HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2014 October 01.

Published in final edited form as:

Nat Genet. 2014 April ; 46(4): 345–351. doi:10.1038/ng.2926.

Systematic evaluation of coding variation identifies a candidate causal variant in *TM6SF2* influencing total cholesterol and myocardial infarction risk

Oddgeir L. Holmen[1,2,*], **He Zhang**[3,*], **Yanbo Fan**[3,*], **Daniel H. Hovelson**[3,4], **Ellen M. Schmidt**[3,4], **Wei Zhou**[3], **Yanhong Guo**[3], **Ji Zhang**[3], **Arnulf Langhammer**[1], **Maja-Lisa Løchen**[5], **Santhi K. Ganesh**[3,6], **Lars Vatten**[7], **Frank Skorpen**[8], **Håvard Dalen**[9,10], **Jifeng Zhang**[3], **Subramaniam Pennathur**[11], **Jin Chen**[3], **Carl Platou**[9], **Ellisiv B. Mathiesen**[12,13], **Tom Wilsgaard**[5], **Inger Njølstad**[5], **Michael Boehnke**[14], **Y. Eugene Chen**[3], **Gonçalo R. Abecasis**[14], **Kristian Hveem**[1,9], and **Cristen J. Willer**[3,4,6]

[1]HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway

[2]St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

[3]Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, Michigan, United States of America

[4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America

[5]Epidemiology of Chronic Diseases Research Group, Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

Corresponding authors Cristen Willer, Division of Cardiovascular Medicine and Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, cristen@umich.edu, Tel: 734-647-6018, Fax: 734-764-4142, Kristian Hveem, HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Trondheim, Norway, kristian.hveem@ntnu.no, Tel: +47-7407-5180, Fax: +47-7359-7577.

*These authors contributed equally to this work.

Author Contributions

A.L., I.N., K.H., H.D., C.P., E.B.M., T.W., L.V., F.S., M.L.L. and O.L.H. obtained, contributed and analyzed the phenotype data. O.L.H. and T.W. was responsible for sample selection. O.L.H. and H.Z. were responsible for genetic data analysis and interpretation. H.Z. and J.C. performed variant calling from sequence data. D.H.H., E.M.S., and W.Z. generated figures and performed secondary analyses. Epidemiological expertise was provided by L.V., M.L.L., S.K.G., A.L., E.B.M., I.N. and K.H. Clinical expertise was provided by H.D. and C.P. Genotyping and genetic epidemiology expertise was provided by G.R.A., F.S. and M.B. Mouse experiments were conducted by Y.F., Y.G., J.Z., S.P. and J.Z. under the supervision of Y.E.C. with assistance from C.J.W. The manuscript was drafted by C.J.W., G.R.A., and O.L.H., with assistance from D.H.H., H.Z., M.B., and K.H, and then critically reviewed, including comments and feedback from Y.F., E.M.S., W.Z., Y.G., J.Z., A.L., M.L.L., S.K.G., L.V., F.S., H.D., J.C., C.P., E.B.M., T.W., I.N. and Y.E.C. The study was conceived by C.J.W., O.L.H. and K.H., and the study was designed by C.J.W., O.L.H., K.H., M.B. and G.R.A. Overall leadership for the project was provided by K.H. and C.J.W.

Competing financial interests

The authors declare no competing financial interests.

URLs

The HUNT study: www.ntnu.edu/hunt

The Tromsø Study: www.tromsoundersokelsen.no

Exome array design webpage: genome.sph.umich.edu/wiki/Exome_Chip_Design

ImageJ software: rsweb.nih.gov/ij/

[6]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, United States of America

[7]Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway

[8]Department of Laboratory Medicine, Children's and Women's Health, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[9]Department of Medicine, Levanger Hospital, Nord-Trøndelag Health Trust, Levanger, Norway

[10]MI Lab, Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway

[11]Department of Internal Medicine, Division of Nephrology, University of Michigan, Ann Arbor, Michigan, United States of America

[12]Brain and Circulation Research Group, Department of Clinical Medicine, Faculty of Health Sciences, UiT The Artic University of Norway, Tromsø, Norway

[13]Brain and Circulation Research Group, University Hospital of North Norway, Tromsø, Norway

[14]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America

Abstract

Blood lipid levels are heritable, treatable risk factors for cardiovascular disease. We systematically assessed genome-wide coding variation to identify novel lipid genes, fine-map known lipid loci, and evaluate whether low frequency variants with large effect exist. Using an exome array, we genotyped 80,137 coding variants in 5,643 Norwegians. We followed up 18 variants in 4,666 Norwegians to identify 10 loci with coding variants associated with a lipid trait ($P < 5\times10^{-8}$). One coding variant in *TM6SF2* (p.Glu167Lys), residing in a GWAS locus for lipid levels, modifies total cholesterol levels and is associated with myocardial infarction. Transient overexpression and knockdown of *TM6SF2* in mouse produces alteration in serum lipid profiles consistent with the association observed in humans, identifying *TM6SF2* as the functional gene at a large GWAS locus previously known as *NCAN/CILP2/PBX4* or 19p13. This study demonstrates that systematic assessment of coding variation can quickly point to a candidate causal gene.

Circulating blood lipids are heritable, treatable, risk factors for cardiovascular disease, a leading cause of death globally[1,2]. Understanding the genetic basis of lipid levels in humans can identify targets for new, improved therapies for cholesterol management and prevention of heart disease[3]. Genome-wide association studies (GWAS) for plasma lipid levels have so far identified association with 157 loci[4,5], primarily represented by one or more common variants (minor allele frequency [MAF] > 5%) with small effect sizes. These GWAS variants together explain ~12–14% of the trait variation in lipid levels, corresponding to 20–30% of the total genetic contribution to these traits[6]. Some of the "missing heritability" may be due to low frequency (MAF 1–5%) and rare (MAF < 1%) variants that are not well tested by GWAS[7–9]. These low frequency and rare variants are plentiful in the genome[10,11], but are difficult to capture on GWAS chips, directly or through imputation[12–14].

Systematic assessment of association between blood lipid levels and coding variants has several potential benefits. First, it could implicate new loci in the regulation of blood lipids. Second, it could lead to the discovery of new lipid modifying alleles at known loci that point to candidate causal genes. In some cases where GWAS signals are shadows of a nearby rare variant with much larger effects, these alleles could be critical in directing follow-up functional experiments. For example, in *PCSK9* a low frequency functional variant explains the nearby common variant GWAS signal[15], suggesting that the GWAS variant has no relevant functional consequence and would not be a productive target for functional experiments. Even when they do not account for the GWAS signal, rare coding variants in known loci can pinpoint specific genes as candidates for follow-up and functional analyses and clarify biology. A good example of the later situation is *IFIH1*, where multiple independently associated loss-of-function variants, identified through GWAS, show that disabling this gene will reduce the risk of type 1 diabetes[16].

In this study, we systematically assess coding variants for association with lipid levels. Through chip based genotyping of >80,000 coding variants (>68,000 with MAF <5%) and follow-up of interesting variants, we identify associated coding variants in 10 loci. Nine of these variants point to genes with well-established roles in lipid metabolism, the tenth (p.Glu167Lys in *TM6SF2*) suggests a causal gene in a previously described GWAS locus[17,18]. Follow-up experiments in the mouse show that over-expression of human *TM6SF2* raises total cholesterol compared to a control construct, and that knockdown of endogenous *Tm6sf2* decreases total cholesterol, consistent with this gene being involved in the regulation of blood lipid levels.

Results

Genotyping of discovery sample and assessment of coverage

To systematically assess the role of coding variants in lipid levels, we successfully genotyped 5,771 Norwegian participants from the population-based Nord-Trøndelag Health Study (the HUNT study)[19] for 234,187 variants using the Illumina HumanExome Beadchip arrays. Among the 5,643 (97.8%) analyzed individuals passing quality control, 80,137 coding variants were polymorphic in our sample, of which 68,615 had a frequency <5% (Table 1). We considered coding variants to refer to protein-altering variants: premature stop, essential splice donor/acceptor, readthrough or missense. Clinical characteristics of the stage 1 study participants are summarized in Supplementary Table 1.

