Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 813912, 7 pages
http://dx.doi.org/10.1155/2013/813912

Research Article

Simpute: An Efficient Solution for Dense Genotypic Data

Yen-Jen Lin,[1] Chun-Tien Chang,[1] Chuan Yi Tang,[1,2] and Wen-Ping Hsieh[3]

[1] *Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan*
[2] *Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan*
[3] *Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan*

Correspondence should be addressed to Chuan Yi Tang; cytang@pu.edu.tw and Wen-Ping Hsieh; wphsieh@stat.nthu.edu.tw

Received 27 November 2012; Accepted 4 January 2013

Academic Editor: Hao-Teng Chang

Single nucleotide polymorphism (SNP) data derived from array-based technology or massive parallel sequencing are often flawed with missing data. Missing SNPs can bias the results of association analyses. To maximize information usage, imputation is often adopted to compensate for the missing data by filling in the most probable values. To better understand the available tools for this purpose, we compare the imputation performances among BEAGLE, IMPUTE, BIMBAM, SNPMStat, MACH, and PLINK with data generated by randomly masking the genotype data from the International HapMap Phase III project. In addition, we propose a new algorithm called simple imputation (Simpute) that benefits from the high resolution of the SNPs in the array platform. Simpute does not require any reference data. The best feature of Simpute is its computational efficiency with complexity of order $(mw + n)$, where n is the number of missing SNPs, w is the number of the positions of the missing SNPs, and m is the number of people considered. Simpute is suitable for regular screening of the large-scale SNP genotyping particularly when the sample size is large, and efficiency is a major concern in the analysis.

1. Background

A single nucleotide polymorphism (SNP) is a genetic variation at a single base-pair position. It is acquired and retained in the population. Most SNPs produce no observable difference between members of a species. These variations in the DNA can occur on both coding and noncoding sequences at a frequency of approximately 1 per 1000 base pairs [1, 2]. This leads to a rate of an estimated 11 million loci that can vary in approximately 0.1% of the population according to neutral theory of population genetics [3].

Studies concerning genetic association examine genetic traits shared among individuals in a population. SNPs have an important role in these studies because they record the history of recombination and are sufficiently dense to form linkage disequilibrium (LD) in nearly all functional genes. However, it is common for data to be missing on the various genotyping platforms. Even for array technology, the rate of missing data can be as high as 0.53% [4]. This is approximately 5300 loci for every million SNPs designed on the arrays. Assuming a random missing mechanism exists, if any locus in a sample is removed, the missing rate can become as high as $1 - (1 - 0.0053)^n$ in an association study of n samples.

Because it is often not financially viable to regenotype the missing data, imputation is used to fill in the missing SNP values, and to maintain low costs. Imputation can be as simple as selecting at random a genotype that already exists in the data or by using a major allele. However, such naive methods normally result in high error rates [4]. Certain other methods are based on haplotypes, which are sets of SNPs that are associated on one chromosome pair. These methods include the Hidden Markov Model (HMM), Markov chain (MC), maximum-likelihood, and neural network. Because a multitude of methodologies exists that can be employed to impute a haplotype, a range of imputation software, consequently, also exists. Examples of imputation software include IMPUTE [5], MaCH [6], SNPMSTAT [7], fastPhase [8], and BEAGLE [9].

Both the IMPUTE and BEAGLE software use the HMM. The HMM is a statistical tool for modelling generative sequences, which are characterised by the use of an underlying process to generate an observable sequence. In the HMM these underlying processes are represented by states, which are considered to be unobserved or hidden. The hidden state used is a pair of haplotypes observed in reference samples from the HapMap project. The observed data are the individual genotypes at the corresponding loci. IMPUTE considers mutation and recombination in its HMM model; this requires additional information from CHIAMO [10] and HAPGEN [6, 10, 11] to determine the probability of the genotypes estimated from the arrays and the predicted haplotypes. MaCH uses a Markov chain-based approach using samples from HapMap as references. Long missing segments are compensated for in MaCH by using haplotypes from the reference samples.

Alternative imputation software and methodologies include SNPMSTAT and FFNN. SNPMSTAT uses a maximum-likelihood framework on the genotype data. It uses HapMap data or other similar data sets to construct the most-likely haplotypes to occur for a missing SNP value. The feed-forward neural networks (FFNNs) proposed by Sun and Kardia [12] were reported to perform well by using a Bayesian criterion to select the predictors. They claimed that the performance is better than that of fastPHASE [8] and the LD-based method, which is used by HelixTree [10].

