COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 9, No. 3, pp. 253-268, 2009
© 2009 International Press
002

AN INTEGER PROGRAMMING APPROACH FOR THE SELECTION OF TAG SNPS USING MULTI-ALLELIC LD*

YANG-HO CHEN† AND TING CHEN†

Abstract. Single Nucleotide Polymorphisms (SNPs) are common among human populations. SNPs that are proximally located within a small human chromosome region are generally strongly correlated that a subset of SNPs, termed tag SNPs, can provide enough information to infer neighboring SNPs. Such correlations are generally known as linkage disequilibrium (LD) and are measured either pair-wise, such as r^2, or multi-to-one (multi-marker). For any given set of SNPs, a variety of algorithms have been proposed to identify a subset of tag SNPs by which the remaining SNPs can be inferred. This paper focuses on finding that number of tag SNPs from which remaining SNPs can be inferred through multi-allelic LD or pair-wise LD with a pre-defined r^2 threshold. We call this the optimal tag SNP selection problem. Although this problem is theoretically NP-hard, it can be formulated as an integer programming (IP) problem under a certain constraint, and the optimal solution can be efficiently found by our newly developed IPMarker program. In addition, the flexibility of the computational framework allows us to formulate and solve the problem of finding common tag SNPs for multiple populations that have different LD patterns. Various datasets, including ENCODE and the Major Histocompatiability Complex (MHC) region, were used to evaluate the performance of IPMarker. We also extended IPMarker to the whole genome HapMap Phase I data. Results showed that IPMarker significantly reduces the number of tag SNPs required when compared to the most widely used program, Haploview, although a significant longer running time is required. Thus, overall, genotyping a selected set of tag SNPs is the most cost-effective way to conduct large-scale genome-wide association studies.

Keywords: Tag SNP, Multi-marker, Integer Programming, Power Analysis

1. Introduction. Single nucleotide polymorphisms (SNPs)[1] are widely used as genetic markers for studies of population genetics and complex diseases. Through the HapMap Project, more than three million common SNPs have been found in the human genome by sequencing and comparing the chromosomes of hundreds of individuals from multiple populations [2-5]. From these data, linkage disequilibrium (LD), which is the association of alleles at two or more SNP loci, can be observed, especially for SNPs proximally located on the human chromosome. As a result, once researchers have genotyped a selected subset of SNPs, called tag SNPs, the alleles at other SNP loci can be inferred through the neighboring tag SNPs [6-11]. This strategy can significantly reduce the genotyping cost.

Many methods have been proposed for tag SNP selection with the aim of improving the power of inferring neighboring SNPs or minimizing the size of the tag SNP set. Most of them are based on the pair-wise LD between two SNP loci [2, 5, 17, 28]. The pair-wise LD is generally measured by the r^2 score which is equal to the square

*Dedicated to Michael Waterman on the occasion of his 67th birthday.

†Program in Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089-2910 USA.

of the correlation coefficient between two SNP loci. Equation 1 shows the definition of the r^2 score over two loci A and B,

$$r^2 = \frac{\mathrm{Cov}(A,B)^2}{\mathrm{Var}(A)\mathrm{Var}(B)} = \frac{(P_{AB} - P_A P_B)^2}{P_A(1-P_A)P_B(1-P_B)}, \tag{1}$$

where P_A is the probability of the major alleles at locus A and P_B is the probability at locus B.

In practice, tag SNPs are selected with some pre-defined r^2 threshold, which guarantees that at least one of the tag SNPs will be associated with another SNP locus. Theoretically, the problem of finding the minimum set of tag SNPs with a pre-defined r^2 threshold is NP-hard, which can be proved by a reduction from the "Set Cover" problem. Therefore, it is impossible to develop an algorithm to find the optimal solution for every case in polynomial time. As a result, many heuristic methods, including greedy, entropy maximization or principle component analysis [8, 10, 18-30], have been proposed to solve the problem in practice.

On the other hand, multiple efforts [10, 24, 26, 27, 30] have been made to extend the pair-wise LD to the multi-allelic LD in order to increase the statistical power of inference. That is, a combination of SNPs, also referred to as haplotypes, can be used to infer the alleles of other SNP loci with a higher accuracy. Earlier works on "block-based tagging" [12-17] are, in fact, based on multi-allelic LD. They first partition long haplotypes into blocks using some LD measures and then select tag SNPs from each haplotype block such that other SNPs can be predicted directly from the tag SNPs. Table 1 shows an example of multi-allelic LD for four haplotypes consisting of three SNPs. In this example, SNP A and SNP B together completely determine SNP C. We define such a pattern as a 2-to-1 perfect LD. For comparison, SNP A or SNP B alone can only determine SNP C with 80% accuracy.

TABLE 1

Four haplotypes, H_1, H_2, H_3, and H_4, with frequencies 0.4, 0.2, 0.2, and 0.2, respectively, are observed for SNP A, SNP B and SNP C. 0 represents the major allele, and 1 is the minor allele. SNP A and SNP B jointly determine SNP C.

