

Published in final edited form as:

*Nat Genet.* 2015 June ; 47(6): 589–597. doi:10.1038/ng.3300.

## The impact of low-frequency and rare variants on lipid levels

*A full list of authors and affiliations appears at the end of the article.*

### Abstract

Using a genome-wide screen of 9.6 million genetic variants achieved through 1000 Genomes imputation in 62,166 samples, we identify association to lipids in 93 loci including 79 previously identified loci with new lead-SNPs, 10 new loci, 15 loci with a low-frequency and 10 loci with missense lead-SNPs, and, 2 loci with an accumulation of rare variants. In six loci, SNPs with established function in lipid genetics (*CELSR2*, *GCKR*, *LIPC*, and *APOE*), or candidate missense mutations with predicted damaging function (*CD300LG* and *TM6SF2*), explained the locus associations. The low-frequency variants increased the proportion of variance explained, particularly for LDL-C and TC. Altogether, our results highlight the impact of low-frequency variants in complex traits and show that imputation offers a cost-effective alternative to re-sequencing.

---

Genome-wide association (GWA) studies have been successful in identifying genetic loci associated with complex diseases and traits. Due to the design of genotyping arrays, most of the associated variants have been common in population samples. While thousands of loci have been associated with complex diseases and traits, they so far typically explain only a fraction of the heritability<sup>1</sup>.

It has now become possible to search for associations with variants that are less frequent than in previous GWA studies by the analysis of large numbers of samples using whole genome or exome sequencing approaches. However, costs have so far limited the possibility for sequencing of tens of thousands of samples likely needed to detect significant associations for low-frequency variants.

Stochastic imputation to individuals genotyped using genotyping arrays in large enough samples offers an alternative and cost-effective design to study associations of low-frequency and rare variants at a genome-wide level. GWA studies of circulating lipids have been highly successful in identifying loci with common variants with small effects<sup>2,3</sup>. In

---

**Corresponding author** Prof. Samuli Ripatti Institute for Molecular Medicine Finland, FIMM P.O. Box 20, 00014 University of Helsinki, Finland samuli.ripatti@fimm.fi.

#### Contributions

IS, MH, RM, APS, AnuM, VasL, LM, TF, EI, OK, VP, CML, UT, AaP, MIM, APM, IP, and SR designed and performed experiments, analysed data and wrote the paper. BM, ST, JoK, MatP, JuK, HJW, EI, JPM, THP and LF performed follow-up experiments and analysed the data. GT, SH, JJH, AI, CL, MarB, TE, JSR, CPN, CW and SG analysed cohort specific data. HS, JE, NJD, JaK, LL, CG, AndM, PES, LG, CMvanD, JGE, AJ, VeS, DIB, CP, OTR, MRJ, and KS designed cohort specific experiments. MatB, AJMdeC, EJdeG, JD, HG, AH, ASH, CH, JJHD, EH, MK, LCK, TL, ValL, PKEM, EM, MMN, NLP, BWJHP, MarP, AnP, JR, JHS, VaS, MDT, NT, EMvanL, JSV, SMW, and GW performed cohort specific experiments and analysed cohort specific data. All authors contributed to the research and reviewed the manuscript.

#### Competing financial interests

UT, GT, VaS and KS are employed by deCODE Genetics/Amgen inc.

previous large scale GWA studies, 157 loci have been shown to associate with lipids<sup>2,3</sup>, but the strongest associations have almost exclusively been reported with common variants (minor allele frequency, MAF > 5%) in European datasets due to the study designs.

In contrast, previously published variants known to cause Mendelian forms of dyslipidemic syndromes and, more broadly, variants with known functional impact on lipids (FL SNPs) typically have low MAF ( $\leq 5\%$ ). While there are almost 40 loci where both FL SNPs and common SNPs, implicated in GWA studies, reside, it is often not known if these associations are driven by the same underlying haplotypes and if the Mendelian variants explain the association in population samples.

We sought to evaluate the impact of common (MAF > 5%), low-frequency ( $0.5\% < \text{MAF} \leq 5\%$ ) and rare (MAF  $\leq 0.5\%$ ) genetic variants on circulating blood lipids in up to 62,166 European samples by imputing variants into the GWA cohorts using the sequence-based 1000 Genomes reference panel<sup>4</sup> (Phase I interim release, June 2011). We aimed to answer the following questions: 1) what is the role of low-frequency and the burden of rare variants in the established lipid loci, 2) can a dense set of markers from 1000 Genomes-based imputation help to identify additional loci undetected in previous studies focused largely on common variants imputed up to less dense reference panels from the HapMap-project, 3) how do low-frequency and FL variants contribute to the overall trait variance compared to common variants.

## RESULTS

### Study Overview

To understand the contribution of low-frequency and rare genetic variation to circulating lipid concentrations, we undertook genome-wide imputation and association analysis in up to 62,166 individuals across 22 GWA cohorts of European ancestry. Within each cohort, we performed sex-stratified inverse-rank normalisation of high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG) and total cholesterol (TC), after adjustment of each trait for age, age<sup>2</sup>, and study-specific covariates, including principal components to account for population structure. Case-control studies were further sub-divided according to original data selection disease status. Each cohort GWA genotype scaffold was imputed at up to ~37.4 million autosomal variants from the 1000 Genomes Project multi-ethnic reference panel<sup>4</sup> (Phase I interim release, June 2011). Across a subset of studies, ~98% and ~95% of variants present in the reference panel with  $1\% < \text{MAF} \leq 5\%$  and  $0.5\% < \text{MAF} \leq 1\%$ , respectively, were well imputed, defined here by an IMPUTE<sup>5,6</sup> info score of at least 0.4 (Supplementary Table 1). However, as expected, imputation of rare variants (MAF  $\leq 0.5\%$ ) proved more difficult, although ~65% of the rare variants polymorphic in the reference dataset were well imputed across the same subset of studies.

