Whole-Exome Sequencing Identifies Rare and Low-Frequency Coding Variants Associated with LDL Cholesterol

Leslie A. Lange,[1,68] Youna Hu,[2,68] He Zhang,[3,68] Chenyi Xue,[4] Ellen M. Schmidt,[4] Zheng-Zheng Tang,[5] Chris Bizon,[6] Ethan M. Lange,[1,5] Joshua D. Smith,[7] Emily H. Turner,[7] Goo Jun,[2] Hyun Min Kang,[2] Gina Peloso,[8,9] Paul Auer,[10,11] Kuo-ping Li,[2] Jason Flannick,[12,13] Ji Zhang,[3] Christian Fuchsberger,[2] Kyle Gaulton,[14] Cecilia Lindgren,[14] Adam Locke,[2] Alisa Manning,[9,11,12,15] Xueling Sim,[2] Manuel A. Rivas,[14] Oddgeir L. Holmen,[16] Omri Gottesman,[17] Yingchang Lu,[18] Douglas Ruderfer,[19] Eli A. Stahl,[19] Qing Duan,[1] Yun Li,[1,5,20] Peter Durda,[21] Shuo Jiao,[10] Aaron Isaacs,[22] Albert Hofman,[23] Joshua C. Bis,[24] Adolfo Correa,[25] Michael E. Griswold,[25] Johanna Jakobsdottir,[26] Albert V. Smith,[26,27] Pamela J. Schreiner,[28] Mary F. Feitosa,[29] Qunyuan Zhang,[29] Jennifer E. Huffman,[30] Jacy Crosby,[31] Christina L. Wassel,[32] Ron Do,[8,9] Nora Franceschini,[33] Lisa W. Martin,[34] Jennifer G. Robinson,[35] Themistocles L. Assimes,[36] David R. Crosslin,[37,38] Elisabeth A. Rosenthal,[37] Michael Tsai,[28] Mark J. Rieder,[7] Deborah N. Farlow,[12] Aaron R. Folsom,[28] Thomas Lumley,[38,39] Ervin R. Fox,[25] Christopher S. Carlson,[10] Ulrike Peters,[10] Rebecca D. Jackson,[40] Cornelia M. van Duijn,[22] André G. Uitterlinden,[41] Daniel Levy,[42,43] Jerome I. Rotter,[44] Herman A. Taylor,[25,45,46] Vilmundur Gudnason, Jr.,[26,27] David S. Siscovick,[24,47,48] Myriam Fornage,[31,49] Ingrid B. Borecki,[29] Caroline Hayward,[30] Igor Rudan,[50] Y. Eugene Chen,[3] Erwin P. Bottinger,[17] Ruth J.F. Loos,[18] Pål Sætrom,[51,52] Kristian Hveem,[16] Michael Boehnke,[2] Leif Groop,[53,54] Mark McCarthy,[55] Thomas Meitinger,[56,57] Christie M. Ballantyne,[58,59] Stacey B. Gabriel,[9] Christopher J. O'Donnell,[42,60] Wendy S. Post,[61] Kari E. North,[33] Alexander P. Reiner,[47] Eric Boerwinkle,[31] Bruce M. Psaty,[24,47,48,62] David Altshuler,[8,9,13,15] Sekar Kathiresan,[8,9,60] Dan-Yu Lin,[5] Gail P. Jarvik,[7,37] L. Adrienne Cupples,[42,63] Charles Kooperberg,[10] James G. Wilson,[64] Deborah A. Nickerson,[7,69] Goncalo R. Abecasis,[2,69] Stephen S. Rich,[65,69] Russell P. Tracy,[21,66,69] Cristen J. Willer,[3,4,67,69,*] and the NHLBI Grand Opportunity Exome Sequencing Project

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [3]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA; [4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA; [5]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [6]Renaissance Computing Institute, Chapel Hill, NC 27517, USA; [7]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [8]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA; [9]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02141, USA; [10]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [11]School of Public Health, University of Wisconsin – Milwaukee, Milwaukee, WI 53201, USA; [12]Broad Institute of Harvard and MIT, Cambridge, MA 02141, USA; [13]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA; [14]Wellcome Trust Centre for Human Genetics, University of Oxford, OX1 2JD Oxford, UK; [15]Department of Genetics, Harvard Medical School, Boston, MA 02138, USA; [16]HUNT Research Center, Department of Public Health, Norwegian University of Science and Technology, 7600 Levanger, Norway; [17]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [18]The Genetics of Obesity and Related Metabolic Traits Program, The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [19]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [20]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [21]Department of Pathology, University of Vermont, Colchester, VT 05446, USA; [22]Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, 3015 DR Rotterdam, the Netherlands; [23]Department of Epidemiology, Erasmus University Medical Center, 3000 DR Rotterdam, the Netherlands; [24]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [25]Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA; [26]Icelandic Heart Association, IS-201 Kopavogur, Iceland; [27]University of Iceland, 101 Reykjavik, Iceland; [28]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN 55454, USA; [29]Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA; [30]Medical Research Center for Human Genetics, Medical Research Center Institute of Genetics and Molecular Medicine, University of Edinburgh, EH4 2XU Edinburgh, UK; [31]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [32]Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA 15261, USA; [33]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [34]Division of Cardiology, George Washington School of Medicine and Health Sciences, Washington, DC 20037, USA; [35]Departments of Epidemiology and Medicine, University of Iowa, Iowa City, IA 52242, USA; [36]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA; [37]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [38]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [39]Department of Statistics, University of Auckland, Auckland 1142, New Zealand; [40]Division of Endocrinology, Ohio State University, Columbus, OH 43210, USA; [41]Department of Internal Medicine, Erasmus University Medical Center, 3000 DR Rotterdam, the Netherlands; [42]Center for Population Studies, National Heart, Lung, and Blood Institute, Framingham, MA 01702, USA; [43]Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA 01702, USA; [44]Institute for Translational Genomics and Population Sciences, Los Angeles BioMedical Research Institute, and Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [45]Tougaloo College, Jackson, MS 39174, USA; [46]Jackson State University, Jackson, MS 39217, USA; [47]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; [48]Department of Medicine, University of Washington Medical Center, Seattle, WA 98195, USA; [49]Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [50]Centre for Population Health Sciences, Medical School, University of Edinburgh, EH8 9YL Edinburgh, UK; [51]Department of Computer and

Elevated low-density lipoprotein cholesterol (LDL-C) is a treatable, heritable risk factor for cardiovascular disease. Genome-wide association studies (GWASs) have identified 157 variants associated with lipid levels but are not well suited to assess the impact of rare and low-frequency variants. To determine whether rare or low-frequency coding variants are associated with LDL-C, we exome sequenced 2,005 individuals, including 554 individuals selected for extreme LDL-C (>98th or <2nd percentile). Follow-up analyses included sequencing of 1,302 additional individuals and genotype-based analysis of 52,221 individuals. We observed significant evidence of association between LDL-C and the burden of rare or low-frequency variants in *PNPLA5*, encoding a phospholipase-domain-containing protein, and both known and previously unidentified variants in *PCSK9*, *LDLR* and *APOB*, three known lipid-related genes. The effect sizes for the burden of rare variants for each associated gene were substantially higher than those observed for individual SNPs identified from GWASs. We replicated the *PNPLA5* signal in an independent large-scale sequencing study of 2,084 individuals. In conclusion, this large whole-exome-sequencing study for LDL-C identified a gene not known to be implicated in LDL-C and provides unique insight into the design and analysis of similar experiments.

Introduction

Elevated low-density lipoprotein cholesterol (LDL-C) is one of the cardinal risk factors for coronary artery disease, the leading cause of death in the United States.[1] LDL-C is a complex trait whose variation is influenced by the environment and genes; approximately 40%–50% of the variation is estimated as heritable.[2,3] Rare mutations have been identified in families affected by Mendelian forms of lipid-related disorders. Family members carrying these rare variants typically demonstrate extreme lipid phenotypes in childhood and, for those with high LDL-C, premature cardiovascular disease. Family studies have shown that extremely high cholesterol levels can result from mutations in *LDLR* (MIM 606945), *PCSK9* (MIM 607786), *APOB* (MIM 107730), *ABCG5* (MIM 605459), *ABCG8* (MIM 605460), and *LDLRAP1* (MIM 605747), whereas extremely low cholesterol levels can result from mutations in *PCSK9*, *MTTP* (MIM 590075), *APOB* (Rahalkar and Hegele[4]), and *ANGPTL3*[5] (MIM 603874). Targeted sequencing studies in subjects with low cholesterol levels have detected rare mutations in *LDLR*,[6] *PCSK9*,[7] and *NPC1L1*[8] (MIM 608010), but the overall contribution of rare and low-frequency variants to population variation in cholesterol levels remains poorly defined.

Genome-wide association studies (GWASs) focused primarily on common variants have identified 157 loci associated with lipid levels, including LDL-C.[9] Although GWASs have identified loci with robust evidence of association with LDL-C, only 10%–12% of the total variance in LDL-C can be attributed to these common variants,[9] despite 40%–50% estimated heritability.[2,3] We evaluated the hypothesis that rare or low-frequency variants, which are not well covered by GWASs and not easily imputed, are also associated with LDL-C.

