The International Journal of Biostatistics

Volume 3, *Issue* 1 2007 *Article* 2

A Resampling-Based Approach to Multiple Testing with Uncertainty in Phase

Andrea S. Foulkes, *University of Massachusetts*
Victor G. DeGruttola, *Harvard*

Recommended Citation:
Foulkes, Andrea S. and DeGruttola, Victor G. (2007) "A Resampling-Based Approach to Multiple Testing with Uncertainty in Phase," *The International Journal of Biostatistics*: Vol. 3: Iss. 1, Article 2.
DOI: 10.2202/1557-4679.1037

A Resampling-Based Approach to Multiple Testing with Uncertainty in Phase

Andrea S. Foulkes and Victor G. DeGruttola

Abstract

Characterizing the genetic correlates to complex diseases requires consideration of a large number of potentially informative biological markers. In addition, attention to alignment of alleles within or across chromosomal pairs, commonly referred to as phase, may be essential for uncovering true biological associations. In the context of population based association studies, phase is generally unobservable. Preservation of type-1 error in a setting with multiple testing presents a further analytical challenge. This manuscript combines a likelihood-based approach to handling missing-ness in phase with a resampling method to adjust for multiple testing. Through simulations we demonstrate preservation of the family-wise error rate and reasonable power for detecting associations. The method is applied to a cohort of 626 HIV-1 infected individuals receiving highly active anti-retroviral therapies, to ascertain potential genetic contributions to abnormalities in lipid profiles. The haplotypic effects of 2 genes, hepatic lipase (HL) and endothelial lipase (EL), on high-density lipoprotein cholesterol (HDL-C) are tested.

KEYWORDS: multiple testing, resampling, genotype, haplotype, covariates, phase, HIV-1, lipids

Author Notes: Support for this research was provided by the National Institute of Allergy and Infection Diseases (NIAID) research awards AI056983 and AI51164 and by the Adult AIDS Clinical Trials Group (ACTG) funded by the NIAID (AI38858). We thank the ACTG New Works Concept Sheet 224 Study Team, particularly Dr. Muredach Reilly, Cardiovascular Division, University of Pennsylvania School of Medicine, for helpful discussions and providing access to data. We also thank the reviewers for helpful comments and Jason Sinnwell, Department of Health Sciences Research, Mayo Clinic for support using the R library haplo.stats.

1 Introduction

Identifying and characterizing the molecular level contributors to complex diseases, such as cardiovascular disease, typically requires tests of large number of hypotheses; such tests must appropriately control for type-1 error. In addition, consideration of how genetic variants within a gene align within or across chromosomal copies (commonly referred to as phase) may prove essential to characterizing genetic contributions to variability in a disease trait. This consideration presents an additional analytical challenge, because haplotype is generally unobservable in association studies of unrelated individuals.

Several methods have been proposed recently to handle unobservable phase in the context of identifying associations between haplotypes and a disease phenotype. These include likelihood-based expectation-maximization (EM)-type methods as described in Excoffier and Slatkin (1995), Zaykin *et al.* (2002), Schaid *et al.* (2002), Zhao *et al.* (2003), Lake *et al.* (2003), and Lin and Zeng (2006), Markov Chain Monte Carlo (MCMC) approaches as described in Stephens *et al.* (2001), and hidden Markov modeling, described in Scheet and Stephens (2006). These methods make use of information from those subjects whose haplotypes are fully determined based on their observed genotypes. Methods for controlling type-1 error rates in the context of high-dimensional data analysis have also been described extensively. These include resampling based approaches as given in Westfall and Young (1993), Yekutieli and Benjamini (1999), Ge *et al.* (2003), and Pollard and van der Laan (2004) as well as a Monte Carlo method (Lin, 2006). A comprehensive discussion of the relative merits of these approaches, with regard to computational efficiency and flexibility for handling covariate adjustments, is provided in Lin (2006).

Our proposed methods combine existing likelihood and resampling-based procedures to adjust for multiplicity of comparisons in the context of phase uncertainty. This approach extends existing methods for analyzing association studies of unrelated individuals. Its novelty lies primarily in the way in which these existing methods are combined; in some cases, modification of existing resampling procedures is required to handle phase uncertainty. Section 2 describes the new methods, and Section 3 describes a simulation study. Section 4 provides an application to a study of anti-retroviral therapy (ART) associated dyslipidemia in HIV-1 infected individuals.

