equal 0. If this cluster has a low population frequency (0.16 and 0.04 are illustrated), then power is less than 80% for ratios of standard deviations of as high as 1. Performance is improved for the dominant model, in which all clusters with at least one copy of the founder haplotype have an effect on the phenotype while the remaining cluster effects are assumed to have no variability. In the examples provided for the dominant model, 4 of the 10 cluster effects and 6 of the 21 cluster effects arise from a normal distribution while the remaining in each case are set to equal 0. In these cases, greater than 80% power is observed for ratios of standard deviations of 0.7 and greater.

Power under specific departures from HWE is also estimated. Consistent with the approach of Satten and Epstein (2004), we let the joint probability of the haplotype pair (H, H') be given by $\alpha_{HH'} = I(H = H')F\delta_H + 2^{I(H\neq H')}(1 - F)\delta_H\delta_{H'}$, where $I(\cdot)$ is the indicator function, F is a scalar, and δ_H is the population-level frequency of haplotype H. Notably, $F = 0$ corresponds to the HWE setting. In this case, we assume wild-type and variant SNP frequencies of 0.80 and 0.20 for each of 2 SNPs. This results in 4 haplotypes with frequencies of (0.64, 0.16, 0.16, 0.04) and 10 corresponding clusters with frequencies under HWE of (0.0496, 0.2048, 0.2048, 0.0512, 0.0256, 0.0512, 0.0238, 0.0258, 0.0128, 0.0016). Ambiguity lies between 2 of these clusters and all individuals within these 2 clusters are ambiguous. We apply the mixed modeling approach under the HWE assumption to 100 simulated data sets of size $n = 100$, where $\sigma_b/\sigma_\epsilon = 0.60$. Values of $F = -0.05$, 0, and 0.05 are considered as described in Satten and Epstein (2004). Resulting power for omnibus test of no variability in cluster effects is 84%, 84%, and 86%, respectively, suggesting reasonable performance under moderate departures from HWE.

Figure 3(c) illustrates power both for the omnibus test of variability in random cluster effects and for an F-test based on a one-way ANOVA model in which clusters are treated as fixed factor levels. The ANOVA approach is extended for unobservable haplotypes in Lake *and others* (2003) and easily implemented in R with the haplo.glm() function of the haplo.stats package. Here, we focus on the observed haplotype setting to characterize overall performance, though a reduction in power is expected using both approaches in the context of missingness in phase, as described in more detail below for the mixed modeling setting. A ratio of standard deviations of 0.40 is assumed. Power is comparable between the 2 approaches when the number of clusters (factor levels) is less than 15, corresponding to 5 haplotypes; however, it tends to deviate as the number of clusters increases to 36 (corresponding to 8 haplotypes.) While power is maintained at near-constant values for the mixed modeling approach, it tends to decrease with increasing haplotypes using ANOVA. This is consistent with reports in the literature suggesting that the increase in degrees of freedom associated with consideration of more haplotypes can reduce power using an ANOVA approach (Tzeng *and others*, 2006). Ultimately, power will also decline using the mixed modeling approach since as the number of clusters (haplotypes) increases, the number of individuals within the clusters will decrease for a fixed sample size.

An additional comparison of power between the ECM approach described herein and a 2-stage MI approach is provided. The later approach and its performance are described in detail in Foulkes *and others* (2007). In this example, the trait is simulated for a sample of size $n = 200$ for each of 100 simulations and model fitting assumes HWE with missingness between 2 clusters (representing about 8% ambiguity for this data example). Power for detecting variability is comparable for the ECM and MI approaches, averaging 79% and 78.6%, respectively, across a range of standard deviations from 0.4 to 0.8; however, an approximate 1.5-fold increase in the 25th, 50th, and 75th quantiles of the test statistic distribution is observed for the ECM versus MI approach. Notably, the computational burden associated with increasing amounts of haplotype ambiguity for ECM (on the order of 2^{N_a}, where N_a is the number of ambiguous individuals between 2 clusters) far exceeds that associated with the MI approach (on the order of $B \times N_a$,

