Haplotypes and Informative SNP Selection Algorithms: Don't Block Out Information

Dedicated to Michael Waterman on his 60th Birthday

Vineet Bafna* Bjarni V. Halldórsson† Russell Schwartz ‡ Andrew G. Clark§
Sorin Istrail¶

ABSTRACT

It is widely hoped that variation in the human genome will provide a means of predicting risk of a variety of complex, chronic diseases. A major stumbling block to the successful identification of association between human DNA polymorphisms (SNPs) and variability in risk of complex diseases is the enormous number of SNPs in the human genome (4,9). The large number of SNPs results in unacceptably high costs for exhaustive genotyping, and so there is a broad effort to determine ways to select SNPs so as to maximize the informativeness of a subset.

In this paper we contrast two methods for reducing the complexity of SNP variation: haplotype tagging, i.e. typing a subset of SNPs to identify segments of the genome that appear to be nearly unrecombined (haplotype blocks), and a new block-free model that we develop in this report. We present a statistic for comparing haplotype blocks and show that while the concept of haplotype blocks is reasonably robust there is substantial variability among block partitions. We develop a measure for selecting an informative subset of SNPs in a block free model. We show that the general version of this problem is NP-hard and give efficient algorithms for two important special cases of this problem.

*Informatics Research, Celera. Current address: The Center for Advancement of Genomics, 1901 Research Blvd., 6th floor, Rockville, MD 20850, *vbafna@tcag.org*

†Informatics Research, Celera/Applied Biosystems, 45 W. Gude Drive, Rockville MD, 20850, *Bjarni.Halldorsson@celera.com*

‡Informatics Research, Celera. Current address: Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, 15213, *russells@andrew.cmu.edu*

§Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, 14853, *ac347@cornell.edu*

¶Informatics Research, Celera/Applied Biosystems 45 W. Gude Drive, Rockville MD, 20850, *Sorin.Istrail@celera.com*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
RECOMB'03, April 10–13, 2003, Berlin, Germany.
Copyright 2003 ACM 1-58113-635-8/03/0004 ...$5.00.

Categories & Subject Descriptors

F.2.0, G.3, G.4, J.3

General Terms

Algorithms, Theory, Experimentation.

Keywords

SNPs, haplotype tagging, haplotype blocks.

1. INTRODUCTION

In anticipation of cost-effective SNP genotyping technology and the availability of a large number of SNPs, many investigators are seriously considering genome-wide SNP scans with the hope of performing association tests. While the cost of SNP genotyping may be rapidly decreasing, it is still infeasible to genotype all available SNPs [5]. This paper deals with the challenging problem of choosing an optimal subset of SNPs to be used in such a study. This subset of SNPs is referred to as the "minimal informative subset". The end objective is to be able to identify DNA variation within the population that is associated with elevated risk of diseases, but there are many unknowns regarding the nature of that variation. We do not know how many genes will be involved, what the frequency of the alleles causing disease is, nor do we know the magnitude of the increased risk associated with those alleles. In the absence of this information, the best we can do is to try to identify SNPs that at least allow us to reconstruct the haplotypes of all the other observed SNPs. In this sense, the problem is like a data compression problem: We have the full pattern of all SNPs for a small sample, and want to devise an algorithm to select SNPs that will allow us to reconstruct the full data set. We hope that even though only a subset of all SNPs would be genotyped, that the statistical power for identifying disease associations would not be seriously compromised.

We note that each SNP defines a partition of the population sample. If a pair of SNPs defines precisely the same partition, they are said to be in absolute linkage disequilibrium (LD), and the information they carry is totally redundant. In this trivial case, it is easy to see that the minimal informative subset would not include both SNPs. Now consider the question of whether the partitioning of any given SNP x can be reconstructed from the partitioning of some set of other SNPs. If we can reconstruct this

partition from other SNPs, then it would not be necessary to genotype SNP x. The state of SNP x could be predicted from the states of the other SNPs. For any one SNP the identification of this partitioning is not challenging, but the task of finding a global solution to minimizing the number of SNPs needed to predict all the others is.

The following method for selecting informative SNPs has been widely suggested [2, 16, 21, 25]: partition a chromosome into contiguous segments called haplotype blocks and select, within each block, a subset of "haplotype tagging" SNPs sufficient to reconstruct the diversity within that block. Haplotype blocks are intuitively defined as a contiguous series of SNPs exhibiting low diversity across individuals that appear to have little evidence for recombination within the sample. Due to the low diversity within a given haplotype block, the SNPs are highly correlated and the value of one SNP can be used to predict the value of another. Despite rough agreement on this intuitive understanding, no consensus has been reached in the community over how they are best practically defined.

In this paper, we consider the problem of defining haplotype tagging SNPs and that problem's relationship to haplotype blocks. We study the nature of haplotype block patterns and use results about potential limitations of the blocks to suggest methods for haplotype tagging that are independent of the block problem. We then present theoretical formalisms and practical algorithms for two variants of the SNP selection problem, one based on a restricted notion of blocks and one on block-free selection.

In Section 2 we statistically assess the concept of haplotype blocks by examining the concordance of block structures derived through different blocking methods. We provide a paradigm for comparing different block definitions; we demonstrate that an optimal haplotype block partition can be calculated efficiently using a randomized dynamic programming algorithm and we give a statistic for comparing two partitions. We apply this statistic to three haplotype blocking methods representing the major principles used to date for block assignment in the literature: four gamete [14], diversity based [21] and an LD based definition [11]. Our results suggest that the concept of haplotype blocks is reasonably robust, and we show that the different block definitions exhibit a high degree of concordance. It is important to note, however, that the concordance is highest for common haplotype blocks, and that there is substantial variability among block partitions, especially for those that are of low frequency in the population. It does not appear to be possible to unambiguously and uniquely infer the "true" block partitioning.