### Genome-wide screen for single variant associations

We first tested for association of over 9.6 million genotyped or successfully imputed SNPs, enabled by the 1000 Genomes imputation, with circulating HDL-C, LDL-C, TG and TC

levels. Overall, we detected 93 loci with genome-wide significant association (Supplementary Figure 1) to one or more lipid traits ( $p$ -value  $< 5 \times 10^{-8}$ ), of which 10 loci have not been associated to lipids before (Table 1, Supplementary Figures 2A-J and Supplementary Figures 3A-J). Out of the 83 previously established lipid loci, 79 had a novel lead-SNP for at least one lipid trait in our analysis (Supplementary Table 2). In 34 out of the 79 loci the linkage disequilibrium (LD) is  $r^2 \leq 40\%$  (15 loci with  $r^2 \leq 5\%$ ) and in 56 loci the newly identified variant has not been present in HapMap2 imputation reference set used in previous studies. In 11 loci the novel lead-SNP had MAF  $\leq 5\%$  and an average effect size of 0.18 (in standard deviation, s.d., units) compared to the average effect size of 0.05 for the previously established common lead-SNP estimated in a cohort independent of the discovery scan to avoid bias due to the winner's curse ( $N = 5,119$ , Figures 1A-B). These include well-known lipid gene *LPA* for LDL-C (rs186696265, MAF = 0.8%, effect size = 0.26,  $p$ -value =  $4.4 \times 10^{-14}$ ,  $r^2 = 0.1\%$ ). In addition we observed high effect lead-SNPs in *PCSK9* for LDL-C (rs11591147, MAF = 1.9%, effect size = 0.53,  $p$ -value =  $2.2 \times 10^{-92}$  and  $r^2 = 0.9\%$ ) and *APOE* for TC (rs7412, MAF = 7.1%, effect size = 0.41,  $p$ -value =  $7.5 \times 10^{-239}$  and  $r^2 = 1.6\%$ ) that were highlighted already in the Global Lipids Genetics Consortium fine-mapping analyses<sup>3</sup>.

Using a formal conditional analysis, in *MAFB* locus, the new low-frequency lead-SNP with large effect size (effect size  $> 0.2$  and MAF  $\leq 5\%$ ) explained the association of the previously identified lead-SNP in seven population cohorts ( $N = 12,834$ ) though the linkage disequilibrium between the variants was less than  $r^2 = 5\%$  (Figures 1A-B). Additionally, there were 7 loci with two or more association lead-SNPs over 1Mb apart and with  $r^2 < 5\%$ , but in all cases the individual level formal conditional analyses showed that the associations were completely explained by the known lipid SNPs in the regions (*ZCCHC11*, *TMEM48* and *PPAP2B* associations explained by rs11591147 in *PCSK9* locus, *OR*-cluster association by rs7395581 in *LRP4-MADD* locus, *CCDC79* association by rs73591976 in *LCAT-RANBP10* locus, and *PSG9* and *IRF2BP1* associations by rs7412 in *APOE* locus).

In five of the 79 loci, the lead-SNP was a missense variant pointing to either a well-established causal gene (*ANGPTL4*, *APOE*, *PCSK9* and *CILP2*) or to a new candidate gene (*ABCA6/8*). The *APOE* lead-SNP for TC, rs7412 (Arg176Cys, MAF = 7.1%,  $r^2 = 0.7\%$ ) has been shown to associate with recessive familial type III hyperlipoproteinemia<sup>7,8</sup> and the *PCSK9* lead-SNP for LDL, rs11591147 (Arg46Leu, MAF = 1.9%,  $r^2 = 0.9\%$ ), with extreme LDL-C values<sup>9</sup>. In the *ANGPTL4* locus, the lead-SNP in our GWA data is a predicted damaging missense variant, rs116843064 (Glu40Lys, MAF = 3.0%) with  $r^2 = 1.8\%$  with the previously associated common lead-SNP. The missense variant is associated with TG and HDL-C, and has previously been associated with extreme TG values<sup>10</sup>. The *CILP2* lead-SNP, rs58542926 (Glu167Lys in *TM6SF2* gene, MAF = 7.8%,  $r^2 = 98\%$ ), was associated to TC, myocardial infarction risk, and nonalcoholic fatty liver disease in two papers appearing while revising this manuscript<sup>11,12</sup>. Our new lead-SNP in *ABCA6/8* locus, rs77542162 (Cys1359Arg in *ABCA6* gene, MAF = 2.0%,  $r^2 = 0.6\%$ ) associates with LDL-C and TC ( $p$ -value =  $1.6 \times 10^{-18}$  and  $p$ -value =  $1.9 \times 10^{-13}$ , respectively).

In the genome-wide screening we identified 10 loci that have not previously been associated to lipids (near *PROX1*, *CEP68*, *PRKAG3*, *ADAMTS3*, *MTHFD2L*, *GPR85*, *RMI2*,

TM4SF5, GATA6 and *ZNF274*), with 4 having a low-frequency variant (MAF < 5%) as the lead-SNP (lowest MAF = 0.7% rs182616603 in *MTHFD2L* locus; Table 1). All except one of the lead-SNPs have not been surveyed in the previous GWA studies based on HapMap 2 imputation. The one lead-SNP that has been present in the HapMap 2 imputation references is in the *PROX1* 5'UTR (rs340839 associated with TG, p-value =  $4.4 \times 10^{-12}$ ) and is correlated with a marker previously associated with fasting glucose and type 2 diabetes<sup>13</sup> (rs340874,  $r^2 = 74.7\%$ ). The lead-SNP in the HDL associated *PRKAG3* locus is located upstream to the gene, close to a transcription factor binding site. *PRKAG3* is a regulatory subunit of the AMP-activated protein kinase (AMPK), which has previously been shown to regulate lipid homeostasis<sup>14</sup>.

### The role of variants with known functional impact on lipids in the general population

In 8 loci (*PCSK9*, *CELSR2-SORT1*, *GCKR*, *HLA*-region, *LPL*, *LIPC*, *CETP* and *APOE*), we tested if the variants known to cause Mendelian forms of dyslipidemic syndromes and, more broadly, with known functional impact on lipids, also explained the associations of the common lipid SNPs. These FL SNPs were identified through the Online Mendelian Inheritance in Man database search (OMIM; [www.omim.org](http://www.omim.org)) and confirmed through literature, and SNPs previously reported to affect gene transcription or translation in cellular and/or animal models were taken forward into conditional analyses in seven population cohorts ( $N = 12,834$ ; Supplementary Figure 4, Supplementary Table 3 and Supplementary Table 3).

The FL SNPs explained the lead-SNP association (with p-value <  $5 \times 10^{-8}$  and conditional p-value > 0.01 for the lead-SNP) in four of the 8 loci (*CELSR2-SORT1*, *GCKR*, *APOE* and *LIPC*; Table 2, Supplementary Figures 5A-G). In *GCKR* and *APOE* loci, the lead-SNPs of our GWA screen were FL SNPs (rs1260326<sup>15</sup>; Pro446Leu and rs7412<sup>7,8</sup>; Arg158Cys, for *GCKR* and *APOE*, respectively). In the *GCKR* locus, rs1260326 explained the population-level association. Similarly, in *APOE* locus, the two FL SNPs rs7412 and rs429358<sup>16</sup> (Cys112Arg) defining the *APOE* isoforms  $\epsilon 2$ ,  $\epsilon 3$  and  $\epsilon 4$ <sup>17</sup> explained the association (Supplementary Figures 5D and 5E). The *LIPC* association was explained by rs1800588<sup>18</sup> (-514C-T, MAF = 25.1%) and rs113298164<sup>19</sup> (Thr383Met, MAF = 1.4%) for TC and TG (Supplementary Figures 5F and 5G) but not for HDL-C (Supplementary Figure 5H). All results for the conditional analyses are presented in Supplementary Tables 5A-D.