Table 3. *Simulation results for differing percents ambiguity and variance ratios*

Ambiguity (%)	σ_b/σ_ϵ	Power (%)	Bias ($\widehat{\text{se}}$)†					CR‡				
			β_0	$\bar{\alpha}_\text{a}$	$\bar{\alpha}_\text{u}$	σ^2_ϵ	σ^2_b	β_0	$\bar{\alpha}_\text{a}$	$\bar{\alpha}_\text{u}$	σ^2_ϵ	σ^2_b
0*	0.2	32	0.0039(0.084)	—	0.00083(0.015)	0.0021(0.10)	0.011(0.043)	0.95	—	0.97	0.95	0.94
	0.4	88	0.00046(0.12)	—	0.00024(0.015)	0.0073(0.11)	0.017(0.082)	0.94	—	0.97	0.96	0.97
	0.6	99	0.0061(0.16)	—	0.00048(0.015)	0.011(0.11)	0.037(0.16)	0.95	—	0.97	0.96	0.95
	0.8	100	0.0019(0.19)	—	0.00048(0.015)	0.0086(0.11)	0.014(0.25)	0.96	—	0.96	0.95	0.95
	1.0	100	0.0036(0.24)	—	0.00048(0.015)	0.013(0.11)	0.025(0.36)	0.96	—	0.97	0.94	0.96
5*	0.2	34	0.011(0.088)	0.00085(0.020)	0.00026(0.015)	0.010(0.11)	0.0020(0.056)	0.96	0.96	0.97	0.95	0.94
	0.4	89	0.0039(0.12)	0.0016(0.020)	0.00026(0.014)	0.0053(0.11)	0.0066(0.095)	0.96	0.99	0.96	0.93	0.96
	0.6	100	0.0046(0.16)	0.00031(0.018)	0.00026(0.014)	0.00071(0.11)	0.0091(0.16)	0.96	0.98	0.97	0.94	0.98
	0.8	100	0.032(0.20)	0.00073(0.019)	0.00026(0.014)	0.0096(0.11)	0.012(0.26)	0.93	0.96	0.97	0.97	0.97
	1.0	100	0.011(0.23)	0.0010(0.018)	0.00039(0.015)	0.0038(0.11)	0.034(0.36)	0.97	0.98	0.97	0.91	0.94
10**	0.2	33	0.0027(0.091)	0.00049(0.019)	0.00084(0.014)	0.00015(0.10)	0.021(0.049)	0.96	0.97	0.96	0.97	0.93
	0.4	87	0.0099(0.14)	0.0011(0.019)	0.00057(0.014)	0.0051(0.11)	0.025(0.093)	0.95	0.97	0.97	0.94	0.95
	0.6	100	0.0011(0.17)	0.00095(0.019)	0.00048(0.014)	0.026(0.11)	0.048(0.17)	0.96	0.97	0.97	0.94	0.91
	0.8	100	0.026(0.19)	0.0017(0.018)	0.00066(0.014)	0.0073(0.10)	0.061(0.31)	0.93	0.97	0.97	0.96	0.94
	1.0	100	0.016(0.22)	0.00068(0.018)	0.00080(0.014)	0.0059(0.12)	0.097(0.39)	0.95	0.97	0.97	0.95	0.93
20**	0.2	30	0.0037(0.089)	0.0022(0.020)	0.00019(0.013)	0.0011(0.10)	0.044(0.071)	0.95	0.97	0.96	0.95	0.90
	0.4	88	0.0076(0.11)	0.0022(0.019)	0.00077(0.013)	0.0083(0.11)	0.083(0.12)	0.97	0.98	0.96	0.95	0.88
	0.6	99	0.013(0.15)	0.0019(0.020)	0.00058(0.013)	0.0042(0.11)	0.13(0.21)	0.95	0.98	0.97	0.95	0.91
	0.8	100	0.017(0.20)	0.0011(0.020)	0.00077(0.013)	0.010(0.11)	0.13(0.29)	0.95	0.98	0.96	0.97	0.92
	1.0	100	0.0014(0.23)	0.0012(0.020)	0.00077(0.013)	0.0029(0.11)	0.27(0.44)	0.94	0.97	0.97	0.95	0.89

