DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene

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

Lipoprotein lipase plays a central role in lipid metabolism and the gene that encodes this enzyme (*LPL*) is a candidate susceptibility gene for cardiovascular disease. Here we report the complete sequence of a fraction of the *LPL* gene for 71 individuals (142 chromosomes) from three populations that may have different histories affecting the organization of the sequence variation. Eighty-eight sites in this 9.7 kb vary among individuals from these three populations. Of these, 79 were single nucleotide substitutions and 9 sites involved insertion-deletion variations. The average nucleotide diversity across the region was 0.2% (or on average 1 variable site every 500 bp). At 34 of these sites, the variation was found in only one of the populations, reflecting the differing population and mutational histories. If *LPL* is a typical human gene, the pattern of sequence variation that exists in introns as well as exons, even for the small number of samples considered here, will present challenges for the identification of sites, or combinations of sites, that influence variation in risk of disease in the population at large.

Introduction

The Human Genome Project is progressing towards the identification of an entire human reference sequence[1,2]. As our knowledge of this sequence grows, so will our need to explore the levels of natural variation relative to this reference sequence, among individuals and between human populations[3–5]. Knowledge of DNA sequence variation among individuals will play an important role in understanding the history and organization of the human genome as well as the aetiology of the complex links between genotypic variation and variation in traits that are determinants of human health and disease susceptibility and resistance[6].

In the absence of complete sequence data, much emphasis has been placed on how to infer the location of causal variants in genes that influence human health and disease susceptibility[3,7,8]. Even with candidate genes, we have relied on our ability to detect the presence of disease susceptibility mutations at a locus by using nearby markers chosen for their regional chromosomal location and informativeness[7,8]. The pattern and magnitude of linkage disequilibrium among such markers and between markers and disease susceptibility loci can affect the efficacy of association studies in populations. Complete sequence data provides the ultimate level of resolution for characterizing these patterns of disequilibrium.

There is considerable literature on the use of polymorphic markers in genetic studies[7–9]. Markers are used in linkage studies to search for candidate genes[7] and in association studies to identify the functional allelic variations in candidate genes that influence inter-individual variation in risk factor levels or risk of developing disease[7–9]. A major weakness of linkage and association studies is that they rely on incomplete data about sequence variation, so that the distribution of linkage disequilibrium among pairs of nucleotide sites across a gene, and the consistency of that linkage disequilibrium (or lack thereof), cannot be adequately assessed.

To gain a greater understanding of the type and amount of human DNA sequence variation, its history, and the relationship of its contemporary organization to the continuous distribution of measures of human health among individuals in the population at large (such as blood pressure or plasma cholesterol levels), we are undertaking a multidisciplinary and multinational project known as MDECODE (*M*olecular *D*iversity and *E*pidemiology of *Co*mmon *D*is*e*ase). In this report, we describe a systematic survey of the natural sequence variation in approximately one-third of the *LPL* gene. Lipoprotein lipase plays an important role in human lipid metabolism, and it is one of three related human lipases, together with the hepatic and pancreatic lipases[10]. Human *LPL* encodes a protein of 448 amino acids that has several well-characterized functional domains including the APOC-II binding site, the heparin-binding clusters used to localize LPL to the endothelial wall and the domains contributing to the active site[10–12]. Similar to many of the genes involved in human lipid metabolism, a number of functional DNA sequence variations in *LPL* have been identified in individuals expressing clinically abnormal lipid profiles that are often associated with coronary artery disease, atherosclerosis and/or obesity[10,13–15]. Little is known about the spectrum of DNA sequence variation in the general population for *LPL* (ref. 10), or any of the genes involved in human lipid metabolism.

The contiguous DNA sequence data presented in this report represents one of the largest bodies of sequence data on multiple individuals for the same gene, obtained from the population at

[1]*Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, Washington 98125, USA.* [2]*Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania 16802, USA.* [3]*Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA.* [4]*Preventive Cardiology, University of Mississippi Medical Center, Jackson, Mississippi 39216, USA.* [5]*National Public Health Institute, Department of Epidemiology and Health Promotion, Helsinki, Finland.* [6]*Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77225, USA.* [7]*Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. Correspondence should be addressed to D.A.N. e-mail: debnick@u.washington.edu*

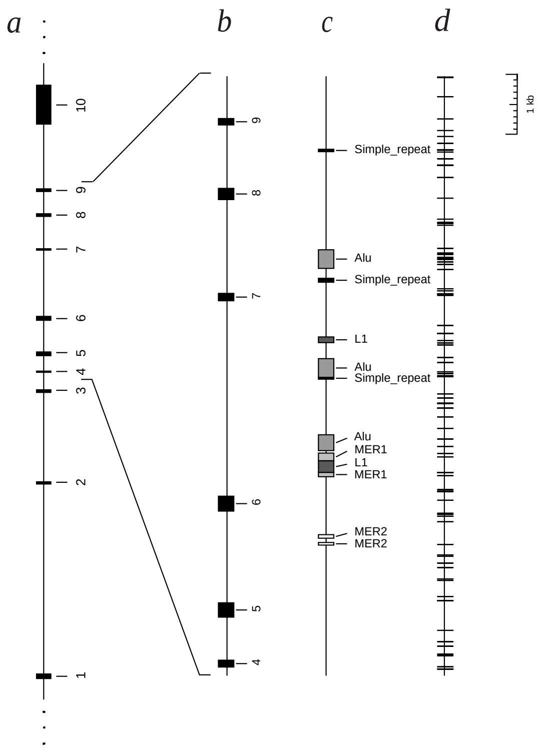


Fig. 1 A schematic diagram of the human *LPL* gene. ***a***, The entire gene spans approximately 30 kb of genomic DNA. ***b***, The region sequenced in 71 individuals spans 9.7 kb of the *LPL* gene from the 3′ end of intron 3 to 5′ end of intron 9. ***c***, The types and distribution of repeat sequences within the target region identified by RepeatMasker. ***d***, The location of the 88 DNA variants identified by sequencing the 71 individuals across the 9.7-kb region.

large. Individuals were selected for sequencing without regard to disease status from three populations: African-Americans (Jackson, Mississippi); Europeans (North Karelia, Finland); and European-Americans (Rochester, Minnesota). These populations were chosen for studies of the genetics of cardiovascular disease risk factors because they represent a range of low (Rochester), intermediate (Jackson) and high (North Karelia) coronary heart disease (CHD) prevalence rates. The availability of sequence data will add a new dimension of information to our ongoing efforts to understand the role of genetic variation in determining the distribution of CHD within and between these populations.

