Replication of genetic associations with plasma lipoprotein traits in a multiethnic sample

Matthew B. Lanktree,* Sonia S. Anand,†,§ Salim Yusuf,†,§ Robert A. Hegele,[1],* and the SHARE Investigators[2]

Robarts Research Institute,* Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada; Population Health Research Institute,† Hamilton Health Sciences, Hamilton, Ontario, Canada; and Departments of Medicine and Clinical Epidemiology,§ McMaster University, Hamilton, Ontario, Canada

Abstract Recent genome-wide association studies (GWAS) have reproducibly identified loci associated with plasma triglycerides (TG), HDL cholesterol, and LDL cholesterol. We sought to replicate these findings in a multiethnic population-based cohort using the curated single nucleotide polymorphism (SNP) set found on the new Illumina cardiovascular disease (CVD) beadchip, which contains approximately 50,000 SNPs densely mapping approximately 2,100 genes, selected based on their potential role in CVD. The sample consisted of individuals with European (n = 272), South Asian (n = 330), and Chinese (n = 304) ancestry. Identity by state clustering successfully classified individuals according to self-reported ethnicities. Associations between TG and *APOA5*, TG and *LPL*, HDL and *CETP*, and LDL and *APOE* were all identified ($P < 2 \times 10^{-6}$). In 13 loci, associations with the same SNP or a proxy SNP were identified in the same direction as previously reported ($P < 0.05$). Assessing the cumulative number of risk-associated alleles at multiple replicated SNPs increased the proportion of explained lipoprotein variance over and above traditional variables such as age, sex, body mass index, and ethnicity. The findings indicate the potential utility of the Illumina CVD beadchip, but they underscore the need to consider meta-analysis of results from commonly studied clinical or epidemiological samples.—Lanktree, M. B., S. S. Anand, S. Yusuf, and R. A. Hegele, and the SHARE Investigators. **Replication of genetic associations with plasma lipoprotein traits in a multiethnic sample.** ***J. Lipid Res.*** **2009.** 50: **1487–1496.**

Supplementary key words Beadchip • ethnicity • genomics • false discovery rate • identity by state • linear regression • lipids • permutation analysis • population

Plasma lipids, including cholesterol and triglyceride (TG), play vital roles in membrane fluidity, hormone and bile synthesis, and energy metabolism. The identification of genetic variants affecting lipoprotein traits, defined as plasma concentrations of TG, HDL, and LDL cholesterol, can give biological insight into both new and old pathways of lipid metabolism. These findings will have potential implications for the diagnosis, prognosis, and treatment of dyslipidemia. In the last two decades, the rare genetic variants responsible for many individually rare dyslipidemia conditions have been discovered, and common variations in many candidate genes have been tested for association with lipoprotein traits. However, in the last year, ten genome-wide association studies (GWAS) have consistently identified association between common genetic variation in multiple novel, as well as previously known, genes and lipoprotein traits in normolipidemic individuals (1–10). The identification of loci with previous evidence for roles in lipoprotein metabolism, such as association between the gene for apolipoprotein E (*APOE)* and LDL, serve as a positive control for the approach, while many new associations between lipoproteins and genes without a priori hypotheses were also uncovered (1–10). In total, 40 loci have been associated at a GWAS significance level with at least one of TG, HDL, or LDL (see supplementary Table I). Of note, in 13 of 15 genes previously identified to contain rare mutations causative for Mendelian lipid abnormalities, common single nucleotide polymorphism (SNP) variation within the same gene has been associated with the same lipoprotein trait that is primarily disturbed in the Mendelian disease. An important next step is replication of the findings in multiple ethnicities, both to ensure the findings are generalizable and to further delineate the effect size and location of the causative variants. This study set out to replicate

This work was supported by a grant from the Canadian Institute of Health Research, Genome Canada through the Ontario Genomics Institute, and by a Pfizer Jean Davignon Cardiovascular Grant. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET; www.sharcnet.ca).

Manuscript received 15 January 2009 and in revised form 13 March 2009.

Published, JLR Papers in Press, March 18, 2009.
DOI 10.1194/jlr.P900008-JLR200

Abbreviations: Apo, apolipoprotein; BMI, body mass index; CVD, cardiovascular disease; FDR, false discovery rate; FWER, family-wise error rate; GWAS, genome-wide association study; IBS, identity-by-state; SNP, single nucleotide polymorphism; TC, total cholesterol; TG, triglyceride.

[1] To whom correspondence should be addressed.
e-mail: hegele@robarts.ca

[2] The SHARE Investigators are listed in the Appendix.

S The online version of this article (available at http://www.jlr.org) contains supplementary data in the form of one table and three figures.

the previously identified GWAS findings in a multiethnic, population-based sample.

As microarray technology improves, the density of the arrays, or the number of SNPs evaluated on a single chip, increases. The improved density presents two possibilities: denser SNP coverage of the human genome or the ability to genotype multiple individuals on the same chip. The Illumina Human CVD beadchip uses the second approach, using multiple "wells" to interrogate approximately 50,000 SNPs in 12 individuals on a single array (11). The SNPs were selected for inclusion on the human CVD beadchip based on (1) early access to lipid, lipoprotein, and cardiovascular disease (CVD) GWAS results; (2) established quantitative trait loci in CVD; (3) genes with a functional link to CVD; and (4) a heavy bias for tagSNPs, nonsynonymous SNPs, and SNPs with known function (11). The selection criteria create a pool of SNPs with a higher prior probability of association, reducing both the false discovery rate (FDR) and the cost per sample. Permutation analysis has been suggested as a more appropriate method to correct for multiple testing (12), but it has not become standard practice for GWAS studies, at least partially due to the computational requirements.

We sought to replicate reported genetic associations with fasting TG, HDL, and LDL using approximately 50,000 SNPs in approximately 2,100 genes in a multiethnic population-based sample using the new Illumina CVD beadchip microarray and performing large-scale permutation analysis to improve signal-to-noise ratio and correct for multiple testing.

METHODS

Study subjects

The study was approved by the ethics boards of McMaster University and the University of Western Ontario. All participants provided informed consent for DNA analysis. The Study of Health Assessment and Risk in Ethnic Groups (SHARE) population was collected as a random prospective population sample in Hamilton, Toronto, and Edmonton as previously described (13). Individuals were classified as South Asian (n = 330) if their ancestors originated from India, Pakistan, Sri Lanka, or Bangladesh; Chinese (n = 304) if their ancestors originated from China, Taiwan, or Hong Kong; and European (n = 272) if their ancestors originated from Europe (13). All participants are between the ages of 35 and 75 years and have lived in Canada for five years or more. Anthropometric data was measured and fasting (over 12 h) and 2-h postglucose load blood samples were collected from study subjects. The following quantitative measures were obtained using established methodology: TG, total cholesterol (TC), apolipoprotein B (apoB), VLDL cholesterol, and HDL cholesterol. LDL cholesterol was calculated via the Friedewald equation. Relevant baseline characteristics are shown in **Table 1** (see supplementary Fig. III for trait distributions). Sixty-seven individuals (7.4%) were on lipid-lowering therapy and were excluded from further analysis.

