Cladistic Structure Within the Human *Lipoprotein Lipase* Gene and Its Implications for Phenotypic Association Studies

Alan R. Templeton,* Kenneth M. Weiss,†,‡ Deborah A. Nickerson,§ Eric Boerwinkle and Charles F. Sing††**

**Department of Biology, Washington University, St. Louis, Missouri 63130-4899, †Institute of Molecular Evolutionary Genetics, Department of Biology, and ‡Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania 16802, §Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730, **Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77225-0334 and ††Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109-0618*

Manuscript received January 10, 2000
Accepted for publication June 29, 2000

ABSTRACT

Haplotype variation in 9.7 kb of genomic DNA sequence from the human *lipoprotein lipase* (*LPL*) gene was scored in three populations: African-Americans from Jackson, Mississippi (24 individuals), Finns from North Karelia, Finland (24), and non-Hispanic whites from Rochester, Minnesota (23). Earlier analyses had indicated that recombination was common but concentrated into a hotspot and that recurrent mutations at multiple sites may have occurred. We show that much evolutionary structure exists in the haplotype variation on either side of the recombinational hotspot. By peeling off significant recombination events from a tree estimated under the null hypothesis of no recombination, we also reveal some cladistic structure not disrupted by recombination during the time to coalescence of this variation. Additional cladistic structure is estimated to have emerged after recombination. Many apparent multiple mutational events at sites still remain after removing the effects of the detected recombination/gene conversion events. These apparent multiple events are found primarily at sites identified as highly mutable by previous studies, strengthening the conclusion that they are true multiple events. This analysis portrays the complexity of the interplay among many recombinational and mutational events that would be needed to explain the patterns of haplotype diversity in this gene. The cladistic structure in this region is used to identify four to six single-nucleotide polymorphisms (SNPs) that would provide disequilibrium coverage over much of this region. These sites may be useful in identifying phenotypic associations with variable sites in this gene. Evolutionary considerations also imply that the SNPs in the 3′ region should have general utility in most human populations, but the 5′ SNPs may be more population specific. Choosing SNPs at random would generally not provide adequate disequilibrium coverage of the sequenced region.

CORONARY artery disease (CAD) is the major cause of death in many countries, and several genetic and nongenetic risk factors have been identified (Sing and Skolnick 1979; Rao *et al.* 1984; Sing *et al.* 1995; Boerwinkle *et al.* 1996). Among the genetic factors is the human *Lipoprotein Lipase (LPL)* locus. For this reason, Nickerson *et al.* (1998) sequenced and described allelic variation in a 9.7-kb region within this locus in 71 individuals from three populations that differ greatly in incidence of CAD. Clark *et al.* (1998) described the haplotype structure of that variation, and Templeton *et al.* (2000) examined the roles of recombination and mutation as sources of haplotype variation. The ultimate purpose of these studies is to investigate associations between phenotypic variation in CAD and in intermediate phenotypes (such as serum cholesterol levels) in the general population and variation in candidate loci such as *LPL*. There are many ways of analyzing such associations (Templeton *et al.* 1987; Templeton and Sing 1993; Templeton 1996; Long *et al.* 1998; Lyman and Mackay 1998). One method is the nested cladistic method in which the evolutionary history of the haplotype variation is estimated and used as a statistical design to test for phenotypic associations (Templeton *et al.* 1987; Templeton and Sing 1993; Templeton 1996). Such an approach requires that a substantial fraction of variation in the gene is cladistically structured and that these evolutionary relationships among the observed haplotypes can be estimated. The nested design is tolerant of some error or ambiguity in the estimated evolutionary structure (Templeton and Sing 1993).

Templeton *et al.* (2000) showed that recombination is common within this 9.7-kb region of *LPL* and that nearly half of the variable sites in this segment of *LPL* occur at a small number of sites with highly mutable motifs. Both of these features can complicate the estimation of cladistic structure, with recombination being

Corresponding author: Alan R. Templeton, Department of Biology, Campus Box 1137, Washington University, St. Louis, MO 63130-4899. E-mail: temple_a@biology.wustl.edu

the more worrisome. However, a cladistic approach to phenotype association studies does not require a complete absence of recombination. Indeed, this approach can yield better biological insights when some recombination is present because in those cases inferences about the physical location of causative variation are possible, whereas in the complete absence of recombination it is possible to localize markers of association only in the temporal/historical framework of the haplotype tree (Templeton *et al.* 1987; Templeton and Sing 1993; Markham *et al.* 1996; Templeton 1996; Keavney *et al.* 1998). However, these previous successful applications of the cladistic approach occurred in DNA regions in which recombination was present but rare. In contrast, our analyses of the 9.7-kb region of the *LPL* gene indicate that recombination is not only present, but is common at particular regions within this gene (Clark *et al.* 1998; Templeton *et al.* 2000). Fortunately, the recombination events are concentrated into a 1.9-kb region within this 9.7-kb segment (Templeton *et al.* 2000). Therefore, subregions in which recombination is absent or rare can be identified, and the evolutionary structure of each subregion can be estimated separately (Templeton and Sing 1993). Templeton *et al.* (2000) also found that three motifs previously discovered to be associated with high mutability display rates of variation one to two orders of magnitude higher than sites not known to be highly mutable. This high frequency of variation at a small number of sites raises the possibility that these sites have experienced more than one mutational event.

The first goal of this article is to show that there is indeed considerable cladistic structure in the 5′ and 3′ portions of this 9.7-kb segment that flank the region of high recombination identified by Templeton *et al.* (2000). A second goal of this article is to reconstruct the evolutionary history of the entire sequenced region by removing the inferred statistically significant recombination and gene conversion events from the evolutionary "tree" estimated in Templeton *et al.* (2000) under the null hypothesis of no recombination or gene conversion. In this manner, we intend to separate that portion of the cladistic structure of this candidate gene region that has never been disrupted by recombination from that portion of the cladistic structure that arose due to mutational accumulation after recombination. The estimated nonrecombinant and postrecombinant cladistic structure will be validated against the evolutionary structure estimated in the 5′ and 3′ flanking subregions that display little evidence for recombination.

A third goal of this article is to use the subregional, nonrecombinant, and postrecombinant trees to examine the possible role of multiple mutational events at highly mutable sites as a source of homoplasy. The fourth goal of this article is to show how the results of such an evolutionary/recombinational analysis can be applied to the problem of genetic/phenotypic associations. A final goal is to show how analyses of the evolutionary history, amount of recombination, and patterns of mutation can be used to identify a small number of single nucleotide polymorphisms (SNPs) that could be used in phenotypic association studies through linkage disequilibrium.

MATERIALS AND METHODS

Population samples: We use the data of Nickerson *et al.* (1998) and Clark *et al.* (1998) on three human samples: (1) a Jackson, Mississippi, sample ($n = 24$) that is part of an ongoing National Heart, Lung, and Blood Institute study of hypertension in African-Americans; (2) a sample from the FINRISK study from North Karelia ($n = 24$), an area in eastern Finland that has had the world's highest known CAD risk and that has been studied for many years (Tunstall-Pedoe *et al.* 1994); and (3) a sample from the Rochester Family Heart Study ($n = 23$), a study of cardiovascular disease risk in the Rochester, Minnesota, area.

DNA sequencing: DNA sequencing was performed on diploid genotypes as described in Nickerson *et al.* (1998), followed by extensive confirmatory resequencing and data validation procedures (Clark *et al.* 1998; Nickerson *et al.* 1998). The *LPL* reference sequence is available in GenBank (http://www.ncbi.nlm.nih.gov/Web/Genbank, accession no. AF050163).

Haplotype determination: Haplotypes were determined by a mixture of allele-specific PCR (AS-PCR) and the haplotype subtraction algorithm of Clark (1990), followed by extensive confirmatory analyses (Clark *et al.* 1998) that indicate a high degree of confidence. Nevertheless, many haplotypes were not observed directly, so not all haplotypes are known with certainty. The analyses in this article are therefore conditional on the haplotype inferences made in Clark *et al.* (1998). Nickerson *et al.* (1998) revealed a total of 88 variable sites in the 9.7-kb region. Length variation at a tetranucleotide repeat was excluded from haplotype determination as this variation does not fall under the same evolutionary models as the other variable sites (Clark *et al.* 1998). Clark *et al.* (1998) determined the linkage phase of 69 of the remaining 87 variable sites for which the rarer variant was found in at least three chromosomes. These 69 variable sites (Table 1) determined a total of 88 distinct haplotypes (Clark *et al.* 1998).

Inference of cladistic structure when recombination is present: Figure 1 presents the overall flowchart for inferring cladistic structure used in this article. Figure 1 also illustrates our inference procedure with a fictional data set that makes the steps easier to follow than with the much more complicated *LPL* data set. Step 1 in Figure 1 starts with known haplotypes, just as we start with the *LPL* haplotypes given in Clark *et al.* (1998).

The second step is to estimate the statistical parsimony (SP) tree for the entire DNA region under the null hypothesis of no recombination or gene conversion. Statistical parsimony was specifically designed for estimating intraspecific haplotype trees (Templeton *et al.* 1992). Standard maximum parsimony minimizes the total number of mutational steps in the tree, but under a neutral model of intraspecific evolution, it is unlikely (in a quantifiable fashion) for haplotypes that are separated by only one or very few nucleotide differences to have experienced multiple mutational hits at the few sites by which they differ. Haplotypes that differ at many nucleotides are more likely to have experienced multiple mutational hits. Hence, statistical parsimony gives precedence to connections between haplotypes that differ by one or a few nucleotides

TABLE 1

Sequence variants used to define haplotypes in the *LPL* locus

Site[a]	Position[b]	Variant[c]	Site[a]	Position[b]	Variant[c]	Site[a]	Position[b]	Variant[c]
1	106	C → A	24	3297	$(A)_4 \rightarrow (A)_5$	47	6718	A → G
2	110	A → C	25	3609	T → C	48	6772	A → G
3	145	G → A	26	3723	T → C	49	6863	C → T
4	325	T → C	27	3843	G → A	50	6939	delAAAT
5	343	$(TG)_3 \rightarrow (TG)_4$	28	4016	C → G	51	7315	G → C
6	479	T → C	29	4343	A → T	52	7344	A → G
7	551	$(A)_3 \rightarrow (A)_2$	30	4346	C → G	53	7360	A → G
8	736	T → C	31	4418	C → T	54	7413	T → C
9	1216	C → G	32	4426	T → C	55	7754	A → C
10	1220	T → C	33	4509	T → C	56	8088	insAG
11	1286	C → T	34	4576	A → T	57	8089	G → T
12	1547	A → C	35	4872	G → A	58	8285	C → G
13	1571	C → G	36	4935	T → C	59	8292	A → C
14	1828	C → G	37	5085	G → A	60	8393	T → G
15	1939	A → G	38	5168	T → C	61	8533	A → C
16	2131	C → T	39	5395	$(A)_8 \rightarrow (A)_9$	62	8537	A → C
17	2500	G → A	40	5441	T → C	63	8538	$(A)_3 \rightarrow (A)_2$
18	2619	A → G	41	5554	A → C	64	8644	T → C
19	2987	T → G	42	5560	A → G	65	8755	G → A
20	2996	C → A	43	5687	T → C	66	8852	T → G
21	3022	G → A	44	6250	C → T	67	9402	A → G
22	3248	C → G	45	6595	G → C	68	9712	G → A
23	3290	$(T)_7 \rightarrow (T)_8$	46	6678	T → G	69	9721	G → A

From Nickerson *et al.* (1998).

