Characterization of Multilocus Linkage Disequilibrium

Alessandro Rinaldo,[1] Silviu-Alin Bacanu,[2] B. Devlin,[2] Vibhor Sonpar,[2] Larry Wasserman,[1] and Kathryn Roeder[1]*

[1]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania*
[2]*Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania*

Linkage disequilibrium (LD) in the human genome, often measured as pairwise correlation between adjacent markers, shows substantial spatial heterogeneity. Congruent with these results, studies have found that certain regions of the genome have far less haplotype diversity than expected if the alleles at multiple markers were independent, while other sets of adjacent markers behave almost independently. Regions with limited haplotype diversity have been described as ''blocked'' or ''haplotype blocks.'' In this article, we propose a new method that aims to distinguish between blocked and unblocked regions in the genome. Like some other approaches, the method analyses haplotype diversity. Unlike other methods, it allows for adjacent, distinct blocks and also multiple, independent single nucleotide polymorphisms (SNPs) separating blocks. Based on an approximate likelihood model and a parsimony criterion to penalize for model complexity, the method partitions a genomic region into blocks relatively quickly, and simulations suggest that its partitions are accurate. We also propose a new, efficient method to select SNPs for association analysis, namely tag SNPs. These methods compare favorably to similar blocking and tagging methods using simulations. *Genet. Epidemiol.* © 2005 Wiley-Liss, Inc.

Key words: ancestry; haplotype block; information theory; model selection; tag SNP

Contract grant sponsor: NIH; Contract grant number: MH057881.
*Correspondence to: Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.
E-mail: roeder@stat.cmu.edu
Received 25 June 2004; Accepted 3 October 2004
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20056

INTRODUCTION

Recent studies of linkage disequilibrium (LD) in human genome [Daly et al., 2001; Reich et al., 2001; Patil et al., 2001; Abecasis et al., 2001; Gabriel et al., 2002; Clark et al., 2003; McVean et al., 2004] have demonstrated definitively what earlier studies seemed to suggest [Jorde, 1995, 2000], that the structure of LD in the genome is highly idiosyncratic. Alleles at polymorphisms separated by as little as a few base pairs can exhibit complete LD or none at all, and this is true regardless of the population examined [Crawford et al., 2004]. When multilocus LD is computed in terms of haplotype distributions, LD is also unpredictable. Certain genomic regions show a diversity of haplotypes close to that expected if alleles at proximate loci were independent, while others show a striking dearth of haplotypes (i.e., high LD). Regions of high LD have been dubbed haplotype blocks. Unfortunately, because LD can be defined in many ways, and because LD is itself the result of stochastic processes, the meaning and extent of blocks depend on how we define them [Cardon and Abecasis, 2003].

Given the complexity of human evolution, including migration, gene flow, bottlenecks, and selection, it is unsurprising that the extent of haplotype blocks varies among populations. For example, haplotype blocks appear to be much more extensive in peoples of European ancestry than they are in peoples of African origin. The origin of haplotype blocks, in general, is unclear. They could arise as the result of hot and cold spots of recombination in the genome. Indeed, that appears to explain the LD structure of the HLA region on chromosome 6p [Jeffreys et al., 2001, McVean et al., 2004], and other well-studied regions of the genome [Lien et al., 2000; May et al., 2002; Schneider et al., 2002]. On the other hand, it is well known that strictly random processes can generate clumping, and so random processes have been evoked to explain observed blocks [Phillips et al., 2003]. The truth must lie somewhere in between [McVean et al., 2004]. Data

from the HapMap [Gibbs et al., 2003] project should lend more insight into their origin.

Clearly, local patterns of LD in the genome can tell us much about evolution. Patterns of LD are also useful for deciphering the etiology of complex disease via association studies. It has long been noted that the level of LD in a candidate gene helps to determine the density of genetic markers required to detect association between a liability allele (LA) and phenotype. More recently, it has become apparent that strong irregularities in these patterns play a key role in the choice of which single nucleotide polymorphisms (SNPs) to genotype [Meng et al., 2003]. Within a haplotype block, the correlation between neighboring SNPs may be sufficiently high to ensure that one or two ''tag SNPs'' describe the common haplotypes over a substantial region. Alternatively, in other regions, neighboring SNPs may be essentially uncorrelated and thus careful study of the region requires genotyping of nearly all the SNPs in these intervals. Even the position of a LA within the correlation structure of a region can help to determine the optimal method of data analysis [e.g., Roeder et al., 2005]. For instance, if the LA is located within a haplotype block, methods that exploit the pattern of the signal within a block can be more powerful than others that ignore the spatial pattern of the markers. Alternatively, methods that target spatial signals fare poorly if the LA occurs within a recombination hotspot. Clearly, statistical methods that explain the local pattern of LD and carefully select tag SNPs are critical to successful studies of genetic epidemiology.

BLOCKING METHODS

Numerous methods for finding haplotype blocks have been proposed thus far. These methods can be loosely classified based upon their objective. (1) Methods designed to minimize haplotype diversity within a block. These include the method of Zhang et al. [2002a], which is based on minimizing a measure of haplotype diversity, and the method of Patil et al. [2001], which chooses blocks with an aim to limit the number of so-called haplotype tagging SNPs (htSNPs) [Johnson et al., 2001] required to identify the most commonly observed haplotypes in a sample. These two methods differ in their implementations: a greedy algorithm [Patil et al., 2001] versus a dynamic programming algorithm [Zhang et al., 2002a, 2004]. (2) Methods designed to locate recombinational hotspots, such as the method devised by Wang et al. [2002], which uses an adaptation of the four-gamete-test (FGT) [Hudson and Kaplan, 1985]. Also in this category are the methods based on the minimum-description-length principle [Anderson and Novembre, 2003; Koivisto et al., 2003]. This principle, while often explained in terms of coding theory, is essentially equivalent to a model selection method, such as the Aikake Information criterion (AIC) and the Bayesian Information criterion (BIC) [Rissanen, 1989]. These approaches seek a model for the data that maximizes the likelihood of the data, subject to a penalty for model complexity. (3) Methods designed to locate only those blocks with sufficient LD to be almost irrefutable. This includes the method of Gabriel et al. [2002], which assesses the strength of the pairwise LD of all loci within a proposed block.

Gabriel et al.'s [2002] approach seems true to nature in the sense that they aim to find blocks in some regions of the genome but no blocks in others, and that the unblocked portion of the genome can be substantial. Their results are corroborated by Wall and Pritchard [2003] and McVean et al. [2004]. By comparing simulated data and existing SNP samples, Wall and Pritchard [2003] conclude that the recombination rate heterogeneity must be extremely pronounced to produce realistic data, and a substantial portion of the genome is not contained in blocks. This is contrary to most haplotype blocking routines, many of which assume that all sequence should be assigned to a block. On the other hand, the method of Gabriel et al. [2002] depends on confidence intervals for pairwise LD and is, therefore, inherently highly dependent on the sample size. With greater sample sizes, the precision of the confidence intervals will increase, more pairs of markers will be in ''strong LD'' or will exhibit ''historical evidence of recombination,'' and consequently the coverage of sequence in blocks can change, possibly substantially based upon the sample size. For small sample sizes, confidence intervals can be quite wide and consequently very few pairwise comparisons will be categorized definitively.