Biochemical analyses

Genomic DNA was extracted from leukocytes as previously described (14). Whole genomic DNA was checked for quality by 1.5% agarose gel electrophoresis. DNA was diluted to 50–70 ng/ul, and the concentration was verified using a Nanodrop spectrophotometer. Standard protocols for hybridization and scanning of the Illumina Human CVD beadchip (version 1) on the Illumina BeadStation 500G were used for genotyping at the Centre for Applied Genomics (TCAG) (Hospital for Sick Children, Toronto, Ontario, Canada; www.tcag.ca). Briefly, approximately 200 ng (4 uL at 50–70 ng/uL) of double-stranded genomic DNA was added to a whole genome amplification reaction producing fragments of approximately 1.5–2 kb in length. Enzymatic fragmentation, followed by purification, produces 200–600 bp fragments for hybridization to the beadchips. Each bead contains many oligonucleotides to measure the presence of a single allele, with approximately 30 replicates of each bead randomly distributed on the beadchip. During chip quality control performed by Illumina (San Diego, CA; www.illumina.com), the location of the bead replicates are identified for each chip and distributed on a DVD with the beadchip. Each beadchip contains wells allowing 12 samples to run concurrently. Genotyping and quality control were performed in Illumina's BeadStudio Genotyping Module v3.2. Sixty-seven individuals (7.4%) were excluded because they were on lipid-lowering therapy, and 11 individuals (1.3%) were excluded from the analysis due to genotype call rates less than 95%. 1,151 SNPs (2.3%) were excluded from the analysis due to genotype call rates less than 95%. SNPs that were not in Hardy-Weinberg equilibrium (HWE) ($P < 0.0001$) or with a minor allele frequency less than 0.01 were excluded, leaving 35,303, 31,751, and 35,018 SNPs in South Asian, Chinese, and Caucasian samples, respectively. Due to the marginal power of the SHARE sample and given the effect size of many of the recently reported associations, only SNPs that were prevalent in all three populations (MAF > 0.01) were included in the analysis. The intersection of these three population sets left 29,377 SNPs, which were studied in the final analyses.

Statistical methods

Pairwise identity-by-state (IBS) distance and multi-dimensional scaling as implemented in PLINK (12) was used to test for popu-

TABLE 1. Baseline clinical characteristics in the multiethnic SHARE study

	South Asian	Chinese	Caucasian	Total sample
n	330	304	272	906
Male (%)	55	51	49	52
Age	49.5 (9.3)	47.8 (8.9)	51.2 (11.0)	49.5 (9.8)
BMI (kg/m^2)	26.3 (4.2)	24.0 (3.6)	27.4 (4.6)	25.9 (4.4)
LDL (mmol/L)	3.30 (0.82)	3.18 (0.81)	3.28 (0.82)	3.21 (0.81)
HDL (mmol/L)	1.03 (0.30)	1.19 (0.38)	1.20 (0.37)	1.13 (0.35)
FTG (mmol/L)	2.00 (1.3)	1.65 (1.2)	1.55 (1.1)	1.77 (1.3)
apoB (g/L)	1.08 (0.26)	1.00 (0.25)	1.01 (0.24)	1.03 (0.26)
NFTG (mmol/L)	1.91 (1.16)	1.67 (1.34)	1.55 (1.20)	1.72 (1.25)

Apo, apolipoprotein; BMI, body mass index; FTG, fasting triglyceride; NFTG, nonfasting triglyceride; SHARE, Study of Health Assessment and Risk in Ethnic Groups. Standard deviation given in parentheses.

lation stratification, sample duplication, or contamination. In IBS, a similarity matrix is produced by computing the proportion of the number of alleles shared between all pairs of individuals. Reducing the number of dimensions by classical (metric) multidimensional scaling enables the subjects to be drawn on a two-dimensional plot. All association analysis was performed in PLINK (12). The reported linear regression significance is for a codominant model, testing for an additive effect of allele dosage, with the asymptotic *P* value of the t-statistic reported using age, sex, BMI, and ethnicity as covariates. The functional relevance or previous associations reported for the SNPs found on the CVD beadchip and the density of markers in the candidate loci renders the Bonferroni correction particularly over-conservative. FDR control is a statistical method, less conservative and more powerful than family-wise error rate (FWER) control, used to correct for multiple comparisons and determine an appropriate significance threshold that reduces the probability of errors in the rejected hypotheses (15). The Benjamini Hochberg FDR procedure (15), as used by Sabatti et al. (10), was used to iteratively examine the most significant associations. However, the Benjamini Hochberg FDR procedure is only valid when the tests are independent (15); therefore, to be conservative, only significant SNPs found on separate chromosomes were included in the procedure. A significance threshold of 2.27×10^{-6} was calculated to control the FDR across the three reported traits (a total of 29,377 SNPs × 3 traits = 88,131 tests) at a $P = 0.05$ level. The additional two traits [apoB and nonfasting TG (see supplementary Fig. II)] were highly correlated with the traits reported here and including them in FDR calculations would artificially reduce power (10). To further assess significance and correct for multiple-testing, 500,000 label-swapping permutations were computed on the "whale" cluster of the Shared Hierarchical Academic Research Computing Network (SHARCNET; www.sharcnet.ca). A shell script was written to call 200 instances of PLINK per trait, each on a different CPU, and the PLINK "mperm" command was used to generate 2,500 permutations (200 × 2,500 = 500,000 permutations/trait). In each permutation, each individual's quantitative trait value is randomly assigned to a different individual's genotype set, and regression is performed on all SNPs. The procedure is repeated to create a distribution of all possible regression *P* values for all SNPs. The actual significance for each SNP is compared with the distribution of possible results from all 500,000 permutations of all the SNPs to calculate an empirical significance value as follows:

$$P = \frac{\{\sum_i p_i(N_i + 1) - 1\} + 1}{\{\sum_i N_i\} + 1} = \frac{\{\sum_i p_i(2501) - 1\} + 1}{500001}$$

where P is the overall empirical *P* value, p_i is the empirical *P* value and N_i is the number of permutations in the i^{th} instance of PLINK (12). Since genotype data is unaffected, linkage disequilibrium between SNPs remains throughout permutations. Results of association and permutation were displayed using WGAViewer (16).

RESULTS

Three distinct clusters were identified by IBS and multidimensional scaling (**Fig. 1**). Points were then colored by self-reported ethnicity, and all but two individuals fell into their respective clusters. The two individuals who did not cluster appropriately were removed from further analysis.