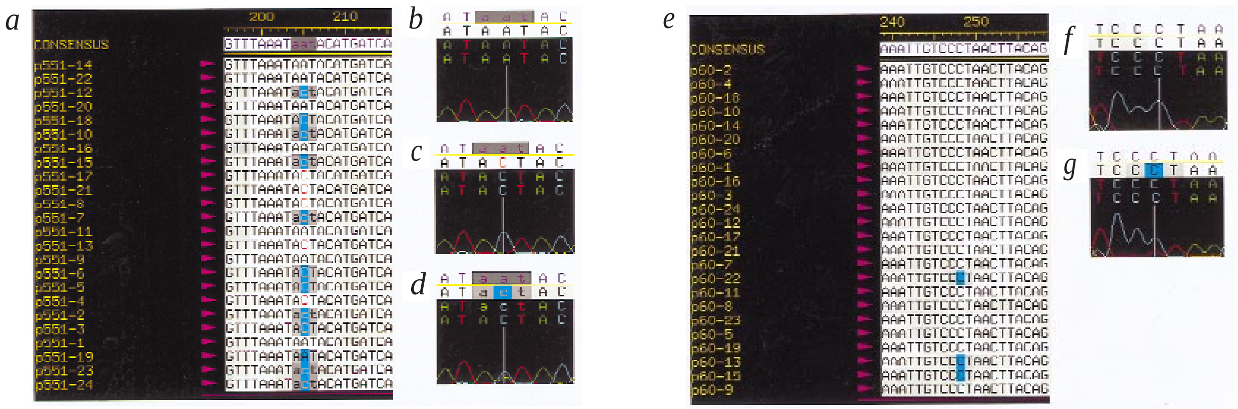
We present 9.7 kb of *LPL* sequence for 142 chromosomes (71 individuals sampled from the three populations; 24 from Jackson, 24 from North Karelia and 23 from Rochester). Although studies of phenotypic variation with larger sample sizes will be done as an extension of these findings, our object in this paper is to report on the type and amount of sequence variation and the distribution of that variation within and among three human populations that differ in CHD prevalence. In a separate report[16], we present an analysis of the organization of the observed variation into haplotypes and test hypotheses about the historic origins and evolutionary forces that most likely shaped the contemporary genomic organization of *LPL*.

Results

High throughput analysis of DNA variations in *LPL*

Human *LPL*, located on chromosome 8p22, has ten exons distributed across approximately 30 kb of genomic DNA (ref. 10; Fig. 1*a*). For this analysis, we focused on the 3′ end of the gene and sequenced a region of 9,734 bp that spans from the 3′ end of intron 3 to the 5′ end of intron 9 (Fig. 1*b*). A contiguous baseline reference sequence for the target region was assembled from existing exon and intron sequences obtained from GenBank, and joined together by sequences obtained from directed primer walking across the gaps. Overall, this 9.7-kb region is typical of most human sequences, with the majority of sequence found in non-coding introns (8,736 bp, approximately 90% of the target sequence), and approximately 10% of the sequenced region (998 bp) involved in encoding a polypeptide (exons 4, 5, 6, 7, 8 and 9). This region also contains a number of common human interspersed repeat sequences (Fig. 1*c*). The positions of all varying sites are indexed with regard to their location on the baseline reference sequence (Fig. 1*d*).

Following amplification, DNA variants were identified by direct fluorescence-based sequencing of PCR products using the PolyPhred program[17]. PolyPhred identifies heterozygous sites by scanning the assembled sequence traces for: (i) the presence of a drop in normalized fluorescence peak height at a position when compared to the respective peak height for all individuals that are homozygous at the position, and (ii) the presence of another base (a second peak) that accompanies the drop in fluorescence peak height. Representative examples of 2 of the 81 new variants identified by PolyPhred among the *LPL* sequences are shown (Fig. 2).



```
CONSENSUS   GTTTAAATaatACATGATCA
p551-14     GTTTAAATAATACATGATCA
p551-22     GTTTAAATAATACATGATCA
p551-12     GTTTAAATactACATGATCA
p551-20     GTTTAAATAATACATGATCA
p551-18     GTTTAAATACTACATGATCA
p551-10     GTTTAAATactACATGATCA
p551-16     GTTTAAATAATACATGATCA
p551-15     GTTTAAATactACATGATCA
p551-17     GTTTAAATACTACATGATCA
p551-21     GTTTAAATACTACATGATCA
p551-8      GTTTAAATACTACATGATCA
p551-7      GTTTAAATactACATGATCA
p551-11     GTTTAAATAATACATGATCA
p551-13     GTTTAAATACTACATGATCA
p551-9      GTTTAAATAATACATGATCA
p551-6      GTTTAAATACTACATGATCA
p551-5      GTTTAAATACTACATGATCA
p551-4      GTTTAAATACTACATGATCA
p551-2      GTTTAAATactACATGATCA
p551-3      GTTTAAATACTACATGATCA
p551-1      GTTTAAATAATACATGATCA
p551-19     GTTTAAATAATACATGATCA
p551-23     GTTTAAATactACATGATCA
p551-24     GTTTAAATactACATGATCA
```

```
CONSENSUS   AAATTGTCCCTAACTTACAG
p60-2       AAATTGTCCCTAACTTACAG
p60-4       AAATTGTCCCTAACTTACAG
p60-18      AAATTGTCCCTAACTTACAG
p60-10      AAATTGTCCCTAACTTACAG
p60-14      AAATTGTCCCTAACTTACAG
p60-20      AAATTGTCCCTAACTTACAG
p60-6       AAATTGTCCCTAACTTACAG
p60-1       AAATTGTCCCTAACTTACAG
p60-16      AAATTGTCCCTAACTTACAG
p60-3       AAATTGTCCCTAACTTACAG
p60-24      AAATTGTCCCTAACTTACAG
p60-12      AAATTGTCCCTAACTTACAG
p60-17      AAATTGTCCCTAACTTACAG
p60-21      AAATTGTCCCTAACTTACAG
p60-7       AAATTGTCCCTAACTTACAG
p60-22      AAATTGTCCCTAACTTACAG
p60-11      AAATTGTCCCTAACTTACAG
p60-8       AAATTGTCCCTAACTTACAG
p60-23      AAATTGTCCCTAACTTACAG
p60-5       AAATTGTCCCTAACTTACAG
p60-19      AAATTGTCCCTAACTTACAG
p60-13      AAATTGTCCCTAACTTACAG
p60-15      AAATTGTCCCTAACTTACAG
p60-9       AAATTGTCCCTAACTTACAG
```

Fig. 2 Novel DNA variants identified in *LPL* using sequence analysis. ***a***, Consed view of a frequent DNA polymorphism identified at position 205 among the assembled sequences (complement of position 8852 in the target sequence). The homozygous nucleotide selected for the consensus sequence is colour-coded by black lettering, while the alternative homozygote at this position is identified as a high-quality base mismatch (gold). Heterozygotes are tagged blue by the PolyPhred program. ***b***, Trace view of the homozygous nucleotide selected for the consensus (A peak). ***c***, Trace view of the homozygote for the alternative allele (C peak) and ***d***, heterozygous position containing both the C and A peaks. ***e***, An example of a less frequent DNA variant identified at position 248 among the assembled traces (position 2131 in the target sequence). ***f***, Trace view of a homozygous sequence *versus* ***g***, a heterozygous sequence.

