REPORT

Common Single-Nucleotide Polymorphisms Act in Concert to Affect Plasma Levels of High-Density Lipoprotein Cholesterol

Victor Spirin,* Steffen Schmidt,* Alexander Pertsemlidis, Richard S. Cooper, Jonathan C. Cohen, and Shamil R. Sunyaev

The identification of DNA sequence variants underlying human complex phenotypes remains a significant challenge for several reasons: individual variants can have small phenotypic effects or low population frequencies, and multiple allelic variants may act in concert to affect a trait. We evaluated the combined effect of allelic variants in seven genes involved in high-density lipoprotein (HDL) metabolism, using forward stepwise regression. Analysis of all known common single-nucleotide polymorphisms (SNPs) in the seven candidate genes revealed four variants that were associated with incremental changes in HDL cholesterol levels in three independent samples. Conversely, analysis of 660 polymorphisms in eight genes that do not appear to be involved in HDL metabolism did not identify any associations with plasma HDL-cholesterol levels. These data indicate that several common SNPs act in concert to influence plasma levels of HDL cholesterol.

The plasma level of high-density lipoprotein cholesterol (HDL-C) provides a useful model system for analysis of the genetic architecture of complex traits: several genes involved in HDL metabolism have been well characterized biochemically,[1] and several Mendelian disorders of HDL metabolism have been elucidated.[2] Multiple rare variants in three genes responsible for Mendelian forms of low HDL (*ABCA1* [MIM 600046], *APOA1* [MIM 107680], and *LCAT* [MIM 606967]) collectively contribute to variation in plasma HDL levels in the population,[3,4] but such variants were identified in a minority of individuals (~15%) with low HDL levels. In this study, we examined the contribution of common sequence polymorphisms to variation in HDL-C levels. Several studies have reported that common sequence variants are associated with plasma levels of HDL-C, but the individual effects of these alleles are small.[5] Earlier work on model organisms provides examples of additive contribution of multiple genetic variants of small individual effects to quantitative trait variation.[6] Accordingly, we evaluated the hypothesis that HDL-C levels reflect the cumulative contributions of multiple common DNA sequence variants, each of which has a small effect.

We focused on all known, common SNPs in seven key genes that mediate HDL metabolism (*APOA1, ABCA1, CETP* [MIM 118470], *LPL* [MIM 609708], *PLTP* [MIM 172425], *LIPC* [MIM 151670], and *LIPG* [MIM 603684]). A total of 797 SNPs in these genes were identified through dbSNP and were genotyped in 3,306 participants in the Dallas Heart Study (DHS),[7] a population-based probability sample of African American, European American, and Hispanic men and women. Before the analysis, we randomly divided the DHS population into two groups. The first set of 1,700 individuals (the training set) was used to develop the statistical model, and the second set of all the remaining individuals (the testing set) was used to test and validate the model. After elimination of poor genotypes and duplicates, the training set was reduced to 1,580 individuals, and the testing set comprised 1,726 individuals. A parallel analysis was also performed on the entire DHS population.

To take into account stratification of the DHS population according to race and sex, we used the analysis of covariance (ANCOVA), which allows for differences in HDL-C levels among population groups, while aiming at identification of alleles with the same effect within groups. HDL-C levels were log-transformed to better approximate a normal distribution. Stepwise regression was used to build a linear model incorporating sex, race, and multiple genetic factors. The initial "seed" model considered only the effects of sex and ethnicity. Each of the 797 SNPs was then added to the model, and the SNP that provided the greatest increase in likelihood was identified. If the increase in likelihood was significant (as determined by a permutation test), then the SNP was incorporated into the model. If the *P* value of any of the previously added SNPs became insignificant, that SNP was removed from the model. The process was repeated using the remaining SNPs. At

From the Genetics Division, Brigham and Women's Hospital and Harvard Medical School (V.S.; S.S.; S.R.S.), and Division of Health Sciences and Technology, Harvard–Massachusetts Institute of Technology (V.S.; S.S.; S.R.S.), Boston; Donald W. Reynolds Cardiovascular Clinical Research Center (A.P.; J.C.C.), Center for Human Nutrition (A.P.; J.C.C.), McDermott Center for Human Growth and Development (A.P.; J.C.C.), and Department of Internal Medicine, University of Texas Southwestern Medical Center (A.P.; J.C.C.), Dallas; and Department of Preventive Medicine and Epidemiology, Loyola University Stritch School of Medicine, Maywood, IL (R.S.C.)

