BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 9 2005, pages 2001–2007
doi:10.1093/bioinformatics/bti261

Genetics and population analysis

Inference of missing SNPs and information quantity measurements for haplotype blocks

Shih-Chieh Su[1], C.-C. Jay Kuo[1] and Ting Chen[2,*]

[1]Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA and
[2]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

Received on October 23, 2004; revised on December 13, 2004; accepted on December 31, 2004.
Advance Access publication February 4, 2005

ABSTRACT

Motivation: Missing data in genotyping single nucleotide polymorphism (SNP) spots are common. High-throughput genotyping methods usually have a high rate of missing data. For example, the published human chromosome 21 data by Patil *et al*. contains about 20% missing SNPs. Inferring missing SNPs using the haplotype block structure is promising but difficult because the haplotype block boundaries are not well defined. Here we propose a global algorithm to overcome this difficulty.

Results: First, we propose to use entropy as a measure of haplotype diversity. We show that the entropy measure combined with a dynamic programming algorithm produces better haplotype block partitions than other measures. Second, based on the entropy measure, we propose a two-step iterative partition-inference algorithm for the inference of missing SNPs. At the first step, we apply the dynamic programming algorithm to partition haplotypes into blocks. At the second step, we use an iterative process similar to the expectation-maximization algorithm to infer missing SNPs in each haplotype block so as to minimize the block entropy. The algorithm iterates these two steps until the total block entropy is minimized. We test our algorithm in several experimental data sets. The results show that the global approach significantly improves the accuracy of the inference.

Availability: Upon request.

Contact: tingchen@usc.edu

1 INTRODUCTION

The study based on the single nucleotide polymorphism (SNP) has drawn wide attraction from the public. An SNP is considered to be a mutation at a single nucleotide position, and it keeps record through heredity thereafter. Human SNP data are very useful for the study of various subjects in human genetics and diseases. Moreover, it is known that SNPs are highly abundant across the entire human genome.

Human chromosomes appear in pairs. A haplotype is an observed sequence of SNPs in one chromosome, and a pair of haplotypes are called a genotype. It has been observed that human haplotypes have a block structure (Daly *et al*., 2001). Within a block, haplotypes have low diversity. Recombination hotspots are considered to be a source of perturbation of the block patterns—which can also be caused by the population structure. However, these boundaries are not well defined and finding them is itself a challenging task. The study of haplotype block partitioning was pioneered by Daly *et al*. (2001) and Patil *et al*. (2001), each offering a human haplotype data set. Systematic partitioning using a dynamic programming (DP) algorithm was first proposed by Zhang *et al*. (2002). Later, the objective function in the algorithm was replaced with minimum description length (MDL) measurements (Koivisto *et al*., 2003; Anderson and Novembre, 2003). These two methods produce different results, depending upon how the MDL describes the data set and how it treats missing data.

Missing data are common in haplotype data sets. The data set in Daly *et al*. (2001) has around 10% missing SNPs, and the data set in Patil *et al*. (2001) has around 20% missing SNPs. The missing SNPs will cause ambiguities in haplotypes and thus seriously affect many SNP-based applications such as disease mapping. In previous works, missing SNPs were either inferred with some simple methods or just ignored.

The problem of the inference of missing SNPs is associated with the problem of haplotype inference, where given a set of observed genotypes, we are asked to estimate the frequencies of all haplotypes. Statistical solutions for the problem of haplotype inference include those by Excoffier and Slatkin (1995), Niu *et al*. (2002), Qin *et al*. (2002), Lin *et al*. (2002), Stephens *et al*. (2001) and Stephens and Donnelly (2003). In general, there are two approaches: expectation-maximization (EM) algorithms and Gibbs sampling algorithms. Two kinds of priors were used: Dirichlet prior and approximate coalescent prior. Based on the estimated frequencies of haplotypes, we can assign values to missing SNPs. The challenge of applying these methods directly to infer missing SNPs is the identification of haplotype block boundaries, which is a difficult task.

In this study, we propose to measure haplotype diversity within a block using an information quantity measure called entropy. We develop a dynamic programming algorithm to partition the SNP data into haplotype blocks so as to minimize the total block entropy. Also, we show that the haplotype block structure produced by this measure is closer to the manually generated block structure suggested in Daly *et al*. (2001) than other measures. Given this haplotype block partition, we develop an EM-like iterative process to infer values of missing SNPs within each block. Combining these two steps, we propose a new algorithm, called the iterative partition-inference (IPI), to infer missing SNPs jointly with haplotype block partitioning. In the first step, we apply a dynamic programming algorithm to partition haplotypes into blocks. In the second step, we use the

*To whom correspondence should be addressed.