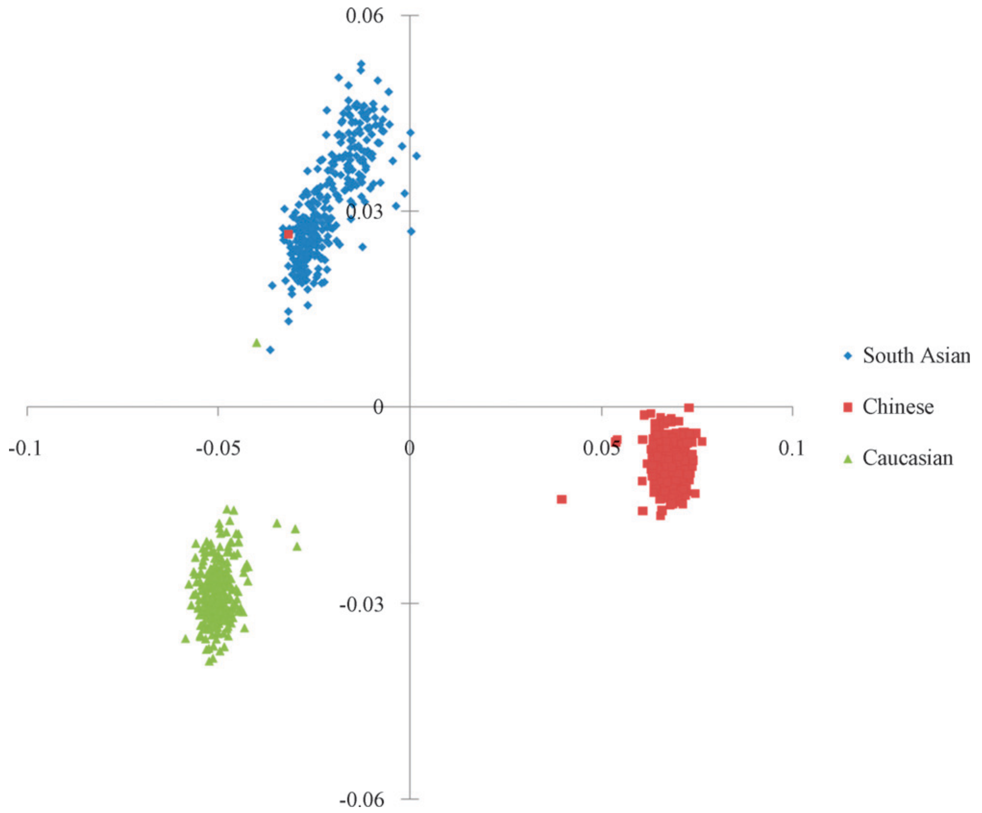


Fig. 1. Clustering of IBS scores after multi-dimensional scaling reveals the three distinct population clusters in the SHARE sample. Two individuals did not cluster with self-reported ethnicity and were excluded from further analysis.

Among the SHARE participants (Table 1), 27 SNPs in 4 loci were associated with a lipoprotein trait at a Benjamini Hochberg FDR–corrected significance ($P < 2.27 \times 10^{-6}$) (**Fig. 2**). The strongest association was seen between two variants just upstream of the *APOA5* gene and TG concentrations (rs651821 and rs662799; $P = 5.5 \times 10^{-12}$). *APOA5* lies within a cluster of apolipoprotein genes (*APOA1/A4/A5/C3*), resulting from an ancestral gene duplication event (17) in which variants have been consistently reported to be associated with TG and HDL (2, 7). Associations between SNPs within the *LPL* gene and TG (lead SNP: rs13702; $P = 1.7 \times 10^{-6}$), the *CETP* gene and HDL concentrations (lead SNP: rs9939224; $P = 6.2 \times 10^{-7}$), and the *APOE* gene and LDL [lead SNP: rs7412; $P = 1.7 \times 10^{-6}$ (see supplementary Fig. I for Q-Q plots)] were also identified below the Benjamini Hochberg FDR threshold. No SNPs located outside of the previously reported loci were associated after Benjamini Hochberg FDR correction. After using max(T) permutation to empirically derive significance corrected for multiple testing, association of the *APOA5* and *LPL* loci with TG, the *CETP* locus and HDL and the *APOE* locus and LDL remained ($P < 0.05$). The apparent signal-to-noise ratio improved after max(T) permutation (Fig. 2). The total CPU time for permutation analysis was over 10 years (480 CPUs × 10 days).