Received March 23, 2007; accepted for publication August 3, 2007; electronically published October 19, 2007.

Address for correspondence and reprints: Dr. Shamil R. Sunyaev, Genetics Division, Brigham and Women's Hospital and Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115. E-mail: ssunyaev@rics.bwh.harvard.edu

* These two authors contributed equally to this work.

Am. J. Hum. Genet. 2007;81:1298–1303. 0002-9297/2007/8106-0017$15.00
DOI: 10.1086/522497

Table 1. ANCOVA *P* Values for the Model and Individual SNPs for All Data Sets

SNP(s)	*P* by Set					Standardized Regression Coefficient
	Training	Test	Validation	DHS	DHS-Maywood	
CETP rs183130	1.34×10^{-6}	4.04×10^{-4}	8.00×10^{-4}	5.18×10^{-9}	7.46×10^{-11}	−1.50
CETP rs5880	8.16×10^{-6}	9.55×10^{-5}	5.95×10^{-3}	1.19×10^{-8}	6.86×10^{-10}	−1.27
LPL rs326	1.27×10^{-2}	3.07×10^{-4}	1.98×10^{-2}	2.17×10^{-5}	9.14×10^{-7}	−1.15
PLTP rs3843763	3.85×10^{-2}	7.07×10^{-3}	2.86×10^{-2}	1.00×10^{-3}	1.51×10^{-4}	.86
Model[a]	2.30×10^{-14}	4.83×10^{-11}	3.28×10^{-7}	8.01×10^{-24}	3.67×10^{-29}	...

NOTE.—A one-sided test was used for the validation set, and two-sided tests were used for other data sets. Standardized regression coefficients are shown for HDL-C levels without logarithmic transformation and for the combined set only.

[a] Includes all four SNPs.

each step, the model considered only individuals with available genotype information for all selected SNPs. The procedure was terminated when the addition of new SNPs no longer significantly improved the model likelihood, with use of a *P* value threshold of 10^{-3}. Simulation studies indicated that, at this threshold, the model would pick about one-half of a false-positive SNP, on average, in the training set, whereas the chance of successful validation for the testing set at the significance level of 10^{-3} is ~1 in 1,000. The model is purely additive and does not incorporate dominance or epistatic interactions. Importantly, the stepwise regression strategy takes into account dependence of SNPs due to linkage disequilibrium (LD). The model will select a minimal combination of SNPs and will not include multiple redundant SNPs in LD. However, the model will include multiple SNPs in LD if all of them are functional. Although the power to add more SNPs to the model is reduced when there is strong LD, this effect is of limited magnitude (table C1). Computer simulations suggest that, at least for additive effects, stepwise regression has higher power to detect multiple functional SNPs in LD than does analysis of variance (ANOVA) of haplotype effects (table C2).

In the training set, the first two iterations resulted in the addition of two *CETP* SNPs (*rs183130* [$P = 9.62 \times 10^{-8}$] and *rs5880* [$P = 1.90 \times 10^{-6}$]) to the model. The third and fourth iterations added SNPs from *PLTP* (*rs6065904* [$P = 5.88 \times 10^{-4}$]) and *LPL* (*rs2197089* [$P = 8.02 \times 10^{-4}$]). All SNPs remained significant after the addition of the subsequent SNPs. To account for possible deviations from normality, we also used permutation tests to confirm that addition of each of the SNPs increased the significance of the model. The *P* value for the model containing all four SNPs was 3.06×10^{-18}, when the effect of race and sex was not considered. The same model was obtained when individuals with plasma triglyceride levels ⩾250 mg/dl, which are strongly associated with low plasma levels of HDL-C,[8] were excluded from the analysis. The model was then evaluated in the testing set. As in the training set, the *P* value for the combined model was highly significant ($P = 3.14 \times 10^{-7}$, one-sided test). This model is not driven solely by *CETP* variants and remains nominally significant for the testing set if *CETP* is excluded from the analysis.