In Section 3 we specifically address the problem of deriving informative subsets of SNPs, defining a measure for how one SNP can be used to predict another, setting up the basic algorithmic problem, and showing that it is NP-hard in the general case. The *informativeness* measure is introduced for how well one SNP predicts another and for how well a set of SNPs predicts a single SNP and a set of SNP. This measure is closely related to the haplotype diversity measure [6] commonly use. Given this measure, we can state the minimum informative SNPs problem (MIS) and we show that the general problem is NP-complete.

In Section 4 we give efficient algorithms for two important special cases of the MIS problem. In practice, the full MIS problem is often overly general, as we only want to select a SNP to be used in the prediction of another SNP if the two SNPs are in strong LD (the SNPs are highly correlated). We consider two such cases: when SNPs are predicted only on recombination-free blocks and when a SNP is only used in the prediction of another if they are in close proximity to each other. We show that this approach leads to substantial improvements over a representative block method in capturing a large fraction of the information in a set of sequences using only a subset of the SNPs.

2. EVALUATING HAPLOTYPE BLOCKS

Many studies (confer [8, 10, 15, 21]) suggest that human haplotypes can be characterized by long stretches of sequence, or *blocks* within which there is high LD across common SNPs, but between which recombination has left little LD. Because of the high LD, there is redundant information within a block, and only a few SNPs might be sufficient to characterize the block. Other studies [24], while not disputing the existence of blocks, critique the value obtained from such sparse maps. We tackle a different question in the paper, regarding the *necessity* of block selection as a prelude to selecting a set of informative SNPs characterizing haplotypes.

There are many computational approaches to predicting block boundaries. We compare three different methods, representing three basic blocking paradigms in common use: recombination-based, diversity-based, and LD-based. The simplest example of a recombination-based test is due to Hudson and Kaplan [14]. Using the infinite sites assumption (each site mutates at most once) and the assumption of no recombination within a block, no block should exhibit all four possible gametes for any pair of sites. Efficient algorithms to predict blocks satisfying the no-four-gamete criteria are known. For a diversity-based test, we use a generalization of the test presented by Patil et al. [21]. In their test a region is a block if at least 80% of the sequences occur in more than one chromosomes. This test is developed for a sample of only 20 chromosomes and does not scale well to larger sample sizes as it will tend to yield larger blocks as more chromosomes are studied. We generalized this test by defining a region as a potential block if sequences within that region (haplotypes) accounting for at least 80% of the sampled population each occur in at least 10% of the sample. We were also unable to find an LD-based test in the literature that would scale well with depth of coverage over the range of datasets available to us, and therefore developed a test based on the D' statistic, similar to that used by Gabriel et al. [11] but with the test of significance tuned to produce more meaningful results on small population samples. By this measure, a region is defined to be a block if the D' value of every pair of SNPs within the block shows significant LD given the individual SNP allele frequencies with a P-value of 0.001.

Following Zhang et al. [25], we consider two optimization criteria for any given block definition. The first criterion is *minimum test cover*, in which for each block the number of SNPs required for haplotype tagging is computed as a minimum test cover problem [12, 5, 4] and sum of the number of SNPs over all the blocks is minimized. The second is *minimum block*, in which the number of predicted blocks is minimized among all partitions that minimize the number

of SNPs. We use a randomized variant of Zhang et al.'s dynamic programming algorithm that samples uniformly at random among all optimal solutions to locate optimal block partitions using all three block definitions and both optimization criteria.

2.1 Statistic for Comparing Block Partitions

We can use the number of shared block boundaries as a statistic for the similarity of two block partitions. We give a simple formula to test whether this number of shared boundaries might have occurred by random chance. Let B_1, B_2 be the number of boundaries in the two partitions, m be the number of boundaries shared by the partitions and S be the total number of SNPs. If the partitions are independent of one another, the probability that they share exactly m boundaries can be calculated as follows:

$$\frac{\binom{B_1}{m}\binom{S-1-B_1}{B_2-m}}{\binom{S-1}{B_2}}$$

The P-value for the intersection of the two partitions being random can then be calculated as follows:

$$\sum_{i=m}^{\min(B_1,B_2)} \frac{\binom{B_1}{i}\binom{S-1-B_1}{B_2-i}}{\binom{S-1}{B_2}} \quad (1)$$

This measure thus allows us to test the hypothesis that two partitions are related.

2.2 Computational Results

For evaluation, we rely on two publicly available datasets. The first is the Perlegen chromosome 21 dataset [21], which consists of 24,047 SNPs typed on 20 phased chromosomes. This dataset contains a large contiguous set of SNPs providing an excellent test of blocking algorithms. We also use a dataset derived from 71 individuals typed at 88 polymorphic sites in the human lipoprotein lipase (LPL) gene [20], from which we ignored one multi-allelic site to simplify our analysis. The fewer SNPs in the LPL dataset makes it more manageable for illustrative purposes. In addition, its greater depth of coverage allows us to draw more confident predictions and provides enough individuals to compare results from distinct subsets of the population sample.

