University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 150

Quantification and Visualization of LD Patterns and Identification of Haplotype Blocks

Yan Wang[*] Sandrine Dudoit[†]

[*] Division of Biostatistics, School of Public Health, University of California, Berkeley, yanw@stat.berkeley.edu

[†] Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

Quantification and Visualization of LD Patterns and Identification of Haplotype Blocks

Yan Wang and Sandrine Dudoit

Abstract

Classical measures of linkage disequilibrium (LD) between two loci, based only on the joint distribution of alleles at these loci, present noisy patterns. In this paper, we propose a new distance-based LD measure, R, which takes into account multilocus haplotypes around the two loci in order to exploit information from neighboring loci. The LD measure R yields a matrix of pairwise distances between markers, based on the correlation between the lengths of shared haplotypes among chromosomes around these markers. Data analysis demonstrates that visualization of LD patterns through the R matrix reveals more deterministic patterns, with much less noise, than using classical LD measures. Moreover, the patterns are highly compatible with recently suggested models of haplotype block structure. We propose to apply the new LD measure to define haplotype blocks through cluster analysis. Specifically, we present a distance-based clustering algorithm, DHPBlocker, which performs hierarchical partitioning of an ordered sequence of markers into disjoint and adjacent blocks with a hierarchical structure. The proposed method integrates information on the two main existing criteria in defining haplotype blocks, namely, LD and haplotype diversity, through the use of silhouette width and description length as cluster validity measures, respectively. The new LD measure and clustering procedure are applied to single nucleotide polymorphism (SNP) datasets from the human 5q31 region (Daly et al. 2001) and the class II region of the human major histocompatibility complex (Jeffreys et al. 2001). Our results are in good agreement with published results. In addition, analyses performed on different subsets of markers indicate that the method is robust with regards to the allele frequency and density of the genotyped markers. Unlike previously proposed methods, our new cluster-based method can uncover

hierarchical relationships among blocks and can be applied to polymorphic DNA markers or amino acid sequence data.

Contents

1 Introduction

1.1 Linkage disequilibrium

Linkage disequilibrium (LD) refers to the allelic association at two chromosomal loci. LD has traditionally been quantified by a series of association measures, based solely on the joint distribution of alleles at the two loci of interest, through a two-locus allele contingency table. Commonly-used LD measures have been studied and reviewed extensively (Hedrick, 1987; Devlin and Risch, 1995; Weir, 1996; Pritchard and Przeworski, 2001). Box 1 provides definitions of the classical LD measures D, D', and r^2 for biallelic markers.

Many evolutionary forces can dramatically impact LD, including recombination, gene flow, gene conversion, inbreeding, genetic drift, population bottleneck, selection, mutation, and population substructure. Moreover, simply an admixed sample may show higher LD than what is expected from a more homogeneous sample. In the absence of other forces, LD between two loci is expected to be attenuated exponentially by recombination, as follows:

$$D_t = D_0(1-\theta)^t, \tag{1}$$

where D_t is a measure of LD (as in Box 1) after t generations since an ancient time $t = 0$, and θ is the recombination fraction between the two loci (Pericak-Vance, 1998). This relationship between LD and recombination fraction (approximately equal to genetic distance for tightly linked loci) suggests that LD can play a powerful role in mapping susceptibility genes for complex diseases. That is, strong LD between a marker and a disease susceptibility gene implies that the gene is located in the neighborhood of the marker. Moreover, the role of LD patterns in the investigation of the biochemical processes of recombination and evolution is also significant (Wall and Pritchard, 2003).

However, in the presence of the evolutionary forces mentioned above, an apparent problem with classical LD measures is that a great amount of deviation from Equation (1) has been observed for empirical data. The resulting LD patterns tend to be highly noisy, and hence of limited use in LD mapping and other studies. In order to play a more effective role in mapping susceptibility genes for complex diseases or other endeavors, LD needs to be quantified in a more reliable manner, so that the LD measure between two loci can better reflect their genetic distance, without being confounded by historical events other than recombination.

Box 1. Classical LD Measures D, D', and r^2.
Given a pair of biallelic markers $\mathcal{A}$ (with alleles A_0 and A_1) and $\mathcal{B}$ (with alleles B_0 and B_1), let p_{ij} denote the frequency of haplotype A_iB_j, $i, j \in \{0, 1\}$. Then, $p_{i\cdot} = \sum_{j=0}^{1} p_{ij}$ and $p_{\cdot j} = \sum_{i=0}^{1} p_{ij}$ denote the marginal frequencies of allele A_i at locus $\mathcal{A}$ and allele B_j at locus $\mathcal{B}$, respectively. The LD measure D is defined by

$$D = p_{00} - p_{0\cdot}p_{\cdot 0}.$$

The LD measure D' is defined by

$$D' = \frac{p_{00} - p_{0\cdot}p_{\cdot 0}}{D_{max}},$$

where D_{max} is the largest value of D, given the marginals:

$$D_{max} = \begin{cases} \min\{p_{0\cdot}p_{\cdot 1}, p_{1\cdot}p_{\cdot 0}\}, \text{ if } D \geq 0 \\ \min\{p_{0\cdot}p_{\cdot 0}, p_{1\cdot}p_{\cdot 1}\}, \text{ if } D < 0. \end{cases}$$

The LD measure r^2 is defined by

$$r^2 = \frac{(p_{00} - p_{0\cdot}p_{\cdot 0})^2}{p_{0\cdot}p_{\cdot 0}p_{1\cdot}p_{\cdot 1}}.$$

A number of new measures have recently been suggested for quantifying LD. Morton et al. (2001) proposed the association probability ρ as a model-based metric for allelic association. With effects such as genetic drift and directional selection being controlled for by model parameters, the ρ metric is a more robust measure of LD than either D' or r^2. Pritchard and Przeworski (2001) proposed another model-based measure, also denoted by ρ, which is a function of the recombination fraction. Instead of measuring LD for pairs of loci, it estimates an average LD for a given chromosomal region by using the expected r^2.

1.2 Block model for LD patterns

The haplotype block structure has been considered as an appealing model for visualizing LD patterns, after a number of recent studies suggested that single nucleotide polymorphisms (SNPs) across a relatively large chromosomal region could be parsed into blocks of various lengths (Daly et al., 2001; Patil et al., 2001; Dawson et al., 2002; Gabriel et al., 2002; Phillips et al., 2003). In this context, *blocks* of SNPs or other dense markers can be loosely defined as sets of contiguous markers that exhibit: (i) low haplotype diversity within blocks; and (ii) strong LD within blocks and sharp decay of LD between blocks (Anderson and Novembre, 2003).

However, the actual visualization of LD patterns throughout the genome depends heavily on how one defines the block structure. At this point, no agreement has been reached upon a universal definition of blocks, and as a result, the block structures identified by different groups via different rules carry different features. Some methods are based on criterion (i) only (Patil et al., 2001; Dawson et al., 2002; Zhang et al., 2002a; Koivisto et al., 2003), some on (ii) only (Daly et al., 2001; Gabriel et al., 2002; Wang et al., 2002), and others on both (i) and (ii) (Anderson and Novembre, 2003). Intuitively the last combined approach should be of more value. In addition, some algorithms allow for overlapping blocks and gaps between blocks, which may lead to ambiguous block boundaries (Dawson et al., 2002; Gabriel et al., 2002), while others produce a unique partition with disjoint and adjacent blocks (Wang et al., 2002; Anderson and Novembre, 2003; Koivisto et al., 2003). According to the three criteria proposed by Wall and Pritchard (2003), a good method of defining block structure should achieve: (i) high coverage of the data by a small number of haplotypes within blocks (i.e., low haplotype diversity); (ii) internally consistent blocks (i.e., the boundaries are consistent with the underlying chromosomal loci where recombinations occur); (iii) unambiguous block boundaries (i.e., a unique solution with disjoint and adjacent blocks is preferred).

Anderson and Novembre (2003) and Koivisto et al. (2003) independently adopted a model selection approach, based on the minimum description length (MDL) principle (Rissanen, 1978, 1989) as a global criterion for selecting the block structure. The MDL principle, originally proposed by Jorma Rissanen in 1978, is a method for inductive inference, i.e., the process of inferring a model from observed data. The MDL conceptualization originated from the notion that the existence of a data generating model necessarily implies redundancy in the information from successive observations. The more we are able to compress the data, the more we learn about the underlying model that generated these data. The "best" model given a set of observations is the one that permits the greatest compression of the data, as indicated by the shortest description length. Anderson and Novembre (2003) and Koivisto et al. (2003) both calculate the description length of SNP data using a two-stage coding scheme, in which one first builds a probabilistic model for the block structure, and next defines description length as the negative log-likelihood penalized by a function of model parameters. However, the two groups consider different probabilistic models for the block structure. Anderson and Novembre (2003) model the dependence between blocks using a Markov model. In contrast, Koivisto et al. (2003) assume independence between markers and between blocks.

