

Contributions of 18 Additional DNA Sequence Variations in the Gene Encoding Apolipoprotein E to Explaining Variation in Quantitative Measures of Lipid Metabolism

Jari H. Stengård,^{1,5} Andrew G. Clark,² Kenneth M. Weiss,³ Sharon Kardia,⁴ Deborah A. Nickerson,⁶ Veikko Salomaa,¹ Christian Ehnholm,¹ Eric Boerwinkle,⁷ and Charles F. Sing⁵

¹National Public Health Institute, Helsinki; ²Department of Molecular Biology & Genetics, Cornell University, Ithaca; ³Department of Anthropology, Pennsylvania State University, University Park; Departments of ⁴Epidemiology and ⁵Human Genetics, University of Michigan, Ann Arbor; ⁶Department of Molecular Biotechnology, University of Washington, Seattle; and ⁷Human Genetics Center, The University of Texas Health Science Center, Houston

Apolipoprotein E (ApoE) is a major constituent of many lipoprotein particles. Previous genetic studies have focused on six genotypes defined by three alleles, denoted $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$, encoded by two variable exonic sites that segregate in most populations. We have reported studies of the distribution of alleles of 20 biallelic variable sites in the gene encoding the ApoE molecule within and among samples, ascertained without regard to health, from each of three populations: African Americans from Jackson, Miss.; Europeans from North Karelia, Finland; and non-Hispanic European Americans from Rochester, Minn. Here we ask (1) how much variation in blood levels of ApoE (lnApoE), of total cholesterol (TC), of high-density lipoprotein cholesterol (HDL-C), and of triglyceride (lnTG) is statistically explained by variation among *APOE* genotypes defined by the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles; (2) how much additional variation in these traits is explained by genotypes defined by combining the two variable sites that define these three alleles with one or more additional variable sites; and (3) what are the locations and relative allele frequencies of the sites that define multisite genotypes that significantly improve the statistical explanation of variation beyond that provided by the genotypes defined by the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles, separately for each of the six gender-population strata. This study establishes that the use of only genotypes defined by the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles gives an incomplete picture of the contribution that the variation in the *APOE* gene makes to the statistical explanation of interindividual variation in blood measurements of lipid metabolism. The addition of variable sites to the genotype definition significantly improved the ability to explain variation in lnApoE and in TC and resulted in the explanation of variation in HDL-C and in lnTG. The combination of additional sites that explained the greatest amount of trait variation was different for different traits and varied among the six gender-population strata. The role that noncoding variable sites play in the explanation of pleiotropic effects on different measures of lipid metabolism reveals that both regulatory and structural functional variation in the *APOE* gene influences measures of lipid metabolism. This study demonstrates that resequencing of the complete gene in a sample of ≥ 20 individuals and an evaluation of all combinations of the identified variable sites, separately for each population and interacting environmental context, may be necessary to fully characterize the impact that a gene has on variation in related traits of a metabolic system.

Introduction

Apolipoprotein E (ApoE) is an integral surface component of triglyceride (TG)-rich chylomicrons, chylomicron

remnants, very-low-density lipoprotein (VLDL), and high-density lipoprotein (HDL) and is involved in the receptor-mediated metabolism of these particles (Davignon et al. 1988; Mahley 1988; Davignon 1993; Mahley and Rall 2000). The ApoE molecule has three common isoforms—E2, E3, and E4—that can be identified by PAGE (Utermann et al. 1977). The observed variation in the isoelectric points of these isoforms is attributable to cysteine-arginine variation at positions 112 and 158 in the 299-amino-acid chain of the ApoE molecule. These two variations are encoded by variable sites in exon 4 (positions 3937 and 4075; fig. 1) of the *APOE* gene (MIM 107741), which is located on chromosome 19q13. Combinations of variations at these two

Received February 25, 2002; accepted for publication June 6, 2002; electronically published August 5, 2002.

Address for correspondence and reprints: Dr. Charles F. Sing, Department of Human Genetics, University of Michigan, 1241 East Catherine Street, 5928 Buhl Building, Ann Arbor, MI 48109-0618. E-mail: csing@umich.edu; or Dr. Jari H. Stengård, National Public Health Institute, Department of Epidemiology and Health Promotion, Mannerheimintie 166, FIN-00300 Helsinki, Finland. E-mail: jari.stengard@ktl.fi

© 2002 by The American Society of Human Genetics. All rights reserved.
0002-9297/2002/7103-0006\$15.00

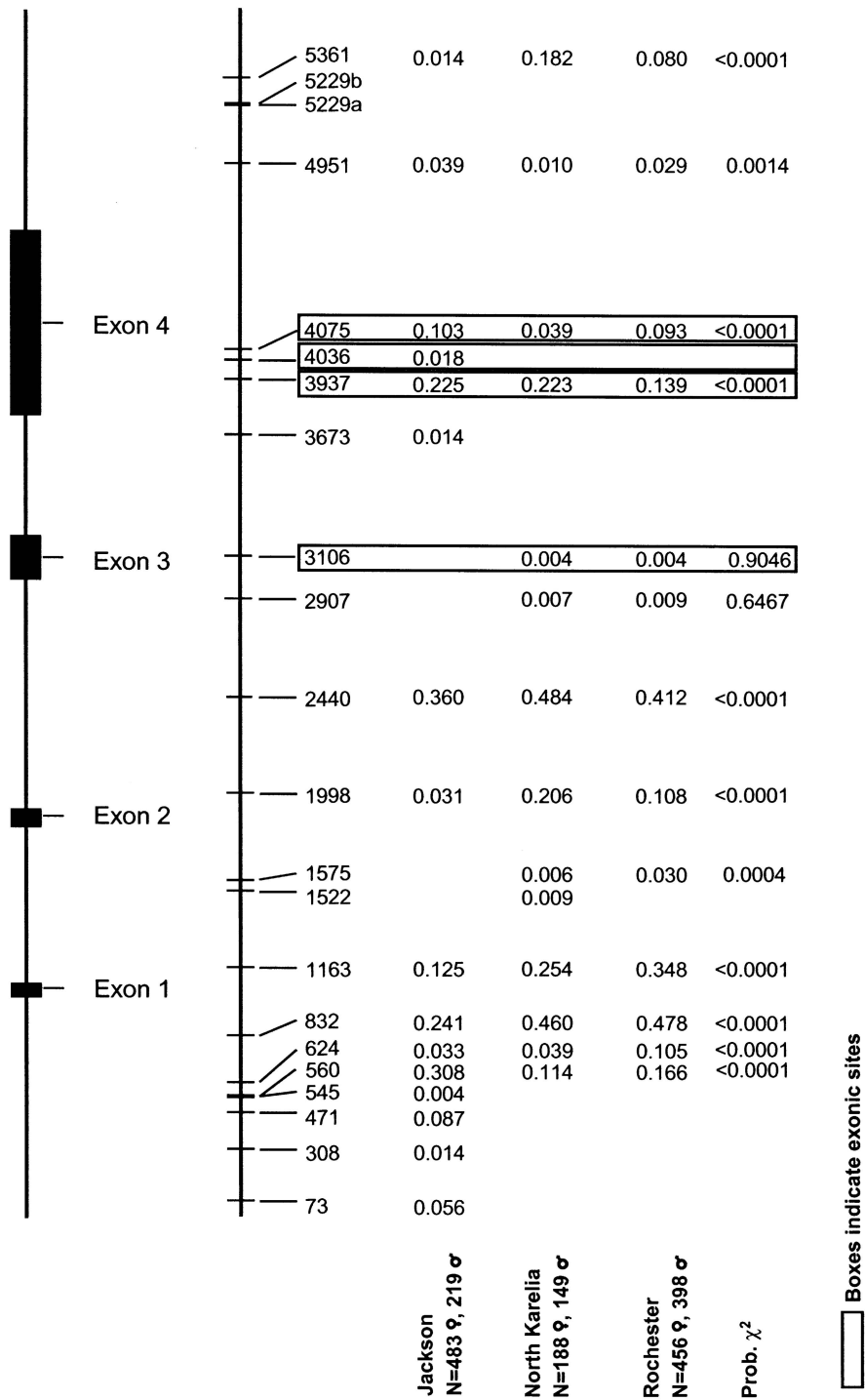


Figure 1 Genomic structure and locations of SNPs in *APOE*, for each population stratum

variable sites define three alleles—denoted “ $\epsilon 2$,” “ $\epsilon 3$,” and “ $\epsilon 4$,”—that encode the three common isoforms. The designations are historical, and the fourth possible combination of these two biallelic sites has not been observed.