In this article, we propose a new haplotype blocking method, called EB for Entropy Blocker, that aims to identify only those blocks that exhibit substantial multilocus disequilibrium and range over a greater number of SNPs. In addition, although discovered blocks may be physically adjacent, EB allows for one or more SNPs separating blocks. For this reason, EB can be

classified in category (3), like that of Gabriel et al. [2002]. As a statistical procedure, however, EB is most similar to the minimum-description blocking procedure (MDB) of Anderson and Novembre [2003]. Like the MDB approach, we evaluate a model by computing the likelihood of the data, minus a penalty for model complexity. The best models are defined as those that describe the data well, but require few parameters to do so. In particular, a model is favored if it selects a blocking structure that satisfies both of the usual criterion for blocks: the resulting blocks have very low diversity and the LD with SNPs outside the block is low. By design, this method is likely to perform well for modest sized samples. For this reason, we see it as a complement to the Gabriel et al. [2002] method.

TAGGING METHODS

To avoid genotyping all the SNPs in a region, several approaches have been developed for selecting a subset of highly informative SNPs. These are often called haplotype tagging SNPs (htSNPs). A good set of htSNPs aims to predict the existing haplotypes in the region [Zhang et al., 2002a; Ackerman et al., 2003; Ke and Cardon, 2003; Meng et al., 2003; Sebastiani et al., 2003]. Several approaches appeared almost simultaneously with algorithms directed at this purpose. Chapman et al. [2003] and Stram et al. [2003] independently proposed methods to assess the coefficient of determination, or R^2, for predicting an individuals[1] pair of haplotypes from the multilocus genotypes of the htSNPs assessed. A stepwise selection technique is suggested for selecting the htSNPs that achieve a preset R^2 across the region under investigation with a minimum of SNPs. While the motivation for these approaches is commendable, the algorithm is quite sensitive to the level of correlation in the gene and other features.

Rather than choosing the SNPs to predict haplotypes, the goal might be to choose a subset of tag SNPs that successfully predict unmeasured SNPs. Carlson et al. [2004] developed a simple algorithm based on the r^2 measure of LD [Devlin and Risch, 1995]. Their algorithm ensures that for any SNP not measured there exists a SNP in the tag SNP set that has an r^2 of at least a preset value η. Although this method performs quite well, it does have one shortcoming. It requires the imputation of haplotype phase for each pair of SNPs in the region to compute r^2. We develop an alternative based on hierarchical clustering procedures, which we call H-clust. This simple method can be implemented with standard statistical software, and does not require haplotypes. By comparing the H-clust method to the r^2 method of Carlson et al. using simulations, we show that the two are virtually indistinguishable even though H-clust analyses genotype data and the r^2 method is based on known haplotypes (i.e., no haplotype uncertainty).

METHODS

DESCRIPTIVE TOOLS

Let S be the number of SNPs in the genomic region of interest. Each SNP is modeled as a Bernoulli random variable for which $p_i \geq 0.5$ denotes the probability of observing the major allele and $1 - p_i$ is the probability of observing the minor allele in the i'th SNP. The corresponding haplotype is an S-dimensional binary random vector. The haplotypes encode the joint distribution of the SNPs.

Pairwise LD measures such as r^2 or $|r|$ [Devlin and Risch, 1995] are frequently displayed in a matrix to illustrate multilocus LD structure. Although such a representation is useful, it cannot directly capture the full multilocus associations of a genomic region. Nothnagel et al. [2003] suggested that higher level association can be captured by the entropy of the relative frequencies of the multilocus haplotypes. (Consistent with statistical usage, we will call this set of relative frequencies the ''haplotype distribution.'') The intuition is as follows: when the haplotype distribution for a set of consecutive SNPs has low entropy, it is usually deviating substantially from randomness and hence the set of SNPs likely comprise a block.

To discover block structure, we investigate sets of $m \leq S$ consecutive SNPs. A haplotype consisting of m SNPs can potentially take on 2^m forms. For notational simplicity, we define $T \equiv 2^m$, where m is implicitly understood. For any haplotype distribution consisting of m SNPs taking on $h = 1, \ldots, T$ distinct forms with probabilities $q_1, \ldots, q_T$, the *entropy* is defined as $H = -\sum_{h=1}^{T} q_h \log_2(q_h)$. Under the assumption of linkage equilibrium, the entropy H_E is computed using the equilibrium haplotype frequencies, $q_h = \prod_{j=1}^{m} p_j^{I_h^j} (1 - p_j)^{1 - I_h^j}$, where I_h^j is an indicator function that assumes the value 1 if in the haplotype indexed by h the j-th SNP exhibits its common type and 0 otherwise. The standardized entropy is computed as

$\epsilon = \frac{H_E - H}{H_E}$. This measure varies from zero, for perfect equilibrium, to close to $(m-1)/m$, for a haplotype block consisting largely of a single haplotype. Heuristically, the measure can be used to detect genomic regions with unusually low entropy, as follows: (1) Compute ϵ for blocks