To quantify array coverage of all coding variation present in our Norwegian sample, we performed low-pass whole genome sequencing with exome enrichment on 152 samples (2.7% of Stage 1 sample). Average sequencing depth was 45× for the exome target capture regions. We identified 46,170 coding variants in our sample via sequencing (5,648 on average per individual). Concordance between non-reference sequencing-based genotypes (>10× depth) and exome array genotypes was >99% for all allele frequencies (see Online Methods for details). Overall, we estimate that 70.9%, 77.4% and 78.0% of rare, low-frequency and common coding variants (MAF <1%, 1–5% and >5%) observed in >1 sequenced samples were successfully genotyped using the array (Table 1). Most of the rare and low-frequency coding variants identified via sequencing and typed on the array have not

been examined in previous lipid GWAS and cannot be imputed accurately using HapMap or 1000 Genome reference panels[4,5], providing unique opportunities for evaluating the effect of low-frequency variants on lipid levels.

Evaluation of known lipid signals

To validate our approach, we tested for association at known GWAS loci. Among the 157 previously described independent lipid-associated SNPs[4], 127 were directly genotyped on the exome array. For the remaining 30 SNPs not genotyped, 17 were discovered after the array was designed, and 13 failed array design. Among the 127 variants examined, we identified genome-wide significant association at 7 (probability that 7 or more loci exceed $P < 5\times10^{-8}$ is $P_{binominal} = 5\times10^{-41}$) and nominal association ($P < 0.05$) at 45 (probability that 45 or more loci exceed $P < 0.05$ is $P_{binominal} = 3\times10^{-27}$). We compared the effect sizes for low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, total cholesterol (TC), and triglycerides (TG) observed in our non-fasting samples to those observed in much larger GWAS reports[4] and found a high degree of correlation (LDL $r^2 = 0.46$, HDL $r^2 = 0.75$, TC $r^2 = 0.47$, TG $r^2 = 0.84$; Supplementary Figure 1).

Furthermore, when lipid traits were examined individually, the direction of effect for associated SNPs matched that reported in GWAS[4] for 79%, 83%, 85% and 90% of variants (HDL, TG, LDL and total cholesterol; Supplementary Table 2), suggesting high reliability of array genotype data for our -downstream analyses.

Follow-up genotyping and summary of association results

We next tested for association between coding variants with at least 6 copies of the minor allele (51,453 coding variants) and lipid measurements in our stage 1 samples (N = 5,643). We successfully genotyped 18 variants for follow-up in 4,666 Norwegian individuals from the population-based Tromsø Study[20] (Stage 2). We saw no evidence for inflation of test statistics (HDL $\lambda_{GC} = 1.06$, LDL $\lambda_{GC} = 1.04$, TC $\lambda_{GC} = 1.08$, TG $\lambda_{GC} = 1.04$, Supplementary Figure 2). We selected for replication coding variants with $P < 2\times10^{-5}$, MAF < 10%, and minor allele count (MAC) > 5 to capture variants less likely to be covered by GWAS but with enough rare alleles to allow for association testing. We performed a joint association analysis in stage 1 and stage 2 samples (N = 10,309). We considered the threshold for significance to be 5×10^{-8} to be consistent with GWAS. Clinical characterization of stage 2 study participants are summarized in Supplementary Table 1.

This design had 80% power to detect variants with MAF 1% and effect size of 0.44 standard deviation (SD) units (Figure 1) at alpha of 5×10^{-8}. As expected, power varies with allele frequency so that we have 80% power to detect common variants with MAF 10% and effect size of 0.15 standard deviations, and equivalent power for variants with MAF 0.5% and effect size of 0.62 standard deviations.

Overall, we identified coding variants in 11 genes that reach genome-wide significance (Table 2). Most of these variants map to genes whose roles in lipid metabolism are now well established (*APOE, APOB, ABCG5, ABCG8, CETP, LIPC, LIPG, APOA5* and *LPL*) but two map to genes (*RNF111* and *TM6SF2*) that were not previously implicated in blood lipid

levels. We will first briefly summarize results at known loci, before dissecting results at *RNF111* and *TM6SF2* in greater detail. Further details of association results are given in Supplementary Table 3 and Supplementary Figure 2.

Variants in previously implicated genes

To explore the relationship between coding variants identified in our experiment and previous GWAS signals, we performed conditional analysis, evaluating association at each coding variant after accounting for each nearby GWAS index SNP and vice-versa. In addition, to define a list of independent association signals at each locus, we performed a conditional analysis adjusting for all previously known lipid-associated SNPs.

LDL cholesterol and variants in *APOE, APOB, ABCG5* and *ABCG8*

We identified LDL-associated coding variants in four previously implicated genes: *APOE, APOB,* and at *ABCG5* and *ABCG8* (which are adjacent genes that share a promoter on chromosome 2). Variants at three of these genes (*APOB, ABCG5* and *ABCG8*) were also associated with total cholesterol levels, consistent with the strong correlation between LDL and total cholesterol levels. Dissection of the association signals at the three loci suggested different relationships between coding variants and previously reported GWAS signals.

The coding variant at *APOB* p.Thr98Ile (rs1367117) was previously identified as the GWAS index SNP for LDL and total cholesterol[4,5]. In contrast, at *APOE*, association signals with the p.Arg176Cys variant (rs7412; Table 2) and the GWAS index SNP (rs4420638) appeared mostly independent (Supplementary Table 4). The p.Arg176Cys variant is one of the two SNPs that determine well known functional haplotypes in *APOE*[21] and the GWAS index SNP is thought to be a shadow of p.Cys130Arg (rs429358), the other functional SNP in the gene, which is difficult to genotype and was not successfully assayed in our samples.

At *ABCG5* and *ABCG8*, coding variants (*ABCG5* p.Arg50Cys and *ABCG8* p.Asp19His, both MAF = 6.5%) were assayed in GWAS[4], but an intronic SNP (rs4299376) was assigned as the GWAS index SNP because of stronger evidence for association. Conditional analysis in our samples shows that signal at the two highly correlated *ABCG5/8* missense variants is only modestly attenuated and thus likely to be independent from the GWAS index SNP (*P* value in stage 1 reduced from $P = 1\times10^{-7}$ to $P_{\text{conditional}} = 3\times10^{-6}$; Supplementary Table 4).

Triglycerides and variants in *APOA5, LPL* and *ANGPTL4*

We identified triglyceride-level associated coding variants in three genes, *APOA5, LPL* and *ANGPTL4*. Although *ANGPTL4* is a locus also identified through GWAS, low frequency variant p.Glu40Lys variant (rs116843064, MAF = 3.2%, $P_{stage1} = 8\times10^{-9}$) shows significant association even in the absence of a significant signal at the nearby GWAS index SNP (rs7255436, $P_{stage1} = 0.066$; Supplementary Table 4), suggesting the two are independent or that the GWAS index SNP is a weak proxy for p.Glu40Lys. Associations at p.Ser19Trp in *APOA5* (rs3135506, MAF = 6.1%) and p.Ser474Stop in *LPL* (rs328, MAF = 9.6%, previously known as p.Ser447Stop) were completely attenuated by conditioning on the nearby GWAS index SNPs (rs964184 and rs12678919, respectively; Supplementary Table 5). In *APOA5*, the GWAS SNP still showed association after adjusting for the coding variant

($P_{conditional} = 2\times10^{-21}$) suggesting p.Ser19Trp as a shadow to the GWAS variant. In *LPL*, the GWAS signal was completely attenuated ($P = 0.96$) by the coding variant suggesting the effects of these variants on triglyceride levels are statistically indistinguishable.

HDL cholesterol and variants in *CETP, LIPC* and *LIPG*

We identified four HDL-associated coding variants in three previously implicated genes: *CETP* (2 variants), *LIPC* and *LIPG* (Table 2). In all three genes, these variants show significant association even after adjusting for nearby GWAS index SNPs (Supplementary Table 4). Three of the HDL-associated variants were low frequency (MAF < 5%) and thus beyond the reach of traditional GWAS genotyping arrays and imputation. The ancestral alleles at two variants in *CETP* (Ala at position 390 and Val at position 422) were associated with increased HDL cholesterol (Supplementary Table 4). The p.Thr405Met variant (rs113298164, $P = 6\times10^{-6}$, MAF = 0.7%, previously known as p.Thr383Met) in the hepatic lipase gene (*LIPC*) was the only rare coding variant (MAF < 1%) we identified to be associated with a lipid trait. The p.Thr405Met variant has previously been observed in families with hepatic lipase deficiency, a disease resulting in elevated lipid levels and triglyceride rich HDL subfractions[22], but – until now – had not been associated with HDL cholesterol in population samples (Supplementary Figure 3a). Conditional analysis demonstrates that this association signal appears to be independent of the two common variants that independently show association 5' of the gene5 ($P_{cond} = 8\times10^{-9}$, adjusting for rs1532085 and rs261334; Supplementary Figure 3c).

