Circulation: Genomic and Precision Medicine

ORIGINAL ARTICLE

Multiplexed Targeted Resequencing Identifies Coding and Regulatory Variation Underlying Phenotypic Extremes of High-Density Lipoprotein Cholesterol in Humans

BACKGROUND: Genome-wide association studies have uncovered common variants at many loci influencing human complex traits, such as high-density lipoprotein cholesterol (HDL-C). However, the contribution of the identified genes is difficult to ascertain from current efforts interrogating common variants with small effects. Thus, there is a pressing need for scalable, cost-effective strategies for uncovering causal variants, many of which may be rare and noncoding.

METHODS: Here, we used a molecular inversion probe target capture approach to resequence both coding and regulatory regions at 7 HDL-C–associated loci in 797 individuals with extremely high HDL-C versus 735 low-to-normal HDL-C controls. Our targets included protein-coding regions of *GALNT2*, *APOA5*, *APOC3*, *SCARB1*, *CCDC92*, *ZNF664*, *CETP*, and *LIPG* (>9 kb) and proximate noncoding regulatory features (>42 kb).

RESULTS: Exome-wide genotyping in 1114 of the 1532 participants yielded a >90% genotyping concordance rate with molecular inversion probe-identified variants in ≈90% of participants. This approach rediscovered nearly all established genome-wide association studies associations in *GALNT2*, *CETP*, and *LIPG* loci with significant and concordant associations with HDL-C from our phenotypic extremes design at 0.1% of the sample size of lipid genome-wide association studies. In addition, we identified a novel, rare, *CETP* noncoding variant enriched in the extreme high HDL-C group ($P<0.01$, score test).

CONCLUSIONS: Our targeted resequencing of individuals at the HDL-C phenotypic extremes offers a novel, efficient, and cost-effective approach for identifying rare coding and noncoding variation differences in extreme phenotypes and supports the rationale for applying this methodology to uncover rare variation—particularly noncoding variation—underlying myriad complex traits.

Sumeet A. Khetarpal, MD, PhD*
Paul L. Babb, PhD*
Wei Zhao, MS
William F. Hancock-Cerutti, MS
Christopher D. Brown, PhD
Daniel J. Rader, MD
Benjamin F. Voight, PhD

*Drs Khetarpal and Babb contributed equally to this work as co-first authors.

Key Words: genetics ■ genetic techniques ■ genetic variation ■ genome-wide association study ■ humans

http://circgenetics.ahajournals.org

Although genome-wide association studies (GWAS) have elucidated the role of common genetic variation to many human complex traits, such as blood lipids, the role of rare genetic variation remains poorly defined.[1] This is especially true for rare noncoding variants, which are not captured by whole-exome sequencing currently being applied to large numbers of participants. One strategy to capture novel variation that may include putatively causal variants is targeted resequencing of genes at candidate loci for lipid traits. Indeed, this approach has been applied to the follow-up of initial GWAS studies for low-density lipoprotein cholesterol and triglycerides.[2,3] These efforts have largely sequenced the coding regions of candidate genes, with the goal of identifying protein-altering variants that may have a profound functional impact. However, given that the majority of GWAS-implicated variants are in the noncoding genome,[4] the contribution of rare noncoding variants to these traits is underexplored.

Plasma levels of high-density lipoprotein cholesterol (HDL-C) are highly heritable. There are >70 loci significantly associated with HDL-C levels through testing of common variants (minor allele frequency [MAF], >0.05) on genome-wide genotyping arrays.[5] However, pinpointing the causal variants and genes from these associated loci is challenging. Current efforts to resolve this have included fine mapping of identified loci to determine causal variants,[6] but these methods are limited in that they focus on common single-nucleotide polymorphisms (SNPs) with generally small effect sizes. Given that common SNPs are estimated to explain only a fraction of the heritability of HDL-C levels,[7] additional variance may be explained by low frequency (MAF, 0.01–0.05) or rare variation (MAF, <0.01) not yet captured in existing genotyping arrays and imputation reference panels. Furthermore, the identification of rare, causal, noncoding variants with strong effect sizes on HDL-C may help to delineate causal and heritable mechanisms governing HDL metabolism that could directly relate to coronary heart disease risk. One limitation hampering targeted sequencing efforts for the noncoding genome is the relatively poor annotation of functional elements most likely to harbor variants of significance. A related issue is that targeted sequencing efforts are costly and scale with the size of the genomic targets, so methods have largely been developed for reliably amplifying and sequencing coding regions of genes. Thus, there is a pressing need for efficient and scalable method for capturing the noncoding genome to apply to large populations to uncover causal variation underlying complex traits, such as HDL-C.

Here, we investigated the feasibility of targeting the noncoding regions of candidate gene loci to identify rare variants underlying HDL-C phenotypic extreme using a novel, cost-effective approach. We adapt a recently reported target capture method involving molecular inversion probes (MIPs)[8,9] utilized for autism spectrum disorder candidate gene sequencing to date. Modifying this method, we show the ability to capture noncoding genomic regions in a cohort of high HDL-C cases versus low HDL-C controls. Our results validate previously reported coding and noncoding SNP associations with HDL-C, identify gene-level associations in these 7 regions with this trait, and also show the promise of expanding large-scale targeted resequencing of noncoding regions via this approach for complex traits.

MATERIALS AND METHODS

The data, analytic methods, and study materials will be made available to other researchers for purposes of reproducing the results or replicating the analyses reported here. Data may be made available on request to the corresponding authors. All recruitment of human participants of this study was approved by the Institutional Review Board of the Perelman School of Medicine at the University of Pennsylvania, and all participants provided informed consent. Full Materials and Methods are available in the Data Supplement of the article.

RESULTS

Multiplexed Targeted Sequencing Approach and Variants Identified

We sought to apply multiplexed targeted sequencing with the purpose of identifying novel and known noncoding variants underlying HDL-C. In doing so, we also sought to test the hypothesis that noncoding variation at HDL-C loci could underlie extreme HDL phenotypes similarly to coding variation traditionally identified by targeted resequencing approaches to date. Thus, we used MIPs to resequence 7 HDL candidate gene regions (Figure 1) in 1532 participants with either extremely high HDL-C versus low HDL-C controls (Table 1; Materials and Methods in the Data Supplement; Figure I in the Data Supplement). After initial sequencing read quality control, alignment, genotyping, filtering, and principal component analysis-based additional sample filtering (Materials and Methods in the Data Supplement; Tables I through IV in the Data Supplement; Figures II through X in the Data Supplement), a total of 1500 of 1532 original samples remained for further variant analysis.