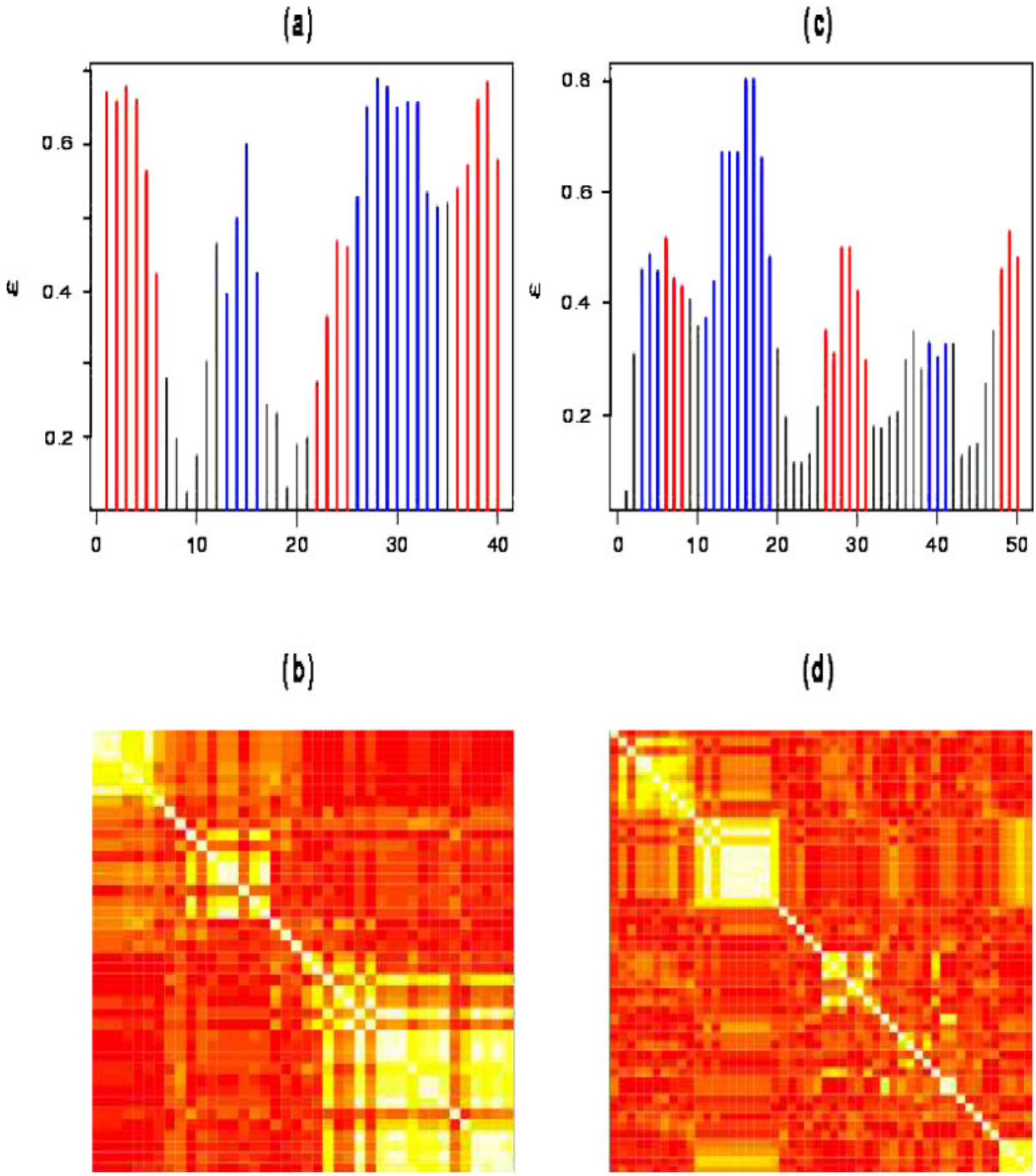

Fig. 1. Standardized Entropy and pairwise LD ($|r|$) for two genes from the SeattleSNP database: IL21R, the interleukin 21 receptor (a,b), and JAK3, Janus kinase 3 (c,d). Red and blue bars in the entropy plots indicate haplotype blocks, and black bars indicate unblocked SNPs. The absolute value of the pairwise correlation ($|r|$) between SNPs is depicted with a color gradient: white/yellow is high and red is low.

consisting of m adjacent SNPs. (2) Plot ϵ for a moving window across the gene (Fig. 1a,c). At the boundaries of the genotyped regions, the entropy is computed using a smaller window. (3) Search for regions for which ϵ is closer to 1. These identify SNPs that possess high multivariate correlation and are more likely to be part of a block.

MODEL SELECTION

Figure 1 provides a heuristic way of locating potential haplotype blocks. Still we need a method for comparing competing partitions of the SNPs into blocked and non-blocked regions. The proposed EB algorithm is based on a working model that treats blocks as mutually independent and allows for sequences of independent SNPs. Partitions of the sequence into blocked and unblocked regions are considered as descriptive models of the genomic region. The favored partition is selected using an AIC model selection procedure that attempts to balance the competing criterion of model fit and parsimony [Rissanen, 1989]. In reality, SNPs in unblocked regions are seldom independent. Indeed, they often possess a notable level of correlation; however, the strength of correlation in unblocked regions it typically far lower than that observed in the blocked regions. Consequently, although the working model is an oversimplification of reality, it provides an effective, parsimonious approximation to reality that allows us to discover the principal block structure in the region.

Formally, let N be the number of sampled haplotypes and S be the number of SNPs, indexed in increasing order according to their positions in the genomic region. The frequency of haplotypes of each possible kind form a T-dimensional random vector $\mathbf{z} = (z_1, \ldots, z_T)$, which is assumed to follow a Multinomial distribution with log probability mass function:

$$\log f(z_1, \ldots, z_T) = \log\left\{\frac{N!}{z_1!...z_T!}\right\} + \sum_{h=1}^{T} z_h \log q_h. \qquad (1)$$

For a given partition P of $\{1, \ldots, S\}$, let $I \subset \{1, ..., S\}$ be the set containing the indexes for the SNPs assumed to be mutually independent and M be the number of disjoint blocks in the partition P. Each block can be identified by a consecutive index set b_l such that $b_l \subseteq \{1, \ldots, S\} - I$ and $b_l \cap b_r = \emptyset$, $l, r = 1, \ldots, M$. Clearly, $\{1, \ldots, S\} = I \cup (\bigcup_{l=1}^{M} b_l)$. The blocks b_l are assumed to be mutually independent and independent from the SNPs in the set I. Therefore, the probability of each possible haplotype sequence h, $h = 1, \ldots, T$, is equal to

$$q_h = \prod_{i \in I} p_i^{I_h^i} (1 - p_i)^{1 - I_h^i} \prod_{l=1}^{M} p_h^{b_l}, \qquad (2)$$

in which $p_h^{b_l}$ is the probability of observing haplotype h in the subregion b_l. In reality the multinomial in (1) is often extremely sparse and $z_h = 0$ for all but k categories. Consequently, we reduce $\mathbf{z}$ to $(z_1, z_2, \ldots, z_k)$, where k is the number of categories with count of at least one.

If haplotypes are measured directly, then calculation of the AIC requires only an estimate of the parameters in the probability model and a penalty for model dimension. These can be computed as follows. (1) Let $|A|$ denote the cardinality of a set A. Within each block b_l, the number of distinct, observed haplotypes is $|b_l| = k_l \leq k$. (2) The parameters θ in the likelihood model are $\{p_i\}$ for $i \in I$ and $\{p_h^{b_l}\}$ for $h = 1, \ldots, k_l$ and $l = 1, \ldots, M$. (3) Let x_i be the sample frequency of the common allele for SNP i. The maximum likelihood estimator (MLE) for p_i, $i \in I$, is $\hat{p}_i = x_i/N$. (4) Let $x_h^{b_l}$ be the sample frequency of haplotypes of type h in block b_l. The MLE for $p_h^{b_l}$ is $\hat{p}_h^{b_l} = x_h^{b_l}/N$. (5) Because the values of k and $(z_1, \ldots, z_k)$ do not depend on the specific partition $\mathcal{P}$, but on the whole region of interest, the term $N!/z_1!...z_k!$ in (1) is constant across partitions and does not enter into the calculation of the AIC. (6) From (1) and (2), it follows that the log-likelihood $\mathcal{L}(\hat{\theta})$, evaluated at the MLE of the parameters in the likelihood model, equals $\sum_h z_h \log \hat{q}_h$. This quantity can be re-expressed as a function of the sufficient statistics for the model, $\{x_i\}$ and $\{x_h^{b_l}\}$:

$$\mathcal{L}(\hat{\theta}) = \sum_{i \in I} [x_i \log(\hat{p}_i) + (N - x_i) \log(1 - \hat{p}_i)] + \sum_{l=1}^{M} \left[\sum_{h=1}^{k_l} x_h^{b_l} \log(\hat{p}_h^{b_l})\right].$$