We specifically asked whether the concept of blocks is robust to the various block measures. We conducted comparisons by running the available algorithms on the two datasets and comparing the outcomes using the shared boundary statistic described above. Table 1 presents the results when run on the chromosome 21 data of Patil et al. [21] minimizing the total number of blocks consistent with each block definition. Table 2 shows the analogous results when choosing blocks so as to minimize the total number of SNPs needed to type the blocks for each definition. Although the percent similarities are stronger than can be explained by chance, there is a fair amount of discrepancy between different measures. The 4-gamete test appears much closer to the diversity- and LD-based tests than either of those is to the other. Furthermore, minimum SNP solutions appear to have more in common with one another than minimum block solutions.

For purposes of illustration, we provide visual comparisons of block assignments for the LPL dataset of Nickerson et al.[20]. Due to the much smaller number of SNPs in the LPL dataset compared to the chromosome 21 dataset, the LPL comparisons do not generally yield statistically significant results, although they show comparable percent agreement to the chromosome 21 comparisons. Figure 1 illustrates the correspondence between the different block measures by showing side-by-side comparisons of block boundaries for each. The results appear to show generally poor yet significant agreement between block boundaries derived from distinct measures. From these results, we conclude that the blocks are capturing general regions of low diversity, but that the boundaries between them are not rigorously defined.

	diversity-based	LD-based
4-gamete	$15.7\%/3.02 \times 10^{-162}$	$12.3\%/1.70 \times 10^{-18}$
diversity-based	-/-	$7.19\%/3.89 \times 10^{-5}$

Table 1: Comparisons of block definitions on the chromosome 21 dataset of Patil et al. minimizing blocks. Each element of the matrix gives the percentage of block boundaries assigned by either method that are shared by both, followed by the p-value of the overlap.

	diversity-based	LD-based
4-gamete	$22.6\%/5.30 \times 10^{-366}$	$16.0\%/1.03^{-54}$
diversity-based	-/-	$9.70\%/6.14 \times 10^{-23}$

Table 2: Comparisons of block definitions on the chromosome 21 dataset of Patil et al. minimizing SNPs.

3. THE SNP SELECTION PROBLEM

Our primary concern in the present work is defining subsets of "haplotype tagging" SNPs that characterize the overall genetic diversity of a region. Many common existing methods for SNP selection are based on measures of LD, or diversity applied to a block at a time. Minimizing the number of SNPs under these measures is computationally hard in the general case, but practical in most cases of interest, as the blocks are typically not very large, and have only a few haplotypes. In the following, we describe a new measure of *informativeness* of a SNP that is related to measures of Haplotype Diversity [6], but provides a direct measure of how a SNP, or a set of SNPs can be used to characterize another SNP, or a set of SNPs. Our measure is also NP-hard in the general case, but is tractable in many cases of practical interest. We also describe the use of this measure to select a minimum informative subset of SNPs, *without* partitioning the SNPs into blocks.

3.1 Defining Informativeness

A set S of m SNPs over a population of n haplotypes can be denoted by an $n \times m$ matrix M_S. The columns of M correspond to SNPs in the population, and rows correspond to haplotypes. For a SNP $s = M_S[, j]$, denote $s_i = M_S[i, j]$. We assume for simplicity that all SNPs are diallelic (taking on only two values); consequently $M_S[i, j] \in \{0, 1\} \forall i, j$. Let a 'target' SNP t be associated with a disease of interest. If t is not typed, its value is predicted using proximal SNPs. We would like to define a measure of 'informative-

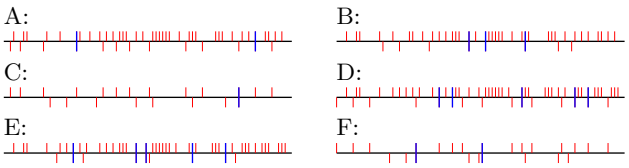


Figure 1: Pairwise comparison of block boundaries for the different block definitions for the LPL dataset of Nickerson et al. Each image shows the block boundaries as vertical lines, with boundaries above the horizontal line coming from the first method and those below the horizontal line coming from the second method. Shared boundaries are drawn as thicker lines and unshared boundaries as thinner lines. The comparisons shown are A: 4-gamete vs. diversity-based minimizing blocks B: 4-gamete minimum blocks vs. LD-based minimizing blocks C: diversity-based vs. LD-based minimizing blocks D: 4-gamete vs. diversity-based minimizing SNPs E: 4-gamete vs. LD-based minimizing SNPs F: diversity-based vs. LD-based minimizing SNPs

ness' of a SNP s w.r.t t to quantify the confidence with which we can make this prediction.
For a SNP s and haplotypes i, j, let $D^s_{i,j}$ be the event that that $M[i,s] \neq M[j,s]$. We define the *Informativeness* of SNP s w.r.t a SNP t as

$$I(s,t) = \text{Prob}_{i \neq j}(D^s_{i,j} | D^t_{i,j}) \quad (2)$$

where i, j are drawn uniformly at random from the set of all distinct haplotype pairs. Observe that $I(s,t) = 1$ implies complete predictability, and $I(s,t) = 0$ when there is no predictability. $I(s,t)$ is estimated easily from a sample as follows: Consider the complete graph G_H on m nodes labeled $1, 2, \ldots, m$. Each SNP s defines a subgraph which is a bipartite clique with m nodes. The edge set $E(s)$ is defined by the rule $(i,j) \in E(s)$ iff $s_i \neq s_j$. Then

$$I(s,t) \simeq \frac{|E(s) \cap E(t)|}{|E(t)|}$$

as the informativeness of s w.r.t t. The definition is easily extended to a subset of SNPs. For $S' \subseteq S$, let $D^{S'}_{i,j}$ be the event that $M[i,s] \neq M[j,s]$ for some $s \in S'$. Likewise, let $E(S') = \cup_{s \in S'} E(s)$. Then,

$$I(S',t) = \text{Prob}_{i \neq j}(D^{S'}_{i,j} | D^t_{i,j}) \simeq \frac{|E(S') \cap E(t)|}{|E(t)|} \quad (3)$$

and, for all $S', T \subseteq S$

$$I(S',T) = \sum_{t \in T} I(S',t)$$

3.2 Basic algorithmic problem

Based on the definition above, a subset of SNPs S' is said to be *informative* w.r.t a subset of SNPs T if $I(S',T) = |T|$.

