Hum Genet
DOI 10.1007/s00439-007-0427-y

ORIGINAL INVESTIGATION

Methods to impute missing genotypes for population data

Zhaoxia Yu · Daniel J. Schaid

Received: 19 June 2007 / Accepted: 30 August 2007

Abstract For large-scale genotyping studies, it is common for most subjects to have some missing genetic markers, even if the missing rate per marker is low. This compromises association analyses, with varying numbers of subjects contributing to analyses when performing single-marker or multi-marker analyses. In this paper, we consider eight methods to infer missing genotypes, including two haplotype reconstruction methods (local expectation maximization-EM, and fastPHASE), two *k*-nearest neighbor methods (original *k*-nearest neighbor, KNN, and a weighted *k*-nearest neighbor, wtKNN), three linear regression methods (backward variable selection, LM.back, least angle regression, LM.lars, and singular value decomposition, LM.svd), and a regression tree, Rtree. We evaluate the accuracy of them using single nucleotide polymorphism (SNP) data from the HapMap project, under a variety of conditions and parameters. We find that fastPHASE has the lowest error rates across different analysis panels and marker densities. LM.lars gives slightly less accurate estimate of missing genotypes than fastPHASE, but has better performance than the other methods.

Z. Yu
Department of Statistics, University of California,
Irvine, CA 92697, USA
e-mail: yu.zhaoxia@ics.uci.edu

D. J. Schaid (✉)
Harwick 775, Division of Biostatistics,
Department of Health Sciences Research,
Mayo Clinic College of Medicine, 200 First Street, SW,
Rochester, MN 55905, USA
e-mail: schaid@mayo.edu

Background

As most common human diseases show complicated etiology of genetic effects, genome wide association studies are becoming widely used. However, missing genetic markers are common, causing single-maker analyses, or multi-marker analyses, to be applied to different subsets of subjects without missing data. Furthermore, excluding subjects with missing genotypes can remove a large portion of subjects and thereby decrease power. Replacing missing genotypes with observed means or the most probable genotypes does not use linkage disequilibrium (LD) information from nearby markers, decreasing statistical efficiency and possibly causing bias. It is therefore important to develop efficient statistical methods to accurately infer missing genotypes.

Estimation of missing genotypes can be a by-product of haplotype reconstruction, with either a maximum likelihood method implemented by the expectation maximization (Dempster et al. 1977) algorithm (Chiano and Clayton 1998; Excoffier and Slakin 1995; Fallin and Schork 2000; Hawley and Kidd 1995; Long et al. 1995; Qin et al. 2002; Scheet and Stephens 2006) or Bayesian methods (Lin et al. 2004; Niu et al. 2002; Stephens and Donnelly 2003; Stephens and Scheet 2005; Stephens et al. 2001). When the number of loci under consideration is large, both approaches are computationally intensive. While the maximum likelihood can lead to computer memory limitations, the Bayesian methods can take a longer time to converge. In both approaches, missing genotypes and missing phase are treated equivalently and inferred simultaneously. The accuracy of imputation of missing genotypes using different haplotype reconstruction methods have been compared in several papers, such as Stephens and Scheet (2005), Scheet and Stephens (2006), and Marchini et al. (2006).

One could also simultaneously estimate missing genotypes and the association parameters in a parametric model that tests genetic association. Lake et al. (2003) allowed for missing genotypes in a generalized linear model for haplotype associations with a trait. Hoti and Sillanpaa (2006) treated missing genotypes as additional parameters in their hierarchical model and estimated all parameters in a Bayesian framework. Souverein et al. (2006) modeled missing genotypes as a function of other markers and phenotypes in a polytomous logistic regression model. Although they varied the set of covariates in the polytomous logistic regression, the optimal set of covariates to be used in general is not clear. In contrast to others, Dai et al. (2006) proposed a classification tree method. For each marker with missing genotypes, they first fit a classification tree, using both disease status and the updated values of other SNPs. They compared it with two EM-based methods: a conventional EM algorithm for maximum likelihood of haplotype frequencies and a weighted EM (WEM) that simultaneously estimates association parameters and haplotype frequencies (Lake et al. 2003). Results based on ten imputations showed that both EM and WEM resulted in smaller imputation errors than their classification tree method; however, the classification tree method was computationally faster than the other methods.

Studies thus far compared only a few methods. In this paper, we describe and examine the performance of eight methods to impute missing genotypes: two haplotype reconstruction methods, three linear regression methods, two *k*-nearest neighbor methods and a regression tree method. They were evaluated using SNP data on chromosome 22 from the HapMap project (The International HapMap Consortium 2005) under the assumption that genotypes are missing completely at random. Their computational complexity was also evaluated.

