Sequence Diversity and Large-Scale Typing of SNPs in the Human Apolipoprotein E Gene

Deborah A. Nickerson,^{1,7} Scott L. Taylor,¹ Stephanie M. Fullerton,^{2,3} Kenneth M. Weiss,² Andrew G. Clark,³ Jari H. Stengård,⁴ Veikko Salomaa,⁴ Eric Boerwinkle,⁵ and Charles F. Sing⁶

¹Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195, USA; ²Department of Anthropology, ³Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA; ⁴National Public Health Institute, Department of Epidemiology and Health Promotion, Helsinki, Finland; ⁵Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77225, USA; ⁶Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA

A common strategy for genotyping large samples begins with the characterization of human single nucleotide polymorphisms (SNPs) by sequencing candidate regions in a small sample for SNP discovery. This is usually followed by typing in a large sample those sites observed to vary in a smaller sample. We present results from a systematic investigation of variation at the human apolipoprotein E locus (*APOE*), as well as the evaluation of the two-tiered sampling strategy based on these data. We sequenced 5.5 kb spanning the entire *APOE* genomic region in a core sample of 72 individuals, including 24 each of African-Americans from Jackson, Mississippi; European-Americans from Rochester, Minnesota; and Europeans from North Karelia, Finland. This sequence survey detected 21 SNPs and 1 multiallelic indel, 14 of which had not been previously reported. Alleles varied in relative frequency among the populations, and 10 sites were polymorphic in only a single population sample. Oligonucleotide ligation assays (OLA) were developed for 20 of these sites (omitting the indel and a closely-linked SNP). These were then scored in 2179 individuals sampled from the same three populations ($n = 843, 884,$ and 452 , respectively). Relative allele frequencies were generally consistent with estimates from the core sample, although variation was found in some populations in the larger sample at SNPs that were monomorphic in the corresponding smaller core sample. Site variation in the larger samples showed no systematic deviation from Hardy-Weinberg expectation. The large OLA sample clearly showed that variation in many, but not all, of OLA-typed SNPs is significantly correlated with the classical protein-coding variants, implying that there may be important substructure within the classical $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles. Comparison of the levels and patterns of polymorphism in the core samples with those estimated for the OLA-typed samples shows how nucleotide diversity is underestimated when only a subset of sites are typed and underscores the importance of adequate population sampling at the polymorphism discovery stage.

[The sequence data described in this paper have been submitted to the GenBank data library under accession no. AF261279.]

The human apolipoprotein E gene (*APOE*) encodes a single chain polymorphic protein composed of 299 amino acids that plays a key role in the transport and metabolism of plasma cholesterol and triglycerides (Mahley and Huang 1999). *APOE* harbors a globally distributed polymorphism that influences variation in disease risk in human populations. There are three common isoforms of apoE that differ in their amino acid sequence at residues 112 and 158, i.e., apoE2 (cysteine-cysteine), apoE3 (cysteine-arginine), and apoE4 (arginine-arginine; Weisgraber 1994). These protein variants are encoded by haplotypes involving two diallelic single nucleotide polymorphisms (SNPs), located

in the 3' exon, that together yield the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles, respectively. Extensive association studies with disease risk have been performed for these alleles (for review, see de Knijff et al. 1994). These analyses reveal that the $\epsilon 4$ allele is associated with an increased risk for cardiovascular disease (CVD; for review, see Davignon et al. 1999) and Alzheimer's disease (Corder et al. 1993; Strittmatter et al. 1993; Meyer et al. 1998; Tang et al. 1998).

The success of association studies with *APOE* has stimulated the development of systematic approaches to find and type sequence variations on a large scale for use in candidate gene or genome-wide association studies (Lander and Schork 1994; Collins et al. 1997; Lai et al. 1998; Martin et al. 2000; Prezworski et al. 2000). Although *APOE* is often presented as a paradigm for SNP analysis in the human genome, there has yet to

⁷Corresponding author.

E-MAIL debnick@u.washington.edu; FAX (206) 685-7301

Article and publication are at www.genome.org/cgi/doi/10.1101/gr.146900.

be a systematic survey of sequence variation within this gene. Furthermore, it has become clear that not all individuals with the same *APOE* protein genotype are at equivalent risk, and variants in the regulatory regions unrelated to the protein isoforms have been identified that may have functional relevance and complicate the simple subdivision into three haplotypes (Mui et al. 1996; Artiga et al. 1998a,b; Bullido et al. 1998; Lambert et al. 1998a,b). Therefore, we have undertaken a comprehensive analysis of the genomic sequence of *APOE* to gain a better understanding of the natural variation in this gene.

In this report, we present the sequence variation observed in *APOE* in 72 individuals (144 chromosomes) sampled from three populations (two of European-descent and one of African descent) currently engaged in epidemiological studies of environmental and genetic factors that influence the risk of cardiovascular disease. This represents a core data set characterizing the relative allele and genotype frequency distributions of variable sites in this gene. Association studies between disease risk and variation in candidate genes require population samples that have been systematically investigated for both phenotypes and genetic variation. Here we use *APOE* to illustrate how the variation identified in the core sample compares with variation in a much larger epidemiological sample of 2179 individuals from the same three populations, typing only those sites observed to be variable in the smaller core sample.

This two-step approach to an association study, in which variation is defined by sequencing a small random sample of individuals, followed by the genotyping of these variations in a larger sample sufficient for epidemiological studies, is becoming a common strategy. Within a given candidate gene, this procedure can result in only a sparse set of markers. Even when a high fraction of variation within a gene is identified in a core sample, however, such a two-step approach raises important statistical problems for analysis of the resulting genetic data. The problem centers around the fact that the larger epidemiological sample is scored only at nucleotide sites that were observed to vary in the core sample. This conditional sampling of genetic variation imposes a bias, in that rare variants are likely to be missed in the larger sample. Although time and expense precluded complete sequence analysis of the larger epidemiological sample here, we were able to examine features of the bias described above by comparison of population genetic statistics estimated from the core and larger samples.

RESULTS

APOE Sequence Variants in the Core Sample

Approximately 5500 bp of DNA containing the *APOE* gene were amplified and scanned for variation (Fig. 1).

The target region contained 1059 bp of 5' flanking sequence, the entire coding sequence and intervening introns of *APOE* (four exons and three introns) spanning 3586 bp, as well as 846 bp 3' to the polyadenylation signal (Fig. 1a). Approximately 20% of the scanned sequence was coding (1156 of 5491 bp), and 80% was noncoding (4335 of 5491 bp). Several putative regulatory elements—i.e., promoter and enhancer elements, which contain protein-binding sites—have also been mapped in the 5' flanking sequence and first intron of this sequence (Fig. 1b; Paik et al. 1988; Smith et al. 1988). In addition to these elements, the noncoding regions associated with *APOE* also contain a number of common interspersed repeats such as *Alu* elements. Interspersed repeats comprised nearly half (46%) of the noncoding sequence (1987 of 4335 bp) examined (Fig. 1c).

In all, a core sample of 72 individuals (144 chromosomes) from three populations was scanned across the target region, and 22 varying sites were identified by comparing the amplified sequences using the PolyPhred program (Fig. 1d and Table 1). Of these, 21 variants (95%) were diallelic single nucleotide substitutions. Among these, transition type substitutions were more common (14 of 21, 67%) than transversions (7 of 21, 33%). One multiallelic insertion/deletion type variant was identified in the 3' end of the *APOE* gene, resulting from a length change in a mononucleotide-G tract (5229A). This position, 5229, was a compound site of variation because a single nucleotide substitution was also detected at this position (5229B).

Four of the 22 varying sites were located in the coding regions of *APOE* (Fig. 1d and Table 1). All four changes lead to nonsynonymous substitutions in the protein. Alleles defined by amino acids at two variant sites, positions 3937 (Cys112Arg) and 4075 (Arg158Cys), determine the polypeptide isoforms originally detected by protein electrophoresis—i.e., apoE2, apoE3, and apoE4—that are now routinely typed by PCR (Hixson and Vernier 1990). Two other coding SNPs, Leu28Pro (3106) and Arg142Cys (4036), were identified in exons 3 and 4, respectively. These also lead to nonconservative amino acid substitutions in the apoE protein and had been reported previously (Havel et al. 1983; de Knijff et al. 1994). Eighteen of the varying sites were located within the noncoding sequences (Fig. 1 and Table 1). Of these, seven were found 5' to exon 1. The majority of 5' sites (75, 471, 545, 560, and 624) were associated with *Alu* or *Mir* repeats (Fig. 1c). Only two of these 5' sites, 308 and 832, were not associated with known repeat elements. Moreover, one of these variants, 832, is located in a region of known enhancer activity. An enhancer sequence has also been identified in intron 1 of *APOE*, and a variant at position 1163 was found in this region as well (Fig. 1b; Mui et al. 1996). Although both of

Figure 1 Genomic structure and locations of single nucleotide polymorphisms in *APOE*. (a) The exon-intron structure. (b) The distribution of mapped protein-binding sites in the 5' flanking sequence and first intron. (c) The types and distribution of repeat sequences. (d) DNA variants identified by sequencing 72 individuals. Coding-region variants are boxed. (e) Percentage identity plot for a comparison of human and mouse genomic sequences for *APOE*. The regions associated with the exons are indicated by number (1–4). The two regions associated with enhancer activities are indicated by letter (a or b).

these sites lie in regulatory regions, neither is located in one of the mapped protein-binding sequences identified in *APOE* (Fig. 1b; Paik et al. 1988; Smith et al. 1988). Comparison of *APOE* sequences from human (positions 441 to 4478) and mouse reveals extended similarity in the coding regions (Fig. 1e). Only two noncoding regions had similarities >60% across >40 nucleotides, and both of these fall in the regions of the known enhancer activity described above.