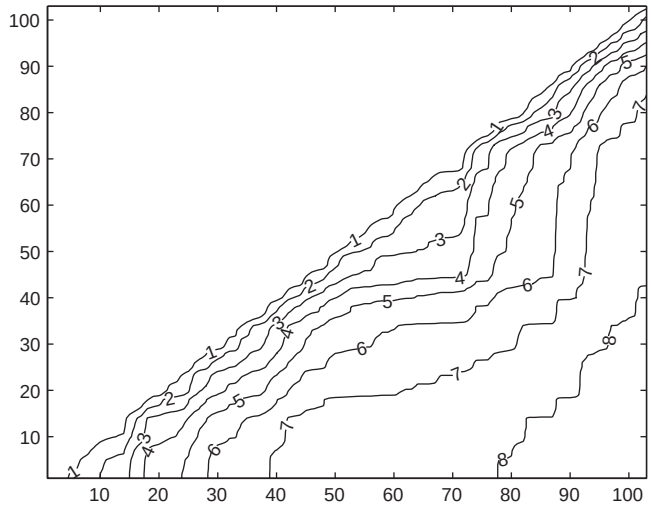

Fig. 1. Entropy map of the data set in Daly *et al*. (2001). Each point (i, j) indicates the entropy of the block from the ith SNP to the jth SNP. The contours connect points having the same entropy.

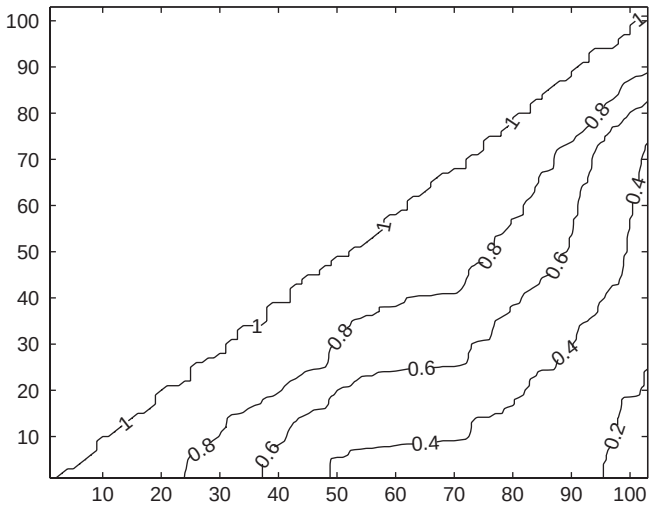

Fig. 2. α map of the data set in Daly *et al*. (2001). Each point (i, j) indicates the α value of the block (i, j), where α is the percentage of the unambiguous haplotypes in the block.

iterative process to infer missing SNPs in each haplotype block so as to minimize the block entropy. The algorithm iterates these two steps until the total block entropy is minimized. We test our algorithm in several experimental data sets. The results show that the global approach significantly improves the accuracy of the inference.

2 METHODS

2.1 Entropy map

Conventionally, the linkage disequilibrium (LD) can be measured in a pairwise manner. The widely used Lewontin's D' is simple and powerful in measuring the LD (Hedrick, 1987). However, the LD measurements are usually too noisy for measuring haplotype blocks.

Low diversity is a common feature of haplotype blocks. Here we use entropy as a measure of haplotype diversity within a block: low entropy indicates low diversity. We define the haplotype block entropy as follows. Let (i, j) denote the SNPs from the ith SNP to the jth SNP. Let $\Phi(i, j)$ denote the set of haplotypes collected. The block entropy is then defined as

$$E(i, j) = \sum_{\phi \in \Phi(i,j)} P_\phi \log \frac{1}{P_\phi}. \quad (1)$$

This block entropy measurement has an important property of consistency: $E(k, m) \geq E(i, j)$ if $k \leq i \leq j \leq m$. In a region of interest, we can draw entropy contours for better visualization. Figure 1 shows an entropy contour plot for Daly's data. In the plot, a contour connects all of the (i, j) pairs with same $E(i, j)$. We call this contour plot an *entropy map*. Since the entropy map is symmetric, we only show the lower triangle of the map.

Entropy has been used as a measure to enhanced the conventional LD. Nothnagel *et al*. (2003) used a sliding window of pre-selected sizes to cumulate entropy-based statistics. In this study, we use information quantities to measure the whole haplotype data through block partitioning.