Statistical methods

Haplotype reconstruction methods

For diploid organisms, haplotype phase information is usually not directly observed but can be estimated by statistical methods. Missing genotypes and missing haplotype phase can be treated in a similar way and be estimated simultaneously using statistical methods. We considered two haplotype reconstruction methods: the EM algorithm (Chiano and Clayton 1998; Excoffier and Slakin 1995; Fallin and Schork 2000; Hawley and Kidd 1995; Lin et al. 2004; Long et al. 1995; Qin et al. 2002; Scheet and Stephens 2006) and the fastPHASE (Scheet and Stephens 2006).

The EM algorithm (Dempster et al. 1977) can handle missing genotypes and unobserved haplotype phase by alternating between an expectation step and a maximization step. The expectation step computes the expected log-likelihood using complete data, including genotypes and haplotype phase from the prior step; the maximization step maximizes the log-likelihood according to its parameters, and as a by product gives posterior probabilities of the complete data, conditional on the observed data. In this paper, we used the haplo.em (Schaid et al. 2002) package, available for either R or Splus, which provides posterior probabilities for all possible pairs of haplotypes for each subject. Throughout this paper, we code a genotype by the number of copies of the rare allele, i.e., 0, 1, or 2. Therefore, we calculated an expected score for a missing genotype of a subject by summing the posterior probabilities of haplotypes that contain the rare allele. When the number of loci in a region is large, the EM algorithm can be very slow. Even worse, it will fail if it requires more computer memory than that available. Therefore, when imputing missing genotypes of a SNP, we only used markers that are close to it. Here, the number of markers on each side of the SNP being imputed is a tuning parameter (n) and we used values 2–5.

The haplotype reconstruction package fastPHASE (Scheet and Stephens 2006) assumes that haplotypes in a population cluster into groups over short chromosome regions, and cluster memberships are allowed to change continuously along a chromosome according to a hidden Markov model (Rabiner 1989). The EM algorithm is used to estimate genetic parameters and haplotype frequencies, and unobserved haplotype phase. For each missing genotype, the posterior mean from fastPHASE was used to predict it. The number of haplotype origins in local regions (n.ho) is a tuning parameter, and according to Scheet and Stephens (2006), n.ho = 8 seemed to perform reasonably well across different scenarios. We used n.ho values of 3, 5, 10, and 15.

Iterative methods

We can infer missing genotypes by iteratively estimating missing values and updating models that formulate the relationship between a SNP and its flanking markers. Assume a data set consists of N subjects and p SNPs. Let $M_{i,j}$ be the genotype of the ith subject at the jth SNP, which is coded to 0, 1, or 2 according to the number of copies of the rare allele of the SNP. Let M_j be the $N \times 1$ vector of the jth SNP, $M_j^{(t)}$ be the corresponding updated vector at the tth iteration, and $M_j^{(0)}$ be the corresponding vector for initial values with missing values replaced by the mean of

observed genotypes at the jth SNP. At each iteration, we imputed missing genotypes in all SNPs in turn:

$$M_1^{(t+1)} \sim f_1^{(t+1)}(M_1 | M_2^{(t)}, M_3^{(t)}, \ldots, M_p^{(t)})$$
$$M_2^{(t+1)} \sim f_2^{(t+1)}(M_2 | M_1^{(t+1)}, M_3^{(t)}, \ldots, M_p^{(t)})$$
$$\ldots$$
$$M_p^{(t+1)} \sim f_p^{(t+1)}(M_p | M_1^{(t+1)}, M_2^{(t+1)}, \ldots, M_{p-1}^{(t+1)}),$$

where f indicates a chosen model. Specifically, at the $(t + 1)$th iteration and for the jth SNP, we first used all subjects with the jth SNP observed to fit a model, then updated the missing values at the jth SNP using the fitted model. The algorithm was considered to have converged if the maximum difference between the imputed values at a current step and those at its previous step was less than 0.01. When f represents a linear regression model, an estimate may be less than 0 or greater than 2. This can lead to inaccurate estimation and occasionally, it may fail to converge. Therefore, when an estimate was below 0 or above 2, we forced it to be 0 or 2, respectively.

We considered six iterative methods: three linear regression methods, two k-nearest neighbor methods, and a regression tree method. For most of these methods, when updating missing values of a SNP, it is not practical to use information from all the other SNPs in a data set, because this can dramatically increase computation time. Therefore, we chose the number of markers on each side of a SNP, n, being updated to be 2, 3, 5, or 10. Because the computation of Eigen values and Eigen vectors of a matrix is relatively fast, for linear regression with singular value decomposition (SVD), n was chosen to be 10, 20, 30, or 50.

Linear regression with backward selection

A difficulty in imputing missing values is to choose the number of predicting SNPs. It is well known that a large number of predictors can cause overfitting and thereby decrease prediction accuracy. For linear regression methods, we considered three strategies: the backward stepwise selection, the least angle regression (LARS), and the SVD. All these methods assumed a linear relationship between the SNP score being predicted and the scores of its flanking markers.

