Contribution of regulatory and structural variations in *APOE* to predicting dyslipidemia

Jari H. Stengård,[1,*,†] Sharon L. R. Kardia,[§] Sara C. Hamon,[†] Ruth Frikke-Schmidt,[**] Anne Tybjærg-Hansen,[**] Veikko Salomaa,[*] Eric Boerwinkle,[††] and Charles F. Sing[1,†]

National Public Health Institute,* Helsinki, Finland, Department of Human Genetics† and Department of Epidemiology,§ University of Michigan, Ann Arbor, MI; Department of Clinical Biochemistry,** Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark; and Human Genetics Center and Institute of Molecular Medicine,†† University of Texas Health Science Center, Houston, TX

Abstract The objective of this study was to evaluate *1*) whether non single nucleotide polymorphisms-coding (non-cSNP) in the apolipoprotein E gene (*APOE*) identified by resequencing studies contribute to statistically explaining dyslipidemia if variations in the two cSNPs in exon 4 that define the ε*2*, ε*3*, and ε*4* alleles are ignored, and *2*) whether the contribution of these additional SNPs persists when variations in the cSNPs are considered. We used an ecological, multiple-population, data-mining strategy to identify single-SNP and two-SNP genotypes that distinguish between high and low levels of plasma lipids in three training samples, European-Americans from Rochester, MN, African-Americans from Jackson, MS, and Europeans from North Karelia, Finland. We found that a pair of SNPs located in the 5′ region define genotypes $A_{560}T_{832}/A_{560}T_{832}$, $A_{560}T_{832}/A_{560}G_{832}$, and $A_{560}T_{832}/T_{560}T_{832}$, which distinguish between high and low levels of HDL-cholesterol (HDL-C), triglycerides (TG), and/or total cholesterol (T-C). The $A_{560}T_{832}$/-genotypes predicted high TG and high T-C in both genders in a large independent test sample from Copenhagen, Denmark. Prediction of high T-C in the Danish females was dependent on genotypes defined by the cSNPs. Our study suggests that both regulatory and structural variations should be considered when evaluating the utility of *APOE* for predicting dyslipidemia in the population at large.—Stengård, J. H., S. L. R. Kardia, S. C. Hamon, R. Frikke-Schmidt, A. Tybjærg-Hansen, V. Salomaa, E. Boerwinkle, and C. F. Sing. **Contribution of regulatory and structural variations in *APOE* to predicting dyslipidemia. *J. Lipid Res.* 2006.** 47: **318–328.**

Supplementary key words apolipoprotein E gene • pleiotropy • data mining • regulation • lipids

Cholesterol accumulation in arterial walls is an important contributing factor in the development of atherosclerotic cardiovascular disease (CVD) (1). Information about the genetic basis of interindividual differences in lipid metabolism is thus expected to be useful in risk assessment, providing clues for the development of nonpharmacological and pharmacological interventions and suggesting population-based disease prevention strategies for CVD (2–5). A plethora of variations in genes involved in lipid metabolism have been characterized (6–10). The statistical evaluation of the contributions of these genomic variations to variation in measures of lipid metabolism and risk of CVD presents one of the most difficult challenges facing CVD research. The biological realities that interactions between gene variations and environmental variations are the primary causes of interindividual differences in lipid metabolism and risk of CVD, and that these interactions are dynamic over the lifetime of the individual, serve as major obstacles for the study of phenotype-genotype relationships (11).

Studies of the influence of the variation in the gene coding for apolipoprotein E (*APOE*) on quantitative blood measures of lipid metabolism have demonstrated both context-dependent genotype effects (12–14) and phenotype-*APOE* genotype relationships that are less sensitive to contexts indexed by time and space (12, 14, 15). In the study reported here, we use data collected from three populations that are ethnically and geographically distinct to identify phenotype-*APOE* genotype relationships that are less sensitive to the influence of genetic and environmental contexts indexed by gender, ethnicity, and geographic location. The utility of the identified phenotype-genotype models is then tested in a sample from a large independent study of a fourth population. We chose this strategy to identify phenotype-*APOE* genotype relationships that are expected to have the greatest utility in predicting dyslipidemia in the broadest range of contexts.

Apolipoprotein E (apoE) is a structural constituent of many atherogenic lipoprotein particles, such as triglyceride (TG)-rich chylomicrons and HDLs, and is involved in their transport from one tissue or cell type to another (16–18). It has three common isoforms, E2, E3, and E4 (19), which are encoded by three alleles, ε*2*, ε*3*, and ε*4*,

Manuscript received 17 May 2005 and in revised form 8 November 2005.

Published, JLR Papers in Press, November 29, 2005.
DOI 10.1194/jlr.M500491-JLR200

[1] To whom correspondence should be addressed.
e-mail: jari.stengard@ktl.fi (J.H.S.); csing@umich.edu (C.F.S.)

defined by two variable sites in exon 4 of *APOE*. Variation in the blood concentration of total cholesterol (T-C) is commonly associated with this structural variation in apoE (16, 20). In a study that resequenced 5.5 kb of *APOE*, including related 5′ and 3′ flanking regions, we identified 10 public biallelic single-nucleotide polymorphisms (SNPs) that segregate in multiple populations (8, 9). These SNPs included the two variations in the fourth exon (denoted cSNPs, at positions 3937 and 4075) that encode the differences between the E2, E3, and E4 isoforms. In this study, we ask two questions. *1*) Do the eight additional non-cSNP variations identified by resequencing studies contribute to statistically explaining differences between individuals with high and low levels of HDL-C, TG, or T-C if variations in the cSNPs at positions 3937 and 4075 are ignored? *2*) Do such contributions persist when variations in the two cSNPs are considered? We chose dichotomous lipid phenotypes as our end points because they are widely used in clinical risk assessment and in public health programs to reduce the health burden of CVD.

We addressed the first of these two questions using an ecological (21), multiple-population, data-mining strategy to identify SNPs, or pairs of SNPs, of *APOE* that define genotypes that statistically distinguish between high and low levels of HDL-C, TG, and/or T-C subgroups in a sample of European-Americans from Rochester, MN. Because heterogeneity in the phenotype-genotype relationship across different populations is an important concern to those seeking context-independent predictors of the risk of disease (22, 23), we then selected only those SNPs, or pairs of SNPs, that define genotypes that distinguish between high and low concentrations of at least two of the three measures of lipid metabolism in both genders in at least one of the two other independent samples collected, in Jackson, MS, and North Karelia, Finland. Our specific questions here are as follows. *1*) How many SNPs, and pairs of SNPs, satisfy the proposed selection criteria? *2*) What are the locations of the selected SNPs? *3*) What are the relative frequencies of the single-SNP alleles and the two-SNP haplotypes defined by selected pairs of SNPs? *4*) What are the high-risk and/or low-risk genotypes defined by the selected single SNPs and pairs of SNPs? We then asked whether *1*) the hypothesized high-risk genotypes predict low HDL-C, high TG, and/or high T-C and *2*) the variations in the 3937 and 4075 cSNP positions are related to the observed discriminative abilities of the proposed phenotype-genotype models using a large population-based sample of Europeans from Copenhagen, Denmark.

METHODS

We used the National Cholesterol Education Program Expert Panel's recommendations for defining dyslipidemic subgroups (24). Dyslipidemia was diagnosed when an individual's blood T-C concentration was >200 mg/dl, TG was >150 mg/dl, or HDL-C was <40 mg/dl.

Our research strategy involved three steps: *1*) SNP selection using three independent samples; *2*) selection of phenotype-genotype models using the information obtained in the SNP selection procedure with these samples; and *3*) evaluation of the utility of the selected models in a fourth independent test sample. In the first SNP selection step, we first used a sample from Rochester to identify single SNPs and pairs of SNPs that defined genotypes that significantly distinguished between high-risk and low-risk subgroups for at least two measures of lipid metabolism in both females and males. The Rochester sample included 854 unrelated individuals (456 females and 398 males) recruited by the Rochester Family Heart Study (25, 26). The participants in the Rochester sample were requested to fast for 12 h before examination.

