Imputation Methods to Improve Inference in SNP Association Studies

James Y. Dai,[1†*] Ingo Ruczinski,[2†] Michael LeBlanc[1,3] and Charles Kooperberg[1,3]

[1]*Department of Biostatistics, University of Washington, Seattle, Washington*
[2]*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland*
[3]*Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington*

Missing single nucleotide polymorphisms (SNPs) are quite common in genetic association studies. Subjects with missing SNPs are often discarded in analyses, which may seriously undermine the inference of SNP-disease association. In this article, we develop two haplotype-based imputation approaches and one tree-based imputation approach for association studies. The emphasis is to evaluate the impact of imputation on parameter estimation, compared to the standard practice of ignoring missing data. Haplotype-based approaches build on haplotype reconstruction by the expectation-maximization (EM) algorithm or a weighted EM (WEM) algorithm, depending on whether case-control status is taken into account. The tree-based approach uses a Gibbs sampler to iteratively sample from a full conditional distribution, which is obtained from the classification and regression tree (CART) algorithm. We employ a standard multiple imputation procedure to account for the uncertainty of imputation. We apply the methods to simulated data as well as a case-control study on developmental dyslexia. Our results suggest that imputation generally improves efficiency over the standard practice of ignoring missing data. The tree-based approach performs comparably well as haplotype-based approaches, but the former has a computational advantage. The WEM approach yields the smallest bias at a price of increased variance. *Genet. Epidemiol.* 30:690–702, 2006. © 2006 Wiley-Liss, Inc.

Key words: EM algorithm; Gibbs sampler; CART; linkage disequilibrium; multiple imputation

[†]Contributed equally to this work.

Contract grant sponsor: NIH; Contract grant numbers: CA 74841, CA105069, CA 53996, CA 90998, and HL74745.

*Correspondence to: James Y. Dai, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N/M3-A410, Seattle, WA 98109, USA. E-mail: yud@u.washington.edu

Received 30 December 2005; Revised 14 June 2006; Accepted 18 July 2006
Published online 19 September 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20180

INTRODUCTION

It is widely recognized that complex diseases are likely caused by multiple susceptible loci, each contributing a small to medium amount to the disease risk, that are potentially interacting with each other [Risch, 1990, 2000; Botstein and Risch, 2003]. While linkage analysis shows to be largely ineffective, association studies, in which the frequencies of marker alleles in affected individuals and controls (either population- or family-based) are compared, may hold the promise of dissecting the genetic susceptibility of complex diseases [Risch, 2000; Botstein and Risch, 2003]. With the explosion of single nucleotide polymorphism (SNP) discovery and the advances in genotyping technologies, numerous SNP-based association studies have been carried out in a scale ranging from a few candidate genes to the whole genome [Barnby et al., 2005; Cope et al., 2005; Hu et al., 2005]. Despite the increasingly improved cost efficiency and call rate in genotyping, missing SNP data are fairly common in these association studies, sometimes with a rate of 5–10%. Although highly desirable, re-genotyping the missing ones is often not practical due to financial constraint.

Depending on the analytical strategy undertaken, the missing SNPs have different impact on association inference. The haplotype approach treats a collection of adjacent SNPs in linkage disequilibrium (LD) all together and models the disease-haplotype association [Schaid et al., 2002; Zhao et al., 2003; Epstein and Satten, 2003; Stram et al., 2003]. Missing SNPs are essentially imputed in the process of haplotype reconstruction with a cost of extra variation. Although haplotype analysis is effective to model SNPs in a tight LD block, it runs into difficulties when there are a large number of SNPs under investigation and LD blocks are not well defined (for example, the 10 SNP developmental dyslexia (DD) data in this article). In view of the polygenic nature of complex diseases, an alternative strategy is to directly regress disease status on the SNP main effects and SNP-SNP interactions. Cordell and Clayton [2002] proposed a stepwise logistic regression procedure for both case-control data and family data. Ruczinski et al. [2003] developed logic regression, an adaptive regression methodology well suited for detecting interactions between binary SNP variables. These SNP-based approaches build the regression model by search algorithms, offer a flexible choice of hypothesis testing, yet remain computationally tractable. However, missing data in SNP genotypes pose a serious problem to regression approaches. This paper is concerned with imputation methods to improve inference in association studies that use SNPs as predictors.

The standard procedure to cope with the missing SNPs is to ignore the individuals that have missing values in the SNP loci under investigation (but to include individuals that may have missing data for other SNPs, in a sense "all-available" data for the model), the so-called complete-case analysis. In general, the complete-case analysis reduces the effective sample size and potentially introduces bias in parameter estimates [Greenland and Finkle, 1995]. In particular, if a large number of SNPs are under investigation simultaneously [for example, logic regression; Ruczinski et al., 2003], the proportion of individuals with at least one missing value can be quite high, even if the rate of missing SNPs is low for each locus. Given that neighboring SNPs are in LD, it is feasible to impute missing SNPs by borrowing information from the observed ones. Furthermore, the imputation may also benefit from incorporating information on disease status and other covariates. For example, when we studied the association between breast cancer and polymorphisms in the XPD gene in a matched case-control study, the imputation frequencies for missing SNPs relied strongly on disease status and whether there was a family history of breast cancer [Brewster et al., 2006]. It is therefore desirable to develop a flexible imputation approach which takes into account LD in neighboring SNPs, as well as disease status and covariates if they are relevant.

The aforementioned haplotype reconstruction can be used for imputation even when the regression modeling involves individual SNPs. Existing expectation-maximization (EM) algorithms [Excoffier and Slatkin, 1995; Qin et al., 2002] accommodate missing SNPs by first replacing the missing locus with all possible alleles. After haplotype reconstruction, the missing SNP genotypes are filled by sampling compatible haplotypes from their conditional distributions given the unphased genotypes. Similarly, Bayesian methods for haplotype reconstruction can be use for imputation [Stephens et al., 2001; Niu et al., 2002; Lin et al., 2002]. All these methods may over-simplify the haplotype distribution in case-control samples, as the frequencies of the disease-associated haplotypes may differ between cases and controls. To alleviate this problem, Lake et al. [2003] used a weighted EM (WEM) algorithm to jointly model the haplotype effects and haplotype frequencies. Epstein and Satten [2003] developed a retrospective likelihood, and estimated haplotype frequencies separately in cases and in controls. These methods can be easily adapted to imputing missing SNPs, with disease status and extra covariates being accounted for.