In the current study, we performed a two-stage association study to evaluate low-frequency variation in protein-coding regions across the genome for association with LDL-C. We examined the spectrum of coding variants in associated genes in an unbiased manner. To address these goals, the NHLBI Grand Opportunity (GO) Exome Sequencing Project (ESP)[10] completed exome sequencing and analysis of 2,005 individuals, including 307 individuals with extremely high and 247 with extremely low LDL-C (>98th percentile or <2nd percentile) from population-based cohorts (stage 1). We followed up with the most promising 17 genes in 1,302 additional sequenced individuals, including 157 individuals with extremely high and 144 with extremely low LDL-C (stage 2). We also performed genotype-based follow-up of variants in 15 genes in up to 52,221 participants from population-based cohorts.

Subjects and Methods

Study-Participant Samples

We selected samples from seven population-based cohorts: Atherosclerosis Risk in Communities (ARIC),[11] Coronary Artery Risk Development in Young Adults (CARDIA),[12] the Cardiovascular Health Study (CHS),[13] the Framingham Heart Study (FHS),[14] the Jackson Heart Study (JHS),[15] the Multi-Ethnic Study of Atherosclerosis (MESA),[16] and the Women's Health Initiative (WHI).[17] Of the 2,005 individuals with exome sequence data in stage 1, 854 (43%) were African American (AA) and the remainder (n = 1,153 [57%]) were European American (EA) (Table S1, available online).

We calculated fasting LDL-C by using the Friedewald formula.[18] For individuals on lipid-lowering medication, we estimated

Information Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway; [52]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, 7489 Trondheim, Norway; [53]Department of Clinical Sciences, Diabetes, and Endocrinology, Lund University, Skåne University Hospital, 221 00 Malmö, Sweden; [54]Glostrup Research Institute, Glostrup University Hospital, 2600 Glostrup, Denmark; [55]Oxford Centre for Diabetes, Endocrinology, and Metabolism and Oxford National Institute for Health Research Biomedical Research Centre, University of Oxford, Churchill Hospital, OX1 2JD Oxford, UK; [56]Institute of Human Genetics, Helmholtz Center Munich, German Research Center for Environmental Health, 85764 Neuherberg, Germany; [57]Institute of Human Genetics, Technical University of Munich, 85764 Neuherberg, Germany; [58]Baylor College of Medicine, Houston, TX 77030, USA; [59]Houston Methodist DeBakey Heart and Vascular Center, Houston, TX 77030, USA; [60]Cardiology Division, Massachusetts General Hospital, Boston, MA 02114, USA; [61]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [62]Group Health Research Institute, Group Health Cooperative, Seattle, WA 98195, USA; [63]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02215, USA; [64]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA; [65]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA; [66]Department of Biochemistry, University of Vermont, Burlington, VT 05405, USA; [67]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[68]These authors contributed equally to this work
[69]These authors contributed equally to this work
*Correspondence: cristen@umich.edu
http://dx.doi.org/10.1016/j.ajhg.2014.01.010.

pretreatment LDL-C values by dividing treated LDL-C values by 0.75 to model a 25% reduction in LDL-C on therapy. We then regressed estimated pretreatment LDL-C levels (or actual LDL-C levels for those not on lipid-lowering therapies) on sex, age, and age squared within both cohort and ethnic (EA and AA) groups. Residuals were then combined across studies, within ethnicity strata, for selection of extreme LDL-C levels.

Participants with extreme levels of LDL-C (Table S1) were selected from four population-based cohorts: ARIC,[11] CHS,[13] FHS,[14] and JHS[15] and represented the 1st and 99th percentile tails in EA individuals (n = 156 high LDL-C and 137 low LDL-C) and the 2nd and 98th percentile tails in AA individuals (n = 151 high LDL-C and 110 low LDL-C). Additional samples not selected for LDL-C levels came from ESP studies (n = 1,451) on the basis of the following phenotypes: early-onset myocardial infarction cases and controls, ischemic stroke cases, blood pressure extremes, and body mass index (BMI); also included was a set of randomly selected samples among participants with near-complete phenotype data across a range of traits.

Stage 2 samples (n = 1,302 [66.2%] AA) were selected from the same seven cohorts as stage 1 and included individuals in the 1st and 99th percentile tails of LDL-C in EA individuals (n = 61 high LDL-C and 63 low LDL-C) and 2nd and 98th percentile tails in AA individuals (n = 96 high LDL-C and 81 low LDL-C). Stage 1 samples included an 18-fold enrichment of extreme samples in EA individuals and a 12-fold enrichment of extreme samples in AA individuals (stage 2 samples were 21-fold and 7-fold for EA and AA individuals, respectively). Additional information about these samples and the distribution of LDL-C are given in Table S1 and Figure S1. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national), and all individuals provided informed consent. Protocols were evaluated by individual institutional review boards.

Exome Sequencing

Exome sequencing was performed at the University of Washington (UW; stage 1, n = 773; stage 2, n = 858) and at the Broad Institute of Harvard and MIT (Broad; stage 1, n = 1,232; stage 2, n = 444). DNA samples were quality controlled by concentration estimation by Pico Green and, in some cases, by gel electrophoresis and real-time-PCR-based genotyping. For the majority of the samples other than those from the WHI, initial quality control (QC) was done centrally at the University of Vermont prior to shipping to the UW and the Broad. Both centers prepared DNA samples by subjecting genomic DNA to shearing and then ligating sequencing adaptors. Exome capture for the samples was performed with the Roche Nimblegen SeqCap EZ (UW) or Agilent SureSelect Human All Exon 50 Mb (Broad) according to the manufacturers' instructions. Paired-end sequencing (2 × 76 bp) was carried out with Illumina GAII and HiSeq sequencing instruments. For QC purposes prior to the release of sequence data, samples were initially converted from real-time base calls to qseq.txt files with the use of Bustard and aligned to the human reference sequence (UCSC Genome Browser, hg19) with the Burrows-Wheeler Aligner.[19] We performed duplicate removal and indel realignment by using the Genome Analysis Toolkit (GATK).[20] After the use of GATK filters, samples were required to reach at least 20× coverage over 70% of the exome target. Prior to the release of individual-level sequence reads, sequence data were required to match known fingerprint genotypes for their respective samples. Variant calls were evaluated on both bulk and per-sample properties for novel (absent from dbSNP) and known variant counts, transition/transversion (Ti/Tv) ratio, heterozygote/homozygote ratio, and insertion/deletion ratio. Both bulk and sample metrics were compared to historical values for exome sequencing projects at the two centers. DNA samples that failed laboratory QC were requeued for library preparation and sequencing.

A subset of these data is available from dbGaP under accession numbers phs000279, phs000401, phs000354, phs000362, phs000285, phs000399, phs000347, phs000546, phs000556, phs000581, phs000398, phs000402, phs000582, phs000422, phs000400, phs000327, phs000403, phs000296, phs000254, phs000518, phs000281, phs000291, phs000290, and phs000335.

Joint Read Mapping, Genotype Calling, and Variant-Level QC for Data from the UW and Broad Sequencing Centers

An average of 130 million mapped reads were generated per sample, and 95.5% of bases reached a recalibrated quality score of Q20 or greater. A total of 63.8% of the reads mapped to the exonic target region, and the mean depth of targeted regions was 127×. To generate high-quality genotype calls for analysis, we removed reads with map quality < 20 prior to variant calling with the University of Michigan's multisample SNP-calling pipeline UMAKE (H.M.K. and G.J., unpublished data). To reduce the number of sequencing variants miscalled because of sequencing and alignment artifacts, the UMAKE pipeline uses a support vector machine (SVM)[21] to exclude likely sequencing artifacts by using a battery of SNP quality metrics (Table S9). These include allelic balance (the proportional representation of each allele in likely heterozygotes), base quality distribution for sites supporting the reference and alternate alleles, and the distribution of supporting evidence between strands and sequencing cycle, among others. We used variants identified by dbSNP or 1000 Genomes as the positive training set and used variants that failed multiple filters as the negative training set. We found this method to be effective at removing sequencing artifacts while preserving good-quality data, as indicated by the Ti/Tv ratio for previously known and newly identified variant sites, the proportion of high-frequency variants overlapping with those in dbSNP, and the ratio of synonymous to nonsynonymous variants, as well as attempts at validation of a subset of sites. With the use of SVM filtering, 19,775 coding variants (5.72%) were removed.[21] The genotype concordance rate among five duplicate pairs blindly sequenced at both sequencing centers was 99.97%, and the concordance of nonhomozygous reference genotype calls was 98.97%. The genotype concordance rate for 289 AA samples genotyped at 5,051 autosomal markers with Metabochip was 98.8%, and genotype concordance was 98.7% for 526 markers with minor allele frequency (MAF) < 1%. Allelic concordance rates were 99.39% for all markers and 99.3% for variants with MAF < 1%.

Stage 2 samples were exome sequenced with the same technical and bioinformatics pipeline as those in stage 1, although variants were called and filtered as a separate batch. We only analyzed stage 2 genes that reached $p < 1 \times 10^{-5}$ in stage 1.

To reduce any differences between samples sequenced at different centers, we called variants only for the targeted region of the sample and marked them as missing if they were outside the target region. Although we initially observed a batch effect between sequencing centers, this was essentially eliminated, as determined from quantile-quantile plots, by the application of a call-rate filter. We used SVM filtering to further refine the results and saw no significant evidence of differences between sequencing centers. All extreme-LDL-C samples were sequenced at the UW.

