European Journal of Human Genetics (2008) 16, 487–495

www.nature.com/ejhg

ARTICLE

Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks

Yan V Sun*,[1] and Sharon LR Kardia[1]

[1]*Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA*

With advances in high-throughput single-nucleotide polymorphism (SNP) genotyping, the amount of genotype data available for genetic studies is steadily increasing, and with it comes new abilities to study multigene interactions as well as to develop higher dimensional genetic models that more closely represent the polygenic nature of common disease risk. The combined impact of even small amounts of missing data on a multi-SNP analysis may be considerable. In this study, we present a neural network method for imputing missing SNP genotype data. We compared its imputation accuracy with fastPHASE and an expectation–maximization algorithm implemented in HelixTree. In a simulation data set of 1000 SNPs and 1000 subjects, 1, 5 and 10% of genotypes were randomly masked. Four levels of linkage disequilibrium (LD), LD $R^2<0.2$, $R^2<0.5$, $R^2<0.8$ and no LD threshold, were examined to evaluate the impact of LD on imputation accuracy. All three methods are capable of imputing most missing genotypes accurately (accuracy $>86\%$). The neural network method accurately predicted 92.0–95.9% of the missing genotypes. In a real data set comparison with 419 subjects and 126 SNPs from chromosome 2, the neural network method achieves the highest imputation accuracies $>83.1\%$ with missing rate from 1 to 5%. Using 90 HapMap subjects with 1962 SNPs, fastPHASE had the highest accuracy ($\sim 97\%$) while the other two methods had $>95\%$ accuracy. These results indicate that the neural network model is an accurate and convenient tool, requiring minimal parameter tuning for SNP data recovery, and provides a valuable alternative to usual complete-case analysis.
European Journal of Human Genetics (2008) **16,** 487–495; doi:10.1038/sj.ejhg.5201988; published online 16 January 2008

Keywords: SNP; neural networks; missing data imputation; genotype prediction

Introduction

The availability of high-throughput single-nucleotide polymorphism (SNP) genotyping platforms has resulted in an exponential increase of measured genotypes. Therefore, genetic epidemiological studies are faced with complex issues of genotyping error and missing data. Methods to assess genotyping error in linkage and family-based association studies, as well as their impact on statistical inferences, has been the focus of many studies.[1–3] The study of missing genotype data, on the other hand, has not been an equally active area of research. For genotyping data, the issue of missing data and its imputation is perhaps a more contentious issue than genotyping error issues since it implies creating genotype data, which are often considered too individualistic to be imputed. Many software packages used in statistical analysis to identify genetic predictors of disease require a complete data subsample for the estimation of model parameters. There is a large literature base to suggest that not imputing missing data may have serious consequences for statistical validity, and that it affects the estimability of statistical parameters that are intended to be generalized to the larger population of inference.[4–6] Most recently, a study suggests

*Correspondence: Dr YV Sun, Department of Epidemiology, School of Public Health, University of Michigan, 109 Observatory no. 4605, Ann Arbor, MI 48109-2029, USA.
Tel: +1 734 615 6279; Fax: +1 734 764 1357;
E-mail: yansun@umich.edu
Received 30 May 2007; revised 21 November 2007; accepted 28 November 2007; published online 16 January 2008

that imputation generally improves efficiency over the standard practice of ignoring missing data in genetic association studies of common human diseases.[7] The compounded effects of missing data are also likely to contribute to wide variability in reported association study results and ultimately contribute to their lack of replicability.

In this paper, we present a neural network method with variable and model selection strategies for imputing missing SNP genotype data. To assess its strengths and limitations, we compared the neural network to two other imputation methods that use haplotype-phasing algorithms. Knowing that most genetic variations such as SNPs are moderately dependent on each other, we can utilize the predictive model to illustrate the dependencies to impute the missing genotypic data.[8,9] The goal of imputing missing SNP genotypes focuses on accurately predicting individual missing values. Because of the amount of correlations between SNPs, especially high-density SNPs, it is feasible to accurately predict missing SNP genotypes.[9]

Feed-forward neural networks (FFNNs) have received considerable attention due to their successful use in a wide variety of statistical applications, including regression and classification problems. A general description of the use of FFNNs can be found in Bishop.[10] Ripley[11] provides examples and applications related to pattern recognition and gives a detailed account of fitting and prediction procedures. The Bayesian approach to fitting FFNNs has gained much acceptance.[12,13] In this study, we use an FFNN model and a Bayesian approach to model selection to classify subjects into one of three genotype categories for each SNP based on the predictive ability of the genotypes of other SNPs.

Model selection is one of the most important considerations when fitting FFNNs. In addition to the complexity of finding the best predictor variables, as in regression problems, FFNN also requires consideration of determining the number of hidden nodes. Including too many predictors, hidden layers or hidden nodes can lead to overfitting, and thus poor classification performance. Including unrelated or too few predictor variables can also result in poor predictive performance. The Bayesian Information Criterion (BIC)[14] has been shown to be an approximation to the log of a Bayes factor[15] and penalizes the likelihood based on the number of parameters in the model and the sample size. In this study, we used an FFNN model with a single hidden layer and a single hidden node to model each SNP variable in turn as the response. A neural network with a single hidden layer and a single hidden node is equivalent to a logistic regression model. Although, in this application, the simplest neural network model is capable of predicting missing SNP genotypes accurately, a more complicated neural network model is always available using the same framework to potentially improve performance under different situations. The predictor variables are chosen from the best candidates among the other SNP variables based on the BIC.