Alternatively, nonparametric regression methods such as classification and regression trees (CART) [Breiman et al., 1984] can be used to model the missing SNPs without having to reconstruct haplotypes. Recently there is growing interest in applying tree methods to association studies with a large number of SNPs [Zhang and Bonney, 2000; Bureau et al., 2005], in an attempt to unravel the interactions of SNP-SNP and SNP-covariate. For the purpose of imputation here, we regress each SNP locus with missing data on the other SNP loci, the covariates and the disease status, build the tree, and predict the missing data at the locus. In order to obtain the joint distribution of missing SNPs at different loci, we employ a Gibbs sampler which iteratively cycles the regression and prediction by CART through loci with missing SNPs. One advantage of CART is that it deals with missing data by surrogate splits. That is, after choosing the best predictor and split point using the available data, a list of surrogate variables and split points are formed by comparing the performance of the alternate predictor with the primary predictor. If a primary predictor is missing for one individual, we use the secondary predictor if available, and so on.

In this article our main goal is to demonstrate the benefit of a reasonable imputation strategy over simply ignoring the missing data in association studies that use individual SNPs as predictors. A secondary goal is to compare the haplotype-based and CART-based imputation approaches we developed. In particular, we consider the EM and WEM algorithm as two representatives of haplotype-based approaches because of their easy implementation. We choose the case-control design as an illustrative example since it is the most commonly used in association studies. By comparing the imputation accuracy, bias, and efficiency in inference, we evaluate the potential of the tree-based approach as compared to the haplotype-based approaches, and assess the WEM approach as opposed to the regular EM approach.

METHODS

Assume we have a case-control study with $i = 1, 2, \ldots, n$ unrelated individuals. Let $\mathbf{D}_i = 1$ if individual i is a case and $\mathbf{D}_i = 0$ otherwise, and let $\mathbf{G}_i = (g_{i1}, g_{i2}, \ldots, g_{iK})$ be the unphased SNP data on individual i at K loci of interest. Some of the g_{iK} may be missing. Assume that in the population there are m possible haplotypes $h_1, h_2, \ldots, h_m$ with (unknown) population frequencies $\mathbf{p} = (p_1, p_2, \ldots, p_m)$. In addition to the genetic information we also have information on r covariates $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$. Throughout this article, we assume missingness at random (MAR), that is, the missing data mechanism depends on observed data, not on unobserved data [Little and Rubin, 1987]. While the hypothesis of MAR cannot formally be tested, it is a lot less stringent than the requirement of missingness completely at random, that is, the missing data mechanism does not depend on any data (neither observed nor unobserved).

HAPLOTYPE-BASED IMPUTATION

Treating haplotypes $\mathbf{H}_i = (h_{l(i)}, h_{l'(i)})$ as missing data, the EM algorithm [Excoffier and Slatkin, 1995] aims to maximize the likelihood

$$\prod_{i=1}^{n} \Pr(\mathbf{G}_i | p_1, p_2, \ldots, p_m) = \prod_{i=1}^{n} \sum\nolimits_{\mathbf{H}_i \in \mathcal{G}_i} p_{l(i)} p_{l'(i)}$$

where $l(i)$ refers to a conformable haplotype to the observed $\mathbf{G}_i$, and $\mathcal{G}_i$ is the set of all possible haplotype pairs that conform to the observed $\mathbf{G}_i$. If the SNP at the k locus is missing for individual i, all possible genotypes at the k locus are filled in to construct conformable haplotypes. In the E step, the conditional probability of each pair of conformable haplotypes is calculated based on the current estimates of the haplotype frequencies

$$\Pr((h_1, h_2)|\mathbf{G}_i) = \frac{\hat{p}_1 \hat{p}_2}{\sum_{\mathbf{H}_i \in \mathcal{G}_i} \hat{p}_{l(i)} \hat{p}_{l'(i)}}. \qquad (1)$$

Note that Hardy-Weinberg equilibrium assumes that $\Pr(h_{l(i)}, h_{l'(i)}) = p_{l(i)} p_{l'(i)}$. The frequency estimates are then re-estimated in the M-step. At convergence, we use the conditional probabilities of all conformable haplotype pairs in (1) to impute the missing SNPs.

The WEM approach is an extension of the EM algorithm which incorporates disease status, as haplotype frequencies may be different between cases and controls [Lake et al., 2003], as well as other covariates that may affect the disease risk and haplotype frequencies. Given $\mathbf{H}_i$ and $\mathbf{X}_i$ we model the disease penetrance by a logistic function

$$\begin{aligned}\Pr(\mathbf{D}_i = 1 | \mathbf{H}_i = (h_{l(i)}, h_{l'(i)}), \mathbf{X}_i) \\ = \frac{\exp[\alpha + \mathbf{1}(h_{l(i)}, h_{l'(i)})\gamma + \mathbf{X}_i \beta]}{1 + \exp[\alpha + \mathbf{1}(h_{l(i)}, h_{l'(i)})\gamma + \mathbf{X}_i \beta]}\end{aligned} \qquad (2)$$

where $\mathbf{1}(h_{l(i)}, h_{l'(i)})$ denotes a length m indicator vector. For simplicity, we assume an additive model so that for a heterozygous individual the $l(i)$th and $l'(i)$th elements of this vector equal to 1 and all other elements are zero; for a homozygous individual, $l(i) = l'(i)$, the $l(i)$th element equals 2. Our interest is not to use (2) to model the haplotype-disease association, but rather we use it as a vehicle to impute the missing SNPs.

Set $\boldsymbol{\theta} = (\alpha, \gamma, \beta, \mathbf{p})$. Assuming $\mathbf{H}$ is independent $\mathbf{X}$ given $\mathbf{G}$, the observed data log-likelihood relevant to $\boldsymbol{\theta}$ can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\sum_{\mathbf{H}_i \in \mathcal{G}_i} \Pr{}_{\alpha,\beta,\gamma}(\mathbf{D}_i | \mathbf{H}_i, \mathbf{X}_i) \Pr{}_{\mathrm{p}}(\mathbf{H}_i) \right).$$

The complete-data log-likelihood is $\sum_{i=1}^{n} (\log \Pr_{\alpha,\beta,\gamma}(\mathbf{D}_i | \mathbf{H}_i, \mathbf{X}_i) + \log \Pr_{\mathrm{p}}(\mathbf{H}_i))$. The expectation of the complete-data log-likelihood given the observed data is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n} \sum_{\mathbf{H}_i \in \mathcal{G}_i} W_{i,(s)} [\ell(\alpha, \beta, \gamma; \mathbf{D}_i | \mathbf{H}_i, \mathbf{X}_i) + \ell(\mathbf{p}; \mathbf{H}_i)],$$

where $\boldsymbol{\theta}^{(s)}$ denotes the parameter estimates in the sth iteration of the algorithm, and $W_{i,(s)} = \Pr(\mathbf{H}_i|\mathbf{D}_i, \mathbf{G}_i, \mathbf{X}_\mathrm{i}, \boldsymbol{\theta}^{(s)})$ is the conditional probability of a haplotype pair given the observed data and the current estimates of parameters.