To validate the variants identified from our MIP sequencing, we genotyped 1114 of the 1532 participants (681 high HDL-C individuals and 433 low HDL-C individuals) on the probe-based Illumina Exome Array. Among the variants genotyped on this array, 38 were within our target regions. We observed a high concordance rate in variant discovery between MIP sequencing and genotyping results, with 32 of 38 SNPs (84.2%) overlapping on the Exome Array called with >90% concordance across all participants, and 987 of 1114 par-

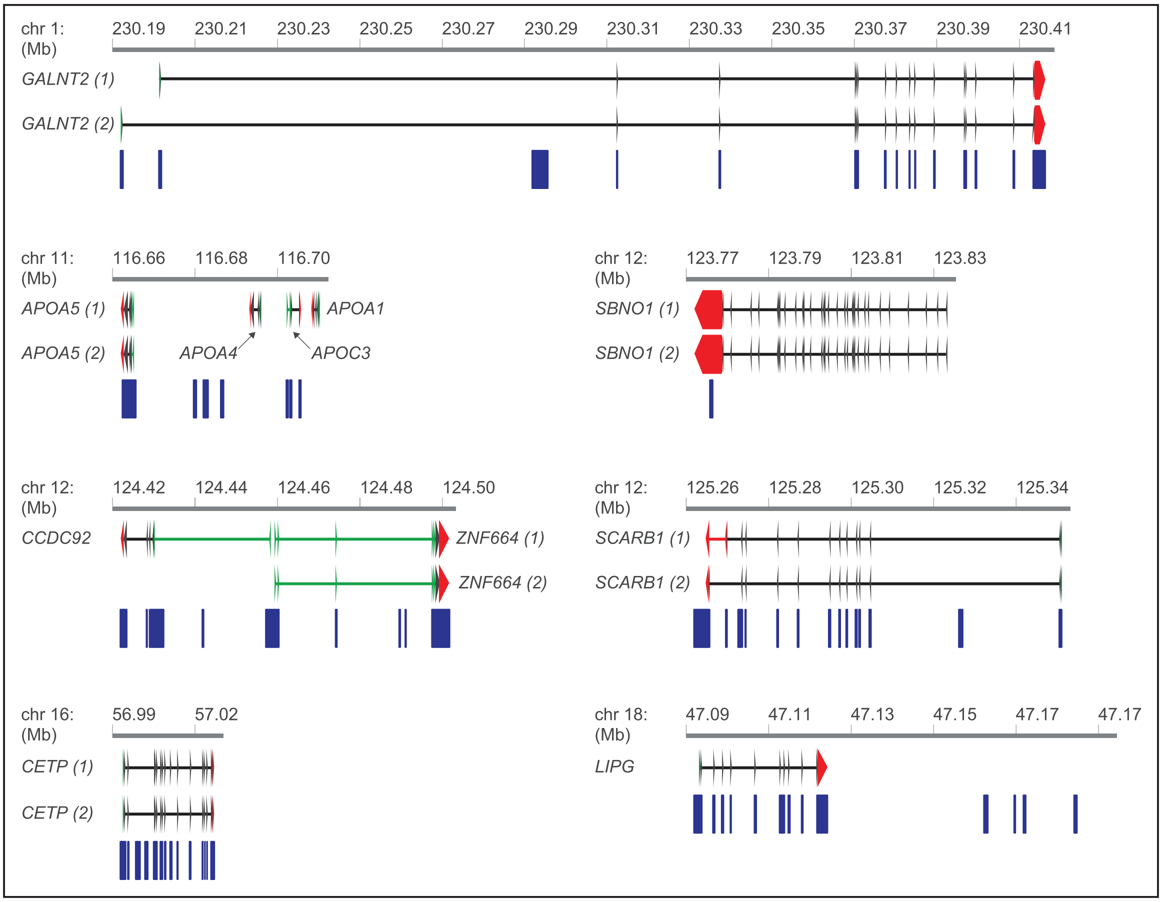

Figure 1. Candidate gene regions for molecular inversion probe (MIP) targeted sequencing.
All coordinates correspond to genomic build GRC37/hg19. Blue boxes correspond to MIP target locations.

ticipants (88.6%) demonstrating >90% concordance of all genotyped SNPs (Figure 2).

The final MIP sequencing variant call set contained 1956 SNPs and 689 distinct insertion/deletion events (indels; 78 insertions and 611 deletions) for a total of 2645 unique variants in 1500 samples. Of these, 556 correspond with previously reported variants in dbSNP (v141), suggesting that the remaining 2089 were novel discoveries without any previous annotation (Tables 2 and 3). We found that our MIP sequencing capture approach did not preferentially identify variants of a given annotation across our selected genomic targets (Figure XI in the Data Supplement). After these efforts, the MIP sequencing

Table 1. Characteristics of Participants for Molecular Inversion Probe Targeted Sequencing

	High HDL Cohort			Low HDL Cohort			High vs Low HDL Cohort (*t* Test)
	All (n=789)	Men (n=228)	Women (n=561)	All (n=743)	Men (n=454)	Women (n=289)	
Age, y (SD)	58 (13)	59 (15)	58 (12)	55 (13)	56 (12)	53 (15)	*P*<0.0001
White, %	86.2	89.9	84.7	61.5	65.0	56.1	NA
Ashkenazi, %	7.9	8.3	7.7	2.6	3.5	1.0	NA
Black, %	4.6	2.2	5.5	27.5	23.3	33.9	NA
Total cholesterol, mg/dL	240 (42)	227 (40)	245 (42)	177 (72)	172 (74)	185 (68)	*P*<0.0001
HDL-C, mg/dL	107 (21)	94 (19)	112 (19)	32 (11)	31 (12)	34 (8)	*P*<0.0001
LDL-C, mg/dL	127 (60)	127 (40)	127 (71)	100 (59)	96 (58)	105 (61)	NS
TG, mg/dL	77 (34)	78 (37)	77 (32)	266 (566)	270 (537)	259 (610)	*P*<0.0001

Participants were recruited from the Penn High HDL Study as described previously. All lipid measurements were performed on plasma collected after participants fasted overnight. Comparisons of absolute measurements were performed using a Student unpaired *t* test of all high HDL cohort participants vs all low HDL cohort participants. All absolute data are reported as mean±SD. HDL indicates high-density lipoprotein; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; NA, not available; NS, nonsignificant; and TG, triglyceride.