In this paper, we propose an algorithm based on observed genotypes and the LD at three neighbouring SNPs, including the SNP under consideration, to impute the missing SNPs, and to reduce the error rate for estimation. This algorithm only considers the two neighbouring SNPs and uses the haplotype information, which is a direct consequence of LD. Jung et al. used the same level of information in their proposed method [4], which phased genotypes by the partition-ligation expectation maximization (PLEM) [11] to impute the missing SNPs. We compare the results using SNPs from the same chromosomal regions in Jung's study and demonstrate the better performance of our algorithm. We also compare the general-purpose methods including BIMBAM v0.99, BEAGLE v3.0.3, IMPUTE v0.5.0, MARCH v1.0.16, PLINK, and SNPMSTAT v3.1. Because Simpute provides the best power at highly linked loci, we compare it to the best method using SNPs with strong LD. We demonstrate that Simpute is a promising tool to provide efficient computation when it comes to the age of massive parallel sequencing.

2. Methods

SNPs could be bi-, tri-, or tetraallelic polymorphisms by definition, but triallelic and tetraallelic SNPs rarely exist in the human population. SNPs are usually considered biallelic, and three genotypes are possible for each SNP locus. They are coded as 0 (homozygous for the wild type), 2 (heterozygote), and 1 (homozygous for mutants) in this study.

Two neighbouring SNP loci of the missing target are considered in the Simpute method. Haplotypes formed by the consecutive pair of loci are constructed and the estimated haplotype probabilities are combined with the LD information from either side of the missing SNP to predict the missing SNP genotype.

2.1. Estimate the Population Proportion of Haplotypes. We first considered genotypes at two loci. The counts of all genotype combinations are summarized in Table 3.

In Table 3, there are nine genotype combinations. The haplotypes for eight of them can be clearly resolved, while those of the $N_{1,1}$ could be either ab/AB or aB/Ab. The proportion of the four haplotypes can be estimated as follows:

$$\begin{aligned} p(\mathrm{ab}) &= \frac{2 \times N_{0,0} + N_{0,1} + N_{1,0} + X_1 \times N_{1,1}}{2 \times N_{P,Q}}, \\ p(\mathrm{aB}) &= \frac{2 \times N_{0,2} + N_{0,1} + N_{1,2} + X_2 \times N_{1,1}}{2 \times N_{P,Q}}, \\ p(\mathrm{Ab}) &= \frac{2 \times N_{2,0} + N_{2,1} + N_{1,0} + X_2 \times N_{1,1}}{2 \times N_{P,Q}}, \\ p(\mathrm{AB}) &= \frac{2 \times N_{2,2} + N_{2,1} + N_{1,2} + X_1 \times N_{1,1}}{2 \times N_{P,Q},}, \end{aligned} \tag{1}$$

where X_1 is the proportion of the phase ab/AB with the observed genotype aAbB, and X_2 is the proportion of the phase aB/Ab.

The initial values for X_1 and X_2 are set to 0.5, and they are iteratively updated to get a more probable estimate. The updating step is

$$\begin{aligned} X_1 &= \frac{p(\mathrm{ab}) \times p(\mathrm{AB})}{p(\mathrm{ab}) \times p(\mathrm{AB}) + p(\mathrm{aB}) \times p(\mathrm{Ab})}, \\ X_2 &= \frac{p(\mathrm{aB}) \times p(\mathrm{Ab})}{p(\mathrm{ab}) \times p(\mathrm{AB}) + p(\mathrm{aB}) \times p(\mathrm{Ab})}. \end{aligned} \tag{2}$$

The estimated X_1 and X_2 are then used to calculate the p(ab), p(aB), p(Ab), and p(AB) in (1). The 10 iterations will stop for either X_1 or X_2. According to (1) and (2), X_1 or X_2 is a cubic function, solved by the cubic formula. Here we use the iterative method to solve X_1 and X_2. The initial value of both is set to 0.5, where the two phases have the same probability (Table 4).

2.2. Linkage Disequilibrium Measurement. We impute the missing genotypes using the LD information and the haplotype probabilities calculated in the previous section. If the LD value between two SNP sites is high, then the two SNPs are close to each other, and there are relatively few recombination events between them. Some measurements are commonly used to evaluate the extent of LD between a pair of SNP sites. Two important pairwise measures of LD are r^2 and $|D'|$ [13–15]. Their range is from 0 to 1, but their interpretation is slightly different. When $|D'|$ is equal to 1, r^2 can be small. For example, when $p(\mathrm{ab}) = 0.9$, $p(\mathrm{aB}) = 0.1$, $p(\mathrm{Ab}) = 0.1$, and $p(\mathrm{AB}) = 0$, $|D'|$ is equal to 1, the r^2 value is 0.012. In this paper, r^2 is derived from the input samples. The $|D'|$ and r^2 can be computed as follows.

