Haplotype Structure and Population Genetic Inferences from Nucleotide-Sequence Variation in Human Lipoprotein Lipase

Andrew G. Clark,[1] Kenneth M. Weiss,[1,2] Deborah A. Nickerson,[3] Scott L. Taylor,[3]
Anne Buchanan,[2] Jari Stengård,[4] Veikko Salomaa,[4] Erkki Vartiainen,[4] Markus Perola,[5]
Eric Boerwinkle,[6] and Charles F. Sing[7]

[1]Institute of Molecular Evolutionary Genetics, Department of Biology, and [2]Department of Anthropology, Pennsylvania State University, University Park; [3]Department of Molecular Biotechnology, University of Washington, Seattle; Departments of [4]Epidemiology and Health Promotion and [5]Human Molecular Genetics, National Public Health Institute, Helsinki; [6]Human Genetics Center, University of Texas Health Science Center, Houston; and [7]Department of Human Genetics, University of Michigan Medical School, Ann Arbor

Summary

Allelic variation in 9.7 kb of genomic DNA sequence from the human lipoprotein lipase gene (*LPL*) was scored in 71 healthy individuals (142 chromosomes) from three populations: African Americans (24) from Jackson, MS; Finns (24) from North Karelia, Finland; and non-Hispanic Whites (23) from Rochester, MN. The sequences had a total of 88 variable sites, with a nucleotide diversity (site-specific heterozygosity) of .002 ± .001 across this 9.7-kb region. The frequency spectrum of nucleotide variation exhibited a slight excess of heterozygosity, but, in general, the data fit expectations of the infinite-sites model of mutation and genetic drift. Allele-specific PCR helped resolve linkage phases, and a total of 88 distinct haplotypes were identified. For 1,410 (64%) of the 2,211 site pairs, all four possible gametes were present in these haplotypes, reflecting a rich history of past recombination. Despite the strong evidence for recombination, extensive linkage disequilibrium was observed. The number of haplotypes generally is much greater than the number expected under the infinite-sites model, but there was sufficient multisite linkage disequilibrium to reveal two major clades, which appear to be very old. Variation in this region of *LPL* may depart from the variation expected under a simple, neutral model, owing to complex historical patterns of population founding, drift, selection, and recombination. These data suggest that the design and interpretation of disease-association studies may not be as straightforward as often is assumed.

Received March 20, 1998; accepted for publication June 11, 1998; electronically published July 29, 1998.

Address for correspondence and reprints: Dr. Andrew G. Clark, Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, PA 16802. E-mail: c92@psu.edu

0002-9297/98/6302-0036$02.00

Introduction

At the heart of the problem of the genetics of common, chronic diseases is a need to understand the quantity, nature, and arrangement of underlying DNA-sequence variation and to formulate and test models for the mapping of that variation into interindividual variation in risk of disease. History influences the way that we approach such difficult challenges, and, beginning with Garrod's (1902) classic work on alkaptonuria, the common approach essentially has been based on the view that for each gene there are only two alleles, normal and defective, segregating in the population. The theoretical basis for this view was the assumption that most mutations were deleterious and were eliminated quickly from the population and that mutations occurred recurrently between "wild-type" and harmful alleles. It of course is commonly appreciated that multiple alleles occur, but the ramifications of allelic complexity often seem to be underappreciated in contemporary biomedical genetic studies.

A starting point for studies explaining the role of genetic variation in the variation in risk of disease is a description of the quality, quantity, and organization of genetic variation within and between human populations. We are now in an era of the study of candidate genes that may influence traits that are measures of the biochemical or physiological features of disease. Most studies of candidate genes rely, at least in part, on ascertainment via clinical cases. This sampling approach introduces a bias toward identification of alleles that are associated with larger phenotypic effects. This is a sensible starting point and has proved to be remarkably effective in the identification of genes with particular alleles that explain Mendelian disorders, such as cystic fibrosis. Extensive family studies have demonstrated that, for each simply inherited, Mendelian disorder, hundreds of allelic variants in a particular gene may be associated with the disease phenotype. Less is known about the variation in the same genes in the unaffected (or not yet affected) subset of the population, and clinical ascertainment precludes an unbiased assessment of the frequency, penetrance, and expressivity of each allele in the population at large.

For the common chronic diseases having a complex multifactorial etiology that aggregate but do not segregate in accordance with Mendelian expectations, the problem is to explain the reason for the aggregation. Gene-mapping and physiological approaches have identified a number of candidate genes or chromosomal regions where candidate susceptibility genes may be found some day. When the physiology is understood, the population can be screened for background levels of variation in candidate genes, to find associations between the genetic variation and the quantitative variation in biochemical, physiological, and anatomic risk factors for disease. However, analysis based solely on clinical cases (the extremes of the phenotype distribution) may tell us little about the effects of genetic variation in the rest of the population.

Clearly, the most exhaustive way to assess DNA-sequence variation in the general population is to obtain full sequence and haplotype information from large numbers of individuals drawn randomly from different populations. Haplotypes specify the phase relationship of all heterozygous sites in a gene in a given individual and should be important to the understanding of the impact of genetic differences on variation in measures of health, because a functional polypeptide is the product of the haplotype (i.e., is coded by a single chromosome).

In this and a second article (Nickerson et al. 1998), we present analyses of a systematic survey of variation in a 9.7-kb region of the gene encoding the lipoprotein lipase (LPL) molecule in 71 individuals from three populations: (1) an African American population from Jackson, MS; (2) a population from the North Karelia region of Finland; and (3) a mixed European-derived cohort from Rochester, MN. The populations were chosen because they differ greatly in risk of coronary heart disease (CHD) and because extensive epidemiological studies of cardiovascular diseases are under way for all three populations. The purpose of our studies is to determine the role that variation in candidate genes, such as *LPL,* plays in causing variation in risk of CHD in the general population, rather than in the clinical population. Our goal in this article, however, is to understand the extent of the variation in this CHD-susceptibility gene at the population level; how that variation is structured into haplotypes and genotypes; and how they are distributed among individuals within and between populations.

Sequence variation in *LPL* is expected to be associated with some appropriate measure of physiological variation, such as the mean or variance of LPL protein levels or LPL kinetics among individuals. In general, regions coding for known functional sites in the LPL protein seem to be conserved with the other lipases (Murthy et al. 1996). The samples included in this study are too small to provide much exonic variation, and not all the *LPL* exons were included in the sequenced region. However, ~75 exonic mutations have been identified already (Santamarina-Fojo 1992; Stocks et al. 1992; Brunzell 1995; Yang et al. 1995; Murthy et al. 1996), and, at least in some instances, the exon mutations have clinical effects (Peacock et al. 1992; Reymer et al. 1995). As determined from similar data and as shown by Nickerson et al. (1998), variation in *LPL* differs among human populations (Chamberlain et at. 1989; Hata et al. 1990; Thorn et al. 1990; Gotoda et al. 1991; Normand et al. 1992; Hegele et al. 1994). For example, mutations have been described that appear to be specific to particular populations of African Americans (Heizmann et al. 1991) and French Canadians (Ma et al. 1991; Murthy et al. 1996).

In our other article (Nickerson et al. 1998), we describe the allelic variation in the *LPL* gene segment sequenced in these individuals. Here, we describe the haplotype structure of that variation and aspects of the evolutionary (population-historical) processes that produced that variation. For this purpose, we also sequenced chimpanzee *LPL,* so that it could serve as an outgroup for evolutionary and population genetic tests. We compare our results with those expected under the infinite-sites model, in which the evolution of DNA sequences is not affected by natural selection and each new mutation occurs at a previously unmutated nucleotide (Kimura 1969). To the extent that aspects of the observed sample fit the expectations of the infinite-sites model, other aspects of the data might be extrapolated to the general population, with some confidence. Given that there is already compelling evidence that human populations deviate substantially from the assumptions of the infinite-sites model (e.g., owing to major population expansions in our history) (Harpending et al. 1998; Jorde et al. 1998), our data raise caveats regarding generalities made about human genetic variation, especially with regard to understanding the distribution of risk of disease within and among populations.

Subjects and Methods

Population Samples

The populations sampled included (1) a Jackson, MS, population sample (n = 24) that is part of an ongoing National Heart Lung and Blood Institute study of hypertension in African Americans; (2) the Finnish risk-study population from North Karelia (n = 24), an area in eastern Finland that has had the world's highest known CHD risk and that has been studied for many years (Tunstall-Pedoe et al. 1994); and (3) the population from the Rochester Family Heart Study ($n = 23$), a longitudinal study of cardiovascular disease risk in the Rochester, MN, area. All three samples were cross-sectional with regard to their health status, and blood samples had been collected previously for epidemiological studies.

DNA Sequencing

DNA sequencing was performed on diploid genotypes, in accordance with procedures described in detail elsewhere (Nickerson et al. 1998). In brief, PCR-amplification products were used as templates in separate cycle-sequencing reactions to prepare fluorescent-tagged products for analysis on an ABI 373 automated sequencer. Each diploid sequencing trace was analyzed by use of procedures and software that can resolve heterozygotes, under replicable quality-control conditions, as described by Nickerson et al. (1997). Extensive confirmatory resequencing also was performed on these samples (Nickerson et al. 1998).

Chimpanzee LPL *Sequencing*

Sequence from the *LPL* gene also had been obtained from chimpanzee (*Pan troglodytes*). PCR was performed with the same array of primer pairs, and dye termination cycle sequencing was performed on the products. Not all genomic PCR reactions were successful, and not all insertion/deletion variations were resolvable; however, of the 9,734 sites scored in humans, the chimpanzee homologue was aligned and scored for 8,091 of these sites, including reference sites 1–1636, 1898–2916, 2938–3307, 3734–4124, 4289–4675, 5222–5849, 6029–6571, and 6618–9734. The GenBank accession numbers for *LPL* reference sequences are AF050163 for human and AF071087–AF071094 for chimpanzee.