[a] A site number is assigned to each variable character used to define haplotypes in order from 5′ to 3′.

[b] Position in the baseline sequence (GenBank accession no. AF050163).

[c] Substitution and insertion/deletion variants are reported as the state in the baseline sequence → alternative state.

and preferentially allocates multiple hits to long branches. The difference between these two methods is illustrated in Figure 2. Step 2 has also been completed for the entire 9.7-kb *LPL* region (Templeton *et al.* 2000) and is shown in Figure 3.

Step 3 in Figure 1 is the estimation and testing of recombination and gene conversion events with the haplotypes given in step 1 through the use of the algorithm given in Crandall and Templeton (1999) with additions given in Templeton *et al.* (2000). This phase of the inference sequence has also been completed with the *LPL* haplotypes (Templeton *et al.* 2000). A total of 29 recombination events statistically significant at the 5% level and 1 significant gene conversion event were detected. The details of the inferred recombination events are available at the MDECODE website (http://www.mdecode.umich.edu/). Twenty-four of the 29 significant recombination events were clustered between variable sites 19 (2987 in the map of Nickerson *et al.* 1998) and 35 (4872 in the *LPL* map). Therefore, sites 1–18 in Table 1 are used to define the 5′ flanking region to the recombinational hotspot, and sites 36–69 are used to define the 3′ flanking region.

In step 4 we remove the impact of recombination upon the cladistic structure. Two different approaches are used, indicated by 4a and 4b in Figure 1. We proceed to step 4a in our inference procedure only if the recombination encountered in step 3 is either rare and/or concentrated into hotspots, as is true for *LPL*. Step 4a is to subdivide the DNA region into smaller segments that show little or no internal recombination. This is then followed by step 5a: estimating separate SP trees for each subregion. Since step 3 indicated that the 5′ subregion of *LPL* defined by variable sites 1–18 and the 3′ subregion defined by variable sites 36–69 have experienced little internal recombination, we estimate separate haplotype trees in the 5′ subregion (variable sites 1–18) and in the 3′ subregion (36–69), thereby excluding the recombinational hotspot (19–35) from the analysis. When using only a subset of the sites, many of the original haplotypes collapse into a common state. Tables 2 and 3 present these collapsed haplotype categories for the 5′ and 3′ regions, respectively. Although the 5′ and 3′ regions have very little internal recombination, there is some (Templeton *et al.* 2000). The haplotypes that are the products of recombination or the descendants of the initial recombinant within each subregion are excluded when the cladistic structure of that subregion is estimated, as recommended by Templeton *et al.* (1987). Two recombination events (numbers 12 and 21, available at the MDECODE website) are contained within the 5′ subregion, leading to the exclusion of the haplotypes found in the resulting recombinant categories 5′-9 and 5′-10 in Table 2. In addition, the inferred gene conversion event took place within the 5′ region, leading to the exclusion of its product, haplotype 20J. The 3′ region had three inferred internal recombination events (recombination events 5, 13, and 29 at the MDECODE website), thereby leading to the exclusion of haplotypes 76R, 15J, 63N, and 71R.

Steps 4b and 5b (Figure 1) show an alternative and novel method for removing the effects of recombination and estimating cladistic structure in a DNA region that has experienced recombination. Step 4b removes all the recombination events inferred in step 3 from the SP tree estimated for the entire region obtained in step 2. By removal, we mean that the recombinant haplotype itself is removed from the SP tree, the homoplasies that were used to identify the recombinant

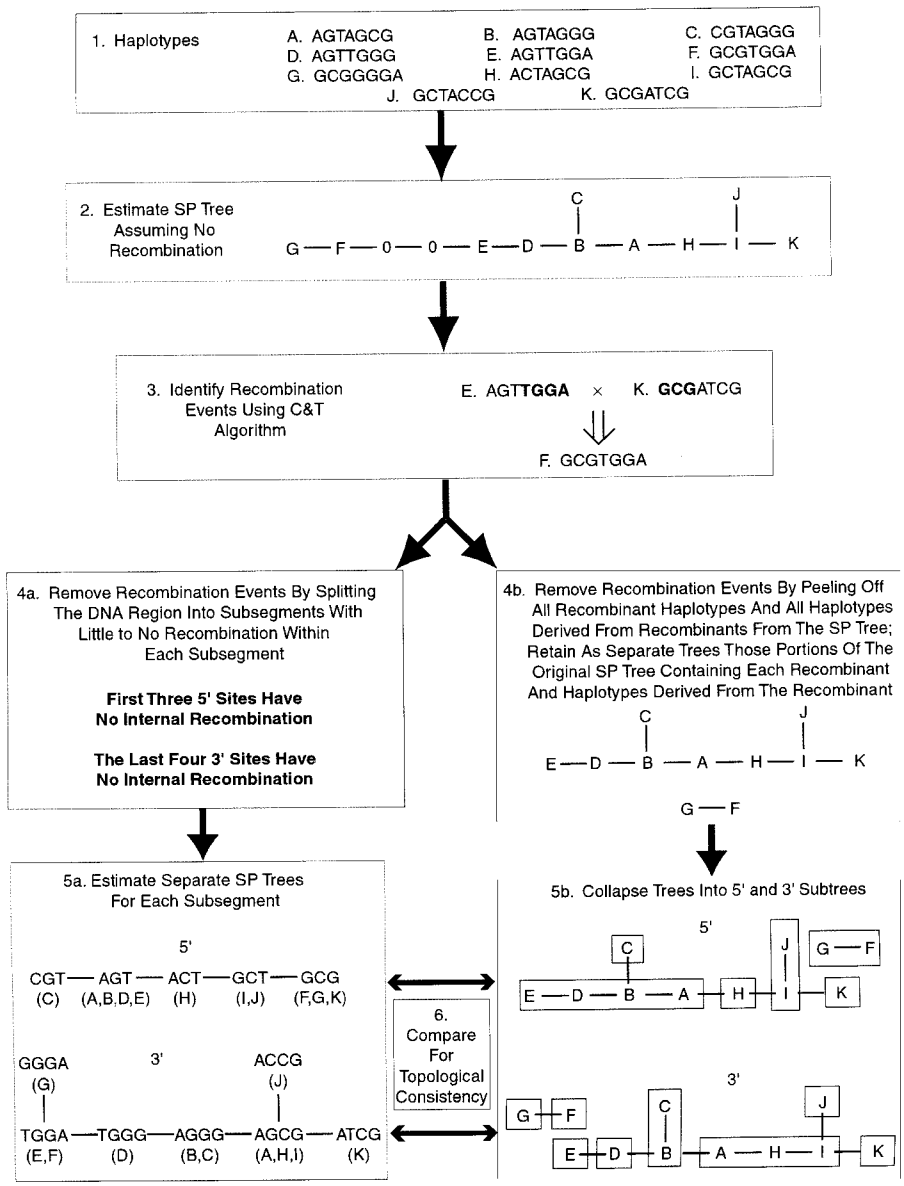


FIGURE 1.—An inference flowsheet illustrated by a fictional data set. Step 1 is to obtain the haplotype data, and in the fictional data set 11 haplotypes (labeled A–K) are present. Step 2 is to estimate the SP tree of the haplotypes found in step 1 under the assumption of no recombination or gene conversion. In the tree for the fictional data, each line represents a single nucleotide substitution, and a 0 indicates an inferred intermediate haplotype that is necessary to interconnect haplotypes found in step 1 but that is not present in the actual sample. In step 3, the CRANDALL and TEMPLETON (1999) algorithm (C&T) is applied to the SP tree to identify statistically significant recombination events. For the fictional data set, a single recombination event is detected, with haplotype F being the recombinant and haplotypes E and K being candidates for the parental types. In step 4a, the gene region is split into subsegments with little or no internal recombination. For the fictional data set, the sequences are split into the three nucleotides 5′ of the single inferred recombination position and the four nucleotides 3′ to the recombination site. In step 5a, separate SP trees are then estimated for each subregion. The haplotypes in the fictional data set that are identical for the 5′ and 3′ subregions are indicated in parentheses below their respective 5′ and 3′ sequence states. In step 4b, the detected recombination events are removed from the SP tree obtained in step 2. For the fictional data set, this step requires discarding the three mutational events interconnecting haplotypes E and F in the step 2 SP tree, as the C&T algorithm used in step 3 indicates that these three homoplasious "mutations" are not mutational events at all but rather are due to the recombination event. This step also requires the removal of the recombinant haplotypes from the SP tree (in this case haplotype F) and any other parts of the step 2 SP tree that are descended from the recombinant (in this case, haplotype G and the mutational transition interconnecting haplotypes F and G). This results in a subset of the step 2 SP tree that is called the peeled tree. The cladistic structure arising after recombination is also retained: in the fictional data set this consists of only haplotype F, the mutational transition connecting it to haplotype G, and haplotype G. In step 5b, the peeled tree and the postrecombinational tree(s) are collapsed into haplotype categories that reflect each subregion defined in step 4a. Boxes indicate haplotypes that are identical in the subregion under consideration. In step 6, the topologies of the trees obtained in step 5a are tested for consistency with the cladistic structures found in the 5′ and 3′ ends of the peeled and postrecombinant trees obtained in step 5b.

under the CRANDALL and TEMPLETON (1999) algorithm are removed and discarded because they do not represent actual mutational events, and all haplotypes and branches that are derived from the original recombinant haplotype by subsequent mutational events are removed. We refer to that portion of the SP tree from step 2 that remains after this removal as the "peeled tree," reflecting the fact that the effects of all inferred recombination events have been peeled off the tree.

The peeled tree should ideally reflect the component of haplotype diversity that has not been affected by recombination during the coalescence of this DNA region. However, additional cladistic structure could have arisen in haplotype lineages derived from the recombinant haplotype. This postrecombinational cladistic structure is estimated in step 4b as those subsets of the SP tree estimated in step 2 that consist of branches and haplotypes derived from each of the original recombinants by subsequent mutational events.