For the subset of single SNPs and pairs of SNPs that significantly discriminated between high and low concentrations of two or more traits in both genders in the Rochester sample, we next considered the replication of the selected SNP effects in the Jackson and North Karelia samples as a second criterion for SNP selection. The Jackson sample included 702 unrelated African-American individuals (483 females and 219 males) who were part of the ongoing Genetic Epidemiology of Atherosclerosis study (27). The North Karelia sample included 337 unrelated individuals (188 females and 149 males) who were ascertained by an ongoing prospective study, the population-based FINRISK study (28, 29). Each participant in the North Karelia sample was measured for three lipid phenotypes twice, once at the baseline survey in 1992 and then in 1995 in connection with a 3 year follow-up examination (28, 30). To minimize the misclassification of dyslipidemia, we considered only those individuals from North Karelia who had high or low HDL-C, high or low TG, and/or high or low T-C at both the baseline and follow-up surveys. The subset of SNPs, considered singly or in pairs, whose ability to distinguish between high and low concentrations of multiple measures of lipid metabolism replicated in males and females in at least one of these two additional samples was then taken as the final set of selected SNPs. The participants in the Jackson study were requested to fast for 12 h, and the participants of the North Karelia study were requested to fast for 4 h, before examination.

In the second step, we identified the haplotypes and genotypes defined by the selected SNPs that were responsible for the observed statistically significant phenotype-genotype associations observed in the first step and whose effects were replicated in both genders in at least one of the two other independent samples collected in Jackson and North Karelia. In the third and final step, we tested the utility of phenotype-genotype models established in step 2 for predicting low HDL-C, high TG, or high T-C in large population-based samples of females and males collected in Copenhagen. The Danish sample included 9,011 unrelated, native-born, non-Hispanic European individuals (4,947 females and 4,064 males) ascertained without regard to health status in connection with the third examination of the Copenhagen City Heart Study (31, 32). The participants in the Danish study were not requested to fast before examination. All participants in the Rochester, Jackson, and North Karelia samples gave informed consent, and the Copenhagen City Heart Study was approved by the Danish Ethics Committee for Copenhagen and Frederiksberg (No. 100.2039/91).

Blood HDL-C, TG, and T-C concentrations for the Rochester and Jackson samples were measured at the Mayo Clinic (Rochester, MN) using published methods (20, 33, 34). The Finnish and Danish samples were measured by standard enzymatic assays (Boehringer Mannheim GmbH Diagnostics, Mannheim, Germany) at the Department of Biochemistry, National Public Health Institute, in Helsinki (28, 35) and at the Department of Clinical Biochemistry, Rigshospitalet, Copenhagen University Hospital (32), respectively. The methods used to genotype

the *APOE* SNPs have been described by Nickerson et al. (9) for the Rochester, Jackson, and North Karelia samples and by Frikke-Schmidt et al. (32) for the Danish sample. The relative frequencies of two-site haplotypes for each population were estimated using an E-M algorithm (36).

In the first SNP selection step, we used the combinatorial partitioning method (CPM) (37) as a data-mining tool to evaluate the ability of genetic variations defined by one- and two-SNP genotypes to distinguish between high and low concentrations of HDL-C, TG, and T-C in the female and male Rochester samples. This method was developed to identify partitions of genotypes that statistically explain interindividual variation in quantitative trait levels. We modified the CPM for this study to identify partitions of single- and two-SNP genotypes that statistically distinguish dichotomized trait levels. In this modified strategy, we first estimated the prevalence of the trait of interest (e.g., low blood HDL-C concentration) for each genotype in the set of genotypes defined by a particular SNP or pair of SNPs. The genotypes were then ranked according to their prevalence estimates. The ranked genotypes were partitioned into groups, and the prevalence was reestimated for each partition. The utility of each set of partitions for distinguishing between high and low trait levels was evaluated using the contingency Chi-square statistic. For each SNP and each pair of SNPs, this strategy selects the set of partitions that maximized similarities of the prevalences associated with genotypes within partitions and minimized similarities of the prevalences assigned to different partitions of genotypes.

At present, there is no formal, widely accepted, statistical strategy for distinguishing statistically significant results from a single study that are a consequence of "true" biological effects from those that are type I errors (11). Hence, we used an ad hoc strategy to minimize the possibility that the significant result of a particular CPM analysis is a type I statistical error by selecting only those SNPs, or pairs of SNPs, that define genotypes that distinguish between high and low blood concentrations of at least two measures of lipid metabolism in both females and males, first in the Rochester sample, and subsequently in both female and male samples from Jackson or North Karelia or from both samples.

We next used a second data-mining strategy to identify the single-SNP and/or two-SNP genotype(s) that are most likely responsible for the statistically significant phenotype-genotype associations in the Rochester, Jackson, and North Karelia samples. This involved identifying those genotypes that have a higher prevalence of the trait of interest (e.g., low HDL-C) than the overall prevalence in the gender/population sample being considered. Again, we selected only those genotypes whose higher ranking was consistent across at least five of the six gender/population samples.

Finally, the utility of the phenotype-genotype models obtained in the two data-mining steps for predicting dyslipidemia was evaluated in the Danish sample using conventional logistic regression analysis (38). Unless noted otherwise, we considered a nominal $\alpha = 0.05$ level of probability to be a statistically significant estimate of the relative odds of dyslipidemia.

RESULTS

Description of the Rochester sample

Gender-specific means and variances of age, basic anthropometric characteristics, and the three blood measures of lipid metabolism, HDL-C, TG, and T-C, are given in **Table 1**. The average age of the female and male samples was similar (48 years), but the variability in age was significantly greater in females. On average, females were significantly leaner, and they were less frequently dyslipidemic (20, 19, and 39% for low HDL-C, high TG, and high T-C, respectively) than males (56, 35, and 47%, respectively). The estimates of interindividual variance of body mass index were significantly greater in females than in males.

TABLE 1. Description of female and male samples collected in Rochester

Anthropometric Characteristics	Females (n = 456)	Males (n = 398)
Age (years)		
Mean	48.46	48.14
Variance	93.09	74.84[a]
Weight (kg)		
Mean	69.2	86.84[b]
Variance	189.12	188.65
Height (cm)		
Mean	163.78	177.44[b]
Variance	33.02	38.33
Body mass index (kg/m^2)		
Mean	25.83	27.59[b]
Variance	26.6	17.84[b]
Plasma HDL-C (mg/dl)		
Mean	51.74	39.64[b]
Variance	192.7	105.45[b]
Percentage of low HDL-C	20.18	55.53[b]
Plasma TG (mg/dl)		
Mean	107.86	142.02[b]
Variance	3,358.39	7,491.25[b]
Percentage of high TG	18.86	35.43[b]
Plasma T-C (mg/dl)		
Mean	192.9	199.46[c]
Variance	1,412.19	1,217.03
Percentage of high T-C	38.82	46.48[a]

HDL-C, HDL-cholesterol; T-C, total cholesterol; TG, triglyceride.
[a] Significant difference between males and females at $\alpha \leq 0.05$.
[b] Significant difference between males and females at $\alpha \leq 0.001$.
[c] Significant difference between males and females at $\alpha \leq 0.01$.

Utility of single-SNP genotype variations for distinguishing between high and low HDL-C, TG, and/or T-C in the Rochester sample

The tests of associations between lipid traits and single-SNP genotype variations are summarized in the diagonal cells of **Fig. 1**, separately for females and males. Only 1 of the 10 SNPs (5361) defined a single-SNP genotypic variation that distinguished between high and low concentrations of more than one blood measure of lipid metabolism in either gender.