The difference between the observed and the expected probability of two loci is measured. The disequilibrium coefficient D is expressed as

$$D = p(\mathrm{ab}) - p(\mathrm{a}\cdot) \times p(\cdot\mathrm{b}). \tag{3}$$

The normalized disequilibrium coefficient is defined as $D' = D/|D|_{\max}$ according the study of Pritchard and Przeworski [14], where

$$D_{\max} = \begin{cases} \min\left(p(\mathrm{a}\cdot) \times p(\cdot\mathrm{B}), p(\mathrm{A}\cdot) \times p(\cdot\mathrm{b})\right), & \text{if } D \geq 0 \\ \min\left(p(\mathrm{a}\cdot) \times p(\cdot\mathrm{b}), p(\mathrm{A}\cdot) \times p(\cdot\mathrm{B})\right), & \text{if } D < 0. \end{cases} \tag{4}$$

The range of the normalized disequilibrium coefficient D' is $[-1, 1]$. D' can be 1 while the P value is not significant. That is, when D' is equal to 1, there can still be no association. Hence, we adopt another popular measurement r^2, where

$$r^2 = \frac{D^2}{p(\mathrm{a}\cdot) \times p(\cdot\mathrm{b}) \times p(\mathrm{A}\cdot) \times p(\cdot\mathrm{B})}. \tag{5}$$

The r^2 value between the sites P and Q is denoted as $r^2_{P,Q}$.

2.3. Imputation Algorithm. Consider three SNP sites P, Q, and R that are in consecutive order. The imputation procedure is as follows.

(1) Use the samples with no missing data at P, Q, and R to calculate the pairwise r^2 at loci P, Q, and R. If the r^2 equals zero, it will be set to a minimum value of 10^{-5} to facilitate the following computation.

(2) Because most haplotypes consisting of three loci are rare in the population, and the population proportion cannot be correctly estimated with the limited samples in most studies, we approximate it with the product of haplotype proportion for the three pairs of loci and put the LD measured between the two loci as the weights. The probability for haplotype $h_1h_2h_3$ is approximated as

$$\begin{aligned} P_{P,Q,R}(h_1h_2h_3) = {} & P_{P,Q}(h_1h_2) \times r^2_{P,Q} \\ & \times P_{Q,R}(h_2h_3) \times r^2_{Q,R} \times P_{P,R}(h_1h_3) \times r^2_{P,R}, \end{aligned} \tag{6}$$

where $P_{P,Q}(h_1h_2)$, $P_{Q,R}(h_2h_3)$, and $P_{P,R}(h_1h_3)$ are the probabilities of haplotype h_1h_2 at loci P, Q, haplotype h_2h_3 at loci Q, R, and haplotype h_1h_3 at loci P, R. These probabilities were generated by (1).

(3) Calculate the weighting score of genotype $\otimes\oplus$ at each pair of loci:

$$W_{(\otimes,\oplus)} = 1 - \left| \frac{N_{\otimes,\cdot} \times N_{\cdot,\oplus}}{N \times N} - \frac{N_{\otimes,\oplus}}{N} \right|, \tag{7}$$

where $\otimes$ and $\oplus$ are the genotypes at the first and the second locus in each pair. If the $W_{(\otimes,\oplus)}$ equals zero, it will be set to a minimum value of 10^{-5} to facilitate the following computation.

(4) Calculate the haplotype pair score

$$\begin{aligned} \text{score} = {} & \left(P_{P,Q,R}(h_1h_2h_3) + P_{P,Q,R}(h'_1h'_2h'_3)\right) \\ & \times \frac{N^{P,Q}_{\otimes,\oplus} \times N^{P,R}_{\otimes,\circ} \times N^{Q,R}_{\oplus,\circ}}{N^{P,Q}_{\otimes,\cdot} \times N^{P,R}_{\cdot,\circ} \times N^{Q,R}_{\oplus,\cdot}} \times W_{(\otimes,\oplus)} \times W_{(\otimes,\circ)} \times W_{(\oplus,\circ)}, \end{aligned} \tag{8}$$

where the probabilities of the haplotype pair $P_{P,Q,R}(h_1h_2h_3)$ and $P_{P,Q,R}(h'_1h'_2h'_3)$ are calculated by (6), and $\otimes$, $\oplus$, $\circ$ represent the same genotypes $(h_1h'_1)$, $(h_2h'_2)$, and $(h_3h'_3)$ at locus P, Q, and R, respectively.

(5) Choose from all legitimate haplotype pairs that maximize the score in (8).

The algorithm also considers the situation when consecutive SNPs are missing. In that case, the two neighbouring loci P and R of the missing locus Q can represent the adjacent two loci on the same side of the Q. For example, when there is a long stretch of missing genotypes from SNP 1 to 4 in a specific sample, we can first impute locus 4 with information from locus 5 and 6 and then sequentially fill in all the missing ones.

