Reconstructing Ancestral Haplotypes with a Dictionary Model

KRISTIN L. AYERS[1], CHIARA SABATTI[2] and KENNETH LANGE[3*]

Departments of Biomathematics[1,3], Human Genetics[2,3], and Statistics[2,3],

University of California, Los Angeles, CA 90095–1766.

UCLA Statistic Department Preprint # 418

March 28, 2005

Running head Dictionary Model for Haplotypes

Keywords linkage disequilibrium; haplotype blocks; minimum description length; forward and backwards algorithms; EM algorithm.

Corresponding author Kenneth Lange

Department of Human Genetics
UCLA School of Medicine
695 Charles E. Young Drive South
Los Angeles, California 90095-7088 (USA)
e-mail: klange@ucla.edu
phone: (310) 206-8076
fax: (310) 825-8685

ABSTRACT

We propose a dictionary model for haplotypes. According to the model, a haplotype is constructed by randomly concatenating haplotype segments from a given dictionary of segments. A haplotype block is defined as a set of haplotype segments that begin and end with the same pair of markers. In this framework, haplotype blocks can overlap, and the model provides a setting for testing the accuracy of simpler models invoking only nonoverlapping blocks. Each haplotype segment in a dictionary has an assigned probability and alternate spellings that account for genotyping errors and mutation. The model also allows for missing data, unphased genotypes, and prior distribution of parameters. Likelihood evaluations rely on forward and backward recurrences similar to the ones encountered in hidden Markov models. Parameter estimation is carried out with an EM algorithm. The search for the optimal dictionary is a particularly difficult because of the variable dimension of the model space. We define a minimum description length criteria to evaluate each dictionary and use a combination of greedy search and careful initialization to select a best dictionary for a given data set. Application of the model to simulated data gives encouraging results. In a real data set, we are able to reconstruct a parsimonious dictionary that captures patterns of linkage disequilibrium well.

1. INTRODUCTION

Recent high-density genotyping of unrelated individuals has revealed limited haplotype diversity and wide variation in linkage disequilibrium levels across the human genome (Patil *et al.*, 2001; Daly *et al.*, 2001; Jeffreys *et al.*, 2001; Reich *et al.*, 2001; Gabriel *et al.*, 2002). In many narrow genome regions, a handful of conserved haplotypes account for almost all chromosomal variation. Haplotypes spanning several such adjacent regions often appear to be formed by concatenating short haplotypes segments drawn from nonoverlapping haplotype blocks. One can explain these observations by postulating that block boundaries coincide with recombination hot spots (Jeffreys *et al.*, 2001; Cullen *et al.*, 2002; Gabriel *et al.*, 2002; Kauppi *et al.*, 2003) and that population bottlenecks and genetic drift act to limit haplotype diversity (Wang *et al.*, 2002). Although the evidence in favor of limited haplotype diversity is compelling, the hypothesis of universal block boundaries between conserved haplotype segments is, in our view, largely untested. The obvious danger is that the sharp boundaries currently seen are artifacts of the simple computational models used to analyze haplotype data. The goal of the present paper is to explore a more complicated statistical model that captures limited haplotype diversity and varying patterns of linkage disequilibrium within the framework of overlapping block boundaries.

A more sophisticated model has important implications for disease gene mapping. A better representation of haplotype conservation will enable a more parsimonious selection of markers for genotyping. It will also promote greater statistical efficiency in association testing with cases and controls, if no other reason than that it will lead to better haplotyping. The dictionary model we introduce is purely phenomenological and does not rely on detailed assumptions about the evolution of the underlying population. Building a model that incorporates evolutionary history is apt to be frustrated by great uncertainties and nearly insurmountable computational barriers. In contrast, our dictionary model permits fast, flexible parsing of long haplotypes.

2. THE DICTIONARY MODEL

The data we seek to model consists of n inferred haplotypes $(h_1, \ldots, h_n)$ gathered on m consecutive linked markers labeled 1 through m. When phase information is lacking, these haplotypes collapse to $t = n/2$ multilocus genotypes $(g_1, \ldots, g_t)$. For the sake of simplicity, we assume temporarily that haplotypes are available. Figure I illustrates the main features of the dictionary model. Each of the observed haplotypes h_1, h_2, and h_3 in the figure is constructed by concatenating a sequence of haplotype segments. One can draw an analogy with an ordinary dictionary by equating haplotypes to sentences and haplotype segments to words. The analogy is only partial because a haplotype segment is always constrained to begin and end with the same pair of markers. The alphabet (set of alleles) also varies from marker to marker. Finally in our haplotype dictionary model, the boundaries separating haplotype segments are hidden. This ambiguity renders statistical inference difficult.

If h is an observed haplotype, we will denote its alleles between markers i and j by the haplotype segment $h[i:j]$. In this notation, $h = h[1:m]$. At the risk of some confusion, we will call a consecutive set of markers

$$[i:j] \quad = \quad \{k : i \leq k \leq j\}$$

a marker segment with boundaries i and j. When a random haplotype H is constructed and one of its concatenated segments exactly spans the marker segment $[i:j]$, the haplotype segment $H[i:j]$ must be drawn from a collection of $\mathcal{B}_{[i:j]}$ of permitted haplotype segments. The collection $\mathcal{B}_{[i:j]}$ is said to be a haplotype block. The block boundaries of H determine a partition π that divides the m markers into consecutive marker segments π_1 through $\pi_{|\pi|}$ with the following properties: (a) if segment π_ℓ ends with marker j, then segment $\pi_{\ell+1}$ begins with marker $j+1$, (b) the first segment π_1 begins with marker 1, and (c) the last segment $\pi_{|\pi|}$ ends with the last marker m. In most haplotype block models, only one partition is feasible. With overlapping blocks, many different partitions are possible.

As concrete examples of these conventions, haplotypes h_1 and h_2 in Figure I share the marker segment $[8:10]$ and the particular haplotype segment $h_1[8:10] = h_2[8:10]$ filling it. Haplotypes h_2 and h_3 display different haplotype segments on marker segment $[1:2]$, but these segments are drawn from the same haplotype block $\mathcal{B}_{[1:2]}$. Finally, haplotype h_2 has the marker segment $[11:12]$ internal to the marker segment $[11:13]$ of haplotype h_3.

Our dictionary model for haplotypes is inspired by an earlier dictionary models used to identify binding sites of regulatory proteins along DNA sequences (Bussemaker *et al.*, 2000; Sabatti and Lange, 2002). To the extent possible, we transfer the probability structure of these motif models to the haplotype model. Thus, in constructing a haplotype by concatenation from left to right, the added marker segments and haplotype segments are independently chosen according to specific probabilities. To achieve maximum model clarity, it is convenient to define marker segment probabilities, haplotype segment probabilities, and genotyping error probabilities. Error probabilities not only cover genotyping error per se but also mutation. Because of errors, we must allow observed haplotypes to differ from theoretical haplotypes constructed by concatenation.

In constructing a random haplotype H, suppose that concatenation has brought us to the point where the current marker segment ends with marker $i-1$. We then choose the next marker segment with conditional probability $q_{[i:j]}$. This mechanism forces the constraint $\sum_{j=i}^{m} q_{[i:j]} = 1$. Once we have chosen the next marker segment $[i:j]$, we choose the next haplotype segment s from the haplotype block $\mathcal{B}_{[i:j]}$ with probability r_s. Again we have the constraint $\sum_{s \in \mathcal{B}_{[i:j]}} r_s = 1$. A haplotype segment can be corrupted by either genotyping error or mutation. The simplest error model postulates a product multinomial distribution and a uniform error rate ϵ_k across the a_k alleles at marker k. If H is constructed using the segment $s \in \mathcal{B}_{[i:j]}$, then these assumptions yield the conditional probability

$$\Pr(H[i:j] = h[i:j] \mid s) \quad = \quad \prod_{k=i}^{j} \Pr(H[k:k] = h[k:k] \mid s),$$

where

$$\Pr(H[k:k] = h[k:k] \mid s) \;=\; \begin{cases} 1 & h[k:k] \text{ is missing} \\ 1-\epsilon_k & h[k:k] = s[k:k] \\ \frac{\epsilon_k}{a_k - 1} & h[k:k] \neq s[k:k]\;. \end{cases}$$

To express the probability $\Pr(H = h)$ succinctly, it is helpful to define

$$p(h[i:j]) \;=\; \sum_{s \in \mathcal{B}_{[i:j]}} r_s \Pr(H[i:j] = h[i:j] \mid s). \tag{1}$$

The quantity $p(h[i : j])$ is just the conditional probability of the observed haplotype segment $h[i:j]$ given H is constructed using marker segment $[i:j]$. With this notation, the joint probability of a partition π and an observed haplotype h can be written as

$$\Pr(H = h, \pi) \;=\; \prod_{[i:j] \in \pi} q_{[i:j]} p(h[i:j]).$$

Because block boundaries are unobserved, the full probability of h is

$$\Pr(H = h) \;=\; \sum_{\pi} \prod_{[i:j] \in \pi} q_{[i:j]} p(h[i:j]), \tag{2}$$

where the sum ranges over all possible partitions π.