2 Methods

In population-based association studies of unrelated individuals, heterozygosity at more than one locus within a gene renders haplotypic phase unobservable. Our approach begins by determining the set of haplotype pairs consistent with each individuals observed genotype, as described by Lin and Zeng (2006). Treating the true, but unobservable, haplotype as "missing," the genetic models can be fit using an EM approach. If haplotype were known, both the free step-down resampling approach of Westfall and Young (1993) and the method of Pollard and van der Laan (2004) could be used to test its effects on a measure of phenotype. The former is based on resampling of model residuals, whereas the latter resamples from the observed data themselves. To use the the Westfall and Young approach in our setting, we would sample the residuals corresponding to each possible haplotype pair for an individual according to a posterior probability associated with that pair. Here we implement the Pollard and van der Laan (2004) approach because it is more generally applicable. Both methods provide for appropriate adjustment for multiple comparisons in the setting of unobservable haplotypes.

2.1 Notation and models

Consider the general linear model in Equation 1, where $\mathbf{Y}_{N\times 1}$ is a vector of responses for N unrelated individuals, $g(\cdot)$ is an appropriately defined link function, $\mathbf{X}_{N\times P}$ represents individual level covariate values, $\mathbf{H}_{N\times M}$ captures the haplotype information and $\epsilon \sim N(0, \sigma^2 I_{N\times N})$. Below we distinguish between two types of models: the genetic model (GM) and the model of association (MA). Notably, the number of columns in $\mathbf{H}$ (given by M) depends on the specific model (GM and/or MA) under investigation, while the number of rows of $\mathbf{H}$ (given by N) will always equal the sample size. In the simplest case, M is equal to the number of possible haplotypes and the $(i,j)th$ element of $\mathbf{H}$ is an indicator that individual i has haplotype j, where $i = 1, \ldots, N$ and $j = 1, \ldots, M$. In this case, Equation 1 reduces to the usual analysis of covariance (ANCOVA) model.

$$g(Y) = \alpha + \mathbf{X}\beta + \mathbf{H}\gamma + \epsilon \qquad (1)$$

For simplicity of presentation, we assume interest lies in assessing main effects of haplotypes and not interactions between covariates (given in the $\mathbf{X}$

matrix) and these haplotypes; however, straightforward extensions of these models would allow for interaction effects. Consider the general null hypothesis $H_0 : \gamma_j = 0$, $j = 1, \ldots, M$ where γ_j is the jth element of γ in Equation 1. Using the EM algorithm given by Lake *et al.* (2003) and Lin and Zeng (2006) and described in detail in Section 2.2, test statistics (such as Wald test statistics) corresponding to these null hypotheses can be calculated. Additional application of a resampling based procedure provides a multiple testing adjustment in this setting and is described in Section 2.3. First we elaborate on specific examples of the model provided in Equation 1.

The GM refers to the combined action of two chromosomal copies of a gene. For example, an additive GM assumes that the effect of having a single copy of a given haplotype h^* is half the effect of having 2 copies of this haplotype. The MA, on the other hand, refers to the interaction among haplotypes across genes. For example, an additive MA assumes that the effect of having a given haplotype h_r^* at gene r and haplotype h_s^* at gene s is the sum of the individual effects. A multiplicative (or synergistic) MA would instead assume this effect to be the product of the individual effects.

Underlying GMs and MAs can be represented in the definition of the matrix $\mathbf{H}$ in Equation 1. To illustrate, we consider an additive GM within a single gene. Let $\mathcal{H}$ be the set of all haploytypes in the data sample under consideration, and suppose (h_k, h_l) is the haplotype pair (diplotype) for a given individual i where $h_k, h_l \in \mathcal{H}$. In this case, the $(i, j)th$ element of the matrix $\mathbf{H}$ is equal to $H_{ij} = I(h_k = h_j) + I(h_l = h_j)$, for $j = 1, \ldots, M$. In other words, $H_{ij} = 0$, 1 or 2 depending on whether individual i has 0, 1 or 2 copies of haplotype $h_j \in \mathcal{H}$.

