Measuring and Partitioning the High-Order Linkage Disequilibrium by Multiple Order Markov Chains

Yunjung Kim,1,2 Sheng Feng,1,2 and Zhao-Bang Zeng1–3

1Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina
2Department of Statistics, North Carolina State University, Raleigh, North Carolina
3Department of Genetics, North Carolina State University, Raleigh, North Carolina

A map of the background levels of disequilibrium between nearby markers can be useful for association mapping studies. In order to assess the background levels of linkage disequilibrium (LD), multilocus LD measures are more advantageous than pairwise LD measures because the combined analysis of pairwise LD measures is not adequate to detect simultaneous allele associations among multiple markers. Various multilocus LD measures based on haplotypes have been proposed. However, most of these measures provide a single index of association among multiple markers and does not reveal the complex patterns and different levels of LD structure. In this paper, we employ non-homogeneous, multiple order Markov Chain models as a statistical framework to measure and partition the LD among multiple markers into components due to different orders of marker associations. Using a sliding window of multiple markers on phased haplotype data, we compute corresponding likelihoods for different Markov Chain (MC) orders in each window. The log-likelihood difference between the lowest MC order model (MC0) and the highest MC order model in each window is used as a measure of the total LD or the overall deviation from the gametic equilibrium for the window. Then, we partition the total LD into lower order disequilibria and estimate the effects from two-, three-, and higher order disequilibria. The relationship between different orders of LD and the log-likelihood difference involving two different orders of MC models are explored. By applying our method to the phased haplotype data in the ENCODE regions of the HapMap project, we are able to identify high/low multilocus LD regions. Our results reveal that the most LD in the HapMap data is attributed to the LD between adjacent pairs of markers across the whole region. LD between adjacent pairs of markers appears to be more significant in high multilocus LD regions than in low multilocus LD regions. We also find that as the multilocus total LD increases, the effects of high-order LD tends to get weaker due to the lack of observed multilocus haplotypes. The overall estimates of first, second, third, and fourth order LD across the ENCODE regions are 64, 23, 9, and 3%. Genet. Epidemiol. 2008.

INTRODUCTION

Assessing the patterns of linkage disequilibrium (LD) along a chromosome has been an important issue in disease mapping studies and in studies of history of humans and other species. In particular, in an association mapping of disease genes, the inference is usually based on an LD test for association between genetic variation at a known set of markers and disease phenotypes. If such an association is detected between a particular marker locus and the phenotype, it suggests that either the variation at that marker locus affects the phenotype of interest, or that the variation of that marker locus is in LD with the true phenotype-related locus, which was not genotyped. The association signal and the LD pattern in the region provide important information about candidate location of disease genes. Many studies have indicated that the levels of disequilibrium vary a lot across genomic regions and populations. To design and interpret disease mapping studies, one needs to refer to a map of the background levels of disequilibrium that can be expected in a given region and in a given population. To construct such an LD map, the levels and patterns of disequilibrium between close-by markers...
need to be measured effectively [Sabatti and Risch, 2002; Muller, 2004; Greenspan and Geiger, 2005].

In association mapping, many investigators have studied the question whether haplotype-based association tests are more powerful than single-locus tests. Akey et al. [2000] found that haplotype-based tests can improve the power of association mapping. In contrast, Long and Langley [1999] and Kaplan and Morris [2001] found that single-locus tests are as powerful as haplotype-based tests. Nielsen et al. [2004] compared the power of haplotype and single-marker tests under different patterns of pairwise and three-locus LD. They found that haplotype-based tests tend to be more powerful when moderate to high levels of three-locus LD exist and single-marker tests tend to prevail when pairwise LD between the markers and the functional site is high.