Association studies, which have focused mainly on quantitative blood measures of lipid metabolism, have estimated that, in the population at large, ~12%–20% of the interindividual variation in the ApoE level, as well as 5%–8% of the interindividual variation in total cholesterol (TC) level, is associated with variation among the six genotypes defined by the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles (Sing and Davignon 1985; Boerwinkle and Utermann 1988; Davignon et al. 1988; Kaprio et al. 1991; Xhignesse et al. 1991). Furthermore, individuals with the $\epsilon 4$ allele are reported to suffer cardiovascular diseases (CVD) more often than do individuals without this allele (Davignon et al. 1988; Kuusi et al. 1989; Davignon 1993; Wald et al. 1994; Stengård et al. 1995). More recently, the six APOE genotypes have been shown to also influence interindividual variation in both the risk of Alzheimer’s disease (Corder et al. 1993; Mahley and Huang 1999) and the progression of neurological disease in those infected with HIV (Corder et al. 1998). There is also evidence that the six APOE genotypes influence interindividual variation in TG and/or HDL cholesterol (HDL-C) levels in some strata (Nelson et al. 2001; Lussier-Cacan et al. 2002).

In a recent study that resequenced 5.5 kb of the APOE gene, including related 5′- and 3′-flanking regions, we identified 21 variable sites in a sample of 72 unrelated individuals, composed of 24 individuals from each of three populations: Jackson, Miss.; North Karelia, Finland; and Rochester, Minn. (Fullerton et al. 2000; Nickerson et al. 2000) (fig. 1). Twenty of the observed sites are biallelic single-nucleotide polymorphisms (SNPs), and one site varied for a multiallelic insertion/deletion polymorphism (site 5229a) and a SNP (site 5229b). Variation at site 5229 was not measured for the study reported here because the nested nature of variation at this site rendered high-throughput SNP genotyping impractical. Of the 21 SNPs, 17 are located in noncoding regions (i.e., are noncoding SNPs [non-cSNPs]), and 4 are located in a coding region (i.e., are coding SNPs [cSNPs]). In addition to the two cSNPs in exon 4 that encode the three common isoforms, a third cSNP is located between these two (position 4036; fig. 1), and the fourth cSNP is located in exon 3 (position 3106; fig. 1). Variations in each of the four cSNPs result in amino acid changes. Of the 20 SNPs included in the study reported here, 16 were variable in the Jackson sample, 14 were variable in the North Karelia sample, and 13 were variable in the Rochester sample. The 10 that varied in all three samples included the two cSNPs that encode the three common ApoE isoforms. This ob-

servations is consistent with previous studies that have established that the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles are found in most populations (Davignon et al. 1988; Hallman et al. 1991; Gerdes et al. 1992; Stengård et al. 1998).

The availability of 18 additional APOE SNPs leads to the question of whether the measurement of one or more of these additional SNPs can improve our ability to statistically explain interindividual variation in quantitative measures of lipid metabolism—and ultimately to explain variation in the risk for CVD. To address this question in the present article, we ask the following, for each gender, in samples from each of three populations: (1) How much trait variation is associated with variation among the genotypes defined by the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles encoded by the 3937 and 4075 cSNPs? (2) How much additional trait variation is associated with genotypes defined by the combination of the 3937 and 4075 cSNPs with one or more additional SNPs? (3) What are the locations and the relative allele frequencies of the sites in the subset of sites that combine with sites 3937 and 4075 to significantly improve the statistical explanation of trait variation?

Material and Methods

Samples

The Jackson sample included 702 unrelated African Americans (483 females and 219 males) who were 45–64 years of age. The participants were measured for fasting (12 h) blood levels of ApoE, of TC, of HDL-C, and of TG, and samples were genotyped for all of the 16 APOE SNPs that were found to segregate in Jackson (Fullerton et al. 2000). Participants were examined according to a standardized protocol and were part of the ongoing GENOA study, which was designed to identify and characterize genes that influence the risk for essential hypertension (FBPP Investigators 2002). For the measurements of weight and height, participants wore lightweight clothing and removed their shoes.

The North Karelia sample included 337 unrelated Europeans (188 females and 149 males) who were 45–64 years of age. The participants were measured for blood levels of TC, of HDL-C, and of TG, and samples were genotyped for all of the 14 APOE SNPs that were found to segregate in North Karelia (Fullerton et al. 2000). A subsample of 162 females and 124 males were measured for ApoE level. In connection with an ongoing prospective study (the population-based FINRISK study), participants were examined, between 11:00 A.M. and 6:00 P.M., according to a standardized protocol (Salomaa et al. 1994; Vartiainen et al. 1994), which was designed to evaluate the utility of established risk factors as predictors of CVD in middle-aged female and male Finns. Participants were told to avoid fatty meals and

to fast for ≥ 4 h prior to the examination and blood sampling. The length of fast and the type of previous meal were recorded. The initial analysis of these recordings indicated good compliance with the instructions (Salomaa et al. 1994). For the measurements of weight and height, participants wore lightweight clothing and removed their shoes.

The Rochester sample included 854 unrelated non-Hispanic European Americans (456 females and 398 males) who were 45–64 years of age and who were from the middle generation of 583 three-generation pedigrees that were recruited, between December 1984 and July 1991, by the Rochester Family Heart Study. These pedigrees were ascertained through school-age children, regardless of the health status of the children, parents, or grandparents (Moll et al. 1986; Turner et al. 1989). Participants were measured for fasting (12 h) blood levels of ApoE, of TC, of HDL-C, and of TG, and samples were genotyped for all of the 13 APOE SNPs that were found to segregate in Rochester (Fullerton et al. 2000). For the measurements of weight and height, participants wore lightweight clothing and removed their shoes.

Laboratory Methods

Quantitative measures of blood-lipid and ApoE levels for the Jackson and Rochester samples were taken at the Mayo Clinic (Rochester, MN), by published methods (National Institutes of Health 1974; Barr et al. 1981; Kaprio et al. 1991). Those for the Finnish sample were measured at the Department of Biochemistry, National Public Health Institute, Helsinki, by standard enzymatic assays (Boehringer Mannheim Diagnostics) (Salomaa et al. 1994; Schiele et al. 2000). The methods used to genotype the APOE SNPs have been described by Nickerson et al. (2000).

Statistical Methods

In all three samples, the distributions of levels of ApoE and TG were significantly positively skewed in both genders. The natural log (ln) transformation of these variables reduced skewness to nonsignificant values < 0.5 in each case. These transformed variables were used in the analyses presented here to accommodate statistical tests that assume normality.

We used Fisher's F ratio, to test whether there was a statistically significant difference, between female and male participants, in the phenotypic variance of a trait (Sokal and Rohlf 1995). Student's t test was used to test the statistical significance of the difference between gender means when the F ratio was not significant, and Satterthwaite's modification of the t test (Sokal and Rohlf 1995) was used when the F ratio was significant. Permutation tests were also done to assess statistical significance (details are given below).

The Boerwinkle and Sing (1986) bias-corrected estimator of genetic variance, σ_G^2 , was used to measure the utility, for the statistical explanation of phenotypic variability, of different sets of multisite genotypes. It estimates the component of a trait's phenotypic variance, in the population at large, that is attributable to deviations of genotype means from the population mean, weighted by the relative frequencies of the genotypes. The estimator is as follows:

$$\begin{aligned} S_G^2 &= \sum_{i=1}^k \frac{n_i (\bar{Y}_i - \bar{Y})^2}{n} - \frac{k-1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_i)^2}{n-k} \\ &= \frac{SS_G}{n} - \frac{k-1}{n} MS_W, \end{aligned}$$

where n is the total sample size, k is the observed number of genotype classes, \bar{Y} is the sample grand mean, n_i is the number of individuals with the i th genotype class, \bar{Y}_i is the sample mean of the i th genotype, Y_{ij} is the phenotype of the j th individual in the i th genotype class, and MS_W is the mean-squared estimate of the phenotypic variability among individuals within genotype classes in the population at large. This statistic adjusts the genotypic sum of squares, SS_G , from the one-way analysis of variance by a quantity that increases with the number of genotypes if the estimate of the mean square within genotypes, MS_W , does not decrease as additional genotypes are considered.

Permutation methods (Fisher 1935; Lehmann 1986; Edgington 1995; Good 2000) were used (1) to test the statistical significance of variation among genotype-class means ($H_0: \sigma_G^2 = 0$) and (2) to test whether the addition of one or more SNPs to the definition of genotype classes, beyond the two cSNPs (3937 and 4075) that define the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles, significantly increases σ_G^2 in the gender-specific population being sampled. To test whether σ_G^2 is significantly greater than 0, we generated a null distribution of 1,000 S_G^2 values from 1,000 samples, each of which was obtained by randomly reassigning the observed phenotypes to individuals without replacement. The null hypothesis is rejected for an α -test criterion if the original S_G^2 value obtained from the non-permuted sample exceeds the S_G^2 value associated with the upper α percentage of the distribution of values obtained from the 1,000 permutations (Good 2000). We considered 1,000 permuted samples because it has been suggested that this number is sufficient to give a stable estimate of the critical value for an α -test criterion (Churchill and Doerge 1994).