1.3 Overview

LD measure. In this paper, we propose a distance-based, nonparametric measure of linkage disequilibrium, the R measure. This new measure, based on phased genotype data, is a function of the length of shared ancestral segments among chromosomes, thereby capturing information on recombination and the geneological relationships among chromosomes. In this way, confounding effects from population history, other than recombination, can be greatly limited. In comparison to classical LD measures, the new measure takes into account multilocus haplotypes in the neighborhoods of the two loci of interest and information on the genetic (or physical) distance between these two loci. Moreover, it can be computed from genotype data with either missing alleles or missing phase. All these factors contribute to making the new measure more informative and robust than classical measures.

Identification of haplotype blocks. LD patterns visualized through the R matrix are highly compatible with the postulated existence of haplotype blocks in the genome. We propose to apply the new LD measure to define haplotype blocks through cluster analysis. Specifically, we present a *distance-based* clustering algorithm, `DHPBlocker`, which performs *hierarchical partitioning* of an ordered sequence of markers into disjoint and adjacent blocks with a hierarchical structure. The proposed method integrates information on both LD and haplotype diversity, through the use of silhouette width (Rousseeuw, 1987) and description length (Rissanen, 1978, 1989) as cluster validity criteria, respectively. Thus, blocks are identified based on global criteria rather than local criteria using a sliding window as in most of the existing methods (Patil et al., 2001; Daly et al., 2001; Gabriel et al., 2002; Wang et al., 2002). Moreover, `DHPBlocker` allows us to uncover the hierarchical structure of haplotype blocks, which is expected intuitively from the tree struture of population genealogy. It should be noted that `DHPBlocker` is not based on models for the haplotype block structure and that no assumptions are made regarding the mechanisms that determine the observed LD patterns.

Imputation of missing data. Missing data is an important and practical issue in methodological development for the analysis of SNP data, because there is usually a significant proportion of missing genotypes due to deficiencies in the SNP genotyping technology. Moreover, some methods involving haplotype analysis require phase information and treat genotypes with unresolved phase as missing. For example, in the SNP dataset from the 5q31 region (Daly et al., 2001), there are 12.7% of ungenotyped SNPs and 7.3% of genotypes with unresolved phase, among 103 SNPs and 516 chromosomes. This may pose a serious problem if there is no reliable approach for imputation of the missing data. Currently, missing genotypes are imputed mostly at random, according to the empirical SNP allele frequencies. Alternatively, haplotype frequencies can be estimated using the EM algorithm (Excoffier and Slatkin, 1995) or Bayesian approaches (Stephens et al., 2001; Niu et al., 2002; Xing et al., 2004), so that there is no need to actually count haplotypes in order to estimate their frequencies. K-nearest neighbor (KNN) imputation, used by Troyanskaya et al. (2001) in distance-based cluster analysis of microarray data, can be conveniently applied in our setting as a useful by-product of the LD quantification process.

The article is organized as follows. Section 2 describes our proposed methods for addressing the above three issues, namely: quantification of LD, identification of haplotype blocks, and imputation of missing data. These methods are assessed in Section 3 using SNP data from the human 5q31 region (Daly et al., 2001) and the class II region of the human major histocompatibility complex (Jeffreys et al., 2001). These datasets have been used repeatedly in the literature for evaluating new measures of LD (Zhang et al., 2002b) and new approaches for identifying block boundaries (Anderson and Novembre, 2003; Koivisto et al., 2003), because their underlying block structure is considered as known, based on either experimental evidence of recombination hotspots or various quantitative methods of validation. Finally, Section 4 summarizes our findings and outlines open questions.

2 Methods

Suppose we have *phased* genotype data on n chromosomes at J markers, in the form of an $n \times J$ matrix, $X = (X(i,j) : i = 1, \ldots, n;\ j = 1, \ldots, J)$, where $X(i,j)$ denotes the allele at marker j for chromosome i.

2.1 Linkage disequilibrium measure

2.1.1 Motivation

Genealogical relationships among chromosomes can be described by a tree-like structure. This structure is locus-specific, because the genealogical relationships vary across the genome due to evolutionary forces, such as recombination and gene conversion. Markers in high LD, attributed to low recombination rates, tend to co-segregate through generations and may therefore possess very similar genealogical trees. Therefore, LD between two loci may be quantified based on extent of similarity of their genealogical trees.

To describe genealogical relationships among chromosomes at a given locus, we use a between-chromosome similarity matrix, where each entry is a measure of the genealogical distance between a chromosome pair. Given a pair of chromosomes, the total time to their most recent common ancestor (MRCA) (measured in number of generations as the unit of time) quantifies their genealogical distance at a given locus. This quantity is related to the length of the common ancestral segment (measured in Morgan as the unit of genetic distance) shared by the chromosome pair around that locus. The closer the chromosomes are related genealogically, the longer the common ancestral segment they share. Since the length of the common ancestral segment cannot be precisely observed, we use the length of the shared haplotype as a proxy measure for the similarity between a chromosome pair.

2.1.2 Locus-specific between-chromosome similarity (L_j matrices)

We denote the length of the haplotype shared by chromosomes i and i' around marker j by $L_j(i,i')$ and the index of the leftmost (respectively rightmost) marker of the shared haplotype by $m_{L,j}(i,i')$ (respectively $m_{R,j}(i,i')$). In the case that chromosomes i and i' share the same allele at marker j, i.e.,

$X(i,j) = X(i',j)$, we let

$$\begin{aligned} m_{L,j}(i,i') &= j - \max_{k=1,\ldots} \{ \sum_{l=1}^{k} I(X(i,j-l) = X(i',j-l)) = k \}, \\ m_{R,j}(i,i') &= j + \max_{k=1,\ldots} \{ \sum_{r=1}^{k} I(X(i,j+r) = X(i',j+r)) = k \}, \end{aligned}$$

where $I(X(i,j-l) = X(i',j-l))$ is the indicator function, equal to one if chromosomes i and i' have the same allele at marker $j-l$, and equal to zero otherwise. Then, we define the lengths of the shared haplotypes as

$$L_j(i,i') = \begin{cases} D(m_{L,j}(i,i'), m_{R,j}(i,i')), & \text{if } X(i,j) = X(i',j), \\ 0, & \text{otherwise,} \end{cases}$$

where $D(x,y)$ denotes a measure of distance between two markers x and y. Theoretically, the genetic distance (in Morgan, M) is the preferred measure of distance. Physical distance (in base-pair, bp) can be used when genetic distance is not available, under the assumption that they have a linear relationship. When the marker density is constant across the chromosomal region of interest, the number of markers in between x and y can also be used. We evaluate the lengths of shared haplotypes $L_j(i,i')$ for every chromosome pair (i,i'), at every marker locus j, so that J $n \times n$ similarity matrices, $L_j = (L_j(i,i') : i,i' = 1,\ldots,n)$, $j = 1,\ldots,J$, are constructed. In the case of missing value(s) for one (or both) chromosome(s) at a marker j, the evaluation of $L_j(i,i')$ is simply skipped. Each L_j matrix is a between-chromosome similarity matrix describing the underlying genealogical patterns among n chromosomes at a marker locus j. Through these L_j matrices, we develop a new measure of LD, the dissimilarity R (Section 2.1.3) and perform KNN missing value imputation (Section 2.3).

2.1.3 Between-marker dissimilarity (R matrix)

Due to the symmetry of the L matrices, only the lower-left (or upper-right) triangles are necessary in the evaluation of a dissimilarity measure between them. The lower-left triangle of L_j, containing $n(n-1)/2$ elements, can be transformed into a vector, $\tilde{L}_j$. Based on J such vectors $\tilde{L}_j$, we define a $J \times J$ between-marker distance matrix R with entries

$$R(j,j') = d(\tilde{L}_j, \tilde{L}'_j),$$

where $d(x, y)$ is any measure of distance between two vectors x and y, such as the correlation distance, Euclidean distance, or Manhattan distance. The distance $R(j, j')$ is evaluated for each pair of markers j and j', so that a $J \times J$ between-marker dissimilarity matrix is constructed as $R = (R(j, j') : j, j' = 1, \ldots, J)$. Subject matter knowledge might come into play when choosing an appropriate distance function d. However, our analyses of SNP data show that the LD patterns and block structures identified based on this new R measure are robust to the choice of distance measure d (Section 2.2). Throughout the paper, we use the correlation distance, which is one minus the Pearson correlation coefficient. Figure 1 shows the relationship between the L similarity matrices and the final R matrix of LD measures.