To optimize the model, we reanalyzed all SNPs that were in strong LD ($D' > 0.9$ in all populations) with the four SNPs originally included in the model in the training set. Using SNPs from these LD blocks, we selected the model with the lowest *P* value in the testing set. The final model was highly significant in both training and testing sets, with $P = 2.30 \times 10^{-14}$ and $P = 4.83 \times 10^{-11}$ (one-sided test), respectively. The individual contributions of each of these four SNPs to the model were statistically significant in both the training and the testing data sets (table 1).

Although partitioning the DHS population into two groups allows training and testing steps to be performed with independent data sets, analysis of the entire sample would provide greater power to detect SNPs associated with HDL-C.[9] Thus, we repeated the forward stepwise regression for the complete DHS population (table 1). The same four SNPs were included in the model, and no additional SNPs were selected.

To validate the model, we assayed the four SNPs in an independent population of 849 African American men and women from Maywood (suburban Chicago). All four SNPs were significantly associated with HDL-C levels (table 1), and their combined effect was highly significant

Table 2. HDL-Lowering SNPs

SNP	Nucleotide Change	Allele			Minor-Allele Frequency		
		Ancestral	Minor	HDL Lowering	Non-Hispanic White	Non-Hispanic Black	Hispanic
CETP rs183130	C/T	T	T	C	.325	.246	.322
CETP rs5880	C/G	G	C	C	.063	.012	.117
LPL rs326	A/G	G	G	A	.289	.558	.250
PLTP rs3843763	C/T	C	T	T	.270	.165	.356