We attempted calling insertion-deletion polymorphisms with SAMtools;[22] however, the concordance rate of the resultant indel calls was only 70.9% among our duplicate pairs, so we excluded these calls from this analysis.

QC of Individual Participant Samples

We identified related individuals by applying a maximum-likelihood algorithm[23] as implemented in RelativeFinder and by examining the mean and SD of the identity-by-state estimate for putative first- or second-degree relatives. In each group of related individuals, we prioritized (1) individuals with extreme LDL-C and (2) the individual with the highest genotype call rate and excluded all putative first- or second-degree relatives ($n = 35$). We determined the number of sequenced reads that mapped to the Y and X chromosomes, grouped the ratio of Y and X chromosome reads into two clusters, and excluded outliers on the basis of their reported sex ($n = 3$, Figure S2). We performed principal-component analysis as implemented in the PLINK software package[24] and used the first and second principal components (PC1 and PC2, respectively) as covariates for all analyses. PC1 had a squared correlation of 0.988 with estimates of European ancestry among AA samples (ancestry estimated with SEQMIX). Of 2,038 individuals with exome sequence data, 2,005 (including 554 LDL-C extremes) passed all sample-level QC and were included in the final analyses (Table S1).

Variant Annotation

We used ANNOVAR[25] with GENCODE genes (v.7; UCSC Genome Browser, hg19) to annotate variants as nonsense, splice, read-through, missense, synonymous, UTR, or noncoding and selected the most deleterious annotation for each variant (i.e., if missense in one transcript and synonymous in another, the variant was considered to be missense). We considered splice variants to be those that altered either the first two or the last two nucleotides of an intron (essential splice donor and acceptor sites). The following RefSeq accession numbers were used for annotating variants in significant genes: NM_000384.2 (*APOB*), NM_001195802.1 (*LDLR*), NM_174936.3 (*PCSK9*), and NM_138814.3 (*PNPLA5*).

In stage 1, we identified 588,226 genetic variants in the protein-coding regions of genes (exome). Of these, 3,093 (0.5%) were splice variants, 6,958 (1.1%) were nonsense variants, 345,569 (58.5%) were missense variants, and 232,182 (39.5%) were synonymous variants (Table 1). On average, each EA individual had 16.6 splice variants, 46.9 nonsense variants (stop-gained), 15.9 read-through variants (loss of stop codon), 5,865 missense variants, and 7,089 synonymous variants. By comparison, each AA individual had, on average, 24.0 splice variants, 53.5 nonsense variants, 17.7 read-through variants, 7,284 missense variants, and 9,113 synonymous variants. The number of unique variants (not seen in any other individual in our study) also differed by ethnic group. Each EA individual had, on average, 1 unique splice variant, 2.5 unique nonsense variants, and 91 unique missense variants. AA individuals had an average of 1 unique splice variant, 2.4 unique nonsense variants, and 110 unique missense variants (Table S8).

Single-Variant Association Tests

Single-variant tests for medication-adjusted LDL-C were performed by linear regression with covariates for age, sex, ethnicity (AA versus EA), primary phenotype (early-onset myocardial infarction cases and controls, ischemic stroke cases, blood-pressure extremes, BMI, and random set), and PC1 and PC2 as implemented in PLINK.[24] We excluded variants with a genotype call rate $< 50\%$ and MAF $< 1\%$.

Burden Tests

Burden tests that aggregated certain classes of variants within each gene across the genome were performed with the combined multivariate and collapsing (CMC) test[26] with multiple frequency thresholds for variant inclusion (MAF $< 5\%$, 1%, 0.5%, and 0.1%) and different classes of variants: (1) nonsynonymous and splice and (2) loss of function (LoF; nonsense, read-through, and splice)—only for MAF $< 5\%$. We used multiple frequency thresholds because the inclusion of low-frequency benign variants might have diluted a signal seen with a small number of functional rare variants, but we also wanted to test low-frequency variants because these would be present in higher numbers of individuals and might be functional. Because we expected most LoF variants to be functional, we opted to use only a single frequency threshold (MAF $< 5\%$). We used a CMC model that assigns samples as carriers or noncarriers of rare variants in a particular gene and tests for association with LDL-C values by using a linear regression model.[26]

$$\mathrm{LDL} = \beta_0 + \beta_1 \text{burden score} + \beta_2 \text{age} + \beta_3 \text{sex} + \beta_4 \mathrm{PC}_1 + \beta_5 \mathrm{PC}_2 + \beta_6 \text{ethnicity} + \beta_7 \text{ESP phenotype}.$$

The effect sizes for the burden of rare variants (shown in Table 2) were estimated in all samples with the same model. We considered estimating effect sizes from the nonextreme samples only, but this resulted in the exclusion of many individuals who carry rare variants at these genes (seven out of eight carriers at *APOB* were excluded, for example). Instead, we estimated the aggregate effect sizes from the entire sample, including extreme individuals, which might have resulted in upwardly biased estimates. Extremely large population-based samples will be required to provide unbiased effect-size estimates for very rare variants, which are currently unavailable. We tested for heterogeneity between aggregate effect sizes in AA and EA samples by performing analyses separately in AA and EA samples and then explicitly testing for heterogeneity (METAL).[27]

We also performed a test that allows for opposite directions of effect for variants within a single gene (optimized sequence kernel association test [SKAT-O]).[28] We analyzed three classes of variants: (1) LoF with MAF $< 5\%$, (2) LoF and missense variants predicted by PolyPhen-2 to be "probably damaging" with MAF $< 5\%$, and (3) missense and LoF variants with MAF $< 5\%$. For each gene, we selected the minimal p value across all burden tests.

We set the threshold for exome-wide significance to be $p < 5 \times 10^{-7}$ to account for a Bonferroni-corrected p value for 88,113 gene tests performed (eight burden tests of up to 16,141 genes, Table S7 and Figure S3). Allele frequencies used for the inclusion cutoffs were estimated from all individuals (AA and EA combined) from the ESP5500,[10] a superset that was drawn from the same cohorts as the samples examined here, had a similar ethnic make-up, and was exome sequenced at the same time as the LDL-C samples. Not all samples had LDL-C available, but we utilized in either stage 1 or stage 2 every ESP sample with LDL-C available. Burden tests were performed with covariates for ethnicity, age, sex, PC1, PC2, and ESP primary ascertainment (extreme LDL-C, BMI, extreme blood pressure, deeply phenotyped resource, early-onset myocardial infarction cases, early-onset myocardial infarction controls, and ischemic stroke). A score test was used for determining significance levels. Where genotypes were missing, we assigned the average allele frequency of the genotyped individuals. The rate of imputed genotypes at rare variants in the four known lipid genes ranged from 0.6% for *APOB* to 7.8% for *PCSK9*.

We observed no inflation of burden-test statistics (Figure S4). To further test the robustness of our results, we applied a variety of

Table 1. Study Sample: Stages 1 and 2

	AA Individuals					EA Individuals					
	Total No. of DNA Samples	Population-Based Samples Sequenced	Extremely High LDL-C Sequenced	Extremely Low LDL-C Sequenced	AA Sequenced Samples	Total No. of DNA Samples	Population-Based Samples Sequenced	Extremely High LDL-C Sequenced	Extremely Low LDL-C Sequenced	EA Sequenced Samples	All Sequenced Samples (EA and AA)
Stage 1											
n	17,628	591	151	110	854	44,987	860	156	137	1,151	2,005
Mean LDL-C (SD)	-	138.9 (34.1)	243.0 (31.9)	52.7 (12.6)	143.7 (60.9)	-	131.8 (31.2)	265.8 (48.6)	50.1 (13.8)	140.0 (64.7)	142.8 (63.7)
Range	-	57–230	195–398	22–78	22–398	-	46–224	180–479	14–80	14–479	14–479
Stage 2											
n	4,422	685	96	81	862	5,000	316	61	63	440	1,302
Mean LDL-C (SD)	-	140.7 (33.4)	246.1 (48.5)	53.7 (16.3)	144.3 (55.7)	-	131.2 (27.6)	228.4 (37.9)	54.9 (15.3)	133.8 (54.0)	140.7 (55.4)
Range	-	70–230	195–472	12–79	12–472	-	74–201	189–406	20–91	20–406	12–472
Stages 1 and 2											
n	22,050	1,276	247	191	1,714	49,987	1,176	217	200	1,593	3,307
Mean LDL-C (SD)	-	139.9 (33.7)	244.2 (39.1)	53.2 (14.3)	145.3 (58.9)	-	131.7 (30.2)	255.3 (48.8)	51.6 (14.4)	138.5 (62.2)	142.0 (60.6)
Range	-	57–230	195–472	12–79	12–472	-	46–224	180–479	14–91	14–479	12–479

Abbreviations are as follows: AA, African American; and EA, European American.

association models with different covariates and found the results to be quite similar (Table S10). The association p value for *APOB* was the most variable, and we expect that this was because the *APOB* signal was driven by only eight individuals with LoF variants in stage 1.