Most previous studies on LD patterns and LD blocks have focused on using pairwise LD measures and have not examined and utilized joint multilocus LD [Nielsen et al., 2004]. It is more advantageous to use multilocus LD to assess the background levels of LD because the combined analysis of all pairwise LD measures across a region is insufficient to detect simultaneous allele associations among multiple loci. An illustrative example can be found in Nielsen et al. [2004] and Muller [2004]. Muller [2004] reviewed various multilocus LD measures. One common property of these methods is to summarize the data with a single multilocus LD measure by first calculating the difference between observed state and the expected one under linkage equilibrium (LE) and then normalizing it to allow comparison across different regions and populations. The extent of normalized difference between observed state and the expected state can be regarded as the degree of overall departure from LE. Although these methods overcome the limit of pairwise LD measure by considering joint multilocus LD, they do not distinguish the type and structures of multilocus LD, such as how much multilocus LD is due to LD between marker pairs and how much due to higher orders. For example, if we consider \( p \) single nucleotide polymorphisms (SNPs) simultaneously, there are \( 2^p - (p + 1) \) LD terms of varying orders that need to be considered to fully characterize the LD structure in them. In this paper, we develop a non-homogeneous, multiple order Markov Chain (MOMC) model for characterizing LD structure and use it to estimate the contributions of hierarchical structure of multiple markers to multilocus LD. This partition of LD will help us to better understand the LD structure and provide more useful information in designing appropriate methods for fine mapping of disease genes.

For this study, we are interested in measuring the extent of the overall departure from LD, called the total multilocus LD, in a chromosome region and partitioning the total multilocus LD into various orders of LD due to two-, three-locus association, and so on. For this purpose, we employ a non-homogeneous, MOMC model as a statistical framework. Our measure of the multilocus LD based on the MOMC model is very similar to that of Nothnagel et al. [2002] which was based on the concept of entropy. Both methods can describe the general LD patterns along a chromosome and detect haplotype blocks as one of applications. However, our MOMC model can provide more detailed information about the structure and detailed patterns of LD. To our knowledge, this is the first time that a model is constructed to partition the total multilocus LD into LD components of lower orders. We apply our method to the phased haplotype data in the ENCODE regions of the HapMap project. By partitioning the multilocus LD into different components, we observe that a great proportion of the total LD can be explained by LD between adjacent marker pairs. We also observe significant variation in the partitions of LD between chromosomal regions and between populations. YRI population has more high order LD as compared to the other three populations (CEU, HCB, and JPT).

**METHODS**

**MOMC MODEL AND LIKELIHOOD**

In this section, we formulate our statistical model and elaborate its theoretical foundation. The model was initially proposed by Feng [2004]. We consider a population in Hardy-Weinberg equilibrium. Our current method assumes that phased haplotypes are available and there are no genotyping errors. The phased haplotype data can be obtained either from laboratory techniques such as long-range PCR or chromosomal isolation [Michalatos-Beloin et al., 1996; Douglas et al., 2001] or reconstructed from diplotype data by some statistical or computational methods [Clark, 1990; Excoffier and Slatkin, 1995; Niu et al., 2002].

A Markov chain is a probabilistic model that can be used to represent dependencies between successive observations of random variables. In this paper, we consider each SNP as a discrete random variable \( (M_i) \) taking values in the finite set \{0, 1\}. For the first-order Markov chain model, denoted as MC1, an observation at a marker depends only on the observation of its adjacent marker on the left or right and is independent of others. Similarly, for the \( r \)th order Markov chain model, denoted as MCR, an observation at a marker, say \( M_i \), depends on the observations of previous \( r \) markers from \( M_{i-1} \) to \( M_{i-r} \), for example. The Markov chain model applied in this study is non-homogeneous in that all
TABLE 1. Estimation of proportional effects of lower order disequilibrium in the 10 ENCODE regions of HapMap data