2.2 Distance-based hierarchical partitioning method for identifying haplotype blocks – `DHPBlocker`

2.2.1 Motivation

Partitioning based on LD through silhouette width criterion. Information on LD is captured by the new LD measure R. Given the dissimilarity nature of R, distance-based cluster analysis can be conveniently applied to group J markers into B disjoint blocks, under the constraint of linear spatial order of *consecutive* markers. As in any clustering problem, a figure of merit must be chosen to determine the optimal number of clusters and to validate the clustering results. According to the silhouette validation method (Rousseeuw, 1987), we seek the partition that maximizes the mean silhouette width in the high-dimensional space of all 2^{J-1} potential partitions (Section 2.2.2). A deletion/substitution/addition (D/S/A) algorithm (Sinisi and van der Laan, 2004) is proposed to efficiently search for a sequence of partitions, $\mathcal{B}_t$, of the markers into t blocks, $t = 1, \ldots, T$, where partition $\mathcal{B}_t$ maximizes the mean silhouette width for a given number of blocks t (Section 2.2.3). This series of partitions is subject to further model selection for the optimal number of blocks.

Model selection based on haplotype diversity through the MDL principle. In information theory, the entropy of a random variable is a measure of the uncertainty for its distribution (Cover and Thomas, 1991). Haplotype diversity for a given block can be evaluated by the entropy of the haplotype distribution; the lower the entropy, the lower the haplotype diversity. Entropy should therefore be minimized to determine the block structure with the optimal number of blocks. Note that the entropy of a distribution corresponds to the risk for the negative log-likelihood loss function. Thus, when using the empirical distribution of the data, minimization of entropy is equivalent to maximization of the likelihood. As in any model selection method based on the maximum likelihood criterion, a model with extra complexity tends to overfit the data. In order to achieve a balance between model complexity and haplotype diversity, we employ the minimum description length (MDL) principle with a two-stage coding scheme (Rissanen, 1989). Anderson and Novembre (2003) provide intuitive illustrations of the MDL principle and two-stage coding scheme in the context of the identification of haplotype blocks. The description length of the data under a model is the negative log-likelihood penalized by the model complexity, which is the code length required for encoding the probabilistic model, including the estimated values of the parameters. A probabilistic model of block structure is described in Section 2.2.4 for the purpose of calculating description lengths. Unlike the block model of Anderson and Novembre (2003), ours does not account for dependence of genotypes between blocks, because this dependence is related to LD and hence captured by the new LD measure. While previous MDL-based approaches rely on dynamic programming to minimize description length over a large space of block models, we only evaluate a limited number of models (a partial space) selected based on LD using the R matrix and the D/S/A algorithm. In other words, we give LD more importance than haplotype diversity when identifying the block structure.

Hierarchical block structure. Block boundaries are expected to correspond to chromosomal sites with either *higher* recombination rates (e.g. hotspots) or crossovers at *more ancient* times. This mechanism produces a hierarchical relationship for the blocks, where *children* blocks are nested under a larger *parent* block. In order to reveal this hierarchical structure, we apply a *divisive* hierarchical clustering algorithm to the markers (Section 2.2.5). In comparison to *agglomerative* approaches, divisive approaches tend to be robust in terms of retaining the overall hierarchical structure, which means that the root or “upper” levels of the *dendrogram* (i.e., graphical representation of the resulting tree-like cluster structure) are representative of the global relationships among the objects.

2.2.2 Silhouette width

In general, for every object in a clustered sample, the silhouette width can be calculated as a measure reflecting how well the object belongs to its assigned cluster relative to other clusters (Kaufman and Rousseeuw, 1990). In our case, given a clustering of the markers into blocks, the *silhouette width* $S(j)$ for marker j is defined by

$$S(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}},$$

where $a(j)$ is the average distance of marker j to all other markers in the same block; $b(j)$ is the average distance of marker j to all the markers in the closest block, which, under the constraint of linear spatial order, should be one of the two adjacent blocks with the smaller average distance. Silhouette widths range from -1 to 1. A large silhouette width $S(j)$ corresponds to a small within-cluster distance and/or a large between-cluster distance, and therefore suggests that marker j is well clustered. The *mean silhouette width* across all markers provides an overall assessment of the clustering results (Kaufman and Rousseeuw, 1990). According to the *silhouette validation method*, an optimal clustering has the highest mean silhouette width. Thus, for clustering procedures based on the R distance measure, markers with strong LD (i.e., small R measures) should be clustered together, while markers having experienced larger LD decay (i.e., large R measures) should be assigned to different blocks.

2.2.3 Deletion/substitution/addition (D/S/A) distance-based ordered partitioning algorithm

A *deletion/substitution/addition* (D/S/A) algorithm (Sinisi and van der Laan, 2004) is used to generate candidate block structures seeking to maximize the mean silhouette width for the LD measure R. As the name suggests, there are three basic moves in searching for a better partition than the current one: deletions, substitutions, and additions. In our context, a deletion move corresponds with merging two adjacent blocks; a substitution move corresponds with shifting block boundaries; and an addition move corresponds with a split of one of the current blocks into two blocks. The algorithm is described step-by-step in Box 2.

Box 2. Deletion/Substitution/Addition Algorithm.

1. **Deletion.** Among all deletion moves, find the one that maximizes the mean silhouette width and see if it improves on the current mean silhouette width. If so, carry out the deletion move and repeat this step until no further merging of two adjacent blocks increases the mean silhouette width. Then, continue to the next step.
2. **Substitution.** Among all substitution moves, find the one that maximizes the mean silhouette width and see if it improves on the current mean silhouette width. If so, carry out the substitution move and go back to the first step. Otherwise, continue to the next step.
3. **Addition.** Among all addition moves, find the one that maximizes the mean silhouette width and see if it improves on the current mean silhouette width. If so, carry out the addition move and go back to the first step. Otherwise, the algorithm stops.

Although the partitioning is done sequentially, the D/S/A algorithm allows constant correction and refinement of the boundaries identified in previous steps, through block merging, boundary shifting, and block splitting, until the mean silhouette width does not increase any more. As a result of applying the D/S/A algorithm, we obtain a series of block structures, $\boldsymbol{\mathcal{B}} = (\mathcal{B}_t : t = 1, \ldots, T)$, where $\mathcal{B}_t$ is a partition of the J markers into t blocks. The set $\boldsymbol{\mathcal{B}}$ provides a model space for selection of the number of blocks using the MDL criterion to be described below.

2.2.4 Minimum description length (MDL)

To describe the probabilistic model corresponding to a partition $\mathcal{B}$ of J markers into B blocks, denote the block boundaries by $e = (e(b) : b = 0, \ldots, B)$, where $e(b)$ is the index of the rightmost marker in block b, $e(0) \equiv 0$, and $e(B) \equiv J$. Given a block b, denote the set of K_b distinct haplotypes by $\mathcal{H}_b = \{H_b(k) : k = 1, \ldots, K_b\}$. Assume that within each block b, haplotypes follow a multinomial distribution with haplotype frequencies $\pi_b = (\pi_b(k) : k = 1, \ldots, K_b)$, where $\sum_{k=1}^{K_b} \pi_b(k) = 1$. Further assume the independence of haplotypes between blocks. A block model can therefore be specified as $\mathcal{B} = (e, \pi, \mathcal{H})$, where $\mathcal{H} = \{\mathcal{H}_b : b = 1, \ldots, B\}$ and $\pi = \{\pi_b : b = 1, \ldots, B\}$

denote, respectively, sets of distinct haplotypes and their corresponding frequencies for each block.

From information theoretical results (Hansen and Yu, 2001; Lee, 2001), the following *code lengths* (in bits) are required to encode the model $\mathcal{B}$: $B \log_2 J$ for encoding the block boundaries e; given a block b, $\frac{K_b - 1}{2} \log_2 n$ for encoding the estimated haplotype frequencies $\hat{\pi}_b$; and $K_b(e(b) - e(b-1))$ for encoding the distinct haplotypes $\mathcal{H}_b$. Under independence of haplotypes between blocks, the overall code length of model $\mathcal{B}$ is

$$\varphi(\mathcal{B}) = B \log_2 J + \sum_{b=1}^{B} \frac{K_b - 1}{2} \log_2 n + \sum_{b=1}^{B} K_b(e(b) - e(b-1)). \tag{2}$$

To compute the *negative* $\log_2$ *likelihood* under model $\mathcal{B}$, we denote the haplotype data within block b as $\mathcal{Y}_b = \{Y(i,b) = (X(i,j) : j = e(b-1) + 1, \ldots, e(b)),\ i = 1, \ldots, n\}$. Then,

$$-\log_2 Pr(X|\mathcal{B}) = -\sum_{i=1}^{n} \sum_{b=1}^{B} \sum_{k=1}^{K_b} I(Y(i,b) = H_b(k)) \log_2 \pi_b(k), \tag{3}$$

where $I(Y(i,b) = H_b(k))$ is the indicator function evaluating whether the observed haplotype $Y(i,b)$ is the same as one of the distinct haplotypes $H_b(k)$ for block b, $b = 1, \ldots, B$. Thus, by adding Equations (2) and (3), we obtain the *description length* of the data given model $\mathcal{B}$, denoted by $\phi_{\mathcal{B}}(X)$. Then, according to the MDL principle, the optimal block model $\mathcal{B}^*$ among a set $\boldsymbol{\mathcal{B}}$ is selected as $\mathcal{B}^* = \arg\min_{\mathcal{B} \in \boldsymbol{\mathcal{B}}} \phi_{\mathcal{B}}(X)$.

