BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i137–i144
DOI: 10.1093/bioinformatics/bth907

High density linkage disequilibrium mapping using models of haplotype block variation

G. Greenspan and D. Geiger*

Computer Science Department, Technion, Haifa 32000, Israel

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: The presence of millions of single nucleotide polymorphisms (SNPs) in the human genome has spurred interest in genetic mapping methods based on linkage disequilibrium. The recently discovered haplotype block structure of human variation promises to improve the effectiveness of these methods. A key difficulty for mapping techniques is the cost involved in separately identifying the haplotypes on each of an individual's chromosomes.

Results: We present a new approach for performing linkage disequilibrium mapping using high density haplotype or genotype data. Our method is based on a statistical model of haplotype block variation, which takes account of recombination hotspots, bottlenecks, genetic drift and mutation. We test our technique on two empirically determined high density datasets, attempting to recover the location of an SNP which was hidden and converted into phenotype information. We compare the results against a mapping method based on individual SNPs as well as a competing haplotype-based approach. We show that our strategy significantly outperforms these other approaches when used as a guide for resequencing and that it can also deal with both unphased genotype data and low penetrance diseases.

Availability: HaploBlock executables for Linux, Mac OS X and Sun OS, as well as user documentation, are available online at http://bioinfo.cs.technion.ac.il/haploblock/

Contact: gdg@cs.technion.ac.il, dang@cs.technion.ac.il

1 INTRODUCTION

The linkage disequilibrium (LD) approach to genetic mapping looks for markers in a candidate region whose alleles are correlated with disease in unrelated individuals. It assumes that each genetic mutation associated with a disease occurred in only a few founding individuals. Any marker allele near a founder mutation is likely to be inherited along with it, due to the low probability of recombination between the two loci. These marker alleles will therefore be more prevalent in contemporary affected individuals than in the rest of the population, generating a correlation that allows the mutation to be mapped. In the wake of the Human Genome Project, millions of single nucleotide polymorphisms (SNPs) have been discovered, opening the way for high density LD studies.

LD mapping studies based on individual SNP markers have met with little success (Risch, 2000; Cardon and Bell, 2001). Even if a single marker is close to the phenotypic site, its degree of correlation with disease will rarely distinguish it from other markers associated by chance. A more powerful approach treats multi-marker haplotypes as the variable for correlation (Botstein and Risch, 2003; Fan and Knapp, 2003). The descendants of a disease founder are more clearly identified by a haplotype than by a single marker since two haplotypes with different lineages are unlikely to be identical at many sites. Nevertheless, tests based on haplotypes must consider the possibility that recombinations and mutations have taken place, complicating the correlation with disease. Many methods for addressing this challenge have been proposed, based on evolutionary trees (Lam *et al*., 2000), haplotype sharing (McPeek and Strahs, 1999; Morris *et al*., 2000), clustering (Liu *et al*., 2001) and the coalescent (Rannala and Reeve, 2001).

In recent years, several studies of human variation have demonstrated the presence of haplotype blocks, defined as genomic regions in which a small number of multi-site variants cover most of the observed variation (Daly *et al*., 2001; Goldstein, 2001; Patil *et al*., 2001; Gabriel *et al*., 2002). These blocks may result from recombination hotspots, which separate between stretches of DNA that are almost never divided during meiosis (Jeffreys *et al*., 2000, 2001). Alternatively, they may be distributed randomly as a result of uniform but rare recombination (Zhang *et al*., 2003). In either case, haplotype blocks are further explained by population phenomena such as bottlenecks and genetic drift, which reduce the amount of variation in a genetic region over time.

A diverse range of block identification criteria have been proposed, based on heterogeneity (Daly *et al*., 2001), haplotype tagging SNP (htSNP) informativeness (Patil *et al*., 2001), linkage disequilibrium (Gabriel *et al*., 2002), the four-gamete test (Wang *et al*., 2002) and statistical model selection (Greenspan and Geiger, 2003; Anderson and Novembre, 2003; Koivisto *et al*., 2003). Each of these methods infers a single block partition for a genomic region, often by dynamic programming (Zhang *et al*., 2002b). Recent research suggests

*To whom correspondence should be addressed.

however that it may be difficult to justify the selection of one partition over another, due to the complex patterns generated by recombination and mutation (Schwartz *et al*., 2003).

Both haplotype block identification and LD mapping in general are made harder by an absence of haplotype phasing information. Standard SNP genotyping processes yield an unordered pair of alleles for each locus, with no information on which alleles are co-located on the same chromosome. A genotype containing s heterozygous sites can be separated into constituent haplotypes in 2^{s-1} different ways. This degeneracy leads to the haplotype resolution problem, which we addressed directly in an earlier work (Greenspan and Geiger, 2003).

In this paper, we proposed a new method for high density LD mapping that takes account of haplotype blocks. For a set of SNPs at known locations, our method analyzes a list of haplotypes or genotypes with corresponding phenotype information, generating a posterior distribution for the position of a phenotype locus over the candidate region. This posterior distribution provides a prioritization for chromosome resequencing, which is the final step required to identify a disease-related mutation.