Rare variant association with *RNF111* due to *LIPC* variant

We were initially intrigued by association between the rare p.Pro836Ser variant in *RNF111* (rs181181625, MAF = 0.46%) and HDL cholesterol ($P = 3\times10^{-9}$). The gene is >500kb from any known GWAS index SNPs (Table 2) and association remained significant after adjusting for the top GWAS association signals in the chromosome, both coding and non-coding. Interestingly, the variant appears at low frequencies in Norwegians (MAF = 0.46% in N = 10,309) and Estonians[23] (MAF = 0.30% in N = 9,328), and is absent in other samples we examined: 2,979 Dutch Caucasians from the Rotterdam Study[24], and 5,338 Hispanic-American and 4,216 African-Americans from the Mt Sinai BioMe study[25].

A detailed conditional association analysis to identify independently associated variants on chromosome 15 revealed that the *RNF111* p.Pro836Ser variant shows no evidence for association after accounting for *LIPC* p.Thr405Met located 522,192 bases away ($P_{conditional} = 0.82$; Supplementary Table 4 and Supplementary Figure 3). Long-range linkage disequilibrium (LD, $r^2 = 0.60$) between the two variants suggests that the *RNF111* variant might have arisen on the *LIPC* p.Thr405Met haplotype in a common ancestor of Norwegians and Estonians.

Having dissected associated variants in familiar loci and recognized the potentially novel signal in *RNF111* as a shadow of rare variant signals in *LIPC*, we turned our attention to coding variants in an unfamiliar gene that showed significant association with total cholesterol.

Identification of *TM6SF2* as causal gene for total cholesterol

In *TM6SF2*, the p.Glu167Lys variant (rs58542926, MAF = 8.9%) reached the threshold for genome-wide significant association with total cholesterol ($P = 4\times10^{-8}$; Supplementary Figure 4). The *TM6SF2* variant was in strong linkage disequilibrium with the GWAS index SNP at the gene-rich locus previously named *NCAN*[4,17], *CILP2*[5], *PBX4*[18] or 19p13[18] (rs10401969, $r^2 = 0.97$; Supplementary Figure 4). This locus also harbors a coding variant in *NCAN* (rs2228603, p.Pro92Ser) in moderately strong LD with the index SNP ($r^2 = 0.71$), but the *TM6SF2* p.Glu167Lys variant shows stronger evidence for association in our sample (Supplementary Table 3d), and higher r^2 with the GWAS index SNP ($r^2 = 0.97$). Both p.Glu167Lys in *TM6SF2* and p.Pro92Ser in *NCAN* are classified as 'probably damaging' by PolyPhen2 (Polyphen score 0.996 and 0.957, respectively)[26]. However, *NCAN* is primarily expressed in the brain[27] and is not a strong candidate for having a role in lipid biology. There are 21 genes within the GWAS association signal and prior to this study it was unknown which of these mediate the observed association signal. In our sample, total cholesterol association for the p.Glu167Lys *TM6SF2* coding variant and the GWAS index SNP are indistinguishable, suggesting the coding variant could be causal. Specifically, conditioning on the coding variant, the total cholesterol association at the GWAS index SNP becomes non-significant ($P_{conditional} = 0.65$; Supplementary Table 4), demonstrating that the TM6SF2 variant accounts for the association signal detected by GWAS. *In silico* look-up in 92,605 European samples genotyped with the Metabochip[4] confirmed strong evidence for association with total cholesterol ($P = 1\times10^{-47}$; Supplementary Table 5), LDL cholesterol ($P = 2\times10^{-38}$) and triglyceride levels ($P = 9\times10^{-50}$).

To further test whether *TM6SF2* might be the functional gene at this locus, we determined the tissue distribution pattern of TM6SF2 in C57BL/6J mouse by Northern blotting and Western blotting. Endogenous *Tm6sf2* was highly expressed in the liver at both mRNA and protein levels, suggesting a potential role of TM6SF2 in hepatic lipid metabolism (Supplementary Figure 5). Next, we transiently overexpressed human *TM6SF2* in C57BL/6J mice using liver-targeting adenovirus containing the human *TM6SF2* coding region and compared lipid levels to mice injected with a control adenovirus construct containing LacZ. Five days after tail vein injection, the protein levels of TM6SF2 in Ad-*TM6SF2* transduced-mouse liver were 2.4-fold increased when compared with Ad-LacZ. In *TM6SF2* overexpressing-mice compared to mice injected with the control construct (LacZ), total cholesterol, LDL cholesterol and triglyceride levels were increased (TC: 2.3 fold, $P = 9\times10^{-4}$; LDL: 5.8 fold, $P = 4\times10^{-4}$, TG: 1.13 fold, $P = 0.031$, respectively), and HDL cholesterol levels were decreased (0.45 fold, $P = 9\times10^{-4}$; Figure 2).

We additionally tested the impact of knockdown of *Tm6sf2* by transient transduction of shRNA-*TM6SF2* in C57BL/6J mice using adenovirus containing short hairpin RNA targeting the *TM6SF2* coding region. Six days after tail vein injection, the protein levels of TM6SF2 were decreased by 49% in the liver of *TM6SF2*-knockdown mice (Figure 2). We found that fasting total cholesterol levels were decreased by 18.2% in *Tm6sf2* knockdown mice compared to controls ($P = 0.013$, N = 16 mice; Figure 2). Given overexpression of *TM6SF2* increases total cholesterol and knockdown of *Tm6sf2* decreases total cholesterol, and that the minor human allele is associated with reduced cholesterol, our results suggest

the presence of a positively charged amino acid (Lys) at codon 167 results in decreased function of TM6SF2 relative to the negatively charged amino acid (Glu) more commonly observed.

Association with myocardial infarction and liver disease

Our stage 1 sample included 2,833 medical-record confirmed myocardial infarction (MI) cases and 2,938 healthy controls matched on sex and birth year. In addition, our stage 2 follow-up sample included 2,349 medical-record confirmed MI cases and 2,317 controls. This allowed us to test for association with MI in the same individuals where the lipid associations were identified.

We observed suggestive association with MI for coding variants in two genes associated with lipid traits: triglycerides (*ANGPTL4*, rs116843064, OR = 0.78, $P = 0.003$) and total cholesterol (*TM6SF2*, rs58542926, OR = 0.87, $P = 0.005$). As expected, alleles associated with reduced lipid levels also conferred a reduced risk of myocardial infarction (Table 2).

At the *TM6SF2* locus, the GWAS index SNP was associated with CAD in a large sample of 20,597 cases and 61,046 controls[28] (rs10401969, OR = 0.90, 95% CI 0.85 – 0.95, $P = 2\times10^{-4}$). Since previous reports describe association with the correlated *NCAN* p.Pro92Ser variant (rs2228603) at this locus and non-alcoholic fatty liver disease (NAFLD, OR = 1.65, $P = 5\times10^{-5}$)[29], we tested for association with serum alanine aminotransferase levels, a marker of liver damage. In human samples, we found modest evidence for association between serum alanine aminotransferase levels and the *TM6SF2* p.Glu167Lys variant (rs58542926, N = 1,481, $P = 7\times10^{-3}$), but not with *NCAN* p.Pro92Ser (rs2228603, $P = 0.15$). This is consistent with the possibility that TM6SF2 p.Glu167Lys may also be responsible for association of this locus with NAFLD[29].

We found no evidence of triglyceride accumulation in mouse liver in any experimental or control mouse, suggesting that neither TM6SF2 expression changes nor adenovirus injection caused any liver damage in our experimental model (Supplementary Figure 6). We also found no significant differences in serum alanine aminotransferase levels in experimental versus control mice after either overexpression or knockdown of *Tm6sf2* (Supplementary Figure 7). We hypothesize that the absence of any liver disease phenotype in mice is due to transient exposure to altered TM6SF2 levels in contrast to a lifetime of exposure to altered TM6SF2 in humans carrying Glu167 in *TM6SF2*.