2.2.5 Hierarchical partitioning

The ordered partitioning algorithm, described in Sections 2.2.2 – 2.2.4, only allows identification of a single level of block structure. In this subsection, we propose a divisive hierarchical clustering method, in which our algorithm is implemented recursively within blocks. The recursive process stops for a given block if the model with no partition is favored based on the MDL principle (i.e., $\mathcal{B}^*= \mathcal{B}_1$). Block boundaries identified at the same level of the block hierarchy are viewed as corresponding to LD decays of similar magnitude; block boundaries at upper levels of the tree correspond to sharper LD decay than those at lower levels.

2.3 KNN imputation of missing data

Our proposed *K-nearest neighbor (KNN)* imputation method relies upon the L matrices of shared haplotype lengths defined in Section 2.1.2. If a chromosome i has a missing allele $X(i,j)$ at marker j, find the K chromosomes which are closest to i according to the genealogical similarity matrix L_j and have phased allele data at marker j. Since SNP data are binary, impute the missing allele by majority vote, that is, choose the most frequent allele among the K closest chromosomes. Votes weighted by the similarities $L_j(i,\cdot)$ can also be considered. The KNN approach can also be used for polymorphic markers, where voting would be over a larger number of alleles. Note that there is no need for imputing missing data to construct the L matrices, and hence the R matrix of LD measures.

3 Applications

3.1 SNP data from the human 5q31 region

Daly et al. (2001) genotyped 103 SNPs, spanning 500 kb of the human 5q31 region, in 129 parent-child trios. The 103 SNPs have minor allele frequencies greater than 5% and are contiguously located with a density of 1 SNP roughly every 4 kb, with the exception of markers 98 and 99 that have a physical distance larger than 100 kb. The LD patterns from this dataset have been well studied in Daly et al. (2001) and subsequent articles. Daly et al. (2001) originally defined 11 blocks in this region, indexed by the following markers: 1-8; 10-14; 16-24; 25-35; 36-40; 41-45; 46-76; 78-84; 86-91; 92-98; 99-103. Their definition allows for a one-marker gap between adjacent blocks, in the case that the single marker does not favor belonging to either of its two adjacent blocks. Such a situation may occur when only one marker is genotyped within a block of a small size. Later, by two independent MDL approaches, Anderson and Novembre (2003) and Koivisto et al. (2003) identified no-gap, no-overlap block structures with boundaries indexed by the rightmost marker in each block: $\{8, 14, 24, 36, 47, 57, 76, 86, 91, 98\}$ and $\{14, 24, 41, 91, 98\}$, respectively. The above three block structures are displayed in Figure 3.

LD measures and visualization of LD patterns. The classical LD measures r^2 and D' and the new measure R were computed for each marker pair and then displayed using gray-scale images in Figure 2. Genotypes for which phase information could not be resolved based on the parent-offspring relationship of the trios were treated as missing and, in particular, were omitted when computing r^2 and D'. Instead of the r^2 and D' measures themselves, matrices of $1 - r^2$ and $1 - |D'|$ values are displayed to facilitate comparison with the R matrix, for which smaller values (i.e., darker shades of gray) correspond to higher LD. As illustrated in panel (A) of Figure 2, a typical problem of the classical LD measures r^2 and D' is that distant markers may present high LD values, whereas markers closely located within the same block (as defined by Daly et al. (2001)) may present low LD values. In comparison, the R distance matrix exhibits a more deterministic behavior and reflects LD patterns that are highly compatible with the previously postulated existence of haplotype blocks in the 5q31 region. Markers within the same block have uniformly low R values (as reflected by the dark submatrices along the diagonal) and the R values increase off-diagonal, with dramatic increments typically occuring at the block boundaries. Moreover, given two blocks, marker pairs consisting of a marker from each of these two blocks tend to have similar R measures, suggesting, as expected, that the same amount of LD is maintained by these marker pairs.

Identification of block structure. `DHPBlocker` was applied to identify the block structure in region 5q31. Missing data were imputed by the KNN method (Section 2.3), with $K = 5$, to allow computation of the MDL criterion. The resulting block structure is highly consistent with previous findings (Daly et al., 2001; Anderson and Novembre, 2003; Koivisto et al., 2003). The identified block boundaries, indexed by the rightmost marker in each block, are $\{8, 9, 14, 24, 36, 42, 44, 76, 77, 91, 98\}$, with the hierarchical relationship displayed in Figure 3. A single boundary between markers 98 and 99 was identified to define the block structure at the highest level. This boundary is consistent with the relatively large gap across these two markers, where LD is expected to decay to the greatest extent. In contrast, the block between markers 37 and 77 was partitioned in the last three steps of the clustering process, suggesting that LD does not decay as sharply within this block as it does elsewhere. This is also evidenced from the fact that most of the disagreements with block boundaries identified in previous studies occur within this region. Moreover, compared to the automatic algorithms of Anderson and Novembre (2003) and Koivisto et al. (2003), which also aim for disjoint and adjacent blocks, `DHPBlocker` is more likely to identify blocks defined by a single marker. For instance, we identified markers 9 and 77 as single-marker blocks. This finding is compatible with the fact that the Daly et al. (2001) method did not support assigning these markers to either one of their neighboring blocks.

Impact of missing data imputation method. `DHPBlocker` was applied with missing data imputed by the KNN method (Section 2.3) and generated the same block structures for $K = 1$, 5, and 9 neighbors. For comparison purposes, we also applied a naive imputation method, whereby missing $X(i, j)$ alleles were randomly imputed based on the empirical allele frequencies of marker j (i.e., the allele frequencies for the Daly et al. (2001) genotype data). Figure 4 shows that for random imputation, all of the previously defined block boundaries were identified and the upper-level block structure was highly similar to that resulting from KNN imputation. The preserved upper-level structure provides evidence for the robustness of `DHPBlocker`. However, many more boundaries were found at lower levels, with a total of 32 blocks in 11 levels, compared to 11 blocks in 7 levels for KNN imputation. This suggests that random imputation of the missing alleles delays termination of the clustering algorithm due to inaccurate calculation of the MDL criterion. Further analyses comparing these two imputation methods consistently suggested that random imputation tends to identify unnecessarily fine levels of hierarchical structure (results not shown). In contrast, the KNN method performed satisfactarily in terms of achieving proper termination of the clustering procedure via the MDL principle. In the remainder of the article, we only report results based on KNN missing data imputation, with $K = 5$ neighbors.

Impact of marker selection. To further study the robustness of `DHPBlocker`, different subsets of the original set of 103 markers were analyzed. First, subsets of SNPs were selected based on their empirical minor allele frequencies: 84 SNPs with minor allele frequencies greater than 10% (Figure 5A); 58 SNPs with minor allele frequencies greater than 20% (Figure 5B); and 36 SNPs with minor allele frequencies greater than 30% (Figure 5C). Next, 80% and 50% of the original markers were selected at random, thus reducing the number of SNPs to 83 and 52, respectively (Figures 5D and E). We found the upper-level block structures to be highly consistent with the one obtained from the original set of SNPs. As expected, subsets of fewer SNPs tended to yield block structures with smaller numbers of boundaries and hierarchical levels. Since some boundary indices shifted due to the absence of the original boundaries from the subset of selected markers, all of the selected SNPs are plotted in Figure 5.

3.2 SNP data from the human MHC

We also applied our new LD measure and `DHPBlocker` to SNP data from the class II region of the human major histocompatibility complex (MHC). The dataset of Jeffreys et al. (2001) contains genotypes for 296 SNPs in a 216 kb segment for 50 individuals. Sperm crossover analysis confirmed six recombination hotspots: the DNA1–3, DMB1–2, and TAP2 hotspots, reported as narrow regions of several hundred up to about one thousand base pairs in length. The approximate centers of these hotspots were obtained from links on the website `http://www.le.ac.uk/genetics/ajj/HLA/`. Although recombination hotspots may not be the only mechanism generating block boundaries, LD is expected to decay sharply at these hotspots.