The rest of this paper is organized as follows. Section 2 describes our statistical model, which summarizes the effects of recombination hotspots, bottlenecks, genetic drift and mutations. Section 3 describes how we infer an ensemble of models from observed data and calculate a posterior distribution for the position of the phenotype locus using this ensemble. Section 4 demonstrates the effectiveness of our technique using real-world SNP data, showing how our method significantly outperforms two other approaches when used as a guide for resequencing. In this section, we also demonstrate that our strategy remains effective for low penetrance diseases and in the absence of haplotype phasing information. Finally, Section 5 presents some further points for discussion.

2 STATISTICAL MODEL

Our model for the haplotype block variation in a genomic region, which was introduced previously (Greenspan and Geiger, 2003), is defined by (a) a partition of the region into blocks, (b) one or more ancestor haplotypes for each block, (c) a Markov chain over the blocks defining the ancestor distributions and (d) site-specific mutation rates reflecting the mutations accumulated since the ancestors were alive. All aspects of the model are inferred from the raw haplotype or genotype data with no other prior knowledge.

We partition a genomic region containing l SNPs into adjacent and contiguous blocks, numbered $1,\ldots,b$, with the indices of the first and last SNP of block k defined by s_k and e_k, respectively. The ancestor haplotypes for block k are numbered $1,\ldots,q_k$. The sequence of ancestor haplotype c of block k is given by $a_{k,c}$, a string of $e_k - s_k + 1$ symbols from the

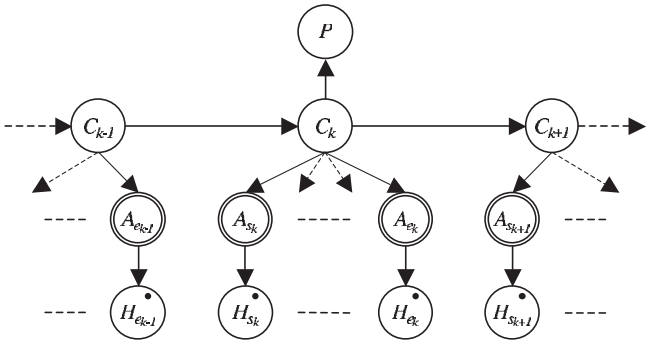

Fig. 1. Bayesian Network for modeling haplotype data.

set $B = \{A, C, G, T, -\}$ of SNP alleles, which contains the four nucleic acids and a deletion. The probability that a haplotype is descended from ancestor c in the first block is defined by the parameter $\theta_{1,c}$. For subsequent blocks, $\theta_{k,c'\to c}$ defines the probability that a haplotype is descended from ancestor c in block k, given that it is descended from ancestor c' in the previous block $k-1$. Note that the directionality of this dependency between adjacent blocks is meaningless, as with any Markov chain. The mutation parameter $\mu_{j,a\to h}$ denotes the probability that ancestral allele a at site j is observed today as allele h. Population genetic considerations led to a constraint of $10^{-6} \le \mu_{j,a\to h} \le 10^{-3}$ for all $a \neq h$ (Greenspan and Geiger, 2003).

The joint distribution defined by our model can be concisely depicted using a Bayesian Network. A Bayesian Network is a directed acyclic graph, in which each node represents a variable and each variable's distribution is dependent on those which point to it (Pearl, 1988; Jensen, 1996). Using this representation, general probability computations can be performed efficiently by bucket variable elimination (Dechter, 1996). Furthermore, suitable parameters for the conditional distributions can be inferred from observed data using the EM algorithm (Lauritzen, 1995).

The Bayesian Network corresponding to our model is shown in Figure 1. It contains a random variable C_k for each block $k = 1,\ldots,b$ and two random variables A_j and H_j for each SNP $j = 1,\ldots,l$. Variable P will be discussed later in Section 3.1. Each variable C_k defines the ancestor from which a haplotype is descended in block k. For the first block, $\Pr(C_1 = c) = \theta_{1,c}$ and for subsequent blocks, $\Pr(C_k = c \mid C_{k-1} = c') = \theta_{k,c'\to c}$. For each block k, variables $A_{s_k},\ldots,A_{e_k}$ define the sequence of the ancestor indicated by the value of C_k. For SNP j in block k, $\Pr(A_j = a \mid C_k = c) = 1$ if $a_{k,c,j} = a$ and 0 otherwise. Variables $H_1,\ldots,H_l$ define the observed haplotype data over loci $1,\ldots,l$, where $\Pr(H_j = h \mid A_j = a) = \mu_{j,a\to h}$ for each SNP j. The double borders in Figure 1 denote that variables A_j are deterministic and the black dots indicate that variables H_j are observed.

Let $\delta(x, y) = 1$ if $x = y$ and 0 otherwise. The Bayesian Network defines the joint distribution over all variables

$\Pr(c_1, \ldots, c_b, a_1, \ldots, a_l, h_1, \ldots h_l)$ as:

$$\theta_{1,c_1} \prod_{k=2}^{b} \theta_{k,c_{k-1} \to c_k} \prod_{k=1}^{b} \prod_{j=s_k}^{e_k} \delta(a_{k,c_k,j}, a_j) \cdot \mu_{j,a_j \to h_j}. \quad (1)$$

The likelihood $\Pr(h_1, \ldots, h_l)$ of a haplotype h is the sum of the joint distribution in Equation (1) over all values of the unobserved variables, calculated efficiently by bucket variable elimination (Dechter, 1996).

