Am. J. Hum. Genet. 77:988–998, 2005

The Effect of Single-Nucleotide Polymorphism Marker Selection on Patterns of Haplotype Blocks and Haplotype Frequency Estimates

Michael Nothnagel[1,2] and Klaus Rohde[1]

[1]Department of Bioinformatics, Max Delbrück Center for Molecular Medicine, Berlin; and [2]Institute for Medical Informatics and Statistics, Christian Albrechts University, Kiel, Germany

The definition of haplotype blocks of single-nucleotide polymorphisms (SNPs) has been proposed so that the haplotypes can be used as markers in association studies and to efficiently describe human genetic variation. The International Haplotype Map (HapMap) project to construct a comprehensive catalog of haplotypic variation in humans is underway. However, a number of factors have already been shown to influence the definition of blocks, including the population studied and the sample SNP density. Here, we examine the effect that marker selection has on the definition of blocks and the pattern of haplotypes by using comparable but complementary SNP sets and a number of block definition methods in various genomic regions and populations that were provided by the Encyclopedia of DNA Elements (ENCODE) project. We find that the chosen SNP set has a profound effect on the block-covered sequence and block borders, even at high marker densities. Our results question the very concept of discrete haplotype blocks and the possibility of generalizing block findings from the HapMap project. We comparatively apply the block-free tagging-SNP approach and discuss both the haplotype approach and the tagging-SNP approach as means to efficiently catalog genetic variation.

Introduction

Work done in recent years has suggested the existence of distinguished chromosomal areas, or haplotype blocks, in the human genome that are, to some extent, independent of their surrounding areas with regard to linkage disequilibrium (LD) or recombination (Daly et al. 2001; Patil et al. 2001; Subrahmanyan et al. 2001; Dawson et al. 2002; Gabriel et al. 2002). In his commentary, Goldstein (2001) painted the most optimistic picture of a genome that is composed of discrete blocks that are separated by hotspots of recombination. The use of haplotypes of SNP markers at these areas as multiallelic markers with increased heterozygosity in association studies has been proposed (Morris and Kaplan 2002; Zhang et al. 2002*a*; Knapp and Becker 2003). Johnson et al. (2001) suggested distinguishing between block haplotypes by a minimal set of haplotype-tagging SNPs (htSNPs), which would then efficiently describe the variation in the human genome by allowing for genotyping of only a subset of SNP marker loci. The usefulness of htSNPs in disease-association studies has recently been questioned (Crawford et al. 2004; Zhai et al. 2004). LD in a chromosomal region is shaped by a complex interplay of factors—including mutation and recombination rates, population and migration histories, selection, and also chance—many of which are unknown. Block-like chromosomal patterns can have a number of sources, which reflects the many factors that influence LD, and they also might occur stochastically (Subrahmanyan et al. 2001; Wang et al. 2002). The International Haplotype Map (HapMap) project is currently genotyping huge numbers of SNP marker loci in samples of individuals of European, Asian, and African descent to better understand human haplotype structure (International HapMap Consortium 2003). By spring 2005, the project has already accomplished a remarkable average marker spacing of 5 kb genomewide. For selected chromosomal regions, this resolution has been increased even more within the Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium 2004).

Several publications have shown that there are differences between populations in LD patterns and block patterns (Goddard et al. 2000; Kidd et al. 2000; Reich et al. 2001; Gabriel et al. 2002; Hinds et al. 2005; Sawyer et al. 2005). Phillips et al. (2003) demonstrated that marker ascertainment and spacing can explain observed block lengths, and variability in recombination rates, bottlenecks, and selection are not required to this end. The crucial influence of the sample SNP density on the length of method-defined blocks has been demonstrated (Ke et al. 2004). Also, Sun et al. (2004)

Received June 13, 2004; accepted for publication September 16, 2004; electronically published October 19, 2005.

Address for correspondence and reprints: Dr. Michael Nothnagel, Christian Albrechts University, Medical Faculty, Institute for Medical Informatics and Statistics, Brunswiker Strasse 10, D-24105 Kiel, Germany. E-mail: nothnagel@medinfo.uni-kiel.de

0002-9297/2005/7706-0010$15.00

Table 1

ENCODE Regions under Investigation

Region (Position)	Length (Mb)	No. of SNPs (Distance[a]) in Population			
		CEU	HCB	JPT	YRI
ENm013 (7q21.13)	1.11	421 (.66, 1.19 ± .38)	305 (1.09, 1.62 ± .38)	262 (1.32, 1.89 ± .40)	338 (.88, 1.46 ± .42)
ENm014 (7q31.33)	1.16	555 (.48, .90 ± .30)	373 (.71, 1.34 ± .44)	290 (.85, 1.71 ± .61)	419 (.61, 1.19 ± .41)
ENr112 (2p16.3)	.5	599 (.53, .83 ± .22)	543 (.51, .92 ± .29)	531 (.51, .94 ± .30)	541 (.55, .92 ± .26)
ENr113 (4q26)	.5	627 (.40, .80 ± .28)	525 (.46, .95 ± .35)	520 (.48, .96 ± .34)	535
ENr131 (2q37.1)	.5	649 (.40, .76 ± .26)	515 (.53, .95 ± .30)	503 (.54, .97 ± .31)	510 (.52, .97 ± .32)
ENr213 (18q12.1)	.5	430 (.66, 1.15 ± .34)	345 (.77, 1.44 ± .47)	354 (.77, 1.40 ± .45)	370 (.77, 1.35 ± .43)
ENr232 (9q34.11)	.5	359 (.55, 1.39 ± .59)	316 (.59, 1.58 ± .70)	311 (.59, 1.60 ± .72)	319 (.59, 1.56 ± .68)
ENr321 (8q24.11)	.5	399 (.63, 1.25 ± .44)	442 (.61, 1.13 ± .36)	413 (.64, 1.21 ± .40)	411 (.64, 1.22 ± .41)

Note.—SNPs were required to have <5% missing alleles in the sample, and both alleles had to have a frequency of at least 0.1, corresponding to a minimum heterozygosity of 0.18. Differences in SNP numbers between the populations were due to different heterozygosities (in the raw data) and to the subsequent filtering of rare SNPs. The median was used as an additional robust characterization of the skewed SNP distance distributions. The regions have comparable SNP densities among the four populations, CEU, HCB, JPT, and YRI.