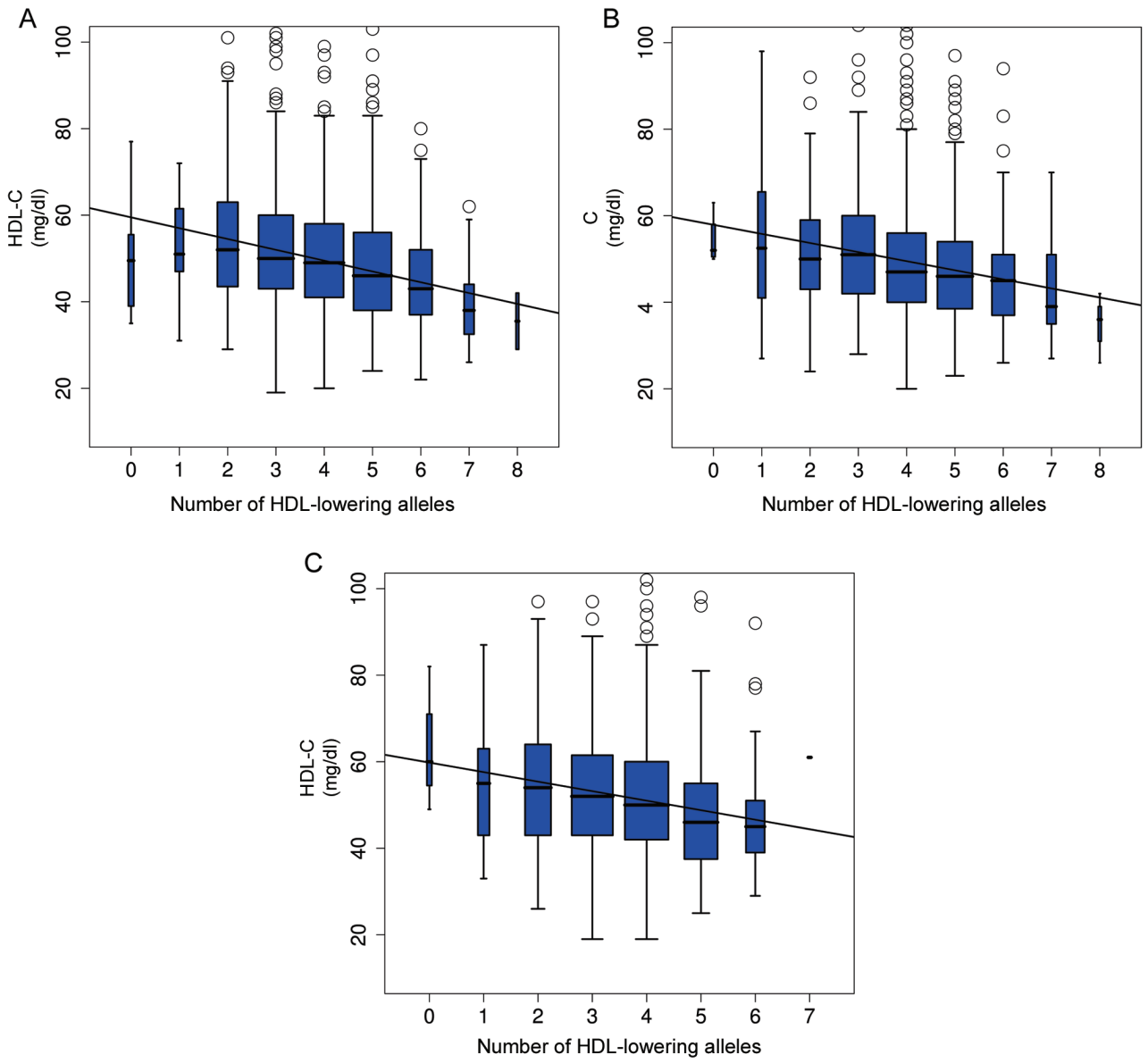

Figure 1. Dependence of HDL-C levels on the total number of HDL-lowering alleles of the four SNPs included in the model. The figure shows HDL-C levels rather than the logarithmic HDL-C levels used in the analysis. For illustration, we show the number of derived alleles without regression coefficients of the corresponding SNPs. *A,* Training set. *B,* Testing set. *C,* Validation set.

($P = 3.28 \times 10^{-7}$, one-sided test). Because the Maywood sample includes only individuals of African American descent, race was not considered a parameter of the model in this population. Similar results were obtained when the model was tested with the merged DHS-Maywood data set (table 1).

The number of HDL-lowering alleles was strongly correlated with plasma HDL-C levels. Of the DHS population, ~10% has two or fewer HDL-lowering alleles, and ~10% has six or more. The mean weighted HDL-C levels in the training set were, respectively, 55.2 mg/dl and 44.1 mg/dl for individuals in these groups (fig. 1*A*). In the testing set, the HDL-lowering alleles reduced the average plasma HDL-C level in the corresponding groups from 52.6 mg/dl to 44.8 mg/dl (fig. 1*B*), and, in the Maywood population, from 55.8 mg/dl to 47.3 mg/dl (fig. 1*C*). However, our analysis of all common SNPs in the reverse cholesterol pathway was able to explain only 2.2% of variance of HDL-C level in the DHS population. Standardized regression coefficients for the four SNPs are shown in table 1.