As described in Lin and Zeng (2006) for the single gene case, recessive, dominant and codominant GMs can be represented by $H_{ij} = I(h_k = h_l = h_j)$, $H_{ij} = I(h_k = h_j) + I(h_l = h_j) - I(h_k = h_l = h_j)$ and $H_{ij} = I(h_k = h_j) + I(h_l = h_j) + I(h_k = h_l = h_j)$, respectively. Haploinsufficiency refers to the GM in which a single functional (wildtype) copy of the gene does not produce enough gene product, resulting in the disease phenotype; this model is also represented by appropriate definition of the $\mathbf{H}$ matrix. Note that each haplotype represented in the columns of $\mathbf{H}$ can have different GMs, though generally these are assumed to be the same.

More generally, in the multiple gene framework, consideration must be given to the joint effects on disease phenotype both of chromosomal copies within genes and of haplotypes across genes. For example, an additive MA across R genes can be described for all the aforementioned GMs. This model is represented by creating additional columns in the **H** matrix corresponding to the haplotypes for each gene under consideration, as in an R-factor ANOVA model; once again, the elements would be indicator variables defined according to the GM.

An interaction model or a model for epistasis (interaction in the absence of main effects) can again be represented through appropriate definition of the columns of the H matrix. For example, a model for epistasis between 2 genes, assuming a recessive GM for each, is given by $H_{ij'} = I(h_{rk} = h_{rl} = h_{rj}) * I(h_{sk} = h_{sl} = h_{sj})$ where (h_{rk}, h_{rl}) and (h_{sk}, h_{sl}) are the diplotypes for individual i at genes r and s, respectively. More sophisticated models, in which, for example, groups of genes in the same pathway to disease act synergistically with additive effects across groups, can also be represented using this framework.

2.2 EM to account for missing phase

In general, **H** of Equation 1 is not observable though for each individual, there is a finite set of haplotype pairs, $\mathcal{H}_g$ that is consistent with the observed genotype g. For example, if an indvidual's genotype is Aa and Bb at two respective SNPs within a gene, then there are two possible haplotype pairs given by the set $[(AB, ab), (Ab, aB)]$. By borrowing from the information on individuals for whom **H** is fully observed, i.e. individuals who are heterozygous at exactly one or no SNPs within the gene(s) under consideration, the posterior probabilities associated with each possible haplotype pair can be estimated.

Several recent manuscripts, including Lake *et al.* (2003) and Lin and Zeng (2006), have described the application of the EM algorithm (Laird and Ware, 1982) for estimation and testing of haplotype effects in this unobservable data setting. Using the notation of Lin and Zeng (2006), the likelihood contribution for an individual i is given by Equation 2 where $\mathcal{G}$ is the set of all possible genotypes, $\mathcal{H}_g$ is again the set of all haplotype pairs that are consistent with genotype $g \in \mathcal{G}$, G_i is the observed genotype for individual i, $\theta = (\alpha, \beta, \gamma, \sigma)$ is the vector of parameters in the model given by Equation 1, π is a parameter vector of true population level haplotype prevalences, and $\Phi = (\theta, \pi)$.

$$L_i(\Phi) = \prod_{g \in \mathcal{G}} \left[\sum_{(h_k, h_l) \in \mathcal{H}_g} Pr_\theta(y|\mathbf{x}, (h_k, h_l)) Pr_\pi(h_k, h_l) \right]^{I(G_i = g)} \tag{2}$$

DOI: 10.2202/1557-4679.1037

In practice, the EM algorithm proceeds by first taking the expectation of the complete data log likelihood conditional on the observed data, (y_i, x_i, g_i) where g_i is the genotype for individual i and $i = 1, \ldots, N$ (E-step). This conditional expectation is a function of the posterior haplotype probabilities, denoted $p_{ikl}(\Phi)$, which are defined simply as the probability of a haplotype given the observed genotype, phenotype and covariate data. The second step of the EM algorithm (M-step) maximizes this conditional expectation where $p_{ikl}(\Phi)$ is evaluated at the current estimate of Φ. This approach is described in detail in Lin and Zeng (2006) and summarized below.