Among the 22 variants, five sites showed only a single copy of the rarer nucleotide in the core sample (i.e., singletons, positions 308, 545, 2907, 3106, and 3673; 3106 leading to a nonsynonymous polymorphism; Table 1). Another five sites had only two copies of the rarer nucleotide (doubletons, sites 73, 471, 1522, 1575, and 4036; 4036 leading to a nonsynonymous polymorphism). Whereas one or two copies of an allele in a sample of 144 chromosomes implies a low overall relative allele frequency (i.e., 0.005 to 0.01), their relative frequency within the

population sample in which they were found is substantial (0.02 to 0.04, $2n = 48$), i.e., frequencies that would not typically be considered rare in SNP identification searches.

A visual representation of the genotypes determined for the core sample of 72 individuals reveals several key features of the sequence variation (Fig. 2). In this representation, the variable sites are color-coded for each individual, with homozygotes for the allele with the highest relative frequency across the samples color-coded blue, homozygotes for the less frequent allele color-coded yellow, and heterozygotes color-coded red. On average, each individual differed from the reference sequence at approximately four positions (range, one to eight positions) either by being heterozygous or homozygous for the rarer allele. Six individuals (three in Jackson, two in North Karelia, and one in Rochester) were homozygous across the entire scanned region and, not surprisingly, were homozygous for the most common *APOE* genotype $\epsilon 3/\epsilon 3$.

Table 1. Relative Frequencies of *APOE* Sequence Variants in the Core and Large OLA-Typed Samples

Relative frequency ^b											
Position ^a	Variant type	Jackson		North Karelia		Rochester		Total		F _{ST}	
		core (2n = 48)	OLA-typed (2n = 1686 ^c)	core (2n = 48)	OLA-typed (2n = 904 ^c)	core (2n = 48)	OLA-typed (2n = 1768 ^c)	core (2n = 144)	OLA-typed (2n = 4358 ^c)	core	OLA-typed
73	C/T	0.042	0.055	0.000	0.000 ^g	0.000	0.000 ^g	0.014	0.021	0.028	0.037
308	C/T	0.021	0.014	0.000	0.000 ^g	0.000	0.000 ^g	0.007	0.006	0.014	0.009
471	A/G	0.042	0.088	0.000	0.000 ^g	0.000	0.000 ^g	0.014	0.034	0.028	0.060
545	C/T	0.021	0.003	0.000	0.000 ^g	0.000	0.000 ^g	0.007	0.001	0.014	0.002
560 ^d	A/T	0.292	0.305 ^h	0.125	0.113 ^h	0.271	0.169	0.229	0.210 ^h	0.031	0.041
624 ^d	T/C	0.000	0.033 ^h	0.021	0.037 ^h	0.250	0.107	0.090	0.064 ^h	0.156	0.021
832 ^d	G/T	0.229	0.244	0.542	0.460 ^h	0.438	0.485	0.403	0.386 ^h	0.070	0.049
1163 ^d	G/C	0.208	0.123 ^h	0.333	0.246 ^h	0.354	0.351	0.299	0.242 ^h	0.020	0.048
1522	G/A	0.000	0.000 ^g	0.042	0.009	0.000	0.000 ^g	0.014	0.002	0.028	0.006
1575	C/T	0.000	0.000 ^g	0.000	0.007 ^h	0.042	0.031	0.014	0.014	0.028	0.014
1998	G/A	0.000	0.030 ^h	0.229	0.210 ⁱ	0.083	0.109	0.105	0.098 ^h	0.096	0.053
2440	G/A	0.521	0.354 ⁱ	0.396	0.475 ^j	0.333	0.407	0.417	0.400 ^h	0.025	0.010
2907	T/G	0.000	0.000 ^g	0.021	0.007 ^h	0.000	0.009	0.005	0.005	0.014	0.003
3106 ^e	T/C	0.000	0.000 ^g	0.000	0.007 ^h	0.021	0.005	0.007	0.003	0.014	0.002
3673	C/G	0.021	0.014 ^h	0.000	0.000 ^g	0.000	0.000 ^g	0.007	0.005	0.014	0.009
3937 ^e	T/C	0.104	0.222 ^h	0.229	0.226	0.125	0.138	0.153	0.189 ^h	0.023	0.010
4036 ^e	C/T	0.042	0.020 ^h	0.000	0.000 ^g	0.000	0.000 ^g	0.014	0.008	0.028	0.013
4075 ^e	C/T	0.042	0.103 ^h	0.042	0.040 ^h	0.187	0.093	0.090	0.086 ^h	0.057	0.011
4951	A/C	0.000	0.038 ^h	0.042	0.009	0.042	0.029	0.028	0.028 ^h	0.014	0.006
5229A	delG	0.042	k	0.250	k	0.125	k	0.139	k	0.030	k
	G ^f	0.542	k	0.417	k	0.333	k	0.431	k		
	insG	0.312	k	0.333	k	0.542	k	0.396	k		
	insGG	0.104	k	0.000	k	0.000	k	0.035	k		
5229B	G/T	0.958	k	0.958	k	0.813	k	0.910	k	0.057	k
5361	T/C	0.021	0.014 ^h	0.167	0.180 ⁱ	0.042	0.079	0.076	0.074 ^h	0.059	0.056

^aPosition of the variant position in the reference sequence, GenBank AF261279.^bFrequency of the least common allele (nucleotide listed second under Variant Type).^cTotal number of chromosomes surveyed; actual *n* varies, as frequencies were calculated relative to nonmissing data only.^dPreviously identified noncoding sites: 560 = -491, 624 = -427, 832 = -219 or Th1/E47cs, 1163 = 1E1.^eCoding-region variant.^fAllele identified in the baseline sequence for this mononucleotide track.^gThese sites were typed in approximately 188 individuals from the population sample without detecting the alternative allele.^hProportion of sites not scored = 1%–5%.ⁱProportion of sites not scored = 5.1%–10%.^jProportion of sites not scored = 10.1%–15%.^kSite not typed.

APOE Variation in the Larger Population-Based Sample Typed by OLA

Of the 21 SNPs identified in the core sequenced sample, 20 diallelic variations were amenable to genotyping in larger epidemiological samples from the same three populations (*n* = 2179 total; Table 1). Although we attempted to type a compound site of variation at position 5229, the presence of a large number of highly variable-sized alleles in close proximity to an adjacent SNP made it impossible to genotype variation at these sites accurately. The associated indel (5229A) and SNP (5229B) were thus excluded from the analysis of the larger sample. In certain cases only a subset of the full sample was typed for a given variant. If no variation was observed in the first 188 individuals (376 chromosomes) surveyed in each population, the site

was regarded as monomorphic and scored accordingly (Table 1). In addition, although minor technical difficulties precluded complete genotyping in all individuals, only four of the typed positions in the three samples had missing genotypes for >5% of the individuals. Most sites had <1% of the individuals left unscored (Table 1). A χ^2 test of homogeneity of the estimates of relative allele frequencies between the core and epidemiological samples was not significant (*P* > 0.05) for each population (Table 1). All sites that varied in a given core sample also varied in the larger epidemiological sample from the same population. However, in six cases, SNPs that showed no variation in a given core sample were found to vary in the larger OLA-typed sample from the same population (sites 624, 1998, and 4951 in Jackson; 1575 and 3106 in

Figure 2 Visual genotypes of *APOE* in the core data. Color codes represent the genotypes at each polymorphic site in each individual in the sample: blue for homozygous for the common allele, red for heterozygous, and yellow for homozygous for the rarer allele. (a) Jackson sample, (b) North Karelia sample, and (c) Rochester sample. Individuals homozygous across the locus are marked with an asterisk.

North Karelia; and 2907 in Rochester). Therefore, if only the sites that varied in a given core sample had been typed in the corresponding epidemiological sample, these sites would have remained undetected.