2.4. Time Complexity. Our algorithm requires the computation complexity at the order of $O(mw + n)$ where n is the number of missing SNPs, w is the number of the SNP loci with at least one missing entry, and m is the number of individual with at least one locus missing. Each sample requires the order of $O(1)$ to count each of the 9 genotype and the order of $O(mw)$ for steps 1 and 2. Hence, the total complexity of the algorithm is $O(mw + n)$.

3. Data Description

In this paper, we used two data sets to compare imputation performance. All data sets are based on the individuals included in the HapMap project [16].

3.1. SNP-Dense Region on Chromosome 22. The first data set was the testing region adopted from Jung et al. They identified a region with dense SNP distribution and demonstrated their performance with only six SNPs, as annotated in HapMap Phase II, release 22. Those SNPs are rs2213329, rs2227029, rs9610029, rs2213331, rs9619447 and rs743726, and are located from positions 33227611 to 33233156 of chromosome 22. This region was selected for its strong linkage of $|D'| > 0.7$. We used the SNP data of 270 people from HapMap to generate the testing data. The data were randomly selected with missing rates of 5%, 10%, 15%, and 20% from the total of $270 \times 6 = 1620$ SNPs. We adopted the settings of the missing rates of Jung et al. for comparison purposes. This random procedure was repeated 100 times, and the average error rates were obtained. A more realistic comparison is demonstrated with the other set of random missing studies described in the following section.

3.2. Random Missing SNPs from the HapMap Phase III on Chromosome 21. We used samples of HapMap phase III as our testing data. Because some of the software we compared required reference data, we provided samples of HapMap Phase II release 22 as the reference samples; those samples were, thus, excluded in our testing set. SNP loci that are tri-alleic or tetraallelic were excluded in the comparison; Tables 1 and 2 show the proportion of this type of loci in the reference samples (HapMap Phase II release 22) and testing samples (HapMap Phase III specific samples), respectively.

Table 1: The nonbiallelic loci proportion in the HapMap phase II release 22.

Population	Individuals	SNPs	Nonbiallelic
CEU	90	48217	1.69%
JPT + CHB	90	50053	1.81%
YRI	90	48541	1.60%

Table 2: The non-bi-allelic loci proportion in the HapMap phase III.

Population	Individuals	SNPs	Non-bi-allelic
CEU	80	19250	0.39%
JPT + CHB	77	17286	0.21%
YRI	80	20198	0.21%

Table 3: A 3 × 3 contingency table for the genotypes at two consecutive loci. A and a are the two alleles in locus 1 while B and b are the two alleles in locus 2.

	0 (bb)	1 (bB)	2 (BB)	Total
0 (aa)	$N_{0,0}$	$N_{0,1}$	$N_{0,2}$	$N_{0,\cdot}$
1 (aA)	$N_{1,0}$	$N_{1,1}$	$N_{1,2}$	$N_{1,\cdot}$
2 (AA)	$N_{2,0}$	$N_{2,1}$	$N_{2,2}$	$N_{2,\cdot}$
Total	$N_{\cdot,0}$	$N_{\cdot,1}$	$N_{\cdot,2}$	$N_{P,Q}$

Table 4: Notation for the haplotype probabilities at the two loci.

Locus $P \setminus Q$	b	B	Total
a	p (ab)	p (aB)	p (a·)
A	p (Ab)	p (AB)	p (A·)
Total	p (·b)	p (·B)	1

Table 5: Error rates* for Simpute, BEAGLE, and Jung's method with random missing study on the six SNPs of chromosome 22.

Missing rate/method	Simpute	BEAGLE	Jung's method
5%	1.358%	1.7531%	16.59%
10%	1.8944%	2.1429%	17.82%
15%	3.0207%	3.4132%	20.25%
20%	4.4472%	4.4907%	20.07%

*Error rates = number of error imputed entries/number of missing entries ∗100%.

The proportion is low and is not crucial for the conclusion. We conducted the experiment on the smallest chromosome to enable easier computation of the less efficient algorithms in the comparison. The results are reported separately for the different ethnic groups because certain interesting differences were observed.

We generated three sets of testing data from the HapMap Phase III specific samples. The first set was derived by randomly masking the genotypes on chromosome 21, called the *complete set*. Because the error rate of genotype calling is less than 1% [17], the missing rates were 0.1%, 0.5%, 1%, and 5%. Ten randomly missing testing data sets were generated for comparison, and the accuracy was calculated as the average of the 10 repeats. The software used to compare the data set included BIMBAM v0.99, BEAGLE v3.0.3, IMPUTE v0.5.0, MARCH v1.0.16, PLINK, and SNPMSTAT v3.1 and used the system Linux kernel version 2.6 on AMD 64 platform.

