Haplotype Block Definition and Its Application

X. Zhu, S. Zhang, D. Kan, and R. Cooper

Pacific Symposium on Biocomputing 9:152-163(2004)

HAPLOTYPE BLOCK DEFINITION AND ITS APPLICATION

X. ZHU[1], S. ZHANG[2,3], D. KAN[1], R. COOPER[1]

[1] Department of Preventive Medicine and Epidemiology, Loyola University Stritch School of Medicine, Maywood, IL 60153, [2] Department of Mathematical Science, Michigan Technological University, Houghton, MI and [3]Department of Mathematics, Heilongjiang University, Harbin, China

We present a simple two-stage procedure to define haplotype blocks and construct a statistic to test whether a polymorphism belongs to a block. Applying this method to the data of Gabriel et al. [2002] yielded longer haplotype blocks than were originally reported with a similar average percentage of common haplotypes in blocks. Furthermore, across regions of the genome and among the four populations that were studied, we found that linkage disequilibrium between a given single nucleotide polymorphism (SNP) and the haplotype block was a monotonic function of distance. This correlation was essentially independent of the minor allele frequency of the putative causal SNP when it fell outside of the block, however it was strongly dependent on the minor allele frequency when the SNP was internal to the block. These results have direct application to the design of candidate gene or region-wide association studies.

1 Introduction

Since SNPs occur about every 300bp they provide much more information than other types of sequence variants in mapping complex diseases. Because the evolutionary history of common diseases is not known, one promising approach is to comprehensively test common genetic variants for association with the trait being studied [1,2]. A large-scale project has recently been initiated to define a "haplotype map" (HapMap) of genomic blocks that are shared in common across continental populations [3]. Knowledge of haplotype structure might make it possible to conduct genome-wide association studies at greatly reduced cost, since haplotype tagging SNPs can be chosen to capture most of the genetic information in a region that has a block structure [4-6]. However, debate continues over the plausibility of the "common disease-common variants" (CDCV) hypothesis as a framework for understanding the genetic contribution to complex diseases [7-9]. Pritchard [10] studied an explicit model for the evolution of complex disease loci by incorporating mutation, random genetic drift, and selection against susceptibility mutations. The results indicate that neutral susceptibility alleles are likely to be rare and contribute little for the most plausible range of mutation rates. Allelic heterogeneity at the associated loci might also be found in complex diseases. Finally, the patterns of LD are often extremely variable within and among loci and populations [11-13].

Substantial evidence has already accumulated that the genome can be parsed into haplotype blocks of variable length [4,14-16]. Within these block-like segments, LD is strong and only a few common haplotypes are observed, while between blocks, LD is disrupted, presumably due to historical recombination. The

statistical methods that can best define this structure are still being developed, however, and relatively little is known about their efficiency in gene mapping.

Gabriel et al. [15] proposed an approach based on the pattern of recombination across a region. They defined a pair of SNPs to be in "strong LD" if the one-sided upper 95% confidence bound on D' is >0.98 (consistent with no historical recombination) and the lower bound is above 0.7. A pair of SNPs were defined as having "strong evidence for historical recombination" if the upper confidence bound on D' is less than 0.9. Furthermore, informative marker pairs are those where both minor allele frequencies exceed 0.2. The haplotype block was then defined as a region over which less than 5% of comparisons among informative SNP pairs show strong evidence of historical recombination. In defining a haplotype block, this approach excludes all the SNPs with minor allele frequencies less than 20%, therefore, may lead to under represent of the genetic variation due to these SNPs. Zhu et al. [17] also proposed a definition of haplotype blocks based on the pairwise D' values. Intervals in which all SNPs have a pairwise D' value > 0.8 are identified and it is assumed that they constitute the basic blocks. These intervals are then expanded by adding SNPs to the ends to find the longest intervals, as follows: The observed haplotypes and 95% confidence intervals are bootstrapped before adding a SNP. If the haplotype frequencies after adding a SNP fall into the corresponding 95% confidence intervals, it is concluded that the added marker belongs to the same block. This procedure was repeated until adding a marker leads to a statistical change in the haplotype distribution. No apparent recombination events would therefore have occurred within a block based on this definition. In this report, we apply a modified method of Zhu et al. [17] to the data produced by Gabriel et al. [15] to define haplotype blocks but include all the SNPs with minor allele frequencies great than 5%. However, our focus is to explore the correlation of LD with the physical distance and minor allele frequency after haplotype blocks are defined.