$$W_{i,(s)} = \frac{\Pr(\mathbf{D}_i|\mathbf{H}_i,\mathbf{X}_i,\boldsymbol{\theta}^{(s)})\ \Pr(h_{l(i)}|\boldsymbol{\theta}^{(s)})\Pr(h_{l'(i)}|\boldsymbol{\theta}^{(s)})}{\sum_{\mathbf{H}_i\in\mathcal{G}_i}\Pr(\mathbf{D}_i|\mathbf{H}_i,\mathbf{X}_i,\boldsymbol{\theta}^{(s)})\ \Pr(h_{l(i)}|\boldsymbol{\theta}^{(s)})\Pr(h_{l'(i)}|\boldsymbol{\theta}^{(s)})}. \tag{3}$$

Note that the first part of expected log likelihood is weighted log-likelihood for a generalized linear model, such as logistic regression in (2) for case-control data. The second part is a weighted multinominal log-likelihood. Both can be readily maximized using existing software. The derivation assumes Hardy-Weinberg equilibrium. We impute the missing SNPs by sampling the conformable haplotype pairs according to (3) at convergence. Although we describe the WEM approach for case-control study with logistic regression, in principle it works for other generalized linear models.

The implementation of the EM and WEM algorithms is an adaptation of the existing **R** package *haplo.stats* [Schaid et al., 2002; Lake et al., 2003]. We applied the *haplo.em* function to perform the EM algorithm. Rather than the regular EM algorithm, this function uses an efficient algorithm which progressively inserts a batch of SNP loci, enumerates possible haplotypes, runs EM, and trims off haplotypes with conditional probabilities below a threshold. We set the batch size to be 3, and the minimal conditional probability to 0.001. Starting from the *haplo.em* function, we develop a WEM algorithm similar to the *haplo.glm* function in *haplo.stats*. The minimum haplotype frequency allowed is set to 10^{-6}.

TREE-BASED IMPUTATION

The tree-based approach is a general algorithm to impute missing data, including missing SNPs and missing covariates in SNP association studies. For each individual i, let $\mathbf{M}_i = (M_{i1}, M_{i2}, \ldots, M_{ip})$ be the vector of p variables consisting of the covariates $\mathbf{X}_i = (x_{i1}, \ldots, x_{ir})$ and the unphased SNP data $\mathbf{G}_i = (g_{i1}, \ldots, g_{iK})$ which have missing entries $(1 \le p \le r + K)$. Let $\mathbf{C}_i$ be the vector of the remaining covariates and unphased SNP data for which all data are available. We assume that the outcome $\mathbf{D}_i$ is always observed. The joint probability distribution of the missing data for individual i given the observed data, $\Pr(M_{i1}, M_{i2}, \ldots, M_{ip}|\mathbf{C}_i, \mathbf{D}_i)$, is difficult to get. An obvious problem is that the sets of missing data $\mathbf{M}_i$ and complete data $\mathbf{C}_i$, respectively, are different for each individual i. Instead of modeling the joint distribution, we use the Gibbs sampler, a Markov chain Monte Carlo technique that uses conditional (low-dimensional) distributions to draw samples from a high-dimensional distribution.

Specifically, we consider iteratively sampling from the following sequence of the full conditional distributions in the $(s+1)$th iteration:

$$M_1^{(s+1)} \sim \Pr(M_1|M_2^{(s)}, M_3^{(s)}, \ldots, M_p^{(s)}, \mathbf{C}, \mathbf{D})$$

$$M_2^{(s+1)} \sim \Pr(M_2|M_1^{(s+1)}, M_3^{(s)}, \ldots, M_p^{(s)}, \mathbf{C}, \mathbf{D})$$

$$\vdots$$

$$M_p^{(s+1)} \sim \Pr(M_p|M_1^{(s+1)}, M_2^{(s+1)}, \ldots, M_{p-1}^{(s+1)}, \mathbf{C}, \mathbf{D})$$

where each full conditional distribution, for example $\Pr(M_1|M_2^{(s)}, M_3^{(s)}, \ldots, M_p^{(s)}, \mathbf{C}, \mathbf{D})$, is modeled by CART. This is easily done even though $M_2, \ldots, M_p$ contain missing observations before imputation has taken place, as CART uses surrogate splits if missing observations are encountered in a node [Breiman et al., 1984]. For example, if M_1 are actual data from an SNP, each terminal leaf in the classification trees provides a multinomial distribution from which we can sample. A convenient property of surrogate splits is that we do not have to guess the initial values of the missing data in M; as a result, only a very short burn-in of the above sampler is required. Under mild regularity conditions, this sequence of conditional variables converge to the joint distribution of missing data.

A similar idea, data augmentation [Tanner and Wong, 1987], has been exploited to deal with missing data in a Bayesian framework. However, data augmentation is only analytically tractable in some simple situations, such as a multivariate normal distribution. The advantage of applying decision trees such as CART [Breiman et al., 1984] is that it can handle variables of any type, such as 3-level factor (0,1,2) coded SNP genotypes in a locus, or a continuous variable such as age. Though lacking a formal proof, it has been demonstrated in simulation studies that the inference in missing data problems is fairly nonsensitive to model mis-specification as long as the distribution of the missing data given the observed data involves the covariates that are ultimately found to be important in the model [Schafer, 1997]. It is therefore natural to investigate the performance of nonparametric regression methods such as decision trees for imputation. This has been suggested in the literature before [Harrell, 2001], though not for SNP association studies.

Our tree algorithm is based on the *rpart* package in **R** [Therneau and Atkinson, 1997]. The nodes in the decision trees generated by this package are split until the improvement of impurity measure (by default, the GINI) for the best possible split is less than 1% of the impurity in the root node. Also, splits are usually only attempted on nodes with at least 5% of the number of total observations. This allows for somewhat larger trees in case-control studies with relatively few observations. Using those parameters, we grow the trees to full size without model selection and pruning. In our simulations, this provided some additional computational benefit as it was not necessary to carry out cross-validation, without compromising the quality of the imputations. By default, we iterate 10 times through the set of missing variables (''sweeps'' through the data) before imputing the missing values. However in data sets with severe missingness, more sweeps might be beneficial.