Our second test data consisted of numerous regions of only three SNPs on chromosome 21, called the *short input*. At most, two of the three SNPs were permitted to be missing in our random sampling process. The error rates are reported, with the averages of 25 repeats of the random missing procedure for missing rates, as 0.1%, 0.5%, 1%, and 5%.

The algorithm we proposed adopted minimum information to complete the missing gaps, and, hence, it is not designed for all purposes. We show that its performance at the highly linked regions is no worse than the best method previously mentioned. The third set of test data consists of missing SNPs with strong linkage ($r^2 > 0.9$) to either one of their adjacent SNPs, called *high LD*. The advantage acquired at the highly linked regions is the most important aspect of Simpute and is why Simpute is the most helpful program in global genome sequencing projects. The error rates are reported from the average of 100 repeats of the random missing procedure for the missing rates at 0.1%, 0.5%, 1%, and 5%.

4. Results

We used samples from HapMap Phase II release 22 as the reference data set, which is required by BEAGLE, BIMBAM, MACH, SNPMStat, IMPUTE, and plink. Because of the intractable computation load of SNPMStat and IMPUTE, we divided the chromosome into segment of 10,000 SNPs for the inputs. Because SNPMStat requires substantial CPU time, only three repeats were performed to derive the average accuracy. All the other programs used 10 repeats to obtain the averages. The results from the *complete set* are shown in Figure 1. BEAGLE gives the best overall accuracy and is also the fastest on our benchmark platform CentOS 5.3 under the VNWare ESX 4i in Figure 2. The following comparisons only address the differences between Simpute and BEAGLE.

The results from the SNP-dense region of chromosome 22 in Jung's study are shown in Table 5. The error rates from the Jung et al. study are copied directly from their report because we did not implement their algorithm. It appears that Simpute has a strong advantage in the SNP-dense regions. Although BEAGLE used the same HapMap samples as reference samples and used all six SNPs together in their complicated algorithms, it still has slightly higher error rates, and the contrast is strong at the lower missing rates.

To understand the relation between the information fed into each method and the power each method gains, we first assessed the sets of three SNPs on chromosome 21. This provided limited information, and the error rates for Simpute and BEAGLE are shown in Table 6. The missing rates were set as 0.1%, 0.5%, 1%, and 5% to better match the actual applications. Because the data are artificial and require repeated initiation processes for BEAGLE to process all the short regions, extensive computation time is required for BEAGLE to process all the data. Hence, comparing the computation time is not feasible, and it is difficult to run the entire set of simulations on all three ethnic groups. We only reported the results for Group CEU with 25 repeats of the random missing