The likelihood of a collection of independent haplotypes $h_1, \ldots, h_n$ factors into the likelihoods of the separate haplotypes. If the data consists of independent multilocus genotypes, then the likelihood factors over genotypes, but corresponding to each genotype there is an outer sum that increases the computational complexity of likelihood evaluation. For a multilocus genotype g, let

$$S_g \;=\; \left\{(h_1, h_2) : g = \frac{h_1}{h_2}\right\}$$

denote the set of maternal-paternal haplotype pairs compatible with g. If we assume that gametes combine at random, then the likelihood formula

$$\Pr(G = g) \;=\; \sum_{(h_1,h_2) \in S_g} \Pr(H_1 = h_1) \Pr(H_2 = h_2) \tag{3}$$

connects a random genotype G and its constituent haplotypes H_1 and H_2. With codominant markers and complete typing, consistent haplotype pairs differ only in phase. If parents are available, then most of the time one can infer phase. Appendix I points out that the probability that a child's phase is ambiguous at a codominant marker is at most $\frac{1}{8}$ when both parents are typed and at most $\frac{1}{4}$ when only one parent is typed. If p heterozygous markers are unphased for the multilocus genotype g, then the sum (3) will range over 2^p haplotype pairs.

The likelihood expression (2) is computationally impractical as it stands. Generalizations of Baum's forward and backwards algorithms from the theory of hidden Markov chains offer a better avenue to evaluation. Let A_i be the event that some marker segment of the random haplotype H ends with marker i. In the forward algorithm, we calculate the joint probability

$$f_i \quad = \quad \Pr(H[1:i] = h[1:i], A_i)$$

of A_i and the event that $H[1:i]$ coincides with the partially observed haplotype $h[1:i]$. The forward algorithm initializes $f_0 = 1$ and computes the remaining f_i from the recurrence

$$f_i \quad = \quad \sum_{k=1}^{\min\{d,i\}} f_{i-k} q_{[i-k+1:i]} p(h[i-k+1:i]) \tag{4}$$

based on equation (1) and an assumed maximum haplotype segment length d. The final probability f_m is the likelihood of the haplotype H. The backwards algorithm computes the conditional probability $b_i = \Pr(H[i:m] = h[i:m] | A_{i-1})$ of the event $H[i:m] = h[i:m]$ given the event A_{i-1}. We initialize $b_{m+1} = 1$ and update the b_i in the reverse order $i = m, \ldots, 1$ via

$$b_i \quad = \quad \sum_{k=1}^{\min\{d,m-i+1\}} q_{[i:i+k-1]} p(h[i:i+k-1]) b_{i+k}. \tag{5}$$

The final term b_1 of the recurrence gives the likelihood of H.

The conditional marker segment probabilities and haplotype segment probabilities do not fully convey how often particular marker segments or haplotype segments are used in haplotype construction. However, the correct probabilities can be computed by a variation of the forward algorithm. As before, let A_i be the event that a random haplotype has a marker segment ending with marker i. We can compute the probabilities $\Pr(A_i)$ via the recurrence

$$\Pr(A_i) \quad = \quad \sum_{k=1}^{\min\{d,i\}} \Pr(A_{i-k}) q_{[i-k+1:i]} \tag{6}$$

starting with the initial condition $\Pr(A_0) = 1$. This is just the forward algorithm with missing alleles at each marker. The probability that marker segment $[i : j]$ appears in a random haplotype is just $\Pr(A_{i-1})q_{[i:j]}$. We will refer to this quantity as the marker segment probability. No backward probability is required here because with probability 1 a partial haplotype ending at marker j is completed by concatenation. Similarly, the probability that a particular haplotype segment s in $\mathcal{B}_{[i:j]}$ appears in a random haplotype is $\Pr(A_{i-1})q_{[i:j]}r_s$. When we refer to the probability of s in the sequel, we will mean $\Pr(A_{i-1})q_{[i:j]}r_s$ rather than $q_{[i:j]}r_s$.

3. DICTIONARY RECONSTRUCTION

The dictionary model presents two fundamental challenges: (a) estimation of the parameter vectors q, r, and ϵ in a static dictionary with known haplotype segments; and (b) assembly of a dictionary of unknown haplotype segments. We take up parameter estimation first.

3.1 *Parameter estimation*

Every observed haplotype h involves missing information represented by the partition π and the uncorrupted haplotype segments that fill the marker segments of π. With obvious missing information of this sort, the EM algorithm is the natural method of parameter estimation (Dempster *et al.*, 1977). The complete data loglikelihood required by the EM algorithm can be written in terms of the following quantities:

(a) $Q_{[i:j]}$, the number of haplotypes using the marker segment $[i:j]$;

(b) R_s, the number of haplotypes using the haplotype segment s;

(c) T_i, the number of typing errors at marker i;

(d) $\theta = (q, r, \epsilon)$, the parameter vector.

Because each independent haplotype is constructed via a sequence of hidden multinomial trials, the complete data loglikelihood reduces to

$$\begin{aligned}\ln L_{\text{com}}(\theta) &= \sum_{i=1}^{m}\sum_{j=i}^{m} Q_{[i:j]} \ln q_{[i:j]} + \sum_{s} R_s \ln r_s + \sum_{i=1}^{m} T_i \ln \epsilon_i \\ &\quad + \sum_{i=1}^{m} (T_i^* - T_i) \ln(1-\epsilon_i)\end{aligned}$$

up to an irrelevant constant, where T_i^* is the number of typing events at locus i. When there is no missing data at marker i, clearly $T_i^* = n$. The E step of the EM algorithm calculates the conditional expectation of $\ln L_{\text{com}}(\theta)$ with respect to the observed data $\mathbf{h} = (h_1, \dots, h_n)$ and the current parameter vector θ^ℓ. For hidden multinomial trials, a simple counting argument correctly suggests that the M-step update for each parameter can be expressed as a ratio of an expected number of successes to an expected number of trials, where all expectations are conditional on $\mathbf{H} = \mathbf{h}$ and θ^ℓ (Lange, 2002). Hence, for $s \in \mathcal{B}_{[i:j]}$

$$
\begin{aligned}
q_{[i,j]}^{\ell+1} &= \frac{\mathrm{E}(Q_{[i:j]} \mid \mathbf{H} = \mathbf{h}, \theta^\ell)}{\sum_{k=i}^m \mathrm{E}(Q_{[i:k]} \mid \mathbf{H} = \mathbf{h}, \theta^\ell)} && (7)\\
r_s^{\ell+1} &= \frac{\mathrm{E}(R_s \mid \mathbf{H} = \mathbf{h}, \theta^\ell)}{\mathrm{E}(Q_{[i:j]} \mid \mathbf{H} = \mathbf{h}, \theta^\ell)} && (8)\\
\epsilon_i^{\ell+1} &= \frac{\mathrm{E}(T_i \mid \mathbf{H} = \mathbf{h}, \theta^\ell)}{T_i^*}. && (9)
\end{aligned}
$$

All of the conditional expectations appearing in these formulas are straightforward to calculate as explained in Appendix II. Since expectations are additive, it clearly suffices to consider the case $n = 1$ of a single haplotype. Using the results of the forward algorithm (4) and backward algorithm (5), we then have, for instance,

$$
\mathrm{E}(Q_{[i:j]} \mid H = h, \theta^\ell) = \frac{f_{i-1} q_{[i:j]} p(h[i:j]) b_{j+1}}{\Pr(H = h)}. \qquad (10)
$$

Similar expressions hold for the other conditional expectations figuring in the EM updates.

In the presence of a multilocus genotype G, we have to amend these formulas slightly. If X is one of the random variables featured in equations (7) through (9), then we can always decompose X as the sum $Y_1 + Y_2$, where Y_1 is the contribution coming from the maternal haplotype H_1 and Y_2 is the contribution coming from the paternal haplotype H_2. The decomposition

$$
\begin{aligned}
\mathrm{E}(X \mid G = g, \theta^\ell) &= \sum_{(h_1,h_2) \in S_g} \frac{\Pr(H_1 = h_1 \mid \theta^\ell) \Pr(H_2 = h_2 \theta^\ell)}{\Pr(G = g \mid \theta^\ell)} \\
&\quad \times \Big[\mathrm{E}(Y_1 \mid H_1 = h_1, \theta^\ell) + \mathrm{E}(Y_2 \mid H_2 = h_2, \theta^\ell) \Big]
\end{aligned}
$$

in conjunction with formulas such as (10) makes it possible to pass from haplotype data to multilocus genotype data in forming EM updates.