The information quantity measurement has some advantages over the conventional LD measurement in that it can measure not only the diversity within a block, but also the pairwise relationship between two blocks. For example, the mutual information between two blocks, (i, j) and (k, m), is defined as

$$I[(i, j), (k, m)] = E(i, j) + E(k, m) - E[(i, j), (k, m)],$$

where $E((i, j), (k, m))$ is calculated over all combinations of haplotypes across two blocks (i, j) and (k, m). The mutual information increases as LD increases. When there is no linkage between the two blocks, which means that the two blocks are independent, $D' = 0$, and the mutual information is zero as well. On the other side, when there is full linkage between the two blocks, $D' = 1$, and the mutual information is maximum. The mutual information is an un-normalized measurement, so the maximum value varies.

The α-value proposed in Zhang *et al*. (2002), defined as the percentage of the unambiguous haplotypes in the block that appear more than once, is another normalized measurement of block diversity. Higher α-values indicate lower diversity. Figure 2 shows the plot of the α map. By definition, the α value bears the consistency property mentioned above. Therefore, the α map has contours similar to those of the entropy map. However, the α-value is not a precise measurement of diversity; it measures the portion of singleton haplotypes in a block.

In conclusion, the information quantities have good properties and yield better measurements of diversity and linkage. We will apply them to the haplotype block partitioning problem.

2.2 Haplotype block partitioning

In general, the haplotype block structure has the following properties that can be measured by information quantities:

(1) Low intra-block diversity, which can be measured by entropy;

(2) High inter-block diversity, which can be measured by joint entropy;

(3) Low inter-block dependency, which can be measured by mutual information.

We use an entropy threshold T to define a block ($\leq T$). A proper T-value can be selected with the help of the entropy map. We assume the haplotype blocks are consecutive along chromosomes. Based on the three properties of the haplotype block structure, we employ three different cost functions to partition haplotypes into blocks.

The minimum entropy (ME) method. The ME method is to minimize the total block entropy. Denote $B(j)$ as the minimum total block entropy from the first SNP to the jth SNP. Let $e(j)$ be the beginning SNP of the last block of the partition that yields $B(j)$. Then we have the DP structure as

$$B(j) = \min_{1 \leq i \leq j} \{B(i-1) + E(i, j); \quad \text{for } E(i, j) \leq T\}. \quad (2)$$

The condition $E(i, j) \leq T$ in Equation (2) defines a block.

The maximum joint entropy (MJE) method. Given a haplotype block partition, the total joint entropy is the sum of the joint entropies of adjacent blocks. Let

$C(j,k)$ denote the maximum total joint entropy from the first SNP to the kth SNP, with the last block beginning at the jth SNP. Then, the maximum total joint entropy can be computed by a two-dimensional DP algorithm using the following recursion:

$$C(j,k) = \max_{1\le i<j\le k}\{C(i,j) + E(i,j-1) + E(j,k); \text{ for } E(j,k) \le T \text{ and } E(i,k) \ge T\}. \quad (3)$$

The minimum mutual information (MMI) method. Given a haplotype block partition, the total mutual information is the sum of the mutual informations of adjacent blocks. Let $D(j,k)$ denote the minimum total mutual information from the first SNP to the kth SNP, with the last block beginning at the jth SNP. Then, the minimum total mutual information can be computed by a two-dimensional DP algorithm using the following recursion:

$$D(j,k) = \min_{1\le i<j\le k}\{D(i,j) + I((i,j-1),(j,k)); \text{ for } E(j,k) \le T \text{ and } E(i,j-1) \le T\}. \quad (4)$$

Information quantities serve as criteria in all of our three methods. The DP algorithms proposed above can produce partitions which yield the optimal information quantities. However, the resulting partitions are sensitive to the threshold T. In Section 4, we will discuss how to choose T.

2.3 Inference of missing SNPs

In this study, we use entropy as the measure for haplotype blocks. Since haplotype blocks lack diversity, the value of a missing SNP can be inferred so as to lower the block entropy. However, there may be multiple missing SNPs within a block, so they have to be assigned jointly to yield the lowest block entropy of that block. Assuming that there are m missing SNPs within a block, let $X = x_1,\ldots,x_m$ be the random variables of these missing SNPs, each of which can be assigned 0 or 1 representing the wild type or the mutant type, respectively. Our goal is to find

$$X = \arg_X \min E(X),$$

where $E(\cdot)$ is the entropy of this block. Note that the number of possible assignments for X can be as large as 2^{m-1}. A brute force search would be impractical because m is generally large. Thus, we develop an EM-like iterative process as follows. In each run of the iterative process, we fix $m-1$ missing SNPs, and update the value of the remaining missing SNP according to the frequencies of the current haplotypes within the block. Gradually, the updating process organizes the haplotype content within a block into lower diversity.