Sequencing-Based Follow-Up in Go-T2D

The Genetics of Type 2 Diabetes (Go-T2D) study aims to characterize the genetic architecture of type 2 diabetes and related quantitative traits through low-coverage (4×) whole-genome sequencing, deep (~70×) exome sequencing, and 2.5M SNP genotyping of 1,425 cases and 1,425 controls from four European cohorts: FUSION, DGI, WTCCC, and KORA. Genotypes were called for each technology and integrated into a single data set. In total, 27.4 million SNPs, 1.5 million indels, and 12,000 structural variants passed the QC filters. For the current analysis, we included 2,084 individuals with LDL-C measurements available.

We estimated the pretreatment LDL-C levels for individual samples (n = 300) reported to be on LDL-lowering medication (14.4%) by dividing by 0.75. We performed a burden test for *PNPLA5* (MIM 611589) variants with an observed MAF < 1% (n = 9, all SNPs) by using the CMC test[26] with covariates for age, sex, type 2 diabetes status, study site, and PC1 and PC2 derived from 2.5M SNP genotypes.

Follow-Up Using HumanExome Beadchip Array

Genotyping-based follow-up was performed with the Illumina HumanExome Beadchip array in 12 studies: ARIC[11] (n = 2,955 AA and 10,488 EA individuals), CROATIA-Korcula[29] (n = 843 EA individuals), JHS (n = 2,139 AA individuals), Family Heart Study[30] (n = 1,862 EA individuals), Age, Gene/Environment Susceptibility-Reykjavik Study [31] (n = 2,972 EA individuals), CHS[13] (n = 750 AA and 4,021 EA individuals), Rotterdam Baseline[32] (n = 1,681 EA individuals), FHS[33] (n = 6,946 EA individuals), CARDIA[12] (n = 1,886 AA and 2,131 EA individuals), WHI[17] (n = 2,142 AA and 4,005 EA individuals), Nord-Trondelag Health Study[34] (n = 5,869 EA individuals), and BioMe Clinical Care Cohort at Mount Sinai Medical Center in New York City (n = 1,091 EA and 1,974 AA individuals). Unfortunately, many individuals from the ESP also had Illumina HumanExome Beadchip genotypes by CHARGE, and these could not be removed from the analysis. To account for this, we did not meta-analyze the exome sequence and HumanExome samples together. When medication use was available, we estimated pretreatment LDL-C measures for individuals on lipid-lowering medication by dividing by 0.75. Burden tests were performed with the same variant definition as the optimal test from stage 1 (e.g., nonsynonymous and splice variants with frequency < 0.1% for *LDLR*, etc.). Frequencies were estimated within each study. Results were combined with a SE-weighted meta-analysis (METAL[27]).

Results

Exome Sequencing in Discovery Sample

We performed exome sequencing in 2,005 individuals in stage 1; 854 (43%) were AA, and the remainder (n = 1,153 [57%]) were EA. We enriched the sample of sequenced individuals by oversampling individuals with extreme LDL-C levels: $<1^{st}$ percentile and $>99^{th}$ percentile for EA and $<2^{nd}$ percentile and $>98^{th}$ percentile for AA individuals (Table 1 and Table S1). After QC, 554 individuals with extreme LDL-C levels were successfully exome sequenced: 151 AA individuals with high LDL-C, 110 AA individuals with low LDL-C, 156 EA individuals with high LDL-C, and 137 EA individuals with low LDL-C. This represents a 9-fold enrichment of individuals with extreme LDL-C levels.

Single-Variant Association Tests Detect Only *APOE* Variants

We initially examined association between LDL-C levels and each exonic variant with MAF > 1%. Only one SNP located near *APOE* and highly correlated with rs7412, which comprises the *APOE* ε2/ε3/ε4 haplotype together with rs429358, was genome-wide significant (rs1160983, $p = 7.6 \times 10^{-14}$, Table S2). Given that this association signal is well known, we focused on other statistical tests that have greater statistical power to detect association with rare variation.

Gene-Based Burden Tests in Discovery Sample Identify Three Known Genes

We hypothesized that functional variants have lower frequencies on average, and we therefore applied gene-based burden tests that evaluate the aggregate effects of variants with a low MAF. We hypothesized that different genes might have different gene-specific underlying genetic architecture, and we thus used eight burden tests. The first four tests were CMC tests[26] of missense and putative LoF variants at four MAF thresholds (MAF < 5%, 1%, 0.5%, and 0.1%), and the fifth was a CMC test of LoF variants only (MAF < 5%). The sixth test was a SKAT-O[28] for missense and LoF variants (MAF < 5%), the seventh was a SKAT-O for "probably damaging" missense and LoF variants (MAF < 5%), and the eighth test was a SKAT-O for LoF variants only (MAF < 5%). Among the 2,005 sequenced exomes, we identified variants in 16,933 genes; 16,141 of these genes had a minimum of five individuals who carried at least one LoF or missense variant with MAF < 5%. We defined putative LoF variants as nonsense, splice donor site, splice acceptor site (within two nucleotides of exon boundary), or read-through variants.

To account for the multiple testing that results from the various gene-based tests, we selected a significance threshold of $p < 5 \times 10^{-7}$ to reflect Bonferroni correction for the total number of gene-burden tests performed (88,113 tests, Table S7), although we expect that this threshold was conservative because the tests were not independent. For each gene-based burden test that reached this significance threshold, we selected as optimal the model that resulted in the most significant p value (Table 2 and Table S7).

In stage 1 alone, burden tests for three genes previously reported to be associated with LDL-C reached the threshold for exome-wide significance. All three associated genes were confirmed in the combined analysis after follow-up sequencing: *PCSK9* ($p = 3 \times 10^{-18}$), *LDLR* ($p = 3 \times 10^{-13}$), and *APOB* ($p = 2 \times 10^{-10}$).

Follow-Up Using Exome Sequencing Implicates *PNPLA5* in LDL-C

For 17 gene-based burden tests that reached $p < 5 \times 10^{-5}$ in the discovery sample, we analyzed sequence data in an

Table 2. Genes with a Burden of Rare or Low-Frequency Variants Significantly Associated with LDL-C by Exome Sequencing

Gene	Genomic Location	Optimal Burden Test	AA Individuals (Stages 1 and 2, n = 1,714)		EA Individuals (Stages 1 and 2, n = 1,593)		Stages 1 and 2 (n = 3,307)		Stages 1–3 (n = 5,391)	
			No. of variants	Burden Frequency (%)	No. of Variants	Burden Frequency (%)	Burden Effect Size in mg/dl (SE)	p Value	Burden Effect Size in mg/dl (SE)	p Value
PCSK9	Chr1 (55.5 Mb)	CMC LoF < 5%	4	2.9	-	0	−70.0 (8.7)	3×10^{-18}	-	-
PCSK9	Chr1 (55.5 Mb)	CMC nonsyn < 5%	39	28.2	19	3.9	−24.5 (2.9)	7×10^{-17}	-	-
LDLR	Chr19 (11.2 Mb)	CMC nonsyn < 0.1%	25	2.5	29	2.6	51.5 (6.9)	3×10^{-13}	-	-
APOB	Chr2 (21.2 Mb)	CMC LoF < 5%	1	0.06	11	0.7	−98.0 (21.3)	2×10^{-10}	-	-
PNPLA5	Chr22 (44.2 Mb)	CMC nonsyn < 0.1%	15	1.6	10	1.1	43.5 (9.6)	3×10^{-7}	41.7 (7.8)	1×10^{-7}

Burden frequency is the percent of individuals with at least one copy of the included rare or low-frequency alleles. Hyphens indicate that data were unavailable. Locations of variants are listed in Tables S5A–S5D. Abbreviations are as follows: AA, African American; CMC, combined multivariate and collapsing method; EA, European American; LOF, loss-of-function variants; and nonsyn, nonsynonymous and splice variants.

additional 1,302 individuals by using the same technology and analysis pipelines. The sequencing follow-up sample had a similar level of enrichment for extreme individuals and a similar proportion of AA individuals (Table 1 and Table S1) as the discovery set. In a meta-analysis of sequencing-based discovery and follow-up samples, 4 of 17 genes reached exome-wide significance, including the burden of variants in three genes identified in the discovery alone. Using a burden test for rare (MAF < 0.1%) LoF and missense variants, we identified one gene that is not known to be associated with lipid levels: *PNPLA5* ($p = 3 \times 10^{-7}$, effect size = 43.1 mg/dl, SE = 8.7, Table 2 and Table S3). Among all individuals with extremely high LDL-C, 3.1% carry a rare missense variant in *PNPLA5*; in contrast, 1.2% of nonextreme individuals and 0.5% of individuals with extremely low LDL-C carry such a variant. The impact of carrying a rare missense variant in *PNPLA5* is an increase in LDL-C by 43.1 mg/dl (SE = 8.7).

PNPLA5 encodes a member of the patatin-like phospholipase-domain-containing family. The proteins in this family share a Gly-X-Ser-X-Gly domain, and members 1–5 act as triacylglycerol lipases. *PNPLA5* is adjacent to *PNPLA3*, and the c.444C>G (p.Ile148Met) variant (rs738409) in *PNPLA3* has been implicated in nonalcoholic fatty-liver disease[35] (NAFLD [MIM 613282 and 613387]). No variants near these genes were identified by GWASs of >94,000 individuals for LDL-C[9] (Figure S5). In the current study, the *PNPLA3* variant previously associated with NAFLD was not associated with LDL-C (p = 0.15 in stage 1).

