JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 3, 2010

Pp. 547–560
DOI: 10.1089/cmb.2009.0199

EMINIM: An Adaptive and Memory-Efficient Algorithm for Genotype Imputation

HYUN MIN KANG,[1,2] NOAH A. ZAITLEN,[3] and ELEAZAR ESKIN [4]

ABSTRACT

Genome-wide association studies have proven to be a highly successful method for identification of genetic loci for complex phenotypes in both humans and model organisms. These large scale studies rely on the collection of hundreds of thousands of single nucleotide polymorphisms (SNPs) across the genome. Standard high-throughput genotyping technologies capture only a fraction of the total genetic variation. Recent efforts have shown that it is possible to "impute" with high accuracy the genotypes of SNPs that are not collected in the study provided that they are present in a reference data set which contains both SNPs collected in the study as well as other SNPs. We here introduce a novel HMM based technique to solve the imputation problem that addresses several shortcomings of existing methods. First, our method is adaptive which lets it estimate population genetic parameters from the data and be applied to model organisms that have very different evolutionary histories. Compared to previous methods, our method is up to ten times more accurate on model organisms such as mouse. Second, our algorithm scales in memory usage in the number of collected markers as opposed to the number of known SNPs. This issue is very relevant due to the size of the reference data sets currently being generated. We compare our method over mouse and human data sets to existing methods, and show that each has either comparable or better performance and much lower memory usage. The method is available for download at http://genetics.cs.ucla.edu/eminim.

Key words: genetic variation, genetics, genomics, statistics.

1. INTRODUCTION

RECENT ADVANCES IN HIGH-THROUGHPUT GENOTYPING technologies are helping to uncover the genetic basis of complex phenotypes in human (Wellcome Trust Case Control Consortium, 2007), mouse (Frazer et al., 2007), rat (STAR Consortium, 2008), dog (Karlsson et al., 2007), arabidopsis (Borevitz et al., 2007), and other model organisms. While the vast majority of positions in a genome are identical among individuals in a population, a significant portion of positions differ. Many of these positions are single

[1]Biostatistics Department and [2]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Ann Arbor, Michigan.
[3]Bioinformatics Program, University of California, San Diego, La Jolla, California.
[4]Departments of Computer Science and Human Genetics, University of California, Los Angeles, Los Angeles, California.

nucleotide polymorphisms (SNPs). In a typical study that attempts to identify variation involved in a trait, variation information (e.g., SNP genotypes) is collected from a set of individuals and the trait is measured in each individual. Each SNP is then correlated/associated with the trait. Any statistically significant associations are reported as possible causal variation with respect to the trait (Risch and Merikangas, 1996; de Bakker et al., 2005).

Genotyping arrays, such as those developed by Affymetrix and Illumina (Matsuzaki et al., 2004; Gunderson et al., 2005), simultaneously probe hundreds of thousands of marker SNPs in an individual's genome. While this is a significant amount of information, it is only a fraction of the millions of SNPs and other genetic variation in the population. Only complete resequencing of individual genomes will guarantee collecting all variation in a study. However, resequencing still remains prohibitively expensive. Array based genotyping is currently the most practical cost-effective method for collecting large amounts of variation information on a set of individuals. Although only a subset of individual genetic variation is collected by a genotyping array, due to the correlation structure of variation in the genome, SNPs on the array can serve as proxies for SNPs which are not collected (Devlin and Risch, 1995; Collins et al., 1998). This property is called linkage disequilibrium (LD), and it greatly extends the coverage of the array since a causal SNP need not be collected, but only strongly correlated with one of the collected markers on the array (Zaitlen et al., 2007). However, if the causal variants are not in LD with one of the SNPs included on the array then the study will not be able to discover the association. Thus, increasing the number of collected SNPs in the study increases the study's power to identify casual variation and is of fundamental importance.

Recently, several studies have proposed methods to increase the ability of a study to identify associations at SNPs which are not collected by "imputing" or predicting the genotypes of SNPs that are not contained in the study data set. These methods work by using a reference sample, such as the HapMap (International HapMap Consortium, 2007) for humans, which has genotyped millions of SNPs at great cost and effort. These reference samples contain both SNPs which are collected in the study as well as other SNPs. An imputation method uses the correlation patterns between the collected and uncollected SNPs inferred from the reference sample to make predictions of the uncollected SNPs in the study sample. This problem is effectively a missing data problem in which partial data is observed in the study and complete data is observed in the reference sample.

Consider the example shown in Figure 1, where there is a set of reference individuals shown on top and a study individual shown on the bottom. In the reference set, all the SNPs are genotyped in all five individuals. In the study individual, some of the SNPs are uncollected and denoted by "?." The goal of imputation is to resolve the genotypes of the uncollected SNPs by using the overlap of the typed SNPs between the reference set and the study set. Our method selects the most likely reference individual for

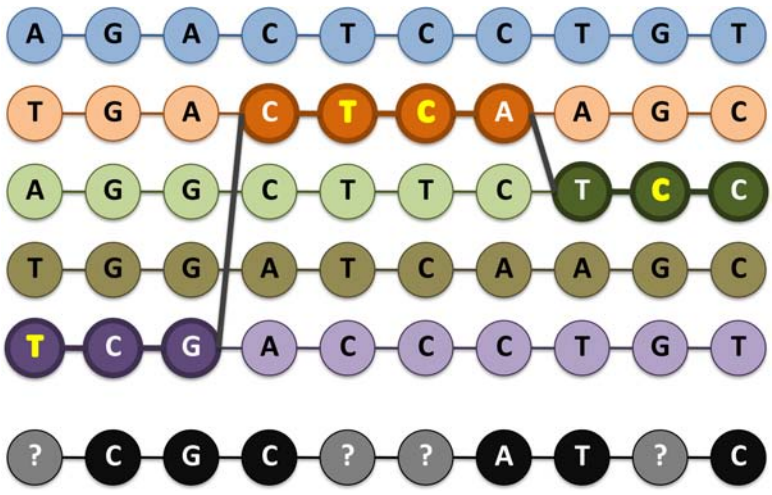

FIG. 1. An example of the imputation problem. The five reference individuals are genotyped on all ten SNPs, while the study individual is genotyped on only six SNPs. The goal of imputation is to resolve the genotypes of the uncollected SNPs.

each marker (both collected and uncollected). The path in bold shown in Figure 1 denotes that the sequence of SNP values in target individual is composed of pieces of the three reference individuals along the six collected and four uncollected markers. The actual imputation procedure is more complicated than the above example due to the diploid nature of human genome and missing genotypes.