The EM algorithm also easily adapts to maximum a posteriori estimation. The simplest approach is to introduce independent Dirichlet priors for each of the hidden multinomial distributions.

These conjugate priors add pseudo-counts to their corresponding multinomial categories (Lange, 2002). For instance, in estimating the error probability ϵ_i at marker i, we add the log beta prior

$$\ln\Gamma(\mu_i+\nu_i)-\ln\Gamma(\mu_i)-\Gamma(\nu_i)+(\mu_i-1)\ln\epsilon_i+(\nu_i-1)\ln(1-\epsilon_i)$$

to the loglikelihood of the complete data. This sum survives intact through the E step of the EM algorithm. At the M step it produces the revised update

$$\epsilon_i^{\ell+1} \quad = \quad \frac{\mathrm{E}(T_i \mid \mathbf{H}=\mathbf{h},\theta^\ell)+\mu_i-1}{n+\mu_i+\nu_i-2}.$$

In other words, the counting interpretation of the EM prevails provided we add $\mu_i - 1$ imaginary typing errors and $\mu_i + \nu_i - 2$ imaginary trials to the hidden multinomial trials determining which allele is read at marker i. Similar considerations apply to maximum a posteriori estimation of the parameter vectors q and r.

3.2 *Dictionary selection*

Choosing the number and kind of haplotype segments to include in a haplotype dictionary is a typical model selection problem. Because fuller dictionaries always provide more flexible descriptions of data and consequently higher likelihoods, one cannot rely simply on the comparison of models through their maximum likelihoods. Common sense suggests that statistical inference should be guided by considerations of parsimony as well. Thus, it is customary to minimize an objective function that incorporates both a model's negative loglikelihood and a penalty for model complexity. The two-stage version of the minimum description length (MDL) approach to inference provides a rationale for this procedure (Rissanen, 1978, 1983; Hansen and Yu, 2001). In fact, MDL criteria have been previously suggested for the selection of block boundaries in the analysis of haplotype data (Anderson and Novembre, 2003; Koivisto *et al.*, 2003; Sheffi, 2004).

In the MDL framework, one selects the probability model that requires the minimum number of bits to describe (or transmit) both the model and the data. The number of bits necessary to transmit a model reflects the complexity of the model. The number of bits necessary to transmit the data given a model depends on how well the model fits the data. Standard information theory arguments show that the shortest code capable of transmitting a random message has length equal to the negative of the logarithm base 2 of the probability of the message. This means that we can measure the number of bits needed to transmit the data given the model by the minimum of the negative loglikelihood of the data. This yields the loglikelihood term in the MDL objective function.

We must also evaluate the description length of the haplotype dictionary $\mathcal{D}$ behind the model. Each haplotype block $\mathcal{B}_{[i:j]}$ of $\mathcal{D}$ involves $|\mathcal{B}_{[i:j]}| - 1$ independent parameters from the r vector and one parameter from the q vector. The q parameters involve m constraints, one for each marker. The loss of these m parameters is compensated by the m error parameters. It follows that the number of independent parameters equals the size

$$|\mathcal{D}| \;=\; \sum_{i=1}^{m}\sum_{j=i}^{m} |\mathcal{B}_{[i:j]}|$$

of $\mathcal{D}$. For n haplotypes, all of the parameters are real numbers measured with finite precision proportional to $1/\sqrt{n}$, so at most $\frac{1}{2}\log_2 n$ bits are required to encode each parameter. This takes care of the parameters. To transmit a haplotype segment s, we must transmit the allele at each of its participating markers. Because marker k has a_k alleles, it takes $\log_2 a_k$ bits to transmit one of its alleles. Although we also need to transmit the beginning and ending markers of each haplotype segment s, we will omit these relatively minor contributions to model complexity. Summarizing both data and model contributions to the MDL, our objective function for comparing dictionaries is

$$\text{Obj}(\mathcal{D}) \;=\; -\frac{1}{\ln 2}\ln L(\hat{\theta} \mid \mathcal{D}) + \frac{1}{2}|\mathcal{D}|\log_2 n + \sum_{i=1}^{m}\sum_{j=i}^{m} |\mathcal{B}_{[i:j]}| \sum_{k=i}^{j} \log_2 a_k. \qquad (11)$$

Here the conversion factor $1/\ln 2$ turns natural logarithms into logarithms base 2.

With this objective function in hand, we still need a strategy for exploring model space. Our prefered strategy combines heuristic construction of an initial dictionary with subsequent alternation of growing and pruning steps. For computational convenience, we retain all single-marker haplotype segments throughout dictionary construction. At marker k, there are a_k such trivial haplotype segments. This constraint ensures the compatibility of the dictionary with every conceivable haplotype even in the absence of genotyping error.

Because of the work involved in finding maximum likelihood estimates, it is unrealistic, on the one hand, to start a search from an exhaustive dictionary containing all possible haplotype segments of length d or less. On the other hand, starting the search from the smallest dictionary with just the haplotype segments of length 1 tends to miss large conserved haplotype segments. It is better to make an educated guess of a fairly large initial dictionary that is not heavily redundant. One way of assembling an initial dictionary is to choose two positive constants α and β and scan the data for common haplotype segments s (Patil *et al.*, 2001; Johnson *et al.*, 2001; Zhang *et al.*, 2002b). When the empiric proportion of a segment s exceeds β and the empiric proportion of its associated haplotype block $\mathcal{B}_{[i:j]}$ exceeds α, s is included in the initial dictionary. If potential haplotype segments are visited in the order of their length and α and β are large, say .8 and .2, then this criterion tends to favor inclusion of just the short haplotype segments. On the other hand, taking α and β small, say .4 and .05, may produce a dictionary with tens of thousands of entries. Such massive dictionaries contain almost every possible block and ensure that important haplotype segments are not missed.

It is useful to prune a massive initial dictionary first by heuristic methods that avoid likelihood evaluation. Without committing ourselves to nonoverlapping blocks, we attempt to locate hard block boundaries in two passes through the markers. The forward pass generates a subdictionary with nonoverlapping blocks; the reverse pass does likewise. However, taking the union of these two subdictionaries invariably yields a pruned dictionary with overlapping blocks. The forward pass starts at marker 1 and progresses to marker m. Once we decide that a block ends with marker $i-1$, we seek the marker segment $[i:j]$ defining the next block. The decision to end the block with marker j depends on the number of haplotype segments n_{ij} in the dictionary that span the marker segment $[i:j]$. At a true block boundary j, the difference $d_{ij} = n_{i,j} - n_{i,j+1}$ should be positive. If this condition holds and $j > i+2$, then we declare j to be the end of the current block. To discourage the creation of short blocks ending with $j = i, i+1, i+2$, and $i+3$ we use the stronger stopping rule $d_{ij} > 1$. When we find the block boundary j, we drop all haplotype segments that start at i and end before j, except, of course, the trivial ones. The reverse pass operates similarly but starts at marker m and progresses to marker 1.

Once the initial dictionary is pruned by this heuristic method, we alternate more targeted growing and pruning steps until the current dictionary stabilizes. Growing may well restore some of the haplotype segments deleted in the initial pruning. To grow the dictionary, we must identify candidate haplotype segments to add. The most fruitful approach is to concatenate two adjacent haplotype segments, s and t, already in the dictionary. Because trivial haplotype segments are always retained, it is possible to grow haplotype segments laboriously one marker at a time. To decide whether to add the concatenated haplotype segment st to the current dictionary, we use the expected number $\mathrm{E}(R_{st} \mid \mathbf{H} = \mathbf{h}, \hat{\theta})$ of times st appears in the haplotype data $\mathbf{h}$. This criterion is given by

$$\mathrm{E}(R_{st} \mid \mathbf{H} = \mathbf{h}, \hat{\theta}) \quad = \quad \sum_{h=h_1}^{h_n} \frac{f_{i-1} q_{[i:k]} p(h_{[i:k]}) q_{[k+1:j]} p(h_{[k+1:j]}) b_{j+1}}{\Pr(H = h)},$$

where s spans the marker segment $[i:k]$ and t spans the marker segment $[k+1:j]$. It is instructive to compare this conditional expectation to the theoretical probability

$$c_{st} \quad = \quad \Pr(A_{i-1}) q_{[i:k]} r_s q_{[k+1:j]} r_t$$

that the two haplotype segments s and t co-occur in a random haplotype. Since the actual number of times the segment st appears follows a binomial distribution, we define the score

$$\tau_{st} \quad = \quad \frac{\mathrm{E}(R_{st} \mid \mathbf{H} = \mathbf{h}, \hat{\theta}) - nc_{st}}{\sqrt{nc_{st}(1 - c_{st})}}$$

and add the segment st to the dictionary whenever τ_{st} falls above a user designated cut-off.