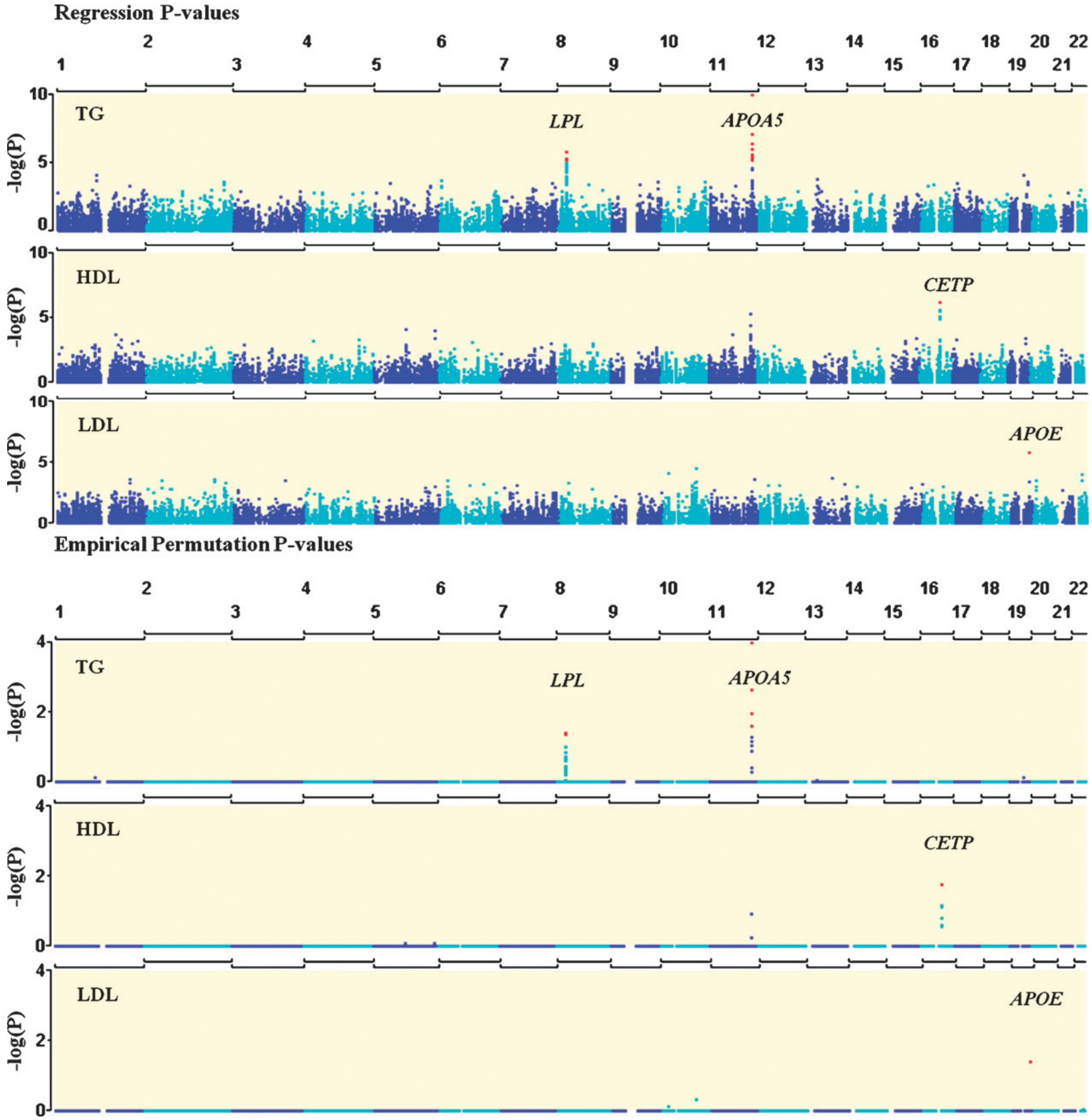


Fig. 2. Manhattan plots of regression and permutation results. Each point represents the −log(*P*) value for a single SNP linear regression including age, sex, BMI, and ethnicity as covariates. Improvement of signal-to-noise ratio upon 500,00 label-swapping max(T) permutations is seen in the bottom three graphs, in which the corrected empirical *P* values are reported. All points in the top graph are in the bottom graph, but the *P* value of many points have approached 1. $P < 2.2 \times 10^{-6}$ are highlighted in red in the top three graphs; $P < 0.05$ are highlighted in red in the bottom three graphs.

TABLE 2. Replication of TG-associated loci in the multiethnic SHARE sample

Gene	No. of SNPs[a]	SNP	Allele	MAF			Effect (β)	*P*
				SA	CH	EC		
APOA5	38	rs662799[b]	G	0.20	0.26	0.04	0.097	5.5×10^{-12}
LPL	46	rs13702[b]	G	0.28	0.25	0.32	−0.057	1.7×10^{-6}
		rs328	G	0.10	0.14	0.08	−0.063	0.00026
APOE	33	rs7412	T	0.01	0.08	0.08	0.094	0.00025
		rs439401	G	0.44	0.43	0.60	0.028	0.013
ANGPTL3-DOCK7	6	rs1748197	A	0.08	0.22	0.33	−0.030	0.0024
		rs1748195	G	0.47	0.24	0.32	−0.030	0.012
GCKR	8	rs1260326	A	0.21	0.48	0.41	0.031	0.0085
MLXIPL	13	rs17145738	T	0.08	0.11	0.19	−0.037	0.023

FDR, false discovery rate; MAF, minor allele frequency; SHARE, Study of Health Assessment and Risk in Ethnic Groups; TG, triglyceride; SA, South Asian; CH, Chinese; EC, European Caucasian. *P* (significance) and β (regression coefficient) for linear regression of lipoprotein trait versus number of minor alleles, including age, sex, BMI, and ethnicity as covariates.

[a] Number of SNPs within 50 kb of loci that passed quality control.

[b] Met significance criteria after FDR correction and permutation analysis.

Of the 40 lipoprotein associations previously reported, 32 were represented on the Illumina CVD beadchip and 22 of these loci contained a nominally associated SNP ($P < 0.05$). In a thorough examination of the linkage disequilibrium surrounding the previously reported SNPs, either the lead SNP or the SNP reported to be strongest associated to a trait in a previously reported study, or its proxy, a nearby SNP correlated with the lead SNP, was associated in the same direction with the same lipid fraction in our multiethnic sample for 13 loci ($P < 0.05$) (**Tables 2**, **3**, and **4**). Strong correlation was observed between fasting and postprandial TG concentrations ($r = 0.96$; $P < 0.0001$) and between LDL and apoB ($r = 0.86$; $P < 0.0001$); thus, similar associations were observed between the pairs of traits (see supplementary Fig. II for Manhattan and Q-Q plots).

Focusing on one locus recently discovered using the GWAS approach, the strength of association and the direction and size of the effect in the three populations in the SHARE sample appears to focus the region of association. In the *CELSR2/PSRC1/SORT1* locus, the strongest association with plasma LDL was found centered over the *PSRC1* and *CELSR2* genes [(rs657420, $P = 0.0047$ (**Fig. 3**)]. Nearby rs646776 and rs12740374 were found to be associated in the same direction with similar effect size to earlier reports (8–10). Between *PSRC1* and *SORT1*, a recombination hotspot has been reported (2) (Fig. 3), and toward *SORT1*, the effect of the minor allele becomes discordant between ethnicities.

To identify the cumulative effect of multiple-risk alleles on plasma lipoprotein concentrations and the total proportion of trait variation that can be explained by the replicated genetic factors in a multiethnic sample, we performed a multivariate linear regression. A model was created to include age, sex, BMI, and ethnicity as covariates, as well as the sum of the number of risk alleles at the replicated SNPs for each of the lipoprotein traits (for TG, rs662799, rs13702, rs7412, rs1748197, rs1260326, and rs17145738; for HDL, rs9939224, rs651821, rs331, rs4775041, rs4149327, rs12726525, and rs7120118; and for LDL, rs7412, rs6413504, rs657420, and rs3761739). It should be noted that many of the SNPs included in the model were not associated at the FDR threshold but were previously reported and nominally associated in the SHARE sample ($P < 0.05$) (Tables 2, 3, and 4). A significant relationship was identified between the number of risk alleles and plasma levels of TG, HDL, and LDL [$P < 0.0001$ (**Fig. 4**)]. The model incorporating age, sex, BMI, ethnicity and the replicated

TABLE 3. Replication of HDL-associated loci in the multiethnic SHARE sample

Gene	No. of SNPs[a]	Lead SNP	Allele	MAF			Effect (β)	*P*
				SA	CH	EC		
CETP	61	rs9939224[b]	A	0.22	0.14	0.19	−0.093	6.2×10^{-7}
		rs3764261	A	0.33	0.18	0.33	0.073	1.2×10^{-5}
APOA5	83	rs651821	G	0.19	0.26	0.04	−0.088	5.3×10^{-6}
LPL	46	rs331	A	0.18	0.23	0.30	0.054	0.00012
		rs328	G	0.10	0.14	0.08	0.06	0.012
LIPC	132	rs4775041	G	0.23	0.21	0.31	0.049	0.0037
ABCA1	107	rs4149327	C	0.19	0.44	0.10	−0.041	0.011
GALNT2	67	rs12726525	A	0.19	0.27	0.29	−0.038	0.029
		rs4846914	A	0.43	0.24	0.58	0.027	0.086
NR1H3-FOLH1	11	rs7120118	G	0.40	0.27	0.27	0.031	0.038

FDR, false discovery rate; MAF, minor allele frequency; SHARE, Study of Health Assessment and Risk in Ethnic Groups; TG, triglyceride; SA, South Asian; CH, Chinese; EC, European Caucasian. *P* (significance) and β (regression coefficient) for linear regression of lipoprotein trait versus number of minor alleles, including age, sex, BMI, and ethnicity as covariates.

