The International Journal of Biostatistics

Volume 3, *Issue* 1 2007 *Article* 12

Multiple Imputation and Random Forests (MIRF) for Unobservable, High-Dimensional Data

Bareng A. S. Nonyane, *University of Massachusetts, Amherst*
Andrea S. Foulkes, *University of Massachusetts, Amherst*

Recommended Citation:
Nonyane, Bareng A. S. and Foulkes, Andrea S. (2007) "Multiple Imputation and Random Forests (MIRF) for Unobservable, High-Dimensional Data," *The International Journal of Biostatistics*: Vol. 3: Iss. 1, Article 12.
DOI: 10.2202/1557-4679.1049

Multiple Imputation and Random Forests (MIRF) for Unobservable, High-Dimensional Data

Bareng A. S. Nonyane and Andrea S. Foulkes

Abstract

Understanding the genetic underpinnings to complex diseases requires consideration of sophisticated analytical methods designed to uncover intricate associations across multiple predictor variables. At the same time, knowledge of whether single nucleotide polymorphisms within a gene are on the same (in cis) or on different (in trans) chromosomal copies, may provide crucial information about measures of disease progression. In association studies of unrelated individuals, allelic phase is generally unobservable, generating an additional analytical challenge. In this manuscript, we describe a novel approach that combines multiple imputation and random forests for this high-dimensional, unobservable data setting. An application to a cohort of HIV-1 infected individuals receiving anti-retroviral therapies is presented. A simulation study is also presented to characterize method performance.

KEYWORDS: recursive partitioning, random forests, haplotype, genotype, phase, HIV-1, lipids

Author Notes: Corresponding author. Tele: 1-413-577-4365, E-mail: aletta@schoolph.umass.edu. Acknowledgements: We thank R. Yucel, R. Balasubramanian for helpful discussions and Andy Liaw for assistance with the RandomForests package in R. We also thank the AIDS Clinical Trials Group (ACTG) New Works Concept Sheet 224 (NWCS224) study team including Dr. M.P. Reilly who played a valuable role in collecting the data presented in this manuscript. Support for this research was provided by a National Institute of Allergy and Infectious Diseases (NIAID) Independent Research Award to ASF (R01 AI056983). This research was also supported in part by an NIH/NIDDK Research Award (R01 DK021224), the Adult ACTG funded by the NIAID (AI38858), a Center for AIDS Research Development Award from the University of Pennsylvania (5-P30 AI45008) and a NWCS224 award from the ACTG.

1 Introduction

Recent advances in sequencing technologies present a new opportunity to incorporate the molecular level characteristics of an individual in clinical decision making. However, characterizing associations among genetic polymorphisms and complex disease phenotypes continues to present several analytical challenges. These arise due to: (1) the large number of potentially informative genetic markers for disease and the complex, uncharacterized associations among them and (2) the unobservable nature of allelic phase in association studies of unrelated individuals.

Several well-described methods allow for discovering associations in the context of high-dimensional data, including semi-parametric as well as model-based approaches. These include classification and regression trees (CART) (Breiman *et al.* (1984); Zhang and Bonney (2000); Segal *et al.* (2001); Foulkes *et al.* (2004)), multi-factor dimensionality reduction and combinatorial partitioning (Ritchie *et al.* (2001); Nelson *et al.* (2001)), random forests (RF) (Breiman (2001)), multivariate adaptive regression splines (MARS) (Friedman (1991)), Bayesian variable selection (George (2000)), mixed modeling (Foulkes *et al.* (2005)), support vector machines (Huang and Kecman (2005)), and an alternative Bayesian approach (Lunn *et al.* (2006)), among others. The strength of each approach depends highly on both the scientific hypotheses under consideration and the plausible biological mechanisms for disease. Machine learning algorithms are particularly well-suiting to uncovering complex structure in high-dimensional data.

In this manuscript we consider random forests (RFs). Random forests involve constructing an ensemble of classification or regression trees and results in variable importance scores for each predictor that are aggregated over all trees (Breiman (2001)). Through resampling predictor variables at each step of the growing procedure, RFs provide a natural setting to account for collinearity among the predictors. Several recent manuscripts describe the application of random forests for discovering associations in high dimensional data settings, for example Bureau *et al.* (2005); Segal *et al.* (2004) and Diaz-Uriarte and de Andres (2006). Straightforward implementation of the RF methodology is achieved using the `randomForest` package in R. Notably, this software includes an imputation procedure for handling missing values in the predictor variables. We distinguish that in our setting, while genotype data may be missing, haplotypic phase is potentially unobservable; however, a set of haplotype pairs will be consistent with the observed genotype data

for each individual

Multiple methods have been described for estimating haplotype frequencies within a single gene and making inference between these haplotypes and a disease phenotype. These include expectation maximization (EM) type approaches (Excoffier and Slatkin (1995); Lake *et al.* (2003); Lin and Zeng (2006)) and Markov chain Monte Carlo methods (Stephens and Smith (2001)). However, methods that handle simultaneously the two analytical challenges arising from high dimensional data and uncertainty in phase, remain underdeveloped. The Bayesian approach described by Lunn *et al.* (2006) is fully likelihood based and incorporates an imputation procedure for simultaneously making inference on a large number of genotype variables and accounting for uncertainty in phase.