Haplotype (freq)	SNP A	SNP B	SNP C
H_1(0.4)	0	0	0
H_2(0.2)	0	1	0
H_3(0.2)	1	0	0
H_4(0.2)	1	1	1

The most notable work using pair-wise and multi-allelic LD for selecting tag SNPs is a program called Haploview [26], the official tagging tool used in the International HapMap project. However, as the number of possible haplotypes to be tested by the LD statistic grows exponentially by the number of SNPs, Haploview slows down

dramatically. Consequently, in order to gain efficiency, Haploview must operate under several restrictions. Nonetheless, the improved prediction accuracy using multi-allelic LD does offer a significant advantage in that fewer tag SNPs are needed to achieve the same prediction accuracy by the increase in correlations among multiple SNPs. Using the example shown in Table 1, if the r^2 threshold is set as 0.9, we only need two tag SNPs (SNP A and SNP B) using multi-allelic LD, but we need all three tag SNPs if we only use pair-wise LD. Huang and Chao [30] formulated a multi-allelic LD based on the tag SNP selection problem (MTMH) as follows: given a set of SNPs, find a minimum subset of tag SNPs which defines a set of haplotypes completely predictive of the alleles of all other SNPs. They divided this problem into three sub-problems, and they solved each of them using exact and approximation algorithms.

Theoretically, we can extend the aforementioned 2-to-1 perfect LD to k-to-1 perfect LD. However, in actual practice, as k increases, identified k-to-1 associations will also increase, but as a result of random chance because of the limited sample size. Therefore, k-to-1 associations could have a counter-effect because of false associations, and this could lead to increased errors in the inference of other SNPs. To reduce such random associations, we consider associations in local regions and further restrict them to the following two types: the 2-to-1 perfect LD and the pair-wise LD with a pre-defined r^2 threshold. We formulate these two types of LD into integer programming and develop a program called IPMarker. Our computational framework can easily incorporate k-to-1 associations if needed, but we find that this will significantly increase the computational time and inference errors with little gain. In fact, our experimental results show that most cases having k-to-1 perfect LD can be recovered by chaining multiple cases having 2-to-1 perfect LD. IPMarker is also extended to select a common subset of tag SNPs for multiple populations. It is well known that different populations have different LD patterns in the same genetic regions. Using common tag SNPs to allow different LD associations in different populations will simplify the genotyping processes for association studies with multiple populations [31].

2. Methods.

2.1. Algorithm for SNP Inference. We will only use the following two types of LD for our tag selections: the 2-to-1 perfect LD and the pair-wise LD. We propose the following two-step algorithm to infer other untyped SNPs using tag SNPs.

1. Use 2-to-1 perfect LD to determine a subset of untyped SNPs, and then
2. Use pair-wise LD to infer the remaining untyped SNPs using both the tag SNPs and the determined SNPs (in Step 1).

2.2. Problem Definition. Based on the above algorithm, we formulate the Di-Markers-Haplotype-Tagging (DMHT) problem as follows:

DMHT: Given a dataset S of m haplotypes over n SNPs and a pair-wise LD threshold R, find a minimum subset of tag SNPs such that the remaining SNPs can

be inferred through 2-to-1 perfect LD and pair-wise LD.

2.3. Association Graph. An association graph G is constructed to represent the two kinds of LD. Each node in G represents a SNP or a combination of two SNPs. There are three types of edges, one representing pair-wise LD and the other two representing 2-to-1 perfect LD. In the example shown in Figure 1, node C1 represents the combination of SNP1 and SNP2, and node C2 represents the combination of SNP3 and SNP5. There are three types of edges in G, as shown in Figure 1. A dash-line edge represents the pair-wise LD between SNP6 and SNP7. The two hyper-directed edges, (SNP1 + SNP2) → C1 and (SNP3 + SNP5) → C2, represent the combinations of SNPs for C1 and C2. Three directed edges, C1 → SNP4, C1 → SNP5 and C2 → SNP6, represent the perfect LD that a SNP can be determined completely by a combination of two SNPs.

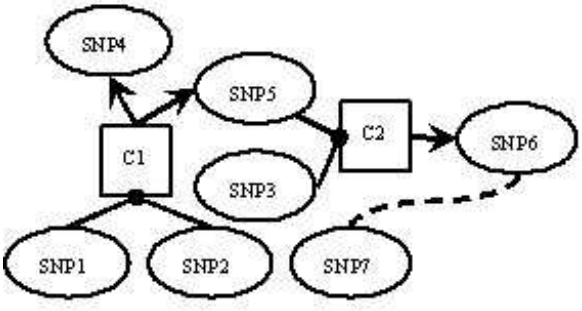

FIG. 1. *An association graph for 7 SNPs. Two nodes, C1 and C2, are introduced to represent the combinations of {SNP1, SNP2} and {SNP3, SNP5}. This graph includes three 2-to-1 associations, {SNP1, SNP2}→ SNP4, {SNP1, SNP2}→SNP5, and {SNP3, SNP5} → SNP6, and one pair-wise association between SNP6 and SNP7.*

Note that chain inference is allowed through the 2-to-1 associations. For example, the chaining of two 2-to-1 associations, {SNP1, SNP2}→ SNP5 and {SNP3, SNP5} → SNP6, leads to a 3-to-1 association, {SNP1, SNP2, SNP3} → SNP6. In this way, 2-to-1 associations can be extended to multi-to-one associations. In the second step, the tag SNPs and the perfectly inferred SNPs are used to infer the alleles of the remaining SNPs through the pair-wise LD (i.e., SNP 7 can be inferred from SNP 6).