Tests for rare variants

To test rare and low frequency variants for association, we performed gene-based burden tests for LDL cholesterol, HDL cholesterol, total cholesterol and triglycerides. Using both CMC[30] and SKAT-O[31] statistical tests, we selected missense and loss-of-function variants using three allele frequency thresholds: MAF < 5%, MAF < 1% and MAF < 0.1%. There were no genes that reached exome-wide significance ($P < 5\times10^{-7}$) that were not already highlighted by single variant association tests (Supplementary Table 6). However, for three genes, the strength of association with a burden of rare variants was more significant than observed for a single variant test, and in each case, a second nominally significant SNP was

observed; *LIPG* p.Arg476Trp for HDL cholesterol (MAF = 0.18%, $P = 9\times10^{-4}$), *LIPC* p.Arg208His for LDL cholesterol (MAF = 0.16%, $P = 0.02$), and *ANGPTL4* p.Arg336Cys for triglycerides (MAF = 0.24%, $P = 0.01$). When we examined 104 additional missense or loss-of-function variants in the ten genes, using a reduced threshold for significance ($P < 5\times10^{-4}$), only one low frequency variant was associated with any lipid trait; *LPL* p.Asn318Ser ($P = 1\times10^{-4}$, MAF = 3.2%) with triglycerides (Supplementary Table 7).

Discussion

By examining a substantial fraction of the coding variants in >10,000 Norwegians, we identified significant association at 10 loci highlighting one novel functional gene (*TM6SF2*) at a GWAS locus for total cholesterol and 9 known functional genes: *APOE, APOB, ABCG5/8* for LDL cholesterol*; APOA5, LPL, ANGPTL4* for triglyceride levels; and *CETP, LIPC*, and *LIPG* for HDL cholesterol.

Of ten genome-wide significant variants, we identified two rare or low-frequency coding variants that had not previously been identified in population based association studies: *LIPC* p.Thr405Met for HDL cholesterol; and *ANGPTL4* p.Glu40Lys for triglycerides. This suggests that there are likely to be additional low frequency variants that were not detected by genome-wide association studies, a subset of which may have been previously identified by candidate gene studies or in dyslipidemia families. While this study had 80% power to detect variants with MAF 1% and effect size of 0.44 SD, and can exclude the possibility that large numbers of such variants exist, we have only modest power for detecting rare variants with smaller effects. Subsequent studies in larger sample sizes will be required to determine the proportion of heritability that can be explained by rare or low frequency variants.

The effect (0.7 SD, 10.5 mg/dL) of the rare missense p.Thr405Met variant in *LIPC* identified for HDL cholesterol was larger than for any lipid variant previously observed from GWAS (maximum 0.3 SD). This variant, with an allele frequency of 0.7% in Norwegians, highlights the importance of evaluating low-frequency variants for large effect sizes on lipid levels. This variant was most likely not captured by GWAS because: i) it is not present in HapMap, ii) variants with <1% frequency were not typically analyzed for association in GWAS, and iii) low frequency variants are difficult to impute accurately. Despite its large effect size, the variant also illustrates challenges in rare variant studies: the observed association was more significant for common variants with smaller effect sizes and, individually, these common variants accounted for more of the trait heritability. Identifying association with rare variants at the population level may require very large sample sizes even when their effect sizes are large.

The coding variants identified in our study (3–4 variants per trait) explained between 1.9% (for triglyceride levels) and 4.5% (for LDL cholesterol) of the trait variance (3.4% for HDL cholesterol and 3.1% for total cholesterol). In future clinical settings where genomic sequence might be available in patients, evaluation of genetic risk scores would be improved by incorporating these variants, and others yet to be discovered, with relatively high impact on lipid levels. However, the population attributable risk (although substantially higher per coding variant than for non-coding variants identified by GWAS) may have no relation to

the potential importance of the gene as a novel drug target. For example, non-coding variants near HMGCR have a small impact on LDL cholesterol levels (~2.4 mg/dL)[5] but disruption of this gene's protein by statins can have a substantial impact on an individual's LDL cholesterol levels (average decrease of 40 mg/dL)[32].

More importantly, the study of coding variants may quickly point to the functional gene, as we have demonstrated for *TM6SF2*. At loci where linkage disequilibrium extends over a large region containing many genes, as is the case for *TM6SF2*, functional experiments on a large number of genes would be extremely challenging[33]. Instead, identification of an associated coding variant allowed us to prioritize one gene previously unknown to be involved in regulation of blood lipid levels as a functional gene.

Functional follow-up by overexpressing the human form of *TM6SF2* in mice confirmed the gene as attractive functional candidate in the locus previously known as *NCAN*[4,17], *CILP2*[5], *PBX4*[18] or 19p13[18]. In mice, overexpression of human *TM6SF2* resulted in higher total cholesterol, LDL cholesterol, triglycerides and lower HDL cholesterol, while conversely, knockdown of *Tm6sf2* resulted in decreased total cholesterol. Consistent with the expectation that increased cholesterol levels cause cardiovascular disease, in humans we found that the Glu167 allele associated with reduced total cholesterol also confers a reduced risk of myocardial infarction (OR = 0.87, 95% CI 0.79–0.93, $P = 0.005$; Table 2). Additional experiments are needed to clarify the value of *TM6SF2* as a potential drug target.

We expect that additional coding variants with association to lipid traits will be identified using larger sample sizes as well as expanded coverage of coding variation not assayed in this study. However, our results, in particular the identification of a common coding variant and functional gene *TM6SF2,* illustrate the value of systematic studies of coding variation in identifying lipid-associated genetic variation relevant for prevention of cardiovascular disease.

Online Methods

Subjects

We selected samples from two large population-based cohorts in Norway. Clinical characteristics of the study participants are provided in Supplementary Table 1.

The HUNT study (Stage 1)—We included 5,771 individuals with at least one lipid measurement from the second survey of the Nord-Trøndelag Health Study (HUNT): 2,833 cases with hospital diagnosed myocardial infarction (primary phenotype) and 2,938 healthy controls without cardiovascular disease matched on sex, birth year (+/− 1 year), and municipality or geographical region to minimize population stratification. HUNT is a population based health study with personal and family medical histories on approximately 120,000 individuals from Nord-Trøndelag County, Norway, collected in three surveys (HUNT 1, 2, and 3)[19,35]. HUNT 2 was conducted in 1995–97, inviting all residents ≥20 years of age in Nord-Trøndelag County, Norway. Self-reported questionnaires, clinical examination, and non-fasting venous blood samples were collected on 62,816 individuals (66.9% of invited). The population of Nord-Trøndelag County is ethnically homogenous (<

3% non-Caucasian ethnicity), making it especially suitable for epidemiological genetic research[35].

The Tromsø Study (Stage 2)—We included for follow up 4,666 individuals from the fourth wave of the Tromsø Study (Tromsø 4): 2,349 hospital-diagnosed myocardial infarction (primary phenotype) cases identified retrospectively and 2,317 controls without cardiovascular disease matched on sex and birth year (+/− 2.5 years). The Tromsø Study is a population based health study with medical histories on approximately 40,000 individuals of the Tromsø municipality, North Norway, collected with six surveys (Tromsø 1–6). Tromsø 4 was conducted in 1994–95 with questionnaires, clinical examination and non-fasting venous blood samples from 27,158 individuals ≥25 years of age[20].

Laboratory measurements

Clinical chemical analyses were conducted on fresh venous non-fasting blood samples (16% of HUNT participants and 25% of Tromsø Study participants reported ≥4 hours fasting prior to blood draw) at Levanger Hospital (Stage 1) or the University Hospital of North Norway (Stage 2). Total cholesterol, high density lipoprotein (HDL) cholesterol, and triglycerides where measured by an enzymatic colorimetric method using Hitatchi 911 Auto-Analyzer (Mito, Japan; Stage 1) applying commercial kits from Boehringer Mannheim (Mannheim, Germany; both stages). HDL cholesterol was measured after precipitation with phosphotungstate and magnesium ions (Stage 1) or heparin and manganese chloride (Stage 2). Day-to-day coefficients of variation were 1.3%–1.9% for total cholesterol, 2.4% for high density lipoprotein-cholesterol, and 0.7% – 1.3% for triglyceride (Stage 1). Alanine aminotransferase was measured by NADH (with P-5'-R) methodology using Architect cSystems ci8200 (Abbot Diagnostics, Ireland) and Activated alanine aminotransferase assays (Abbot Laboratories, IL; Reagent kit 8D36-30, 3913/R3). Day-to-day coefficients of variation were 3.4% in the low range and 1.8% in the high range.

Phenotypes

Lipids—Associations are reported for four lipid traits: low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, total cholesterol (TC), and triglycerides (TG). LDL cholesterol was calculated using Friedewald formula36. Information on lipid lowering medication was not available for the cohorts studied. The use of statins for primary prevention purposes was rare in Norway in the 1990s and the early 2000. *Liver enzyme.* Association is reported for alanine aminotransferase.