To see if these two methods of estimating cladistic structure in a recombining DNA region yield compatible results, we next collapse the cladistic networks estimated in step 4b by first considering only the 5′ sites (1–18) and then the 3′ sites (36–69). This results in separate 5′ and 3′ tree topographies

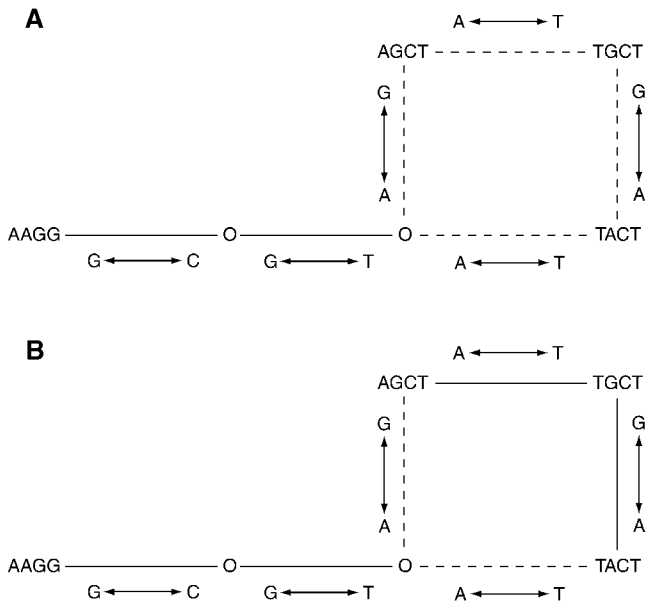


FIGURE 2.—A hypothetical example illustrating the difference between maximum parsimony (A) and statistical parsimony (B). (A) The maximum parsimony networks that interconnect four haplotypes. An O indicates an inferred intermediate haplotype not present in the sample. Solid lines indicate connections that are unique in the maximum parsimony tree, whereas dashed lines indicate ambiguity. The lines of ambiguity form a loop that can be broken at any of its four sides to yield four distinct maximum parsimony networks, each with a total length of five mutations. (B) The statistical parsimony networks for the same haplotypes, under the assumption that the limit of parsimony (as calculated in TEMPLETON *et al.* 1992) is one. This limit means that single nucleotide transitions that interconnect haplotypes that differ only at a single nucleotide are regarded as certain. Thus, haplotypes AGCT and TGCT differ by only a single nucleotide, so the direct connection between them is regarded as certain under statistical parsimony, as is the connection between haplotypes TGCT and TACT. Note that breaking the loop of ambiguity by discarding the branch between the AGCT and TGCT haplotypes is not allowed under statistical parsimony because such a break would require these haplotypes to be interconnected evolutionarily by three mutations, including a homoplasy. Instead, homoplasies are allocated to the longer branches under statistical parsimony (in this case the two possible connections of haplotype AAGG to either haplotype AGCT or haplotype TACT). Hence, there are only two statistically parsimonious networks compatible with this fictional data set, in contrast to four maximum parsimony solutions.

(step 5b in Figure 1) that are contained within the cladistic structure estimated in step 4b but that correspond to the subregions identified in step 4a. This allows a direct comparison of the two methods (steps 4a and 4b) for removing the effects of recombination. Step 6 in Figure 1 is therefore to test the concordance of the tree topologies estimated in step 5a *vs.* those estimated in step 5b. The null hypothesis that a given data set fits equally well into two alternative evolutionary trees or networks is tested with a Wilcoxon matched pair, signed rank tests according to the procedures given in TEMPLETON (1983, 1987) and the Templeton test option of PAUP* (SWOFFORD 1997).

Testing for an association between mutagenic sites and homoplasy: TEMPLETON *et al.* (2000) showed that nearly half of the polymorphic sites in *LPL* have mutagenic motifs. This high polymorphism frequency raises the possibility that such sites are also likely to experience multiple mutational events, leading to true homoplasies in the haplotype tree. Because the recombination rate is high in the *LPL* region, most homoplasies have been inferred to be caused by recombination and not mutation, and this obscures the role of mutation as a contributor to homoplasy. However, steps 4a and 4b are designed to eliminate the homoplasies due to inferred recombination events. Hence, we examine every mutational event in the trees estimated in steps 4b, 5a, and 5b and note whether or not the mutated site was one of the mutagenic sites identified previously (TEMPLETON *et al.* 2000). Because of sample size constraints, we then classify each site appearing in this estimated cladistic structure as having 0–1 homoplasies or >1 homoplasy. A Fisher's exact test is then applied to the resulting two-by-two table to test the null hypothesis no association between mutagenic category (highly mutable *vs.* nonmutagenic) and homoplasy category (0–1 homoplasies *vs.* ≥2).

RESULTS

The first three steps in the inference chain (Figure 1) have already been completed and published in previous articles for *LPL* (CLARK *et al.* 1998; TEMPLETON *et al.* 2000). Moreover, our previous analysis (TEMPLETON *et al.* 2000) had already identified the subregions that would show little or no recombination (step 4a), as indicated in the previous section. Hence, we begin our presentation of the results with step 5a in Figure 1.

Tree estimation of the subregions flanking the recombinational hotspot (step 5a): Figures 4 and 5 show, respectively, the SP trees estimated for the 5′ and 3′ regions that flank the recombinational hotspot. The 5′ region has two statistically parsimonious solutions (Figure 4) for the haplotype categories given in Table 2. Figure 5 shows the resulting statistical parsimony networks for the remaining haplotypes as grouped into the 3′ categories shown in Table 3. The long branches leading to the four major termini that frequently serve as candidates for parental types in recombination (TEMPLETON *et al.* 2000) are well defined and have no loops, but the 3′ haplotypes at each of the four termini show much homoplasy, including several loops, indicating multiple statistically parsimonious solutions (Figure 5).

Estimation of cladistic structure for the entire 9.7-kb region with no detected recombination (step 4b): Figure 3 shows the SP tree estimated by TEMPLETON *et al.* (2000) under the null hypothesis of no recombination or gene conversion. Figure 6 shows the cladistic structure that remains after the removal of all 30 significant recombination and gene conversion events, the homoplasies attributable to them, and any cladistic structure that evolved from the original recombinant/converted haplotypes.

Estimation of cladistic structure for the entire 9.7-kb region that evolved from recombinant haplotypes (step 4b): Figure 7 shows the 29 recombination events and 1 gene conversion event (given at the MDECODE web-

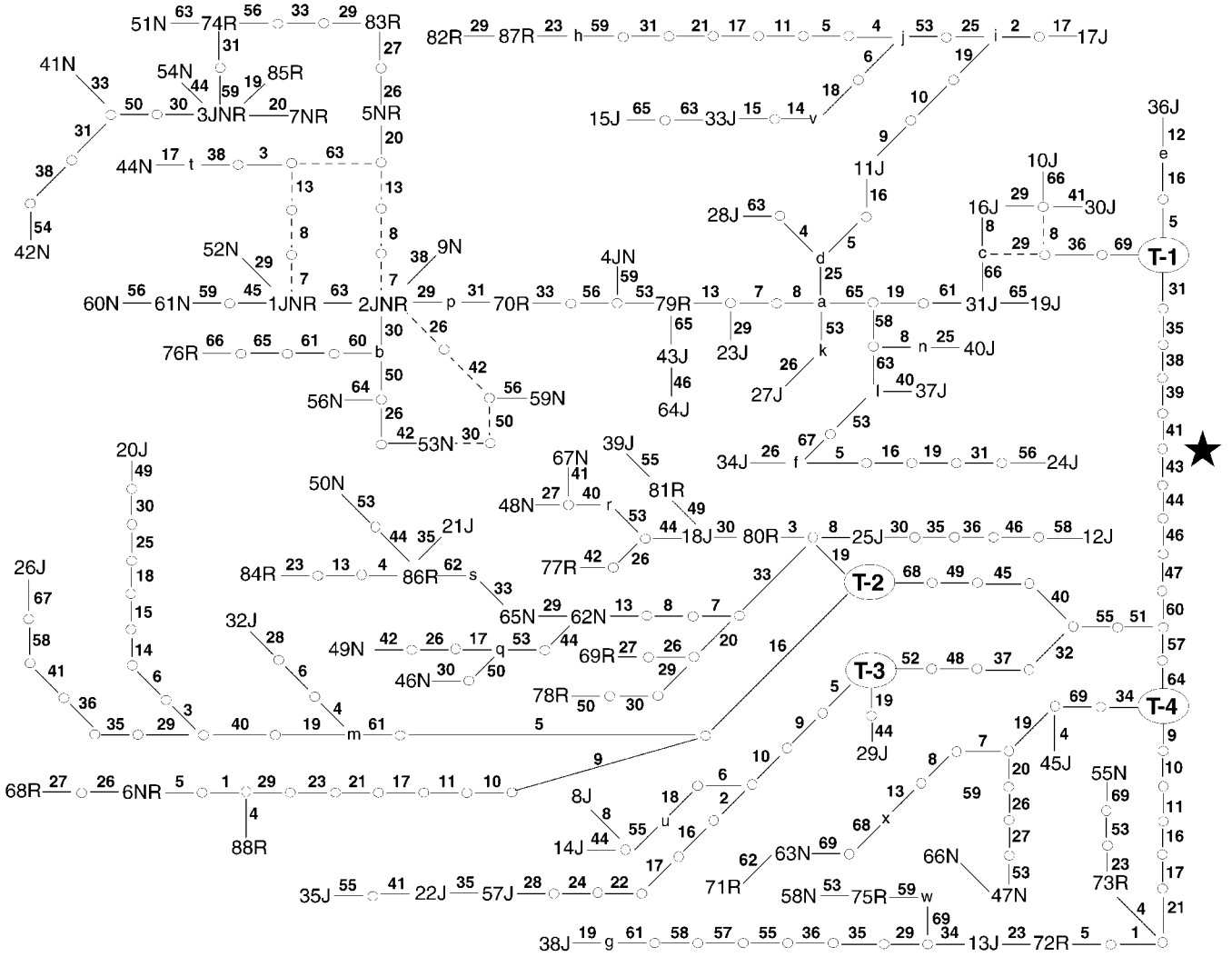


FIGURE 3.—The estimated total haplotype network under the null hypothesis of no recombination or gene conversion. Haplotypes are indicated by a number followed by one or more letters indicating the populations in which that particular haplotype was present (J, Jackson; N, North Karelia; R, Rochester). Small circles indicate nodes in the tree that represent intermediate haplotype states not found in the sample. Each line (solid or dashed) represents a single mutational event. The site involved in the mutation is indicated near the line by a small boldface number, with the numbers corresponding to the variable site numbers given in Table 1. Dashed lines indicate where loops or alternatives create ambiguity in the topology of the tree. Nodes that define four major clades are indicated by an oval containing T-i, where i can be 1, 2, 3, or 4. Other nodes involved in recombination events are indicated by the lowercase letters a–n and p–x. A star next to the branch between T-1 and the node connecting to T-2, -3, and -4 indicates that this is the branch that connects to the chimpanzee sequence and hence indicates the rooting of the tree.

site), along with the cladistic structure estimated to have arisen as recombinant haplotypes and their descendants accumulated subsequent mutations. Sometimes one of these descendant haplotypes served as an inferred parental type in a subsequent recombination event. This results in an interlocking of the cladistic structure that evolved from one recombinant with that of another recombinant, as is also shown in Figure 7. In other cases, neither the recombinant nor any of its descendant haplotypes engaged in any subsequent recombination events. Such cases are indicated in Figure 7 by the absence of any connection to any other recombination event or its postrecombinational cladistic structure.

Collapsing the nonrecombinant and postrecombinant cladistic structure into separate 5′ and 3′ subregional trees (step 5b): We collapsed the peeled tree of nonrecombinant cladistic structure obtained in step 4b into its 5′ and 3′ subsets as defined by the subregions identified in step 4a by simply removing all variable characters not in the subregion of interest (step 5b, Figure 1). Figure 8A shows the resulting haplotype network when the peeled nonrecombinant tree for the entire 9.7-kb region includes only variable characters 1–18 (the 5′ region flanking the recombinational hotspot), and Figure 8B shows the corresponding 3′ collapsed tree (characters 36–69). Similarly, we collapsed the postrecombi-

national cladistic structure shown in Figure 7 by considering only the 5′ characters (1–18), with the result shown in Figure 9A, and by considering only the 3′ characters (36–69), with the result shown in Figure 9B.