We present a semi-parametric approach that combines multiple imputation and random forests (MIRF) to characterize haplotype-phenotype associations. This approach involves estimation of allelic phase, imputing data according to these estimates and then combining the results of random forests across multiply imputed datasets. Notably, fitting RFs requires relatively few model assumptions. Furthermore, this combined approach can be implemented easily with slight modification of existing software tools.

The data motivating our research arise from a cross-sectional study of $N = 626$ individuals infected with Human Immunodeficiency Virus Type-1 (HIV-1) who are at risk for anti-retroviral therapy (ART) associated dyslipidemia. ARTs have demonstrated a profound effect on delaying the onset of clinical disease and death in HIV-1 infected individuals. Unfortunately, long term exposure to ARTs may lead to a number of serious health complications including the accelerated onset of cardiovascular disease. We aim to determine whether haplotypes in four candidate genes, apolipoprotein-C-III (ApoCIII), apolipoprotein-E (ApoE), endothelial lipase (EL) and hepatic lipase (HL), are associated with changes in high density lipoprotein cholesterol (HDL-c) in this population. Ultimately, characterizing the genetic polymorphisms that predispose individuals to abnormal lipid profiles may provide clinicians with the tools for making more informed treatment decisions and appropriate interventions.

DOI: 10.2202/1557-4679.1049

2 Methods

Classification and regression trees (CART) allow for characterizing associations among a large number of predictor (independent) variables and a categorical or continuous response (dependent) variable. This approach involves recursively partitioning data according to the values of the predictors in a way that minimizes the within group impurity as described in Breiman *et al.* (1984). Random forests (RFs) represent an extension of the CART methodology and are comprised of an ensemble of classification or regression trees (Breiman (2001)). In this manuscript we focus on a continuous response though straightforward extensions of the methods described can be applied to categorical responses (e.g. case/control status). We begin in this section by outlining our notation and providing a brief summary of RFs. We then introduce a multiple imputation procedure for calculating random forests in the context of phase unknown data.

2.1 Notation and Random Forests

Let $\mathbf{Y} = (y_1, y_2, ...y_n)$ be a vector of responses for the n subjects in our sample. Further suppose $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is a vector of p predictor variables for individual i, $i = 1, \ldots, n$. The root node of a tree consists of the entire sample and is split into right and left daughter nodes according the values of a predictor. The *best* split is generally defined as the one that maximizes the reduction in node impurity, given by $\phi(s, \tau)$ in Equation 1. Here τ represents node and $\tau_{s,L}$ and $\tau_{s,R}$ are the left and right daughter nodes corresponding to the split indicated by s. For a continuous Y, node impurity is commonly defined as the mean squared error (MSE) and given by $\mathcal{I}(\tau) = \frac{1}{n_\tau} \sum_{i \in \tau} (y_i - \bar{y}_\tau)^2$ where $\bar{y}_\tau$ is the mean response among individuals in τ. Extensions to multi-way splits have been described, though we limit our discussion here to the more common implementation of CART based on binary splits.

$$\phi(s, \tau) = \mathcal{I}(\tau) - \mathcal{I}(\tau_{s,L}) - \mathcal{I}(\tau_{s,R}) \tag{1}$$

Splitting is done recursively until a stopping rule is met. Usually a node consisting of less than 5 observations is not split further. Typically, some nodes of the tree are then removed, a process commonly referred to as *pruning* that minimizes over-fitting. In the CART setting, minimal cost-complexity

cross validation pruning is generally used to achieve balance between model complexity and predictive accuracy as described in Breiman *et al.* (1984). The random forests procedure for determining variable importance is described in detail in Breiman (2001) and summarized in Algorithm 1 below. Notably, this approach does not require pruning of each tree since multiple trees are created using bootstrapped samples. Furthermore, through sampling a subset of variables at each split (step (2) of Algorithm 1), RFs allow for capturing information on multiple, collinear variables. This method can be applied in R using the `randomForest` package.

ALGORITHM 1

Initialize $b = 0$.