MIS: Minimum Informative SNPs

Input: A set of n SNPs S, with subset $T \subseteq S$, $0 < k \leq n$.

Output: Does there exist a subset $S' \subseteq S \setminus T$ such that $I(S',T) = |T|$, and $|S'| \leq k$.

The MIS problem is closely related to the problem where we can only type k SNPs, and we need to select the k SNPs that are most informative.

k-MIS: k Most Informative SNPs

Input: A set of n SNPs S, subset SNP $T \subseteq S$, $0 < k \leq n$.

Output: Find the subset $S' \subseteq S \setminus T$ such that $I(S',T) = \max_{S_T \subseteq S, |S_T| \leq k} I(S_T, T)$

We first show that the MIS, and k-MIS problems are NP-complete in the general case.

3.3 Tractability

LEMMA 3.1. *The* Minimum Informative SNPs *problem is* NP*-complete.*

PROOF. We reduce the set cover problem to MIS. Recall the definition of set cover: Given a collection C of subsets of a finite set X, and positive integer $k \leq |C|$, does there exist $C' \subseteq C$ with $|C'| \leq k$ such that every element of X belongs to at least one member of C'.
We construct a SNP matrix M_S with $|X| + 1$ haplotypes, and $|C|+1$ SNPs. Label the elements of X arbitrarily from 1 to $|X|$. For each subset $C_j \in C$, define a SNP $M[*, j]$ such that

$$M[i,j] = \begin{cases} 0 & \text{if } i \leq |X| \text{ and } X_i \in C_j \\ 1 & \text{otherwise} \end{cases}$$

SNP $t = M[*, |C|+1]$ is defined by the vector $[0, 0, \ldots, 0, 1]$ with exactly $|X|$ zeros and a single one. It is not hard to see that $C' \subseteq C$ covers X if and only if the corresponding subset of SNPs S' are informative w.r.t t. □

As the MIS problem is a restriction of the k-MIS problem, we have

COROLLARY 3.1. *The k-MIS problem is NP-hard.*

4. SNP SELECTION ALGORITHMS

While the general MIS problem is NP-hard, we show that it is tractable in two special cases that are of practical interest. In Section 4.1 we give an efficient algorithm for the case when the SNPs are to be selected on blocks in which there is no recombination. In Section 4.2 we give an efficient algorithm for the case when predictive SNPs are in close proximity to their target.

4.1 Block-based selection

In this section, we give an efficient algorithm for the case when SNPs are selected on blocks which are defined on the basis that there is no evidence of recombination. Consider two diallelic SNPs s and t. Under the infinite-sites model of evolution and the assumption that no recombination has occurred between these two sites, it can be shown [14] that all four allele pairs will not be observed (the four gamete property). We define such a pair of SNPs to be in *complete LD*. This condition is fairly restrictive. Observe however, that the MIS haplotype problem remains NP-complete even if we restrict our selection to SNPs that are in complete LD with the target SNP t.
Consider the case of a *complete* LD block, defined as a set of SNPs S such that every pair in S is in complete

LD. The resulting SNP matrix M_S has a lot of structure. Specifically, a perfect phylogeny can be constructed of the set of haplotypes (rows in M_S) (See for example, [7]).

blMIS *Minimum informative SNPs on a block in complete LD*

Input: A set of n SNPs S that define a complete LD block, subset $T \subset S$, $0 < k \leq n$.

Output: Does there exist a subset $S' \subseteq S \setminus T$ such that $I(S', T) = |T|$, and $|S'| \leq k$.

k-blMIS *k Most Informative SNPs on a block in complete LD*

Input: A set of n SNPs S that define a complete LD block, subset SNP $T \subset S$, $0 < k \leq n$.

Output: Find the subset $S' \subseteq S \setminus T$ such that $I(S', T) = \max_{S_T \subseteq S, |S_T| \leq k} I(S_t, T)$

LEMMA 4.1. *The blMIS problem is solved in $O(mn)$ time, when $|T| = 1$.*

PROOF. For SNP j, let H_j^0 be the set of rows (haplotypes) i, such that $M[i, j] = 0$. Correspondingly, let H_j^1 be the set of rows i with $M[i, j] = 1$. We need to find a set of SNPs that are informative w.r.t t. As a consequence of the 'no-four-gamete' condition, note that for each SNP s with alleles a, and b, that either $H_s^a \subseteq H_t^0$, or $H_s^b \subseteq H_t^1$, or $H_s^a \subseteq H_t^1$, or $H_s^b \subseteq H_t^0$.
For each SNP s, relabel a, and b to 0, and 1 so that either $H_s^0 \subseteq H_t^0$, or $H_s^1 \subseteq H_t^1$. Consequently, any set S' of SNPs that is informative w.r.t t has the property that $\cup_{s \in S'} H_s^0 = H_t^0$, or $\cup_{s \in S'} H_s^1 = H_t^1$. To compute MIS, we find a minimum set S_0 such that $\cup_{s \in S_0} H_s^0 = H_t^0$, and a minimum set S_1, such that $\cup_{s \in S_1} H_s^1 = H_t^1$, and choose the smaller of the two sets.
To compute S_0, we employ a greedy strategy. At each step, we pick a SNP s such that $H_s^0 \subseteq H_t^0$, and H_s^0 covers the maximum number of haplotypes not covered by previously chosen SNPs. To see that S_0 is minimal, observe for two sets H_0^r, H_0^q that either one is contained in the other or the two set are non-overlapping, due to the four gamete property.