The large OLA-typed samples give us the opportunity to detect smaller deviations from Hardy-Weinberg proportions than would be possible using the core sequence data alone. As shown in Table 2, there is a good fit with expectation in most cases. Whereas several sites appear to have an excess of heterozygotes (i.e., 73, 832, and 3937 in Jackson; 3937 in North Karelia; and 624, 832, and 3937 in Rochester), others showed a relative deficit of heterozygosity (1998 in Jackson and 5361 in North Karelia). None of these deviations were large enough to be considered significant at an experiment-wide $\alpha = 0.05$, based on a Monte Carlo procedure like that of McIntyre et al. (2000).

Relationship of the $\epsilon 2/\epsilon 3/\epsilon 4$ Alleles and Flanking SNPs

Full analysis of linkage disequilibrium among variable sites in the *APOE* gene region requires knowledge of the linkage phase of the individuals who are heterozygous at two or more sites. Lacking this information for the OLA data, we can still examine the important issue of the extent to which the flanking SNPs occur homogeneously across the $\epsilon 2$ - $\epsilon 3$ - $\epsilon 4$ genotypes. Because of the sequence differences involved at the two determinative sites, these genotypes can be scored unambiguously without specific haplotype phasing. The major classical genotypes are $\epsilon 3/\epsilon 3$, $\epsilon 3/\epsilon 2$, and $\epsilon 4/\epsilon 3$; so for each of these three genotypes, we tallied the relative frequencies of the rare and common nucleotide in each population sample (Table 3). In 26 of 37 cases, Fisher's exact tests showed that SNP frequencies were significantly heterogeneous ($P < 0.05$) across the $\epsilon 2$ - $\epsilon 3$ - $\epsilon 4$ genotypes, and these inferences were not dependent on the inclusion of the rare genotypes, i.e., $\epsilon 2/\epsilon 2$, $\epsilon 4/\epsilon 2$, and $\epsilon 4/\epsilon 4$. In many cases, relative SNP frequencies differ strikingly among $\epsilon 2$ - $\epsilon 3$ - $\epsilon 4$ genotypes; for example the rare allele frequency for site 1998 in Rochester was 0.002, 0.004, and 0.391 in the $\epsilon 3/\epsilon 3$, $\epsilon 3/\epsilon 2$, and $\epsilon 4/\epsilon 3$ genotypes, re-

Table 2. Counts of Genotypes and Tests of Fit to Hardy-Weinberg Proportions (OLA-Typed Samples)

Site	g ₁₁	g ₁₂	g ₂₂	ObsHet	ExpHet	X ² 1d.f.
Jackson						
73	749	92	0	0.109	0.103	2.816
308	818	24	0	0.029	0.028	0.176
471	699	136	6	0.162	0.160	0.048
545	836	5	0	0.006	0.006	0.008
560	405	343	81	0.414	0.424	0.450
624	773	55	0	0.066	0.064	0.977
832	465	333	37	0.399	0.369	5.593
1163	637	180	12	0.217	0.216	0.031
1998	775	43	3	0.052	0.058	7.481
2440	338	356	105	0.446	0.457	0.543
3673	803	23	0	0.028	0.027	0.165
3937	492	312	29	0.375	0.346	5.875
4036	798	34	0	0.041	0.040	0.362
4075	666	155	8	0.187	0.185	0.094
4951	749	58	2	0.072	0.074	0.600
5361	806	24	0	0.029	0.028	0.179
North Karelia						
560	341	84	7	0.194	0.201	0.476
624	409	31	1	0.070	0.072	0.256
832	130	217	95	0.491	0.497	0.063
1163	251	160	28	0.364	0.371	0.135
1522	444	8	0	0.018	0.018	0.036
1575	439	6	0	0.013	0.013	0.020
1998	258	141	17	0.339	0.332	0.172
2440	112	198	92	0.493	0.499	0.063
2907	438	6	0	0.014	0.013	0.021
3106	439	6	0	0.013	0.013	0.020
3937	263	172	16	0.381	0.350	3.617
4075	408	33	1	0.075	0.076	0.148
4951	441	8	0	0.018	0.018	0.036
5361	287	115	18	0.274	0.295	2.147
Rochester						
560	610	250	24	0.283	0.280	0.072
624	699	178	5	0.202	0.190	3.149
832	226	455	199	0.517	0.500	1.082
1163	372	395	111	0.450	0.456	0.149
1575	827	55	0	0.062	0.060	0.914
1998	695	174	9	0.198	0.195	0.269
2440	310	422	147	0.480	0.483	0.028
2907	865	16	0	0.018	0.018	0.074
3106	874	8	0	0.009	0.009	0.018
3937	648	218	12	0.248	0.238	1.763
4075	727	146	9	0.166	0.169	0.302
4951	833	49	1	0.055	0.056	0.100
5361	750	128	6	0.145	0.146	0.044

plies that the degree of association between the $\epsilon 2$ - $\epsilon 3$ - $\epsilon 4$ sites and the flanking SNPs varies markedly across populations.

Population Distribution of *APOE* Diversity

Genetic variation at the *APOE* locus is clearly not uniformly distributed among the three populations surveyed. For example, only nine of the 22 variable sites identified in the core sample (560, 832, 1163, 2440, 3937, 4075, 5229A, 5229B, and 5361) were found to vary in all three populations (Table 1). The proportion of shared variation was slightly higher among the OLA-typed samples, with 10 out of the 20 sites showing variation in all three samples (560, 624, 832, 1163, 1998, 2440, 3937, 4075, 4951, and 5361; Table 1). Fifteen of the 22 core variable sites and 16 of the 20 OLA-typed variable sites were observed in the Jackson population, 14 of the core sites and 14 of the OLA-typed sites in the North Karelia population, and 14 core sites and 13 OLA-typed sites were observed in the Rochester population (Table 1). The level of sequence polymorphism in the samples, and its distribution among different subregions of the 5500 bp surveyed, did not differ among the three populations surveyed as assessed by the extensive overlap of the confidence intervals of the estimates

of nucleotide diversity (π in Table 4).

spectively. For such sites, the genotype at the $\epsilon 2$ - $\epsilon 3$ - $\epsilon 4$ sites provides information to predict the genotype at the SNPs (or vice versa). In addition to revealing marked intergenotypic heterogeneity observed for nearly all the sites investigated, Table 3 also illustrates the extent to which the relative frequency of an allele associated with a particular genotype can vary among population samples. For example, in those with the $\epsilon 4/\epsilon 3$ genotype, the C variant of site 1163 is found at relative frequencies of 0.082, 0.185, and 0.217 among the Jackson, North Karelia, and Rochester groups, respectively. This im-

of nucleotide diversity (π in Table 4).

Average site heterozygosity within the populations was 0.134, 0.169, and 0.182 for the Jackson, North Karelia, and Rochester core samples and 0.129, 0.138, and 0.140, respectively, for the equivalent OLA-typed samples. For the 22 sites that varied in the core sample, the mean observed heterozygosity across all samples was 0.161; in the sample typed by OLA, the equivalent value was 0.136. Although only the two cSNPs (3937 and 4075) are commonly typed in surveys of *APOE* variation, nine other sites were also common in our samples, yielding heterozygosities >0.10.

Table 3. Relative Frequencies of *APOE* Sequence Variants Within Major Genotype Classes (OLA-Typed Samples)

Position ^a	Variant type	Population	Relative SNP frequency ^b by $\epsilon 2/\epsilon 3/\epsilon 4$ genotype			Prob ^c
			$\epsilon 3/\epsilon 3$	$\epsilon 3/\epsilon 2$	$\epsilon 4/\epsilon 3$	
73	C/T	J	0.089	0.030	0.036	***
		N ^d	—	—	—	—
		R ^d	—	—	—	—
308	C/T	J	0.018	0.022	0.011	NS
		N ^d	—	—	—	—
		R ^d	—	—	—	—
471	A/G	J	0.006	0.004	0.185	***
		N ^d	—	—	—	—
		R ^d	—	—	—	—
545	C/T	J	0.004	—	0.002	NS
		N ^d	—	—	—	—
		R ^d	—	—	—	—
560	A/T	J	0.258	0.193	0.378	***
		N	0.109	0.333	0.085	***
		R	0.163	0.312	0.075	***
624	T/C	J	0.025	0.077	0.019	***
		N	0.028	0.188	0.013	***
		R	0.076	0.270	0.058	***
832	G/T	J	0.175	0.142	0.353	***
		N	0.337	0.120	0.641	***
		R	0.483	0.298	0.605	***
1163	G/C	J	0.180	0.136	0.082	***
		N	0.333	0.120	0.185	***
		R	0.455	0.246	0.217	***
1522	G/A	J ^d	—	—	—	—
		N	0.013	—	0.006	NS
		R ^d	—	—	—	—
1575	C/T	J ^d	—	—	—	—
		N	0.011	—	0.003	NS
		R	0.043	0.016	0.013	*
1998	G/A	J	—	—	0.068	***
		N	0.005	—	0.469	***
		R	0.002	0.004	0.391	***
2440	G/A	J	0.532	0.225	0.271	***
		N	0.654	0.375	0.317	***
		R	0.527	0.256	0.281	***
2907	T/G	J ^d	—	—	—	—
		N	0.011	—	0.003	NS
		R	0.013	—	0.007	NS
3106	T/C	J	—	—	—	—
		N	—	—	0.016	NS
		R	0.001	—	0.018	**
3673	C/G	J	0.017	0.022	0.011	NS
		N ^d	—	—	—	—
		R ^d	—	—	—	—
4036	C/T	J	0.039	0.004	0.009	***
		N ^d	—	—	—	—
		R ^d	—	—	—	—
4951	A/C	J	—	—	0.084	***
		N	—	—	0.016	***
		R	—	—	0.102	***
5361	T/C	J	0.020	0.004	0.015	NS
		N	0.243	0.083	0.137	***
		R	0.112	0.039	0.035	***

"—" means site did not vary. SNP = single nucleotide polymorphisms; J = Jackson; N = North Karelia; R = Rochester.