### Search for novel functional candidate SNPs

We then searched for potential candidate causal SNPs in the lipid-associated (157 established and 10 novel) loci with a similar predicted function to well-characterized FL SNPs. We identified possible functional variants in four loci without known functional variants at the time of analysis (*MLXILP*, *LRP4-MADD*, *SOST-DUSP3* and *CILP2*), and tested whether the identified variants explained the significant association seen in the locus (Supplementary Table 6). The results of the conditional regression analyses for these four loci are presented in Table 3 and Supplementary Figures 6A-F. In the *SOST-DUSP3* and *CILP2* loci, the functional candidates explained the genome-wide associations of the lead-SNPs in the region in the test set (in both loci, conditional p-value > 0.01). In the *SOST-DUSP3* locus (Figure 2A), a single low-frequency deleterious missense variant, rs72836561

(Arg82Cys, MAF = 2.7%, p-value =  $1.36 \times 10^{-8}$ , effect size = 0.23) in the *CD300LG* gene, explained the whole regional association indicating *CD300LG* as a likely candidate gene for TG in the locus. The same variant has also recently been shown to associate with HDL-C and with fasting serum triacylglycerol in exome-wide association studies<sup>20,21</sup>.

In the *CILP2* locus for LDL-C, TC and TG, two independent missense variants ( $r^2 = 0$ ) in the *TM6SF2* gene, a deleterious missense variant rs187429064 (MAF = 3.6%, Leu156Pro; for TC effect size = -0.25 and p-value =  $2.03 \times 10^{-11}$ ) and a probably damaging missense variant rs58542926 (MAF = 6.3%, Glu167Lys; for TC effect size = -0.18 and p-value =  $6.47 \times 10^{-12}$ ), explained the lead-SNP association for LDL-C, TC and TG (Table 3, Supplementary Figures 6D-F, and Figure 2B and Supplementary Figure 7 illustrates the result of the conditional analysis for TC).

### Biological profiling of *CD300LG* and *TM6SF2* genes in lipid metabolism

CD300LG (CD300 Molecule-Like Family Member G; also called nepmucin) is a type I cell surface glycoprotein that contains a single immunoglobulin (Ig) V-like domain<sup>22</sup> and plays a role in lymphocyte binding and transmigration<sup>23</sup>. The predicted damaging mutation (Arg82Cys) in our TG/HDL-C-associated variant rs72836561 is located in the Ig domain of *CD300LG*, which binds to lymphocytes. *CD300LG* is expressed in the vascular endothelial cells of various tissues, and is located both at the plasma membrane and intracellular vesicles<sup>23,24</sup>. While CD300 family members have been demonstrated to bind lipids<sup>25</sup>, the function of *CD300LG* in lipid metabolism has not been studied. TM6SF2 (Transmembrane 6 Superfamily Member 2)<sup>26</sup>, is a multi-pass membrane protein, in which the predicted deleterious missense mutation (rs187429064; Leu156Pro) locates to the predicted 5<sup>th</sup> transmembrane domain, and the probably damaging missense mutation (rs58542926; Glu167Lys) in the exposed non-transmembrane domain. *TM6SF2* gene has been shown to localize to endoplasmic reticulum (ER) compartment/ER-Golgi intermediate compartment (ERGIC) and influence TG secretion in liver cells<sup>27</sup>. Additionally, the Glu167Lys missense mutation was shown to alter serum lipid profiles in humans and the knockdown of TM6SF2 in mice was shown to lead to increased liver triglyceride content and decreased very-low-density lipoprotein (VLDL) secretion<sup>11,12</sup>.

We further characterized the two genes by using the Gene-Network database<sup>28</sup> (<http://genenetwork.nl/genenetwork>, see **online Methods** for details) for tissue specific expression, pathway analysis, and prediction of mice knockout phenotype, based on Mouse Genome Informatics (MGI; <http://www.informatics.jax.org>)<sup>29</sup>. We found that *CD300LG* gene is co-expressed with genes where knockout increases circulating very low-density lipoprotein (VLDL) particle levels in mice (prediction p-value =  $1.4 \times 10^{-9}$ ), in line with our phenotype of higher TG levels in humans carrying the deleterious missense variant of *CD300LG*. For *TM6SF2*, the MGI-based predictions, using co-expression of genes, show abnormal lipid levels (decreased LDL-C: prediction p-value =  $8.6 \times 10^{-19}$ , decreased VLDL: prediction p-value =  $2.5 \times 10^{-29}$  and decreased TC: prediction p-value =  $6.3 \times 10^{-24}$ ) amongst the most highly significant predictions, in line with the recent publications and our association results. All associated MGI-based knockout predictions (p-value <  $1e-6$ ) are shown in the

Supplementary Tables 7A-B and lists of genes with same and stronger MGI-based predictions can be found in the Supplementary Table 8.

Both genes were found to be amongst the most highly expressed genes in tissues important for lipid absorption and/or metabolism based on the analysis using the Gene Network database (Supplementary Tables 9A-B). *CD300LG* is highly expressed in muscles, plasma, and adipose tissue and *TM6SF2* in liver, plasma, and intestines. Furthermore, based on the gene expression network analysis, *TM6SF2* likely interacts with proteins involved in intestinal absorption (Supplementary Table 10), and it is most highly predicted to function as lipid transporter (p-value =  $1.05 \times 10^{-14}$ , prediction is based on co-expressed genes, Supplementary Table 11).

### Contribution of low-frequency variants to population lipid variation

We estimated the proportion of the variance of lipid traits explained by variants in the 157 previously established and 10 novel loci in an additional cohort of 5,119 individuals from Finrisk cohort (FRCOREXOME9702), not included in our discovery meta-analysis. The lead-SNPs from all three GWA screens (Teslovich et al.<sup>2</sup>, Willer et al.<sup>3</sup> and this study) together with the FL SNPs and new functional candidate SNPs were divided into two groups based on their allele frequency in the FRCOREXOME9702 dataset. Common SNPs explained 8.2% (TG), 11.9% (HDL-C), 16.3% (LDL-C), and 16.2% (TC) of the variance in lipid levels (Figure 3). Together with the low-frequency variants we now explain 9.3%, 12.8%, 19.5% and 18.8% of the variance in TG, HDL-C, LDL-C and TC, respectively.