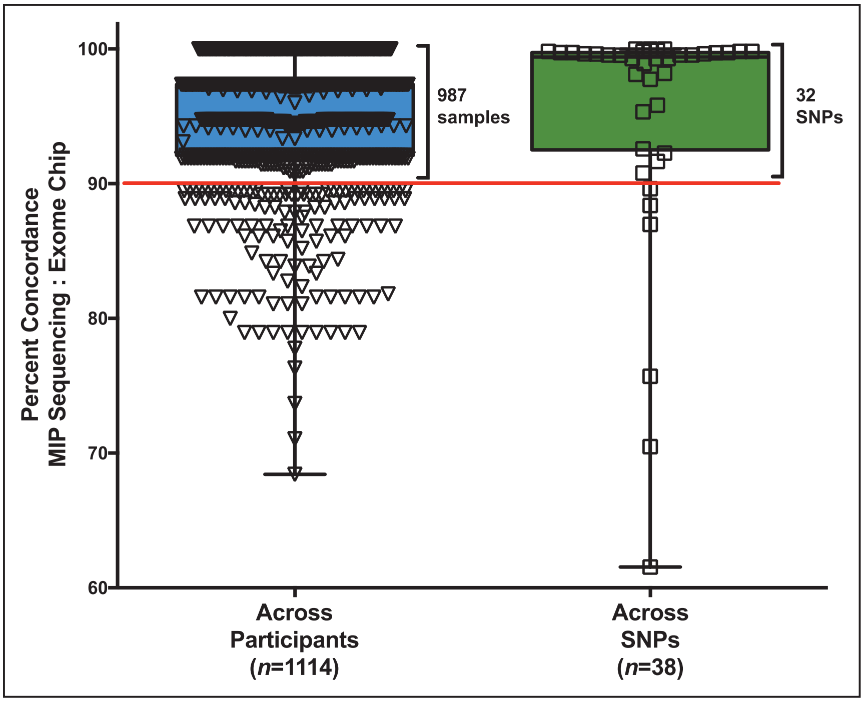

Figure 2. Concordance of variants identified from molecular inversion probe (MIP) sequencing with exome chip genotyping. Single-nucleotide variants identified in the targeted regions by MIP sequencing were compared with the discovery of those variants by genotyping on the exome chip in a subset of 1114 participants who were included in both variant discovery efforts. A total of 38 single-nucleotide polymorphisms (SNPs) that were included in the exome chip were found to overlap the targeted regions by MIPs. Box plot on the (**left**) shows the percentage of total SNPs that were found by both discovery methods for each individual (n=1114 participants). Box plot on the (**right**) shows the percentage of individuals for which a given SNP was found to be concordant across the 2 discovery methods (n=38 SNPs). Red line indicates those samples (**left**) and SNPs (**right**) for which concordance between MIP sequencing and the exome chip genotyping was >90%. Thirty-two of 38 SNPs (84.2%) overlapping on the exome array were called with >90% concordance across all participants. Nine hundred eighty-seven of 1114 participants (88.6%) demonstrated >90% concordance of all genotyped SNPs.

variants were then tested for association with HDL-C using a framework sensitive to MAF and protein coding status of the different variants (Figure XII in the Data Supplement).

Association of Single Variants From Targeted Sequencing With Extremely High HDL-C

We tested the association of 336 common and low-frequency (MAF, ≥0.01) SNPs and indels identified with high versus low HDL levels and observed 34 alleles at significantly greater frequencies among the high HDL-C participants ($P<1.49\times10^{-4}$, score test; Table 4). Of these, 17 were previously reported by the Global Lipids Genetics Consortium GWAS.[7]

Replication of HDL-C Associations From GWAS Through MIP Sequencing

In addition to rare, noncoding variants identified from MIP sequencing, we also recovered common variants previously associated with HDL-C through the Global Lipids Genetics Consortium plus MetaboChip GWAS.[7] In the Global Lipids Genetics Consortium study, 49 variants that exceeded genome-wide significance ($P<5\times10^{-8}$) in their associations with HDL-C are located in regions that overlap with MIP sequencing targets.

Table 2. SNPs and INDELs Identified by Molecular Inversion Probe Sequencing of 1500 Extreme HDL-C Participants

			SNPs		INDELs		
Chromosome	Genic Region	Target Size, bp	Common+Low Frequency	Rare	Common+Low Frequency	Rare	Total Variants
1	*GALNT2*	9636	63	271	14	110	458
11	*APOA5-APOC3*	6550	44	274	6	98	422
12	*SBNO1*	530	7	33	…	6	46
12	*CCDC92-ZNF664*	11 955	49	355	4	130	538
12	*SCARB1*	7815	38	247	7	102	394
16	*CETP*	5739	37	202	6	87	332
18	*LIPG*	9095	56	280	5	114	455
	Total	51 320	294	1662	42	647	2645

SNPs and INDELs were assessed for each gene region (GRC37/hg19) and were processed using sample-level and variant-level quality control filters (Materials and Methods). Minor alleles of identified variants were compared for frequency in the high vs low HDL cohort by the score test statistic. Noncoding variants included any variants that were not present in protein-coding regions of the gene regions, including splice-site, intronic, 5′ UTR, 3′ UTR, and intergenic variants. ExAC indicates Exome Aggregation Consortium; HDL, high-density lipoprotein; HDL-C, high-density lipoprotein cholesterol; INDEL, insertion/deletion event; and SNP, single-nucleotide polymorphism.