[a] Distance values are median, mean ± SD (in kb).

pointed to the impact of sample size and marker selection on the number of haplotype blocks. This effect is predominantly the result of blocks that fall apart into subblocks when more SNPs are added to the sample and of newly detected blocks too small to be resolved by the previous SNP resolution. Rarely, haplotype blocks also disappear with increasing SNP density (i.e., the "flip-flop" effect). Simultaneously, the overall block-covered sequence and the number of blocks increase while the average block length decreases. Thus, detected block structures are preliminary and dependent on the sample. However, it is usually suspected (by the authors of previous publications) that a higher SNP density will resolve this problem and give a more accurate picture of the "true" underlying block structure.

Here, we focus on the block-covered sequence as the part of the genome in which information about the chromosomal sequence is supposed to be neatly summarizable by information on a few haplotypes in a distinct genomic region. If haplotype blocks are a sound concept, they should be detected regardless of the particular choice of SNPs, at least for high marker densities. Various block methods are consistent with regard to block-covered sequence—that is, the part of some genomic sequence that is included in blocks—when SNPs are successively *added* to the sample (unpublished data). In this case, blocks at a lower density are, to a great extent, included in blocks at a higher density. But the block borders are usually not stable, and blocks continue to fall apart into subblocks with growing SNP density. For the current study, we were interested in investigating the effect of different but comparable sets of SNPs in the same region—that is, the influence of marker selection—on the block definition and the observed haplotypic pattern.

Data Sets and Methods

Data Sets

We analyzed 32 publicly available high-quality data sets of eight genomic regions in four populations that were provided by the ENCODE project, to avoid any bias resulting from the use of a particular genomic region or population. We downloaded the genotype data files as of December 2, 2004 (available at the ENCODE Web site). The selected samples had comparatively high SNP densities that were very similar among the populations (see table 1). The data sets contained genotypic information on 30 trios of European American descent from Utah (CEU), 45 Chinese individuals from Beijing (HCB), 44 unrelated Japanese individuals from Tokyo (JPT), and 30 Yoruban trios from Nigeria (YRI). We required each SNP to have a minimum heterozygosity of 0.18.

To assess the effect of marker selection on LD and block patterns in a particular region, we split each data set in half by allocating the first SNP (in order of physical location), the third, the fifth, and so on to the first subset and allocating the second SNP, the fourth, the sixth, and so on to the second subset. In this way, we generated two complementary data sets with interdigitated marker positions that were almost identical with regard to their average marker distance and the genomic region they covered.

Block Algorithms and LD Assessment

We employed three different algorithms for the definition of blocks. Two methods primarily target absent recombination events in blocks. The four-gamete test (Hudson and Kaplan 1985; Wang et al. 2002) defines blocks as areas between consecutive SNPs where one or

more haplotypes of each marker pair have a frequency <0.01. The approach by Gabriel et al. (2002) imposes limits on the CIs of D' (Lewontin 1964); they must be exceeded by at least 95% of all pairs. We used the implementation of both algorithms in HaploView 3.2 (Barrett et al. 2004). The third algorithm aims for elevated levels of multilocus LD within blocks. We used sliding windows of four consecutive SNPs, estimated the haplotype frequencies for these SNPs, and calculated the normalized entropy difference (NED), ε (Nothnagel et al. 2002), as a measure of multilocus LD. Blocks were defined as the union of consecutive windows whose ε values exceeded a threshold of 0.5 (Nothnagel 2004). This algorithm and the succeeding analysis (designated "NED(4;0.5)") were implemented in C, Perl, and R (R Development Core Team 2004).

Haplotype frequencies were estimated using an expectation-maximization algorithm for trios in which information on the children is used only to infer the phase of the parents (Becker and Knapp 2004). We further normalized ε analogously to D': if m markers are considered, ε reaches its maximum of $(m-1)/m$ if exactly two haplotypes are present (Nothnagel 2004). We therefore defined $\varepsilon' = \varepsilon/[(m-1)/m]$, which assumes a value between 0 (linkage equilibrium) and 1 (highest possible disequilibrium), and used it to comparatively describe multilocus LD with different numbers of SNPs.

Assessment of Block Concordance

We wanted to investigate the effect of marker selection while keeping the marker density and the covered area comparable. We therefore repeatedly generated pairs of data subsets with complementary, interdigitated SNP sets of thinned density. First, a subset of the full SNP set was randomly selected, with each SNP having a probability of being selected of 100%, 50%, 20%, and 10%. Second, the selected SNP set was split by alternating division into two subsets, as described above. Thus, the resulting SNP sets contained a SNP from the full set with a probability of 50%, 25%, 10%, and 5%, respectively (i.e., the thinning level). We generated 100 replications for each thinning level, population, and region, except for the 50% level for which the full SNP set was split only once.

We assessed the concordance in block definition between the two subsets as the portion of the total physical length of chromosomal sequence included in blocks by at least one subset (S_{union}) that was also included in blocks by both subsets ($S_{\text{intersection}}$). The ratio $S_{\text{intersection}}/S_{\text{union}} \in (0,1)$ was averaged over all replications. This measure focuses on block-covered sequence. For an explicit description of block border similarity, we employed the measure SB_2 by Liu et al. (2004), which can assume values between 0 and 1. In short, this measure describes how well a block partition in a particular region is matched by another, with 1 indicating a perfect match. It focuses on the block borders and takes only the number of SNPs within blocks into account, not their physical distance. It favors longer blocks shared between the partitions over shorter ones. Since SB_2 is not symmetric, we report the average over both directions of projection.