Table 1 • Sequence variants identified in *LPL*

			Relative frequency[a]			
			Jackson	North Karelia	Rochester	
			MS	Finland	MN	Total
Position[b]	Variant[c]	Coverage[d]	(n=24)	(n=24)	(n=23)	(n=71)
106	C→A	2.0[g]	0.96	0.96	0.87	0.93
110	A→C	1.6[g]	0.92	1.00	1.00	0.97
145[e]	G→A	2.3[g]	0.94	0.94	0.93	0.94
325	T→C	3.5[g]	0.94	0.98	0.89	0.94
343	(TG)3→(TG)4	3.3[g]	0.60	0.96	0.86	0.81
479	T→C	2.7[g]	0.83	1.00	1.00	0.94
551	(A)3→(A)2	2.8[g]	0.83	0.40	0.46	0.56
736	T→C	2.4[g]	0.63	0.40	0.46	0.49
1216	C→G	2.5[g]	0.74	0.96	0.78	0.83
1220	T→C	2.5[g]	0.75	0.96	0.78	0.83
1286	C→T	2.4[g]	0.96	0.94	0.77	0.89
1547	A→C	1.7[g]	0.96	1.00	1.00	0.99
1571	C→G	2.4[g]	0.85	0.40	0.50	0.58
1756	A→T	1.8[g]	0.98	1.00	1.00	0.99
1828	C→G	2.4[g]	0.94	1.00	1.00	0.98
1939	A→G	3.3[g]	0.94	1.00	1.00	0.98
1961	A→T	3.3[g]	0.96	1.00	1.00	0.99
2131	C→T	2.6[g]	0.69	0.94	0.78	0.80
2500	G→A	2.0[g]	0.88	0.90	0.78	0.85
2589	C→G	2.0	0.98	1.00	1.00	0.99
2619	A→G	2.3	0.85	1.00	1.00	0.95
2638	C→T	2.4	0.98	1.00	1.00	0.99
2849[e]	A→G	1.4	1.00	1.00	0.98	0.99
2987	T→G	2.1[g]	0.56	0.94	0.76	0.75
2996	C→A	2.1[g]	0.92	0.77	0.85	0.85
3022	G→A	2.7[g]	0.96	0.94	0.78	0.89
3248	C→G	2.8[g]	0.94	1.00	1.00	0.98
3290	(T)7→(T)8	2.4[g]	0.96	0.94	0.80	0.90
3297	(A)4→(A)5	2.4[g]	0.94	1.00	1.00	0.98
3553	C→T	2.9[g]	0.96	1.00	1.00	0.99
3609	T→C	2.4[g]	0.87	0.00	0.00	0.96
3723	T→C	2.0[g]	0.87	0.75	0.83	0.81
3843	G→A	1.9[g]	0.91	0.79	0.85	0.85
4016	C→G	2.1[g]	0.91	1.00	1.00	0.97
4202	G→A	2.0[g]	0.96	1.00	1.00	0.99
4343	A→T	2.5[g]	0.50	0.69	0.67	0.62
4346	C→G	1.7[g]	0.92	0.85	0.91	0.89
4418	C→T	2.4[g]	0.48	0.83	0.89	0.73
4426	T→C	2.2[g]	0.83	1.00	1.00	0.94
4509	T→C	2.7[g]	0.94	0.40	0.45	0.60
4576	A→T	2.8[g]	0.96	0.92	0.95	0.94
4823	(ATTT)8	1.8[g]	0.17	0.02	0.00	0.06
	(ATTT)9[f]		0.40	0.52	0.43	0.45
	(ATTT)10		0.13	0.13	0.24	0.16
	(ATTT)11		0.23	0.27	0.22	0.24
	(ATTT)12		0.06	0.06	0.11	0.08
	(ATTT)13		0.02	0.00	0.00	0.01
4872	G→A	1.5[g]	0.26	0.27	0.35	0.29
4907	C→A	1.5[g]	1.00	1.00	0.98	0.99
4935	T→C	2.0[g]	0.75	0.74	0.65	0.71
5085	G→A	1.7[g]	0.85	0.00	0.00	0.95
5168	T→C	3.1[g]	0.64	0.65	0.65	0.65
5371	G→A	5.3[g]	0.96	1.00	1.00	0.99
5395	(A)8→(A)9	3.4[g]	0.63	0.73	0.65	0.67
5441	T→C	2.6[g]	0.91	0.89	0.74	0.85
5554	A→C	2.7[g]	0.63	0.75	0.65	0.68
5560	A→G	2.6[g]	0.00	0.06	0.02	0.03
5687	T→C	3.1[g]	0.60	0.74	0.65	0.66
6176[e]	G→A	1.5[g]	1.00	1.00	0.98	0.99
6196[e]	C→T	1.8[g]	0.98	1.00	1.00	0.99
6203[e]	A→G	1.4[g]	0.98	1.00	1.00	0.99
6250	C→T	1.7[g]	0.35	0.19[h]	0.33	0.29
6283	C→T	1.7[g]	1.00	1.00	0.95[h]	0.99
6595	G→C	1.0	0.83	0.82	0.73	0.80
6678	T→G	1.5[g]	0.62	0.73	0.61	0.66
6718	A→G	1.6	0.60	0.73	0.65	0.66
6760	G→A	1.6	0.96	1.00	1.00	0.99
6772	A→G	1.6	0.83	1.00	1.00	0.94
6793	T→G	1.6[g]	1.00	0.98	0.98	0.99
6840	T→C	1.7[g]	1.00	0.98	1.00	0.99
6863	C→T	1.6[g]	0.88	0.83	0.74	0.82
6939-6942	delAAAT	2.0[g]	1.00	0.90	0.98	0.96
7315	G→C	1.8[g]	0.67	0.83	0.75	0.75
7344	A→G	1.2[g]	0.83	1.00	1.00	0.94
7360	A→G	1.6[g]	0.73[h]	0.08[h]	0.37	0.39
7413	T→C	1.9[g]	1.00	0.98	1.00	0.99
7754[e]	A→C	1.4[g]	0.23	0.17	0.24	0.21
8088-8089	insAG	1.5[g]	0.92	0.54	0.52	0.66
8089	G→T	1.5[g]	0.96	0.90	0.91[h]	0.92
8285	C→G	2.3[g]	0.85	1.00	1.00	0.95
8292	A→C	2.2[g]	0.90	0.75[h]	0.86	0.84
8393	T→G	2.9[g]	0.60	0.73	0.63	0.65
8502	G→C	2.6[g]	0.96	1.00	1.00	0.99
8533	A→C	1.7[g]	0.57	0.74	0.63	0.65
8537	A→C	1.7[g]	0.98	0.98	0.93	0.96
8538	(A)3→(A)2	1.7[g]	0.16	0.28	0.20	0.21
8644	T→C	2.2[g]	0.94	0.88	0.91[h]	0.91
8755	G→A	1.7[g]	0.67	0.27	0.37	0.44
8852	T→G	2.2[g]	0.54	0.73	0.63	0.64
9040[e]	C→G	1.7[g]	1.00	1.00	0.96[h]	0.99
9402	A→G	2.1[g]	0.94	1.00	1.00	0.98
9712	G→A	1.2[g]	0.83	0.81	0.72	0.79
9721	G→A	1.2[g]	0.63	0.81	0.67	0.70

[a] Relative frequency of the baseline allele for base substitutions and diallelic insertion/deletion variations, and as reported for the tetranucleotide repeat. [b] Position in the baseline sequence. [c] Substitution and diallelic insertion/deletion variants are reported as the allele in the baseline sequence(alternative allele). [d] Average number of times each site was covered with a confirming sequence read. [e] DNA variant in coding region. [f] Allele for tetranucleotide repeat found in the baseline sequence. [g] Indicates opposite strand confirmation. [h] Chi-square tests of Hardy-Weinberg equilibrium at *Pr* (0.05 level of statistical significance).