Replication of *PNPLA5* in an Independent Study

The association between LDL-C and rare variants in *PNPLA5* was evaluated in an additional large-scale sequencing study; the Go-T2D project performed low-pass whole-genome, deep exome sequencing and 2.5M SNP genotyping of 2,084 individuals of Northern European descent with estimated pretreatment LDL-C available. In this sample, ten individuals carried nine different *PNPLA5* nonsynonymous or splice variants with a frequency < 0.1%. In this additional replication sample, the burden of these rare variants was associated with LDL-C levels (p = 0.040). Compared to those without rare variants, individuals with rare variants showed a substantial mean increase in LDL-C with the same direction as observed in the samples from ESP stages 1 and 2 (Go-T2D: effect = 26.0 mg/dl, SE = 12.8; ESP combined stages 1 and 2: effect = 43.1 mg/dl, SE = 8.7).

Genetic Architecture of Associated Genes

We examined the variants that contributed to each gene-level association with LDL-C. In *PCSK9*, which encodes a convertase that mediates degradation of the LDL-C receptor,[36] we observed the strongest association with LoF variants with MAF < 5%. LoF variants in *PCSK9* were present in 13.0% of AA individuals with extremely low LDL-C yet only 0.4% with extremely high LDL-C and 1.9% with nonextreme LDL-C. The burden of LoF variants was associated with decreased LDL-C levels ($p = 3 \times 10^{-18}$,

effect = −72.2 mg/dl, SE = 8.3). A larger proportion of individuals (19.1%) carried either a missense or a LoF variant, and carrier status for this broader class of variants was also strongly associated with LDL-C levels ($p = 7 \times 10^{-17}$), although with a smaller effect size (−22.3 mg/dl, SE = 2.9). Sequencing identified nine previously unreported variants found exclusively in individuals with extremely low LDL-C (eight missense variants and one nonsense variant) and not present in databases of clinically relevant mutations[37]: c.214T>G (p.Trp72Gly), c.470A>G (p.Asn157Ser), c.721G>A (p.Val241Met), c.1180G>A (p.Gly394Ser), c.1427G>A (p.Arg476His), c.1492G>A (p.Glu498Lys), c.1496G>T (p.Arg499Leu), c.1855C>T (p.Gln619Ter), and c.1991C>T (p.Thr664Ile). Two low-frequency *PCSK9* alleles with modest effect sizes have been previously described[38]—c.137G>T (p.Arg46Leu), which has a frequency of 1.6% in EA individuals, and c.1327G>A (p.Ala443Thr), which has a frequency 8.4% in AA individuals—and burden tests that included these variants revealed significant association with LDL-C (Table 2 and Table S3).

In contrast, association with *LDLR* ($p = 3 \times 10^{-13}$) was driven by a burden of rare missense and LoF variants with MAF < 0.1% (burden frequency = 2.4%) (Table 2 and Table S3). Although *LDLR* is one of the most frequently resequenced genes and has over 1,122 variants previously identified (LOVD database[37]), we identified three previously unreported variants in the 307 individuals with extremely high LDL-C (Table S5B). Among individuals with extremely high LDL-C, 6.9% carried a rare missense or LoF variant in *LDLR*; in contrast, 2.0% of individuals with nonextreme LDL-C and 0.8% of individuals with extremely low LDL-C had such a variant. The impact of the burden of these rare variants was an increased LDL-C (effect = 40.8 mg/dl, SE = 6.5).

We identified association with *APOB* when we included a burden of LoF variants ($p = 2 \times 10^{-10}$), but not when we included missense variants (stage 1 p = 0.35, Table S7). *APOB* encodes apolipoprotein B, which is the main apolipoprotein for chylomicrons and LDL-C. We identified 12 LoF variants, and 11 were seen in individuals with extremely low LDL-C (11/12, observed effect = −98 mg/dl, Table S5C). None of these 12 variants were included in dbSNP, 1000 Genomes, or the International HapMap Project prior to our sequencing efforts.

Examining Other Genes Associated with Dyslipidemia

We also examined five candidate genes recognized as associated with Mendelian forms of high or low LDL-C.[4] Two genes, *ABCG5*[39] and *NPC1L1*,[8] exhibited suggestive evidence ($p < 4 \times 10^{-4}$) of association with LDL-C with the use of burden tests in stage 1 (Table S3). Individuals with low-frequency (MAF < 5%) missense or LoF variants in *ABCG5* had increased LDL-C ($p = 2 \times 10^{-4}$), whereas those with rare *NPC1L1* variants (MAF < 0.5%) had decreased LDL-C ($p = 3 \times 10^{-4}$). However, the evidence of association decreased after the inclusion of stage 2 results for both *ABCG5* ($p = 8 \times 10^{-3}$) and *NPC1L1* (p = 0.02). Burden tests for *LDLRAP1*, *MTTP*, *ANGPTL3*, and *ABCG8* exhibited no evidence of association (p > 0.20).

We separately examined genes near GWAS regions. We selected 192 genes that fall within the associated region at 54 LDL-C-associated loci.[40] We examined the five primary gene-based burden-test results (with the CMC test) for this subset of genes, and after performing Bonferroni correction (for 757 tests), we identified significant association ($p < 7 \times 10^{-5}$) only with *LDLR* and *PCSK9*.

Heterogeneity in Frequency of Variants by Ethnicity

To identify genes with heterogeneous effects or burden-frequency differences between EA and AA individuals, we also examined the association between coding variants and LDL-C within each ethnicity. We did not observe any genes that reached exome-wide significance in ethnic-specific analyses in either stage 1 or the combined stages 1 and 2, although *APOB* LoF mutations were primarily observed in EA individuals (11/12). We explicitly tested for heterogeneity between the burden effect size observed in AA and EA individuals and found no significant difference (all p > 0.05).

To quantify the advantages of a multiethnic study design, we estimated the power to detect significant evidence of association if our study design had included only a single ethnic group but still had the same stage 1 sample size and enrichment of extremes (n = 2,005, Table S6). Power estimates were based on the observed ethnic-specific effect sizes for the rare-variant burden in each gene, as well as the rare-variant burden frequencies in each ethnicity, for the four LDL-C-related genes (Table S3). A sample consisting entirely of AA individuals has >80% power to detect association with two genes at a threshold of 1×10^{-6}: *PCSK9* and *PNPLA5*. An entirely EA sample of the same size has >80% power to detect association with *LDLR* and *APOB* (Table S6). We also demonstrated the power to detect association in the combined samples from stages 1 and 2 for each ethnicity, and we note the substantially higher burden effect size for these genes than for variants discovered by GWASs (Figure 1).

Genotyping-Based Follow-Up in Large Sample Sizes Confirms Two Genes

To perform additional follow-up, we repeated our two-stage study of gene-based burden tests and focused only on variants that appear on the Illumina Infinium HumanExome BeadChip. Using this subset of variants (seven genes overlapped with the sequencing-based follow-up), we identified 15 genes that reached $p < 5 \times 10^{-5}$ in our stage 1 samples. We used the Illumina HumanExome Beadchip to perform genotype-based follow-up in 52,221 individuals (n = 11,912 [22.8%] AA individuals) for these 15 genes (Table S4) by using a SE-weighted meta-analysis. When we considered variants that were present on the array in these additional population-based samples, only two genes reached Bonferroni-adjusted significance (p < 0.003): *PCSK9* ($p = 2 \times 10^{-36}$) and *LDLR* ($p = 5 \times 10^{-17}$). In several of the contributing cohorts (n = 44,783), we repeated burden tests

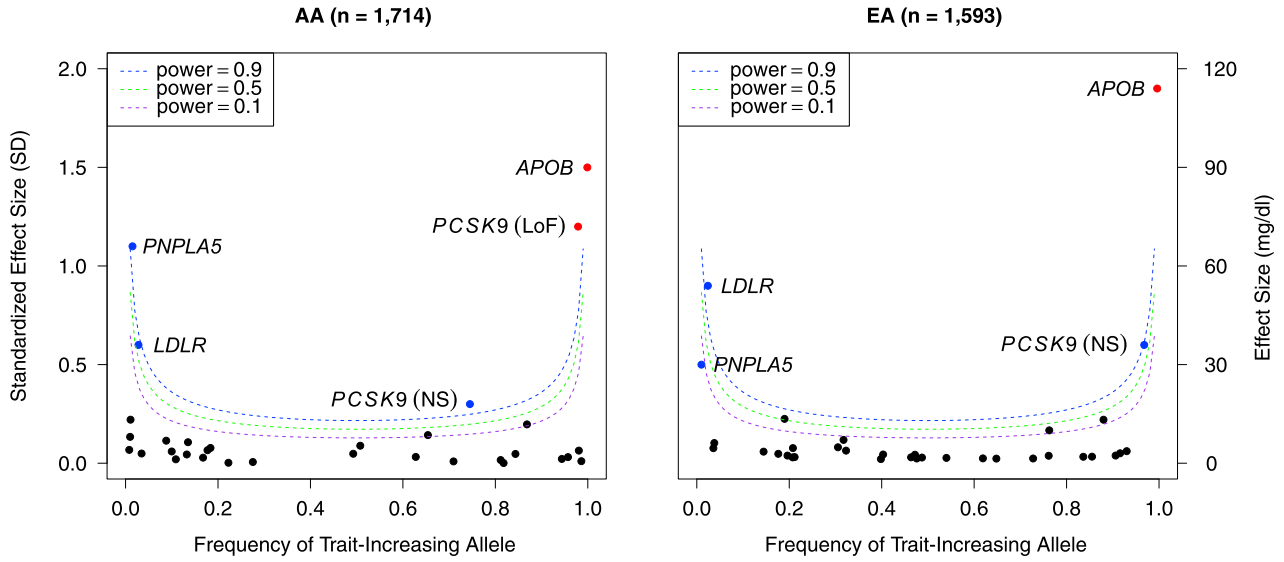

Figure 1. Effect Sizes Observed for Gene-Based Burden Tests Relative to GWAS Variants
Effect sizes are shown in SD units for genes identified by burden tests of nonsynonymous (NS) and splice variants (blue), LoF variants only (red), and GWAS variants (black) from Willer et al.[40] For genes, the burden effect sizes and burden frequencies are plotted. For GWAS variants, the observed effect sizes and MAFs are plotted. The burden frequencies for the gene-based tests (red and black) were observed in this study, whereas the GWAS-variant frequencies are plotted as reported in Willer et al.[40] The alpha level was set to 5×10^{-7} to reflect the significance threshold used for gene-based burden tests.