The genotype phase was inferred using the PHASE 2.0.2 program (Stephens et al., 2001; Stephens and Donnelly, 2003). As suggested by the software documentation, PHASE was run five times using different seeds and the result of the run (seed 3838) with the largest goodness-of-fit measure was used for our analysis. For markers where the phase or the missing alleles were difficult to infer, we used the PHASE inferred results when phase certainty and genotype certainty were both $\geq$ 90%. That is, markers with either genotype certainty or phase certainty less than a 90% threshold were regarded as missing. As a result, empirical allele frequencies in the phased dataset may differ from those in the original unphased dataset. In particular, 14 markers were considered as monomorphic in the phased dataset. The allele frequencies used in the analyses below were computed based on the phased dataset. A separate analysis performed for a 50% threshold yielded very similar results. Therefore, only the results for a 90% threshold are reported here.

LD measures and visualization of LD patterns. Figure 6 displays gray scale images of the classical $1 - r^2$ and $1 - D'$ measures and the new R measure, after removal of the 14 monomorphic markers from the phased dataset. In the R matrix, the locations of sharp decay of LD correspond well with the recombination hotspots identified experimentally in Jeffreys et al. (2001). Considering the block structure defined by these hotspots, R generally reports high LD (i.e., darker gray) for marker pairs from the same block and low LD (i.e., lighter gray) for marker pairs from different blocks. In comparison, the classical LD measures present a much noisier picture.

Identification of block structure. We applied DHPBlocker to identify the block structure in the MHC region, based on subsets of markers with varying minor allele frequencies. Specifically, we constructed four subsets of markers, (a) – (d), containing 282, 247, 193, and 144 SNPs, with empirical minor allele frequencies larger than 0%, 5%, 10%, and 20%, respectively. The numbers of levels for the four resulting hierarchical block structures are 4, 6, 4, and 3, respectively. The upper levels of all four block structures are in good agreement with the blocks defined by the recombination hotspots. This finding is emphasized in Figure 7A, by plotting only the upper two levels of the block structures. Using different SNP subsets, we consistently found that DNA2, DNA3, DMB2, and TAP2 were among the strongest boundaries, defining the highest levels of block structure. DMB1 tended to be a weaker boundary than DMB2, but was identified at least at the next level of the block structure. The DNA1 hotspot was found much weaker in terms of decay of LD than the other hotspots. A block boundary at about 1.7 kb upstream of the approximate center of DNA1 was identified at the 4th level when using SNP subsets (a) and (b), at the 2nd level when using SNP subset (c), and at the 3rd level when using SNP subset (d).

Figure 7A shows that, at the upper two levels, there were other loci consistently identified as block boundaries using different SNP subsets, including one at 15 kb upstream of the DNA1 hotspot center, one at 2.3 kb upstream of the DMB1 hotspot center, and one at 29 kb upstream of the TAP2 hotspot center. We do not yet have any experimental or other evidence to justify these findings. In addition, these loci may not necessarily correspond to recombination hotspots.

Figure 7B displays all block boundaries identified by `DHPBlocker`, irrespective of their level in the block hierarchy, for the four different SNP subsets. Most of the block boundaries were found to be located within a narrow region of the approximate centers of the recombination hotspots of Jeffreys et al. (2001), rather than spread out evenly. This finding is compatible with the previous observation that historical recombinations tend to cluster in the genome (Jeffreys et al., 2001).

4 Discussion

LD measure, R. The proposed LD measure R facilitates both qualitative and quantitative description of LD patterns. Color images (or other graphical displays) of matrices of R measures allow direct visualization of LD patterns and identification of haplotype blocks. LD patterns revealed by the R measure are less noisy than those obtained from classical LD measures: the R measure shows low variance for marker pairs located within the same blocks and within the same block pair (with one marker belonging to each block in the later case). The reliability of the new LD measure is evidenced by the observed robustness of the block structures derived from it using different SNP subsets. The new measure is not restricted to binary alleles and can therefore be applied to polymorphic markers, such as microsattelites, and to amino acid sequence data.

Distance-based hierarchical partitioning method, `DHPBlocker`. We have proposed a distance-based hierarchical partitioning method, `DHPBlocker`, to identify haplotype blocks. The three main components of `DHPBlocker`, distance, partition, and hierarchy, are summarized in Box 3. In the particular implementation of `DHPBlocker` considered here, we used as distance matrix the R matrix of LD measures, which is based on the correlation distance between marker-specific L_j matrices of shared haplotype lengths (Section 2.1). The partitioning component seeks the partition of markers into blocks (or clusters) that maximizes the mean silhouette width as the objective function (Section 2.2.2), using a deletion/subsitution/addition (D/S/A) search algorithm (Section 2.2.3). The optimal number of blocks is chosen based on the MDL principle (Section 2.2.4). In order to uncover the hierarchical nature of haplotype blocks, a divisive hierarchical clustering approach is adopted, whereby the partitioning algorithm is applied recursively within blocks (Section 2.2.5).

Note that our proposed distance-based hierarchical partitioning clustering approach is very flexible, in the sense that investigators can incorporate subject matter knowledge in choosing: (i) the matrix of pairwise distances between markers; (ii) the objective function for the partitioning algorithm; (iii) the optimization algorithm; (iv) the criterion for selecting an optimal number of blocks. In this article, the particular choices for each of the components described above were made specifically for the analysis of binary SNP data, which motivated `DHPBlocker`. In the case of more polymorphic markers, other choices may be preferrable. Moreover, `DHPBlocker` is a geneal clustring method that can be applied to address other problems.

Box 3. `DHPBlocker` Components.

Distance: A matrix of pairwise distances between markers, e.g., LD measures R, $1 - r^2$, $1 - D'$.

Partition: A distance-based ordered partitioning method.

- Objective function: e.g., mean silhouette width, median silhouette width.
- Optimization algorithm: e.g., D/S/A algorithm, steepest descent.
- Criterion for selecting number of blocks: e.g., MDL principle, mean silhouette width, median silhouette width, mean split silhouette (MSS) (Pollard and van der Laan, 2002).

Hierarchy: A divisive hierarchical clustering method, i.e., recursive application of the partitioning algorithm to blocks identified from previous steps.

Application of `DHPBlocker` to SNP data from the human 5q31 region and the human MHC region yielded block sructures that were highly consistent with previously published results. Furthermore, these analyses suggest that the block structure identified by `DHPBlocker` is robust to the selection of markers, in terms of number of markers and marker allele frequencies. An analysis based on bootstrapped chromosome samples also suggested robustness of the block defining procedure (results not shown).

Ongoing work includes the application of the R measure and the `DHPBlocker` procedure to investigate the block structure in the highly polymorphic HLA region using microsatellite and amino acid sequence data. We are also interested in between population comparisons of the hierarchical block structures revealed by `DHPBlocker`.

Acknowledgements. We would like to thank Mark van der Laan for insightful discussions on D/S/A algorithms and cluster analysis. We are also grateful to Eric Anderson, Glenys Thomson, Montgomery Slatkin, Bin Yu, and Kristie Mather for stimulating exchanges and constructive comments on an earlier version of the manuscript.

Software. The new LD measure R and `DHPBlocker` algorithm will be available in an R package at the time of publication of this manuscript.

References

E. C. Anderson and J. Novembre. Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics*, 73:336–354, 2003.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2):229–232, 2001.

E. Dawson, G. R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D. M. Beare, J. Pabial, T. Dibling, E. Tinsley, S. Kirby, D. Carter, M. Papaspyridonos, S. Livingstone, R. Ganske, E. Lohmussaar, J. Zernant, N. Tonisson, M. Remm, R. Magi, T. Puurand, J. Vilo, A. Kurg, K. Rice, P. Deloukas, R. Mott, A. Metspalu, D. R. Bentley, L. R. Cardon, and I. Dunham. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418:544–548, 2002.

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

L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12:921–927, 1995.

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

M. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96:746–774, 2001.

P. W. Hedrick. Gametic disequilibrium measures: Proceed with caution. *Genetics*, 117:331–341, 1987.

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

L. Kaufman and P. J. Rousseeuw. *Finding group in data: An introduction to cluster analysis*. John Wiley & Sons, 1990.

M. Koivisto, M. Perola, T. Varilo, W. Hennah, J. Ekelund, M. Lukk, L. Peltonen, E. Ukkonen, and H. Mannila. An mdl method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. *Pacific Symposium on Biocomputing*, 8:502–513, 2003.

T. C. M. Lee. An introduction to coding theory and the two-part minimum description length principle. *International statistical review*, 69:169–183, 2001.

N. E. Morton, W. Zhang, P. Taillon-Miller, S. Ennis, P. Y. Kwok, and A. Collins. The optimal measure of allelic association. *Proceedings of the National Academy of Science*, 98:5217–5221, 2001.