We also compared the contribution of our SNPs to the additive genetic variance estimated by a linear mixed model (LMM) applied to 10,472 individuals from six Finnish GWA cohorts (**Online Methods**) with those obtained from a large twin study<sup>30</sup>. The narrow sense heritability estimates from the twin study were 40%, 51%, 51% and 33% and the mixed linear model estimates derived from the Finnish subset were 26%, 29%, 27% and 19% (for HDL-C, LDL-C, TC and TG, respectively. See **Online Methods** for details). We estimate that the SNP set explain at least 28.1%, 32.0%, 38.2% and 36.7% (narrow sense heritability) and at most 48.9%, 49.2%, 67.2% and 69.6% (LMM heritability estimate) of the additive genetic variance of TG, HDL-C, LDL-C and TC, respectively.

### Gene-based association analysis

To complement the single-variant tests for low-frequency variation we used GRANVIL<sup>31</sup> to test for association of each lipid trait with accumulations of minor alleles (“mutational load”) at well imputed rare variants within genes in a subset of 30,463 individuals from 15 cohorts (Online Methods, Supplementary Table 1).

We observed genome-wide significant evidence of association (p-value <  $1.7 \times 10^{-6}$ , Bonferroni correction for 30,000 genes) of HDL-C with the mutational load of rare non-synonymous variants in *LIPC* (p-value =  $2.1 \times 10^{-7}$ , mean MAF = 0.26%, Supplementary Figure 8). To further investigate the relationship between gene-based and single SNP association signals at this locus, we performed conditional analysis, adjusting the effect of the mutational load for the lead-SNP in our study (rs261291). The association of HDL-C

with rare non-synonymous variants in *LIPC* remains relatively unchanged (conditional p-value =  $3.6 \times 10^{-6}$ ), suggesting that the mutational load of the gene is independent of the GWA signal at this locus.

We identified two genes for which the mutational load of rare variants (irrespective of annotation) was associated with TG at genome-wide significance, both mapping to the APO-cluster: *ZNF259* (p-value =  $1.5 \times 10^{-11}$ , mean MAF = 0.25%) and *APOA5* (p-value =  $5.0 \times 10^{-8}$ , mean MAF = 0.24%). Conditional analyses, adjusting for the association lead-SNP (rs964184) at the APO-cluster, reduced the strength of association of rare variants in both *ZNF259* and *APOA5* with TG, but could not fully explain the effect of the mutational load of these genes (Supplementary Table 12). As *ZNF259* and *APOA5* map within 2kb of each other, we further investigated the impact of LD on the association signal at the region with conditional analyses adjusting for the mutational load of each gene for that at the other (**Online Methods**). The strength of association of both genes was reduced, but not fully attenuated, after adjusting for the effect of the other (*ZNF259* conditional p-value =  $1.4 \times 10^{-5}$ ; *APOA5* conditional p-value =  $6.3 \times 10^{-4}$ ), suggesting the effects of rare variants in these two genes to be only partially correlated with each other.

## DISCUSSION

Using 1000 Genomes imputed data with a dense SNP set, we were able to impute 9.6M common and low-frequency SNPs with good quality in 62,166 European samples. With GWA meta-analysis on these data, we identified 10 novel loci associated to blood lipids and new lead-SNPs in 79 previously known lipid loci. In 11 previously known loci, the new lead-SNP had a minor allele frequency  $\leq 5\%$  and, on average the newly identified low-frequency variants showed 3.6 times larger effect size compared to the corresponding lead-SNP in previous meta-analysis studies. Moreover, in four of the ten novel loci, the lead-SNPs were low-frequency variants.

Our association results reveal that low-frequency variants have a much larger contribution to lipid variation in the general population than has previously been shown<sup>2,3</sup>. In several cases, the association that has previously been tagged by common variants is now led by variants with 0.5 – 5% allele frequency and larger effect sizes. The large effect sizes also show in the population lipid variance explained, where low-frequency variants add 3.2% to the LDL-C variance explained when adding on top of the common variants identified in previous reports or in our study, even though there are relatively few carriers of low-frequency variants in the general population.

While GWA studies have typically identified associations to lipid levels in cohorts with normal population variation, the known functional variants, some causing Mendelian forms of lipid syndromes and others changing the protein structure or disturbing the gene transcription, often been identified in patients and families with extreme lipid values. We found four regions where the population-level association was explained by known Mendelian and/or functional SNPs, suggesting that the effects of FL SNPs seem to generalize to European samples with normal lipid variation. Taken together, the successfully

imputed and tested functional SNPs together with the new functional candidate variants explained 2.2 – 6.7% of the lipid variation in the population level.

As the FL SNPs explained the population-level association in four of the studied eight loci through LD-structure, we reversed this connection to identify potential functional candidate genes through SNPs with similar functional profile to the FL SNPs in lipid loci with no previous strong functional candidates. Using this strategy we identified two loci where missense variants with predicted damaging or deleterious functions explained the lead-SNP associations from the GWA meta-analysis, thus, together with previous evidence, supporting the role of *CD300LG* (TGs) and *TM6SF2* (TC, LDL-C, TGs) in lipid metabolism together with evidence from gene network analysis, gene expression correlations, predicted functions in mice, and expression patterns across organs each suggest potential links to lipid metabolism. *TM6SF2* was recently listed among genes potentially affecting LDL-C uptake in a recent siRNA screen focused on cellular lipid phenotypes within previously published blood lipid-associated GWA-loci.<sup>32</sup> Additionally, two reports showing strong evidence for one of the two *TM6SF2* missense variants, Glu167Lys, on VLDL and TG metabolism were published during this study<sup>11,12</sup>. However, in our data, this mutation alone does not explain the whole regional association, but together with a second missense mutation, with lower MAF and larger effect, the association was explained. Overall, our results reinforce the importance of *CD300LG* and *TM6SF2* for blood lipid levels in the general population.