2.4. Formulation of the Integer Programming Problem. We formulated the DMHT problem as an integer programming problem (IP) problem. Assume that we have n SNPs: SNP1, SNP2…SNPn. We define three kinds of binary variables, $\{T_i\}$, $\{P_i\}$, and $\{C_j\}$ as the following:

- $T = \{T_1, T_2 \ldots T_n\}$: $T_i = 1$ if SNPi is selected as a tag SNP. Otherwise, $T_i = 0$.

- $C = \{C_1, C_2, \dots\}$: Each C_j corresponds to a combination of two SNPs. $C_j = 1$ if both its members, say SNP1 and SNP2, are either selected as tag SNPs or are perfectly determined. That is, each of its SNP members is either selected as a tag SNP or determined by some combination k, with $C_k = 1$. The latter case is based on the chain inference of 2-to-1 associations.
- $P = \{P_1, P_2, \dots, P_n\}$: $P_i = 1$ if SNPi is a tag SNP or can be determined perfectly. This happens if SNPi is selected as a tag SNP, $T_i = 1$, or a "perfectly determined" SNP by some combination where $C_k = 1$ and $C_k \rightarrow$ SNPi.

Note that the definitions of C and P are recursive, which accounts for the circles in the association graph. We will discuss the circular structure of the association graph later. Here, we define two types of symbols:

1. $\{S(1), S(2), \dots, S(n)\}$: S(i) is the set of SNPs whose pair-wise LD with SNPi is above the pre-defined threshold R.

2. $\{K(1), K(2), \dots, K(n)\}$: K(i) is defined as the set of SNP combinations which can determine SNPi using "2-to-1" perfect associations.

In Figure 1, DMHT is solved by selecting only SNP1, SNP2 and SNP3 as tag SNPs. Correspondingly, $T_1 = T_2 = T_3 = 1$, $T_4 = T_5 = T_6 = T_7 = 0$, and $P_1 = P_2 = P_3 = 1$ and $C_1 = 1$ by definition. Through C_1, SNP_4 and SNP_5 can be perfectly determined; $P_4 = P_5 = 1$. Thus $C_2 = 1$ because $T_3 = 1$ and $P_5 = 1$. Through C_2, SNP6 can be determined, $P_6 = 1$, and SNP 7 can be inferred from SNP6 through the pair-wise LD, but $P_7 = 0$.

The DMHT problem can be formulated as an integer programming problem as follows:

$$(2) \quad \text{Minimize} \sum_{i=1}^{n} T_i$$

Subject to

I. $\forall C_k : P_{k_1} \geq C_k,$ and $P_{k_2} \geq C_k, k = 1, 2, \cdots,$ where$C_k = \{\text{SNP}k_1, \text{SNP}k_2\}$.

II. $\forall P_i : T_i + \sum_{C_j \in K(i)} C_j \geq P_i, i = 1, \cdots, n,$

III. $\forall \text{SNP}i : \sum_{\text{SNP}j \in \text{S}(i)} P_j \geq 1, i = 1, \cdots, n,$

IV. Binary Constraints: $T_i, C_k, P_i = \{0, 1\}, i = 1, \cdots, n, k = 1, 2, \cdots$

The target function is to minimize the number of selected tag SNPs. There are four types of inequalities in the IP formulation: I, II, III and IV. Type I inequalities indicate that $C_k = 1$ if, and only if, all of its SNP members are determined. Type II inequalities say that $P_i = 1$ only if SNPi is a tag SNP or it can be determined by some combination. Type III inequalities force each SNP to be inferred by itself or its neighboring SNPs through pair-wise LD. Type IV inequalities limit each variable

to a binary value. The IP problem can be efficiently solved by integer programming software [32]. In theory, integer programming is also NP-hard; however, in practice, a combination of the linear programming solution plus rounding and a branch-and-bound strategy can efficiently find the optimal or sub-optimal solution. Table 2 shows the IP formulation of the example in Figure 1.

TABLE 2
The set of IP constraints for the example in Figure 1.