Three of the sequence variants in the model are very common, noncoding SNPs with minor-allele frequencies >25% (table 2). The fourth SNP (*CETP* A390P [*rs5880*]) is less common (minor-allele frequency 4.6%) and is the only nonsynonymous SNP that was significantly and reproducibly associated with HDL-C. The two *CETP* variants

Table 3. ANCOVA *P* Values for the Model That Incorporates Race, Sex, Age, and BMI

SNP	*P* by Set	
	DHS	DHS-Maywood
CETP rs183130	5.49×10^{-10}	4.73×10^{-12}
CETP rs5880	1.71×10^{-8}	7.48×10^{-10}
LPL rs326	1.30×10^{-6}	3.49×10^{-8}
HU*PLTP rs3843763*	4.37×10^{-4}	1.19×10^{-4}
Model[a]	2.99×10^{-26}	6.29×10^{-32}

[a] Includes all four SNPs.

are not in strong LD with each other ($D' = 0.462$ for whites, $D' = 0.0073$ for blacks, and $D' = 0.642$ for Hispanics). One of the two *CETP* SNPs (*rs183130*) is in strong LD with a *Taq*I SNP (*rs708272*) that was associated with HDL-C levels earlier.[10]

We further considered the effect of additional nongenetic variables: BMI and age. BMI was strongly correlated with HDL-C level, and age showed a weak though significant correlation with HDL-C in the DHS population. The effect of the four SNPs on HDL-C remains significant in a model that includes race, sex, age, and BMI (table 3).

To test for epistatic interactions among the four allelic variants, we introduced pairwise interaction terms in the linear model. These terms would account for the contribution of pairs of interacting SNPs to plasma HDL-C level beyond a simple additive effect. All pairwise interactions between SNPs in our model appeared to be insignificant. Thus, we did not identify any epistatic interactions. However, two SNPs (*CETP rs183130* and *LPL rs326*) had nominally significant interactions with BMI ($P = .033$ and $P = .031$, respectively).

We attempted to include additional SNPs in the model by relaxing the threshold for inclusion. In every case, however, the SNPs that were added in the training set negatively affected performance of the model for the testing and validation sets. Also, the additional SNPs did not have consistent individual effects on HDL-C levels in the training and testing sets. Our study certainly does not exclude the possibility that other allelic variants with small effects on HDL-C were not included in our model because of insufficient power. For example, a SNP in the hepatic lipase gene (−514C→T) that contributed to variation in HDL-C levels in other populations[11,12] was associated with modest increases in HDL-C in all three data sets in our study but was not selected by the model at our conservative threshold.

For a negative control, we analyzed 660 SNPs from eight genes that do not appear to play a significant role in HDL metabolism—*AACS*, *ABCG5* (MIM 605459), *ABCG8* (MIM 605460), *ACACA* (MIM 200350), *CART* (MIM 602606), *GHRL* (MIM 605353), *MC3R* (MIM 155540), and *SIM1* (MIM 603128). These SNPs were analyzed using the ANCOVA procedure described above. No SNP was added to the model at a significance threshold .001. Although one SNP had a *P* value close to this threshold on the training set, this SNP together with all SNPs in LD with it were not significant in either the training or the testing set individually. We then relaxed the threshold so at least three SNPs were included in the model. The resulting model and individual SNPs were not significant in the testing set.

This study suggests that the combined effect of frequent SNPs of small individual effects can be important in the inheritance of complex phenotypes. The observation that multiple SNPs have an effect on HDL-C level, including rare SNPs with large effects and frequent SNPs of smaller effects, is consistent with an evolutionary model of purifying selection. The action of purifying selection would keep allelic variants of larger effect at low frequency while allowing smaller-effect variants to reach higher frequencies.

Acknowledgments

We thank the DHS Investigators, for providing the clinical material for this study; Perlegen Sciences, for genotyping; Tommy Hyatt and Cheng Lee, for technical assistance; and Helen Hobbs, for helpful discussions. The study was funded by the International HDL Research Award by Pfizer (support to S.R.S.), the Donald W. Reynolds Cardiovascular Clinical Research Center at Dallas, the Le Ducq Foundation, and by National Institutes of Health Roadmap for Medical Research grant U54LM008748.

Appendix A

Population Study Subjects

The African American sample was recruited from Maywood, IL, a working-class community near Chicago, ~10 miles from downtown. The survey enrolled a representative random sample of the population, aged 18–74 years.