1. Let $b = b + 1$. Draw a random sample of size n_1 with replacement and call this a learning sample. This is about $2n/3$ of the observed sample data where n is the total number of subjects in the sample. The remaining observations of size n_2 (about $n/3$ of data) form the out-of-bag (OOB) data.

2. Generate an unpruned regression tree, using a randomly selected subset of the p predictors at each node and a prespecified measure of node purity.

3. Record the MSE for the OOB data and call this π_b.

4. Using the OOB data, for each $j = 1, \ldots, p$, permute the predictor variable $\mathbf{x}_j$ and record the MSE. Denote this by π_{bj} and define importance as $d_{bj} = (\pi_{bj} - \pi_b)$.

5. Repeat steps (1)-(4) B times to obtain B trees and $d_{1j}, \ldots, d_{Bj}$ for $j = 1, \ldots, p$. Let $\widehat{\theta}_j = \frac{1}{B} \sum_{b=1}^{B} d_{bj}$ be the jth predictor variable's mean importance score across all trees and let s_j be the corresponding standard error.

The parameters of interest in a random forest are the mean variable importance scores for a predictor variable x_j, estimated by $\widehat{\theta}_j$ and the corresponding standardized importance given by $z_j = \widehat{\theta}_j / s_j$ for $j = 1, \ldots, p$ where s_j is the sample standard deviaion. The interpretation and use of these

DOI: 10.2202/1557-4679.1049

importance scores varies across studies. One approach is to treat the standardized scores as standard normal deviates and to determine significance level according to a normal distribution as described in Breiman (2004). In this case, appropriate adjustment for the inflation of type-1 error resulting from testing p predictors is required. For example, the bootstrap-based resampling approach to multiple testing of van der Laan *et al.* (2004a,b) and Pollard and van der Laan (2004) could be applied. Alternatively, the rank order of resulting importance scores can be reported for a single or repeated applications of RFs as described in Bureau *et al.* (2005); Segal *et al.* (2004) and Diaz-Uriarte and de Andres (2006). In this manuscript we focus on the latter approach since it requires fewer assumptions and can be useful for identifying candidate predictors for further investigation.

2.2 Accounting for uncertainty in phase

In the case that haplotype data are fully observed, application of the RF procedure described above is straightforward. For example, the $(x_{i1}, x_{i2}, ..., x_{ip})$ can be defined as indicators for the presence of haplotypes $1, \ldots, p$, respectively for individual i. These haplotype indicators can be within a single gene or across multiple genes. In general, however, the x_{ij} are unobservable in data arising from unrelated individuals in which allelic phase is not known. Specifically, if an individual is heterozygous at two or more single nucleotide polymorphisms (SNPs) within a gene, then the corresponding haplotype pair (referred to as diplotype) for this individual is not known with certainty. However, haplotype frequencies can be estimated from the observed genotype data and in turn, used to calculate posterior probabilities for all haplotype pairs that are consist with an individual's observed genotype.

In order to address haplotype uncertainty, we propose estimating these posterior haplotype probabilities, multiply imputing haplotype data using these estimates and fitting a random forest for each imputation. This process is repeated multiple times in order to account for the variability derived from the imputation procedure. The combined imputation and RF procedure is given in Algorithm 2 below. Note that the length of $\mathbf{r}_{ik}$ in step (1) equals the size of $\mathcal{H}_{ik}$ where $\mathcal{H}_{ik}$ is the set of all haplotypes that are consistent with individual i's observed genotype for gene k. Furthermore, if an individual's haplotype is fully observable for a given gene, then $\mathbf{r}_{ik}$ will reduce to a scalar equal to 1 corresponding to this diplotype. Straightforward implementation

of this initial step is achieved using the `haplo.em()` function of the R package `haplo.stats`.

ALGORITHM 2

Initialize $m = 0$

1. Apply the EM approach of Excoffier and Slatkin (1995) to arrive at posterior haplotype probabilities for each subject in our dataset, for each gene under consideration. Denote these individual level probabilities by the vector $\mathbf{r}_{ik}$ where i indicates individual, k indicates gene and the elements of this vector correspond to the posterior probabilities of the diplotypes consistent with the individual's observed genotype.

2. Let $m = m + 1$. For each individual i and gene k, sample a single diplotype with probabilities $\mathbf{r}_{ik}$, until a complete data set is obtained.

3. Fit a random forest according to ALGORITHM 1 using the dataset imputed in step (2) and record importance scores $\widehat{\theta}_j^m$ and their standard errors s_j^m, $j =, \ldots, p$.

4. Repeat steps (2)-(3) M times to arrive at $\widehat{\theta}_j^1, \ldots, \widehat{\theta}_j^M$ and $s_j^1, \ldots, s_j^M$ for each predictor variable x_j, $j = 1, \ldots, p$.