T. Niu, Z. S. Qin, X. Xu, and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157–169, 2002.

N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294:1719–1723, 2001.

M. A. Pericak-Vance. Linkage disequilibrium and allelic association. In J. L. Haines and M. A. Pericak-Vance, editors, *Approaches to gene mapping in complex human diseases*, chapter 15. Wiley-Liss, 1998.

M. S. Phillips, R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding, M. A. Donaldson, J. F. Studebaker, W. M. Ankener, S. V. Alfisi, F.-S. Kuo, A. L. Camisa, V. Pazorov, K. E. Scott, B. J. Carey, J. Faith, G. Katari, H. A. Bhatti, J. M. Cyr, V. Derohannessian, C. Elosua, A. M.

Forman, N. M. Grecco, C. R. Hock, J. M. Kuebler, J. A. Lathrop, M. A. Mockler, E. P. Nachtman, S. L. Restine, S. A. Varde, M. J. Hozza, C. A. Gelfand, J. Broxholme, G. R. Abecasis, M. T. Boyce-Jacino, and L. R. Cardon. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genetics*, 33:382–387, 2003.

K. S. Pollard and M. J. van der Laan. A method to identify significant clusters in gene expression data. In *Proceedings of SCI*, volume II, pages 318–325, 2002.

J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, 69:1–14, 2001.

J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

J. Rissanen. Stochastic complexity in statistical inquiry. In *Series in computer science*, volume 15. World Scientific, London, 1989.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

S. Sinisi and M. J. van der Laan. Loss-based cross-validated deletion/substitution/addition algorithms in estimation. Technical Report 143, Division of Biostatistics, U.C. Berkeley, 2004.

M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73:1162–9, 2003.

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction. *American Journal of Human Genetics*, 68:978–989, 2001.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17:520–525, 2001.

J. D. Wall and J. K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4:587–597, 2003.

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

B. S. Weir. *Genetic Data Analysis II.* Sinauer Associates, 1996.

E. P. Xing, R. Sharan, and M. I. Jordan. Bayesian haplotype inference via the dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*, volume II, 2004.

K. Zhang, M. Deng, T Chen, M. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Science*, 99:7335–7339, 2002a.

W. Zhang, A. Collins, N. Maniatis, W. Tapper, and N. E. Morton. Properties of linkage disequilibrium (ld) maps. *Proceedings of the National Academy of Science*, 99:17004–17007, 2002b.

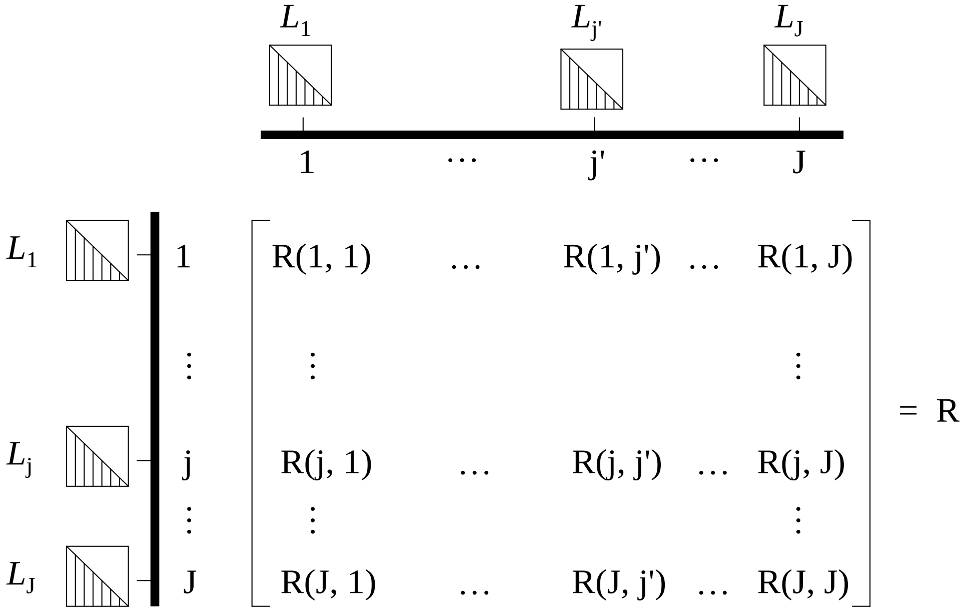
L_1
$L_{j'}$
L_J
1
…
j'
…
J
L_1
1
R(1, 1) … R(1, j') … R(1, J)
⋮
L_j
j
R(j, 1) … R(j, j') … R(j, J)
⋮
L_J
J
R(J, 1) … R(J, j') … R(J, J)
= R

Figure 1: *Construction of the R matrix of LD measures based on the L matrices of shared haplotype lengths.* The matrices L_j are $n \times n$ matrices of marker-specific shared haplotype lengths among n chromosomes. Entry $R(j, j')$ of the $J \times J$ matrix R is a measure of LD for markers j and j', based on a distance measure between the lower-left triangles of matrices L_j and L'_j.

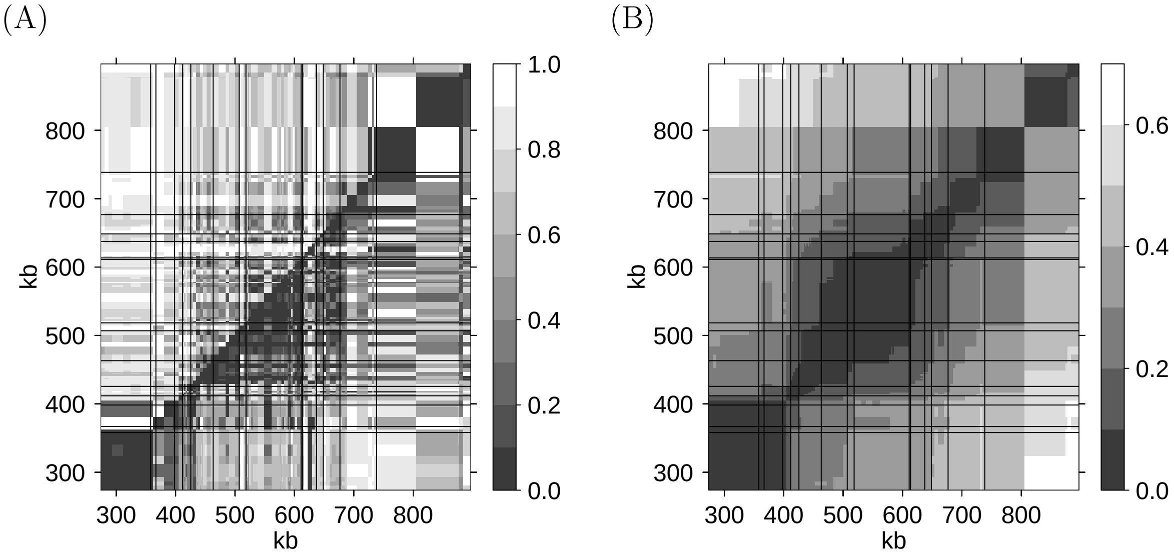

Figure 2: *LD patterns in the 5q31 region.* Visualization of LD patterns by: (A) two classical LD measures, $1 - r^2$ (upper-left triangle) and $1 - |D'|$ (lower-right triangle), and (B) the new R measure. Darker shades of gray correspond to higher LD (i.e., lower $1-r^2$, $1-D'$, and R values). All measures are computed based on unimputed data. The block boundaries defined by Daly et al. (2001) are shown by the solid lines. Markers are displayed using their physical distance.