^aPosition of the variant position in the reference sequence, GenBank AF261279.

^bFrequency of the least common allele (nucleotide listed second under Variant Type).

^cProbabilities were assessed by a Fisher exact test for the resulting 2×3 table: * $P < 0.05$;

** = $P < 0.01$; *** = $P < 0.001$; NS = nonsignificant.

^dThese sites were typed in approximately 188 individuals from the population sample without detecting the alternative allele.

The degree of population subdivision in *APOE* nucleotide variation was quantified using classical F statistics (Weir 1996). For the core set of individuals who were fully sequenced, the estimate of the proportion of the variation that was attributable to among-population differences was $F_{ST} = 0.045$. Site-specific estimates of F_{ST} for the core sample ranged from 0.014 to 0.156 (Table 1). The equivalent estimate for the OLA-typed data ($n = 2179$) was $F_{ST} = 0.034$ (with site-specific values ranging from 0.002 to 0.060). As the larger sample is expected to have more unmeasured rare variants that would most likely show differences among the populations, the estimate of F_{ST} for the samples typed by OLA is likely to be an underestimate of the total differentiation among the populations; the theoretically expected degree of underestimation is under investigation. Although the differences were small, site-specific core F_{ST} values were, in most cases, larger than estimates based on variation in the OLA-typed samples (Table 1).

As shown in Figure 3, sites found in the 3' half of the 5.5-kb sequenced region have, on average, lower F_{ST} estimates than sites found more 5' in the sequence, and this is true whether estimates based on the core or OLA-typed samples are considered (only F_{ST} values for the OLA-typed samples are shown in Fig. 3; see Table 1 for corresponding core sample estimates). The low estimates of F_{ST} for sites in and around the fourth exon (including and surrounding the cSNPs at sites 3937 and 4075) are consistent with previous reports of low global variation in the frequencies of the sites responsible for the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles (e.g., Hallman et al. 1991; Gerdes et al. 1992).

Table 4. Genomic Distribution of *APOE* Sequence Diversity in the Core Sample

Sequence ^a	Jackson (2n = 48)		North Karelia (2n = 48)		Rochester (2n = 48)		Total (2n = 144)	
	S ^b	π ($\times 10^{-4}$) ^c	S	π ($\times 10^{-4}$)	S	π ($\times 10^{-4}$)	S	π ($\times 10^{-4}$)
mRNA	3	3.06 $\pm$ 3.04	2	3.83 $\pm$ 3.48	3	4.98 $\pm$ 4.10	4	4.04 $\pm$ 3.56
Non-mRNA	11	4.71 $\pm$ 2.69	11	6.10 $\pm$ 3.31	10	6.74 $\pm$ 3.60	17	6.07 $\pm$ 3.26
5'	6	9.72 $\pm$ 6.61	3	7.29 $\pm$ 5.43	3	12.17 $\pm$ 7.77	7	10.28 $\pm$ 6.80
3'	2	1.47 $\pm$ 2.29	3	5.28 $\pm$ 4.78	3	5.61 $\pm$ 4.96	3	4.28 $\pm$ 4.16
Intron	3	3.66 $\pm$ 2.60	5	5.87 $\pm$ 3.65	4	4.87 $\pm$ 3.13	7	4.87 $\pm$ 3.14
Repeat	6	3.77 $\pm$ 2.84	5	3.58 $\pm$ 2.74	5	6.34 $\pm$ 4.07	8	4.79 $\pm$ 3.30
Nonrepeat	5	5.50 $\pm$ 3.51	6	8.23 $\pm$ 4.76	5	7.08 $\pm$ 4.23	9	7.16 $\pm$ 4.22
Total	14	4.36 $\pm$ 2.41	13	5.62 $\pm$ 2.97	13	6.37 $\pm$ 3.30	21	5.65 $\pm$ 2.95

^aSequence scanned for each region was mRNA: 1156 bp, non-mRNA: 4335 bp, 5': 1059 bp, 3': 846 bp, intron: 2430 bp, repeat: 1987 bp, and nonrepeat: 2348 bp, and total: 5491 bp.

^bS is the number of variable sites detected, excluding the indel at 5229A.

^cAverage pairwise sequence difference; SE derived from stochastic and sampling variance, assuming no recombination (Tajima 1993).

Nucleotide Diversity in the Core Versus OLA-Typed Samples

For each population surveyed, an equal or greater number of polymorphic variants was observed in the OLA-typed sample than in the core sample (Table 5). However, the overall level of nucleotide diversity in the samples typed by OLA, summarized either as expected per nucleotide heterozygosity (θ ; Watterson 1975) or average pairwise sequence difference (π ; Tajima 1993), was consistently lower than the core sample. θ for the total core sample was 0.000690 ± 0.000214 , whereas it was estimated as $0.000407 \pm$

0.000107 for the combined OLA-typed sample (Table 5). Similarly, π for the core data set was estimated as 0.000565 ± 0.000295 and 0.000492 ± 0.000290 for the OLA-typed data set. These estimates are all lower than, but not significantly different from, estimates of diversity reported for other autosomal (Li and Sadler 1991; Harding et al. 1997; Nickerson et al. 1998; Rana et al. 1999; Rieder et al. 1999) and X-linked (Zietkiewicz et al. 1997; Harris and Hey 1999; Jaruzelska et al. 1999; Kaessmann et al. 1999) human loci.

The two estimates of nucleotide diversity, which are expected to be equal if the data are drawn from an equilibrium population of fixed size and constant mutation rate and all mutations are selectively neutral, may be compared using Tajima's (1989) test statistic D . For the total core data set, π was slightly lower than θ , resulting in a negative, but nonsignificant, Tajima's D . For the total OLA-typed sample, π was larger than θ , resulting in a nonsignificant positive D value. Similar levels and patterns of polymorphism were observed when data for each of the populations were considered separately (Table 5). In all but the Jackson core sample, estimates of D were positive, suggesting a slightly greater level of nucleotide site heterozygosity than expected, consistent with previous studies of human nuclear DNA sequence variation (as discussed in Hey 1997).

Figure 3 F_{ST} estimates for the OLA-typed single nucleotide polymorphisms (SNPs). The site-specific estimate for each of the 20 SNPs typed in the large epidemiological samples is shown, plotted relative to the genomic location of each site in the 5.5-kb sequenced region. The location of the *APOE* exons within the sequenced region, as presented in Fig. 1, are shown directly below the plot.

Table 5. Summary of Sequence Diversity in the Core and OLA-Typed Samples

Population	Sample	Sample size	No. of variable sites ^a	θ ($\times 10^{-4}$) ^b	π ($\times 10^{-4}$) ^c	D (Tajima 1989)
Jackson	Core	48	14	5.75 ± 2.17	4.36 ± 2.41	-0.736
	OLA-typed	1686	16	3.64 ± 1.07	4.52 ± 2.70	0.521
North Karelia	Core	48	13	5.33 ± 2.05	5.62 ± 2.97	0.163
	OLA-typed	904	14	3.45 ± 1.09	4.56 ± 2.72	0.701
Rochester	Core	48	13	5.33 ± 2.05	6.37 ± 3.30	0.586
	OLA-typed	1768	13	2.94 ± 0.93	5.06 ± 2.96	1.467
Total	Core	144	21	6.90 ± 2.14	5.65 ± 2.95	-0.507
	OLA-typed	4358	20	4.07 ± 1.07	4.92 ± 2.90	0.449

^aExcluding indel at 5229A in the core samples and 5229A and 5229B in the OLA-typed samples.

^bExpected heterozygosity per nucleotide; SE derived from variance estimate, assuming no recombination (Watterson 1975).

^cAverage pairwise sequence difference; SE derived from stochastic and sampling variance, assuming no recombination (Tajima 1993). All values of D are nonsignificant.

In all cases, estimates of D were larger for the equivalent OLA-typed sample.