Myocardial infarction—We identified acute myocardial infarction (MI) retrospectively by linking our study participants to the electronic diagnosis registry of the two local hospitals in Nord-Trøndelag County (Stage 1) and the University Hospital of North Norway (Stage 2) using the 11-digit national identify number unique to every citizen of Norway. For Stage 1, we identified MI events registered as principal diagnosis (ICD-10 I21 or ICD-9 410) in a medical department from December 1987 to June 2011. For Stage 2, MI events were registered from the time of enrollment in the Tromsø Study to December 2010, and an adjudication of hospitalized and out-of hospital events was performed by an independent endpoint committee using data from both hospital and out-of hospital journals, autopsy

records, and death certificates (when applicable). Stage 1 controls were selected among HUNT 2 study participants with DNA available (N = 62,816) after excluding those with self-reported and/or hospital diagnosed; MI, MI in 1st or 2nd degree family members, cardiovascular disease, diabetes and hypertension. Stage 2 controls were selected among the Tromsø 4 study participants with DNA available (N = 27,158) after excluding those with cardiovascular disease.

Stage 1 genotyping

We attempted genotyping of 5,771 HUNT individuals at the Genomic Core Facility, Norwegian University of Science and Technology, Norway, using the HumanExome-12v1_A Beadchip (Illumina, CA) and the Infinium HD ultra protocol. The exome array includes 247,870 markers focused on protein-altering variants (nonsynonymous, splicing, and stop-altering) selected from >12,000 exome and genome sequences, including variants associated with complex traits in previous GWAS, HLA tags, ancestry-informative markers, markers for identity-by-descent estimation, and random synonymous SNPs. Details about SNP content and selection strategies can be found at the exome array design webpage.

Quality control of genotypes

Each 96-well plate included both cases and control individuals at random order and one internal blind sample present on each plate. Genotype calling was done on GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina, CA) in combination with zCall version 2.2[37]. Samples below 99% genotype completion rate and samples expressing gender discrepancy or high level of heterozygosity ($-0.05 < F$ (inbreeding coefficient) <0.1) were excluded from further analysis (N = 128, 2.22%). In addition, SNPs that did not meet 99% genotyping threshold or showed deviation from Hardy-Weinberg equilibrium ($P < 1\times10^{-5}$) were removed (13,683 SNPs). After quality assessment, 234,187 SNPs (94.5%) on 5,643 (97.8%) individuals remained for further analysis of which 99,854 (42.6%) were not monomorphic. Of these SNPs, 80,137 were coding variants and 68,615 of these coding variants had a frequency < 5% (Table 1).

Whole-genome sequencing

We performed low-pass whole genome sequencing with exome enrichment (SeqCap EZ Human Exome Library v3.0, Roche NimbleGen Inc., WI) on 76 MI cases and 76 controls at the University of Michigan DNA Sequencing Core using Illumina Hi-Seq 2500. Average sequence coverage was 45× for the capture regions. Concordance for non-reference genotypes between genome sequenced samples (> 10× depth) and exome array genotypes was 99.4% for variants >5% frequency, 99.7% for variants with 1–5% frequency and 99.5% for variants with frequency <1%.

Stage 2 genotyping

Based on preliminary results from Stage 1, we selected for replication single variants with $P < 2\times10^{-5}$, MAF < 10%, and MAC > 5 to capture variants less likely to be covered by GWAS but with enough variants to have a reasonable possibility of replication. We

attempted SNP-based follow-up in 4,666 individuals at the Centre for Integrative Genetics, Ås, Norway, using iPLEX Gold MassARRAY technology (Sequenom, CA). Of the 22 SNPs meeting our initial replication threshold, 4 failed array design and were excluded prior to genotyping.

Statistical analysis

Assuming an additive genetic model, we tested for trait-SNP association with lipid levels using linear regression, with covariates for age, sex, primary phenotype (myocardial infarction status; yes/no), and principal components 1 and 2 as implemented in PLINK[38]. We evaluated the impact of cryptic relatedness using a linear mixed model (EMMAX)[39] and found the results to be highly correlated with linear regression (Pearson $r^2 = 0.85$, $P < 2.2\times10^{-16}$). Lipid values were transformed to rank-based-inverse-normal-residuals to reduce the impact of outliers. Each individual's rank was calculated based on the total number of individuals with available lipid measurements and DNA in the two cohorts (5,643 individuals for Stage 1 and 4,666 individuals for Stage 2). We also performed conditional logistic regression including the subset of lipid GWAS SNP(s) reported by Willer et al.[4] within 500 kb of our lead SNP as covariate(s) to identify additional association signals accounting for the effects at known and newly discovered lipid loci. We estimated the linkage disequilibrium (LD) metric r^2 using 5,643 individuals from Stage 1 who passed genotyping quality control. LD with SNPs not included on the exome array was estimated from 1000 Genomes EUR individuals. For lipids, we declared a single variant association significant if $P < 5\times10^{-8}$, to be consistent with GWAS studies. For myocardial infarction, we declared a single variant association significant if $P < .005$, corresponding to Bonferroni correction for the 10 variants tested in Table 2. Differences in mouse lipid levels were assessed using Mann-Whitney *U*-test, of which $P < 0.05$ was declared significant. Statistical power of association tests were estimated using two samples t-tests model implemented in R package pwr.

Annotation of genetic variants

Variants were annotated as missense, splice, premature stop, read-through, synonymous, or non-coding using ANNOVAR (Version 2012-05-25)[34]. Variant identifiers and chromosomal positions are listed with respect to the hg19 genome build. The following RefSeq accession numbers were used to annotate variants in associated genes: NM_000041.2 (*APOE*); NM_000384.2 (*APOB*); NM_022437.2 (*ABCG8*); NM_022436.2 (*ABCG5*); NM_000078.2 (*CETP*); NM_000236.2 (*LIPC*); NM_006033.2 (*LIPG*); NM_001001524.2 (*TM6SF2*); NM_001166598.1 (*APOA5*); NM_000237.2 (*LPL*) and NM_139314.2 (*ANGPTL4*).

Animal procedures

Overexpression and knockdown of TM6SF2 in mouse—We created adenovirus overexpressing human *TM6SF2* as previously described[40]. In brief, we amplified the coding region sequence corresponding to human *TM6SF2* from human cDNA by high-fidelity *pfu* polymerase (Agilent Technologies, CA). PCR products were sequenced and cloned into pCR®8/GW/TOPO vector (Invitrogen, CA), and then the gene coding sequences were

cloned from Entry vector to the pAd/CMV/V5-DEST vector by LR recombination (Invitrogen, CA). The recombinant adenoviruses were purified by $CsCl_2$ density gradient ultracentrifugation. 8- to 10-week-old male C57BL/6J mice were transduced with purified adenovirus through tail vein injection (0.1 OD per mouse). Overexpression of *TM6SF2* was performed in 8 mice, and compared to 8 mice given a control construct containing only LacZ. Mice were randomly assigned to the experimental (overexpressing *TM6SF2*) or control (overexpressing LacZ) groups. At day 5, mice were fasted overnight and serum total cholesterol, HDL cholesterol, LDL cholesterol and triglyceride were measured with a Cobas Mira Plus chemistry analyzer (Roche, Basel) at the Michigan Diabetes Research and Training Center Chemistry Laboratory, University of Michigan. Mice were purchased from Jackson Laboratory (Bar Harbor, ME), and fed a standard diet (22.5% protein, 11.8% fat, and 52% carbohydrate by mass). Lipids were measured by technicians blinded to the mouse experimental/control status.

To knockdown endogenous TM6SF2 in mice, adenovirus encoding a short hairpin RNA (shRNA) targeting the mouse *Tm6sf2* gene (target sequence: CATCCTTGGTAAATACAGT) driven by the U6 polymerase III promoter was generated using the BLOCK-iT™ U6 RNAi Entry Vector and LR recombination system (Invitrogen, CA). A similar construct was designed to target the LacZ gene as a control. Tail-vein injection was performed as described above with the shRNA vector (0.15 OD per mouse) in 8 mice, and with a control vector containing shLacZ in 8 control mice. At day 6, serum lipid levels were measured following an overnight fast. Lipids were measured by technicians blinded to the mouse experimental/control status.