Assessment of Haplotype Concordance

We introduce a new measure for LD, ϕ', that measures the correlation in the haplotypic structure between two subsets of SNPs. To this end, we consider the subsets as two multiallelic markers and consider the resultant haplotypes as their possible alleles. Again, we use the entropy concept to describe the state of these systems: $S_i = -\sum_k p_k^{(i)} \log p_k^{(i)}$ $(i = 1,2)$, where $p_k^{(i)}$ denotes the frequency of the kth haplotype in subset i and the summation is over all occurring haplotypes. The joint occurrence of these two markers forms the haplotypes of the full SNP set—that is, the union of the two subsets—and the corresponding entropy is S_{F}. Under independence, S_{F} would simply equal the entropy sum of the subsets: $S_{\text{I}} = S_1 + S_2$. On the other hand, the minimum value for S_{F} is $\max(S_1,S_2)$. In analogy to ε, we normalize the difference between the entropy expected under independence and the observed entropy by the maximum possible entropy: $\phi = (S_{\text{I}} - S_{\text{F}})/S_{\text{I}}$. It is easy to prove that $\phi \in (0,0.5)$, with $\phi = 1/2$ if the haplotype frequencies in all three SNP sets are identical. We therefore scale the measure to the standard interval: $\phi' = \phi/0.5$. Thus ϕ' is 0 if the two SNP subsets are in linkage equilibrium, and 1 if they are in perfect LD.

TagSNP-Based Prediction

For a comparative analysis of LD-based but block-free concepts for describing genetic variation, we also applied the tagging approach suggested by Carlson et al. (2004). This approach targets networks, or bins, of SNPs that are in high LD. We used the implementation in HaploView 3.2 (Barrett et al. 2004) that selects the minimum set of tagging SNPs (tagSNPs) in such a way that the r^2 value with respect to all SNPs in the sample is greater than a specified cutoff for at least one of those tagSNPs. We used the standard pairwise tagging option and cutoffs of 0.5, 0.8, and 1.0. We then assessed the ability of the tagSNPs to predict the SNP alleles in the complementary set as the portion of all SNPs that have an r^2 value, with respect to at least one of the tagSNPs, that is greater than the tagging threshold.

Results

LD and Block Structure

Figure 1 illustrates the $|D'|$ values for all pairs within the two complementary SNP subsets in the European

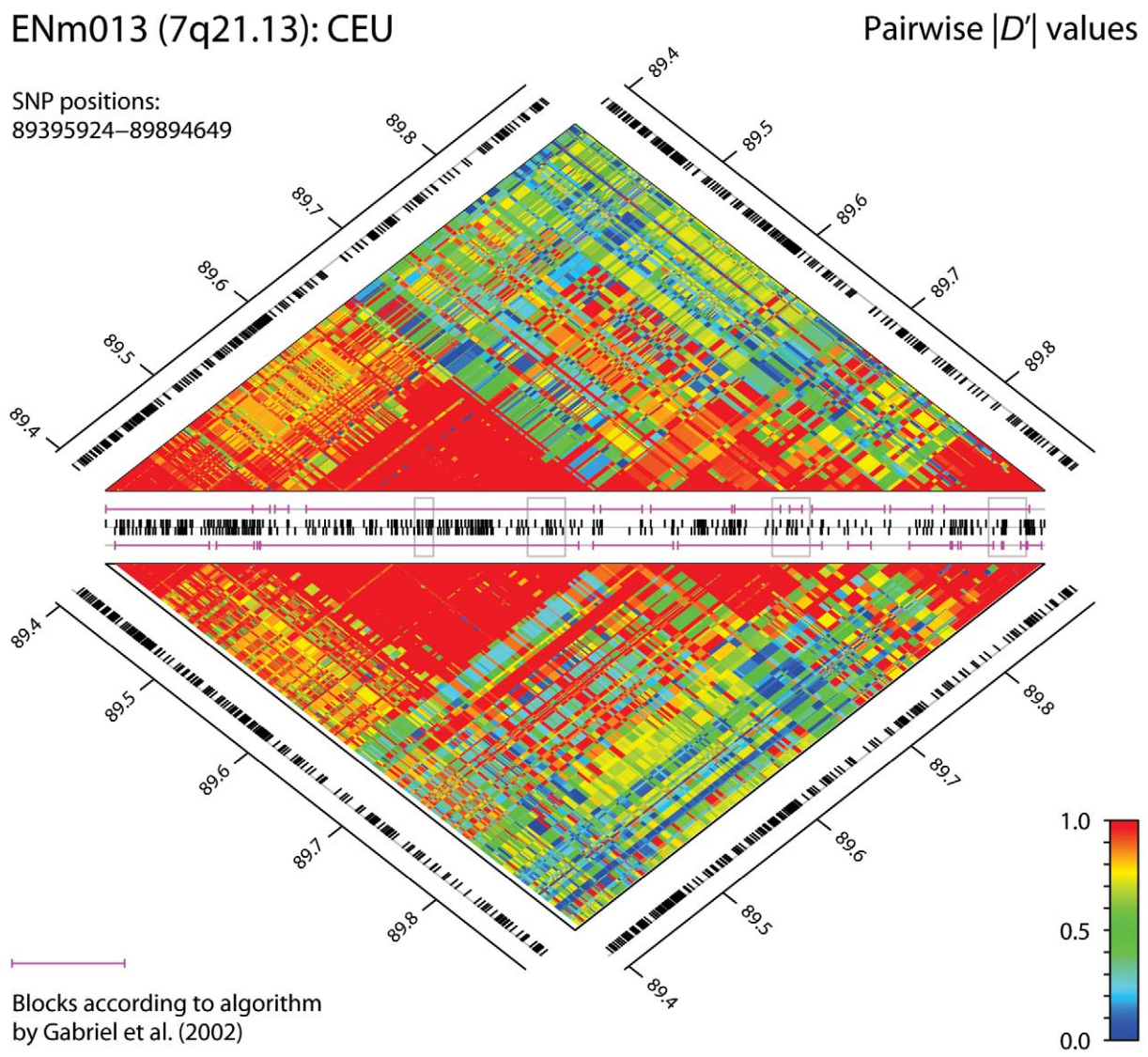


Figure 1 Pairwise $|D'|$ values and standard Gabriel block definition for the pair of complementary SNP subsets (50% level) in the CEU sample of the ENm013 region

sample for the ENm013 region and the blocks defined by the standard Gabriel method. Although the broad picture of LD was very similar for the two sets, the fine structure was different, and so was the block structure. Results for r^2 and for other regions and populations were similar and are provided, together with a description of the distances between SNPs in the subsets and the distribution of their less frequent alleles, at our Web site (see authors' Web site in Web Resources).