Multiple techniques have been successfully employed to solve the imputation problem, and Hidden Markov models (HMMs) have been amongst the most popular, and have been used in several studies in both human and model organisms such as mouse. However, each of these existing HMM techniques fails to address at least one of several important problems. IMPUTE (Marchini et al., 2007) models an individual haplotype as a result of mosaic recombination and mutation among a set of reference haplotypes. The recombination probability follows a Continuous-Time Markov Process (CTMP) with respect to the genetic distance, which is essentially the same model with traditional genetic model of meiosis with different number of haploids as reference haplotype. This model is a reasonable approximation of the standard population genetics model (Marchini et al., 2007; Kingman, 1982). However, the method requires to have predefined recombination and mutation parameters, as well as the genetic distance between markers. This leads to several limitations to apply this method to a general context of imputation problem. First, the method is not useful when the genetic map in a certain population of reference individuals is not known a priori. Second, the predefined parameters do not allow individual-specific difference in the imputation procedure. In model organisms association mapping, where the genetic variations between individual strains are more dramatic than the variations in one human population, this problem is manifested more clearly. Another problem with IMPUTE is that it has memory requirements that grow linearly in the size of the reference data set, which may become prohibitively large as more SNPs are discovered. This becomes more important as the next-generation sequencing technologies discover orders of magnitude larger number of SNPs in the next few years (Kaiser, 2008).

On the other hand, MACH (Scott et al., 2007) uses a similar hidden Markov model with a very different idea from IMPUTE. Instead of imposing predefined recombination and mutation parameters like IMPUTE, MACH allows different recombination and mutation parameters per each SNP differently. These parameters are learned from the data using a Monte-Carlo Haplotyping procedure, which essentially simulates the haplotypes from the HMM and estimates the parameters iterative over multiple rounds. The advantage of this method is that it does not depend on the predefined parameters or external resources such as genetic map. However, this model does not account for the distance between markers when estimating the transition probabilities and it solely relies on the empirical correlation structure among the markers. Such an estimation procedure of transitional probabilities per marker-interval can be accurate only if an enough number of samples are imputed simultaneously. In addition, assigning the same transition and mutation parameters per each marker across all individuals makes it impossible to account for the individual-specific differences in a heterogeneous population. For inbred mouse imputation, when the number of samples is small and the population structure is highly heterogeneous, such features of MACH will substantially degenerate the accuracy of imputation.

As briefly discussed above, we are initially motivated by the problem of inbred mouse imputation that is very different from classical context of human genotype imputation problem. First, the genealogy of inbred mouse strains are complex and heterogeneous (Beck et al., 2000), so each individual mouse strain may have very different recombination and mutation parameters when the HMM is applied. For example, the wild-derived strains are genetically far apart from the classical inbred strains, and they must have substantially higher recombination and mutation probabilities than the classical inbred strains. Second, the number of inbred mouse strains are very small, so the methods that take advantages of the availability of a large number target subjects for imputation may not be applicable. Third, the inbred mouse strains have genetically identical pair of chromosomes via subsequent brother-sister mating for the sake of reproducibility, so an haploid implementation of imputation algorithm is required instead of a diploid algorithm for inbred mouse imputation. Although the haploid models are simpler than the diploid ones, there are few methods flexibly implemented to allow haploid imputation. A recently developed algorithm similar to fastPHASE (Scheet and Stephens, 2006) is specifically designed for inbred mouse imputation, employing HMMs with a predefined sets of clusters with Dirichlet prior distribution and Viterbi training algorithm (Szatkiewicz et al., 2008). The average imputation error reported was 10.4% and 4.4% for high-confidence genotypes, which is higher than the error rates in most of the human imputation studies.

In this article, we propose EMINIM (Expectation-Maximized INtegrative IMputation), an adaptive genotype imputation method that learns HMM parameters using an Expectation-Maximization (EM)

Table 1. Feature Comparisons of EMINIM with Widely Used Imputation Methods: IMPUTE and MACH

Features	*IMPUTE*	*MACH*	*EMINIM*
Provide inbred/haploid interface	X	X	O
Learn HMM parameters from data	X	O	O
Run without genetic map	X	O	O
Allow different HMM parameters per target	X	X	O
Recombination model accounts distance between markers	O	X	O
Accurate with a small number of targets	O	X	O
Memory-efficient	X	O	O

algorithm (Dempster et al., 1977) under the sophisticated recombination model on which IMPUTE is based. Our method robustly estimates the HMM parameters for each individual separately, accounting for heterogeneous sample structure. Our method utilizes various types of silent states in the HMM to estimate the EM parameters, to increase memory-efficiency, to impute genotypes at collected SNPs, and to obtain a leave-one-snp-out estimate of imputation accuracies. Our method is also more memory efficient allowing much larger data sets to be imputed. Table 1 summarizes the comparisons among EMINIM, IMPUTE, and MACH in several aspects discussed above. In this article, we describe the algorithmic detail of our method that improves accuracy and computational efficiency in addition to the core statistical model, in order to facilitate further progress in the area of genotype imputation.

We applied our method to the imputation of 8.27 million SNPs that have been discovered from a resequencing of 15 inbred mouse strains (Frazer et al., 2007), based on the 138,980 SNPs collected from the mouse HapMap project over 94 inbred strains. Imputation in mouse strains differs from human imputation because the reference datasets have drastically different linkage properties. Using several different measures, we estimated the error rate of our method and compared it to the recently published mouse imputation article (Szatkiewicz et al., 2008). Our method's overall error rate is less than half the error rate of the previous method, and for high confidence genotypes our error rate is ten time smaller (Table 2).

Next, we applied our method to the imputation of human HapMap SNPs from the Wellcome Trust Case-Control Consortium (WTCCC) genotypes. The adaptive nature of our method can help to improve the flexibility of the method and the accuracy imputation in human samples, too. Our results show that our method consistently achieves similar or better imputation accuracy over different populations than other state-of-the-art methods without requiring predefined parameters for each population. Our method also shows a significant increase of memory-efficiency which can be an important technical issue when imputing genotypes of a dense SNP sets over a large number of reference samples. For example, in a recent study, EMINIM used only 508 MB of memory to impute chromosome 22 while IMPUTE required 6.6 GB. This problem will become even more severe on larger chromosomes and data sets. Our method is publicly available at http://genetics.cs.ucla.edu/eminim.

A shorter version of this article can be found in Kang et al. (2009). This article includes many more details on the mathematical description of the model and its derivation as well as more details on the experimental results and comparisons with other methods.