MULTIPLE IMPUTATION

The uncertainty of imputations is addressed by multiple imputation [Little and Rubin, 1987; Schafer, 1997]. Multiple imputation is a Monte Carlo technique which draws multiple samples from the probability distribution of predicted missing values. In essence, multiple imputation acknowledges the uncertainty due to missing data, instead of simply ignoring it: several complete data sets are generated, and the uncertainty in the model parameter estimates incorporates the standard errors of the parameter estimates as well as the variability between the parameter estimates from the replicate data sets. We draw 10 samples from the resulting joint distribution of missing data at convergence, whether it is from EM, WEM, or tree algorithm. Each imputed sample is analyzed by standard methods, and the results are combined across 10 samples to get parameter estimates and their standard errors. The details of multiple imputation have been documented in Little and Rubin [1987] and Schafer [1997].

SIMULATIONS

Our simulation studies involved drawing case-control samples from a population, randomly masking a proportion of SNPs as missing, and imputing them by the methods under investigation. We adopted an eight-haplotype distribution based on four SNPs in the progesterone receptor (PGR) gene [Kraft et al., 2005]. Previously, De Vivo et al. [2002] found that a G/A polymorphism in the PGR gene may be associated with an increased risk of endometrial cancer. Kraft et al. [2005] genotyped four haplotype tagging SNPs in case-control data in order to compare several methods currently used in haplotype-disease association studies. Table I shows the distribution of eight haplotypes estimated in Kraft et al. [2005]. Based on these frequencies and assuming Hardy-Weinberg equilibrium, we created a population of 100,000 individuals with diploid genotypes.

We added a disease-association signal to haplotype **1000** through a logistic penetrance function

$$\operatorname{logit}(\Pr(\mathbf{D}=1|\mathbf{H})) = -3 + \beta \cdot (\text{number of copies of } h_{\mathbf{1000}}) \qquad (4)$$

with $\beta = 0$, 1, or 2. **D** is the dichotomous disease status, and **H** refers to the haplotype pair for an individual. We randomly sampled 100 cases and 300 controls from the population. Either 10% or 20% of the SNPs were made missing completely at random. These missing SNPs were imputed 10 times using the EM, WEM, and tree approach. To construct a baseline for the imputation comparison, we used the observed marginal SNP genotype distribution to impute the missing ones. We call this method the ''naive'' approach, as it uses no information of other SNPs or the response. We calculated the imputation error probability for each SNP using a 0/1 loss function. That is, we coded each genotype as 0 (homozygous wide

TABLE I. PGR haplotype frequencies [Kraft et al., 2005] used in the simulation study

Haplotype	Frequency
0000	0.3265
0001	0.1327
0100	0.0306
0101	0.0408
1000	**0.1633**
1010	0.0408
1100	0.0204
1110	0.2449

type), 1 (heterozygotes), and 2 (homozygous mutant) and any difference in imputed genotype was counted as an error. We explored a variety of other error functions, reaching similar conclusions about the various approaches.

While we predisposed disease risk on the haplotype level, we analyzed the imputed data using SNP-based logistic regression models. We first considered marginal SNP association with disease by fitting a logistic regression model of the form

$$\text{logit}(\Pr(\mathbf{D}|\text{SNPs})) = \alpha_0 + \alpha_1 x \quad (5)$$

where x denotes the number of variant alleles (0,1,2) for a particular SNP. For simplicity we treated x as a continuous variable so that having two copies doubles the effect of having one copy. We also investigated the effect of imputation on simple interactions using the model

$$\text{logit}(\Pr(\mathbf{D}|\text{SNPs})) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2. \quad (6)$$

Similarly, x_1 and x_2 are the coding variables for SNP1 and SNP2, respectively. γ_3 is the interaction parameter of interest. We compared the parameter estimates using various approaches of imputing SNP data with the true values, that can be computed by fitting (5) and (6) to the whole population.

DATA APPLICATION

We used a recently published case-control data set on DD to compare various imputation approaches. Cope et al. [2005] performed a high-density LD screen in a 575-kb region of chromosome 6p22.2 with both case-control and family data. After removing redundant SNPs with pairwise correlation $r^2 \leq 0.8$, Cope et al. [2005] genotyped 10 SNPs in 223 cases and 273 controls. These 10 SNPs are in weak LD with average pairwise $r^2 = 0.16$. Seven SNPs showing significant results in case-control data were genotyped in 143 parent-proband trios. Table II shows the number of missing values for these 10 SNPs, separated by cases and controls. All but 25 probands in the trios were included in the case-control sample. Those 25 probands were added later to cases and thus they do not have genotypes for SNPs *rs6911855*, *rs6939068*, and *rs2143340*. Cope et al. [2005] ignored missing data and analyzed the data in a SNP by SNP fashion.

We re-analyzed the marginal SNP association in DD data via the multiple imputation approaches under investigation. The SNP-disease association was modeled as in (5). To compare the imputation errors, we randomly generated extra missing values and computed the probability of false imputation for the additional missing data. Parameter estimates for models (5) and (6) were computed with the extra missing data imputed, but the original missing data from Table II were left unimputed. The "true values" of parameter estimates are therefore computed from the original data with missing values. We compared the bias and sampling variance as in the simulation study.

TABLE II. Percentage of missing SNPs in the case-control study of Cope et al. [2005]

	Case ($n = 248$)		Control ($n = 273$)	
SNP	♯	%	♯	%
rs2793422	27	10.8	22	8.1
rs4504469	8	3.2	9	3.3
rs6911855	30	12.1	8	2.9
rs6939068	48	19.4	25	9.2
rs2179515	16	6.5	16	5.9
rs6935076	17	6.9	18	6.6
rs2038137	19	7.8	16	5.9
rs2143340	45	18.1	23	8.4
rs3777664	24	9.6	17	6.2
rs1053598	23	9.2	15	5.5

Note: rs6911855, *rs6939068*, and *rs2143340* have extra missing values in cases since 25 probands from the later family study are included. By design these 25 cases do not have the genotypes for these three.