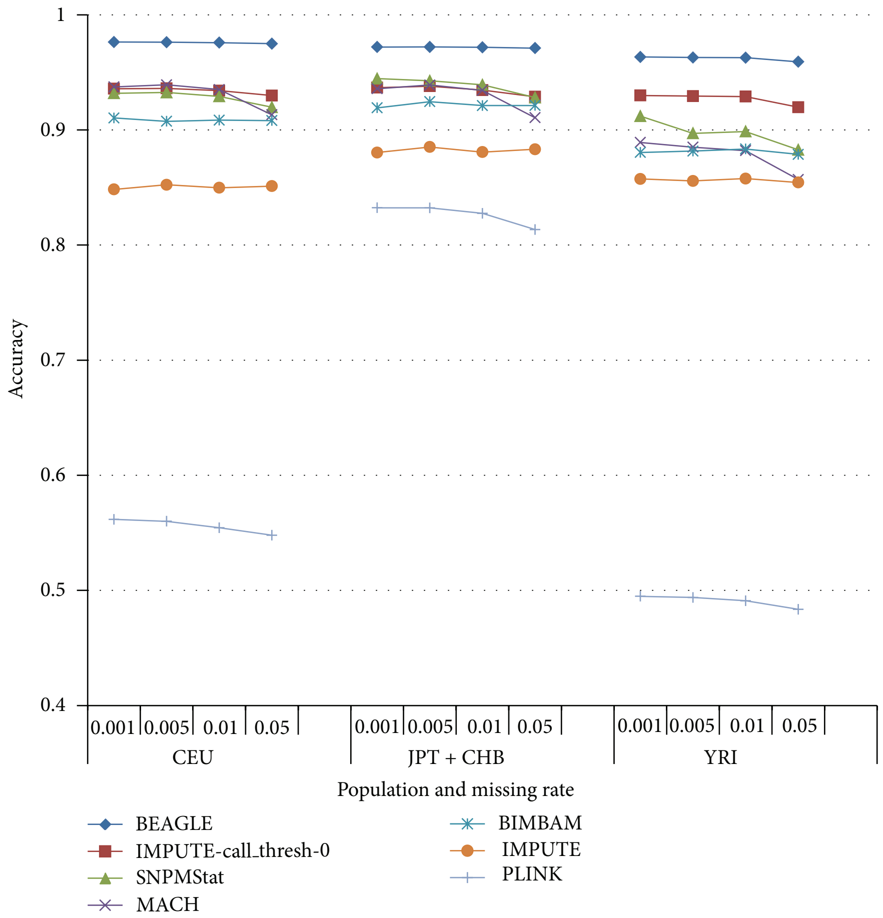


Figure 1: Imputation accuracy compared across BEAGLE, IMPUTE, BIMBAM, SNPMStat, MACH, and plink using the *complete set*. The curve with IMPUTE-call_thresh-0 stands for the best setting (call thresh = 0) we found for Impute rather than the default setting. Accuracy = number of correctly imputed entries/number of missing entries $*100\%$.

procedure, as shown in Table 6. The error rate of Simpute is approximately the same as that of BEAGLE and matches our expectations.

Tables 7, 8, and 9 show the evaluation of Simpute and BEAGLE using the *high LD* testing data on chromosome 21. The default setting of BEAGLE used the same 270 people from HapMap Phase II as the reference data. In contrast, Simpute used the two neighboring SNPs of the missing one. The error counts are the averages of 100 repeats of the random missing procedure. BEAGLE performed better than Simpute but the difference is negligible when the missing rate is low. In addition, BEAGLE requires substantially more processing time.

Table 6: The error rates* for random missing SNPs of short input at $r^2 \geq 0.9$ from the HapMap phase III on chromosome 21 of short input for the CEU.

Method/ missing rate	Simpute	BEAGLE
0.1%	37.136/483 (7.69%)	38.09/483 (7.89%)
0.5%	188/2412 (7.79%)	183.6364/2412 (7.61%)
1%	378.333/4823 (7.84%)	376.762/4823 (7.81%)
5%	1913.632/24111 (7.94%)	1892.053/24111 (7.84%)

*Error rates = number of error imputed entries/number of missing entries $*100\%$.

5. Conclusion and Discussion

In this study we developed a simple strategy to impute missing genotypes for SNPs that have a high resolution. Our method requires only two neighbouring loci of a missing SNP. Furthermore, we show in our study that for highly linked loci, our algorithm has comparable performance to BEAGLE, a system that incorporates data from various sources of information, as has been suggested in recent studies. These sources of information include reference samples and long-range LD.

The algorithm we introduced in our study has a complexity of $O(mw + n)$, where n is the number of missing SNPs, w is the number of the positions of the missing SNPs, and m is the sample size. Because of the design of our algorithm, and the reduction of the prerequisite input incorporated into the imputation algorithm, we were able to significantly reduce the computation time.

Table 7: Error rates* and computation time for random missing SNPs of high LD for the CEU samples.

Method/missing rate	Simpute		BEAGLE	
	Error rate	Running time (sec)	Error rate	Running time (sec)
0.1%	5.52/483 (1.14%)	12.88	4.49/483 (0.93%)	164.17
0.5%	27.94/2412 (1.16%)	13.09	21.01/2412 (0.87%)	164.82
1%	57.22/4823 (1.19%)	14.07	44.07/4823 (0.91%)	168.47
5%	321.9/24111 (1.33%)	18.24	224.65/24111 (0.974%)	168.69

*Error rates = number of error imputed entries/number of missing entries ∗100%.

Table 8: Error rates* and computation time for random missing SNPs of high LD for the CHB + JPT samples.

Method/missing rate	Simpute		BEAGLE	
	Error rate	Running time (sec)	Error rate	Running time (sec)
0.1%	5.15/493 (1.04%)	10.90	4.64/493 (0.94%)	138.40
0.5%	27.29/2463 (1.10%)	11.08	24/2463 (0.97%)	139.79
1%	55.07/4925 (1.11%)	11.77	47.69/4925 (0.97%)	138.96
5%	322.38/24622 (1.31%)	16.113	242.26/24622 (0.98%)	140.96

*Error rates = number of error imputed entries/number of missing entries ∗100%.

Table 9: Error rates* and computation time for random missing SNPs of high LD for the YRI samples.

Method/missing rate	Simpute		BEAGLE	
	Error rate	Running time (sec)	Error rate	Running time (sec)
0.1%	2.57/271 (0.95%)	12.42	2.23/271 (0.82%)	187.80
0.5%	13.54/1353 (1.00%)	12.925	11.2/1353 (0.83%)	188.41
1%	27.19/2705 (1.00%)	13.08	22.89/2705 (0.85%)	187.45
5%	161.02/13525 (1.19%)	15.921	119.94/13525 (0.887%)	191.29

*Error rates = number of error imputed entries/number of missing entries ∗100%.