The study population included all participants in the DHS from whom fasting venous blood samples were obtained ($n = 3{,}543$). The DHS is a multiethnic probability-based sample of Dallas County (Texas) weighted to include 50% African Americans subjects, as described elsewhere.[7] The blood samples were maintained at 4°C until the plasma and serum were separated, aliquoted, and stored at −80°C. Genomic DNA was isolated from the leukocytes with use of Pure Gene (Gentra Systems).

Measurement of Plasma HDL-C Levels

Plasma HDL-C concentrations were determined using commercial enzymatic reagents, as described elsewhere.[13]

Assay of Mutations

DHS SNPs were assayed by PCR-based amplification of genomic DNA, followed by hybridization to high-density oligonucleotide arrays (Perlegen Sciences). The four SNPs analyzed in the Maywood samples were assayed by real-time PCR.

Appendix B

Stepwise Regression

This procedure involves (1) identification of a "seed" model, (2) iterative alteration of the model by addition of a SNP in accordance with the *P* value criteria, (3) backward elimination—that is, removal of previously added SNPs if their *P* values are no longer significant, and (4) termination of the search when addition no longer significantly improves the likelihood of the model.

The seed model is an ANCOVA linear model that incorporates race and sex only. The first iteration attempts to add a SNP to this model. The log-likelihood change as a result of this addition obeys a χ^2 distribution with 1 df. To choose the best candidate for the addition to the model, all SNPs are ranked by the *P* value of the corresponding log-likelihood change. The highest-ranked SNP is added to the model if its *P* value is significant. To take into account possible deviation from normality, we also test whether the addition of a SNP to the model is significant according to the permutation test. The next iteration attempts to add the next SNP to the model. SNPs are ranked again according to log-likelihood *P* values. The highest-ranked SNP is added to the model on the basis of the same criteria. If the *P* value of any of the previously added SNPs is not significant, that SNP is eliminated, although no elimination step was necessary in our study. The process stops when the addition of a SNP does not significantly improve the model.

Appendix C

Power to Include Multiple Functional SNPs in LD

The stepwise regression strategy does not incorporate multiple redundant SNPs in LD into the model. At the same time, it is expected to incorporate all functional SNPs, even if they are in LD. Strong LD, however, reduces power to incorporate multiple functional SNPs. We attempted to quantify this reduction of power via simulation. We simulated 1,000 individuals with two functional SNPs in LD that had a given D'. A quantitative phenotype was modeled as the sum of the effects of two SNPs and Gaussian noise. We simulated a range of D' values, allele frequencies, and effect sizes but assumed no epistatic interactions. We estimated power as the probability that both SNPs will be included in the model by stepwise regression and that the increase of the likelihood due to the inclusion of the second SNP will be significant at the level of $P = .05$. As seen in table C1, the power to add the second SNP to the model indeed decreases, although the decrease is limited.

Table C1. Power to Include the Second Functional SNP in the Model, as a Function of LD between Two Functional SNPs

	Power by D'					
SNP	.0	.25	.5	.75	.9	1.0
1[a]	.45	.47	.44	.36	.33	.30
2[b]	.94	.95	.89	.82	.69	.61

NOTE.—The table represents a subset of a broader range of parameters.

[a] With minor-allele frequencies .15 and .3, coefficients .5, and SD 15.

[b] With minor-allele frequencies .3 and .4, coefficients .75, and SD 15.

Table C2. Power of Regression versus ANOVA Analysis of Haplotype Effects

	Power by D'					
Analysis	.0	.25	.5	.75	.9	1.0
Regression	.76	.82	.87	.91	.94	.93
Haplotype	.51	.58	.66	.71	.78	.86

NOTE.—With SNP minor-allele frequencies .15 and .3, coefficients .5, and SD 15.

Efficiency of Haplotype Analysis

We also tested whether an alternative strategy of ANOVA of all haplotypes is more efficient for detecting multiple functional SNPs in LD. We found that, in this simple case, the regression model has higher power to reject the hypothesis that the phenotype is independent of SNPs than does the ANOVA model for haplotypes (table C2).