Haplotype Determination

The samples were sequenced on diploid DNA, which yields unphased genotype data, and then the haplotype phases were determined. Traditionally, pedigree analysis has been used to infer linkage phase. Haplotypes also can be determined directly, for each pair of sites, by allele-specific PCR (AS-PCR). An important question is whether it is necessary to test in this manner every pair of sites in multiple-site heterozygotes or whether the data have enough structure to allow some haplotypes to be inferred. In this study, we employed an iterative approach that determines some phases by AS-PCR, then calls phases by an inferential procedure, and then tests the calls by further AS-PCR until no calls are violated by the empirical tests.

Complete homozygosity across a region or heterozygosity at a single site allows unambiguous assignment of a haplotype for that region. Our algorithm identifies a set of putative haplotypes in a sample by “subtracting” such unambiguous haplotypes in individuals having more than one heterozygous site (Clark 1990). The rationale for this approach is that, for any haplotype that is common enough that homozygotes can be found in the sample, the sample is expected to have several heterozygotes bearing one copy of that haplotype. The two types of potential errors are failure to resolve all haplotypes, which occurs if there is an individual heterozygous for two unique haplotypes in the sample, and misclassification. Both errors are corrected empirically by determination of the linkage phase by AS-PCR. When used alone, the accuracy of the inferential procedure depends on several population genetic parameters, including the population size, the sample size, the mutation rate, and the rate of intragenic recombination, but this inferential procedure can be surprisingly effective at inferring the haplotype structure hidden in a set of unphased genotype data (Clark 1990). In conjunction with AS-PCR, the method can be iterated until accuracy is assured. Maximum-likelihood methods for haplotype inference (Excoffier and Slatkin 1995; Hawley and Kidd 1995) do not work with data of this type, because they build the likelihood based on the fit to Hardy-Weinberg proportions. Such a likelihood has no information in the context of data such as those for our *LPL* sequences, because every sampled individual is a heterozygote.

The *LPL* sequences required numerous AS-PCR tests for resolution and verification of the haplotypes. For several different subsets of AS-PCR phase calls, the information from those phase calls was used to seed the inferential algorithm. Each run gave inferred haplotypes for different but overlapping sets of sites. Confidence in the inferred haplotypes came from two observations. First, a sufficient number of molecular calls had been made such that the multiple runs of the algorithm (using different subsets of sites) gave identical calls. Second, we were able to verify the calls by performing runs seeding the algorithm with a subset of AS-PCR calls and by comparing the resulting inferred haplotypes versus those that already had been tested molecularly. For example, the AS-PCR calls for five sites were omitted from the input, and the algorithm was run. These five sites occurred in 22 five-site heterozygotes, 1 four-site heterozygote, 2 three-site heterozygotes, and 3 two-site heterozygotes, for a total of 235 site pairs for which the linkage phase needed to be called. The algorithm produced calls for these site pairs, and subsequent checking with the molecular calls showed that all 235 site pairs were correct. Finally, several inferred haplotypes were verified post hoc by targeted AS-PCR for particular pairs of sites.

This method of haplotype inference has two weaknesses. First, a lot of effort is required for the AS-PCR: for the *LPL* sample, >560 AS-PCR reactions were performed. Second, the only way to call the phase for singleton and doubleton sites is to perform direct molecular tests, such as AS-PCR. Because singleton and doubleton sites are at such low frequency in the population, they yield very little information about the gene's history, and we have not yet determined their linkage phase.

Data Validation

The procedure of calling heterozygous sites in the sequencing of chromatograms is not applied widely, and the PolyPhred program plays a key role in optimizing this procedure. As outlined by Nickerson et al. (1998), several lines of evidence were used to assure accuracy of the sequences, including scoring of both strands, replication of a large fraction of the runs, and scoring of all heterozygous sites by AS-PCR for the haplotype calls. In addition, the structure of the data is consistent with only an exceptionally low error rate. We cite eight reasons for this: (1) Most of the coding variants identified in this study had been seen by others, as was expected for such small samples. (2) The coding region had lower nucleotide diversity than the noncoding regions, as was expected. (3) For subsamples of the data, the number of segregating sites increased with an increase in sample size, in close approximation to the infinite-sites model. (4) Overall, the nucleotide diversity was within the range observed by others. (5) Individual sites that were analyzed singly revealed a good fit with Hardy-Weinberg proportions of genotypes. Although the error rate would have to be inordinately high to cause a rejection based on any one site, the lack of a trend in departures from Hardy-Weinberg proportions, across sites, suggests that the aggregate test has reasonable power. (6) The chimpanzee sequence—which was obtained at an independent laboratory, by K.M.W. and A.B.—showed that every polymorphic site in humans had one or the other nucleotide present in chimpanzee. (7) If errors were frequent, they most likely would occur as singleton calls (false calls of heterozygotes). This would result in an excess of singletons, and the observed count of singletons was slightly below that expected under the infinite-sites model. (8) False calls of heterozygotes also would be expected to generate sites with three or four different segregating nucleotides, which were not observed. Finally, in work that is in progress, K.M.W. and A.B. independently have sequenced a comparable sample of Amerindians and have identified variations at >50 of the same sites, with all other sites either being invariant for one of the nucleotides observed in the original data set or showing a new variant in addition to the observed nucleotide.

Results

The Amount of Genotypic Variation

A detailed description of the amount, quality, and distribution of allelic variation in the *LPL* gene segment sequenced in the sample of 71 individuals from three populations appears in the article by Nickerson et al. (1998). In brief, 88 variable sites in the 9,734 bp of *LPL* were observed. With the exception of a polymorphic tetranucleotide repeat, all segregating positions, including the nine insertion/deletion differences, had only two alternative nucleotides. Ten sites in our sample of 142 chromosomes had only a single copy of the rarer nucleotide (singletons), and another 10 sites had only two copies of the rarer nucleotide (doubletons). Overall, the nucleotide diversity (i.e., the average heterozygosity, including unvarying sites) for the 9,734-bp region was .002 ± .001, which is within the range of previous estimates but is somewhat higher than average estimates (Li and Sadler 1991). Of the variable sites, 9 were insertions or deletions, and 79 were single-nucleotide substitutions. Only 7 of the single-nucleotide polymorphisms (SNPs) were in exons (sites 145, 2849, 6176, 6196, 6203, 7754, and 9040). Three mutations changed an amino acid (2849, Asn→Ser [Gagne and Gaudet 1995]; 6176, Val→Met; and 6203, Thr→Ala), and one mutation (9040) introduced a premature termination codon in the mRNA (Hata et al. 1990). The coding variants at positions 9040 and 2849 had been reported previously (by Reina et al. [1992] and Wiebusch et al. [1996], respectively), whereas those at 6176, 6203, and 7754 had not. We did not observe any individuals who are completely homozygous for this 9.7-kb region, and only two individuals varied at only a single site, whereas two individuals varied at 38 sites. The average individual is heterozygous at 17 sites.

Two estimates of nucleotide variation commonly are used (Hartl and Clark 1997). A central parameter in population genetic models for the balance between mutation and random genetic drift is $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide site. This parameter summarizes the rate at which processes of mutation and random genetic drift generate and maintain variation within a gene, under conditions in which natural selection has not been operating. The degree to which characteristics of the data conform to the predictions of the infinite-sites model has been used as a test of goodness of fit of the model. Such a test begins with estimation of the relevant model parameters, the most important being θ. From the number of segregating sites, S, θ has been estimated as $\hat{\theta} = S/\sum_{i=1}^{n-1}\frac{1}{i}$, where n is the sample size (Watterson 1975). Although the count of segregating sites does not capture all the information in the sample, under the infinite-sites

Table 1

Population Genetic Statistics for *LPL* Variation

Population	No. of Sites	θ	π	Tajima's D^*	C/θ
Jackson	78	.0018 ± .0006	.0020 ± .0010	.812	1.443
North Karelia	56	.0013 ± .0004	.0016 ± .0008	1.007	.371
Rochester	59	.0014 ± .0004	.0020 ± .0010	1.775	.335
Total	88	.0016 ± .0004	.0020 ± .0010	.909	.693

model the frequency spectrum of sites is determined by θ, which in turn is determined by S. Note that violation of the assumptions of the infinite-sites model will lead to bias in the estimate of θ. In particular, if multiple mutations occur at individual nucleotide sites, use of estimates derived from the finite-sites model would be more appropriate (Tajima 1996).

A second measure of sequence variability is π, the average heterozygosity per site, where heterozygosity is the probability that a site will be heterozygous in an individual sampled from a randomly mating population. If the population sample fits the infinite-sites model, θ and π have equal expectation. Table 1 gives the estimates of θ and π for each of the three populations and for the pooled sample. In each case, $\hat{\pi}$ is greater, suggesting that, for the given number of varying sites, there is an excess of heterozygosity. Comparison of these two measures was formalized, by Tajima (1989), in the test statistic D, which is $\hat{\pi} - \hat{\theta}$ divided by the SD of this difference. Despite the consistency of excess heterozygosity, we cannot reject the null hypothesis of neutrality.

Under the infinite-sites model, the expected number of singletons is simply θ (Fu and Li 1993). On the basis of the observed number of segregating sites, our estimate of θ for the entire 9.7-kb region is 15.94 (the estimates in table 1 were calculated on a per-site basis). We observed 10 singletons instead of 16 and applied the appropriate test for the significance of this departure (Fu and Li 1993). The test statistic for this aspect of the data is D^*, for which the expectation is 0 and the SD is 1.0. For our data, $D^* = 0.757$, which is not significant ($P > .05$). Under this model, the expected number of heterozygous sites per individual is also equal to $\theta = 15.94$, which is close to the observed count of 17. To the extent that there is similarity between observed and expected values for these characteristics, one could infer that the populations are evolving in accordance with the infinite-sites model.