I	II	III	IV
$P_1 \geq C_1$; $P_2 \geq C_1$	$T_1 \geq P_1$; $T_2 \geq P_1$	For $i = 1$ to 5	For $i = 1$ to 7
$P_3 \geq C_2$; $P_5 \geq C_2$	$T_3 \geq P_3$; $C_1 + T_4 \geq P_4$	$P_i \geq 1$	For $j = 1, 2$
	$C_1 + T_5 \geq P_5$	$P_6 + P_7 \geq 1$	T_i, P_i and $C_j = 0$ or 1
	$C_2 + T_6 \geq P_6$		

2.5. Cycles and Cycle Breaking. Unfortunately, this IP formulation may encounter problems when a cycle formed by chaining multiple 2-to-1 associations in the association graph exists. Figure 2 shows an example where a cycle is formed by SNP1, C_1, SNP2, C_2, SNP3, and C_3. The IP solution does not return the correct solution. To solve this problem, a modified depth first search (DFS) algorithm can be used to break all cycles. In the DFS, we remove one edge each time we explore a new edge that causes a cycle. Note that breaking all of the cycles by removing a minimum number of edges is an NP-hard problem and thus has no polynomial time solution. In addition, removing the minimum number of edges does not guarantee finding a minimum set of tag SNPs. Therefore, we applied several cycle-breaking heuristics and found that the DFS-based cycle removal method combined with IP performs extremely well in practice.

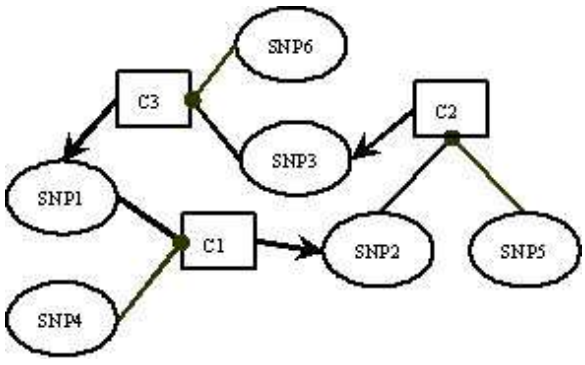

FIG. 2. *The integer programming solver will set the binary variables $T_4 = T_5 = T_6 = 1$ and $C_1 = C_2 = C_3 = 1$ to minimize the target function. However, at least one SNP out of SNP1, SNP2 and SNP3 must be selected as the tag SNP.*

THEOREM. *The IP returns the optimal solution to the DMHT problem if there is no cycle in the association graph.*

Proof. If there are cycles in the association graph, the solutions returned by the IP may not be correct as we have shown in Figure 2. However, if IP returns a correct solution, then it is optimal.

2.6. Selecting Common Tag SNPs for Multiple Populations. The LD patterns vary in different populations. This inspires us to select a common set of tag SNPs and apply different LD associations in different populations to infer the alleles of untyped SNPs. The IP formulation is similar to the previous one, but each population has its own Type I, II, and III inequalities. The following is an example of the IP formulation for two populations, where $P^{(1)}$ and $C^{(1)}$ are parameters for population 1, $P^{(2)}$ and $C^{(2)}$ for population 2, and $K^{(1)}$, $S^{(1)}$ and $K^{(2)}$, $S^{(2)}$ are pre-calculated. This formulation can be easily extended to multiple populations.

$$
\begin{aligned}
&\text{Minimize} \sum_{i=1}^{n} T_i \\
&\text{Subject to} \\
&\textit{Population } (1): \\
&\quad P_{k_1}^{(1)} \geq C_k^{(1)} \text{ and } P_{k_2}^{(1)} \geq C_k^{(1)},\ k = 1, 2, ... \\
&\quad T_i + \sum_{C_k^{(1)} \in K^{(1)}(i)} C_k^{(1)} \geq P_i^{(1)},\ i = 1, ..., n \\
&\quad \sum_{SNP j \in S^{(1)}(i)} P_j^{(1)} \geq 1,\ i = 1, ..., n \\
&\textit{Population } (2): \\
&\quad P_{k_1}^{(2)} \geq C_k^{(2)} \text{ and } P_{k_2}^{(2)} \geq C_k^{(2)},\ k = 1, 2, ... \\
&\quad T_i + \sum_{C_k^{(2)} \in K^{(2)}(i)} C_k^{(2)} \geq P_i^{(2)},\ i = 1, ..., n, \\
&\quad \sum_{SNP j \in S^{(2)}(i)} P_j^{(2)} \geq 1,\ i = 1, ..., n.
\end{aligned} \tag{3}
$$

2.7. Scaled up for the Whole Chromosome. To manage whole genome haplotype data, SNPs are divided into different sizes of non-overlapping blocks. The conventional LD span is about 400kb – 500kb, as now used in Haploview [27]. This genome distance roughly covers about 200 SNPs in HapMap Phase I data [2]. We do not use longer LD spans because errors increase as the LD decreases. We have also attempted other strategies to handle the whole genome data, i.e., dividing the

TABLE 3
Performance of IPMarker on different datasets.

Data	#Samples	SNP Density 1/kb	# SNPs	# tag SNPs ($r^2 = 1.0$)	#Tag SNPs/ # SNPs ($r^2 = 1.0$)	# tag SNPs ($r^2 = 0.8$)	#Tag SNPs/ # SNP ($r^2 \geqq 0.8$)
ENCODE (Average)	478	1/0.5kb	9,523	2,592	27.21%	1,945	20.42%
MHC (Average)	478	1/1.3kb	6,298	1,510	23.98%	1,152	18.3%
Chr2 (CEU)	180	1/3.7kb	69,753	29,968	42.96%	21,557	30.90%
Chr4 (CEU)	180	1/3.5kb	48,952	20,744	42.38%	16,338	33.37%
Chr6 (CEU)	180	1/3.2kb	53,458	22,070	41.28%	16,970	31.74%
Chr8 (CEU)	180	1/3.0kb	60,203	21,181	35.18%	15,432	25.63%
Chr10 (CEU)	180	1/3.5kb	38,940	17,810	45.74%	13,890	35.67%

whole genome into overlapping fragments, but the results do not show significant improvement.