E-step: The complete data log likelihood for individual i is given by $Pr_\theta(y_i|x_i, H_i)Pr_\pi(H_i)$. The expectation of this likelihood conditional on the observed data, (y_i, x_i, g_i) is given by (3) where $p_{ikl}(\Phi)$ is defined as in (4).

$$\sum_{i=1}^{N} \sum_{(h_k,h_l)\in\mathcal{H}_i} p_{ikl}(\Phi)\left[logPr_\theta\left(y_i|x_i,(h_k,h_l)\right) + logPr_\pi(h_k,h_l)\right] \tag{3}$$

$$p_{ikl}(\Phi) = \frac{Pr_\theta\left(y_i|x_i,(h_k,h_l)\right)Pr_\pi(h_k,h_l)}{\sum_{(h_k,h_l)\in\mathcal{H}_i} Pr_\theta\left(y_i|x_i,(h_k,h_l)\right)Pr_\pi(h_k,h_l)} \tag{4}$$

M-step: At the $(m+1)$st iteration, evaluate $p_{ijk}(\Phi)$ at $\widehat{\Phi}^{(m)}$ and maximize the conditional expected log likelihood, expression (3). This maximization requires solving the system of equations given in (5) where ∇_θ and ∇_π are the partial derivatives with respect to θ and π respectively. This can be achieved through a Newton-Raphson algorithm, or under certain assumptions closed form solutions can be obtained as described in detail in Lin and Zeng (2006).

$$\begin{aligned} \sum_{i=1}^{N} \sum_{(h_k,h_l)\in\mathcal{H}_i} p_{ikl}(\widehat{\Phi}^{(m)})\nabla_\theta \log\ Pr_\theta\left(y_i|x_i,(h_k,h_l)\right) = 0 \\ \sum_{i=1}^{N} \sum_{(h_k,h_l)\in\mathcal{H}_i} p_{ikl}(\widehat{\Phi}^{(m)})\nabla_\pi \log\ Pr_\pi(h_k,h_l) = 0 \end{aligned} \tag{5}$$

As described in Lake *et al.* (2003), assuming an exponential family distribution and Hardy Weinberg equilibrium (HWE), the EM algorithm reduces to a weighted regression ("EM by the methods of weights") where the weights are set equal to subjects' posterior haplotype probabilities. Note that HWE implies $\pi_{kl} = \pi_k \pi_l$ where π_{kl} is the prevalence of the diplotype (h_k, h_l), and π_k and π_l are the prevalences of h_k and h_l, respectively. Estimation procedures in the presence of specific departures from HWE are described in Lin and Zeng (2006). Inference on haplotype effects generally requires an estimate of the variance/covariance matrix of Φ, which can be obtained using the information matrix. Lake et al. also note that the observed information can be computed using Louis' formula (Louis, 1982), and that test statistics corresponding to tests of single haplotype effects can be calculated using the estimated parameters and variance/covariance matrix (Lake *et al.*, 2003). Straightforward implementation of this approach is achieved using the `haplo.glm()` function of the `haplo.stats` library in *R*, version 2.2.1.

2.3 Multiple testing adjustment with uncertainty in phase

We propose coupling the EM approach described in Section 2.2 with a resampling based multiple testing procedure (MTP). For the latter, we make use of the approach of Pollard and van der Laan (2004) since its implementation is straightforward and requires fewer assumptions than do others, such as the approach of Westfall and Young (1993). Note that our method is appropriate to carry out analyses of covariance when, as described above, the group assignments ($\mathbf{H}$ of Equation 1) are not known with certainty. A modified version of the free step-down resampling approach of Westfall and Young (1993), discussed in Section 5, can also be used, but requires an assumption of subset pivotality, unlike this approach of Pollard and van der Laan. Subset pivotality should hold in settings similar to ours, but without uncertainty in phase; estimates of the conditional probability of a haplotype, however, may depend on the pattern of true and false null hypotheses. Therefore subset pivotality may not necessarily hold in small samples.

Consider the null hypotheses $H_0 : \gamma_j = 0$ for $j = 1, \ldots, M$ and corresponding Wald test statistics. We begin by fitting the model in Equation 1 using the method outlined in Section 2.2. The coefficient estimates for each haplotype are recorded and denoted by the $M \times 1$ vector $\tilde{\mu}_n$ where the jth element of $\tilde{\mu}_n$ equals the corresponding effect. Testing is based on the following algorithm, proposed by Pollard and van der Laan (2004):

1. Bootstrap Y, $\mathbf{X}$ and $\mathbf{H}$ with replacement, preserving the within individ-

ual link and recalculate coefficient estimates. Denote these by the vector $\mu_n^{\#}$ where again the jth element corresponds to the jth test.