In two established GWAS loci with common lead SNPs, our analyses revealed associations of the mutational load of rare variants with lipids. The association of HDL-C with rare variants in *LIPC* has been previously reported<sup>33</sup>, and we also demonstrate that this signal is independent of the common lead GWAS SNP at this locus. We identified association on TG for the accumulation of *APOA5* rare variants as significant, but conditional analysis on the GWAs lead-SNP suggested that the single variant and gene-based associations are partially correlated. However, the GWA lead-SNP alone was not sufficient to fully explain the gene-based signal. An excess of minor alleles in *APOA5* has previously been associated with hypertriglyceridemia<sup>34</sup>, but we report here an impact of this gene on TG at a population level. Although imputation enables recovery of ~65% of rare variants that are present in the 1000 Genomes haplotypes, many will not be represented in the reference panel. Re-sequencing in large sample sizes will be required to fully elucidate the role of rare variation at these GWA loci on HDL-C and TG and to inform functional studies to determine the underlying mechanisms mediated through these genes for the regulation of lipids.

In addition to the 93 loci identified, there were seven loci showing two or more association signals that were more than 1Mb distance from each other and the linkage disequilibrium between the lead SNPs were small ( $r^2 \leq 0.05$ ). However, in formal conditional analyses of these loci using individual level data the most strongly associated SNPs in the locus explained also the other associations, even over a physical distance of 1Mb or more, or low level of LD. As these observations were only revealed after careful conditional testing of individual level data, they also highlight how challenging it is to interpret the association patterns using only summary level results on single SNP analyses.



There are some potential limitations to our genetic study. Although we used a dense sequence-based global imputation panel, it does not cover all low-frequency and rare variants in Europe. Similarly, although the imputation reference set included a large number of low-frequency SNPs and other variants with known functional impact on lipids, some were either missing from the panel or they were not polymorphic in our test sets of seven Finnish cohorts. Therefore we are likely missing some additional effects in our data. As more individuals are being sequenced and made available as imputation reference panels, more variants can also be imputed with high confidence and tested for associations.

In conclusion, our study shows that low-frequency variants contribute significantly to population variance in lipid levels. The variants known to cause Mendelian forms of lipid syndromes and variants with known functional effects on lipid levels explain the common variant association in overlapping loci revealing a similar role of these variants in extreme patient series and in general populations. In addition, we found 10 new lipid loci for further investigations and for two previously known lipid loci we identified new candidate missense variants with predicted damaging function. When combining all the accumulated genetic evidence, we could explain up to 19.5% of the trait lipid variation. By considering the aggregate effects of rare variants within genes, we identified three transcripts associated with lipids in already established GWA loci that could not be fully explained by the common lead-SNPs reported in this study. Together, these observations show the important role of low-frequency functional SNPs in lipid level variation in the general population and present new therapeutic opportunities for treating dyslipidemias and preventing cardiovascular diseases. They also highlight that imputation is a cost-effective approach to assessing association with low-frequency and rare variants, without the need for costly re-sequencing experiments.

## ONLINE METHODS

### Genotype quality control and imputation

Before imputation all cohorts (see Supplementary Note for cohort information) went through a quality control (QC) pipeline with the following criteria: samples with genotype call rate < 95%, sex discrepancies, excess heterozygosity and cryptic relatedness were removed. Additionally, ethnic outliers and MDS outliers were excluded. SNPs with minor allele frequency (MAF) < 1%, call rate < 95% (or < 99% if the SNP has MAF < 5%), failure of the Hardy-Weinberg Equilibrium (HWE) exact test (precise threshold depending on study) and sex chromosome SNPs were removed. Genotyping platforms, study-specific QC criteria and other details are presented in Supplementary Table 13. The imputation of the datasets was performed using IMPUTE v2.0<sup>5,6</sup> (unless stated otherwise) with 1000 Genomes June 2011 imputation reference panel with 2,188 haplotypes<sup>4</sup> ([www.1000genomes.org](http://www.1000genomes.org)).

### Phenotype measures

All four lipids, HDL-C, LDL-C, TC and TG were measured using basic enzymatic methods. Summary statistics of phenotypes in each cohort are presented in Supplementary Table 14. Individuals with lipid-lowering medication were excluded and measures deviating more than 5 s.d. were set to missing. All four phenotypes were adjusted for age, age<sup>2</sup> and the first three

genetic principal components. Principal components were derived from the GWA data using principal component analysis for the IBS sharing matrix for each study separately<sup>35</sup>. Both the removal of outliers and the adjustments were done for males and females separately in each of the studies for all four traits. The residuals resulting from the adjustments were then inverse normal transformed to the  $N(0,1)$  distribution. The GenMets and DGI cohorts were additionally stratified by the Metabolic syndrome and Type 2 Diabetes case status, respectively. Only men were available in GerMIFS I and II and ULSAM. As NTR has related samples, males and females were analysed together in order to account for the relatedness.

### Single variant association- and Meta-analysis methods

A genome-wide association analysis was run in each of the cohorts separately (see Supplementary Table 13 for software details). The association results were quality controlled centrally to have as harmonized dataset as possible. In the procedure, the following SNPs were removed: SNPs with minor allele count < 3; SNPs with imputation quality Proper\_INFO < 0.4; duplicates; genotyped SNPs with HWE p-value <  $1 \times 10^{-4}$ . The meta-analysis was run using the GWAMA software tool<sup>36,37</sup>, which uses fixed-effects inverse-variance weighted meta-analysis. Genomic control was applied to each of the cohorts in the meta-analysis. SNPs with < 50% of the cohorts contributing or SNPs showing between-study heterogeneity of effect size (Cochran's Q test statistics,  $I^2 < 50\%$ ) were discarded from the meta-analysis results. After these QC steps, the maximum number of SNPs in the analysis was 9,657,952.

SNP associations with p-value <  $5 \times 10^{-8}$  were considered genome-wide significant and lead-SNPs were required to be at least 1Mb away from adjacent lead-SNPs. In areas with long-spanning linkage-disequilibrium, formal conditional analysis was performed in a subset of 12,834 Finnish samples to ensure the independence of the lead-SNPs.

### Search for known functional lipid SNPs

We searched the Online Mendelian Inheritance of Man (OMIM; [www.omim.org](http://www.omim.org)) database for information on 167 loci, which had been found to associate with one of the studied traits (HDL-C, LDL-C, TC and TG) in either the two previously published GWA studies<sup>2,3</sup> or in our genome-wide screening. In each of these 167 loci, every gene in a 2Mb area around the published lead variant was looked up for in the OMIM database and the variants associated with lipid related syndromes or population extreme lipid values were collected. Out of the 167 loci, 38 had OMIM-listed lipid SNP variants within the searched window. As our genotype data only includes SNP variants, deletions, insertions and other copy number variations could not be studied. Each of the OMIM-listed lipid SNPs was subsequently mapped to genome build 37 using dbSNP database for the rsID identification. Of the 38 loci, 18 had at least one polymorphic OMIM SNP in the imputed Finnish test set of seven cohorts (Corogene controls, FTC, GenMets, HBCS, NFBC1966, YFS and PredictCVD, combined  $N = 12,834$ ). To be sure about the functionality of these SNPs, additional literature search was performed to find evidence of the effect on gene transcription or translation. Out of the 18 loci, 8 showed genome-wide significant association in the Finnish meta-analysis and had at least one variant with evidence for functional impact on lipid levels in cell or animal models.