The upper panel of figure 1 provides the distribution of relative nucleotide frequency, in order of descending rank of the 87 variable sites (the tetranucleotide repeat violates the assumptions of the infinite-sites model and was not considered). Also shown are the expected frequencies of alleles at the 87 sites, under the infinite-sites model. These were generated by sampling from the distribution given by $P(j$ copies in a sample of n genes$) = 1/ja$, where $a = \sum_{i=1}^{n-1} \frac{1}{i}$ (eq. 9.69 in Ewens 1979). Consistent with the Tajima (1989) and Fu and Li (1993) tests, the site-frequency distribution is in good accordance with the infinite-sites model, although there is some excess heterozygosity at sites of intermediate rank. The good fit to the infinite-sites model is consistent with the very small effect of a correction under the finite-sites model. Even if we allow for extensive variation, among sites, in the mutation rate (following a gamma distribution with parameter $\alpha = .1$), the estimate of θ, on a per-site basis, is changed from 0.002 (infinite-sites model) to 0.00206 (finite-sites model) (eq. 24 in Tajima 1996).

Population Subdivision

Variation in the gene was distributed unevenly among the three populations. Fifty-one sites varied in all three populations. However, 5 sites were found to vary in only the Rochester population, 2 sites in only the North Karelia population, 3 sites in only the two European-derived populations, and 27 sites in only the Jackson population. One way to quantify the degree of population diversity is with the statistic F_{ST} (Wright 1931), which essentially is the genetic variance (heterozygosity) among populations, divided by the genetic variance of the total population. The nucleotide-site data yielded an estimate of $F_{ST} = .065 \pm .014$, with a 95% confidence interval of (.041–.0922), obtained by bootstrapping over sites (Weir 1996). The data allow strong rejection of the null hypothesis of homogeneity, by permutation tests (Hudson et al. 1992). The level of population subdivision is typical of findings in studies of other genes (Cavalli-Sforza et al. 1994).

The F_{ST} analysis is based on individual nucleotide-site data. As the size of the examined region increases, more and more haplotypes become unique to each population, and eventually every haplotype will be unique to each individual. This kind of scaling problem is important to keep in mind when haplotypic variation over regions as large as or larger than that examined here is considered. As described in Subjects and Methods, we obtained linkage-phase information from AS-PCR tests and by inference, to obtain 88 distinct haplotypes in the sample.

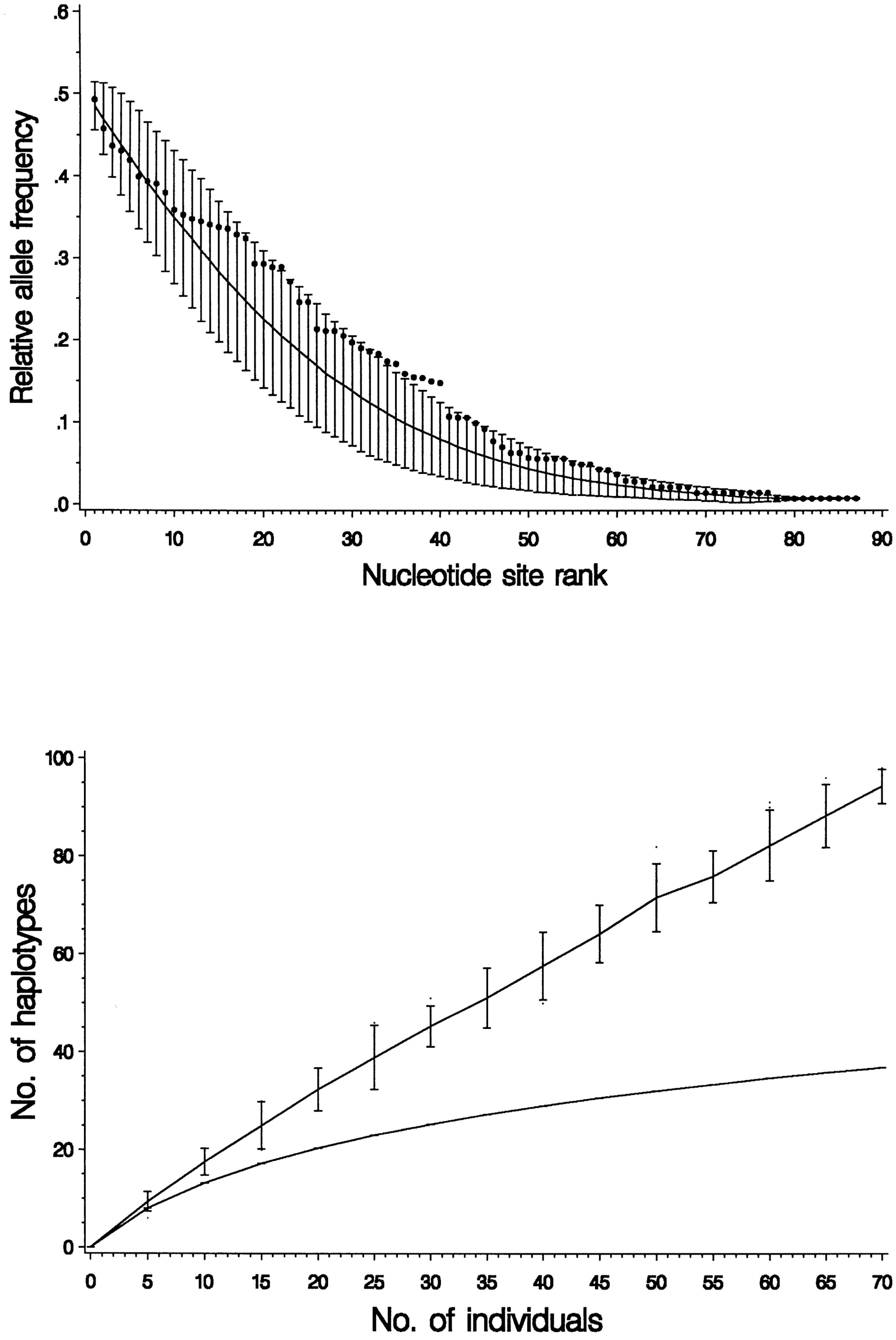

Figure 1 *Top,* Observed relative frequencies of the rarer nucleotide at each varying site (indicated by dots). In general, these frequencies are close to the frequencies expected under the infinite-sites model (indicated by the solid line), for the given sample size and observed number of segregating sites. Error bars are ±1 standard error and were obtained by repeated sampling from the expected distribution given in the text. *Bottom,* Distribution of observed haplotypes when subsets of the data were resampled. The number of distinct haplotypes (indicated by the solid line, with error bars of ±1 standard error) are far in excess of the number expected under the infinite-alleles model (lower solid line), for the given sample size and the estimate $\theta = 4N_e\mu = 15.94$.

Only 3 of the 88 haplotypes were present in all three populations, whereas 1 haplotype was present in the Jackson and North Karelia populations and 3 were present in the North Karelia and Rochester populations. Fully 21 haplotypes were unique to the North Karelia population, 25 haplotypes were unique to the Rochester population, and 35 haplotypes were unique to the Jackson population. When the data are treated as a single locus with 88 alleles, F_{ST} for the *LPL* haplotype data is .0197 ± .0133. This is lower than the nucleotide-site estimate of F_{ST}, because the probability of haplotype identity remains very low whether alleles within the same population are examined or alleles from two different populations are compared.

Haplotype Variation and Intragenic Recombination

Whereas the infinite-sites model yields an expected number of segregating sites for a given sample size, population size, and mutation rate, the infinite-alleles model provides an expected number of distinct alleles (in this case, haplotypes) for these same parameters. Both models serve as a null model in which the only forces acting on the sequence variation are mutation and random genetic drift. The difference lies in the manner in which mutations are thought to occur. In the infinite-alleles model, each mutation is assumed to generate a distinct allele, and no recombination is assumed to occur. In applying this model, we assumed initially that the population is not subdivided and is at equilibrium. The lower panel of figure 1 shows the correspondence between the infinite-alleles model and the mean and variance of the numbers of distinct haplotypes observed when we repeatedly drew samples of various sizes from the observed data. The data show many more haplotypes than the model predicts for the value of θ estimated from the number of segregating sites. One possible cause for the excess is intragenic recombination. Gene conversion is formally distinct from intragenic recombination, but, for purposes of discussion, we refer to the consequences of both processes. We can begin to assess the likelihood that intragenic recombination has played a role in the generation of haplotype diversity by directly examining the variation. Figure 2 shows the 88 haplotypes by plotting a dash (–) for the more common variant and a circle (○) for the rarer variant at each variable site. The haplotypes are arranged in the order in which they appear in a tree of sequence similarities (by use of the neighbor-joining method; Saitou and Nei 1987), which clusters together the haplotypes that are most similar. One feature of the haplotypes is clear: there is a cluster of sites, in the 3′ half of the sequenced region of *LPL,* that partitions the haplotypes into two major clades. Some haplotypes, however, appear to have patterns that have been interchanged in the 5′ and 3′ ends. Such patterns reflect a history of intragenic recombination, which we consider below.

With 88 segregating sites, there is an astronomical number of potential haplotypes, but only a small fraction of these haplotypes would be observed in any finite population. In the absence of intragenic recombination, repeated mutation, or back mutation, the maximum number of haplotypes that one expects to see for *s* sites is $s + 1$ haplotypes. A simple graph indicating intragenic recombination is a sweeping window plot showing an arbitrary window size of five sites at a time and counting the number of distinct haplotypes for each set of five sites (fig. 3). Many sets of five adjacent sites in the *LPL* gene exhibit more than six haplotypes, which is indicative of intragenic recombination. When 10 sites are considered at one time, the same regions show >11 haplotypes. Both window sizes show a marked decrease in the number of distinct haplotypes in a region in the 3′ half of the gene, suggesting less intragenic recombination in this region.