We don't know how many sites this incomplete two-tiered approach misses, but we can roughly estimate the number of variable sites expected to vary in samples the size of our OLA-typed samples, assuming that the data fit the infinite sites model. (If the data conform more closely to a finite sites model of mutation, the expected number of polymorphic sites in the OLA-typed samples would be smaller [Tajima 1996], and the bias caused by two-staged sampling would be less pronounced.) Under this model, the expected number of polymorphic sites in a sample is given by the formula $E(S) = \theta \sum 1/i$, in which θ is the estimate of expected heterozygosity based on variation observed in a fully sequenced sample, and the sum runs from 1 to $2n - 1$, with $2n$ being the sample size (Watterson 1975). Using estimates of θ for *APOE* based on variation observed in each of the separate core samples (Table 5) and given sample sizes of 1686, 904, and 1768 for the Jackson, North Karelia, and Rochester populations, the expected numbers of polymorphic sites in the larger population samples (i.e., the number we would expect to observe if all individuals in the larger samples had been fully sequenced) are estimated as 25.3, 21.6, and 23.6, respectively. Of the 22 sites identified in the core sample, we observed variation in 16, 14, and 13 of OLA-typed sites, respectively, for Jackson, North Karelia, and Rochester (Table 5). Therefore, our calculations suggest that approximately 9, 8, and 11 additional variable sites exist at *APOE* in our epidemiological samples that were missed by genotyping sites found in the core sample only. Note that these estimates of missing sites are highly model dependent and that selection operating on *APOE* could drive the true number of missed SNPs either up or down from these estimates. Furthermore, the relative frequency of the rarer alleles at such sites (not observed in the core samples) in the equivalent larger samples is expected to be low.

DISCUSSION

There has been widespread interest in recent years in the use of SNPs as markers in the search for candidate loci, and ultimately alleles, underlying complex genetic disorders. *APOE* is one of the most intensively investigated of all human loci, in large part because of its role in lipid transport and metabolism, as well as its involvement in modulating cell growth and differentiation, tissue repair, and immunoregulation (Davignon et al. 1999; Mahley and Huang 1999). Studies of the three major alleles of *APOE*—i.e., $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ —have revealed that the $\epsilon 2$ allele is associated with higher levels of apoE and lower levels of plasma cholesterol, low-density lipoprotein cholesterol, apoB, and Lp(a) and have suggested $\epsilon 2$ plays a protective role against CVD, whereas the $\epsilon 4$ allele is associated with lower levels of apoE and higher levels of plasma total cholesterol, low-density lipoprotein cholesterol, apoB, and Lp(a), as well as increased risk of CVD (de Knijff et al. 1994; Stengård et al. 1995; Davignon et al. 1999) and, in some populations, of Alzheimer's disease (Corder et al. 1993; Strittmatter et al. 1993; Meyer et al. 1998; Martin et al. 2000). These well-documented associations have rested almost exclusively on the consideration of the variant protein isoforms alone, or the two cSNPs which determine them. Despite recent efforts to characterize polymorphism in the promoter region of the gene (e.g., Artiga et al. 1998b; Bullido et al. 1998; Lambert et al. 1998a,b) and a broader scan for SNPs in the region (Lai et al. 1998), no systematic investigation of variation in the locus at the nucleotide level has thus far been reported.

The genomic sequence analysis reported here, of 72 individuals (144 chromosomes) for the *APOE* locus and its flanking regions, identifies the extent of *APOE* genetic diversity more comprehensively than has been performed to date and underscores the heterogeneity that remained undetected by earlier investigations. In all, 22 variable sites were observed in 5.5 kb, corre-

sponding to an overall average level per nucleotide heterozygosity of 0.0007 (Tables 4 and 5). In other words, approximately 1 in every 1400 bp varies on average between two randomly sampled chromosomes in the core sample. The four nonsynonymous coding region variants we identified had all been reported previously: the two most common cSNPs (at sites 3937 and 4075) are those that define the major isoforms of apoE, and the rarer variant at 4036 (Arg142Cys) has been previously associated with type III hyperlipoproteinemia in a single family (Havel et al. 1983; Rall et al. 1989). The other nonsynonymous variant at 3106 (Leu28Pro) is not associated with any known lipid disorder (de Knijff et al. 1994). However, 14 of the remaining 18 noncoding variants (at sites 73, 308, 471, 545, 1522, 1575, 1998, 2440, 2907, 3673, 4951, 5229A, 5229B, and 5361) have not been observed previously. Several of these sites have levels of heterozygosity comparable with the normally-assayed cSNPs at 3937 and 4075, yet their effects, if any, with respect to phenotype variation remain to be investigated (C. Sing, in prep.; J. Stengård, in prep.).

A number of recent studies have focused attention on SNPs in the 5' flanking region of *APOE* that could alter gene expression and be involved in the phenotypic associations with Alzheimer's disease and CVD risk (Mui et al. 1996; Artiga et al. 1998a,b; Bullido et al. 1998; Lambert et al. 1998a,b; Lambert et al. 2000). Several of these SNPs, known as -491A/T (position 560) and -427 C/T (position 624), have been associated with an increased risk of Alzheimer's disease that is independent of the $\epsilon 4$ status of the individual (Artiga et al. 1998a,b; Bullido et al. 1998; Martin et al. 2000). Another regulatory region variant, denoted Th1/E47c or -219 G/T (832), has also been found to be associated with the risk of Alzheimer's disease and myocardial infarction (Lambert et al. 1998a,b, 2000). Interestingly, the SNPs at -471 A/T (560) and -427 C/T (624) are associated with an *Alu* sequence. Substitutions in *Alu* sequences are not usually involved in gene function or regulation. Nonetheless, site-directed mutagenesis of the -471 position does lead to changes in *APOE* promoter activity and differential binding to nuclear extracts (Bullido et al. 1998), although the constructs used in these studies included sequences containing the -427, -219, and 1E1 (1163) sites. Therefore, the effects ascribed to -491 A/T could be related to a combination of the alleles at one or more of these sites (e.g., Lambert 1998b).

Because our ultimate aim is to investigate the relationship of variation in measures of lipid metabolism that may play a role in CVD risk to the *APOE* polymorphism, 20 of the 22 variable sites were subsequently typed by OLA in much larger epidemiological samples from the same populations. Relative allele frequency estimates based on OLA genotyping were consistent

with estimates derived from the fully-sequenced core samples. Sites that did not vary in the core sample of a particular population but were subsequently found to be polymorphic in the large epidemiological sample of the same population typically had relative frequencies of the rare allele <0.03 . This illustrates one of several dangers inherent to two-tiered strategies. The core sample from a given population will identify some, but not all, of the sites that vary in that population. Further, the characteristics of variation in the larger sample may be well reflected by those of the same sites in the core sample but cannot accurately address the overall variation in the larger sample, because not all sites that vary in the latter will be known. Finally, of course, the potential efficacy of variation identified in the core sample for use in epidemiological association studies cannot fully be assessed, because the sites of etiological importance, the relative frequencies of their alleles, and their association with variation in the core sites will all be unknown.

Tests of site-specific Hardy-Weinberg equilibrium for the data from the OLA-typed samples suggested no significant departures from expected proportions. All three populations had a small deficit of homozygotes for the rare allele at site 3937, which defines in part the $\epsilon 4$ allele, consistent with a moderate deficit of $\epsilon 4$ homozygous genotypes relative to expectation. Although there is independent evidence for a decline in the frequency of the $\epsilon 4$ allele with age (Miettinen et al. 1994; Schächter et al. 1994; Haviland et al. 1995; Stengård et al. 1995), it is not clear what causes the deficit observed here. There was also weak deviation for the regulatory site, 832, in two populations, and this has been associated with phenotypic effects in some studies (Paik et al. 1988; Smith et al. 1988). Although Table 4 reflects the existence of linkage disequilibrium among sites in *APOE*, the same data also show that without directly typing the site(s) of etiological importance, a random SNP cannot be relied on to reliably predict the $\epsilon 2/\epsilon 3/\epsilon 4$ genotype. Similar observations have been reported for a recent study of 1.5 Mb of sequence surrounding the *APOE* gene (Martin et al. 2000). In that case, only certain close and some distant sites could pick up signal due to the two-site $\epsilon 4$ haplotype.

That the core diversity is representative of the same alleles in the larger sample is shown by the similarity of the nucleotide diversity (π) values in both samples. That rare sites are missing in the OLA sample is reflected in the θ values for the OLA data relative to the core and the consequent inflation of the Tajima's D values that compare the two statistics. To the extent that the underlying theoretical assumptions hold, our calculations suggest that approximately ten more sites may vary in each of the large epidemiological samples. Although relative allele frequencies at these are likely to be rare, because they were not seen in the rather

substantial core samples, their details remain unknown, and thus their relevance for predicting phenotypic variability cannot be evaluated.