Figure 2 graphs the average block-coverage concordance between complementary SNP subsets for all considered regions, populations, methods, and thinning levels. The strong discrepancies between the block-covered sequences for sparse resolutions with an average SNP spacing $\geq$15 kb could be expected. However, it is surprising to observe average concordances near or below 75% and often <50% for all methods, populations, and regions, even for SNP maps with an average spacing $\leq$2 kb. More-stringent thresholds for the Gabriel approach and for the NED approach did not improve these numbers (not shown). Although the concordance tended to increase with growing SNP density, the picture of block-covered sequence still showed inconsistencies for SNP resolutions several times higher than the current HapMap resolution. Discrepancies tended to be greater in the African samples.

Table 2 reports the similarity in block borders between the two interdigitated SNP subsets at the 50% thinning level for the Gabriel algorithm. Results for the four-gamete test and for NED(4;0.5) were similar (data not

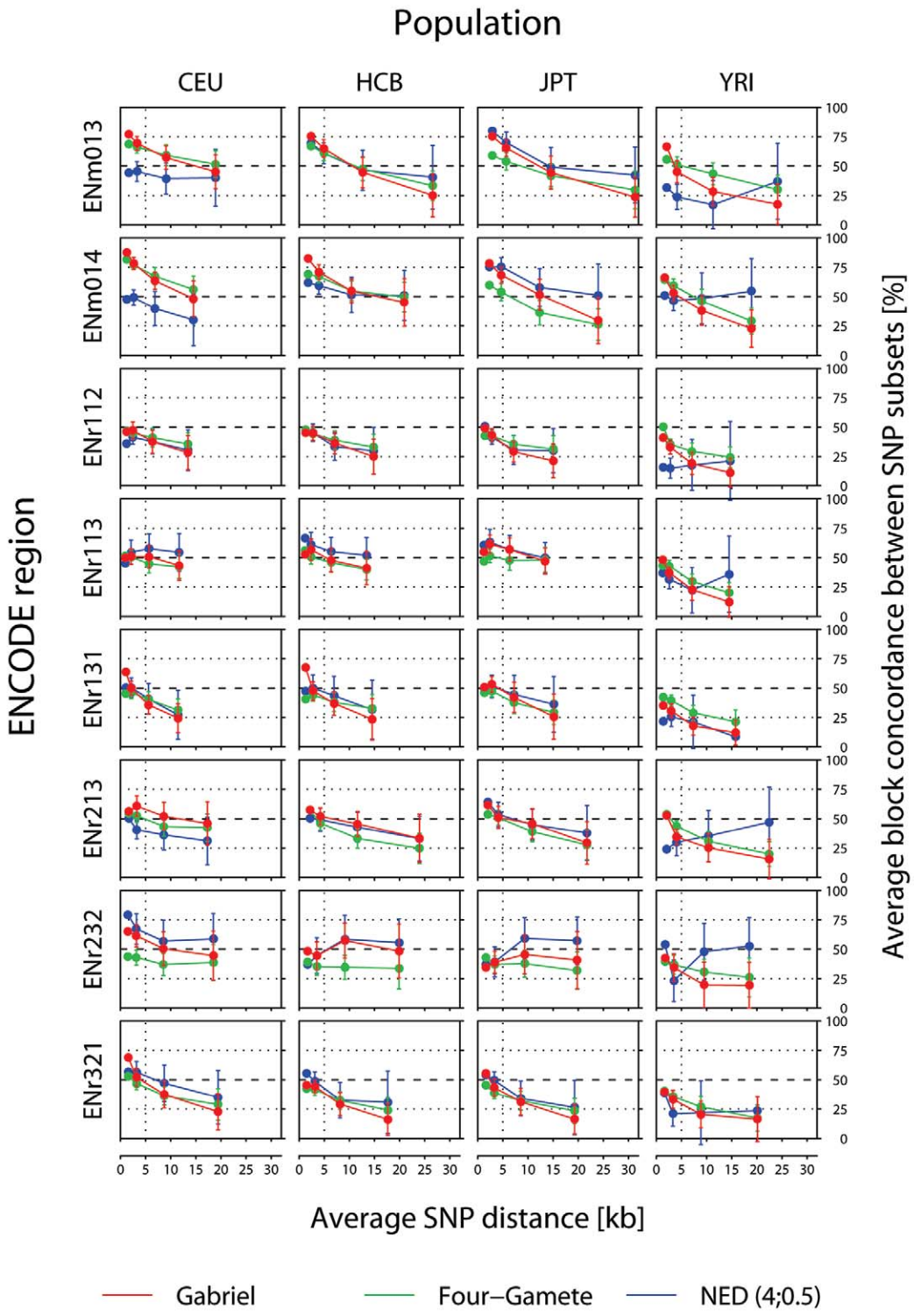


Figure 2 Average concordance in block-covered sequence for three block methods in pairs of complementary SNP subsets. The mean and SD of the concordance over all replications at the different thinning levels are graphed with respect to the average median distance between adjacent SNPs in both sets, separately for each region and population. The concordance for the 50% level was calculated only once. All regions, populations, and methods considered show a similar picture of improved concordances with increasing marker density, but these are usually <75% and often average <50%. In a few cases, predominantly in the African samples, concordance even decreased with increasing marker density. The vertical dotted line within each graph marks the current HapMap resolution of ~5 kb.