Several statistical tests for intragenic recombination have been devised. The tests of Sawyer (1989) and Stephens (1985) appear to be best suited to identification of rare recombination or conversion events (also see Betrán et al. 1997; Jakobsen et al. 1997). Both tests clearly show that there is a large amount of variation, but they yield little useful quantitative information, because there appears to be so much recombination. The test of Hudson and Kaplan (1985) is based on the four-gamete test, which is a two-site analogue to figure 3. A population that has only one haplotype—for example, *A-B*—may mutate to give *A-b* and *a-B*. Haplotype *a-b* can arise only through recombination or repeated mutation. We will assume that repeated mutation is a less likely explanation and that all site pairs having four gametes must have had a recombination occur between them at some time in their ancestral history. Repeated mutation will be considered further in another article, but the initial assumption that repeated mutation is rare stems from the good fit with the infinite-sites model and the lack of even a single observed site with more than two segregating nucleotides. When these caveats are considered, simply tallying the fraction of site pairs that have all four gametes present will provide an indication of intragenic recombination. Note that a single recombination event can result in multiple site pairs having all four gametes in the sample; thus, the inference of numbers and locations of recombination events requires consideration of the locations of the site pairs that have four gametes present.

Figure 4 shows the site pairs for which four gametes were present in the sample of 142 chromosomes. The distribution of such pairs shows that intragenic recombination is likely to have occurred throughout the *LPL* gene. Hudson and Kaplan (1985) devised an estimate

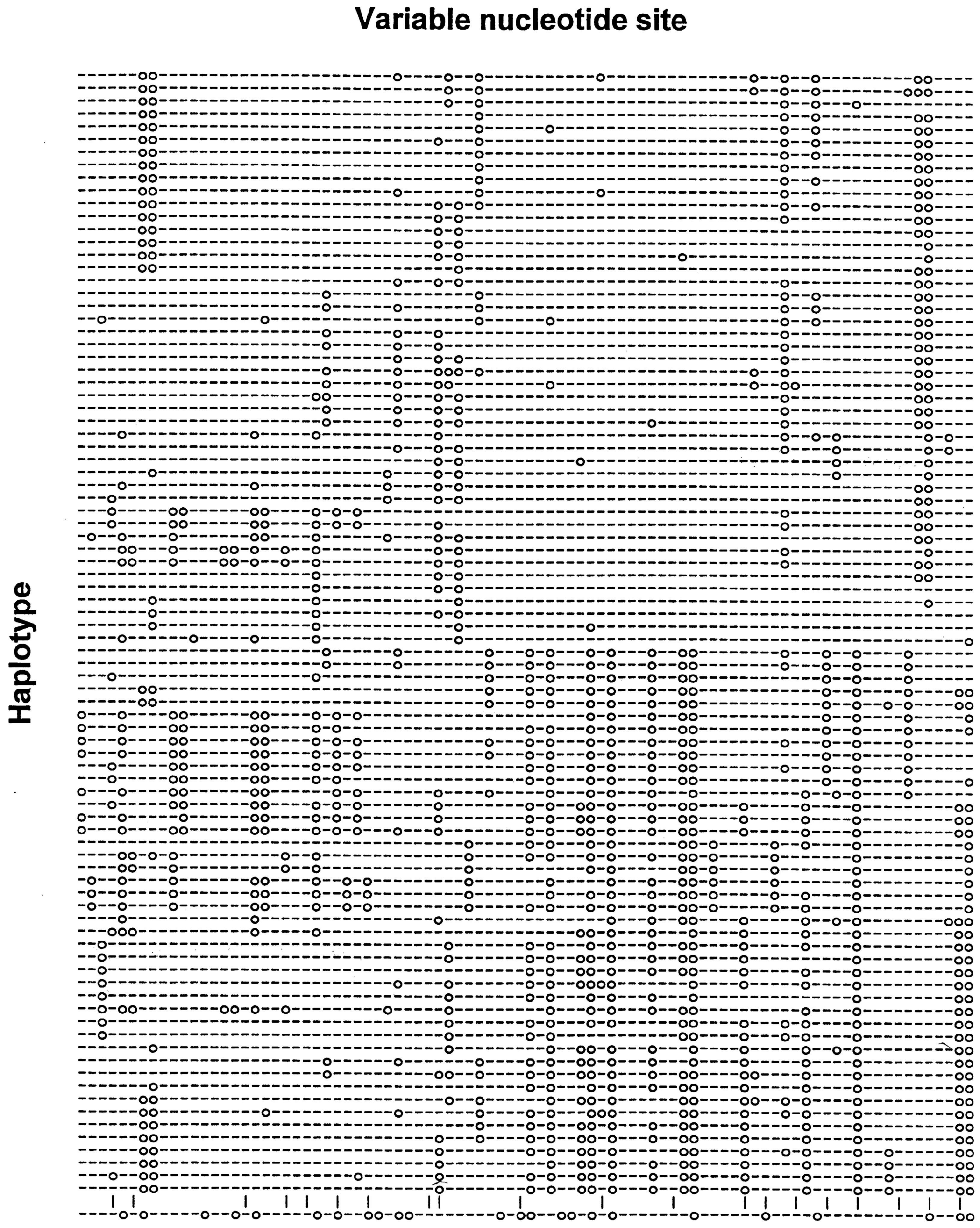

Figure 2 Illustration of the 88 haplotypes. For each site, a dash (–) represents the common nucleotide, and a circle (○) represents the rarer nucleotide. The haplotypes are arranged in order of the tips of a neighbor-joining tree, so that more-similar haplotypes are clustered. Note that the haplotypes fall into two major groups, divided roughly in the middle of the list, and that some intragenic recombination events can be identified by sight. The penultimate line indicates the positions of the inferred minimum recombination events. The last line indicates the chimpanzee haplotype.

for the minimum number of recombination events that is consistent with the matrix of four-gamete results. For our data, the estimate of the expected number of recombination events is 45. Our estimate of the minimum number is 20. Although this method yields a uniform distribution of recombination events along the gene (fig. 2, *bottom*), the methods of Jakobsen et al. (1997) and Crandall and Templeton (1998) identify a hot spot in the middle of the sequenced region. In applying the four-gamete tests to the Jackson, North Karelia, and Rochester samples separately, we found evidence for a minimum of 17, 14, and 13 past recombination events,

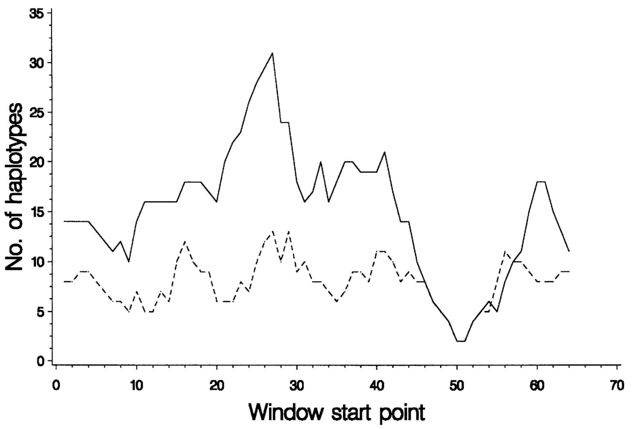

Figure 3 Observed number of distinct haplotypes, in sliding windows of 5 bp (dashed line) and 10 bp (solid line). For a window of size *s,* the maximum number of haplotypes in a population, in the absence of intragenic recombination, back mutation, or recurrent mutation, is $s + 1$. Observation of regions with a number of haplotypes greater than $s + 1$ indicates numerous recombination events in the ancestral history of that segment of the gene.

respectively. Because the Jackson sample had more segregating sites and likely is more admixed than the other populations, there is greater power to discriminate recombination events in the Jackson sample; thus, these data do not suggest variation among populations in levels of recombination.

To determine whether the method by which haplotypes were inferred might be responsible for this plethora of four-gamete site pairs, we evaluated the unphased (raw) genotype data. When the unphased data were considered by two sites at a time, the double homozygotes and single heterozygotes each provided unambiguous information about haplotypes. If we did not have linkage-phase information, double heterozygotes did not allow direct counts of haplotypes; therefore, they were not used in this analysis. Thus, the unphased data provide a minimal bound on the true number of site pairs with all four gametes. Of the 2,211 site pairs that could be tallied, 1,187 had all four gametes present. When we employ the complete phase-known data, some of the site pairs that had only three unambiguous haplotypes now will have all four haplotypes present. The phase-known data added only 223 four-gamete site pairs (for a total of 1,410). This result also indicates that, in general, the recombination events that produced these gametes occurred far enough in the past for the gametes to become frequent enough to be present in unambiguous genotype combinations even in a small sample.

Intragenic recombination results in a reduction in the variance of the pairwise mismatch distribution. Hudson (1987) made use of this fact to devise an estimate of the ratio of C/θ, where $C = 4N_ec$ is a compound measure of effective population size multiplied by the recombination rate between flanking sites, *c,* and where $\theta = 4N_e\mu$, as before. For the *LPL* data, application of Hudson's estimate gives $\hat{C}/\hat{\theta} = 0.693$ for the entire sample, and, for the Jackson, North Karelia, and Rochester samples, the estimates were 1.443, 0.371, and 0.335, respectively (table 1). These numbers suggest that the rate of intragenic recombination between a pair of sites is within a factor of two of the mutation rate (again with the caveat that repeated mutations are assumed not to have been important in producing this variation). It again appears that the Jackson population has a higher recombination rate than the other populations, probably owing to admixture, but the statistical significance of this difference was not determined. Recent refinement of this estimation procedure gave essentially the same results (Hey and Wakeley 1997; Wakeley 1997). High levels of intragenic recombination in human genes had been inferred in previous studies (Chakravarti et al. 1984, 1986).