To test whether the addition of sites significantly increased the σ_G^2 estimate, we computed the difference, ΔS_G^2 , between the S_G^2 value associated with the variation in the six APOE genotypes defined by the 3937 and

Table 1**Phenotypic Distributions of Anthropometric Characteristics and Measures of Lipid Metabolism in Each Gender-Population Stratum**

	AGE			HEIGHT			WEIGHT			BMI				
	Mean (years)	P		Mean (cm)	P		Mean (kg)	P		Mean (kg/m ²)	P			
		Variance (years)	Mean		Variance	Mean		Variance	Mean		Variance	Mean	Variance	
Jackson, Miss., sample: Female participants (N = 483)	56.17	101.04	0.044	0.531	164.51	38.68	<0.0001	86.42	339.10	0.381	0.006	31.98	48.07	<0.0001
Male participants (N = 219)	57.84	108.40			178.27	41.82		87.60	244.63			27.52	20.06	
North Karelia, Finland, sample: Female participants (N = 188)	54.61	33.51	0.683	0.290	160.25	32.31	<0.0001	72.24	192.37	<0.0001	0.872	28.18	30.50	<0.0001
Male participants (N = 149)	54.34	39.47			172.21	44.89		83.54	187.44			28.13	16.42	
Rochester, Minn., sample: Female participants (N = 456)	48.46	93.09	0.609	0.025	163.78	33.02	<0.0001	69.20	189.12	<0.0001	0.981	25.83	26.60	<0.0001
Male participants (N = 398)	48.14	74.84			177.44	38.33		86.84	188.65			27.59	17.84	
			InTG			HDL-C			TC			InApoE		
	Mean (mg/dl)	Variance (mg/dl)	P		Mean (mg/dl)	P		Mean (mg/dl)	P		Mean (mg/dl)	P		
Jackson, Miss., sample: Female participants (N = 483)	4.85	0.18	0.111	0.272	54.80	252.11	<0.0001	207.96	2,226.40	0.004	0.018	1.60	0.10	0.188
Male participants (N = 219)	4.91	0.20			47.03	312.98		197.91	1,680.22			1.55	0.12	
North Karelia, Finland, sample: Female participants (N = 188 ^a)	4.77	0.24	0.002	0.370	58.30	163.45	<0.0001	230.53	1,731.41	0.300	0.448	1.34	.08	0.142
Male participants (N = 149 ^b)	4.94	0.28			50.26	177.12		235.14	1,536.47			1.38	0.07	
Rochester, Minn., sample: Female participants (N = 456)	4.56	0.23	<0.0001	0.165	51.74	192.70	<0.0001	192.90	1,412.19	0.009	0.127	1.62	0.10	0.428
Male participants (N = 398)	4.82	0.26			39.64	105.45		199.46	1,217.03			1.63	0.11	

NOTE.—For all *P* values shown, the mean was determined by *t* test, and the variance was determined by *F* test.^a For InApoE, N = 162.^b For InApoE, N = 124.

Table 2

Statistically Significant Estimates of Percent Trait Variation Explained by Two-Site APOE Genotypes and by the Best Sets of Three-Site and Multisite APOE Genotypes

	TWO-SITE GENOTYPES		BEST SET OF THREE-SITE GENOTYPES				BEST SET OF MULTISITE GENOTYPES		
	Estimate A (% variance)	n_G^a	Estimate B (% variance)	Added SNP	n_G^a	Test of Improvement, ^b B vs. A	Estimate C (% variance)	n_G^a	Test of Improvement, ^c C vs. B
Jackson, Miss., sample:									
Female participants (N = 483):									
lnTG	4.3**	560	17		9.8**	111	*
HDL-C	3.0**	624	12		8.3**	77	*
TC	4.0***	6	4.9***	1163	10		9.1*	94	
lnApoE	15.5***	6	19.3***	4036	7	***	23.0***	112	
Male participants (N = 219):									
lnTG	12.1*	56	
HDL-C	9.9*	26	
TC	3.2*	6	7.9**	560	13	**	15.4**	38	*
lnApoE	14.6***	6	24.6***	4036	9	***	30.2***	37	*
North Karelia, Finland, sample:									
Female participants (N = 188 ^d):									
lnTG
HDL-C	5.5*	560	11		15.3***	38	**
TC	4.3*	6	4.3*	1575	7		6.0*	15	
lnApoE	11.1***	6	13.9***	2440	11	*	23.4***	31	*
Male participants (N = 149 ^e):									
lnTG	4.2*	1575	6		8.2*	28	
HDL-C
TC	5.3*	5	6.7*	560	10		14.5**	29	*
lnApoE	8.8**	5	17.2***	1575	6	***	24.9***	23	
Rochester, Minn., sample:									
Female participants (N = 456):									
lnTG
HDL-C	2.6*	560	14	
TC	3.2**	6	3.9**	5361	10		7.7*	82	
lnApoE	20.0***	6	20.4***	2907	8		20.8***	51	
Male participants (N = 398):									
lnTG	3.5*	28	
HDL-C
TC	2.6**	6	3.5*	624	13		5.9*	38	
lnApoE	12.7***	6	16.3***	832	14		23.3***	39	**

NOTE.—The two-site genotypes are defined by sites 3937 and 4075; the three-site genotypes are defined by the combination of sites 3937 and 4075 with one additional site; and the multisite genotypes are defined by the combination of sites 3937 and 4075 with multiple additional sites.

* $\alpha \leq 0.05$.

** $\alpha \leq 0.01$.

*** $\alpha \leq 0.001$.

^a Number of observed genotypes.

^b The best three-site genotype model is compared to the two-site genotype model defined by only the 3937 and 4075 cSNPs, if both explained a statistically significant amount of interindividual trait variation.

^c The best statistically significant multisite genotype model (“Estimate C”) is compared to the best statistically significant three-site genotype model (“Estimate B”).

^d For lnApoE, N = 162.

^e For lnApoE, N = 124.

4075 cSNPs and the $\Delta\sigma_G^2$ value associated with the genotypes defined by the array of SNPs being considered (which always included the 3937 and 4075 cSNPs). The null hypothesis (i.e., $H_0: \Delta\sigma_G^2 = 0$) of no improvement in the σ_G^2 estimate was rejected for an α -test criterion if the original $\Delta\sigma_G^2$ value obtained from the nonpermuted sample exceeded the $\Delta\sigma_G^2$ value associated with the upper

α percentage of the distribution of values obtained from 1,000 permuted samples. Each permuted sample was created by the random shuffling of the observed phenotypes among genotypic classes within each of the six APOE genotype classes defined by the 3937 and 4075 sites. The expected $\Delta\sigma_G^2$ value is influenced by (1) the separate marginal effects of the genotypes defined by

additional SNPs and (2) statistical interactions between the effects associated with the genotypes defined by the added SNPs and the effects associated with the six genotypes defined by sites 3937 and 4075. Rejection of the null hypothesis indicates that the genotypes defined by the additional SNPs have marginal effects that are independent of the effects of the six genotypes and/or the effects of genotypes defined by the additional SNPs are heterogeneous among the APOE genotypes. It is not possible to distinguish between these two possibilities when there is linkage disequilibrium between one or more of the additional sites and site 3937 and/or site 4075. Unless denoted otherwise, we considered the 0.05 level of probability as the criterion for significance of a test statistic.

Results

Gender-Specific Distributions of Quantitative Measures of Body Size and Lipid Metabolism

A description of the female and male samples from each of the three populations is given in table 1. Within each of the three populations sampled, the age distributions and at least two of the three anthropometric characteristics—height, weight, and/or BMI (in kg/m²)—were significantly different between female and male participants. Similarly, the distributions of at least two of the four quantitative measures of lipid metabolism were significantly different, between genders, within each of the three samples. These observations are consistent with widely recognized gender and ethnic differences that document the rationale for stratified analyses of the contribution that genetic variability makes to interindividual variation in trait variability presented below, in which the strata are defined by gender and population sampled.

Population-Specific Distributions of the Relative Allele Frequencies

The relative frequency of the least common allele at a variable site ranged from 0.004 to 0.484 (fig. 1). Five of the 20 SNPs had an extremely rare allele. Their relative frequencies were <0.01, and they segregated in only one (Jackson [at site 545] and North Karelia [at sites 1522 and 1575]) or two (North Karelia and Rochester [at sites 2907 and 3106]) samples. One of the cSNPs (site 3937) involved in the definition of the ϵ_2 , ϵ_3 , and ϵ_4 alleles is in the middle of the distribution of relative frequencies in all three populations (ranges between 0.139 and 0.225), whereas the other (site 4075) is in the lower tail of this distribution (ranges between 0.039 and 0.103). The relative frequencies of the least common allele for the other two cSNPs (site 4036 in the Jackson sample and site 3106 in the North Karelia and Rochester

samples) were also in the lower tail of the distribution. Two of the non-cSNPs (site 832 in the promoter region and intronic site 2440) have the most common alleles in all three samples. Common allelic variations were also observed in promoter-region sites 560 and 624 and intronic sites 1163 and 1998, in all three samples.