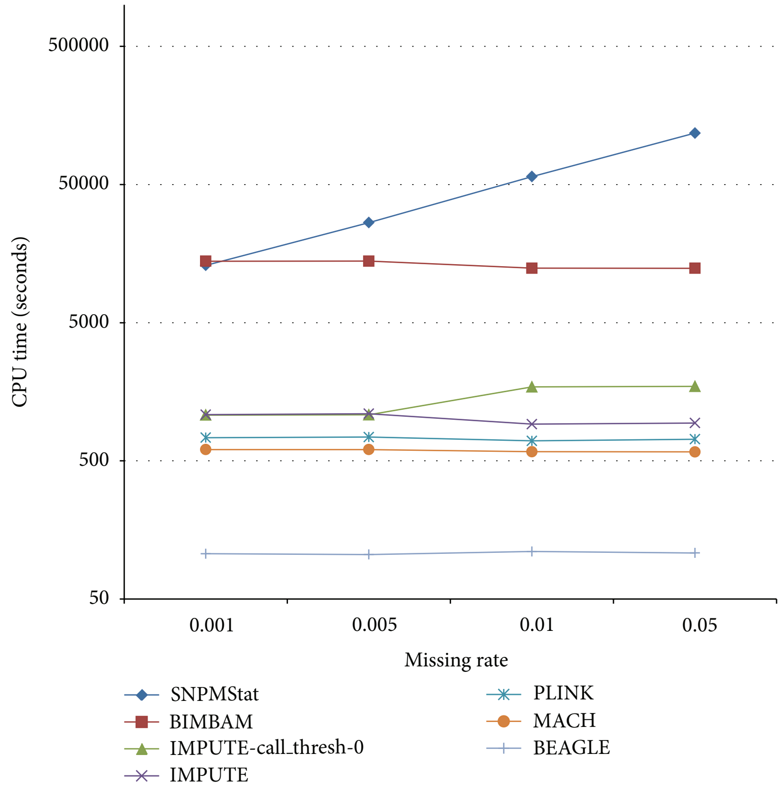


Figure 2: CPU Time.

Although Simpute is unable to outperform most software for general purposes, it has shown its potential for specific purposes. With the current trend of mass parallel-sequencing technologies, SNPs will soon be discovered with ease, without requiring the use of predefined positions for their detection. Furthermore, the availability of samples will accumulate in the following few years. Thus, it is expected that most SNPs will be highly linked in samples of moderate size.

Simpute has a strong advantage over more complicated algorithms that use high LD regions. Moreover, it demonstrates a distinct advantage in efficiency when handling large data sets. This efficiency is of great benefit to genome centers, which have increasing demands in the number of personal genomes that must be sequenced and analyzed through a real-time system.

Availability

Simpute is available from the following website: http://www.cs.nthu.edu.tw/~dr928307/Simpute.htm. We provide an integrated interface to run all of these softwares. It can be downloaded at http://kitty.2y.idv.tw/~tcs/ASHG2009/ and performed under Linux kernel 2.6 amd64 platform.

Authors' Contribution

Y.-J. Lin and C. T. Chang contributed equally to this work.

References

[1] J. I. Bell, "Single nucleotide polymorphisms and disease gene mapping," *Arthritis Research*, vol. 4, supplement 3, pp. S273–S278, 2002.

[2] R. Sachidanandam, D. Weissman, S. C. Schmidt et al., "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–933, 2001.

[3] L. Kruglyak and D. A. Nickerson, "Variation is the spice of life," *Nature Genetics*, vol. 27, no. 3, pp. 234–236, 2001.

[4] H. Y. Jung, Y. J. Park, Y. J. Kim, J. S. Park, K. Kimm, and I. Koh, "New methods for imputation of missing genotype using linkage disequilibrium and haplotype information," *Information Sciences*, vol. 177, no. 3, pp. 804–814, 2007.

[5] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature Genetics*, vol. 39, no. 7, pp. 906–913, 2007.

[6] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genetic Epidemiology*, vol. 34, no. 8, pp. 816–834, 2010.

[7] D. Y. Lin, Y. Hu, and B. E. Huang, "Simple and efficient analysis of disease association with missing genotype data," *American Journal of Human Genetics*, vol. 82, no. 2, pp. 444–452, 2008.

[8] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," *American Journal of Human Genetics*, vol. 78, no. 4, pp. 629–644, 2006.

[9] B. L. Browning and S. R. Browning, "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals," *American Journal of Human Genetics*, vol. 84, no. 2, pp. 210–223, 2008.

[10] M. N. Chiano and D. G. Clayton, "Fine genetic mapping using haplotype analysis and the missing data problem," *Annals of Human Genetics*, vol. 62, no. 1, pp. 55–60, 1998.