Endogenous Expression Patterns of *Tm6sf2* mRNA and protein

Northern blotting—Total RNA from tissues was extracted using the Trizol (Invitrogen, CA). Ten μg of total RNA per lane was fractionated on formaldehyde-agarose gels, transferred to nylon membranes, and hybridized to cDNA probes for mouse *TM6SF2* gene. The cDNA probes were synthesized by RT-PCR and labeled with 32P-dCTP as previously described[41]. Autoradiography was scanned, and the signal intensity was analyzed by use of National Institutes of Health ImageJ software.

Western Blotting—Tissue proteins were prepared with lysis buffer (Thermo Fischer Scientific, MA) and a protease inhibitor cocktail (Roche Applied Science, Germany). Protein extracts were resolved in 10% SDS-PAGE gels and electroblotted to PVDF membranes (BioRad, CA). Membranes were blocked in TBST containing 5% (weight/volume) non-fat dry milk at room temperature for 1 hour, and incubated overnight with primary antibody against TM6SF2 (Antibody Verify, NV, Cat No: AAS00444C). After washing, membranes were incubated with an IRDye conjugated secondary antibody (Li-Cor Odyssey, NE) diluted 1:5000 for 1 hour. Blots were scanned and quantitatively analyzed using an image-processing program (Li-Cor Odyssey, NE). For internal control of TM6SF2 tissue distribution, the membranes were stained with Fastgreen (Santa Cruz Biotech, TX). For adenovirus-mediated TM6SF2 overexpression or knockdown in liver, the band density on Western blot was quantitatively analyzed and normalized relative to GAPDH (antibody from Santa Cruz Biotech, CA, Cat No: SC-25778). Western blotting was performed for both

TM6SF2 overexpression (5 mice in each group) and TM6SF2 knockdown experiments (3 mice in each group).

Oil Red O Staining—Mouse Livers were fixed by 4% Paraformaldehyde and then impregnated with 30% sucrose. Frozen liver sections (8 μm) were stained with Oil Red O (Sigma-Aldrich, MO) for 30 min. Sections were then examined under light microscopy. A total of 10 tissue sections were analyzed for each animal.

Alanine aminotransferase activity assay—The liver enzyme activity in mouse serum was determined with an alanine aminotransferase assay kit according to the manufacturer's instructions (Cayman Chemical Company, MI).

Ethics

Both the HUNT and Tromsø Study were conducted according to the principles expressed in the Declaration of Helsinki. Attendance was voluntary, and each participant signed a written informed consent including information on genetic analyses. This study was approved by the Regional Committees for Medical and Health Research Ethics (REC Central), Norway. All animal work was performed in accordance with the University of Michigan Animal Care and Use Committee.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology NTNU), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health. CJW is supported by HL094535 and HL109946. MB is supported by DK062370. YEC is supported by HL068878 and HL117491.

For frequency look-up for the *RNF111* variant, we thank the following: Ruth Loos and Kevin Lu of the BioMe Clinical Care Cohort operated by The Charles Bronfman Institute for Personalized Medicine (IPM) at the Mount Sinai Medical Center (the Mount Sinai IPM Biobank Program is supported by The Andrea and Charles Bronfman Philanthropies); Andres Metspalu of the Estonian Genome Center (the Estonian Biobank data was provided by Evelin Mihailov from the Estonian Genome Center of University of Tartu, Estonia); and André Uitterlinden, Fernando Rivadeneira and Karol Estrada of Erasmus University Rotterdam.

References

1. Go AS, et al. Heart disease and stroke statistics--2013 update: a report from the American Heart Association. Circulation. 2013; 127:e6–e245. [PubMed: 23239837]
2. LipidResearchClinicProgram. The Lipid Research Clinics Coronary Primary Prevention Trial results. II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. JAMA. 1984; 251:365–374. [PubMed: 6361300]
3. Shen L, Peng H, Xu D, Zhao S. The next generation of novel low-density lipoprotein cholesterol-lowering agents: proprotein convertase subtilisin/kexin 9 inhibitors. Pharmacol Res. 2013; 73:27–34. [PubMed: 23578522]
4. Willer CJ, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013; 45:1274–1283. [PubMed: 24097068]
5. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]

6. Pilia G, et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. 2006; 2:e132. [PubMed: 16934002]
7. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]
8. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010; 11:415–425. [PubMed: 20479773]
9. Eichler EE, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450. [PubMed: 20479774]
10. Genomes Project, C. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
11. Genomes Project, C. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
12. Huyghe JR, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet. 2013; 45:197–201. [PubMed: 23263489]
13. Jostins L, Morley KI, Barrett JC. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. Eur J Hum Genet. 2011; 19:662–666. [PubMed: 21364697]
14. Musunuru K, Kathiresan S. HapMap and mapping genes for cardiovascular disease. Circ Cardiovasc Genet. 2008; 1:66–71. [PubMed: 20031544]
15. Sanna S, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet. 2011; 7:e1002198. [PubMed: 21829380]
16. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324:387–389. [PubMed: 19264985]
17. Willer CJ, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008; 40:161–169. [PubMed: 18193043]
18. Kathiresan S, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet. 2008; 40:189–197. [PubMed: 18193044]
19. Krokstad S, et al. Cohort Profile: The HUNT Study, Norway. Int J Epidemiol. 2012
20. Jacobsen BK, Eggen AE, Mathiesen EB, Wilsgaard T, Njolstad I. Cohort profile: the Tromso Study. Int J Epidemiol. 2012; 41:961–967. [PubMed: 21422063]
21. Golledge J, et al. Apolipoprotein E genotype is associated with serum C-reactive protein but not abdominal aortic aneurysm. Atherosclerosis. 2010; 209:487–491. [PubMed: 19818961]
22. Hegele RA, et al. A hepatic lipase gene mutation associated with heritable lipolytic deficiency. J Clin Endocrinol Metab. 1991; 72:730–732. [PubMed: 1671786]
23. Nelis M, et al. Genetic structure of Europeans: a view from the North-East. PLoS One. 2009; 4:e5472. [PubMed: 19424496]
24. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol. 1991; 7:403–422. [PubMed: 1833235]
25. Gottesman O, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013; 15:761–771. [PubMed: 23743551]
26. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
27. Inatani M, et al. Upregulated expression of neurocan, a nervous tissue specific proteoglycan, in transient retinal ischemia. Invest Ophthalmol Vis Sci. 2000; 41:2748–2754. [PubMed: 10937593]
28. Schunkert H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011; 43:333–338. [PubMed: 21378990]
29. Speliotes EK, et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet. 2011; 7:e1001324. [PubMed: 21423719]

30. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–321. [PubMed: 18691683]
31. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13:762–775. [PubMed: 22699862]
32. Cholesterol Treatment Trialists, C. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. Lancet. 2012; 380:581–590. [PubMed: 22607822]
33. Blattmann P, Schuberth C, Pepperkok R, Runz H. RNAi-based functional profiling of loci from blood lipid genome-wide association studies identifies genes with cholesterol-regulatory function. PLoS Genet. 2013; 9:e1003338. [PubMed: 23468663]
34. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164. [PubMed: 20601685]

References in Online Material only

35. Holmen J, et al. The Nord-Trøndelag Health Study 1995–97 (HUNT 2): objectives, contents, methods and participation. Norsk Epidemiologi. 2003:19–32.
36. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem. 1972; 18:499–502. [PubMed: 4337382]
37. Goldstein JI, et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics. 2012; 28:2543–2545. [PubMed: 22843986]
38. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]
39. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354. [PubMed: 20208533]
40. Fan Y, et al. Kruppel-like factor-11, a transcription factor involved in diabetes mellitus, suppresses endothelial cell activation via the nuclear factor-kappaB signaling pathway. Arterioscler Thromb Vasc Biol. 2012; 32:2981–2988. [PubMed: 23042817]
41. Fan Y, et al. Suppression of pro-inflammatory adhesion molecules by PPAR-delta in human vascular endothelial cells. Arterioscler Thromb Vasc Biol. 2008; 28:315–321. [PubMed: 18048767]