Utility of two-SNP genotype variations for distinguishing between high and low HDL-C, TG, and T-C in the Rochester samples

The tests of associations between lipid traits and two-SNP genotype variations are summarized in the off-diagonal cells of Fig. 1, separately for females and males. Twelve pairs of SNPs in females (26%; denoted by red, blue, green, or purple in Fig. 1) and 23 pairs in males (51%; also denoted by red, blue, green, or purple in Fig. 1) defined two-SNP genotype variations that distinguished between high and low concentrations of more than one lipid

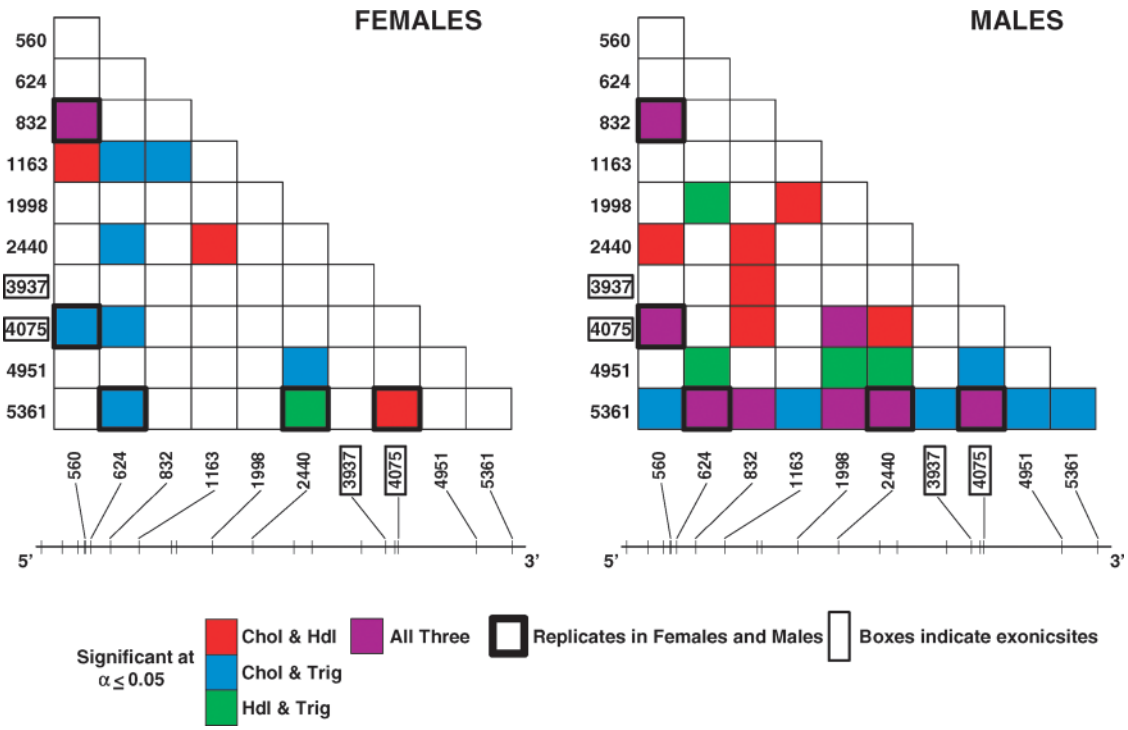

Fig. 1. Utilities of one- and two-SNP genotype variations for distinguishing between high and low HDL-cholesterol [HDL-C (Hdl)], triglyceride [TG (Trig)], and/or total cholesterol [T-C Chol)] in the Rochester sample. The tests of associations between lipid traits and single-SNP genotype variations are summarized in the diagonal cells, and the tests of associations between lipid traits and two-SNP genotype variations are summarized in the off-diagonal cells. Different colors indicate the number of high and low concentrations of lipid traits that can be distinguished by a particular genotype.

trait. Of these pairs, only five (560-832, 560-4075, 624-5361, 2440-5361, and 4075-5361; denoted by black boxes in Fig. 1) distinguished between high and low trait concentrations in both genders. For each of these pairs, we next considered the replication of the phenotype-genotype association in the Jackson and North Karelia samples as a second criterion for SNP selection.

Utility of selected two-SNP genotype variations for distinguishing between high and low HDL-C, TG, and T-C in the Jackson and North Karelia samples

The tests of associations between the lipid traits and two-SNP genotype variations defined by each of the five selected two-SNP combinations in the Jackson and North Karelia samples are summarized in **Table 2**. Four of the

TABLE 2. Probabilities for the two-SNP combinations that discriminate between high and low blood concentrations for more than one of the three measures of lipid metabolism in both females and males in the Rochester sample

	Hypothesis about Discriminative Pairs			Site selection					
	Rochester			Jackson			North Karelia		
SNPs	HDL-C	TG	T-C	HDL-C	TG	T-C	HDL-C	TG	T-C
560/832									
Females	<0.01*,†	0.02*,†	0.01*,†	0.12	0.03*,†	<0.01*,†	<0.01*	0.05*	0.06
Males	0.01*,†	0.05*,†	0.03*,†	0.25	0.04*,†	0.01*,†	0.06	0.24	0.11
560/4075									
Females	0.09	0.04*,†	<0.01*,†	0.03*	0.02*	0.02*,†	0.11	0.25	0.25
Males	0.03*	0.05*,†	<0.01*,†	0.29	0.22	<0.01*,†	0.11	0.28	0.44
624/5361									
Females	0.22	0.02*,†	0.03*,†	0.13	0.51	0.83	0.49	0.23	0.17
Males	0.03*	0.02*,†	0.02*,†	0.05*	0.29	0.20	0.22	0.05*	0.40
2440/5361									
Females	<0.01*,†	0.03*,†	0.55	<0.01*,†	0.25	0.17	0.02*	0.41	0.03*
Males	0.05*,†	0.02*,†	0.02*	0.05*,†	0.11	0.04*	0.06	0.25	0.22
4075/5361									
Females	<0.01*,†	0.22	<0.01*,†	0.09	0.05*	0.12	0.49	0.35	0.05*
Males	0.05*,†	0.01*	<0.01*,†	0.05*	0.13	<0.01*	0.18	0.38	0.36

Probabilities that are considered statistically significant are denoted by asterisks. Daggers indicate consistent test results in both genders within the particular population sampled. The SNPs are labeled according to the nomenclature of Fullerton et al. (8) and Nickerson et al. (9).

five pairs of SNPs (560-832, 560-4075, 2440-5361, and 4075-5361) define two-SNP genotype variations that significantly ($P < 0.05$) distinguished between high and low concentrations of multiple measures of lipid metabolism in one or more of the four additional gender- and population-specific samples (denoted by asterisks in Table 2). Only one pair (560-832) satisfied the criterion that the discrimination between high and low subgroups is statistically significant in both females and males for two or more traits. This result suggests further investigation of the 560-832 pair to identify the haplotypes and genotypes responsible for the statistically significant phenotype-genotype associations.

Relative frequencies of the two-SNP haplotypes defined by variations in the non-cSNPs at positions 560 and 832

Adenine (A_{560}) and guanine (G_{832}) are the most common nucleic acids at the 560 and 832 sites, respectively, in all three samples (**Table 3**). Estimates of the relative frequencies of these alleles, however, were heterogeneous among the three populations. The relative frequency of the A_{560} allele was ~20% lower, and that of the G_{832} allele 150% higher, in the Jackson sample than in the Rochester and North Karelia samples. The A_{560} and G_{832} alleles define the most common two-SNP haplotype in all three populations. The A_{560} allele together with thymine at the 832 position (T_{832}) define the second most common haplotype in the Rochester and North Karelia samples, whereas in the Jackson sample this haplotype was the least common. The T_{560} and G_{832} alleles define the second most common two-site haplotype in the Jackson sample.

Identification of the most informative two-SNP genotypes defined by SNPs at positions 560 and 832

Prevalence estimates of low HDL-C, high TG, and high T-C in each of the six gender/population samples are denoted by red lines in **Fig. 2A, B, C**, respectively. These estimates ranged between 555:1,000 and 29:1,000 for low HDL-C (Fig. 2A), between 434:1,000 and 189:1,000 for high TG (Fig. 2B), and between 881:1,000 and 388:1,000 for high T-C (Fig. 2C). The test of heterogeneity of the prevalences among the six gender/population samples was statistically significant at $P < 0.001$ for each of the three lipid traits.