Relative allele frequencies were significantly different ($P < 0.05$) among samples for all but two sites, 2907 and 3106 (fig. 1). The largest interpopulation difference in the relative frequency of the least common allele is seen at site 5361, where it ranges from 0.182 (in the North Karelia sample) to 0.014 (in the Jackson sample). The smallest interpopulation difference is seen at site 2440, where the relative frequency of the least common allele ranges from 0.360 (in the Jackson sample) to 0.482 (in the North Karelia sample). The interpopulation differences in the relative frequency of the common allele at sites 3937 and 4075 are between these two extremes—ranging, respectively, from 0.225 and 0.103 (in the Jackson sample) to 0.139 and 0.093 (in the Rochester sample) and 0.223 and 0.039 (in the North Karelia sample). The observed population differentiation of allele frequencies further justifies stratified analyses of the contribution that genetic variability makes to the statistical explanation of interindividual trait variability presented below (see “Contribution of the APOE Genotypes Defined by the 3937 and 4075 cSNPs to a Statistical Explanation of Interindividual Variation in Quantitative Measures of Lipid Metabolism in Each Gender-Population Stratum”), where strata are defined by the populations sampled.

Contribution of the APOE Genotypes Defined by the 3937 and 4075 cSNPs to a Statistical Explanation of Interindividual Variation in Quantitative Measures of Lipid Metabolism in Each Gender-Population Stratum

The estimates of the amount of interindividual variation in lnApoE and TC levels explained by the two-site genotype model defined by the combination of the 3937 and 4075 cSNPs were statistically significant in all six gender-population strata (table 2, under “Estimate A”). In female participants, 11%–20% of the variation in lnApoE, as well as 3%–4% of the variation in TC level, was explained by variability among the genotypes defined by these two cSNPs. In male participants, estimates ranged from 9% to 15% of variation in lnApoE and from 3% to 5% of variation in TC. Genotypes defined by the 3937 and 4075 cSNPs did not explain a significant amount of interindividual variation in HDL-C or lnTG, in any of the gender-population strata.

Consequences of Combining One SNP with the 3937 and 4075 cSNPs for Statistically Explaining Interindividual Variation in Quantitative Measures of Lipid Metabolism in Each Gender-Population Stratum

For each trait in each gender-population stratum, we determined the sets of three-site genotypes (defined by the combination of one additional SNP with the 3937 and 4075 cSNPs) that explained the greatest amount of trait variation and that were statistically significant. These sets are presented in table 2, under “Estimate B.”

In the Jackson sample, two different SNPs combine with the 3937 and 4075 cSNPs to define three-site genotypes (table 2, under “Estimate B”) that explained significantly larger proportions of interindividual variation in lnApoE and TC than the APOE genotypes encoded by the 3937 and 4075 cSNPs (table 2, under “Estimate A”). One set, defined by the 3937, 4036, and 4075 cSNPs, improved the statistical explanation of lnApoE variation in both female (19% vs. 16%) and male (25% vs. 15%) participants. A second set of genotypes, defined by the combination of site 560 with the 3937 and 4075 cSNPs, improved the statistical explanation of TC variation in male (8% vs. 3%) participants. In addition, we found that, in female participants, the combination of sites 560 or 624 with the 3937 and 4075 cSNPs explained a significant amount of lnTG (4%) and HDL-C (3%) variation, respectively, whereas the genotypes defined by sites 3937 and 4075 did not.

In the North Karelia sample, two different SNPs combined with the 3937 and 4075 cSNPs to define three-site genotypes that gave a significant improvement in the statistical explanation of variation in lnApoE when compared to the genotypes defined by the 3937 and 4075 cSNPs. The set defined by the addition of site 2440 improved the statistical explanation of lnApoE variation in female (14% vs. 11%) participants, whereas the set defined by the addition of site 1575 improved the explanation of lnApoE variation in male (17% vs. 9%) participants. None of the sets of three-site genotypes provided statistically significant improvement in the explanation of variation in TC in either gender. In addition, we found that the combination of site 560 with the 3937 and 4075 cSNPs significantly explained 6% of HDL-C variation in females and that the combination of site 1575 with these two-sites significantly explained 4% of lnTG variation in male participants, whereas the genotypes defined by sites 3937 and 4075 did not statistically explain variation in either case.

In the Rochester sample, none of the sites in combination with sites 3937 and 4075 defined three-site genotypes that explained significantly larger proportions of interindividual variation in lnApoE or TC in either gender than did the APOE genotypes encoded by sites 3937 and 4075. In female participants, we found that

genotypes defined by the combination of site 560 with the 3937 and 4075 cSNPs explained a significant fraction (3%) of HDL-C variation, whereas the genotypes defined by sites 3937 and 4075 did not.

Consequences of Combining Two or More SNPs with the 3937 and 4075 cSNPs for Statistically Explaining Interindividual Variation in Quantitative Measures of Lipid Metabolism in Each Gender-Population Stratum

The best sets of multisite genotypes defined by the combination of sites with the 3937 and 4075 cSNPs that explained a statistically significant amount of trait variation are given, in table 2 (under “Estimate C”), for each trait in each gender-population stratum. The site that was combined with the 3937 and 4075 cSNPs to define the best set of three-site genotypes that explained a statistically significant amount of trait variation (table 2, under “Estimate B”) was always found to be included in the set of sites that were combined with the 3937 and 4075 cSNPs to define the best multisite genotype model that explained a significant amount of trait variation (figs. 2A and 2B).

In the Jackson sample, there was no statistically significant evidence that the addition of sites beyond those included in the best three-site model improves the explanation of lnApoE or TC (table 2, under “Test of Improvement, C vs. B”) in female participants. In male participants, two different additional sets of sites defined multisite genotypes that explained significantly more variation in lnApoE or TC than the best three-site combinations did (fig. 2B). The magnitude of improvement in the explanation of the lnApoE variation by the first set of multisite genotypes was modest (30% vs. 25%), whereas the amount of TC variation that was explained by a second set (15%) was almost twice the amount explained by the best set of three-site genotypes (8%). In female participants, there was one set of multisite genotypes that explained significantly more variation in lnTG, as well as another set that explained significantly more variation in HDL-C (fig. 2A), than that obtained using the best sets of three-site genotypes. Furthermore, in male participants, we found one set of multisite genotypes that significantly explained 12% of variation in lnTG and a second set that significantly explained 10% of variation in HDL-C (fig. 2B), whereas neither the APOE genotypes defined by sites 3937 and 4075 nor any of the three-site genotypes defined by the combination of a site with the 3937 and 4075 cSNPs explained a significant amount of variation.

In the North Karelia sample, one set of multisite genotypes significantly improved the explanation of variation in lnApoE in female participants (23% vs. 14%), and a second set of multisite genotypes significantly improved the explanation of TC variation in male partic-

ipants (15% vs. 7%). A third set of multisite genotypes gave a significant improvement in the explanation of HDL-C variation in females compared to the best set of three-site genotypes (15% vs. 6%). No combination of sites explained a significant amount of variation in either HDL-C, in male participants, or lnTG, in female participants.

In the Rochester sample, none of the sets of multisite genotypes significantly improved the explanation of trait variation in female participants. One multisite combination significantly improved the explanation of lnApoE variation in male participants. Furthermore, we found a set of multisite genotypes that significantly explained 4% of lnTG variation in male participants, whereas neither the APOE genotypes defined by sites 3937 and 4075 nor any of the three-site genotypes defined by the combination of a site with the 3937 and 4075 cSNPs explained a significant amount of variation.

Discussion

One promise of the postgenomic era is that fast and inexpensive gene-measurement technologies will empower medical and public health care delivery systems to incorporate variation at the DNA level into algorithms for the identification of individuals and populations that are at increased risk for common chronic diseases that have a complex multifactorial etiology (Sing et al. 1996). At present, the widely accepted research strategy for the evaluation of the utility that DNA information has in risk assessment considers only one or two variable sites in each candidate gene. In line with this research strategy, nearly all earlier studies of the APOE gene have considered only the two cSNPs (3937 and 4075) that encode nonsynonymous amino acid changes that were discovered as a consequence of the characterization of electrophoretic variations of the gene product. This top-down strategy is the typical approach for the selection of measures of gene variation. It uses prior knowledge about the biology of intermediate biochemical and physiological traits that connect genome variation with variation in the risk for disease. Variations are selected because they change the amino acid sequence or because they mark an established promoter region. Such a top-down strategy de-emphasizes the study of the possible pleiotropic effects of the gene and ignores the fact that phenotypic effects of each variable DNA site are contingent on the context defined by other variable sites in the gene and interacting agents, including other genes and exposures to environments both internal and external to the organism. Furthermore, the impact that heterogeneity has, among populations, in both the structure of gene variation (wherein it is defined by the number of alleles and the relative allele and genotype frequencies) and genetic architecture (wherein it is defined

by the contribution that variations in the gene make to trait variation) cannot be evaluated unless all variable sites within the gene are considered. Hence, to obtain a comprehensive, bottom-up evaluation of the contribution that variation in a candidate gene makes to variation in intermediate biochemical and physiological traits—and, ultimately, to risk for disease—the contribution of genotypes determined by all known variable sites must be evaluated in samples that are representative of the interacting environmental contexts and populations of interest. To do otherwise requires assumptions about gene variation and homogeneity of gene effects that are likely to be false and that cannot be tested.