Table 2

Block Border Concordance

Region	Similarity of Block Borders: SB_2 (P)			
	CEU	HCB	JPT	YRI
ENm013	.69 (.008)	.70 (.001)	.73 (.001)	.55 (.502)
ENm014	.85 ($<10^{-4}$)	.75 ($<10^{-4}$)	.69 ($<10^{-4}$)	.66 ($<10^{-4}$)
ENr112	.68 ($<10^{-4}$)	.60 (.048)	.65 ($<10^{-4}$)	.62 ($<10^{-4}$)
ENr113	.78 ($<10^{-4}$)	.74 ($<10^{-4}$)	.83 ($<10^{-4}$)	.68 ($<10^{-4}$)
ENr131	.72 ($<10^{-4}$)	.64 (.003)	.61 (.021)	.62 (.003)
ENr213	.86 ($<10^{-4}$)	.73 ($<10^{-4}$)	.80 ($<10^{-4}$)	.66 (.001)
ENr232	.79 ($<10^{-4}$)	.67 (.001)	.65 (.005)	.62 (.012)
ENr321	.79 ($<10^{-4}$)	.71 ($<10^{-4}$)	.65 (.008)	.61 (.021)

Note.—The similarity between block borders defined by the Gabriel algorithm in the two complementary SNP subsets (50% level) are shown for the ENCODE regions under investigation. For each region and population, the similarity between the block partitions was measured by SB_2 (see the "Data Sets and Methods" section). Empirical P values for the nonrandomness of the observed similarity are given in parentheses.

shown). Although the block borders show similarities that are, in most cases, significantly different from those expected by chance, these similarities are comparable to the observed concordance in block-covered sequence (fig. 2).

Underlying Haplotype Structure

To examine the underlying haplotypic structure more closely, we selected four exemplary areas in the CEU sample for the ENm013 region that extended over 10–20 kb and contained 13–15 SNPs each. Two areas showed perfect concordance in block definition, whereas the other two showed inconsistencies (see gray boxes in fig. 1). For each area, we estimated the haplotype frequencies in each SNP subset separately and in both sets jointly (see fig. 3).

Area 1 represents the almost ideal case: both SNP sets gave an identical block pattern and the same haplotypic picture, with the minor exception of two SNPs (*rs42610* and *rs42618*) that split the rarest haplotype into a group of three. This is the picture that we would expect to see in a haplotype block. Area 2 also showed also an identical block pattern and four common haplotypes; however, their frequencies differed substantially. We could not establish a perfect match between the subset haplotypes as we did for area 1. Depending on the SNP set we use in area 2, we get different haplotype frequencies for subsequent analyses.

In area 3, the block patterns and the frequencies of the common haplotypes differed between the two subsets; although the area was defined as a single block in one set, there are several interruptions in the other set. But the haplotype groups—consisting of one common haplotype and one or more very similar rare haplotypes—still matched well, with one group being mapped to two in one case. In area 4, the situation deteriorated further, with differences in the block pattern and haplotype frequencies and only a partial match between the haplotype groups in the two SNP subsets.

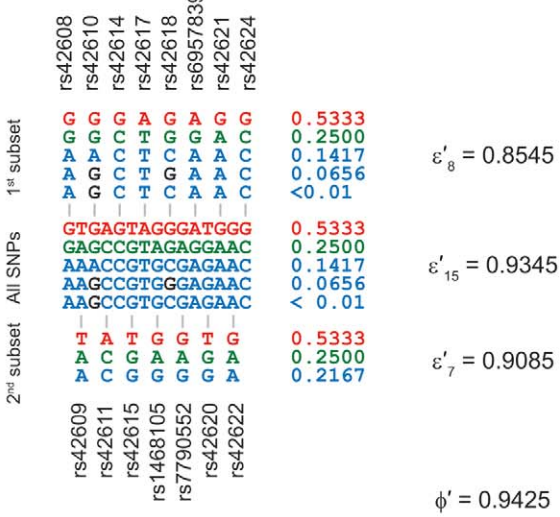
Multilocus LD decreased in the series of these four areas. But all areas showed elevated levels of LD—that is, they were highly structured and far from being random combinations of the single-SNP alleles. To quantify how representative the four exemplary regions are of the investigated regions, we employed sliding windows of 5 and 10 SNPs, thus representing situations of both nesting and overlapping SNP sets. We used ϕ' to assess haplotypic correlation between the two complementary subsets within these windows and assessed the physical coverage of each region by windows whose ϕ' values equaled or exceeded a specified threshold. Table 3 lists the results for the most stringent threshold, 1.0 (the most interesting case, when the haplotype frequencies in all three SNP sets are identical), and for 0.9 as the threshold separating area 1 from the other three. The portions of identical haplotype patterns differ considerably between the regions and populations. Although areas of identical haplotype patterns in the subsets are not common, extensive areas show high correlations. The African samples persistently showed much smaller portions, but there was no clear trend for the other three populations. For region ENm013 in the CEU sample, less than one-sixth of the SNP sequence gave identical haplotype patterns, despite a block concordance >75%. The percentage of a region with identical haplotype patterns was strongly correlated with the average of ε'_4 in the full SNP set (0.86) across the regions and populations but was less so with the average r^2 value (0.47). This reflects the local LD assessment by ε'_4.

In summary, our results demonstrate that it is difficult, if not impossible, to reliably infer the underlying haplotype structure from the block structure. Only for the highest values of LD is such an inference possible, but, even then, a consideration of additional SNPs might destroy the well-ordered picture. Depending on the chosen SNP set, we might get different haplotype block patterns and different haplotype frequencies, even at high marker densities.

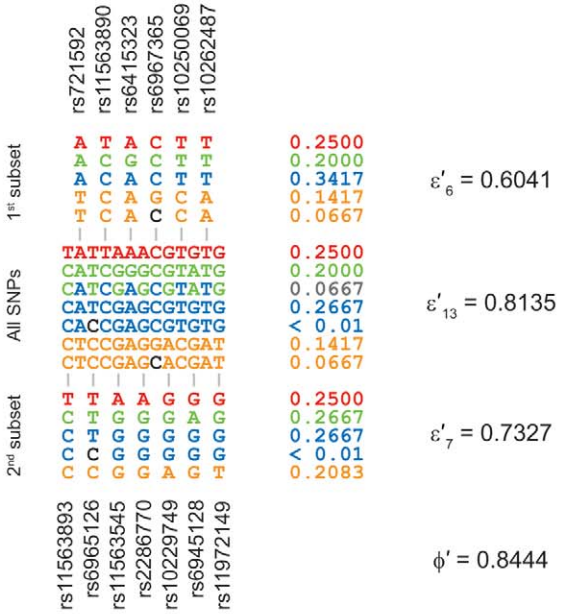
TagSNP Prediction

Although our analysis demonstrated the fragility of haplotypes and of haplotype blocks, SNPs are often not independent, and dependencies between them need to be considered. One way to discard the rigid concept of blocks but exploit SNP associations is to consider networks of SNPs that leave in-between SNPs aside and do not claim to be representative of them. Carlson et al. (2004) recently proposed such an approach to tagSNP selection for association analysis. We wanted to assess