Table 3. Variants Identified by Molecular Inversion Probe Sequencing of 1500 Extreme HDL-C Participants

Chrom	Genic Region	Known* Coding	Known* Noncoding	Novel† Coding	Novel† Noncoding	Significant HDL-C Associations (Score Test)
Common±low-frequency variants (MAF≥0.01)						
1	*GALNT2*	4	56	…	17	…
11	*APOA5-APOC3*	4	32	3	11	…
12	*SBNO1*	…	2	…	5	…
12	*CCDC92-ZNF664*	4	36	5	8	9‡
12	*SCARB1*	6	26	1	12	…
16	*CETP*	5	32	2	4	20‡
18	*LIPG*	4	45	…	12	5‡
	Total	27	229	11	69	34‡
Rare variants (MAF≤0.01)						
1	*GALNT2*	10	44	37	290	…
11	*APOA5-APOC3*	8	26	52	286	…
12	*SBNO1*	…	3	…	36	…
12	*CCDC92-ZNF664*	10	45	66	364	…
12	*SCARB1*	13	30	45	261	…
16	*CETP*	18	26	42	203	2§
18	*LIPG*	8	59	44	283	…
	Total	67	233	286	1723	2§

SNPs and INDELs were assessed for each gene region (GRC37/hg19) and were processed using sample-level and variant-level quality control filters (Materials and Methods). Minor alleles of identified variants were compared for frequency in the high vs low HDL cohort by the score test statistic. Noncoding variants included any variants that were not present in protein-coding regions of the gene regions, including splice-site, intronic, 5′ UTR, 3′ UTR, and intergenic variants. ExAC indicates Exome Aggregation Consortium; HDL, high-density lipoprotein; HDL-C, high-density lipoprotein cholesterol; INDEL, insertion/deletion event; MAF, minor allele frequency; and SNP, single-nucleotide polymorphism.

*Known variants were those for which an reference SNP cluster ID existed in Single Nucleotide Polymorphism database (v141) or were able to be ascertained in publically available variant databases, including 1000 Genomes, the National Heart, Lung and Blood Institute Exome Variant Server, and the ExAC database.

†Novel variants were all other variants not listed as known above.

‡No. of single-variant associations with HDL-C using the score test,[9] at or below the experimental significance threshold of $P<1.49\times10^{-4}$ (testing only 336 common and low-frequency variants).

§No. of single-variant associations with HDL-C using the score test,[9] at or below the experimental significance threshold of $P<1.89\times10^{-5}$ (testing all 2645 variants).

We observed all of the 49 variants in the MIP sequencing variant call set and likewise observed all of them at common or low frequencies (MAF, >0.01) in the 1500 samples. A total of 17 of the 49 exceeded an experimental statistical threshold (score test $P<1.49\times10^{-4}$), with an additional 10 that were nominally significant (score test $P<0.01$; Table V in the Data Supplement; Figures XIII through XVII in the Data Supplement). All of the experiment-wide significant and nominally significant associations we identified were directionally consistent with prior reports of SNPs as those loci with HDL-C levels and with comparable MAFs to those reported for each variant from 1000 Genomes Project (phase 3 v5a, European sample set).[10,11] An additional 17 SNPs and 2 indels not previously identified in the Global Lipids Genetics Consortium GWAS were significantly associated with HDL-C (Table VI in the Data Supplement).

Rare, Novel, Noncoding Variants With Nominally Significant Associations With HDL-C

Because of the small sample size of our study, we expected modest power to demonstrate association beyond a reasonable doubt. Thus, we examined variants that exhibited nominally significant associations ($P<0.01$, score test) with elevated HDL-C and identified 68 such SNPs and indels (Table VII in the Data Supplement). These included 54 noncoding variants (ie, located outside of protein-coding sequence) and 14 coding variants. Among the coding and noncoding variants identified with nominally significant associations were 11 rare variants (MAF, ≤0.01), 6 low-frequency variants (0.01<MAF<0.05), and 8 variants not previously described in Single Nucleotide Polymorphism database. Of the noncoding variants identified, 12 were found to have Combined Annotation

Table 4. Significant Single-Variant Associations With High HDL-C

Region	Chrom	Position	Variant	dbSNP rsID*	Type	Variant Call Rate	MAF	Score Statistic	Score *P* Value†
CCDC92-ZNF664	12	124421453	T/C	rs9863	Noncoding	0.9993	0.4109	4.0051	6.20E-05
	12	124427306	T/A	rs11057401†	Coding	1	0.3407	4.7169	2.40E-06
	12	124428162	T/A	rs4930725	Noncoding	0.9987	0.3632	4.2263	2.38E-05
	12	124428331	T/C	rs4930726†	Noncoding	0.9873	0.3754	4.4037	1.06E-05
	12	124429279	G/A	rs3186071	Noncoding	0.9973	0.3259	4.1498	3.33E-05
	12	124430612	G/A	rs4765305	Noncoding	0.9660	0.4824	4.1397	3.48E-05
	12	124430812	G/A	rs4765335	Noncoding	0.9953	0.3985	4.0566	4.98E-05
	12	124431049	G/A	rs11835839	Noncoding	0.9740	0.4182	4.8946	9.85E-07
	12	124499839	C/T	rs3768	Noncoding	0.9993	0.2255	3.9722	7.12E-05
CETP	16	56995236	C/A	rs1800775†	Noncoding	0.8893	0.3212	7.3626	1.80E-13
	16	56995814	G/A	rs34498052	Noncoding	0.9580	0.0010	5.3163	1.06E-07
	16	56996158	T/C	rs3816117	Noncoding	0.9920	0.4755	9.7746	1.45E-22
	16	56996211	G/A	rs711752†	Noncoding	0.9880	0.4295	7.8694	3.56E-15
	16	56996288	G/A	rs708272	Noncoding	0.9887	0.4413	7.6112	2.72E-14
	16	56998918	A/G	rs12720926	Noncoding	0.9360	0.3650	7.8158	5.46E-15
	16	56999258	A/C	rs7203984†	Noncoding	0.9747	0.2309	–8.3195	8.83E-17
	16	56999328	C/T	rs11508026†	Noncoding	0.9873	0.3964	9.4414	3.68E-21
	16	57001254	T/TCACA	rs12720908	Noncoding	0.9780	0.1953	–7.8050	5.95E-15
	16	57001274	AC/A	rs200751500	Noncoding	0.8853	0.1325	5.9512	2.66E-09
	16	57001438	G/A	rs12444012	Noncoding	0.2433	0.4932	4.5664	4.96E-06
	16	57004889	G/A	rs7205804†	Noncoding	0.9753	0.3568	6.9836	2.88E-12
	16	57005301	C/T	rs1532625†	Noncoding	0.9840	0.3581	8.2715	1.32E-16
	16	57005883	G/A	rs374409989	Noncoding	0.8733	0.0023	5.3838	7.29E-08
	16	57007353	C/T	rs5883†	Coding	0.9847	0.0735	5.4895	4.03E-08
	16	57007446	T/G	rs11076176	Noncoding	0.9940	0.1851	–6.8759	6.16E-12
	16	57015091	G/C	rs5880 (Ala390Pro)†	Coding	1	0.0350	–4.9197	8.67E-07
	16	57016092	G/A	rs5882†	Coding	0.9973	0.3737	–4.9708	6.67E-07
	16	57017319	G/A	rs1800777 (Arg468Gln)	Coding	0.9973	0.0247	–5.4589	4.79E-08
	16	57017474	G/A	rs289741†	Noncoding	0.9347	0.3574	–5.2733	1.34E-07
	16	57017662	G/A	rs1801706†	Noncoding	0.9913	0.1725	4.8147	1.47E-06
	16	57017796	G/A	rs289743	Noncoding	0.9440	0.2256	–3.8050	1.42E-04
LIPG	18	47096016	G/A	rs1320700	Noncoding	0.9693	0.2775	4.1477	3.36E-05
	18	47158186	T/C	rs10438978†	Noncoding	1	0.1920	4.7214	2.34E-06
	18	47158234	C/T	rs9304381†	Noncoding	1	0.1767	4.6760	2.92E-06
	18	47167214	T/C	rs4939883†	Noncoding	1	0.2073	4.6702	3.01E-06
	18	47179516	G/A	rs1943973†	Noncoding	0.9947	0.1079	3.8348	1.26E-04