When a frequent variant is identified in a population (Fig. 2*a*), homozygotes for each of the alternative alleles are usually present as high-quality base mismatches (with regard to the consensus sequence; Fig. 2*b*,*c*) as well as a number of heterozygotes containing these alternative alleles (colour-coded blue by the PolyPhred program; Fig. 2*a*,*d*). In comparison, when the relative frequency of a variant is low, they are usually identified by infrequent heterozygotes among homozygotes (Fig. 2*e*,*f*, homozygote; Fig. 2*g*, heterozygote, note the drop in the C peak and the underlying T peak, alternative allele for this heterozygous position). These DNA variants would be missed without the use of specialized software that can accurately detect heterozygotes among homozygotes, as both base-calling programs (Phred and ABI) applied to these individual samples miscalled the heterozygous sites as homozygous (Fig. 2*e*,*g*; ref. 17).

Nature and distribution of sequence variation in *LPL*

In total, 88 variable sites were identified among 71 individuals across the scanned sequence (Fig. 1*d*). Each variable site has been confirmed by resequencing the PCR products from the same or opposite strand (Table 1; gives the average number of confirmations at each of these sites across the 71 individuals). Single base substitutions were the most frequent form of DNA variation (79 of 88 variants, 90%). Among these sites, transition substitutions were more prevalent (47 of 79, 59%) than transversions (32 of 79, 41%; ref. 18). Insertion/deletion variants occurred less frequently (9 of 88 variants, 10%), which is consistent with other studies of human DNA sequence variation[19]. Furthermore, the majority of insertion/deletion variants (8 of 9) reflect copy number changes in regions known to be highly polymorphic, for example short tracts of polynucleotide (A or T), dinucleotide, or tetranucleotide sequences[20–22].

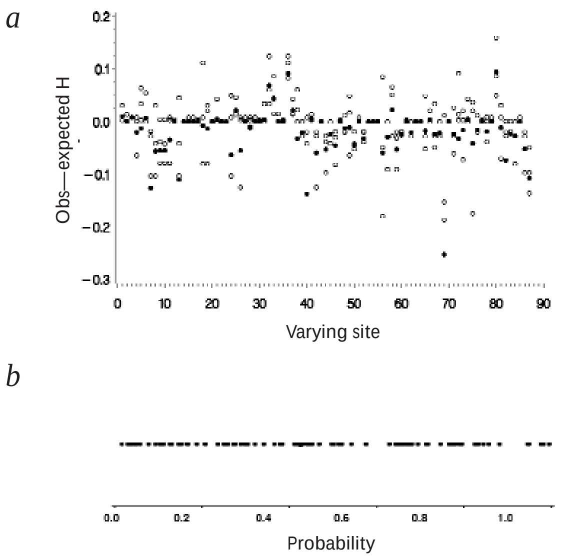


Fig. 3 Comparison of observed and expected heterozygosities. ***a***, Each circle represents the difference between observed heterozygosity at a nucleotide site and the expected Hardy-Weinberg heterozygosity calculated as 1 - Σp_i^2, where p_i is the frequency of each nucleotide in each population. Solid dots are for the data pooled across populations. The solid dots show some tendency toward a deficit in observed heterozygosity due to the Wahlund effect, whereas the open circles show no such trend. ***b***, Uniform distribution of *P*-values for the significance tests of departures from Hardy-Weinberg reflect consistency with the null hypothesis.

Most of the scanned sequence was non-coding, and as expected, the majority of variable sites were found in non-coding regions (81 sites in 8,736 bp). In the coding regions, 7 out of 998 sites varied, and the relative number of varying sites in coding (1 in every 142 bp) and non-coding (1 in 108 bp) regions were similar. This is so despite the lower nucleotide diversity in coding regions due to the lower heterozygosity of the varying sites within the coding regions. Among the sites identified in the coding region, three had variations from the baseline sequence that would lead to amino-acid substitutions (2849: 291Asn→Ser, refs 14,15; 6176: 370Val→Met; 6203: 379Thr→Ala), and one variable site would lead to a premature truncation of the protein by two amino acids (9040: 447Ser→Ter, ref. 23). The remaining variable sites detected in the coding region (positions 145, 6196, 7754) were silent, third-base substitutions (Table 1).

Distribution of the sequence variation

Except for a tetranucleotide repeat (position 4823), all of the variable sites were diallelic in samples from the three populations. The relative genotype frequencies in each sample fitted the Hardy-Weinberg expectation with remarkable fidelity (Fig. 3*a*). Only 6 out of 193 tests had a chi-square significant at the 5% level (Table 1), and the overall distribution of *P*-values was close to uniform (Fig. 3*b*). In this regard, there was no systematic deviation towards an excess or deficit of heterozygosity and no tendency of the deviation to cluster along the sequence (Fig. 3*a*). Any tendency toward over- or under-calling of heterozygous sites would be expected to result in a skew in the pattern of deviations from Hardy-Weinberg.

The mean observed heterozygosity for the 88 varying sites across all samples was 0.20 (Fig. 4). Mean observed heterozygosities for the 88 sites within Jackson, North Karelia and Rochester samples were 0.22, 0.17 and 0.20, respectively. Seventy-eight of the eighty-eight variations identified in *LPL* were found in the Jackson sample, whereas 56 of 88 and 59 of 88 variants were identified in the samples from North Karelia and Rochester, respectively (Fig. 4). Fifty-one of the sequence variations were detected in all three samples. Of the 34 population-specific sites, 27 were found in the sample from Jackson, 2 in the sample from North Karelia and 5 in the sample from Rochester. These findings are not unexpected since it is known that Africans are generally genetically more variable than people from other regions of the world, and also because African-Americans are admixed between Africans and Europeans, yielding an additional source of variability[4]. As expected, many of the variants found in only one sample were rare. Only a few of the variants associated with a single sample had heterozygosities greater than 0.10 (13 of 34), and all of these were associated with the sample from Jackson.

Sequence diversity in *LPL*

In terms of total sequence diversity (including nucleotide substitutions, insertion/deletion and repeat variations), the three populations revealed similar levels of nucleotide diversity (diversity is defined as the expected heterozygosity averaged across all sites). Nucleotide diversity is equivalent to the expected number of nucleotide differences per site between a random pair of chromosomes drawn from the population. Across the three populations, the estimate of nucleotide diversity was 0.002, which is approximately one variant every 500 bp (Table 2). As the majority (five of seven) of the coding region variations were rare in the three samples (Table 1), the sequence diversity of coding-region sequences was lower (0.0005, or 1 in every 2,000 bp) than that found in non-coding regions (0.0021, or 1 in every 476 bp). Several inter-

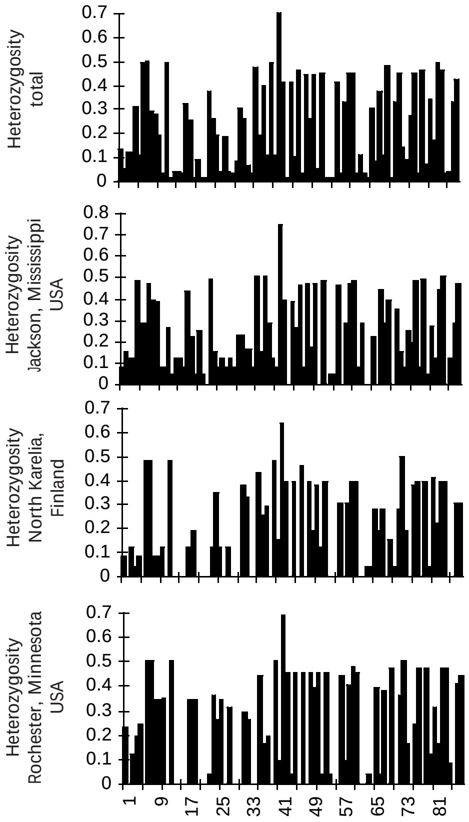


Fig. 4 A plot of the heterozygosity for the 88 variant sites found in the *LPL* sequence. The majority of sites (87 of 88) were diallelic, and there was one tetranucleotide repeat with heterozygosity greater than 65% present among the variants. More than one-half of the variant sites exceed a heterozygosity of 10% among the individuals examined.