[11] Z. S. Qin, T. Niu, and J. S. Liu, "Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms," *American Journal of Human Genetics*, vol. 71, no. 5, pp. 1242–1247, 2002.

[12] Y. V. Sun and S. L. R. Kardia, "Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks," *European Journal of Human Genetics*, vol. 16, no. 4, pp. 487–495, 2008.

[13] B. Devlin and N. Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping," *Genomics*, vol. 29, no. 2, pp. 311–322, 1995.

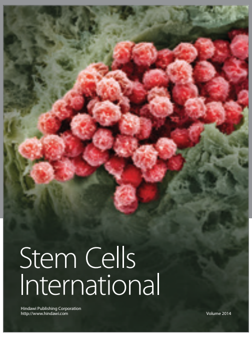
[14] J. K. Pritchard and M. Przeworski, "Linkage disequilibrium in humans: models and data," *American Journal of Human Genetics*, vol. 69, no. 1, pp. 1–14, 2001.

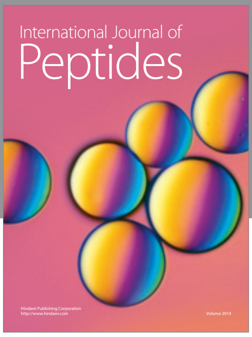
[15] R. C. Lewontin, "Interaction of selection and linkage. I. General considerations; heterotic models," *Genetics*, vol. 49, no. 1, pp. 49–67, 1964.

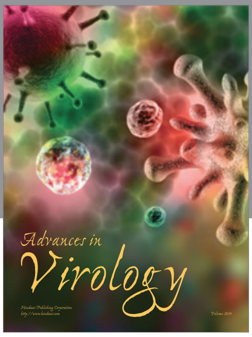
[16] K. A. Frazer, D. G. Ballinger, D. R. Cox et al., "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 7164, pp. 851–861, 2007.

[17] M. J. Huentelman, D. W. Craig, A. D. Shieh et al., "SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data," *BMC Genomics*, vol. 6, article 149, 2005.

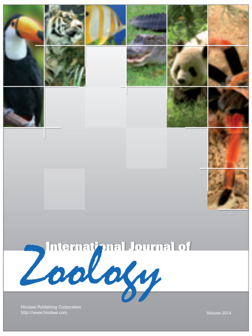
BioMed
Research International

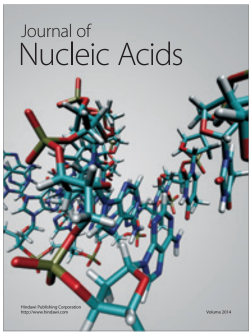

Stem Cells
International


International Journal of
Peptides


Advances in
Virology

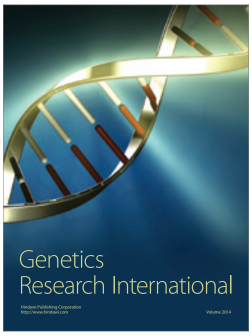
International Journal of
Genomics

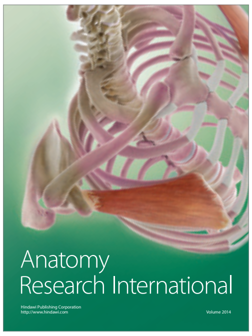

International Journal of
Zoology


Journal of
Nucleic Acids

Journal of
Signal Transduction

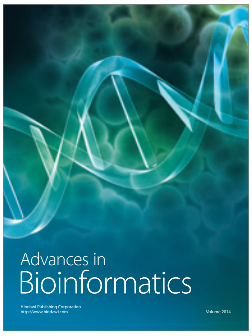
The Scientific
World Journal

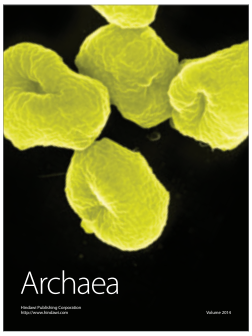

Genetics
Research International

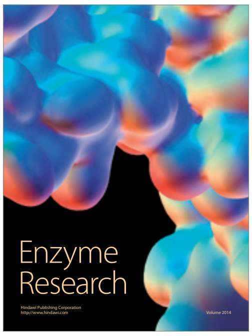

Anatomy
Research International

International Journal of
Microbiology

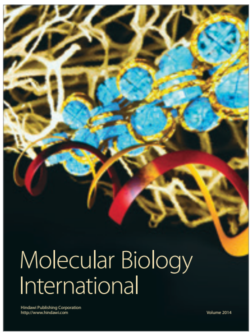
Biochemistry
Research International


Advances in
Bioinformatics


Archaea


Enzyme
Research

International Journal of
Evolutionary Biology


Molecular Biology
International

Journal of
Marine Biology