Combining importance scores, $\widehat{\theta}_j^1, \ldots, \widehat{\theta}_j^M$ across imputed datasets requires consideration of both the between and within-imputation variance. We use the approach described in Little and Rubin (2002) Section 5.4 to combine these scores. Specifically, for each variable x_j, we let $\bar{\theta}_{jM} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}_j^m$ be the average importance score across the M imputations and define $T_j = \left(\bar{\theta}_{jM}\right) V_{jM}^{-1/2}$ where V_{jM} is the sum of the within ($\bar{W}_{jM}$) and the between (B_{jM}) imputation variances as given in Equation 2. The resulting T_j are ordered and the maximum (or maximum subset) reported as potentially informative. In the remainder of this manuscript, these are referred to as adjusted variable importance scores.

DOI: 10.2202/1557-4679.1049

$$
\begin{aligned}
V_{jM} &= \bar{W}_{jM} + \frac{M+1}{M} B_{jM} \\
\bar{W}_{jM} &= \frac{1}{M} \sum_{m=1}^{M} (s_j^m)^2 \\
B_{jM} &= \frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}_j^m - \bar{\theta}_{jM} \right)^2
\end{aligned}
\tag{2}
$$

3 A Simulation Study

A simulation study is presented to describe the performance of RFs in both the observed and unobservable haplotype settings assuming a variety of underlying biological and clinical models. Presently, to our knowledge, an alternative machine learning algorithm for unobservable predictors has not been described. Thus, for the purpose of comparison, we present the results of a more standard application of the generalized linear modeling (GLM) for unobservable phase approach of Lake *et al.* (2003). This approach employs an EM-type algorithm that iterates between estimation of haplotype population frequencies and haplotype effects on the phenotype of interest. Straightforward implementation of this method is available with the `haplo.glm()` function in the `haplo.stats` package in R. Notably, unlike machine learning algorithms, the GLM approach is not designed specifically to uncover complex structure and therefore we focus on relatively simple (2-gene) models of associations for the purpose of illustration.

A summary of the components to each assumed model is given in Table 1 for the MIRF approach and Table 2 for the GLM approach. These include additive models assuming independent predictors (Models 1-3), additive models assuming haplotypes from associated genes as predictors (Models 4 and 5), interaction models with and without main effects (Models 6 and 7) and a conditional dependence model (Model 8). In all cases, 4 haplotypes within each of 4 genes are simulated with frequencies of 0.2, 0.2, 0.2 and 0.4. A continuous phenotype y is generated according to the indicated model for $i = 1, \ldots, n$ for $n = 500$ as well as $n = 1000$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The intercept is set equal to 0 for simplicity.

Table 1: Simulation study results for MIRF

Model	Observed Haplotypes Max_1	Observed Haplotypes Max_2	Observed Haplotypes $Max_{1,2}$	Unobservable Haplotypes Max_1	Unobservable Haplotypes Max_2	Unobservable Haplotypes $Max_{1,2}$
ADDITIVE (INDEPENDENT PREDICTORS) MODELS						
1. Single haplotype effect $y_i = \beta_1 I(H_{11} \in h_{1i}) + \epsilon_i$ Condition: $\beta_1/\sigma = 0.5$	0.88/ 0.99	-	-	0.69/ 0.99	-	-
2. Two varying haplotype effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = 0.6, \beta_2/\sigma = 0.4$	0.52/ 0.60	0.12/ 0.13	0.64/ 0.79	0.31/ 0.40	0.11/ 0.14	0.42/ 0.54
3. Two equal haplotype effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = \beta_2/\sigma = 0.5$	0.50/ 0.54	0.50/ 0.46	0.92/ 0.95	0.39/ 0.46	0.52/ 0.50	0.55/ 0.70
ADDITIVE (ASSOCIATED PREDICTORS) MODELS						
4. Dependence with single haplotype effect $y_i = \beta_1 I(H_{11} \in h_{1i}) + \epsilon_i$ Condition: h_{1i} and h_{2i} associated, $\beta_1/\sigma = 0.5$	0.87/ 1.00	0.01/ 0.00	0.03/ 0.06	0.59/ 0.91	0.01/ 0.01	0.02/ 0.02
5. Dependence with two haplotype effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: h_{1i} and h_{2i} associated, $\beta_1/\sigma = \beta_2/\sigma = 0.5$	0.45/ 0.53	0.55/ 0.49	0.85/ 0.98	0.41/ 0.46	0.49/ 0.51	0.46/ 0.70
MULTIPLICATIVE (INTERACTION) MODELS						
6. Interaction without main effects $y_i = \beta_{12} I(H_{11} \in h_{1i}) I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_{12}/\sigma = 0.5$	0.21/ 0.27	0.28/ 0.44	0.16/ 0.35	0.19/ 0.17	0.16/ 0.22	0.01/ 0.06
7. Interaction with main effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) +$ $\beta_{12} I(H_{11} \in h_{1i}) I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = \beta_2/\sigma = 0.4$ $\beta_{12}/\sigma = 0.6$	0.44/ 0.53	0.56/ 0.47	1.00/ 1.00	0.45/ 0.62	0.53/ 0.37	0.65/ 0.90
CONDITIONAL MODEL						
8. Conditional dependence $y_i = \beta_1 I(H_{11} \in h_{1i}) +$ $\beta_{12} I(H_{11} \in h_{1i}) I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = \beta_{12}/\sigma = 0.5$	1.00/ 1.00	0.00/ 0.00	0.38/ 0.58	0.92/ 0.97	0.02/ 0.01	0.19/ 0.18