The nature and size of the core sample can also have important consequences for subsequent large-scale analyses in the geographic apportionment of the observed diversity. At *APOE*, estimates of F_{ST} were 0.045 for the core and 0.034 for the OLA-typed samples. Although the second value is underestimated, both estimates are low compared with an average estimate of 0.139 ± 0.010 reported earlier for a collection of 100 diallelic human genetic markers (Bowcock et al. 1991). Although some of the apparent difference may be because our study did not sample worldwide geographic variation, our values do agree with the previous suggestion that interpopulation differences in *APOE* diversity are low relative to other loci (Gerlenter et al. 1998). Yet, despite this fact, readily apparent differences in site variability exist among the three populations surveyed here. Seven of the 22 variable sites were observed to vary in one population only, for example, and all but one of these (site 1522) were restricted to the Jackson African-American sample. These population-specific variants attain relative frequencies of as much as 0.09 in the OLA-typed Jackson sample and would generally be considered common polymorphisms by most human geneticists, hence well worth consideration with respect to phenotypic variance. Such low-frequency population-specific variants would have remained uninvestigated, however, if the core analysis had focused either on a smaller sample from the same population or, as has been proposed recently (Collins et al. 1999), relied solely on a small mixed panel of anonymous individuals from different geographic regions. As such frequency differences are only likely to be greater at more polymorphic loci, the implications of poor sample choice at the SNP discovery stage are considerable. Clearly, small core samples underrepresent the true number of variable sites, missing even those with an appreciable relative frequency of the rarer allele.

Genomic sequence analysis is an important prerequisite for designing and implementing large-scale genotyping studies. As our survey of *APOE* nucleotide diversity shows, however, care must be exercised in the way such core analysis is conducted and interpreted. With adequate population sampling at the sequence level, the likelihood of characterizing sites with high information content increases, and the ability to draw inferences about underlying variation left untyped in large-scale genotyping surveys is enhanced. Even then, it is not clear a priori what sampling design would be optimum for detecting disease association by disequilibrium with untyped sites, because that question involves the unknown number, arrangement, and frequency of the etiologically relevant variations. How-

ever, our results suggest that those who rely solely on a small core panel of SNPs, ascertained from a limited number of individuals with poorly defined population affiliation, may miss important underlying variation, with possible adverse consequences for the power of subsequent large-scale genotype-phenotype analyses. It is clear, for example, that there is much more to *APOE* genetic diversity than two cSNPs and their well-known resulting isoforms; there is considerable haplotypic variation within each of those isoforms as well (Fullerton et al., in press). The extent to which these newly discovered polymorphisms may explain additional variation in phenotypes is currently being investigated.

METHODS

Population Samples

Individuals from three populations were sampled: (1) Europeans from North Karelia, Finland ($n = 24$), (2) European-Americans from Rochester, Minnesota ($n = 24$), and (3) African-Americans from Jackson, Mississippi ($n = 24$). All subjects were selected for this survey without respect to their disease status or their levels of any risk factor trait. After the variable nucleotide sites were ascertained in this set of 72 individuals (144 chromosomes), a larger sample from North Karelia ($n = 452$), Rochester ($n = 884$), and Jackson ($n = 843$) was scored by OLA.

DNA Amplification

The *APOE* gene (reference sequence: GenBank AF261279) was amplified from each individual in nine overlapping segments. Either a universal forward (-21M13, TGTAACGACGGC-CAGT) or reverse (M13reverse, CAGGAAACAGCTATGACC) sequence was added to each *APOE* specific primer (forward to forward; reverse to reverse) before synthesis. The following specific primer pairs were used to amplify the *APOE* gene (listed as the forward and reverse primers with PCR product size and primer annealing temperature in parentheses): (1) CTTGATGCTCAGAGAGGACAAG and GGCATAGAGTCTTTTGACCA (1122 bp, 63°C), (2) GGTCAGGAAAGGAGGACTCT and GTCCCAGTCTCGCATTCCTC (1072 bp, 58°C), (3) GGCAGCGACACGGTAGCTAG and AACCGAGGCCAGAGAGCGT (672 bp, 61°C), (4) GTTGCTGGTCACATTCCTGG and GAGTCGGTTAATCACTTG (940 bp, 63°C), (5) AGCCCTGCCTGGGGCACAC and GGACACTCACCTCAGTTCCT (744 bp, 58°C), (6) GAGTGGCAGAGCGGCCAGCG and CCTTCACTCCTTCATGGTCTC (1143 bp, 63°C), (7) CTAGCTCCTCTTCGTCTCTG and GCTCGAACCAGCTCTTGAGG (694 bp, 58°C), (8) GCCAGCCGCTACAGGAGCG and CCAGCTACTGAGGCAGCAG (638 bp, 58°C), and (9) GTGTGTATCTTTCTCTTGCC and GGCAGGCCGCTCGGAGCCCAT (751 bp, 63°C). All amplification reactions were performed in 96-well microtiter plate thermal cyclers (PTC 100, MJ Research). PCRs were assembled in 20- μ L total volume containing 50 ng of genomic DNA using the advantage GC genomic polymerase system (Clontech). Following assembly, 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 sec, primer annealing for 30 sec (temperatures above), and primer extension at 72°C for 2 min. After 35 cycles, a final extension was performed 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 previously (Nickerson et al. 1997). Cycle sequencing was performed according to the manufacturer's instructions using ABI PRISM Dye Primer Sequencing Kits with AmpliTaq FS DNA polymerase (PE Biosystems). Dye primer sequencing using the universal forward and reverse primers attached to the gene-specific primers was performed by assembling four separate reactions as follows: 1 μ L each of the PCR sample mixed with 4 μ L of the PRISM ready premix for the A and C reactions, and 2 μ L each of the PCR sample mixed with 8 μ L of the PRISM ready premix for the G and T reactions. Sequencing reactions were denatured for 1 min at 96°C and subjected to 15 cycles at 96°C for 10 sec, 55°C for 5 sec, and 70°C for 1 min and 15 cycles at 96°C for 10 sec and 70°C for 1 min. Then, the A, C, G, and T reactions were pooled and subjected to ethanol precipitation, resuspended in 1.5 μ L of loading buffer (5:1, 1% deionized formamide/50 mM EDTA at pH 8.0), heated for 2 min at 90°C, and loaded onto an Applied Biosystems 377 sequencer according to the manufacturer's directions.

Sequence Analysis and Polymorphism Identification

The ABI sequence software (version 2.1.2) was used for lane tracking and first-pass base-calling (PE Biosystems). Chromatograms were transferred to a Sun UNIX workstation, base-called with Phred (Ewing et al. 1998; Ewing and Green 1998), assembled with Phrap (Green 1999), and scanned by PolyPhred (Nickerson et al. 1997). The results were viewed with the Consed program (Gordon et al. 1998). Interspersed repeats in the target sequence were identified by the program RepeatMasker (Smit and Green 1999). Specific descriptions and documentation for Phred, Phrap, Consed, and RepeatMasker, are available at <http://bozeman.mbt.washington.edu/index.html>; for PolyPhred, <http://droog.mbt.washington.edu>.

DNA polymorphisms were identified using the PolyPhred program (version 3.0; Nickerson et al. 1997). Once identified, the variants were visually inspected and automatically entered into a database for subsequent analysis. Each variant position was confirmed by reamplifying and resequencing the variant site from the same or opposite strand. In addition, because of the sequence overlap within the analyzed regions, more than one call for each genotype was obtained for each position in a sample. In regard to data quality and accuracy, it is important to note that (1) the base-calling program we applied, Phred, has a significantly higher accuracy in calling bases correctly, i.e., a lower error rate, than even the ABI software (Ewing et al. 1998); (2) the genotype accuracy was estimated to be >99% based on genotype confirmation obtained from multiple or opposite strand sequencing (data not shown); and (3) all the identified variants were confirmed by genotyping the PCR products independently by OLAs. Information on these SNPs is available in AF261279 and in dbSNP (Sherry et al. 1999). Sequence comparisons between human (AF261279) and mouse (D00466) were performed with Advanced Pipmaker (<http://bio.cse.psu.edu/pipmaker/>) using the chaining option (Schwartz et al. 2000).

