Sciencexpress Report

Haplotype Variation and Linkage Disequilibrium in 313 Human Genes

J. Claiborne Stephens,* Julie A. Schneider, Debra A. Tanguay, Julie Choi, Tara Acharya, Scott E. Stanley, Ruhong Jiang, Chad J. Messer, Anne Chew, Jin-Hua Han, Jicheng Duan, Janet L. Carr, Min Seob Lee, Beena Koshy, A. Madan Kumar, Ge Zhang, William R. Newell, Andreas Windemuth, Chuanbo Xu, Theodore S. Kalbfleisch, Sandra L. Shaner, Kevin Arnold, Vincent Schulz, Connie M. Drysdale, Krishnan Nandabalan, Richard S. Judson, Gualberto Ruaño, Gerald F. Vovis

Genaissance Pharmaceuticals, Inc., Five Science Park, New Haven, CT 06511, USA.

*To whom correspondence should be addressed. E-mail: c.stephens@genaissance.com

Variation within genes has important implications for all biological traits. We identified 3899 single nucleotide polymorphisms (SNPs) that were present within 313 genes from 82 unrelated individuals of diverse ancestry, and we organized the SNPs into 4304 different haplotypes. Each gene had several variable SNPs and haplotypes that were present in all populations, as well as a number that were population-specific. Pairs of SNPs exhibited variability in the degree of linkage disequilibrium that was a function of their location within a gene, distance from each other, population distribution, and population frequency. Haplotypes generally had more information content (heterozygosity) than did individual SNPs. Our analysis of the pattern of variation strongly supports the recent expansion of the human population.

Large-scale investigations of sequence variation within the human species have only just begun (*1–3*). Initial estimates are that sequence differences between an individual's maternal and paternal genomes occur on average at about every 500 to 2000 bases (*2–4*), with the most frequently cited value being one difference approximately every kilobase (*5*). However, relatively little is known about the pattern of DNA sequence variation among humans, within a population and between different populations. In particular, the pattern of linkage disequilibrium among closely spaced SNPs, for example, those that are less than 20 kb apart, is known only for a few well-studied genes, and the results from these studies are highly discordant (*6–9*).

We have undertaken a systematic discovery of gene-based sequence variation in 82 unrelated individuals, whose ancestors were from various geographical origins. The sample size and composition were sufficient to detect, with high certainty, globally distributed variants present at a frequency of at least 2% and population-specific variants present at a frequency of at least 5%. Our population sample, using the definitions of the U.S. Census Bureau, was comprised of approximately an equal number of self-described Caucasians, African-Americans, Asians, and Hispanic-Latinos (*10*). Our goal was to identify SNPs and to organize them into their gene-specific allelic haplotypes. A haplotype is the specific combination of the nucleotides, one from each of the polymorphic sites that are present on an individual chromosome. We sequenced the exons (coding regions, 5′UTR and 3′UTR), up to 100 bases into the introns from the exon-intron boundaries (including the splice junctions), and the 5′ upstream genomic region (*11*). The 313 genes were chosen from those genes for which complete genomic organization was publicly available. To assist in assessing the quality of the sequence information and to validate the construction of haplotypes, we also included a three-generation Caucasian family and a two-generation African-American family. For evolutionary comparisons, we also sequenced the corresponding genomic regions from a chimpanzee. The position and sequence of the human polymorphisms have been deposited in GenBank.

We discovered 3899 polymorphic sites in nearly 720 kb of genomic sequence or an average of one SNP approximately every 185 bases. Less than 2% of these polymorphic sites were previously described (*12*). The average number of polymorphic sites per kilobase of DNA was 3.4 in the coding regions, 5.3 in the 5′UTR, 5.9 in the 5′ upstream region, 6.5 in the exon-intron boundaries, and 7.0 in the 3′UTR (*13*). Fifty-one of these polymorphisms were within splice sites (*14*). Of the 1033 polymorphic sites within the coding regions, 565 coded for an amino acid change, 459 did not result in an amino acid change, and nine changed an amino acid codon to a termination codon. In addition, we observed proportionately fewer polymorphisms that resulted in an amino acid change than have been observed in pseudogenes (Table 1), which are presumably subject to less stringent natural selection than are functional genes (*15*). Furthermore, of the mutations that resulted in an amino acid change, we identified fewer that caused either a radical change or a termination codon than have been observed in pseudogenes.