$I(H_{11} \in h_{1i})$ and $I(H_{21} \in h_{2i})$ are indicators for the presence of haplotypes H_{11} and H_{21}, respectively in the observed haplotype pairs for genes 1 and 2 in the ith individual. The proportions Max_1, Max_2 and $Max_{1,2}$ are given in pairs corresponding to sample sizes $n = 500$ and $n = 1000$.

DOI: 10.2202/1557-4679.1049

Relative rankings of haplotypes within each model across simulations are reported in Table 1. These are based on the importance score of the RF as discussed in Section 2.2. In this table Max_1 and Max_2 are defined as the proportion of simulations for which H_{11} or H_{21} respectively has the maximum associated ranking across all haplotypes. $Max_{1,2}$ is defined as the proportion of simulations for which the rankings corresponding to H_{11} and H_{21} are both greater than all other haplotype importance scores. Results are reported assuming haplotypes are fully observed and under the assumption of potential ambiguity, using MIRF and a GLM. $S = 100$ simulations are performed for each model. An additional 10 samples are drawn per simulation for the MIRF unobservable setting in order to characterize the between-imputation variance as described in Section 2.2. The proportions Max_1, Max_2 and $Max_{1,2}$ in Table 1 are given in pairs corresponding to sample sizes $n = 500$ and $n = 1000$.

Model 1 assumes a single haplotype effect among the 16 available haplotype variables (4 in each of 4 genes.) A moderate effect size of 0.5 is assumed. In the fully observed setting with a sample size of $n = 1000$, the correct haplotype is selected in 99% of the simulations. As expected, corresponding power decreases in the unobservable setting to 91%, reflecting an efficiency of about 92%. Models 2 and 3 assume two haplotypes, 1 in each of 2 genes are predictive of the phenotype. In both cases, the two predictor variables were simulated assuming independence between them. In Model 2, the effect sizes of these two variables are assumed to differ, equalling 0.6 and 0.4 respectively, while in Model 3, we assume equal effect sizes of 0.5. In this context of two additive haplotype effects, RFs perform better with equal effects (Model 3) than with varying effects (Model 2) with power measured by $Max_{1,2}$ for $n = 1000$ equal to 95% in the observed setting and 70% in the unobservable setting.

Models 5 and 6 both assume correlation between two haplotypes across two different genes. That is, these models assume that the presence of a specific haplotype in one gene is correlated with the presence of a specific haplotype in another gene. In Model 4, only one of these haplotypes is additionally associated with the outcome and is assumed to have a moderate effect size of 0.5. Model 5, on the other hand, assumes the two correlated haplotypes are both predictive of the outcome with equal effect sizes of 0.5. A sample size of $n = 100$ yields 100% and 91% power for detecting the single effect (Model 4) in the observed and unobservable settings respectively. In the case of two predictor variables that are correlated (Model 5), the results

are similar to the additive model with independent predictors (Model 3).

Model 6 assumes that the presence of two independent haplotypes across two genes are predictive of the outcome though neither haplotype alone is predictive. Notably, RFs do not perform well in this setting of interaction in the absence of main effects (Model 6). This finding is consistent with the fact that trees are generated recursively by first splitting on single variables; thus if haplotypes are not predictive singly, then they will not be detectable. Model 7 assumes interaction in the more standard statistical sense in that main effects as well as an interaction term are included in the model. In this setting, reasonable power for detecting the two haplotypes (100% and 90%) is achieved for both the observed and unobservable setting. Finally, Model 8 is a conditional model in which a haplotype in gene 1 is predictive of the outcome and another haplotype in gene 2 is only predictive in the presence of the haplotype in gene 1. In this case, the power is reasonable for detecting the haplotype in gene 1 though the detection rates for the haplotype in the second gene are greatly reduced.