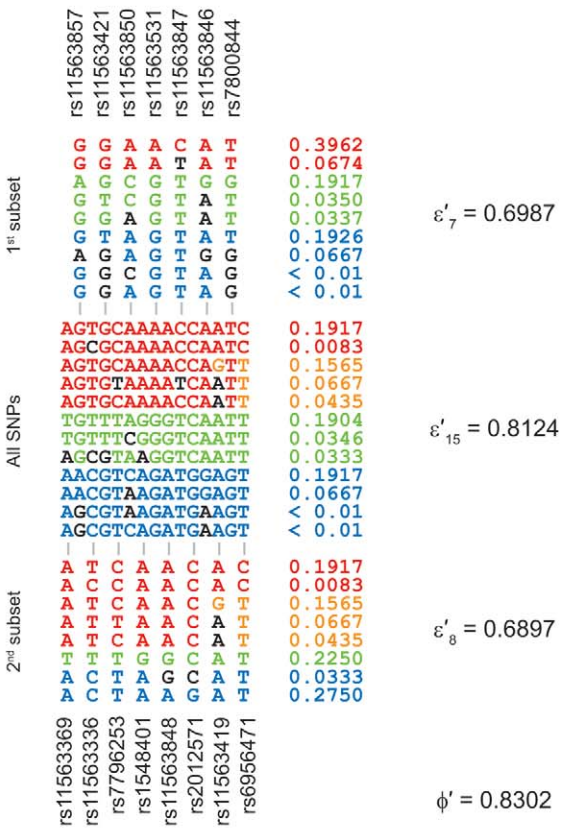
Area 1: 89.56-89.57 Mb



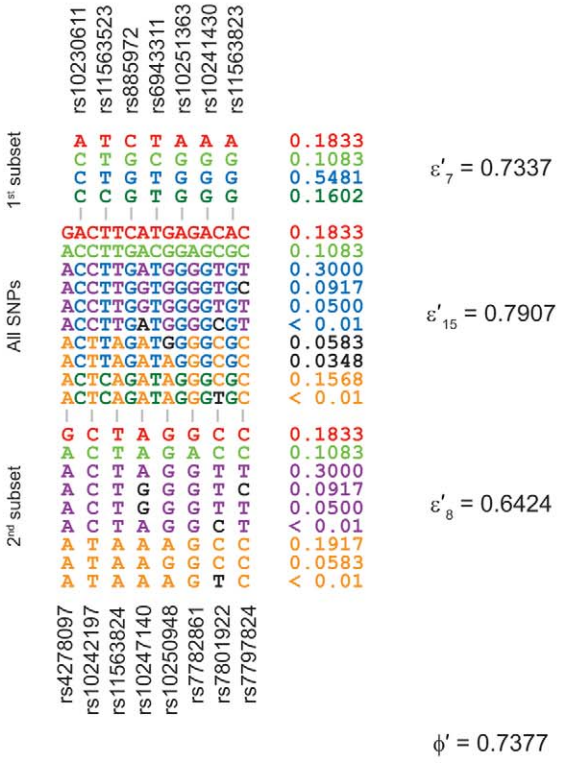
Area 2: 89.62-89.64 Mb



Area 3: 89.75-89.77 Mb



Area 4: 89.865-89.885 Mb



how well tagSNPs in one subset can predict SNPs in the complementary set. Since the results for both directions between the two subsets were highly similar, table 4 lists the average values of both directions.

A cutoff of 0.5 for r^2 yielded successful predictions of usually >90% in the complementary SNP subset, with the exception of the African samples, which showed markedly lower percentages. A more stringent threshold of 0.8 led to an ~50%–100% increase in the number of tagSNPs. Simultaneously, the success in predicting the complementary SNP set declined, but the percentages were still, with few exceptions, well over 80% in the non-African samples. For the most stringent cutoff of 1.0—that is, with no loss of information on genetic variation at all—between one- and two-thirds of all markers were needed for tagging in the non-African samples, and even more in the African ones. Prediction of complementary (i.e., unknown) SNPs further deteriorated and ranged from ~50%–85% in the non-African samples. So, although we could reduce genotyping efforts by one- to two-thirds for SNP samples of the given density in these populations, we could not predict with certainty 15%–50% of the SNPs that lie in between the sample SNPs. The situation was worse for the African samples.

Table 3

ϕ Regions for the Four Populations

	Physical Percentage of Region with							
	$\phi' = 1$ (Identical Haplotype Patterns)				$\phi' \geq .9$			
Region	CEU	HCB	JPT	YRI	CEU	HCB	JPT	YRI
ENm013	14.7	11.4	33.0	3.9	30.5	41.3	49.2	11.5
ENm014	17.7	28.0	16.5	12.0	30.9	52.0	55.0	20.5
ENr112	18.4	20.8	18.2	.2	36.9	44.2	40.6	4.2
ENr113	16.4	35.3	31.9	8.0	41.9	53.9	67.2	22.0
ENr131	13.7	16.5	18.7	3.4	33.7	30.1	32.7	12.1
ENr213	17.2	30.8	25.5	4.7	46.0	54.8	49.1	9.4
ENr232	31.9	7.1	12.6	5.3	53.2	21.9	20.5	10.8
ENr321	20.0	16.1	17.5	6.8	53.1	49.7	43.4	17.0

Note.—Portions of the ENCODE regions under investigation with elevated levels of haplotypic association between the complementary SNP subsets (50% level) are shown. A five-SNP sliding window was used along the sequence. Within each window, the SNP set was split into two interdigitated subsets, and ϕ' was calculated. Listed is the physical portion of the SNP sequence that is covered by windows with $\phi' = 1$ (i.e., identical haplotype frequency patterns in both complementary subsets and the full set) and $\phi' \geq 0.9$. Results for the 10-SNP window differed, with few exceptions, by only a few percentage points for the threshold 1.0 but differed considerably for 0.9 (not shown). These numbers are therefore only a rough assessment of the covered sequence. African samples consistently showed a much smaller portion of identical haplotype patterns than did the other three populations. Almost all the sequence in all regions and populations showed ϕ' values ≥ 0.5 (not shown).