Table 2. Imputation Error Rates of Inbred Strains

	Confidence cutoff		
	High (>0.98)	*High + medium (>0.8)*	*All (≥ 0)*
LOOCV with unknown reference	0.37% (81%)	0.81% (90%)	2.40% (100%)
LOOCV without unknown reference	0.52% (76%)	1.24% (93%)	2.46% (100%)
36 non-wild WTCHG strains	0.35% (89%)	1.00% (96%)	2.25% (100%)
all 47 WTCHG strains	0.37% (72%)	1.98% (89%)	4.86% (100%)

First two rows (LOOCV) use leave-one-strain-out estimation using 15 reference strains and 138,980 combined SNPs of the target strain across 12 classical inbred strains. The unknown reference strain is used in the first but not in the second. Last two rows uses WTCHG genotypes as validation set, and impute those genotypes in 47 WTCHG strains not included in the reference strains. The fraction of imputed genotypes in each category is shown within a parenthesis.

2. METHODS

The imputation problem

In this section, we formalize the problem of imputing missing genotype data in an individual using a reference population. The terms and definitions described here will be used throughout the text. We classify the imputation problem into two categories—haploid (or inbred) imputation and diploid imputation. Suppose that we genotype m SNP *markers* on an individual (*target individual*) and wish to determine the genotypes of additional *"uncollected" SNPs*. We will employ a set of *reference haplotype* that are genotyped on the m collected markers as well as an additional set of uncollected SNPs. In diploid model, we assume that each reference individuals is already phased into two reference haplotypes. The allele of the i-th reference haplotype at the t-th marker is represented as $G_{i,t} \in \{0, 1, 2\}$, where 0 represents a missing reference genotype and 1, 2 represents two alleles of biallelic SNP marker. Let n be the number of reference haplotype collected at m markers in the target individual. Let $\mathbf{d} = \{d_1, d_2, \cdots, d_{m-1}\}$ be the physical or genetic distance between consecutive markers i - 1 and i for d_i. In the target individual, the genotype at the t-th marker is represented as $g_t \in \{1, 2\}$ in haploid model, and $g_t \in \{\{1, 1\}, \{1, 2\}, \{2, 2\}\}$ in diploid model, where 1 and 2 represents two alleles of the biallelic SNP marker.

In our model, we assume that in each region, the target individual has similar haplotypes to the reference haplotypes. The goal of imputation can then be rephrased as assigning one (haploid model) or a pair of (diploid model) the n reference haplotypes to each of the m markers. From these assignments, we will assign the individual SNP genotype values to each uncollected SNP in the target individual by using the alleles of the assigned reference haplotypes at the nearby markers. We employ HMM to assign the reference haplotypes to the markers of the target individual. We first describe the detail of the HMM in the haploid case, and describe the extension to the diploid model later on.

Hidden Markov model for imputation

In the following section, we describe the algorithm for performing imputation for haploid or inbred organisms. In this case, the reference haplotypes (or individuals) are called reference strains and we assume that we do not observe any genotypes where an individual has both alleles 1 and 2 of the SNP (i.e., no heterozygous genotypes).

The goal is then to assign one of the n reference strains to each of the m markers as a mosaic combination as described in the previous section. To accomplish this, we use an HMM like that shown in Figure 2. For each of the m markers there are n states corresponding to each of the reference strains. From each state

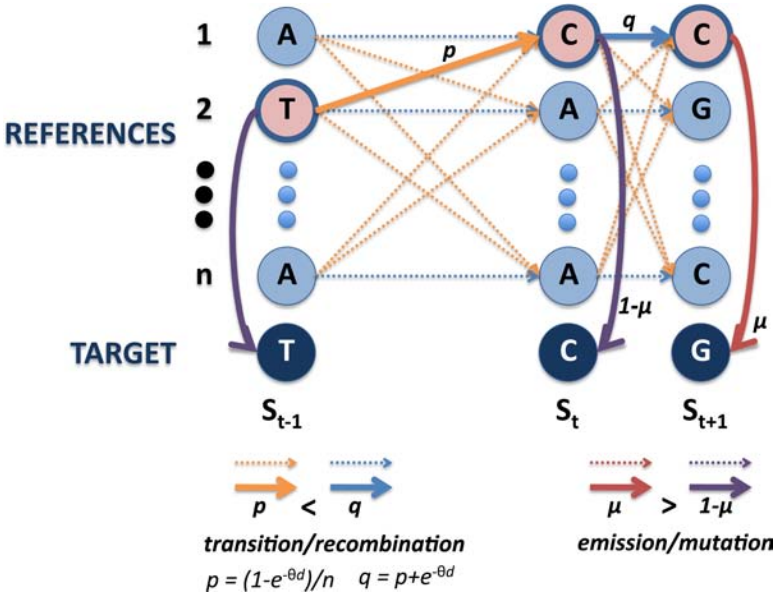

FIG. 2. An example of the hidden Markov model for the imputation problem. p and q represent transition probabilities to another state; μ and $1 - \mu$ represent the emission probability of an observed genotype given a state.

there are n edges (with the exception of the states representing the final marker) directed towards the n states for the next marker. The edges corresponding to a change in the reference strain are called transitions or a recombination. Each state can also emit one of the two possible alleles for that marker. Emitting an allele that does not match the target strain is called a mutation.

Let $S_t \in \{1, 2, \ldots, n\}$ be the reference strain assigned to the target strain at marker $t \in \{0, 1, \ldots, m-1\}$. Since some of genomic segments may not be represented by any of the reference strains, we introduce another reference strain consisting of only missing genotypes to represent the unknown reference state. We let the initial probability that marker S_0 is assigned to strain i be $\Pr(S_0 = i) = \pi_i$, with $\pi = \{\pi_1, \ldots, \pi_n\}$. Similar to many other methods designed for genotype imputation and haplotype phasing (Marchini et al., 2007; Scheet and Stephens, 2006), our HMM relies on a typical Markov chain model of population genetics based on neutral Wright-Fisher model (Kingman, 1982). We use two parameters μ, and θ to be the mutation and recombination parameters, and use the standard distributions for computing the probability of transitioning between states $p_n(x) = (1 - e^{-x})/n$ and $q_n(x) = p_n(x) + e^{-x}$ based on the continuous-time Markov process. Figure 2 shows how the these probabilities correspond to the edges in the HMM. The transition probabilities are computed from the recombination parameter and the distance between markers as follows.

$$\Pr(S_t = j | S_{t-1} = i, \theta) = \begin{cases} q_n(-\theta d_t) & i = j \\ p_n(-\theta d_t) & i \neq j \end{cases} \tag{1}$$

The probability of an observed genotype given a state is computed from the mutation parameter and the allele observed at the reference strain at the state. If the reference strain has missing genotype at the marker, then the probabilities are equally assigned between the alleles. Figure 2 shows examples of matching and mutated genotypes in the emission states of the HMM.

$$\Pr(g_t | S_t, \mu) = \begin{cases} 1 - \mu & g_t = G_{S_t, t} > 0 \\ \mu & g_t \neq G_{S_t, t} > 0 \\ 0.5 & G_{S_t, t} = 0 \end{cases} \tag{2}$$

We assume a uniform distribution of initial state probabilities, $\pi_1 = \cdots = \pi_n = 1/n$, and learn the mutation and transition parameters from the data using the EM algorithm presented below.