Overall, application of multiple imputation in combination with RFs for the unobservable haplotypic phase setting results in a loss of efficiency (compared to RF alone) for detecting true underlying associations. This finding is expected and a common phenomenon resulting from the loss of information due to missing data. Interestingly, the GLM for unobservable phase performs relatively well for these simple models. These results are provided in Table 2 and the ranks are based first on F-statistics for overall gene effects and then Wald statistics for individual haplotype effects. Models are fitted separately for each gene in this setting. Specifically, detection rates for Model 6 are remarkably better in the GLM framework. As noted above, however, RFs are not well suited to interaction in the absence of main effects.

The global null model of no association between haplotypes and the phenotype is also simulated $S = 500$ times again assuming sample sizes of $n = 500$. The error rate defined as the maximum proportion of times a single haplotype has the largest rankings across the simulations is 0.11 and 0.12 in the observed and unobservable settings respectively using RFs. A similar error rate of 0.11 for the GLM approach was observed.

DOI: 10.2202/1557-4679.1049

Table 2: Simulation study results for GLM

Model	Unobservable Haplotypes Max_1	Max_2	$Max_{1,2}$
ADDITIVE (INDEPENDENT PREDICTORS) MODELS			
1. Single haplotype effect $y_i = \beta_1 I(H_{11} \in h_{1i}) + \epsilon_i$ Condition: $\beta_1/\sigma = 0.5$	0.92/1.00	-	-
2. Two varying haplotype effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = 0.6, \beta_2/\sigma = 0.4$	0.84/0.90	0.41/0.43	0.28/0.46
3. Two equal haplotype effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = \beta_2/\sigma = 0.5$	0.40/0.54	0.45/0.43	0.60/0.95
ADDITIVE (ASSOCIATED PREDICTORS) MODELS			
4. Dependence model with single haplotype effect $y_i = \beta_1 I(H_{11} \in h_{1i}) + \epsilon_i$ Condition: h_{1i} and h_{2i} associated, $\beta_1/\sigma = 0.5$	0.95/1.00	0.00/0.00	0.00/0.02
5. Dependence model with two haplotype effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: h_{1i} and h_{2i} associated, $\beta_1/\sigma = \beta_2/\sigma = 0.5$	0.45/0.53	0.44/0.47	0.64/0.97
MULTIPLICATIVE (INTERACTION) MODELS			
6. Interaction without main effects $y_i = \beta_{12} I(H_{11} \in h_{1i}) I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_{12}/\sigma = 0.5$	0.26/0.33	0.21/0.42	0.08/0.41
7. Interaction with main effects $y_i = \beta_1 I(H_{11} \in h_{1i}) + \beta_2 I(H_{21} \in h_{2i}) +$ $\beta_{12} I(H_{11} \in h_{1i}) I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = \beta_2/\sigma = 0.4, \beta_{12}/\sigma = 0.6$	0.48/0.60	0.52/0.39	0.97/0.97
CONDITIONAL MODEL			
8. Conditional dependence $y_i = \beta_1 I(H_{11} \in h_{1i}) +$ $\beta_{12} I(H_{11} \in h_{1i}) I(H_{21} \in h_{2i}) + \epsilon_i$ Condition: $\beta_1/\sigma = \beta_{12}/\sigma = 0.5$	1.00/1.00	0.00/0.00	0.12/0.33

4 Example: ART associated dyslipidemia in HIV

We present an application of MIRF using data on $n = 626$ individuals, collected as part of AIDS Clinical Trials Group (ACTG) New Works Concept Sheet 224 (NWCS224). A primary aim of this study is to identify haplotypes within and across multiple candidate genes that are associated with an increased risk of dyslipidemia in HIV-1-infected individuals who are on combination antiretroviral therapy (ART). First stage analysis results, including complete demographic and clinical information on this cohort are presented in Foulkes *et al.* (2006).

Single nucleotide polymorphisms in each of four genes, ApoC-III [-482C/T (rs2854117), -455T/C (rs2854116), intron 1 (466) G/C (rs2070669), Gly34Gly C/T (rs4520), exon 4 SstI 4348(5) C/G (rs5128)], ApoE [Arg112Cys T/C (rs429358), Arg158Cys T/C (rs7412)], EL [rs12970066, Asn396Ser, rs3829632 (-1309A/G)] and HL [rs2070895, rs12595191, rs690, rs6084] and their effects on HDL-C are considered. Haplotype frequency estimates are obtained within each racial/ethnic group separately due to potential violations of Hardy Weinberg equilibrium when racial/ethnic groups are combined. Resulting estimates are provided in Table 3. All haplotypes with estimated frequencies of greater than 0.01 are included in analysis and those with an estimated frequency of greater than 0.05 in at least one racial/ethnic group is presented.