One way to select predictors from flanking SNPs is to use the backward stepwise selection. We used the "step" function with the "backward" direction in Splus, which is based on the Akaike information criterion (Akaike 1974). However, the backward stepwise procedure is computationally expensive. To reduce computational burden, we fixed the predictors for each SNP after five iterations and only reevaluated regression coefficients in later iterations. The tuning parameter for this method is the number of candidate markers on each side of a SNP being imputed, chosen to be 2, 3, 5, or 10.

Linear regression with LARS

As a key step in linear regression methods is model selection, we considered different strategies. Similar to the least absolute shrinkage selection operator (Tibshirani 1996), LARS (Enfron et al. 2004) usually provides more robust estimators of regression coefficients than ordinary linear regression. Like forward stepwise regression, at each step, LARS picks the predictor that is most correlated with the current residuals. Instead of bringing the chosen predictor completely into the model, LARS works in a less greedy manner: it takes the largest step possible in the least angle direction (the direction equiangular among all current predictors) until some other predictor is equally correlated with the current residuals. The authors (Enfron et al. 2004) determined an approximation to the degrees of freedom. We chose the fit that minimized Mallows' Cp (Mallows 1973). The tuning parameter for this method is the number of candidate markers on each side of a SNP being imputed, chosen to be 2, 3, 5, or 10.

Linear regression with SVD

An alternative way of using SNPs as covariates is to use their Eigen vectors, which are linear combinations of SNP scores (Alter et al. 2000). By singular value decomposition, the subjects × SNPs matrix M can be decomposed into a product of three matrices:

$$M_{N\times p} = U_{N\times N} S_{N\times p} V'_{p\times p},$$

where S is a matrix with nonnegative values on the diagonal and zeros off the diagonal, and U and V are orthonormal matrices. Similar to Troyanskaya et al. (2001), the columns of matrix U represent "EigenSNPs". When imputing missing values for the ith SNP, we can use "EigenSNPs" as covariates in linear regressions. EigenSNPs that only explain a small portion of the total variance of the data matrix M usually represent background noise and do not provide useful information about the underlying data structure. We chose the smallest number of EigenSNPs that explained at least a certain percentage of the total variance of the data matrix. Therefore, a tuning parameter is the percentage of variance explained by selected EigenSNPs (p.var), chosen to be 50, 60, 70, or 80%. Because the computation of the singular value

decomposition of a matrix is very fast, we chose the number of SNPs on each side to be 10, 20, 30, or 50.

k-Nearest neighbor

In contrast to assuming linear relationship among the SNP scores, the nearest neighbor method is a nonparametric model. To simplify the presentation of the *k*-nearest neighbor (KNN) methods, let Y be the SNP being updated and x be its flanking markers. Let i be the index for the ith subject. If the flanking markers of Y can predict it accurately, then subjects with similar flanking marker profiles should have similar Y values. Therefore, we can use subjects with similar flanking marker profiles to predict the missing values of a SNP. The original *k*-nearest neighbor fit for $\hat{Y}$ is defined as (Hastie et al. 2001)

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} Y_i,$$

where $N_k(x)$ represents the k nearest neighbors of x based on Euclidean distance. A way to improve the original KNN is to weight the contribution of Y_i to $\hat{Y}$ is

$$\hat{Y}(x) = \frac{\sum_{x_i \in N_k(x)} e^{-||x_i - x||} y_i}{\sum_{x_i \in N_k(x)} e^{-||x_i - x||}},$$

where $||x_i - x||$ is the Euclidean distance between points x_i and x. Notice that we use person-neighbor instead of SNP-neighbor. In addition to the number of markers on each side (chosen to be 2, 3, 5, or 10), the number of nearest neighbors k is a tuning parameter. We considered k = 3, 5, 10, or 15.

Regression tree

Regression tree is another nonparametric model we considered. Here, f indicates a partition tree that is built by recursively applying binary splits to a data set into subsets. An advantage of the tree model is that predictors can be used multiple times in the recursive splitting process, thereby allowing nonlinear effects of predictive variables as well as complex interactions among them. We used the RPART package in R (Atkinson and Therneau 1997). In addition to the number of markers on each side (n = 2, 3, 5, 10), we also varied two other tuning parameters: the complexity parameter (cp) that prevents splitting that does not improve model fit, and the minimum number of observations in a node for which to be split (minsplit). Three values for the complexity parameter were evaluated (0.01, 0.05 and 0.10); and two values for minsplit were evaluated (5 and 10% of the sample size). Using each set of tuning parameters, similar to Dai et al. (2006), a tree was grown to its maximum size allowed by cp and minsplit, without pruning. Assume the jth SNP of the ith subject was missing, its estimate is the mean of the observed genotype scores of M_j that are assigned to the same tree node based on its flanking markers.