2 Methods

We modify the block definition by Zhu et al. [17]. In detail, we begin by generating a block definition in two steps: first, we define a basic block, then we extend it by sequential addition of the SNPs. Assuming there are N phase known chromosomes, at step 1 we bootstrap the original N chromosomes. For each pair of SNPs i and j, we calculate the bootstrapped α^{th} quantile of $D_{ij}^{'}$ [18]. A segment of consecutive SNPs is a basic block if all the α^{th} quantiles of the pairwise $D_{ij}^{'}$ in the interval are greater than 0.95. Here we choose 0.95 because any recurrent and/or backward mutation, genotype error or sampling variation can affect the value of D' [15]. To choose the value α, we consider two SNPs with alleles (A, a) and (B, b), respectively. We further assume there is no historical recombination between the

two SNPs and no recurrent mutation. The true haplotype frequencies for haplotypes AB, Ab, aB, ab are n_1, n_2, n_3 and 0 ($n_1 + n_2 + n_3 = N$), with corresponding D'=1. Since any genotype error can lead to the change of haplotype frequencies, the haplotype frequencies that are actually observed for AB, Ab, aB, ab may be $n_1^{'}$, $n_2^{'}$, $n_3^{'}$ and $n_4^{'}$, with $n_1^{'} + n_2^{'} + n_3^{'} + n_4^{'} = N$ and $n_4^{'}$>0. Therefore, the probability of D'=1 (which is the true D' value) in the bootstrapped sample is $(1 - \frac{n_4^{'}}{N})^N \approx e^{-n_4^{'}}$, when N is large. Assuming our genotyping is highly accurate, we may allow for a small $n_4^{'}$. We then choose $\alpha = 1 - e^{-n_4^{'}}$. Thus, α is closely related to the possible genotype error rate from two common haplotypes to the rare haplotype (For example, the error rate from Ab->ab, aB->ab, or AB->ab).

In the second step, we examine whether or not a SNP nearest the basic block also falls within its boundaries. To do this, we assume that there are K SNPs in a basic block with the haplotypes $H_1, H_2, ..., H_{n_K}$, and the corresponding number of observed haplotypes $n_1, n_2, ..., n_{n_K}$. We then define a statistic S_K as

$$S_K = \sum_{i=1}^{n_K} \frac{n_i(n_i - 1)}{N(N-1)}.$$

Since $n_1, n_2, ..., n_{n_K}$ follow a multinomial distribution with corresponding population haplotype frequencies $p_1, p_2, ..., p_{n_K}$, we have $E(S_K) = \sum_{i=1}^{n_K} p_i^2$. It can be proven that S_K is the minimum variance unbiased estimate of homozygosity $\sum_{i=1}^{n_K} p_i^2$. Now we add the K+1 th SNP and denote the haplotypes across the K+1 SNPs by $H_{11}, H_{12}, H_{21}, H_{22} ..., H_{n_K 1}, H_{n_k 2}$ with the corresponding numbers of observed haplotypes $n_{11}, n_{12}, n_{21}, n_{22} ..., n_{n_K 1}, n_{n_k 2}$ and corresponding haplotype frequencies $p_{11}, p_{12}, p_{21}, p_{22}, ..., p_{n_k 1}, p_{n_k 2}$. Denote $T_{K+1} = \sum_{i=1}^{n_K} \sum_{j=1}^{2} p_{ij}^{\ 2}$ and $S_{K+1} = \sum_{i=1}^{n_K} \sum_{j=1}^{2} \frac{n_{ij}(n_{ij} - 1)}{N(N-1)}$. Then S_{K+1} is a minimum variance unbiased estimate of T_{K+1}. If the K+1 th SNP falls into the basic block, we would expect that no new haplotypes would be created. Therefore, we have

$T_K = T_{K+1}$. Thus, a reasonable test to determine if the K+1 th SNP falls into the basic block is to test the null hypothesis H_0: $T_{K+1} = T_K$ vs H_1: $T_{K+1} < T_K$. Since $T_{K+1} \leq T_K$ is always true, this test is one-sided. If we do not reject H_0, the K+1 th SNP belongs to the block consisting of SNPs 1, 2,…K. We then continue to add the neighboring SNPs as long as we do not reject the null hypothesis. When the K+1th SNP is rejected, this SNP is regarded as the putative starting point of a new block and a new basic block is again sought and expended. This procedure is repeated until all the SNPs are examined, leading to the initial block partition. We next examine the SNPs in the blocks consisting of less than 3 SNPs. We would test if these SNPs fall into the next block by above method to further expend the blocks. The block size is determined as the sequence length from the beginning to the ending SNP.