OLA

A colorimetric single-well OLA was used to type the identified SNPs as described previously in detail (Tobe et al. 1996). Regions of the *APOE* gene containing SNPs were amplified as

described above, and the products (~20 μ L) were diluted with 50 μ L of distilled H₂O containing 0.1% Triton X-100. A 10- μ L aliquot of the diluted product was then mixed with 10 μ L of a solution containing 2 $\times$ ligase buffer (40 mM Tris-HCl [pH 8.0]/20 mM MgCl₂/2 mM dithiothreitol), 2 mM nicotinamide adenine dinucleotide, 25 mM KCl, 0.167 U Ampligase DNA Ligase (Epicentre), and 200 fmol of each of the ligation primers (the two allele-specific primers each labeled at its 5' end with a specific hapten [digoxigenin or fluorescein] and the joining primer for the SNP being tested phosphorylated and labeled at its 3' end with biotin). Ligation reactions were overlaid with mineral oil and placed in a thermocycler for 20 cycles at 93°C for 30 s and 58°C for 2 min. After cycling, the reactions were stopped by the addition of 10 μ L of 0.1 M EDTA in 0.1% Triton H₂O and transferred in their entirety (including the mineral oil) to a 96-well flat bottom microtiter plate (Falcon) that had been coated with streptavidin (Sigma; 50 μ L of 25 μ g/ml incubated 1 hr at 37°C). Ligation products were allowed to capture on the streptavidin plate at room temperature (RT) for 1 hr, and the plate was washed twice with an NaOH buffer (0.01 M NaOH/0.05% Tween 20) followed by two washes with Tris buffer (100 mM Tris-HCl [pH 7.5]/150 mM NaCl/0.05% Tween 20). An antibody mixture (40 μ L in 1 $\times$ PBS with 0.5% BSA) consisting of a 1:1000 dilution of alkaline phosphatase-labeled anti-fluorescein antibodies and 1:1000 dilution of horseradish peroxidase-labeled anti-digoxigenin antibodies was added to each well. After 30 min at RT, plates were washed six times with Tris buffer. After washing, an alkaline phosphatase substrate (25 μ L, Bethesda Research Laboratories enzyme-linked immunosorbent assay amplification system) was added to the wells, the plates were incubated for an additional 10 min at RT, and then 25 μ L of amplifier were added to each well.

Spectrophotometric absorbances were taken at 490 nm using a microplate reader (Bio-Rad 3550) and saved as optical density (OD) readings in the attached computer. After detection of the fluorescein reporter, the plates were washed again six times with Tris buffer, and 50 μ L of the horseradish peroxidase substrate, 3,3',5,5'-tetramethylbenzidine (TMB; Sigma), were added to each well to detect the digoxigenin reporter. Spectrophotometric absorbances were taken at 655 nm for this reporter and saved in the attached computer. Sequences of the OLA primers and a detailed protocol for the assay are available at <http://droog.mbt.washington.edu>.

Genotypes were automatically derived by applying a simple threshold to call a positive (OD > 0.150) or negative (OD < 0.150) reaction. Duplicate genotypes were obtained for ~10% of the individuals assayed at each site, and genotype concordance of >98.6% was detected, a concordance rate that was similar to prior studies on other SNPs (Delahunty et al. 1996; Tobe et al. 1996). Individuals with discordant genotypes were reassayed by OLA or by sequencing analysis, and the genotypes concordant for two of the three assays were accepted as the final genotype. Genotyping of site 5229 in *APOE* (varying sites 5229a and 5229b) in the larger population samples revealed that the core sequencing sample did not reveal the entire spectrum of allelic variation (number of G's in the tract) at this position. Also, this position could not be accurately interpreted by length because of the difficulties normally encountered for typing mononucleotide tracts, which is also compounded by the presence of substitution variation. Therefore, this position could only be accurately called by sequence analysis combined with manual interpretation of the position, precluding its typing on a large scale.

Statistical Analyses

Allele frequencies for each variable site (with or without regard to the observed $\epsilon 2/\epsilon 3/\epsilon 4$ genotype) were estimated by gene counting, because genotypes were scored directly by sequencing or OLA typing.

Several standard statistics were estimated to characterize the amount and pattern of nucleotide polymorphism in *APOE*. Nucleotide diversity, π , was estimated as the average heterozygosity for all nonindel sites in the sequence (invariant sites counted as having heterozygosity of zero); standard errors of this estimate included both stochastic and sampling variance and were calculated with the conservative assumption of no recombination between sites (Tajima 1993). A related estimator, $\theta = 4N_e\mu$, characterizes the variation in terms of that expected in a standing population in mutation-drift equilibrium, with mutation rate, μ , per sequence per generation and N_e as the effective population size, and represents the expected average heterozygosity. θ was estimated from the observed number of nonindel segregating sites, S , in a sample of $2n$ chromosomes, according to the formula $\theta = S/\epsilon(1/i)$, summed from $i = 1$ to $2n - 1$ (Watterson 1975). The standard error of this estimate, also following Watterson (1975), was calculated assuming no recombination. The equality of the estimates of θ and π was tested by Tajima's D statistic (Tajima 1989).

Hardy-Weinberg tests for genotype frequency distributions were performed on the observed genotype frequencies for each site and population, with significance based on a standard observed-expected χ^2 with 1 df. The degree to which subdivision into separate populations is reflected in the amount of allelic variation was measured by the parameter F_{ST} , the ratio of the variance of allele frequencies among the population to the genetic variance of the pooled data (Weir 1996). Pairs of sites showing significant linkage disequilibrium were identified by application of a likelihood ratio test that compares the likelihood of the data assuming linkage equilibrium (calculated as the product of the allele frequencies at each site) with the likelihood of the data assuming haplotype frequencies estimated by the expectation-maximization algorithm (Slatkin & Excoffier 1996), implemented by the program Arlequin v. 2 (<http://anthropologie.unige.ch/arlequin/>).

ACKNOWLEDGMENTS

We thank Cheryl Thayer, Christa Broers and Barney Gill for their assistance in obtaining the human *APOE* sequences and OLA typings. This work was accomplished with support from the National Heart, Blood, and Lung Institute (HL58238, HL58239, HL58240, and HL39107).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Artiga, M.J., Bullido, M.J., Frank, A., Sastre, I., Recuero, M., Garcia, M.A., Lendon, C.L., Han, S.W., Morris, J.C., Vazquez, J., et al. 1998a. Risk for Alzheimer's disease correlates with transcriptional activity of the *APOE* gene. *Hum. Mol. Genet.* **7**: 1887-1892.
- Artiga, M.J., Bullido, M.J., Sastre, I., Recuero, M., Garcia, M.A., Aldudo, J., Vazquez, J., and Valdivieso, F. 1998b. Allelic polymorphisms in the transcriptional regulatory region of apolipoprotein E gene. *FEBS Lett.* **421**: 105-108.
- Bowcock A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K., and Cavalli-Sforza, L.L. 1991. Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. *Proc. Natl. Acad. Sci.* **88**: 839-843.
- Bullido, M.J., Artiga, M.J., Recuero, M., Sastre, I., Garcia, M.A., Aldudo, J., Lendon, C., Han, S.W., Morris, J.C., Frank, A., et al. 1998. A polymorphism in the regulatory region of *APOE* associated with risk for Alzheimer's dementia. *Nat. Genet.* **18**: 69-71.
- Collins F.S., Guyer, M.S., and Chakravarti, A. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* **278**: 1580-1581.
- Collins F.S., Brooks, L.D., and Chakravarti, A. 1999. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**: 1229-1231.
- Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., and Pericak-Vance, M.A. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**: 921-923.
- Davignon, J., Cohn, J.S., Mabile, L., and Bernier, L. 1999. Apolipoprotein E and atherosclerosis: Insight from animal and human studies. *Clin. Chim. Acta.* **286**: 115-143.
- de Knijff, P., van den Maagdenberg, A.M.J.M., Frants, R.R., and Havekes, L.M. 1994. Genetic heterogeneity of apolipoprotein E and its influence on plasma lipid and lipoprotein levels. *Hum. Mut.* **4**: 178-194.
- Delahunty, C., Ankener, W., Deng, Q., Eng, J., and Nickerson, D.A. 1996. Testing the feasibility of DNA typing for human identification by PCR and an oligonucleotide ligation assay. *Am. J. Hum. Genet.* **58**: 1239-1246.
- Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**: 186-194.
- Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**: 175-185.
- Fullerton, S. M., Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., et al. 2000. Apolipoprotein E variation at the sequence haplotype level: Implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* (In press.).
- Gelernter, J., Kranzler, H., and Lacobelle, J. 1998. Population studies of polymorphisms at loci of neuropsychiatric interest (tryptophan hydroxylase [TPH], dopamine transporter protein [SLC6A3], D3 dopamine receptor [DRD3], apolipoprotein E [APOE], mu opioid receptor [OPRM1], and ciliary neurotrophic factor [CNTF]). *Genomics* **52**: 289-297.
- Gerdas, L., Klausen, I., Sihm, I., and Faergeman, O. 1992. Apolipoprotein E polymorphism in a Danish population compared to findings in 45 other study populations around the world. *Genetic Epidemiol.* **9**: 155-167.
- Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**: 195-202.
- Green, P. 1999. <http://www.genome.washington.edu/UWGC/analysis-tools/phrap.htm>.
- Hallman, D., Boerwinkle, E., Saha, N., Sandholzer, C., Menzel, H.J., Csazar, A., and Utermann, G. 1991. The Apolipoprotein A polymorphism: A comparison of allele frequencies and effects in nine populations. *Am. J. Hum. Genet.* **49**: 338-349.
- Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S., and Clegg, J.B. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**: 772-789.
- Harris, E.E. and Hey, J. 1999. X chromosome evidence for ancient human histories. *Proc. Natl. Acad. Sci.* **96**: 3320-3324.
- Havel, R.J., Kottie, L., Kane, J.P., Tun, P., and Bersot, T. 1983. Atypical familial dysbetalipoproteinemia associated with apolipoprotein phenotype $\epsilon 3/\epsilon 3$. *J. Clin. Invest.* **72**: 379-387.
- Haviland, M.B., Lussier-Cacan, C., Davignon, J., and Sing, C.F. 1995. Impact of apolipoprotein E genotype variation on means, variances, and correlations of plasma lipid, lipoprotein, and apolipoprotein traits in octogenarians. *Am. J. Med. Genet.* **58**: 315-331.