The AIC formula for a model with d-dimensional parameter space is: $-2\mathcal{L}(\hat{\theta}) + 2d$. For a given partition, the dimension of the model, d, is estimated as: $d = |I| + \sum_{l=1}^{M} [2^{(k_l - 1)} - 1]$. This data dependent estimate of model dimension is motivated below. For any given partition, the AIC of the corresponding model is computed and the model with minimal AIC is retained.

Typically genotypes, not haplotypes, are measured. Although it would be ideal to have a genotype-based blocking algorithm, it is possible to utilize the haplotype-based method with some modifications. Note that the input for the EB algorithm is the frequency of haplotypes, rather than the particular phase for indiviual genotypes. The former quantity can be estimated much more reliably than the latter. For example, the haplotype frequencies ($p_h^{b_l}$) can be estimated with an algorithm such as PHASE [Stephens et al., 2001, Stephens and Donnelly, 2003] or HAPLOTYPER [Niu et al., 2002]. It is not difficult to infer the haplotype frequencies within a block because there is limited haplotype diversity. Outside the blocks, the haplotypes are more challenging to estimate, but the EB algorithm will still function well because it is robust to some errors in the haplotype frequencies when the true state of nature is unblocked. Provided the algorithm infers that there are a large number of distinct haplotypes present in this region, then the region will be properly classified.

If genotypes, not haplotypes, are measured, then we modify the EB algorithm as follows. We approximate the haplotype count, $x_h^{b_l}$, by using the estimated haplotype frequencies $N\hat{p}_h^{b_l}$. A complication occurs when estimating k. Many algorithms designed to estimate haplotype frequencies assign small probabilities to numerous haplotype categories that are consistent with the observations, but the haplotypes do not exist in the sample. To handle this problem, we suggest trimming the count to those haplotypes that account for the bulk of the probability mass in the estimated haplotype distribution, say 95%. For instance, Daly et al. [2001] support the haplotype blocks in their sample by illustrating the number of distinct haplotype forms necessary to obtain approximately 90% of the observations in each block [see also Patil et al., 2001].

MEASURE OF MODEL COMPLEXITY IN THE AIC

The parameter d in the AIC formula measures model complexity and is crucial to obtain meaningful results. If the penalty for model complexity describing the blocks is too harsh, then the algorithm will not include any blocks. Likewise, if the penalty is too lenient, then the algorithm will choose a single inclusive block. Traditionally, for the two major approaches to model selection, AIC and BIC, the penalty is a function of the dimension of the parameter space. Statistically, the dimension of a multinomial on $\{0,1\}^m$ is $2^m - 1$, which is typically much greater than N. Furthermore, given the strength of the LD between SNPs within a block, the full model space is unlikely to be realized even if $N \to \infty$. Therefore, a different, biologically relevant, description of the model dimension is required. Ideally, the model dimension should reflect a priori knowledge about the nature of the evolution of haplotypes in the population.

The proposed specification relates d to k_l via an evolutionay argument. The dimension is chosen to be the minimal number of mutations or branches in an evolutionary tree required to produce a sample of k_l haplotypes in the absence of recombination. Next, because the polymorphisms are binary, $2^{(k_l-1)} - 1$ is a natural measure of the dimension of the resulting multinomial for the block. For the actual observed data, the required number of branches and mutations could be higher, or lower, with recombination. We estimate the dimension of the model using only the count of distinct haplotype forms.

HEURISTIC FOR HOW THE BLOCKING ALGORITHM WORKS

To understand how EB discovers block structure, consider splitting the last SNP from a block called A and add this SNP to the set of singleton SNPs. If this split does not collapse any of the existing categories in A, then the log likelihood adds a term of the form $x \log \hat{p} + (N - x) \log(1 - \hat{p})$ and d increases by one. Consequently, the AIC score for this model is higher than the original model and the block will not be split. On the other hand, if the SNP at the contested position is approximately independent of the SNPs within block A, then it is likely that one or more categories will collapse causing an overall decrease in the AIC. Specifically, for a collapse of categories (j, j'), terms like $(x_{Aj} + x_{Aj'}) \log(p_{Aj} + p_{Aj'})$ replace terms like $x_{Aj} \log p_{Aj} + x_{Aj'} \log p_{Aj'}$, causing an increase in the likelihood for the block. This increase is counterbalanced by the decrease in likelihood introduced by the new term for the SNP of the form $x \log \hat{p} + (N - x) \log(1 - \hat{p})$. Overall, the likelihood will either be unchanged, or more likely decrease. Overall, the AIC tends to decrease because the dimension of the model will decrease if one or more categories are combined. This heuristic explains why introducing a necessary split will decrease the AIC score. Overall, this

suggests that an AIC criterion will identify the key block structure, provided the working model is approximately correct.

Another difficulty is how to approach this massive computational problem. It is clearly not possible to examine all possible block solutions. To visualize the data, we compute the standardized entropy of consecutive sets of 5 SNPs and plot this quantity versus the physical location of the central marker in the window. The peaks in the plot indicate the location and length of blocks. The dips are suggestive of areas of historical recombination (Fig. 1).

The number of ways of partitioning the region of interest into blocks and singleton SNPs grows extremely fast as the dimension S of the region of interest increases. The complexity of the searching procedure was reduced by designing a greedy iterative algorithm that searches only for partitions corresponding to regions with high standardized entropy. See the Appendix for details on the implementation of the algorithm.

TAGGING

Because SNPs in a block are highly correlated, complete genotyping of all SNPs in a gene is often not practical. For a thorough analysis of the variation in a gene and its potential effects on a phenotype, one would want to genotype a subset of the highly correlated SNPs and all SNPs that are only weakly correlated with other SNPs. We describe how a simple clustering method can be used to rapidly identify the desired set of tag SNPs.

Clustering methods partition observed data into more homogeneous classes. Consequently, they have great potential as a tag SNP selection tool. Clustering methods can be applied directly to genotype data or they can be applied to linkage disequilibrium (LD) matrices derived by estimating the haplotype structure of the sample. Because estimating haplotype frequencies can be laborious and error prone, especially for a large number of SNPs, we favor selecting SNPs directly from the genotype data. Moreover, the clustering approach yields essentially the same set of SNPs as methods that use the haplotype structure. By basing the clustering methods on the squared correlation matrix of the genotype data, haplotype estimation can be eliminated and the process of selecting SNPs becomes near-instantaneous.