Our statistical model represents a series of multiple star genealogies, one for each haplotype block. Each block ancestor corresponds to the center of one star, while the haplotypes descended from that ancestor correspond to the star's points. The Markov chain expresses the dependencies between the block genealogies, reflecting the fact that linkage disequilibrium exists between blocks as well as within them. Tests on real-world data confirm that a Markov chain provides a close approximation to true haplotype distributions observed (data not shown). By allowing mutation rates to be site-specific and allele-specific, we consider the possibility of mutation hotspots and coldspots with different substitution patterns. Other biological assumptions underlying our model were discussed previously (Greenspan and Geiger, 2003).

To assess the suitability of a particular model M for representing an observed dataset $\mathcal{D}$, we use the minimum description length (MDL) criterion, which considers both the complexity of M and the likelihood of $\mathcal{D}$ under M (Rissanen, 1978). This criterion is based on the total amount of information required to transmit data $\mathcal{D}$ using model M, as denoted by $DL(\mathcal{D}, M)$. If $\mathrm{DL}(M)$ bits are required to represent a model M then $\mathrm{DL}(\mathcal{D}, M) = \mathrm{DL}(M) - \log_2 \Pr(\mathcal{D}|M)$. The description length $\mathrm{DL}(M)$ of model M is calculated using an efficient representation of the model parameters, as described previously (Greenspan and Geiger, 2003). Assuming independence, the likelihood $\Pr(\mathcal{D} \mid M)$ is the product of the likelihoods of each sample in $\mathcal{D}$ under model M.

For genotype data, we require an extended model to deal with the lack of phasing information. The genotype model, depicted in Figure 2, contains two identical copies of the haplotype model, where the mirrored copy has variables renamed to C'_k, A'_j and H'_j. The new deterministic variable G_j corresponds to the joint observation at site j, taking values from the set D of unordered pairs of SNP alleles, where $D = \{[b_1, b_2] : b_1, b_2 \in B\}$. The value of G_j is fixed by the alleles present on each chromosome at site j, so that $\Pr(g_j | h_j, h'_j) = 1$ if $g_j = [h_j, h'_j]$ and 0 otherwise.

3 MAPPING USING MODELS

A high density LD mapping study is based on a list $\mathcal{H} = \{h^1, \ldots, h^n\}$ of n phased haplotypes or a list $\mathcal{G} = \{g^1, \ldots, g^n\}$ of n unphased genotypes over the entire region of interest. We use the symbol $\mathcal{D}$ to refer to input $\mathcal{H}$ or $\mathcal{G}$ as appropriate. The other inputs are a list $\mathcal{P} = \{p^1, \ldots, p^n\}$ of phenotypes

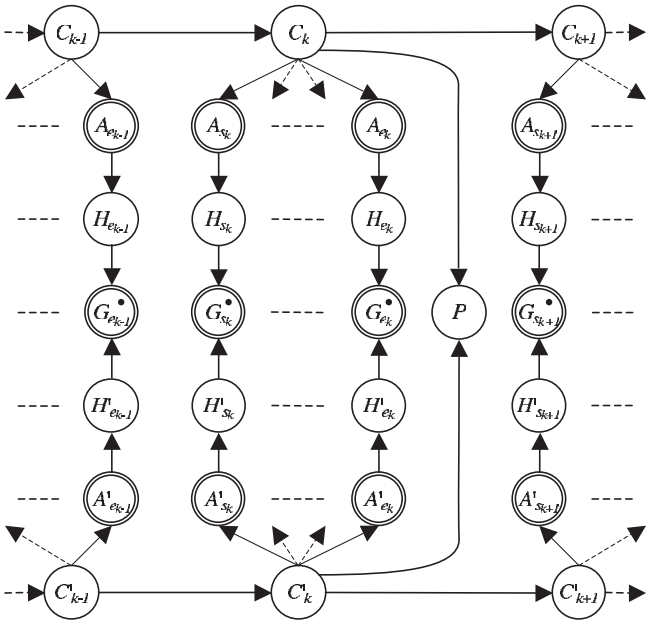

Fig. 2. Bayesian Network for modeling genotype data.

associated with each haplotype or genotype and the distances d_j in base pairs between adjacent SNPs j and $j+1$ over $j = 1, \ldots, l-1$. For haplotype mapping, each haplotype h^i is a string of l symbols from the set B of SNP alleles, where l is the number of loci examined. For genotype mapping, each genotype g^i is a string of l elements from the set D of unordered SNP allele pairs. Each p^i is in the range $1, \ldots, p_{\max}$, where $p_{\max}$ is the total number of phenotypes observed. In a simple case–control study, $p_{\max} = 2$.

We are searching for an unobserved genetic locus within the candidate region that affects the phenotypes observed. Let L_j denote the hypothesis that this locus is situated in the interval between SNPs j and $j+1$, so that we consider the set of hypotheses $\{L_1, \ldots, L_{l-1}\}$. We express the output of a mapping study as a posterior distribution $\Pr(L_j \mid \mathcal{P}, \mathcal{D})$ over these alternatives, normalized so that $\sum_{j=1}^{l-1} \Pr(L_j \mid \mathcal{P}, \mathcal{D}) = 1$. This distribution is calculated in the following four stages.