- Hey, J. 1997. Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* **14**: 166–172.
- Hixson, J.E. and Vernier, D.T. 1990. Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with *Hha*I. *J. Lipid Res.* **31**: 545–548.
- Jaruzelska, J., Zietkiewicz, E., Batzer, M., Cole, D.E., Moisan, J.P., Scozzari, R., Tavaré, S., and Labuda, D. 1999. Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: Analysis of the haplotype structure and genealogy. *Genetics* **152**: 1091–1101.
- Kaessmann, H., Heissig, F., von Haeseler, A., and Pääbo, S. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.* **22**: 78–81.
- Lai, E., Riley, J., Purvis, I., and Roses, A. 1998. A 4.4 Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54**: 31–38.
- Lambert, J.-C., Pasquier, F., Cottel, D., Frigard, B., Amouyel, P., and Chartier-Harlin, M.-C. 1998a. A new polymorphism in the APOE promoter associated with risk of developing Alzheimer's disease. *Hum. Mol. Genet.* **7**: 533–540.
- Lambert, J.-C., Berr, C., Pasquier, F., Delacourte, A., Frigard, B., Cottel, D., Perez-Tur, J., Mouroux, V., Mohr, M., Cecyre, D., et al. 1998b. Pronounced impact of the Th1/E47cs mutation compared with-491 AT mutation on neural APOE gene expression and risk of developing Alzheimer's disease. *Hum. Mol. Genet.* **7**: 1511–1516.
- Lambert, J.-C., Brousseau, T., Defosse, V., Evans, A., Arveiler, D., Ruidavets, J.-B., Haas, B., Cambou, J.-P., Luc, G., Ducimetière, P., et al. 2000. Independent association of an APOE gene promoter polymorphism with increased risk of myocardial infarction and decreased APOE plasma concentrations—the ECTIM Study. *Hum. Mol. Genet.* **9**: 57–61.
- Lander, E.S. and Schork, N.J. 1994. Genetic dissection of complex traits. *Science* **265**: 2037–2048.
- Li, W.H. and Sadler, L.A. 1991. Low nucleotide diversity in man. *Genetics* **129**: 513–523.
- Mahley, R.W. and Huang, Y. 1999. Apolipoprotein E: From atherosclerosis to Alzheimer's disease and beyond. *Curr. Opin. Lipidol.* **10**: 207–217.
- Martin E.R., Lai, E.H., Gilbert, J.R., Rogala, A.R., Afshari, A.J., Riley, J., Finch, K.L., et al. 2000. SNPping away at complex diseases: Analysis of single-nucleotide polymorphisms around APOE in Alzheimer Disease. *Am. J. Hum. Genet.* **67**: 383–394.
- McIntyre, L.M., Martin, E.R., Simonsen, K.L., and Kaplan, N.L. 2000. Circumventing multiple testing: A multilocus monte carlo approach to testing for association. *Genet. Epidemiol.* **19**: 18–29.
- Meyer, M.R., Tschanz, J.T., Norton, M.C., Welsh-Bohmer, K.A., Steffens, D.C., Wyse, B.W., and Breitner, J.C. 1998. APOE genotype predicts when—not whether—one is predisposed to develop Alzheimer disease. *Nat. Genet.* **19**: 321–322.
- Miettinen, H.E., Korpela, K., Hamalainen, L., and Kontula, K. 1994. Polymorphisms of the apolipoprotein and angiotensin converting enzyme genes in young North Karelian patients with coronary heart disease. *Hum. Genet.* **94**: 189–192.
- Mui, S., Briggs, M., Chung, H., Wallace, R.B., Gomez-Isla, T., Rebeck, G.W., and Hyman, B.T. 1996. A newly identified polymorphism in the apolipoprotein E enhancer gene region is associated with Alzheimer's disease and strongly with the ϵ 4 allele. *Neurol.* **47**: 196–201.
- Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25**: 2745–2751.
- Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengård, J., Salomaa, V., Vartiainen, E., Boerwinkle, E., and Sing, C.F. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat. Genet.* **19**: 233–240.
- Paik, Y.-K., Chang, D.J., Reardon, C.A., Walker, M.D., Taxman, E., and Taylor, J.M. 1988. Identification and characterization of transcriptional regulatory regions associated with expression of human apolipoprotein E gene. *J. Biol. Chem.* **263**: 13340–13349.
- Prezowski, M., Hudson, R.R., and Di Rienzo, A. 2000. Adjusting the focus on human variation. *Trends Genet.* **16**: 296–302.
- Rall, S.C., Newhouse, Y.M., Clarke, H.R.G., Weisgraber, K.H., McCarthy, B.J., Mahley, R.W., and Bersot, T.P. 1989. Type III Hyperlipoproteinemia associated with apolipoprotein E phenotype ϵ 3/3. Structure and genetics of an apolipoprotein ϵ 3 variant. *J. Clin. Invest.* **83**: 1095–1101.
- Rana, B.K., Hewett-Emmett, D., Jin, L., Chang, B.H., Sambuughin, N., Lin, M., Watkins, S., Bamshad, M., Jorde, L.B., Ramsay, M., et al. 1999. High polymorphism at the human melanocortin 1 receptor locus. *Genetics* **151**: 1547–1557.
- Rieder, M.J., Taylor, S.L., Clark, A.G., and Nickerson, D.A. 1999. Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.* **22**: 59–62.
- Schächter, F., Faure-Delanef, L., Guenot, F., Rouger, H., Froguel, P., Lesueur-Ginot, L., and Cohen, D. 1994. Genetic associations with human longevity at the APOE and ACE loci. *Nat. Genet.* **6**: 29–32.
- Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMarker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**: 577–586.
- Sherry, S.T., Ward, M., and Sirotkin, K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**: 677–679.
- Slatkin, M., and Excoffier, L. 1996. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity* **76**: 377–383.
- Smit, A.F.A. and Green, P. 1997. RepeatMasker at <http://ftp.genome.washington.edu/RM/RepeatMasker.html>.
- Smith, J.D., Melian, A., Leff, T., and Breslow, J.L. 1988. Expression of the human apolipoprotein E gene is regulated by multiple positive and negative elements. *J. Biol. Chem.* **263**: 8300–8308.
- Stengård, J.H., Zerba, K.E., Pekkanen, J., Ehnholm, C., Nissinen, A., and Sing, C.F. 1995. Apolipoprotein E polymorphism predicts death from coronary heart disease in a longitudinal study of elderly Finnish men. *Circulation* **91**: 265–269.
- Strittmatter, W.J., Saunders, A.M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G.S., and Roses, A.D. 1993. Apolipoprotein E: High-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl. Acad. Sci.* **90**: 1977–1981.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- . 1993. Measurement of DNA polymorphism. In *Mechanisms of Molecular Evolution. Introduction to Molecular Paleopopulation Biology* (eds. N. Takahata and A.G. Clark), pp. 37–59. Sinauer Associates Inc, Sunderland, Maryland.
- . 1996. The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* **143**: 1457–1465.
- Tang, M.X., Stern, Y., Marder, K., Bell, K., Gurland, B., Langigua, R., Andrews, H., Feng, L., Tycko, B., and Mayeux, R. 1998. The APOE- ϵ 4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *JAMA* **279**: 751–755.
- Tobe, V.O., Taylor, S.L., and Nickerson, D.A. 1996. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. *Nucleic Acids Res.* **24**: 3728–3732.
- Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7**: 256–276.
- Weir, B. 1996. *Genetic Data Analysis*. Sinauer Associates Inc., Sunderland, Maryland.
- Weisgraber, K.H. 1994. Apolipoprotein E: Structure-function relationships. *Adv. Protein Chem.* **45**: 249–302.
- Zietkiewicz, E., Yotova, V., Jarnik, M., Korab-Laskowska, M., Kidd, K.K., Modiano, D., Scozzari, R., Stoneking, M., Tishkoff, S., Batzer, M., et al. 1997. Nuclear DNA diversity in worldwide distributed human populations. *Gene* **205**: 161–171.

Received May 3, 2000; accepted in revised form August 17, 2000.

Sequence Diversity and Large-Scale Typing of SNPs in the Human Apolipoprotein E Gene

Deborah A. Nickerson, Scott L. Taylor, Stephanie M. Fullerton, et al.

Genome Res. 2000 10: 1532-1545

Access the most recent version at doi:[10.1101/gr.146900](https://doi.org/10.1101/gr.146900)

References

This article cites 54 articles, 21 of which can be accessed free at:
<http://genome.cshlp.org/content/10/10/1532.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Affordable, Accurate
Sequencing.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>