Cross-validation of the cladistic structures emerging from steps 5a and 5b (step 6): We first compare the cladistic structures estimated for the 5′ and 3′ regions flanking the recombinational hotspot (with any recombination events involving those flanking regions excluded) as shown in Figures 4 and 5 with the nonrecombinant cladistic structure given in Figure 8. There are fewer haplotype categories in Figure 8 than in Figures 4 and 5 because all crossover events anywhere in the 9.7-kb region were peeled off in obtaining Figure 8, whereas the trees given in Figures 4 and 5 excluded only crossover events that were internal to either the 5′ or 3′ flanking regions, respectively. Consequently, the contrast of alternative topologies is limited only to that portion of the topology defined by the shared haplotype categories. The Templeton test for the 5′ region is 1, with only 1 observation out of 18 not tied. A minimum of 5 untied observations is required for significance at the 5% level, so this result is not even close to significance at the 5% level. The Templeton test for the 3′ region is 3 with only 2 untied observations out of 34. This result is also not significant at the 5% level.

We next cross-validate the postrecombinational cladistic structure given in Figure 9 by comparing it with the 5′ and 3′ trees shown in Figures 4 and 5, respectively. There are no topological inconsistencies for the 5′ region, so there was no need to perform a Templeton test. For the 3′ region, the Templeton test is 2 with only 1 untied observation out of 34, a result not significant at the 5% level.

Association between homoplasies in the cladistic structure with mutagenic categories after the removal of the effects of detected recombination: The cladistic structure is probably most accurately reconstructed for the 5′ and 3′ flanking regions, which show considerable cross-validation through two different estimation techniques (step 6 above). Hence, we count the number of homoplasies at sites 1–18 and 36–69 as displayed in Figures 4 and 5, but using the resolved loops from Figures 8 and 9 that are topologically consistent with Figures 4 and 5. Making use of this added resolution is justified because the trees in Figures 8 and 9 are based on more character state information than those in Figures 4 and 5. In addition, six recombination/gene conversion events were excluded in estimating the cladistic structures shown in Figures 4 and 5, as mentioned earlier. Figure 9 shows the mutations inferred to have occurred after these recombination/gene conversion events, and the sites found in the relevant flanking regions given in Figure 9 are also included in this analysis. In this fashion, all 18 sites in the 5′ region and all 34 sites in the 3′ region are included. Each of these 52 sites was then characterized as being highly mutable [CG dinucleotides, mononucleotide runs of length five or greater, and DNA polymerase α arrest sites having the motif TG(A/G)(A/G)GA] or not, as detailed in Templeton *et al.* (2000). The number of homoplasies associated with each of these sites was then counted. The above counts exclude the 17 variable sites found in the recombinational hotspot. To include these sites, we used the cladistic structure shown in Figures 6 and 7 to count the number of homoplasies for sites 19–35. Because there is no cross-validation for these sites, our analysis of the association between highly mutable sites and homoplasy is done both including and excluding these 17 sites.

The results are shown in Table 4, and a Fisher's exact test reveals a significant pattern in which homoplasies are disproportionately found at highly mutable sites, when the recombinational hotspot sites (19–35) are excluded (a two-tailed probability of 0.0089 under the null hypothesis of homogeneity) or included (a two-tailed probability of 0.0013 under the null hypothesis of homogeneity).

TABLE 2

The sets of haplotypes that are identical for variable sites 1–18

Haplotype set name	Haplotypes in set
5′-1	1JNR, 2JNR, 4JN, 9N, 21J, 43J, 46N, 50N, 52N, 53N, 56N, 59N, 60N, 61N, 62N, 63N, 64J, 65N, 70R, 71R, 76R, 79R, 86R
5′-2	10J, 12J, 16J, 25J, 30J, 40J
5′-3	3JNR, 5NR, 7NR, 19J, 27J, 31J, 34J, 37J, 41N, 42N, 47N, 51N, 54N, 66N, 69R, 74R, 78R, 83R, 85R
5′-4	28J, 45J
5′-5	18J, 39J, 48N, 67N, 77N, 80R, 81R
5′-6	11J, 24J, 26J
5′-7	15J, 33J
5′-8	17J, 22J, 35J, 57J
5′-9	6NR, 13J, 38J, 58N, 68R, 72R, 75R
5′-10	55N, 73R, 82R, 87R, 88R

Any haplotypes not listed in this table have a unique 5′ sequence in the 1–18 interval.

DISCUSSION

Cladistic structure: Although recombination is common in the *LPL* gene, the inference that most of this recombination is concentrated into a hotspot (Templeton *et al.* 2000) implies that there should be much cladistic structure to the haplotype variation in the flanking regions having few inferred recombination or gene conversion events. This prediction is borne out by Figures 4 and 5. The 5′ region has only two statistically parsimonious solutions (Figure 4), so it represents a well-resolved cladistic network. The 3′ region has more

TABLE 3

The sets of haplotypes that are identical for variable sites 36–69

Haplotype set name	Haplotypes in set
3′-1	1JNR, 52N
3′-2	2JNR, 5NR, 70R, 83R
3′-3	27J, 33J, 74R
3′-4	3JNR, 4JN, 7NR, 82R, 85R, 87R
3′-5	11J, 17J, 23J, 79R
3′-6	6NR, 18J, 25J, 62N, 65N, 68R, 69R, 80R, 88R
3′-7	21J, 84R, 86R
3′-8	22J, 57J
3′-9	13J, 72R, 73R
3′-10	47N, 55N
3′-11	58N, 66N
3′-12	10J, 31J

Any haplotypes not listed in this table have a unique 3′ sequence in the 36–69 interval.

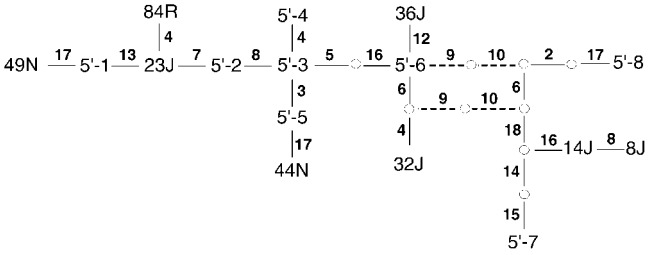


FIGURE 4.—The haplotype network for the 5′ region encompassing variable sites 1–18. The 5′ haplotype categories are given in Table 2. The layout of the network is the same as that given in Figure 3.

ambiguity, as shown by the multiple loops in Figure 5. One loop (associated with haplotype 49N) is due to the fact that a critical variable nucleotide was not scored (position 40) in haplotype 49N, but the others are the result of apparent homoplasy that could be due to either multiple mutational events or undetected recombination/gene conversion events. Nevertheless, the long branches leading to the four major termini (T-1 through T-4 in Figure 5) remain well defined and with no loops, as are many of the branches leading from these termini to specific haplotypes or haplotype clusters. Therefore, there is also much cladistic structure in the 3′ region. Moreover, the long branch lengths among the four major termini imply that this evolutionary structure is quite old. This conclusion is reinforced by the fact that when the chimpanzee sequence (NICKERSON *et al.* 1998) is added as an out-group, it roots the 3′ tree within the long branch connecting T-1 to the other three termini (data not shown). Thus, there is ancient and deep evolutionary structure in the gene region despite frequent overall recombination, particularly for sites that flank the recombinational hotspot and that are preferentially found in these long branches.

These conclusions are reinforced by the peeled tree shown in Figure 6, which provides an estimate of the cladistic structure contained within the entire 9.7-kb region sequenced that has not been altered by any detected recombination or gene conversion events. This peeled tree also contains the long branches among the four major termini that are defined primarily by sites 3′ to the recombinational hotspot. Indeed, the peeled tree in Figure 6 shows that some nonrecombinant states have persisted all the way back to the root of the *LPL* gene tree. The peeled tree also reinforces the conclusion of a recombinational hotspot because the peeled tree has only one current haplotype associated with termini 2, 4, and 5, thereby implying that virtually all the haplotypes defined by the evolutionary old 3′ structure on the long branch between T-1 and the remaining termini have undergone recombination. Figure 7 also reveals that substantial cladistic structure has also arisen after recombination events have occurred. Thus, there is both nonrecombinant and postrecombinant cladistic structure in this region of the *LPL* locus. We anticipate that cladistic structure will exist for other nuclear DNA regions showing recombination when the recombination is concentrated into hotspots (CHAKRAVARTI *et al.* 1984).

Partitioning cladistic structure into nonrecombinational and postrecombinational components: Peeling off inferred recombinants from a tree estimated through statistical parsimony under the null hypothesis of no recombination is a novel method of estimating cladistic structure in a DNA region subject to recombination. Such peeling partitions cladistic structure into a component that has never been influenced by detectable recombinational events and a component that evolved after recombination. To check the accuracy of this novel approach, we cross-validated it with the haplotype networks estimated in a more traditional fashion within subregions with little evidence for recombination (Figures 4 and 5) and found no significant differences. The few topological differences that did exist were minor. Starting with the 5′ region, the only discrepancy between Figure 8A and Figure 4 is in the position of 5′-2. There is an additional homoplasy at site 8 in the peeled tree that is not present in the more traditionally estimated tree. The single discrepancy between the topologies could be due to a recombination event that placed site 8 upon a new 3′ background but had too few markers to be statistically significant under the recombination test given by CRANDALL and TEMPLETON (1999).

Similarly, the 3′ region of the peeled nonrecombinant tree shown in Figure 8B differs from the topology given in Figure 5 by only two additional homoplasies involving sites 53 and 65 and by a slightly different but equally parsimonious (when inference is restricted to characters 36–69) placement of haplotype 59N. Once again, these