Linkage Disequilibrium

Despite the apparently high rate of intragenic recombination, considerable linkage disequilibrium still exists among pairs of sites from these *LPL* data. Figure 5 shows results of pairwise tests, indicating those site pairs that exhibited significant disequilibrium by a Fisher's

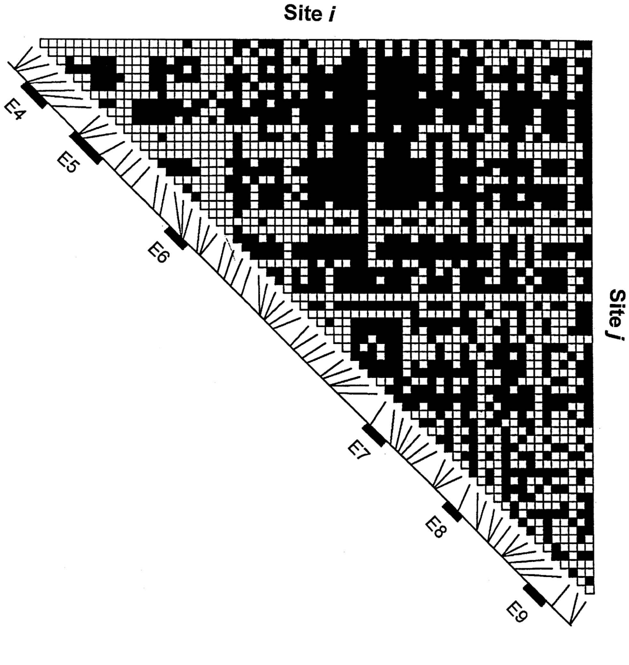

Figure 4 Plot for the four-gamete test. A blackened square indicates that the given site pair had all four possible gametic phases. The presence of all four gametes in site pairs also suggests intragenic recombination. The diagonal line, with exons 4–9 labeled, indicates the location of each varying site along the gene.

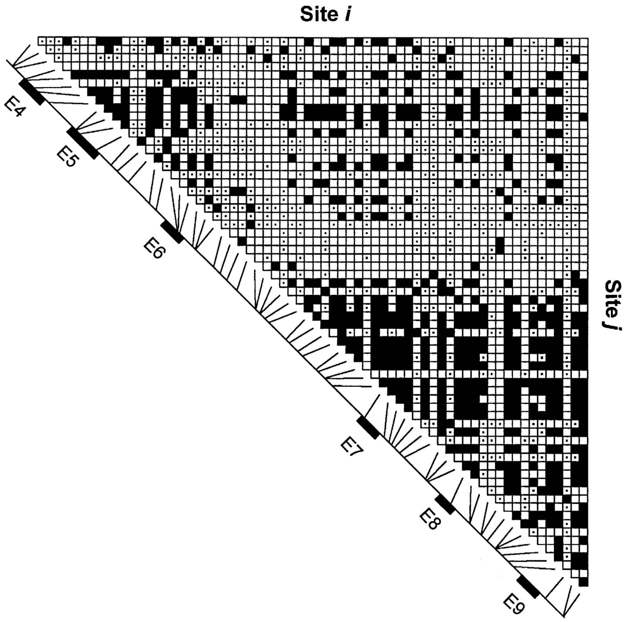

Figure 5 Plot showing pairwise linkage disequilibrium, indicated by a blackened square, for site pairs with a significant Fisher's exact test ($P < .001$) and no correction for multiple comparisons. In comparisons of site pairs in which both sites have rare nucleotides, there can be complete disequilibrium (one of the four possible gametes having a count of 0), yet the Fisher's exact test can be not significant. Site pairs that lack the power to test a significant association are indicated by a dot in the center of the square. The layout of the figure is the same as that for figure 4.

exact test (Weir 1996). Figure 5 shows a tendency for more cases of significant disequilibrium in the 3′ half of the gene, corresponding to the somewhat lower apparent recombination rate in this region, indicated by figure 3. For each pair of sites, we calculated the linkage-disequilibrium statistic *D*, defined as $P_{ij} - p_i p_j$, where P_{ij} is the frequency of the haplotype with nucleotides *i* and *j* at the two sites and where p_i and p_j are the frequencies of nucleotides *i* and *j*, respectively. A test of overall linkage disequilibrium, based on the numbers of observed and expected signs of *D* (Lewontin 1995), revealed a significant excess of cases of disequilibria in which rare alleles were associated.

The linkage disequilibria calculated for all pairs of sites were correlated significantly between the samples (Jackson vs. North Karelia, $r = .792$; Jackson vs. Rochester, $r = .592$; and North Karelia vs. Rochester, $r = .673$; $P < .0001$ for all correlations), and there were no cases of significant disequilibrium in opposite directions, from one sample to another. This suggests that the basic pattern of disequilibrium that will be detected in samples of this size reflects events that occurred long ago in the ancestral population common to all of the current samples. The replicability of this pattern may be a useful tool for mapping strategies. Of the 2,211 site pairs for which linkage disequilibrium could be assessed, 795 were in complete disequilibrium (having only three of the four gametes present), and six site pairs were in absolute disequilibrium (having only two of the four gametes present). In addition, there was a highly significant higher-order linkage disequilibrium, as assessed by Brown et al.'s (1980) test. In this test, the observed variance of the pairwise mismatch distribution is compared with the variance for randomly permuted data (the null hypothesis of linkage equilibrium). The observed variance (125.72) was greater than the variance of 1,000 permuted trials (for which the mean variance was 18.22); thus, the evidence for multisite disequilibrium is strong.

Pairwise disequilibrium also was inferred from the unphased genotype data (again, in an effort to see if any attribute of the pattern of disequilibrium might be an artifact of the haplotype-inference procedure). Following the method of Hill (1974), we assumed that genotypes are formed by random union of gametes and obtained maximum-likelihood estimates of gametic frequencies and of Δ, the composite linkage-disequilibrium parameter from the genotype counts. The values of Δ and *D* were highly correlated ($r = .983$; $P < .0001$).

Similarity Relationships among LPL *Haplotypes*

Gene trees typically are drawn to portray the sequence similarity of a sample of haplotypes. Without recombination or recurrent mutation, a series of such sequences will form a cladistic, or hierarchical, relationship that reflects the ancestral historical order of the mutations that have occurred. Intragenic recombination, as reflected in the *LPL*-gene data, makes it impossible to do this, because, when gene segments are exchanged among alleles, the sequences no longer have a branchlike cladistic structure (the tree has loops as well as branches). The full network of all 88 *LPL* haplotypes is a complicated tangle of loops and is not shown.

One feature of the relationships among haplotypes that is clear without drawing a tree is the tendency of the haplotypes to cluster into two major clades. This pattern is evident by inspection of the raw data presented in figure 2. The distribution of pairwise mismatches between alleles has been examined in several genes, for the purpose of drawing inferences about past demographic history (Slatkin and Hudson 1991; Rogers and Harpending 1992). Figure 6 shows that the mismatch distribution for *LPL* is clearly bimodal, reflecting the two deep clades apparent in figure 2. In the absence of intragenic recombination, the expected mismatch distribution under the infinite-sites model is geometric (Watterson 1975), but the sampling properties of the model are such that wide differences from this expectation are not uncommon.

In order to test the significance of the departure of the observed mismatch distribution from the neutral expectation, we performed a coalescence simulation. By applying the algorithm of Hudson (1990), for the generation of samples of genes with intragenic recombination, we generated 1,000 simulated sets of 142 haplotypes having the observed number of segregating sites and having two clades, of sizes 93 and 49. For each such sample, we scored the ratio of the sum of the squared number of mismatches between clades over that within clades (analogous to an *F* test for the analysis of variance). This gave the null distribution for a score of how distinct the two major clades were. The observed value of this test statistic was 6.44, which fell into the bottom 0.7% of the distribution. We conclude that the degree of separation of the two major clades significantly exceeds the separation expected by chance.

One obvious cause for such clade separation is geographic isolation. Table 2 shows the counts of haplotypes in clades 1 and 2, for each of the three populations studied. For table 2, the χ^2 test of heterogeneity yielded $\chi^2 = 0.561$, which is not significant. Thus, the two major clades are represented equally in all three population samples, and geographic isolation cannot explain the depth of the split. We conclude that these two major clades are very old and have been maintained in the human population for a longer period than would be expected if their variation had no physiological effect. Additional tests for evolutionary forces that may impact both this variation and the dating of when the two major clades split require comparison with an outgroup sequence such as that of chimpanzee.

Table 2

Frequencies of Haplotypes in Each of the Two Major *LPL* Clades

	Frequency, by Population			
Clade	Jackson	North Karelia	Rochester	Total
1	29	34	30	93
2	19	14	16	49
Total	48	48	46	142

Comparison of Human Polymorphism versus Divergence from Chimpanzee

Chimpanzee and human *LPL* differ at 112 sites, and, for every site that is variable in humans, one of the nucleotides is present in the chimpanzee sequence (see the bottom line of fig. 2). One of the most powerful means of detecting the role of natural selection in shaping the distribution of genetic variation is a systematic comparison of levels of polymorphism within a species versus the divergence between species (Hudson et al. 1987). If two genes are undergoing neutral evolution, so that the fate of new mutations is determined entirely by random genetic drift, and if their levels of interspecific divergence are similar (suggesting that the rates of mu-

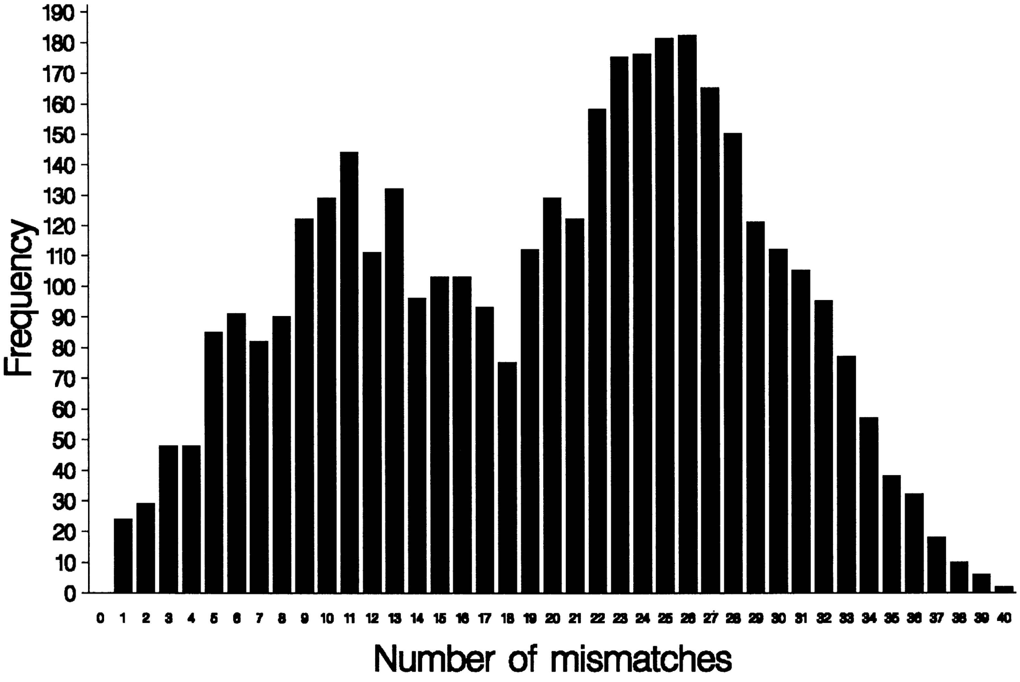

Figure 6 Pairwise mismatch distribution for *LPL,* constructed by counting the number of differences between all pairs of the 142 chromosomes in the entire sample. The pairwise mismatch distribution has a bimodal distribution, consistent with the two major clades identified in figure 2.