Table 2 • Sequence Diversity in *LPL*

Sequence[a] type	No. of sites	Jackson, Mississippi, USA	North Karelia, Finland	Rochester, Minnesota, USA	Total[b]	Sequence Diversity in bp
All Sites	88	0.0020 ± 0.0010	0.0016 ± 0.0008	0.0020 ± 0.0010	0.0020 ± 0.0010	1 in 500 bp
Coding	7	0.0006 ± 0.0005	0.0004 ± 0.0004	0.0007 ± 0.0006	0.0005 ± 0.0005	1 in 2,000 bp
Non-coding	81	0.0022 ± 0.0011	0.0017 ± 0.0008	0.0021 ± 0.0010	0.0021 ± 0.0010	1 in 476 bp
Repeat	26	0.0035 ± 0.0019	0.0027 ± 0.0016	0.0033 ± 0.0018	0.0032 ± 0.0018	1 in 312 bp
Non-repeat	62	0.0017 ± 0.0009	0.0013 ± 0.0007	0.0017 ± 0.0008	0.0017 ± 0.0008	1 in 588 bp

[a]Repeat refers to interdispersed sequences identified using RepeatMasker. Non-repeat includes coding and non-repeat, non-coding sequences. [b]All 71 individuals.

spersed repeats are located in the target sequence (*Mer*, *Alu*, *L1* and simple repeat sequences; Fig. 1*c*). We found that the nucleotide diversity associated with these sequences is high (0.0032, or 1 in every 312 bp) when compared with non-repeat containing sequence (0.0017, or 1 in every 588 bp).

A visual representation of this diversity across the *LPL* sequence among the 71 individuals is shown (Fig. 5). In this representation, the variable sites are colour-coded for each of the individuals that are homozygotes for the allele with highest relative frequency across the samples (blue), homozygotes for the less frequent allele (yellow) or heterozygotes (red). A consensus sequence was constructed using the most common nucleotide at each site. On average, each individual differed from this consensus sequence at approximately 22 positions (range 4 to 40 positions) either by being heterozygous or homozygous for the rare allele. In fact, nearly all the individuals sequenced (67 of 71), could be uniquely identified based on their sequence across this region. Two pairs of individuals from different populations had identical genotypes. In one case an individual from Rochester had the same genotype as one from Jackson, and in the other case, an individual from North Karelia had the same genotype as an individual from Jackson.

Discussion

The amount of DNA variation found in *LPL* highlights the potential problems associated with interpreting genetic studies that consider only a few marker sites per gene. The variation found in this gene is considerably greater than generally appreciated in biomedical genetics. Even in a small sample of 71 adults unselected for health status, we found 88 variable sites, 78 of them with relative allele frequencies greater than 1% in the pooled sample. These data suggest that the levels and patterns[16] of sequence variation found in the human genome could pose a challenge in identifying the sites, or combination of sites, that influence variation in risk of disease within and among populations.

DNA sequencing is the most sensitive method for finding DNA polymorphisms and mutations[24]. In the present study, 81 of the 88 variable sites that were not previously reported (two new sites in coding and seventy-nine new sites in non-coding regions) may now be used in genetic analyses of this region. The remaining seven sites had been identified by scans to find coding-region variants or restriction fragment length polymorphisms (RFLPs). Screens for RFLPs in *LPL* previously identified 2 of the 88 variable sites (*Pvu*II, variant at position 4509, ref. 25; *Hin*dIII, variant at position 8393, ref. 26). Even if every known restriction enzyme with five- or six-base specificities[27] were applied to the target *LPL* sequence, only 34 of the 88 variable sites would have been discovered.

In addition to its sensitivity, full DNA sequence analysis provides complete knowledge of the type, position and context of every variation in a single pass, regardless of whether it is a single nucleotide substitution or insertion/deletion variation[19]. This information is key in searching for functional variations which could reside in coding or non-coding regions, or represent a combination of variants such as those observed in the *APOE* gene[28]. By sequencing just 71 individuals unselected for disease status, two previously reported coding region variants were identified. These include one that leads to a premature truncation of the enzyme by two amino acids but does not appear to influence *LPL* function[29], and another previously reported variant (2849: 291Asn→Ser) that has been associated with premature atherosclerosis[15]. The latter was identified in only 1 of the 71 sequenced individuals as a heterozygote. This demonstrates the sensitivity of the approach as well as its potential to uncover functional variants, as the variant at 291 (Asn→Ser) has been shown to decrease the catalytic function of *LPL* (ref. 15). Two new missense variations were also identified in the coding region (6176: 370Val→Met; 6203: 379Thr→Ala). The functional significance of these new *LPL* variants remains to be determined. Four of the seven variable sites (57%) identified in the coding region of *LPL* would lead to changes in the primary sequence of the protein. Analysis of sequence differences between 47 human and rodent genes revealed 4.7 synonymous substitutions for every nonsynonymous substitution[30]. One might expect that the proportion of

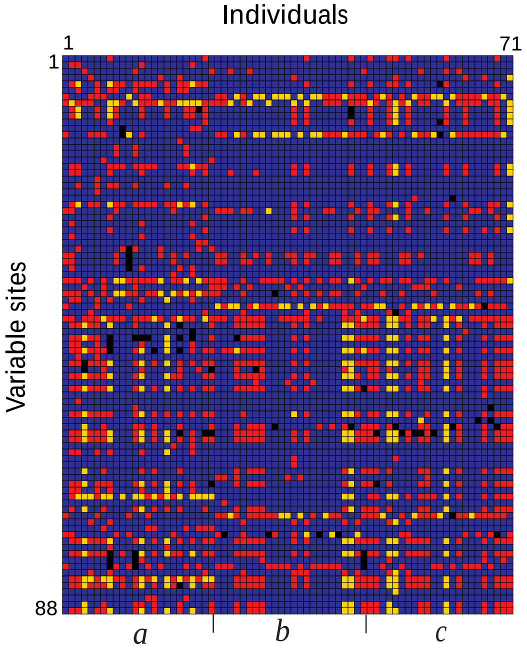


Fig. 5 Comparison of the 71 individuals to the reference *LPL* sequence for the 88 variant sites. Homozygotes for the allele with highest relative frequency across the populations (most common allele) for a variable site (blue), homozygotes for alternative alleles (gold) and heterozygotes (red) are indicated. Black denotes positions where genotypes were not obtained. Genotypes from ***a***, the 24 samples from Jackson, MS, ***b***, the 24 samples from North Karelia, Finland, and ***c***, the 23 samples from Rochester, MN.

nonsynonymous variants in the population would be higher than the proportion of nonsynonymous substitutions between species, because some of these variants are likely to be deleterious and have a lower probability of drifting to fixation. The proportions of synonymous and nonsynonymous variants in coding regions of human genes vary widely across different types of genes[31], and it is an important question to determine the causes for this variation.