Prevalences of low HDL-C, high TG, and high T-C for each of the observed two-SNP genotypes defined by the 560-832 pair of SNPs are given in Fig. 2A, B, C, respectively, separately for each of the six gender/population samples. Prevalences of high and low lipid concentrations in subsamples of carriers of the $T_{560}T_{832}$ and $T_{560}G_{832}$ haplotypes tended to deviate more from the prevalences of the respective gender/population samples than did prevalences in subsamples of individuals who were either homozygous or heterozygous for the two common haplotypes $A_{560}T_{832}$ and $A_{560}G_{832}$. Rankings of genotype-specific prevalences vary from one lipid trait to another within a particular gender/population sample, as well as from one gender/population sample to another for a particular lipid trait. There are exceptions, however. The prevalence of low HDL-C in the subsample of $A_{560}T_{832}/A_{560}T_{832}$ homozygous individuals was higher than the sample prevalence in five of the six gender/population samples. Furthermore, the prevalence of high T-C in this subsample of homozygotes was higher than the sample prevalence in all six gender/population samples. Using a Sign's test (39), the probability of observing the observed ranking of the $A_{560}T_{832}/A_{560}T_{832}$ genotype with respect to the prevalence in each of the gender/population samples, assuming that there is no association between this genotype and prevalence, is 0.109 for five of six rankings and 0.035 for six of six rankings. The prevalences of high TG in the subsample of $A_{560}T_{832}/A_{560}G_{832}$ and $A_{560}T_{832}/T_{560}T_{832}$ heterozygous individuals was lower than the sample prevalence in five of six gender/population samples, whereas the prevalence of high T-C in subsamples of $A_{560}T_{832}/A_{560}G_{832}$ heterozygous individuals was higher than the sample prevalence in five of the six gender/population samples.

In summary, we conclude from the analyses of the Rochester, Jackson, and North Karelia samples that the $A_{560}T_{832}$ haplotype-containing genotypes are the most informative predictors of dyslipidemia. Individuals who are homozygous for the $A_{560}T_{832}$ haplotype have an increased risk of low HDL-C that is consistent among samples that differ in gender, ethnicity, and geographic location. A subsample of $A_{560}T_{832}/A_{560}T_{832}$ homozygous and $A_{560}T_{832}/A_{560}G_{832}$ and $A_{560}T_{832}/T_{560}T_{832}$ heterozygous individuals (denoted as $A_{560}T_{832}/$-) have a decreased risk of high TG but an increased risk of high T-C. We next tested the utility of these recessive and dominant genetic models in distinguishing between low HDL-C and high TG and T-C, respectively, using data from large population-based samples of females and males collected in Copenhagen.

TABLE 3. Relative frequencies of the 560 and 832 alleles and haplotypes in the three samples

Allele/Haplotype	Rochester	Jackson	North Karelia
Single-SNP allele			
A_{560}	0.83	0.69	0.89
T_{560}	0.17	0.31	0.11
G_{832}	0.52	0.76	0.54
T_{832}	0.48	0.24	0.46
Two-SNP haplotype			
$A_{560}G_{832}$	0.45	0.60	0.47
$A_{560}T_{832}$	0.38	0.09	0.42
$T_{560}T_{832}$	0.10	0.15	0.04
$T_{560}G_{832}$	0.07	0.16	0.07

The SNP alleles are labeled according to the nomenclature of Fullerton et al. (8) and Nickerson et al. (9).

A test of the utility of the selected two-SNP genotypes in predicting dyslipidemia in large population-based samples of females and males from Copenhagen

The relative odds of low HDL-C and high TG and T-C in those individuals with the hypothesized phenotype-genotype models are given in **Table 4**, separately for females and males from Copenhagen. The association of low HDL-C with the $A_{560}T_{832}/A_{560}T_{832}$ genotype observed in the Rochester, Jackson, and North Karelia samples tended to survive further testing in the Danish samples. The estimated odds of low HDL-C were higher in $A_{560}T_{832}/A_{560}T_{832}$ homozygous females and males than in carriers

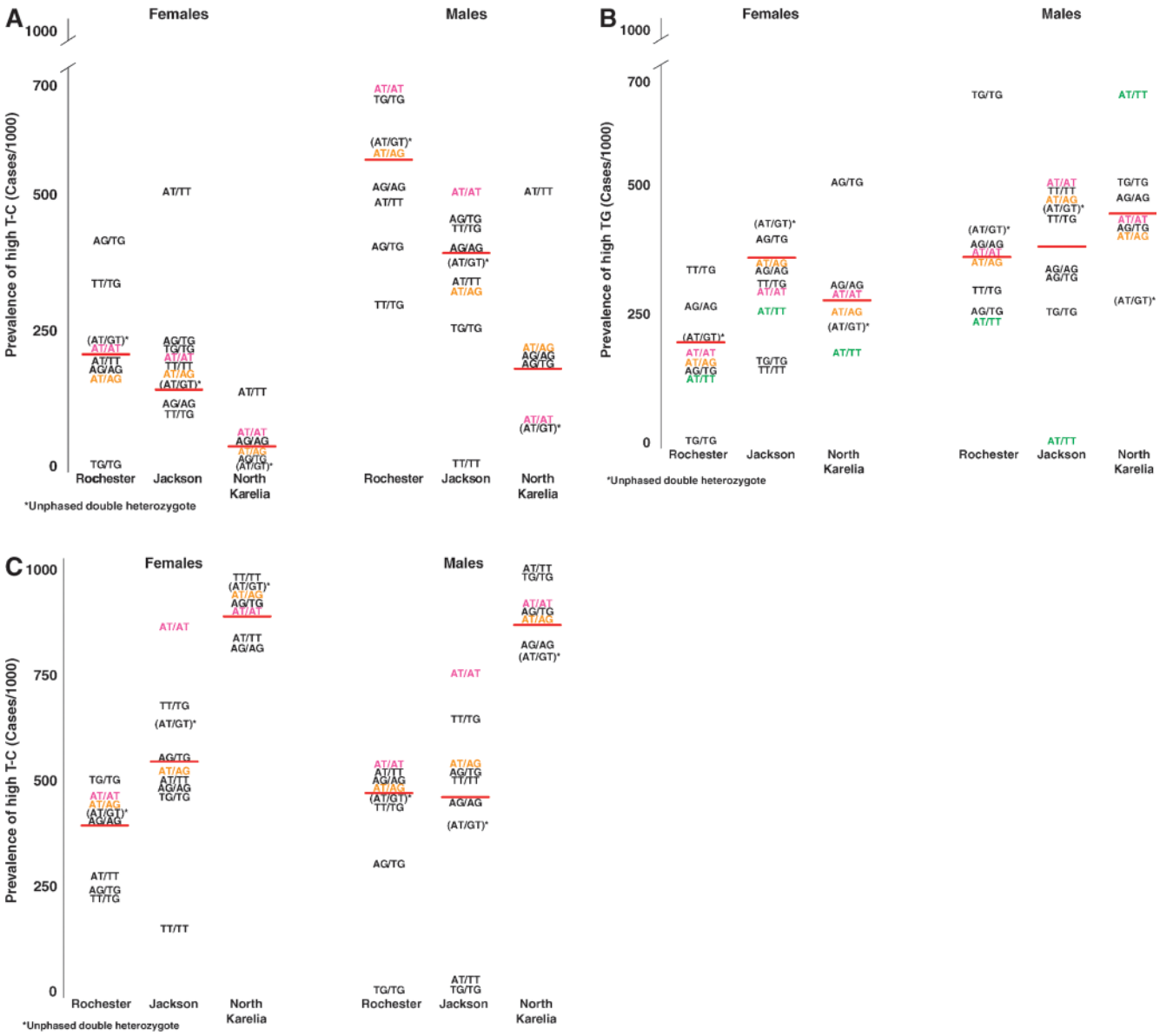

Fig. 2. A: Overall and genotype-specific prevalences of low HDL-C level, separately for each of the six gender/population strata. The prevalence for the AT/AT genotype is higher than the overall prevalence in five of six gender/population samples ($P = 0.109$). The genotypes are labeled according to the nucleic acids determined by the alleles of the 560 and 832 SNPs. B: Overall and genotype-specific prevalences of high TG level, separately for each of the six gender/population strata. The prevalence for both the AT/AG and AT/TT genotypes was lower than the overall prevalence in five of six gender/population samples ($P = 0.109$). The genotypes are labeled according to the nucleic acids determined by the alleles of the 560 and 832 SNPs. C: Overall and genotype-specific prevalences of high T-C level, separately for each of the six gender/population strata. The prevalence for the AT/AT genotype was higher than the overall prevalence in all six gender/population samples ($P = 0.035$). The prevalence for the AT/AG genotype was higher than the overall prevalence in five of the six gender/population samples ($P = 0.109$). The genotypes are labeled according to the nucleic acids determined by the alleles of the 560 and 832 SNPs.