Our expectation that noncoding variable sites, in different combinations with variation at sites 3937 and 4075, will have utility in the explanation of phenotypic variation in different contexts defined by gender and population was the motivation for the study reported here. This bottom-up strategy for the evaluation of the contributions of additional variable DNA sites began with the resequencing of a sample of individuals ascertained without regard to health, to characterize the variation in the APOE gene in each population. We then asked whether we can statistically explain significantly greater variation in measures of lipid metabolism in samples of six gender-population strata by combining variation at the 3937 and 4075 sites with 1 or more of the 18 additional SNPs revealed by the initial resequencing step. Our study suggests that the commonly held assumptions, which are based on studies of variation in sites 3937 and 4075, that APOE gene variation and the connection between this variation and phenotypic variation in measures of lipid metabolism are homogeneous among populations should be questioned.

The Role of Gender and Population Structure in the Definition of the Population of Inference

Ignoring the gene variations that do not vary in all populations can result in the definition of a population of inference that is not representative of any population. Studies that pool female and male data and data from different populations assume that genotype-phenotype relationships are independent of context indexed by gender and population. Both the heterogeneity observed in the trait distribution between genders within samples from a particular population and among samples from different populations for a particular gender and the heterogeneity of relative allele frequencies among samples from different populations suggest that an analysis of pooled data will produce inferences that are not representative of either gender or of any population. Heterogeneity of genetic structure, defined by relative allele and genotype frequencies, is of particular concern because the contribution that variation in a gene makes to

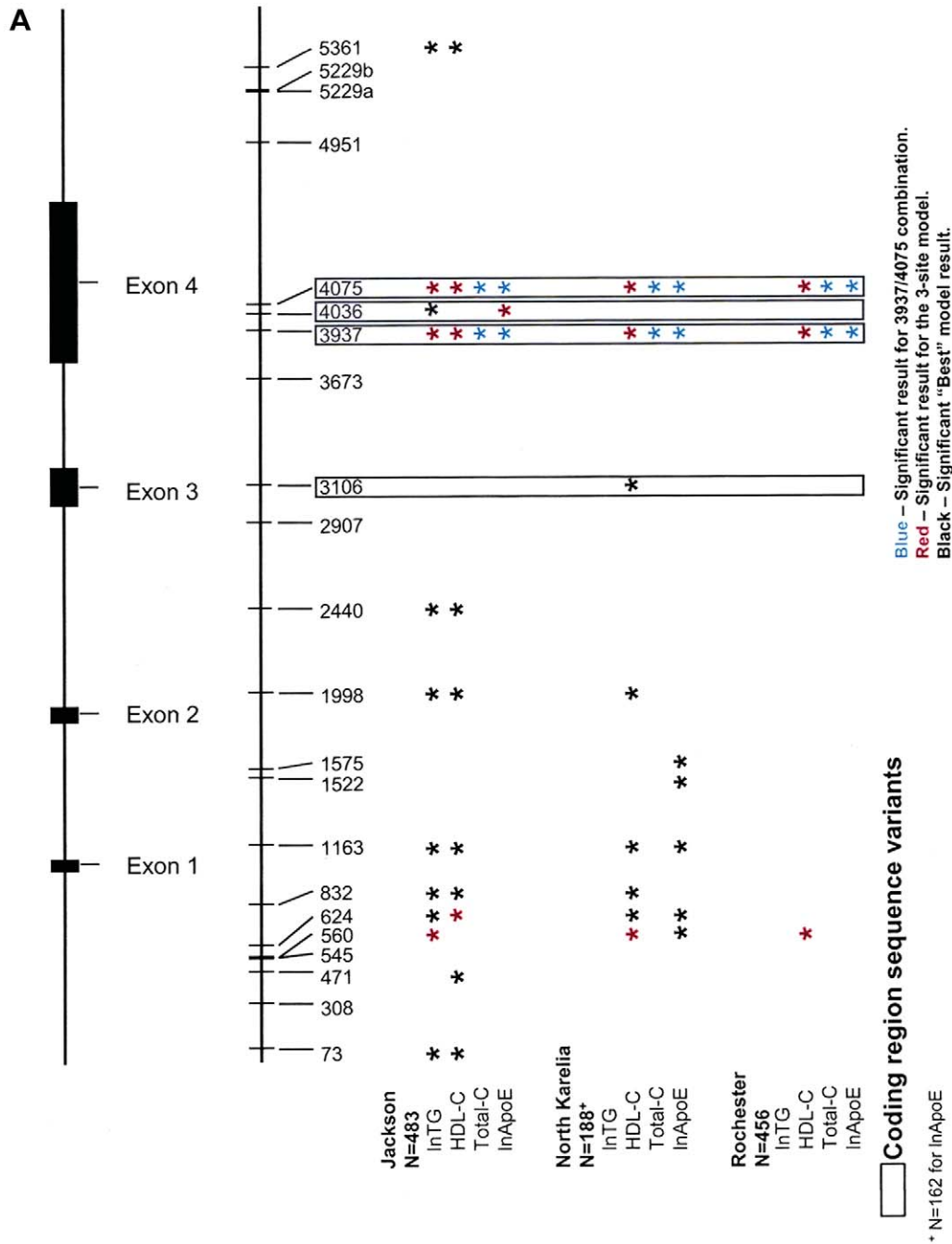
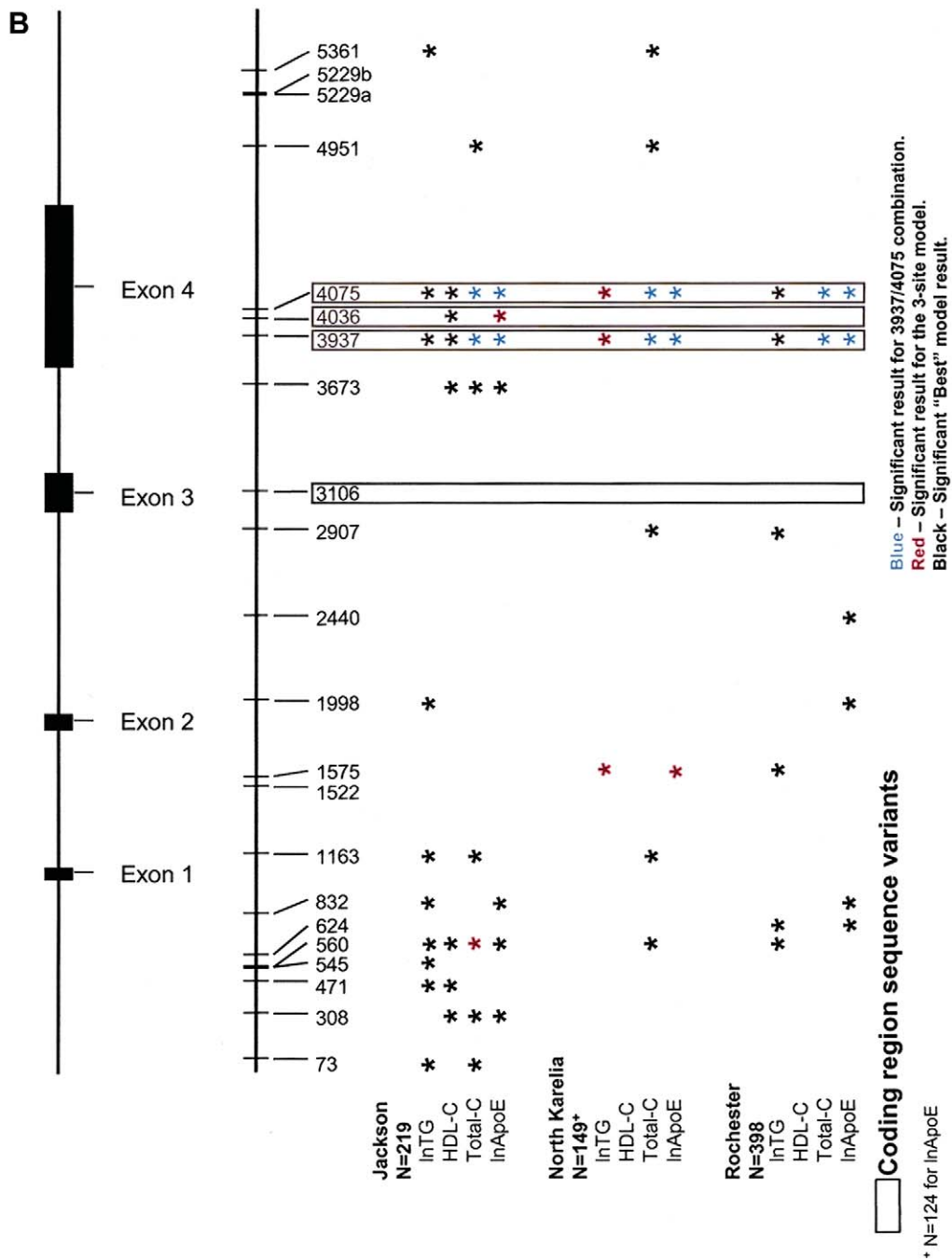


