Am. J. Hum. Genet. 66:216–234, 2000

Linkage Disequilibrium and Allele-Frequency Distributions for 114 Single-Nucleotide Polymorphisms in Five Populations

Katrina A. B. Goddard,[1] Penelope J. Hopkins,[2] Jeff M. Hall,[2] and John S. Witte[1]

[1]Case Western Reserve University, Cleveland; and [2]PPGx, Inc., La Jolla, CA

Summary

Single-nucleotide polymorphisms (SNPs) may be extremely important for deciphering the impact of genetic variation on complex human diseases. The ultimate value of SNPs for linkage and association mapping studies depends in part on the distribution of SNP allele frequencies and intermarker linkage disequilibrium (LD) across populations. Limited information is available about these distributions on a genomewide scale, particularly for LD. Using 114 SNPs from 33 genes, we compared these distributions in five American populations (727 individuals) of African, European, Chinese, Hispanic, and Japanese descent. The allele frequencies were highly correlated across populations but differed by >20% for at least one pair of populations in 35% of SNPs. The correlation in LD was high for some pairs of populations but not for others (e.g., Chinese American or Japanese American vs. any other population). Regardless of population, average minor-allele frequencies were significantly higher for SNPs in noncoding regions (20%–25%) than for SNPs in coding regions (12%–16%). Interestingly, we found that intermarker LD may be strongest with pairs of SNPs in which both markers are nonconservative substitutions, compared to pairs of SNPs where at least one marker is a conservative substitution. These results suggest that population differences and marker location within the gene may be important factors in the selection of SNPs for use in the study of complex disease with linkage or association mapping methods.

Received August 26, 1999; accepted for publication November 4, 1999; electronically published January 11, 2000.

Address for correspondence and reprints: Dr. Katrina A. B. Goddard, Department of Epidemiology and Biostatistics, Case Western Reserve University, 2500 MetroHealth Drive, Cleveland, OH 44109-1998. E-mail: katrina@darwin.cwru.edu

0002-9297/2000/6601-0000$02.00

Introduction

Traditional methods that have been successful in the mapping of genes for Mendelian disorders, such as parametric linkage analysis, have not been as successful in studies of complex genetic traits, indicating a need for alternative approaches. Linkage disequilibrium (LD) mapping (Risch and Merikangas 1996) and model-free methods of linkage analysis (e.g., see Kruglyak et al. 1996; Elston et al. 1999) have been suggested as alternative approaches. Unfortunately, these methods may require a substantial number of markers, as well as large sample sizes, for detection of linkage, for a variety of reasons. First, the sample heterogeneity that is often present in studies of complex traits reduces the power and increases the sample size necessary to detect linkage (e.g., see Goldin and Gershon 1988; Risch 1990; Goldin and Weeks 1993). Also, LD mapping may require a very high density of markers, since, in many populations, LD is detectable only across small regions. Finally, the model-free methods of linkage analysis usually require a large number of individuals, even in light of the power improvement arising from the development of multipoint analysis and other modifications (Kruglyak et al. 1996; Elston et al. 1999).

Regardless of the study design used, single-nucleotide polymorphisms (SNPs) may provide an important alternative to conventional markers, for genetic mapping studies of complex traits. SNPs are sites in the genome that have nucleotide differences. These polymorphisms are highly abundant, occurring approximately ~1/1,000 bp (Wang et al. 1998). Methods for the genotyping of SNPs are more easily automated and potentially less expensive per marker than are conventional methods such as microsatellite markers (Nickerson et al. 1990; Pease et al. 1994). Given the large number of markers and individuals that must be genotyped for studies of complex traits, SNPs could substantially reduce the cost of a genetic mapping study. For these reasons, SNPs may become a key component in future studies of complex traits.

Several studies have evaluated SNP characteristics that are important for both linkage and association mapping studies, including the allele frequencies and the LD be-

tween markers. In the context of LD mapping, recent work demonstrates that the power and sample size necessary for mapping studies depends on the allele frequencies of the SNP markers (Chapman and Wijsman 1998; Xiong and Jin 1999). The power of linkage-analysis methods also depends on the allele frequencies, although frequencies of the major allele that are between ~.5 and ~.8 provide essentially equivalent power to detect linkage (Kruglyak 1997; Goddard 1999). Clusters of SNPs have been considered as an alternative to uniformly spaced markers in a linkage-based genome screen. Here, multiple SNPs with essentially no recombination among them are used as a single marker to provide more information than would be available with single-SNP markers (Nickerson et al. 1992; Goddard 1999). For the clustered SNP map structure, intermarker LD generally reduces the information content of the cluster (Goddard 1999) by shifting the haplotype frequencies away from the most informative case of equal frequencies (e.g., under complete LD, only two haplotypes are observed).

Little information is available about the actual distribution of these marker characteristics for SNPs on a genomewide scale. Previous reports on allele frequency and LD distributions for SNPs have focused on only one gene or region, including lipoprotein lipase (Clark et al. 1998; Nickerson et al. 1998), apolipoprotein E (Lai et al. 1998), and the single-minded homolog 2 (SIM2) gene (Carlson and Cox 1998). It is unclear whether these results can be generalized to the whole genome. Recently, Cargill et al. (1999) and Halushka et al. (1999) evaluated the allele-frequency distribution for SNPs in 106 and 75 genes, respectively; however, these studies considered relatively small sample sizes—57 and 74 individuals, respectively—from multiple populations. Cambien et al. (1999) evaluated allele-frequency and LD distributions for SNPs in 36 genes from individuals of European descent, but they did not consider population differences in these distributions.

The distribution of allele frequencies and LD may be substantially different among populations. Numerous studies have indicated—by use of multiple types of polymorphisms, such as blood-group markers (Cavalli-Sforza et al. 1994), microsatellites (Bowcock et al. 1991, 1994; Jorde et al. 1997; Destro-Bisol et al. 1999), and RFLPs (Dean et al. 1994)—that the distribution of allele frequencies differs among populations, so it is reasonable to expect population differences in the allele frequencies for SNPs as well. Despite the small sample sizes considered in previous reports, population differences in the allele frequencies of SNPs were observed (Nickerson et al. 1998; Lai et al. 1998; Cargill et al. 1999; Halushka et al. 1999). Little information is available about population differences in the LD distribution.

If one wants to develop a panel of SNPs for mapping to be used across populations, as currently exists with microsatellites, population differences in the distribution of allele frequencies and LD will limit the choice of markers. Population differences in the information content of markers alter the power to detect linkage among the populations. Compared to microsatellites, SNPs are more likely to have large differences in the marker information content, since SNPs have relatively few alleles that may not be observed in all populations. It may be possible to include multiple markers for each gene or region in a screening set of SNPs to increase the probability that variability is observed in all populations under consideration; however, this redundancy increases the cost of using SNP markers compared to microsatellite markers.

In the present paper we evaluate the allele-frequency and intermarker LD distributions for SNPs. To investigate these distributions on a genomewide scale, we consider 114 SNP markers that are located in 33 genes on 16 chromosomes. Our study sample consists of 727 individuals from five American populations of African, European, Chinese, Hispanic, and Japanese descent. We find important differences in the distribution of allele frequencies and LD among different populations and among different locations within the gene (e.g., coding vs. noncoding regions). We consider the influence of these differences in the allele-frequency and intermarker LD distributions on marker selection for a genome screen using association or linkage analysis, and we discuss using the distribution of the intermarker LD as a surrogate for the distribution of trait-marker LD.

Subjects and Methods

Samples

The study sample consisted of individuals from five populations. In particular, we enrolled in the study 190 European American, 190 Hispanic American, 190 African American, 79 Chinese American, and 78 Japanese American volunteers from Southern California, all apparently healthy. It is important to note that, in contrast to panels such as the human genetic diversity project, the individuals in this study do not necessarily represent the aboriginal populations of the associated geographic regions and, therefore, do not necessarily reflect country- or region-specific data such as are typically studied by population geneticists. However, these population groups are representative of self-reported ethnicity in the United States, which is often used to define populations in genetic mapping studies. Each subject provided a blood sample, after providing informed consent and self-report of his or her ethnicity. Among the 114 SNPs evaluated, 1%–2% of the marker genotypes were missing for each population. With few exceptions, most indi-

viduals had missing information for <10 of the 114 markers. In addition, there were only two markers with missing data for >12 individuals from a single population. For these two markers, only half of the individuals were genotyped for the African American, Hispanic American, and European American populations. However, since the remaining sample size of ~85 individuals was still larger than the sample size for the Chinese American and Japanese American populations, these markers were included to maximize the number of SNPs and genes that were considered. Individuals with missing data at a particular marker were removed from any analysis that included that marker.