<table>
<thead>
<tr>
<th>Region</th>
<th>YRI</th>
<th>CEU</th>
<th>HCB</th>
<th>JPT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\phi_1)</td>
<td>(\phi_2)</td>
<td>(\phi_3)</td>
<td>(\phi_4)</td>
</tr>
<tr>
<td>ENr112</td>
<td>53 (25)</td>
<td>29 (22)</td>
<td>13 (16)</td>
<td>6 (10)</td>
</tr>
<tr>
<td>ENr131</td>
<td>56 (26)</td>
<td>26 (22)</td>
<td>13 (17)</td>
<td>5 (11)</td>
</tr>
<tr>
<td>ENr113</td>
<td>61 (26)</td>
<td>25 (22)</td>
<td>10 (14)</td>
<td>4 (9)</td>
</tr>
<tr>
<td>ENm1010</td>
<td>57 (25)</td>
<td>26 (21)</td>
<td>13 (15)</td>
<td>4 (9)</td>
</tr>
<tr>
<td>ENm1015</td>
<td>51 (26)</td>
<td>29 (24)</td>
<td>15 (18)</td>
<td>5 (11)</td>
</tr>
<tr>
<td>ENr113</td>
<td>56 (26)</td>
<td>25 (22)</td>
<td>10 (14)</td>
<td>4 (9)</td>
</tr>
<tr>
<td>ENm1014</td>
<td>55 (27)</td>
<td>26 (23)</td>
<td>14 (17)</td>
<td>5 (11)</td>
</tr>
<tr>
<td>ENm1013</td>
<td>61 (26)</td>
<td>26 (22)</td>
<td>13 (15)</td>
<td>4 (9)</td>
</tr>
<tr>
<td>ENr113</td>
<td>53 (25)</td>
<td>29 (22)</td>
<td>13 (16)</td>
<td>6 (10)</td>
</tr>
<tr>
<td></td>
<td>71 (24)</td>
<td>21 (21)</td>
<td>7 (12)</td>
<td>2 (5)</td>
</tr>
<tr>
<td></td>
<td>67 (24)</td>
<td>21 (21)</td>
<td>7 (12)</td>
<td>2 (5)</td>
</tr>
<tr>
<td></td>
<td>62 (26)</td>
<td>24 (22)</td>
<td>9 (14)</td>
<td>4 (9)</td>
</tr>
<tr>
<td></td>
<td>71 (25)</td>
<td>21 (21)</td>
<td>6 (12)</td>
<td>2 (6)</td>
</tr>
<tr>
<td></td>
<td>65 (27)</td>
<td>26 (24)</td>
<td>7 (13)</td>
<td>1 (5)</td>
</tr>
<tr>
<td></td>
<td>63 (24)</td>
<td>24 (21)</td>
<td>7 (13)</td>
<td>3 (8)</td>
</tr>
<tr>
<td></td>
<td>68 (25)</td>
<td>21 (20)</td>
<td>8 (14)</td>
<td>3 (8)</td>
</tr>
<tr>
<td></td>
<td>63 (24)</td>
<td>24 (21)</td>
<td>9 (14)</td>
<td>3 (8)</td>
</tr>
<tr>
<td></td>
<td>63 (25)</td>
<td>25 (22)</td>
<td>9 (13)</td>
<td>3 (9)</td>
</tr>
<tr>
<td></td>
<td>78 (24)</td>
<td>17 (20)</td>
<td>4 (9)</td>
<td>1 (6)</td>
</tr>
<tr>
<td></td>
<td>67 (23)</td>
<td>21 (23)</td>
<td>8 (13)</td>
<td>2 (6)</td>
</tr>
<tr>
<td></td>
<td>66 (24)</td>
<td>22 (21)</td>
<td>7 (12)</td>
<td>2 (7)</td>
</tr>
</tbody>
</table>

The average proportional effects from adjacent pairs of markers (\(\phi_1\)), from adjacent three and non-adjacent two markers (\(\phi_2\)), from adjacent four, non-adjacent three, and non-adjacent two markers (\(\phi_3\)), from adjacent five, non-adjacent four, non-adjacent three, and non-adjacent two markers (\(\phi_4\)) to the total LD are estimated across each ENCODE region in each population. The numbers in parentheses are standard deviation.

transition probabilities from one state to another state are locus specific (or marker specific).

Consider a data set containing \(N\) independent phased haplotypes of a number of markers from a natural population. Suppose each phased haplotype \(c (c = 1, 2, \ldots, N)\) has \(L\) SNP markers, with two alleles at each SNP \(M_i (i = 1, 2, \ldots, L)\) coded as 0 or 1. Let \(m_{i,c}\) (\(m_{i,c} = 0\) or 1) denote the observed allele of \(M_i\) on the haplotype \(c\). Let \(H(i, r)\) denote the haplotype fragment containing \(r\) consecutive markers starting from the \(i\)th marker (i.e. \(M_i, M_{i+1}, \ldots, M_{i+r-1}\)). Let \(h_i(i, r)\) denote the observed \(H(i, r)\) on a particular chromosome \(c\). Note that \(h_i(i, r)\) is a \(r \times 1\) vector of 0 and 1. Then, \(P[H(i, r) = h_i(i, r)]\) represents the population frequency of the specific haplotype fragment \(h_i(i, r)\). Let \(P[M_i = m_{i,c} | H(i-r, r) = h_i(i-r, r)]\) denote the conditional probability that a marker \(M_i\) takes value \(m_{i,c}\) (\(m_{i,c} = 0\) or 1) given its previous \(r\) markers \(M_{i-r}, \ldots, M_{i-1}\) taking values \(h_i(i-r, r)\). If we assume that the haplotype counts follow a multinomial distribution, the maximum likelihood estimate of \(P[H(i-r, r) = h_i(i-r, r)]\) is

\[
\hat{P}[H(i-r, r) = h_i(i-r, r)] = [\text{count of } h_i(i-r, r)]/N.
\]