3. Experimental Results.

3.1. Datasets. The following datasets are used in this study:

1. Phased ten ENCODE regions with minor allele frequency (MAF) $> 5\%$ (478 haplotype samples with 9,523 common SNPs)

2. Phased Major Histocompatibility Complex (MHC) SNP data [33] (478 haplotype samples with an average of 6,298 SNPs in four populations)

3. Phased central European human Chromosome data selected from phase I HapMap data (Chromosomes 2, 4, 6, 8 and 10 from 180 haplotypes with 271,337 SNPs per haplotype)

We first remove SNPs that have the same distribution as others. This simple filtering process significantly reduces the number of SNPs in the datasets. For example, it reduces the size of the ENCODE data down to about 1/2 to 1/3 of the original size.

3.2. Performance of IPMarker. A software package lp_solved[32] is called by IPMarker to solve the integer programming problem. The r^2 threshold is set as 0.8 and 1.0 in all tests. Table 3 shows the SNP density, the number of SNPs, the number of tag SNPs found by IPMarker, and the compression ratio, which is the ratio of the number of tag SNPs over the total number of SNPs, on the haplotype data from ENCODE, MHC and the five selected human chromosomes from the central European population. As expected, the compression ratio decreases as the SNP density increases. Both the ENCODE data and the MHC data have higher SNP densities than the human chromosome data, while, at the same time, they show lower compression ratios.

3.3. Comparison between IPMarker and Haploview. IPMarker is compared with Haploview using the ENCODE (Table 4), the MHC (Table 5) and the phase I HapMap human chromosome data (Table 6) as the test data. Under the

TABLE 4

Number of tag SNPs selected by IPMarker and Haploview on ten ENCODE datasets (CEU).

ENCODE Data Set (CEU)	Original # SNPs	Haploview (Pair-wise LD)	Haploview (multi-allelic LD)		IPMarker (2-to-1 LD)	
		$r^2 = 1.0$	$r^2 = 1.0$	$r^2 \geq 0.8$	$r^2 = 1.0$	$r^2 \geq 0.8$
ENm010	618	307	251	178	181	145
ENm013	1,042	264	201	140	163	131
ENm014	1,124	319	264	200	209	173
ENr112	1,157	424	323	212	240	182
ENr113	1,393	403	321	215	236	185
ENr123	1,052	358	271	176	198	157
ENr131	1,221	457	357	235	266	208
ENr213	796	282	235	162	176	140
ENr232	614	293	257	185	194	141
ENr321	768	305	245	179	186	149
Total	9,795	3,414	2,725	1,882	2,049	1,611

same r^2threshold and LD span, IPMarker selected significantly fewer tag SNPs than Haploview 4.0, especially given high LD regions as input. The tradeoff for this better optimization is a significantly longer running time.

Table 4 compares the results of 10 ENCODE regions between Haploview and IPMarker. With r^2=1.0, IPMarker selects a total of 2,049 tag SNPs from 9,795 SNPs, about 25% fewer than the 2,705 tag SNPs found by Haploview using the multi-allelic LD. With a lower r^2 threshold of 0.8, IPMarker reduces the number of tag SNPs down to 1,611, or 78.6% of the 2,049 tag SNPs found using r^2=1.0. It is 14.4% less than the tag SNPs found by Haploview under the same threshold.

In Table 4, note that both Haploview and IPMarker demonstrate that significantly fewer tag SNPs will be needed if we use multi-allelic LD with r^2=1.0 rather than pair-wise LD with r^2=1.0. Haploview reduces the 3,414 tag SNPs found by the pair-wise LD method down to 2,725 (or 80.6%), while IPMarker has a bigger reduction at 60%.

Figure 3 compares the results of the MHC dataset between Haploview and IP-Marker. The SNP density of the MHC data is about 1.27 kb per SNP. Under the threshold r^2= 1.0, IPMaker selects an average of 1,510 tag SNPs for each of the four populations, about 18.5% fewer tag SNPs than the number required by Haploview using multi-allelic LD. By relaxing the r^2 threshold to 0.8, the number of tag SNPs found by IPMarker reduces to 1,152. Again, we observe a big reduction in the number

TABLE 5

Number of tag SNPs ($\times 10^3$) selected by IPMarker and Haploview on five HapMap chromosome datasets.