RESULTS

For the simple LD block with four SNPs as shown in Table I, both the EM and WEM approaches yield better predictions of the missing SNPs than the tree approach (Table III), while all three approaches show a marked improvement over the naive approach. When there is no association between the SNPs and the outcome ($\beta = 0$), the EM and WEM approaches perform equally well and the tree approach makes roughly 2–3% more errors on average. However, when there is a disease risk associated with haplotype **1000** the WEM approach yields more accurate imputations than the EM approach. As β increases to 2, the advantage of WEM for SNP1 imputation becomes substantial: WEM produces almost 5–6% less errors than EM. This was expected since the case-control status influences the estimation of

TABLE III. Mean imputation errors in the simulated data of four SNPs on the PGR gene

		10% missing data				20% missing data			
	Approach	SNP1	SNP2	SNP3	SNP4	SNP1	SNP2	SNP3	SNP4
$\beta = 0$									
	Naive[a]	0.625	0.596	0.568	0.449	0.625	0.595	0.567	0.449
	EM	0.412	0.390	0.243	0.379	0.427	0.407	0.271	0.385
	WEM	0.412	0.390	0.243	0.379	0.427	0.406	0.271	0.385
	Tree	0.440	0.397	0.260	0.399	0.461	0.411	0.292	0.415
$\beta = 1$									
	Naive	0.627	0.589	0.560	0.441	0.627	0.589	0.560	0.441
	EM	0.433	0.383	0.245	0.369	0.448	0.399	0.273	0.375
	WEM	0.415	0.381	0.241	0.369	0.431	0.396	0.269	0.375
	Tree	0.449	0.389	0.263	0.389	0.471	0.407	0.296	0.403
$\beta = 2$									
	Naive	0.628	0.587	0.557	0.438	0.627	0.588	0.557	0.438
	EM	0.443	0.380	0.246	0.365	0.457	0.397	0.273	0.371
	WEM	0.386	0.375	0.233	0.363	0.402	0.391	0.257	0.370
	Tree	0.422	0.388	0.262	0.385	0.443	0.398	0.292	0.399

Each number is the average of imputation error probabilities from 1,000 simulations.

[a] This method imputes the missing by the marginal distribution of available SNP genotypes and ignores other SNPs.

haplotype frequencies. When the association is absent ($\beta = 0$) or small ($\beta = 1$), the tree approach performs comparably to the EM and WEM approach. When the association is strong ($\beta = 2$), the tree approach even outperforms EM for SNP1, presumably because the strength of the association now overcomes the incorrect model. Given that the tree algorithm treats individual SNPs as 0/1/2 categorical variables, and only indirectly models the correlation between SNPs, such performance is impressive. A graphical representation of Table III as well as the rest of the tables can be found as supplementary material at `http://biostat.jhsph.edu/∼iruczins/supplements/05.comparison`.

In Table IV we compare the effects of the different imputation approaches on estimating association parameters when 10% of the SNPs are missing. The left four columns depict the estimates of log-odds ratio for SNP3, as modeled by α_1 in (5); the right four columns show the estimate of the interaction effect between SNP2 and SNP3 as modeled by γ_3 in (6). The lines "True data" refer to the case-control data before SNPs were made missing, as these are the best imputations one can ever obtain. Compared to the complete-case analysis, all three imputation methods decrease sampling variances of parameter estimates and hence mean square error (MSE). The reduction of MSE ranges from 5–15% for the marginal effect, to 15–30% for the interaction effect. For SNP interactions, the complete-case analysis hurts more because 10% missing values in each SNP may result in up to 20% missing in either one of two SNPs. Among the three imputation methods, WEM has the smallest bias particularly when there is strong association between haplotype 1000 and disease, yet it incurs the largest sampling variances. This is an example of bias-variance tradeoff as the WEM approach jointly models the haplotype frequencies and haplotype risks, and thus has more parameters (eight more, to be exact) than the EM approach. Although the EM approach yields more bias as β increases, it yields a smaller MSE than the WEM approach. The tree approach reduces some bias, as it takes disease status into account, and yields a smaller sampling variance than the WEM approach. This is probably because the tree algorithm does model selection inherently when building a tree, thus incurs less parameters than the WEM approach. For this particular set of parameters, the tree approach yields the smallest MSE. For other SNPs, we found a similar pattern that the WEM approach is most effective in eliminating bias with a cost of large variance. The magnitude of improvement and the comparison between the three methods vary with SNPs. Interestingly, the EM and the tree approach produce a smaller sampling variance and RMSE than the true data. Further inspection of simulation results suggests that the imputation of missing data always shrinks the parameter estimates slightly toward null. This shrinkage effect may lower the sampling variance, and thus generate a smaller RMSE than using the true data.

TABLE IV. The effect of different imputation approaches on association parameters in the simulation study with 10% missing data

	Marginal[a]				Interaction[b]			
Approach	Bias	SD($\hat{\alpha}_1$)	RMSE[c]	%[d]	Bias	SD($\hat{\gamma}_1$)	RMSE[c]	%[d]
$\beta = 0$								
True data	−0.007	0.181	0.181	—	−0.017	0.265	0.265	—
Complete-case	−0.008	0.191	0.192	—	−0.034	0.305	0.307	—
EM	−0.006	0.179	0.179	13.1	−0.021	0.258	0.259	28.8
WEM	−0.007	0.185	0.185	7.2	−0.021	0.274	0.274	20.3
Tree	−0.006	0.179	0.179	13.1	−0.020	0.257	0.257	29.9
$\beta = 1$								
True data	−0.014	0.193	0.193	—	−0.017	0.300	0.301	—
Complete-case	−0.012	0.201	0.202	—	−0.033	0.346	0.347	—
EM	0.016	0.187	0.187	14.3	−0.051	0.289	0.294	28.2
WEM	−0.009	0.196	0.196	5.8	−0.024	0.309	0.310	20.1
Tree	0.006	0.188	0.188	13.4	−0.038	0.290	0.293	28.7
$\beta = 2$								
True data	−0.008	0.215	0.215	—	−0.011	0.315	0.315	—
Complete-case	−0.007	0.231	0.231	—	−0.016	0.350	0.350	—
EM	0.047	0.209	0.214	14.1	−0.066	0.301	0.308	22.5
WEM	−0.007	0.222	0.222	7.6	−0.023	0.326	0.327	12.7
Tree	0.026	0.212	0.213	14.9	−0.039	0.301	0.304	24.6

[a] Marginal effect: logistic model logit(Pr($\mathbf{D}$ | SNPs)) = $\alpha_0+\alpha_1\, x$ was fitted; x is the continuous coding variable for SNP3. The true value of the parameter is obtained by fitting the model to the population data before data were made missing; Bias, SD, and RMSE are computed from 1,000 iterations.

[b] Interaction: logistic model logit(Pr($\mathbf{D}$ | SNPs)) = $\gamma_0+\gamma_1\, x_1+\gamma_2\, x_2+\gamma_3\, x_1\, x_2$ was fitted. x_1 and x_2 are the 0/1/2 continuous coding variables for SNP2 and SNP3.