Missing SNP genotypes can also be imputed by the phasing algorithms that estimate the haplotype phase from genotype data using population genetic models. Several studies have implied that such phasing algorithms are capable of imputing the missing SNP genotypes as well as imputing haplotypes.[16,17] The PHASE (v2.1) algorithm is reported to be the most accurate phasing method[16] but is known to be very computationally intensive and slow. Recently a new algorithm, fastPHASE,[17] was developed to estimate haplotypes and can impute haplotypes and the missing SNP data at a much faster rate in samples of unrelated individuals from natural populations. As a comparison, we tested another linkage disequilibrium (LD)s-based algorithm which is an extension of the expectation–maximization (EM) algorithm which has been implemented in the genetic association software HelixTree.[18]

Materials and methods

Data

To test the SNP prediction accuracy, we used complete simulated SNP data sets generated by the *ms* program.[19] We implemented a standard coalescent model of 1000 SNPs across a 6-Mbp region, which has the same density as the Affymetrix human 500K SNP genotyping array, assuming an effective population size of 10 000 individuals with recombination and mutation rates both equal to 10^{-8} per generation and per bp. Three other tagSNP data sets with LD $R^2<0.8$ (680 tagSNPs), LD $R^2<0.5$ (552 tagSNPs) and LD $R^2<0.2$ (288 tagSNPs)[20] were produced as subsets of the 1000 SNP data to determine the LD effects on the imputation accuracy. There is no minor allele frequency cutoff used to generate these data sets. From each complete set of SNP genotypes, 1, 5 and 10% of data were randomly masked five times to estimate the imputation accuracy. Each of the FFNN, fastPHASE and EM methods was applied to the 60 data sets (five replicates of three missing rates and the four LD levels) to impute the missing genotypes. The imputed genotypes were then compared to the actual genotypes to estimate the SNP imputation accuracy.

To compare the missing SNP genotype imputation methods in a dense real data set, we downloaded SNP genotypes in a 6-Mb region on chromosome 22 measured on 90 HapMap CEPH participants. After removing monomorphic SNPs and SNPs with missing rate more than 0.1, we obtained a complete genotype data set with 1962 SNPs and 90 individuals. To estimate the imputation accuracy 1, 5 and 10% of data were randomly masked five times. Each of the FFNN, fastPHASE and EM methods was applied to the five data sets (five replicates of three missing rates) to impute the missing genotypes. The imputed genotypes were then compared to the actual genotypes (masked) to estimate the SNP imputation accuracy using means and SD from the five replicate data set.

As part of the Genetic Epidemiology Network of Arteriopathy (GENOA) study, 126 chromosome 2 SNPs (on 25 candidate genes) spanning 42.5 Mbp were measured on 1458 subjects collected in sibships.[21,22] In this study, we used a subset of 126 SNPs measured on 419 unrelated subjects as a real data example. Overall, 0.92% of the SNP data were missing. The number of missing observations per SNP ranges from 0 to 13 out of 419 subjects. Additional 1, 2 and 5% of SNP genotypes were randomly masked five times to create samples for imputation method comparison in a real data set. The true genotypes were compared to the imputed genotypes from the three methods, FFNN, fastPHASE and HelixTree EM, to calculate the imputation accuracy.

Neural network model

FFNNs are often developed in a nonparametric regression context with normal error terms. They have been shown to be universal approximators in the sense that they can approximate any continuous function with any specified degree of accuracy by increasing the number of hidden nodes.[23,24] To extend the model to classification, each categorical response is modeled as a multinomial random variable. This model is often referred to as the softmax model in the computer science literature.[25]

A large enough network must be used to fit the data well, but small enough to avoid overfitting and poor predictive performance. Therefore, a single hidden layer was chosen for the FFNN model which is the simplest but capable of predicting missing genotypes in the study. For the same reason, for each SNP variable modeled as the response, we chose five SNP predictors with the smallest *P*-values in a series of 3×3 contingency tables based on the χ^2-test of independence. Among those five SNPs, we fit FFNN models to all possible combinations taken one at a time, two at a time, three at a time and so on, resulting in a total of 31 models. For each model, a data set with complete cases was constructed by removing the missing values. Of those 31 models, the BIC[14] was used to select a final model. The BIC has been shown to be an approximation to the log of the Bayes factor,[15] and penalizes the likelihood based on the number of parameters in the model and the sample size.