Thus, the conditional probability \(P[M_i = m_{i,c} | H(i-r, r) = h_i(i-r, r)]\) can be estimated as

\[
\hat{P}[M_i = m_{i,c} | H(i-r, r) = h_i(i-r, r)] = \frac{\hat{P}[H(i-r, r + 1) = h_i(i-r, r + 1)]}{\hat{P}[H(i-r, r) = h_i(i-r, r)]}.
\]

We illustrate our model using a small hypothetical dataset in which each chromosome \(c\) consists of only 3 SNP markers of \(M_1, M_2, M_3\). Let \(L_c(MCr)\) be the likelihood of observing \((M_1, M_2, M_3)\) assuming \(MCr\) on chromosome \(c\).

\[
L_c(MC0) = P(M_1) \times P(M_2) \times P(M_3)
\]

\[
L_c(MC1) = P(M_1) \times P(M_2 | M_1) \times P(M_3 | M_2) = P(M_1)
\]

\[
\times \frac{P(M_1, M_2)}{P(M_2)}
\]

\[
L_c(MC2) = P(M_1, M_2) \times P(M_3 | M_1, M_2) = P(M_1, M_2)
\]

\[
\times \frac{P(M_1, M_2, M_3)}{P(M_2, M_3)}.
\]

For \(C\) independent chromosomes, the likelihood of \(MC\) model is

\[
L(MC) = \prod_c L_c(MCr).
\]

This likelihood will be the same regardless of the orientation to which a Markov chain moves. That is, \(L(MC)\) calculated from left to right is equal to that calculated from right to left.

There are two types of parameters in the likelihood of \(MC\) model: the \(2^r\) chain initiating probabilities and \(2^r\) \((\text{number of markers} - r)\) transition probabilities. Assuming that all the possible haplotypes are observed, the total number of parameters in \(MC\) model will be the sum of two types of parameters (see Supplementary Table 1). Maximum likelihood estimates of both types of parameters can be obtained by assuming that the haplotype counts follow a multinomial distribution. Then, the likelihood \(L(MC)\) is calculated by replacing all parameters with their maximum likelihood estimates. If some of multilocus haplotypes are unobserved due to finite sample size, the associated parameters cannot be estimated and thus 0 will be assigned to them. However, the unobserved haplotypes do not contribute to the
likelihood of different orders of MC model by the convention of $0 \times \log(\text{non-negative value}) = 0$ [Liu and Lin, 2005].

A sliding-window approach is used to compute the magnitude of multilocus LD along the region of interest and to partition the overall departure from LE into lower order disequilibria. Within each window of size $w$, Markov chain models for different orders from 0 to $w-1$ are applied to fit the haplotype data. Then, the log-likelihoods corresponding to different MC orders are computed. In each window, MC0 model represents the random association of alleles from different marker loci (i.e., LE) while MC$w-1$ model represents the full haplotype model with non-random association of alleles among all markers within a window. In between, the other MC models involve different levels of Markov properties (conditional independence). Note that the log-likelihood of a lower MC order is always smaller or equal to that of a higher MC order. That is, $LL(\text{MC0}) \leq LL(\text{MC1}) \leq \cdots \leq LL(\text{MCw - 1})$.

**MULTILOCUS LD MEASURE**

Using the log-likelihoods from different MC orders in each window, we define $\delta_1, \delta_2, \ldots, \delta_{w-1}$ as follows:

\[
\delta_1 = \frac{LL(\text{MC0}) - LL(\text{MC1})}{LL(\text{MC0})} \times \frac{w}{w-1},
\]

\[
\delta_2 = \frac{LL(\text{MC0}) - LL(\text{MC2})}{LL(\text{MC0})} \times \frac{w}{w-1},
\]

\[
\vdots
\]

\[
\delta_{w-1} = \frac{LL(\text{MC0}) - LL(\text{MCw - 1})}{LL(\text{MC0})} \times \frac{w}{w-1},
\]