Data

The performance of the eight methods was assessed using SNP data on chromosome 22 from the HapMap project (The International HapMap Consortium 2005). We used uncorrelated subjects, i.e., 60 founders from the Centre d'Etude du Polymorphisme Humain samples (CEU), 60 founders from the Yoruba in Ibadan, Nigeria (YRI), 45 Japanese from Tokyo, Japan, and 45 Han Chinese from Beijing, China (J/C). For each analysis panel, all SNPs were required to have minor allele frequencies (MAFs) no less than 5% and P values for the Hardy–Weinberg equilibrium (HWE) test greater than 0.01. The original data set contains about 2–4% of missing values. To compare the estimation accuracy of the eight methods, we generated 100 data sets. In each data set, 5% of observed genotypes were randomly chosen and treated as missing. Each method with a set of chosen tuning parameters was used to estimate the missing values in each generated data set. We calculated the mean square error (MSE) to compare the difference between the inferred and observed genotypes. The mean of the 100 MSEs was then used to calibrate the accuracy of the different methods with different sets of tuning parameters.

We also evaluated the eight methods and the effect of their tuning parameters under different LD levels: strong LD, weak LD, and no LD. We treated the first 100 SNPs that passed the MAF and HWE thresholds on the leftmost region on chromosome 22 as SNPs in strong LD. To select 100 SNPs in weak or no LD, we further required the r^2 value (the square of Pearson's correlation) between any two adjacent SNPs to be less than 0.1 for weak LD and less than 0.0001 for no LD. To describe the LD structure of these data sets, we measured LD levels using another measure: the absolute value of Lewontin's D' (Lewontin 1964) for adjacent pairs of SNPs. $|D'|$ standardizes the deviation of an observed haplotype frequency from its expectation under the assumption of linkage equilibrium. It ranges from 0 to 1, regardless of the allele frequencies of the two SNPs that make up the haplotypes.

Results

Strong LD

The LD strength based on r^2 and $|D'|$ are summarized in Table 1. The entries of the SNPs in strong LD represent the LD levels in the original data. The performance of the different methods, including their tuning parameters that minimized MSE is summarized in Table 2. The accuracy of each method was evaluated based on its smallest MSE, over all sets of parameters. Across all three analysis panels, the fastPHASE method had smaller MSEs than the other methods. Of the remaining methods, LM.lars gave smallest MSEs. While neither KNN nor wtKNN outperformed most of the other methods, the improvement by weighting is consistent. Similar to Dai et al. (2006), we also found the MSEs of EM were smaller than those of Rtree.

Although the optimal tuning parameters were not consistent for the three analysis panels, we can see some common optimal parameters across them. When the number of markers on each side was chosen to be five for the EM algorithm, ten for LM.lars and ten for Rtree, their minimum MSEs were reached. The number of haplotype origins that minimizes the MSEs of fastPHASE was 15 for all three analysis panels. We also found that both KNN and wtKNN gave their best performance when five markers on each side were used. Another trend shown in Table 2 is that the MSEs of YRI were greater than the other two analysis panels. This is because that the YRI data had less LD and hence there was less information to borrow from nearby SNPs (Fig. 1).

Table 1 Summary of SNP data

LD	r^2		$\|D'\|$	
	Mean	Median	Mean	Median
Strong				
CEU	0.32	0.09	0.80	1.00
J/C	0.29	0.09	0.80	1.00
YRI	0.20	0.05	0.80	1.00
Weak 0.1				
CEU	0.05	0.03	0.74	1.00
J/C	0.04	0.03	0.76	1.00
YRI	0.03	0.02	0.69	0.79
None 0.0001				
CEU	<0.01	<0.01	0.02	0.01
J/C	<0.01	<0.01	0.02	0.01
YRI	<0.01	<0.01	0.02	0.01

The means and medians of r^2 and $|D'|$ were measured on adjacent pairs of SNPs

Weak LD

The SNPs in weak LD were restricted to SNPs with adjacent r^2 less than 0.1. Despite this restriction, we found that there was still substantial LD based on $|D'|$. The relative accuracy of the eight methods was similar to those when SNPs were in strong LD. For example, the advantage of fastPHASE and LM.lars over the other methods, the benefit of weighting when using nearest neighbor methods, and the relatively small MSEs of EM to those of Rtree. As mentioned earlier, the relatively small values of $|D'|$ in YRI may explain why the MSEs for YRI were greater than those of the other two analysis panels (Fig. 2; Table 3).

No LD

When SNPs were nearly in linkage equilibrium, the MSEs were much greater than when SNPs were in strong or weak LD. When replacing missing values with observed mean genotype scores, it is not difficult to show that under the assumption of HWE, the theoretical MSE of a SNP is the variance of its genotype score, i.e., $2p(1-p)$, where p is the MAF of the SNP. Based on the allele frequencies of SNPs in no LD, we found the theoretical MSEs of using observed mean scores to impute missing genotypes were 0.34 for CEU, 0.32 for J/C, and 0.31 for YRI. The results in Table 4 show that the minimum MSEs of fastPHASE, LM.lars, and LM.back were close to those when using observed mean scores to impute missing genotypes. This is because when no predictor was chosen in LM.back and LM.lars, we used the observed mean score of a SNP to infer its missing values. In LM.svd, we evaluated its MSEs using the smallest number of EigenSNPs that explained a certain percentage of variance, with no further model selection. This can cause serious overfitting when the data contain no signal, which results in large MSEs.