Marker Selection

Genes were initially selected for analysis on the basis of their known or potential pharmacological relevance to an individual's response to drugs. SNPs were identified in the genes on the basis of existing sequence information or by resequencing in 10–16 individuals from each of the European American, African American, and Hispanic American populations (except for three markers that were resequenced in 16 individuals from each of the European American, African American, and Chinese American populations). With regard to SNPs identified by resequencing, a site was considered a SNP if there was a base-pair difference for at least one individual in the reference set of 30–48 individuals. This detection method is more likely to identify SNPs with a minor-allele frequency close to .5 for at least one of the populations in the reference set; however, the detection method does not tend to increase the similarity of allele frequencies among the populations in the reference set. For inclusion here, both alleles of an SNP had to be observed in the study sample in at least one population (described below), and at least two markers had to be observed in the same gene. We evaluated a total of 114 autosomal, diallelic markers (44 from existing information, 70 from resequencing) that were genotyped in all populations. These SNPs were distributed among 33 genes located on 16 chromosomes. We observed 2–13 markers per gene, resulting in 215 pairs of markers within genes that were evaluated for intermarker LD.

Marker Genotyping

DNA was extracted from blood by use of a kit from Gentra Systems, Inc. SNP genotypes were determined by use of the *Taq*Man assay (Heid et al. 1996). Samples were assayed in triplicate in a Robbins 96-well plate. The primers for each SNP were either derived from published sequence information or developed at PPGx, Inc. Fragments were amplified by PCR in reactions containing 20 ng genomic DNA, 900 nM forward unlabeled inner primer, 900 nM reverse unlabeled inner primer, 200 nM 6-carboxy-fluorescein (FAM)-labeled probe, 200 nM tetrachloro-6-carboxy-fluorescein (TET)-labeled probe, and 1 × *Taq*Man reagent mix 43C4447 (PE Biosystems). PCR reactions were preincubated at 50°C for 2 min, then at 95°C for 10 min. Two-step thermocycling was performed for 45 cycles of denaturation at 95°C for 30 s and annealing at 64°C for 30 s. On completion of thermocycling, the fluorescence was read on an ABI 7700 Sequence Detector using the allelic discrimination software. FAM:TET ratios for each sample DNA, normalized against the TAMRA signal, indicated the genotype of each patient and were further confirmed by similar signals from known control DNAs.

Statistical Methods

The allele frequencies for each marker were estimated by use of the allele-counting method. We used the χ^2 approximation to test Hardy-Weinberg equilibrium (HWE) at each locus (Weir 1996) and used the EM algorithm to estimate pairwise haplotype frequencies (Excoffier and Slatkin 1995). *P* values for a test of intermarker LD were obtained by use of a randomization test for the test statistic, $S = 2ln(L*/L_o)$, where L^* is the likelihood computed by use of the haplotype frequencies estimated from the EM algorithm and L_o is the likelihood under the assumption of no disequilibrium (Slatkin and Excoffier 1995). This randomization test using the estimated haplotype frequencies performed well compared with Fisher's exact test using the actual haplotype frequencies in simulations (Slatkin and Excoffier 1995). Nine measures of LD were initially considered, including the composite disequilibrium for genotype data (Weir 1996) and eight measures for haplotype data that were suggested in Devlin and Risch (1995). We obtained similar results with the different measures investigated, so the only measure presented here is the difference in proportions, $d = \pi_{11}/\pi_{.1} - \pi_{12}/\pi_{.2}$, where π_{ij} is the frequency of haplotypes with allele i at the first marker and allele j at the second marker and $\pi_{.j}$ is the frequency of haplotypes with either allele at the first marker and with allele j at the second marker (Nei and Li 1980). This measure has a range of -1 to 1, and is equal to 0 when there is no disequilibrium. The difference in proportions was less dependent on allele frequencies than were the other measures investigated, on the basis of empirical observations of the relationship between allele frequencies and the measures of LD in this data set. Here we define $q_{\min}$ as the smallest allele frequency for a pair of SNPs (i.e., $q_{\min} = \min(q_1,q_2)$, where q_i is the minor-allele frequency at locus i).

Results

Population Differences in Allele Frequencies

We observed different levels of variation across populations, with regard to the SNP allele frequencies. The African American population had the most variation, with both alleles observed for 92% of the SNPs (table 1). The Chinese American and Japanese American populations had the least variation, with both alleles observed for only 60% and 62% of the SNPs, respectively. (Appendix A provides all of the allele frequencies and tests for HWE.) Alleles that were not observed in one population tended to have small allele frequencies in the other populations. Thus, populations with the largest number of SNPs with variability also had the largest number of SNPs with rare alleles (e.g., minor-allele frequencies 0–.05). For example, 32% of the SNPs had a minor-allele frequency of 0–.05 in the African American population, compared to only 12% of the SNPs with a minor-allele frequency of 0–.05 in the Chinese American population (table 1). We observed a similar pattern when we considered only the SNPs detected by resequencing. Approximately 80%–95% of the SNPs with one allele fixed in the Chinese American and Japanese American populations had a minor-allele frequency <.05 in the African American, European American, and Hispanic American populations. This implies that, even though the African American population has more sites with variability than the Chinese American and Japanese American populations, under most circumstances many of these sites may have little information for linkage or association studies because of the low allele frequencies. Although the greater variability observed among the African American, European American, and Hispanic American populations may be the result of an ascertainment bias in the selection of markers, our observations are consistent with other studies of SNPs (Zietkiewicz et al. 1997; Nickerson et al. 1998) where there was no ascertainment bias in the selection of markers.

Allele frequencies were generally highly correlated ($\rho > .8$) among the populations (fig. 1, above diagonal). The Japanese American and Chinese American populations had the most similar allele frequencies, with a correlation of .99. The Hispanic American population had relatively high correlations ($\rho > .87$) with all of the other populations, whereas the remaining pairs of populations had lower correlations ($\rho < .83$). Despite these high correlations, there were still important allele-frequency differences among the populations. For 35% of the SNPs, the allele frequencies differed by >.2 for at least one pair of populations. Furthermore, 54% of the SNPs with a major-allele frequency of .5–.8 in one population had a major-allele frequency >.8 for at least one other population. This latter observation is important because, as noted above, in linkage analysis, markers with a major-allele frequency of .5–.8 are essentially equivalent in information content, whereas markers with a major-allele frequency >.8 have reduced information content (Kruglyak 1997; Goddard 1999).

Table 1

SNPs for Each Range of Minor-Allele Frequency

Population	No. (Proportion) of SNPs When Allele Frequency Is			
	$q = 0$	$0 < q < .01$	$.01 \leq q < .05$	$q \geq .05$
African American	9 (.08)	16 (.14)	21 (.18)	68 (.60)
European American	20 (.18)	14 (.12)	13 (.11)	67 (.59)
Hispanic American	16 (.14)	16 (.14)	18 (.16)	64 (.56)
Chinese American	46 (.40)	4 (.04)	9 (.08)	55 (.48)
Japanese American	44 (.38)	6 (.05)	11 (.10)	53 (.46)

Population Differences in the Distribution of LD

When LD was considered, similarities in the LD measure suggested categorizing the five populations into two groups (fig. 1). In particular, populations in the same group had a high correlation in the measure of LD ($\rho > .87$), whereas populations in different groups had a lower correlation in the measure of LD ($\rho < .65$) (fig. 1, below diagonal). The first group was composed of the Chinese American and Japanese American populations, and the second group was composed of the African American, European American, and Hispanic American populations. Appendix B presents both the measure of LD for each pair of SNPs within a gene and the corresponding *P* value. It is interesting to note that the allele frequencies and the measure of LD have a similar pattern in the correlation for the populations considered here. The similarity in allele frequencies and the measure of LD may reflect a more recent common population history for some of the populations, such as may exist for the Chinese American and Japanese American populations (Bowcock et al. 1994; Cavalli-Sforza et al. 1994; Jorde et al. 1997; Zietkiewicz et al. 1997).