EM algorithm for learning maximum-likelihood parameters

Many of the previous methods suggested for HMM-based imputation of missing genotypes use either predefined transitional parameters (Marchini et al., 2007) or Viterbi training which may not converge to a local maximum (Szatkiewicz et al., 2008). Other methods estimate the transition probabilities per marker interval independently to avoid the computational complexity in constraining the transition parameters consistently across different states (Scheet and Stephens, 2006; Scott et al., 2007).

In the genotype imputation of inbred mouse strains, independent unconstrained estimation of parameters at each marker is prone to inherent bias and inaccuracy because of two reasons. First, the total number of target strains is small, so estimating parameters per marker independently may be highly inaccurate. Second, these strains have complex genetic relationship, so the transitional and mutational parameters vary greatly across different strains. We constrain the parameters to be equal over the genome, but allows different transition and mutation parameters for each strain. Instead of the simple Viterbi training algorithm, we present an EM algorithm based on the exact conditional probabilities obtained from the forward-backward algorithm.

Let us denote $X_t^- = (X_1, \ldots, X_t)$ and $X_t^+ = (X_{t+1}, \ldots, X_m)$ as observed data, and $\lambda = (\pi, \mu, \theta)$ be the initial, mutational, and transitional parameters of the hidden Markov model. The forward-backward algorithm estimates $\alpha_t(i) = \Pr(X_t^-, S_t = i | \lambda)$ and $\beta_t(i) = \Pr(X_t^+ | S_t = i, \lambda)$ using a dynamic programming.

$$\alpha_t(S_t) = \sum_{S_{t-1}} \alpha_{t-1} S_{t-1} \Pr(S_t | S_{t-1}) \Pr(X_t | S_t) \tag{3}$$

$$\beta_t(S_t) = \sum_{S_{t+1}} \beta_{t+1} S_{t+1} \Pr(S_{t+1} | S_t) \Pr(X_{t+1} | S_{t+1}) \tag{4}$$

Let $X = (g, G)$. The EM algorithm starts with initial parameters (μ_0, θ_0). At the E-step of r-th iteration, $\Pr(S_t | X, \lambda_r)$ are computed from the forward-backward algorithm. Let $S = \{S_0, \cdots, S_{m-1}\}$. At the M-step, the expected likelihood function can be written as follows.

$$Q(\mu,\theta) = \sum_{S} \Pr(S|X,\lambda_r)\log\Pr(g,S|\lambda) \tag{5}$$
$$= \sum_{t=1}^{m-1}\sum_{(S_t,S_{t-1})} \log\Pr(S_t|S_{t-1},\theta)\Pr(S_t,S_{t-1}|X,\lambda_r)$$
$$+ \sum_{t=0}^{m-1}\sum_{S_t} \log\Pr(g_t|S_t,\mu)\Pr(S_t|X,\lambda_r) - \log n$$

The expectation-maximized parameters used for the next round of E-step can be obtained as follows.

$$\mu_{r+1} = \frac{\sum_{t=1}^{m}\sum_{S_t} I(g_t \neq G_{i,t})\Pr(S_t|X,\lambda_r)}{\sum_{t=1}^{m}\sum_{S_t} I(G_{i,t}>0)\Pr(S_t|X,\lambda_r)} \tag{6}$$

$$\theta_{r+1} = \arg\max_{\theta}\left[\sum_{t=1}^{m-1}\sum_{(S_t,S_{t-1})} \log\Pr(S_t|S_{t-1},\theta)\Pr(S_t,S_{t-1}|X,\lambda_r)\right] \tag{7}$$

$$= \arg\max F_h(\theta) \tag{8}$$

In order to estimate the joint probability $\Pr(S_t,S_{t-1}|X,\lambda_r)$, we introduce a silent state J_t between S_{t-1} and S_t with the following transition probabilities which keeps $\Pr(S_t|S_{t-1})$ unchanged.

$$\Pr(J_t=(i,b)|S_{t-1}=i,\theta) = \begin{cases} q_n(-\theta d_t) & b=0 \\ (n-1)p_n(-\theta d_t) & b=1 \end{cases} \tag{9}$$

$$\Pr(S_t=j|J_t=(i,b),\theta) = \begin{cases} 1 & b=0,\ i=j \\ 1/(n-1) & b=1,\ i\neq j \end{cases} \tag{10}$$

The probabilities of the other transitions are set to zero. As illustrated in Figure 3, such a procedure temporarily creates $2n$ silent states between a pair of consecutive states. When a recombination occurs, the original state i makes a transition to a different state via the silent state $(i,1)$. If no recombination occurs, it remains at the same state via the silent state $(i,0)$. The model of introducing a intermediate state is somewhat similar to Scheet et al. (2006), but their EM algorithm is more trivial due to the independence assumption between the recombination parameters across markers. The idea of constraining the parameters across markers are only briefly mentioned, and here we present an non-trivial algorithmic detail to perform EM parameter learning at a linear time complexity with respect to the number of states using silent states.

In general, the marginal probabilities of these silent states can be computed, in the following way. Let L be a silent state connecting S_{t-1} and S_t satisfying the following condition.

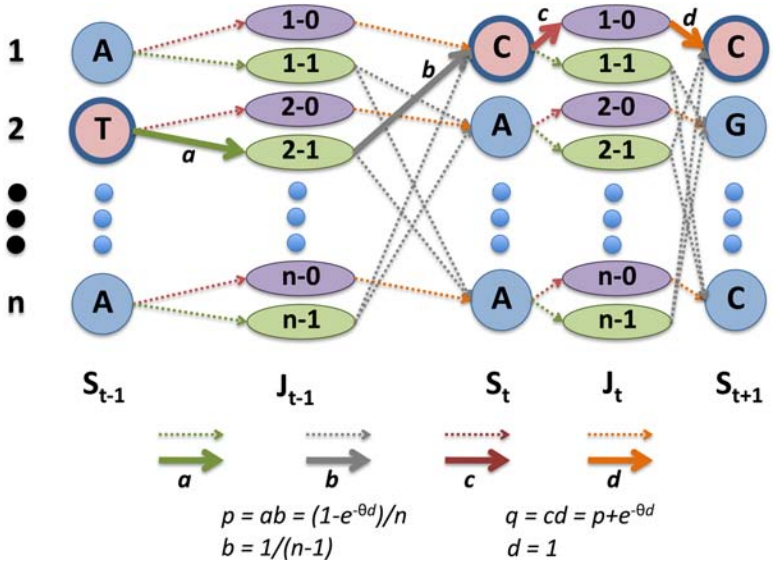

FIG. 3. Adding a silent state for EM parameter estimation. p and q represent original transition probabilities; *a, b, c*, and *d* represent the transition probabilities of modified HMM after adding silent states. The emission probabilities remain the same.