Most of the optimal tuning parameters in Table 4 brought the estimate of missing values toward observed mean scores. For example, EM, LM.back, LM.lars, and Rtree reached their minimum MSEs when two markers on each side were used, which was the smallest value among the four values considered in our paper. As mentioned in Scheet and Stephens (2006), the number of haplotype origins reflects model complexity—the larger the n.ho, the more parameters in the model. The results in Table 4 show that when SNPs were in no LD, the smallest value of n.ho resulted in the smallest MSEs; therefore simpler models were preferred over more complex models. It is also easy to see from Scheet and Stephens (2006) that, when n.ho = 1 is chosen (which was not considered in our paper), posterior means estimated by fastPHASE is in fact observed mean genotype scores. The percentage of

Table 2 Mean square error for SNPs in strong LD

(a)	EM	fastPHASE	LM.back	LM.lars	LM.svd
Range of tuning parameter 1	n = 2–5	n.ho = 3, 5, 10, 15	n = 2, 3, 5, 10	n = 2, 3, 5, 10	n = 10, 20, 30, 50
Range of tuning parameter 2	–	–	–	–	p.var = 50–80%
Range of tuning parameter 3	–	–	–	–	–
Range of MSE (over tuning parameters)					
CEU	0.087–0.105	0.049–0.082	0.069–0.109	0.058–0.108	0.090–0.159
J/C	0.065–0.116	0.044–0.074	0.062–0.116	0.057–0.116	0.080–0.105
YRI	0.149–0.187	0.092–0.123	0.129–0.166	0.107–0.165	0.158–0.243
Optimal tuning parameter					
CEU	n = 5	n.ho = 15	n = 10	n = 10	n = 30, p.var = 60%
J/C	n = 5	n.ho = 15	n = 5	n = 10	n = 30, p.var = 70%
YRI	n = 5	n.ho = 15	n = 5	n = 10	n = 20, p.var = 50%

(b)	KNN	wtKNN	Rtree
Range of tuning parameter 1	n = 2, 3, 5, 10	n = 2, 3, 5, 10	n = 2, 3, 5, 10
Range of tuning parameter 2	k = 3, 5, 10, 15	k = 3, 5, 10, 15	cp = 1, 5, 10%
Range of tuning parameter 3	–	–	minsplit = 5, 10%
Range of MSE (over tuning parameters)			
CEU	0.101–0.145	0.093–0.136	0.105–0.185
J/C	0.083–0.146	0.076–0.143	0.090–0.152
YRI	0.150–0.210	0.140–0.208	0.155–0.254
Optimal tuning parameter			
CEU	n = 5, k = 5	n = 10, k = 3	n = 10, cp = 10%, minsplit = 10%
J/C	n = 5, k = 3	n = 5, k = 5	n = 10, cp = 5%, minsplit = 10%
YRI	n = 5, k = 5	n = 5, k = 10	n = 10, cp = 10%, minsplit = 10%

Notation for tuning parameters: n is the number of markers on each side of a SNP being imputed; n.ho is number of haplotype origins; p.var is the percentage of variance explained by EigenSNPs; k is the number of nearest neighbors; cp is the complexity parameter, and minsplit is the minimum number of observations in a node to be split

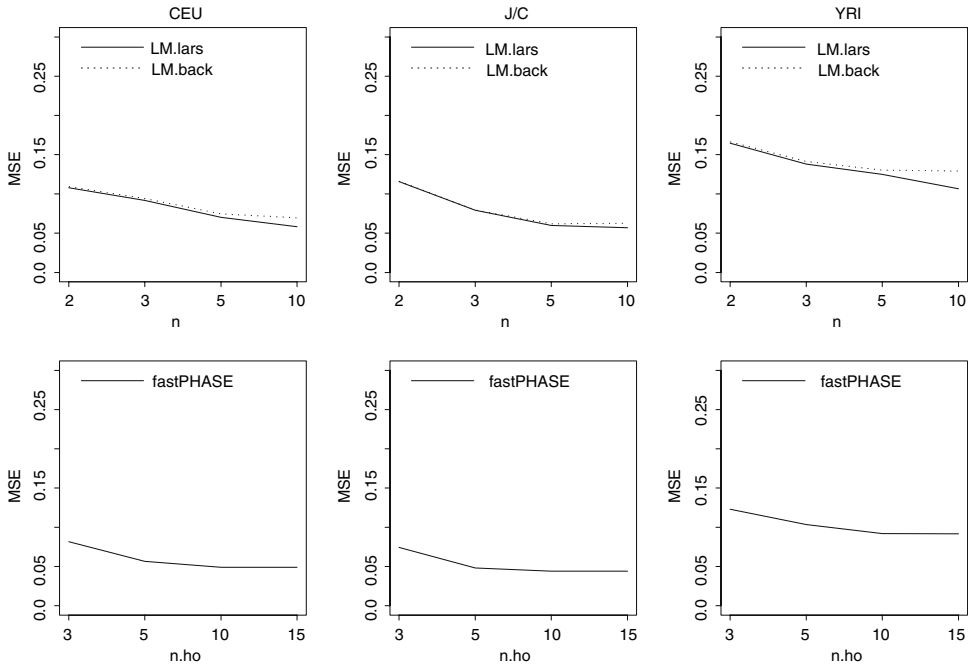

Fig. 1 Mean square error of LM.back, LM.lars, and fastPHASE for SNPs in strong LD