where $LL(\text{MC0}), LL(\text{MC1}), \ldots, LL(\text{MCw - 1})$ denote the log-likelihoods computed from MC0, MC1, ..., MC$w-1$ model, respectively, and $w$ denotes a window size. They measure the deviation from LE when the loci are modeled by different MC orders. $w/(w-1)$ is multiplied to make $\delta$s range between 0 and 1. Before the normalization, $\delta$s depend on the window size ($w$) and $(w - 1)/w$ is the maximum value that they can reach. By multiplying $w/(w-1)$, the parameters are bounded as $0 \leq \delta_1 \leq \delta_2 \leq \cdots \leq \delta_{w-1} \leq 1$ regardless of window size $w$, with 0 indicating LE and 1 indicating the complete LD. It is interesting to note that $\delta_{w-1}$ coincides with the normalized entropy difference ($\epsilon'$) developed by Nothnagel et al. [2002].

**PARTITION OF THE MULTILOCUS LD**

The overall deviation from the LE ($\delta_{w-1}$) in each window can be partitioned into the amount of contribution from each lower order MC model in the following way:

\[
\phi_1 = \frac{\delta_1}{\delta_{w-1}} = \frac{LL(\text{MC1}) - LL(\text{MC0})}{LL(\text{MCw - 1}) - LL(\text{MC0})},
\]

\[
\phi_2 = \frac{\delta_2 - \delta_1}{\delta_{w-1}} = \frac{LL(\text{MC2}) - LL(\text{MC1})}{LL(\text{MCw - 1}) - LL(\text{MC0})},
\]

\[
\vdots
\]

\[
\phi_{w-1} = \frac{\delta_{w-1} - \delta_{w-2}}{\delta_{w-1}} = \frac{LL(\text{MCw - 1}) - LL(\text{MCw - 2})}{LL(\text{MCw - 1}) - LL(\text{MC0})}.
\]

To show the parametric compositions of $\phi$s in terms of marker LD measures, we can approximate the expected values of numerators and denominators of different $\phi_i (i = 1, 2, \ldots, w - 1)$ in equations (1) by using a Taylor series expansion of order 2. This approximation can help us to understand and interpret the meanings of different quantities. It turns out that the numerators and denominators of different $\phi_i (i = 1, 2, \ldots, w - 1)$ are related to the sum of the corresponding squared correlation coefficients of different markers at different orders. The detailed derivation of the approximate expectation of the log-likelihood difference between different MC order models (MC1, MC2, ..., MC$w-1$) and the independent model (MC0) is given in Appendix A.

For example, for a window of size 2 consisting of two markers ($M_i, M_j$), the expected difference between log-likelihoods of MC1 and MC0 is related to a half of the square of the correlation coefficient between the two markers

\[
\frac{1}{N}E[LL(\text{MC1}) - LL(\text{MC0})] = \sum_i \sum_j P_{ij} \log \left( \frac{P_{ij}}{P_{i} \times P_{j}} \right) \approx \frac{1}{2} r_{ij}^2,
\]

where $N$ denotes the sample size, $E$ denotes expectation,

\[
r_{ij}^2 = (D_{ij})^2 / P_i \times P_j \times (1 - P_i) \times (1 - P_j),
\]

the two-locus LD coefficient ($D_{ij}$) is defined as $P_{ij} - P_i P_j$. Similar approximation results for a pair of markers are found in two papers [Nothnagel et al., 2002; Liu and Lin, 2005]. The second term in equation (2) is the same as the mutual information between two systems in information theory [Kullback, 1978; Goebel et al., 2005]. Note that if we multiply 2N to equation (2), the expressions in (2) become the standard test of independence in a $2 \times 2$ contingency table, $\chi^2_1$ test. Using the moment-generating function, we can easily prove that the expressions in (2) follow a gamma distribution with two parameters ($\alpha = 1/2, \beta = 1/N$).

For a window of size 3 consisting of three markers ($M_i, M_j, M_k$) in this order, we have the following
Fig. 1. Comparison of LD patterns obtained from our multilocus LD measure and pairwise LD measures using HaploBlockFinder version 0.7. In the top diagrams, multilocus total LD in ENm010 region is plotted against the physical location of the central marker in each window for four different populations (YRI, CEU, HCB, JPT). Window size 5 is used for this analysis. The locations of blocks using a threshold of 0.5 are depicted by red bold lines below the graph. In the bottom diagrams, the LD values are plotted using two pairwise LD measures, $|D|$ (top right) and $r^2$ (bottom left) and the haplotype blocks are shown on the top side and the left side with orange color. Short little white lines on the top and left in the bottom diagram represent the position of SNPs. LD, linkage disequilibrium; SNPs, single nucleotide polymorphisms.
approximations:
\[
\frac{1}{N} E[LL(MC1) - LL(MC0)] \approx \frac{1}{2} r_{ij}^2 + \frac{1}{2} r_{jk}^2,
\]