of other genotypes [odds ratios (ORs) = 1.46 for females and 1.10 for males]. This observed increase was statistically significant in females but not in males. Similarly, associations between the two other measures of dyslipidemia, high TG and high T-C, and the $A_{560}T_{832}$/- genotypes tended to replicate in the Danish sample. The estimated odds of high TG were reduced, and odds of high T-C were increased, in individuals in the group with the $A_{560}T_{832}$/- genotypes compared with odds for the group with the other genotypes (ORs = 0.92 and 1.21 for high TG and high T-C, respectively, in females and 0.85 and 1.18 for high TG and high T-C, respectively, in males). All but one of these four ORs was statistically significant. The lower prevalence of high TG for the group with the $A_{560}T_{832}$/- genotypes was not significant in females.

Genotypes defined by the two cSNPs were statistically significant predictors of high T-C and high TG in both genders and of low HDL-C in females only. There was no evidence of a statistically significant interaction between the effects of the group of $A_{560}T_{832}$/- genotypes and the

TABLE 4. Relative odds (95% confidence interval) of low HDL-C, high TG, and high T-C in the Copenhagen sample for high-risk genotypes selected by analyses of the Rochester, North Karelia, and Jackson samples

Gender	Low HDL-C[a]	High TG[b]	High T-C[b]
Females			
Unadjusted	1.46 (1.08–1.98)	0.92 (0.82–1.03)	1.21 (1.05–1.39)
Adjusted[c]	1.41 (1.03–1.94)	0.86 (0.75–0.97)	—[d]
Males			
Unadjusted	1.10 (0.90–1.35)	0.85 (0.75–0.96)	1.18 (1.03–1.36)
Adjusted[c]	1.00 (0.81–1.24)	0.83 (0.72–0.95)	0.96 (0.82–1.12)

[a] $A_{560}T_{832}/A_{560}T_{832}$ genotype contrasted with other genotypes.
[b] $A_{560}T_{832}/A_{560}T_{832}$ and $A_{560}T_{832}/-$ combined and contrasted with other genotypes.
[c] Adjusted for variation in the two cSNPs 3937 and 4075, grouped as follows: (*ε2/2*, *ε3/2*), (*ε4/2*, *ε4/3*, *ε4/4*), and *ε3/3*.
[d] Relative odds dependent upon variation in the two cSNPs 3937 and 4075, grouped as follows: (*ε2/2*, *ε3/2*), (*ε4/2*, *ε4/3*, *ε4/4*), and *ε3/3*, as presented in Table 5.

TABLE 5. Relative odds of high T-C in the Copenhagen female sample, separately for the three genotype groups defined by variation in the two cSNPs 3937 and 4075

Genotype Groups	*ε2/2*, *ε3/2*	*ε4/2*, *ε4/3*, *ε4/4*	*ε3/3*
$A_{560}T_{832}/-$	0.38 (0.26–0.56)	1.24 (1.00–1.53)	0.78 (0.64–0.94)
Others	0.38 (0.31–0.48)	0.97 (0.69–1.38)	1

The 5′ genotypes defined by the 560 and 832 SNPs are labeled according to the nucleic acids determined by the SNP alleles.

effects of genotypes defined by variations in the two cSNPs 3937 and 4075 in predicting low HDL-C and high TG (Table 4). The ORs for low HDL-C and high TG when the two cSNPs are ignored were in the same range as the adjusted ORs estimated when the two cSNPs are included in the prediction model. There was a statistically significant interaction between the effect of the group of $A_{560}T_{832}/-$ genotypes and the genotypes defined by variations in the two cSNPs in the prediction of high T-C in females (**Tables 4, 5**). The estimated OR for high T-C is significantly higher (1.24; 95% confidence interval = 1.00–1.53) for the *ε4* allele-carrying females in the $A_{560}T_{832}/-$ genotypes group and significantly lower (0.78; 95% confidence interval = 0.64–0.94) for the *ε3/3* group of females compared with females with the *ε3/3* genotype who did not have the $A_{560}T_{832}/-$ genotypes. The group with the $A_{560}T_{832}/-$ genotypes was not identified as a statistically significant predictor of high T-C in males when variations in the two exon 4 cSNPS were included in the prediction model.

DISCUSSION

An alternative research strategy

A commonly used strategy for identifying genetic variations that are predictors of phenotypic variation is to collect a large representative sample from a particular population, use statistical summaries to test phenotype-genotype hypotheses, and turn to Baconian induction to infer the generality of genetic effects (40–42). An integral part of such a strategy is that a statistically significant, empirically derived hypothesis must survive further testing in other studies of other samples to become a universal "truth" (42). The expectation is that the surviving hypothesis can then be used to predict future events in any population (43). Genetic analyses of phenotypes that have a complex multifactorial etiology, such as dyslipidemia, challenge this induction/deduction paradigm because it ignores the possibility that the hypothesis generated is dependent on the context of the population studied. The predictions of the proposed hypothesis simply may not survive further testing because of the heterogeneity of the phenotype-genotype relationship among populations or the lack of statistical power associated with small samples. As likely is the possibility that it may not survive further testing in any population because a hypothesis derived from the study of only one population may be a type I error (11). We suggest here an alternative strategy to this induction/deduction paradigm that reduces the possibility that the initial hypothesis is a type I error by applying an ecological data-mining strategy to samples collected from multiple populations to generate a hypothesis that is expected to be less sensitive to context. This multiple-population data-mining strategy sorts out those hypothesized phenotype-genotype relationships that are less likely to be type I errors and more likely to be of utility in unstudied populations that differ for genetic and environmental contexts indexed by gender, ethnicity, and geographic locations. Although this strategy increases the likelihood that a particular genetic variation may have utility in predicting phenotypic variation in an unstudied population, we emphasize that the predictive utility realized in independent samples of Danish females and males must be reevaluated anew in subsequent populations of interest because of the anticipated role of context dependence in the etiology of measures of lipid metabolism. We discuss below *1*) the limitations of this research strategy for modeling the genetic architecture of measures of lipid metabolism; *2*) the relationship between phenotypic variation in lipid traits and variation in *APOE* identified by this strategy; *3*) how the proposed phenotype-genotype model reflects current knowledge about the biology of *APOE* and lipid metabolism at the cellular level; and *4*) how this phenotype-genotype model can be used in medical practice and/or public health programs in a particular population of interest.

Limitations of the research strategy for characterizing the genetic architecture of lipid and lipoprotein traits

The genetic architecture of a complex trait is a function of the relative frequencies of genotypes (genetic structure) and phenotype-genotype relationships. The SNPs, genotypes, and phenotype-genotype models identified by an ecological data-mining strategy may not be the best choices for predicting variation in lipid traits in any one of the three populations considered, or in any other particular population, because this approach considers only the public genetic variations that are shared by all pop-

ulations. The number of private, population-specific SNPs in *APOE* varies: there are three in Rochester, four in North Karelia, and six in Jackson (8, 9). Furthermore, there is significant heterogeneity in the relative frequencies of the alleles at the 10 public SNPs among these three populations (14). A research strategy that ignores the microdifferentiation of the relative allele frequencies that results in heterogeneity of the genetic structure among human populations may underestimate the contribution of variation in a candidate gene to variation in intermediate biochemical and physiological traits and, ultimately, to the risk assessment of correlated disease end points in any particular population. Hence, using an ecological datamining strategy, one forgoes the search for the best genetic predictors in any particular population to identify a model that is less likely to be a false-positive (type I) statistical error and is expected to have greater general applicability across the populations of interest.