In a sense, FFNNs can be regarded as a nonlinear prediction model. Consider an FFNN with one hidden layer and M hidden nodes. Let $y_i = (y_{i1}, y_{i2}, y_{i3})$ be the multinomial response for subject i, $i = 1,\ldots,n$, where $y_{ij} = 1$ when a subject belongs to category j, $j = 1,2,3$, and $y_{ij} = 0$ otherwise. Denote by π_{ij} the true underlying probability $P(y_{ij} = 1,|\pi_{ij})$, and by π_{ij}, the fitted value under the model. The likelihood is

$$f(y|\pi) = \prod_{i=1}^{n}\prod_{j=1}^{3}\pi_{ij}^{yij}$$

and the fitted probabilities are found from the neural network by

$$\hat{\pi}_{ij} = \frac{\exp(\hat{\eta}_{ij})}{\sum_{r=1}^{3}\exp(\hat{\eta}_{ij})}$$

$$\eta_{ij} = \beta_{0j} + \sum_{m=1}^{M}\beta_{mj}\,\Psi(x_i^T\gamma_m)$$

$$\Psi(\omega) = \frac{e^{\omega}}{1+e^{\omega}}$$

where $x_i^{\mathrm{T}} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ are the explanatory variables, and $\beta^{\mathrm{T}} = (\beta_{01}, \beta_{02}, \beta_{03}, \ldots, \beta_{M3})$ and $\gamma_M^{\mathrm{T}} = (\gamma_{M0}, \gamma_{M1}, \gamma_{M2}, \ldots, \gamma_{Mp})$ are called weight parameters. For this model with p predictor variables and M hidden nodes there are $3 + M(2p + 4)$ weight parameters. The function Ψ is often referred to as the activation function and can be any smooth sigmoidal function. The logistic activation function is commonly used and is the one we chose for this application. To improve the imputation accuracy for the poorly predicted SNPs, we considered more complex models by attempting to increase the number of hidden nodes ($M = 1$–7) using the R software package which uses a quasi-Newton method of optimization known as the Broyden, Fletcher, Goldfarb and Shanno (BFGS) algorithm (for details, see Venables and Ripley[26]). These attempts did not significantly improve classification performance. As a result, all FFNN models have one ($M = 1$) hidden node. Since the SNPs are categorical variables with three levels, each predictor was coded as two indicator variables and each model required $2p + 7$ weight parameters. Thus, a model with three predictor SNPs required estimation of 13 parameters.

The BIC that we used for model selection is a modification of the likelihood ratio statistic (LRS). It is well known that for large samples the LRS rejects good models in favor of models with more parameters.[27] To take this into account the BIC penalizes models based on large sample sizes that have large numbers of parameters. The BIC can be calculated using

$$\mathrm{BIC} = L - \frac{p}{2}\log n$$

where L is the log-likelihood evaluated at the parameter estimates, p is the number of parameters in the model and n is the sample size.

Once a final set of predictors was selected for a particular SNP using BIC, a complete subset including all predictor SNPs and the outcome SNP was generated by removing the missing values to test the accuracy of FFNN model. Note that the number of cases in each complete data set varies due to different patterns of missing data on the variables included in each model. In some cases, the subject carries missing values for both predictors and outcome. In that case, the missing genotype cannot be imputed. Such situations impair the recovery rate, especially for data sets

with higher missing data rates. To address this issue and improve the recovery rate, we sequentially applied the FFNN algorithm multiple times for data sets with higher missing rates. Because the input data are different for each imputation round, the model selection process can pick separate models with an alternative set of predictors that have different missing data patterns. As a result, more missing genotypes can be recovered using additional rounds.

An R program was implemented for neural network classification of SNP genotypes and for prediction of missing SNP genotypes. First, for each SNP variable the program selects the five most associated SNPs based on a χ^2-test of association. Then it fits FFNNs (nnet library in R) to all SNP variables using up to five predictor SNPs based on the BIC criterion for model selection. For each FFNN model, a complete subset was created using the selected predictor(s). Finally, the missing genotypes are imputed by using the model and the values of the predictors from the observed data. Multiple rounds of the imputation are done by using the imputed data from the last round as the input data.

Additional SNP imputation methods

We intended to apply PHASE (v2.1) in this study because of its high reported imputation accuracy. However, it is not computationally practical to process the simulation data sets under our current setup. For one of the simulation data sets (1000 SNPs and 1000 subjects), it took about 2 weeks to calculate approximately 20% of the task. Therefore, we only compared fastPHASE, which has similar imputation accuracy as PHASE (v2.1).[17] To impute the missing SNP genotypes with fastPHASE, the simulated data sets, HapMap chromosome 22 data set and the GENOA chromosome 2 data were ordered by their genomic locations and were transformed to the compatible file format. Using a built-in utility of fastPHASE, the cluster number was optimally selected. The iteration number of 50 was used to achieve optimal results.

Using an extension of the EM algorithm implemented in HelixTree (Golden Helix Inc., Bozeman, MT, USA), we inferred missing values from neighboring markers.[18] Different parameter settings, such as the size of marker window, the number of highest LD markers and the number of EM iterations, were examined. The best results were obtained when the 20 SNPs with the highest LD R^2 value are selected within a window of 30 markers centered around the marker of interest. For this paper, missing values were computed through the 20-marker haplotypes with the EM convergence tolerance of 0.001 and the maximum EM iteration number of 50.