Discussion

Recent work on haplotypes blocks demonstrated the crucial influence that the population, sample SNP density, and sample size have on detected block patterns and haplotypic structure. In the present work, we studied the effect of marker selection on the detected block structure. To this end, we used high-density SNP data sets that were complementary but otherwise nearly identical with regard to covered sequence and marker distance. We employed various methods for the definition of haplotype blocks in a series of real data sets from different genomic regions and populations to avoid any bias resulting from a particular method, chromosomal region, or population. We focused on the criterion of block-covered sequence as the area for which information reduction is potentially feasible.

We were able to demonstrate that the choice of markers for an analysis can have a profound effect on the pattern of haplotype blocks and haplotype frequencies that we see in a particular region. Although large-scale LD patterns were usually similar between interdigitated SNP sets, haplotype patterns and LD fine structure was not. Haplotype patterns are fragile and make sense only in a fine-structure view. Regions of identical haplotype patterns in the complementary SNP subsets do exist, but they are not common and might break up if more SNPs are considered. Depending on the marker set, we might get different blocks or common haplotypes that differ in frequency or even in number between subsets, with unknown consequences for subsequent analysis steps, such as association analysis or htSNP definition. This holds true even for high SNP densities $\leq$2 kb.

The underlying reason for this complicated picture is presumably the complex interplay of factors that shape LD and whose magnitude is usually not known or not very well known. This includes the different ages of mutations; the potential differences in local recombination rates; the repeated population history events of bottlenecks, migration, and genetic drift; selection; and other factors. Extensive simulation studies to determine

Figure 3 Haplotype patterns in the complementary SNP subsets (50% level) of the CEU sample in four exemplary areas of the ENm013 region. The graphs detail the haplotypic structure seen in the gray boxes of figure 1. The frequency estimation was used for each SNP subset separately and both sets together. LD was assessed separately for each SNP set (ε') and between the SNP subsets (ϕ'). The upper two areas had identical block patterns in the subsets, whereas the lower two differed. Only area 1 met the expectation of a consistent haplotype block, showing nearly identical patterns of blocks and frequencies and also the highest LD values of the four areas. The haplotype frequency changes in areas 2 and 4 were due to rare entwined haplotypes in the full SNP set.

Table 4

TagSNPs and Prediction of SNPs in the Complementary Subset for the Four Populations

	CEU		HCB		JPT		YRI	
r^2 VALUE AND REGION	TagSNPs (%)	SNPs Predicted (%)	TagSNPs (%)	SNPs Predicted (%)	TagSNPs (%)	SNPs Predicted (%)	TagSNPs (%)	SNPs Predicted (%)
$r^2 \geq 0.5$:								
ENm013	10.2	96.4	10.8	96.4	10.3	95.0	23.7	86.7
ENm014	13.2	94.8	13.7	96.0	12.8	92.4	27.2	84.5
ENr112	14.0	93.7	14.9	96.3	15.8	96.0	27.7	81.5
ENr113	12.1	91.1	8.0	96.4	7.7	97.3	31.4	84.7
ENr131	20.3	90.1	25.8	84.3	25.2	88.5	42.4	76.5
ENr213	15.8	96.0	15.1	97.7	15.5	93.2	37.0	78.9
ENr232	18.9	91.1	32.9	82.0	30.5	85.5	54.2	66.1
ENr321	16.8	94.0	19.2	93.7	18.6	92.5	40.1	82.0
$r^2 \geq 0.8$:								
ENm013	21.1	86.5	23.9	91.1	26.0	89.3	43.2	76.9
ENm014	23.8	88.5	22.8	92.5	24.1	83.8	44.4	75.4
ENr112	27.9	86.1	24.7	90.8	29.4	89.5	53.8	64.7
ENr113	21.4	86.8	15.0	94.3	15.2	93.1	48.4	71.2
ENr131	33.9	86.4	42.5	77.7	41.4	77.5	61.0	60.4
ENr213	25.3	90.7	29.0	91.3	29.1	88.4	56.8	62.7
ENr232	31.5	85.8	49.7	70.6	49.8	70.7	73.0	52.4
ENr321	28.3	89.0	33.3	86.7	35.1	87.6	61.6	62.5
$r^2 = 1.0$:								
ENm013	36.3	78.1	43.3	77.7	42.0	77.1	68.9	53.6
ENm014	36.6	79.6	38.3	79.3	45.9	74.8	64.2	58.0
ENr112	52.1	61.9	49.5	75.7	52.0	72.1	82.1	33.6
ENr113	42.3	70.8	27.6	87.4	27.3	86.5	66.7	53.8
ENr131	52.7	59.0	57.1	63.3	55.5	65.0	73.7	44.3
ENr213	43.7	75.1	47.8	74.8	47.5	74.3	75.9	41.4
ENr232	54.6	64.1	70.3	51.6	65.3	53.7	84.3	29.5
ENr321	51.1	70.9	58.4	66.7	56.4	68.0	79.3	41.9

NOTE.—The average percentage of tagSNPs in the two interdigitated subsets (50% level) of the ENCODE regions under investigation and the average success in predicting the complementary subset are shown. The table lists the percentage of all SNPs in one subset that were identified as tagSNPs with the use of different r^2 cutoffs and the percentage of all SNPs in the complementary subset that they predict, averaged over both subsets for each region and population. For example, in the CEU sample of region ENm013, 10.2% of all SNPs in a subset are tagSNPs, and they predict 96.4% of the complementary SNPs with an accuracy of $r^2 \geq 0.5$. Percentages of tagSNPs were persistently higher in the African samples, whereas the prediction of complementary SNPs was less successful, compared with the non-African samples.

the shaping factors in each region and population—as were done, for example, by Phillips et al. (2003)—would be desirable but are beyond the scope of the present study. Missing knowledge of the factors that shaped LD in a particular region is the situation that a "gene-hunting" researcher usually confronts. Only for regions with high haplotypic similarity between SNP subsets can we assume recent bottlenecks with absent or very few successive recombination events or strong selective pressure as likely sources. This is consistent with the observations of a lower concordance with regard to block coverage and block borders, lower percentages of identical frequency patterns in the SNP subsets, and higher proportions of tagSNPs with lower prediction success in the African samples. Our results confirm previous reports of higher genomic variability and a smaller average extent of LD in African populations than in other populations.