[a] Number of SNPs within 50 kb of loci that passed quality control.

[b] Met significance criteria after FDR correction and permutation analysis.

TABLE 4. Replication of LDL-associated loci in the multiethnic SHARE sample

Gene	No. of SNPs[a]	Lead SNP	Allele	MAF			Effect (β)	P
				SA	CH	EC		
APOE	33	rs7412[b]	T	0.01	0.08	0.08	−0.43	1.7×10^{-6}
		rs2075650	G	0.09	0.05	0.18	0.15	0.019
LDLR	35	rs6413504	G	0.47	0.32	0.48	0.11	0.0029
		rs6511721	G	0.45	0.27	0.45	0.10	0.013
CELSR2-SORT1	61	rs657420	G	0.48	0.49	0.47	−0.11	0.0047
		rs646776	G	0.24	0.04	0.20	−0.12	0.025
		rs12740374	A	0.24	0.04	0.20	−0.12	0.027
HMGCR	14	rs3761739	A	0.14	0.20	0.13	0.11	0.024
		rs12654264	A	0.40	0.46	0.64	−0.07	0.066

FDR, false discovery rate; MAF, minor allele frequency; SHARE, Study of Health Assessment and Risk in Ethnic Groups; TG, triglyceride; SA, South Asian; CH, Chinese; EC, European Caucasian. P (significance) and β (regression coefficient) for linear regression of lipoprotein trait versus number of minor alleles, including age, sex, BMI, and ethnicity as covariates.

[a] Number of SNPs within 50 kb of loci that passed quality control.

[b] Met significance criteria after FDR correction and permutation analysis.

SNPs accounted for 25% of the variation in TG concentrations, 33.5% of the variation in HDL concentrations, and 14.2% of the variation in LDL concentrations (**Fig. 5**). The average TG concentration increased from 1.28 mmol/L for those with 3 or less risk alleles to 2.33 mmol/L for those with 9 or more risk alleles. The HDL concentration decreased from 1.59 mmol/L for those with 2 or less risk alleles to 0.85 mmol/L for those with 9 or more risk alleles. Finally, the average LDL concentration increased from 2.97 mmol/L for those with 2 or less risk alleles to 3.87 mmol/L for those with 7 or more risk alleles.

DISCUSSION

In a multiethnic, population-based sample, we attempted to replicate genetic associations with plasma lipoprotein traits using the recently developed, curated Illumina CVD beadchip. SNPs within 4 loci (*APOA5*, *LPL*, *CETP*, and *APOE*) were robustly associated with lipoprotein traits, using either a Benjamini Hochberg FDR–corrected significance or permutation analysis. These 4 loci have established roles in inter-individual differences in plasma lipoprotein concentration from Mendelian, candidate gene and GWAS investigations. Of 40 loci previously associated with at least one plasma lipoprotein fraction, 32 were represented on the Illumina CVD beadchip, and 22 contained a SNP that was nominally associated at an uncorrected significance ($P < 0.05$). The lead SNP or proxy was associated in the same direction as previously reported in 13 loci ($P < 0.05$). The major limitation, and potentially the cause of the non-replication of more loci, was the limited sample size. We believe that the this work illustrates two important concepts: (1) Common variants represented in multiple ethnicities often show associations with the same directionality, indicating differences in association between ethnicities are more likely to be due to differences in allele frequencies or haplotype structures than differences in the real effect;

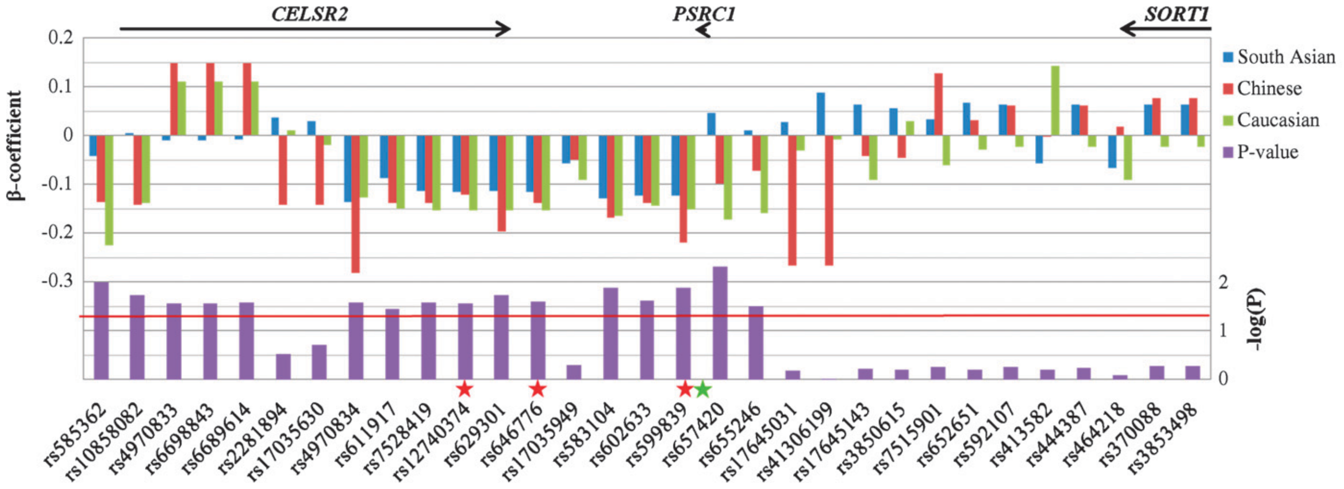
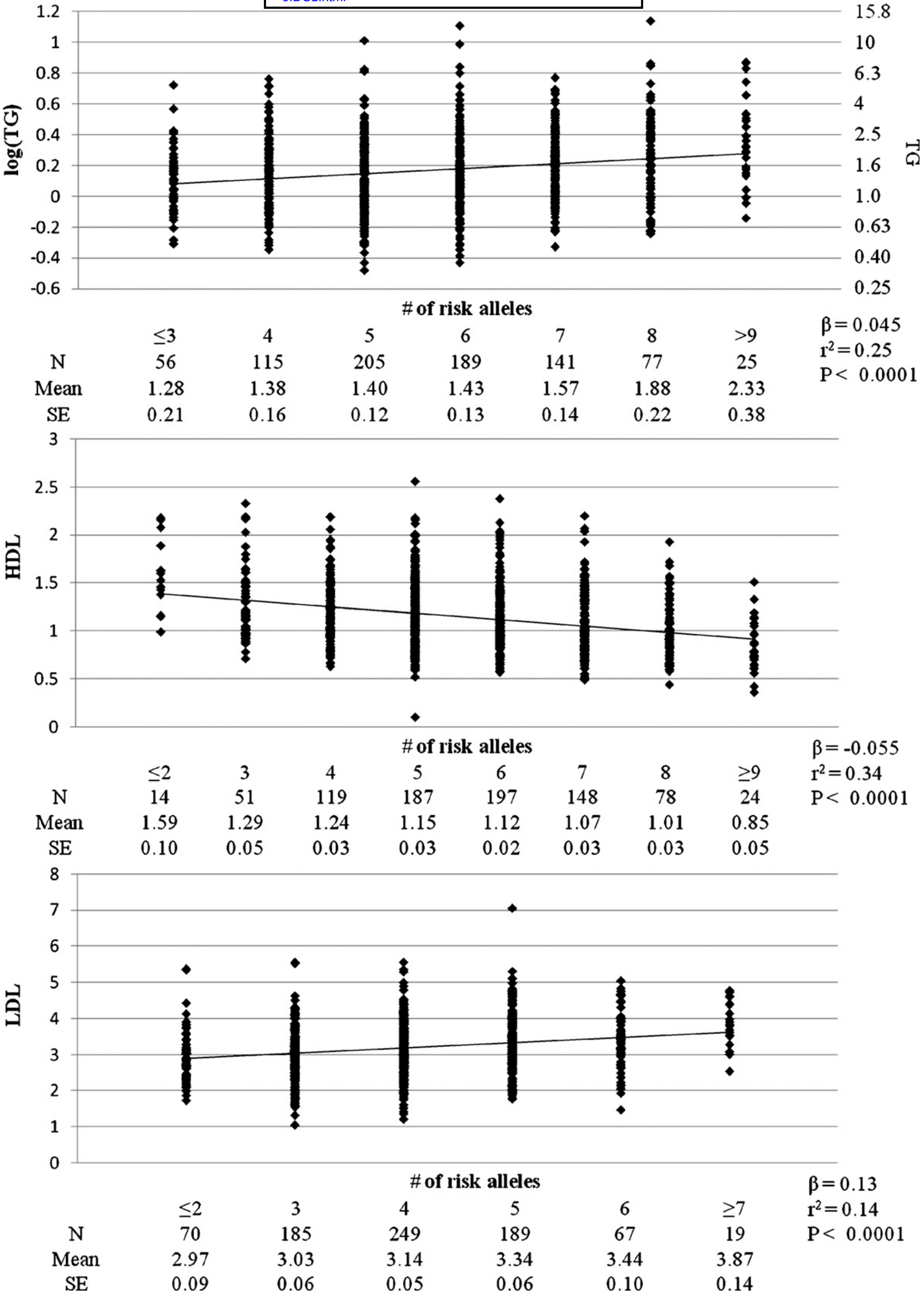