There were two cases with extreme differences, in the measure of LD, among different populations (Appendix B, markers 103 and 107 and markers 106 and 107). Extreme differences in the measure of LD occur when the "A" allele at one locus is associated with the "A" allele at the second locus in some populations, whereas it is associated with the "B" allele at the second locus in other populations. Both pairs of SNPs with extreme differences in the measure of LD were in the CYP2D6 gene. Both a wide range in the allele frequencies and *P* values $\leq .05$ for the test of HWE were observed for some SNPs in this gene (Appendix A, markers 103, 106, and 107). However, across all of the populations, 30/570 (5%) of the tests for HWE had a significant result at the 5% significance level, indicating that these markers are

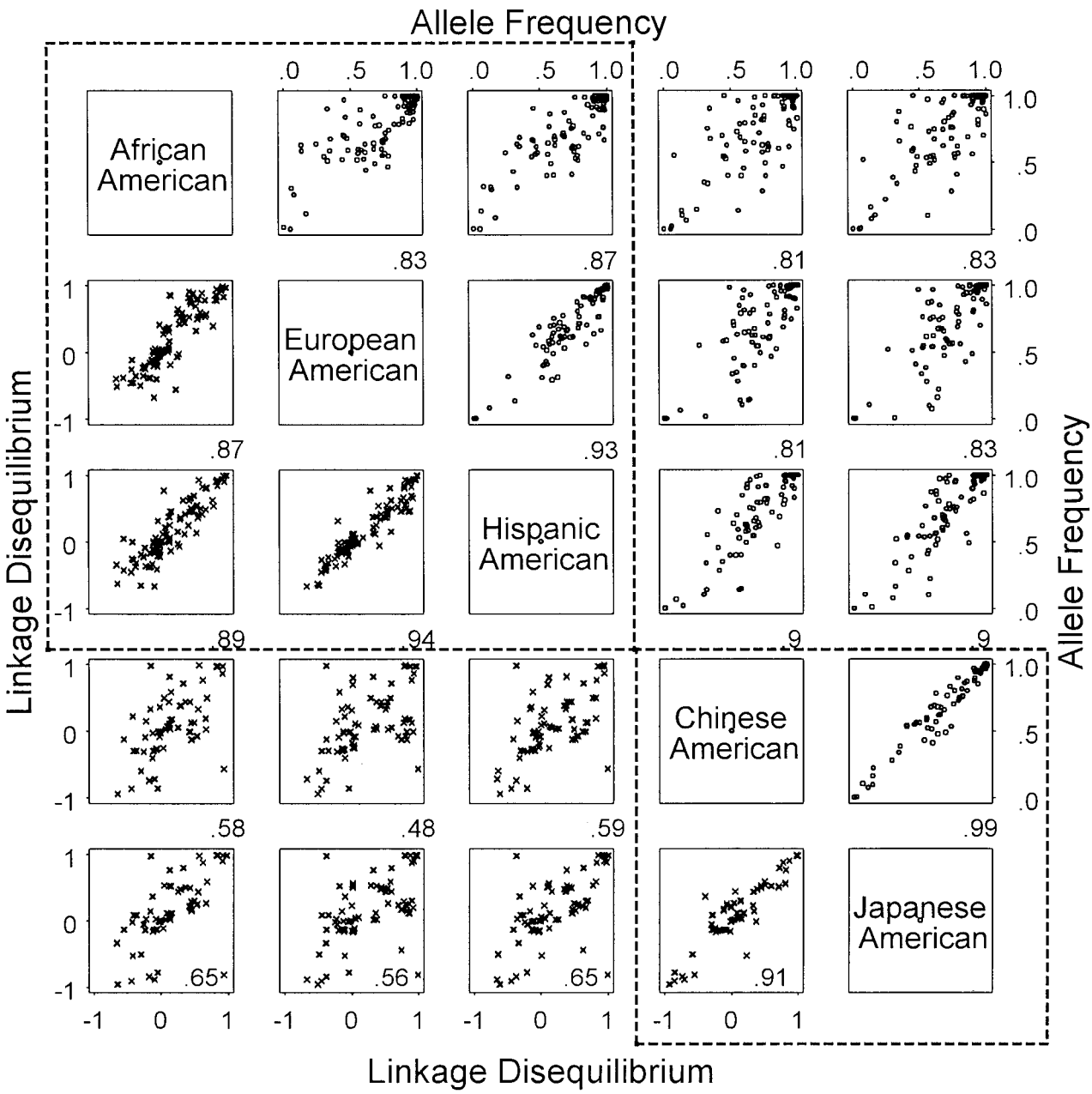

Figure 1 Comparison of allele frequencies and LD among populations. The upper triangle corresponds to the allele frequencies (0), and the lower triangle corresponds to the LD measure, d (×). The correlation is indicated in the lower right corner of each graph. The P value for the correlation was <.0001 in all cases.

consistent with HWE. These differences among the populations may at least partially explain the extreme differences in the measure of LD for these SNPs. Removing these two pairs of SNPs did not considerably change the correlation in the measure of LD.

Low allele frequencies ($q_{min} < .05$) accounted for 80% of the situations in which LD was not detected ($P > .05$) for SNPs within the same gene (table 2). The power to detect LD is low when the allele frequencies for at least one of the SNPs are very extreme, and, in fact, it may be impossible to achieve significance under certain circumstances with very rare alleles (Lewontin 1995). When both SNPs had high minor-allele frequencies (i.e., $q_{min} \geq .05$), LD was detected 82% of the time. In contrast, when $q_{min} < .05$, LD was detected only 26% of the time. The percentage of observations in which LD was not detected when $q_{min} \geq .05$ ranged between 3% (Hispanic Americans) and 36% (Chinese Americans) for

Table 2

Pairs of SNPs within the Given Range of q_{min} and *P* Value for Test of LD

	No. (Proportion) of SNP Pairs within Range								
		$0 < q_{min} < .05$				$q_{min} \geq .05$			
Population	$q_{min} = 0$	$P \leq .001$	$.001 < P \leq .05$	$P > .05$	Total	$P \leq .001$	$.001 < P \leq .05$	$P > .05$	Total
African American	40	15 (.14)	13 (.12)	78 (.74)	106	46 (.67)	13 (.19)	10 (.14)	69
European American	88	14 (.23)	6 (.10)	41 (.67)	61	54 (.82)	5 (.08)	7 (.11)	66
Hispanic American	66	13 (.15)	10 (.11)	65 (.74)	88	55 (.90)	4 (.06)	2 (.03)	61
Chinese American	138	1 (.06)	1 (.06)	16 (.88)	18	35 (.59)	3 (.05)	21 (.36)	59
Japanese American	135	4 (.17)	1 (.04)	19 (.79)	24	35 (.62)	6 (.11)	15 (.26)	56
Total	467	47 (.16)	31 (.10)	219 (.74)	297	225 (.72)	31 (.10)	55 (.18)	311

individual populations. However, these were not all independent observations, since many instances in which we failed to detect pairwise LD when $q_{min} \geq .05$ occurred in the same gene, UGT1, for the Chinese Americans (15/21) and the Japanese Americans (11/15). (Fig. 2 shows the *P* values for LD for each pair of SNPs within the same gene and for each population.) The power to detect LD is also low when the minor allele at each locus is on a separate haplotype (i.e., the repulsion phase) (Thompson et al. 1988), which accounts for some of the cases in which LD is not observed. The intermarker distance is one possible explanation for the remaining situations where LD is not observed, since LD is generally only detectable for a small region near each site. However, in many instances where LD was not detected for one population, it was detected in other populations (fig. 2), suggesting additional explanations, such as factors associated with population history (e.g., population size and growth), for the lack of detectable LD.

As expected, LD was generally not detected for pairs of SNPs on different chromosomes, although the proportion of significant tests across all populations (7% [114/1598 pairs]) was slightly higher than would be expected by chance at the 5% significance level. We did not consider all possible pairs of SNPs on different chromosomes, because of computational constraints. Instead, we evaluated LD for markers on different chromosomes within a subset consisting of one SNP randomly selected from each gene. The proportion of pairs of markers in which LD was detected ($P \leq .05$) was .06 (22/386), .09 (39/416), .07 (27/414), .05 (10/201), and .09 (16/181) for the African Americans, European Americans, Hispanic Americans, Chinese Americans, and Japanese Americans, respectively (LD was not tested if one marker had a minor-allele frequency equal to 0). This background rate of LD is much lower than the rate of LD that we observed for linked markers, although it is higher than would be expected under the null hypothesis. This suggests that a low level of background LD may exist in these populations.

Allele-Frequency Differences in Terms of SNP Location

There were several important differences in the allele frequencies for SNPs, in terms of the functional class and the location of the SNP within the gene (table 3). Most (70%) of the SNPs in this sample were nonsynonymous substitutions located in the coding region of the genes. The average minor-allele frequencies for SNPs located in noncoding regions (20%–25%) were significantly higher than the average minor-allele frequencies for SNPs located in coding regions (12%–15%) ($P \leq .05$ for all populations except European Americans, according to the Wilcoxon rank-sum test). This may reflect the deleterious effect of mutations in the coding regions of genes, suggesting that the low minor-allele frequencies of SNPs in coding regions are caused by the young age of the mutations. The average minor-allele frequencies for SNPs located within the promotor region (23%–27%) were higher than the average minor-allele frequencies for SNPs within other noncoding regions (12%–19%). The five SNPs with either a frameshift mutation or the deletion of an entire amino acid had the lowest average minor-allele frequencies (0%–1%). These mutations may produce more-deleterious alterations to the gene product, which may result in a high selection against maintaining these polymorphisms in the population. The average minor-allele frequencies for synonymous substitutions (15%–20%) were higher than the average minor-allele frequencies for nonsynonymous substitutions (12%–16%). The synonymous substitutions do not alter the gene product, so we may expect a reduced effect of selection for synonymous substitutions contributing to the higher minor-allele frequencies for these SNPs. Finally, the conservative substitutions (11%–15%) had slightly lower average minor-allele frequencies than did the nonconservative substitutions (13%–17%), although these differences were not statistically significant ($P > .1$, Wilcoxon rank-sum test). A test of the difference in the minor-allele frequencies was not performed for some of the above comparisons because of small sample sizes, unless indicated otherwise.