Results are based on *400 and **200 simulations per condition (σ_b/σ_ϵ) with samples of size $n = 200$ and $m = 21$ clusters.

†Bias is defined as the absolute difference between the median of the estimate over the simulations and the true parameter value. $\bar{\alpha}_\text{a}$ and $\bar{\alpha}_\text{u}$ are the average bias across the ambiguous and unambiguous clusters, respectively. Standard errors ($\widehat{\text{se}}$) are calculated based on all simulations within a condition.

‡CR is defined as the proportion of simulations for which the true parameter value is within the corresponding 95% confidence interval.

where B is the number of imputations). While the theoretical advantage of ECM is that it incorporates information on the trait of interest (as reflected in our simulation study by the overall distribution of test statistics being greater for ECM compared to MI), we believe that a complete investigation of the relative performance is warranted in light of the trade-off in computational efficiency.

More detailed simulation results are given in Table 3 for varying degrees of cluster ambiguity and ratios of standard deviations. In all cases, a sample size of $n = 200$ is again assumed. The ECM approach described in this manuscript is applied for settings in which the ambiguity is greater than 0%. Bias is defined as the absolute difference between the median parameter estimate over the simulations and the true value. The estimated standard error of the parameter estimates based on the simulations is given by $\widehat{\text{se}}$. β_0 is a fixed intercept and is set equal to 0 across all simulations. $\bar{\alpha}_{\text{a}}$ and $\bar{\alpha}_{\text{u}}$ are the average biases over the ambiguous and unambiguous clusters, respectively. CR is defined as the percentage of simulations for which the true parameter value is within the 95% confidence interval. These intervals are constructed for each simulation based on the current parameter estimate and the estimated standard error ($\widehat{\text{se}}$) across all simulations. Although the variance estimates appear to be slightly rightwardly skewed, transformations are not applied since they result in more pronounced leftward skewness. Both logarithmic and square root transformations were considered (results not shown).

Cluster-level ambiguity is sampled as follows. For 5% ambiguity, 10 of the $n = 200$ observations are ambiguous between a pair of clusters. In the case of 10% ambiguity, 5% of the sample is ambiguous between a pair of clusters and another 5% is ambiguous between a second pair of clusters. Likewise for 20% ambiguity, 4 sets of individuals, each consisting of 5% of the sample, are ambiguous between different pairs of clusters. In all cases, 21 clusters are assumed and frequencies ranging from 0.01 to 0.08 are determined as described for Figure 3 above. In total, 200–400 simulations are performed per condition as indicated, and model fitting is based on the general case in which we do not make any assumption about HWE. Standard deviation ratios of 0.2–1.0 are presented. Finally, FDRs are reported. In this case, 100 simulations are performed for each level of ambiguity under the assumption that $\sigma_b^2 = 0$, and we report the percentage of times we reject the null hypothesis $H_0 : \sigma_b^2 = 0$ based on the likelihood ratio test.

These results suggest modest reductions in power for detecting overall variability in cluster effects with ambiguity as high as 20%. While CRs are maintained at between 0.93 and 0.97 for the fixed effect, β_0, and the haplotype frequencies, α_{a} and α_{u}, CRs for the variance parameters decline with increasing ambiguity. This is most marked for σ_b^2 for which the CRs drop to between 0.88 and 0.92 for 20% ambiguity. In general, σ_b^2 is slightly overestimated when there is ambiguity in the cluster assignments. This is reflected further in the FDRs that tend to increase as the rate of ambiguity increases from 0% to 20%. Specifically, for 0%, 5%, 10%, and 20% ambiguity, the estimated FDRs are 5%, 6%, 9%, and 7%, respectively.