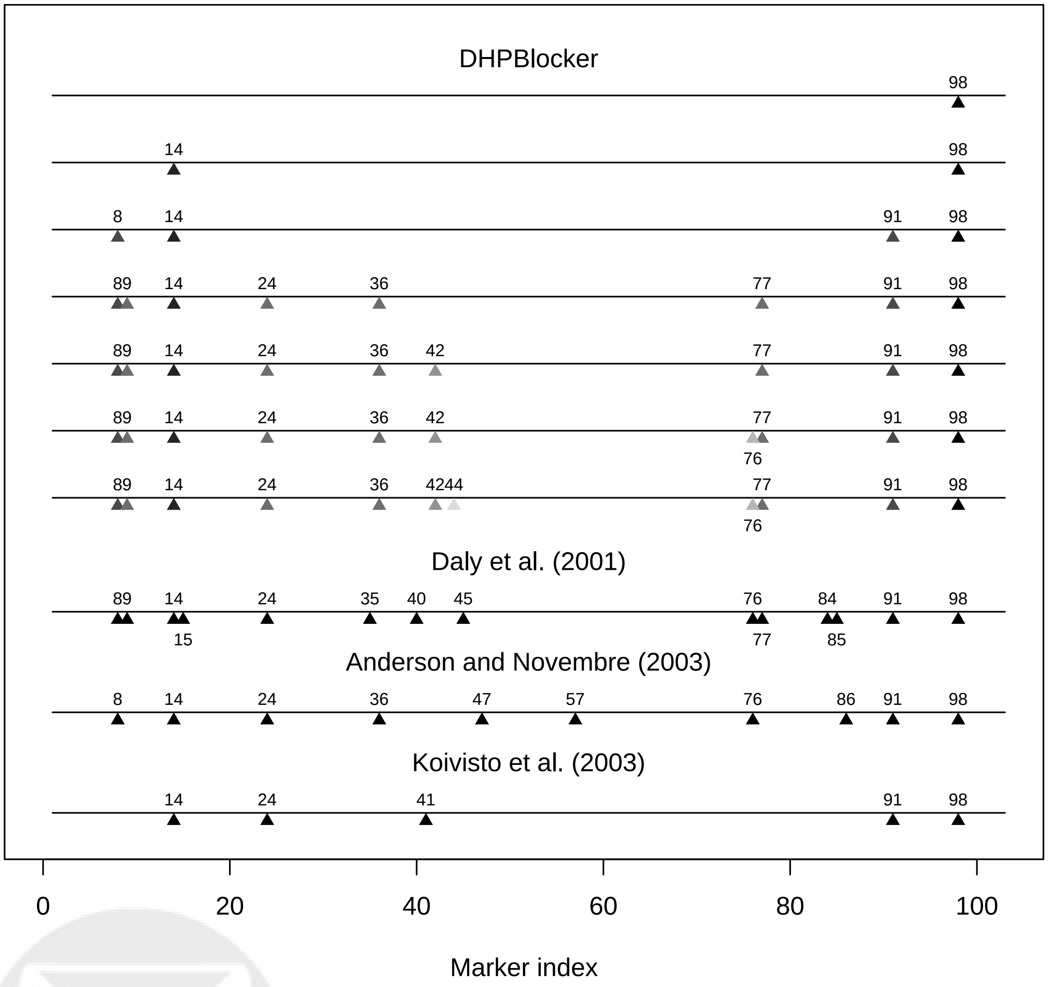

Figure 3: *Block structures in the 5q31 region identified by four methods.* For `DHPBlocker`, missing data were imputed by the KNN method, with $K = 5$ neighbors. The hierarchical structure is displayed for each level identified sequentially by the clustering procedure. Block boundaries are labelled by triangular plotting symbols, corresponding to the index of the rightmost marker of the block to the left. Different shades of gray are used for the plotting symbols in order to summarize the block hierarchy: the darker the plotting symbol, the earlier the corresponding block boundary was identified by `DHPBlocker`.

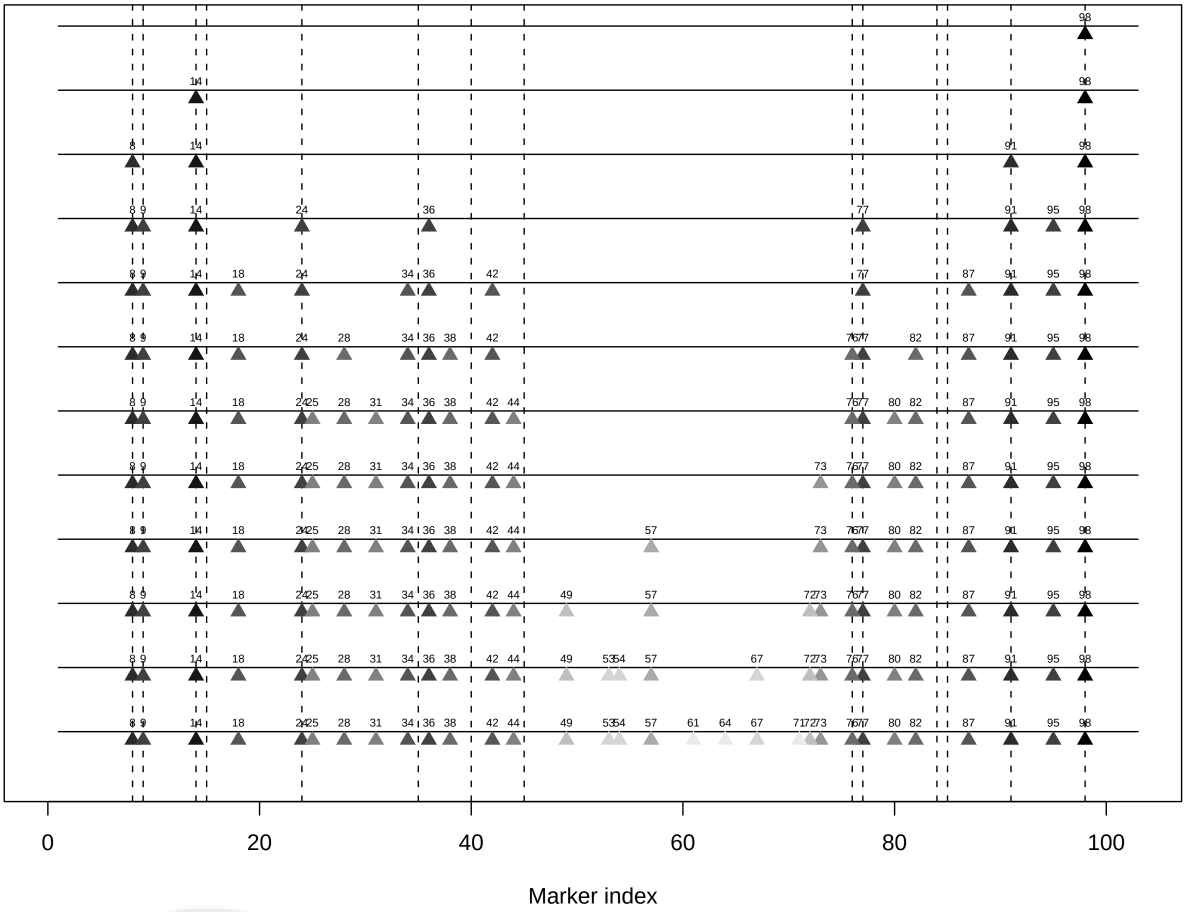

Figure 4: *Block structures in the 5q31 region identified by* `DHPBlocker` *when missing data were imputed at random.* The hierarchical structure is displayed for each level identified sequentially by the clustering procedure. Block boundaries are labelled by triangular plotting symbols, corresponding to the index of the rightmost marker of the block to the left. Different shades of gray are used for the plotting symbols in order to summarize the block hierarchy: the darker the plotting symbol, the earlier the corresponding block boundary was identified by `DHPBlocker`. The dashed vertical lines indicate the block boundaries defined by Daly et al. (2001).