Results

For each SNP in the simulation data, we fit one FFNN model to the training data and then used the fitted values to classify test subjects using the R nnet function. To improve the missing data recovery rate, multiple rounds of FFNNs were applied sequentially. Table 1 summarizes the imputation accuracy for missing rates of 1, 5 and 10% at four levels of LD. Two, three and four rounds of FFNNs (data with higher missing rates need more rounds of imputation) were applied for missing rate of 1, 5 and 10%, respectively, to obtain the optimal recovery rate. For all LD levels we have tested, the multiple-run strategy improves the imputation accuracy. After running five rounds of imputation, we found that although multiple rounds of FFNN prediction increase the accuracy, especially for higher missing rate data sets, the accuracies peak after a limited number of rounds. At missing rate of 1, 5 and 10%, the necessary rounds are 2, 3 and 4, respectively, to obtain the best results. For each combination of missing rate and LD level, we created five missing data sets to compare the three methods. In the data set without LD threshold, the mean accuracies are 95.9, 94.7 and 94.7% at missing rate of 1, 5 and 10%, respectively. For SNPs with LD $R^2<0.8$, the mean accuracies drop to 93.1, 92.9 and 92.1% at 1, 5 and 10% missing rate, respectively. For SNPs with LD $R^2<0.5$, about 92.5, 92.7 and 91.6% of missing data can be accurately recovered with 1, 5 and 10% missingness, respectively. For SNPs with LD $R^2<0.2$, the mean accuracies of 93.2, 92.3 and 92.0% are obtained for data with missing rate of 1, 5 and 10%, respectively. Overall, the lower level of the inter-SNP dependency (indicated by

Table 1 Running multiple rounds of FFNN improves the imputation accuracy at three missing rates and four LD levels with the simulation data set

	1% missing		*5% missing*			*10% missing*			
	Round 1 (%)	*Round 2 (%)*	*Round 1 (%)*	*Round 2 (%)*	*Round 3 (%)*	*Round 1 (%)*	*Round 2 (%)*	*Round 3 (%)*	*Round 4 (%)*
All SNP	94.6	**95.9**	86.4	92.4	**94.7**	77.0	87.6	92.6	**94.7**
LD $R^2<0.8$	92.3	**93.1**	81.8	88.3	**92.9**	69.6	80.6	87.8	**92.1**
LD $R^2<0.5$	91.2	**92.5**	80.6	87.2	**92.7**	68.4	79.5	86.6	**91.6**
LD $R^2<0.2$	91.4	**93.2**	81.3	88.3	**92.3**	69.5	81.4	89.1	**92.0**

FFNN, feed-forward neural network; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.
Bold values indicate the final values for FFNN accuracy.

LD R^2) mildly decreases the imputation accuracy. However, when LD $R^2 < 0.5$, the further reduction of the LD does not have obvious impact on the imputation accuracy. The results in Table 1 also indicate that the FFNN method has relative stable performance at missing rates from 1 to 10% but requires more computation rounds for a data set with a higher missing rate to reach the best imputation accuracy.

The results in Figure 1 provide the comparison of the imputation accuracy (mean and SD) of FFNN, fastPHASE and the HelixTree EM methods in the five simulated data sets. After increasing the window size and the EM iterations, the EM algorithm achieves the best accuracies at the three missing data rates across the four LD levels. Its imputation accuracies are relatively stable when