The proposed method, H-clust, consists of two stages. The first stage uses hierarchical clustering to determine the clusters. In the second stage, for each cluster, the method chooses the SNP most correlated with all the other SNPs in the cluster as the tag SNP for that cluster. Code the genotypes as 0, 1, 2 to denote the number of copies of a particular variant. Let $\mathbf{X}$ be the coded genotype matrix for the sample of multilocus genotypes, and let Σ be the squared correlation matrix associated with $\mathbf{X}$. Each entry in the matrix is the square of Pearson's correlation coefficient between allele counts at pairs of SNPs. It is worth noting that the numerator of this correlation measure is identical to the disequilibrium coefficient proposed by Weir [1979] and recently studied by Schaid [2004] and Zaykin [2004].

H-clust uses hierarchical clustering based on a dissimilarity matrix $\mathbf{D} = 1 - \Sigma$. Hierarchical clustering can be represented as a dendogram in which any two SNP groups diverge at a height that is a function of the dissimilarity between members of the two groups. The clusters are obtained by declaring SNPs to be in the same cluster when they are found in the same sub-tree below a preset value, say $1 - \eta$. Due to the relationship between height of divergence and the squared correlation between observations, the cut-off value η ensures that a minimum level of pairwise correlation exists between each unmeasured SNP and at least one tag SNP. The clustering method yields SNPs that are in multiple SNP clusters and SNPs that form clusters of size one. The next step chooses the representative tag SNP for each block. This is done by simply choosing the SNP that is most correlated with all the other SNPs in the block. If multiple SNPs show equal correlation, then the one in the middle is chosen as the tag SNP.

Hierarchical clustering can be achieved by multiple methods, such as complete, single, and average linkage [Kauffman and Rousseeuw, 1990]. We prefer the complete linkage method because it finds compact, spherical clusters. It iteratively merges the two clusters with the smallest maximum pairwise dissimilarity between any two of their members. The single linkage method finds larger clusters. It iteratively merges the two clusters with the smallest minimum pairwise dissimilarity between any two of their members. Average linkage can be regarded as yielding clusters with characteristics between the single and complete link methods. In all our simulations, we used the R/Splus hclust function with its default method, complete linkage

[Venables and Ripley, 1994; Kauffman and Rousseeuw, 1990].

RESULTS

EXAMPLES

The standardized entropy (with blocking) and pairwise correlation matrices ($|r|$) for the interleukin 21 receptor (IL21R) and Janus kinase 3 (JAK3) genes from the SeattleSNPs compilation, African-American sample (NHLBI Program for Genomic Applications; see Electronic-Database), are shown in Figure 1. Notice how the peaks in the standardized entropy correspond to the blocks of high pairwise correlation in the correlation matrices. From these two diagnostic plots, some of the block boundaries are quite obvious, while others are not. The solution found by the EB algorithm is depicted by the color coding of the standardized entropy plot: red and blue sections denote blocked intervals, while black denotes unblocked SNPs.

IL21R has two clear-cut blocks on the left-hand side and one somewhat ambiguous block directly adjacent to a large region of high correlation on the right-hand side of the matrix. The EB algorithm splits this large unit of high correlation at SNP 35. In the correlation matrix, this appears to be an outlier within a much larger block. In the standardized entropy plot, the split into two blocks appears to be supported by a dip in the level of standardized entropy. However, due to the use of a moving window of size 5, this dip could also be an artifact induced by an outlier SNP. Follow-up analysis without the suspect SNP supports this conjecture. The algorithm chooses to merge the two right-most blocks into a single large block.

The H-clust plot provides further information about patterns of LD in IL21R (Fig. 2), clearly identifying those SNPs in weak LD with all other common SNPs $\{7, 8, 11, 15, 18, 20, 21, 35\}$. With

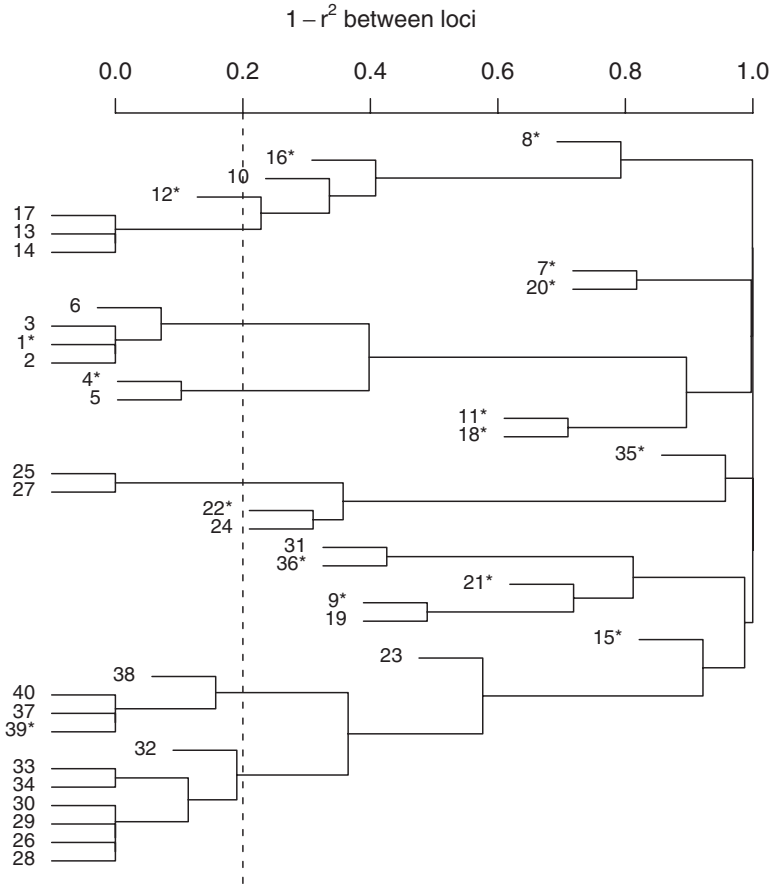

Fig. 2. Hierarchical clustering of common SNPs from the IL21R gene from the SeattleSNP database. The distance measure plotted on the vertical axis equals one minus Pearson's squared correlation coefficient.