$$\Pr(S_t|S_{t-1}, \lambda) = \sum_L \Pr(S_t|L, \lambda) \Pr(L|S_{t-1}, \lambda) \tag{11}$$

Then the transition probability between marker $t-1$ and t will remain unchanged, and the results of forward-backward algorithm can still be used because L do not emit any variables. The forward and backward probability of any silent state L are computed as

$$\alpha(L) = \sum_{S_{t-1}} \alpha_{t-1}(S_{t-1}) \Pr(L|S_{t-1}) \tag{12}$$

$$\beta(L) = \sum_{S_t} \beta_t(j) \Pr(S_t|L) \Pr(X_t|S_t) \tag{13}$$

The computational complexity of forward and backward probability can be reduced to a constant time as described in the upcoming section, so this procedure does not increase the time complexity of the inference in HMM. Then the computation of posterior marginal probability of any silent state L is trivial, and thus the marginal probability of $\Pr(J_t|X,\lambda_r)$ can be computed. The objective function of M-step transitional parameter becomes

$$F_h(\theta) = \sum_{t=1}^{m-1} \left[\log q_n(-\theta d_t) \sum_{i=1}^{n} \Pr(J_t = (i,0)|X, \lambda_r) + \log p_n(-\theta d_t) \sum_{i=1}^{n} \Pr(J_t = (i,1)|X, \lambda_r) \right] \tag{14}$$

This function can be numerically optimized using a Newton-Raphson algorithm. Although the complex function $F(\theta)$ may not be convex in general, we were able to find a consistent estimator of θ_{r+1} with various initial parameters in the all experiment described in the results section.

Imputation of uncollected genotypes

Let h_t be the number of "uncollected SNPs" that need to be imputed between marker $(t-1)$ and t. An uncollected SNP is represented as (t,s), where $t \in \{1, 2, \ldots, m-1\}$ and $s \in \{1, \ldots, h_t\}$. Let $T_{t,s} \in \{1, \ldots, n\}$ be the state at an uncollected SNP (t,s).

We again modify the HMM by adding a silent state $T_{t,s}$ to the original HMM. The transition probabilities between S_{t-1},$T_{(t,s)}$, and S_t is defined in the same way to to Equation 1 based on the distance between them. Let $H_{i,t,s} \in \{0,1,2\}$ be the genotypes of i-th reference strain at the uncollected SNP (t,s). Then distribution of the imputed genotype $z_{t,s}$ at the uncollected SNP is estimated as follows.

$$\Pr(z_{t,s}|X, \lambda) = \begin{cases} (1-\mu) \sum_{i=1}^{n} I(H_{i,t,s} = z_{t,s}) p_{t,s}(i) & \\ \quad + \mu \sum_{i=1}^{n} I(H_{i,t,s} \neq z_{t,s}) p_{t,s}(i) & z_{t,s} > 0 \\ \sum_{i=1}^{n} I(H_{i,t,s} = 0) p_{t,s}(i) & z_{t,s} = 0 \end{cases} \tag{15}$$

where $p_{t,s}(i)$ denotes the marginal probability of state i at the uncollected SNP (t,s). When estimating leave-one-snp-out imputation accuracy, the same imputation methods are applied by pretending marker t has a silent state by ignoring the observed genotype.

Improving computation time complexity

A standard HMM implementation requires squared time complexity with respect to the number of individuals when computing the forward and backward probabilities. It is possible to reduce the time complexity to be linear in the number of states, by leveraging the fact that the transition probabilities are uniform over different states. When $t > 1$, $\alpha_t(i)$ and $\beta_t(i)$ follows that

$$\alpha_t(i) = \left[\exp(-\theta d_t) \alpha_{t-1}(i) + p_n(-\theta d_t) \sum_{x=1}^{n} \alpha_{t-1}(x) \right] \Pr(X_t|S_t = i) \tag{16}$$

$$\beta_t(i) = \exp(-\theta d_{t+1}) \beta_{t+1}(i) \Pr(X_{t+1}|S_{t+1} = i) + p_n(-\theta d_{t+1}) \sum_{x=1}^{n} \beta_{t+1}(x) \Pr(X_{t+1}|S_{t+1} = x) \tag{17}$$

Each $\alpha_t(i)$ and $\beta_t(i)$ can be computed in a constant time if we precomputed $\sum_{x=1}^{n} \alpha_{t-1}(x)$ and $\sum_{x=1}^{n} \beta_{t+1}(x) \Pr(X_{t+1}|S_{t+1} = x)$, which takes linear time in the number of states. So the computation of

$\alpha_t(i)$ and $\beta_t(i)$ over all states can be performed in linear time. Such a reduction in the time complexity applies the same way in the computation of forward and backward probabilities of silent states presented in Equations 12 and 13.

When computing the forward and backward probabilities, precision problems will occur when the number of markers are large since $\alpha_t(i)$ and $\beta_t(i)$ can have extremely small values. Storing the values in a logarithmic scale would resolve this problem, but the computational cost will substantially increase due to the repetitive execution of exponential and logarithmic functions. Since the forward and backward probabilities are meaningful only relatively because they will have to be normalized in order to obtain posterior marginal probabilities or joint probabilities. Since $\sum_{x=1}^{n} \alpha_{t-1}(x)$ and $\sum_{x=1}^{n} \beta_{t+1}(x) \Pr(X_{t+1}|S_{t+1}=x)$ are precomputed at each step, we can avoid the precision problem by maintaining the normalized values divided by those summation instead. The HMM procedure would not be affected by such a normalization procedure.

Extension to diploid model

When imputing human genotype data, the reference individuals are typically provided as phased haplotypes across a dense set of SNPs, and a number of unphased genotypes of a target individual are provided as a subset of SNPs. In this case, the state at each collected marker $Z_t = (i, j)$ represents the combined states of each haplotype. Here we assume there are no missing alleles in the reference haplotypes because they are phased. However, missing alleles can also be handled in a similar way presented as in the haploid model. Their initial state probabilities are defined as $\Pr(Z_0 = (i,j)) = \pi_{ij} = 1/n^2$, and the transition probabilities are defined as follows.

$$\Pr(Z_t = (i, j)|Z_{t-1} = (k, l), \theta) = \begin{cases} q_n(-\theta d_t)^2 & i = k,\ j = l \\ p_n(-\theta d_t) q_n(-\theta d_t) & i = k \ \oplus\ j = l \\ p_n(-\theta d_t)^2 & i \neq k,\ j \neq l \end{cases} \tag{18}$$

where $\oplus$ denotes exclusive OR operator. Let $Z_t = (Z_{t,0}, Z_{t,1})$ be the individual states of each chromosome, then the imputed genotype $(g_{t,0}|Z_{t,0}, G, g_{t,1}|Z_{t,1}, G)$ independently follows the mutational distribution in the inbred imputation.