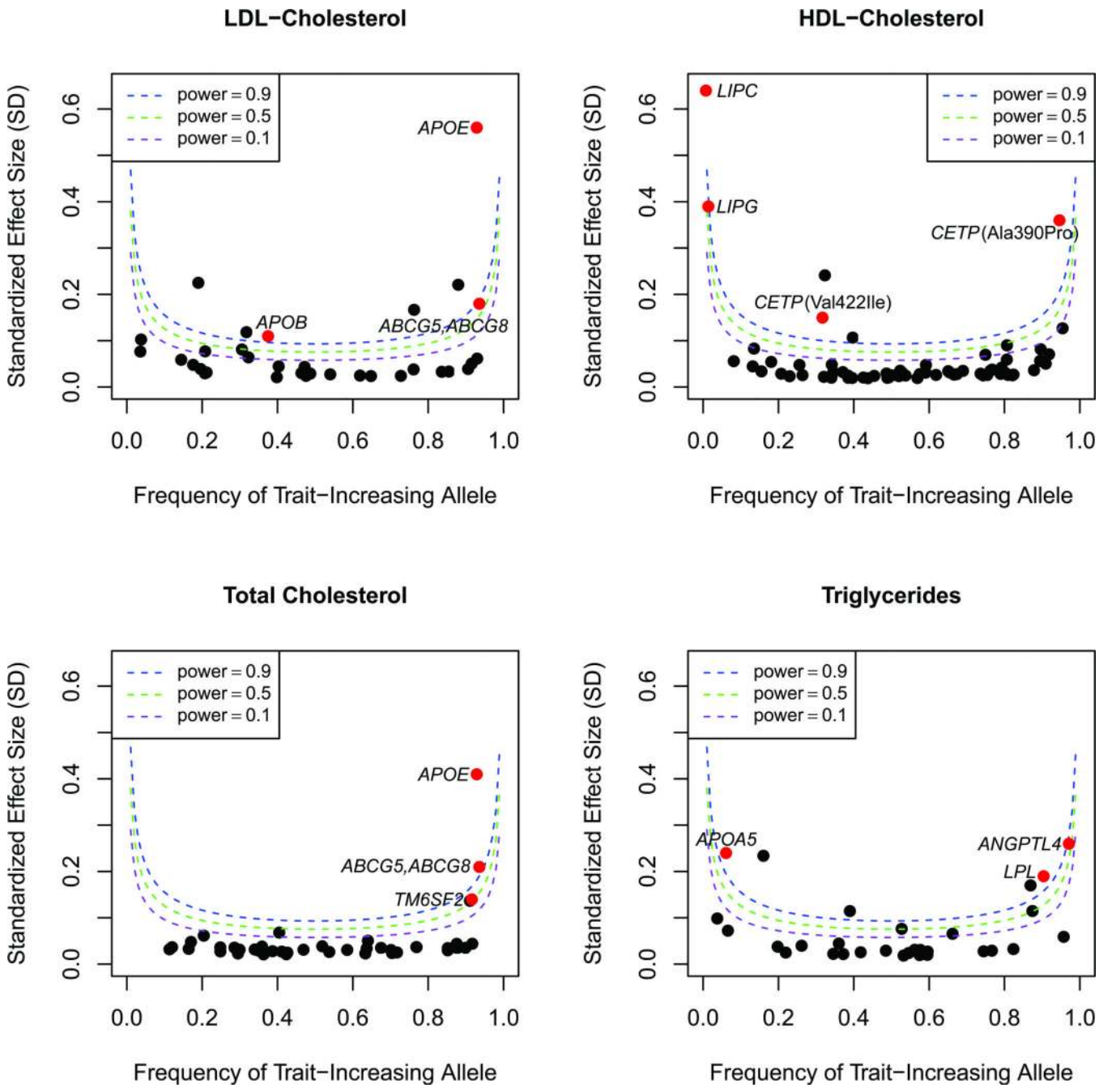

Figure 1. Power estimates for current study compared to estimated effect sizes for coding variants and GWAS index SNPs

This figure shows effect size estimates for the coding variants identified in this study and previous GWAS results. Estimated power curves are shown (as dotted lines) for the minimum standardized effect sizes (in standard deviation units) that could be identified for a given effect-allele frequency with 10% (pink), 50% (green), and 90% (blue) power assuming sample size 10,000 and alpha level 5×10^{-8}. Observed coding variants reaching genome-wide significant association with Stage 1 lipid levels from Table 2 are shown in red.

Previously known lipid marker effect sizes and frequencies as identified by the Global Lipid Genetic Consortium[4] are shown in black.

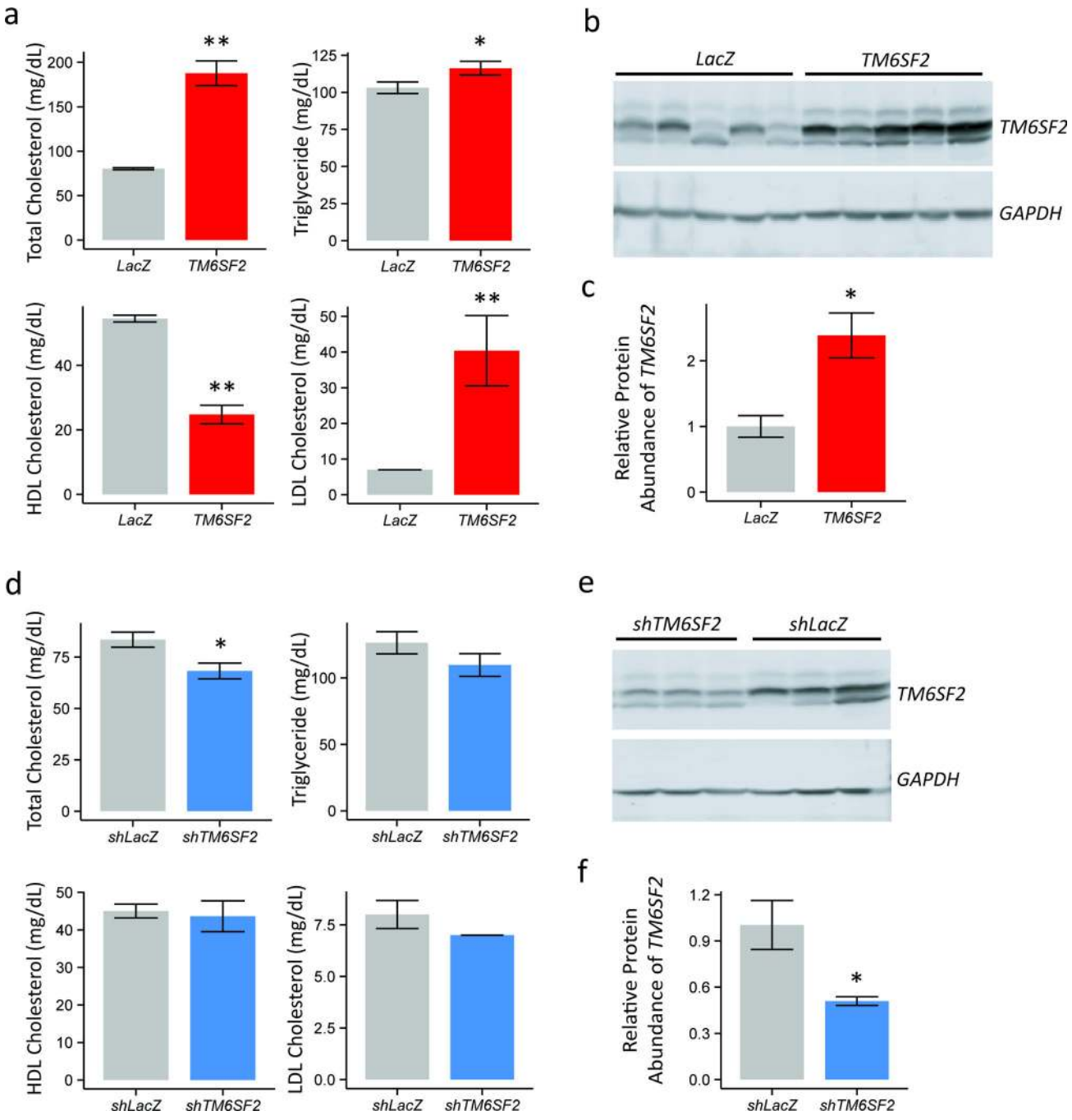

Figure 2. Functional follow-up in C57BL/6J mice implicates *TMF6SF2* in lipid metabolism

This figure demonstrates the functional impact of altered TM6SF2 levels in C57BL/6J mice. (**a**) Transient overexpression of human *TM6SF2* in C57BL/6J mice demonstrate an increase in fasting total cholesterol (2.3 fold, $P = 9\times10^{-4}$), triglyceride levels (1.1 fold, $P = 0.031$), LDL cholesterol levels (5.8 fold, $P = 4\times10^{-4}$) and a decrease in HDL cholesterol levels (0.45 fold, $P = 9\times10^{-4}$) five days after tail-vein injection, compared to LacZ controls (8 mice in each group). (**b**) Western blot demonstrating increased expression of TM6SF2 in mice injected with Ad-*TM6SF2* compared to the Ad-*LacZ* control mice from panel **a**. (**c**)

The mean increase in expression of TM6SF2 was quantitated as 2.4-fold ($P = 0.0066$, 5 mice in each group). (d) Knockdown of *Tm6sf2* by tail-vein injection of adenovirus containing sh*Tm6sf2* in C57BL/6J mice (0.15 OD per mouse) showed decreased levels of fasting total cholesterol (0.81 fold, $P = 0.013$) six days after injection, compared to Ad-shLacZ controls (8 mice in each group). (**e**) Western blot demonstrating decreased expression of TM6SF2 in mice injected with Ad-sh*TM6SF2* compared to control mice. (**f**) The mean decrease in expression of *TM6SF2* was quantitated as 0.51 fold ($P = 0.0375$, 3 mice in each group). *P* values from Mann-Whitney *U*-test. Standard error of the mean indicated. *, $P < 0.05$; **, $P < 0.01$.