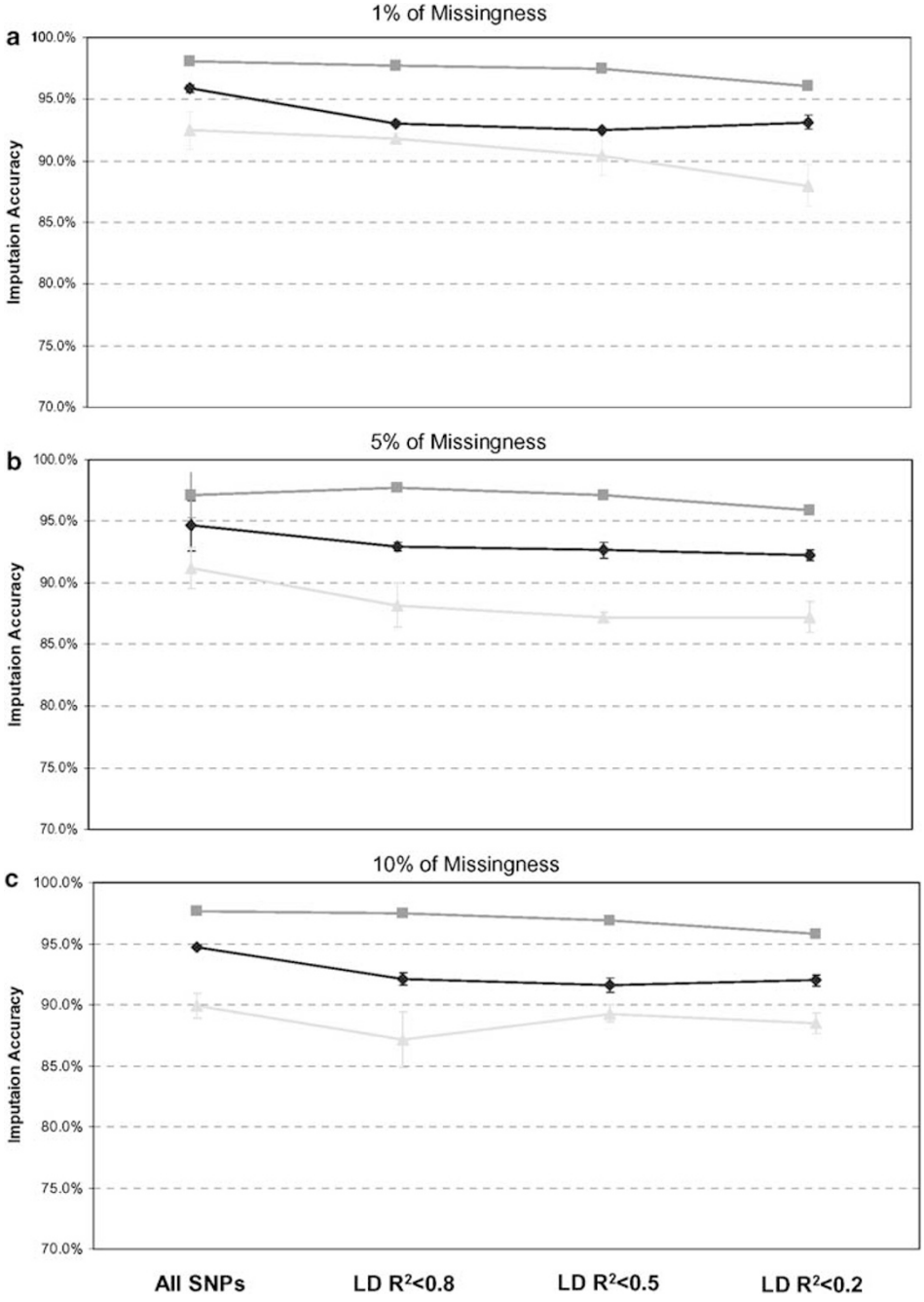

Figure 1 The comparison of mean imputation accuracy in various linkage disequilibrium (LD) structure at missing rate of 1% (**a**), 5% (**b**) and 10% (**c**) with simulated data. The mean imputation accuracies of FFNN (◆), EM (■) and fastPHASE (▲) methods are presented with four LD levels at each missing rate. The error bars indicate the SD of the imputation accuracies based on five samples.

LD $R^2 > 0.5$ and decrease when LD $R^2 < 0.2$. Similar to the FFNN method, the missing rate does not have clear impact on the recovery rate for the four tested LD levels. The fastPHASE program provides very good but relatively less accurate SNP imputation results. The fastPHASE method also shares the same trend that low LD correlation decreases imputation accuracy. In Figure 1, the LD impact of imputation accuracy is summarized at missing rate of 1% (Figure 1a), 5% (Figure 1b) and 10% (Figure 1c).

Figure 2a summarizes the imputation accuracy for 1, 5 and 10% missing data of HapMap chromosome 22 data. For each missing rate, five samples were generated to estimate the SD. Three rounds of FFNN were applied to obtain the optimal recovery rate. The mean accuracies for FFNN method are 96.2, 95.4 and 95.1% at missing rate of 1, 5 and 10%, respectively. At the same level of missing rate, the mean accuracies of the HelixTree EM and fastPHASE methods are 95.7, 97.4% (1% missing); 95.4, 97.2% (5% missing) and 95.2, 96.9% (10% missing). The HelixTree EM and fastPHASE methods also have slightly larger SD than the FFNN method. In this HapMap data comparison, fastPHASE has the highest imputation accuracy and all three methods have very high mean accuracy above 95%.

Figure 2b summarizes the imputation accuracy for 1, 2 and 5% additional missing data of GENOA chromosome 2 data. For each missing rate, five samples were generated to estimate the SD. Three rounds of FFNN were applied to obtain the optimal recovery rate. Same as the simulation results, the accuracies maximize after a limited number of rounds of FFNN imputation procedure. The mean accuracies for FFNN method are 86.8, 86.5 and 83.1% at missing rate of 1, 2 and 5%, respectively. At the same level of missing rate, the mean accuracies of the HelixTree EM and fastPHASE methods are 73.6, 82.1% (1% missing); 73.0, 81.6% (2% missing) and 72.9, 80.7% (5% missing). The HelixTree EM and fastPHASE methods also have slightly larger SD than the FFNN method. In this real data comparison, the FFNN method has superior performance over the other two methods.