To prune the dictionary, we use a combination of global and local tactics. After each round of growing, we first attempt to prune haplotype segments en masse by pruning blocks. The current blocks are ordered by their corresponding conditional expectations $\mathrm{E}(Q_{[i:j]} \mid \mathbf{H} = \mathbf{h}, \theta)$ evaluated at the maximum likelihood estimate $\theta = \hat{\theta}$. An exception is made for the trivial blocks, which are forced to come first. The resulting order defines a sequence of nested subdictionaries, and bisection is applied to find the subdictionary with the lowest value of the objective function (11). Assuming that the objective function first declines and then rises as we progress from the smallest to the largest subdictionary, bisection makes it possible to prune several haplotype blocks simultaneously. We also apply bisection to drop haplotype segments in chunks rather than haplotype blocks in chunks. The nontrivial haplotype segments s are first ordered by their conditional expectations $\mathrm{E}(R_s \mid \mathbf{H} = \mathbf{h}, \hat{\theta})$.

Local pruning is achieved by dropping individual haplotype segments rather than entire haplotype blocks. We first prune all haplotype segments s with low conditional probabilities r_s, say less than 10^{-3}. We then consider haplotype segments u nested within larger segments v. For example, if the concatenated segment st is added during the previous round of growing, then s and t will be nested within st unless they have already been pruned. Let $\mathcal{C}_u$ denote the collection of segments v strictly containing u. If n_u counts the number of observed haplotypes that agree with u on the marker segment defining u and m_u counts the number of observed haplotypes that agree with some v in $\mathcal{C}_u$ on the marker segment defining v, then comparison of n_u and m_u conveys how redundant u is. Hence, we drop u from the current dictionary if the ratio n_u/m_u falls below a user defined cutoff γ. For nested pairs $u \subset v$ of comparable length, it may be better to drop v rather than u. In such cases we use the values Obj$(\mathcal{D} \cup \{u\} \cup \{v\})$, Obj$(\mathcal{D} \cup \{u\})$, and Obj$(\mathcal{D} \cup \{v\})$ of the objective function to reach a decision, where $\mathcal{D}$ is the current dictionary omitting u and v. Alternation of growing and pruning steps stops when there are no further promising haplotype segments to add.

The search strategy just described is clearly heuristic. It is based on experience and tinkering, and by no means is it guaranteed to converge to the optimal dictionary. However in practice, it successfully identifies parsimonious dictionaries that offer accurate descriptions of the data.

3.3 *Software implementation*

Our computer program HAPPY implements the EM algorithm and the dictionary selection procedure. The user designates values for the parameters α, β, the threshold τ for adding haplotype segments, and the threshold γ for dropping nested haplotype segments. One may construct a better dictionary by playing with these parameters. HAPPY outputs the final marker segments $[i : j]$, their probabilities $\Pr(A_i)q_{[i:j]}$, the alleles of each haplotype segment s in the corresponding block $\mathcal{B}_{[i:j]}$, and the conditional probability r_s of s given the block $\mathcal{B}_{[i:j]}$. To provide a snapshot of the linkage disequilibrium patterns in the data as explained by the model, HAPPY computes the co-occurrence probability m_{ij} that markers i and j are found on the same haplotype segment. These probabilities are output as a matrix $M = (m_{ij})$ ready for 2-dimensional display. Last of all, when the data consists of multilocus genotypes, HAPPY provides for each observed genotype the most likely pair of haplotypes and its probability. The program runs fairly fast on phased data, constructing a dictionary in approximately 5 minutes for 20 markers and 200 genotypes and in 1 hour for 100 markers and 100 genotypes. For fully unphased data, the program has to sum over every possible phase in likelihood evaluation. A data set of 100 genotypes and 20 markers accordingly takes several days to construct a dictionary.

4. RESULTS

To illustrate the main features of our model and analysis strategy, we start with a simple simulated example. As depicted in Figure II, the example involves 22 biallelic markers and a dictionary of 10 nontrivial haplotype segments grouped in 5 partially overlapping blocks. Each horizontal stripe in the figure corresponds to a haplotype segment in the dictionary. All markers are biallelic, and the colors red and blue represent the two alleles. A haplotype segment's color intensity is proportional to the probability of the segment.

Note that the nontrivial haplotype segments either always span or always exclude the markers segments [1:5], [5:10], and [13:22]; no partial overlaps occur. Also no haplotype segment spans the marker segment [9:13]. Given the specified dictionary, it is possible to calculate the co-transmission probability for the marker pair i, j; this is simply the probability that in a random haplotype, marker i and marker j belong to the same haplotype segment. The matrix of co-transmission probabilities for the example dictionary is given in Figure IIIa.

We generated 200 phased haplotypes with a genotyping error rate of 0 and no missing data and were able to reconstruct the dictionary perfectly. We then used a genotyping error rate of 5% and 5% missing data under the model. Figure IIIb shows the linkage disequilibrium structure of the generated data, using the linkage disequilibrium measure D' between marker pairs. Despite the relatively large sample, there is a considerable noise in these rough measures. However, they clearly demonstrate that the first 10 markers are in approximate linkage equilibrium with the last 10 markers and suggest a recombination hot spot in the vicinity of markers 11 and 12. When we analyze the data under the dictionary model, we obtain the empirical co-transmission probabilities described in Figure IIIc. The smooth structure of this matrix is a consequence of our explicit modeling of linkage disequilibrium caused by conserved haplotype segments. The block structure is well detected.

A co-occurrence probability matrix not only provides a fast visual display of the linkage disequilibrium implications of the dictionary model, it also offers a useful device for comparing different data sets and different algorithms for estimation and selection. To measure the distance between two co-occurrence matrices, corresponding to analyses of the same dataset or analyses of different datasets under the same model, one can compute their correlations as random vectors whose entries are read from left to right and top to bottom. This is analogous to conventional measures of agreement between partitions or clusters, the difference being that the partitions described by co-transmission probabilities are fuzzy rather than sharp (Fowlkes and Mallows, 1983).

To test the performance of our model in a real setting, we reanalyzed a data set that lead to one of the first reports of block structure in SNP haplotypes (Daly *et al.*, 2001). Our analysis focused on 129 phased genotypes spanning 103 markers, all single nucleotide polymorphisms. Figure IV illustrates the extent of pairwise disequilibrium in this dataset using the absolute value of D'. Figure V presents the haplotype dictionary we reconstruct (top) and the one suggested by the original investigators (Daly *et al.*, 2001) (bottom), using the conventions adopted earlier in our description of the simulated dictionary. We should clarify that in the Daly dictionary we used the haplotypes segments identified by Daly *et al.* (2001) and re-estimated their probabilities via the EM algorithm. Our treatment incidentally loses any information on the dependencies across blocks that the authors describe with a hidden Markov model. Figure VI further compares the Daly dictionary and ours in terms of co-transmission probabilities. In our reconstruction, the long range dependencies are captured via longer haplotype segments. This is particularly clear for the markers in the Daly dictionary described by the first three blocks, as it can be seen from the upper left corner of the co-transmission probability matrix in Figure VI. Note that the overlapping haplotype segments identified by our model suggest a possible phylogeny; with long segments ancestral, and shorter segments the result of subsequent recombination. In particular, the haplotype segment labeled (a) in Figure V can be interpreted as ancestral to the other haplotype segments indicated with an asterisk (*). Furthermore, the haplotype segment labeled (b) in our reconstructed dictionary substitutes for haplotype segment (1) of the Daly dictionary. The MDL criterion strongly favors our dictionary with overlapping haplotypes to the Daly dictionary, the difference in the objective function (11) being 1,163. This suggests that a dictionary with overlapping blocks offers a substantially better description of the data than one with non-overlapping blocks.

5. DISCUSSION

A few years ago geneticists noticed block-like patterns of linkage disequilibrium in several sequences of closely spaced SNPs (Daly *et al.*, 2001; Patil *et al.*, 2001; Reich *et al.*, 2001; Gabriel *et al.*, 2002). This observation spurred development of novel statistical methods and massive data gathering, eventually prompting founding of the International Hapmap Project (2003). The extensive literature on haplotype blocks defies brief summary. Scientists have approached the subject of haplotype geography with different scientific objectives, leading to varied definition of blocks, contrasting interpretations of their nature, and different strategies to identify them (Schwartz *et al.*, 2003). The definitions of a block include: (a) a contiguous set of markers for which the average value of the disequilibrium index D' exceeds some predetermined threshold (Daly *et al.*, 2001; Reich *et al.*, 2001; Gabriel *et al.*, 2002), (b) a contiguous set of markers which plausibly show no ancestral recombination (Bafna *et al.*, 2003), and (c) a contiguous set of markers that can be represented parsimoniously using a limited number of SNPs (Patil *et al.*, 2001). The three adjectives disequilibrium, recombination, and diversity succinctly summarize these distinctions.