□

We note that the k-blMIS problem also has the same time complexity when $|T| = 1$. For every $l, 0 \leq l \leq k$ find a set $S_{0l}, |S_{0l}| = l$ such that $\cup_{s \in S_{0l}} H_s^0 \subseteq H_t^0$ and $\cup_{s \in S_{0l}} H_s^0$ has maximum cardinality, using the greedy algorithm. Define S_{1l} similarily. An optimal solution for the k-blMIS can be found by finding the $S_{0l} \cup S_{1(k-l)}$ with maximum informativeness.
The complexity of the general blMIS and k-blMIS problems remains open.

4.2 Bounded-width algorithm

While significant LD can occur between SNPs physically distant on the genome, such distant relationships are likely to reflect selection bias or random chance, rather than recent common ancestry. In practice, measured high LD occurring at greater distances than a fixed constant, such as 50 kb, is commonly ignored [11] and is considered to be an artifact of a small sample size. Our primary concern in locating SNPs in LD with one another is finding those sets of SNPs that are predictive not only of other SNPs of which we are aware, but also of those SNPs that were not typed in our sample population. We therefore wish primarily to discover those SNPs that characterize regions of common recent ancestry (conserved haplotypes) rather than those that characterize isolated distant SNPs due to selection bias or random chance. SNPs of recent common ancestry are likely to be those that are sufficiently close that recombination has not occurred frequently between them. We therefore believe it reasonable to restrict our search for predictive SNPs to those that are in relatively close proximity to the targets for which they might be predictive.
Consider again the informativeness of a subset S' w.r.t a SNP t. For simplicity, we define the distance between SNPs s and t simply by the number of SNPs in between s and t. Informally, we would like to find a most informative subset of SNPs given that only SNPS that are a distance w apart can be used in the prediction. The abstraction described here is used only for ease of exposition. In practice, important efficiency gains are made by a proper choice of useful SNPs, and the general approach will be described in the following subsection.
Given a positive integer w, the w-bounded informativeness of S w.r.t t, denoted by $I(w, S', t)$, is defined as follows:

$$I(w, S', t) = \frac{|\cup_{s \in S', d(s,t) \leq \lfloor \frac{w}{2} \rfloor} E(s) \cap E(t)|}{|E(t)|} \quad (4)$$

The w bounded informativeness w.r.t. a set of SNPs T we define is the sum of the informativeness w.r.t. the individual SNPs:

$$I(w, S', T) = \sum_{t \in T} I(w, S', t) \quad (5)$$

The problem of finding a bounded width informative subset is defined as follows:

(k, w)-MIS *The k most informative SNPs with bounded width w*

Input: A SNP matrix M_S

Output: A set of k SNPs S_k that maximizes $I(w, S_k, S)$

We need to set out some notation to show how the (k, w)-MIS can be solved efficiently. Number the n SNPs from 1 to n. Suppose a subset S' of SNPs within distance $\lfloor \frac{w}{2} \rfloor$ of s are used to predict s. Define the corresponding *assignment* A_s as follows

$$A_s[i] = \begin{cases} 1 & \text{if SNP } s - \lfloor \frac{w}{2} \rfloor + i \in S' \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Correspondingly, define a the subset of SNPs $S(A_s)$ to contain all SNPs s' such that $A[s' + \lfloor \frac{w}{2} \rfloor - s] = 1$.

THEOREM 4.1. *The (k, w)-MIS problem can be solved $O(nk2^w)$ time, and $O(k2^w)$ space.*

PROOF. A solution to the (k, w)-MIS problem can be described by an $O(n)$ size bit-vector, such that $B[i] = 1$ if SNP i is selected, and 0 otherwise. At most k entries are

For s **from** 1 **to** n
 For l **from** 1 **to** k
  Forall assignments A_s
   $A_s^0 \leftarrow A_s >> 1$
    Right shift by 1
   $A_s^1 \leftarrow A_s >> 1 + 2^{w-1}$
    Right shift and assign 1 to the leftmost bit
   $I_w(s, l, A_s) \leftarrow I(S(A_s), s)+$
    $\max(I_w(s-1, l-A_s[w], A_s^0), I_w(s-1, l-A_s[w], A_s^1)$

Figure 2: An $O(nk2^w)$ algorithm for the (k, w)-MIS problem

1 in any solution. The solution also implies an assignment A_s for each SNP s as $A_s[i] = B[s - \lfloor \frac{w}{2} \rfloor + i]$.
Let $A_s^0 (= A_s >> 1)$ be the vector obtained by right shifting A_s. Note that in any solution $A_s[i] = B[s - \lfloor \frac{w}{2} \rfloor + i] = B[(s-1) - \lfloor \frac{w}{2} \rfloor + (i+1)] = A_{s-1}[i+1]$. Therefore, depending on whether the $A_{s-1}[0]$ is 0 or 1, $A_{s-1} = A_s^0$, or $A_{s-1} = A_s^0 + 2^{w-1}$.
Let $I_w(s, l, A_s)$ be the score of most informative subset of l SNPs chosen from SNPs 1 through s, such that A_s described the assignment for SNP s. The score obtained for informing SNP s is exactly $I(S(A_s), s)$, and $I_w(s, l, A_s)$ is given by $I(S(A_s), s)$ plus score of the best assignment for SNPs 1 through $s-1$ that is consistent with A_s.
By the argument above, there are only two possibilities for the assignment to SNP $s-1$, described by A_s^0, or A_{s-1}^1. Finally, the assignment to SNPs 1 through $s-1$ cannot use SNP $s + \lfloor \frac{w}{2} \rfloor$. Therefore, the number of SNPs available to SNPs 1 through $s-1$ are $l-1$ if $A_s[w] = 1$, and l otherwise. Thus