Tagging Method	LD used	Block Size	Threshold	# of tag SNPs (x 10^3)				
				Chr2	Ch4	Chr6	Chr8	Chr10
Haploview	pair-wise LD	200 SNPs	$r^2 = 1.0$	41.0	29.3	31.7	30.6	24.7
Haploview	Multi-allelic LD	800 SNPs	$r^2 = 1.0$	34.1	24.8	26.5	25.1	21.1
IPMarker	2-to-1 perfect LD	200 SNPs	$r^2 = 1.0$	30.0	20.7	22.1	21.2	17.8
Haploview	Multi-allelic LD	800 SNPs	$r^2 \geq 0.8$	22.7	17.1	18.1	16.2	14.6
Haploview	Multi-allelic LD	200 SNPs	$r^2 \geq 0.8$	22.9	17.2	18.2	16.4	14.7
IPMarker	2-to-1 perfect LD	200 SNPs	$r^2 \geq 0.8$	21.5	16.3	16.9	15.4	13.9
Total number of SNPs (x 10^3)				69.8	49.0	53.5	60.2	38.9

of tag SNPs using multi-allelic LD.

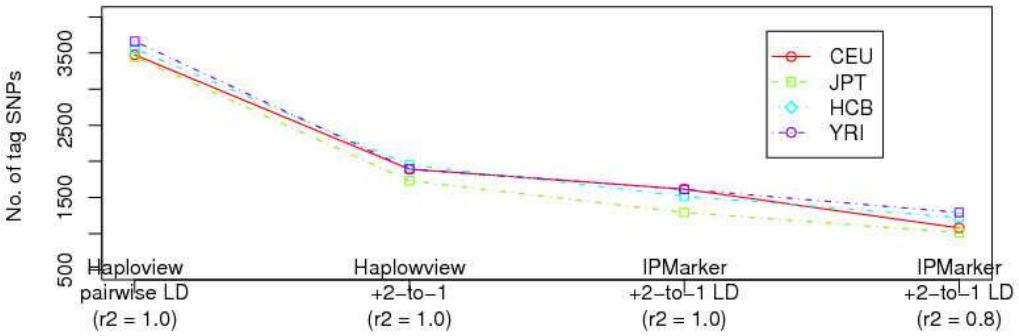

FIG. 3. *Comparison of the number of tag SNPs found by IPMarker and Haploview on the MHC dataset.*

Table 5 shows the results of five human chromosomes from the HapMap phase I dataset using Haploview and IPMarker. IPMarker consistently selects significantly fewer tag SNPs than Haploview under the same criteria. On average, IPMarker selects 16% fewer tag SNPs than Haploview with $r^2 = 1.0$, and 6% fewer tag SNPs with $r^2 \geq 0.8$. Note that both block sizes of 200 and 800 SNPs have been used in Haploview, but the results are similar. Again, we observe a big reduction in the number of tag SNPs using multi-allelic LD as opposed to pair-wise LD.

3.4. Evaluation of Prediction Power. We perform a 20-fold cross validation to assess the prediction error rate using the tag SNPs. That is, we use 95% of the samples to find tag SNPs and use the remaining 5% of the samples for the following test. If the selected tag SNPs in these test samples are genotyped, we then asked how well they could be used to predict untyped SNPs. The accuracy of the prediction is

evaluated by comparing the real alleles with the predicted alleles in the test dataset. The error rate is defined as the ratio of the number of prediction errors over the total number of SNPs in the test, including both the genotyped tag SNPs and the untyped SNPs. Thus, the error rate is proportional to the number of prediction errors and does not depend on the number of tag SNPs or the number of untyped SNPs. For comparison, we also implement a method to select tag SNPs using only pair-wise LD in IPMarker using a similar IP formulation.

Table 6 shows the prediction error rates using the multi-allelic LD and the pair-wise LD on the MHC data. Under the same r^2 threshold, the prediction error rates of the method using the multi-allelic LD are always higher than those using only the pair-wise LD: 0.44% *vs.* 0.11%, respectively, for $r^2 = 1.0$ and 0.83% *vs.* 0.57%, respectively, for $r^2 \geq 0.8$. This is because the number of tag SNPs selected using multi-allelic LD is much less than the number achieved by using pair-wise LD. The consequence is that many more untyped SNPs are predicted by the multi-marker tagging method, which drives up the prediction errors. However, with a comparable number of tag SNPs, multi-allelic LD has a lower prediction error rate. For example, with $r^2 = 1.0$, IPMarker finds an average of 1,509 tag SNPs using multi-allelic LD, compared to the 1,593 tag SNPs selected with the pair-wise LD with $r^2 \geq 0.8$, but the average prediction error rate for the multi-allelic LD is 0.44%, better than 0.57% for the pair-wise LD. As another example, with $r^2 \geq 0.8$, IPMarker finds an average of 1,136 tag SNPs using multi-allelic LD, compared to the 1,252 tag SNPs found by the pair-wise LD with $r^2 \geq 0.7$, but the average prediction error rate for the multi-allelic LD is 0.83%, better than 1.13% for the pair-wise LD. This line of evidence shows that multi-allelic LD methods have a higher prediction power than those of pair-wise LD.

3.5. Selecting a Common Set of Tag SNPs for Multiple Populations. In many association studies, it is important to determine whether the tag SNPs selected from one or a limited number of sample populations, such as those genotyped in the HapMap project, can be used in the populations being studied without losing significant prediction power.