### Formal conditional association analysis in loci containing known functional lipid SNPs

Formal conditional analyses were run using the Finnish test set of 7 cohorts ( $N = 12,834$ ). Each of the cohorts was analysed separately with linear regression analysis implemented by SNPTest software ([http://mathgen.stats.ox.ac.uk/genetics\\_software/snptest/snptest.html](http://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html)). In each cohort, imputation quality threshold of Proper\_INFO > 0.4 was applied. Each locus was analysed only for the trait(s) it had been previously reported to associate with in already published GWA studies. In the conditional analysis on a particular SNP(s), the phenotype was first adjusted with the SNP(s) and then a linear regression model was fitted for the remaining residuals. When we performed iterative conditional analyses in a locus, the signal was first conditioned on the most significant variant followed by conditioning on the top variant from the initial conditional analysis and so forth. Loci where the initial lead-SNP association in the conditional analyses was conditional p-value < 0.01 and no further significant associations (conditional p-value <  $5 \times 10^{-8}$ ) were found within the 2Mb window were considered to be explained.

The results from the seven Finnish cohorts were combined using GWAMA. Because the conditional analyses were run for the pre-selected 2Mb windows only, genomic inflation factor ( $\lambda$ ) correction could not be applied. However, we did not see substantial inflation in the genome-wide association analysis of all four traits in the seven Finnish cohorts ( $\lambda$  range 0.992 – 1.029 depending of the trait and cohort).

### Search for functional candidate SNPs

In order to explore suggestive functional variants causing association signals, which do not have lipid related OMIM listed variants in the locus, we selected 9 loci: *GALNT2*, *MLXILP*, *PPP1R3B*, *TRIB1*, *ADAMTS3*, *LRP4-MADD*, *SOST-DUSP3*, *CILP2* and *HNF4A*. These loci had been significantly associated with lipid traits in either previously published GWA studies<sup>2,3</sup> or in our genome-wide screening as well as in the meta-analysis using 7 Finnish cohorts ( $N = 12,834$ ). In each of these loci, 2Mb windows were searched for functional variants that had association p-value <  $5 \times 10^{-4}$ . Candidate SNPs were annotated using the Ensembl database and functional effects were predicted using the Provean<sup>38</sup>, SIFT<sup>39</sup> and PolyPhen<sup>40</sup> databases. If a variant was annotated as a missense mutation with damaging prediction in at least one of the prediction databases, it was treated as a FL variant and formal conditional analysis was performed to investigate if it explains the association.

### Gene-network analysis

We used 2,206 principal components that had been derived from 77,840 Affymetrix microarrays (54,736 human, 17,081 mouse and 6,023 rat). Since gene-set enrichment analysis showed that each of these components are enriched for at least one biological pathway we used these components to developed a gene function prediction algorithm. To do so we first determine whether each of the components are enriched for a given gene-set, by performing a T-test (contrasting genes, known to be part of this pathways with all other genes), and transform the T-statistics into Z-scores. Subsequently we can the eigenvector coefficients of the 2,206 components for individual genes of interest with the Z-score profile of this gene-set, to predict the gene's involvement in a specific pathway (details provided in Fehrmann et al.<sup>28</sup>, see Cvejic et al.<sup>41</sup> and Wood et al.<sup>29</sup> for a short description). We used a

permutation strategy to determine significance of the predictions, controlling the false discovery rate at 5%. See <http://genenetwork.nl/genenetwork> for the predictions. Based on the Mouse Genome Informatics (MGI; <http://www.informatics.jax.org>) mouse knockout database, we predicted *CD300LG* to increase circulating VLDL cholesterol levels. For *TM6SF2*, the most significantly predicted biological process was intestinal absorption. Only highly significant predictions (permuted p-value <  $1 \times 10^{-6}$ ) were taken into account when profiling the two genes.

We text-mined the sample descriptions provided by the experimenters who uploaded the microarray data to GEO. This text mining allowed us to determine the tissue or cell type for the majority of the samples. We subsequently used Wilcoxon-Mann-Whitney tests in the human samples from the Affymetrix U133 Plus 2.0 platform to ascertain how highly each gene was expressed in samples of a certain tissue or cell type as compared to samples in other tissues and cell types. We found that *CD300LG* is highly expressed in adipose tissue, heart, muscle and plasma and that *TM6SF2* is highly expressed in ileum and intestinal mucosa. See <http://genenetwork.nl/genenetwork> for the expression of genes in different tissues and cell types.

### Modelling proportion of variance explained

To estimate the phenotypic variance explained by different types of SNPs we ran multiple linear regression models in  $R^{42}$  using the FRCORExome9702 dataset ( $N = 5,119$ ), an independent sample set from the Finrisk cohort. For the models all lead-SNPs (Teslovich et al.<sup>2</sup>, Willer et al.<sup>3</sup> and our study) together with FL SNPs and new functional candidates were divided into two groups based on the MAF of the variants in the FRCORExome9702 dataset. The tested SNP sets were:

1. Common (MAF > 5%) lead-SNPs and functional SNPs
2. Adding low-frequency (MAF  $\leq 5\%$ ) lead-SNPs and functional SNPs to SNP-set 1.
3. FL SNPs and the three identified functional candidates.

These SNP-sets were used to explain the variation of the sex, age, age<sup>2</sup> and population stratification adjusted trait residuals. In order to apply linear models, TG was log-transformed before adjustments.

### Linear mixed model estimate of the variance explained by common SNPs

We estimated how much phenotypic variance a panel of 319,445 directly genotyped SNPs with MAF > 1% in the autosomes explain using the linear mixed model approach implemented in GCTA<sup>43</sup> (v.1.13). This estimate is a lower bound of the total additive genetic variance, because it only includes the contribution of the variants tagged by the panel of common SNPs that was used in the analysis. The analysis included samples from six Finnish cohorts (NFBC1966, Corogene controls, GenMets, YFS, HBCS and PredictCVD) for which we had access to the individual genotype data. All mixed model analyses excluded individuals in such a way that none of the remaining pairs of individuals had an estimated relatedness coefficient  $r > 0.05$  and the same trait values were used as with

the individual SNP analyses. The sample sizes for the traits were 10,466 for HDL-C, 10,383 for LDL-C, 10,472 for TC and 10,451 for TG.