We start with the ith missing SNP x_i. Let $h(x_i)$ be the haplotype containing x_i. Define H_{-i} to be the set of haplotypes excluding $h(x_i)$. Similarly, define X_{-i} as the set of missing SNPs excluding x_i. The conditional probability for x_i being the majority ($x_i = 0$) is $P(x_i = 0|X_{-i})$. Let $\mathbb{D}$ be the non-missing SNPs in this block. Then, in the first step (similar to the E-step in EM) of the iterative process, we estimate the frequency of haplotype $h(x_i)$ for $x_i = 0$,

$$f_0 = P(h(x_i = 0)|H_{-i},\mathbb{D}), \quad (5)$$

and the frequency of haplotype h for $x_i = 1$,

$$f_1 = P(h(x_i = 1)|H_{-i},\mathbb{D}). \quad (6)$$

In the second step (similar to the M-step in EM), we assign $x_i = 0$ if the new conditional majority probability

$$P(x_i = 0|X_{-i}) = \frac{f_0}{f_0 + f_1} \ge 0.5, \quad (7)$$

and $x_i = 1$ otherwise.

The haplotypes within a block can be organized into clusters within which haplotypes are identical. The block entropy is measured by the size of each cluster and the number of clusters. Thus, each run of the iterative process incurs a cluster movement of the following:

- *Death*: current haplotype $h(x_i)$, itself a cluster, merges into another haplotype cluster, or
- *Migration*: current haplotype $h(x_i)$ moves from a smaller-sized cluster to a larger-sized cluster, or
- *Keep*: current haplotype does not move anywhere.

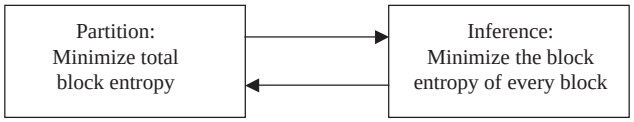

Fig. 3. The IPI algorithm for the inference of missing SNPs.

There are two impossible movements: *Birth* and *Migration from larger cluster to smaller cluster*. For all of the possible movements, we show the increment of block entropy $\Delta E \le 0$. The proof is shown in supplementary data. Whenever there is a cluster movement, the block entropy will decrease. The movement will eventually stop as the block entropy reaches its local minimum.

2.4 Iterative partition and inference

The aforementioned haplotype block partitioning and inference of missing SNPs can be combined together to reduce the diversity in all blocks. Thus, we propose the IPI scheme shown in Figure 3. In the haplotype partitioning module, we choose the *ME method*, which computes the partition to minimize the total block entropy. In the inference module, we employ the iterative process to minimize the entropy locally within a block. The whole system will converge to a locally minimal block entropy and a locally minimal error rate of the inference, given the current parameters in the partitioning step. We can further feed the inference performance back to the partitioning step. Thus the parameters can be adjusted dynamically so that the total block entropy reduces.

3 RESULTS

3.1 Data sources

We use two data sets for testing, one from Daly *et al*. (2001), and the other from Patil *et al*. (2001). Here we use 0 and 1 to denote the majority value (wild type) and the minority value (mutant) of an SNP, and we introduce 2 for the missing SNP. The data set in Daly *et al*. (2001) contains 387 samples of 103 SNPs in genotype format. To create a ground truth haplotype data set for testing, we pre-process the data set in Daly *et al*. (2001) as follows. Every heterozygous allele is inferred through trios (father, mother and child). If it cannot be determined, we assume it as missing. These originally missing SNPs are assigned to the majority value at their loci. This haplotype data set will serve as the ground truth for the test of the inference error rate. The data set in Patil *et al*. (2001) contains 20 samples of 24,047 SNPs in haplotype format. Thus, no reduction needs to be made on it. However, the quality varies across its 24,047 SNPs. We select a high-quality region (8461–8720) in our test. This data set contains 683 missing SNPs (13%) among the total of 5200 SNPs. Similarly, these missing SNPs are assigned to the majority value at their loci. Both data sets are used in the following experiments.