There are several shortcomings of the ecological datamining strategy for modeling the biological relationships between phenotype and genotype. Statistical models that have general applicability across populations cannot be expected to capture the biological complexities of the connections known to be involved. The role of population-specific gene-gene and gene-environment interactions and population-specific age-dependent exposures to specific environmental agents can only be studied on a population-by-population basis. More importantly, in common with all association studies, most single-gene effects on the phenotype of interest in a particular population cannot be estimated because *1*) they are too small to measure; *2*) they cannot be accurately estimated; *3*) they are confounded with the effects of unmeasured genetic and/or environmental agents (44) and/or even chance (45); *4*) the effects are inseparable from the effects of closely linked gene variations; and *5*) the complexities of the cause-and-effect connections through the intermediate pathways to the phenotype result in no detectable association between phenotype and genotype (11, 46, 47). In addition, genetic influences are distributed throughout multiple intermediate pathways that lead to the dyslipidemia phenotype. A linear statistical model cannot capture the nonlinear processing of genetic effects through the pathways that connect genotype with phenotype. Such impenetrable features introduce uncertainty into the application of any strategy for modeling genetic predictors of phenotypes that have a complex multifactorial etiology.

Lipid phenotype-*APOE* genotype statistical models

Most association studies have focused on the phenotypic effects of variations in the cSNPs located at positions 3937 and 4075 in exon 4 of *APOE* that determine the ε*2*, ε*3*, and ε*4* alleles. Increased T-C has been repeatedly associated with carriers of the ε*4* allele (14, 16, 20). Some population-based studies have also reported that ε*4* carriers are at increased risk for CVD (48). In contrast, only a few studies have reported pleiotropic effects of the six genotypes determined by the cSNPs at positions 3937 and 4075 on HDL-C and TG (13, 32, 37, 49). The percentage of interindividual variation in HDL-C explained by these genotypes in a sample of Danish females is relatively small and depends on age (32). Consistent with this finding, the six *APOE* genotypes defined by the two exon 4 cSNPs were not identified as being associated with higher or lower risk of dyslipidemia in the Rochester sample. In contrast, the three genotypes defined by the SNPs at 5′ positions 560 and 832 ($A_{560}T_{832}$/$A_{560}T_{832}$, $A_{560}T_{832}$/$A_{560}G_{832}$, and $A_{560}T_{832}$/$T_{560}T_{832}$) were repeatedly associated with low HDL-C, high TG, and/or high T-C across contexts defined by gender, ethnicity, and/or area of residence. These statistical findings suggest that the 5′ regulatory region of *APOE* has a larger domain of biological functionality than the nonsynonymous variations in exon 4. We return below to discussing the possible biological basis for this statistical finding.

The statistically significant associations between the three measures of dyslipidemia and the three genotypes in the test samples of Danish females and males are consistent with the hypothesis that variation in the 5′ promoter region of *APOE* has pleiotropic effects on lipid metabolism. This finding is also consistent with an earlier observation reported in young and middle-aged Danish females by Frikke-Schmidt et al. (32) that combining SNP variations in the 5′ promoter region and in the exon 4 structural region doubled the estimated proportions of HDL-C variation that could be statistically explained compared with the proportion explained by the six exon 4 genotypes considered separately.

The biological reality that structural variation in exon 4 of *APOE* has an important role in lipid and lipoprotein metabolism (16–18) raises the question of whether the observed abilities of the 5′ genotypes to distinguish between high and low HDL-C, TG, or T-C are attributable to the effects of variation in the 5′ promoter region or to an association attributable to linkage disequilibrium (LD) with the structural variation in exon 4. Frikke-Schmidt et al. (32) reported statistically significant pair-wise LD between SNPs in the 5′ promoter region and in the exon 4 structural region in the Danish sample. However, the magnitudes of the relevant LD estimates were low. The r^2 measure of LD ranged from 0.023 to 0.079 for the 560-4075 and 832-4075 pairs of SNPs, respectively. It is unlikely that such weak pair-wise LD between these two regions could be responsible for the association of measures of lipid metabolism with particular 5′ genotypes observed here. The statistical independence of the 5′ genetic effects is consistent with our observations that low HDL-C, or high TG, was significantly associated with particular 5′ genotypes in three of the four analyses that included the exon 4 variation. The small role of LD is further supported by a statistically significant interaction between the effects of 5′ genotypic variation and the exon 4 structural variation in predicting high T-C in females.

Biological inferences from phenotype-genotype statistical models

Population genetic analyses (8, 50) and experimental laboratory studies (51–54) provide an evolutionary and

molecular basis for interpreting the statistical association between the three measures of dyslipidemia and the 5′ promoter region. Fullerton et al. (8) and Templeton et al. (50) have demonstrated that multisite *APOE* haplotypes fall into four major lineages, or clades, that correspond to variation in the three structural isoforms defined by variations in the two cSNPs in exon 4. The E2 and E4 isoforms are associated with haplotypes that group into two phylogenetically distinct lineages, whereas the haplotypes that code the E3 isoform fall into two separate phylogenetically distinct clades (8, 50). Variation in blood T-C level is associated with the structural variations in *APOE* that code the three isoforms of the apoE molecule (16, 20). Variation in the 560 and 832 sites divides the four clades into two groups of haplotypes. The first group includes one of the clades that codes the E3 isoform and the clade that codes the E4 isoform. Sixty percent of the *APOE* haplotypes are included in these two clades. Haplotypes in this grouping of clades have an A at position 560 and a T at position 832, whereas none of the haplotypes falling into the second clade that codes the E3 isoform and the clade that codes the E2 isoform have this pair of bases at the 560 and 832 sites. These 5′ sites subdivide the haplotypes that code the E3 structural isoform into those having decreased risk of high TG (Table 4), similar to the effects of haplotypes coding the E4 isoform, and those that have increased risk of high TG, similar to the effects of haplotypes coding the E2 isoform. The separate functional effects of the 5′ regulatory and exon 4 structural variations on lipid metabolism are further suggested by the statistical evidence for interaction of the effects of these two regions in predicting increased risk of high T-C in Danish females (Tables 4, 5). The 5′ variations subdivide the individuals bearing the E4 isoform into a high-risk group that carry the $A_{560}T_{832}$ haplotype and a group of individuals whose risk of high T-C is similar to that estimated for the group of individuals homozygous for the allele coding the E3 isoform. In summary, our statistical analyses suggest that mutational changes in the 5′ region of *APOE* have separate effects on lipid levels that modify the effects of the nonsynonymous changes in exon 4 that determine structural variations in the apoE molecule. This conclusion would not be possible using only the two 5′ SNPs, 560 and 832, to tag variation (55) in the four most common haplotype variations reported by Fullerton et al. (8), because the majority of these haplotypes fall within the E3 isoform group. Similarly, this conclusion would not be possible by measuring only the two cSNPs, 3937 and 4075, to tag the variation in the three common two-site ε*2*, ε*3*, and ε*4* haplotypes that define the structural variation of the apoE molecule, because they do not capture the functional variation among haplotypes that code the E3 isoform.