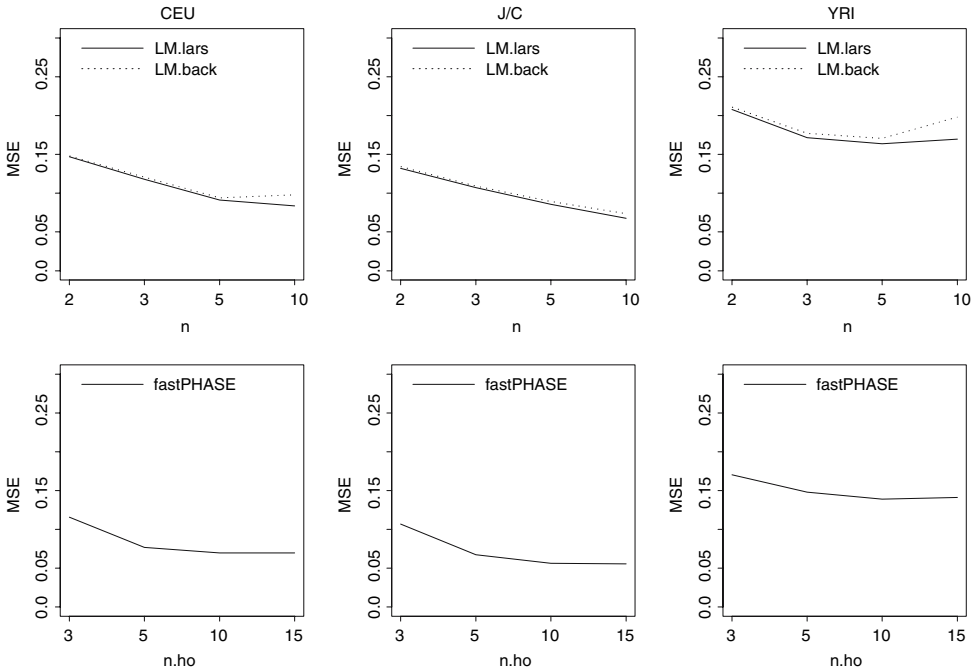

Fig. 2 Mean square error of LM.back, LM.lars, and fastPHASE for SNPs in weak LD

Table 3 Mean square error for SNPs in weak LD

(a)	EM	fastPHASE	LM.back	LM.lars	LM.svd
Tuning parameter 1	n = 2–5	n.ho = 3, 5, 10, 15	n = 2, 3, 5, 10	n = 2, 3, 5, 10	n = 10, 20, 30, 50
Tuning parameter 2	–	–	–	–	p.var = 50–80%
Tuning parameter 3	–	–	–	–	–
Range of MSE (over tuning parameters)					
CEU	0.120–0.146	0.070–0.116	0.094–0.148	0.084–0.147	0.128–0.198
J/C	0.094–0.144	0.056–0.107	0.073–0.134	0.069–0.132	0.098–0.145
YRI	0.188–0.231	0.139–0.170	0.170–0.211	0.164–0.208	0.228–0.372
Optimal tuning parameter					
CEU	n = 5	n.ho = 10	n = 5	n = 10	n = 10, p.var = 60%
J/C	n = 5	n.ho = 15	n = 10	n = 10	n = 30, p.var = 60%
YRI	n = 3	n.ho = 10	n = 5	n = 5	n = 10, p.var = 50%

(b)	KNN	wtKNN	Rtree
Tuning parameter 1	n = 2, 3, 5, 10	n = 2, 3, 5, 10	n = 2, 3, 5, 10
Tuning parameter 2	k = 3, 5, 10, 15	k = 3, 5, 10, 15	cp = 1, 5, 10%
Tuning parameter 3	–	–	minsplit = 5, 10%
Range of MSE (over tuning parameters)			
CEU	0.131–0.184	0.120–0.183	0.129–0.216
J/C	0.111–0.173	0.103–0.172	0.102–0.166
YRI	0.190–0.257	0.182–0.257	0.197–0.280
Optimal tuning parameter			
CEU	n = 10, k = 5	n = 10, k = 5	n = 5, cp = 5%, minsplit = 10%
J/C	n = 10, k = 5	n = 10, k = 5	n = 10, cp = 5%, minsplit = 10%
YRI	n = 5, k = 10	n = 5, k = 10	n = 3, cp = 10%, minsplit = 10%