Figure 2 Combinations of variable sites that define the best sets of genotypes—requiring the inclusion of the 3937 and 4075 cSNPs—that explain a statistically significant amount of trait variation, for each of the four lipid traits among females (A) and males (B) from each population stratum.

the genetic architecture of a quantitative trait (Boerwinckle et al. 1986) is determined by deviations of genotype means from the population mean, weighted by the relative frequencies of the genotypes. Obviously, the six sites (four of which are 5' to the first exon) that vary only in the Jackson sample cannot make a contribution

to the statistical explanation of trait variation in the other two samples. Only half of the 20 SNPs that we studied were variable in all three samples. In all of these 10 cases, relative allele frequencies were significantly different among samples—the maximum difference being 0.24 (at site 832, when Jackson and Rochester samples



are compared). These differences in the organization of genetic variation document that these three populations have very different demographic histories (Fullerton et al. 2000). Such heterogeneity in the genetic structure of variation in the gene of interest, as well as in other unmeasured interacting genes, could lead to very different

contributions that a gene makes to the genetic architecture of phenotypic variation in different populations.

The implications of microdifferentiation of relative allele frequencies for the evaluation of the impact that gene variation has on phenotypic variation in human populations have been largely ignored because most candi-

date-gene studies assume that only the public polymorphisms (i.e., those that vary in all populations) are relevant for the prediction of phenotypic variability in any population (Reich et al. 2001). By using only common public variations (i.e., those that have an ancient common ancestor), one assumes that the DNA variations of recent origin, which vary only in particular populations, are of little importance in the determination of phenotypic variation. The potential bias in the evaluation of the influence that the *APOE* gene has on phenotypic variation could be major, because only half of the variable sites were found to segregate in all three samples. Unless all variations, common and rare, are measured, it is not possible to evaluate the consequences of the bias associated with this assumption. Our study documents that this bias can be large. For instance, by consideration of the rare allele (relative frequency 0.018; at site 4036), which varies only in the Jackson sample, it is possible to statistically explain 25% and 68% more \ln ApoE variation in female and male participants, respectively, than is explained by genotypes defined by the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles that occur in all three samples. The significant contribution that a rare variation at site 1575 (which does not vary in the participants from Jackson) makes to \ln ApoE and \ln TG variation in male participants from North Karelia (table 2, under “Estimate B”) serves as a second example.

Experimentwise Error Rate

Our general, global null hypothesis is that none of the sets of multi-SNP genotypes defined by combining one or more additional SNPs with the 3937 and 4075 cSNPs statistically explain a greater proportion of quantitative variation in any of the four measures of lipid metabolism than the proportion that is explained by the genotypes defined by sites 3937 and 4075, in any of the six gender-population samples. The construction of a null distribution for this hypothesis is complicated by the differences across samples from different populations, discussed above (see “The Role of Gender and Population Structure in the Definition of the Population of Inference”), in both trait distributions and relative allele frequencies, and by the expected gender-specific differences among populations in correlations between traits. Furthermore, the fact that different subsets of the *APOE* SNPs segregate in different samples dictates that we cannot exhaustively evaluate the utility of all possible sets of multi-SNP genotypes across the strata defined by population. There will always be sets that could explain trait variation only within a particular population. To our knowledge, when these complications hold, there are no appropriate statistical methods, traditional or otherwise, for the testing of the global null hypothesis against the alternative hypothesis that there is at least one multi-

SNP combination that improves the statistical explanation of trait variation in at least one of the six strata. In the absence of replicate samples of the many populations of inference represented here, an alternative approach—which does not consider a specific statistical model or the heterogeneity in statistical power, to reject the null hypothesis across strata and traits—is the calculation of the probability that the observed significance tests are randomly distributed among the six strata for each trait. Under the assumption of a 0.05 test criterion and equal statistical power among tests, for the three SNP genotype analyses, this probability is <0.01 . Given that, in many of the individual tests, the probability was <0.05 and given the observation that in no case did combining a SNP with the 3937 and 4075 cSNPs improve the statistical explanations of variation in any of the traits in either gender in the Rochester sample (whereas the analyses of the Jackson and North Karelia samples resulted in significant improvements in several comparisons, in both female and male participants), this probability is likely much smaller. Actually, in the Jackson sample, combining a SNP with the 3937 and 4075 cSNPs, to define the best set of three-SNP genotypes, improved the statistical explanation of \ln ApoE variation, in 5,000 permutations in both genders, at $P < 0.0002$, which is less than the Bonferroni $P < 0.0005$ criteria associated with 100 tests. Furthermore, the additional SNPs that defined the best sets (for explanation of variation) of three-SNP genotypes are not randomly distributed across the resequenced region of the gene, but cluster in the 5' regulatory region. Hence, we feel confident in rejecting the global null hypothesis in favor of the alternative hypothesis that there is at least one set of multi-SNP genotypes that statistically explain variation in at least one trait, in at least one strata, better than do the genotypes defined by sites 3097 and 4075. The distribution among gender-population strata of the significant tests of the improvement in the explanation of variation by use of the best combination of all sites with the 3937 and 4075 cSNPs compared to the best combination of three-site genotypes was very similar to the distribution of the significant three-SNP genotype results. In most cases, the combination of sites with the 3937 and 4075 cSNPs significantly improved the statistical explanation of trait variability only in the samples from Jackson and North Karelia. We expect that a replication of our analysis in longitudinal follow-up studies in progress will help to serve as further evidence for or against these conclusions.

Contribution of APOE SNPs to the Genetic Architecture of Measures of Lipid Metabolism

Seventeen years ago, Sing and Davignon (1985) reported the first estimates of the contribution that the

two variable sites in exon 4 that determine the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles make to interindividual variation in measures of lipid metabolism in the population at large. Since that time, hundreds of studies have established that these three alleles vary in most populations (Gerdes et al. 1992), and the six genotypes they define explain a statistically significant fraction of interindividual variation in lipid and lipoprotein levels (Hallman et al. 1991; Kaprio et al. 1991; Xhingnesse et al. 1991). The present study establishes that these two cSNPs do not capture all of the genetic variation, in *APOE*, that influences variation in measures of lipid metabolism. Generally, the addition of sites resulted in improved statistical explanations of variation for every trait in at least one gender-population stratum. Two- and three-fold increases were observed in many cases.

Different regions of the gene statistically explain variation in different components of lipid metabolism. It is of great importance to note that the identification and characterization of pleiotropic effects that the *APOE* gene has on measures of lipid metabolism depend on the measurement of non-cSNP variations. The genotypes defined by only the 3937 and 4075 cSNPs explain significant amounts of variation in lnApoE and TC (figs. 2A and 2B) but not in HDL-C and lnTG. This result is consistent with findings of studies that have measured only the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles (Breslow 2000). The present study suggests that the addition of genotype variation defined by the 5' region of the gene measured by sites 560 and 624 results in a statistically significant explanation of 3%–5% of variation in HDL-C in each of the three female samples. This pleiotropic effect of the *APOE* gene is supported by a significant ($P < 0.05$) marginal effect of either site 560 or the closely linked 624 regulatory site, in each case (Stengård et al. 2000). These effects are of the same magnitude as the influence that variation in sites 3937 and 4075 has on the explanation of variation in TC. The addition of non-cSNP sites also resulted in the definition of genotypes that explained statistically significant amounts of variation in lnTG in particular strata. These results suggest that the nonstructural, quantitative variations in the *APOE* gene product may influence the levels of HDL-C and TG. Furthermore, measuring only cSNPs, a strategy advocated by some investigators, would miss as much as 50% of the variation in ApoE.

The observation that the SNPs that confer additional information about the tendency that interindividual differences in lipid metabolism have to cluster in the regulatory region is consistent with recent laboratory observations (Artiga et al. 1998a, 1998b), as well as with our earlier observation that variation in these sites is involved in the separation, undetected by protein-level variation, of phylogenetically distinct lineages, or clades, of the *APOE* gene tree (Fullerton et al. 2000). Artiga et

al. (1998a, 1998b) have found that sites 560 and 624 combine to define three different haplotypes that have different transcriptional activities in cell cultures, owing to their different affinity for nuclear proteins. Both of these sites, together with sites 832, 1163, and 2440, define the four most common haplotypes that carry the $\epsilon 3$ allele (Fullerton et al. 2000). The difference, between populations, in the relative frequencies of these four haplotypes may explain why different combinations of these SNPs define multisite genotypes that best explain variations in different traits in different population strata. The observation that site 560 is seen more often in these combinations than are the other SNPs may be attributable to its position in the *APOE* gene tree, which suggests that it may be particularly susceptible to recurrent mutations and/or gene conversions, placing it in association with different allelic background, with different functional effect, in different populations. An analysis of the effects that the *APOE* haplotypes have on measures of lipid metabolism are in progress (A. R. Templeton, personal communication).