[c] square root of mean square error (MSE).

[d] % reduction of MSE compared the complete-case analysis.

Cope et al. [2005] estimated the odds ratios for the allele effect of 10 SNPs ignoring missing data [see Table III in Cope et al., 2005]. They concluded that 7 out of 10 SNPs are associated with DD at a significance level of 0.05. Haplotype analysis using all 10 SNPs yields approximately 50 haplotypes, none of which is significantly associated with disease (results not shown). The large number of haplotypes is produced by weak correlation due to removal of redundant SNPs and irregular LD structure [see Table II in Cope et al., 2005]. In this situation, a SNP-based regression approach seems to be a better strategy. We carried out a multiple imputation analysis to see whether imputation of missing data influences their conclusion. We first verified that the SNP effect is indeed additive, and applied the univariate logistic regression (5) to each SNP. Due to space limit, we only show the results for four SNPs (*rs4504469*, *rs6911855*, *rs6939068*, and *rs6935076*). SNPs *rs6911855* and *rs6939068* are in tight linkage ($r^2 = 0.79$), whereas *rs4504469* and *rs6935076* are in a weak LD with the other SNPs. Table V compares the log-odds ratio estimates and *P*-values after multiple imputation. For SNPs *rs6911855* and *rs6939068*, standard errors become smaller and point estimates are enlarged by both the EM and WEM imputations. Hence *P*-values are smaller than in the complete-case analysis. This is probably driven by the 25 proband cases, who have both SNPs missing. The WEM approach seems to capture the missing pattern depending on the disease status, and therefore yields more significant results than the other approaches (P-value of *rs6939068* < 0.05). For SNPs *rs4504469*, *rs6911855*, the imputation has little effect on the standard errors. The effect of the sample size increment after imputation seems to be canceled by the extra variability raised by multiple imputation. This is perhaps because the rate of missing values is small for these loci and LD is weak.

To compare the accuracy of the three imputation approaches, we randomly removed an extra 5%, 10%, and 15% of the SNPs from the data set of

TABLE V. A comparison of the log-odds ratio estimates by different imputation methods for the developmental dyslexia data

Approach	Complete-case			EM			WEM			Tree		
SNP	log-OR	SE	P-val	log-OR	SE	P-val	log-OR	SE	P-val	log-OR	SE	P-val
rs4504469	0.417	0.018	0.002	−0.399	0.017	0.002	−0.416	0.018	0.002	−0.417	0.018	0.002
rs6911855	0.658	0.138	0.076	0.659	0.136	0.074	0.720	0.140	0.054	0.621	0.142	0.100
rs6939068	0.637	0.123	0.070	0.642	0.108	0.051	0.688	0.112	0.040	0.572	0.120	0.098
rs6935076	0.396	0.019	0.005	0.376	0.019	0.006	0.404	0.019	0.004	0.340	0.019	0.014

Note: The estimates are based on 10 imputations.

Cope et al. [2005]. Table VI shows the comparison of imputation error probabilities for the additional missing SNPs stratified by SNP, imputation approach, and missing percentage. Similar to Table III, all three approaches work much better than the naive approach. The WEM approach performs consistently better than the EM and tree approaches, although the improvement of WEM over EM is for most scenarios less than 1%. This, again, may be explained by the weak LD among the 10 SNPs, as haplotype ambiguity is so dominant that knowing case-control status does not gain much in imputation. On the other hand, the accuracy of the tree approach is only 1–2% lower than two haplotype-based approaches, suggesting that in practice the tree approach may be sufficiently accurate to characterize the inter-SNP correlation in a modest LD block.

Table VII shows the biases, sample standard deviations and MSE for marginal effects and the interaction of two SNPs found to be most significantly associated with DD in Cope et al. [2005]. These statistics are conditional on the original data from Cope et al. [2005]. That is, the "true values" of the parameters are obtained from the original data with the original missing values and no imputation. Likewise, the "Complete-case" here refers to data sets with both the original missing SNPs and extra missing data removed, again serving as the baseline for comparison. The first eight columns compare the estimates of SNP marginal effects. Cope et al. [2005] found a significant interaction between *rs4504469* and *rs6935076*. In the last four columns in Table VII, we compared the effects of imputation on the interaction parameter in the logistic regression model (6). It appears that all three imputation approaches significantly improve the SD and RMSE over the complete-case analysis. The reduction of MSE is 20–30% in estimating the marginal effect of SNP *rs4504469*, 40–50% in estimating the marginal effect of SNP *rs6935076*, and 50–60% in estimating the interaction. The WEM approach is effectively unbiased with the cost of increased sample variance, which is similar to the results in Table IV. The variance of WEM estimators for interaction is apparently much larger than other two approaches, so that their MSEs are the worst. This is probably caused by the extra parameters (approximately 50) in modeling haplotype-disease association. Interestingly, the tree approach yields a smaller variance than EM and achieves the best performance in MSE in the majority of scenarios. This seems contradictory

TABLE VI. The comparison of imputation error probabilities for the developmental dyslexia data

Approach	*rs4504469*	*rs6911855*	*rs6939068*	*rs6935076*	Average of 10 SNPs
5% missing					
Naive	0.609	0.117	0.133	0.596	0.486
EM	0.367	0.014	0.028	0.300	0.199
WEM	0.364	0.012	0.028	0.296	0.197
Tree	0.379	0.032	0.034	0.324	0.223
10% missing					
Naive	0.609	0.114	0.137	0.597	0.486
EM	0.372	0.020	0.035	0.309	0.206
WEM	0.367	0.019	0.033	0.307	0.205
Tree	0.390	0.039	0.044	0.339	0.235
15% missing					
Naive	0.610	0.114	0.136	0.595	0.486
EM	0.375	0.024	0.040	0.319	0.214
WEM	0.368	0.024	0.038	0.314	0.211
Tree	0.396	0.041	0.049	0.353	0.248

Note: The numbers are the averages of imputation error probabilities from 200 simulations.

to the comparison in Table VI, where the tree approach makes more imputation errors than the other two approaches. Since we count 1 error whenever the imputed SNP genotype is different from the true genotype, the impact of different genotype errors on association parameter estimates may be different. For example, imputing a missing SNP with true genotype "2" to be "1" has less effect than imputing it to be "0". It is possible that the impact of imputation errors on estimating association parameters is smaller in the tree approach than that in the EM approach, since the former use case-control status in the imputation. Taken collectively, all three imputation methods improve MSE considerably in comparison to complete-case analysis. The tree approach seems to have an advantage in dealing with 10 SNPs in weak correlation. It offers bias reduction without incurring too much variance.