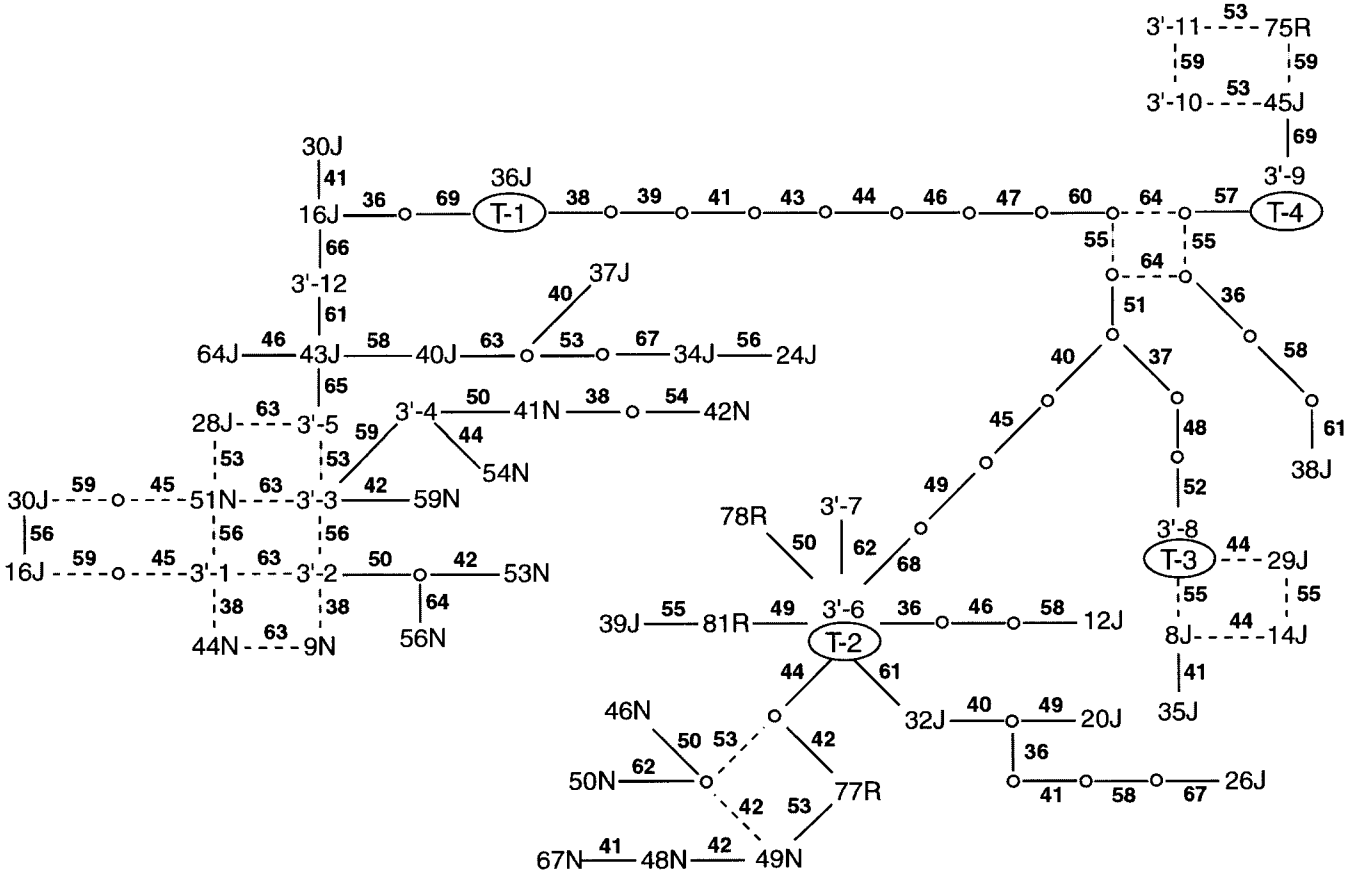


FIGURE 5.—The haplotype network for the 3′ region encompassing variable sites 36–69. The 3′ haplotype categories are given in Table 3. The layout of the network is the same as that given in Figure 3. In addition, nodes that define four major clades that served as common termini for parental types involved in recombination events are indicated by an oval containing T-i, where i can be 1, 2, 3, or 4.

topological differences are not significant by the Templeton test and can be explained by two additional recombination events that were not statistically significant because of too few markers. Therefore, the peeled tree shown in Figure 6 accurately reflects both the 5′ and 3′ evolutionary structure that flanks the recombinational hotspot found in this 9.7-kb region. The few topological discrepancies that are detected are explicable by three additional recombination events, but these three events did not involve enough markers to achieve statistical significance. Given the overall number of recombination events that were statistically significant, it is reasonable to expect that some additional recombination events occurred but were not detected (TEMPLETON *et al.* 2000).

The results given in Figure 7 imply that much of the cladistic structure observed in the 9.7-kb region arose due to mutational accumulation in haplotype lineages that were initially created by a recombination or gene conversion event. In addition, Figure 7 implies that a single haplotype lineage could have been affected by multiple recombination events during its evolutionary history. For example, haplotypes 63N and 71R have 10 inferred recombination events in their evolutionary history, as well as postrecombination mutational accumulation. Overall, Figure 7 reveals a complex history of interlocking recombination events as a major force in shaping the haplotype diversity found in this region of the *LPL* gene. At this point, we have not yet determined how accurately this complex recombinational history has been reconstructed, although we do know that two events are not robust to alternative tree topologies (recombination events 23 and 24 in Figure 7) and that events 26 and 28 are collapsed into a single recombination event under some alternative tree topologies, as are events 16 and 20 (TEMPLETON *et al.* 2000). In light of these uncertainties, Figure 7 is best regarded as a hypothesis of the recombinational history of the DNA region that explains much of the haplotype complexity and that is based upon individual recombination events that all have statistical support. However, even if all the details implied by Figure 7 are not accurate, we feel that it does portray the complexity of the interplay among many recombinational and mutational events that would be needed to explain the patterns of haplotype diversity in this gene. This qualitative conclusion is ro-

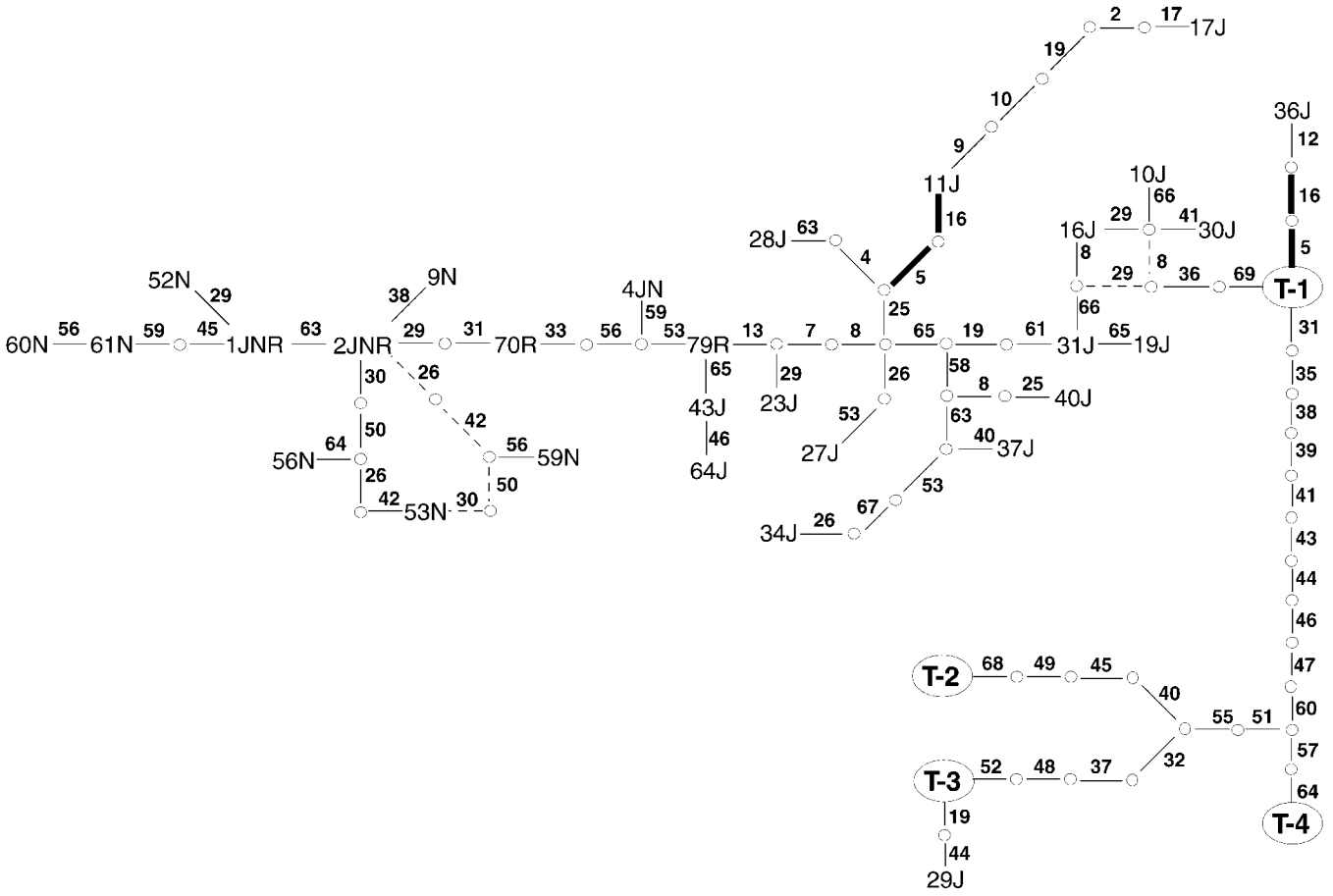


FIGURE 6.—The estimated haplotype tree after all recombinant and gene conversion clades have been removed. The layout of the network is the same as that given in Figure 3. The thick lines indicate the two possible resolutions of recombination event 6 (the second event in Figure 2 of TEMPLETON *et al.* 2000). One of these pairs of thick lines should be removed, but the symmetry of this recombination event prevents any inference on which one leads to a parental type and which to a recombinant.

bust to alternative tree topologies and exclusion of likely false positives identified in TEMPLETON *et al.* (2000).

As with the peeled nonrecombinant tree, we found that the postrecombinational cladistic structure suggested by Figure 7 was topologically consistent with the 5′ and 3′ trees shown in Figures 4 and 5, respectively, as ascertained by the Templeton test. There were no topological inconsistencies for the 5′ region between Figures 3 and 9A, and there is only a single inconsistency involving two additional homoplasies at sites 53 and 69 in the 3′ region.

All this topological consistency indicates that this peeling method has promise in partitioning the cladistic structure in gene regions with recombination into nonrecombinant and postrecombinant components. We used the existence of a recombinational hotspot in the *LPL* region as a tool to provide a cross-validation test, but the inference scheme given in steps 1, 2, 3, 4b, and 5b in Figure 1 could also be applied to regions with uniform recombination. In contrast, step 4a would be difficult to implement under uniform recombination unless recombination were rare because it would be impossible to find large subregions that show little or no internal recombination. Hence, the peeling method of inferring cladistic structure has a broader range of applicability than the method of subdividing a gene into smaller regions that show little or no internal recombination.