$$I_w(s, l, A_s) = I(S(A_s), s)+$$
$$\max(I_w(s-1, l-A_s[w], A_s^0), I_w(s-1, l-A_s[w], A_s^1)$$

Figure 2 describes the algorithm for computing this recurrence using dynamic programming. The space saving trick of Hirschberg [13] can be used to reduce the space requirements to $O(k2^w)$.

□

4.3 SNP Neighborhood selection

In practice, it is neither efficient nor desirable to have a fixed neighborhood in which to look for SNPs. Due to variability in SNP density and recombination rate, we note that for most SNPs, the neighborhood of SNPs that are in LD with it, or are otherwise informative for it is highly *variable*. We exploit this by making a neighborhood graph with SNP sites as vertices. Two vertices are connected by an edge only if they are in high LD w.r.t each other. For any SNP s, the computation is then performed only on subsets of neighboring SNPs that are connected to s, and are proximal to it.
Even in the modified approach, the large memory required for modest values of w makes this approach hard to implement. An interesting open problem is to use a Hirschberg like trick on the larger dimension (2^w). In our formulation, the problem reduces to finding a minimum Bandwidth layout for a *De Bruijn* graph on 2^w nodes. As the Bandwidth is known to be greater than $\frac{2^w}{w}$, this is unlikely to offer great improvement. However, we can reduce the memory demands further by not considering SNPs whose addition improves the informativeness marginally.

4.4 Computational Results

We now compare the number of SNPs that need to be typed to capture α% of the information based for the block a block based method and for our block free method. We follow Zhang et al. [25] for the block method and require that α% of the diversity of each individual block be captured. For the block free method we require that α% of the sum of the informativeness of the SNPs be captured.
For a judicial comparison between the block and block free methods, in the block free method we will only allow a SNP i to be used in the prediction of the informativeness of SNP j if i and j are in a putative block. We note that this restriction is in addition to the restriction that only SNPs within a fixed window size be used in prediction.

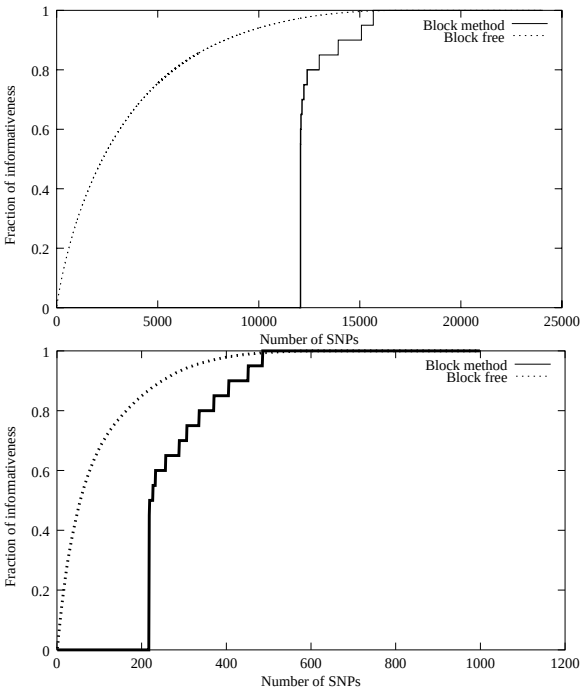


Figure 3: The x-axis shows the number of SNPs typed and on the y-axis informativeness attained. The dotted lines represent the block free method and the solid the block method. The upper picture shows results for the Perlegen dataset where an LD based method was used for block determination (and determining predictive SNPs). The lower shows results for the first 1000 SNPs of the Perlegen dataset where blocks where determined using the method presented in Patil et al.

Figure 3 shows the informativeness attained as a function of the number of SNPs typed using the Perlegen data set and both a conservative LD block detection method and the less conservative diversity based method of Patil et al. [21]. We see in both cases that the block free method requires fewer SNPs to be typed.
Figure 4 shows the information attained as a function of

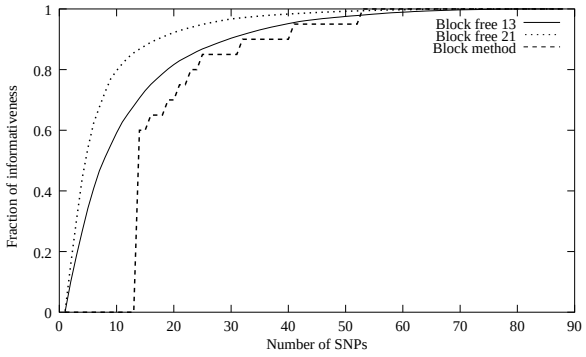


Figure 4: The x-axis shows SNPs typed and the y-axis the information attained for the LPL data set, using the diversity based block definition of Patil et al. The dotted line represent the block free method with w limited to 21, the solid line the block free method with w limited to 13 and the dashed line shows the block method.

the number of SNPs typed using the LPL data set and the Cox block detection method. As the Cox method allows for large putative blocks we considered the effect of increasing the maximum w for the block free method. Again, our results show that the block free SNP detection can produce similar informativeness with fewer SNPs as we can take advantage of informative SNPs in neighboring 'blocks'.