For 38% of the SNPs, the minor allele was observed only once as a single heterozygote (*16*). The occurrence of these variants was different among the four population samples. The African-American sample had 662 as compared with 294, 223, and 273 in the Asian, Caucasian, and Hispanic-Latino samples, respectively (*17*). In addition, the African-American sample had the greatest number of population-specific rare alleles that occurred only two, three, or four times (Table 2). For 32% of all SNPs, the minor allele frequency was between 1 and 5% (*17*). For about 17%, the minor allele frequency was between 5 and 20%, and for the remaining 13%, the minor allele frequency was between 20 and 50%.

Population genetics theory suggests that rare variants are more likely to be recently derived than are the common variants and are, therefore, more likely to be population-specific (*18*, *19*). Hence, rare variants are sensitive indicators of recent migration and the relationships among various populations (*20*). Our Hispanic-Latino sample shared a substantial number of rare variants with the African-American and Caucasian samples. By comparison, fewer rare variants were shared between the Caucasian and African-American samples, and even fewer were shared between the Asian and non-Asian samples (Table 2). This pattern of

sharing is consistent with the history of these populations in the United States and with the self-identification given when our samples were collected (*10*, *21*).

Not all population-specific alleles were observed at a low frequency. In the African-American and Asian samples, some population-specific alleles were found at frequencies >25%. Highly frequent population-specific alleles are particularly useful in mapping genes responsible for disease susceptibility and other traits in populations of mixed ancestry (*22–24*). On the other hand, SNPs, in which both alleles are present in all populations ("cosmopolitan" SNPs), are useful in conventional multigeneration linkage studies, as well as in genome-wide scans that use smaller family units (*25–27*). Of the 3899 SNPs, 21% were found to be cosmopolitan (Fig. 1A).

Nucleotide diversity provides a measure of genetic variation that is normalized by the number of chromosomes sampled. We calculated two conventional measures of nucleotide diversity for each gene: π, the average heterozygosity per site (*28*, *29*), and θ, the population mutation parameter (*13*, *30*). The average nucleotide diversity for the 292 autosomal genes ($\pi = 0.058\%$ and $\theta = 0.096\%$) and the 21 X-linked genes ($\pi = 0.028\%$ and $\theta = 0.045\%$) were within the range of values previously described (*2–4*). The fact that the average nucleotide diversity for the X-linked genes was reduced compared with the average autosomal nucleotide diversity is consistent with an equal number of males and females in the human population, in which males have only a single copy of the X chromosome.

We also calculated the autosomal nucleotide diversity values separately for each functional gene region and for each population. Exon-intron boundaries showed significantly higher average π values ($P < 0.01$ by single-factor ANOVA) than did the coding regions (0.088 and 0.034%, respectively), and the African-American sample had a significantly higher average π value than did the other population groups (0.068% and a range of 0.047 to 0.053%, respectively; $P << 0.0001$).

In our analyses of nucleotide diversity, π was consistently and significantly lower than θ ($P < 0.0001$). The difference between π and θ forms the basis of Tajima's D, a statistic that is used to detect departures from the standard neutral model (*9*, *31*, *32*). For an individual gene, a positive Tajima's D value is evidence for heterozygotes having a selective advantage, whereas a negative value is evidence for selection of one specific allele over alternate alleles. If, however, a majority of genes have a negative Tajima's D value, the simplest explanation for such results would be that the human population underwent a recent expansion. Previous studies of sixteen nuclear genes reported an even distribution between positive and negative Tajima's D values (*9*), results that offer no support for a recent population expansion. In contrast, of the 313 genes analyzed in our study, 281 showed a negative Tajima's D value. We interpret this result as strong evidence for a recent expansion of the human population.