Table 7 shows the prediction error rates in all pair-wise inter-population tests using tag SNPs selected on the basis of the LD in one population (rows) to predict alleles of other SNPs in another population (columns). It is clear that the intra-population prediction error rate (always $< 0.5\%$) is much smaller than the inter-population prediction error rate (always > 6 %). Therefore, the results discourage using tag SNPs selected from one population to be used in another population.

A better strategy is to mix haplotype data samples from all populations, treating them as if they were from one population, and then applying an algorithm to select a common set of tag SNPs, which can be used to design a single SNP chip for all studies involving multiple populations. Since tag SNPs are selected to capture the common

TABLE 6

Number of tag SNPs (x 10^3) selected by IPMarker and Haploview on five HapMap chromosome datasets.

IPMarker		CEU 180 Haps 6150 SNPs		JPT 90 Haps 6391 SNPs		HCB 88 Haps 6380 SNPs		YRI 120 Haps 6270 SNPs		Average
LD used	r^2 Threshold	Error Rate	Ave # of Tag SNPs	Error Rate	Ave # of Tag SNPs	Error Rate	Ave # of Tag SNPs	Error Rate	Ave # of Tag SNPs	Error Rate
2-to-1	1	0.42%	1641.5	0.47%	1359.0	0.38%	1499.4	0.47%	1587.4	0.44%
2-to-1	0.8	0.80%	1055.0	0.82%	1002.0	0.87%	1217.0	0.82%	1271.0	0.83%
pair-wise	1	0.09%	2535.0	0.14%	2086.0	0.10%	2348.0	0.10%	2616.7	0.11%
pair-wise	0.8	0.64%	1399.7	0.55%	1411.2	0.57%	1692.6	0.51%	1871.3	0.57%
pair-wise	0.7	1.22%	1104.9	0.97%	1114.1	1.05%	1275.8	1.29%	1513.8	1.13%

TABLE 7

Prediction error rates using tag SNPs selected in one population (rows) to predict untyped SNPs in another population (columns) using $r^2 = 1$ and 2-to-1 perfect LD.

Testing **Training**	CEU	JPT	HCB	YRI
CEU	0.42%	7.23%	6.26%	7.74%
JPT	9.32%	0.47%	4.85%	10.34%
HCB	7.34%	4.39%	0.38%	8.37%
YRI	6.73%	7.52%	6.93%	0.47%

LD patterns in the mixed populations, we call this the "mixing" approach. In this paper, we propose anther approach called the "splitting" approach that selects a common set of tag SNPs, but infers untyped SNPs using different LD patterns in each population separately. The "splitting" approach requires knowing the population for each sample in order to apply the proper LD pattern. In fact, in most association studies, such information is known. Table 8 shows the comparison of these two approaches on two populations of the MHC data using IPMarker. The result shows that for any two populations, the "splitting" approach selects an average of 37.5% fewer tag SNPs than the "mixing" approach, but the average prediction error rate is only slightly higher, 0.47% *vs.* 0.33%.

3.6. Running Time. With the r^2 threshold equal to 1.0, IPMarker runs as fast as Haploview. Nevertheless, the running time becomes slower for smaller thresholds. It is also clear that the high LD regions need a longer running time. For most of the 200 SNP fragments in the HapMap Phase I data, the optimal IP solution can be found within 10 minutes, while some other fragments take hours. Thus, the running time of selecting tag SNPs for the whole chromosome takes about 40-70 hours on a desktop computer.

TABLE 8

Comparisons of the "mixing" approach (results on the left of "/") and the "splitting" approach (results on the right of "/") for two populations (one in the column and one in the row) on the MHC data.

Mixing/Splitting	CEU		JPT		HCB	
$R^2 = 1.0$	Error Rate	# tag SNPs	Error Rate	# tag SNPs	Error Rate	# tag SNPs
JPT	0.34% / 0.37%	2542 / 1995				
HCB	0.35% / 0.36%	2525 / 1945	0.61% / 0.33%	2333 / 2125		
YRI	0.21% / 0.46%	2825 / 1710	0.22% / 0.46%	2757 / 1854	0.25% / 0.47%	2804 / 1987

4. Conclusion. With the rapid development of sequencing techniques [34], many scientists think that selecting tag SNPs for genotyping is not necessary. However, we argue that genotyping SNPs on a genome-wide scale [35] still costs much less than sequencing the whole genomes. With accumulated haplotype information and techniques for designing a dense genotyping array, tag SNPs provide a cost-effective way to capture most of the information needed in a large-scale association study, especially in a well-studied sub-population. Using programs like Phase[36, 37] to directly evaluate genotype data is certain to be the direction of the future. The source code and the preliminary documents are available at http://code.google.com/p/ipmarker.

Acknowledgements. We thank Dr. Yao-Ting Huang for explaining the idea of the multi-marker selection. This work is supported by NIH grants P50 HG 002790 and R01 LM008991-01 and the Alfred P. Sloan Research Fellowship.