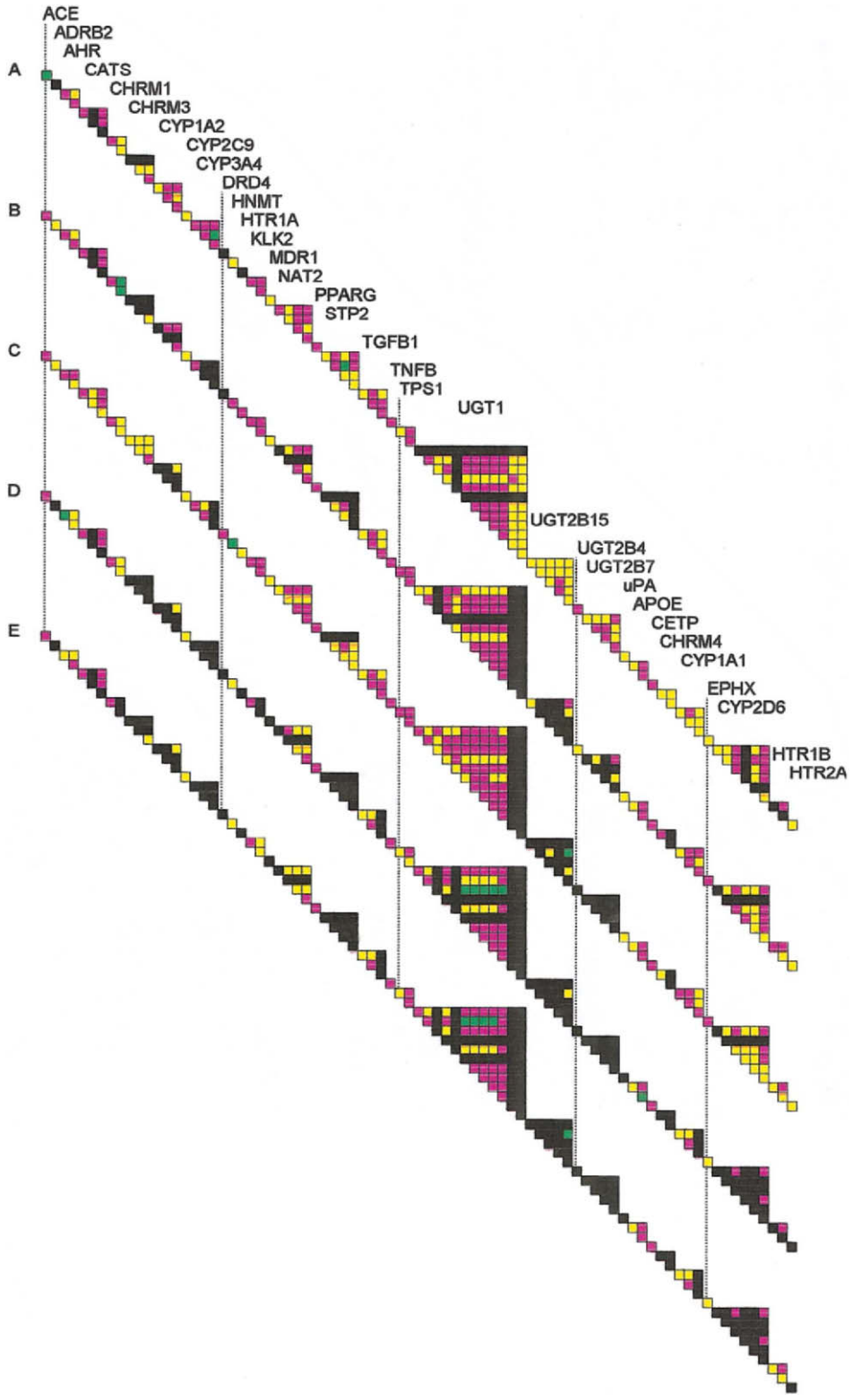

Figure 2 *P* values for the intermarker LD measure, *d*. Each graph represents a single population: African American (A), European American (B), Hispanic American (C), Chinese American (D), and Japanese American (E). Colors indicate the following categories: *pink,* significant *P* value for test of LD ($P \leq .001$); *yellow,* low power to detect LD ($P > .05$ and $q_{min} < .05$ or minor alleles in repulsion phase); *green,* high power to detect LD but LD not detected ($q_{min} > .05$, $P > .05$, minor alleles in coupling phase); and *black,* no variability for at least one locus, so unable to test for LD ($q_{min} = 0$). Pairwise LD shown only for markers within the same gene. The label above each group of SNPs indicates the gene.

Table 3

Minor-Allele Frequencies Dependent on Functional Class and Location within the Gene

Functional Mutation Class	No. of SNPs	Mean (SD) of the Minor-Allele Frequency in African Americans	European Americans	Hispanic Americans	Chinese Americans	Japanese Americans
Coding:	90	.14 (.16)	.15 (.17)	.15 (.16)	.12 (.15)	.12 (.16)
Insertion/deletion	5	.00 (.00)	.01 (.01)	.01 (.01)	.00 (.00)	.00 (.00)
Synonymous	9	.16 (.16)	.18 (.18)	.20 (.16)	.15 (.14)	.17 (.18)
Nonsynonymous:	76	.15 (.16)	.16 (.17)	.15 (.16)	.12 (.16)	.13 (.16)
Conservative	28	.14 (.18)	.14 (.16)	.15 (.17)	.11 (.16)	.12 (.18)
Nonconservative	48	.15 (.15)	.17 (.18)	.15 (.16)	.13 (.15)	.13 (.16)
Noncoding:	18	.23 (.15)	.22 (.17)	.25 (.17)	.22 (.19)	.20 (.18)
Promoter	13	.25 (.16)	.26 (.16)	.27 (.16)	.26 (.18)	.23 (.18)
Other	5	.17 (.13)	.15 (.19)	.19 (.22)	.12 (.17)	.12 (.17)

For all of the categories, the variability of the minor-allele frequency was high, indicating the influence of factors such as genetic drift on the allele frequencies.

Differences in the Distribution of LD, in Terms of the SNP Location

For nonsynonymous substitutions, our results suggest a relationship between the strength of LD and whether the substitutions were conservative or nonconservative (table 4). Although the sample size is small for some cells, we find that, when both SNPs are nonconservative, the measure of disequilibrium tends to be high, and the test of LD is more likely to be significant ($P \leq .05$). For example, for the African American population, none of the pairs of SNPs had a $P \leq .05$ when both SNPs were conservative, 28% of the pairs of SNPs had a $P \leq .05$ when one SNP was conservative, and the other SNP was nonconservative, and 48% of the pairs of SNPs had a $P \leq .05$ when both SNPs were nonconservative. This pattern is consistent for all of the populations considered here and does not appear to be caused by a difference in the proportion of SNPs with a low minor-allele frequency ($q_{\min} \leq .05$).

The relationship between the strength of disequilibrium and the location of the SNPs within the coding versus the noncoding region is less clear. The strength of intermarker LD for SNPs in coding versus noncoding regions is not consistent across populations. For example, in the African American and European American populations, the mean of the magnitude of intermarker LD is higher when both SNPs are in noncoding regions (.33–.51) than when both SNPs are in coding regions (.23–.32). However, for the Chinese American and Japanese Americans, the mean of the magnitude of intermarker LD is higher when both SNPs are in coding regions (.40–.41) than when both SNPs are in noncoding regions (.06–.12). Small sample sizes may contribute to the lack of consistency among the populations. Alternatively, this may reflect differences in the population histories for these markers. Additional markers should be evaluated to determine whether any general conclusions can be made about the strength of disequilibrium for coding versus noncoding SNPs.