See footnotes of Table 2

Table 4 Mean square error for SNPs in no LD

(a)	EM	fastPHASE	LM.back	LM.lars	LM.svd
Tuning parameter 1	n = 2–5	n.ho = 3, 5, 10, 15	n = 2, 3, 5, 10	n = 2, 3, 5, 10	n = 10, 20, 30, 50
Tuning parameter 2	–	–	–	–	p.var = 50–80%
Tuning parameter 3	–	–	–	–	–
Range of MSE (over tuning parameters)					
CEU	0.407–0.594	0.365–0.385	0.359–0.477	0.355–0.373	0.696–0.919
J/C	0.376–0.579	0.339–0.351	0.333–0.388	0.331–0.338	0.638–0.894
YRI	0.360–0.462	0.309–0.321	0.305–0.396	0.301–0.311	0.639–0.886
Optimal tuning parameter					
CEU	n = 2	n.ho = 3	n = 2	n = 2	n = 10, p.var = 50%
J/C	n = 2	n.ho = 3	n = 2	n = 2	n = 20, p.var = 50%
YRI	n = 2	n.ho = 3	n = 2	n = 2	n = 10, p.var = 50%

(b)	KNN	wtKNN	Rtree
Tuning parameter 1	n = 2, 3, 5, 10	n = 2, 3, 5, 10	n = 2, 3, 5, 10
Tuning parameter 2	k = 3, 5, 10, 15	k = 3, 5, 10, 15	cp = 1, 5, 10%
Tuning parameter 3	–	–	minsplit = 5, 10%
Range of MSE (over tuning parameters)			
CEU	0.371–0.477	0.373–0.484	0.360–0.662
J/C	0.349–0.441	0.350–0.446	0.336–0.601
YRI	0.312–0.412	0.312–0.418	0.303–0.513
Optimal tuning parameter			
CEU	n = 5, k = 15	n = 5, k = 15	n = 2, cp = 10%, minsplit = 10%
J/C	n = 5, k = 15	n = 5, k = 15	n = 2, cp = 10%, minsplit = 10%
YRI	n = 5, k = 15	n = 5, k = 15	n = 2, cp = 10%, minsplit = 5%

See footnotes of Table 2

variance explained by EigenSNPs that minimized the MSEs of LM.svd was 50%, which was also the smallest value among all considered values. The optimal number of nearest neighbors for KNN and wtKNN was the largest values that had been considered. And it is also interesting to see that weighting increased the MSEs. All these indicate that when SNPs are in linkage equilibrium, learning nothing from neighbors is preferred.

Computational complexity

We evaluated the computational complexity of each method using the 100 SNPs in strong LD from the CEU samples at its optimal set of tuning parameters. Each method was tested on a Sun workstation (Sun 420r) with four 450 MHz CPUs and 4 GB of memory. The computation time (minutes) is displayed in Table 5. The three nonparametric methods, KNN, wtKNN, and Rtree, and the two haplotype reconstruction methods, EM and fastPHASE took less time than the three linear regression methods. Among the three linear regression methods, LM.back took longer time than the others.

Discussion

We considered eight methods to impute missing genotype data. Our results showed that when SNPs were in LD, by incorporating information from nearby SNPs, we can

Table 5 Computation time based on 100 SNPs in strong LD from the CEU samples

Methods	EM	fastPHASE	LM.back	LM.lars	LM.svd	KNN	wtKNN	RTree
Time (min)	6.96	10.10	49.94	33.36	25.84	1.57	5.83	1.68

The data set had 3% missing genotypes

impute missing genotype data much more accurately than replacing missing genotypes with observed mean genotype scores. For each method, we evaluated both MSEs and computation time. The results suggest that fastPHASE outperforms the other methods in terms of both accuracy and efficiency. Here we compared methods to impute missing genotypes at typed SNPs. Based on typed tag SNPs and the multi-locus LD pattern of a reference panel, e.g., data from the HapMap project, genotypes at untyped SNP can also be imputed. Some recent work (Marchini et al. 2007; Nicolae 2006; Servin and Stephens 2007) shows that untyped SNPs can be accurately estimated based on typed tag SNPs. As a result, the power to detect association is greatly improved.

Becker and Knapp (2005) reported that when missing genotypes exist, using most likely inferred haplotype pairs for each subject tends to inflate Type I error rates. Multiple imputations are usually preferred to reduce the bias caused by a single imputation (Little and Rubin 1987) and therefore improve the accuracy of an imputation method. Unfortunately, multiple imputations might not be practical for large-scale genome-wide association analysis. In this paper, we aim to compare the departure of imputed genotype scores from true genotypes for general purpose. We realize that in practice, especially in genetic association studies, imputed genotype scores, i.e., posterior means, do not provide enough information for estimation uncertainties. From this point of view, it would be beneficial to use posterior probabilities of each missing genotype from EM or fastPHASE. In addition, in the future, it would be useful to compare the effects of different imputation methods on association studies.