REFERENCES

CHIU, W. F., YUCEL, R. M., ZANUTTO, E. AND ZASLAVSKY, A. M. (2005). Using matched substitutes to improve imputations for geographically linked databases. *Survey Methodology* **31**, 69–72.

DEMIDENKO, E. (2004). *Mixed Models: Theory and Applications*. Hoboken, NJ: John Wiley & Sons.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (C/R: p22-37). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–22.

DIGGLE, P., LIANG, K.-Y. AND ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

EXCOFFIER, L. AND SLATKIN, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.

FITZMAURICE, G. M., LAIRD, N. M. AND WARE, J. H. (2004). *Applied Longitudinal Analysis*. New York: John Wiley & Sons.

FOULKES, A. S. AND DEGRUTTOLA, V. (2002). Characterizing the relationship between HIV-1 genotype and phenotype: prediction based classification. *Biometrics* **58**, 145–156.

FOULKES, A. S., REILLY, M., ZHOU, L., WOLFE, M. AND RADER, D. J. (2005). Mixed modeling to characterize genotype-phenotype associations. *Statistics in Medicine* **24**, 775–789.

FOULKES, A. S., WOHL, D. A., FRANK, I., PULEO, E., RESTINE, S., WOLFE, M. L., DUBE, M. P., TEBAS, P. AND REILLY, M. P. (2006). Associations among race/ethnicity, APOC-III genotypes and lipids in HIV-1 infected individuals on antiretroviral therapy. *PLoS Medicine* **3**, e52.

FOULKES, A. S., YUCEL, R. AND REILLY, M. P. (2007). Mixed modeling and multiple imputation for unobservable genotype clusters. *Statistics in Medicine*, doi: 10.1002/sim.3051, 1–18.

JAMSHIDIAN, M. AND JENNRICH, R. I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* **88**, 221–228.

JENNRICH, R. I. AND SCHLUCHTER, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.

LAIRD, N., LANGE, N. AND STRAM, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association* **82**, 97–105.

LAIRD, N. M. AND WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

LAKE, S. L., LYON, H., TANTISIRA, K., SILVERMAN, E. K., WEISS, S. T., LAIRD, N. M. AND SCHAID, D. J. (2003). Estimation and testing of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity* **55**, 56–65.

LIN, D. Y. AND ZENG, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association* **101**, 89–104.

LINDSTROM, M. J. AND BATES, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**, 1014–1022.

LITTLE, R. J. A. AND RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

MCCULLOCH, C. E. AND SEARLE, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons.

MENG, X.-L. AND RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.

MUTHEN, B. AND SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**, 463–469.

PINHEIRO, J. C. AND BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

SANCHEZ, B., BUDTZ-JORGENSEN, E., RYIAN, L. M. AND HU, H. (2005). Structural equation models: a review with applications to environmental epidemiology. *Journal of the American Statistical Association* **100**, 1443–1455.

SATTEN, G. A. AND EPSTEIN, M. P. (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genetic Epidemiology* **27**, 192–201.

SCHAFER, J. L. AND YUCEL, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* **11**, 437–457.

TZENG, J. Y., WANG, C. H., KAO, J. H. AND HSIAO, C. K. (2006). Regression-based association analysis with clustered haplotypes through use of genotypes. *American Journal of Human Genetics* **78**, 231–242.

VERBEKE, G. AND LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

VERBEKE, G. AND MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

VONESH, E. F. AND CHINCHILLI, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker, Inc.

WOLFINGER, R., TOBIAS, R. AND SALL, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific and Statistical Computing* **15**, 1294–1310.

[*Received July 20, 2007; revised December 6, 2007; accepted for publication December 14, 2007*]