First, we infer an ensemble $\mathcal{M}$ of statistical models which are locally optimal in terms of the MDL criterion, i.e. those which provide a compact explanation of the observed data $\mathcal{D}$. We ignore the phenotypes $\mathcal{P}$ during this process, since they barely affect the data likelihood. We explore the search space of models using Gibbs-style iterations, in which the existence and location of each block divider constitute the variable for resampling. The initial model has dividers distributed evenly over the region. During a sampling iteration, each of the dividers in the current model is removed in turn to create a larger block, into which we attempt to add up to three new dividers at optimal locations, so long as this improves the MDL score. All model parameters are optimized at each stage of this process, using the local search and modified

EM algorithms described previously (Greenspan and Geiger, 2003).

Second, for each model M in the ensemble $\mathcal{M}$, we calculate the posterior probability that each block contains the phenotypic locus. Let U_k denote the hypothesis that the locus is in block k of M. The posterior distribution $\Pr(U_k \mid \mathcal{P}, \mathcal{D}, M)$ is calculated using the method described in Section 3.1 or 3.2 as appropriate. Note that at this stage the phenotype data is used to assess hypotheses relating to blocks, rather than SNP intervals, since each model inferred assumes that the alleles within each block segregate together.

Third, the posterior distribution $\Pr(U_k \mid \mathcal{P}, \mathcal{D}, M)$ over the blocks in model M is converted into a posterior $\Pr(L_j \mid \mathcal{P}, \mathcal{D}, M)$ over SNP intervals. For an interval $(j, j+1)$ in block k, for which $s_k \leq j < e_k$, we allocate the posterior in proportion to the length d_j of the interval, setting $\Pr(L_j \mid \mathcal{P}, \mathcal{D}, M) = d_j \cdot \Pr(U_k \mid \mathcal{P}, \mathcal{D}, M)/V_k$, where V_k is the total length of block k. For an interval $(j, j+1)$ on the boundary between blocks k and $k+1$, for which $j = s_{k+1} - 1 = e_k$, we assume that half of the interval lies in each block, setting $\Pr(L_j \mid \mathcal{P}, \mathcal{D}, M) = d_j \cdot \Pr(U_k \mid \mathcal{P}, \mathcal{D}, M)/2V_k + d_j \cdot \Pr(U_{k+1} \mid \mathcal{P}, \mathcal{D}, M)/2V_{k+1}$. The block length V_k is obtained by summing the interlocus distances d_j within the block and half of those at either end, i.e. $V_k = \sum_{j=s_k}^{e_k-1} d_j + \frac{1}{2}(d_{s_k-1} + d_{e_k})$. Note that V_1 and V_b lose elements d_{s_k-1} and d_{e_k} respectively from this sum, where b is the number of blocks in the model.

In the fourth and final stage, the individual posterior distributions $\Pr(L_j \mid \mathcal{P}, \mathcal{D}, M)$ obtained from each model M in the ensemble $\mathcal{M}$ are combined into a single statistic by uniform model averaging, so that $\Pr(L_j \mid \mathcal{P}, \mathcal{D}) = \sum_{M \in \mathcal{M}} \Pr(L_j \mid \mathcal{P}, \mathcal{D}, M)/1/|\mathcal{M}|$. We use a uniform prior for the averaging since the sampling process has already introduced a strong bias toward models with a low MDL score.

3.1 Haplotypes posterior

Recall that hypothesis U_k states that the phenotypic locus is located in block k of a model. Under Bayes' Rule, the posterior probability of hypothesis U_k is given by

$$\Pr(U_k \mid \mathcal{P}, \mathcal{H}, M) = \frac{\Pr(\mathcal{P} \mid U_k, \mathcal{H}, M) \Pr(U_k \mid \mathcal{H}, M)}{\Pr(\mathcal{P} \mid \mathcal{H}, M)}.$$

Since $\Pr(\mathcal{P} \mid \mathcal{H}, M)$ is the same for all k and we assume that the prior $\Pr(U_k \mid \mathcal{H}, M)$ does not depend on $\mathcal{H}$, this can be rewritten as

$$\Pr(U_k \mid \mathcal{P}, \mathcal{H}, M) \propto \Pr(\mathcal{P} \mid U_k, \mathcal{H}, M) \Pr(U_k \mid M). \quad (2)$$

In this equation, $\Pr(U_k \mid M)$ is the prior probability that block k of model M contains a locus which affects the observed phenotypes, while $\Pr(\mathcal{P} \mid U_k, \mathcal{H}, M)$ is the posterior probability of phenotypes $\mathcal{P}$ given haplotypes $\mathcal{H}$ under that assumption.

Phenotype information is expressed as the variable P in our model. Under hypothesis U_k, P is directly dependent only on variable C_k, as depicted in Figure 1. This simple dependence is sufficient because the differences in ancestry reflected by variable C_k capture the ancestral variation at all loci within block k, including those which are not observed.