Among the three linear regression methods, LM.svd had the greatest MSEs. This is not surprising because predictors were selected based on whether they can improve the overall model fit in both LM.back and LM.lars. However, for each chosen p.var for LM.svd, we selected the minimum number of EigenSNPs and then used all the selected EigenSNPs to fit a linear model without further model selection. This might cause overfitting. Similarly, for each chosen number of flanking markers, the EM method estimated haplotype frequencies based on a SNP being imputed and all its flanking markers. Using a fixed number of flanking markers might not be the optimal strategy because recombination events are not uniformly distributed on chromosomes (Lichten and Goldman 1995). In contrast, in fastPHASE, we used the number of haplotype origins as a tuning parameter, which allows SNPs to borrow information from a variable number of flanking markers. Despite some agreements, for each method, the optimal set of tuning parameters for one analysis panel was not necessary the optimal set of another. Even for a given analysis panel, the set of optimal tuning parameters changed when different tag markers were retained. Therefore, as discussed by Scheet and Stephens (2006), cross-validations are useful to find the optimal set of tuning parameters.

When evaluating imputation errors, we randomly selected observed genotype to create artificial missing data. Souverein et al. (2006) investigated the performance of a multiple imputation method based on polytomous logistic regression under different missing mechanisms: missing completely at random, missing at random, and not missing at random. They found that the genotype imputation errors under missing at random and not missing at random were greater than that under missing completely at random. Realizing that informative missing of genotypes can lead to bias in haplotype frequency estimation, Liu et al. (2006) proposed a method to reduce bias in both haplotype frequency estimation and association analysis for two loci when missing was not at random. This approach might not be applicable to genome wide data because of its computational expense. Further studies might be necessary to investigate methods that can handle data that are not missing at random.

Another assumption we made is that the data came from a population. In practice, especially in genetic association studies, data were usually not sampled randomly from a population. As discussed by Dai et al. (2006), ignoring phenotypes, such as disease status, could introduce bias to association studies. Other confounding factors, such as population stratification, can also lead to inaccurate estimation of missing genotypes, which consequently affects the results of association analyses. For fastPHASE, a recent work by Servin and Stephens (2007) suggests that sampling based methods can be used to combine phenotype information. When ancestry information is available, population structure can be taken into account by specifying a subpopulation variable in the fastPHASE package. For regression methods (the three linear regression based methods and the regression tree method), this can be solved by adding phenotypes and other possible confounding factors into models, such as the axes of genetic variation based on principal component analysis (Price et al. 2006) when ancestry information is unknown or indicators of subpopulations when ancestry information is available.

Acknowledgments The authors are grateful to the three anonymous reviewers for their constructive suggestions. This work was supported by the U.S. Public Health Service, National Institutes of Health, contract grant number GM065450.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automatic Control 19:716–723

Alter O, Brown P, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 97:10101–10106

Becker T, Knapp M (2005) Impact of missing genotype data on Monte–Carlo simulation based haplotype analysis. Hum Hered 59:185–189

Chiano MN, Clayton DG (1998) Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 62:55–60

Dai JY, Ruczinski I, LeBlanc M, Kooperberg C (2006) Imputation methods to improve inference in SNP association studies. Genet Epidemiol 30:690–702

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Enfron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32:407–451

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

Fallin D, Schork N (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation–maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, NY

Hawley M, Kidd K (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Hoti F, Sillanpaa MJ (2006) Bayesian mapping of genotype expression interactions in quantitative and qualitative traits. Heredity 97:4–18

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

Lewontin R (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 120:849–852

Lichten M, Goldman A (1995) Meiotic recombination hotspots. Annu Rev Genet 29:423–444

Lin S, Chakravarti A, Cutler D (2004) Haplotype and missing data inference in nuclear families. Genome Res 14:1624–1632

Little R, Rubin D (1987) Statistical analysis with missing data. Wiley, New York

Liu N, Beerman I, Lifton R, Zhao H (2006) Haplotype analysis in the presence of informatively missing genotype data. Genet Epidemiol 30:290–300

Long J, Williams R, Urbanek M (1995) An E–M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810

Mallows C (1973) Some comments on Cp. Technometrics 15:661–675

Marchini J, Culter D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P, the International HapMap Consortium (2006) A comparison of phasing algorithm for trios and unrelated individuals. Am J Hum Genet 78:437–450

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913

Nicolae DL (2006) Testing untyped alleles (TUNA)—applications to genome-wide association studies. Genet Epidemiol 30:718–727

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

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

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

Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286

Schaid D, Rowland C, Tines D, Jacobson RM, Poland G (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotype phase. Am J Hum Genet 78:629–644

Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3(7):e114

Souverein OW, Zwinderman AH, Tanck MWT (2006) Multiple imputation of missing genotype data for unrelated individuals. Anna Hum Genet 70:372–381

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462

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

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Therneau T, Atkinson E (1997) An introduction to recursive partitioning using the RPART routines. Tech Rep 61:52

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525