3.2 Haplotype block partitioning

We compare the three information quantity-based measures with those proposed by Daly *et al*. (2001), Koivisto *et al*. (2003), Anderson and Novembre (2003) and Zhang *et al*. (2002), using the data set offered by Daly *et al*. (2001). All except Daly *et al*. (2001) used dynamic programming techniques to optimize different objective functions. The dynamic programming methods assume blocks to be consecutive, while the partition proposed by Daly *et al*. (2001)

Table 1. Information measurements of different partitioning methods

	Daly-RP[a]	Daly-LP[a]	Anderson	IQB (method, T[c])	Daly-O[a]	Koivisto	Zhang[d]
Number of blocks	11	11	11	11	11	6	5
Total block entropy	31.7722	32.1279	33.2721	31.1801 (ME, 3.8)	30.0693	23.3165	25.6301
Average block entropy	2.8884	2.9207	3.0247	2.8346 (ME, 3.8)	2.7360	3.8861	5.1260
Total joint entropy	48.4251	48.6917	49.4006	52.8365 (MJE, 4.8)	—[b]	32.1616	28.3691
Average joint entropy	4.8425	4.8692	4.9401	5.2836 (MJE, 4.8)	—[b]	6.4323	7.0923
Total mutual information	11.3103	11.4455	13.3346	1.4183 (MMI, 5.1)	—[b]	9.7992	11.7832
Average mutual information	1.1310	1.1455	1.3335	0.1418 (MMI, 5.1)	—[b]	1.9598	2.9458

[a]RP: right-passed; LP: left-passed; O: original.
[b]Daly's original partition contains gaps; thus we omit all these information quantities.
[c]The threshold that defines a block.
[d]The partition was performed in Anderson and Novembre (2003).

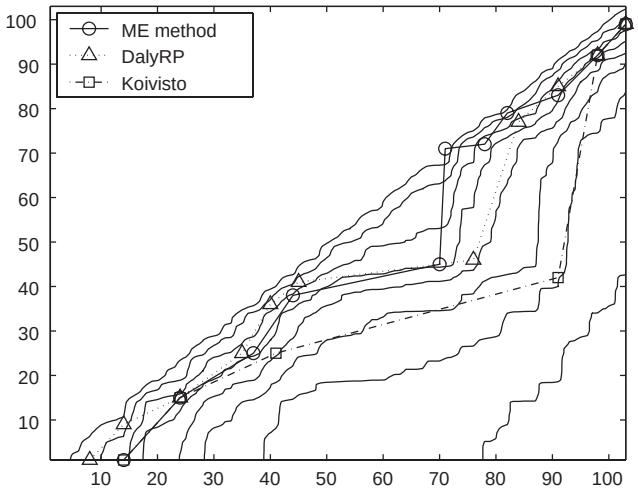

Fig. 4. The block entropy points in the partition of Daly-RP, Koivisto and ME with $T = 3.8$.

contains gaps. There are a total of four gaps in Daly *et al*. (2001), each with one SNP. To make it compatible with others, we merge the gaps to the right blocks in the right-passed (RP) partition of Daly *et al*. (2001), and to the left blocks in the left-passed (LP) partition. We calculate the information measurements on each of the partitioning methods, as shown in Table 1, which compares the methods of Daly-RP, Daly-LP, Anderson, and the information quantity-based (IQB) methods.

Since our methods optimize the haplotype block partitions by three information measurements, it is no surprise that our algorithms out-perform others in each of the three measurement categories. However, if we consider the result of Daly *et al*. (2001) as the human expert ground truth, our ME method yields a closer partition to it. This suggests that 'low intra-block diversity' is likely to be a more important property than the other two.

In Figure 4, we visualize the block partitions on the entropy map (Daly's data) obtained by Daly-Rp, the method in Koivisto *et al*. (2003), and the ME method. The entropy map suggests a low-diversity region from SNP 46 to SNP 76, which clearly forms a large block as indicated by Daly *et al*. (2001). The partition obtained by the ME method is closer to that by Daly than that by Koivisto *et al*. (2003). The low-diversity intra-block entropy optimization will further help to infer missing SNPs in the next step.

3.3 Inference of missing SNPs

First we randomly generate missing SNPs on the preprocessed haplotype data set. We partition the haplotypes into blocks and infer every missing SNP locally within the block. We run the EM-like iterative process on the entire data set. At each run, we choose a missing SNP, identify the block it is located in, calculate its probability of being a majority, and update its value. Then the error rate is computed according to the original data set.