DISCUSSION

Despite the fact that missing SNPs are quite common in genetic association studies, the impact of imputation on SNP association inference has not been adequately studied. In this article, we developed and assessed the impact of haplotype-based and the tree-based imputation approaches with case-control data. Our results suggest that in general there is substantial benefit from imputation over the commonly used complete-case analysis when association studies are analyzed using SNPs as predictors. As we expected, the benefit of imputation is greater in estimating interaction parameters than in estimating marginal parameters (Tables IV and VII).

Imputing missing SNPs usually helps association inference in increasing the efficiency without adding noticeable bias, at the price of some extra variability from the uncertainty in the imputation. With LD structure existing between SNPs, the added sample size usually outweighs the imputation uncertainty and the standard error decreases. This is seen in our simulation study (Tables IV) and the data application (Table VII). The advantage of imputation can be substantial when a regression model with multiple SNPs is involved (Tables IV and VII). We acknowledge that in some cases where the missingness rate is low and LD is weak, the imputation may not help to gain efficiency for marginal parameters (Table V, *rs4504469* and *rs6935076*). On the other hand, imputation of the missing SNPs could also help to correct bias. In our data application, the fractions of missing values for SNPs *rs6911855* and *rs6939068* differ between cases and controls. The parameter estimates and association inferences for these two SNPs were changed greatly by multiple imputation using the WEM approach (Table V), indicating complete-case analysis may cause bias in this scenario. Our overall assessment is that performing multiple imputation up-front yields better inferences than the complete-case analysis in SNP association studies, particularly when regression models with multiple SNPs are involved.

Haplotype analysis is becoming increasingly popular in genetic association studies, whereas tree-based approaches start to draw attention in

TABLE VII. Comparisons of the marginal log-odds ratio estimates for SNPs *rs4504469*, *rs6935076*, and their interaction

	rs4504469[a]				*rs6935076*[a]				Interaction[b]			
Approach	Bias	SD($\hat{\alpha}_1$)	RMSE	%[c]	Bias	SD($\hat{\alpha}_1$)	RMSE	%[c]	Bias	SD($\hat{\gamma}_3$)	RMSE	%[c]
5% missing												
Complete-case	0.0024	0.0291	0.0291	—	−0.0021	0.0306	0.0306	—	−0.0010	0.0701	0.0699	—
EM	0.0120	0.0222	0.0251	25.6	−0.0108	0.0201	0.0228	44.5	0.0076	0.047	0.0475	53.8
WEM	0.0034	0.0228	0.0230	37.5	−0.0005	0.0231	0.0231	43.0	0.0048	0.0513	0.0514	45.9
Tree	0.0094	0.0217	0.0236	34.2	−0.0069	0.0184	0.0196	58.9	0.0029	0.0473	0.0473	54.2
10% missing												
Complete-case	0.0034	0.0438	0.0438	—	−0.0035	0.0479	0.0479	—	−0.0073	0.1047	0.1047	—
EM	0.0201	0.0320	0.0377	25.9	−0.0219	0.0297	0.0368	40.9	0.0119	0.0624	0.0634	63.3
WEM	0.0028	0.0379	0.0379	25.1	0.0006	0.0369	0.0369	40.7	0.0075	0.0783	0.0785	43.8
Tree	0.0168	0.0304	0.0347	37.2	−0.0139	0.0301	0.0330	52.5	0.0054	0.0681	0.0682	57.6
15% missing												
Complete-case	0.0011	0.0555	0.0555	—	−0.0001	0.0621	0.0621	—	−0.0069	0.1298	0.1298	—
EM	0.0268	0.0374	0.0460	31.3	−0.0314	0.0357	0.0475	41.5	0.0184	0.0733	0.0754	66.3
WEM	0.0021	0.0486	0.0486	23.3	−0.0012	0.0466	0.0466	43.7	0.0106	0.0963	0.0966	44.6
Tree	0.0235	0.0393	0.0458	31.9	−0.0215	0.0394	0.0448	47.9	0.0083	0.0774	0.0776	64.3

[a] Marginal effect: logistic model logit(Pr(**D** | SNPs)) = $\alpha_0+\alpha_1 x$ was fitted; x is the continuous coding variables for the targeted SNP. The true value of the parameter is obtained by fitting the model to the original developmental dyslexia data; Bias, SD, and RMSE are computed from 200 simulations.

[b] Interaction: logistic model logit(Pr(**D** | SNPs)) = $\gamma_0+\gamma_1 x_1+\gamma_2 x\gamma_2+\gamma_3 x_1 x_2$ was fitted. x_1 and x_2 are the 0/1/2 continuous coding variables for *rs4504469* and *rs6935076*.

[c] % reduction of MSE compared to complete-case analysis.

studies with a large number of SNPs. Haplotypes can be used directly to model the association, or indirectly to impute SNPs of SNPs are used as predictors. To our knowledge, this is the first paper directly studying haplotype and tree-based imputation approaches. The imputation accuracy can be viewed as an indicator of how well the inter-SNP correlation structure is captured by non-parametric tree regression. Evidently, the tree approach produces slightly more imputation errors than the haplotype approaches. However, its advantages are apparent: it is computationally efficient, it easily accommodates disease status, extra covariates, and a large number of SNPs. In many cases (Tables IV and VII), the tree approach even outperforms the EM approach in both bias reduction and MSE. In many genetic epidemiological studies subjects complete a questionnaire, which may contain dozens of relevant environmental and demographic variables. The tree algorithm can handle an arbitrary number of these variables, as the splits in the decision trees are completely data driven. Computing time so far has never been an issue in our analyses (typical data we see have up to a few thousand observations and a few hundred variables). Considering the increasing number of genome-wide SNP association studies carried out today, we believe that the tree approach provides a competitive alternative for the imputation of missing SNP values.

For the benefit of imputation, the information of disease status is secondary to the correlation between adjacent SNPs in the data we studied. That is perhaps why the improvement of the WEM approach over the EM approach is much smaller compared to that of the EM approach over naive imputation (Tables III and VI). In virtually every situation we examined, the WEM approach produces minimal bias. On the other hand, the WEM approach incurs more variability than the EM approach since it involves more parameters. That said, we recognize that sometimes one may prefer the unbiased estimates even if they have a larger variance.