2. Record $Z_n^{\#b} = \left(\mu_n^{\#} - \tilde{\mu}_n\right)/sd(\mu_n^{\#})$.

3. Repeat steps (1) and (2) B times to get $Z_n^{\#1}, \ldots, Z_n^{\#B}$. The distribution of $Z_n^{\#b}$ is given by $Q_{0n}^{\#}$, which converges to Q_0 conditional on the data, where Q_0 is the distribution of the test statistics under the null.

4. Perform single-step method to determine significance cut-off:

   (a) Let $c = c(Q, \alpha, P_n) \in \mathcal{R}^p$ be a vector cut-off such that $MT(c)$ preserves type-1 error, α.

   (b) Define

   $$R_{0n}^{\#}(c) = R(c|Q_{0n}^{\#}) = \sum_{j=1}^{J} I(|Z_{jn}^{\#}| > c_j) \qquad (6)$$

   and let $c_{0n} = c(Q_{0n}^{\#}, \alpha)$ be the common quantiles of $Q_{0n}^{\#}$ such that $Pr\left[R_{0n}^{\#}(c) \geq k\right] = \alpha$.

The resulting constant c is a multiple-comparisons-adjusted cutoff to which each observed test statistic may be compared. Because the observed test statistics have an additional layer of error introduced by estimation of posterior haplotype probabilities, we use an approximation to the formula of Louis (1982) in calculating the standard deviation, $sd(\mu_n^{\#})$ at step (2) in the above algorithm. The approximation is for computation efficiency as described in Lake *et al.* (2003) and is straightforward to derive using the `haplo.glm()` function in the R library `haplo.stats`.

3 A simulation study

The method described in Section 2 is evaluated through consideration of type-1 family-wise error (FWE). We consider an analysis of multiple haplotypes within a single gene and the null hypotheses $H_{0j} : \gamma_j = \gamma_1$, $j = 2, \ldots, M$ where M equals the number of observed haplotypes for the gene under consideration and the columns of the H matrix of Equation 1 are indicators for the presence of the corresponding haplotypes. Note that this model is an additive genetic model. Without loss of generality, we let $\gamma_1 = 0$; this parameter corresponds

Table 1: Estimated family-wise error and power for detecting haplotype effects

	FWEC		Power	
γ/σ	$N = 200$	$N = 400$	$N = 200$	$N = 400$
0.0	0.044	0.044	-	-
0.2	-	-	0.10	0.28
0.4	-	-	0.58	0.92
0.6	-	-	0.94	1.00
0.8	-	-	1.00	1.00

to the effect of the most prevalent or referent haplotype. Extension to other models is straightforward and requires only modification of the matrix H of Equation 1. The FWE for the complete null is then defined as $FWEC = Pr(\text{reject at least one } H_{0j} \mid \text{all } H_{0j} \text{ are true})$ (Westfall and Young, 1993). A simulation study is performed to investigation the control of FWEC.

Diplotypes for each individual are simulated from assumed haplotype prevalences $\pi_1, \ldots, \pi_M$ and corresponding genotype data are determined to represent the observed data. FWE is estimated by first simulating $\mathbf{Y}$ according to the null model (i.e. no haplotype effect): $\mathbf{Y} = \alpha + \epsilon$. We apply the algorithm described in Section 2.3 and record $\mathcal{K} = I\left[\sum_j I\left(|\tilde{T}_j| \geq c\right) \geq 1\right]$ where $\tilde{T}_j$ is the observed data test statistic for $j = 2, \ldots, M$. This is repeated S times to obtain $\mathcal{K}_1, \ldots, \mathcal{K}_S$ and an estimate of type 1 FWEC is given by $\widehat{FWEC} = \sum_s \mathcal{K}_s/S$.

We assume an additive genetic model and consider $M = 4$ possible haplotypes within a single gene, with prevalences equal to 0.4, 0.2, 0.2 and 0.2. The haplotype with the highest estimated prevalence is treated as the referent and Wald test statistics are calculated for the null hypotheses that each of the remaining haplotype effects is 0. We simulate $S = 500$ datasets; for each, the algorithm outlined in Section 2.3 is performed where $B = 500$ resampled data sets are drawn. Assuming $\alpha = 0.05$, sample sizes of both $N = 200$ and $N = 400$ yield $\widehat{FWEC} = 0.044$.