Fig. 3. Replication of association between the *CELSR2/PSRC1/SORT1* locus and plasma LDL cholesterol concentrations. The upper bars depict the size and direction of the effect of the minor allele, colored by population (using the β-coefficient on the left y-axis for scale). The lower bars depict the probability of SNP association (transformed by −log) in the complete sample with age, sex, BMI, and ethnicity as covariates. The red line indicates a P value of 0.05. The green star indicates a recombination hotspot, and the red stars indicate SNPs which were previously reported as the holding the strongest association with LDL cholesterol.



	≤3	4	5	6	7	8	>9
N	56	115	205	189	141	77	25
Mean	1.28	1.38	1.40	1.43	1.57	1.88	2.33
SE	0.21	0.16	0.12	0.13	0.14	0.22	0.38

$\beta = 0.045$
$r^2 = 0.25$
$P < 0.0001$

	≤2	3	4	5	6	7	8	≥9
N	14	51	119	187	197	148	78	24
Mean	1.59	1.29	1.24	1.15	1.12	1.07	1.01	0.85
SE	0.10	0.05	0.03	0.03	0.02	0.03	0.03	0.05

$\beta = -0.055$
$r^2 = 0.34$
$P < 0.0001$

	≤2	3	4	5	6	≥7
N	70	185	249	189	67	19
Mean	2.97	3.03	3.14	3.34	3.44	3.87
SE	0.09	0.06	0.05	0.06	0.10	0.14

$\beta = 0.13$
$r^2 = 0.14$
$P < 0.0001$

Fig. 4. Regression analyses of the relationship between plasma lipoprotein concentration and the cumulative number of risk alleles from multiple replicated loci. Risk alleles were tallied from genotypes of rs662799, rs13702, rs7412, rs1748197, rs1260326, and rs17145738 for TG; rs9939224, rs651821, rs331, rs4775041, rs4149327, rs12726525, and rs7120118 for HDL; and rs7412, rs6413504, rs657420, and rs3761739 for LDL. Lipoprotein measurements are in mmol/L. β-coefficient is the effect size of each additional risk allele standardized to the standard deviation of the trait; r^2 is the correlation between the lipid trait and the model, including age, sex, BMI, ethnicity, and number of risk alleles. The *P* value represents the significance of the regression.

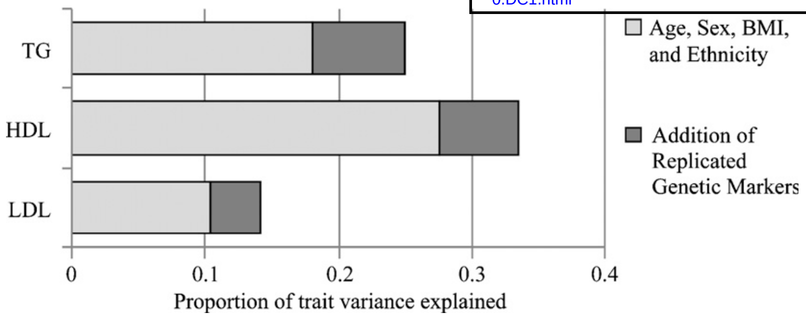


Fig. 5. Proportion of lipoprotein trait variation explained by age, sex, BMI, ethnicity, and with or without replicated genetic markers as shown in Fig. 4.

and (2) the recently reported findings, while of great biological interest, are often not replicable in clinical- or modest-sized epidemiological samples, especially if a genome-wide significance threshold is required. We do not believe that nonreplication of the associations identified in the much larger GWAS studies reduces the credibility of their findings, but we underscore the importance of the associations that are robust to ethnicity and identifiable in a well-studied clinical research cohort of approximately 900 individuals.

The *APOE* ε2 / ε3 / ε4 isoform system has long been associated with plasma lipid concentrations (18, 19). Within the *APOE* locus, the most significantly associated SNP was rs7412, of which a T allele codes for an amino acid change from arginine to cysteine at position 158, guaranteeing an ε2 isoform. Unfortunately rs429358, which differentiates between ε3 and ε4 isoforms, was not included on the Illumina CVD beadarray, and thus, rs7412 was evaluated in the same unbiased manner as all other SNPs on the beadchip. With each additional copy of the T allele at rs7412 and each additional copy of the ε2 isoform, we observe a decrease in LDL cholesterol and an increase in TG concentration ($P <$ 0.0005), consistent with previously reported results of ε2 being associated with lower LDL cholesterol and higher TG concentrations compared with ε3 or ε4 isoforms. In a separate report, the *APOE* ε2 / ε3 / ε4 genotype was determined in the SHARE sample through restriction isotyping and was found to be associated with plasma lipoprotein concentrations in all three ethnicities, essentially following the ε2 / ε3 / ε4 gradient that has been observed historically in many other samples (20). Interestingly, neither rs7412 nor rs429358 has been mentioned in any of the previous GWAS reports, and they are not included on Affymetrix Genome-Wide SNP 6.0 array.