To identify the accurate imputation coverage across the three methods, the imputed genotypes from all three methods are compared to each other as well as the actual genotypes. As an example, the overlapping results of the data set with 1% missing and no LD threshold are summarized in Figure 3. Out of 10 000 missing genotypes

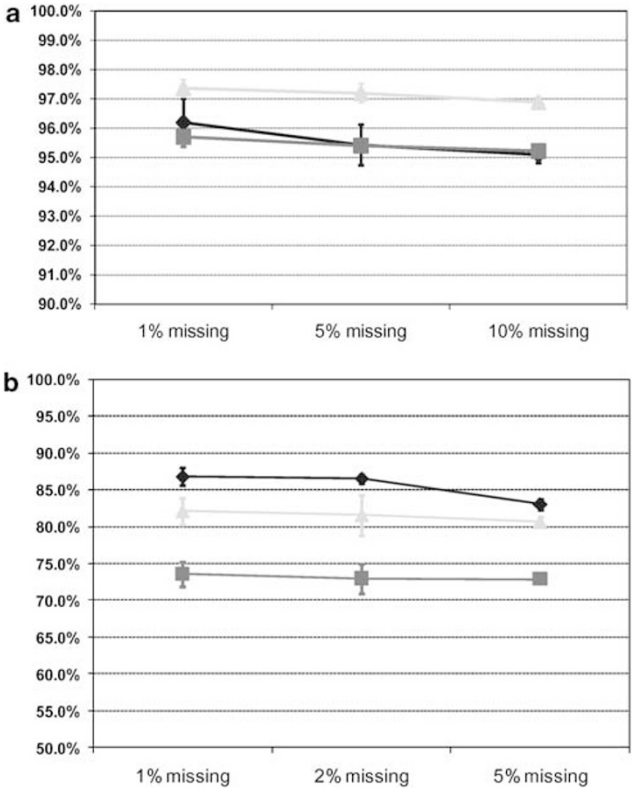

Figure 2 The comparison of mean imputation accuracy on HapMap chromosome 19 data and GENOA chromosome 2 data with additional missingness of 1, 2 and 5%. The mean imputation accuracies of FFNN (◆), EM (■) and fastPHASE (▲) methods are calculated by comparing the imputed genotypes to the randomly masked true genotypes of five samples. The FFNN results are based on the imputed genotypes by running three rounds of the FFNN algorithm. The error bars indicate the SD of the imputation accuracies based on five samples. (**a**) HapMap chromosome 22 data; (**b**) GENOA chromosome 2 data.

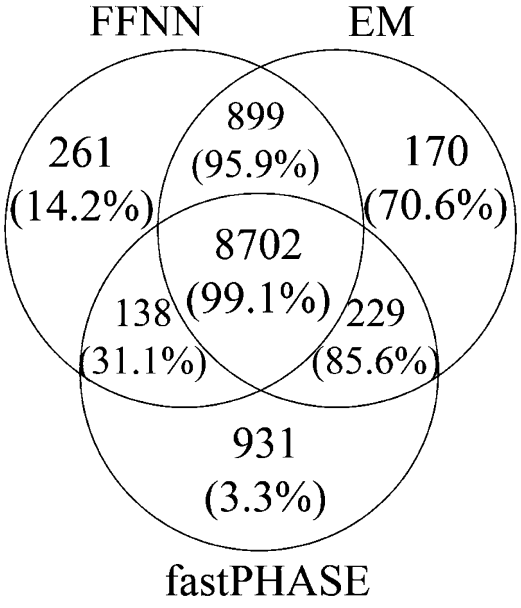

Figure 3 The imputation accuracy of the overlapped genotype inference from the FFNN, EM and fastPHASE methods. As an example, the overlap of the imputed genotypes of the three methods is presented in a Venn diagram at missing rate of 1% for all SNPs without LD restriction. The numbers of imputed genotypes are listed in all overlapping regions. The imputation accuracies for these regions are calculated by comparing to the true genotypes and are listed in the parenthesis.

8702 are imputed identically by all the three methods with accuracy of 99.1%. Among 899 missing genotypes, which are imputed by the FFNN and the EM methods, but not by the fastPHASE method, 95.9% of them are correct. The accuracy of 229 missing genotypes, which are consistently imputed by EM and fastPHASE methods, but not by the FFNN method, is 85.6%. The accuracy of 138 missing genotypes, which are consistently imputed by FFNN and fastPHASE method, but not by EM method, is 31.1%. Among the 261 FFNN imputed genotypes, which are imputed differently from either of the EM or fastPHASE method, the accuracy rate is 14.2%. For the EM method, the nonoverlapped SNP imputation has accuracy of 70.6% among total 170 genotypes. At last, the 931 genotypes imputed only by fastPHASE method have imputation accuracy of only 3.3%. We also observed that using two methods together to identify commonly imputed genotypes has better accuracy than using each method alone. For all 9601 missing genotypes consistently inferred by both FFNN and EM methods, the imputation accuracy is 98.8%, which is higher than using either method alone, 95.7% for the FFNN method and 98.0% for the EM method. For the 8931 missing genotypes consistently inferred by both EM and fastPHASE methods, the imputation accuracy is 98.8%, which is higher than using either method alone, 98.0% for the EM method and 89.0% for the fastPHASE method. Similarly, among the 8840 genotypes consistently inferred by both FFNN and fastPHASE methods, the imputation accuracy is 98.0%, which is higher than using FFNN or fastPHASE methods alone.

Discussion

The results obtained from fitting any statistical model can be seriously affected by patterns of missing data, and steps must be taken to formulate an effective strategy for dealing with this situation, especially in multigenic studies where missing data issues are compounded by different levels of missing SNP data.