Each definition has inspired the development of interesting algorithms for identifying blocks. For instance, the diversity perspective has motivated dynamic programming algorithms (Zhang *et al.*, 2003) and led to application of minimum description length criteria (Anderson and Novembre, 2003; Koivisto *et al.*, 2003). Focusing on the detectability of ancestral recombination has motivated interested developments in computer science (Bafna *et al.*, 2003; Eskin *et al.*, 2003). The different definitions also reflect different scientific constituencies. For instance, some scientists are primarily interested in the biological forces responsible for block formation. These forces include recombination hot spots, population history, and genetic drift. (Wang *et al.*, 2002; Kauppi *et al.*, 2003). Other scientist have emphasized the fact that the low haplotype diversity minimizes the number of markers needed for association mapping of common diseases (Akey *et al.*, 2001; Cardon and Abecasis, 2003; Johnson *et al.*, 2001; Zhang *et al.*, 2002a, 2004). Thus, the utility and validity of the haplotype block model depends to some extent on the background of the individual commentator (Schwartz *et al.*, 2003; Wall and Pritchard, 2003).

Our foray into this research area is motivated by dissatisfaction with the current models. We prefer to construct a model that parsimoniously represents the data without imposing the rigid notion of disjoint blocks. Although one can always argue the merits of mechanistic versus phenomenological models, the purely phenomenological nature of the dictionary model avoids the difficult problems of quantifying population evolution and recombination hot spots that afflict mechanistic models. Marrying the dictionary model with the MDL criterion promotes parsimony.

The notion of overlapping blocks naturally arises as a generalization of segments identical by descent (Chapman and Thompson, 2002). Overlapping haplotypes can be identified by a number of heuristic methods (Cardon and Abecasis, 2003). The extend of block overlap is intimately tied to linkage disequilibrium in the genome (Wall and Pritchard, 2003). Only by including the possibility of block overlap can researchers test the hypothesis that non-overlapping blocks provide an adequate description of linkage disequilibrium in the human genome. This is one of the strengths of the dictionary model. We are aware that a group of researchers from MIT has also been working on another model that allows overlapping ancestral segments (Sheffi, 2004), and we thank them for fruitful exchanges. Our applications of the MDL criterion in model selection parallels the efforts of Koivisto, Anderson, and Sheffi (Koivisto *et al.*, 2003; Anderson and Novembre, 2003; Sheffi, 2004) who resort to other information theory criteria in selecting and testing block boundaries.

One of the attractive features of the dictionary model is that it handles unphased data. A few other recent papers also discuss the reconstruction of haplotype blocks from genotypes (Schwartz *et al.*, 2002; Zhang *et al.*, 2004; Halperin and Eskin, 2004; Greenspan and Geiger, 2004). Our implementation of the dictionary model is limited to about 16 markers on completely unphased data. In the absence of phase information, likelihood evaluation is hindered by the need to sum over all possible phases. Fortunately, the EM algorithm scales no worst than likelihood evaluation. Excoffier and Slatkin (Excoffier and Slatkin, 1995) have noted similar computational barriers. The phase problem and the presence of overlapping blocks make exploration of dictionary space computationally intensive despite the remarkable efficiency of likelihood evaluation via the forward and backward algorithms and search via the EM algorithm. As a consequence, the dictionary model is best suited for analyzing in detail a specific chromosome region rather than tiling an entire genome. This should not be a serious limitation because most geneticists are in interested in narrow regions of linkage disequilibrium.

REFERENCES

Akey, J., Jin, L., and Xiong, M. 2001. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur. J. Human Genet.* 9, 291–300.

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

Bafna, V., Gusfield, D., Lancia, G., and Yooseph, S. 2003. Haplotyping as perfect phylogeny: a direct approach. *J. Comp. Biol.* 10, 323–340.

Bussemaker, H. J., Li, H., and Siggia, E. D. 2000. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* 97, 10096–10100.

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

Chapman, N. H. and Thompson, E. A. 2002. The effect of population history on the lengths of ancestral chromosome segments. *Genetics* 162, 449–458.

Cullen, M., Perfetto, S. P., Klitz, W., Nelson, G., and Carrington, M. 2002. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am. J. Human Genet.* 71, 759–776.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximumlikelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* 39, 1–38.

Eskin, E., Halperin, E., and Karp, R. 2003. Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology.* 1, 1–20.

Excoffier, L. and Slatkin, M. 1995. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12(5), 921–7.

Fowlkes, E. and Mallows, C. 1983. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* 78, 553–569.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. 2002. The structure of haplotype blocks in the human genome. *Science* 296, 2225–9.

Greenspan, G. and Geiger, D. 2004. Model-based inference of haplotype block variation. *J. Comp. Biol.* 11, 495–506.

Halperin, E. and Eskin, E. 2004. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics* 20, 1842–9.

Hansen, M. and Yu, B. 2001. Model selection and the principle of minimum description length. *J. Am. Stat. Assoc.* 96, 746–74.

Jeffreys, A. J., Kauppi, L., and Neumann, R. 2001. Intensly punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* 29, 217–221.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Genova, G. D., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughs, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., and Todd, J. A. 2001. Haplotype tagging for the identification of common disease genes. *Nature Genet.* 29, 233–237.

Kauppi, L., Sajantila, A., and Jeffreys, A. J. 2003. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Human Mol. Genet.* 12, 33–40.

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. *Proc. 8th Pac. Symp. Biocomputing (PSB '03)* 12, 502–13.

Lange, K. 2002. *Mathematical and Statistical Methods For Genetic Analysis*. Spring-Verlag, New York.

Lange, K. 2004. *Optimization*. Spring-Verlag, New York.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A., and Cox, D. R. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human Chromosome 21. *Science* 294, 1719–1724.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. 2001. Linkage disequilibrium in the human genome. *Nature* 411, 199–204.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14, 465–71.

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

Sabatti, C. and Lange, K. 2002. Genomewise motif identification using a dictionary model. *IEEE Proceedings* 90, 1803–1810.

Schwartz, R., Clark, A. G., and Istrail, S. 2002. Methods for inferring block-wise ancestral history from haploid sequences. *Algorithms in Bioinformatics. Lecture Notes in Computer Science* 2452, 44–59, Springer-Verlag, Berlin.

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

Sheffi, J. 2004. An HMM-based boundary-flexible model of human haplotype variation. *Master's thesis, MIT, Departments of Electrical Engineering and Computer Science* .

The International Hapmap Consortium. 2003. The International HapMap Project. *Nature* 426, 789–796.

Wall, J. D. and Pritchard, J. 2003. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Human Genet.* 73, 502–515.

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

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

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F. 2002b. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. USA* 99, 7335–9.

Zhang, K., Qin, Z. S., Liu, J. S., Chen, T., Waterman, M. S., and Sun, F. 2004. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res.* 14, 908–916.

Zhang, K., Sun, F., Waterman, M. S., and Chen, T. 2003. Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am. J. Human Genet.* 73, 63–73.

ACKNOWLEDGMENTS

This research supported in part by NIH training grant HG02536(KLA), NSF grant DMS0239427(CS) and USPHS grants GM53275(CS,KL) and MH59490(KL).

APPENDIX I: MULTILOCUS GENOTYPES

Consider a marker locus with n codominant alleles labeled $1, \ldots, n$. Most genotypic combinations of a mother–father–child trio make it possible to infer which allele the mother contributes to the child and which allele the father contributes to the child. It turns out that the only ambiguous case occurs when all three members of the trio share the same heterozygous genotype i/j. The probability of such a configuration is $(2p_ip_j)^2\frac{1}{2}$, where p_i and p_j are the population frequencies (proportions) of alleles i and j. This formula follows from: (a) the independence of the parents' genotypes, (b) Hardy-Weinberg equilibrium, and (c) a probability of $\frac{1}{2}$ that one of them transmits to the child an i allele and the other transmits a j allele. Using these configuration formulas, we can express the probability that the trio's genotypes do not permit inference of the origin of both of the child's alleles as the sum

$$f(p) \quad = \quad \sum_{i<j} (2p_ip_j)^2\frac{1}{2}.$$

Lange (2004) shows that $f(p)$ attains it maximum of $\frac{1}{8}$ when two of the frequencies p_i equal $\frac{1}{2}$ and the remaining frequencies equal 0.