The observed genotype $g_t = \{g_{t,0}, g_{t,1}\} \in \{\{1,1\}, \{1,2\}, \{2,2\}\}$ represents one of homozygous base alleles, heterozygous alleles, or homozygous mutant alleles. Based on these probability models, an HMM can be constructed with n^2 states for each collected marker.

Let b_t be the number of state changes between marker $t-1$ and t. In order to estimate EM parameters, we introduce a silent state J_t connecting Z_{t-1} and Z_t, with the following transition probabilities.

$$\Pr(J_t = (k, l, b_t)|Z_{t-1} = (i,j)) = \begin{cases} q_n(-\theta d_t)^2 & k = i,\ l = j,\ b_t = 0 \\ 2(n-1) p_n(-\theta d_t) q_n(-\theta d_t) & k = i,\ l = j,\ b_t = 1 \\ (n-1)^2 p_n(-\theta d_t)^2 & k = i,\ l = j,\ b_t = 2 \end{cases} \tag{19}$$

$$\Pr(Z_t = (i,j)|J_t = (k, l, b)) = \begin{cases} 1 & k = i,\ l = j,\ b_t = 0 \\ 1/(2n-2) & (k = i \oplus l = j),\ b_t = 1 \\ 1/(n-1)^2 & k \neq i,\ l \neq j,\ b_t = 2 \end{cases} \tag{20}$$

From the expectation maximized parameters in the M-step it follows that

$$\mu_{r+1} = \frac{1}{m} \sum_{t=0}^{m-1} \sum_{Z_t} \eta(g_t, G_{Z_{t,0},t}, G_{Z_{t,1},t}) \Pr(Z_t|X, \lambda_r) \tag{21}$$

$$\theta_{r+1} = \arg\max_{\theta} \left[\sum_{t=1}^{m-1} \sum_{(Z_t, Z_{t-1})} \log \Pr(Z_t|Z_{t-1}, \theta) \Pr(Z_t, Z_{t-1}|X, \lambda_r) \right] \tag{22}$$

$$= \arg\max_{\theta} F_d(\theta) \tag{23}$$

where $\eta(g, h_1, h_2)$ is the number of mismatched alleles between genotype g and $\{h_1, h_2\}$. $F_d(\theta)$ can be rewritten to be numerically optimized using a Newton-Raphson algorithm as follows.

$$\sum_{t=1}^{m}[2\log q_n(-\theta d_t)\Pr(b_t=0|X,\lambda_r)+\log(p_n(-\theta d_t)q_n(-\theta d_t))\Pr(b_t=1|X,\lambda_r) + 2\log(p_n(-\theta d_t))\Pr(b_t=2|X,\lambda_r)] \tag{24}$$

Similar to the haploid model, each uncollected genotype is imputed without increasing memory by adding a silent state. For example, $\alpha_t(Z_t=(i,j))$ can be computed in a linear time in the number of possible states as follows:

$$\alpha_t(i,j)=\left[\exp(-\theta d_t)^2\alpha_{t-1}(i,j)+\exp(-\theta d_t)p_n(-\theta d_t)\sum_{x=1}^{n}(\alpha_{t-1}(i,x)+\alpha_{t-1}(x,j)) + p_n(-\theta d_t)^2\sum_{x=1}^{n}\sum_{y=1}^{n}\alpha_{t-1}(x,y)\right]\Pr(X_t|Z_t=(i,j)) \tag{25}$$

$$\begin{aligned}\beta_t(i,j)=&\exp(-\theta d_{t+1})^2\beta_{t+1}(i,j)\Pr(X_{t+1}|Z_{t+1}=(i,j))\\ &+\exp(-\theta d_{t+1})p_n(-\theta d_{t+1})\sum_{x=1}^{n}\beta_{t+1}(i,x)\Pr(X_{t+1}|Z_{t+1}=(i,x))\\ &+\exp(-\theta d_{t+1})p_n(-\theta d_{t+1})\sum_{x=1}^{n}\beta_{t+1}(x,j)\Pr(X_{t+1}|Z_{t+1}=(x,j))\\ &+p_n(-\theta d_t)^2\sum_{x=1}^{n}\sum_{y=1}^{n}\beta_{t+1}(x,y)\Pr(X_{t+1}|Z_t=(x,y))\end{aligned} \tag{26}$$

In order to ensure linear time complexity overall, the following values needs to be precomputed to avoid redundancy:

$$\sum_{x=1}^{n}\alpha_t(i,x)=\sum_{x=1}^{n}\alpha_t(x,i) \tag{27}$$

$$\sum_{x=1}^{n}\sum_{y=1}^{n}\alpha_t(x,y) \tag{28}$$

$$\sum_{x=1}^{n}\beta_t(i,x)\Pr(X_t|Z_t=(i,x))=\sum_{x=1}^{n}\beta_t(x,i)\Pr(X_t|Z_t=(x,i)) \tag{29}$$

$$\sum_{x=1}^{n}\sum_{y=1}^{n}\beta_t(x,y)\Pr(X_t|Z_t=(x,y)) \tag{30}$$

Genotype resources

We used 138,980 mouse HapMap SNPs which combine 121,433 Broad SNPs over 94 strains, 7,570 NIEHS/Perlegen gap-filling SNPs over 40 strains, and 13,094 Wellcome Trust SNPs collected in 67 strains among the 94 strains. We imputed the genotypes of 8,223,225 NIEHS/Perlegen SNPs successfully mapped to NCBI mouse build 37 from the genotypes of 15 resequenced strains in addition to the sequenced reference strain C57BL/6J. For the imputation of human genotypes, we used the chromosome 22 data in the WTCCC control set of 1,367 individuals, which consist of 4,367 distinct collected markers. A total of 42,101 polymorphic HapMap SNPs in chromosome 22 are imputed from the phased haplotypes of 60 CEU unrelated samples. We use the genetic map estimated from HapMap CEU population to model transitional probabilities.

3. RESULTS

Genotype imputation of 94 inbred mouse strains

A recent NIEHS/Perlegen mouse resequencing project identified 8.27 million SNPs among 16 inbred mouse strains (Frazer et al., 2007). The Broad mouse HapMap project collected genotypes over 94 strains

at 138,980 SNPs, which is only 1.7% of the number of SNPs identified in the resequencing project. We can achieve high imputation accuracy even with such a small fraction of the SNPs because of the very long regions of linkage disequilibrium.