$\eta = 0.50$, the algorithm selects 15 tag SNPs; those just listed, plus one from each cluster occurring below the $\eta = 0.5$ line, $\{2, 13, 19, 23, 25, 31, 34\}$. SNPs 1–6, which form the first block (Fig. 1), are clustered together (Fig. 2), and a single tag SNP is recommended. Within the region of strong LD extending from SNPs 26–40, only SNPs $\{27, 31, 35, 36\}$ are separated from a single large cluster. Of these, SNP 35 is clearly delineated as an outlier. This SNP does not fit in with the LD structure in the region, having Pearson's squared correlation coefficient less than 0.2 with all other common SNPs. Hence, it is included in any tagging scheme. Alternatively, SNP 27, which also creates an ''orange stripe'' in the pairwise correlation matrix (Fig. 1), clusters with some neighboring SNPs in the adjacent block $\{22, 24, 25\}$. Although tightly related, tag SNPs do not directly correspond to block structure. To cover the remaining region encompassing SNPs 22–34,36–40, which appears to consist of two slightly overlapping blocks, 3 tag SNPs are needed. SNP 25, selected from the cluster $\{22, 24, 25, 27\}$, appears to help tag two blocks.

For JAK3, the surprise in the blocking structure is the relatively strong standardized entropies displayed for the third and fourth blocks (Fig. 1). The presence of these blocks is barely evident in the correlation matrix.

PERFORMANCE OF BLOCKING METHODS EVALUATED BY SIMULATIONS

We compare the EB blocking algorithm with MDB using data with known block structure. The test structure consists of genomic regions with variable recombination rates. To generate haplotype blocks, we used a modification of Hudson's [2002] MS program [Wall and Pritchard, 2003] to produce ''hot spots'' and ''cold regions'' for recombination. For the first simulation, we simulated five cold regions, each of length 10 kb, separated by hot spots, each of length 1 kb. The mutation rate was chosen to be $\theta = 4 N_e \mu = 5.6 \times 10^{-4}$/bp, where μ is the per basepair, per generation, mutation rate, and $N_e = 10,000$ is the effective population size; μ was chosen to yield, on average, the number of common SNPs per kb typically observed in the SeattleSNP database. The recombination rate was chosen to be $r = 4 N_e \delta$, where $\delta = 2.5 \times 10^{-8}$ is the per generation, per basepair, recombination rate. This value was suggested in Nordborg and Tavare [2002]. The scaled recombination rate over the entire region was r times the length of the region, in basepairs. The per basepair rate varied so that it was R_h times greater in hot spots than in cold regions, for $R_h = 50, 100, 200$. One hundred data sets were produced for each scenario. Then, the resulting ''common SNPs'' (with minor allele frequency ≥ 0.10) defined the haplotypes. The resulting haplotypes had on average 12 SNPs per cold spot and 0–2 SNPs per hot spot.

The performance of the blocking algorithm was evaluated by two statistics. To determine if the block boundaries defining the hot and cold regions were discovered, we computed the fraction of times that these boundary points were approximately identified. If a hot spot includes no SNPs, then a single boundary exists between the SNPs in the adjacent cold spots. Otherwise, there are two boundaries, one to the left and right of the SNPs in the hot spot. The algorithm declares success if an estimated block boundary is located within the interval defined by the last SNP in the region to the left of a boundary and the first SNP to the right. We call this sensitivity because the boundaries must occur very near to a change in the recombination rate. One can think of this as a measure of the power to detect change points in the recombination rate. Alternatively, non-concordance aims to determine if the procedure frequently splits a cold spot into two or more blocks. It is computed as the fraction of edges discovered that are misplaced. This measure could be thought of as the false discovery rate. If the sensitivity and non-concordance are one and zero, respectively, the true blocks are discovered and these blocks are not split into sub-blocks.

In this simulation, the EB and MDB methods showed different comparative advantages

TABLE I. Sensitivity and non-concordance rates of two model selection procedures for detecting recombinational hotspots[a]

		Sensitivity		*Non-concordance*	
N	R_h	EB	MDB	EB	MDB
50	50	0.77	0.66	0.31	0.09
50	100	0.76	0.64	0.31	0.11
50	200	0.76	0.66	0.30	0.08
100	50	0.80	0.67	0.23	0.06
100	100	0.74	0.65	0.27	0.06
100	200	0.79	0.64	0.23	0.06

One hundred genes were simulated and analyzed by the proposed method (EB) and Anderson and Novembre's [2003] method (MDB). N is the number of haplotypes, per gene, and R_h, the multiplicative increase in recombinational rate over the background rate.

(Table I). EB is designed to discover interruptions in blocks and, hence, this method has a higher sensitivity than MDB. The cost of this sensitivity, however, is a higher non-concordance rate than MDB. Alternatively, MDB has difficulty recognizing that one or more SNPs have been observed within a hot spot. It often places a single change point where two were needed to describe the structure of the haplotypes. Consequently, MDB's sensitivity is reduced compared to EB. Finally, in terms of computational time, EB runs much more quickly than MDB.

To understand the nature of the false discoveries (misplaced boundaries within blocks) detected by EB, we investigated these using H-clust. For this investigation, we call a SNP an ''outlier'' when it is not clustered with the remainder of a contiguous block by H-clust (at level $\eta = 0.8$). For each misplaced break, we determined if the break was adjacent to an outlier and discovered that, indeed, the correspondence between the false breaks and presence of an outlier was quite high. Under all simulation conditions depicted in Table I, this was true 95% of the time or more. Furthermore, as in IL21R described above, removing the outlier usually leads to correct identification of the block delineated by the underlying cold spot.

We also simulated data with cold regions separated by regions (15 kb), not spots, of high recombination. For these simulations we used $\delta = 9\times10^{-8}$/bp for the baseline recombination rate. This choice led to overall patterns of LD more similar to those observed in the SeattleSNP database. The resulting hot regions contained on average 19 common SNPs. Some clusters of SNPs tend to retain a high degree of LD within a hot region because the realized recombination events have been insufficient to break down LD throughout the region. This occurs even though the recombination rate is high within the entire hot region. Consequently, most block-finding methods are likely to detect some small blocks within the hot regions. Therefore, we compare the blocks found by the EB and the MDB procedures. (Because the MDB procedure does not specifically look for unblocked SNPs, we declare any cluster of two or more SNPs between discovered MDB block boundaries to be blocks. This definition is necessary because MDB ''blocks'' everything. Unblocked material can only be defined as those blocks of length one.)

In an analysis of 100 simulated samples, each with $N = 100$, EB found no blocks in 15 samples. On average, 35% of the SNPs were blocked, usually with two short blocks in the hot region. The length of the blocks ranged from 3–7 SNPs with 92% of them of length 3–4 SNPs. In contrast, the MDB method usually blocked most of the SNPs in the hot region. On average, it found 4–5 blocks per sample. These blocks ranged in length from 2–12 SNPs, with 82% of them of length 2–5 SNPs. For EB, the number of unblocked SNPs per sample ranged from 2–30, with 86% having between 6 and 17, whereas only 2% of the samples had 6 or more SNPs unblocked for MDB.