while excluding *LDLR* variants that were too rare for inclusion on the exome array but that had been specifically nominated by the ESP-LDL working group for the array because they had a single copy present in our data in a known gene. Excluding these rare variants attenuated the association with *LDLR* ($p = 2 \times 10^{-3}$) versus the burden of all variants on the chip ($p = 3 \times 10^{-12}$) in the same samples, suggesting that genotyping arrays are not ideal for replication of genes identified by sequencing and driven by rare variants with MAF $< 0.1\%$, such as *PNPLA5*.

Discussion

We report the results of comprehensive, high-coverage exome sequencing of a total of 2,005 individuals with LDL-C levels and present clear evidence that uncommon and rare variants contribute to variation of LDL-C levels in the general population. Our major specific findings are (1) the identification of *PNPLA5*, (2) the identification of known and previously unidentified variants in three known LDL-C-associated genes (*LDLR*,[6] *PCSK9*,[7] and *APOB*[41]), and (3) the observation that associated variants have a range of MAF and putative functional importance necessitating a variety of analytic approaches for optimizing gene discovery from sequence data.

The genetic architecture of rare variants underlying the association with LDL-C differed for each of the four LDL-C-associated genes (Figure 2). The associated variants ranged from single-copy nonsense variants in *APOB* to low-frequency missense variants in *PCSK9*. The types of variants associated with LDL-C in the four genes varied with respect to the consequence to the protein (missense versus splice or premature stop), direction of effect, allele frequency, and effect size. As a result, no single gene-based association test detected all four of the associated genes (Table S7). These findings suggest that a variety of burden tests that examine different categories of putatively functional variants defined by frequency and predicted function will have the most success at finding association with complex diseases and related phenotypes. For example, a simple burden test considering missense and LoF variants with frequency less than 1% would have identified only one gene in stage 1 (*PCSK9*, $p = 8 \times 10^{-7}$). Not surprisingly, the effect sizes observed for variants predicted to cause LoF of the protein were higher than those observed for missense mutations in both *PCSK9* and *LDLR*.

The effect sizes (ranging from 0.4 to 1.9 SDs, corresponding to ~24–116 mg/dl) observed for burden tests of the four associated genes were substantially higher than those observed for variants discovered by GWASs.[40] Furthermore, the proportion of trait variance explained by these four genes was ~5.4%, which is substantial when compared to 10%–12% explained by ~95 variants discovered by GWASs.[9] Furthermore, ~16% of the population-based samples carried a putatively functional variant in one of these four genes. Although we expect additional rare and low-frequency variants with more modest effect sizes to be discovered in larger sequencing studies, we can be reasonably confident that there are no common variants with effect sizes in the range of what we observed here. Such variants would have had a very high probability of being observed by GWASs.

We identified variants in *LDLR* and *PCSK9* in both EA and AA individuals, whereas in *APOB* and *PNPLA5*, the evidence of association with LDL-C was primarily within one ancestry group. These observations suggest that for

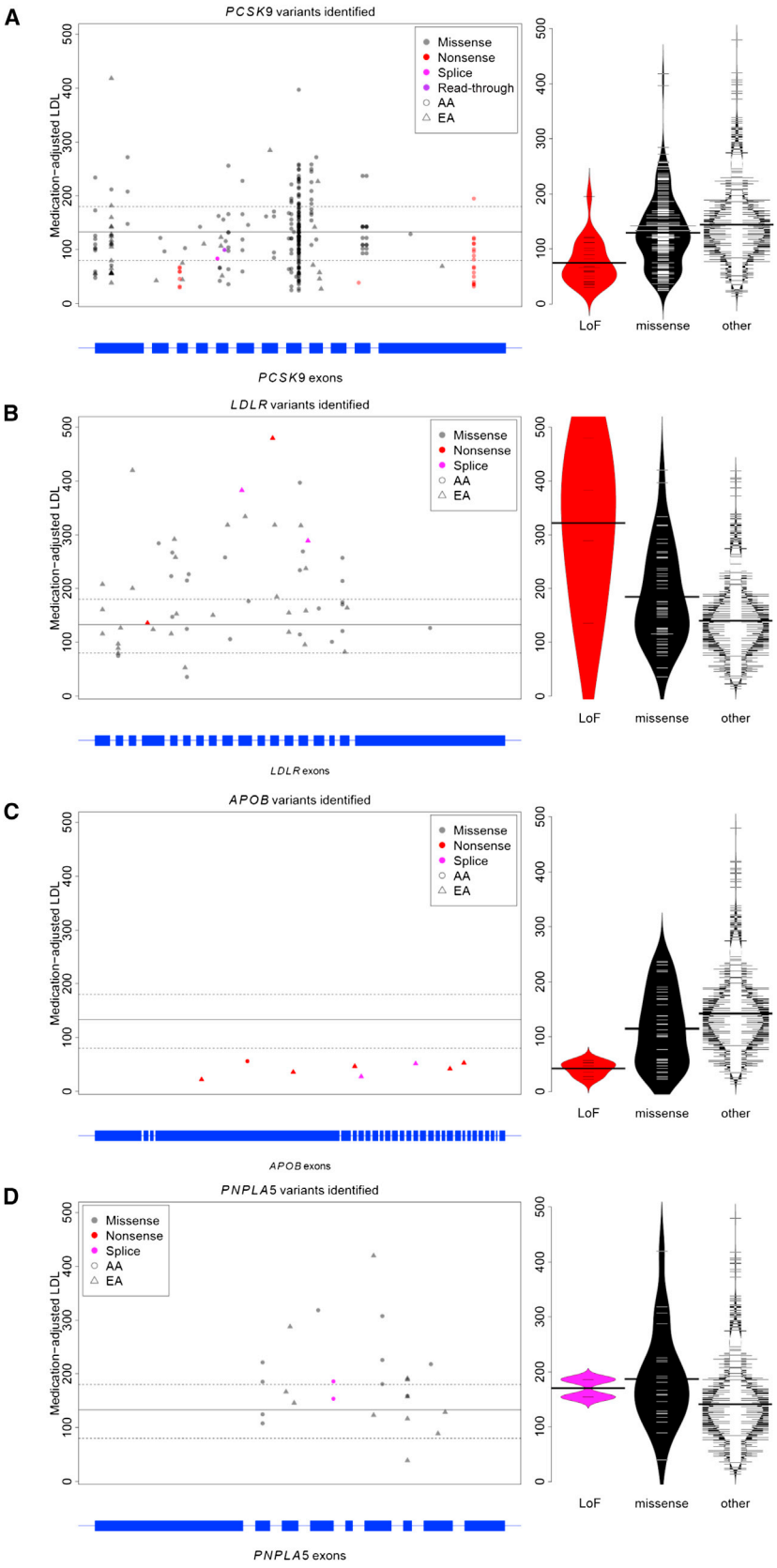

Figure 2. LDL-C Values for Individuals with Different Types of Genetic Variants in Four LDL-C-Associated Genes

A representation of LDL-C values for each individual with a rare variant in (A) *PCSK9*, (B) *LDLR*, (C) *APOB*, or (D) *PNPLA5*. The left side of the figure shows LDL-C levels per individual with a rare allele in the gene. On the right, bean plots indicate the mean (black line) and distribution (bean shape) of LDL-C values for individuals with a LoF variant, a missense variant, or no rare variant.

(A) Individuals classified on the basis of variants in *PCSK9* (MAF < 5% is considered rare).

(B) Individuals classified by carrier status of variants in *LDLR* (MAF < 0.1% is considered rare on the basis of the most significant burden test for this gene).

(C) Individuals classified by genetic status at *APOB* (MAF < 5% is considered rare).

(D) Individuals with a rare variant in *PNPLA5* (nonsynonymous or splice with MAF < 0.1%) are shown.

any complex trait, there will most likely be a subset of genes whose contribution will be most effectively identified in particular ancestry groups. We did not observe heterogeneity in effect sizes between ancestry groups but did observe differences in the proportion of individuals carrying risk alleles between ancestry groups (on average, higher levels of rare variants were observed in AA samples).