Further characterization of this approach is provided through investigation of estimated power. For this investigation, Y is simulated according to the alternative model in Equation 1 for a range of effect sizes (β/σ) and sample sizes and assuming an additive genetic model. In this case $S = 100$ datasets for each condition are simulated, and the data are resampled $B = 500$ times to arrive at adjusted p-values. Again, $M = 4$ haplotypes are assumed with prevalences equal to $(0.4, 0.2, 0.2, 0.2)$. Power estimates are provided in Table 1 for a single haplotype effect size ranging from 0.2 to 1.0 and sample sizes equal to $N = 200$ and 400. These results suggest a sample size of $N = 400$ will have greater than 90% power to detect a moderate effect size of 0.4.

4 Example

Potent ARTs delay the onset of clinical disease and death in HIV-1 infected individuals, but can lead to a host of drug related complications, including abnormalities in lipid profiles and possibly an accelerated risk of cardiovascular disease. The large number of available drug regimens, however, allows the potential for tailoring treatment decisions to individual patient characteristics. Our investigation aims to identify genetic polymorphisms in the infected host that modify the risk of cardiovascular disease related outcomes. Ultimately, understanding the genetic determinants of changes in lipid profiles will help guide choice of long-term treatment strategies.

We use the method described above to test for haplotypic effects within two genes, hepatic lipase (HL) and endothelial lipase (EL), on high density lipoprotein cholesterol (HDL-C) in an HIV-1 infected cohort ($N = 626$). In this section we assume an additive genetic model, but discuss alternative models in Section 5. Genotyping was performed for 4 single nucleotide polymorphisms (SNPs) in HL (rs2070895, rs12595191, rs690, rs6084) and 3 SNPs in EL (rs3829632 (-1309A/G), rs12970066, Asn396Ser). The genetic data were generated as part of the AIDS Clinical Trials Group (ACTG) New Works Concept Sheet 224 (NWCS224), and the clinical data were collected across several ACTG trials. A description of the data and genotyping methods, as well the primary analysis results, can be found in Foulkes *et al.* (2006). Due to population admixture and potential effect modification by race/ethnicity, as described for this cohort in Foulkes *et al.* (2006), all analyses are stratified by race/ethnicity. Estimated haplotype prevalences for HL and EL within racial/ethnic strata are provided in Table 2.

Fully-adjusted multivariable models are fitted for natural log transformed HDL-C assuming an additive GM for haplotype effects within EL and HL. Models are fitted for each gene separately; therefore, the MA is not specified. Covariates include age, sex, use of lipid lowering therapy, study and current drug exposures. Drug exposure variables were created for each of the 3 class

Table 2: Estimated haplotype effects on log-HDL-C and adjusted test results

	White/Non-Hispanic ($N = 317$)			**Black/Non-Hispanic** ($N = 92$)		
	Est Prev	Est Effect (se)	Test Statistic	Est Prev	Est Effect (se)	Test Statistic
EL						
AGA	0.281	0.031 (0.026)	1.19	0.211	-0.021 (0.074)	-0.29
GCA	0.202	0.074 (0.032)	**2.33***	-	-	-
ACA	0.502	REF	REF	0.742	REF	REF
HL						
GGCG	0.184	-0.007 (0.037)	-0.18	-	-	-
AGCT	0.124	-0.043 (0.045)	-0.95	0.184	-0.257 (0.133)	-1.94
AGCG	-	-	-	0.159	-.118 (0.114)	-1.03
AGTT	-	-	-	0.157	0.054 (0.122)	0.44
GGTG	-	-	-	0.116	-0.153 (0.121)	-1.26
GGCT	0.308	REF	REF	0.123	REF	REF

	Hispanic ($N = 103$)		
	Est Prev	Est Effect (se)	Test Statistic
EL			
AGA	0.205	-0.021 (0.049)	-0.42
GCA	0.325	0.030 (0.039)	0.77
ACA	0.456	REF	REF
HL			
GGCG	0.119	0.107 (0.077)	1.38
AGCT	0.102	-0.034 (0.086)	-0.40
AGCG	0.285	0.107 (0.054)	1.97
AGTT	-	-	-
GGTG	-	-	-
GGCT	0.211	REF	REF

*Indicates significance at the $\alpha = 0.05$ level based on the Pollard and van der Laan (2004) MTP. Rare haplotypes (estimated prevalence $< 10\%$) are indicated by -. REF refers to the referent haplotype and is determined based on the highest estimated prevalence within Whites/Non-Hispanics.