To test the null hypothesis $H_0 : T_K = T_{K+1}$, we can use the bootstrap technique to estimate the empirical distribution of S_K. That is, we bootstrap the original N chromosomes and calculate the empirical distribution of S_K. If our observed S_{K+1} falls in the left 5% tail of the empirical distribution of S_K, we consider K+1th SNP falls outside the block comprised of SNPs 1,2…,K. Otherwise, the K+1th SNP is included within the block.

3 Results

To conduct an empirical test we applied this method to the data from Gabriel et al. [15]; obtained from the public access website of the Whitehead Institute]. The genotype data were obtained from four population samples: 30 parent-offspring trios (90 individuals) from Nigeria, 93 individuals from 12 multigenerational CEPH pedigrees of European ancestry, 42 unrelated individuals of Japanese and Chinese origin, and 50 unrelated African Americans. A total of 3738 SNPs in 54 autosomal regions were successfully genotyped in all four groups. The average size of a region was 250 kb. For family data, we first used MERLIN [19] to reconstruct the haplotypes and then to estimate the haplotype frequencies vie EM algorithm, while we directly applied EM algorithm [20] to infer the haplotype frequencies for unrelated data in each region. Then we applied the proposed method to define haplotype blocks. Table 1 presents the characteristics of haplotype blocks for the four populations using α =0.5 and 0.638. Our definition identified more haplotype blocks in Nigerians and African-Americans than in Europeans and Asians (the blocks are limited to those covered by more than two SNPs), and encompassed around two third of the total sequence. On average we obtained block sizes somewhat longer than those reported by Gabriel et al. [15], who reported averaged 9 kb in the Nigerian and African-American samples and 18 kb in the European and Asian samples. The common haplotypes accounted for most of the information on

heterozygosity in a block, representing on average 93% to 96% of all haplotypes (Table 1). Perhaps somewhat surprising, our method gave similar numbers of common haplotypes in the different populations. However, as in previous analyses, the Europeans and Asians consistently had fewer haplotype blocks and LD extended over longer intervals than among Africans and African-Americans. Because the overall results of using α =0.5 and 0.638 are very similar, we only performed our next analyses based on α =0.5.

Table I. Characteristics of haplotype blocks using the proposed definition

	∀=0.5				∀=0.638			
	Nig	AA	EA	As	Nig	AA	EA	As
# of blocks	510	470	420	324	503	459	396	306
% of sequence covered by blocks	59.2	52	76	69	62.4	56.7	77.9	71.0
Average block size (kb)	12.7	13.6	22	26.2	13.6	15.1	23	28.5
Average # of common haplotypes	3.1	3.4	3.3	3.5	3.2	3.6	3.4	3.7
Average % of common haplotypes in blocks	96.5	95.1	95.8	93.4	96.3	94.4	95.2	93.0

Common haplotypes : frequencies>5%; Nig: Nigerian; AA: African American; EA: European American; As: Asian

To estimate the distribution of block sizes, we performed simulations following the procedures of Gabriel et al. [15] in which block sizes were exponentially distributed and markers were randomly spaced. The simulations provided an almost perfect fit to the observed data for both the African and European samples (Table 2). The definition applied by Gabriel et al. [15], on the other hand, overestimated the incidence of blocks with sizes less than 5 kb. We also compared the block boundaries defined in the four populations and found that most of the boundaries observed among Europeans and Asians were also present among Nigerians. To obtain a summary of this phenomenon we examined whether a block boundary among Europeans, Asians and African-Americans was consistent with that found among Nigerians across all 54 regions. We assumed consistency if the ending SNP of a block or the beginning SNP of the adjacent block in the three non-African populations fell between two SNPs that define the end and the beginning of the corresponding segment among the Nigerians. Our calculation suggests that 61%, 71% and 72% of block boundaries in the Europeans, Asians and African-Americans are consistent with those among the Nigerians. The results also suggest that most of the historical recombination breakpoints are shared across the four populations.