Variants (SNPs and INDELs) across targets were compared for frequency of the minor allele in high vs low HDL participants by score test statistic. Score test *P* values, where $P<0.01$ were considered nominally statistically significant, whereas *P* values $<1.27\times10^{-5}$ were considered to exceed the experimental significance threshold accounting for all 2645 variants called in this study. MAF refers to MAF within the sequencing cohort. Call rate refers to the fraction of 1500 samples for which a particular variant position was sequenced and passed sample-level and variant-level quality filtering. dbSNP indicates Single Nucleotide Polymorphism database; GLGC, Global Lipids Genetics Consortium; GWAS, genome-wide association study; HDL, high-density lipoprotein; HDL-C, high-density lipoprotein cholesterol; INDEL, insertion/deletion event; MAF, minor allele frequency; MIP, molecular inversion probe; rsID, reference SNP cluster ID; and SNP, single-nucleotide polymorphism.

*Variants identified from MIP sequencing with significant associations with HDL-C ($P<1.49\times10–4$ by score test[9]) that were previously found to have significant associations with HDL-C in the GLGC GWAS[7] are designated with

†Single-variant associations with HDL-C using the score test.[9] Only variants with *P* values below experimental significance threshold of $P<1.49\times10^{-4}$ are shown.

Dependent Depletion scores of ≥10, suggestive of deleteriousness to gene expression or function (Table VII in the Data Supplement).[12] We evaluated the putative impact of the noncoding variants we identified across our regions by exploring overlap between these SNPs and transcription factor binding sites and microRNA seed sites, which identified multiple common noncoding variants across our loci that overlapped such regulatory features (Table VII in the Data Supplement). Among the noncoding SNPs with potential functional impact on gene expression is a proximal variant 21 bp upstream of the transcription start site of *CETP*, rs34498052 (chr16:56995814 G>A), that was previously identified in a resequencing study of 68 genes in French Canadian myocardial infarction cases and controls. Although this variant overlaps multiple epigenetic marks from Encyclopedia of DNA Elements, including CpG methylation marks in HepG2 hepatocytes and Human Microvascular Endothelial Cells endothelial cells, it was extremely rare (MAF, 0.001; allele count [AC], 3), which made statistical interpretation challenging because the score test is not intended or calibrated for that end of the frequency spectrum given our sample size. More conservatively, for variants identified with >5 allelic copies among the 1500 participants, we identified a single rare, novel, noncoding SNP in a splice region of the *CETP* gene (chr16:57005300 G>A) that was nominally associated with high HDL-C (*P*=0.009, score test; AC, 8).

We also investigated the association of these SNPs with expression of genes as expression quantitative trait loci from the Genotype-Tissue Expression Project (Tables VII and VIII in the Data Supplement).[13] Analysis of expression quantitative trait loci across human tissues identified 21 of the 54 noncoding SNPs with at least 1 significant expression quantitative trait loci in a human tissue. Among these are a set of noncoding SNPs at the *CCDC92* locus associated with reduced *CCDC92* expression and that of other genes in subcutaneous adipose tissues, consistent with the recent identification of a sentinel SNP at this locus in linkage disequilibrium with our identified SNPs that was associated with coronary artery disease and also with decreased *CCDC92* expression in the same tissue.[14] As another example, we show that another set of SNPs downstream of the *LIPG* gene are associated with *LIPG* gene expression in skeletal muscle and skin tissues. These SNPs are in linkage disequilibrium with other GWAS-implicated SNPs downstream of *LIPG* that we previously showed to reduce endothelial lipase protein levels.[15] Thus, our MIP sequencing experiment identified multiple regulatory variants underlying high HDL-C that also correlated with *cis*-regulatory effects on gene expression across human tissues.

Rare Variant Burden Associations With Extremely High HDL-C

Lastly, we tested the hypothesis that the genomic regions we targeted harbor rare variants that collectively contribute to the relationship of these genes with HDL-C levels. We performed aggregate rare variant burden using a framework that categorized rare variants (MAF, <0.01) on the basis of their coding status, deleteriousness, and genic region (Tables 5 and 6; Figure XII in the

Table 5. Rare Variant Burden Test Associations of Molecular Inversion Probe Sequencing Variants With High High-Density Lipoprotein Cholesterol (Coding)

Chr	Genic Region	Disruptive*			Disruptive and Missense†			Loss of Function‡		
		Variants	β (SE)	*P* Value	Variants	β (SE)	*P* Value	Variants	β (SE)	*P* Value
1	*GALNT2*	15	0.03 (0.33)	0.92	34	0.10 (0.22)	0.64	15	0.03 (0.33)	0.92
11	*APOA5*	15	−0.36 (0.30)	0.23	33	−0.37 (0.23)	0.10	14	−0.35 (0.31)	0.25
11	*APOC3*	5	−0.38 (0.71)	0.59	7	−0.60 (0.65)	0.36	5	−0.38 (0.71)	0.59
12	*CCDC92*	20	0.25 (0.27)	0.36	33	0.00 (0.22)	0.98	20	0.25 (0.27)	0.36
12	*ZNF664*	2	2.11 (1.38)	0.13	14	0.37 (0.32)	0.25	2	2.11 (1.38)	0.13
12	*SCARB1*	24	0.28 (0.27)	0.29	37	0.27 (0.21)	0.20	23	0.30 (0.27)	0.27
16	*CETP*	23	−0.10 (0.27)	0.71	33	−0.06 (0.22)	0.77	12	−0.52 (0.45)	0.25
18	*LIPG*	14	−0.10 (0.29)	0.73	32	0.01 (0.21)	0.95	13	−0.12 (0.29)	0.67
	All coding§	118	0.04 (0.14)	0.78	223	−0.09 (0.12)	0.44	104	0.01 (0.14)	0.95

VEP indicates Ensembl Variant Effect Predictor.