We approximate the term $\Pr(\mathcal{P} \mid U_k, \mathcal{H}, M)$ of Equation (2) by assuming sample independence and inferring maximum-likelihood parameters for $\Pr(P \mid C_k, M)$. These parameters are obtained using the EM algorithm with the haplotypes $\mathcal{H}$ and phenotypes $\mathcal{P}$ as evidence (Lauritzen, 1995). The subsequence of each haplotype for block k is usually compatible with only one value of C_k, so the EM algorithm converges uniquely and quickly.

The prior probability $\Pr(U_k \mid M)$ of Equation (2) is based on two elements. The first element assigns probability in proportion to V_k, the length of block k. The second element adjusts for the fact that blocks with more ancestors have more parameters for maximizing the likelihood $\Pr(\mathcal{P} \mid U_k, \mathcal{H}, M)$. We compensate by considering the optimal number of bits W_k required to represent $\Pr(P \mid C_k, M)$. Using a standard encoding, $W_k = q_k \cdot (p_{\max} - 1) \log_2 n/2$, where q_k is the number of ancestors for block k, $p_{\max}$ is the number of phenotypes and n is the number of samples observed (Rissanen, 1983). Applying the MDL schema, elements V_k and W_k are combined to obtain $\Pr(U_k \mid M) \propto V_k \cdot 2^{-W_k}$ (Rissanen, 1978).

3.2 Genotypes posterior

For genotype data, the posterior distribution $\Pr(U_k \mid \mathcal{P}, \mathcal{G}, M)$ is obtained in a similar manner as for haplotypes. Equation (2) is trivially rewritten as

$$\Pr(U_k \mid \mathcal{P}, \mathcal{G}, M) \propto \Pr(\mathcal{P} \mid U_k, \mathcal{G}, M) \Pr(U_k \mid M). \quad (3)$$

As before, we represent phenotype information as the variable P in our model. For dominant, recessive and co-dominant disease models, the phenotype is affected by genetic variation in both chromosomes. Therefore, under hypothesis U_k, P depends on both variables C_k and C'_k, as depicted in Figure 2. The differences between haplotype and genotype posterior calculations stem only from this more complex dependency.

Element $\Pr(\mathcal{P} \mid U_k, \mathcal{G}, M)$ of Equation (3) is calculated as before by assuming sample independence and inferring the parameters of $\Pr(P \mid C_k, C'_k, M)$ by EM. This distribution is symmetrical for the two variables C_k and C'_k, reflecting the functional symmetry between the maternal and paternal chromosomes in a cell.

The prior probability $\Pr(U_k \mid M)$ of Equation (3) is also calculated as before, based on the length V_k and the number of bits W_k required to represent $\Pr(P \mid C_k, C'_k, M)$. Since the distribution $\Pr(P \mid C_k, C'_k, M)$ is symmetrical, we set $W_k = q_k \cdot (q_k + 1) \cdot (p_{\max} - 1) \log_2 n/4$. The two elements are combined as before so that $\Pr(U_k \mid M) \propto V_k \cdot 2^{-W_k}$.

Table 1. Outcomes for full penetrance haplotype tests

Data set and SNP range	Target SNP	Individual		BLADE		HaploBlock	
		Rank	Sequence (kb)	Rank	Sequence (kb)	Rank	Sequence (kb)
5q31	3	1	7	8	71	3	7
	7	5	43	1	80	1	68
	21	5	14	18	17	1	5
	80	69	336	54	277	7	111
	84	54	255	9	273	1	9
	Mean	13.0	131	9.6	144	2.6	40
Chromosome 21							
3877–4077	4063	2	2	114	140	3	12
8538–8738	8597	28	101	7	20	1	17
15 510–15 710	15 607	1	17	104	267	1	24
15 855–16 055	15 870	2	8	9	52	1	10
16 807–17 007	16 918	36	38	27	60	33	57
	Mean	13.8	33	16.4	107	7.8	24

For each algorithm, the 'Rank' column shows the position of the interval containing the target SNP in a ranking of intervals according to the posterior probability assigned. The 'Sequence' column shows how much of the region would be resequenced before finding the target SNP when resequencing intervals in descending order of posterior density.

4 RESULTS

4.1 Full penetrance haplotype mapping

We assessed our mapping technique using two large sets of empirically determined human haplotypes: (a) 258 transmitted haplotypes for 98 SNPs over 464 kb in the 5q31 region (Daly *et al.*, 2001) and (b) 20 haplotypes over the whole of chromosome 21 (Patil *et al.*, 2001).

Each test set was generated from a set of haplotypes by randomly selecting a target SNP to be converted into phenotype information. Each haplotype was assigned the phenotype corresponding to the allele it possessed for this SNP, which was then removed from the marker data—the goal of the mapping algorithm was to recover its location. Since all SNPs were biallelic, haplotypes which had the more common allele for the target SNP were labeled as 'healthy' while the others were labeled 'diseased'. This mirrors the LD mapping problem for high penetrance diseases, where a hidden locus which determines phenotypic differences must be found.

For the 5q31 data, we created five separate test sets, selecting SNPs as the target with probability in proportion to the distance between their neighboring SNPs. For chromosome 21, we used five randomly selected contiguous subsets of 201 SNPs from the NT_002836 contig, then created a single test set from each subset as before. We removed those few haplotypes from test sets for which the target SNP allele was unknown.