Table II. Observed and predicted proportion of sequence found in haplotype blocks. Block span is based on an exponentially distributed random variable with mean size of 22 kb in the European and 13 kb in the Nigerian samples

	Nigerians		Europeans	
Block size (kb)	Obs.	Pred.	Obs.	Pred.
0-5	6.1 (12.4)	5.6 (6.3)	2.0 (4.4)	2.2 (1.8)
5-10	11.8 (15.3)	11.9 (15.1)	5.7 (7.4)	5.5 (5.2)
10-20	26.3 (20)	27.1 (31.5)	15.3 (14.9)	15.3 (15.2)
20-30	19.6 (12.8)	22.6 (21.8)	16.7 (16.6)	16.9 (16.6)
30-50	18 (22.2)	22.2 (19.1)	25.9 (18)	26.7 (26.9)
>50	18.2 (17.4)	10.6 (6.3)	34.4 (38.7)	33.4 (34.2)

() are the values from Gabriel et al. [2002]

We further looked at the strength of LD between a given SNP and the haplotype blocks in each region. For two biallelic markers, there are currently a number of measures for LD, as reviewed by Devlin and Risch [21]. It may be difficult to obtain an entirely satisfactory LD measure for two multiallelic markers using these procedures, however. As an alternative Zhao et al. [22] have proposed a permutation-based method to measure the LD between two multiallelic markers. This measure is based on the likelihood ratio test statistic t, which asymptotically follows a noncentral χ^2 distribution with γ degrees of freedom, and is defined as

$$\xi = \frac{\sqrt{2\gamma}}{n}\left(\frac{t-\mu}{\sigma}\right),$$

where μ and σ^2 are the mean and variance of the empirical distribution of the likelihood ratio test statistic t from the permuted samples, n is the number of individuals in the sample. Consequently ξ is the measure of the overall deviation from random association. The test of LD using ξ to detect overall deviation from random association is therefore more powerful than one based on asymptotic distributions [22].

In Figure 1 the LD measure ξ is presented as a function of distance between a given SNP and a block in a selected region for the four populations. Each line in figure 2 represents a given SNP and the data in the figure is centered on that SNP. ξ is close to a montonic function of distance and potential noise caused by marker history almost disappears. The observed decrement of ξ is presumably due to the breakdown of LD by recombination. For comparison, we also present LD measured by the correlations r^2 and D' using the same marker set (figures 2 and 3). As expected, very substantial noise was observed as a result of the marker history for pairwise r^2 or D' (figures 2 and 3). The results for D' are particularly erratic, presumably due to its dependence on the allele frequencies as well as marker

history. We extended this analysis across all the markers by summarizing the percentage of SNPs where this monotonic property is preserved. According to Zhao et al. [22], ξ > 0.59 indicates at least a weak degree of LD and we used this definition to determine how often ξ was a montonic function of the distance between a SNP and haplotype blocks in regions where ξ >0.59. In the European, African American, Asian and Nigerian population samples, respectively, this finding was observed in 77% (1352/1758), 77% (1286/1677), 69% (1158/1679) and 90% (1030/1146) of the instances examined. These results demonstrate that haplotype blocks can potentially be very useful in efforts to localize disease loci using LD mapping.

While it is necessary to establish consistent patterns of LD under conditions likely to apply in association studies, variation in the frequency of the marker alleles being studied can also threaten the validity of the statistical analyses. We therefore examined the LD between a SNP with a minor allele frequency >5% and a block. Here we considered a block as a supermarker and haplotypes as the alleles. We looped the haplotypes with frequencies less than 5% together. Figure 4 presents the scatter plot of LD measured by ξ vs the minor allele frequency of a SNP when this SNP falls within the block for the four populations. The scatter plot of ξ vs the minor allele frequency of a SNP when the SNP is outside of the boundaries of the block is presented in figure 5. In general, these data demonstrate that the strength of LD is significantly dependent on the minor allele frequency. This association is obviously much stronger if the SNP falls within the block rather than outside it. Table 3 presents the variance expressed by the minor allele frequency - represented by R^2 when fitting a linear regression of ξ on minor allele frequencies. As can be seen, the R^2 values in the four populations are very similar, ranging from 0.49-0.76 when a SNP falls within a block, and they decrease to 0.02-0.04 when the SNP resides outside the block. Since the association between a SNP and a block is dependent on distance (defined as the number of bp between the SNP and the middle position of the block), we added distance as an independent variable in the regression model for external SNPs. R^2 values in this model increased to 0.10-0.14 for the four populations. However, for internal SNPs, R^2 values remain virtually unchanged even after distance was added to the regression model (data not shown). Our results indicate that the LD between an internal SNP and a block is strongly dependent on the minor allele frequency, while distance is the primary determinant for external SNPs. Furthermore, our defined blocks are also valid because the LD between an internal SNP and a block does not depend on the position of the SNP. We also observed that the average values of ξ for internal SNPS are very similar among the four populations, ranging from 0.89 to 0.93. This result seems reasonable since there is presumably little or no historical recombination within these