Discussion

To investigate the marker characteristics that may affect the value of SNPs for linkage and association mapping, we compared the allele frequency and the LD distribution for 114 SNP markers in five populations. The African Americans had the largest number of SNPs in which both alleles were observed, whereas the Japanese Americans and Chinese Americans had the smallest number of SNPs with variability. The correlation of the allele frequencies was high ($\rho > .8$) between all of the populations, although the Japanese American and Chinese Americans had the most similar allele frequencies. The correlation in the LD measure was high ($\rho > .87$) among the Japanese Americans and Chinese Americans and also among the African Americans, European Americans, and Hispanic Americans. However, the correlation in LD across these two groupings (e.g., Japanese American vs. African Americans) was substantially lower ($\rho < .65$). LD was detected ($P \leq .05$) for pairs of SNPs within the same gene 82% of the time when the minor-allele frequency was high for both markers ($q_{\min} \geq .05$). If $P \leq .001$ is used as a criterion for detection of LD, as in the study by Clark et al. (1998), LD is detected for pairs of SNPs within the same gene in 72% of the cases when the minor-allele frequency is high for both markers ($q_{\min} \geq .05$). The deficiency in detectable LD for pairs of SNPs where the minor-allele frequency is low for at least one of the markers ($q_{\min} < .05$) does not necessarily indicate a lack of LD for these markers but more likely reflects a low power to detect LD in these situations. The intermarker distances alone could not explain the lack of observed LD for the remaining situations, since LD was observed for these pairs of SNPs in other pop-

Table 4

LD between Conservative and Nonconservative Pairs of SNPs

	Mean (SD) Absolute Value of d			No. of SNP Pairs/Total SNP Pairs (Proportion) with P Value ≤.05 in LD Test					
				Cn/Cn[a]		Cn/NCn[b]		NCn/NCn[c]	
Population	Cn/Cn[a]	Cn/NCn[b]	NCn/NCn[c]	$q_{min} \leq .05$	$q_{min} > .05$	$q_{min} \leq .05$	$q_{min} > .05$	$q_{min} \leq .05$	$q_{min} > .05$
African American	.05 (.12)	.18 (.22)	.29 (.27)	0/11 (0)	0/1 (0)	5/41 (.12)	11/16 (.69)	3/30 (.10)	25/28 (.89)
European American	.15 (.22)	.37 (.32)	.34 (.32)	0/2 (0)	1/1 (1)	1/15 (.07)	13/15 (.87)	1/23 (.04)	25/27 (.92)
Hispanic American	.21 (.20)	.30 (.27)	.26 (.30)	0/2 (0)	1/1 (1)	1/16 (.06)	13/14 (.93)	11/33 (.33)	24/26 (.92)
Chinese American	…	.27 (.33)	.45 (.31)	…	…	2/6 (.33)	2/10 (.20)	1/4 (.25)	19/30 (.63)
Japanese American	…	.33 (.31)	.40 (.32)	…	…	1/5 (.20)	6/10 (.60)	2/5 (.40)	19/29 (.66)

[a] Both SNPs conservative.
[b] One SNP conservative, one SNP nonconservative.
[c] Both SNPs nonconservative.

ulations. These results suggest that population differences in the allele-frequency and LD distributions should be considered when SNPs are selected for an association or linkage mapping study.

Our comparison of allele-frequency and LD distributions for different locations within the gene revealed some interesting observations. We found significantly lower average minor-allele frequencies for SNPs that are located in coding versus noncoding regions. Cargill et al. (1999) also found lower average minor-allele frequencies for SNPs in coding (12%) versus noncoding (13%) regions. In addition, we found higher average minor-allele frequencies for synonymous versus nonsynonymous mutations, which is consistent with the results of both Cargill et al. (1999) and Halushka et al. (1999). The higher average minor-allele frequency for synonymous mutations suggests a stronger selection against polymorphisms that cause an amino acid change in the protein product. Although we found higher average minor-allele frequencies for nonconservative versus conservative substitutions, Cargill et al. (1999) found the opposite result, with slightly higher minor-allele frequencies for conservative (11%) versus nonconservative substitutions (7%). Low power to detect a statistically significant difference in allele frequencies for conservative versus nonconservative substitutions probably contributes to the inconsistent results among studies. In the present study, LD appears to be stronger when both SNPs are nonconservative substitutions for all of the populations evaluated. However, the sample sizes were particularly small when both SNPs were conservative substitutions. The results were less clear on the strength of LD for coding versus noncoding pairs of SNPs. Additional data are needed for clarification of whether the strength of LD varies depending on the location of the SNPs within the gene.

Our observations on population genetic diversity parallel the results from other studies. The African American population had the largest number of SNPs with variability and the largest number of markers in which the major-allele frequency was high (>.95). The greater genetic diversity observed among the African Americans is consistent with the findings of other studies using SNPs (Zietkiewicz et al. 1997; Nickerson et al. 1998) or microsatellites (Bowcock et al. 1994; Jorde et al. 1995; Jorde et al. 1997; Pérez-Lezaun et al. 1997). Many of these studies have also reported a greater genetic diversity among European populations than among Asian populations, although this difference was not shown to be statistically significant. Several hypotheses have been suggested to explain the greater genetic diversity among African populations, including admixture, an older population (e.g., see Cavalli-Sforza et al. 1994), a larger effective population size (e.g., see Relethford and Jorde 1999), and gene flow (e.g., see Zietkiewicz et al. 1997). As noted above, the Japanese American and Chinese American populations had very similar allele frequencies and LD measures. Several reports have indicated that the Japanese and Chinese populations may have a more recent common population history than do the other populations in this study, which would increase the similarity of the allele frequencies and LD in these populations (Cavalli-Sforza et al. 1994; Bowcock et al. 1994; Jorde et al. 1997; Pérez-Lezaun et al. 1997; Zietkiewicz et al. 1997). Furthermore, the European American, African American, and Hispanic American populations also had similar allele frequencies and LD measures. Admixture among the European American, Hispanic American, and African American populations could at least partially explain this observation. For Hispanic American populations, estimates of the admixture proportions are 45%–68% for the European contribution and 3%–37% for the African contribution (Hanis et al. 1991; Long et al. 1991; Tseng et al. 1998). For the African American population, the admixture proportion is

~25% for the European contribution (Chakraborty et al. 1992; Destro-Bisol et al. 1999).

There are a few limitations to our analysis that are worth noting. First, the SNPs evaluated here were not randomly selected across the genome. Therefore, the distributions of allele frequencies and LD observed here may not necessarily be representative of the corresponding distributions for all sites in the genome. In particular, these markers were primarily in genes that may have potential pharmacological effects, and they were preferentially chosen from or detected in the coding regions of these genes. Selection and mutation may behave differently in coding regions, compared with other sites in the genome, which may alter the distribution of allele frequencies or LD. In addition, in this data set there were a large number of SNPs with a minor-allele frequency <.05 for at least one of the populations. Markers with equally frequent alleles are the most informative for linkage, so the SNPs in this data set do not represent the optimal distribution of allele frequencies. It is interesting to note that, although this was not a random sample of sites, the distribution of allele frequencies that we observed is very similar to the distribution observed in studies that included all SNPs detected in regions with both coding and noncoding sequences (Nickerson et al. 1998; Cargill et al. 1999; Halushka et al. 1999). Nevertheless, the markers in this sample may be representative of SNPs that one might use in either linkage analysis or an association study. Most (71%) of the polymorphic sites that we considered cause amino acid changes in the gene product and are good candidates to consider as disease-causing mutations. In addition, this sample reflects the distribution of allele frequencies and LD for a wide variety of locations in the genome, compared with many previous studies that focused on a small genomic region.

Second, ascertainment bias of the polymorphisms may also influence the distribution of allele frequencies and LD that we observed. One might expect that populations used to identify the polymorphisms generally have more SNPs with variability than do populations that were not used to identify the polymorphisms. The 44 SNPs identified from the literature were detected in numerous different populations, so it is unclear how ascertainment bias of the polymorphisms affected the distributions of allele frequencies and LD for these markers. Of the 70 SNPs identified by resequencing, 67 were detected in a defined sample composed of individuals from the African American, European American, and Hispanic American populations. These populations had more SNPs with variability, compared with the Chinese American and Japanese American populations, which may reflect an ascertainment bias. However, our results are consistent with other studies using either microsatellites (Jorde et al. 1997) or SNPs (Zietkiewicz et al. 1997; Cargill et al. 1999) that do not have an ascertainment bias. We did not find a difference in our results on allele frequencies when we stratified on the method used to identify the SNPs.

Finally, we did not evaluate the relationship between LD and physical distance, since the distance between the SNPs was not known for most of the markers. Clark et al. (1998) found that intermarker LD was not always detectable for SNPs within a 9.7-kb region near the human lipoprotein lipase gene, which is consistent with our results. In a review of 19 disequilibrium studies, Jorde et al. (1994) showed that there is a low correlation between physical distance and measures of LD, for markers that are <75 kb apart. The physical distances for SNPs in this study may be within this range, since we considered intermarker LD only for SNPs within the same gene. Furthermore, the presence or absence of LD did not correspond to the relative order of SNPs, determined from coding sequence information, within two genes with numerous markers (UGT1 and CYP2D6, marker order as in fig. 2). Although, in our study, physical distance is unlikely to be a major explanation for the situations in which LD was not observed, an evaluation of the relationship between LD and physical distance will provide important information for association mapping studies. For example, Kruglyak (1999) suggested that a SNP marker density of one marker per 3 kb might be necessary for LD mapping in complex diseases. Empirical observations on the relationship between LD and physical distance are needed to determine the optimal marker density for LD mapping studies.