Pleiotropic effects that the *APOE* gene has on multiple traits in the same metabolic system suggest a complex cause-and-effect genotype-phenotype relationship mediated by posttranslational modifications in the gene product. The observation that such effects occur primarily in female participants further supports the role that unmeasured interacting agents play in the determination of a complex, context-dependent connection between variation in the *APOE* gene and interindividual variation in components of the lipid-metabolism system. These results clearly establish that a top-down strategy that selects variable DNA sites because they influence protein sequence (the cSNPs) can underestimate the role of the gene and miss pleiotropic effects on traits that are influenced by variation in non-cSNPs.

The present study clearly establishes that it takes more than one additional site, beyond sites 3937 and 4075, to capture the contribution that the *APOE* gene makes to the variation in the ApoE levels in the Jackson and North Karelia female and male samples. Furthermore, the present study also suggests that there may be no single best set of sites that captures the variation in a gene that influences variation in a particular trait, in all samples from different gender-population strata. Even though the maximum amount of ApoE and TC variation explained by the combination of multiple sites with the 3937 and 4075 cSNPs is approximately the same (20%–25% and 6%–10%, respectively) across gender-population strata, different combinations of additional sites are involved (figs. 2A and 2B). The differences in sites selected to explain trait variation is a consequence of heterogeneity in the structure of gene variation among populations complemented by heterogeneity in the genotype-phenotype relationships attributable to interac-

tions with environments internal and external to the organism indexed by gender and population. Gender and population are representative of only two of the many strata that may index such interacting environments. Inferences about the utility of any particular SNP—or of combination of SNPs in risk assessment—therefore need to be made with caution because (1) rare allele, haplotype, or genotype variations with large effects can explain variation in only a subset of populations (Weiss and Clark 2002); (2) the effect that a particular SNP has on trait variation may be conditional on the genotype of another SNP(s) either in the *APOE* gene (Hamon et al. 2001) or in another linked or unlinked gene (Templeton 2000; Nelson et al. 2001); (3) the genetic architecture of trait variability may depend on the synergistic effects, of a combination of SNPs (that each may not have separate marginal effects), that will be different in different populations (Hamon et al. 2001); and (4) there are gene \times environment interactions whereby the influence that a SNP has on phenotype level varies among environments (Reilly et al. 1992, 1994; Stengård et al. 1999, 2001; Humphries et al. 2001; Lussier-Cacan et al. 2002). These considerations prompt studies in progress to ask whether there can be invariant criteria for the selection of variable sites in candidate-gene studies.

Utility of Genetic Information in Clinical Practice and Prevention Programs

An ultimate goal of human genetic studies is to provide information that improves our ability to identify individuals who have—or are at increased risk for—a particular disease. Advanced gene-measurement technologies have made an invaluable contribution to the identification and characterization of genetic variations that are involved in the determination of monogenic diseases that are inherited in a Mendelian manner, as well as to the identification of individuals who carry such variations (Guo and Lange 2000). In contrast, use of such technologies for the identification of individuals who are at increased risk for CVD or other common chronic diseases has been less successful. The fact that a DNA sequence variation in a particular gene can be related to variation in a risk-factor trait and/or in the risk for disease in a sample from one population while other variations of the gene, or variations in other gene(s) are found to be associated with variation in the very same traits in other populations has limited the utility of the identified genetic variations in clinical practice and public health programs. Furthermore, the fact that the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles of the *APOE* gene that are present in most populations (Davignon et al. 1988; Gerdes et al. 1992; Davignon 1993; Stengård et al. 1998) explain only a small fraction of interindividual variation in intermediate traits that connect genome variation with

variation in onset, in progression, and in severity of disease also makes them unattractive for clinical and public health applications. The association between an endpoint phenotype and a predictor is expected to be much stronger before it is a worthwhile consideration in risk assessment (Wald et al. 1999).

The present study argues that, for most intermediate quantitative traits whose phenotypes emerge as a consequence of interactions between many genetic and environmental factors, it may not be possible to identify either a particular genetic variant or a particular subset of variants that are specific and/or sensitive enough for an identification of individuals who are at increased risk for common chronic diseases. Different populations or environmental strata within populations may require a different set of genetic-risk indicators. Such a conclusion follows from complexity research that has shown that it is extremely difficult to explain a detailed outcome of a highly interactive system in terms of the behavior of either a particular agent or subsets of agents (Solé and Goodwin 2000)—it may even be theoretically impossible (Axelrod and Cohen 2000). Despite this expectation, complexity research suggests that we can learn much about the functioning of the component interacting agents and about how to use such information in risk assessment if we ask nontraditional questions about genotype-phenotype relationships. For example, rather than asking whether a particular variant site is associated with trait variation, the research strategy employed in the present study empowers one to identify several sets of multi-SNP genotypes of the *APOE* gene that explained variations in HDL-C and in TG that could not be explained by genotypes defined by the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ alleles. Furthermore, we found several sets of multi-SNP genotypes that statistically explained two to three times more TC and lnApoE variation than these genotypes did. All of the identified sets were, however, either gender-specific or population-specific, and, in many cases, they were both gender-specific and population-specific. These results clearly argue that there is a need for a shift in the strategy for the incorporation of genetic variation at the DNA level into algorithms for the identification of individuals and populations that are at increased risk for common chronic diseases that have a complex, multifactorial etiology. Instead of asking which genetic variants are invariant risk indicators of a given trait in all populations, we should be asking which genetic variants are risk indicators of which traits in which populations and strata within a particular population.

Conclusions

Measuring only cSNPs fails to capture all the variation in the *APOE* gene that influences lipid metabolism. Microdifferentiation of *APOE* SNP alleles among popu-

lations limits general inferences that can be made about the impact that gene variation has on trait variation. Variation in many sites may explain variation in a particular trait, and the more variation we measure, the better we can document the pleiotropic effects that the *APOE* gene has on multiple measures of lipid metabolism. Establishing whether there are different sets of sites that explain variation in different populations and environment strata will require confirmation in longitudinal follow-up studies and replicate samples of particular environmental strata of particular populations. The role that non-cSNPs play in the explanation of pleiotropic effects on HDL-C and TG suggests that the *APOE* gene has both regulatory and structural functional effects on lipid metabolism.

Acknowledgments

We wish to thank Kenneth G. Weiss for his persistent attention to the details of the data management and statistical analyses. The technical support of Lynn Illeck and Debbie Theodore in developing this manuscript is also deeply appreciated. This work was supported by National Heart, Lung, and Blood Institute grants HL54481, HL51021, HL39107, HL58238, HL58239, and HL58240 and National Institute of General Medical Science grant GM65509.

Electronic-Database Information

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

Online Mendelian Inheritance in Man (OMIM), <http://www.ncbi.nlm.nih.gov/Omim/> (for APOE [MIM 107741])