We categorized the SNPs according to their 5′ to 3′ orientation on the sense strand of DNA (*17*). Seventy-one percent of the SNPs were transitions (35.8% G⇔A and 35.4% C⇔T), even though transitions represent only one third of the total possible types of mutation. Additionally, for the four categories of SNPs, G⇔A, C⇔T, C⇔A, and G⇔T, there was a pronounced bias observed. SNPs, in which the G or the C allele was more common than the alternate allele, were observed 1.9 to 2.4 times more frequently than were SNPs, in which the G or C allele was the less common allele. This bias was most prominent for rare (observed once or twice) and population-specific variants and was statistically significant ($P = 0.001$) for transitions. Population genetics theory predicts that the more frequent allele is usually the ancestral allele (*18*); hence, the bias observed suggests that the predominant direction of mutation was a change from G or C to A or T. For the two categories of SNPs, G⇔C and A⇔T, there was no apparent bias and, therefore, no predominant direction of mutation was apparent.

Methylation of CpG dinucleotides is thought to account for a large number of mutations, most of which would involve G⇔A or C⇔T transitions (*7*, *33*). Nearly 40% of the SNPs were consistent with mutation of either base in a CpG dinucleotide. Evolutionary pressure to relax the strength of base pairing would favor the conversion of G and C to A or T, which was the pattern of bias that was observed. On the other hand, the changes G⇔C and A⇔T, for which we observed no bias, are the only changes that do not alter the strength of base pairing.

Instead of using the frequency of an allele as a surrogate for ancestry, an alternative approach is to compare the alleles seen in humans with the corresponding sequence of a chimpanzee. With this approach, the human allele that matches the chimpanzee sequence is assumed to be the ancestral allele (*34*, *35*). There was general agreement between the two approaches of inferring ancestry for an allele, namely, the more common human allele generally matched the chimpanzee sequence (*36*).

We identified an average of approximately 12.5 biallelic SNPs per gene (*13*). If SNPs were randomly associated with each other within a gene, there would be about 2^{12} possible haplotypes. In fact, without recombination or recurrent mutation, the number of haplotypes should be less than the number of SNPs (*7*, *8*). In each gene studied, the combination of alleles present at each site of polymorphism in each individual was analyzed by a computer program (*37*) that assigned a specific pair of haplotypes to each individual, as well as a score reflecting the confidence in that assignment.

We observed an average of approximately 14 different haplotypes per gene, which was about 1.1 times the average number of individual SNPs identified per gene. Although the number of SNPs and haplotypes varied considerably among the genes studied, there was a linear relationship between the number of individual SNPs identified within a gene and the number of different haplotypes assigned per gene [$r^2 = 0.74$ (*38*)]. The fact that the number of haplotypes was greater than the number of SNPs is an indication that some level of recombination and recurrent mutation occurred within these genes (*7*, *8*).

The number of different haplotypes identified for a gene ranged from 2 to 53 in our sample of 313 genes (*13*). We estimated the heterozygosity at each gene by treating each haplotype as an individual allele (*39*). The haplotype heterozygosity of the 313 genes ranged from 0.012 to 0.929 in the pooled population sample and had an average of 0.534 (*13*). The average haplotype heterozygosity ranged from 0.437 in Asians to 0.584 in African-Americans. The maximum attainable heterozygosity for a single biallelic SNP is only 0.50. Of the 313 genes, 199 had a haplotype heterozygosity greater than 0.50. Thus, in general, the higher heterozygosity and multiallelic nature make haplotypes more informative than biallelic SNPs.

Sixteen percent of the 4304 haplotypes were cosmopolitan (Fig. 1B). If, however, the frequency of occurrence of each individual haplotype is considered, the cosmopolitan haplotypes accounted for nearly 82% of the total haplotypes

observed, whereas population-specific haplotypes accounted for only about 8% of the total (*40*). Nearly 4% of the total haplotypes were present in two populations, and almost 6% were present in three populations.

We determined the extent of allele sharing among individuals both within and among our four populations. The neighbor-joining algorithm (*41*) was used to cluster individuals on the basis of their haplotype pairs for each of the 313 genes. The results confirmed the integrity of the self-described ancestry of these individuals (*42*). The Asians and the African-Americans each formed separate clusters. Individual Hispanic-Latinos clustered with Caucasians or were connected to the base of either the Asian or African-American clusters. The lack of a defined cluster of Hispanic-Latinos is consistent with some Hispanic-Latinos being either primarily of European or Amerindian descent and others being combinations of European, Amerindian, and African descent. The extent of allele sharing between individuals in the same population was only slightly greater than that observed between individuals from different populations. This result presumably reflects Lewontin's (*43*) observation that the majority of human genetic variation occurs among individuals within a local population group with only a small additional variation occurring between individuals from different populations.