TABLE II. LD Structure of blocked and unblocked snps in a region of high recombination rate[a]

Type of SNPs	$\lvert r\rvert$	$\lvert D'\rvert$
EB blocked	0.524 (0.129)	0.852 (0.126)
EB unblocked	0.168 (0.029)	0.401 (0.085)
EB all unblocked[b]	0.17 (0.032)	0.413 (0.062)
MDB blocked	0.373 (0.189)	0.75 (0.222)
MDB unblocked	0.131 (0.186)	0.239 (0.303)

[a]One hundred genes were simulated, each with $N = 100$, and analyzed by the proposed method (EB) and Anderson and Novembre's [2003] method (MDB). A fraction of a gene was declared blocked/unblocked by each method, for each gene. Average values of $|D'|$ and $|r|$, the correlation coefficient, were obtained for pairs of SNPs within blocks and outside of blocks. The standard deviation of the LD measures is given in parentheses.
[b]Of 100 simulated genes, 15 were entirely unblocked by EB.

To assess the nature of the discovered blocks, we examined their LD structure using $|r|$ and D' as measures (Table II). The LD of SNPs blocked and unblocked by EB differed substantially ($|r| = 0.52$ vs. 0.17). On average, SNPs blocked by MDB had considerably less LD ($|r| = 0.37$) and this result held regardless of the size of the MDB blocks (Table II).

We can conclude from these simulations that, if the hot spots are small (e.g., 1 Kb), MDB provides a better reconstruction of the location of the hotspots boundaries. This method is less likely than EB to break cold regions into disparate blocks due to a single SNP with an inconsistent pattern of LD within a large block pattern. If there are hot regions, however, the results of MDB are difficult to interpret because many regions of low LD will be blocked. Consequently, if the goal of the analysis is to discover the sets of contiguous SNPs with high LD, EB is a better choice than MDB. Finally, EB, used in conjunction with H-clust, permits the identification of outlier SNPs within a block. These SNPs, with their dissimilar correlation structure, are of inherent interest (see also Dawson et al. [2002]).

TABLE III. Performance of Two Tag SNP Selection Procedures

			SNPs		Min R^2		Mean R^2	
N	R_h	η	Carlson	H-clust	Carlson	H-clust	Carlson	H-clust
50	50	0.8	28	27	96.5	95.8	99.7	99.7
50	100	0.8	28	28	96.7	96.5	99.4	99.5
50	200	0.8	27	27	94.6	94.1	99.6	99.6
100	50	0.8	27	27	94.7	94.3	99.5	99.6
100	100	0.8	27	26	94.2	93.8	99.4	99.5
100	200	0.8	27	27	94.6	94.2	99.5	99.6
50	50	0.6	25	23	90.7	88.5	99.1	98.9
50	100	0.6	28	25	91.4	89.5	99.2	99.0
50	200	0.6	27	25	91.6	89.9	99.2	99.1
100	50	0.6	27	25	87.8	85.4	99.6	98.5
100	100	0.6	26	24	85.6	84.6	98.3	98.3
100	200	0.6	27	24	86.7	85.5	98.3	98.5

[a]The data generated for Table I are analyzed by the proposed tag SNP selection method (H-clust) and the method of Carlson et al. [2004]. N is the number of haplotypes, R_h is the multiplicative increase in recombinational rate, η is the minimum r^2 required, and R^2 is the coefficient of determination for predicting a SNP genotype, using only the tag SNP genotypes.

COMPARISON OF TAGGING METHODS

To evaluate the H-clust tag SNPs, we compared them to Carlson et al.'s [2004] tag SNPs by the simulation scheme used to assess blocking (Table I). With the correlation level for both methods set at a common value ($\eta = 0.8, 0.6$), we computed the number of tag SNPs selected. On average, the Carlson et al. method requires slightly more genotyping, choosing up to three more polymorphisms for the tag SNPs. Within a set of highly correlated SNPs, the choice of tag SNPs is not unique. Hence, it is difficult to directly compare among tag SNP sets. We contrast the selection indirectly by comparing their ability to predict the genotype of each common SNP in the region by using only the genotypes of the selected tag SNPs. For each level of η, the coefficient of determination, or R^2 for predicting an unknown or ungenotyped SNP is measured using the approach described by Stram et al. [2003]. This approach uses the multilocus genotypes of the tag SNPs, coded as 0, 1, 2 to record the count of a particular form, to predict the genotype of the SNP in question. The predictive ability of each tag set was surprisingly high and almost identical across methods (Table III). These results indicate that the simpler H-clust approach, is either identical or superior in performance to the method of Carlson et al. [2004], except in terms of the minimum R^2 (Table III). This is notable because the latter method requires imputation of pairwise haplotype phase while the former is genotype-based.

DISCUSSION

Human phenotypic variation, including the expression of disease, is determined by the interplay of genetic variation and environmental factors. At present, there is a massive effort to understand what genetic variants, alter liability to common diseases, such as heart disease, diabetes, and major depression, and how they have an impact. To find these genetic variants, researchers first survey the variation in targeted genes or genomic regions, and then relate this variation to phenotypic variation at the populational and intra-familial levels. As Clark et al. [1998] point out, this can be a challenging task for all but the smallest of genes. Indeed even ''average'' size genes exhibit substantial variability (Fig. 2, SeattleSNPs site), and larger genomic regions multiply the complexity.

How one should screen a large number of SNPs and other genetic variants for their impact on liability is still a matter of ongoing research [Chapman et al., 2003; Zhang et al., 2002b, 2003; Carlson et al., 2004]. An obvious first step would be to quantify the structure of linkage disequilibrium in the gene or region. Then, for subregions with a limited set of haplotypes, or haplotype blocks, one could test for association between haplotypes and phenotypes [Clayton, 1999; Clayton and Jones, 1999; Seltman et al., 2003; Fallin et al., 2001]; in regions without haplotype blocks, single SNPs could be tested. For this approach to work, one needs a method to discriminate between blocked and unblocked regions.

We propose one such method here, called EB, which is structurally related to the methods proposed by Anderson and Novembre [2003] and is conceptually related to the methods of Gabriel et al. [2002]. Unlike most methods for discovering haplotype-blocks, EB does not aim to discover "haplotype tagging SNPs" to be used to identify the full underlying haplotypes. Rather, it aims to differentiate between regions populated by weakly correlated SNPs and regions populated by at least several SNPs in strong LD. As a result, it "blocks" only those regions of minimal diversity in the haplotype distribution, relative to the potential diversity expected from SNPs with little pairwise correlation. Viewed under a range of conditions, the method is quite successful in achieving this goal, relative to competing methods. In addition to being extremely simple to implement and fast to deliver its solution, simulations show that the EB method is quite successful at discriminating between blocked and unblocked regions.