The distribution of trait-marker LD may be substantially different for complex traits than for simple traits. There are numerous examples of the distribution of trait-marker LD for genomic regions near loci that influence simple traits such as diastrophic dysplasia (Hästbacka et al. 1992), Huntington disease (Huntington's Disease Collaborative Group 1993), and Werners syndrome (Goddard et al. 1996). LD mapping studies such as these are often conducted in genetically homogeneous populations, for a rare trait with a high penetrance. Under these circumstances, LD has been found for markers $\geqslant$500 kb away from the mutation (Jorde et al. 1994). In contrast, genetic polymorphisms that influence complex traits may have major alleles with a small effect and, potentially, multiple different mutations represented in the study population. Therefore, factors that influence the presence and extent of LD—such as the age of the mutation, selection, and the number of independent mutations—will differ for complex and simple traits. These factors may be more similar for SNPs and complex-trait loci, suggesting that the intermarker LD distribution for SNPs may be indicative of the trait-marker LD distribution for complex traits. In particular, the effect of selection will be small for both SNPs and complex-trait loci, since SNPs are thought to be neutral

mutations in most cases and since complex-trait loci may have only a small effect on the disease phenotype. In addition, both SNPs and complex-trait loci have major-allele frequencies and are observed in multiple populations, indicating that both types of polymorphisms may have a similar (old) age.

Our results have important implications for the selection of SNP markers for association and linkage mapping studies. When possible, markers should be selected to have high allele frequencies (e.g., minor-allele frequency $\geq .05$) in all of the populations under consideration for association studies. As suggested by previous studies, the higher allele frequencies increase power to detect LD—although, as we observed here, this does not guarantee that LD is detectable for all sites within the same gene. In addition, we have noted potential differences in the strength of LD in terms of the functional class of the mutation, indicating that the location of the SNP within the gene may be an important factor in the selection of SNPs for association or linkage studies. For genetic linkage mapping studies using clustered or uniform marker spacing, our results indicate that SNPs should be carefully considered for inclusion in a screening set that might be used with multiple populations. In particular, for the clustered marker spacing, the inter-marker LD that we observed for pairs of markers within the same gene generally reduces the information content of the cluster (Goddard 1999). Moreover, for both uniform and clustered marker spacing, it may be necessary to include multiple SNPs for each region in a genome screen, to ensure that at least one marker is informative for each population under consideration. These factors may reduce the potential value of the use of SNPs, compared with the current genotyping methods.

Acknowledgments

We thank R. Elston, E. Wijsman, and two anonymous reviewers for their helpful comments. This work was supported in part by National Institutes of Health grant CA73270 (to J. S. W.) and Department of Defense grant DAMD17-98-1-8589 (to J. S. W. and K. A. B. G.).

Appendix A

Table A1

SNP Allele Frequencies and Hardy-Weinberg Equilibrium for Each Population

SNP	Gene	African Americans	European Americans	Hispanic Americans	Chinese Americans	Japanese Americans
1	ACE	.30	.54	.43	.35	.38
2		.68	.60	.70	.65	.62
3	ADRB2	.48	.59	.57	.40	.54
4		1.00	.98	.99	1.00	1.00
5	AHR	.86	.78	.68	.60	.69
6		.58	.89[a]	.86[a]	.69	.49
7		.97	.93	.97[a]	.97	.98
8	CATS	.14	.58	.30	.10	.08
9		.74	.66	.61	.57	.68
10		1.00	1.00	.98	1.00	1.00
11		.74	.66	.61	.63	.68
12	CHRM1	.97	.94	.77	.92	.95
13		.97	.93	.76	.92	.94
14		.91	.95	.97	.98	1.00
15	CHRM3	1.00	1.00	.99	1.00	1.00
16		.05	.00	.01	.00	.00
17		.93[a]	.96	.98	1.00	.99
18		.78	.47	.55	.59	.41
19	CYP1A2	.90	.98	.97	.90	.97
20		.99	1.00	1.00	1.00	1.00
21		.91	1.00	1.00	1.00	1.00
22		.13	.63	.32	.14	.16
23	CYP2C9	.95	.87	.90	.99	.99
24		.98	.93	.96	.99	.96
25	CYP3A4	.45	.97	.91	1.00	1.00
26		.33	.90[a]	.64	.75	.80
27		.76	1.00	.99	1.00	1.00

(*continued*)

Table A1 (continued)

SNP	Gene	African Americans	European Americans	Hispanic Americans	Chinese Americans	Japanese Americans
28		.94	1.00	1.00	1.00	1.00
29	DRD4	1.00	.99	1.00	1.00	1.00
30		.95	1.00	.99	1.00	1.00
31	EPHX	.69	.77	.88	.85	.80
32		.79	.74	.62	.61	.56
33	HNMT	.61	.63	.67	.65	.67
34		.97	.91	.90	.96[a]	.96
35	HTR1A	1.00	.99	.99	1.00	1.00
36		1.00	.99	.99	1.00	1.00
37	KLK2	.35	.68	.69	.68	.66
38		.57	.79	.77	.79	.73
39		.97	.82	.91[a]	1.00	.99
40	MDR1	.98	.94	.94	1.00	1.00
41		.98	.99	.99	1.00	1.00
42	NAT2	.72	.78	.81	.72	.83
43		.92	1.00	1.00	1.00	1.00
44		.98	.97	.89	.91	.89
45		.68	.49	.69	.99	.97
46		.56	.60	.64	.96	.97
47	PPARG	.94	.90	.90	.81	.81
48		.97	.88	.90	.96	.94
49	STP2	.997	1.00	1.00	1.00	1.00
50		.75	.65	.62	.93	.85
51		.95	1.00	.99	1.00	1.00
52		.93	.87	.91	1.00	1.00
53		.98	1.00	.99	1.00	1.00
54	TGFB1	.94[a]	.92	.96	.99	.99
55		.77	.68	.53	.40	.53
56		.58	.60	.49	.40	.54
57		.99	.99	.99	1.00	1.00
58	TNFB	.45	.70	.67	.54	.58[a]
59		.46[a]	.69	.66	.54	.60[a]
60	TPS1	.56	.57	.56	.14	.10
61		.97	.94	.97	.99	.99
62		.67	.84	.70	.91	.94
63	UGT1	1.00	1.00	.97	.83	.86
64		.44	.52[a]	.54	.70	.76
65		.35	.59	.62	.90	.88
66		1.00	1.00	.99	1.00	1.00
67		.94	.90	.89	.78	.90
68		1.00	.997	.997	1.00	1.00
69		.68	.64[a]	.72	.79	.74
70		.61	.57[a]	.70	.79	.75
71		.75	.67[a]	.77	.82	.76
72		.74	.60	.75	.82	.76
73		.08	.26	.32	.55	.52
74		.997	1.00	1.00	1.00	1.00
75		.98	1.00	1.00	1.00	1.00
76	UGT2B15	.99	.98	1.00	1.00	1.00
77		.62	.44[a]	.63	.58	.51
78		.997	1.00	1.00	1.00	1.00
79		.98	1.00	1.00	1.00	1.00
80		.997	1.00	1.00	1.00	1.00
81		.24	.64	.57	.15	.22
82	UGT2B4	.83	.79	.72	.99	.99
83		.95	1.00	1.00	1.00	1.00
84	UGT2B7	.33	.51	.31	.34	.34
85		.99	1.00	1.00	1.00	1.00

(*continued*)

Table A1 (continued)

SNP	Gene	African Americans	European Americans	Hispanic Americans	Chinese Americans	Japanese Americans
86		.99	.99	1.00	1.00	1.00
87		.997	1.00	1.00	1.00	1.00
88		.97	1.00	1.00	1.00	1.00
89	uPA	.94	.79	.72	.62	.78
90		.84	1.00	.99	1.00	1.00
91	APOE	.90	.92	.96	.92	.9512fna[a]
92		.16	.12	.09	.06	.10
93		.74	.55	.41	.28	.28
94	CETP	.78	.5512fna[a]	.52	.63	.62
95		.47	.6812fna[a]	.62	.63	.47
96	CHRM4	.89	1.00	.98	1.00	1.00
97		.89	.83	.91	.95	.93
98		.9912fna[a]	1.00	1.00	1.00	1.00
99	CYP1A1	.06	.31	.14	.02	.01
100		.98	.95	.6712fna[a]	.75	.70
101		.47	.84	.5612fna[a]	.5312fna[a]	.43
102		.98	.97	.9612fna[a]	1.00	1.00
103	CYP2D6	.9012fna[a]	.80	.85	.47	.5612fna[a]
104		.99	1.00	1.00	1.00	1.00
105		1.00	.97	.97	1.00	1.00
106		.5512fna[a]	.5212fna[a]	.40	.73	.53
107		.00	.02	.01	.00	.00
108		.99	1.00	.99	1.00	1.00
109		.6012fna[a]	.52	.41	.45	.5512fna[a]
110	HTR1B	.76	.72	.61	.49	.56
111		1.00	.98	.99	1.00	.99
112		.76	.56	.58	.85	.87
113	HTR2A	.86	.92	.93	.99	.99
114		.99	.98	1.00	1.00	1.00

[a] Marker is not in Hardy-Weinberg equilibrium ($P \leq .05$).