Evidence for multiple mutational events: TEMPLETON *et al.* (2000) showed that sites in the *LPL* region that have known mutagenic motifs have much higher levels of polymorphism than the remaining sites. This high polymorphic frequency raises the possibility that such sites are also likely to experience multiple mutational events, leading to true homoplasies in the evolutionary structure of the haplotype tree. Table 4 reveals that mutagenic motifs are also associated with increased homoplasy. The homoplasy counts in Table 4 eliminate the impact of all 30 significant recombination/gene conversion events detected previously (TEMPLETON *et al.* 2000). As mentioned above and discussed in TEMPLETON *et al.* (2000), it is quite likely that additional but nonsignificant recombination/gene conversion events have occurred. To the extent that some of the remaining

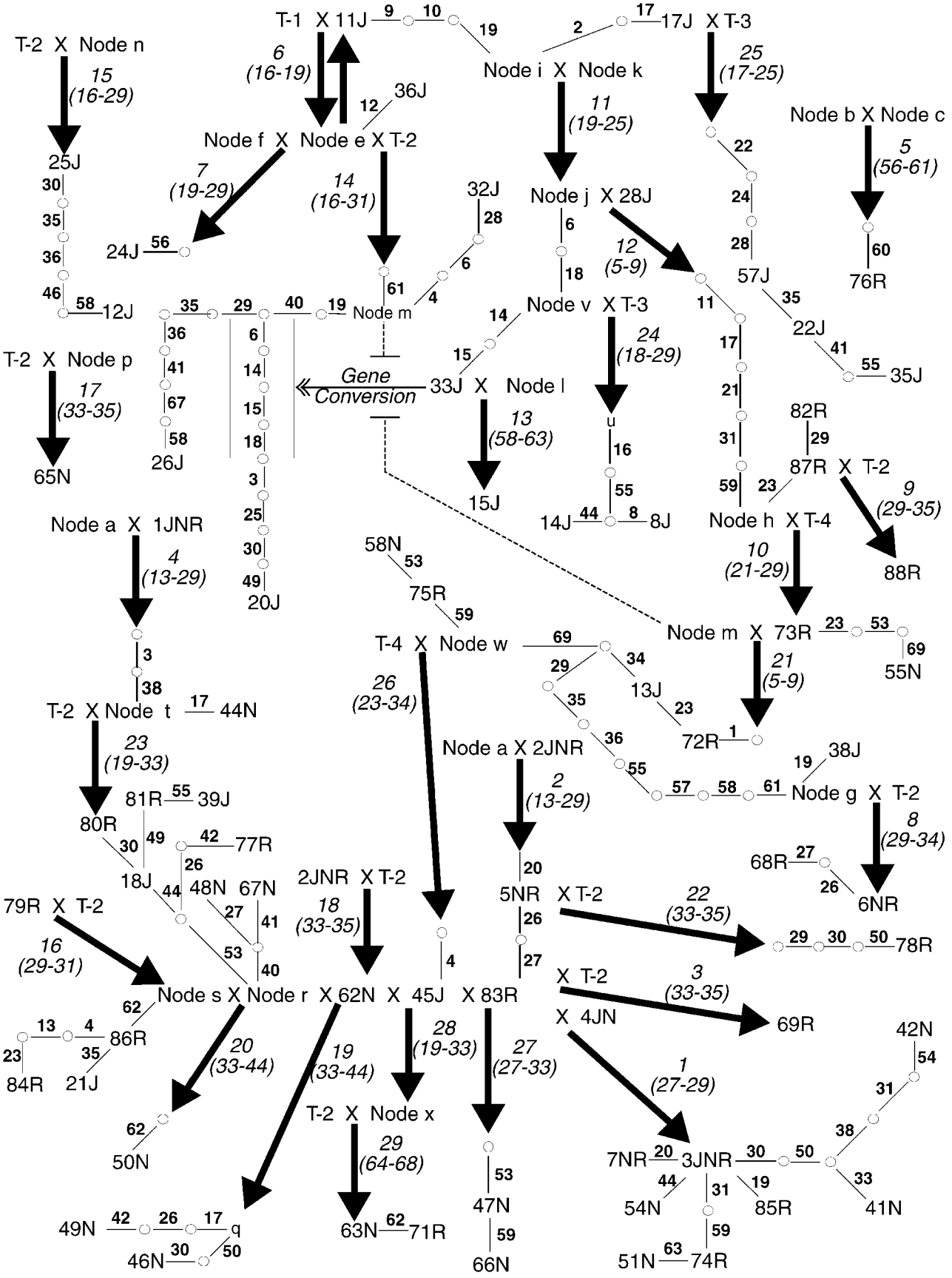
homoplasies are still caused by recombination rather than mutation, the potential effect of highly mutable sites upon patterns of apparent homoplasy will be diminished because apparent homoplasies due to recombination depend only upon physical position and not upon sequence motif at a particular site. Hence, the homoplasy counts in Table 4 are conservative for testing the null hypothesis that highly mutable sites have the same incidence of homoplasy as do the remaining sites.

Because undetected recombination or gene conversion would be expected to weaken this association, these results indicate that the highly mutable sites have indeed been subjected to multiple mutational events. Hence, the infinite sites mutation model, which assumes that multiple events can never occur, is not strictly applicable to the *LPL* region. This is an important conclusion because many of the statistics commonly used to analyze DNA sequence data are based upon the infinite sites model, and some of these statistics, but perhaps not all (Clark *et al.* 1998), may be inappropriate for this region (further discussion on this point is in Templeton *et al.* 2000). Statistics based upon the infinite sites model must therefore be used with caution in the *LPL* region and in other DNA regions in the human genome that have mutagenic motifs associated with a substantial proportion of the variation (Rideout *et al.* 1990; Krawczak and Cooper 1991; Krawczak *et al.* 1995; Reiss *et al.* 1991; Jones *et al.* 1992; Magewu and Jones 1994; Todorova and Danieli 1997; Agarwal *et al.* 1998; Nakagawa *et al.* 1998; Tvrdik *et al.* 1998).

Implications for genotype/phenotype association studies: As shown in Figures 3–7, there is much cladistic structure in the *LPL* locus despite common recombination, primarily because the recombination is concentrated into a small hotspot. This cladistic structure is important for future studies on associations between genetic variation and phenotypic variation. Indeed, the *LPL* gene is an ideal candidate for a cladistic analysis because the recombinational hotspot preserves much cladistic structure while simultaneously allowing positional inferences to be made on any detected phenotypic associations (Templeton and Sing 1993; Keavney *et al.* 1998). One would perform separate nested analyses using the permutational methods developed for diploid genotypic data (Templeton *et al.* 1988) on the two subregions using the subregional haplotype trees given in Figures 4 and 5. The nesting rules given in Templeton and Sing (1993) can deal with the loops found in the 3′ end while making use of the extensive cladistic structure found in this subregion. Alternatively, one could use the results shown in Figures 6 and 7 to resolve most of the loops before creating the nested design from the subregional haplotype networks. Any significant phenotypic effects detected within either subregion would most likely be due to a mutation or mutations on the appropriate 5′ or 3′ side of the hotspot, and the many recombination events observed within the hotspot could be used to confirm this hypothesis. Moreover, the six recombination/gene conversion events that occur within the subregions (three in each) would be excluded from the initial cladistic analysis of phenotypic associations, but could then be used to physically localize within the flanking regions any detected phenotypic associations in a manner similar to the use of recombinants in the *Alcohol Dehydrogenase* locus of *Drosophila melanagaster* by Templeton *et al.* (1987). However, two of these six recombination events may be false positives (Templeton *et al.* 2000), so recombination events 6 and 12 in Figure 7 should be used with caution in attempts to localize physically any detected phenotypic associations. Even with this caution, a nested cladistic analysis of this region of the *LPL* locus should yield more information than a cladistic analysis in a region that had no recombination, as already demonstrated by Templeton *et al.* (1987), Templeton and Sing (1993), and Keavney *et al.* (1998) for other nuclear regions showing some recombination.

Choosing a small number of SNPs for disequilibrium mapping in the *LPL* region: One of the major new goals of the Human Genome Project is to develop a map of more than 100,000 SNPs distributed over the entire human genome (Collins *et al.* 1998). Although the SNPs are frequently described as being randomly chosen, they are not because an effort is made to find SNPs that have common allelic alternatives. Because common alleles tend strongly to be old alleles (Kimura and Ohta 1973; Castelloe and Templeton 1994), the current strategy will preferentially use SNPs that are evolutionarily old. The justification for creating a SNP map is the expectation that linkage disequilibrium, at least in relatively isolated, homogeneous human populations, should exist between at least one of these SNPs and the common alleles at any locus in the genome. Given that the current strategy has about one SNP per gene (although recent recommendations increase this number), this claim requires that a single SNP marker near or in the locus of interest would show significant disequilibrium with nearly all common alleles at this locus. This is certainly not the case for even the third of the *LPL* gene analyzed in this article.

It is virtually impossible for any single SNP to show disequilibrium across the entire 9.7-kb region sequenced because of the recombinational hotspot. First, SNPs in the hotspot show little disequilibrium either with one another or with any markers in the 5′ and 3′ flanking regions (see Figure 8 in Templeton *et al.* 2000). This is particularly true for SNPs with common alleles: by selecting such SNPs one also selects for old polymorphisms, thereby providing much time for recombination to occur if the SNP is located in the hotspot. Hence, ~20% of the sequenced portion of the *LPL* gene represents a disequilibrium blind spot. If a "randomly" chosen SNP fell into this hotspot region, it would most probably be useless for any analysis requiring disequilibrium. The


T-1 X 11J
Node i X Node k
17J X T-3
T-2 X Node n
15 (16-29)
6 (16-19)
25 (17-25)
Node b X Node c
5 (56-61)
Node f X Node e X T-2
11 (19-25)
7 (19-29)
14 (16-31)
Node j X 28J
25J
24J
32J
12 (5-9)
57J
76R
12J
Node m
Node v X T-3
22J
35J
T-2 X Node p
17 (33-35)
Gene Conversion
33J X Node l
24 (18-29)
13 (58-63)
65N
26J
15J
14J
8J
82R
87R X T-2
9 (29-35)
Node h X T-4
10 (21-29)
88R
Node a X 1JNR
4 (13-29)
20J
58N
75R
Node m X 73R
55N
T-4 X Node w
21 (5-9)
13J
26 (23-34)
72R
T-2 X Node t
44N
23 (19-33)
Node a X 2JNR
2 (13-29)
38J
Node g X T-2
8 (29-34)
81R
39J
80R
77R
68R
6NR
18J
48N
67N
2JNR X T-2
18 (33-35)
5NR X T-2
22 (33-35)
79R X T-2
16 (29-31)
78R
Node s X Node r X 62N X 45J X 83R
X T-2
3 (33-35)
69R
X 4JN
86R
84R
21J
20 (33-44)
19 (33-44)
28 (19-33)
27 (27-33)
1 (27-29)
42N
50N
T-2 X Node x
29 (64-68)
7NR
3JNR
54N
85R
41N
49N
47N
63N
71R
66N
51N
74R
46N

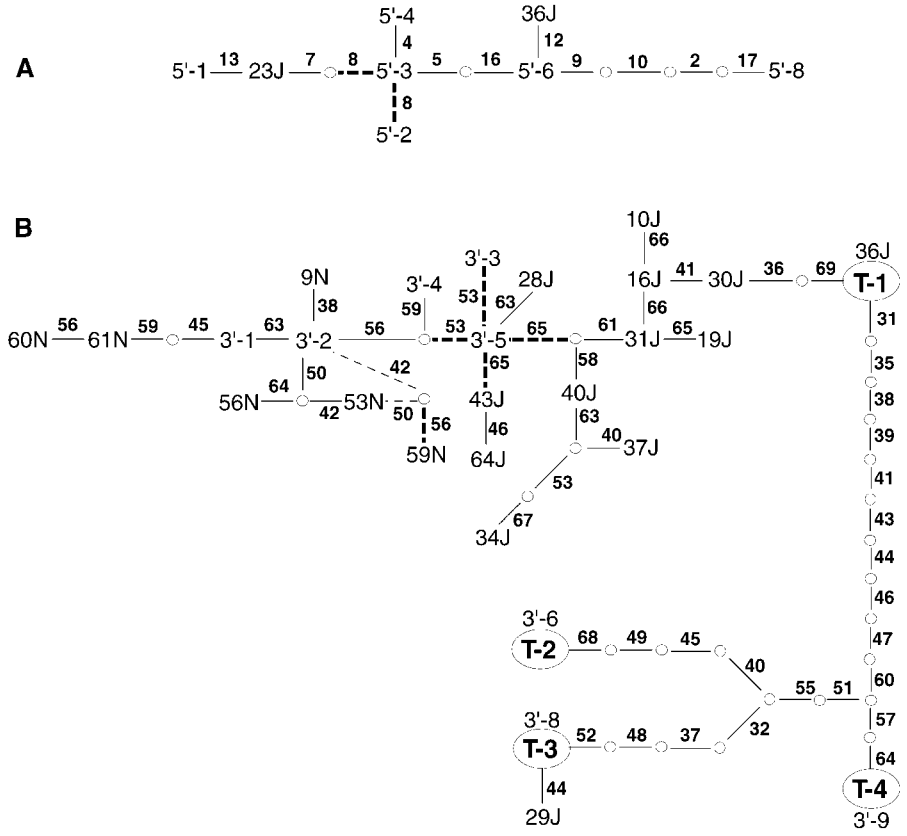


FIGURE 8.—(A) The topology that results when only sites 1–18 are retained in the peeled tree shown in Figure 6; (B) the topology that results when only sites 36–69 are retained in the peeled tree shown in Figure 6. Thick dashed lines indicate those portions of these haplotype networks that are topologically incompatible with the networks given in Figures 1 and 2.

only SNPs that would be expected to provide some disequilibrium coverage are those falling into the 5′ and 3′ regions that flank the recombinational hotspot. Obviously, significant disequilibrium is expected only within each of the flanking regions and not between them. Therefore, a single SNP could not provide adequate disequilibrium coverage for even this third of the *LPL* gene. Rather, separate SNPs would be needed for each flanking region.