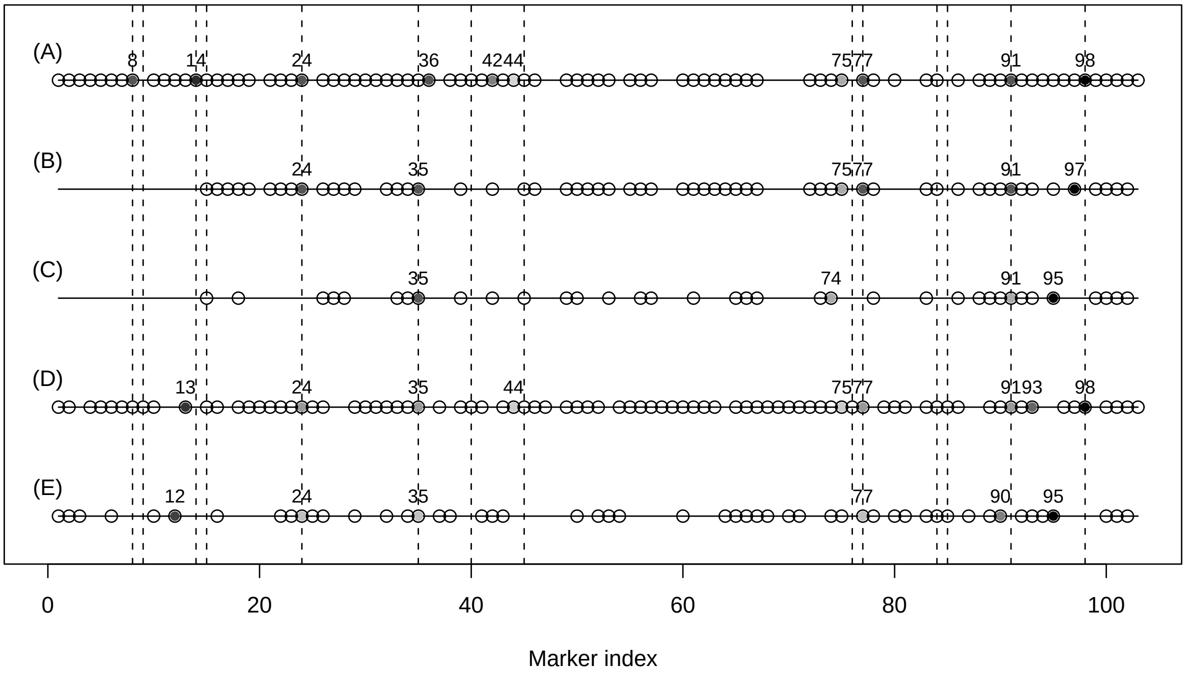

Figure 5: *Block structures in the 5q31 region identified by* `DHPBlocker` *using different subsets of SNPs.* From top to bottom, the SNP subsets are: (A) SNPs with minor allele frequencies greater than 10%; (B) SNPs with minor allele frequencies greater than 20%; (C) SNPs with minor allele frequencies greater than 30%; (D) a random subset of 80% of the original SNPs; (E) a random subset of 50% of the original SNPs. The selected SNP subsets are indicated by circles. Missing data were imputed by the KNN method, with $K = 5$ neighbors. Different shades of gray are used for the plotting symbols in order to summarize the block hierarchy: the darker the plotting symbol, the earlier the corresponding block boundary was identified by `DHPBlocker`. The dashed vertical lines indicate the block boundaries defined by Daly et al. (2001).

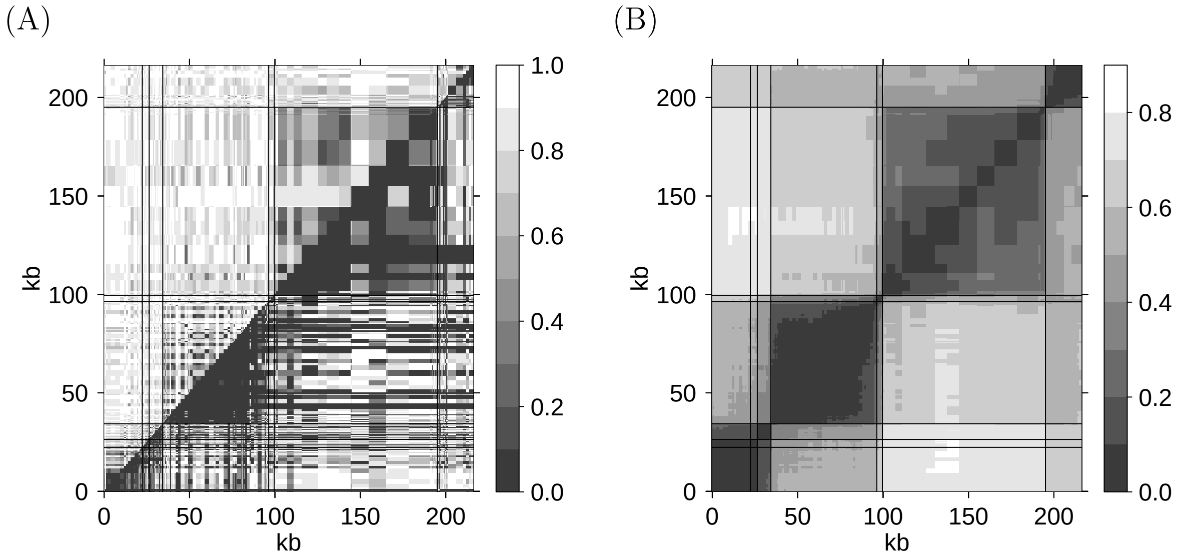

Figure 6: *LD patterns in the class II region of the MHC.* Visualization of LD patterns by: (A) two classical LD measures, $1 - r^2$ (upper-left triangle) and $1 - |D'|$ (lower-right triangle), and (B) the new R measure. Darker shades of gray correspond to higher LD (i.e., lower $1 - r^2$, $1 - D'$, and R values). The horizontal and vertical lines indicate the approximate centers of the DNA1–3, DMB1–2, and TAP2 recombination hotspots. All measures are computed based on phased data inferred by the `PHASE` program. Markers are displayed using their physical distance from the leftmost marker.

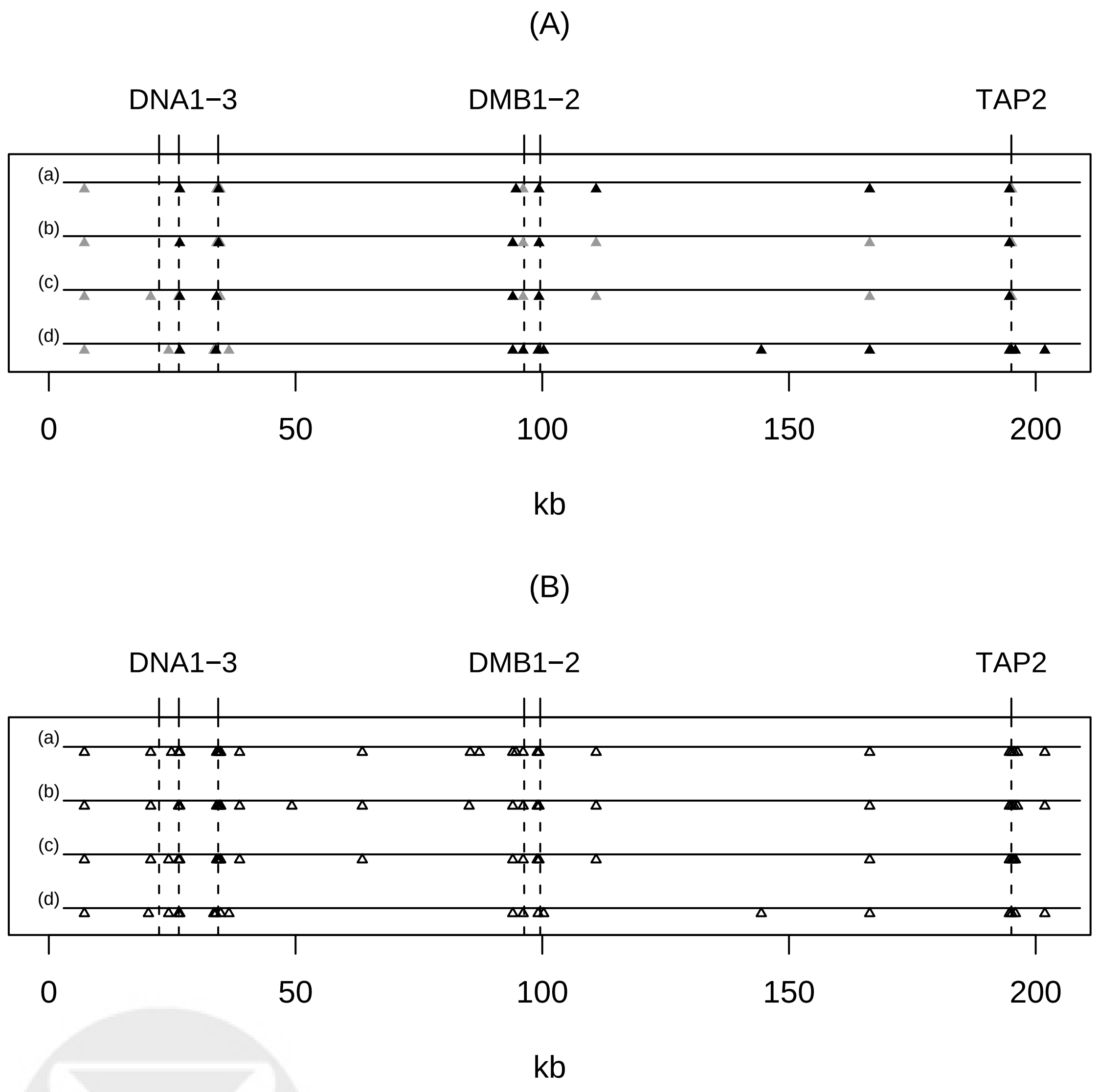

Figure 7: *Block structures in the class II region of the MHC identified by `DHPBlocker` using different subsets of SNPs.* (A) Only the first two levels of the block hierarchy are shown. The boundaries at the highest level are indicated by black triangles, those at the next level by gray triangles. (B) The overall block structure is shown without displaying the depth. From top to bottom, the SNP subsets are those with minor allele frequencies: (a) > 0%; (b) > 5%; (c) > 10%; (d) > 20%. The dashed vertical lines indicate the approximate centers of the DNA1–3, DMB1–2, and TAP2 recombination hotspots.