Also, the chance that the second SNP will be added to the regression model by stepwise regression is higher than the chance that the haplotype of the largest effect will be identified by ANOVA as the most significant (that it will achieve the nominal significance level and will have a *P* value lower than any other haplotype). Even for $D' = 1.0$ in the case of SNP allele frequencies 0.15 and 0.3 and strength of the effect of 0.5 for each SNP, stepwise regression has a 30% chance to add the second SNP to the model, whereas the haplotype of two functional alleles is detected by ANOVA as the most significant only in 22% of cases. We admit, however, that our model is purely additive and that a simple ANOVA test possibly is not the most efficient way to conduct the haplotype analysis, because heterozygotes and homozygotes for the same haplotype are treated as independent groups. We note, however, that applying the Cochran-Armitage test with true effect sizes of haplotypes (unknown in practice) provided only a small increase in power compared with the stepwise regression model (data not shown).

Web Resources

The URLs for data presented herein are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for (*ABCA1, APOA1, LCAT, CETP, LPL,*

PLTP, LIPC, LIPG, ABCG5, ABCG8, ACACA, CART, GHRL, MC3R and *SIM1*)

S.R.S.'s Web site, http://genetics.bwh.harvard.edu/genetics/labs/Sunyaev/HDL/index.html (for information about allele frequencies and LD for all SNPs included in this study)

References

1. Lewis GF, Rader DJ (2005) New insights into the regulation of HDL metabolism and reverse cholesterol transport. Circ Res 96:1221–1232
2. Hovingh GK, de Groot E, van der Steeg W, Boekholdt SM, Hutten BA, Kuivenhoven JA, Kastelein JJ (2005) Inherited disorders of HDL metabolism and atherosclerosis. Curr Opin Lipidol 16:139–145
3. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872
4. Frikke-Schmidt R, Nordestgaard BG, Jensen GB, Tybjaerg-Hansen A (2004) Genetic variation in ABC transporter A1 contributes to the HDL cholesterol in the general population. J Clin Invest 114:1343–1353
5. Brousseau ME (2004) Common variation in genes involved in HDL metabolism influences coronary heart disease risk at the population level. Rev Endocr Metab Disord 5:343–349
6. Schork NJ, Krieger JE, Trolliet MR, Franchini KG, Koike G, Krieger EM, Lander ES, Dzau VJ, Jacob HJ (1995) A biometrical genome search in rats reveals the multigenic basis of blood pressure variation. Genome Res 5:164–172
7. Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA, et al (2004) The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am J Cardiol 93:1473–1480
8. Gotto AM Jr (1990) Interrelationship of triglycerides with lipoproteins and high-density lipoproteins. Am J Cardiol 66: 20A-23A
9. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213
10. Boekholdt SM, Sacks FM, Jukema JW, Shepherd J, Freeman DJ, McMahon AD, Cambien F, Nicaud V, de Grooth GJ, Talmud PJ, et al (2005) Cholesteryl ester transfer protein TaqIB variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment: individual patient meta-analysis of 13,677 subjects. Circulation 111:278–287
11. Guerra R, Wang J, Grundy SM, Cohen JC (1997) A hepatic lipase (*LIPC*) allele associated with high plasma concentrations of high density lipoprotein cholesterol. Proc Natl Acad Sci USA 94:4532–4537
12. Andersen RV, Wittrup HH, Tybjaerg-Hansen A, Steffensen R, Schnohr P, Nordestgaard BG (2003) Hepatic lipase mutations, elevated high-density lipoprotein cholesterol, and increased risk of ischemic heart disease: the Copenhagen City Heart Study. J Am Coll Cardiol 41:1972–1982
13. Vega GL, Grundy SM (1991) Influence of lovastatin therapy on metabolism of low density lipoproteins in mixed hyperlipidaemia. J Intern Med 230:341–350