Appendix B

Table B1

Measure of LD for Each Pair of SNPs within a Gene

			Population				
Gene	SNP Pair		African American	European American	Hispanic American	Chinese American	Japanese American
ACE	1	2	−.11[†]	−.68[**]	−.67[**]	−.72[**]	−.89[**]
ADRB2	3	4	…	−.42	−.43	…	…
AHR	5	6	−.24[**]	−.24[**]	−.37[**]	.14[†]	.02
	5	7	−.14	−.11	−.33[*]	−.41	.36
	6	7	−.43[*]	−.11[*]	.00	.22	−.53[*]
CATS	8	9	.15[*]	.85[**]	.45[**]	.18[**]	.11[**]
	8	10	…	…	.29	…	…
	8	11	.13[*]	.86[**]	.46[**]	.16[**]	.11[**]
	9	10	…	…	.11	…	…
	9	11	.91[**]	.92[**]	.94[**]	.86[**]	.90[**]
	10	11	…	…	.02	…	…
CHRM1	12	13	…	…	.97[**]	…	.88[**]
	12	14	−.03	.04[†]	−.02	−.09	…
	13	14	−.03	.02[†]	−.03	−.08	…

(*continued*)

Table B1 (continued)

Gene	SNP Pair		Population				
			African American	European American	Hispanic American	Chinese American	Japanese American
CHRM3	15	16	…	…	.01	…	…
	15	17	…	…	.11	…	…
	15	18	…	…	−.01	…	…
	16	17	.00	…	.01	…	…
	16	18	.10	…	.02	…	…
	17	18	−.09*	.05	.05*	…	.02
CYP1A2	19	20	−.10	…	…	…	…
	19	21	−.11*	−.02**	…	…	…
	19	22	.11*	.04**	.03	.11	.04
	20	21	.09**	…	…	…	…
	20	22	−.01	…	…	…	…
	21	22	.10*	.01**	…	…	…
CYP2C9	23	24	−.05	−.14	−.11	−.01	−.01
CYP3A4	25	26	.52**	.32**	.13*	…	…
	25	27	.40**	…	.24	…	…
	25	28	.46*	…	…	…	…
	26	27	.44**	…	.65*	…	…
	26	28	.05†	…	…	…	…
	27	28	.53*	…	…	…	…
DRD4	29	30	…	…	.00*	…	…
EPHX	31	32	−.14	−.13*	−.12*	−.03	−.17
HNMT	33	34	.25	.58**	.15†	.05	.44
HTR1A	35	36	…	−.01*	−.01	…	…
KLK2	37	38	.32**	.59**	.63**	.71**	.79**
	37	39	−.67**	−.40**	−.35**	…	−.34
	38	39	−.44*	−.26**	−.26**	…	−.02
MDR1	40	41	−.02	.41	−.06	…	…
NAT2	42	43	−.31**	…	−.19	…	…
	42	44	−.27	−.23	−.22*	−.31*	−.06
	42	45	−.42**	−.46**	−.27**	−.30	.08
	42	46	−.38**	−.35**	−.32**	−.29	.08
	43	44	−.08	…	.00	…	…
	43	45	−.12*	…	.00	…	…
	43	46	−.08*	…	.00	…	…
	44	45	−.03*	−.07**	−.16**	.36	−.01
	44	46	−.02	−.06	−.13**	−.03	−.01
	45	46	.65**	.78**	.81**	.19**	…?
PPARG	47	48	.67**	.57**	.81**	.49*	.59**
STP2	49	50	.00	…	…	…	…
	49	51	.05**	…	…	…	…
	49	52	.00	…	…	…	…
	49	53	.00*	…	…	…	…
	50	51	−.27**	−.36	−.23	…	…
	50	52	−.07†	—.40**	−.41**	…	…
	50	53	.76**	…	.62	…	…
	51	52	−.04	.00	−.01	…	…
	51	53	−.05	…	−.01	…	…
	52	53	−.07	…	−.09	…	…
TGFB1	54	55	−.03	−.03	−.05	−.01	.00
	54	56	.13**	.17**	.06*	…	…
	54	57	.24	−.07	−.03	…	…
	55	56	.56**	.80**	.93**	.99**	.97**
	55	57	.78**	.39	.54**	…	…
	56	57	.59*	.60*	.50**	…	…
TNFB	58	59	.97**	.97**	.99**	…	.98**
TPS1	60	61	−.45	−.46**	−.46*	−.87	−.91

(*continued*)

Table B1 (continued)

Gene	SNP Pair		Population				
			African American	European American	Hispanic American	Chinese American	Japanese American
(TPS1)	60	62	−.66**	−.51**	−.63**	−.94**	−.96**
	61	62	.08**	.35**	.12**	.07	.11**
UGT1	63	64	…	.01	−.05*	−.25*	−.17*
	63	65	…	.00	−.04	−.18	−.15
	63	66	…	…	−.03*	…	…
	63	67	…	.03**	−.03	−.22*	−.16
	63	68	…	.00**	−.03	…	…
	63	69	…	.01	.09**	.70**	.51**
	63	70	…	.01	.09**	.70**	.50**
	63	71	…	.01	.12**	.80**	.54**
	63	72	…	.01	.11**	.80**	.56**
	63	73	…	−.01	.04	.31**	.22*
	63	74	…	…	…	…	…
	63	75	…	…	…	…	…
	64	65	.62**	.89**	.82**	.76**	.87**
	64	66	−.56	…	.04	…	…
	64	67	.12	.31*	.37**	.81**	.76**
	64	68	…	.52	.54**	…	…
	64	69	.52**	.69**	.52**	−.08	.24†
	64	70	.54**	.79**	.55**	−.08	.25†
	64	71	.45**	.73**	.46**	−.13	.17†
	64	72	.44**	.81**	.50**	−.13	.15†
	64	73	.55**	.50**	.36**	.33**	.43**
	64	74	.44	…	…	…	…
	64	75	−.35	…	…	…	…
	65	66	.35	…	.12	…	…
	65	67	−.56**	−.38**	−.43**	−.12	−.14
	65	68	…	.59	.62**	…	…
	65	69	.44**	.77**	.66**	.13†	.29**
	65	70	.48**	.89**	.69**	.13†	.30**
	65	71	.41**	.82**	.61**	−.01†	.21*
	65	72	.42**	.91**	.65**	−.01†	.20*
	65	73	.66**	.54**	.48**	.02†	.26**
	65	74	.35	…	…	…	…
	65	75	−.43	…	…	…	…
	66	67	.00	…	−.01	…	…
	66	68	…	…	1.00**	…	…
	66	69	.00	…	.01	…	…
	66	70	.00	…	.00	…	…
	66	71	.00	…	.01	…	…
	66	72	.00	…	.01	…	…
	66	73	.00	…	.01	…	…
	66	74	.00*	…	…	…	…
	66	75	.00	…	…	…	…
	67	68	…	−.10	−.11	…	…
	67	69	−.08**	−.12*	−.15**	−.28	−.14
	67	70	−.07*	−.14**	−.13**	−.28	−.14
	67	71	−.08**	−.14*	−.14**	−.27	−.14
	67	72	−.08**	−.16**	−.14**	−.27	−.14
	67	73	−.17*	−.16**	−.18**	.33**	.14*
	67	74	−.06	…	…	…	…
	67	75	.06	…	…	…	…
	68	69	…	.01	.01	…	…
	68	70	…	.01	.01**	…	…
	68	71	…	.01	.01**	…	…
	68	72	…	.01	.01	…	…

(*continued*)

Table B1 (continued)