of drugs: protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs) and non-NRTIs (NNRTIs). PI exposure is coded as a 3-level factor for no exposure, exposure to a non-ritonavir (RTV) containing PI regimen and exposure to a RTV-containing PI-regimen. NRTI exposure is also coded as a 3-level factor for no exposure, exposure to a non-thymidine containing regimen (3TC, ABC and/or DDI) and exposure to a thymidine containing regimen (ZDV, D4T, Combivir, and/or Trizivir.) Finally, NNRTI is coded as a 2-level factor indicating any current exposure to a drug in this class. In all cases, individuals who have exposure to a drug class within 14 days prior to a blood draw for lipid measurements, but who are not receiving drug from that class at the time of the blood draw, are excluded from all analysis. Individuals with short-term exposure to any of the drug class ($<$ 21 days) at the time of the lipid measurements are also excluded.

Haplotypes with estimated prevalences of less than 10% within race / ethnicity groups are pooled as rare haplotypes for the purpose of model fitting and are included in the models but not tested. The referent haplotype is the haplotype with the highest estimated prevalence within Whites/non-Hispanics. The same referent haplotype is used across all race/ethnicities to improve interpretability of results. The referent is determined within Whites/non-Hispanics because this group has the largest number of observations and provides the most stable estimates. Adjustments for multiple testing are done within each race/ethnicity category and within each gene model. Estimated haplotype effects are provided in Table 2. These results are based on evaluation of $B = 1000$ bootstrapped data sets.

As higher levels of HDL-C are considered to be beneficial, positive coefficient estimates suggest a better HDL profile. Although the direction and magnitude of the effects appear to vary across racial/ethnic groups, this interaction could not be tested formally, because of sample size limitations. These results suggest that in White/non-Hispanics, carrying one copy of the GCA haplotype in EL results in a fold increase in HDL-C of $exp(0.074) = 1.08$ compared to the most common ACA haplotype. This increase is significantly different than 1 at the 0.05 level. Genetic associations between HL variants and HDL-C is consistent with previous reports, including those of Reilly *et al.* (2005), Ma *et al.* (2003), and deLemos *et al.* (2002). The lack of observed haplotype effects in Black/non-Hispanics and Hispanics may be due to sample size limitations and low power as evidenced in the simulation results.

5 Discussion

The multiple testing settings we consider require adjustment for covariates. When there is no justifiable choice for a parametric model, the semi-parametric approach of Yang and DeGruttola (2006) can be used. To reduce the impact of mis-modeling of the covariate effect, patients with the same haplotypes can be matched on the covariate (either exactly or within a caliper) and permutations performed on matched pairs rather than on individuals. In the context of no covariates, alternative multiple comparison adjustments, such as the maxT approach of Westfall and Young (1993) could be considered.

We use an approximation to the formula of Louis (1982) to adjust the test statistics properly for the uncertainty in haplotype probabilities, both for the observed data test statistics and the test statistics calculated using the resampled data. Estimation of the null distribution of the test statistics also requires consideration of the impact of this uncertainty on the resampling. Our simulation study shows that using the Louis formula to correct standard deviations in the denominator of the test statistics appears to adjust adequately for this uncertainty because type 1 error is preserved. For confirmation, we applied the double bootstrap approach of Beran (1987) to a few additional data sets simulated under the same model described in Section 3 and found that no futher adjustment was necessary.

We also applied the free step-down resampling approach of Westfall and Young (1993) to adjust for multiple testing. Notably, this approach requires subset pivotality, i.e. that the distribution of p-values is the same under the complete null and any subset of true nulls. While this condition holds in the context of fully observed haplotype data, it may not be valid in the context of unknown phase and small sample settings. Application of this method requires a weighted resampling of the residuals in order to account for uncertainty in haplotypes. We set these weights equal to the estimated posterior haplotype probabilities given in Equation 4. Through an additional bootstrap, we found the impact of disregarding the uncertainty in estimation of posterior haplotype probabilities for the weighted resampling to be minor. Under the simulation conditions described in Section 3 and a sample size of $N = 400$, the estimated family wise error for the complete null using the Westfall and Young approach is 0.056, with a range of 0.048 to 0.056 across 5 additional bootstraps. Power is similar for the two methods.