As so few regions of the genome have been systematically scanned for variation in more than just a handful of individuals, our knowledge about sequence diversity is still in its infancy. Sequence analysis of *LPL* detected a somewhat higher level of diversity (nucleotide diversity equal to 0.002, or about 1 variable site per 500 bp) than had been estimated from a conservative analysis of allelic variation obtained from a scan of the human genome sequence database[31]. The diversity in *LPL* falls within the range of reported values, 0.001 to 0.003 (refs 3,31,32), and is similar to the reported estimate for the human β-globin gene which has been sequenced across 3 kb for a larger number of individuals[33]. It is likely that the amount of sequence diversity will vary widely across different regions of any given gene, and across the genome. In this regard, a higher level of sequence diversity was found associated with the interspersed repeat sequences in *LPL*, 1 in 312 bp. Therefore, these sequences in other genes may be a rich source of new polymorphic markers[3,7,34].

One consequence of the observed sequence diversity is that each individual is heterozygous at an average of 17 sites, even in just the sequenced third of the *LPL* gene. This finding has implications for the study of the biological and statistical relationships between genotype variation and trait variation in all areas of biology and medicine. If the rest of the *LPL* gene is as variable, an average individual would be heterozygous at more than 50 sites. If this is truly the case, a definitive study of the impact of *LPL* on the risk of disease would require typing nearly 300 variable sites to explore which alleles defined by specific sites, or combinations of sites, might influence trait variation. Clearly this possibility poses a daunting challenge for characterizing causation, if for no other reason than the sheer size of the samples required to search through the huge number of genotypes present.

One approach to this inferential problem is to take advantage of the population-historic processes that generate the variation. Mutation arises rarely and stochastically over time at individual sites along a gene. For a mutation to achieve high frequency takes many generations, during which the mutation can also disperse geographically from its origin. Most mutations occur in novel sites relative to an ancestral consensus sequence. This generates a hierarchical pattern of variation along individual chromosomes (for example, haplotypes; refs 35,36). Related haplotypes are often found in a similar location; thus, human population history affects the distribution of variation. In a separate report on the gene structure of this region of *LPL* (ref. 16), we identify 88 unique haplotypes in the 142 chromosomes sampled (coincidentally equaling the number of variable sites). We find that recombination along the chromosomal region may be as important as mutation in generating the observed pattern of variation. As a result, the sequence relationships do not fall in a neat, cladistic (hierarchical) structure that would result in the absence of recombination. However, there is considerable linkage disequilibrium in the data and it exhibits complex patterns within and between populations[16].

For most chronic diseases having a complex multifactorial aetiology, research is still in the gene-discovery stage. That is, genome scans are currently underway to identify candidate genes like *LPL*. It is interesting to consider the variation revealed by *LPL* in light of the gene discovery problem, as well as its implications with respect to recently proposed strategies for identifying candidate genes for complex human phenotypes[3,9,34]. One recent suggestion has been to develop a dense, single nucleotide polymorphism (SNP) map because of their frequency within the genome and ease of genotyping[3,34]. Generally, the idea is to use a high-density map of SNPs that could be genotyped on a single hybridization chip. Even the proposed density of such a map, containing about 100,000 markers, or one every 33,000 bp, would only include one marker in the entire *LPL* gene, and thus no more than 1 of the 88 variable sites we have found in our sample.

It has been suggested that moderately polymorphic SNP markers, with allele frequencies of 0.2 to 0.5 for the rarest allele, be considered for the development of higher density genetic maps[34]. This would be analogous to the search in the past decade to develop a genetic map of highly polymorphic microsatellite markers for mapping studies using pedigrees or families. However, the requirements for markers for disequilibrium mapping of disease susceptibility genes in populations may be somewhat different. A site that is heterozygous in one population may be monomorphic in others, so a SNP marker map must be developed for multiple populations. The ultimate value of an SNP map for identifying candidate genes by population-based association studies depends on the haplotype structure of the variation[16]. For example, sites that are variable in all human populations and/or have the highest heterozygosity will tend to be older, while rare single-site variations that have larger effects on an individual's risk of disease may have arisen more recently. Such a high-risk susceptibility mutation may exhibit stronger population-specific linkage disequilibrium with SNP sites that may have arisen more recently and may thus have lower heterozygosity. A map of exclusively high-heterozygosity SNPs may lose valuable information compared to a high density map that includes lower-heterozygosity markers.

Previously, genetic studies have relied on regional or random DNA markers, and therefore, incomplete data on DNA variation. Our analysis of *LPL* has employed technologies that permit the resolution of diploid sequence data on a practical scale. We have applied a process for the simultaneous analysis of diploid sequences from large numbers of individuals, using newly developed software combined with confirmatory repeat sequencing, allele-specific PCR or PCR followed by restriction digestion; this makes it possible to identify essentially all the variation in a candidate gene from a chosen sample population. Future improvements in this technology will increase its power, yielding the desired level of data completeness[37]. Our work shows that such advances will reveal a level of complexity that has not yet been contemplated, one that necessitates a sobering but much more realistic view of the biological and statistical relationships that exist between genotypic and phenotypic variation.

For example, based on the variation seen in *LPL*, it seems unlikely that only a small number of alleles in individual candidate genes[7] may influence the natural variation of disease risk factors. This may be true for rare alleles with very strong effects on risk, but it is unlikely to apply to alleles of more modest effect which, by being more common in the population, are of higher public health importance[3,38]. Even without new mapping studies, the lists of candidate genes for many important diseases are already long, especially where our understanding of the biochemical aetiology of these diseases is well-known. A spot sampling of only one or even a few varying sites in or near each candidate gene is likely to miss important causal relationships that exist.

The structure of human genetic variation in the population at large is based on the influence of mutation, recombination and differential reproduction of particular genetic combinations which is shaped by the demographic history associated with the dispersal of our species across all continents of the world. Over time and space, population genetic forces have produced the amount, geographic distribution, chromosomal arrangement and phenotypic effects that we seek to understand. Maturing sequencing technologies have made possible the complete characterization of the extant genetic variation. This is a long-awaited but necessary step in our efforts to understand how such variation originated and how it might be used to predict and understand variation in measures of human health. The complexity of the problems we will face in exploring these issues is being brought into sharper focus by the type of genetic data reported here.

Methods

Population samples. Individuals from three populations were sampled: (i) Europeans from North Karelia, Finland (n=24), who are participants in ongoing studies within a region having one of the world's highest frequencies of CHD; (ii) European-Americans from Rochester, Minnesota (n=23), who are participants in the Rochester Family Heart Study of cardiovascular risk and risk factor variation; and (iii) African-Americans from Jackson, Mississippi (n=24) who are participants in ongoing studies of the genetics of essential hypertension and its complications. All subjects were selected for this survey of DNA sequence variation without respect for their disease status or the level of any risk factor trait.