With 1% missing rate on Daly's data and given partitions, the error rate is 9.98% for Daly *et al*. (2001), 9.60% for Koivisto *et al*. (2003) and 6.03% for the ME method. It should be noted that the error rate for Anderson's partition is 5.65%, lower than the three methods. It is because the MDL method contains a step to infer missing SNPs jointly with the block partition. The results show that the block partition obtained by the ME method yields the lowest overall error rate. In conclusion, we show that (1) the lower the diversity is within a block, the more powerful the inference can be, and (2) the intra-block entropy measure is a better measure than others in terms of defining the block structure. In addition, the inferred missing SNP then can be used to generate better haplotype block partition, and this process can iterate. Among all measurements, only the ME method fits into this iterative scheme due to its focus on minimizing block entropy, which matches the goal of the inference of missing SNPs.

3.4 Iterative partition and inference

We use the data sets in both Daly *et al*. (2001) and Patil *et al*. (2001) to test the IPI system. The iteration begins as we assume every missing SNP to be the majority at its location. This assignment is called 'majority assignment'. The majority assignment uses only the single location to compute the likelihood, without using information from its neighborhood. The error rate of the majority assignment is shown for comparison. Then we partition the data set into haplotype blocks using current assignment. After partitioning, the assignment is updated for every missing SNP. The step of the inference of missing SNPs will converge. In the next round of iteration, the updated assignment is again employed for partitioning. We compare the performance of the iterative system in Tables 2 and 3, each with 1% missing rate and five rounds of iteration.

Table 2. IPI for Daly's data, with 1% missing rate (majority assignment error rate: 16.95%)

Round	Error rate (%)	Number of blocks	Total block entropy
1	6.02	11	33.05
2	5.08	11	32.37
3	5.08	11	32.12
4	5.08	11	32.12
5	5.08	11	32.12

Table 3. IPI for the selected region (8461–8720) in Patil's data, with 1% missing rate (majority assignment error rate: 19.23%)

Round	Error rate (%)	Number of blocks	Total block entropy
1	9.61	42	68.49
2	7.69	41	66.71
3	7.69	40	66.25
4	7.69	40	66.25
5	7.69	40	66.25

Table 4. Error rates for the inference of missing SNPs according to different missing rates, using full data in Daly *et al*. (2001)

Missing rates	1%	5%	10%
Majority assignment error rates	16.95%	19.01%	19.35%
1st round	6.02%	8.51%	8.73%
2nd round	5.08%	7.75%	8.02%
3rd round	5.08%	7.38%	8.02%
4th round	5.08%	7.34%	7.98%
5th round	5.08%	7.34%	7.98%

Table 5. Error rates for the inference of missing SNPs according to different missing rates, using region 8461–8720 in Patil *et al*. (2001)

Missing rates	1%	5%	10%
Majority assignment error rates	19.23%	20.38%	23.08%
1st round	9.61%	10.38%	9.81%
2nd round	7.69%	9.62%	9.62%
3rd round	7.69%	9.23%	9.42%
4th round	7.69%	9.23%	9.42%

Both results show a significant improvement due to the iteration. Since the data set in Patil *et al*. (2001) has a much smaller number of samples (20), its error rate is notably higher than that in Daly *et al*. (2001). It should also be noted that 5.08% error rate for Daly's data is lower than the 5.85% error rate obtained through the partition by Anderson and Novembre (2003). In addition, we test our algorithm for different missing rates, 1, 5, and 10%, on the data sets in Daly *et al*. (2001) and Patil *et al*. (2001). The results are shown in Tables 4 and 5. It is obvious that the inference becomes more difficult as there are more missing SNPs.