5. DISCUSSION

We show two things in this paper. First, while the concept of block as a region with low-recombination is robust, the notion of block boundaries is not. This conclusion is consistent with the observation of Gabriel et al. [11] that there can be significant LD between blocks even when they are chosen by an LD-based measure. Therefore, in selecting a subset of SNPs to characterize a large fraction of haplotypes, it is not essential to limit the analysis to computed blocks.

We also present a new measure for the identification of subsets of SNPs that are predictive of other SNPs in a common population. As we argue above (and in the appendix), this measure avoids some of the difficulties traditional linkage disequilibrium measures have experienced when applied to SNP selection, particularly when dealing with small population samples. The concept of pairwise LD does not reliably captures the higher-order dependencies implied by observed haplotype structures, while the extension to 3-way LD and in general k-ary LD would be tremendously complex analytically. Our notion of informativeness provides a practical framework for formalizing the problem of SNP selection suitable for generalized structures of higher-order local dependency. We have characterized this measure theoretically and discussed the hardness of general optimization problems based on the measure. We have also presented algorithms for efficiently solving two practical variants of the informative SNP selection problem: block-based and bounded-width SNP selection. The block-based method provides a new basis for performing SNP selection using the block paradigm. The bounded-width method provides a means of solving the problem without resorting to a block decomposition, which may be an important advantage given the uncertainty about the validity and utility of haplotype blocks.

There are a number of avenues for future work in this area. The informativeness measure remains to be characterized as a statistic, which would be expected to yield more rigorous measures and algorithms for dealing with uncertainty in small population samples. We note that our basic algorithms can be extended to alternate measures. In fact, the algorithm for (k, w)-MIS problem can be used for any measure that grows monotonically in the predictive set (i.e. if more data gives a better prediction).

Additional work with empirical data is also required to establish practical window sizes, w, and assess the effectiveness of the methods in reducing assay sizes and characterizing SNPs unknown at the time of the selection of informative subsets. In particular, the dynamic program presented above requires large memory making it impractical for larger values of w. Cutting the memory requirements of this method is an active area for further research. We can also consider extensions of the above methods to noisy or incomplete data, where techniques that we have developed in [18, 19, 22] may be of use.

Furthermore, we have not considered how to deal with genotype data. One way this could be done is to construct haplotypes from genotypes computationally [1, 3]. Computationally phasing the genotype data however, inherently adds a level of indirection to the problem. An alternate method, of greater interest to us, is to avoid this phasing step and define the algorithm in the space of all possible phasings of the genotype data and optimize over a probability distribution. Preliminary work in this direction has been done in a concurrent paper analyzing the genetic diversity in human chromosomes 6,21,22 [23], written partially by the same set of authors. In that paper, we assume that for any pair of SNPs there is equal probability of the two different phasings of heterozygous pairs.

Finally, these methods must be applied to real problems in assay design and disease inference to characterize their true potential to contribute to the effective application of genome polymorphism data to the diagnosis and treatment of human disease.

6. REFERENCES

[1] Gonçalo R. Abecasis, Stacey S. Cherny, William O. Cookson, and Lon R. Cardon. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30:97–101, 2002.

[2] Hadar I. Avi-Itzhak, Xiaoping Su, and Francisco M. De La Vega. Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity. In *Proceedings of Pacific Symposium on Biocomputing*, pages 466–477, 2003.

[3] V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as a perfect phylogeny. a direct approach. *Journal of Computational Biology*, 2003. To appear.

[4] K.M.J. De Bontridder, B.V. Halldórsson, M.M. Halldórsson, C.A.J. Hurkens, J.K. Lenstra, R. Ravi, and L. Stougie. Approximation algorithms for the minimum test cover problem. *Mathematical Programming-B*, 2003. To Appear.

[5] K.M.J. De Bontridder, B.J. Lageweg, J.K. Lenstra, J.B. Orlin, and L. Stougie. Branch and bound algorithms for the test cover problem. In *Proceedings of the 10th Annual European Symposium on Algorithms (ESA)*, pages 223–233, 2002.

[6] D. Clayton. Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf, 2001.

[7] Gusfield D. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.

[8] M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.

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

[10] D. E. Reich et al. Linkage disequiblirium in the human genome. *Nature*, 2001.

[11] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altschuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[12] B.V. Halldórsson, M.M. Halldórsson, and R. Ravi. On the approximability of the test collection problem. In *Proceedings of the 9th Annual European Symposium on Algorithms (ESA)*, pages 158–169, 2001.

[13] D. S. Hirschberg. A linear space algorithm for computing maximal common subsequence. *Communications of the ACM*, 18:341–343, 1975.

[14] R.R. Hudson and N.L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.

[15] A.J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctute meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29:217–222, 2001.

[16] R. Judson, B. Salisbury, J. Schneider, A. Windemuth, and J. C. Stephens. How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics*, 3:379–391, 2002.

[17] L. Kruglyak. Prospects for whole-genome linkage mapping of common disease genes. *Nature Genetics*, 22:139–144, 1999.

[18] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz. SNPs problems, complexity and algorithms. In *Proceedings of the 9th Annual European Symposium on Algorithms (ESA)*, pages 182–193, 2001.

[19] R. Lippert, R. Schwartz, G. Lancia, and S. Istrail. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in Bioinformatics*, 3(1):23–31, 2002.