Rather than repeat that proof here, we would like to consider the slightly different situation where the genotype of one of the parents, say the father, is unknown. What is the probability that we cannot determine the parental contributions to the child given only the mother's genotype and the child's genotype? Again the only ambiguous case occurs when the mother and the child have the same heterozygous genotypes i/j. It is clear that a mother with heterozygous genotype i/j is equally likely to pass either allele i or allele j. If she passes allele i, then the father must pass allele j. Under the random union of gametes model that we have implicitly adopted, the father passes allele j with probability p_j. Hence,

$$\Pr(\text{child is } i/j|\text{mother is } i/j) \;=\; \frac{1}{2}p_i + \frac{1}{2}p_j,$$

and the probability that the mother and child have the same heterozygous genotype reduces to the sum

$$\begin{aligned} g(p) &= \sum_{i<j} 2p_ip_j\Big(\frac{1}{2}p_i + \frac{1}{2}p_j\Big) \\ &= \sum_{i<j}[p_i^2p_j + p_ip_j^2] \\ &= \sum_i\sum_{j\neq i} p_i^2p_j \\ &= \sum_i p_i^2(1-p_i). \end{aligned}$$

For $n = 2$, the function $g(p) = p_1^2(1-p_1) + (1-p_1)^2p_1 = p_1(1-p_1)$ attains its maximum of $\frac{1}{4}$ when $p_1 = p_2 = \frac{1}{2}$. For general n, this suggests that the maximum occurs when all $p_i = \frac{1}{n}$. However, at this point $g(p)$ equals $\frac{n-1}{n^2}$, which is strictly less than $\frac{1}{4}$ for $n \geq 3$. This failure leads us to guess that the maximum of $\frac{1}{4}$ occurs on a boundary where all but two of the $p_i = 0$. By symmetry we can assume that the sequence p_i is increasing in i. Under this assumption, we now demonstrate that we can increase $g(p)$ by incrementing p_2 by $q \in [0, p_1]$ at the expense of decrementing p_1 by q. If we define the perturbed value of $g(p)$ by

$$h(q) \;=\; (p_1-q)^2(1-p_1+q) + (p_2+q)^2(1-p_2-q) + \sum_{j=3}^{n} p_j^2(1-p_j),$$

then it is evident that its derivative

$$\begin{aligned} & h'(q) \\ = \;& (p_1-q)^2 - 2(p_1-q)(1-p_1+q) - (p_2+q)^2 + 2(p_2+q)(1-p_2-q) \\ = \;& (p_2-p_1+2q)[2-3(p_1+p_2)]. \end{aligned}$$

For $n \geq 3$, our order assumption $0 < q \leq p_1 \ldots \leq p_n$ implies that $p_1 + p_2 \leq \frac{2}{3}$, so both of the factors defining $h'(q)$ are nonnegative. It follows that we can reduce p_1 to 0 and increase p_2 to

$p_1 + p_2$ without decreasing $g(p)$. In general, we should keep discarding the lowest positive p_j until only p_{n-1} and p_n survive. At this juncture we maximize $g(p)$ by taking $p_{n-1} = p_n = \frac{1}{2}$.

The computational complexity of likelihood evaluation under the dictionary model is dominated by the multilocus genotype with the maximum number of possible phases. Hence, it is instructive to examine the worst case of m SNPs, each with two equally frequent alleles. Given fully typed parents, the number of phase ambiguities X_i for the ith multilocus genotype is binomially distributed with m trials and success probability $\frac{1}{8}$. Assuming that t independent multilocus genotypes are observed, the maximum $Z = \max\{X_1, \ldots, X_t\}$ has distribution function

$$\Pr(Z \le z) \quad = \quad \Pr(X_1 \le z)^t.$$

If m is reasonable large, then X_1 is approximately normal with mean $\frac{m}{8}$ and variance $\frac{7m}{64}$ and

$$\Pr(X_1 \le z) \quad \approx \quad \Phi\left(\frac{z - m/8}{\sqrt{7m/64}}\right),$$

where $\Phi(z)$ is the standard normal distribution function. The median value of Z therefore approximately satisfies the identity

$$\frac{1}{2} \quad = \quad \Phi\left(\frac{z_{\text{median}} - m/8}{\sqrt{7m/64}}\right)^t,$$

whose solution is

$$z_{\text{median}} \quad = \quad \frac{m}{8} + \sqrt{\frac{7m}{64}}\Phi^{[-1]}\left(2^{-1/t}\right),$$

where $\Phi^{[-1]}(z)$ is the functional inverse of $\Phi(z)$. Figure VII depicts the level curves of z_{median} as a function of m and t. Computation tends to bog down when the maximum number of phase uncertainties exceeds 12. This optimistic conclusion must be tempered by the fact that most data sets contain a high percentage of partially typed markers.

APPENDIX II: DERIVATION OF EM UPDATES

Because the complete data likelihood $L_{com}(\theta)$ is a product over the observations, we examine the likelihood of a single observed genotype g. The complete data reveal the haplotype pair (h_1, h_2) underlying g and the partitions π^1 and π^2 segmenting h_1 and h_2. The complete data likelihood can be written in terms of the following indicator random variables:

$$
\begin{aligned}
W_{(h_1,h_2)} &= \begin{cases} 1 & \text{if observation } g \text{ has phased genotype } (h_1, h_2) \\ 0 & \text{otherwise,} \end{cases} \\
V^c_{[i:j]} &= \begin{cases} 1 & \text{if marker segment } [i:j] \text{ occurs in haplotype } h_c \\ 0 & \text{otherwise,} \end{cases} \\
F^c_s &= \begin{cases} 1 & \text{if haplotype segment } s \text{ occurs in haplotype } h_c \\ 0 & \text{otherwise,} \end{cases} \\
U^c_{s,k} &= \begin{cases} 1 & \text{if } h_c[k:k] = s[k:k] \\ 0 & \text{otherwise,} \end{cases} \\
Z^c_{s,k} &= \begin{cases} 1 & \text{if } h_c[k:k] \neq s[k:k], \\ 0 & \text{otherwise .} \end{cases}
\end{aligned}
$$

If locus k is untyped, then we put $U^c_{s,k} = Z^c_{s,k} = 0$. For a particular observation g, we can write the complete data likelihood as

$$
\begin{aligned}
L_g(\theta) &= \prod_{(h_1,h_2)} \Big(\prod_{c=1}^{2} \prod_{\pi^c} \prod_{[i:j]\in\pi^c} \Big\{ q_{[i:j]} \prod_{s\in\mathcal{B}_{[i:j]}} \\
&\quad \Big[r_s \prod_{k=i}^{j} (1-\epsilon_k)^{U^c_{s,k}} \Big(\frac{\epsilon_k}{a_k - 1} \Big)^{Z^c_{s,k}} \Big]^{F^c_s} \Big\}^{V^c_{[i:j]}} \Big)^{W_{(h_1,h_2)}} .
\end{aligned}
$$

Taking logarithms yields the complete data loglikelihood

$$
L_g(\theta) = \sum_{(h_1,h_2)} W_{(h_1,h_2)} \sum_{c=1}^{2} \sum_{\pi^c} \sum_{[i:j]\in\pi^c} V^c_{[i:j]} \Big\{ \ln q_{[i:j]} + \sum_{s\in\mathcal{B}_{[i:j]}} F^c_s
$$

$$\times \Big[\ln r_s + \sum_{k=i}^{j} U_{s,k}^c \ln(1-\epsilon_k) + Z_{s,k}^c \Big(\frac{\epsilon_k}{a_k - 1}\Big)\Big]\Big\}.$$

The E step of the EM algorithm is carried to completion by noting that the product of two indicators is an indicator and the conditional expectation of an indicator is a conditional probability.

The surrogate function produced by the E step separates the parameters and allows us to deal with them in batches. Each batch gives rise to a multinomial or binomial loglikelihood of the form $\sum_i u_i \ln b_i$, where u_i and b_i are the fractional count and probability assigned to category i. It is well known that such a sum is maximized subject to $b_i \geq 0$ and $\sum_i b_i = 1$ by taking

$$\hat{b}_i = \frac{u_i}{\sum_j u_j}.$$

This specifies the M step of the EM algorithm in broad outline. In the specific cases of the updates (8), (9), and (9), the reader can check the identities

$$\mathrm{E}(Q_{[i:j]} \mid G = g, \theta^\ell) = \sum_{(h_1,h_2)} \sum_{c=1}^{2} \sum_{\pi^c} \mathrm{E}(W_{(h_1,h_2)} V_{[i:j]}^c \mid G = g, \theta^\ell) \quad (12)$$

$$\mathrm{E}(R_s \mid G = g, \theta^\ell) = \sum_{(h_1,h_2)} \sum_{c=1}^{2} \sum_{\pi^c} \mathrm{E}(W_{(h_1,h_2)} V_{[i:j]}^c F_s^c \mid G = g, \theta^\ell) \quad (13)$$

$$\mathrm{E}(T_k \mid G = g, \theta^\ell) = \sum_{(h_1,h_2)} \sum_{c=1}^{2} \sum_{\pi^c} \sum_{[i:j]\in\pi^c} \sum_{s\in\mathcal{B}_{[i:j]}} \mathrm{E}(W_{(h_1,h_2)} V_{[i:j]}^c F_s^c Z_{s,k}^c \mid G = g, \theta^\ell) \quad (14)$$

connecting the conditional expected counts to the conditional probabilities for a single genotype g. It is worth stressing that the conditional probabilities shown above are efficiently computed by the forward-backward sandwich equations featured in the text. All three formulas (12), (13), and (14) must be summed over the t possible genotypes.