*Disruptive coding variants included nonsense (stop-gained), frameshift, splice-donor, splice-acceptor, stop lost, start lost, inframe insertion, and inframe deletion variants as annotated from VEP tool.

†Collection of disruptive plus missense coding variants. Missense variants were defined as nonsynonymous amino acid-altering variants using the dbNSFP database (v2.9.1). Variants were included in this grouping if they were identified as deleterious or damaging by 1 of the 5 in silico prediction tools: Sorting Intolerant From Tolerant (deleterious), PolyPhen2 HDIV (possibly damaging or probably damaging), PolyPhen2 HVAR (possibly damaging or probably damaging), MutationTaster, and likelihood ratio test (disruptive).

‡Loss-of-function variants were defined based on loss of function prediction flags (high confidence, low confidence) generated by the VEP plugin Loss-Of-Function Transcript Effect Estimator. This set was then filtered to remove variants that were situated in unlikely open-reading frames, single-exon genes, or had weak phylogenetic conservation scores.

§No coding regions of the *SBNO1* region were sequenced in this study.

Table 6. Rare Variant Burden Test Associations of Molecular Inversion Probe Sequencing Variants With High High-Density Lipoprotein Cholesterol (Noncoding)

Chr	Genic Region	Nonrare Alleles at Multiallelic Positions Retained*			Nonrare Alleles at Multiallelic Positions Removed†		
		Variants	□ (SE)	*P* Value	Variants	□ (SE)	*P* Value
1	*GALNT2*	335	0.04 (0.12)	0.72	333	0.10 (0.12)	0.40
11	*APOA5*	100	−0.10 (0.14)	0.51	100	−0.10 (0.14)	0.51
11	*APOA5-APOC3* intergenic	151	0.31 (0.12)	0.009	149	−0.02 (0.14)	0.89
11	*APOC3*	62	0.30 (0.20)	0.13	62	0.30 (0.20)	0.13
12	*SBNO1*	39	0.21 (0.21)	0.31	39	0.21 (0.21)	0.31
12	*CCDC92*	251	−0.08 (0.12)	0.48	251	−0.08 (0.12)	0.48
12	*ZNF664*	156	0.03 (0.13)	0.84	156	0.03 (0.13)	0.84
12	*SCARB1*	292	0.20 (0.14)	0.15	290	0.16 (0.12)	0.17
16	*CETP*	229	−0.06 (0.12)	0.63	229	−0.06 (0.12)	0.63
18	*LIPG*	343	−0.34 (0.14)	0.02	341	−0.16 (0.12)	0.19
	All noncoding	1958	−0.63 (0.94)	0.50	1950	−0.27 (0.38)	0.48

*Aggregation of rare noncoding variants included nonrare alleles at multiallelic positions also harboring rare variants.
†Aggregation of rare noncoding variants with nonrare alleles at multiallelic positions removed.

Data Supplement). We first identified rare coding variants believed to be nonbenign in their putative functional consequence (n=213), organized them based on their predicted impact on protein function (ie, [1] disruptive, [2] disruptive plus missense, or [3] loss-of-function; see Materials and Methods for definitions), and then tested aggregate rare coding variant burden across all targeted genic regions for each predicted impact category. We found that for each predicted impact category, the collection of all rare coding variants did not exhibit a level of rare variant burden that was significantly associated with HDL-C. Similarly, variant aggregation over the coding regions of the individual gene targets separately (n=8) did not identify any individual region with significant variant burden associated with high versus low HDL-C (collapsing test; Tables 5 and 6).

We then asked whether the burden of rare noncoding variants across all targets contributed to extremely high HDL-C. Because of the fact that a methodological framework for predicting the potential regulatory impact of noncoding variants genome wide has yet to be widely accepted, the rare noncoding variants were not subdivided into putative functional categories like the coding variants described above. Thus, we first analyzed all rare noncoding variants as a single group, which resulted in a variant burden that was not significantly associated with high HDL-C in our cohort (P=0.5; Tables 5 and 6). We next grouped rare noncoding variants by physical genic region (n=10) and performed variant burden analyses separately on each region. This approach identified a collection of 151 rare variants in the *APOA4-APOA5* intergenic region that were nominally significantly associated with extremely high HDL-C (P=9.43×10^{-3}, collapsing test; Tables 5 and 6). Within this region, we noted a collection of 3 different indels as multiple alternative alleles at the position chr11:116678249 (hg19). Of these, a rare deletion CAA>C (MAF, 0.003; AC, 7) exhibited nominally significant association with high HDL-C (P=0.0427, score test). The second allele was a common deletion (MAF, 0.06; AC, 138) that was not associated with high HDL-C (P=0.75, score test). The third allele was the same common (MAF, 0.26; AC, 605) yet previously unreported insertion of CAA>CAAA at chr11:116678249 that was significantly associated with high HDL-C (P=8.9×10^{-4}, score test) in the single-variant analysis.

We hypothesized that these particular common alternative alleles were driving the nominally significant rare variant burden association signal for the *APOA4-APOA5* intergenic region. To test this, we removed it (and all other nonrare variants at multiallelic sites) and reassessed rare variant burden and found a complete attenuation of the association (P=0.43; Tables 5 and 6), thus suggesting that the originally significant association of the cluster of *APOA4-APOA5* intergenic variants with HDL-C was driven by common alleles alone.