tation are the same), then their levels of intraspecific polymorphism likewise are expected to be similar. This is because, under the neutral model, polymorphism is determined by the mutation rate and the effective population size. A formal test comparing intraspecific polymorphism to interspecific divergence can be performed by estimation of θ and divergence time by the least-squares method (Hudson et al. 1987). The test of the data's goodness of fit to the model provides a test of neutrality. For the *LPL* data, we first divided the sequenced portion of the gene into the 5′ and 3′ halves and asked whether the two halves were consistent in their levels of polymorphism and divergence. The estimates of θ for the two halves were not found to be heterogeneous, and the overall χ^2 test for homogeneity of polymorphism and divergence was $\chi^2 = 0.234$, which is not significant. We conclude that the effects of mutation, drift, and possibly selection that are at work on the variation in *LPL* are not measurably different in the 5′ and 3′ halves of the sequenced portion of the gene.

Next, we performed the test by Hudson et al. (1987), comparing polymorphism and divergence levels in the entire 9.7 kb of the *LPL* sequence versus those found in β-globin (Harding et al. 1997) and testing the null hypothesis that the same neutral parameters fit both loci. β-Globin divergence between human and chimpanzee in was quantified by alignment of chimpanzee sequence (GenBank accession number X02345), which contains 5′ sequence and most of the coding region, with the human sequence (GenBank accession number L26463). In the 3 kb examined by Harding et al. (1997), 35 variable nucleotides were found in a sample of 349 chromosomes. In the 2,207 bp of the sequence aligned with chimpanzee, there was an average of 26 differences. The test yielded $\chi^2 = 0.40$ (1 df; $P = .53$), reflecting no departure from uniformity across loci. Use of only the 17 varying sites in the region of low recombination in β-globin also yielded no significant departure ($\chi^2 = 0.06$, 1 df; $P = .79$). This result tells us only that β-globin and *LPL* have similar patterns of polymorphism and divergence; so, variation in both genes could be neutral, or it could be undergoing the same pattern of natural selection. As more data to assess levels of polymorphism and divergence at the nucleotide level become available, this will be an informative test for comparison of homogeneity across different genes.

Another test that makes use of interspecific divergence tallies two (or more) classes of substitutions within and between species and tests their homogeneity (Templeton 1987). The chimpanzee sequence exhibited 112 fixed differences from humans, over 8,091 sites examined. Six of these differences were insertion/deletion differences (including an absence of the AAAT tetranucleotide repeat, seen in humans), and 106 were single-nucleotide differences. This contrasts with the 9 insertion/deletion polymorphisms and 79 SNPs within humans. A simple 2 × 2 contingency table yields a Fisher's–exact-test tail probability of .152 (table 3). Furthermore, 4 of the human/chimpanzee differences were coding, and the remaining 108 were noncoding, compared with 7 coding polymorphisms and 81 noncoding polymorphisms within the human sample. This contrast also is not significant. Finally, we can test the divergence and polymorphism levels of silent versus replacement differences, a test used by McDonald and Kreitman (1991) on *Adh* sequenced in *Drosophila* species, to show that interspecific comparisons revealed a striking excess of amino acid replacements. Within the human data, 4 sites were replacement polymorphisms, and 84 were silent, whereas the human/chimpanzee comparison found no replacements and 112 silent differences. This test yielded a significant result (for Fisher's exact test, $P = .036$), indicating that, relative to the amount of amino acid divergence between the two species, there is a slight excess of replacement polymorphism in humans. This pattern is the opposite of the most common pattern of departure found by use of the McDonald-Kreitman test.

Discussion

The four major conclusions that may be drawn from the data presented here all have important consequences for the design of studies in which the goal is to identify genetic factors associated with the risk of common diseases having a complex multifactorial etiology. First, *LPL* exhibits a level of nucleotide variability that is somewhat higher than the average reported values (Li and Sadler 1991; Hey 1997). For a nucleotide diversity of .002, we expect that, between a random pair of chromosomes, two nucleotides will differ for every 1 kb that is compared. Comparisons between pairs of alleles from different individuals will identify additional varying sites; thus, the number of varying sites in a sample clearly will increase with an increase of sample size. The ob-

Table 3

Contrasts between Human Polymorphism and Human/Chimpanzee Divergence in *LPL*

	No. of Sites		
Type of Variation	Poly-morphic	Human/ Chimpanzee Divergent	Prob-ability
Insertion/deletion	9	6	.152
Single-nucleotide difference	79	106	
Coding	9	6	.150
Noncoding	81	108	
Replacement	4	0	.036
Silent	84	112	

servation of 88 variable sites in a sample of 142 chromosomes fits the infinite-sites model well. The observation of such a complex pattern of variation across this gene and among populations underscores the utility of sequencing the gene of interest, in a representative sample of individuals, early in a study, to evaluate the influence of genetic polymorphism on variation in risk factors and risk of disease. Obtaining an accurate, quantitative picture of the distribution of site variation and the structure of linkage disequilibrium will be necessary for optimization of the design of subsequent association studies.

The second major conclusion drawn from these data is that, at the nucleotide-sequence level, the human population does exhibit considerable population structure. By this we mean that there are segregating nucleotide variants with relative frequencies that differ widely from one group to another. For example, not unexpectedly, the Jackson population contains more variation; in part, this is because Africans, in general, are genetically more variable than people from other regions of the world (Cavalli-Sforza et al. 1994) and also because African Americans are admixed between Africans and Europeans, which adds an additional source of variability. In fact, five haplotypes found in the Jackson sample also were found in the European-derived populations; although it is premature to declare that these haplotypes are the result of past admixture, the proportion (11%) is consistent with estimates of European admixture in the African American population (e.g., see Chakraborty et al. 1992). At the same time, there will be many varying sites with frequencies that are similar across populations. The classic means of measuring population subdivision is to use *F* statistics, which partition the total variation in the population into within- and between-population components (Wright 1931). *F* statistics calculated from blood groups or from allozymes are known to differ quantitatively from those estimated from nucleotide data (Cavalli-Sforza et al. 1994): in general, the nucleotide data are better able to discriminate those differences that are unique to each population.

When the population structure of haplotypes is considered, the issue of scale becomes important. Haplotypes constructed from a region of 1–2 kb are likely to show some degree of population overlap. However, the larger the region considered, the smaller the overlap will be. As the region examined becomes larger, eventually virtually every individual in the sample will be unique. With a scale of just 9,734 bp, we found that 81 of the 88 haplotypes are unique to each population sample. Such population-specific differences seem to provide a powerful tool for identification, including the tracing of admixed alleles. Some of this variation will appear as fixed differences; that is, two groups may be monomorphic for different nucleotides. If a phenotype of interest differs markedly between these same two groups, the cause of the phenotypic differences will be virtually impossible to ascribe to particular genetic differences, unless the biochemical mechanism is understood. Of course, this raises the question of what is the appropriate scale on which to define functional allelic units. Historically, the study of genetics has been effective because single-nucleotide changes that disrupt function of single genes may have unambiguous phenotypes. For complex traits, such a simple view cannot be the most effective way to understand the contribution of genetic variation to variation in the underlying network of interactions that are responsible for causation. Our finding that the haplotypes are organized into two clades that transcend population subdivision suggests that the unit of function for the *LPL* gene is larger than a few single-site variations.

The third key point is that a thorough population genetic analysis can be informative with regard to the forces that have acted to shape the variation that we observe today. Variation at individual nucleotide sites in *LPL* is consistent with a neutral gene in balance between mutation and random genetic drift, but, when haplotypes are examined, the data are not compatible with a simple neutral model. Equilibrium models often are rejected when human genetic data are used, mostly because human populations (including each of our sampled populations) have been in a state of expansion and historically have faced a complex hierarchy of migration and founding events (Cavalli-Sforza et al. 1994). The observation of two major clades could not have been caused by demographic isolation, because the three populations have roughly the same relative frequencies of the two major clades. This implies that the two major clades are very old.

Comparisons with sequences from our nearest relative, the chimpanzee, serve to illuminate many features of our own variation. The chimpanzee sequence, as well as sequences from other hominoid primates, serve to indicate which of the nucleotides segregating in humans is ancestral. Such sites may be conserved because, by chance, no mutations have hit those sites or because the mutations that did occur were sufficiently deleterious that they were eliminated from the population. It is remarkable that one of the nucleotides present in every segregating site in humans also is present in chimpanzee. If we root a neighbor-joining tree to the time of common ancestry of the human and chimpanzee lineages, at 4.6 million years ago, the node joining the *LPL* lineages indicates a coalescence time of 1.2 million years ago. More formal estimates of this date of common ancestry, based on coalescence theory, are in progress. The intragenic recombination requires a more elaborate analysis than that of Tavaré et al. (1997). Many studies have shown an effective population size, revealed by extant human genetic variation, to be on the order of 10^4, and our estimate of the *LPL* coalescence time is somewhat older than the expected $4N_e$ for a neutral gene. Although it is tempting to conclude that selection may be maintaining the two deep clades in the population, many other nuclear genes exhibit deep ancestry and an excess of heterozygosity, relative to the infinite-sites model (Takahata 1995; Takahata et al. 1995; Hey 1997). This suggests that the departure may have a demographic cause, and one explanation that is consistent with the data is an effective population size of 10^5 for a period of time after the human/chimpanzee split.