The population distribution of haplotypes was similar to the population distribution of SNPs, with many of the haplotypes being rare and population-specific (Fig. 1B). Of the 2782 population-specific haplotypes, a significant fraction (48%) was seen only in the African-American sample. The Asian sample had the second largest number of population-specific haplotypes, followed by the Caucasian and Hispanic-Latino samples. As with SNPs, the African-American sample had the largest number of distinct haplotypes, and the Asian sample had the smallest.

Our results indicate that many genes do not have one predominant haplotype. For 35% of the genes, no single haplotype had a frequency that was $\geq$50%. Therefore, the concept that there is one predominant or "wild-type" form of a gene and various rare or "mutant" forms is overly simplistic and misleading. Instead, there are multiple haplotypes, each of which is observed in multiple populations, that account for a large fraction of human genomic variability. This variety of different forms, or haplotypes, for most genes constitutes an opportunity for functional adaptation and diversification.

The large number of SNPs identified per gene facilitated the investigation of heritable associations (linkage disequilibrium) between SNPs within a gene. To estimate the relationship between linkage disequilibrium and physical distance, we calculated all the $|D'|$ values for pairs of SNPs with sufficiently high frequencies (*44*, *45*) in the four different population samples as a function of the distance separating them (Fig. 2). Of the 313 genes, 235 had a pair of SNPs whose minor allele was sufficiently frequent to estimate linkage disequilibrium in at least one population (*13*). There were pairs of SNPs that did not agree with the general concept that linkage disequilibrium decreases as a function of distance. For example, of the pairs that were separated by <1 kb, 6% had a $|D'| < 0.3$ and, hence, were only minimally associated (in linkage equilibrium). On the other hand, of the pairs that were separated by >20 kb, nearly 30% had a $|D'| = 1$ and, therefore, were maximally associated (in linkage disequilibrium). These results demonstrate that the probability of any particular pair of SNPs being in linkage disequilibrium is not predictable, and, as a result, linkage disequilibrium should be determined empirically for any specific genomic region.

We further analyzed the SNPs from those pairs that exhibited either extremely high or extremely low levels of association with each other. First, the SNPs from pairs for which the level of association was maximal were compared with the SNPs from pairs for which the level of association was minimal. There was no significant difference between the SNPs from the two groups regarding their distribution among either functional regions of a gene or the specific base pair defining the polymorphism. Second, we examined those pairs of SNPs that were exceptions to the general concept that the shorter the distance the greater the level of linkage disequilibrium. The SNPs from pairs, for which the level of association was maximal and which were separated from each other by >20 kb, were compared with the SNPs from pairs, for which the level of association was minimal and which were separated from each other by <1 kb. There was no significant difference between the SNPs from these two groups in the specific base pair defining the polymorphism. Likewise, there was no apparent difference in the proportion of SNPs found in the 5′ upstream, 5′UTR, coding, or 3′UTR regions. However, the proportion of SNPs from within exon-intron boundaries was significantly different ($P = 0.0121$) between these two categories (54 and 28%, respectively). These results suggest that SNPs found in exon-intron boundaries are the most likely SNPs to be in strong linkage disequilibrium at long distances.

Recently, Reich *et al.* determined the linkage disequilibrium relationship of 272 SNPs distributed over 19 genomic regions, each approximately 160 kb in size, in a population of 44 individuals of Northern European ancestry (*46*). In their study, D′ varied from no apparent association beyond 5 kb to maximal association at the longest distance measured. The extensive variability in linkage disequilibrium, which they observed for regions defined by a specific size, agrees with our results obtained for genomic regions defined by the transcriptional unit of a gene.