We evaluated the accuracy of our genotype imputation method through leave-one-out analysis. For each of the 16 resequenced strains, we ran our EMINIM algorithm to impute the genotypes at NIEHS/Perlegen SNPs using the mouse HapMap genotypes and the NIEHS/Perlegen SNPs for the other 15 strains. Singleton SNPs polymorphic only in the target strain were removed in the evaluation of accuracy since they are not able to be imputed using the rest of strains. The leave-one-strain-out validation provides a conservative estimate of the genome wide imputation accuracy of a unresequenced strain using 16 resequenced strains.

The overall average imputation error over 12 classical strains is 2.40%. We classified the imputed genotypes into the "high-confidence" category if the posterior probability is >0.98, and "medium-confidence" for 0.8–0.98. When considering only high-confidence imputed genotypes after discarding 18.9% of low and medium confidence genotypes, the average imputation error significantly reduces to 0.37%. When including wild-derived strains, the imputation error significantly increases. The average imputation error between four wild-derived strains was 19.2%, each of them ranging from 13.0% (WSB/EiJ) to 34.0% (CAST/EiJ). None of the wild-derived strains have high-confidence imputed genotypes due to high estimates of mutation rates.

Unlike previous imputation methods based on hidden Markov models, we introduce an additional state to account for genomic regions that are not explained by any of the reference strain. We compared this model to one without an additional state, by computing imputation accuracy using leave-one-strain-out cross-validation. The results over 12 classical inbred strains show that the overall imputation error increased from 2.40% to 2.46%. More notably, the average imputation errors in high confidence category increased from 0.37% to 0.52%, and the coverage of high-confidence category reduced from 81.1% to 75.7%. This suggests that our model with additional state for unknown reference strains significantly affects the imputation accuracy probably because some genomic segments are not well characterized by any of the 16 reference strains.

Next, we evaluated the imputation accuracy by comparing the genotypes typed for 78 non-resequenced strains. We used the Wellcome Trust genotypes as a validation set and evaluated how accurately our method can impute the genotypes in the validation set using the 16 resequenced strains as reference strains. Sixty-two strains out of 94 strains were genotyped by Wellcome Trust, and 47 of them were not included in the 16 reference strains. A total of 493,033 genotypes in the validation set were evaluated for imputation accuracy, and the overall imputation error was 4.86%. 353,704 (71.7%) genotypes fall into high-confidence genotypes, and the imputation errors on these high-confidence genotypes are 0.37%. Figure 4 demonstrates that our imputation errors for high-confidence category is more than ten times smaller than the recently published results which used a different imputation method at a similar level of call-rate (Szatkiewicz et al.,

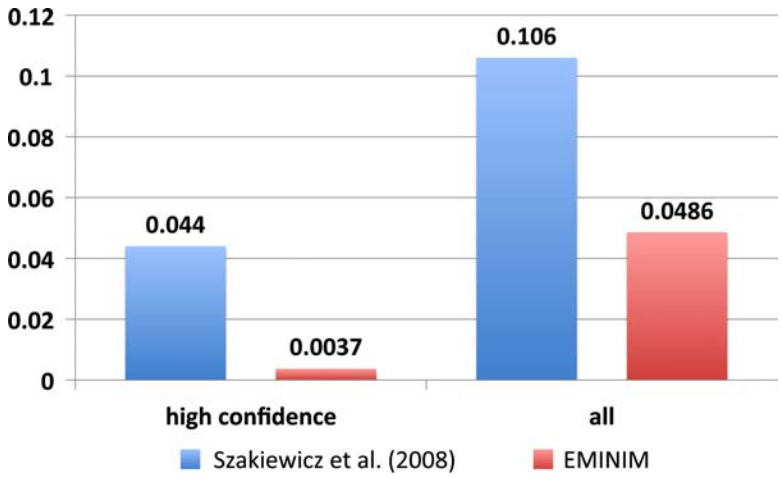

FIG. 4. Comparison of imputation accuracies of classical inbred mouse strains between Szatkiewicz et al. (2008) (blue) and EMINIM (red).

2008). Their imputation errors at high-confidence genotypes were reported to be 4.4% with 69.5% call rate. When excluding eleven wild-derived strains the average error reduced to 2.26%, which is slightly lower than what we observed in 12 classical inbred strains with leave-one-strain-out cross-validation. Among the rest of 36 non-wild strains, 344,747 (88.9%) genotypes out of 387,817 fall into the high-confidence category with an average imputation error of 0.35%, suggesting a high coverage of the mouse HapMap SNP sets with high imputation accuracy.

Imputation of HapMap SNPs in WTCCC samples

We applied EMINIM to impute the uncollected HapMap SNPs of the 1,376 WTCCC control samples. We compared the imputation accuracy and memory efficiency with other published methods to demonstrate the robustness of our method. Our evaluation of chromosome 22 can be extrapolated to the estimate the performance of each method on a genome-wide scale.

First, we evaluated the accuracy of our method by randomly choosing 25% of SNPs out of the collected SNPs and imputing them from the rest of the collected SNPs. We varied the initial HMM transitional parameters of each method and observed the changes of the imputation accuracy to compare the adaptivity of the methods against the bias of the initial parameter. While IMPUTE shows a considerable change of imputation accuracy based on the transitional parameters, EMINIM shows almost the same accuracy regardless of the initial values of HMM parameters, because the optimal parameters are learned from the genotype and haplotype data using EM algorithm. The accuracy table consistently shows that our method has a higher accuracy than the previous methods (Table 3). The imputation accuracy of MACH was outperformed by EMINIM and IMPUTE. Since MACH does not use genetic map as input, we ran EMINIM using physical map instead of genetic map to compare their performance in the absence of genetic map, and EMINIM still showed a higher accuracy. To ensure comparable computational time, MACH is run only for 10 rounds. However, if the number of rounds substantially increases, the imputation accuracy of MACH outperforms EMINIM with physical map because the availability of a large number of target individuals provides advantages to MACH to estimate the transition probability specific to each marker quite accurately.