segments. Figure 3 and 4 also suggest that the variance of ξ is larger in African-Americans and Asians than in Europeans and Nigerians. Two aspects of the data set used in these analyses could potentially explain this result. First, twice as many Europeans and Nigerians were studied, compared to the African-Americans and Asians. Secondly, family members were included among the Europeans and Nigerians and therefore haplotypes can be more reliably inferred.

Table III. The total linkage disequilibrium variance between a SNP and a block expressed by the minor allele frequency and the distance between the SNP and the block. The numbers in the table are the R^2 values from fitting a linear regression.

	European	African-American	Asia	Nigerian
SNP within a block	0.65	0.56	0.48	0.76
SNP outside a block	0.04	0.02	0.04	0.03
SNP outside a block, adding distance	0.14	0.10	0.13	0.13

4 Discussion

Limitations of the empirical test of the method presented here must be recognized. Our analyses are based on an average marker density of 1 SNP/ 7 kb and a relatively small number of individuals (less than 100 unrelated chromosomes). As noted earlier, sample size, frequency and inter-marker interval may alter the scale on which patterns are discerned. As reported recently by Phillips et al. [23], block length may be dependent on the density of SNPs that are typed. Increased density of SNPs may yield shorter haplotype blocks in fact if such dense SNPs exist across the genome. That issue will clearly require further study in large empirical samples. Our proposed method requires that haplotype phase information be available. It may affect the haplotype block partition if only unrelated individuals are studied. However such an effect should be limited because our studied regions are small and therefore strong LD exists [24]. Furthermore, it should be noted that haplotype block partition only serves a tool for mapping a complex disease, which is our ultimate goal.

Although the selection of α may be interpreted as the genotype error rate, it is ad hoc. Our results suggest that the selection of α is not sensitive to the definition of haplotype blocks. In conclusion, our results provide a robust statistical method to define the haplotype structure of the human genome using SNP markers. The method can include the SNPs with minor allele frequencies >5%. By applying this method to a large empirical data set we obtained a highly consistent description of the properties of blocks across 54 genomic regions. Our results support the contention that in most instances LD between a SNP and neighboring haplotype blocks is a monotonic function of the distance. Using this strategy a disease locus could be mapped with high resolution in an appropriately designed association

study if the distribution of haplotype blocks in the region has been well defined. On the other hand, to localize a functional SNP within a block, additional considerations may be critical, especially if the assumption of CDCV fails and the putative causal mutation is infrequent. This is because the LD between a SNP and the block is strongly dependent on the minor allele frequencies. In this case, a design that enriches the sample for the rare disease variants will be an important determinant of the chances of success. Although our results demonstrate that the individual block boundaries overlap across populations, this conclusion should be further investigated using SNPs at higher density.

5 Acknowledgments

We thank Mark Daly for helpful comments. We thank Fang Yang for her assistance in programming. This work was supported by grants from the National Heart, Lung and Blood Institute (HL53353 and HL65702), and the Reynolds Clinical Cardiovascular Research Center at UT Southwestern Medical Center, Dallas, TX.