Our observations demonstrate the necessity of understanding patterns of human genomic evolution if genomic variability is to be used as a tool in human health research. The processes underlying genomic evolution are obviously subject to varying levels of natural selection, which invalidates overly simplistic theoretical models. Additionally, complex or unknown patterns of human migration complicate the distribution and interpretation of genomic variation. In particular, the pattern of linkage disequilibrium within genes is more complicated than was previously estimated. The fluidity of genomic parameters, such as linkage disequilibrium, questions the applicability of genome-wide studies that assume that SNPs, randomly distributed throughout the genome, will be sufficient to detect an association with a phenotype. Haplotypes, on the other hand, can correlate a specific phenotype with a specific gene in a small population sample even when individual SNPs cannot (*47*). Thus, attempts to draw associations between phenotypes and genomic variation are more likely to succeed when the SNPs used in such studies have been confirmed to be in linkage disequilibrium by methods such as haplotyping.

References and Notes

1. D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
2. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
3. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
4. W. H. Li, L. A. Sadler, *Genetics* **129**, 513 (1991).

5. A. Chakravarti, *Nature Genet.* **19**, 216 (1998).
6. R. M. Harding *et al.*, *Am. J. Hum. Genet.* **60**, 772 (1997).
7. S. M. Fullerton *et al.*, *Am. J. Hum. Genet.* **67**, 881 (2000).
8. A. G. Clark *et al.*, *Am. J. Hum. Genet.* **63**, 595 (1998).
9. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000).
10. Blood samples were obtained from approximately 200 individuals recruited in two U.S. locations (Miami, FL, and Anaheim, CA) and immortalized as cell lines by standard procedures. Data on family medical history and geographic origin of self, parents, and grandparents were obtained for each individual. From this collection, individuals were prioritized for sequencing according to the degree of homogeneous origin of their grandparents and to achieve an equal proportion of African-Americans, Asians, Caucasians, and Hispanic-Latinos. Additionally, the medical data were examined so as not to introduce a bias for any known diseases. A 96-well plate was constructed, containing DNA from 76 of these individuals, 10 individuals from a three-generation Caucasian family, 7 individuals from a two-generation African-American family and one chimpanzee, as well as positive and negative DNA controls. Immortalized cell lines derived from the two families were obtained commercially. The plate contained DNA from 82 unrelated individuals: 20 African-Americans, 20 Asians, 21 Caucasians, 18 Hispanic-Latinos, and 3 Native Americans.
11. Regions targeted for sequencing were amplified from genomic DNA isolated from the immortalized cell lines. Polymerase chain reaction (PCR) primer pairs were designed using the sequence and genomic organization in GenBank. The PCR products were purified using a Whatman Polyfiltronics 100 µl 384-well unifilter plate, essentially according to the manufacturer's protocol. The purified DNA was eluted in 50 µl of distilled water. Sequencing reactions were set up using Applied Biosystems Big Dye Terminator chemistry, essentially according to the manufacturer's protocol. The DNA primer used for the sequencing reaction was the M13 forward primer (5´-TGTAAAACGACGGCCAGT-3´) or the M13 reverse primer (5´-AGGAAACAGCTATGACCAT-3´). Reaction products were purified by isopropanol precipitation and analyzed on an ABI Prism 3700 DNA Analyzer. Sequences obtained were examined for the presence of polymorphisms by using the Polyphred program (*48*). The presence of a polymorphism was confirmed by sequencing both strands of DNA.