What is the advantage of using a multiethnic sample? Differences in minor allele frequency and background trait variation between ethnicities lead to differences in power and, subsequently, differences in the significance of the association test. Gene-gene and gene-environment interactions are capable of reducing the effect of a variant, theoretically producing a possible mechanism for nonreplication in a different ethnicity. Many examples of the "flip-flop" phenomenon, in which alternate alleles are associated with disease susceptibility in different ethnicities, have been reported (21). Initially thought to be the result of spurious association, mechanisms to explain the "flip-flop" phenomenon include differences in genetic background or environment between ethnicities, but they more likely result from differences in linkage disequilibrium between ethnicities (21). In general, if the variant queried is truly the responsible variant affecting the expression level, protein structure, or protein function, then the association should be in the same direction across ethnic groups (22). Regions where SNPs are associated in the same direction in a multiethnic sample are likely to be close to the causative variant. If an ancestral recombination occurred between the tested SNP and the causative variant, the opposite allele of the tested SNP would become linked to the causative variant. The likelihood of recombination occurring between a SNP and a causative variant increases with genetic distance. Therefore, for SNPs close to the functional variant, the same allele should be associated in the same direction, with the "flip-flop" phenomenon occurring for SNPs a greater distance from the causative variant. In a multiethnic sample, more ancestral recombination has occurred, producing different LD structures and the possibility of more accurately mapping risk loci.

For example, in the *CELSR2/PSRC1/SORT1* locus, a recombination hotspot has been reported (2) in the same region where discordance in the direction of association between ethnicities occurs in Fig. 3. One might expect the discordance of effect direction if an ancestral recombination occurred leaving opposite alleles linked with the causative variant in subsequent generations. Despite the previously reported associations over *CELSR2* and *PSRC1* (rs646776 and rs12740374) (3, 5, 9, 10) due to a more plausible functional role for *SORT1* in LDL metabolism (3), attention toward *CELSR2/PSRC1/SORT1* locus has been primarily focused on *SORT1* (3). Caution is required due to the marginal nature of the *P* values, but our work suggests that the causative variants responsible for the *CELSR2/PSRC1/SORT1* association with LDL is more likely to lie within or near the *CELSR2* or *PSRC1* genes, approximately 25kb upstream of *SORT1*. If *SORT1* is responsible for the variation in LDL cholesterol, it appears likely that the associated variant acts through a distant *cis*-acting element, telomeric to the putative recombination hotspot.

Computational requirements pose a clear impediment to permutation analysis of GWAS datasets. A benefit of label-swapping permutations is that the LD structure between SNPs remains, thus removing the assumption that each SNP is an independent test, theoretically creating a less conservative correction (12). Adaptive permutation has been proposed, in which SNPs are dropped from further permutation if they are clearly not associated. However, determining the appropriate significance threshold is still difficult, and the problem of multiple-testing remains. In max(T) permutation using the PLINK "mperm" command, the significance can be compared with the distribution of possible regression results for all SNPs, more directly addressing the problem of multiple testing (12, 23). Serial farming, in which hundreds (or thousands) of CPUs perform the permutations on a large cluster, creates an environment in which large-scale permutation analysis of GWAS datasets is possible. The true validity of permutation analysis is whether it can separate biologically valid associations from spurious associations more effectively than simply picking the next most significant *P* value from a list of test statistics. Further application of permutation testing and downstream examination of results will be required. However, in this dataset, the results of 500,000 label-swapping max(T) permutations and a simple multiple-testing correction were not very different. Our findings suggest that permutation analysis, while clarifying the signal-to-noise ratio of *P* values across the range of SNPs evaluated, did not fundamentally change the conclusions from a Benjamini Hochberg FDR–corrected significance generated by the linear regression in PLINK. This suggests that such a computationally intensive analysis may not necessarily yield a substantial improvement.

Studies of coronary artery disease (24), type 2 diabetes (1), and hypertriglyceridemia (25) have reported an increased odds ratio for disease development with the cumulative number of risk alleles. The number of individuals observed decreases with each additional risk allele, as predicted by the minor allele frequencies, and the risk for disease increases until the risk approaches Mendelian proportions (25). One study in Caucasians showed that the number of lipoprotein risk alleles was associated with the quantitative differences in plasma LDL and HDL cholesterol concentrations (24). In a similar fashion, using our multiethnic sample and SNPs from replicated loci specific for the lipoprotein trait instead of a general lipid panel, we were able to show that a relationship exists between the cumulative number of risk alleles and plasma lipoprotein concentrations, with slightly larger effect than the earlier report. Moreover, the inclusion of the genetic markers increased the proportion of trait variance that could be explained versus age, sex, BMI, and ethnicity alone.

The major advantages of the Illumina CVD beadchip is the reduced cost per sample and reduced number of tests due to the curated nature of the SNPs. However, with the recent report of additional lipid associated loci (8–10), the chip is already becoming obsolete. As discussed above, the most substantial limitation of this investigation was the power of the study. The studies used to originally identify the GWAS findings typically included samples of over 2,000 and up to 20,000 individuals (3–10). A multiethnic sample of 900 individuals has the same power as a sample of 900 individuals of homogenous ethnicity only if the size and direction of the effect is similar between populations. As we were concerned with the restricted power afforded by the individual ethnicity samples, we examined only SNPs that were prevalent in all populations. Our goal was to identify the variants that display a common effect across ethnicities, not to identify variants that account for the differences between ethnicities. Hopefully future collaborations and pooling of results will enable the joint analysis of this dataset with the results of additional CVD beadchip studies. The approach used in this study, comparing the effect direction between ethnicities, holds the potential to narrow associated regions and assist in the identification of the functionally responsible variants.

In conclusion, this study successfully replicated associations between *APOA5* and TG, *LPL* and TG, *CETP* and HDL, and *APOE* and LDL in a multiethnic sample using the new curated Illumina CVD beadchip. Associations ($P < 0.05$) were identified in 22 of the 32 previously identified lipid loci represented on the CVD beadchip, and the previously reported lead SNP or its proxy was associated in 13 of 32 loci. Label-swapping max(T) permutation performed in a cluster environment is a feasible method for multiple-testing correction in GWAS, but further studies will be required to determine its benefit over standard Bonferroni correction. Comparison of effect size and direction between multiple ethnicities could potentially be used to refine associated regions, as demonstrated here in the *CELSR2/PSRC1/SORT1* locus. Finally, relationships between the cumulative number of risk alleles and TG, HDL cholesterol, and LDL cholesterol were observed. The findings indicate the potential utility of the Illumina CVD beadchip, but they underscore the importance of both sample and effect size, and the need to consider meta-analysis of results from commonly studied clinical or epidemiological samples.