6 References

1. E.S. Lander, Science, 274:536-9 (1996)
2. N. Risch, K. Merikangas, Science. 273:1516-7 (1996)
3. D.E. Reich DE, Nature. 411:199-204 (2001)
4. M.J. Daly, Nat Genet. 29:229-32 (2001)
5. R. Judson R, Pharmacogenomics. 3:379-91 (2002)
6. G.C. Johnson, Nat Genet. 29:233-7 (2001)
7. J.D. Terwilliger et al, Curr Opin Genet Dev. 12:726-34 (1998)
8. J.D. Terwilliger, K.M. Weiss, Curr Opin Biotechnol. 9:578-94 (1998)
9. K.M. Weiss, A.G. Clark, Trends Genet. 18:19-24 (2002)
10. J.K. Pritchard, Am J Hum Genet. 69:124-37 (2001)
11. J.K. Pritchard, M. Przeworski, Am J Hum Genet. 69:1-14 (2001)
12. L.B. Jorde, Genome Res. 10:1435-44 (2000)
13. M. Boehnke, Nat Genet. 25:246-7 (2000)
14. N. Patil et al, Science. 294:1719-23 (2001)
15. S.B. Gabriel et al. Science. 296:2225-9 (2002)
16. A.J. Jeffreys et al, Nat Genet. 29:217-22 (2001)
17. X. Zhu et al, Genome Research. 13: 171-181 (2003)
18. R. Lewontin, Genetics. 49:49-67.(1964)
19. G.R. Abecasis et al. Nat Genet; 30:97-10119 (2002)
20. L. Excoffier, M. Slatkin, Mol Biol Evol;12:921-927 (1995)
21. B. Devlin, N. Risch, Genomics. 29:311-22 (1995)
22. H. Zhao et al, Ann Hum Genet. 63:167-79. (1999)
23. M.S. Phillips et al. Nat Genet. 33:382-387 (2003)
24. S. Lin et al. Am J Hum Genet. 71:1129-1137 (2002)

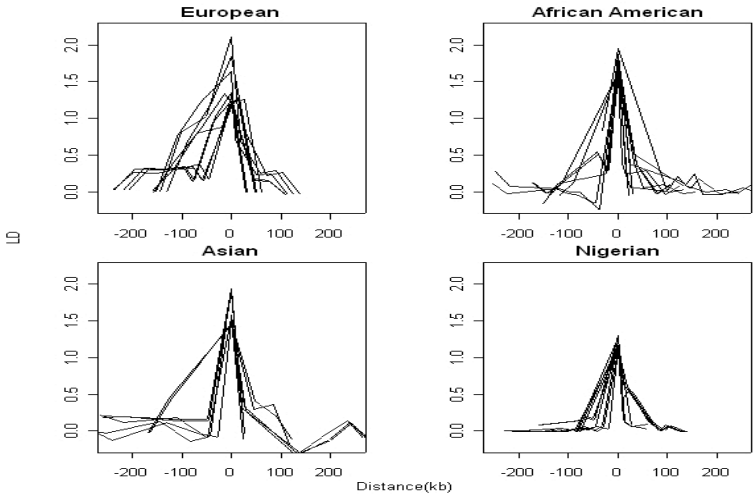

Figure 1. Linkage disequilibrium measured by ξ [21] between a SNP and haplotype blocks. 10 SNPs, represented by 10 lines, were used in each sample. The lines are centered on the selected SNPs. The distance measure is the physical distance between a SNP and the center of a haplotype block. A montonic relationship between the linkage disequilibrium and physical distance is observed.

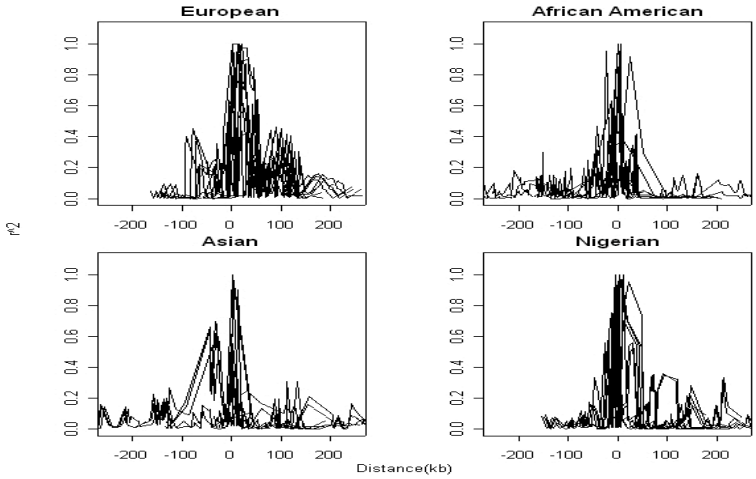

Figure 2. Linkage disequilibrium measured by correlation r^2 between a pair of SNPs. The 10 SNPs used in figure 1 are plotted. All the lines are centered by the 10 SNPs. Much variation is observed due to the marker history.