12. Each polymorphic site was compared with the public databases HGBASE (release 8 2000-11-1) and dbSNP (build 92 February 2001) to determine whether that site had been previously described.
13. A Web site containing the list of genes, base pair coverage, the number of SNPs, the number of haplotypes, and other characteristics described in this paper may be found at www.genaissance.com/genecharacteristics/genecharacteristics.pdf.
14. Splice sites were defined as the first 8 bases at the 5´ end of an intron and the last three bases at the 3´ end of an intron.
15. W. H. Li, C. I. Wu, C. C. Luo, *J. Mol. Evol.* **21**, 58 (1984).
16. To confirm the validity of SNPs that were observed only once as a single heterozygote, we resequenced the genomic regions from a subset of singleton SNPs. For 150 of 159 SNPs (94%), the presence of the polymorphism was confirmed by resequencing both strands of DNA.
17. J. C. Stephens *et al.*, unpublished data.
18. G. A. Watterson, H. A. Guess, *Theor. Popul. Biol.* **11**, 141 (1977).
19. The population-specific SNPs in our sample could occur in low frequency in one or more of the other populations but might not be detected because of the sample size. For instance, a SNP with a minor allele frequency of 0.0172 has only a 50% probability of being observed if the sample size is 20 people.
20. M. Slatkin, *Evolution* **39**, 53 (1985).
21. R. M. Cerda-Flores *et al.*, *Ann. Hum. Biol.* **19**, 347 (1992).
22. J. C. Stephens, D. Briscoe, S. J. O'Brien, *Am. J. Hum. Genet.* **55**, 809 (1994).
23. E. J. Parra *et al.*, *Am. J. Hum. Genet.* **63**, 1839 (1998).
24. M. D. Shriver *et al.*, *Am. J. Hum. Genet.* **60**, 957 (1997).
25. E. S. Lander, N. J. Schork, *Science* **265**, 2037 (1994).
26. N. Risch, K. Merikangas, *Science* **273**, 1516 (1996).
27. N. J. Risch, *Nature* **405**, 847 (2000).
28. W. H. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).
29. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
30. G. A. Watterson, *Theor. Popul. Biol.* **7**, 256 (1975).
31. F. Tajima, *Genetics* **123**, 585 (1989).
32. The standard neutral model (*9*) assumes that: a population is of a constant size; mating occurs randomly within the population; and mutations arise randomly at sites that are not already polymorphic and confer no selective advantage for any allele.
33. A. R. Templeton *et al.*, *Am. J. Hum. Genet.* **66**, 69 (2000).
34. A. G. Clark, *Nature Genet.* **22**, 119 (1999).
35. J. G. Hacia *et al.*, *Nature Genet.* **22**, 164 (1999).
36. J. A. Schneider *et al.*, unpublished data.
37. The haplotypes for each gene were constructed from the unphased genotypes using an extension of Clark's algorithm (*49*). We compared the haplotype assignments made for multiply heterozygous individuals in the two families. The agreement between our algorithm and those assignments made by Mendelian principles was at least 95%. In addition, we obtained similar results by comparing our algorithm with haplotype assignments made by molecular cloning.
38. T. Acharya *et al.*, unpublished data.
39. M. Nei, F. Tajima, *Genetics* **97**, 145 (1981).
40. The total number of haplotypes observed (50,471) was determined as follows: $(2 \times 82 \times 313) - (21 \times 41)$, in which the subtracted term is an adjustment for X-linked genes in males.
41. N. Saitou, M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
42. Z. Mu *et al.*, unpublished data.
43. R. C. Lewontin, *Evol. Biol.* **6**, 381 (1972).
44. Not all populations had variability at the same position. Therefore, not all SNPs could be compared in the four populations. In addition, only those SNPs, with a minor allele frequency of at least five observations, were used. With these restrictions, 7607 SNP pairs were analyzed.
45. R. C. Lewontin, *Genetics* **140**, 377 (1995).
46. D. E. Reich *et al.*, *Nature* **411**, 199 (2001).
47. C. M. Drysdale *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 10483 (2000).