DISCUSSION

Translating GWAS trait- and disease-associated common variants to bona fide causal variants, genes, and biological mechanisms has been a major challenge for human genetics. This is due in part to small effect sizes of GWAS variants, and thus resequencing of candidate genes at GWAS loci at the phenotypic extremes of complex traits has become a leading approach to identify rare variants with larger effects. To date, this approach has been applied to coding regions of GWAS candidate genes, yet coding variants account for only a small fraction (≈11%) of all variants tagged complex trait GWAS studies,[16] underscoring the need to search the noncoding genome for rare, putatively causal variants. Here, we utilized an inexpensive, modular, and scalable targeted sequencing approach for identifying rare noncoding variants in candidate genes influencing HDL-C—a complex trait with 72 associated loci from GWAS.[7] Our proof-of-principle resequencing study of 7 candidate gene regions in 797 extremely high HDL-C versus 735 low HDL-C participants rediscovered and validated nearly all prior GWAS-implicated tag SNPs and revealed ≈2000 variants in noncoding regions of targets, including rare, novel noncoding variants that were nominally associated with HDL-C in our study. As such, our findings provide one of the first applications of a multiplexed targeted resequencing study of noncoding variants across multiple loci at the phenotypic extremes of a complex trait.

We selected 7 candidate loci for targeted sequencing of coding and noncoding regions in our cohorts (Figure 1). Four of the targeted loci, *APOC3*, *SCARB1*, *CETP*, and *LIPG*, have known roles in HDL metabolism for which loss of function has been shown to elevate HDL-C in humans.[17] To explore the hypothesis that rare noncoding variants may underlie GWAS-implicated loci for HDL-C levels, we selected 3 HDL-C loci newly identified through GWAS, *GALNT2*, *SBNO1*, and the *CCDC92-ZNF664* region for our targeted sequencing. Some sequencing efforts have suggested that *GALNT2* coding variants segregate with elevated HDL-C, whereas a recent report from our group found an opposite result for 2 rare coding variants.[18] Similarly, the contribution of either coding or noncoding rare variants at the *CCDC92-ZNF664* and *SBNO1* loci to HDL metabolism remains completely unexplored. Therefore, we evaluated these loci for rare coding and noncoding variants to better determine the directional relationship of these genes with HDL-C beyond the initial common variant associations.

We rediscovered previously implicated variants in our cohort, along with the initial discovery of a few novel candidates requiring statistical support. Most notably, we found significant or nominally significant associations for a majority (55%) of GWAS-implicated HDL-C variants overlapping our targeted regions with consistent directionality to prior associations of these variants. However, we replicated these associations at <1/100th the cohort size of the most recent GWAS for HDL-C (188 577 participants[7] versus 1532 participants in our study) through our phenotypic extremes design. We also identified 3 rare (MAF, <0.01) or low-frequency (MAF, <0.05) nonsynonymous coding variants associated with HDL-C levels with directionalities consistent with previous reports (*CETP* Ala390Pro,[19] *CETP* Arg468Gln,[20] and *LIPG* Asn396Ser[21]). Collectively, these findings support the utility of candidate gene and noncoding locus resequencing at the extremes of a continuous trait distribution to enrich for trait-associated alleles, which may allow ascertainment of genetic associations in smaller populations than historical sizes for complex trait GWAS, such as understudied ethnicities and population isolates.

Our study also has important methodological implications for future targeted resequencing efforts. To date, MIP sequencing has been applied to targeted sequencing of coding regions of candidate genes with a sample preparation cost of <$1 per participant.[8] Our use of MIPs to interrogate noncoding regions of HDL-C candidate genes represents one of the first applications of this methodology for regulatory DNA regions. Our sequencing efforts were completed at a comparable cost with the prior applications, with similar target-coverage depths across coding and noncoding targets. Additionally, our modified dual-barcoding approach allowed us to multiplex all 1532 samples for sequencing in a single lane of an Illumina HiSeq2500 sequencing run with a median base coverage per participant of 110-fold—a robust depth for novel and rare variant identification at a sequencing cost of ≈$2000. Thus, our study highlights the utility of an MIP-based approach for sequencing of noncoding regions at a low per-sample cost.

Among the variants we identified were multiple noncoding variants in *CETP* and *LIPG* associated with high HDL-C in our cohort. CETP is a circulating regulator of HDL metabolism, and genetic inactivation of CETP is associated with increased HDL-C in humans.[17] Similarly, *LIPG*, which encodes an enzyme catabolizing HDL called endothelial lipase, contributes to heritable differences in HDL-C. *LIPG* loss-of-function genetic variants are causal contributors to elevated HDL-C in humans.[18,22] Here, we expanded the allelic spectrum of rare noncoding variation in these 2 HDL-C modulating genes contributing to high HDL-C levels in humans. In both cases, the frequency of these variants and our limited sample size requires further analysis in follow-up cohorts to demonstrate conclusive association of these rare alleles with HDL-C.

Our current study has limitations, which serve as opportunities for further study. First, our total cohort size of 1532 participants limits both the ascertainment of the full spectrum of rare variants that may underlie extremely high HDL-C levels, as well as the power of our statistical tests of common variant association and rare variant burden. Second, our population of high and low HDL-C participants was largely of European ancestry, thus limiting our ability to extrapolate the variants discovered to other populations. Additionally, our ability to validate variant discovery in this population for our MIP sequencing-identified variants was limited to SNPs largely in coding regions as we utilized the Exome Array for confirmatory genotyping; thus we were restricted in our ability to validate indels overlapping these regions. Also, we used conventional strategies for rare-variant grouping, which focused on gene-level aggregation. However, for noncoding sequences, it was not obvious which variant grouping strategy is optimally powered, which remains an open question in the field. Finally, because we selected a finite sequence of noncoding genome with genomic annotations that we believed a priori would be functional and lipid related (eg, enhancer marks in liver) and focused on annotations in HepG2 cell lines for this selection, it remains possible that rare-variant burden either exists in other sequences we did not target here and that such selection could have differed if we had used alternative cell sources for regulatory annotation, such as primary human hepatocytes.