To avoid losing valuable information and decreasing the power of a study by selecting a complete subset (ie, no missing data) of data, we demonstrate that missing SNP genotype data can be imputed with high reliability using the naturally occurring correlations between SNP frequency distributions due to LD. As with any data imputation method it is important to first consider the amount of missing data in the data set on a variable-by-variable basis. Our procedures worked successfully (mean accurate classification rates $>92\%$) in cases where missing data ranged between 1 and 10%. In cases where there are substantially more missing data, there should be some caution in using data imputation methods without rigorous investigation into the amount of genotyping error it may introduce. The impact of genotyping error on linkage and association has been investigated by several researchers,[28–30] and we suggest applying an upper bound misclassification error rate for an imputation procedure of 10–15% to minimize the introduction of genotyping errors into data sets. For a typical SNP that is missing approximately 5% of the data, a 10% misclassification rate will result in recovering 90% of the data without error and will introduce approximately 0.5–0.75% new error into the whole data set, which is equivalent to most accepted genotyping error rates. Of course, this is assuming that the original genotyping data are collected without error, which is not true. However, efforts to reduce genotyping error in high-throughput laboratories is increasing by applying quality control standards in genotyping laboratories, doing diagnostic statistical tests such as Hardy–Weinberg equilibrium tests, using available family data to test for Mendelian consistency, and applying good epidemiological practices of submitting blind duplicates to actually estimate error rates.

With even low levels of inter-SNP correlation, classification tools such as neural networks can be utilized to identify these correlations to impute missing genotypic data. In our study, the FFNN had the best overall imputation accuracy on the real data from the chromosome 2 fine mapping project. Other alternative tools for missing genotype imputation are the phasing algorithms, which impute the unknown haplotypes for each subject, and the missing genotypes can be inferred by combining the imputed haplotypes. Comparing the fastPHASE algorithm, which is reported as one of the most accurate phasing algorithms,[17] and the EM phasing algorithm implemented in HelixTree, we observed that the fastPHASE algorithm had the best imputation accuracy using HapMap data set with 90 subjects and 1962 SNPs, which confirmed the previous report,[17] and the EM algorithm had the best overall imputation accuracy on the simulated data sets. The FFNN method was the best imputation strategy on the real data set, and was the second best performer in both the HapMap data and the simulation data. It should be noted that the majority of the imputed genotypes were identical from the three methods. In all of the three data sets, the genotypes are missing at complete random. The different performance probably relies on the different structure of inter-SNP dependency. The real data include small number of SNPs in a relative large region. Therefore, the pairwise LD-based phasing algorithms may not be able to capture the multi-SNP dependency, which could be the type of dependency critical in this real data. The FFNN model can take advantage of both pairwise and multi-SNP dependencies to predict missing genotypes. However, in the simulated data, even for data with LD $R^2<0.2$, it seems that the pairwise SNP correlations are sufficient to predict the missing genotypes accurately. Additionally, as a variable selection criterion, the BIC seems to perform well in developing parsimonious predictive models. We found that for some SNPs the FFNN classification rates were

excellent with only one or two predictors. In other cases additional predictors (up to five total SNPs) were added to the model in an attempt to balance goodness of fit with parsimony.

In the current FFNN method, we used the χ^2-test to select the top five SNPs correlated with the predicted SNP to build the FFNN model. It is possible that the five best predictors selected by pairwise correlation are not the most informative predictors carrying the maximal predictive power as a group because of potential higher dimensional interactions. An improved method of selecting the most informative SNPs will increase the imputation accuracy of the FFNN method. In addition, our FFNN method only considers the best predictive model for an SNP variable but classification rates may be even further improved by considering multiple predictive models (which all meet the BIC cutoff) when the SNP predictors also contain missing values. Alternatively, combining different imputation methods together, as we did in this study, is another way to maximize accurate SNP imputation. For example, we found that combining the FFNN and the EM algorithm results together provided the highest predictive accuracy of the real data set. For the simulated data set of 1% missing genotypes without LD threshold, combining the imputation results of the two methods delivers 98.8% accuracy and 96.0% recovery rate (4% of the missing genotypes remain missing).

However, such good performance can only be obtained by tuning the parameters of the EM algorithm. The performance of this EM method is sensitive to the selection of the parameters and data sets. With the default parameter setting, the imputation accuracies for the simulated data sets were often lower than 90% and even below 85% in one test case. Because the optimal settings for a real data set are hardly predictable, it is highly recommended to create a subdata set for validation to help identify the optimal settings. Using the validation procedure, different combinations of settings are evaluated by the comparison between masked and imputed genotypes.

To further minimize introducing genotype errors through imputation, one option is to identify those imputed genotypes that are the same (ie, overlapping) across methods. While this may introduce some bias, it is better than not imputing at all, and significantly reduces the missing data problem. Additionally more intensive modeling for SNP genotypes that are difficult to impute could then be performed.