Artiga et al. (51) have found that variations in the 560 and 832 positions are associated with a significant heterogeneity in promoter activity in cell cultures. The $A_{560}T_{832}$ haplotype is associated with ~60% lower promoter activity than the $A_{560}G_{832}$ haplotype. Heterogeneity in *APOE* expression in vitro has been associated with heterogeneity in promoter activity (52). Endogenously synthesized apoE is important for cholesterol efflux from macrophages, which plays an essential role in lipid metabolism and the development of atherosclerosis (53, 54). Through its activity in macrophages, heterogeneity in 5′ promoter activity and expression of *APOE* may cause heterogeneity in reverse cholesterol transport, one of the main functions of HDL-C particles, which may result in pleiotropic effects on multiple lipid traits that are not a consequence of the structural changes in the apoE protein determined by variations in exon 4, which are primarily involved in binding by hepatic receptors of circulating atherogenic lipoprotein particles (16–18).

Applicability to clinics and public health

Knowledge that variation in a gene explains a statistically significant fraction of interindividual variation in a trait of interest highlights its pathogenetic involvement but may have limited usefulness in medical practice. Medical decisions follow from consideration of measures of health that are naturally dichotomous, or are dichotomized according to consensus statements that are based on interpretations of research findings, which is the case for the high-risk and low-risk categories defined by the National Cholesterol Education Program Expert Panel (24). The relevant question is how much greater, or smaller, is the prevalence of the discrete end point outcome in carriers of the proposed high-risk or low-risk genetic, or nongenetic, factor compared with the prevalence in the background population ignoring the risk factor information? The answer to this question is critically important in the evaluation of the medical utility of a hypothesized predictor in a particular population (56). Here, we used an ecological database-mining strategy and analyses of dichotomous end points to estimate genotype-specific probabilities of the end points of interest that are consistent with the needs of physicians and public health decision-makers. We found that variation in the 5′ promoter region of *APOE* has significant pleiotropic effects on multiple measures of lipid metabolism in three different populations. Three candidate high-risk 5′ genotypes identified subgroups of individuals at increased risk of low HDL-C, high TG, or high T-C in a fourth independently ascertained Danish population. As expected for a trait that has a complex multifactorial etiology, the increase in odds of the high-risk lipid classification in the carriers of high-risk genotypes, compared with the odds in individuals who do not carry the proposed high-risk genotypes, was modest (e.g., the odds of low HDL-C was 1.4 times higher in $A_{560}T_{832}/A_{560}T_{832}$ phenotype-carrying females than in other females). Similarly, the genotype-specific prevalences of low HDL-C, high TG, and high T-C did not differ markedly from the prevalences in the Danish population at large when genotype information is ignored. For instance, the prevalence of low HDL-C in Danish females who carry the $A_{560}T_{832}/A_{560}T_{832}$ genotype is 0.073, compared with 0.055 in the overall sample. Such a small difference is consistent with the small contribution of single-gene variation to the etiology of the lipid traits in the population at large and argues that the identified high-risk genotypes are of limited

value in medicine and public health in predicting dyslipidemia in the Danish population. Larger differences in prevalences might be expected in particular contexts defined by other genetic and environmental risk factors, in which case there would be fewer individuals considered to be at higher risk.

Conclusions

We used an ecological, multiple-population, data-mining strategy to identify variations in two SNPs in the 5′ promoter region of *APOE* that define genotype variations that distinguish between high and low HDL-C, TG, and/or T-C. These findings suggest that *APOE* has variations, not considered by the huge number of studies that have measured only the 3097 and 4075 cSNP variations in exon 4, that may be important in predicting dyslipidemia and, ultimately, the age of onset, progression, and severity of atherosclerotic disease. The evaluation of the hypothesized high-risk genotypes for predicting dyslipidemia in large population-based samples of females and males collected from Copenhagen established that prediction of low HDL-C or high TG is independent of whether variation in the two cSNPs is considered in the prediction model. The ability of these high-risk genotypes to predict high T-C was dependent on gender and the genotype defined by variation in the two cSNPs. Because of the role of context in the etiology of measures of lipid metabolism, the utility of the hypothesized high-risk genotypes for predicting dyslipidemia and related disease end points in other populations remains to be elucidated.

The authors thank Kenneth G. Weiss for his persistent attention to the details of the data management and statistical analyses. The technical support of Lynn Illeck in developing this article is also deeply appreciated. This work was supported by National Institutes of Health Grants HL-072905, HL-072810, GM-066509, HL-054481, HL-051021, HL-039107, HL-058238, HL-058239, and HL-058240.

REFERENCES

1. Steinberg, D. 2004. An interpretive history of the cholesterol controversy: part I. *J. Lipid Res.* **45:** 1583–1593.
2. Rader, D. J., and C. Maugeais. 2000. Genes influencing HDL metabolism: new perspectives and implications for atherosclerosis prevention. *Mol. Med. Today.* **6:** 170–175.
3. Rong, J. X., and E. A. Fisher. 2000. High-density lipoprotein: gene based approaches to the prevention of atherosclerosis. *Ann. Med.* **32:** 642–651.
4. Wang, X., and B. Paigen. 2002. Quantitative trait loci and candidate genes regulating HDL cholesterol: a murine chromosome map. *Arterioscler. Thromb. Vasc. Biol.* **22:** 1390–1401.
5. O'Rahilly, S., I. Barroso, and N. J. Wareham. 2005. Genetic factors in type 2 diabetes: the end of the beginning? *Science.* **307:** 370–373.
6. Crawford, D. C., C. S. Carlson, M. J. Rieder, D. P. Carrington, Q. Yi, J. D. Smith, M. A. Eberle, L. Kruglyak, and D. A. Nickerson. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74:** 610–622.
7. Nickerson, D. A. Molecular Diversity and Epidemiology of Common Disease (MDECODE). Common MDECODE SNP data. Accessed May 17, 2005 at http://droog.gs.washington.edu/mdecode/data.
8. Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor, J. H. Stengård, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, 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.* **67:** 881–900.
9. Nickerson, D. A., S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark, J. H. Stengård, V. Salomaa, E. Boerwinkle, and C. F. Sing. 2000. Sequence diversity and large scale typing of SNPs in the human apolipoprotein E gene. *Genome Res.* **10:** 1532–1545.
10. Fullerton, S. M., A. V. Buchanan, V. Sonpar, S. L. Taylor, J. D. Smith, C. S. Carlson, V. Salomaa, J. H. Stengård, E. Boerwinkle, A. G. Clark, et al. 2004. The effects of sale: variation in the *APOA1/C3/A4/A5* gene cluster. *Hum. Genet.* **115:** 36–56.
11. Sing, C. F., J. H. Stengård, and S. L. Kardia. 2003. Genes, environment, and cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.* **23:** 1190–1196.
12. Frikke-Schmidt, R. 2000. Context-dependent and invariant associations between APOE genotype and levels of lipoproteins and risk of ischemic heart disease: a review. *Scand. J. Clin. Lab. Invest. Suppl.* **233:** 3–25.
13. Lussier-Cacan, S., A. Bolduc, M. Xhignesse, T. Niyonsenga, and C. F. Sing. 2002. Impact of alcohol intake on measures of lipid metabolism depends on context defined by gender, body mass index, cigarette smoking, and *Apolipoprotein E* genotype. *Arterioscler. Thromb. Vasc. Biol.* **22:** 824–831.
14. Stengård, J. H., A. G. Clark, K. M. Weiss, S. Kardia, D. A. Nickerson, V. Salomaa, C. Ehnholm, E. Boerwinkle, and C. F. Sing. 2002. Contributions of 18 additional DNA sequence variations in the gene encoding apolipoprotein E to explaining variation in quantitative measures of lipid metabolism. *Am. J. Hum. Genet.* **71:** 501–517.
15. Hallman, D. M., E. Boerwinkle, N. Saha, C. Sandholzer, H. J. Menzel, A. Csazar, and G. Utermann. 1991. The apolipoprotein E polymorphism: a comparison of allele frequencies and effects in nine populations. *Am. J. Hum. Genet.* **49:** 338–349.
16. Davignon, J., R. E. Gregg, and C. F. Sing. 1988. Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis.* **8:** 1–21.
17. Mahley, R. W. 1988. Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. *Science.* **240:** 622–630.
18. Mahley, R. W., and S. C. Rall, Jr. 2000. Apolipoprotein E: far more than a lipid transport protein. *Annu. Rev. Genomics Hum. Genet.* **1:** 507–537.
19. Utermann, G., M. Hees, and A. Steinmetz. 1977. Polymorphism of apolipoprotein E and occurrence of dysbetalipoproteinemia in man. *Nature.* **269:** 604–607.
20. Kaprio, J., R. E. Ferrell, B. A. Kottke, M. I. Kamboh, and C. F. Sing. 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.
21. Last, J. M. 1988. A Dictionary of Epidemiology. 2nd edition. Oxford University Press, New York.
22. Ioannidis, J. P., E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis. 2001. Replication validity of genetic association studies. *Nat. Genet.* **29:** 306–309.
23. Freimer, N., and C. Sabatti. 2004. The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. *Nat. Genet.* **36:** 1045–1051.
24. National Cholesterol Education Program, National Heart, Lung, and Blood Institute. 2002. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). National Institutes of Health, Bethesda, MD. Publication No. 02-5215.
25. Moll, P. P., V. V. Michels, W. H. Weidman, and B. A. Kottke. 1989. Genetic determination of plasma apolipoprotein AI in a population-based sample. *Am. J. Hum. Genet.* **44:** 124–139.
26. Turner, S. T., W. H. Weidman, V. V. Michels, T. J. Reed, C. L. Ormson, T. Fuller, and C. F. Sing. 1989. Distribution of sodium-lithium countertransport and blood pressure in Caucasians five to eighty-nine years of age. *Hypertension.* **13:** 378–391.
27. Boerwinkle, E., C. A. Brown, M. Carrejo, R. Ferrell, C. Hanis, R. Hutchinson, S. Kardia, C. Sing, S. Turner, A. Weder, et al. 2002. Multi-center genetic study of hypertension: the Family Blood Pressure Program (FBPP). *Hypertension.* **39:** 3–9.
28. Salomaa, V., V. Rasi, J. Pekkanen, E. Vahtera, M. Jauhiainen, E. Vartiainen, G. Myllyla, and C. Ehnholm. 1994. Haemostatic factors