In the *LPL* subregion sequenced in chimpanzee, 112 sites were found to show fixed differences; that is, chimpanzee and human have different nucleotides, and there was no variation at those sites in the sample of humans. Such fixed divergent sites also may occur because of either random drift or, possibly, selection. One of the most powerful tests of evolutionary forces is to compare levels of intraspecific polymorphism versus levels of interspecific sequence divergence (Hudson et al. 1987). If mutation and drift are the only forces driving polymorphisms and differences between species, then the relative amount of divergence and polymorphism should be the same across loci. The DNA-sequence varations in *LPL* and β-globin exhibit patterns of within-human polymorphism and human/chimpanzee divergence that are not significantly different from one another, suggesting a failure to detect heterogeneity in forces that act on the variability in these two genes.

The fourth feature of the *LPL* data that merits attention is the inference that there has been abundant intragenic recombination. The rate of recombination (or gene conversion) may seem surprisingly high: almost as many recombination events as mutation events have occurred in the history of the sample of haplotypes that we examined. This conclusion is based on an assumption that warrants further examination—namely, that repeated mutation can be assumed to be absent in these data. If, despite both the absence of any sites with more than two nucleotides and the low nucleotide heterozygosity (.002), there were repeated mutations in these data, then rates of intragenic recombination would be lower. However, evidence for intragenic recombination has been detected in many genes in *Drosophila* (Kreitman and Hudson 1991; Schaeffer and Miller 1993; Richter et al. 1997). Furthermore, both local rates of intragenic recombination and local levels of DNA-sequence variation have been shown to vary across the genome, by more than an order of magnitude, in several organisms, and these two quantitative measures have been shown to be correlated positively in both *Drosophila* (Begun and Aquadro 1992) and humans (Nachman et al., in press). The consequences of recombination and gene conversion on multiple-site disequilibrium can be quite different (Clark and Zheng 1997), and, despite good estimates of conversion-tract lengths in *Drosophila* (Curtis et al. 1989), the relative rates of recombination and conversion are not even known for this model organism.

Despite the evidence for recombination in *LPL*, it should be emphasized that there still is considerable linkage disequilibrium among pairs of sites. The inference of recombinant haplotypes within *LPL* is particularly relevant to the use of anonymous SNPs for detection of associations with complex diseases. It certainly is not the case that all such SNPs will give reliable information about flanking sites. The effectiveness of mapping studies based on disequilibrium among relatively common alleles such as these (Collins et al. 1997), as well as the use of such alleles to find disease association in larger epidemiological samples, may depend on whether the causal variants arose early or late relative to the polymorphism marked by the typed sites. Several studies already have shown that there is little association between linkage disequilibrium and physical distance (Haviland et al. 1991; Zerba et al. 1991). For example, there are cases of immediately flanking sites that are in linkage equilibrium, and examples of population heterogeneity that produces linkage disequilibrium between genes on different chromosomes have been cited (Sinnock and Sing 1972; Zerba et al., in press). The complex historical processes that gave rise to the contemporary patterns of linkage disequilibrium suggest that the simple relation $D' = (1 - r)^t D$ may have little value in the prediction of the fate of linkage disequilibrium in human populations.

The issues raised by the complete sequences from *LPL* have led us to revisit a number of articles published in the 1980s, when RFLPs were first used to identify haplotypes based on multiple markers in short human chromosome regions. An example of this work is a study of a 61-kb region of chromosome 11, including the *apoAI/CIII/AIV* apolipoprotein gene complex, screened by restriction enzymes (Antonarakis et al. 1988). Eleven RFLP sites were used to identify haplotypes, from nuclear families, for 129 Mediterranean and 67 African American chromosomes. Whereas only a minority of these sites were in pairwise disequilibrium, most of these were in maximum disequilibrium. When complete DNA-sequence data are considered, this picture seems to change considerably. Compared with the *apoAI/CIII/AIV* gene region, the *LPL*-sequence data appear to exhibit more recombination and less linkage disequilibrium. However, the high prevalence of four-gamete relationships in the RFLP-study data, as in our *LPL* sample, suggests that recombination was common among restriction sites as well. Because so many additional varying sites were included, the *LPL* data provide sufficient detail to make nearly all haplotypes unique. If we were to consider only small numbers of *LPL* sites, we also would identify fewer haplotypes, as well as a more tractable frequency distribution, evidence for disequilibrium, and broader sharing of the same haplotype among populations.

A number of other studies of RFLP haplotypes at candidate susceptibility loci were performed in the 1980s (e.g., Chakravarti et al. 1984, 1986; Leitersdorf et al. 1989). The details vary among studies, but the basic findings are qualitatively similar. Many more haplotypes were found than were expected, but not nearly as many per kilobase as were found in the *LPL* data. Multiple occurrences of four-gamete states were present among the restriction sites, in samples of roughly the same scale and population diversity as those for the *LPL* data, with only partial and patchy disequilibrium among the RFLP sites. These RFLP studies revealed evidence for recombination hot spots between major regions of the genes under consideration, just as the Jakobsen et al. (1997) test did for the *LPL* data. In the RFLP era, disequilibrium was considered to be undesirable, because the objective was to identify multiple variant sites as independent potential markers for intervening causal variants. For population association studies based on linkage disequilibrium, investigators recently have suggested developing dense SNP maps (Collins et al. 1997). However, our *LPL* data show that, at the sequence level, this may not be so simple. Although it is true that many site pairs are found in linkage disequilibrium, the structure of disequilibrium would lead to unacceptably high type II error rates; that is, a closely linked functional locus would not be detected, because of linkage equilibrium between the SNP and the functional locus.

A more recent study used DNA sequencing to identify all variation in a 3-kb region of the human β-globin gene in 349 chromosomes obtained from individuals from many different populations (Harding et al. 1997). Different regions of the gene had different densities of haplotypes (as also was seen by our sliding window of 5 or 10 sites in *LPL*), suggesting different levels of past recombination effects, including a known recombination hot spot. In 349 sequences in a 2.67-kb subregion, 23 haplotypes clearly were recombinants. When these recombinants were excluded from further analysis, the remaining haplotypes fitted a cladistic structure that did not suggest the kind of global two-clade structure that we found in *LPL*. Overall, the globin-sequence variation revealed an average nucleotide diversity of .0018, indicating levels of variation very similar to those in *LPL*.

Many of our findings may seem to be inconsistent with the success of a number of investigators in the identification of genes for disease phenotypes in populations founded relatively recently by a small number of individuals. Examples are the Amish population, in which Hirschprung disease has been studied (Chakravarti 1996), and studies of North Karelia populations, in which a number of diseases have been mapped successfully (Peltonen 1996). Such successes have led to the view that a dense SNP map—being even more definitive, in terms of a simpler underlying mutational process—would be useful in an even broader population context (e.g., the United States). The explanation for this apparent inconsistency seems to be straightforward. First, the disease phenotypes mapped in this way, to date, usually are rare, often are of single-locus etiology, and exhibit clear Mendelian inheritance in families. In a small population, such traits are likely to have only a small number of causal alleles, so that samples ascertained by such "phenotypes" are likely to comprise clones of descendants of the founding allele(s).

In principle, we should face an easier situation with a known candidate gene, such as *LPL*, for which the key role in lipid physiology has been clearly established. However, we have documented sufficient variation in a complex pattern of linkage disequilibrium that association studies would require inordinate sample sizes in order to ascribe cause to particular nucleotide sites. At present, it is not practical to consider the sequencing of complete candidate genes in thousands of individuals; therefore, some form of data reduction is necessary. Our objective is to use the sample reported here to identify a subset of variations that can be typed in thousands of individuals measured for CHD risk factors in these populations, in order to detect risk-factor phenotypes associated with the chosen subset of sites used as markers.

The data presented in this article suggest that a blind association or disequilibrium test, with three or four random markers chosen from variable sites even *within* 10 kb of the *LPL* gene, would not be a reliable method of detection of nearby causal variation. An obvious alternative is to find a subset of sites that capture the cladistic, or disequilibrium, structure of variation in the whole gene (Templeton 1996). Our inference of intragenic recombination in *LPL* is based on the assumption that recombination, rather than repeated mutation, at a site caused site pairs to have all four gametes present. If, instead, we assume that repeated mutation occurs frequently and only accept recombination when such site pairs are clustered significantly (Crandall and Templeton 1998), then the apparent number of recombination events will be decreased. Application of cladistic methods may be possible in this case, but it will require novel approaches. For the same reason, large association or disequilibrium-mapping studies probably should be undertaken only after a clear picture of the underlying structure of DNA-sequence variation in the targeted populations has been obtained. Investigators contemplating such studies should be aware that the development of reliable association tests may be limited by the complexities of intragenic recombination, population substructure, and an intricate, hierarchical pattern of linkage disequilibrium.

Acknowledgments

We thank R. Travis Merritt for his assistance in obtaining chimpanzee data reported in this article; Terry Melton for chimpanzee DNA; Ingrid Jakobsen for compatability analysis; and Brian Lazzaro, Michael Nachman, Alan Templeton, Sarah Tishkoff, and two anonymous reviewers for their critical comments on the manuscript. We acknowledge support from the National Heart, Blood, and Lung Institute (grants HL39107, HL58238, HL58239, and HL58240).