For each test set, we obtained the distribution $\Pr(L_j \mid \mathcal{P}, \mathcal{D})$ by inferring an ensemble of 100 models as described in Section 3. For comparison, we also obtained posteriors from the BLADE algorithm, allowing it to optimize the number of founders using the MAP criterion (Liu *et al.*, 2001). We further calculated a distribution using a version of our model with no interlocus dependencies, considering each SNP individually as an independent 'block'. We tried to include three other software packages in our comparison, however each proved unobtainable or unsuitable for datasets with a large number of SNPs (Lam *et al.*, 2000; McPeek and Strahs, 1999; Rannala and Reeve, 2001).

Table 1 lists the results for each test set. For each algorithm, the first column shows the position of the interval containing the target SNP, in a ranking of intervals according to their posterior probability. The ranking compared 96 intervals for the 5q31 data, and 199 intervals for each chromosome 21 test set. Note that larger intervals rank higher under any algorithm, so this statistic is not ideal for comparative study.

To generate a better statistic, we used the posterior density of each interval [i.e. $\Pr(L_j \mid \mathcal{P}, \mathcal{D})/d_j$] to determine a resequencing prioritization. We assumed that SNP intervals would be resequenced in descending order of posterior density until the target SNP was found. The second column for each algorithm shows how much of the candidate region would have to be resequenced under this scheme. In the absence of any mapping information, we would expect this to be half of the region's length, i.e. 232 kb for the 5q31 data and 99, 82, 248, 167 and 201 kb, respectively, for each of the chromosome 21 test sets.

In 6 out of the 10 tests, HaploBlock ranked the target SNP interval first, whereas the individual SNP and BLADE approaches did so twice and once, respectively. In terms of the resequencing prioritization, HaploBlock also comfortably outperformed the other two approaches. This is particularly notable for the 5q31 region, in which it required an average of 40 kb instead of 131 and 144 kb, saving ∼70% in resequencing costs.

It is instructive to examine the results for the 5q31 dataset with target SNP 21, in which all three algorithms performed

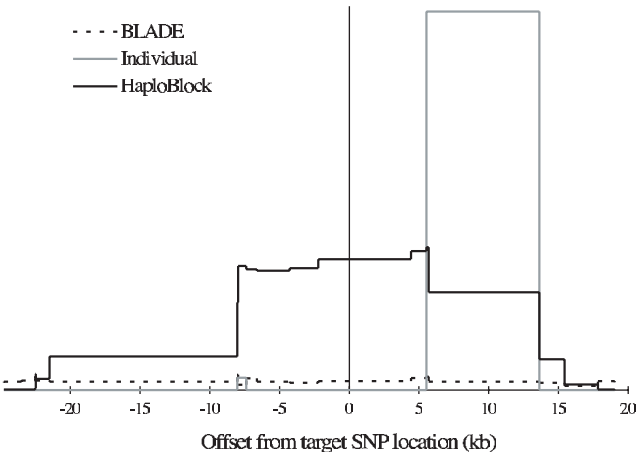

Fig. 3. Posterior densities for SNP 21 in haplotype dataset 5q31.

reasonably well. Figure 3 depicts the posterior density curve assigned by each algorithm in the immediate vicinity of the hidden target SNP. The BLADE algorithm failed to find any significant peak in this area, although it did assign a posterior density to a 50 kb window containing the target that was slightly higher than in the rest of the region. The individual SNP method assigned a peak window between SNP 23 (5.5 kb downstream of target) and SNP 25 (13.5 kb downstream), reflecting a strong association between the phenotypes and SNP 24. While close by, this window failed to include SNP 21, since both SNPs 22 and 23 were poorly correlated with the phenotypes. By contrast, HaploBlock assigned a wider peak which was well centered around the target, reflecting its location within a block whose haplotypes were strongly associated with the phenotypes. It is interesting to note that the original 5q31 analysis assigned a block from SNP 16 (8 kb upstream of target) to SNP 24 (6 kb downstream) (Daly *et al*., 2001). Similarly, 84 of the 100 models sampled by HaploBlock placed SNPs 16–23 in a single block.

HaploBlock performed relatively poorly in terms of resequencing length for SNPs 7 and 80 in the 5q31 region and the last test set for chromosome 21. In all three cases, the target SNP was strongly associated with several haplotype blocks in the surrounding region, reducing the resolution that HaploBlock was able to achieve. Nonetheless, it is encouraging to note that the target was always included in the window of high posterior density output by HaploBlock, while this was not the case for the other two approaches (graphs not shown).

4.2 Genotypes and partial penetrance

We also assessed the effectiveness of our LD mapping method using unphased genotype marker measurements and/or a partial penetrance model. We based the genotype tests on the 129 offspring in region 5q31, while the haplotype tests used the same 258 haplotypes from 5q31 as before. The chromosome 21 data were not used since they contain too few samples for partial penetrance mapping to be viable.

For a phenotype with penetrance p, the disease status was assigned with probability p to haplotypes with the rare allele for the target SNP, while all others were assigned healthy. For genotypes, this model was applied independently to both alleles before combining the results under a co-dominant model to generate three phenotype assignments.