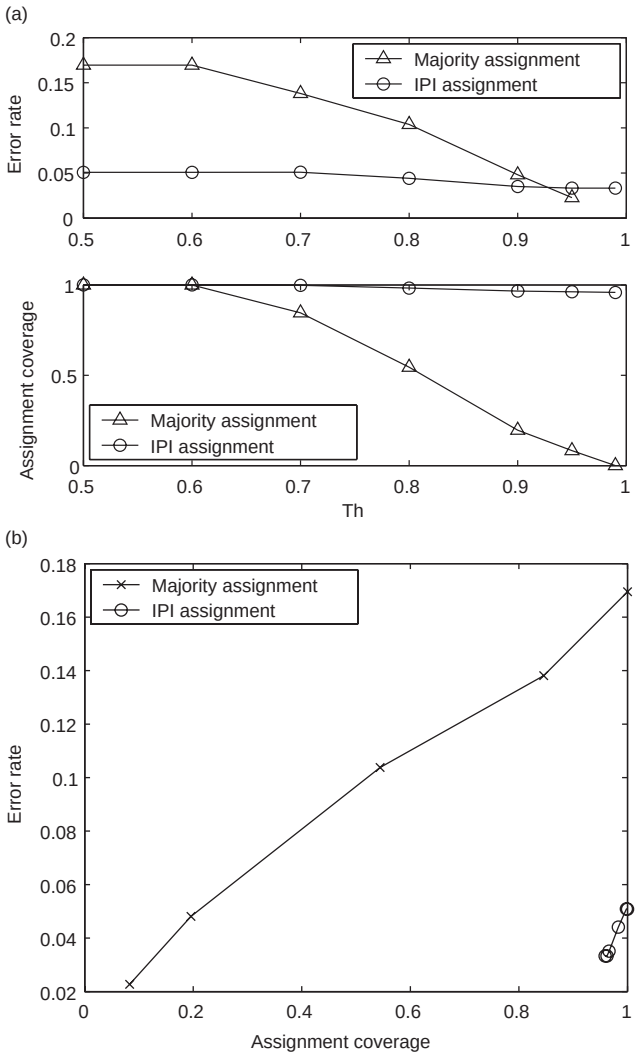

Fig. 5. Using threshold T_h to assign values to missing SNPs (hard decision): (**a**) error rate versus T_h in the upper plot, assignment coverage versus T_h in lower plot; (**b**) error rate versus assignment coverage.

3.5 Hard and soft decisions in iterations

As described in the previous section, the inference of each missing SNP is stored as the probability of being the majority at its location. This is called a *soft decision*, because the decision is in probability format and not yet finalized. When we partition the haplotypes into blocks, we have to make a *hard decision* for the missing SNPs, claiming each of them to be either the majority or the minority. Upon making the hard decision, we can define a threshold to cut-off. Let T_h denote this threshold. If the probability of being the majority for a missing SNP is $\geq T_h$, we assign it to the majority. Similarly, if it is $<1-T_h$, we assign it to the minority. In our previous tests, T_h was set to 0.5, and every missing SNP had an assignment. During the iteration, we perform the hard decision for two purposes: partition and inference, separately. At the partition, every missing SNP is assigned a value: if we fail to assign a missing SNP a value using the current T_h, we use 0.5 as the threshold instead for this particular SNP. Finally, when we calculate the error rate, we only consider the

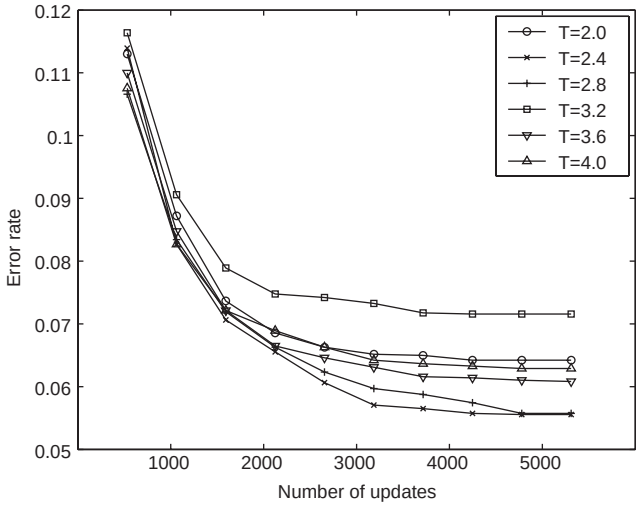

Fig. 6. The error rate of the inference of missing SNPs converges after runs of the iterative process that minimize the entropy using different T-values on Daly's data set. Different T values result in different error rates converged.