Fig. 1. Continued.

where
\[
r_{ij}^2 = \frac{(P_{ij} - P_{ij} \times P_{ij} \times (1 - P_{ij}) \times (1 - P_{ij}))}{P_{ij} \times P_{ij} \times (1 - P_{ij}) \times (1 - P_{ij})}
\]
was resequenced in 48 unrelated individuals and genotyped a representative collection of 10 individuals from each population in the 10 ENCODE regions. We use a software package called HaploBlockFinder [Zhang and Jin, 2003] version 0.7 to perform pairwise LD analysis. For each of four populations, Figure 1 compares the magnitudes of total LD ($D^2$) for window size 5 along the marker sequence in the top plots with the pairwise LD measures of $r^2$ in the bottom plots. We define ad hoc haplotype blocks as an union of windows over which the level of total LD is 0.5 or above. The blocks identified from our approach is depicted by red bold lines in the top plots and compared with those from pairwise LD measures (see Fig. 1). In the bottom plots, the blocks are defined as a consecutive set of markers in which minimal $D^2$ is 0.95 or above and high LD regions appear as different sizes of red triangles. Using the threshold
of 0.5, we identified many short blocks (35 blocks) in YRI, fewer but longer blocks in the other three populations—21 blocks in CEU, 11 blocks in HCB, and 15 blocks in JPT. The average block lengths are 5.5 kb in YRI, 17.6 kb in CEU, 36.3 kb in HCB, and 26.6 kb in JPT. The size of each block varies a lot, from 0.01 to 32 kb in YRI, 0.1 to 168 kb in CEU, 0.5 to 94 kb in HCB, and 0.3 to 94 kb in JPT. The large-scale LD patterns show that high LD regions in the top plots are similar to the ones specified in the bottom plots and the sites of recombination hot spots around high LD regions in the bottom plots are also similar to the regions where total LD drops down sharply in the top plots. Both top and bottom plots agree that the LD pattern in YRI population is very different from those observed in the other three populations in which LD extends to a similar and long extent. This fact is consistent with the results discovered by other research groups [Gabriel et al., 2002; Hinds et al., 2005].

**PARTITIONING OF TOTAL LD**

So far, we have demonstrated that the LD profile created by our total LD measure reasonably agrees with that created by pairwise LD measures. Nothnagel et al. [2002] and Nothnagel [2004] claimed that the normalized entropy difference could be considered as a generalization of \( r^2 \) to haplotypes of more than two bi-allelic loci because the normalized entropy difference and \( r^2 \) share a number of similarities. However, the reason why there is a good agreement between pairwise and multilocus approaches is not discussed in those papers. By decomposing the total LD into various lower orders of disequilibria, we could understand the reason better. Using the log-likelihoods from different MC orders for a fixed window size \( w \), we partition the overall departure from LE (\( d_{w,1} \)) into effects of lower order disequilibria such as two-, three-, up to \( w \)-locus disequilibria. Figure 2 illustrates an

---

**Fig. 2.** Partitioning of the total linkage disequilibrium by the contribution from \( \phi_1, \phi_2, \phi_3, \phi_4 \) using a window size 5 in ENm010 region and JPT population. Black represents the contribution from \( \phi_1 \) (effect from adjacent pairs of markers); red represents the contribution from \( \phi_2 \) (effect from adjacent three markers and non-adjacent two markers); green is from \( \phi_3 \) (effect from the adjacent four markers, non-adjacent three markers, and non-adjacent two markers); blue is from the contribution from \( \phi_4 \) (the residual effect which were not included in previous \( \phi_s \)).

*Genet. Epidemiol.*
example of partitioning when a window size 5 is applied to the JPT population in ENm010 region. The average percentage contribution of $f_1$, $f_2$, $f_3$, and $f_4$ across the region is 67, 23, 8, and 2%, respectively. Table 1 shows a comprehensive analysis of average percentage contribution of various orders of LD in the 10 ENCODE regions and four populations. Regardless of genome regions and populations, it is obvious from the analysis that the great bulk of the total disequilibrium is from the contribution of MC1 ($f_1$) which estimates the association between adjacent two markers. Compared to the other three populations, YRI population consistently shows smaller effect of pairwise association in all regions, i.e. having proportionally more high order LD.