FIGURES

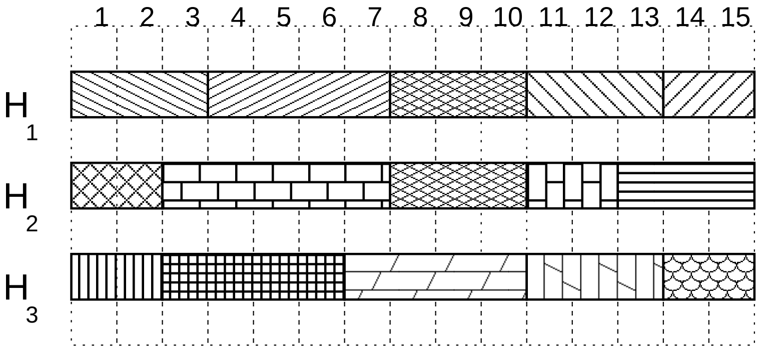


Figure I:

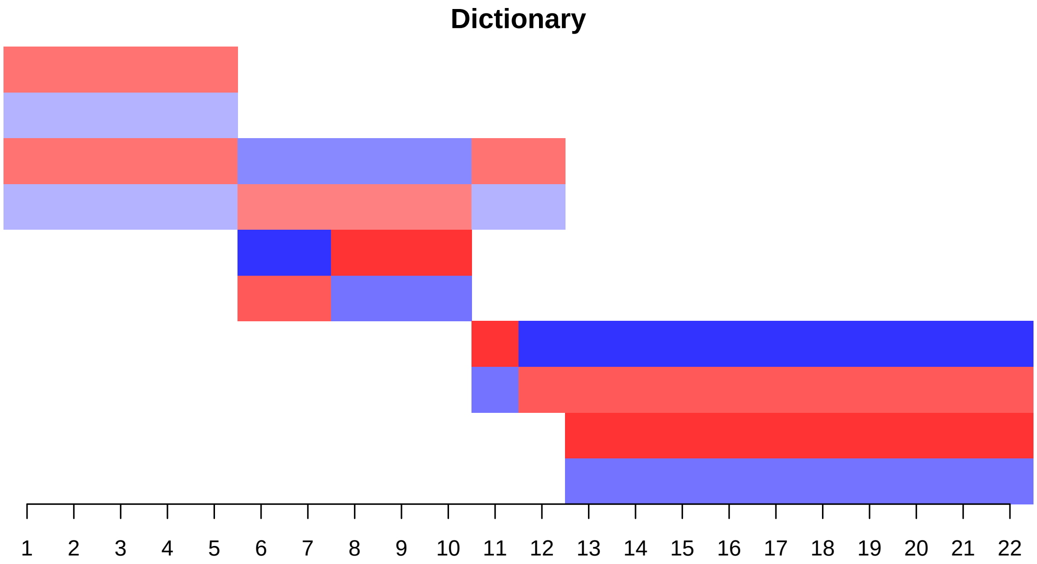
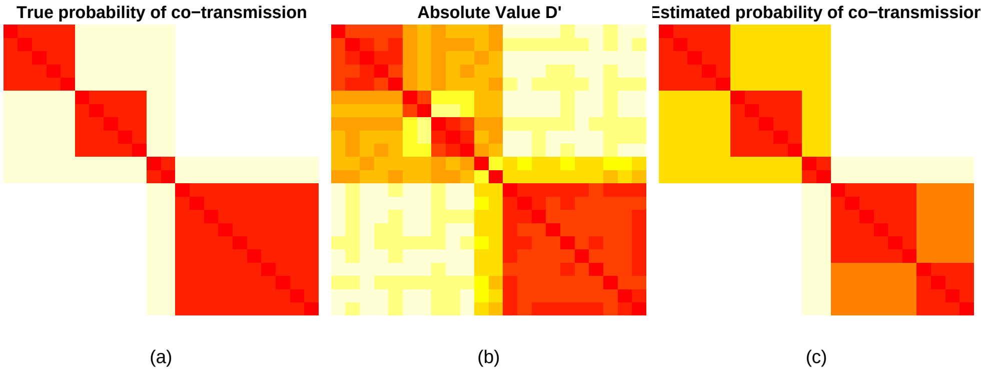


Figure II:



(a)

(b)

(c)

Figure III:

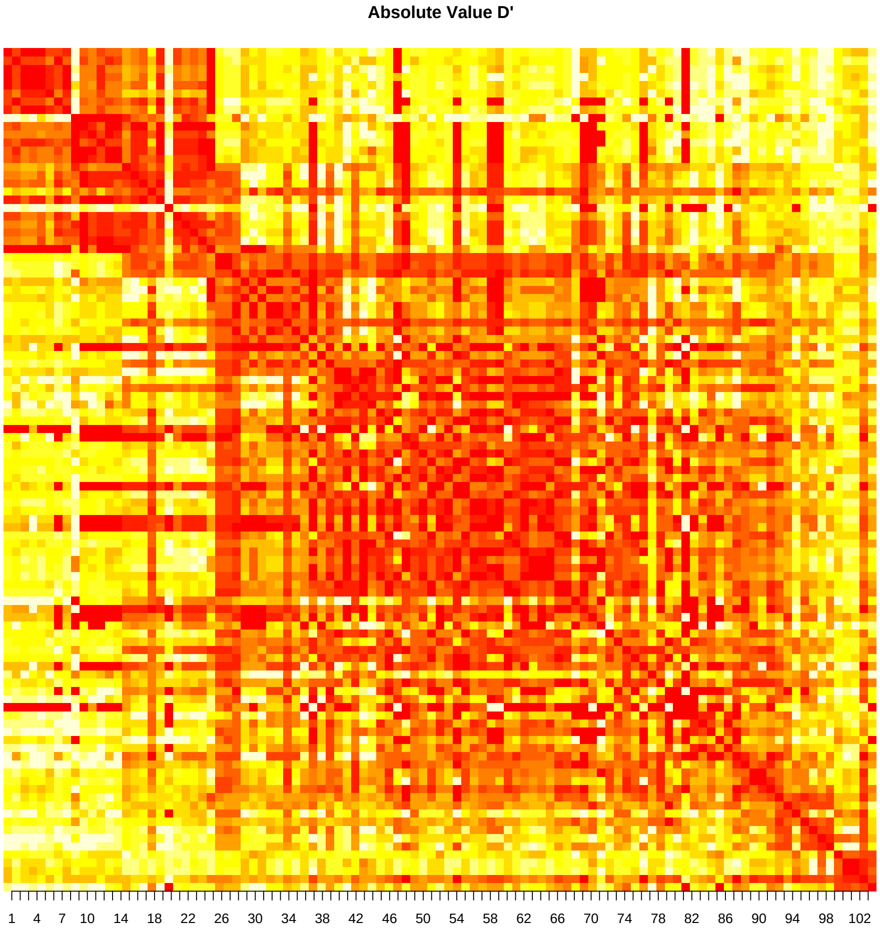
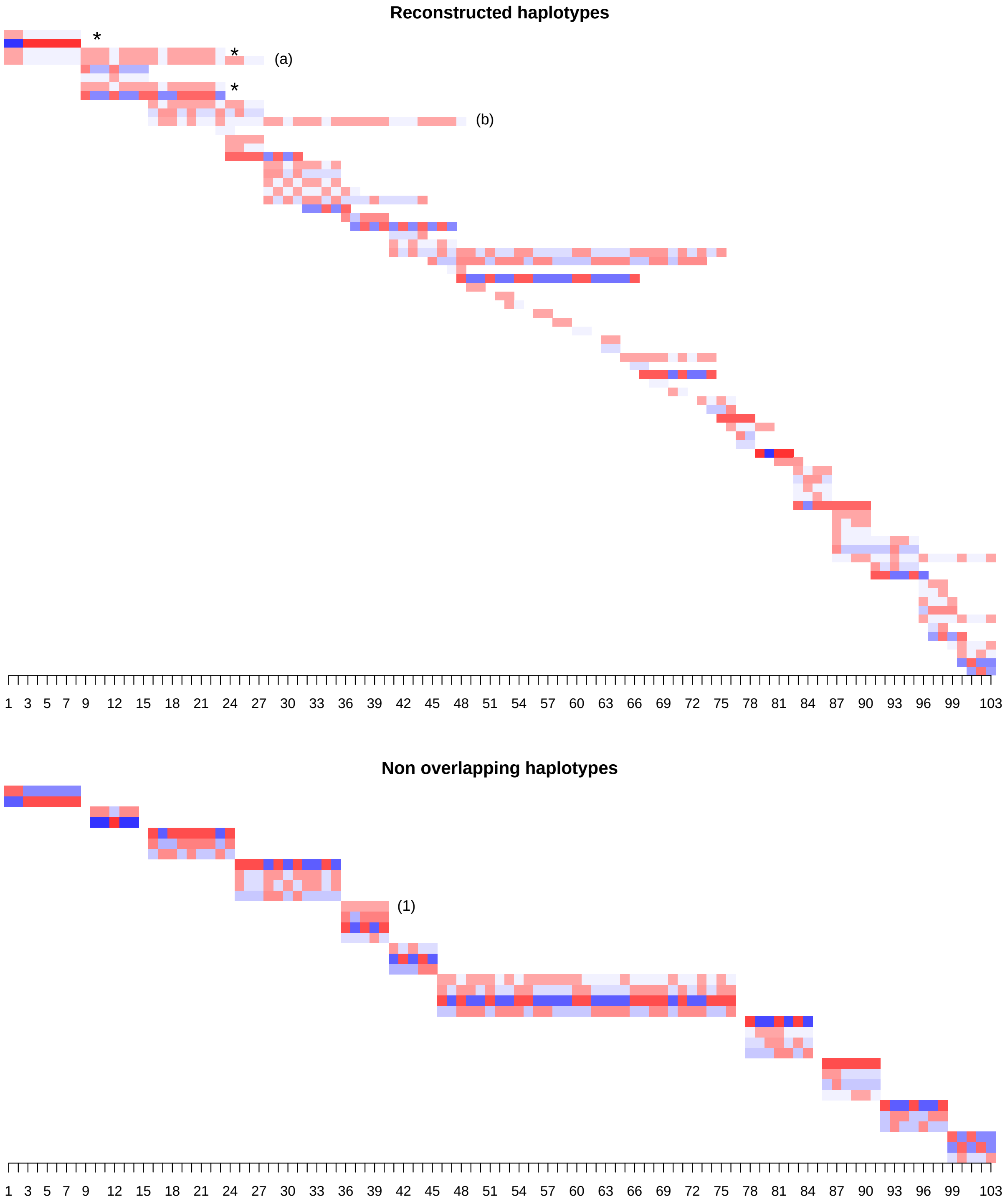
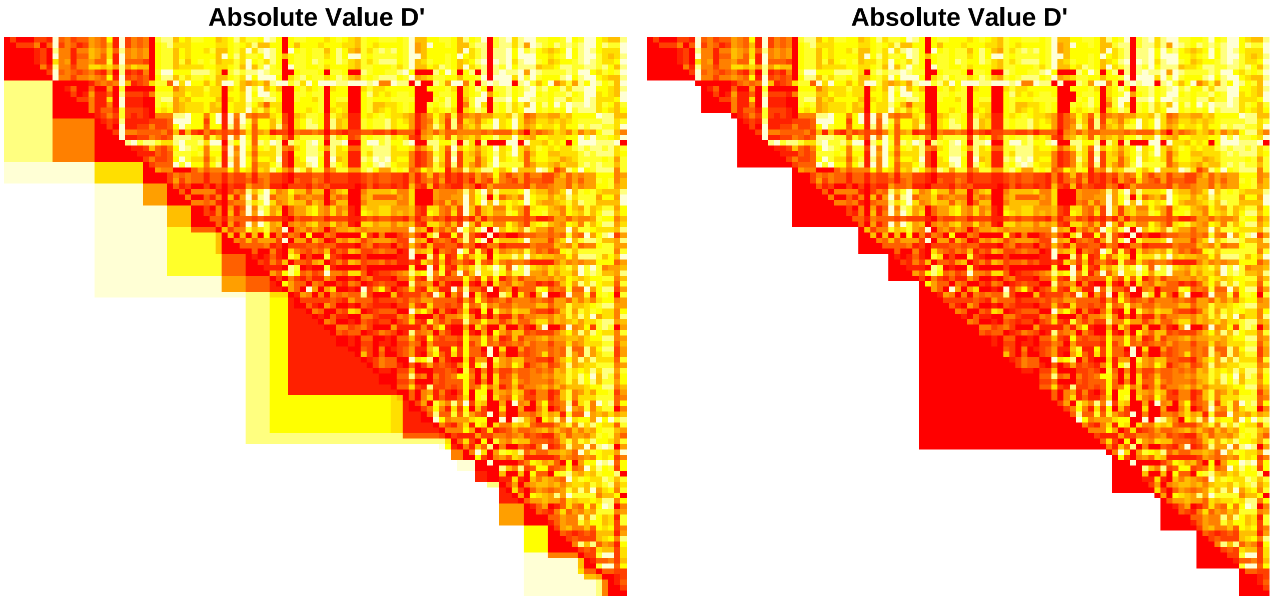

Absolute Value D'
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
98
102

Figure IV:


Reconstructed haplotypes

Non overlapping haplotypes

Figure V:



Absolute Value D'

Absolute Value D'

Co−transmission in the dictionary model

Co−transmission in Daly et al.

Figure VI:

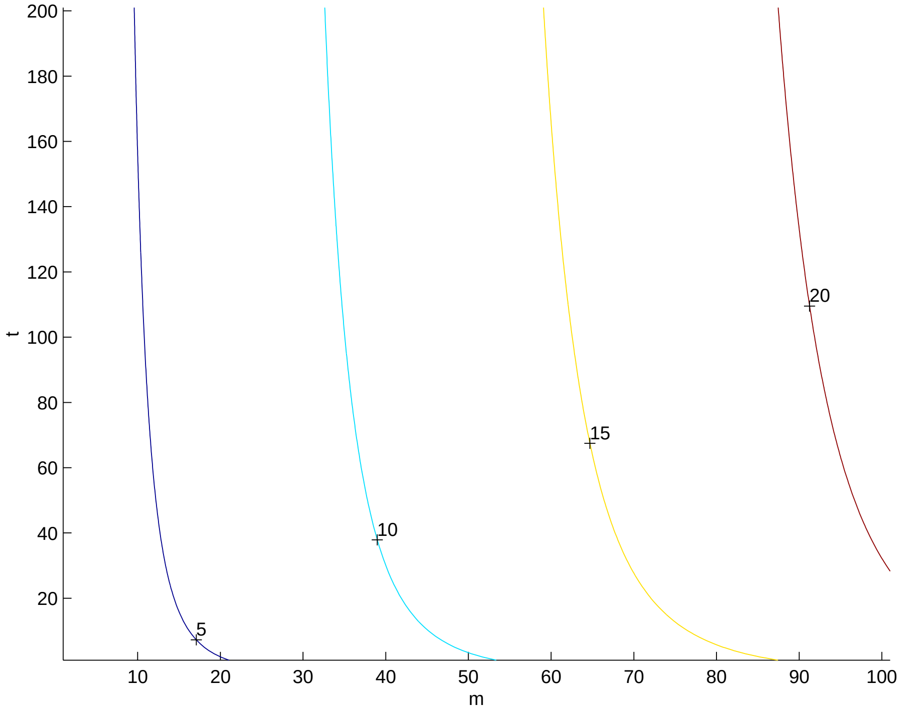

200
180
160
140
120
100
80
60
40
20
t
10
20
30
40
50
60
70
80
90
100
m
5
10
15
20

Figure VII:

Figure I: Graphical representation of $n = 3$ three haplotypes spanning $m = 15$ markers. Each haplotype segment is shaded according the markers contained within it.

Figure II: Dictionary of overlapping blocks considered in the illustrative example. Blue and red are used to represent the two alleles at each SNP. The intensity of the colors is proportional to the probability with which the depicted haplotype appears in a random chromosome in the population (darker being more probable).

Figure III: (a) Co-occurrence probabilities in the true dictionary. Each row and column represent a marker. Red represents a probability of 1, and white represents a probability of 0. (b) Absolute values of D' calculated on the simulated dataset. Again, the more intense the red, the higher the numerical value. (c) Co-occurrence probabilities as estimated by the dictionary model.

Figure IV: Absolute values of the D' measure on the dataset obtained from Daly *et al.* (2001).

Figure V: Pictorial representation of the haplotype dictionary reconstructed by our model (top) and by the analysis presented in Daly *et al.* (2001) (bottom). Blue and red are used to represent the two alleles at each SNP. The color intensity of a haplotype segment is proportional to the probability with which the depicted haplotype appears in a random chromosome in the population (darker being more probable).

Figure VI: Co-transmission probabilities corresponding to our model (left) and the block structure of Daly *et al*. (right) compared to the distribution of D' in the dataset. The upper-right portion of each picture represents the absolute values of D' for the marker pair identified by row and column numbers. The lower left portion corresponds to the co-transmission probability for the same pair.

Figure VII: Level curves of z_{median} for m markers and t multilocus genotypes.