In conclusion, our MIP-based targeted sequencing approach has demonstrated the successful capture of noncoding regions for the discovery of rare, noncoding variants associated with HDL-C in a cohort of extremely high versus low HDL-C participants. Though efforts to better identify the spectrum of noncoding variants underlying complex traits have initiated, including denser genotyping of noncoding variants[20] and whole-genome sequencing,[22] these approaches remain expensive and not readily applicable to the study of large populations or large case-control designs. Our results offer a scalable and cost-effective targeted approach that complement future, larger candidate loci resequencing efforts for the discovery of putatively causal noncoding variants. These efforts, coupled with appropriate functional investigation of identified variants for impact on gene regulation, may substantially refine the causal genes at loci implicated from GWAS studies and also help further explain the missing heritability underlying complex traits, such as HDL-C.

ARTICLE INFORMATION

Received December 18, 2017; accepted May 9, 2018.

Guest Editor for this article was Christopher Semsarian, MBBS, PhD, MPH.

The Data Supplement is available at http://circgenetics.ahajournals.org/lookup/suppl/doi:10.1161/CIRCGEN.117.002070/-/DC1.

Correspondence

Benjamin F. Voight, PhD, Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine at the University of Pennsylvania, 10–126 SCTR, Bldg 421, 3400 Civic Center Blvd, Philadelphia, PA 19104, E-mail bvoight@pennmedicine.upenn.edu or Daniel J. Rader, MD, Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, 11–125 SCTR, Bldg 421, 3400 Civic Center Blvd, Philadelphia, PA 19104, E-mail rader@pennmedicine.upenn.edu

Affiliations

Department of Genetics (S.A.K., P.L.B., W.Z., W.F.H.-C., C.D.B., D.J.R.), Department of Medicine (S.A.K., W.Z., W.F.H.-C., D.J.R.), Department of Systems Pharmacology and Translational Therapeutics (P.L.B., B.F.V.), and Institute for Translational Medicine and Therapeutics (D.J.R., B.F.V.), Perelman School of Medicine at the University of Pennsylvania, Philadelphia. Albert Einstein College of Medicine, Bronx, NY (W.Z.).

Acknowledgement

The authors would like to thank J. Shendure, B. Martin, and J. Schug for their expertise and guidance in the initiation and optimization of our molecular inversion probe sequencing pipeline.

Sources of Funding

This work is supported by the National Institutes of Health grants F30 HL124967 (Dr Khetarpal), R37HL055323, R01 HL111398 and R01 HL089309 (Dr Rader), and R01DK101478 (Dr Voight).

Disclosures

Dr Voight currently serves as an associate editor at *Circulation: Genomics and Precision Medicine.* The other authors report no conflicts.

REFERENCES

1. Bomba L,et al. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol*. 2017;18:77. doi: 10.1186/s13059-017-1212-4.
2. Patel AP, et al. Targeted exonic sequencing of GWAS loci in the high extremes of the plasma lipids distribution. *Atherosclerosis*. 2016;250:63–68. doi: 10.1016/j.atherosclerosis.2016.04.011.
3. Johansen CT, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet*. 2010;42:684–687. doi: 10.1038/ng.628.
4. Lee SH,et al. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet*. 2011;88:294–305. doi: 10.1016/j.ajhg.2011.02.002.
5. Liu DJ, et al; Charge Diabetes Working Group; EPIC-InterAct Consortium; EPIC-CVD Consortium; GOLD Consortium; VA Million Veteran Program. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat Genet*. 2017;49:1758–1766. doi: 10.1038/ng.3977.
6. Roman TS, et al. Multiple hepatic regulatory variants at the GALNT2 GWAS locus associated with high-density lipoprotein cholesterol. *Am J Hum Genet*. 2015;97:801–815. doi: 10.1016/j.ajhg.2015.10.016.
7. Willer CJ, et al; Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45:1274–1283.
8. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*. 2012;338:1619–1622. doi: 10.1126/science.1227764.
9. Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. *Nat Genet*. 2014;46:200–204. doi: 10.1038/ng.2852.
10. Auton A, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
11. Sudmant PH, et al; 1000 Genomes Project Consortium. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526:75–81. doi: 10.1038/nature15394.
12. Kircher M,et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46:310–315. doi: 10.1038/ng.2892.
13. Battle A, et al. Genetic effects on gene expression across human tissues. *Nature*. 2017;550:204–213.
14. Zhao W, et al; CHD Exome+ Consortium; EPIC-CVD Consortium; EPIC-Interact Consortium; Michigan Biobank. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat Genet*. 2017;49:1450–1457. doi: 10.1038/ng.3943.
15. Khetarpal SA, et al. Mining the LIPG allelic spectrum reveals the contribution of rare and common regulatory variants to HDL cholesterol. *PLoS Genet*. 2011;7:e1002393. doi: 10.1371/journal.pgen.1002393.
16. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337:1190–1195. doi: 10.1126/science.1222794.
17. Larach DB,et al. Monogenic causes of elevated HDL cholesterol and implications for development of new therapeutics. *Clin Lipidol*. 2013;8:635–648. doi: 10.2217/clp.13.73.
18. Vitali C,et al. HDL cholesterol metabolism and the risk of CHD: new insights from human genetics. *Curr Cardiol Rep*. 2017;19:132. doi: 10.1007/s11886-017-0940-0.
19. Spirin V, et al. Common single-nucleotide polymorphisms act in concert to affect plasma levels of high-density lipoprotein cholesterol. *Am J Hum Genet*. 2007;81:1298–1303. doi: 10.1086/522497.
20. Surakka I, et al; ENGAGE Consortium. The impact of low-frequency and rare variants on lipid levels. *Nat Genet*. 2015;47:589–597. doi: 10.1038/ng.3300.
21. Edmondson AC, et al. Loss-of-function variants in endothelial lipase are a cause of elevated HDL cholesterol in humans. *J Clin Invest*. 2009;119:1042–1050. doi: 10.1172/JCI37176.
22. Helgadottir A, et al. Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat Genet*. 2016;48:634–639. doi: 10.1038/ng.3561.

Circulation
Genomic and Precision Medicine

American
Heart
Association®

Multiplexed Targeted Resequencing Identifies Coding and Regulatory Variation Underlying Phenotypic Extremes of High-Density Lipoprotein Cholesterol in Humans

Sumeet A. Khetarpal, Paul L. Babb, Wei Zhao, William F. Hancock-Cerutti, Christopher D. Brown, Daniel J. Rader and Benjamin F. Voight

Circ Genom Precis Med. 2018;11:
doi: 10.1161/CIRCGEN.117.002070
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1942-325X. Online ISSN: 1942-3268