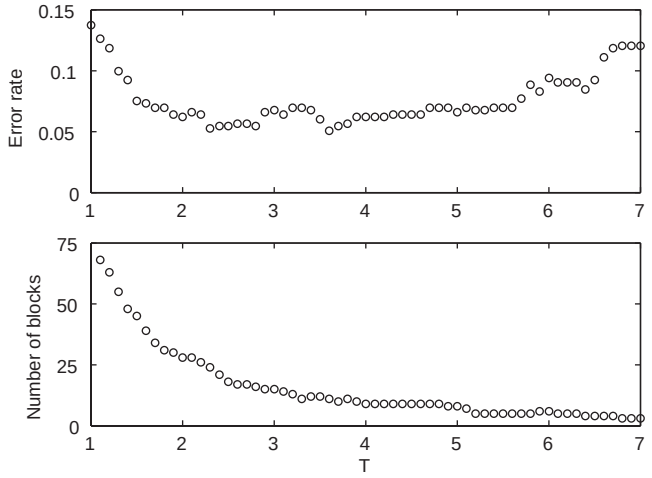

Fig. 7. The selection of T-values. Upper plot: inference error rate versus T; lower plot: number of haplotype blocks versus T.

missing SNPs with assigned values according to the T_h while keeping the un-assigned SNPs out of the assignment coverage.

In Figure 5, we compare the majority assignment and our IPI system on Daly's data set with 1% missing rate. In the upper plot of Figure 5(a), we see that both the majority and the IPI assignments improve as T_h increases. The majority assignment yields a lower error rate when $T_h \geq 0.95$. However, in the lower plot we can also observe that this low error rate has a very poor coverage, $<10\%$, whereas the IPI assignment has more than 95% coverage in all T_h's. The hard decision behavior of both assignments can be concluded in Figure 5(b). The T_h can serve as an index for the confidence level. Our IPI assignment covers 98% of the missing SNPs to a confidence level of 0.999, at an error rate of approximately 3%. At the same error rate, there are only 10% of the missing SNPs covered for the majority assignment, with a confidence level of 0.95.

4 DISCUSSION

Our three information quantity-based partitioning methods are targeted to minimize the total block entropy, to maximize the total joint entropy, or to minimize the total mutual information. Although all the three properties reflect some dimensions of haplotype blocks, they are not equally important. If we assume that the partition in Daly *et al*. (2001) is true, the partition by the ME method is closer to it than the two others. This suggests that the property of low intra-block diversity is a better measure of a haplotype block. In a future study, the two other properties of haplotype blocks can be applied to either synthesize a new cost function, or shape the inter-block requirement for the ME method.

The T-value is the maximum entropy allowed within a block. The selection of the T-values affects the resulting partitions, which in turn affects the accuracy of the inference of missing SNPs. An example is shown in Figure 6 using the data set in Daly *et al*. (2001) with 1% missing SNPs. We run the iterative process on the entire data set. At each run, we choose a missing SNP, identify the block it is located in, calculate its probability of being a majority, and update its value. Figure 6 shows the convergence of the error rate after runs of the iterative process, according to different entropy thresholds T.

The preprocessed data set in Daly *et al*. (2001) has an entropy of 8.71261. We test the range $1 \leq T \leq 7$ on a 1% missing rate, with five IPI iterations. The corresponding error rates and the number of haplotype blocks are shown in Figure 7. When T is too small, the data set is partitioned into many small blocks. In this case, not enough neighboring information is employed to infer missing SNPs. However, when T is too large, the data set is partitioned into too few blocks. Thus, distant neighboring information may be involved in the inference. Normally, the proper T can be selected with the help of an entropy map. In this case, $2.2 \leq T \leq 3.8$ is a suitable range.

ACKNOWLEDGEMENTS

This work is partially supported by NIH Center of Excellent in Genomic Sciences: Implications of Haplotype Structure in the Human Genome, Grant No. P50 HG002790.

SUPPLEMENTARY DATA

Supplementary data for this paper are available on *Bioinformatics* online.

REFERENCES

Anderson,E. and Novembre,J. (2003) Finding haplotype block boundaries by using the minimum-description-length principle. *Am. J. Hum. Genet.*, **73**, 336–354.

Daly,M. *et al*. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.

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

Hedrick,P. (1987) Gametic disequilibrium measures: proceed with caution. *Genetics*, **117**, 331–341.

Koivisto,M. *et al*. (2003) An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. *Pac. Symp. Biocomput.*, **8**, 502–513.

Lin,S. *et al*. (2002) Haplotype inference in random population samples. *Am. J. Hum. Genet.*, **71**, 1129–1137.

Niu,T. *et al*. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157–169.

Nothnagel,M. *et al*. (2003) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Human Heredity*, **54**, 186–198.

Patil,N. *et al*. (2001) Blocks of limited haplotype diversity revealed by highresolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Qin,Z. *et al*. (2002) Partitioning-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide Ploymorphisms. *Am. J. Hum. Genet.*, **71**, 1242–1247.

Stephens,M. *et al*. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

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

Zhang,K. *et al*. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.