Next, we compared the memory efficiency between different algorithms. Since each imputation method requires a significantly large amount of memory to impute a large genomic region, the memory efficiency is an important issue when practically using the methods. The methods requiring too large amount of memory need to partition the chromosome into smaller segments at the expense of increased error rates and redundant computation. While IMPUTE requires 6.6GB of memory space, to impute all the polymorphic HapMap SNPs in chromosome 22, EMINIM requires only 508MB of memory, and MACH used 502MB of memory. This is mainly due to the fact that IMPUTE consumes memory space for each uncollected SNP while EMINIM requires memory space only for collected SNPs using a silent state when imputing each uncollected SNP. Such a difference may be substantial in a larger chromosome such as chromosome

Table 3. Estimated Error Rates of Each Imputation Method with Different Transition Parameters across Different Confidence Cutoffs

	Confidence cutoff			
	>0.9	*>0.8*	*>0.7*	*All*
EMINIM ($N_e^0 = 11{,}400$)	1.23% (81%)	2.09% (87%)	3.07% (91%)	6.57% (100%)
EMINIM ($N_e^0 = 1{,}140$)	1.23% (81%)	2.09% (87%)	3.07% (91%)	6.57% (100%)
EMINIM ($N_e^0 = 114$)	1.23% (81%)	2.09% (87%)	3.07% (91%)	6.57% (100%)
IMPUTE ($N_e = 11{,}400$)	1.35% (81%)	2.25% (87%)	3.21% (91%)	6.61% (100%)
IMPUTE ($N_e = 1{,}140$)	2.79% (87%)	4.00% (91%)	5.04% (94%)	7.52% (100%)
IMPUTE ($N_e = 114$)	3.97% (88%)	5.19% (92%)	6.16% (95%)	8.23% (100%)
EMINIM (physical map)	1.50% (79%)	2.62% (86%)	3.78% (91%)	7.29% (100%)
MACH	N/A	N/A	N/A	7.69% (100%)

$N_e = 11{,}400$ is suggested by Marchini et al. (2007), corresponding $\theta = 3.8$, and different parameters are applied to demonstrate the effect of the initial parameters. The values in parenthesis represents the fraction of imputed genotypes with confidence above the threshold. Note that MACH provides only expected dosage of genotypes, so per-confidence cutoff comparisons are not performed directly.

1 which has more than five times as many SNPs as chromosome 22. In this case, EMINIM is expected to use 2.5GB of memory while IMPUTE may require 33GB of memory space. Such a difference may be more crucial as the number of reference samples increases. The overall CPU time of EMINIM was 4.7 hours with Intel Xeon E5320 Processor, which was faster than IMPUTE 0.5.0 (6.5 hours) and MACH 1.0 (7.2 hours with only 10 rounds), despite the fact that EMINIM runs HMM multiple times per individuals to estimate the EM parameters.

4. DISCUSSION

We have proposed an adaptive and memory efficient imputation method EMINIM. Our method adaptively learns HMM parameters using an EM algorithm. As a result, both in the human and inbred mouse strain imputation problems, our method is shown to outperform previous imputation methods specifically designed for each organism. In addition, the memory requirement of our method is independent of the number of uncollected SNPs by utilizing silent states, which significantly increase the scalability and computational efficiency of our method to genome-wide imputation.

There are several imputation algorithms that are not directly compared against our method. In the inbred model organism mapping, Roberts et al. (2007) proposed a nearest-neighbor search algorithms over sliding windows mainly to fill-in the missing information in the SNPs. This method can be also used for genotype imputation. Their simulation results suggest that the method may achieve a comparable imputation accuracy to the fastPHASE. However, the main drawback of the nearest-neighboring method is that they do not provide a posterior probability of the imputed genotypes. Accounting for the probabilities can be used for quality control of imputed genotypes, and for improving the power of association-based statistics (Marchini et al., 2007).

Other HMMs for genotype imputation algorithms (Browning and Browning, 2009; Guan and Stephens, 2008) use similar heuristic approaches with MACH (Scott et al., 2007), relying on the availability of large imputation target samples enabling us to estimate the parameters per marker differently. Such advantages of marker-specific parameter estimation are orthogonal to the advantages that our method that allows individual-specific parameter estimation. Our method can be combined with one of these heuristics to account for different types of heterogeneity presents in the data as sample structure and recombination hotspots.

ACKNOWLEDGMENTS

H.M.K., N.A.Z., and E.E. are supported by the National Science Foundation (grants 0513612, 0731455, and 0729049) and the National Institutes of Health (grant 1K25HL080079). N.A.Z is supported by the Microsoft Research Fellowship. H.M.K is supported by the Samsung Scholarship. This research was supported in part by the UCLA subcontract of contract N01-ES-45530 from the National Toxicology Program/National Institute of Environmental Health Sciences to Perlegen Sciences.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Beck, J.A., Lloyd, S., Hafezparast, M., et al. 2000. Genealogies of mouse inbred strains. *Nat. Genet.* 24, 23–25.

Borevitz, J.O., Hazen, S.P., Michael, T.P., et al. 2007. Genome-wide patterns of single-feature polymorphism in arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* 104, 12057–12062.

Browning, B.L., and Browning, S.R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223.

Collins, F.S., Brooks, L.D., and Chakravarti, A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* 8, 1229–1231.

de Bakker, P.I.W., Yelensky, R., Pe'er, I., et al. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* 37, 1217–1223.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. R. Statist. Soc. Ser. B Methodol.* 1–38.

Devlin, B., and Risch, N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311–322.

Frazer, K.A., Eskin, E., Kang, H.M., et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* 448, 1050–1053.

Guan, Y., and Stephens, M. 2008. Practical issues in imputation-based association mapping. *PLoS Genet.* 4, e1000279.

Gunderson, K.L., Steemers, F.J., Lee, G., et al. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* 37, 549–554.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861.

Kaiser, J. 2008. DNA sequencing. a plan to capture human diversity in 1000 genomes. *Science* 319, 395.

Kang, H.M., Zaitlen, N.A., Han, B., et al. 2009. An adaptive and memory efficient algorithm for genotype imputation. *Lect. Notes Comput. Sci.* 482–495.

Karlsson, E.K., Baranowska, I., Wade, C.M., et al. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* 39, 1321–1328.

Kingman, J.F.C. 1982. On the genealogy of large populations. *J. Appl. Probabil.* 27–43.

Marchini, J., Howie, B., Myers, S., et al. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913.

Matsuzaki, H., Dong, S., Loi, H., et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* 1, 109–111.

Risch, N., and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.

Roberts, A., McMillan, L., Wang, W., et al. 2007. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* 23, i401–i407.

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

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., et al. 2007. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science* 316, 1341–1345.

STAR Consortium. 2008. SNP and haplotype mapping for genetic analysis in the rat. *Nat. Genet.* 40, 560–566.

Szatkiewicz, J.P., Beane, G.L., Ding, Y., et al. 2008. An imputed genotype resource for the laboratory mouse. *Mamm. Genome* 19, 199–208.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678.

Zaitlen, N., Kang, H.M., Eskin, E., et al. 2007. Leveraging the hapmap correlation structure in association studies. *Am. J. Hum. Genet.* 80, 683–691.

Address correspondence to:
Dr. Eleazar Eskin
Department of Computer Science
UCLA
Los Angeles, CA 90095

E-mail: eeskin@cs.ucla.edu