### Gene-based association analysis

Transcript boundaries were defined according to the UCSC human genome database. Within each study, GRANVIL<sup>31</sup> was used to test for association of each trait with accumulations of minor alleles (“mutational load”) at successfully imputed rare variants (MAF  $\leq 1\%$  and info  $\geq 0.4$ , Supplementary Table 1) within genes in a linear regression framework: (i) irrespective of annotation; and (ii) restricted to non-synonymous changes. Fixed-effects meta-analysis was performed by combining directed Z-scores from the regression analysis across studies, weighted by sample size. The significance threshold was set to p-value  $< 1.7 \times 10^{-6}$  corresponding to a Bonferroni correction for 30,000 genes. Conditional analyses were performed to assess the evidence of association of traits the mutational load of a gene after accounting for the lead-SNP by including the genotype (under an additive model) of this variant as a covariate in the regression model. Conditional analyses were also performed to assess the independence of effects of rare variants in two genes by including the mutational load of one as a covariate in the regression model for the trait association with the other.

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

### Authors

Ida Surakka<sup>1,2</sup>, Momoko Horikoshi<sup>3,4</sup>, Reedik Mägi<sup>5</sup>, Antti-Pekka Sarin<sup>1,2</sup>, Anubha Mahajan<sup>3</sup>, Vasiliki Lagou<sup>3,4</sup>, Letizia Marullo<sup>6</sup>, Teresa Ferreira<sup>3</sup>, Benjamin Miraglio<sup>1</sup>, Sanna Timonen<sup>1</sup>, Johannes Kettunen<sup>1</sup>, Matti Pirinen<sup>1</sup>, Juha Karjalainen<sup>7</sup>, Gudmar Thorleifsson<sup>8</sup>, Sara Hägg<sup>9,10</sup>, Jouke-Jan Hottenga<sup>11</sup>, Aaron Isaacs<sup>12,13</sup>, Claes Ladenvall<sup>14</sup>, Marian Beekman<sup>15,16</sup>, Tõnu Esko<sup>5,17,18,19</sup>, Janina S Ried<sup>20</sup>, Christopher P Nelson<sup>21,22</sup>, Christina Willenborg<sup>23,24</sup>, Stefan Gustafsson<sup>9,10</sup>, Harm-Jan Westra<sup>7</sup>, Matthew Blades<sup>25</sup>, Anton JM de Craen<sup>15,26</sup>, Eco J de Geus<sup>11</sup>, Joris Deelen<sup>15,16</sup>, Harald Grallert<sup>27,28,29</sup>, Anders Hamsten<sup>30</sup>, Aki S. Havulinna<sup>31</sup>, Christian Hengstenberg<sup>32,33</sup>, Jeanine J Houwing-Duistermaat<sup>34</sup>, Elina Hyppönen<sup>35,36,37</sup>, Lennart C Karssen<sup>12</sup>, Terho Lehtimäki<sup>38</sup>, Valeriya Lyssenko<sup>14,39</sup>, Patrik KE Magnusson<sup>9</sup>, Evelin Mihailov<sup>5</sup>, Martina Müller-Nurasyid<sup>20,33,40,41</sup>, John-Patrick Mpindi<sup>1</sup>, Nancy L Pedersen<sup>9</sup>, Brenda WJH Penninx<sup>42</sup>, Markus Perola<sup>1,2</sup>, Tune H Pers<sup>17,18,43</sup>, Annette Peters<sup>27,29,32</sup>, Johan Rung<sup>44</sup>, Johannes H Smit<sup>42</sup>, Valgerdur Steinthorsdottir<sup>8</sup>, Martin D Tobin<sup>45</sup>, Natalia Tsernikova<sup>5</sup>, Elisabeth M van Leeuwen<sup>12</sup>, Jorma S Viikari<sup>46</sup>, Sara M Willems<sup>12</sup>, Gonneke Willemsen<sup>11</sup>, Heribert Schunkert<sup>32,33</sup>, Jeanette Erdmann<sup>23,24</sup>, Nilesh J Samani<sup>21,22</sup>, Jaakko Kaprio<sup>1,47,48</sup>, Lars Lind<sup>49</sup>, Christian Gieger<sup>20</sup>, Andres Metspalu<sup>5,50</sup>, P Eline Slagboom<sup>15,16</sup>, Leif Groop<sup>1,14</sup>, Cornelia M van Duijn<sup>12,13</sup>, Johan G Eriksson<sup>51,52,53,54</sup>, Antti Jula<sup>55</sup>, Veikko Salomaa<sup>31</sup>, Dorret I Boomsma<sup>11</sup>, Christine Power<sup>35</sup>, Olli T Raitakari<sup>56,57</sup>, Erik Ingelsson<sup>3,10</sup>, Marjo-Riitta Järvelin<sup>58,59,60,61,62</sup>, Kari Stefansson<sup>8,63</sup>, Lude Franke<sup>7</sup>, Elina Ikonen<sup>64,65</sup>, Olli Kallioniemi<sup>1</sup>, Vilja Pietiäinen<sup>1</sup>, Cecilia M Lindgren<sup>3,18</sup>, Unnur Thorsteinsdottir<sup>8,63</sup>, Aarno Palotie<sup>1,2,18,66</sup>, Mark I

McCarthy<sup>3,4,67</sup>, Andrew P Morris<sup>3,5,68</sup>, Inga Prokopenko<sup>69</sup>, and Samuli Ripatti<sup>1,47,70</sup>  
**for the ENGAGE Consortium**