PCR and sequencing primers. Primers for PCR amplification of genomic DNA and for sequencing of the amplified products were assembled using standard phosphoramidite chemistry on an Applied Biosystems 394 DNA synthesizer. The *LPL* target region was amplified from each genomic DNA sample in nine overlapping segments and sequenced using the amplification primers as sequencing primers as well as internal sequencing primers for long PCR fragments as follows: (i) 5´–CACTAGAGAATATTTTCTCTCTC–3´ and 5´–AAAAGCAGGAACACAGATTTGG–3´ (936 bp, 60 °C), and sequencing primers 5´–TGGCATTGCAGGAAGTCTGAC–3´ and 5´–GTGTTCTTCCCGGAGACATGAC–3´; (ii) 5´–GGCCAAATGTGTATATGAAAAC–3´ and 5´–CCATGACTGTAGAATAGGAGC–3´ (475 bp, 60 °C); (iii) 5´–AGAGGACTTGGAGGTAAATATT–3´ and 5´–GACTCCTTGGTTTCCTTATTTA–3´ (1783 bp, 63 °C), and sequencing primers 5´–ATGTTACTGGAACAGAAGATG–3´, 5´–GACATTGGGACCTACCTAGG–3´, 5´–TCCAAGGTGGAGGTTGATTTC–3´ and 5´–CTGGTCCACATCTGGGTAAA–3´; (iv) 5´–AGGCTGGAGACTGTTGTAAAT–3´ and 5´–CTCAGGTTTCCATCTCTGGATTC–3´ (1229 bp, 60 °C) and sequencing primer 5´–CTATCAACTCTGTTATGGTGGC–3´; (v) 5´–CAGACCTAAAAGGAATGGTCGG–3´ and 5´–CTGTAATCCCAGTTACTCAGG–3´ (952 bp, 60 °C) and sequencing primer 5´–CAGATCTCTTAAGACTGATGGT–3´; (vi) 5´–ACCATCAGTCTTAAGAGATCTG–3´ and 5´–ATCTCAAAGGCCTGATTGGTA–3´ (1466 bp, 55 °C) and sequencing primers 5´–CTGTAATCCCAGTTACTCAGG–3´, 5´–GCTCAGTTACCCAGGCTGGAG–3´, 5´–CTTTCCTTTTAACCTCTCCCC–3´ and 5´–CTAGGGAGATCAAATCAGAAG–3´; (vii) 5´–GGTTCTGAATTGCCTGACTATT–3´ and 5´–TTCCAGCATTTATCCTAATGAC–3´ (1533 bp, 55 °C) and sequencing primers 5´–TGAGGTCTTTCCTCTATTTGATAT–3´ and 5´–GAGATGGAGTCTCTCTCATTG–3´; (viii) 5´–CATCAATTACAGTCGTACCTAT–3´ and 5´–TCAGCTTTAGCCCAGAATGC–3´ (1881 bp, 62 °C) and sequencing primers 5´–TTCTGAATGATTGACTTCAGG–3´, 5´–GGGAAGTGACAGTATTTTTG–3´, 5´–GAGCAGTCTTATGTTACTGGGC–3´ and 5´–TACACTAGCAATGTCTAGCTGA–3´; (ix) 5´–TCATTTGCAGAAAGGAAAGG–3´ and 5´–AATTCAGAACAGGAGTAGTG–3´ (794 bp, 62 °C).

DNA amplification. DNA samples from 24 African-Americans from Jackson, MS, 24 individuals from North Karelia, Finland and 23 individuals of mixed European descent from Rochester, MN (71 individuals altogether) were amplified across the target region (9.7 kb) in *LPL*. All amplification reactions were performed in a 96-well microtitre plate thermal cycler (PTC 100, MJ Research). The PCRs were assembled (50 µl total volume) with a standard buffer (10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_{2,}$ 0.001% gelatin), the four deoxynucleotide triphosphates (40 µM each), primer (0.5 µM of each), *Taq* polymerase (0.5 U; Perkin-Elmer-Cetus) and genomic DNA (20 ng). Following assembly, the reactions were covered with mineral oil (50 µl). Thermal cycling was performed with an initial denaturation at 94 °C for 1 min followed by 35 cycles of denaturation at 95 °C for 20 s, primer annealing for 30 s (temperatures specified above with primer sequences) and primer extension at 72 °C for 2 min. After 35 cycles, a final extension was carried out at 72 °C for 5 min.

DNA sequencing. Following DNA amplification, PCR products were purified by cutting the specific product from a 1% low-melt agarose gel and isolating the product with the Wizard PCR preps purification system (Promega) as described[19], or by treating the amplified product with enzymes to inactivate the unincorporated primers and deoxynucleotide triphosphates in the samples. Enzymatic treatment was accomplished by mixing the PCR product (6 µl) with exonuclease I (1 µl, 10 U/µl; Amersham) and shrimp alkaline phosphatase (1 µl, 2 U/µl; Amersham) and incubating at 37 °C for 15 min followed by 80 °C for 15 min to inactivate the exonuclease and alkaline phosphatase enzymes prior to sequencing. In our hands, PCR products treated with these enzymes sequence as well as those isolated by agarose gel electrophoresis coupled with column purification[17]. Cycle sequencing was performed according to the manufacturer's instructions using ABI PRISM Dye Terminator Sequencing Kits with Amplitaq DNA polymerase, FS (Perkin Elmer). The entire enzyme-treated PCR sample (8 µl), or the gel-purified sample (8 µl), was used as the sequencing template. The sequencing primer (3.2 pmol, PCR primer or internal primer, see above) and dye terminator (8 µl) ready-reaction sequencing premix was added to each template. Following a denaturation step at 96 °C for 2 min, dye terminator reactions were incubated at 96 °C for 15 s, 50 °C for 1 s and 60 °C for 4 min for 25 cycles. Excess dye-terminators were removed by ethanol precipitation. The extension products were evaporated to dryness under vacuum (Savant Instruments), resuspended in loading buffer (3 µl, 5:1, 1% deionized formamide/50 mM EDTA, pH 8.0), heated for 2 min at 90 °C and loaded onto an Applied Biosystems 373 sequencer.

Sequence analysis and polymorphism identification. The ABI sequence software (version 2.1.2) was used for lane tracking and first pass base-calling (Perkin-Elmer). Chromatograms were transferred to a UNIX workstation (Sun Microsystems Inc.), base-called with Phred (version 0.961028; refs 39,40), assembled with Phrap (version 0.960731), scanned by PolyPhred (version 0.970312; ref. 17) and the results viewed with Consed (version 4.0; ref. 41). Interspersed repeats in the target sequence were identified by RepeatMasker. Specific descriptions and documentation on Phred, Phrap, Consed and RepeatMasker are available at http://www.genome.washington.edu. DNA polymorphisms within the sequencing traces were identified using PolyPhred[17]. Polyphred has been designed to parse information from Phred and Phrap output files, and to provide input to Consed to aid in the identification of heterozygous single nucleotide substitutions by colour-coding potential sites. Sequencing traces were analysed using command line parameters requiring a peak drop ratio of 0.65, and a second peak ratio of 0.25. Once identified by PolyPhred, variants were visually inspected, scored and automatically entered into a database for subsequent analysis. Each variant position has been confirmed by reamplifying and resequencing the variant site from the same or opposite strand. Furthermore, because of the sequence overlap within the analysed regions, more than one call for each genotype was often obtained for each position in a sample (Table 1). With regard to data quality and accuracy, it is important to note that: (i) the base-calling program we applied, Phred, has a significantly higher accuracy in calling bases correctly, in other words, a lower error rate, than even the ABI software[39]; (ii) the genotype accuracy was estimated to be greater than 99.9% based on genotype confirmation obtained from multiple or opposite strand sequencing (6,248 genotypes in the initial data set and a total of 7,115 additional confirming genotypes for 87 of the 88 sites; Table 1), and by genotype confirmations obtained by allele-specific PCR assays, allele-specific PCR combined with sequence analysis of the specific products and/or RFLP analysis (an additional 384 confirming genotypes covering 42 of 88 sites); (iii) we detected several previously known variants as well as new variants in *LPL* (refs 10,14,25,26); (iv) the nucleotide diversity in coding regions was lower than found in non-coding regions as expected and the overall nucleotide diversity was comparable to another large-scale study; and (v) there was no significant deviation from expected estimates of heterozygosity at each site.