We also propose an extremely simple method, called H-clust, for choosing tag SNPs, which is based on hierarchical clustering. H-clust does not utilize the inferred block structure of the region under investigation; instead, it analyses the distribution of multilocus genotypes. In so doing, it is more robust to the true underlying LD structure of the region. It should also be more robust to missing genotype data than equivalent methods that use principle components. Compared to the method of Carlson et al. [2004], H-clust predicts unmeasured SNPs at least as well, using either the same or a smaller number of tagging SNPs. Because it tends to be more parsimonious, it will also tend to be more economical because it chooses fewer SNPs to genotype, and even results in slightly more power for association tests [Roeder et al., 2005]. Remarkably, H-clust does this without knowledge of haplotype phase, which is required for many competing methods, including Carlson et al.'s [2004]. The method uses standard clustering software and the process can be depicted graphically. In this way, it allows for a more natural choice of the number of tag SNPs to genotype for an association study by visualizing the clustering of SNPs.

ELECTRONIC-DATABASE INFORMATION

SeattleSNPs Database, NHLBI Program for Genomic Applications, UW-FHCRC, Seattle, WA, http://pga.gs.washington.edu/.

Software for EB, H-clust, and other descriptive plots is posted at http://wpicr.wpic.pitt.edu/WPICCompGen/.

REFERENCES

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO. 2001. Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191–197.

Ackerman H, Usen S, Mott R, Richardson A, Sisay-Joof F, Katundu P, Taylor T, Ward R, Molyneux M, Pinder M, Kwiatkowski DP. 2003. Haplotypic analysis of the TNF locus by association efficiency and entropy. Genome Biol 4:R24.

Anderson EC, Novembre J. 2003. Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336–354.

Cardon LR, Abecasis GR. 2003. Using haplotype blocks to map human complex trait loci. Trends Genet 19:135–140.

Carlson CS, Eberle, AM, Rieder JM, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120.

Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31.

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612.

Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E. 2003. Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. Am J Hum Genet 73:285–300.

Clayton DG. 1999. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177.

Clayton DG, Jones H. 1999. Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65:1161–1169.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74:610–622.

Daly MJ, Rioux JD, Schaffner SF, Hudson TH, Lander ES. 2001. High resolution haplotype structure in the human genome. Nat Genet 29:229–232.

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

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

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. 2001. Genetic analysis of case/control data using

estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–151.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang LY, et al. 2003. The International HapMap Project. Nature 426:789–796.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237.

Jorde LB. 1995. Linkage disequilibrium as a gene mapping tool. Am J Hum Genet 56:11–14.

Jorde LB. 2000. Linkage disequilibrium and the search for complex disease genes. Genome Res 10:1435–1444.

Kauffman L, Rousseeuw PJ. 1990. Finding groups in data. An introduction to cluster analysis. New York: John Willey & Sons.

Ke X, Cardon LR. 2003. Efficient selective screening of haplotype tag SNPs. Bioinformatics 19:287–288.

Koivisto M, Perola M, Varilo T, Hennah W, Ekelund J, Lukk M, Peltonen L, Ukkonen E, Mannila H. 2003. An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. Pac Symp Biocomput 502–513.

Lien S, Szyda J, Schechinger B, Rappold G, Arnheim N. 2000. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. Am J Hum Genet 66: 557–566.

May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys AJ. 2002. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. Nat Genet 31: 272–275.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169.

Nordborg M, Tavare S. 2002. Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90.

Nothnagel M, Furst R, Rohde K. 2002. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered 54:186–198.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP and Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, et al. 2003. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Rissanen R. 1989. Stochastic complexity in statistical inquiry. London: World Scientific.

Roeder K, Bacanu S-A, Sonpar V, Zhang X, Devlin B. 2005. Analysis of single-locus tests to detect gene/disease associations. Genet Epidemiol 28: in press. doi://10.1002/gepi.20050.

Schaid DJ. 2004. Linkage disequilibrium testing when linkage phase is unknown. Genetics 166:505–512.

Schneider JA, Peto TE, Boone RA, Boyce AJ, Clegg JB. 2002. Direct measurement of the male recombination fraction in the human beta-globin hot spot. Hum Mol Genet 11:207–215.

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF. 2003. Minimal haplotype tagging. Proc Natl Acad Sci USA 100:9900–9905.

Seltman H, Roeder K, Devlin B. 2003. Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. 2003. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 55:27–36.

Venables WN, Ripley BD. 1994. Modern applied statistics with Splus. New York: Springer Verlag.

Wall JD, Pritchard JK. 2003. Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73:502–516.

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

Weir BS. 1979. Inferences about linkage disequilibrium. Biometrics 35:235–254.

Zaykin DV. 2004. Bounds and normalization of the composite linkage disequilibrium coefficient. Genet Epidemiol 27(3): 252–257.

Zhang K, Deng M, Chen T, Waterman MS, Sun F. 2002a. A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339.

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

Zhang X, Roeder K, Wallstrom G, Devlin B. 2003. Integration of association statistics over genomic regions using Bayesian adaptive regression splines. Human Genom 1:20–29.

Zhang K, Qin ZS, Liu JS, Chen T, Waterman M, Sun F. 2004. Haplotype block paritioning and tag SNP selection using genotype data and their applications to association studies. Genome Res 14:908–916.

APPENDIX

MODEL SELECTION PROCEDURES: IMPLEMENTATION

An iterative algorithm, which searches only for partitions corresponding to regions with high standardized entropy, is described as follows:

1. Let $M = 0$, $b_0 = \emptyset$, $l = 1$.
2. If $\{1, \ldots, S\} - \bigcup_{j<l} b_l$ has cardinality less than 3, then $\bigcup_{j<l} b_l$ is the optimal blocking scheme with $M = l - 1$ and the algorithm terminates, otherwise do the following:

   - Set $k = 0$ and define $b_l^k \subset \{1, \ldots, S\} - \bigcup_{j<l} b_l$ to be the block formed by the 3 contiguous SNPs whose standardized entropies are maximal, where the value of the standardized entropy at each SNP i is the standardized entropy computed with the SNP i as the middle one. Compute the AIC corresponding to the model with blocks $(\bigcup_{j<l} b_l) \cup b_l^k$.
   - For each i, with $i \in \{1, \ldots, S\} - (\bigcup_{j<b_j}) \cup b_l^k$ being the index corresponding to the independent SNP contiguous to the block b_l^k having maximal standardized entropy, set $k = k + 1$ and $b_l^k = b_l^k \cup \{i\}$ and compute the AIC corresponding to the partition with blocks $(\bigcup_{j<l} b_j) \cup b_l^k$.
   - Set $b_l = b_l^{k^*}$, where $b_l^{k^*}$ is the block with minimal AIC among the ones with common index l.
   - If the AIC of $\bigcup_{j\le l} b_j$ is bigger than the AIC of $\bigcup_{j<l} b_j$, then $\bigcup_{j<l} b_j$ is the optimal blocking scheme with $M = l - 1$ and the algorithm terminates, otherwise set $l = l + 1$ and go to step 2.