MULTILOCUS TOTAL LD AND HAPLOTYPE DIVERSITY

Within high LD regions or haplotype blocks, the diversity of haplotypes is very limited and only a few haplotypes are observed [Daly et al., 2001]. Thus, it is expected that there should be a strong negative correlation between magnitude of total LD ($\delta_{w-1}$) and the number of haplotypes. To find out the relationship between the multilocus total LD and haplotype diversity, we count the total number of distinct haplotypes in each window for a fixed window of size $w$. For moderate window sizes 4–7, mean of total LD is computed for different numbers of haplotypes. We observe that mean of total LD decreases linearly as the number of haplotypes increases. Therefore, there is a clear inverse relationship between magnitude of total LD and number of haplotypes. Figure 3 illustrates this relationship after all ENCODE regions are combined in each population. Window size 5 is used for this analysis. We get very similar results for other window sizes (data not shown).

HIGH LEVEL OF LD VS HIGH ORDER OF LD

One of misunderstandings about LD measures is that high level of LD implies high order of LD which describes the disequilibrium among alleles at more than two loci. We were interested in determining whether high total LD regions necessarily have high order of LD and found that this is not always the case. Methods based on the pairwise LD measures can identify regions with high level of LD, but not...
high order of LD because LD is measured only between pairs of loci. From Figure 2, it was noticed that the contribution of MC1(\(\phi_1\)) tends to be larger in high total LD regions as opposed to the contribution of higher MC orders (\(\phi_2, \ldots, \phi_{w-1}\)) in those regions. To see if we can detect this pattern in other regions and populations, we divide the whole region into three groups (low, medium, and high haplotype diversity) depending on the number of distinct haplotypes. For a window of size 5, if the number of distinct haplotypes is 2–4, the window is categorized as a low haplotype diversity or high LD region; if the number of distinct haplotypes is 5–7, as a medium haplotype diversity or medium LD region; if the number of distinct haplotypes is \(\geq 8\), as a high haplotype diversity or low LD region. Supplementary Table 3 shows that in all categories of haplotype diversity, the effect of \(\phi_1\) or adjacent pairwise marker association accounts for the most variation in the total LD followed by \(\phi_2, \phi_3, \phi_4\) in this order. As haplotype diversity increases, however, the relative contribution of adjacent pairwise association reduces significantly while the relative contribution of higher MC orders increases. The explanation is that detection of higher MC orders requires many distinct haplotypes. A large number of distinct haplotypes observed in low LD regions can increase the likelihoods of higher MC orders. Similarly, detection of high order of LD requires many different haplotypes and large sample size.

**DISCUSSION**

LD between loci is defined as the deviation of the haplotype frequencies from their expectation under independence. Our approach characterizes multilocus LD using a multiple order Markov Chain model in a sliding window. Different orders of Markov Chain model in a window produce different log-likelihoods, with the smallest log-likelihood from the independent (MC0) model to the largest log-likelihood from the full haplotype (MC\(w-1\)) model. Our multilocus LD measure \(\hat{\delta}_{w-1}\) is based on the log-likelihood difference between the full haplotype model (MC\(w-1\)) and the independent model (MC0) for a certain window size \(w\). If the log-likelihood of the full model is more deviated from the independent model in one region than other regions, it indicates that there is higher LD in the region. That is, \(\hat{\delta}_{w-1}\) can measure the extent of the total local LD for multiple markers.

In addition to measuring the total LD in a local region, our method further partitions the total disequilibrium \(\hat{\delta}_{w-1}\) to lower order disequilibria of two-, three-, etc., and all-marker association in each window. From the approximation of \(\frac{1}{N}[\text{LL}(\text{MCw-1}) - \text{LL}(\text{MC0})]\), we can see that the total multilocus disequilibrium can be decomposed into all combinations of lower order disequilibria. However, the relative contribution from each of lower order disequilibria is quite different depending on the extent of total multilocus LD. We observe that in high total LD regions, a great proportion of the total multilocus disequilibrium is accounted for by adjacent two marker associations. Contrarily, in low total LD regions, the contribution due to adjacent two marker association is reduced a lot, and high orders of LD such as three or more marker associations become more noticeable. This fact implies that the extent of total multilocus disequilibrium does not necessarily indicate the orders of LD. Paradoxically, high level of LD in a region does not mean a high proportion of high-order LD. The reason is that high-level LD is usually associated with low number of haplotypes and that detection of high-order LD requires a large sample size and many diverse haplotypes. Thus, in small samples, the estimates of high-order LD are not quite stable. We should use caution when interpreting the estimates of high-order LD, particularly in small samples.