A linear multivariable model for log(HDL-C) is fitted with independent variables for current drug exposures, race/ethnicity, gender, age, study and use of lipid-lowering therapy. The resulting model residuals are calculated and treated as the outcome in the RF analysis in order to assess genetic effects after accounting for these traditional risk factors for cardiovascular disease. Individuals excluded from analysis include those with unknown durations of drug exposure, short washout periods or short drug exposures (n=60), individuals for whom race/ethnicity is self-reported as other (n=13), and those individuals with missing HDL-c (n=41).

RFs are applied after first stratifying by race/ethnicity due to potential effect modification by race/ethnicity, as described for this cohort in Foulkes *et al.* (2006). A total of $M = 500$ multiply imputed datasets are generated assuming the estimated haplotype frequencies given in Table 3. The resulting adjusted variable importance scores T_j are reported in Table 3. Inter-

Table 3: Estimated haplotype frequencies and importance scores by race/ethnicity

Gene	Haplotype†	White/Non-Hispanic ($N = 317$) Est Freq	T_j	Black/Non-Hispanic ($N = 92$) Est Freq	T_j	Hispanic ($N = 103$) Est Freq	T_j
ApoC3							
	CCCCC	0.006	0.04	< 0.001	$-$	0.007	0.75
	CCGCC	0.123	0.34	0.013	0.09	0.045	0.93
	CTCTC	0.034	0.40	0.052	0.32	0.010	0.30
	CTCTG	< 0.001	$-$	0.059	0.06	0.005	0.07
	CTGCC	0.448	0.20	0.102	1.10^{**}	0.358	0.03
	CTGTC	0.094	0.05	0.062	0.33	0.181	1.34
	TCCCC	0.134	0.16	0.579	0.69	0.194	0.43
	TCCTG	0.079	1.19	0.064	0.32	0.155	0.00
ApoE							
	CC	0.146	0.79	0.222	0.19	0.15	1.41
	TC	0.784	0.72	0.710	0.09	0.82	0.01
	TT	0.069	2.24^{**}	0.068	0.57	0.03	0.03
HL							
	AGCG	0.096	1.47	0.158	0.27	0.281	1.52
	AGCT	0.124	0.94	0.181	0.33	0.108	0.59
	AGTG	0.014	0.38	0.040	0.36	0.052	0.51
	AGTT	< 0.001	$-$	0.160	0.36	0.018	0.35
	GACG	0.054	0.27	0.017	0.09	0.040	0.02
	GACT	0.075	0.07	0.005	0.18	0.047	0.67
	GGCG	0.184	1.71	0.069	1.18^{*}	0.123	1.09
	GGCT	0.318	1.35	0.135	0.57	0.206	1.34
	GGTG	0.068	1.19	0.131	0.54	0.026	0.42
	GGTT	0.050	0.27	0.080	0.36	0.057	0.01
EL							
	ACA	0.502	4.11^{*}	0.750	0.46	0.470	2.75^{**}
	AGA	0.281	2.13	0.211	0.20	0.205	4.28^{*}
	GCA	0.201	1.80	0.027	0.41	0.325	2.42

† Haplotypes with estimated frequencies of greater than 0.05 in at least one race/ethnicity group are presented. The number of haplotypes with frequencies greater than 0.01 across all four genes are 38 in Whites, 36 in Blacks and 34 in Hispanics. $-$ indicates T_j is undefined due to small sample size. $*$ indicates largest adjusted variable importance score within race/ethnicity group. $**$ indicates second largest adjusted variable importance score within race/ethnicity group.

estingly, haplotypes EL-ACA, EL-AGA and ApoE-TT are the most predictive of HDL-C in Whites and EL-ACA, EL-AGA and EL-GCA in Hispanics while haplotypes within two other genes, HL-GGCG and ApoC3-CTGCC are most predictive in Blacks. The adjusted importance statistics for EL-ACA in Whites and EL-AGA in Hispanics were the maximum in 77% and 85% respectively, of the imputations.

5 Discussion

In this manuscript we present a combination of two existing analytical techniques, random forests and multiple imputation. Together these methods allow for assessing a large number of potential genetic effects on a complex trait while accounting for the unobservable nature of haplotypic phase in association studies of unrelated individuals. In our investigation, relative variable importance scores are reported and characterized for both the observed and unobservable settings under different model assumptions. As expected, a loss of power is observed in the context of missing haplotype information; however, detection rates remain reasonable in relation to the standard application of RFs. In the context of simple models of association, application of the GLM method for unobservable phase perform relatively well.