References

- Artiga MJ, Bullido MJ, Frank A, Sastre I, Recuero M, García MA, Lendon CL, Han SW, Morris JC, Vazquez J, Goate A, Valdivieso F (1998a) Risk for Alzheimer's disease correlates with transcriptional activity of the *APOE* gene. *Hum Mol Genet* 7:1887–1892
- Artiga MJ, Bullido MJ, Sastre I, Recuero M, García MA, Al-dudo J, Vázquez J, Valdivieso F (1998b) Allelic polymorphisms in the transcriptional regulatory region of apolipoprotein E gene. *FEBS Lett* 421:105–108
- Axelrod R, Cohen MD (2000) *Harnessing complexity*. Basic Books, New York
- Barr SE, Kottke BA, Map SJT (1981) Improved method for determination of triglycerides in plasma lipoproteins by an enzymic kit method. *Clin Chem* 27:1142–1144
- Boerwinkle E, Chakraborty R, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann Hum Genet* 50:181–194
- Boerwinkle E, Sing CF (1986) Bias of the contribution of single locus effects to the variance of a quantitative trait. *Am J Hum Genet* 39:137–144
- Boerwinkle E, Utermann G (1988) Simultaneous effects of the Apolipoprotein E polymorphism on Apolipoprotein E, Apolipoprotein B, and Cholesterol Metabolism. *Am J Hum Genet* 42:104–112
- Breslow JL (2000) Genetics of lipoprotein abnormalities associated with coronary heart disease susceptibility. *Annu Rev Genet* 34:233–254
- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138:963–971
- Corder EH, Robertson K, Lannfelt L, Bogdanovic N, Eggersten G, Wilkins J, Hall C (1998) HIV-infected subjects with the E4 allele for ApoE have excess dementia and peripheral neuropathy. *Nat Med* 4:1182–1184
- Corder EH, Saunders AM, Strittmatter WJ, Schmechel DR, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (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 (1993) Apolipoprotein E polymorphism and atherosclerosis. In: Born GVR, Schwartz CJ (eds) *New horizons in coronary heart disease*. Current Science, London, pp 5.1–5.21
- Davignon J, Gregg RE, Sing CF (1988) Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis* 8:1–21
- Edgington ES (1995) *Randomization tests*, 3rd ed. Marcel Dekker, New York
- FBPP Investigators (2002) Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). *Hypertension* 39:3–9
- Fisher RA (1935) *Design of experiments*. Hafner, New York
- Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengård JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* 67:881–900
- Gerdes LU, Klausen IC, Sihm I, Faergeman O (1992) Apolipoprotein E polymorphism in a Danish population compared to findings in 45 other study populations around the world. *Genet Epidemiol* 9:155–167
- Good P (2000) *Permutation tests: a practical guide to resampling methods for testing hypotheses*, 2nd ed. Springer, New York
- Guo S-W, Lange K (2000) Genetic mapping of complex traits: promises, problems, and prospects. *Theor Popul Biol* 57:1–11
- Hallman DM, Boerwinkle E, Saha N, Sandholzer C, Menzel HG, Csazar A, Utermann G (1991) The apolipoprotein E polymorphism: a comparison of allele frequencies and effects in nine populations. *Am J Hum Genet* 49:338–349
- Hamon SC, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Stengård JH, Sing CF (2001) Evidence for interaction between nucleotides within the *ApoE* gene. *Am J Hum Genet Suppl* 69:183
- Humphries SE, Talmud PJ, Hawe E, Bolla M, Day IN, Miller GJ (2001) Apolipoprotein E4 and coronary heart disease in middle-aged men who smoke: a prospective study. *Lancet* 358:115–119
- Kaprio J, Ferrell RE, Kottke BA, Kamboh MI, Sing CF (1991) Effects of polymorphisms in apolipoproteins E, A-IV, and

- H on quantitative traits related to risk for cardiovascular disease. *Arterioscler Thromb* 11:1330-1348
- Kuusi T, Nieminen MS, Ehnholm C, Yki-Jarvinen H, Valle M, Nikkila EA, Taskinen MR (1989) Apolipoprotein E polymorphism and coronary artery disease: increased prevalence of apolipoprotein E4 in angiographically verified coronary patients. *Arteriosclerosis* 9:237-241
- Lehmann EL (1986) *Testing statistical hypotheses*, 2nd ed. Wiley, New York
- Lussier-Cacan S, Bolduc A, Xhignesse M, Niyonsenga T, Sing CF (2002) Impact of alcohol intake on measures of lipid metabolism depends on context defined by gender, BMI, cigarette smoking and *apolipoprotein E* genotype. *Arterioscler Thromb Vasc Biol* 22:824-831
- Mahley RW (1988) Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. *Science* 240:622-630
- Mahley RW, Huang Y (1999) Apolipoprotein E: from atherosclerosis to Alzheimer's disease and beyond. *Curr Opin Lipidol* 10:207-217
- Mahley RW, Rall SC Jr (2000) Apolipoprotein E: far more than a lipid transport protein. In: Lander E, Page D, Lifton R (eds) *Genomics and human genetics: annual review*, vol 1. Palo Alto, CA, pp 507-538
- Moll PP, Sing CF, Williams RR, Mao SJT, Kottke BA (1986) Genetic determination of plasma Apolipoprotein A-I levels measured by radioimmunoassay: a study of high-risk pedigrees. *Am J Hum Genet* 38:361-372
- National Institute of Health (1974) *Lipid research clinics program manual of laboratory operations*. Publication No. 75-628, Department of Health, Education and Welfare, Washington, DC
- Nelson MR, Kardia SL, Ferrell R, Sing CF (2001) A CPM to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 11:458-470
- Nickerson DA, Taylor S, Fullerton SM, Weiss KM, Clark AG, Stengård J, Salomaa V, Boerwinkle E, Sing CF (2000) Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res* 10:1532-1545
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. *Nature* 411:199-204
- Reilly SL, Ferrell RE, Kottke BA, Kamboh MI, Sing CF (1992) The gender-specific apolipoprotein E genotype influence on the distribution of lipids and apolipoproteins in the population of Rochester, MN. II. Regression relationships with concomitants. *Am J Hum Genet* 51:1311-1324
- Reilly SL, Ferrell RE, Sing CF (1994) The gender-specific apolipoprotein E genotype influence on the distribution of plasma lipids and apolipoproteins in the population of Rochester, MN. III. Correlations and covariances. *Am J Hum Genet* 55:1001-1018
- Salomaa VV, Rasi V, Pekkanen J, Vahtera E, Jauhiainen M, Vartiainen E, Myllylä G (1994) Haemostatic factors and prevalent coronary heart disease: the FINRISK haemostasis study. *Eur Heart J* 15:1293-1299
- Schiele F, De Bacquer D, Vincent-Viry M, Beisiegel U, Ehnholm C, Evans A, Kafatos A, Martins MC, Sans S, Sass C, Visvikis S, De Backer G, Siest G (2000) Apolipoprotein E serum concentration and polymorphism in six European countries: The ApoEurope Project. *Atherosclerosis* 152:475-488
- Sing CF, Davignon J (1985) Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. *Am J Hum Genet* 37:268-285
- Sing CF, Haviland MB, Reilly SL (1996) Genetic architecture of common multifactorial diseases. In: Chadwick DJ, Cardew G (eds) *Variation in the human genome (Ciba Found Symp 197)*. John Wiley & Sons, Chichester, England, pp 211-232
- Sokal RR, Rohlf FJ (1995) *Biometry: the principles and practice of statistical research in biological research*, 3rd ed. WH Freeman and Co, New York
- Solé R, Goodwin B (2000) *Signs of life: how complexity pervades biology*. Basic Books, New York
- Stengård J, Kardia S, Salomaa V, Vartiainen E, Ehnholm C, Puska P, Clark A, Weiss K, Nickerson D, Sing CF (2000) Knowledge about complete DNA sequence variation in the ApoE gene improves the prediction of the variation in measures of lipid metabolism. Paper presented at Satellite Symposium of the 12th International Symposium on Genetics and Atherosclerosis, Århus, Denmark, June 22-24
- Stengård JH, Kardia SL, Tervahauta M, Ehnholm C, Nissinen A, Sing CF (1999) Utility of the predictors of coronary heart disease mortality in a longitudinal study of elderly Finnish men aged 65 to 84 years is dependent on context defined by Apo E genotype and area of residence. *Clin Genet* 56:367-377
- Stengård JH, Salomaa V, Rasi V, Vahtera E, Ehnholm C, Krusius T, Perola M, Vartiainen E (2001) Utility of the Arg/Gln polymorphism of the factor VII (FVII) gene, serum lipid levels and body mass index in the prediction of the FVII:C and FVII:Ag in North Karelia: a cross-sectional and prospective study. *Blood Coagul Fibrinolysis* 12:445-452
- Stengård JH, Weiss KM, Sing CF (1998) An ecological study of association between coronary heart disease mortality rates in men and the relative frequencies of common allelic variations in the gene coding for apolipoprotein E. *Hum Genet* 103:234-241
- Stengård JH, Zerba KE, Pekkanen J, Ehnholm C, Nissinen A, Sing CF (1995) Apolipoprotein E polymorphism predicts death from coronary heart disease in a longitudinal study of elderly Finnish men. *Circulation* 91:265-269
- Templeton AR (2000) Epistasis and complex traits. In: Wolf JB, Brodie ED 3rd, Wade MJ (eds) *Epistasis and the evolutionary process*. Oxford University Press, New York, pp 41-57
- Turner ST, Weidman WH, Michels VV, Reed TJ, Ormson CL, Fuller T, Sing CF (1989) Distribution of sodium-lithium countertransport and blood pressure in Caucasians five to eighty-nine years age. *Hypertension* 13:378-391
- Utermann G, Hees M, Steinmetz A (1977) Polymorphism of apolipoprotein E and occurrence of dysbetalipoproteinemia in man. *Nature* 269:604-607
- Vartiainen E, Puska P, Pekkanen J, Tuomilehto J, Jousilahti P (1994) Changes in risk factors explain changes in mortality from ischemic heart disease in Finland. *Br Med J* 309:23-27
- Wald NJ, Hackshaw AK, Frost CD (1999) When can a risk

- factor be used as a worthwhile screening test? *Br Med J* 319: 1562–1565
- Wald NJ, Law M, Watt HC, Wu T, Bailey A, Johnson AM, Craig WY, Ledue TB, Haddow JE (1994) Apolipoproteins and ischemic heart disease: implications for screening. *Lancet* 343:75–79
- Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 18:19–24
- Xhignesse M, Lussier-Cacan S, Sing CF, Kessling AM, Davignon J (1991) Influences of common variants of apolipoprotein E on measures of lipid metabolism in a sample selected for health. *Arterioscler Thromb* 11:1100–1110