Electronic-Database Information

Accession numbers and URL for data in this article are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Web/Genbank (for human *LPL* reference sequence [AF050163], chimpanzee *LPL* reference sequences [AF071087–AF071094], chimpanzee β-globin reference sequence [X02345], and human β-globin sequence [L26463])

References

Antonarakis S, Oettgen P, Chakravarti A, Halloran S, Hudson R, Feisee L, Karathanasis K (1988) DNA polymorphism haplotypes of the human apolipoprotein APOA1-APOC3-APOA4 gene cluster. Hum Genet 80:265–273

Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster.* Nature 356:519–520

Betrán E, Rozas J, Navarro A, Barbadilla A (1997) The estimation of the number and the length of gene conversion tracts from population DNA sequence data. Genetics 146: 89–99

Brown AHD, Feldman MW, Nevo E (1980) Multilocus structure of natural populations of *Hordeum spontaneum.* Genetics 96:523–536

Brunzell J (1995) Familial lipoprotein lipase deficiency and other causes of the chylomicronemia syndrome. In: Scriver C, Beaudet A, Sly W, Valle D (eds) The metabolic and molecular bases of inherited disease, 7th ed. McGraw-Hill, New York, pp 1913–1933

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

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

Chakravarti A (1996) Endothelin receptor-mediated signaling in Hirschsprung disease. Hum Mol Genet 5:303–307

Chakravarti A, Beutow KH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH (1984) Nonuniform recombination within the human β-globin gene cluster. Am J Hum Genet 36:1239–1258

Chakravarti A, Elbein S, Permutt M (1986) Evidence for increased recombination near the human insulin gene: implication for disease association studies. Proc Natl Acad Sci USA 83:1045–1049

Chamberlain JC, Thorn JA, Oka K, Galton DJ, Stocks J (1989) DNA polymorphisms at the lipoprotein lipase gene: associations in normal and hypertriglyceridaemic subjects. Atherosclerosis 79:85–91

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122

Clark AG, Zheng Y (1997) Dynamics of linkage disequilibrium in bacterial genomes. J Evol Biol 10:663–676

Collins F, Guyer M, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Crandall LA, Templeton AR (1998) Statistical approaches to detecting recombination. In: Crandall KA (ed) Evolution of HIV: implications for biology and society. Johns Hopkins University Press, Baltimore

Curtis D, Clark SH, Chovnick A, Bender W (1989) Molecular analysis of recombination events in Drosophila. Genetics 122:653–661

Ewens WJ (1979) Mathematical population genetics. Springer-Verlag, Berlin

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

Fu Y, Li WH (1993) New statistical tests of neutrality for DNA samples from a population. Genetics 133:693–709

Gagne C, Gaudet D (1995) Dyslipoproteinemias in Quebec: primary deficit in lipoprotein lipase and familial hypercholesterolemia. Union Med Can 124:61–67

Garrod A (1902) The incidence of alkaptonuria: a study in chemical individuality. Lancet 11:1616–1620

Gotoda T, Yamada N, Kawamura M, Kozaki K, Mori M, Ishibashi S, Shimano H, et al (1991) Heterogeneous mutations in the human lipoprotein lipase gene in patients with familial lipoprotein lipase deficiency. J Clin Invest 88: 1856–1864

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, et al (1997) Archaic African *and* Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772–789

Harpending H, Batzer M, Gurven M, Jorde L, Rogers A, Sherry S (1998) Genetic traces of ancient demography. Proc Natl Acad Sci USA 95:1961–1967

Hartl DL, Clark AG (1997) Principles of population genetics, 3d ed. Sinauer Associates, Sunderland, MA

Hata A, Robertson M, Emi M, Lalouel JM (1990) Direct detection and automated sequencing of individual alleles after electrophoretic strand separation: identification of a common nonsense mutation in exon 9 of the human lipoprotein lipase gene. Nucleic Acids Res 18:5407–5411

Haviland MB, Kessling AM, Davignon J, Sing CF (1991) Estimation of Hardy-Weinberg and pairwise disequilibrium in the apolipoprotein AI-CIII-AIV gene cluster. Am J Hum Genet 49:350–365

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Hegele RA, Evans AJ, Tu L, Ip G, Brunt JH, Connelly PW (1994) A gene-gender interaction affecting plasma lipoproteins in a genetic isolate. Arterioscler Thromb 14:671–678

Heizmann C, Kirchgessner T, Kwiterovich PO, Ladias JA, Derby C, Antonarakis SE, Lusis AJ (1991) DNA polymorphism haplotypes of the human lipoprotein lipase gene: possible association with high density lipoprotein levels. Hum Genet 86:578–584

Hey J (1997) Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol Biol Evol 14: 166–172

Hey J, Wakeley JA (1997) A coalescent estimator of the population recombination rate. Genetics 145:833–846

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239

Hudson RR (1987) Estimating the recombination parameter of a finite population without selection. Genet Res 50: 245–250

——— (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–44

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. Mol Biol Evol 9:138–151

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159

Jakobsen IB, Wilson SR, Easteal S (1997) The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. Mol Biol Evol 14:474–484

Jorde LB, Bamshad M, Rogers AR (1998) Using mitochondrial and nuclear DNA markers to reconstruct human evolution. Bioessays 20:126–136

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903

Kreitman M, Hudson RR (1991) Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics 127:565–582

Leitersdorf E, Chakravarti A, Hobbs HH (1989) Polymorphic DNA haplotypes at the LDL receptor locus. Am J Hum Genet 44:409–421

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

Li WH, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129:513–523

Ma Y, Henderson HE, Ven Murthy MR, Roederer G, Monsalve MV, Clarke LA, Normand T, et al (1991) A mutation in the human lipoprotein lipase gene as the most common cause of familial chylomicronemia in French Canadians. N Engl J Med 324:1761–1766

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in Drosophila. Nature 351:652–654

Murthy V, Julein P, Gagne C (1996) Molecular pathobiology of the human lipoprotein lipase gene. Pharmacol Ther 70: 101–135

Nachman MW, Bauer VL, Crowell SL, Aquadro CF. DNA variability and recombination rates at X-linked loci in humans. Genetics (in press)

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson TG, Stengård J, Salomaa V, et al (1998) Genome resequencing and variation analysis in a 9.7 kb region of the human lipoprotein lipase gene. Nat Genet 19:233–240

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25:2745–2751

Normand T, Bergeron J, Fernandez-Margallo T, Bharucha A, Ven Murthy MR, Julien P, Gagne C, et al (1992) Geographic distribution and genealogy of mutation 207 of the lipoprotein lipase gene in the French Canadian population of Quebec. Hum Genet 89:671–675

Peacock RE, Hamsten A, Nilsson-Ehle P, Humphries SE (1992) Associations between lipoprotein lipase gene polymorphisms and plasma correlations of lipids, lipoproteins and lipase activities in young myocardial infarction survivors and age-matched healthy individuals from Sweden. Atherosclerosis 97:171–185

Peltonen L (1996) Identification of disease genes in genetic isolates. Methods 9:129–135

Reina M, Brunzell JD, Deeb SS (1992) Molecular basis of familial chylomicronemia: mutations in the lipoprotein lipase and apolipoprotein C-II genes. J Lipid Res 33: 1823–1832

Reymer PWA, Gagne E, Groenemeyer BE, Zhang H, Forsyth I, Jansen H, Seidell JC, et al (1995) A lipoprotein lipase mutation (Asn291Ser) is associated with reduced HDL cholesterol levels in premature atherosclerosis. Nat Genet 10: 28–34

Richter B, Long M, Lewontin RC, Nitasaka E (1997) Nucleotide variation and conservation at the *dpp* locus, a gene controlling early development in Drosophila. Genetics 145: 311–323

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Santamarina-Fojo S (1992) Genetic dyslipoproteinemias: role of lipoprotein lipase and apolipoprotein C-II. Curr Opin Lipidol 3:186

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Schaeffer SW, Miller EL (1993) Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics 135:541–552

Sinnock P, Sing CF (1972) Analysis of multilocus genetic systems in Tecumseh, Michigan. II. Consideration of the correlation between nonalleles in gametes. Am J Hum Genet 24:393–415

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Stephens JC (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol Biol Evol 2:539–556

Stocks J, Thorn JA, Galton DJ (1992) Lipoprotein lipase genotypes for a common premature termination codon mutation detected by PCR-mediated site-directed mutagenesis and restriction digestion. J Lipid Res 33:853–857

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphisms. Genetics 123: 585–595

——— (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics 143:1457–1465

Takahata N (1995) A genetic perspective on the origin and history of humans. Annu Rev Ecol Syst 26:343–372

Takahata N, Satta Y, Klein J (1995) Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48:198–221

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518

Templeton AR (1987) Genetic systems and evolutionary rates. In: Campbell KSW, Day MF (eds) Rates of evolution. Allen & Unwin, London

——— (1996) Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome. In: Variation in the human genome. John Wiley & Sons, Chichester, pp 259–276

Thorn JA, Chamberlain JC, Alcolado JC, Oka K, Chan L, Stocks J, Gaiton DJ (1990) Lipoprotein and hepatic lipase gene variants in coronary atherosclerosis. Atherosclerosis 85:55–60

Tunstall-Pedoe H, Kuulasmaa K, Amoyel P, Arveiler D, Rajakangas AM, Pajak A (1994) Myocardial infarction and coronary deaths in the World Health Organization MONICA Project: registration procedures, event rates and case fatality in 38 populations from 21 countries in 4 continents. Circulation 90:583–612

Wakeley J (1997) Using the variance of pairwise differences to estimate the recombination rate. Genet Res 69:45–48

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland, MA

Wiebusch H, Funke H, Santer R, Richter W, Assmann G (1996) A novel missense (E163G) mutation in the catalytic subunit of lipoprotein lipase causes familial chylomicronemia. Hum Mutat 8:392

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Yang WS, Nevin DN, Peng R, Brunzell JD, Deeb S (1995) A mutation in the promoter of the lipoprotein lipase (LPL) gene in a patient with familial combined hyperlipidemia and low LPL activity. Proc Natl Acad Sci USA 92:4462–4466

Zerba, KE, Ferrell RE, Sing CF. Genetic structure of five susceptibility gene regions for coronary artery disease: disequilibria within and among regions. Hum Genet (in press)

Zerba KE, Kessling AM, Davignon J, Sing CF (1991) Genetic structure and the search for genotype-phenotype relationships: an example from disequilibrium in the *ApoB* gene region. Genetics 129:525–533