48. D. A. Nickerson, V. O. Tobe, S. L. Taylor, *Nucleic Acids Res.* **25**, 2745 (1997).
49. A. G. Clark, *Mol. Biol. Evol.* **7**, 111 (1990).
50. R. Grantham, *Science* **185**, 862 (1974).
51. We thank C. F. Aquadro, M. E. Kreitman, R. Myers, and N. Risch for insightful comments and M. Athanasiou, R. R. Denton, P. R. DiBello, A. E. Duda, G. Faith, M. G. Greene, C. Harris-Kerr, A. Kaul, S. Kliem, Z.-H. Lan, E. Lanz, S. A. Leonard, L. M. Malnick, Z. Mu, D. Murallo, C. L. Nicholson, C. Wan, C. Wilcox, A. J. Wood, and W. Zhao for their contributions.

30 January 2001; accepted 7 June 2001
Published online 12 July 2001; 10.1126/science.1059431
Include this information when citing this paper.

Fig. 1. (**A**) The distribution of SNPs among the four population samples. The SNPs were categorized as to whether they were variable in one, two, three, or all four populations. (**B**) The distribution of haplotypes among the four population samples. The haplotypes were categorized as to whether they were observed in one, two, three, or all four populations. Population codes are AF, African-American; AS, Asian; CA, Caucasian; and HL, Hispanic-Latino.

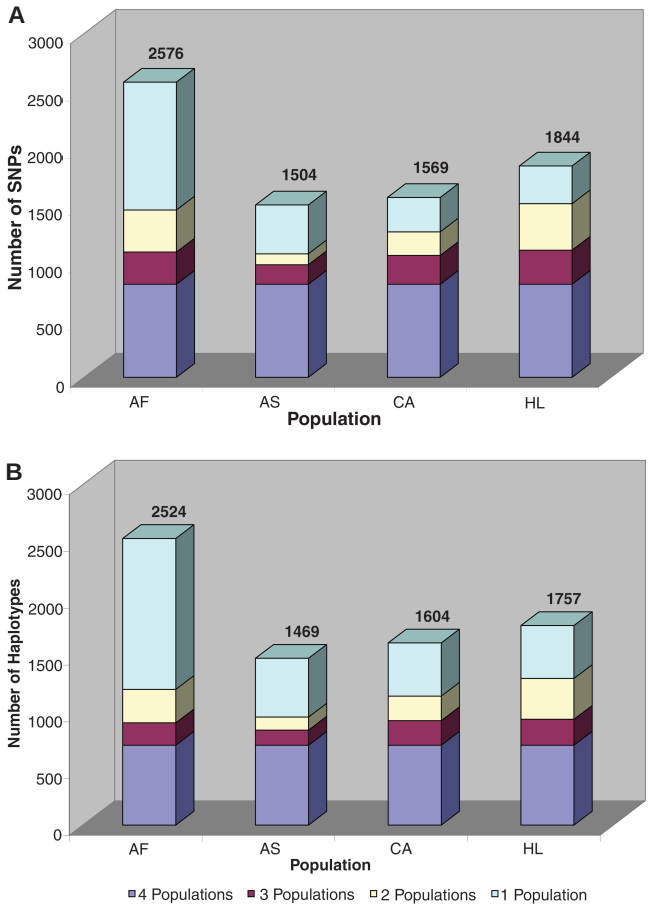
Fig. 2. Linkage disequilibrium for pairs of SNPs within a gene. Linkage disequilibrium (|D´|) was estimated separately for each population for each pair of SNPs within each of the genes. Only SNPs for which a rare variant was observed at least five times in a population were used in this calculation.