The mechanism of missing data is an important issue in a missing data imputation study. In this study, we used the paradigm of missing completely at random to evaluate and compare the imputation accuracies. However, in real data sets, the mechanism of missing data may not be obvious depending on various factors, such as genotyping methods, quality control procedures and DNA quality. Knowing the actual mechanism of missing data will be helpful to improve the imputation methods and obtain more accurate results. For association studies, it is well-known that multiple imputation is superior to single imputation.[6] The direct use of the predicted values as observed values may underestimate the variance in the association analysis. Especially for predicted SNP genotypes with low confidence, such bias is not trivial and needs to be estimated and corrected by multiple imputation methods.

Although the simulated data sets and one set of real data investigated here may not be representative of all types of genotyping data, the results indicate that the neural network model is an accurate and convenient tool requiring minimal parameter tuning for SNP data recovery and provides a valuable alternative to usual complete case analysis. The HelixTree EM algorithm performed the best on the simulated data in contrast to the neural network method, which performed the best on the real data set. Factors such as the population structure, admixture, LD, sampling design and laboratory methods are likely to affect both the patterns of missing data and the ease of imputation. Additional studies of SNP imputation accuracy using larger scale (large number of subjects and SNPs) data sets from different study designs (eg, case–control, sib pair, pedigree and cohort) are needed as more genome-wide SNP association studies are now being conducted. Using training–testing methods within a single data set, it should be possible for investigators to assess whether data imputation is a reasonable solution to their missing data problem. In our study, the balance between low error rate and high recovery rate was achieved by combining two methods, namely the FFNN and the EM methods. In general, imputation should be approached like any other analysis procedure and should not be just implemented without evaluation and testing.

Acknowledgements

We thank Paul Green, Zhaohui Steve Qin and Paul Scheet for their insightful comments on neural networks, data simulation and fastPHASE algorithm. We also thank three anonymous reviewers for their helpful comments. This work was supported by the National Institute of Health Grants HL54457 and HL68737.

References

1 Gordon D, Ott J: Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac Symp Biocomput* 2001; **6**: 18–29.
2 Lincoln SE, Lander ES: Systematic detection of errors in genetic linkage data. *Genomics* 1992; **14**: 604–610.
3 Sobel E, Papp JC, Lange K: Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet* 2002; **70**: 496–508.
4 Efron B: Missing data, imputation, and the bootstrap. *J Am Stat Assoc* 1994; **89**: 463–478.
5 Little RJA: Regression with missing X's: a review. *J Am Stat Assoc* 1992; **87**: 1227–1237.

6 Rubin DB: Multiple imputation after 18 years. *J Am Stat Assoc* 1996; **91**: 473–489.
7 Dai JY, Ruczinski I, LeBlanc M, Kooperberg C: Imputation methods to improve inference in SNP association studies. *Genet Epidemiol* 2006; **30**: 690–702.
8 Huang J, Lin A, Narasimhan B *et al*: Tree-structured supervised learning and the genetics of hypertension. *Proc Natl Acad Sci USA* 2004; **101**: 10529–10534.
9 Roberts A, McMillan L, Wang W, Parker J, Rusyn I, Threadgill D: Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* 2007; **23**: i401–i407.
10 Bishop CM: *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press, 1995.
11 Ripley BD: *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
12 Muller P, Insua DR: Issues in Bayesian analysis of neural network models. *Neural Comput* 1998; **10**: 749–770.
13 Neal RM: *Bayesian Learning for Neural Networks*. New York: Springer, 1996.
14 Schwarz G: Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**: 461–464.
15 Kass RE, Wasserman L: A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J Am Statist Assoc* 1995; **90**: 928–934.
16 Marchini J, Cutler D, Patterson N, *et al*, International HapMap Consortium: A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 2006; **78**: 437–450.
17 Scheet P, Stephens M: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 2006; **78**: 629–644.
18 Chiano MN, Clayton DG: Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet* 1998; **62** (Part 1): 55–60.
19 Hudson RR: Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 2002; **18**: 337–338.
20 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
21 FBPP Investigators: Multi-center genetic study of hypertension: the Family Blood Pressure Program (FBPP). *Hypertension* 2002; **39**: 3–9.
22 Barkley RA, Chakravarti A, Cooper RS, *et al*, Family Blood Pressure Program: Positional identification of hypertension susceptibility genes on chromosome 2. *Hypertension* 2004; **43**: 477–482.
23 Cybenko GR: Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)* 1992; **5**: 455.
24 Hornik K, Stinchcombe M, White H: Multilayer feedforward networks are universal approximators. *Neural Networks* 1989; **2**: 359–366.
25 Bridle JS: Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing: Algorithms, Architectures and Applications* 1990; 227–236.
26 Venables WN, Ripley BD: *Modern Applied Statistics with S*. New York: Springer, 2002.
27 Raftery AE: Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* 1996; **83**: 251–266.
28 Kang SJ, Gordon D, Finch SJ: What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 2004; **26**: 132–141.
29 Pompanon F, Bonin A, Bellemain E, Taberlet P: Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* 2005; **6**: 847–859.
30 Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC: Effects of differential genotyping error rate on the type I error probability of case–control studies. *Hum Hered* 2006; **61**: 55–64.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)