Given that choosing SNPs nonrandomly by population frequency criteria implies a nonrandom choice by evolutionary history as well, it is far better to use the evolutionary history explicitly. We now address how the cladistic and homoplasy structure estimated in this article can be used for choosing SNPs in the two flanking regions.

Starting with the 3′ end, there are four major clades (the four termini in Figures 5 or 6). Given that the 3′ cladistic structure implies that these are old and long-surviving clades, these four clades capture the majority of mutational divergence found in the 3′ end of the gene. Moreover, because the branches interconnecting these four major clades are long, we have great confidence that this portion of the estimated *LPL* haplotype tree is accurate and fully resolved. Cladistic analyses of phenotypic associations do not require that the haplotype tree be estimated in a completely accurate, resolved fashion (TEMPLETON and SING 1993), but ambiguities in the estimated haplotype tree structure do weaken the power of this evolutionary approach. Therefore, SNPs in the 3′ region should be chosen from mutations occurring on the long branches that discriminate among these four major clades. In performing this task, we will focus upon those polymorphic sites that display no homoplasy at all, even before the inferred recombination and gene conversion events have been peeled off. From Figure 3, 12 polymorphic sites exist that show no homoplasy at all. The simplest explanation for why a

FIGURE 7.—The estimated recombinant clades. Each of the 29 recombination events at the MDECODE website (http://mdecode.umich.edu/) is indicated with an arrow showing the immediate product of recombination. The number in italics next to the arrow corresponds to the number of the recombination event given at the MDECODE website, with the numbers in parentheses below the recombination number indicating the inferred site interval in which the crossover event occurred. After recombination, mutations can accumulate and new haplotypes are created, as indicated by the haplotype networks attached to the original recombinant type. The layout of these networks is the same as that given in Figure 3. A box in the recombinant 14 clade encloses the set of variable sites associated with the single statistically significant gene conversion event. For ease of display, node m of recombinant clade 14 is shown twice in the figure, with a broken dotted line connecting these two occurrences. No mutations or recombination events are associated with this dotted line.

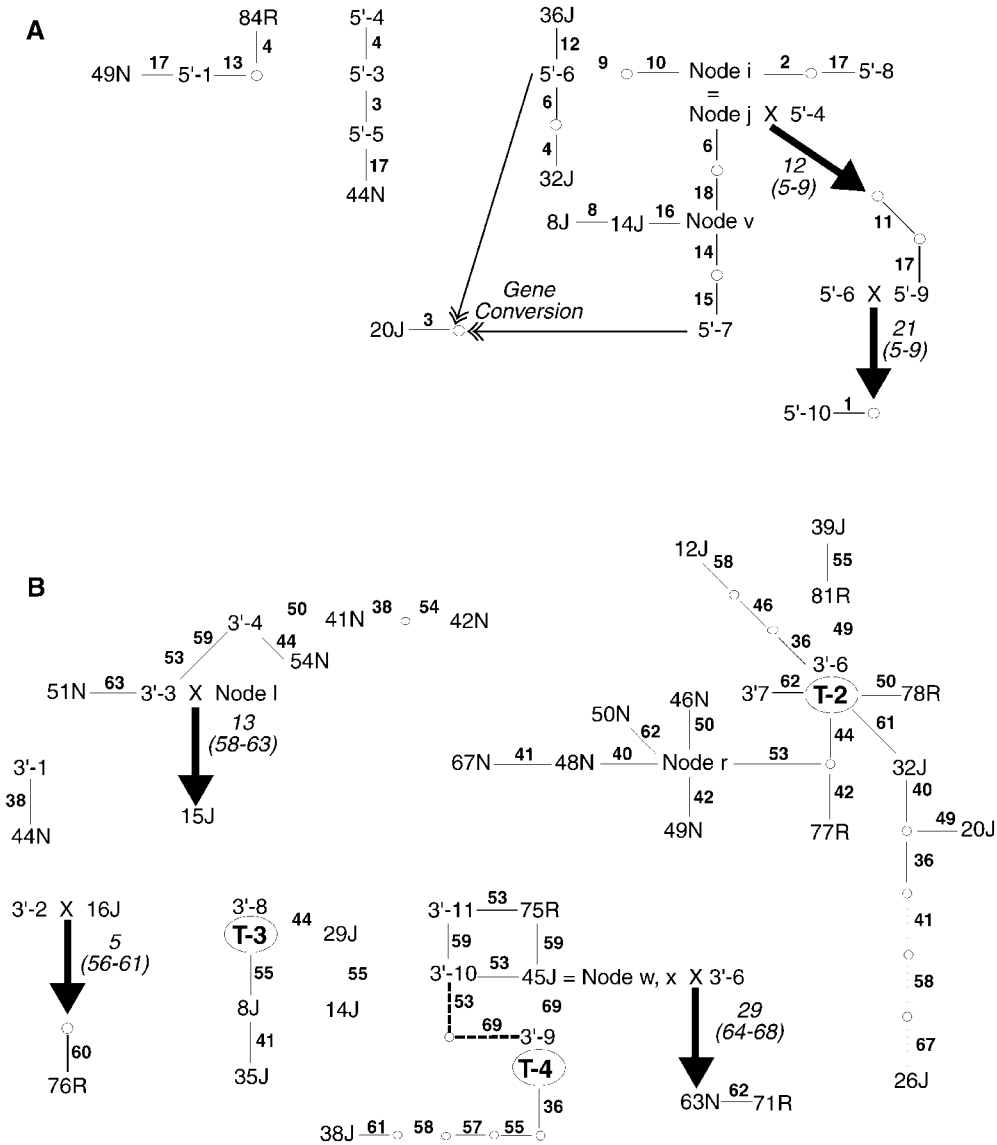


FIGURE 9.—(A) The topology that results when only sites 1–18 are retained in the recombinant clades shown in Figure 7; (B) the topology that results when only sites 36–69 are retained in the recombinant clades shown in Figure 7. Thick dashed lines indicate those portions of these haplotype networks that are topologically incompatible with the networks given in Figures 4 and 5.

site would show no homoplasy is that it is a recent mutation that has had insufficient time and is sufficiently rare to make it unlikely to have been involved in a recombination or gene conversion event. Three of the 12 sites not affected by homoplasy fit this pattern (sites 12, 22, and 24), but 9 of these sites with no homoplasy are in the long branches that mark the four major 3′ clades (sites 32, 37, 39, 43, 47, 48, 51, 52, and 54). Hence, most of the sites with no homoplasy are old and common. Their lack of homoplasy is therefore attributed to a lack of recombination (or a lack of evolutionary persistence of any recombinants), and they are clean cladistic markers for the four major clades.

Three of these nine SNPs (at variable sites 39, 43, and 47) unambiguously mark the oldest and most extensive divergence found within this gene region, the distinction between terminus 1 *vs.* termini 2, 3, and 4. Another important consideration in choosing among these three candidate SNPs is whether or not they are at highly mutable sites. Highly mutable sites should tend to show high levels of polymorphism, which makes them highly informative of their direct phenotypic effects and useful in defining additional haplotype variation when used in conjunction with other nearby variable sites. However, highly mutable sites are not ideal as single-site markers of association (the primary purpose of choosing SNPs in this context) because they could show complicated patterns of disequilibrium due to the fact that identity by state may not reflect identity by descent if multiple mutational events occurred. Our demonstration that mutable sites are strongly associated with increased homoplasy indicates that this possibility cannot be ignored. Given that almost half of the SNPs in the sequenced portion of the *LPL* gene come from the highly mutable classes, this consideration imposes another constraint upon choosing SNPs for disequilibrium mapping. For the particular task of choosing among the three SNPs to discriminate T-1 from the remaining 3′ termini, the

TABLE 4

The distribution of homoplasy across highly mutable and non-highly-mutable sites after the effects of all detected recombination and gene conversion events have been removed

Region	Type of site	0–1 homoplasies	≥2 homoplasies
5′ and 3′ flanking regions	Highly mutable	11	14
	Non-highly mutable	22	5
Recombinational hotspot	Highly mutable	3	7
	Non-highly mutable	5	2
All	Highly mutable	14	21
	Non-highly mutable	27	7

SNP at site 39 is at a highly mutable site and should be excluded, leaving two candidates (the SNPs at sites 44 and 57). To discriminate among the other termini, the SNP at site 51 unambiguously discriminates between terminus 4 and the node leading to termini 2 and 3 and is not at a highly mutable site. Finally, any one of the SNPs at sites 32, 37, 48, or 52 unambiguously discriminate between T-2 and T-3, but the SNP at site 37 would be the best because it is not in one of the highly mutable categories. Hence, three SNPs (at sites 43 or 47, site 51, and site 37) are needed to identify the four major 3′ clades that represent the oldest and most extensive genetic diversity within the region sequenced.

The 5′ end consists of evolutionarily closer haplotypes, and there are no sites totally lacking homoplasy within this end with respect to the total data set (Figure 3). However, much of this homoplasy is due to recombination with the 3′ end, and there are many sites that have no homoplasy within the 5′ end (Figures 4 and 9A). Hence, the 5′ sites with no homoplasy in the 5′ tree can be used to mark the major sources of haplotype divergence in the 5′ end, but should always be used in conjunction with the minimal set of three 3′ markers identified above because all sites show apparent homoplasy in the total data set due to recombination. The single branch in Figure 4 that captures most of the 5′ variation is the branch between the 5′-3 and 5′-6 haplotype sets. To the left of this branch are a set of closely related 5′ haplotype categories that differ from their nearest neighbors by only a single site, and to the right of this branch is a set of more distantly related 5′ haplotype categories (Figure 4). This branch is marked by two sites, and of those, site 5 shows no 5′ homoplasy and is not highly mutable. Hence, the SNP at site 5 captures most of the 5′ variation. With additional sites, even greater 5′ resolution is possible. The right half of Figure 4 consists of three clusters of 5′ haplotypes (5′-6, 32J, and 36J; 5′-8; and 5′-7, 8J, and 14J), all of which are at least two mutational steps from a common node. Variable sites 9 and 10 show no 5′ homoplasy and are not in highly mutable categories, and either would discriminate the [5′-6, 32J, 36J] cluster from the remaining right-side clades. Similarly, sites 2 and 18 show no 5′ homoplasy, but only site 2 is not in a highly mutable category. Hence, the SNP at site 2 should be used for discriminating between the remaining two right-side clades. In summary, scoring three SNPs at sites 5, 9 or 10, and 2 would mark the major sources of haplotype divergence in the 5′ region of the sequenced portion of this gene.