As with most genetic analyses in small sample settings, it would be desirable to obtain independent confirmation of the findings in Section 4. Larger

DOI: 10.2202/1557-4679.1037

sample sizes are also required to achieve adequate power to assess more complex multi-gene and gene-environment interaction models. This is particularly true in multi-ethnic populations where there is the potential for both population admixture and effect modification by race/ethnicity. Primary analysis of the data presented in this manuscript Foulkes *et al.* (2006) found differential gene and drug effects on lipids across racial/ethnic groups. This is indicative of interactions among race/ethnicity, genes and drug exposures, a phenomenon distinct from confounding by race/ethnicity. For this reason, we present all analyses stratified by race/ethnicity. Notably, a stratified analysis also addresses potential population admixture, and corresponding violations of HWE, across racial/ethnic groups.

Furthermore, consideration of alternative GMs is important, particularly when a clear GM has not been described, as is the case with the genes under consideration in this manuscript. In addition to the additive GM for each gene, we also explored dominant and recessive GM models in application of our methods. For these models, no significant haplotype associations with HDL-C within racial/ethnic strata were detectable. Our findings suggest that certain genetic characteristics put patients at increased risk for ART associated cardiovascular complications. Such analyses may ultimately inform changes in medical practice, such as a modifying ART regimens or administering lipid lowering therapy in patients at high risk of elevated HDL-C.

Interaction effects among genes in the absence of main effects are possible. While the example provided did not include this interaction analysis due to sample size limitations, investigation of the MA we considered is a step in the development of more complex models. The method presented in this manuscript will similarly ensure control of type 1 error rates in application of these models.

References

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, **74**(3), 457–468.

deLemos, A., Wolfe, M., Long, C., Sivapackianathan, R., and Rader, D. (2002). Identification of genetic variants in endothelial lipase in persons with elevated high-density lipoprotein cholesterol. *Circulation*, **106**(11), 1321–1326.

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

Foulkes, A., Wohl, D., Frank, I., Puleo, E., Restine, S., Wolfe, M., Dube, M., Tebas, P., and Reilly, M. (2006). Associations among race/ethnicity, APOC-III genotypes and lipids in HIV-1 infected individuals on antiretroviral therapy. *PLoS Medicine*, **3**(3), e52.

Ge, Y., Dudoit, S., and Speed, T. (2003). Resampling-based multiple testing for micro-array data analysis (with discussion). *Test*, **12**, 1–77.

Laird, N. M. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lake, S., Lyon, H., Tantisira, K., Silverman, E., Weiss, S., Laird, N., and Schaid, D. (2003). Estimation and testing of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity*, **55**, 56–65.

Lin, D. (2006). An efficient monte carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, **21**(6), 781–787.

Lin, D. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, **101**(473), 89–104.

Louis, T. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B*, **44**(2), 226–233.

Ma, K., Cilingiroglu, M., Otvos, J., Ballantyne, C., Marian, A., and Chan, L. (2003). Endothelial lipase is a major genetic determinant for high-density lipoprotein concentration, structure, and metabolism. *Proc. Natl. Acad. Sci.*, **100**(5), 2748–2753.

Pollard, K. and van der Laan, M. (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, **125**, 85–100.

Reilly, M., Foulkes, A., Wolfe, M., and Rader, D. (2005). Higher-order lipase gene association with plasma triglycerides. *Journal of Lipid Research*, **46**(9), 1914–22.

Schaid, D., Rowland, C., Tines, D., Jacobson, R., and Poland, G. (2002). Score tests for association between traits and haplotypes when linkage phase in ambiguous. *Am. J. Hum. Genet.*, **70**, 425–34.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, **78**(4), 629–44.

Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Westfall, P. and Young, S. (1993). *Resampling-based multiple testing*. John Wiley & Sons.

Yang, Y. and DeGruttola, V. (2006). Resampling-based multiple testing with covariate adjustment: Application to investigation of antiretroviral drug susceptibility. *in progress*.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, **82**, 171–196.

Zaykin, D., Westfall, P., Young, S., Karnoub, M., Wagner, M., and Ehm, M. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, **53**, 79–91.

Zhao, L., Li, S., and Khalid, N. (2003). A method for the assessment of disease associations with single nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am. J. Hum. Genet.*, **72**, 1231–50.