Rare coding variants in two of the four genes (*PCSK9* and *APOB*) were associated with low LDL-C, whereas rare variants in the remaining two genes (*PNPLA5* and *LDLR*) were associated with high LDL-C. This finding demonstrates the importance of selecting from both tails of the distribution to identify variants that predispose to high LDL-C and a presumed increased risk of cardiovascular disease, as well as variants associated with extremely low LDL-C and a lowered risk of cardiovascular disease. Although we could not detect significant association in the extreme samples alone, the addition of samples and analysis of LDL-C as a quantitative trait improved power.

The combined sample size of 3,302 sequenced samples was required for identifying significant association with a gene not known to be implicated in LDL-C: *PNPLA5*. We anticipate that even larger sample sizes will be required for identifying additional genes with rare variants. High-throughput genotyping will allow testing of low-frequency, and possibly rare, coding variation in large samples but will most likely fail to detect genes that have allelic heterogeneity of very rare variants—as was observed for *APOB* and *PNPLA5*. Although we did observe association with

LDLR by using the genotyping array, this might have been due to sample overlap between the genotyping follow-up and ESP samples. We specifically included all discovered *LDLR* variants on the genotyping array, irrespective of their frequency, which substantially increased the evidence of association in these samples ($p = 3 \times 10^{-12}$ with these variants and $p = 2 \times 10^{-3}$ without).

In summary, this exome sequencing study establishes that among ~17,000 genes examined, *LDLR*, *PCSK9*, *APOB*, and *PNPLA5* show the strongest evidence of a burden of rare or low-frequency coding variants influencing LDL-C. Elements of our study design, such as samples from different ethnicities, categorizing variants on the basis of different frequencies and function, and enriching our sample with both high and low phenotypic extremes, improved our ability to identify genes that contribute to LDL-C. Ultimately, functional studies will be required for uncovering the biological roles of these genes and variants in determining LDL-C levels.

Supplemental Data

Supplemental Data include affiliations and subgroup designations for the NHLBI GO ESP members, Supplemental Acknowledgments, 5 figures, and 13 tables and can be found with this article online at http://www.cell.com/AJHG.

Consortia

The members of the NHLBI GO ESP are Stacey B. Gabriel, David M. Altshuler, Gonçalo R. Abecasis, Hooman Allayee, Sharon Cresci, Mark J. Daly, Paul I.W. de Bakker, Mark A. DePristo, Ron Do, Peter Donnelly, Deborah N. Farlow, Tim Fennell, Kiran Garimella, Stanley L. Hazen, Youna Hu, Daniel M. Jordan, Goo Jun, Sekar Kathiresan, Hyun Min Kang, Adam Kiezun, Guillaume Lettre, Bingshan Li, Mingyao Li, Christopher H. Newton-Cheh, Sandosh Padmanabhan, Gina Peloso, Sara Pulit, Daniel J. Rader, David Reich, Muredach P. Reilly, Manuel A. Rivas, Steve Schwartz, Laura Scott, David S. Siscovick, John A. Spertus, Nathaniel O. Stitziel, Nina Stoletzki, Shamil R. Sunyaev, Benjamin F. Voight, Cristen J. Willer, Stephen S. Rich, Ermeg Akylbekova, Larry D. Atwood,* Christie M. Ballantyne, Maja Barbalic, R. Graham Barr, Emelia J. Benjamin, Joshua Bis, Eric Boerwinkle, Donald W. Bowden, Jennifer Brody, Matthew Budoff, Greg Burke, Sarah Buxbaum, Jeff Carr, Donna T. Chen, Ida Y. Chen, Wei-Min Chen, Pat Concannon, Jacy Crosby, L. Adrienne Cupples, Ralph D'Agostino, Anita L. DeStefano, Albert Dreisbach, Josée Dupuis, J. Peter Durda, Jaclyn Ellis, Aaron R. Folsom, Myriam Fornage, Caroline S. Fox, Ervin Fox, Vincent Funari, Santhi K. Ganesh, Julius Gardin, David Goff, Ora Gordon, Wayne Grody, Myron Gross, Xiuqing Guo, Ira M. Hall, Nancy L. Heard-Costa, Susan R. Heckbert, Nicholas Heintz, David M. Herrington, DeMarc Hickson, Jie Huang, Shih-Jen Hwang, David R. Jacobs, Nancy S. Jenny, Andrew D. Johnson, Craig W. Johnson, Steven Kawut, Richard Kronmal, Raluca Kurz, Ethan M. Lange, Leslie A. Lange, Martin G. Larson, Mark Lawson, Cora E. Lewis, Daniel Levy, Dalin Li, Honghuang Lin, Chunyu Liu, Jiankang Liu, Kiang Liu, Xiaoming Liu, Yongmei Liu, William T. Longstreth, Cay Loria, Thomas Lumley, Kathryn Lunetta, Aaron J. Mackey, Rachel Mackey, Ani Manichaikul, Taylor Maxwell, Barbara McKnight, James B. Meigs, Alanna C. Morrison, Solomon K. Musani, Josyf C. Mychaleckyj, Jennifer A. Nettleton, Kari North, Christopher J. O'Donnell, Daniel O'Leary, Frank Ong, Walter Palmas, James S. Pankow, Nathan D. Pankratz, Shom Paul, Marco Perez, Sharina D. Person, Joseph Polak, Wendy S. Post, Bruce M. Psaty, Aaron R. Quinlan, Leslie J. Raffel, Vasan S. Ramachandran, Alexander P. Reiner, Kenneth Rice, Jerome I. Rotter, Jill P. Sanders, Pamela Schreiner, Sudha Seshadri, Steve Shea, Stephen Sidney, Kevin Silverstein, Nicholas L. Smith, Nona Sotoodehnia, Asoke Srinivasan, Herman A. Taylor, Kent Taylor, Fridtjof Thomas, Russell P. Tracy, Michael Y. Tsai, Kelly A. Volcik, Chrstina L. Wassel, Karol Watson, Gina Wei, Wendy White, Kerri L. Wiggins, Jemma B. Wilk, O. Dale Williams, Gregory Wilson, James G. Wilson, Phillip Wolf, Neil A. Zakai, John Hardy, James F. Meschia, Michael Nalls, Andrew Singleton, Brad Worrall, Michael J. Bamshad, Kathleen C. Barnes, Ibrahim Abdulhamid, Frank Accurso, Ran Anbar, Terri Beaty, Abigail Bigham, Phillip Black, Eugene Bleecker, Kati Buckingham, Anne Marie Cairns, Daniel Caplan, Barbara Chatfield, Aaron Chidekel, Michael Cho, David C. Christiani, James D. Crapo, Julia Crouch, Denise Daley, Anthony Dang, Hong Dang, Alicia De Paula, Joan DeCelie-Germana, Allen DozorMitch Drumm, Maynard Dyson, Julia Emerson, Mary J. Emond, Thomas Ferkol, Robert Fink, Cassandra Foster, Deborah Froh, Li Gao, William Gershan, Ronald L. Gibson, Elizabeth Godwin, Magdalen Gondor, Hector Gutierrez, Nadia N. Hansel, Paul M. Hassoun, Peter Hiatt, John E. Hokanson, Michelle Howenstine, Laura K. Hummer, Jamshed Kanga, Yoonhee Kim, Michael R. Knowles, Michael Konstan, Thomas Lahiri, Nan Laird, Christoph Lange, Lin Lin, Xihong Lin, Tin L. Louie, David Lynch, Barry Make, Thomas R. Martin, Steve C. Mathai, Rasika A. Mathias, John McNamara, Sharon McNamara, Deborah Meyers, Susan Millard, Peter Mogayzel, Richard Moss, Tanda Murray, Dennis Nielson, Blakeslee Noyes, Wanda O'Neal, David Orenstein, Brian O'Sullivan, Rhonda Pace, Peter Pare, H. Worth Parker, Mary Ann Passero, Elizabeth Perkett, Adrienne Prestridge, Nicholas M. Rafaels, Bonnie Ramsey, Elizabeth Regan, Clement Ren, George Retsch-Bogart, Michael Rock, Antony Rosen, Margaret Rosenfeld, Ingo Ruczinski, Andrew Sanford, David Schaeffer, Cindy Sell, Daniel Sheehan, Edwin K. Silverman, Don Sin, Terry Spencer, Jackie Stonebraker, Holly K. Tabor, Laurie Varlotta, Candelaria I. Vergara, Robert Weiss, Fred Wigley, Robert A. Wise, Fred A. Wright, Mark M. Wurfel, Robert Zanni, Fei Zou, Deborah A. Nickerson, Mark J. Rieder, Phil Green, Jay Shendure, Joshua M. Akey, Carlos D. Bustamante,