These considerations indicate that a single SNP cannot provide adequate disequilibrium coverage for even this third of the *LPL* gene and that at least four to six SNPs are needed to mark most of the mutational divergence found in the regions flanking the recombinational hotspot. Note that choosing these four to six SNPs required detailed analysis of the entire 9.7-kb region with respect to cladistic structure, recombination, and mutation. This reinforces the conclusion of Clark *et al.* (1998) that four random markers within 9.7 kb of the *LPL* gene would not be a reliable method of detection of nearby causal variation. But with choices informed by an estimate of evolutionary history, highly mutable sites and recombinational events, four SNPs (site 5 plus three 3′ markers) would capture most of the mutational divergence found in this region, and with just two more SNPs in the 5′ region even greater resolution is possible.

However, even these cladistically informed choices may not always be meaningful because of genetic differences among populations. The four to six SNPs indicated above are informative for the three specific populations we have sampled, but these SNPs may not be polymorphic in all human populations, and hence may not be informative for all human populations. Evolutionary considerations are also relevant to the problem of choosing SNPs in light of potential genetic heterogeneity among populations. Detailed phylogeographic analyses of human mitochondrial DNA, Y-linked DNA, and autosomal DNA all indicate that the primary pattern in recent human evolution has been one of gene flow constrained by isolation by distance (Templeton 1998). Under isolation by distance, it takes time for a new mutation to spread geographically, with the older mutations tending to be the most geographically widespread (Templeton *et al.* 1995). With regard to *LPL*, we have

detected old evolutionary structure in the 3′, but not 5′, flanking region. In particular, the oldest branch is the one connecting the T1 terminus to the other three termini (Figure 5). As expected under an isolation-by-distance model, haplotypes on either end of this oldest branch are found in all three populations sampled. Hence, the SNP chosen to mark this branch (the SNP at either sites 44 or 57) is useful in any of the three populations sampled, and under the evolutionary predictions of the isolation-by-distance model, this SNP should be useful for almost all human populations. The other long branches in the 3′ region are also likely to be old (although not as old as the longest branch that connects to the out-group), and hence should also tend to identify geographically widespread clusters of haplotypes. Indeed, the haplotypes associated with T2 and T4 are also found in all three populations sampled. Only the haplotypes associated with T3 (Figure 5) are restricted to a single population (the Jackson sample). Hence, the remaining 3′ SNPs identified above are also expected to be useful in most human populations, although not as many as the SNP marking the longest branch.

The evolutionary situation is quite different at the 5′ end, which does not reveal any deep cladistic structure. Hence, the SNPs in this part of the *LPL* gene may have little utility beyond the populations actually sampled. For example, as discussed above, the single most informative SNP in the 5′ region for the three populations sampled is the SNP at site 5 (Figure 4). The haplotypes found to the left of the SNP at site 5 in Figure 4 are found in all three populations sampled, but the haplotypes to the right of this SNP in Figure 4 are found only in the Jackson population. Hence, if we had sampled just the Rochester and North Karelian populations and not the Jackson population, the SNP at site 5 would have been uninformative. On the basis of these evolutionary considerations, we conclude that the SNPs identified in the 3′ flanking region are likely to have general utility for most human populations, whereas the SNPs in the 5′ flanking region are likely to be informative only for specific populations.

The ability to identify such highly informative SNPs for particular populations illustrates the utility that simultaneous estimation of recombinational, mutational, and evolutionary structure can play in human genetic epidemiology. Moreover, the ability to identify SNPs that may have general utility in most human populations is another important application of evolutionary features that can arise from such a cladistic analysis of haplotype variation. However, as the 5′ flanking region shows, the evolutionary analysis may also indicate that it may be unlikely to identify a SNP that will be informative for most human populations. When dealing with such DNA regions, it would be better to obtain sequence data for the entire region in the populations of interest rather than relying upon one or a few SNPs.

We thank Jody Hey and two anonymous reviewers for their excellent suggestions for improving an earlier draft of this article. This work was supported by the National Heart, Blood, and Lung Institute grants HL39107, HL58238, HL58239, and HL58240.

LITERATURE CITED

Agarwal, S. K., L. V. Debelenko, M. B. Kester, S. C. Guru, P. Manickam *et al.*, 1998 Analysis of recurrent germline mutations in the *Men1* gene encountered in apparently unrelated families. Hum. Mutat. **12:** 75–82.

Boerwinkle, E., D. L. Ellsworth, D. M. Hallman and A. Biddinger, 1996 Genetic analysis of atherosclerosis—a research paradigm for the common chronic diseases. Hum. Mol. Genet. **5:** 1405–1410.

Castelloe, J., and A. R. Templeton, 1994 Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogenet. Evol. **3:** 102–113.

Chakravarti, A., K. H. Buetow, S. E. Antonarakis, P. G. Weber, C. D. Boehm *et al.*, 1984 Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. **36:** 1239–1258.

Clark, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111–122.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. Am. J. Hum. Gen. **63:** 595–612.

Collins, F. S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland *et al.*, 1998 New goals for the US Human Genome Project—1998–2003. Science **282:** 682–689.

Crandall, K. A., and A. R. Templeton, 1999 Statistical approaches to detecting recombination, pp. 153–176 in *The Evolution of HIV*, edited by K. A. Crandall. The Johns Hopkins University Press, Baltimore.

Jones, P. A., W. M. Rideout, J. C. Shen, C. H. Spruck and Y. C. Tsai, 1992 Methylation, mutation and cancer. Bioessays **14:** 33–36.

Keavney, B., C. A. McKenzie, J. M. C. Connell, C. Julier, P. J. Ratcliffe *et al.*, 1998 Measured haplotype analysis of the *Angiotensin-I Converting Enzyme* gene. Hum. Mol. Genet. **7:** 1745–1751.

Kimura, M., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics **75:** 199–212.

Krawczak, M., and D. N. Cooper, 1991 Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum. Genet. **86:** 425–441.

Krawczak, M., P. H. Reitsma and D. N. Cooper, 1995 The mutational demography of protein C deficiency. Hum. Genet. **96:** 142–146.

Long, A. D., R. F. Lyman, C. H. Langley and T. F. C. Mackay, 1998 Two sites in the *delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. Genetics **149:** 999–1017.

Lyman, R. F., and T. F. C. Mackay, 1998 Candidate quantitative trait loci and naturally occurring phenotypic variation for bristle number in *Drosophila melanogaster*—the *delta-hairless* gene region. Genetics **149:** 983–998.

Magewu, A. N., and P. A. Jones, 1994 Ubiquitous and tenacious methylation of the CpG site in codon 248 of the p53 gene may explain its frequent appearance as a mutational hot spot in human cancer. Mol. Cell. Biol. **14:** 4225–4232.

Markham, R. B., D. H. Schwartz, A. Templeton, J. B. Margolick, H. Farzadegan *et al.*, 1996 Selective transmission of human immunodeficiency virus type 1 variants to SCID mice reconstituted with human peripheral blood mononuclear cells. J. Virol. **70:** 6947–6954.

Nakagawa, H., K. Koyama, Y. Miyoshi, H. Ando, S. Baba *et al.*, 1998 Nine novel germline mutations of Stk11 in ten families with Peutz-Jeghers syndrome. Hum. Genet. **103:** 168–172.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson *et al.*, 1998 DNA sequence diversity in a 9.7-kb region of the human *Lipoprotein Lipase* gene. Nat. Genet. **19:** 233–240.

Rao, D. C., R. C. Elston, L. H. Kuller, M. Feinleib, C. Carter and

R. Havlik (Editors), 1984 *Genetic Epidemiology of Coronary Heart Disease, Past, Present, and Future: Proceedings of a Workshop Held in St. Louis, Missouri August 10–12, 1983*. Alan R. Liss, New York.

Reiss, J., D. N. Cooper, J. Bal, R. Slomski, G. R. Cutting and M. Krawczak, 1991 Discrimination between recurrent mutation and identity by descent: application to point mutations in exon 11 of the cystic fibrosis (CFTR) gene. Hum. Genet. **87:** 457–461.

Rideout, W. M., 3rd, G. A. Coetzee, A. F. Olumi and P. A. Jones, 1990 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science **249:** 1288–1290.

Sing, C. F., and M. Skolnick (Editors), 1979 *Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease*. Alan R. Liss, New York.

Sing, C. F., M. B. Haviland, A. R. Templeton and S. L. Reilly, 1995 Alternative genetic strategies for predicting risk of atherosclerosis, pp. 638–644 in *Atherosclerosis X: Excerpta Medica International Congress Series*, edited by F. P. Woodford, J. Davignon and A. D. Sniderman. Elsevier Science, Amsterdam.

Swofford, D., 1997 PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, MA.

Templeton, A. R., 1983 Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37:** 221–244.

Templeton, A. R., 1987 Nonparametric phylogenetic inference from restriction cleavage sites. Mol. Biol. Evol. **4:** 315–319.

Templeton, A. R., 1996 Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome, pp. 259–283 in *Variation in the Human Genome*, edited by G. Cardew. John Wiley & Sons, Chichester.

Templeton, A. R., 1998 Human races: a genetic and evolutionary perspective. Am. Anthropol. **100:** 632–650.

Templeton, A. R., and C. F. Sing, 1993 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics **134:** 659–669.

Templeton, A. R., E. Boerwinkle and C. F. Sing, 1987 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics **117:** 343–351.

Templeton, A. R., C. F. Sing, A. Kessling and S. Humphries, 1988 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics **120:** 1145–1154.

Templeton, A. R., K. A. Crandall and C. F. Sing, 1992 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics **132:** 619–633.

Templeton, A. R., E. Routman and C. Phillips, 1995 Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. Genetics **140:** 767–782.

Templeton, A. R., A. G. Clark, K. M. Weiss, D. A. Nickerson, E. Boerwinkle *et al.*, 2000 Recombinational and mutational hotspots within the human *Lipoprotein Lipase* gene. Am. J. Hum. Genet. **66:** 69–83.

Todorova, A., and G. A. Danieli, 1997 Large majority of single-nucleotide mutations along the dystrophin gene can be explained by more than one mechanism of mutagenesis. Hum. Mutat. **9:** 537–547.

Tunstall-Pedoe, H., K. Kuulasmaa, P. Amouyel, D. Arveiler, A. M. Rajakangas *et al.*, 1994 Myocardial infarction and coronary deaths in the World Health Organization MONICA Project: registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. Circulation **90:** 583–612.

Tvrdik, T., S. Marcus, S. M. Hou, S. Falt, P. Noori *et al.*, 1998 Molecular characterization of two deletion events involving Alu-sequences, one novel base substitution and two tentative hotspot mutations in the *Hypoxanthine Phosphoribosyltransferase* (*Hprt*) gene in five patients with Lesch-Nyhan syndrome. Hum. Genet. **103:** 311–318.

Communicating editor: J. Hey