TABLE 1

Coverage of coding variation by exome array

Variant type	Frequency	Number of variants genotyped by array‡	Percentage of variants discovered by sequencing§ genotyped by array	
			All samples combined, %	Average for each individual, %
Loss-of-function*	> 5%	174	66.7	65.7
	1–5%	121	72.0	70.3
	6 copies - 1%	875	56.5	57.0
	1–5 copies	945	39.0	39.0
Missense	> 5%	11,348	78.1	78.3
	1–5%	7,976	77.5	77.3
	6 copies - 1%	30,959	71.2	72.1
	1–5 copies	27,739	45.0	45.0
Loss-of-function	≥6 copies	1,170	64.4	65.8
Missense	≥6 copies	50,283	76.0	78.1
Missense + LoF	≥6 copies	51,453	75.9	78.0
GWAS (noncoding)	All frequencies	53/47/32/62†	96.7	-

This table shows the annotation and frequency of variants successfully genotyped. We estimated the coverage of variants in each category by exome-sequencing 152 Norwegians (average 45× coverage of target region). Lipid GWAS markers as published by the Global Lipids Genetics Consortium (2013)[4] were also genotyped. We separated variants with 6 or more copies (in 5,643 individuals) to demonstrate the characteristics of the genetic variants we examined using single variant association tests.

*Loss-of-function (LoF) refers to splice, nonsense, and read-through.

†Number of variants with a primary association with LDL cholesterol, HDL cholesterol, total cholesterol and triglycerides, respectively.

‡Variants in this column were categorized by their frequency in the discovery sample (N=5,643).

§Variants in these two columns were categorized by their frequency in 152 samples as follows: > 5%, 1–5%, 2 copies – 1% and 1 copy.

TABLE 2

Coding variants thatreached genome-wide significant association with lipid levels

					Lipid Association									MI Association	
					Stage 1 (N=5,643)			Stage 2 (N=4,666)			Combined (N=10,309)			Stage 1 or Stages 1+2	
Gene	**Position**	**Rsid**	**Allele**	**Variant**	**MAF, %**	**Effect (s.e.), mg/dL**	***P* value**	**MAF, %**	**Effect (s.e.), mg/dL**	***P* value**	**MAF, %**	**Effect (s.e.), mg/dL**	***P* value**	**OR (95% CI)**	***P* value**
LDL Cholesterol															
APOE	19:45412079	rs7412	T/C	p.Arg176Cys	7.9	−21.2 (1.5)	2.2×10^{-46}	6.1	−28.6 (2.0)	7.3×10^{-48}	7.1	−24.3 (1.2)	2.2×10^{-90}	0.88 (0.79 – 0.98)	0.017
APOB	2: 21263900*	rs1367117	A/G	p.Thr98Ile	37.5	4.8 (0.9)	3.0×10^{-8}	--	--	--	--	--	--	1.02 (0.94 – 1.09)	0.70
ABCG8	2:44066247	rs11887534	C/G	p.Asp19His	6.5	−8.7 (1.7)	1.3×10^{-7}	6.3	−6.5 (2.0)	9.8×10^{-4}	6.4	−7.8 (1.3)	8.0×10^{-10}	0.93 (0.84 – 1.04)	0.22
ABCG5	2:44065090	rs6756629	A/G	p.Arg50Cys	6.5	−8.7 (1.7)	1.5×10^{-7}	6.3	−6.5 (2.0)	9.5×10^{-4}	6.4	−7.8 (1.3)	8.3×10^{-10}	0.93 (0.83 – 1.04)	0.21
HDL Cholesterol															
CETP	16:57015091*	rs5880	C/G	p.Ala390Pro	5.4	−5.4 (0.6)	9.6×10^{-19}	--	--	--	--	--	--	1.16 (0.99 – 1.36)	0.075
CETP	16:57016092	rs5882	G/A	p.Val422Ile	31.7	2.2 (0.3)	2.5×10^{-13}	--	--	--	--	--	--	0.94 (0.87 – 1.02)	0.13
LIPC	15:58855748	rs113298164	T/C	p.Thr405Met	0.7	10.5 (1.7)	5.6×10^{-10}	0.8	8.9 (1.9)	1.9×10^{-6}	0.75	9.6 (1.2)	1.4×10^{-14}	0.91 (0.66 – 1.25)	0.56
LIPG	18:47109955	rs77960347	G/A	p.Asn396Ser	1.6	5.7 (1.1)	4.0×10^{-7}	1.2	6.1 (1.5)	3.1×10^{-5}	1.4	5.8 (0.9)	3.6×10^{-11}	0.87 (0.69 – 1.09)	0.24
Total Cholesterol															
APOE	19:45412079	rs7412	T/C	p.Arg176Cys	7.9	−18.0 (1.6)	1.4×10^{-27}	6.1	−20.9 (2.1)	3.3×10^{-23}	7.1	−19.4 (1.3)	2.1×10^{-48}	0.88 (0.79 – 0.98)	0.017
ABCG8	2:44066247	rs11887534	C/G	p.Asp19His	6.5	−10.0 (1.8)	4.4×10^{-8}	6.3	−9.5 (2.1)	4.3×10^{-6}	6.4	−10.0 (1.4)	5.7×10^{-13}	0.93 (0.84 – 1.04)	0.22
ABCG5	2:44065090	rs6756629	A/G	p.Arg50Cys	6.5	−10.0 (1.8)	4.7×10^{-8}	6.3	−9.5 (2.1)	5.3×10^{-6}	6.4	−10.0 (1.4)	9.3×10^{-13}	0.93 (0.83 – 1.04)	0.21
TM6SF2	19:19379549	rs58542926	T/C	p.Glu167Lys	8.9	−8.1 (1.6)	2.7×10^{-7}	7.9	−4.7 (1.9)	0.010	8.4	−6.6 (1.2)	4.4×10^{-8}	0.87 (0.79 – 0.95)	5.0×10^{-3}
Triglyceride Levels															
APOA5	11:116662407*	rs3135506	C/G	p.Ser19Trp	6.1	24.8 (4.0)	4.1×10^{-10}	--	--	--	--	--	--	0.92 (0.80 – 1.08)	0.33
LPL	8:19819724*	rs328	G/C	p.Ser474X	9.6	−19.7 (3.3)	1.3×10^{-9}	--	--	--	--	--	--	0.96 (0.84 – 1.08)	0.47
ANGPTL4	19:8429323	rs116843064	A/G	p.Glu40Lys	3.2	−32.1 (5.5)	8.2×10^{-9}	2.5	−18.6 (7.5)	0.014	2.9	−26.9 (4.6)	4.2×10^{-9}	0.78 (0.66 – 0.92)	3.1×10^{-3}

Variants that reached significant association ($P < 5\times10^{-8}$) in stage 1 or stage 1+2 are listed, with positions given for hg19, [chromosome]:[base pair position]. Alleles are reported as minor/major on the forward strand of the reference genome. The "Variant" column lists predicted changes in amino acid sequence using ANNOVAR[34]. Effects are reported for the minor allele in mg/dL (±s.e.). LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; MAF, minor allele frequency; s.e, standard error; MI, myocardial infarction; OR, odds ratio. MI association results were determined from 2,350 medical-record confirmed MI cases and 2,318 healthy controls in our Stage 1 samples. At the APOA5 locus, the p.Ser19Trp variant showed weaker evidence for association with triglyceride levels than the GWAS index SNP, indicating that the coding variant may not be functional.

Author Manuscript

Author Manuscript

Author Manuscript

* variants intended for follow-up which failed array design.