APPENDIX

The SHARE Investigators are as follows: S. S. Anand, S. Yusuf, V. Vuksan, S. Devanesen, P. Montague, L. Kelemen, C. Sigouin, K. K. Teo, E. Lonn, H. C. Gerstein, R. A. Hegele, and M. McQueen.

M.B.L. is supported by the Canadian Institute of Health Research vascular research fellowship program and MD/PhD Studentship Award, as well as the University of Western Ontario MD/PhD program. R.A.H. is a Career Investigator of the Heart and Stroke Foundation of Ontario and holds the Edith Schulich Vinet Canada Research Chair (Tier I) in Human Genetics, the Martha G. Blackburn Chair in Cardiovascular Genetics, and the Jacob J. Wolfe Distinguished Medical Research Chair at the University of Western Ontario.

REFERENCES

1. Saxena, R., B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. **316:** 1331–1336.
2. Chasman, D. I., G. Pare, R. Y. L. Zee, A. N. Parker, N. R. Cook, J. E. Buring, D. J. Kwiatkowski, L. M. Rose, J. D. Smith, P. T. Williams, et al. 2008. Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and apolipoprotein B among 6382 white women in genome-wide analysis with replication. *Circ. Cardiovasc. Genet.* **1:** 21–31.

3. Kathiresan, S., O. Melander, C. Guiducci, A. Surti, N. P. Burtt, M. J. Rieder, G. M. Cooper, C. Roos, B. F. Voight, A. S. Havulinna, et al. 2008. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* **40:** 189–197.
4. Kooner, J. S., J. C. Chambers, C. A. Aguilar-Salinas, D. A. Hinds, C. L. Hyde, G. R. Warnes, F. J. Gomez Perez, K. A. Frazer, P. Elliott, J. Scott, et al. 2008. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat. Genet.* **40:** 149–151.
5. Sandhu, M. S., D. M. Waterworth, S. L. Debenham, E. Wheeler, K. Papadakis, J. H. Zhao, K. Song, X. Yuan, T. Johnson, S. Ashford, et al. 2008. LDL-cholesterol concentrations: a genome-wide association study. *Lancet.* **371:** 483–491.
6. Wallace, C., S. J. Newhouse, P. Braund, F. Zhang, M. Tobin, M. Falchi, K. Ahmadi, R. J. Dobson, A. C. Marcano, C. Hajat, et al. 2008. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am. J. Hum. Genet.* **82:** 139–149.
7. Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, S. C. Heath, N. J. Timpson, S. S. Najjar, H. M. Stringham, et al. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40:** 161–169.
8. Aulchenko, Y. S., S. Ripatti, I. Lindqvist, D. Boomsma, I. M. Heid, P. P. Pramstaller, B. W. Penninx, A. C. Janssens, J. F. Wilson, T. Spector, et al. 2009. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.* **41:** 47–55.
9. Kathiresan, S., C. J. Willer, G. M. Peloso, S. Demissie, K. Musunuru, E. E. Schadt, L. Kaplan, D. Bennett, Y. Li, T. Tanaka, et al. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41:** 56–65.
10. Sabatti, C., S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, et al. 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41:** 35–46.
11. Keating, B. J., S. Tischfield, S. S. Murray, T. Bhangale, T. S. Price, J. T. Glessner, L. Galver, J. C. Barrett, S. F. Grant, D. N. Farlow, et al. 2008. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS One.* **3:** e3583.
12. Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81:** 559–575.
13. Anand, S. S., S. Yusuf, V. Vuksan, S. Devanesen, K. K. Teo, P. A. Montague, L. Kelemen, C. Yi, E. Lonn, H. Gerstein, et al. 2000. Differences in risk factors, atherosclerosis, and cardiovascular disease between ethnic groups in Canada: the Study of Health Assessment and Risk in Ethnic groups (SHARE). *Lancet.* **356:** 279–284.
14. Wang, J., H. Cao, M. R. Ban, B. A. Kennedy, S. Zhu, S. Anand, S. Yusuf, R. L. Pollex, and R. A. Hegele. 2007. Resequencing genomic DNA of patients with severe hypertriglyceridemia (MIM 144650). *Arterioscler. Thromb. Vasc. Biol.* **27:** 2450–2455.
15. Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. [Ser A].* **57:** 289–300.
16. Ge, D., K. Zhang, A. C. Need, O. Martin, J. Fellay, T. J. Urban, A. Telenti, and D. B. Goldstein. 2008. WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res.* **18:** 640–643.
17. Pennacchio, L. A., and E. M. Rubin. 2003. Apolipoprotein A5, a newly identified gene that affects plasma triglyceride levels in humans and mice. *Arterioscler. Thromb. Vasc. Biol.* **23:** 529–534.
18. Bennet, A. M., E. Di Angelantonio, Z. Ye, F. Wensley, A. Dahlin, A. Ahlbom, B. Keavney, R. Collins, B. Wiman, U. de Faire, et al. 2007. Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA.* **298:** 1300–1311.
19. Ehnholm, C., M. Lukka, T. Kuusi, E. Nikkila, and G. Utermann. 1986. Apolipoprotein E polymorphism in the Finnish population: gene frequencies and relation to lipoprotein concentrations. *J. Lipid Res.* **27:** 227–235.
20. Burman, D., A. Mente, R. A. Hegele, S. Islam, S. Yusuf, and S. S. Anand. 2009. Relationship of the ApoE polymorphism to plasma lipid traits among South Asians, Chinese, and Europeans living in Canada. *Atherosclerosis.* **203:** 192–200.
21. Lin, P. I., J. M. Vance, M. A. Pericak-Vance, and E. R. Martin. 2007. No gene is an island: the flip-flop phenomenon. *Am. J. Hum. Genet.* **80:** 531–538.
22. Ioannidis, J. P., E. E. Ntzani, and T. A. Trikalinos. 2004. “Racial” differences in genetic effects for complex diseases. *Nat. Genet.* **36:** 1312–1318.
23. Doerge, R. W., and G. A. Churchill. 1996. Permutation tests for multiple loci affecting a quantitative character. *Genetics.* **142:** 285–294.
24. Kathiresan, S., O. Melander, D. Anevski, C. Guiducci, N. P. Burtt, C. Roos, J. N. Hirschhorn, G. Berglund, B. Hedblad, L. Groop, et al. 2008. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N. Engl. J. Med.* **358:** 1240–1249.
25. Wang, J., M. R. Ban, G. Y. Zou, H. Cao, T. Lin, B. A. Kennedy, S. Anand, S. Yusuf, M. W. Huff, R. L. Pollex, et al. 2008. Polygenic determinants of severe hypertriglyceridemia. *Hum. Mol. Genet.* **17:** 2894–2899.