Our results raise the question of how useful and reliable the concept of haplotype blocks is. This concept postulates that the essential variation at a block as a chromosomal segment with limited diversity can be described by the allelic combinations of some of its (common) SNPs. But the sample only provides information on these SNPs; inference about the variation located in between these SNPs is not reliably possible. The definition of haplotype blocks is "soft" in the sense that one does not know what the underlying structure looks like—the notion of haplotype blocks as delineating regions of high LD is wrong because a single additional SNP can destroy this pattern by splitting a "solid" block into two parts. Haplotype structure is fragile; one cannot predict the outcome for blocks and haplotype patterns when additional SNPs or other SNPs than those in the given sample are used. We see "patterns of long segments of strong LD" (Hirschhorn and Daly 2005, p. 98) in the investigated ENCODE regions, but "blocks" are dependent on—besides the population and

the method used—the chosen SNP sample. So are the underlying SNP haplotype structures. Consequently, we need to question the aim of the HapMap project to catalog the human haplotypic variation in discrete blocks. As we and other authors have shown, blocks are preliminary at any given SNP density. The term "block" creates the illusion that a chromosomal region can be segmented into clear disjointed blocks with regard to recombination, LD, or some other feature, which is true only in extraordinary cases.

Nevertheless, many SNPs in the investigated high-density ENCODE regions were in high LD with each other, and the haplotypes in the SNP subsets always showed associations with one another. Block-free tagSNPs in one subset predicted SNPs in the complementary set remarkably well for intermediate r^2 cutoffs, but less so for complete LD. Thus, both the haplotype approach and the tagSNP approach, as means to condense information on genetic variation, face the same problem: both approaches aim to limit the loss of information and both cannot predict the effect of differing SNP density and selection. Although additional SNPs might create additional haplotypes and destroy the picture of an LD block, they can also lead to an increase in the number of required tagSNPs. Both approaches can only describe variation with a given marker selection. The tagSNP approach is more robust in the sense that sets of existing tagSNPs are only expanded by additional ones, with growing marker density, but it also ignores simultaneous correlations between multiple markers, as is inherent in a haplotype consideration. We expect that information on additional SNPs will affect LD maps (Zhang et al. 2002*b*) as well as recombination maps, because of the underlying problem of missing data on genomic variation. On the other hand, genotyping *all* genetic variation in a region would lead to a complete and final picture of haplotypes and tagSNPs without any more changes.

Condensing the information on variation in a genomic region is therefore a trade-off between the accuracy of description—that is, how much loss of information is acceptable—and the genotyping efforts needed to achieve this accuracy. Given the aim of the HapMap project to provide a high-quality reference for the common human haplotypic variation, a complete genotyping of all genetic variation and a subsequent condensation of this information appears to be the only reliable way to achieve this aim; any map based on noncomplete genotyping will be preliminary. For association studies with single markers, a genomewide tagSNP map with an r^2 cutoff <1.0 will presumably suffice.

At high SNP densities, a number of rare haplotypes differed from a common, or major, haplotype by only one or a few SNP alleles. For a description of the essential variation in a particular area, one could consider the use of clustering algorithms or the construction of haplotypes from tagSNPs. Both approaches would presumably lead to a comparable result, but this needs further investigation. The pattern of LD in a particular region is often complex, and any description needs further verification.

Acknowledgments

This work was supported by German National Genome Research Network grant NGFN 01GR0463. We thank Nianjun Liu for providing the source code for the calculation of the block border concordance measure. We also thank two anonymous reviewers for their thoughtful and extensive comments.

Web Resources

The URLs for data presented herein are as follows:

Authors' Web site, http://capella.uni-kiel.de/hapmap/supplements.htm (for results for r^2 and for other regions and populations, as well as a description of the distances between SNPs in the subsets and the distribution of their less frequent alleles)

ENCODE, http://www.hapmap.org/genotypes/2004-12/ENCODE/

References

Barrett JC, Fry B, Maller J, Daly MJ (2004) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265

Becker T, Knapp M (2004) Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 27:21–32

Carlson CS, Eberle MA, Rieder MJ, 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

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 TJ, 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, et al (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418: 544–548

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640

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

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216–234

Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–111

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

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

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Genova DG, 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

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet 13:577–588

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH,* in a global representation of populations. Am J Hum Genet 66: 1882–1899

Knapp M, Becker T (2003) Family-based association analysis with tightly linked markers. Hum Hered 56:2–9

Lewontin R (1964) The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49:49–67

Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, Kidd KK, Brookes AJ, Zhao H (2004) Haplotype block structures show significant variation among populations. Genet Epidemiol 27:385–400

Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233

Nothnagel M (2004) The definition of multilocus haplotype blocks and common diseases. PhD thesis, Humboldt University, Berlin (http://edoc.hu-berlin.de/dissertationen/nothnagel-michael-2004-12-13/PDF/Nothnagel.pdf) (accessed October 3, 2005)

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, 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

R Development Core Team (2004) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.r-project.org/) (accessed September 30, 2005)

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

Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, Kidd KK (2005) Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet 13:677–686

Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor β (*TCRB*) locus. Am J Hum Genet 69:381–395

Sun X, Stephens JC, Zhao H (2004) The impact of sample size and marker selection on the study of haplotype structures. Hum Genomics 1:179–193

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

Zhai W, Todd MJ, Nielsen R (2004) Is haplotype block identification useful for association mapping studies? Genet Epidemiol 27:80–83

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

Zhang W, Collins A, Maniatis N, Tapper W, Morton NE (2002*b*) Properties of linkage disequilibrium (LD) maps. Proc Natl Acad Sci USA 99:17004–17007