and prevalent coronary heart disease: the FINRISK Haemostasis Study. *Eur. Heart J.* **15:** 1293–1299.
29. Vartiainen, E., P. Puska, J. Pekkanen, J. Tuomilehto, and P. Jousilahti. 1994. Changes in risk factors explain changes in mortality from ischemic heart disease in Finland. *BMJ.* **309:** 23–27.
30. Stengård, J. H., V. Salomaa, V. Rasi, E. Vahtera, C. Ehnholm, T. Krusius, M. Perola, and E. Vartiainen. 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.
31. Schnohr, P., G. Jensen, P. Lange, H. Scharling, and M. Appleyard. 2001. The Copenhagen City Heart Study. Tables with data from the third examination 1991–1994. *Eur. Heart J.* **3 (Suppl. H):** 1–83.
32. Frikke-Schmidt, R., C. F. Sing, B. G. Nordestgaard, and A. Tybjaerg-Hansen. 2004. Gender- and age-specific contributions of additional DNA sequence variation in the 5′ regulatory region of the APOE gene to prediction of measures of lipid metabolism. *Hum. Genet.* **115:** 331–345.
33. National Institutes of Health. 1974. Lipid Research Clinics Program Manual of Laboratory Operations. Department of Health, Education, and Welfare, Washington, DC. Publication No. 75-628.
34. Barr, S. I., B. A. Kottke, and S. J. T. Mao. 1981. Improved method for determination of triglycerides in plasma lipoproteins by an enzymic kit method. *Clin. Chem.* **27:** 1142–1144.
35. Schiele, F., D. De Bacquer, M. Vincent-Viry, U. Beisiegel, C. Ehnholm, A. Evans, A. Kafatos, M. C. Martins, S. Sans, C. Sass, et al. 2000. Apolipoprotein E serum concentration and polymorphism in six European countries: the ApoEurope Project. *Atherosclerosis.* **152:** 475–488.
36. Long, J. C., R. C. Williams, and M. Urbanek. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56:** 799–810.
37. Nelson, M. R., S. L. Kardia, R. Ferrell, and C. F. Sing. 2001. A CPM to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* **11:** 458–470.
38. Kahn, H. A. 1983. An Introduction to Epidemiologic Methods. Oxford University Press, New York.
39. Sokal, R. R., and F. J. Rohlf. 1995. Biometry: The Principles and Practice of Statistics in Biological Research. 3rd edition. W. H. Freeman and Company, New York.
40. Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science.* **273:** 1516–1517.
41. Risch, N. 2005. 2004 Curt Stern Award address. The SNP endgame: a multidisciplinary approach. *Am. J. Hum. Genet.* **76:** 221–226.
42. Morowitz, H. 2001. Bacon, popper, and the human genome. *Complexity.* **6:** 14–15.
43. Hempel, C. G., and P. Oppenheim. 1948. Studies in the logic of explanation. *Philos. Sci.* **15:** 135–175.
44. Clark, A. G. 2000. Limits to prediction of phenotype from knowledge of genotypes. *In* Limits to Knowledge in Evolutionary Genetics. M. T. Clegg, M. Hecht, and R. J. MacIntyre, editors. Kluwer Academic/Plenum Publishers, New York. 205–224.
45. Rea, S. L., D. Wu, J. R. Cypser, J. W. Vaupel, and T. E. Johnson. 2005. A stress-sensitive reporter predicts longevity in isogenic populations of *Caenorhabditis elegans*. *Nat. Genet.* **37:** 894–898.
46. Elsasser, W. M. 1998. Reflections on a Theory of Organisms: Holism in Biology. Johns Hopkins University Press, Baltimore, MD.
47. Simon, H. A. 1996. The Sciences of the Artificial. 3rd edition. MIT Press, Cambridge, MA.
48. Stengård, J. H., K. E. Zerba, J. Pekkanen, C. Ehnholm, A. Nissinen, and C. F. Sing. 1995. Apolipoprotein E polymorphism predicts death from coronary heart disease in a longitudinal study of elderly Finnish men. *Circulation.* **91:** 265–269.
49. Frikke-Schmidt, R., B. G. Nordestgaard, B. Agerholm-Larsen, P. Schnohr, and A. Tybjaerg-Hansen. 2000. Context-dependent and invariant associations between lipids, lipoproteins, and apolipoproteins and apolipoprotein E genotype. A study of 9,060 women and men from the population at large. *J. Lipid Res.* **41:** 1812–1822.
50. Templeton, A. R., T. Maxwell, D. Posada, J. H. Stengård, E. Boerwinkle, and C. F. Sing. 2005. Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics.* **169:** 441–453.
51. Artiga, M. J., M. J. Bullido, I. Sastre, M. Recuero, M. A. Garcia, J. Aldudo, J. Vazquez, and F. Valdivieso. 1998. Allelic polymorphisms in the transcriptional regulatory region of apolipoprotein E gene. *FEBS Lett.* **421:** 105–108.
52. Lambert, J. C., T. Brousseau, V. Defosse, A. Evans, D. Arveiler, J. B. Ruidavets, B. Haas, J. P. Cambou, G. Luc, P. Ducimetiere, 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.
53. Bellosta, S., R. W. Mahley, D. A. Sanan, J. Murata, D. L. Newland, J. M. Taylor, and R. E. Pitas. 1995. Macrophage-specific expression of human apolipoprotein E reduces atherosclerosis in hypercholesterolemic apolipoprotein E-null mice. *J. Clin. Invest.* **96:** 2170–2179.
54. Davignon, J. 2005. Apolipoprotein E and atherosclerosis: beyond lipid effect. *Arterioscler. Thromb. Vasc. Biol.* **25:** 267–269.
55. The International HapMap Consortium Group. 2003. The International HapMap Project. *Nature.* **426:** 789–796.
56. Fletcher, R. H., S. W. Fletcher, and E. H. Wagner. 1988. Clinical Epidemiology: The Essentials. 2nd edition. Williams & Wilkins, Baltimore, MD.