GenBank accession number. *LPL*, AF050163

Acknowledgements

We thank Q. Deng and V. Tobe for their assistance in obtaining the human LPL sequences, and M. Olson, D. Burke and P. Green for their advice and encouragement. This work was accomplished with support from the National Heart, Blood, and Lung Institute (HL58238, -39 & -40, HL39107) and the National Science Foundation (DIR 8809710).

Received 23 March; accepted 27 May, 1998.

1. Olson, M. V. A time to sequence. *Science* **270**, 394–396 (1995).
2. Rowen, L., Mahairas, G. & Hood, L. Sequencing the human genome. *Science* **278**, 605–607 (1997).
3. Collins F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
4. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes.* (Princeton University Press, Princeton, New Jersey, 1994).
5. Cavalli-Sforza, L.L. Opinion: How can one study individual variation for 3 billion nucleotide of the human genome? *Am. J. Hum. Genet.* **46**, 649–651 (1990).
6. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
7. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
8. Weiss, K.M. *Genetic Variation and Human Disease: Principles and Evolutionary Approaches* (Cambridge University Press, Cambridge, England, 1997).
9. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
10. Murthy, V., Julien, P. & Gagne, C. Molecular pathobiology of the human lipoprotein lipase gene. *Pharmacol. Ther*. **70**, 101–135 (1996).
11. Kirchgessner, T.G., Svenson, K.L., Lusis, A.J. & Schotz, M.C. The sequence of the cDNA encoding lipoprotein lipase. A member of a lipase gene family. *J. Biol. Chem*. **262**, 8463–8466 (1987).
12. Kirchgessner, T.G. *et al.* Organization of the human lipoprotein lipase gene and evolution of the lipase gene family. *Proc. Natl Acad. Sci. USA* **86**, 9647–9651 (1989).
13. Brunzell, J.D. Familial lipoprotein lipase deficiency and other causes of chylomicronemia syndrome. in *The Metabolic and Molecular Basis of Inherited Diseases* (eds Scriver, C.R., Beaudet, A.L., Sly, W.S. & Valle, D.) 1913–1932 (McGraw-Hill Inc., New York, 1995).
14. Wiebusch, H. *et al.* Mutations in the lipoprotein lipase gene are not restricted to patients with type 1 hyperlipidemia. *Circulation* **86**, I–609 (1992).
15. Reymer, P.W. *et al.* A lipoprotein lipase mutation (Asn291Ser) is associated with reduced HDL cholesterol levels in premature atherosclerosis. *Nature Genet.* **10**, 28–33 (1995).
16. Clark, A.G. *et al.* Haplotype structure and population genetic inferences from the nucleotide sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* submitted.
17. Nickerson, D.A., Tobe, V.O. & Taylor, S.L. Polyphred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **14**, 2745–2751 (1997).
18. Vogel, F. & Kopun, M. Higher frequencies of transitions among point mutations. *J. Mol. Evol*. **9**, 159–180 (1977).
19. Kwok, P.-Y., Deng, Q., Zakeri, H., Taylor, S.L. & Nickerson, D.A. Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics* **31**, 123–126 (1996).
20. Economou, E.P., Bergen, A.W., Warren, A.C. & Antonarakis, S.E. The polydeoxyadenylate tract of *Alu* repeat elements is polymorphic in the human genome. *Proc. Natl Acad. Sci. USA* **87**, 2951–2954 (1990).
21. Weber, J.L. & May, P.E. Abundant class of human DNA polymorphism which can be typed by the polymerase chain reaction. *Am. J. Hum. Genet*. **44**, 388–396 (1989).
22. Edwards, A., Civitello, A., Hammond, H.A. & Caskey, C.T. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet*. **49**, 746–756 (1991).
23. Hata, A., Robertson, M., Emi, M. & Lalouel, J.M. 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 (1990).
24. Eng, C. & Vijg, J. Genetic testing: The problems and the promise. *Nature Biotechnol.* **15**, 422–426 (1997).
25. Fisher, K.L., Fitzgerald, G.A. & Lawn, R.M. Two polymorphisms in the human lipoprotein lipase (*LPL*) gene. *Nucleic Acids Res*. **15**, 7657 (1987).
26. Heinzmann, C. *et al.* RFLP for the human lipoprotein lipase (*LPL*) gene: *Hin*dIII. *Nucleic Acids Res.* **15**, 6763 (1987).
27. Roberts, R.J. & Macelis, D. REBASE-restriction enzymes and methylases. *Nucleic Acids Res*. **25**, 248–262 (1997).
28. Davignon, J., Gregg, R.E. & Sing, C.F. Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis* **8**, 1–21 (1988).
29. Zhang, H. *et al.* Common sequence variants of lipoprotein lipase: standardized studies of in vitro expression and catalytic function. *Biochim. Biophys. Acta* **1302**, 159–166 (1996).
30. Li, W.-H. *Molecular Evolution.* 178–182 (Sinauer Associates, Inc., Sunderland, Massachusetts, 1997)
31. Li, W.-H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
32. Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S. & Schmidtke, J. An estimate of unique DNA sequence heterozygosity in the human genome. *Hum. Genet.* **69**, 201–205 (1985).
33. Harding, R. M. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
34. Kruglyak, L. The use of a genetic map of biallelic markers in linkage studies. *Nature Genet*. **17**, 21–24 (1997).
35. Crandall, K.A. Identifying links between genotype and phenotype using marker loci and candidate genes. in *The Impact of Plant Molecular Genetics* (ed. Sobral, B.W.S.) 137–157 (Birkhauser Press, Basel, Switzerland, 1996).
36. Templeton, A.R. 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* (eds Chadwick, D.J. & Cardew, G.) 259–283 (John Wiley and Sons, Chichester, England, 1996).
37. Rieder, M.J., Taylor, S.L., Tobe, V.O. & Nickerson, D.A. Automating the identification of DNA variations using quality-based fluorescence resequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.* **26**, 967–973 (1998).
38. Sing, C.F., Haviland, M.B. & Rielly, S.L. Genetic architecture of common multifactorial diseases. in *Variation in the Human Genome* (eds Chadwick, D.J. & Cardew, G.) 211–232 (John Wiley and Sons, Chichester, England, 1996).
39. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. **8**, 175–185 (1998).
40. Ewing, B. & Green, P. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
41. Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res*. **8**, 195–202 (1998).