REFERENCES

[1] A.J. BROOKES ed. *The Essence Of SNPs.* Gene. 1999. 234.

[2] *A haplotype map of the human genome.* Nature, 437:7063(2005), pp. 1299-1320.

[3] K.A. FRAZER ET AL., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 449:7164(2007), pp. 851-861.

[4] *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 306:5696(2004), pp. 636-640.

[5] M. JAKOBSSON ET AL., *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature, 451:7181(2008), pp. 998-991003.

[6] B. DEVLIN AND N. RISCH, *A comparison of linkage disequilibrium measures for fine-scale mapping.* Genomics, 29:2(1995), pp. 311-322.

[7] D.B. GOLDSTEIN AND M.E. WEALE, *Population genomics: linkage disequilibrium holds the key.* Curr Biol, 11:14(2001), pp. 576-579.

[8] C.S. CARLSON ET AL., *Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.* Am J Hum Genet, 74:1(2004), pp. 106-120.

[9] M.M. MIRETTI ET AL., *A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms.* Am J Hum Genet, 76:4(2005), pp. 634-646.

[10] P. Sebastiani et al., *Minimal haplotype tagging.* Proc Natl Acad Sci U S A, 100:17(2003), pp. 9900-9905.

[11] J.C. Barrett and L.R. Cardon, *Evaluating coverage of genome-wide association studies.* Nat Genet, 38:6(2006), pp. 659-662.

[12] Y. Zhao et al., *A better block partition and ligation strategy for individual haplotyping.* Bioinformatics, 24:23(2008), pp. 2720-2725.

[13] N. Patil et al., *Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.* Science, 294:5547(2001), pp. 1719-1723.

[14] Y. Guo et al., *Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies.* Eur J Hum Genet, 2008.

[15] K. Zhang et al., *HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms.* Bioinformatics, 21:1(2005), pp. 131-134.

[16] K. Zhang et al., *Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies.* Genome Res, 14:5(2004), pp. 908-916.

[17] N. Sazonova and E.J. Harner, *Haplotype inference and block partitioning in mixed population samples.* J Bioinform Comput Biol, 6:6(2008), pp. 1177-1192.

[18] C.-J. Chang, Y.-T. Huang, and K.-M. Chao, *A greedier approach for finding tag SNPs.* Bioinformatics, 22:6(2006), pp. 685-691.

[19] Z. Lin and R.B. Altman, *Finding haplotype tagging SNPs by use of principal components analysis.* Am J Hum Genet, 75:5(2004), pp. 850-861.

[20] B.V. Halldorsson et al., *Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies.* Genome Res, 14:8(2004), pp. 1633-1640.

[21] P. Paschou et al., *Intra- and interpopulation genotype reconstruction from tagging SNPs.* Genome Res, 17:1(2007), pp. 96-9107.

[22] P. Nicolas, F. Sun, and L.M. Li, *A model-based approach to selection of tag SNPs.* BMC Bioinformatics, 7(2006), pp. 303-303.

[23] J. Hampe, S. Schreiber, and M. Krawczak, *Entropy-based SNP selection for genetic association studies.* Hum Genet, 114:1(2003), pp. 36-43.

[24] K. Hao, *Genome-wide selection of tag SNPs using multiple-marker correlation.* Bioinformatics, 23:23(2007), pp. 3178-3184.

[25] D.O. Stram et al., *Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals.* Hum Hered, 55:4(2003), pp. 179-190.

[26] P.I.W. de Bakker et al., *Efficiency and power in genetic association studies.* Nat Genet, 37:11(2005), pp. 1217-1223.

[27] J.C. Barrett et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 21:2(2005), pp. 263-265.

[28] E. Halperin, G. Kimmel, and R. Shamir, *Tag SNP selection in genotype data for maximizing SNP prediction accuracy.* Bioinformatics, 21: Suppl 1(2005), pp. 195-203.

[29] Z.S. Qin, S. Gopalakrishnan, and G.R. Abecasis, *An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria.* Bioinformatics, 22:2(2006), pp. 220-225.

[30] Y.-T. Huang and K.-M. Chao, *A new framework for the selection of tag SNPs by multimarker haplotypes.* J Biomed Inform, 41:6(2008), pp. 953-961.

[31] B.N. Howie et al., *Efficient selection of tagging single-nucleotide polymorphisms in multiple populations.* Hum Genet, 120:1(2006), pp. 58-68.

[32] M. Berkelaar, J.D., K. Eikland, P. Notebaert, and J. Ebert , *lp_solve* , 2007.

[33] P.I.W. de Bakker et al., *A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.* Nat Genet, 38:10(2006), pp. 1166-1172.

[34] D.R. Smith et al., *Rapid whole-genome mutational profiling using next-generation sequencing*

technologies. Genome Res, 18:10(2008), pp. 1638-1642.

[35] *This time it's personal.* Nature, 453:7196(2008), pp. 697-697.

[36] L. Excoffier and M. Slatkin, *Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.* Mol Biol Evol, 12:5(1995), pp. 921-927.

[37] 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.* Am J Hum Genet, 78:4(2006), pp. 629-644.