Gene	SNP Pair		Population				
			African American	European American	Hispanic American	Chinese American	Japanese American
(UGT1)	68	73	...	−.01	.00	...	...
	68	74	...	...	...	...	...
	68	75	...	...	...	...	...
	69	70	.84**	.85**	.92**	...	...
	69	71	.91**	.96**	.93**	.97**	.98**
	69	72	.84**	.78**	.84**	.97**	.98**
	69	73	.35**	.44**	.41**	.43**	.50**
	69	74	−.33	...	...	...	...
	69	75	−.06	...	...	...	...
	70	71	.82**	.86**	.90**	.97**	.98**
	70	72	.83**	.94**	.93**	.97**	.98**
	70	73	.42**	.53**	.45**	.43**	.49**
	70	74	−.39	...	...	...	...
	70	75	−.07	...	...	...	...
	71	72	.94**	.83**	.90**	...	...
	71	73	.27**	.42**	.33**	.38**	.47**
	71	74	−.25	...	...	...	...
	71	75	−.12	...	...	...	...
	72	73	.29**	.52**	.37**	.38**	.47**
	72	74	−.27	...	...	...	...
	72	75	−.17	...	...	...	...
	73	74	.08	...	...	...	...
	73	75	.01	...	...	...	...
	74	75	.00	...	...	...	...
UGT2B15	76	77	.00	−.01	...	...	...
	76	78	−.01	...	...	...	...
	76	79	−.01	...	...	...	...
	76	80	−.01	...	...	...	...
	76	81	.01	.04**	...	...	...
	77	78	.62	...	...	...	...
	77	79	−.39	...	−.38	...	...
	77	80	.62	...	...	...	...
	77	81	.02	.03	.11†	−.26	−.11†
	78	79	.00	...	...	...	...
	78	80	.00**	...	...	...	...
	78	81	.00	...	...	...	...
	79	80	−.02**	...	...	...	...
	79	81	−.01	...	.01	...	...
	80	81	−.01	...	...	...	...
UGT2B4	82	83	.81**	.79	...	...	...
UGT2B7	84	85	−.02	...	...	...	...
	84	86	.33	.52	...	...	...
	84	87	−.67	...	...	...	...
	84	88	.34**	.51**	...	...	...
	85	86	−.01*	...	...	...	...
	85	87	−.01*	...	...	...	...
	85	88	−.01	...	...	...	...
	86	87	−.01*	...	...	...	...
	86	88	−.01	...	...	...	...
	87	88	.00*	...	...	...	...
uPA	89	90	−.05	−.21	.31	...	...
APOE	91	92	.11	.09	−.01	.09	.05
	91	93	−.13**	−.14**	−.10**	−.30**	−.14**
	92	93	−.24**	−.20**	−.09*	−.09†	−.15*
CETP	94	95	.11*	.36**	.38**	.43**	.52**
CHRM4	96	97	−.12	.02*	−.02	...	...

(*continued*)

Table B1 (continued)

Gene	SNP Pair		Population				
			African American	European American	Hispanic American	Chinese American	Japanese American
(CHRM4)	96	98	−.11	…	…	…	…
	97	98	−.11	…	…	…	…
CYP1A1	99	100	.06	.32	.2124fnb**	.03	.01
	99	101	.1224fnb**	.3724fnb**	.2524fnb**	.04	.01
	99	102	.06	.3224fnb**	.15	…	…
	100	101	.04	.2724fnb**	.7724fnb**	.5024fnb**	.5324fnb**
	100	102	−.02	−.05	−.23	…	…
	101	102	.47	.8724fnb**	.4624fna*	…	…
CYP2D6	103	104	−.10	…	…	…	…
	103	105	−.11	−.2	−.01	…	…
	103	106	−.1924fnb**	−.3824fnb**	−.3824fnb**	−.7524fnb**	−.8524fnb**
	103	107	…	.09	−.04	…	…
	103	108	−.11	−.20	−.16	…	…
	103	109	−.1624fnb**	−.3824fnb**	−.3824fnb**	.9824fnb**	.9724fnb**
	104	105	−.01	…	…	…	…
	104	106	.5524fnb**	…	…	…	…
	104	107	…	…	…	…	…
	104	108	−.0124fna*	…	…	…	…
	104	109	.6124fna*	…	…	…	…
	105	106	.5524fna*	.5324fnb**	.41	…	…
	105	107	…	.02	.01	…	…
	105	108	−.01	.00	−.01	…	…
	105	109	.6024fnb**	.5424fnb**	.4124fna*	…	…
	106	107	…	−.0324fna*	−.01	…	…
	106	108	.01	.01	.01	…	…
	106	109	.9224fnb**	.9824fnb**	.9924fnb**	−.5724fnb**	−.8224fnb**
	107	108	…	.00	.01	…	…
	107	109	…	−.0324fna*	−.01	…	…
	108	109	.01	.01	.01	…	…
HTR1B	110	111	…	.7324fna*	.62	…	-.44
	110	112	−.3124fnb**	−.5024fnb**	−.6624fnb**	−.6024fnb**	−.5124fnb**
	111	112	…	−.03	−.02	…	−.01
HTR2A	113	114	.26	−.02	−.07	…	…

Note.—The measure of LD was not calculated (…) when only one allele was observed for at least one SNP in the pair. Apparent discrepancies between the measure of LD and the *P* value for the test of LD (e.g., the measure of LD is −.02 and the *P* value for the test of LD is ⩽.001) generally occur when the minor-allele frequency is very low for at least one of the markers.

* $P \leqslant .05$ for test of LD.

** $P \leqslant .001$ for test of LD.

† $P > .05$ for test of LD but high power to detect LD.

References

Bowcock AM, Hebert JM, Mountain JL, Kidd JR, Kidd KK, Cavalli-Sforza LL (1991) Study of an additional 58 DNA markers in 5 populations. Gene Geogr 5:151–173

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Cambien F, Poirier O, Nicaud V, Hermann S, Mallet C, Ricard S, Behague I, et al (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders. Am J Hum Genet 65: 183–191

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, et al (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238

Carlson CS, Cox DR (1998) Linkage disequilibrium of SNPs on human chromosome 21. Am J Hum Genet 63:A284

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Chakraborty R, Mohammad KI, Nwankwo M, Ferrell RE (1992) Caucasian genes in American blacks: new data. Am J Hum Genet 50:145–155

Chapman NH, Wijsman EM (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics. Am J Hum Genet 63:1872–1885

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, et al (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63: 595–612

Dean M, Stephens JC, Winkler C, Lomb DA, Ramsburg M, Boaze R, Stewart C, et al (1994) Polymorphic admixture typing in human ethnic populations. Am J Hum Genet 55: 788–808

Destro-Bisol G, Maviglia R, Caglia A, Boschi I, Spedini G, Pascali V, Clark A, et al (1999) Estimating European admixture in African Americans by using microsatellites and a microsatellite haplotype (CD4/Alu). Hum Genet 104: 149–157

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

Elston RC, Buxbaum S, Jacobs KB, Olson JM. Haseman and Elston revisited. Genet Epidemiol (in press)

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Goddard KAB (1999) Design issues in the analysis of complex genetic traits. PhD diss, University of Washington, Seattle

Goddard KAB, Yu C-E, Oshima J, Miki T, Nakura J, Piussan C, Martin GM, et al (1996) Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. Am J Hum Genet 58:1286–1302

Goldin LR, Gershon ES (1988) Power of the affected sib-pair method for heterogeneous disorders. Genet Epidemiol 5: 35–42

Goldin LR, Weeks DE (1993) Two-locus models of disease: comparison of likelihood and nonparametric methods. Am J Hum Genet 53:908–915

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, et al (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22:239–247

Hanis CL, Hewett-Emmett D, Bertin TK, Schull WJ (1991) Origins of U.S. Hispanics: implications for diabetes. Diabetes Care 14:618–627

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Heid CA, Stevens J, Livak KJ, Williams PM (1996) Real time quantitative PCR. Genome Res 6:986–994

Huntington's Disease Collaborative Research Group, The (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72:971–983

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, et al (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, et al (1997) Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA 94:3100–3103

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A, Leppert M (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 54:884–898

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–24

——— (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22: 139–144

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Lai E, Riley J, Purvis I, Roses A (1998) A 4-Mb high-density single nucleotide polymorphism-based map around human ApoE. Genomics 54:31–38

Lewontin RC (1995) The detection of linkage disequilibrium in molecular sequence data. Genetics 140:377–388

Long JC, Williams RC, McAuley JE, Medis R, Partel R, Tregellas WM, South SF, et al (1991) Genetic variation in Arizona Mexican Americans: estimation and interpretation of admixture proportions. Am J Phys Anthropol 84:141–157

Nei M, Li WH (1980) Nonrandom association between electromorphs and inversion chromosomes in finite populations. Genet Res 35:65–83

Nickerson DA, Kaiser R, Lappin S, Stewart J, Hood L, Landergren U (1990) Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. Proc Natl Acad Sci USA 87:8923–8927

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, et al (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19:233–240

Nickerson DA, Whitehurst C, Boysen C, Charmley P, Kaiser R, Hood L (1992) Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly informative and automatable markers for genetic linkage mapping. Genomics 12:377–387

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor P (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Nat Acad Sci USA 91: 5022–5026

Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J (1997) Microsatellite variation and the differentiation of modern humans. Hum Genet 99:1–7

Relethford JH, Jorde LB (1999) Genetic evidence for larger African population size during recent human evolution. Am J Phys Anthropol 108:251–260

Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Slatkin M, Excoffier L (1995) Testing for linkage disequilib-

rium in genotypic data using the expectation-maximization algorithm. Heredity 76:377–383

Thompson EA, Deeb S, Walker D, Motulsky AG (1988) The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII Apolipoprotein genes. Am J Hum Genet 42:113–124

Tseng M, Williams RC, Maurer KR, Schanfield MS, Knowler WC, Everhart JE (1998) Genetic admixture and gallbladder disease in Mexican Americans. Am J Phys Anthropol 106: 361–371

Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Weir BS (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA, pp 125–128

Xiong M, Jin L (1999) Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. Am J Hum Genet 64:629–640

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, et al (1997) Nuclear DNA diversity in worldwide distributed human populations. Gene 205:161–171