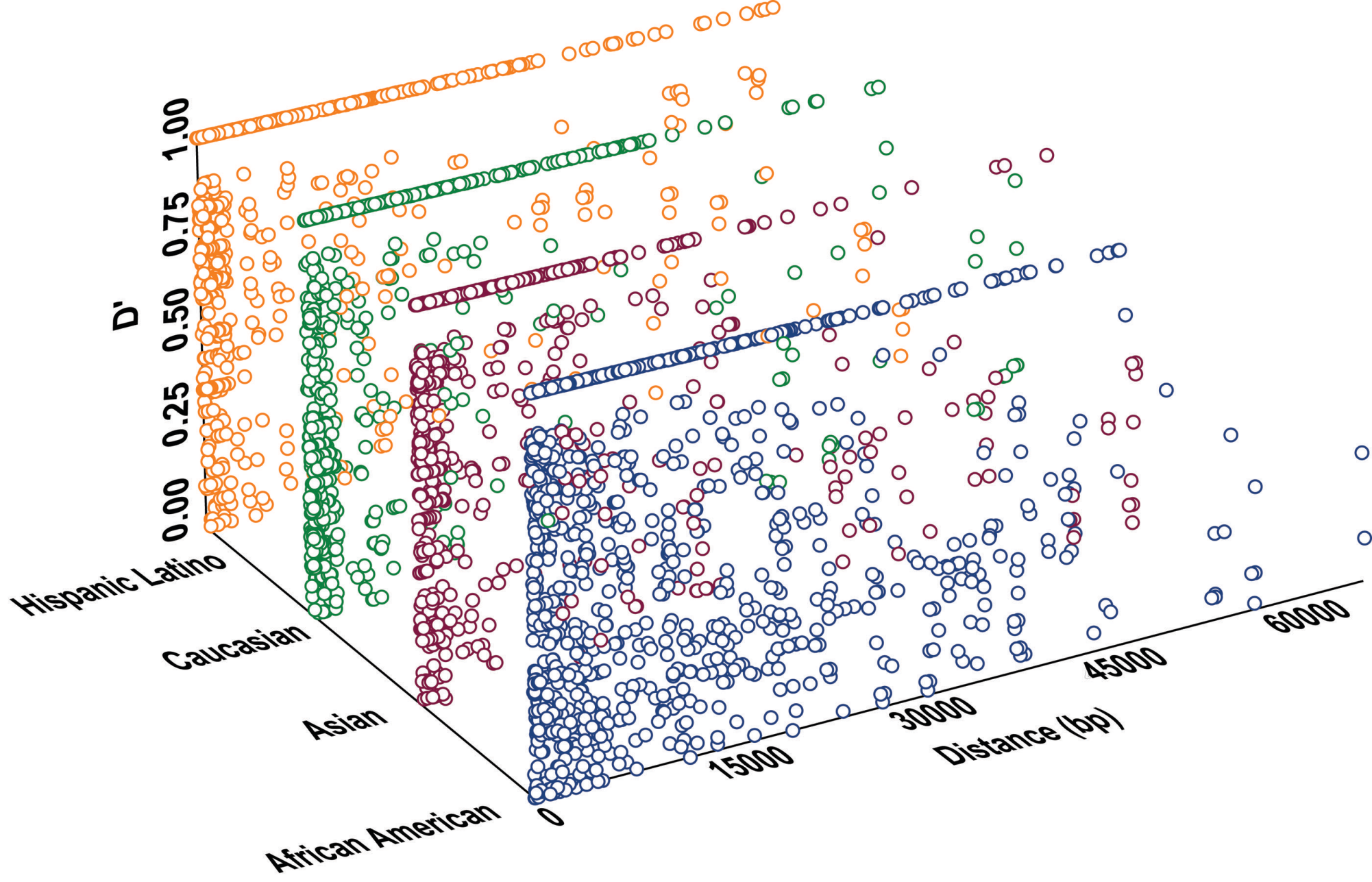
Table 1. The functional consequences of coding region polymorphisms. None means a silent nucleotide substitution. For amino acid changes, the type of change is categorized based upon Grantham values (*50*), which are derived from physiochemical considerations. The range that was used for Grantham values corresponds to that of Li *et al.* (*15*) and is as follows: conservative is <50; moderately conservative is between 51 and 100; moderately radical is between 101 and 150; and radical is ≥151.

Type of amino acid change	This study (%)	Mammalian pseudogenes (%)
None	44.4	27.9
Conservative	19.2	20.1
Moderately conservative	23.6	28.2
Moderately radical	8.2	12.9
Radical	3.7	6.6
To stop codon	0.9	4.4

Table 2. The distribution of rare SNPs among the four populations. SNPs with minor alleles observed two, three, or four times in the sample of 20 African-Americans (AF), 20 Asians (AS), 21 Caucasians (CA), and 18 Hispanic-Latinos (HL). Values on the diagonal are the number of population-specific SNPs. The other values are the number of SNPs shared between specific pairs of populations.

	AF	AS	CA	HL
AF	344	19	41	136
AS		115	7	18
CA			69	87
HL				40


A
3000
2500
2000
1500
1000
500
0
Number of SNPs
2576
1504
1569
1844
AF
AS
CA
HL
Population
B
3000
2500
2000
1500
1000
500
0
Number of Haplotypes
2524
1469
1604
1757
AF
AS
CA
HL
Population
4 Populations
3 Populations
2 Populations
1 Population


D'
1.00
0.75
0.50
0.25
0.00
Hispanic Latino
Caucasian
Asian
African American
0
15000
30000
45000
60000
Distance (bp)