Table 2 compares the results of mapping haplotypes and genotypes with varying degrees of penetrance. The results show that our approach remains effective in the absence of phasing information. For genotypes with full penetrance, a mean rank and resequencing length of (3.2, 37 kb) was achieved, compared with (2.6, 40 kb) for haplotypes. Furthermore, our technique exhibits a similar deterioration in performance for haplotypes and genotypes, achieving (8.8, 118 kb) and (8.2, 113 kb) respectively at 10% penetrance.

The running time for HaploBlock is highly dependent on the parameters of the models inferred. If a bound is placed on the maximum number of SNPs and ancestors in any block, the time complexity is $O(l \cdot n \cdot s)$, where l is the number of SNPs, n is the number of haplotypes or genotypes and s is the number of models to be sampled. In practice, unphased genotypes take much longer to analyze than haplotypes, due to the extra complexity of the calculations involved. On a 2 GHz Pentium IV workstation, HaploBlock took about 15 min of CPU time to analyze each chromosome 21 test set (200 SNPs, 20 haplotypes) and about 3 and 40 h respectively for each set of 5q31 haplotypes and genotypes (97 SNPs, 258 haplotypes or 129 genotypes).

5 DISCUSSION

There is an ongoing debate over whether haplotype blocks are generated by recombination hotspots, or arise from other population processes (Phillips *et al*., 2003; Zhang *et al*., 2003). Our Markov model is neutral on this question, since it captures the dependencies between adjacent blocks in either case. Nonetheless, a future direction for research is to examine high density datasets in order to address the issue directly.

Although we demonstrated our method using real-world haplotypes and genotypes, we were forced to simulate phenotypes using a target SNP, since we could locate no publicly available datasets which combine high density SNP data with phenotype information. We wish to apply our approach to such data in future, either as part of a new mapping study or to confirm the effects of a locus whose position is known.

The experiments performed in this paper were based on a model in which phenotypes were affected by a single locus in the region of interest. However, it is expected that LD mapping techniques will also prove useful for mapping complex diseases, in which phenotypes are the product of interactions between multiple loci as well as non-genetic factors. To fully address this problem, our model would have to be extended to allow multiple haplotype blocks to influence the phenotypes, via an explicit model of interaction that would reduce the number of parameters to be inferred. Nonetheless, the

Table 2. Outcomes for HaploBlock for genotype and partial penetrance tests

Data type (statistic)	Penetrance (%)	Index of target SNP in 5q31 dataset					
		3	7	21	80	84	Mean
Haplotypes (rank)	100	3	1	1	7	1	2.6
	50	3	1	1	10	3	3.6
	25	3	1	1	4	17	5.2
	10	3	1	13	16	11	8.8
Genotypes (rank)	100	3	1	2	7	3	3.2
	50	5	1	2	17	11	7.2
	25	5	1	2	18	16	8.4
	10	5	1	3	21	11	8.2
Haplotypes (sequence, kb)	100	7	68	5	111	9	40
	50	66	68	5	117	42	60
	25	78	68	5	244	51	89
	10	78	68	104	229	109	118
Genotypes (sequence, kb)	100	7	55	5	76	42	37
	50	123	68	5	133	130	92
	25	39	55	13	217	179	100
	10	61	87	13	237	167	113

results for the partial penetrance tests indicate that our method is already useful for individually detecting loci with simple additive or multiplicative interactions.

Two other MDL approaches to modeling haplotype block variation have recently been published, either of which could be used as a basis for our LD mapping technique. Koivisto *et al*. (2003) identify up to 10 haplotype clusters within each block using k-means clustering. Each haplotype cluster defines an independent distribution for the alleles at each SNP, with no constraint on the distribution's parameters. This contrasts with our ancestor plus mutation model, which expresses the variation within each cluster as the result of mutations since a founding bottleneck event. Koivisto *et al*. (2003) consider the ancestry for each block independently, allowing the optimal partition to be identified using dynamic programming (Zhang *et al*., 2002b). In the language of our approach, their model conflates variables A_j and H_j in Figure 1 and removes the Markov chain connecting variables C_k.

Anderson and Novembre (2003) apply a different model, in which they enumerate the different haplotypes observed within each block without clustering by similarity or ancestry. As in our technique, Anderson and Novembre represent the dependencies between adjacent haplotype blocks using a Markov chain. However, since their enumeration approach is liable to identify a large number of different haplotypes for each block, they save space in their model description by storing only selected parameters of this chain, setting the probability of the other haplotypes according to their marginal frequencies. Interestingly, Anderson and Novembre (2003) develop a dynamic programming algorithm to infer the globally optimal block partition in the presence of dependencies between adjacent blocks, which may be applicable with some modifications to our own work.

One clear advantage of our statistical model over these others is its ability to represent unphased genotype data. This allows it to be applied for haplotype resolution and LD mapping in the absence of phasing information. Another of its strengths is that missing data are dealt with naturally within the Bayesian Network framework, by summing over the variables for loci that are not observed. Both Koivisto *et al*. (2003) and Anderson and Novembre (2003) use dynamic programming to infer a single globally optimal partition for a genomic region. By contrast, we infer an ensemble of locally optimal models to allow for the ambiguity of block partitioning. Further research is required to determine which of these approaches is more fruitful.