Notably, MIRF has the advantage of requiring fewer parametric assumptions than traditional modeling techniques such as GLM. In addition, machine learning algorithms such as RFs are well suited to identify higher order structure. The proposed method also allows for discovery of multiple haplotypes associated with the disease phenotype within and across racial/ethnic groups. In the analysis presented, stratification by race/ethnicity prior to tree fitting was done to account for potential effect modification. While the HWE assumption is needed to estimate posterior haplotype probabilities, once these probabilities are determined, multiply imputing data and applying RFs does not require the HWE assumption. Since haplotypic structure varies significantly across race/ethnicity, we recommend stratified analysis in general.

Accounting for potential confounding and effect modification by demographic and clinical factors is an important component to the analysis of population level data. In the example provided in Section 4, the primary outcome of interest is a continuous variable (HDL cholesterol). We were therefore able to apply a first stage linear regression to the data and use the

residuals from this model fitting procedure as the outcome in the application of MIRF. This allows us to assess the variability explained by genetic factors that is above and beyond traditional risk factors for disease. In general, and in the case/control data setting, these covariates can be used as additional predictor variables in the RF procedure.

The approach described herein is based on relative importance of haplotype indicators. Alternatively, a resampling-based approach such as described in van der Laan *et al.* (2004a,b) and Pollard and van der Laan (2004) could be applied to assess significance. Resampling procedures represent a natural approach for making inference in the context of a large number of potentially informative, correlated predictors while not requiring distributional assumptions. Notably, however, the combination of resampling and multiple imputation would result in a computationally intensive procedure. Finally, our presentation does not account for the error introduced from estimation of the posterior probabilities used for resampling; however, bootstrapping the data and re-estimating these probabilities showed minimal variation in the estimates. The proposed extension of RFs will allow for further exploratory investigations for high-dimensional unobservable data.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5–32.

Breiman, L. (2004). http://www.stat.berkeley.edu/∼breiman/RandomForests/.

Breiman, L., Friedman, J., Ohlsen, R. A., and J., S. C. (1984). *Classification and Regression Trees*. Chapman and Hall/ CRC.

Bureau, A., Dupuis, J., Lunetta, K. L., Hayward, B., Keith, T. P., and Van Eerdewegh, P. (2005). Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, **28**(2), 171–182.

Diaz-Uriarte, R. and de Andres, A. S. (2006). Gene selection and classification of microarray data using random forests. *BMC Bioinformatics*, **7**(3).

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. S., De Gruttola, V., and Hertogs, K. (2004). Combining genotype groups and recursive partitioning: an application to Human Immunodeficiency Virus type-1 genetics data. *Applied Statistics*, **53**, 311–323.

Foulkes, A. S., Reilly, M., Zhou, L., and Rader, D. J. (2005). Mixed modelling to characterize genotype-phenotype associations. *Statistics in Medicine*, **24**, 775–789.

Foulkes, A. S., Wohl, D. A., Frank, I., Puleo, E., a. R. S. D. M. P., Tebas, P., and Reilly, M. P. (2006). Associations among race/ethnicity, ApoC-III genotype, and lipids in HIV-1-infected individuals on antiretroviral therapy. *Plos Medicine*, **3**, 1–11.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**(1), 1–141.

George, E. (2000). The variable selection problem. *Journal of the American Statistical Association*, **95**, 1304–1308.

Huang, T. M. and Kecman, V. (2005). Gene extraction for cancer diagnosis by support vector machines - an improvement. *Artificial Intelligence in Medicine*, **35**, 185–194.

Lake, S., Lyon, H., Tantisira, K., Silverman, E., Weiss, S., and Schaid, D. (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 (with discussion). *Journal of the American Statistical Association*, **101**, 89–104.

Little, R. J. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc. Hoboken, New Jersey.

Lunn, D., Whittaker, J. C., and Best, N. (2006). A Bayesian toolkit for genetic association studies. *Genetic Epidemiology*, **30**, 231–247.

Nelson, M., Kardia, S., Ferrell, R., and Sing, C. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, **11**, 458–470.

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

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, R., Dupont, W. D., Parl, F., and Moore, J. H. (2001). Multifactorial dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, **69**(1), 138–147.

Segal, M. R., Barbour, J. D., and Grant, R. M. (2004). Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), Article 2.

Segal, R., Cummings, M. P., and Hubbard, A. (2001). Relating amino acid sequence to phenotype: Analysis of peptide-binding data. *Biometrics*, **57**, 632–643.

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

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004a). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), Article 15.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004b). Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), Article 14.

Zhang, H. and Bonney, G. (2000). Use of classification trees for association studies. *Genetic Epidemiology*, **19**, 323–332.