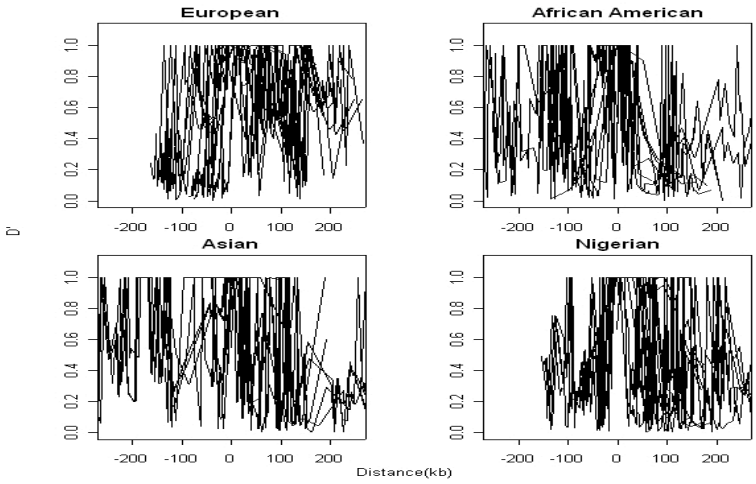

Figure 3. Linkage disequilibrium measured by correlation D' between a pair of SNPs. The 10 SNPs used in figure 2 are plotted. All the lines are centered by the 10 SNPs. Considerable variation is observed due to marker history.

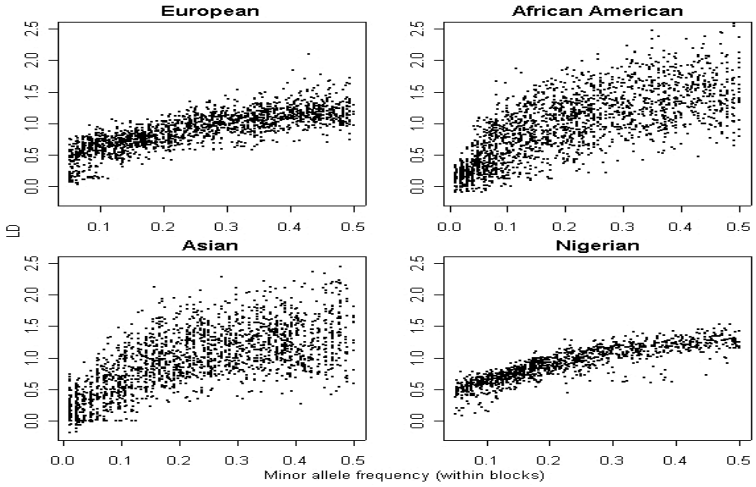

Figure 4. Linkage disequilibrium measured by ξ [21] between a SNP and a block as a function of the SNP's minor allele frequency when the SNP falls within the index block. A substantial proportion of the variance is accounted for by the minor allele frequency.

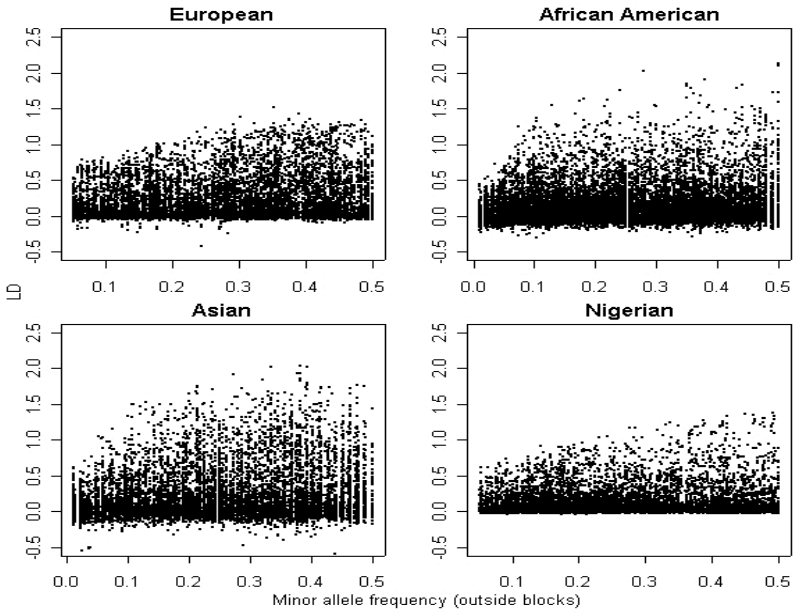

Figure 5. Linkage disequilibrium measured by ξ [21] between a SNP and a block as a function of the SNP's minor allele frequency when the SNP falls outside of the index block. Only a small part of the variance is accounted for by the minor allele frequency.