[20] D. A. Nickerson, S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark, J. H. Stengaard, V. Salomaa, E. Boerwinkle, and C. F. Sing. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Research*, 10:1532–1545, 2000.

[21] N. Patil et al. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294:1719–1722, 2001.

[22] R. Rizzi, V. Bafna, S. Istrail, and G. Lancia. Practical algorithms for the single individual SNP haplotyping problem. In *Workshop on Algorithms in Bioinformatics*, pages 29–43, 2002.

[23] F. M. De La Vega, X. Su, H. Avi-Itzhak, B. V. Halldórsson, D. Gordon, A. Collins, R. A. Lippert, R. Schwartz, C. Scafe, Y. Wang, M. Laig-Webster, R. T. Koehler, J. Ziegle, L. Wogan, J.F. Stevens, K.M. Leinen, S.J. Olson, K.J. Guegler, X. You, L. Xu., H.G. Hemken, F. Kalush, A. G. Clark, S. Istrail, M. W. Hunkapiller, E. G. Spier, and D. A. Gilbert. The profile of linkage disequilibrium across human chromosomes 6, 21, and 22 in African-American and Caucasian populations. *In preparation*, 2003.

[24] K. Weiss and A. Clark. Linkage diseuilibrium and the mapping of comples human traits. *Trends in Genetics*, 18(1):19–24, 2002.

[25] K. Zhang, M. Deng, T. Chen, M.S. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences*, 99(11):7335–7339, 2002.

7. APPENDIX A

Connecting Informativeness with measures of LD, and Diversity

Zhang et al. [25] also consider the problem of minimizing SNPs based on linkage with other SNPs. They limit their prediction to *blocks*, or regions of low recombination. They formulate the problem as one of predicting 'common' haplotypes (haplotypes that occur at least twice), a measure clearly related to ours. Recall our formula for informativeness of set S' in predicting the set T

$$I(S', T) = \frac{|E(S') \cap E(T)|}{|E(T)|}$$

Let S be the set of SNPs in a haplotype block described by the matrix M_S after eliminating all haplotypes that occur exactly once. The rows of this matrix form the set of common haplotypes H. Observe that $E(S) = H \times H$, the complete graph on the vertex set H. Zhang et al. [25] define an informative set of SNPs $S' \subseteq S$, such that

$$\exists V \subseteq S \text{ s.t. } E(S') = V \times V \text{ and } \frac{|V|}{|H|} = \beta > 0.8 \qquad (7)$$

Note that if $\beta = 1$, the problem would be identical to picking $S' \subseteq S$ such that $E(S') = H \times H = E(S)$. Thus the two measures of informativeness of SNPs are expressed on the same graph, one based on vertices, and the other based on edges. One advantage of our method is that it can be used with respect to a single SNP, or a predetermined group of SNPs, which makes it comparable to other measures in population genetics.

In the population genetics community, various measures for Linkage Disequilibrium and association also measure the informativeness of a SNP with respect to another. A commonly employed measure of linkage disequilibrium between allele A_i

and B_j (at loci A and B, respectively) is D_{ij} defined as

$$D_{ij} = p_{ij} - p_{i.}p_{.j} \tag{8}$$

Where $p_{i.}$ is the probability of seeing allele A_i at loci A, $p_{.j}$ is the probability of seeing allele B_j at loci B and p_{ij} is the probability of seeing allele A_i at loci A and allele B_j at loci B. Measures related to this are

$$D^2 = \sum_{i,j} D_{ij}^2 \tag{9}$$

$$D' = \begin{cases} D_{00}/min(p_{0.}p_{.1}, p_{.0}p_{1.}) & D_{00} > 0 \\ D_{00}/min(p_{0.}p_{.0}, p_{.1}p_{1.}) & D_{00} < 0 \end{cases} \tag{10}$$

While these measures are useful in measuring linkage between SNPs, they are less useful in measuring informativeness, which is the extent to which one SNP can help predict the other. Observe for example that $|D'| = 1$, its maximum value whenever any of the four possible alleles is not present. Assuming the infinite-sites model of evolution, this implies that $|D'| = 1$ when there is no recombination between the two loci. This is not, however, enough for a SNP to predict the other. A more useful measure of LD that also describes informativeness is the d^2 measure [9, 17]. Following Kruglyak [17] denote the variant allele in the target SNP t as v, and the normal allele as $+$. Then the d^2 measure for a SNP s containing alleles 0, and 1 with respect to t is estimated as

$$\left(\frac{n_{v0}}{n_v} - \frac{n_{+0}}{n_+}\right)^2 \tag{11}$$

This measure is 1 exactly when s can completely predict t, and 0 when s provides no information w.r.t to t ($\frac{n_{v0}}{n_v} = \frac{n_{+0}}{n_+} = 0.5$). The d^2 measure does not, however, extend easily to the informativeness of a set of SNPs S' w.r.t t. Let $|S'| = k$ so that 2^k haplotypes are possible. For each haplotype H, its informativeness w.r.t t can be denoted by $\left(\frac{n_{vH}}{n_v} - \frac{n_{+H}}{n_+}\right)^2$. The d^2 measure for S' w.r.t t is given by

$$\max_H \left(\frac{n_{vH}}{n_v} - \frac{n_{+H}}{n_+}\right)^2 \tag{12}$$

However, it is easy to see that this measure is actually quite restrictive. A small subset of SNPs S' could easily provide complete information about t even though the d^2 measure for any one pair of SNPs is quite low. The measure we propose is thus a better measure of informativeness, and also provides insight into linkage between two arbitrary sets of SNPs.