In this paper, we described a mapping method which uses a full set of SNP measurements taken from a group of subjects. However, it is hoped that haplotype blocks will lead to cost savings in LD studies by reducing the number of SNP measurements required (Zhang *et al*., 2002*a*; Cardon and Abecasis, 2003). A pilot study is initially performed on a few subjects, from which the structure of haplotype block variation is inferred. The htSNPs are then selected to identify the common variants within each block. Measurements taken at these htSNPs from the full set of subjects are extrapolated into full haplotypes based on the pilot study. Our statistical model could be applied to this strategy, using the full SNP measurements taken in the pilot study to infer an ensemble of models. The htSNPs would then be identified as the most informative SNPs in the context of this ensemble. Measurements taken at these htSNPs would be used with our technique

by setting the alleles at all other SNPs to be unknown. Since our Bayesian Network model deals naturally with any number of unobserved variable values, ancestry would be inferred from the htSNPs as intended and the unmeasured loci would be ignored.

ACKNOWLEDGEMENTS

We wish to thank Xin Lu, Mary Sara McPeek, Jeff Reeve, Kathryn Roeder and Andrew Strahs for replying to issues which arose with their programs. We thank Perlegen Sciences for making their chromosome 21 data available online. This research was supported by the Israel Science Foundation.

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.

Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33** (Suppl.), 228–237.

Cardon,L. and Abecasis,G. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet.*, **19**, 135–140.

Cardon,L. and Bell,J. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.*, **2**, 91–99.

Daly,M., Rioux,J., Schaffner,S., Hudson,T. and Lander,E. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.

Dechter,R. (1996) Bucket elimination: a unifying framework for probabilistic inference. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence* (*UAI-96*), Morgan Kaufmann, San Francisco, CA. pp. 211–219.

Fan,R. and Knapp,M. (2003) Genome association studies of complex diseases by case–control designs. *Am. J. Hum. Genet.*, **72**, 850–868.

Gabriel,S., Schaffner,S., Nguyen,H., Moore,J., Ray,J., Blumensteil,B., Higgings,J., De Felice,M., Loohner,A., Fagger,M. *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Goldstein,D. (2001) Islands of linkage disequilibrium. *Nat. Genet.*, **29**, 109–111.

Greenspan,G. and Geiger,D. (2003) Model-based inference of haplotype block variation. *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology* (*RECOMB 2003*), pp. 131–137.

Jeffreys,A., Kauppi,L. and Neumann,R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, **29**, 217–222.

Jeffreys,A., Ritchie,A. and Neumann,R. (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.*, **9**, 725–733.

Jensen,F. (1996) *An Introduction to Bayesian Networks*. Springer Verlag, New York.

Koivisto,M., Perola,M., Varilo,T., Hennah,W., Ekelund,J., Lukk,M., Peltonen,L., Ukkonen,E. and Mannila,H. (2003) An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. *Proceedings of the eighth Pacific Symposium on Biocomputing* (*PSB'03*), pp. 502–513.

Lam,J., Roeder,K. and Devlin,B. (2000) Haplotype fine mapping by evolutionary trees. *Am. J. Hum. Genet.*, **66**, 659–673.

Lauritzen,S. (1995) The EM algorithm for graphical association models with missing data. *Comp. Stat. Data Anal.*, **19**, 191–201.

Liu,J., Sabatti,C., Teng,J., Keats,B. and Risch,N. (2001) Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.*, **11**, 1716–24.

McPeek,M. and Strahs,A. (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.*, **65**, 858–875.

Morris,A., Whittaker,J. and Balding,D. (2000) Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am. J. Hum. Genet.*, **67**, 155–169.

Patil,N., Berno,A., Hinds,D., Barrett,W., Doshi,J.M., Hacker,C.R., Kautzer,C.R., Lee,D.H., Marjoribanks,C., McDonough,D.P. *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Phillips,M., Lawrence,R., Sachidanandam,R., Morris,A., Balding,D., Donaldson,M., Studebaker,J., Ankener,W., Alfisi,S., Kuo,F. *et al*. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.*, **33**, 382–387.

Rannala,B. and Reeve,J. (2001) High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.*, **69**, 159–178.

Risch,N. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.

Rissanen,J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465–471.

Rissanen,J. (1983) A universal prior for integers and estimation by minimum description length. *Ann. Stat.*, **11**, 416–431.

Schwartz,R., Halldorsson,B., Bafna,V., Clark,A. and Istrail,S. (2003) Robustness of inference of haplotype block structure. *J. Comput. Biol.*, **10**, 13–19.

Wang,N., Akey,J., Zhang,K., Chakraborty,R. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.*, **71**, 1227–1234.

Zhang,K., Akey,J., Wang,N., Xiong,M., Chakraborty,R. and Jin,L. (2003) Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Hum. Genet.*, **113**, 51–59.

Zhang,K., Calabrese,P., Nordborg,M. and Sun,F. (2002a) Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.*, **71**, 1386–1394.

Zhang,K., Deng,M., Chen,T., Waterman,M. and Sun,F. (2002b) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.