There are other algorithms available to reconstruct haplotypes that can be used to impute the missing SNPs. For example, PHASE employs an MCMC approach to infer the haplotypes from unphased genotypes, using priors based on coalescent theory and taking account of the decay of LD [Stephens et al., 2001; Stephens and Scheet, 2003]. However, PHASE, as well as other Bayesian approaches, is designed for inferring haplotype in a population and thus does not incorporate the case-control status to estimate the haplotype frequencies. This may introduce bias to association parameter estimates after imputing the missing SNPs, similarly to the EM algorithm. We tried PHASE to impute the missing SNPs in the DD data. The imputation accuracy was about the same as the EM algorithm, yet the computing time was significantly longer.

The methods presented here depend on the assumption that SNP genotypes are missing at random (MAR). However for some technologies, it is not uncommon to have non-MAR data in SNP genotypes. For example, the Dynamic model-based (DM) algorithm [Di et al., 2005], the original genotype calling algorithm for the Affymetrix Gene Chip Human Mapping Arrays, has a larger fraction of no calls among the heterozygous SNPs than the homozygous SNPs (see for example, Table I of the BRLMM technical manual, Affymetrix Corporation, 2006). The reason for this discrepancy is that the DM algorithm assigns the genotype call AA (homozygous), BB (homozygous), AB (heterozygous), or no call if the pattern is ambiguous. As the acceptance region for the heterozygous AB call is "wedged" between the acceptance regions for the homozygous AA and BB calls, it is much harder for the heterozygote to pass, resulting in a higher missingness rate [Rafael Irizarry and Robert Welch, personal communication]. Affymetrix recently proposed the BRLMM algorithm [the BRLMM technical manual, Affymetrix Corporation, 2006], an extension of RLMM algorithm [Rabbee and Speed, 2006]. This new algorithm improves the overall SNP call rates and (supposably) eliminates the difference between the call rates for homozygous and heterozygous SNPs. That said, non-MAR creates bias in parameter estimates, regardless of whether the missing data are handled by complete-case analysis or by multiple imputation. To account for non-MAR, it is straightforward to derive a Bayesian-type imputation model with a prior that acknowledges different call rates for different genotypes. For the imputation methods considered here, we can compute the posterior probability of missing SNPs by weighting the imputation probabilities (using the prior) outputted from various methods.

ACKNOWLEDGMENTS

We want to thank Dr. Peter Holmans and the other members of the developmental dyslexia (DD) study for allowing us to use the case-control data. We thank Rafael Irizarry, Robert Welch, and Sonja Berndt for helpful discussions. We also thank two reviewers for their constructive comments. This research was supported in part by NIH grants CA 74841 (J.Y.D., I.R., M.L.B., C.K.), CA 105069 (I.R.), CA 53996 (M.L.B., C.K.), CA 90998 (M.L.B.), and HL 74745 (C.K.).

REFERENCES

Affymetrix Corporation. 2006. BRLMM: an improved genotype calling method for the GeneChip$^{\circledR}$ human mapping 500 K array set. `http://www.affymetrix.com/support/technical/product_up dates/brlmm_algorithm.affx`

Barnby G, Abbott A, Sykes N, Morris A, Weeks DE, Mott R, Lamb J, Bailey AJ, Monaco AP, and the International Molecular Genetics Study of Autism Consortium (IMGSAC). 2005. Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT. Am J Hum Genet 76: 950–966.

Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past success for Mendelian disease, future approaches for complex disease. Nat Genet Suppl 33:228–237.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. Classification and Regression Trees. Belmont, CA: Wadsworth International Group.

Brewster AM, Jorgensen TJ, Ruczinski I, Huang HY, Hoffman S, Thuita L, Newschaffer C, et al. 2006. Polymorphisms of the DNA repair genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): relationship to breast cancer risk and familial predisposition to breast cancer. Breast Cancer Res Treat 95:73–80.

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van E. 2005. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28:171–182.

Cope N, Harold D, Hill G, Moskvina V, Stevenson J, Holmans P, Owen MJ, et al. 2005. Strong evidence that KIAA0319 on chromosome 6p is a susceptibility gene for developmental dyslexia. Am J Hum Genet 76:581–591.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124–141.

De Vivo I, Huggins GS, Hankinson SE, Lescault PJ, Boezen M, Colditz GA, Hunter DJ. 2002. A functional polymorphism in the promoter of the progesterone receptor gene associated with endometrial cancer risk. Proc Natl Acad Sci USA 99:12263–12268.

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, et al. 2005. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. Bioinformatics 21:1958–1963.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Greenland S, Finkle W. 1995. A critical look at methods for handling missing covariates in epidemiologic regression analysis. Am J Epidemiol 142:1255–1264.

Harrell FE. 2001. Regression Modelling Strategies. New York: Springer.

Hu N, Wang C, Hu Y, Yang HH, Giffen C, Tang ZZ, Han XY, et al. 2005. Genome-wide association study in esophageal cancer using GeneChip Mapping 10 K Assay. Cancer Res 65: 2542–2546.

Kraft P, Cox DG, Paynter R, Hunter F, De Vivo I. 2005. Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet Epidemiol 28:261–272.

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. 2003. Estimation and tests of haplotype-environmental interaction when linkage phase is ambiguous. Hum Hered 55:56–65.

Lin S, Cutler DJ, Zwick ME, Chakravarti A. 2002. Haplotype inference in random population samples. Am J Hum Genet 71:1129–1137.

Little RJA, Rubin DB. 1987. Statistical Analysis With Missing Data. New York: John Wiley & Sons.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. Am J Hum Genet 70:157–169.

Qin ZS, Niu T, Liu JS. 2002. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 71:1242–1247.

Rabbee N, Speed TP. 2006. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12.

Risch N. 1990. Linkage strategies for genetically complex traits. I. Multi-locus models. Am J Hum Genet 46:222–228.

Risch N. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Ruczinski I, Kooperberg C, LeBlanc M. 2003. Logic regression. J Comput Graph Stat 12:475–511.

Schafer JL. 1997. Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425–434.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Stephens M, Scheet P. 2003. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462.

Stram DO, Leigh PC, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, et al. 2003. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190.

Tanner MA, Wong WH. 1987. The calculation of posterior distributions by data augmentation. JASA 82:528–540.

Therneau TM, Atkinson EJ. 1997. An introduction to recursive partitioning using the RPART routines. Technical Report Series no. 61, Department of Health Science Research, Mayo Clinic, Rochester, Minnesota.

Zhang H, Bonney G. 2000. Use of classification trees for association studies. Genet Epidemiol 19:323–332.

Zhao LP, Li SS, 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 72:1231–1250.