*Deceased

David R. Crosslin, Evan E. Eichler, P. Keolu Fox, Wenqing Fu, Adam Gordon, Simon Gravel, Gail P. Jarvik, Jill M. Johnsen, Mengyuan Kan, Eimear E. Kenny, Jeffrey M. Kidd, Fremiet Lara-Garduno, Suzanne M. Leal, Dajiang J. Liu, Sean McGee, Timothy D. O'Connor, Bryan Paeper, Peggy D. Robertson, Joshua D. Smith, Jeffrey C. Staples, Jacob A. Tennessen, Emily H. Turner, Gao Wang, Qian Yi, Rebecca Jackson, Ulrike Peters, Christopher S. Carlson, Garnet Anderson, Hoda Anton-Culver, Themistocles L. Assimes, Paul L. Auer, Shirley Beresford, Chris Bizon, Henry Black, Robert Brunner, Robert Brzyski, Dale Burwen, Bette Caan, Cara L. Carty, Rowan Chlebowski, Steven Cummings, J. David Curb,[*] Charles B. Eaton, Leslie Ford, Nora Franceschini, Stephanie M. Fullerton, Margery Gass, Nancy Geller, Gerardo Heiss, Barbara V. Howard, Li Hsu, Carolyn M. Hutter, John Ioannidis, Shuo Jiao, Karen C. Johnson, Charles Kooperberg, Lewis Kuller, Andrea LaCroix, Kamakshi Lakshminarayan, Dorothy Lane, Norman Lasser, Erin LeBlanc, Kuo-Ping Li, Marian Limacher, Dan-Yu Lin, Benjamin A. Logsdon, Shari Ludlam, JoAnn E. Manson, Karen Margolis, Lisa Martin, Joan McGowan, Keri L. Monda, Jane Morley Kotchen, Lauren Nathan, Judith Ockene, Mary Jo O'Sullivan, Lawrence S. Phillips, Ross L. Prentice, John Robbins, Jennifer G. Robinson, Jacques E. Rossouw, Haleh Sangi-Haghpeykar, Gloria E. Sarto, Sally Shumaker, Michael S. Simon, Marcia L. Stefanick, Evan Stein, Hua Tang, Kira C. Taylor, Cynthia A. Thomson, Timothy A. Thornton, Linda Van Horn, Mara Vitolins, Jean Wactawski-Wende, Robert Wallace, Sylvia Wassertheil-Smoller, Donglin Zeng, Deborah Applebaum-Bowden, Michael Feolo, Weiniu Gan, Dina N. Paltoo, Phyliss Sholinsky, and Anne Sturcke.

Acknowledgments

The authors wish to acknowledge the support of the NHLBI and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research. Funding for the NHLBI Grand Opportunity (GO) Exome Sequencing Project was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO), and RC2 HL-102924 (Women's Health Initiative Sequencing Project). Exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). G.J. is supported by R01 HL67406 and the Northwest Institute of Genomic Medicine, funded by the Washington State Life Sciences Discovery Fund. S.K.'s effort is funded through National Institutes of Health grant R01HL107816. C.J.W. is supported by R00 HL94535 and R01 HL109946. The University of Iowa receives financial support from Amarin, Amgen, Astra-Zeneca, Daiichi-Sankyo, Esperion, F. Hoffman-La Roche, Glaxo-Smith Kline, Merck, Regeneron and Sanofi, and Takeda and Zinfandel for J.G.R's research. B.M.P serves on the data and safety monitoring board of a clinical trial for Zoll LifeCor. Additional acknowledgements are provided in the Supplemental Data.

Received: November 6, 2013
Accepted: January 14, 2014
Published: February 6, 2014

Web Resources

The URLs for data presented herein are as follows:

dbGaP, http://www.ncbi.nlm.nih.gov/gap
RelativeFinder, http://genome.sph.umich.edu/wiki/RelativeFinder
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
UMAKE, http://www.sph.umich.edu/csg/kang/umake/download/index.html

References

1. Roger, V.L., Go, A.S., Lloyd-Jones, D.M., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S., et al.; American Heart Association Statistics Committee and Stroke Statistics Subcommittee (2012). Heart disease and stroke statistics—2012 update: a report from the American Heart Association. Circulation *125*, e2–e220.
2. Pilia, G., Chen, W.M., Scuteri, A., Orrú, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. *2*, e132.
3. de Miranda Chagas, S.V., Kanaan, S., Chung Kang, H., Cagy, M., de Abreu, R.E., da Silva, L.A., Garcia, R.C., and Garcia Rosa, M.L. (2011). Environmental factors, familial aggregation and heritability of total cholesterol, low density lipoprotein-cholesterol and high density lipoprotein-cholesterol in a Brazilian population assisted by the Family Doctor Program. Public Health *125*, 329–337.
4. Rahalkar, A.R., and Hegele, R.A. (2008). Monogenic pediatric dyslipidemias: classification, genetics and clinical spectrum. Mol. Genet. Metab. *93*, 282–294.
5. Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J., et al. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N. Engl. J. Med. *363*, 2220–2227.
6. Brown, M.S., and Goldstein, J.L. (1986). A receptor-mediated pathway for cholesterol homeostasis. Science *232*, 34–47.
7. Cohen, J., Pertsemlidis, A., Kotowski, I.K., Graham, R., Garcia, C.K., and Hobbs, H.H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat. Genet. *37*, 161–165.
8. Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc. Natl. Acad. Sci. USA *103*, 1810–1815.
9. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
10. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
11. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am. J. Epidemiol. *129*, 687–702.
12. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R., Jr., Liu, K., and Savage, P.J. (1988).

CARDIA: study design, recruitment, and some characteristics of the examined subjects. J. Clin. Epidemiol. *41*, 1105–1116.
13. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., et al. (1991). The Cardiovascular Health Study: design and rationale. Ann. Epidemiol. *1*, 263–276.
14. Dawber, T.R., Meadors, G.F., and Moore, F.E., Jr. (1951). Epidemiological approaches to heart disease: the Framingham Study. Am. J. Public Health Nations Health *41*, 279–281.
15. Taylor, H.A., Jr., Wilson, J.G., Jones, D.W., Sarpong, D.F., Srinivasan, A., Garrison, R.J., Nelson, C., and Wyatt, S.B. (2005). Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn. Dis. *15* (*Suppl 6*), S6-4–S6-17.
16. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Jr., Kronmal, R., Liu, K., et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881.
17. The Women's Health Initiative Study Group (1998). Design of the Women's Health Initiative clinical trial and observational study. Control. Clin. Trials *19*, 61–109.
18. Friedewald, W.T., Levy, R.I., and Fredrickson, D.S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin. Chem. *18*, 499–502.
19. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
20. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
21. Joachims, T. (1999). Making large-scale Support Vector Machine Learning Practical. In Advances in Kernel Methods: Support Vector Learning, B. Schölkopf, C.J.C. Burges, and A.J. Smola, eds. (Cambridge: MIT Press), pp. 169–184.
22. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
23. Epstein, M.P., Duren, W.L., and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. Am. J. Hum. Genet. *67*, 1219–1231.
24. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
25. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
26. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
27. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
28. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.
29. Zemunik, T., Boban, M., Lauc, G., Janković, S., Rotim, K., Vatavuk, Z., Bencić, G., Dogas, Z., Boraska, V., Torlak, V., et al. (2009). Genome-wide association study of biochemical traits in Korcula Island, Croatia. Croat. Med. J. *50*, 23–33.
30. Higgins, M., Province, M., Heiss, G., Eckfeldt, J., Ellison, R.C., Folsom, A.R., Rao, D.C., Sprafka, J.M., and Williams, R. (1996). NHLBI Family Heart Study: objectives and design. Am. J. Epidemiol. *143*, 1219–1228.
31. Harris, T.B., Launer, L.J., Eiriksdottir, G., Kjartansson, O., Jonsson, P.V., Sigurdsson, G., Thorgeirsson, G., Aspelund, T., Garcia, M.E., Cotch, M.F., et al. (2007). Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am. J. Epidemiol. *165*, 1076–1087.
32. Hofman, A., Grobbee, D.E., de Jong, P.T., and van den Ouweland, F.A. (1991). Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur. J. Epidemiol. *7*, 403–422.
33. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N., and Stokes, J., 3rd. (1961). Factors of risk in the development of coronary heart disease—six year follow-up experience. The Framingham Study. Ann. Intern. Med. *55*, 33–50.
34. Krokstad, S., Langhammer, A., Hveem, K., Holmen, T.L., Midthjell, K., Stene, T.R., Bratberg, G., Heggland, J., and Holmen, J. (2013). Cohort Profile: The HUNT Study, Norway. Int. J. Epidemiol. *42*, 968–977.
35. Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L.A., Boerwinkle, E., Cohen, J.C., and Hobbs, H.H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. *40*, 1461–1465.
36. Horton, J.D., Cohen, J.C., and Hobbs, H.H. (2009). PCSK9: a convertase that coordinates LDL catabolism. J. Lipid Res. Suppl. *50*, S172–S177.
37. Fokkema, I.F., den Dunnen, J.T., and Taschner, P.E. (2005). LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum. Mutat. *26*, 63–68.
38. Kotowski, I.K., Pertsemlidis, A., Luke, A., Cooper, R.S., Vega, G.L., Cohen, J.C., and Hobbs, H.H. (2006). A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. Am. J. Hum. Genet. *78*, 410–422.
39. Lee, M.H., Lu, K., Hazard, S., Yu, H., Shulenin, S., Hidaka, H., Kojima, H., Allikmets, R., Sakuma, N., Pegoraro, R., et al. (2001). Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. Nat. Genet. *27*, 79–83.
40. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
41. Young, S.G., Bertics, S.J., Curtiss, L.K., Dubois, B.W., and Witztum, J.L. (1987). Genetic analysis of a kindred with familial hypobetalipoproteinemia. Evidence for two separate gene defects: one associated with an abnormal apolipoprotein B species, apolipoprotein B-37; and a second associated with low plasma concentrations of apolipoprotein B-100. J. Clin. Invest. *79*, 1842–1851.