The log-likelihood difference between MC0 model and MC \(w-1\) model is partitioned into components for the log-likelihood difference between two consecutive MC order models. We estimated the proportional effect of each component, but did not explicitly test the significance of it. The results of the test of each component will be published elsewhere. When the log-likelihood difference between MC0 model and MC \(w-1\) model is completely partitioned, there are \(2^w-(w+1)\) LD coefficients with varying orders. We can test a hypothesis whether \(w\) markers are in LE using a \(\chi^2\) distribution with \(2^w-(w+1)\) degree of freedom if all haplotypes are observed. If the hypothesis is rejected, our next step is to find which LD coefficients are significant and which are not. However, due to finite sample size, many haplotypes for multiple markers are usually unobserved. This imposes significant problems for the test of high-order LD coefficients. In general the \(\chi^2\) approximation performs reasonably well for the statistical test of some coefficients, such as pairwise LD, in a sample with reasonable size, such as HapMap samples. But, our preliminary result indicates that it is not appropriate to use \(\chi^2\) approximation for the test of high-order LD coefficients unless sample size is very large. There is a need for an alternative approach for the statistical test.

Our idea of partitioning the total LD into various components of lower order disequilibria is related to Smouse’s method (1974) but overcomes a critical computational problem associated with his method. In his log-linear approach, a series of multiplicative models with different numbers of disequilibrium terms are first constructed. Then, the difference in deviances for two models that differ only by whether a particular disequilibrium term is included.
provides a $\chi^2$ test statistic for that term. However, a critical computational problem arises when one or more haplotypes are unobserved (with zero observed frequency for the haplotypes). In this case, one is faced with the problem of $\log(0) = -\infty$ and the subsequent tests cannot be performed. In contrast, our approach does not need to observe all the possible multilocus haplotypes to measure the multilocus total LD and estimate the effects from lower order disequilibria because the likelihoods of different orders of MC model are computed using the observed haplotypes. Needless to say, this property is very useful in small samples.

Large-scale LD profiles show reasonable agreement between our multilocus method and traditional pairwise LD measures. However, we expect that fine-scale LD patterns can be somewhat different between the two methods. For example, the number of blocks and block boundaries can be different from method to method. Ke et al. [2004] compared different methods of block definition with respect to number of blocks, average block length, and proportion of sequence contained within blocks. They found that there does not appear to be a strong convergence of block-detection methods. Since the true underlying block structure is unknown, it is difficult to compare which method is more appropriate than the other. Blocks are heavily dependent on the factors related to the sample—SNP density, sample size, and marker selection [Ke et al., 2004; Sun et al., 2004; Nothnagel and Rohde, 2005].

A large number of markers can be used in a sliding-window theoretically, but the number of markers in a window should be limited in practice. As the number of markers in a window increases, the number of possible haplotypes can increase dramatically, but we cannot observe many of them in small samples. Rare long haplotypes which are present in the population but are unobserved in small samples can cause the estimates of LD to be inflated since low numbers of haplotypes falsely indicate high LD. This problem becomes more severe for large windows.

Like most sliding-window approaches, finding an optimal window size to capture the LD pattern accurately over a large region is quite challenging. Window size is negatively correlated with the variability in LD trend. If too small window size is used throughout the whole region, the values of multilocus LD measures are so fluctuated that it is hard to separate high LD regions from low LD regions. Haplotype blocks are also fragmented with only a few SNPs in each block. Applying the definition of haplotype-tagging SNPs within these blocks may be useless. On the other hand, too large window size can introduce other problems such as excessive noise, computer memory problem, and smoothing effect. Nothnagel [2004] recommended medium window sizes such as 4–6 after extensive investigation on this matter.

ACKNOWLEDGMENTS

This work was partially supported by NIH GM45344, UL1 RR024128, and by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number 2005-00754. Y.K. is supported by a University Genomics Science Fellowship and S.F. is supported by NIH grant UL1 RR024128.

REFERENCES


Genet. Epidemiol.