## Affiliations

<sup>1</sup>Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland <sup>2</sup>Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland <sup>3</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK <sup>4</sup>Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK <sup>5</sup>Estonian Genome Center, University of Tartu, Tartu, Estonia <sup>6</sup>Department of Life Sciences and Biotechnology, Genetic Section, University of Ferrara, Ferrara, Italy <sup>7</sup>University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands <sup>8</sup>deCODE Genetics/Amgen inc., Reykjavik, Iceland <sup>9</sup>Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden <sup>10</sup>Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden <sup>11</sup>Department of Biological Psychology, EMGO institute for Health and Care research, VU University & VU medical center, Amsterdam, The Netherlands <sup>12</sup>Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands <sup>13</sup>Centre for Medical Systems Biology, Leiden, the Netherlands <sup>14</sup>Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Centre, Skåne University Hospital, Malmö, Sweden <sup>15</sup>Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands <sup>16</sup>Netherlands Consortium for Healthy Ageing, Leiden, The Netherlands <sup>17</sup>Divisions of Endocrinology and Center for Basic and Translational Obesity Research, Children's Hospital, Boston, Massachusetts, USA <sup>18</sup>The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA <sup>19</sup>Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA <sup>20</sup>Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany <sup>21</sup>Department of Cardiovascular Sciences, University of Leicester, Leicester, LE3 9QP, UK <sup>22</sup>National Institute for Health Research (NIHR) Leicester Cardiovascular Disease Biomedical Research Unit, Glenfield Hospital, Leicester, UK <sup>23</sup>Institut für Integrative und Experimentelle Genomik, Universität zu Lübeck, Lübeck, Germany <sup>24</sup>Deutsches Zentrum für Herz-Kreislauf-Forschung e. V. (DZHK), Partnersite Hamburg, Lübeck, Kiel, Germany <sup>25</sup>Bioinformatics and Biostatistics Support Hub (B/BASH), University of Leicester, University Road, Leicester, UK <sup>26</sup>Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands <sup>27</sup>Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany <sup>28</sup>German Center for Diabetes Research (DZD), Neuherberg, Germany <sup>29</sup>Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany <sup>30</sup>Cardiovascular Genetics and Genomics Group, Atherosclerosis Research Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden <sup>31</sup>Unit of

Chronic Disease Epidemiology and Prevention, National Institute for Health and Welfare, Helsinki, Finland <sup>32</sup>Deutsches Herzzentrum München, Technische Universität München, München, Germany <sup>33</sup>DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany <sup>34</sup>Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands <sup>35</sup>Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health, London, UK <sup>36</sup>South Australian Health and Medical Research Institute, Adelaide, Australia <sup>37</sup>School of population Health and Sansom Institute, University of South Australia, Adelaide, Australia <sup>38</sup>Department of Clinical Chemistry, Fimlab Laboratories and School of Medicine, University of Tampere, Tampere, Finland <sup>39</sup>Steno Diabetes Center A/S, Gentofte, Denmark <sup>40</sup>Department of Medicine I, University Hospital Grosshadern, Ludwig-Maximilians-Universität, Munich, Germany <sup>41</sup>Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany <sup>42</sup>Department of Psychiatry, VU University Medical Center, Amsterdam, The Netherlands <sup>43</sup>Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark <sup>44</sup>European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK <sup>45</sup>Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, University Road, Leicester, UK <sup>46</sup>Department of Medicine, University of Turku and Division of Medicine, Turku University Hospital, Turku, Finland <sup>47</sup>Hjelt Institute, University of Helsinki, Helsinki, Finland <sup>48</sup>Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, Helsinki, Finland <sup>49</sup>Department of Medical Sciences, Uppsala University, Akademiska Sjukhuset, Uppsala, Sweden <sup>50</sup>The Institute of Molecular and Cell Biology of the University of Tartu, Tartu, Estonia <sup>51</sup>Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland <sup>52</sup>Folkhälsan Research Centre, Helsinki, Finland <sup>53</sup>Helsinki University Hospital, Unit of Primary Health Care, Helsinki, Finland <sup>54</sup>Department of Health Promotion and Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland <sup>55</sup>Department of Chronic Disease Prevention, National Institute for Health and Welfare, Turku, Finland <sup>56</sup>Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Finland <sup>57</sup>Department of Clinical Physiology and Nuclear Medicine, University of Turku and Turku University Hospital, Finland <sup>58</sup>Department of Epidemiology and Biostatistics, MRC Health Protection Agency (HPE), Centre for Environment and Health, School of Public Health, Imperial College London, UK <sup>59</sup>Institute of Health Sciences, University of Oulu, Oulu, Finland <sup>60</sup>Biocenter Oulu, University of Oulu, Oulu, Finland <sup>61</sup>Unit of Primary Care, Oulu University Hospital, Oulu, Finland <sup>62</sup>Department of Children and Young People and Families, National Institute for Health Welfare, Oulu, Finland <sup>63</sup>Faculty of Medicine, University of Iceland, Reykjavik, Iceland <sup>64</sup>Institute of Biomedicine, Anatomy, University of Helsinki, Helsinki, Finland <sup>65</sup>Minerva Foundation Institute for Medical Research, Helsinki, Finland <sup>66</sup>Haartman Institute, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland

<sup>67</sup>Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Churchill Hospital, Oxford, UK <sup>68</sup>Department of Biostatistics, University of Liverpool, Liverpool, UK <sup>69</sup>Department of Genomics of Common Disease, School of Public Health, Imperial College London, Hammersmith Hospital, London, UK <sup>70</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

## Acknowledgements

IS was partly funded by the Helsinki University Doctoral Programme in Biomedicine (DPBM). MH was funded by Manpei Suzuki Diabetes Foundation Grant-in-Aid for the young scientists working abroad. APM and AnuM acknowledged funding from the Wellcome Trust under awards WT098017, WT090532 and WT064890. VasL, LM, SH and IP were funded in part through the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project, grant agreement HEALTH-F4-2007- 201413. LM was in part sponsored by "5 per mille" contribution assigned to the University of Ferrara, income tax return year 2009 and in part from the ENGAGE Exchange and Mobility Program for ENGAGE training funds Program for ENGAGE training funds. MDT holds a Medical Research Council Senior Clinical Fellowship (G0902313). ST is supported by Sigrid Juselius foundation. JSR and CG have received funding by a grant of the RFBR (Russian Foundation for Basic Research)-Helmholtz Joint Research Group (12-04-91322). CG received funding from the European Union's Seventh Framework Program (FP7-Health-F5-2012) under grant agreement No 305280 (MIMOmics). NJS holds a BHF Chair funded by the British Heart Foundation. CPN was funded by the NIHR Leicester Cardiovascular Biomedical Research Unit. VeS was supported by Finnish Foundation for Cardiovascular research, and the Finnish Academy (grant number 139635). EI was supported by the Academy of Finland Centre of Excellence in Biomembrane Research (272130), Academy of Finland (263841) and Sigrid Juselius Foundation. VP was supported by the University of Helsinki Postdoctoral researcher grant, Magnus Ehrnrooth foundation and the Kymenlaakso Cultural foundation. OK, JP and VP have received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°258068; EU-FP7-Systems Microscopy NoE. SR was supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (213506 and 129680), Academy of Finland (251217), the Finnish foundation for Cardiovascular Research and the Sigrid Juselius Foundation. The High Throughput Biomedicine Unit of Institute for Molecular Medicine Finland, Ruusu Kovanen and Anna Uro are acknowledged for technical expertise. Jauhainen M. is acknowledged for sharing his expertise in the article writing process. Cohort specific acknowledgements are in the Supplementary Note.

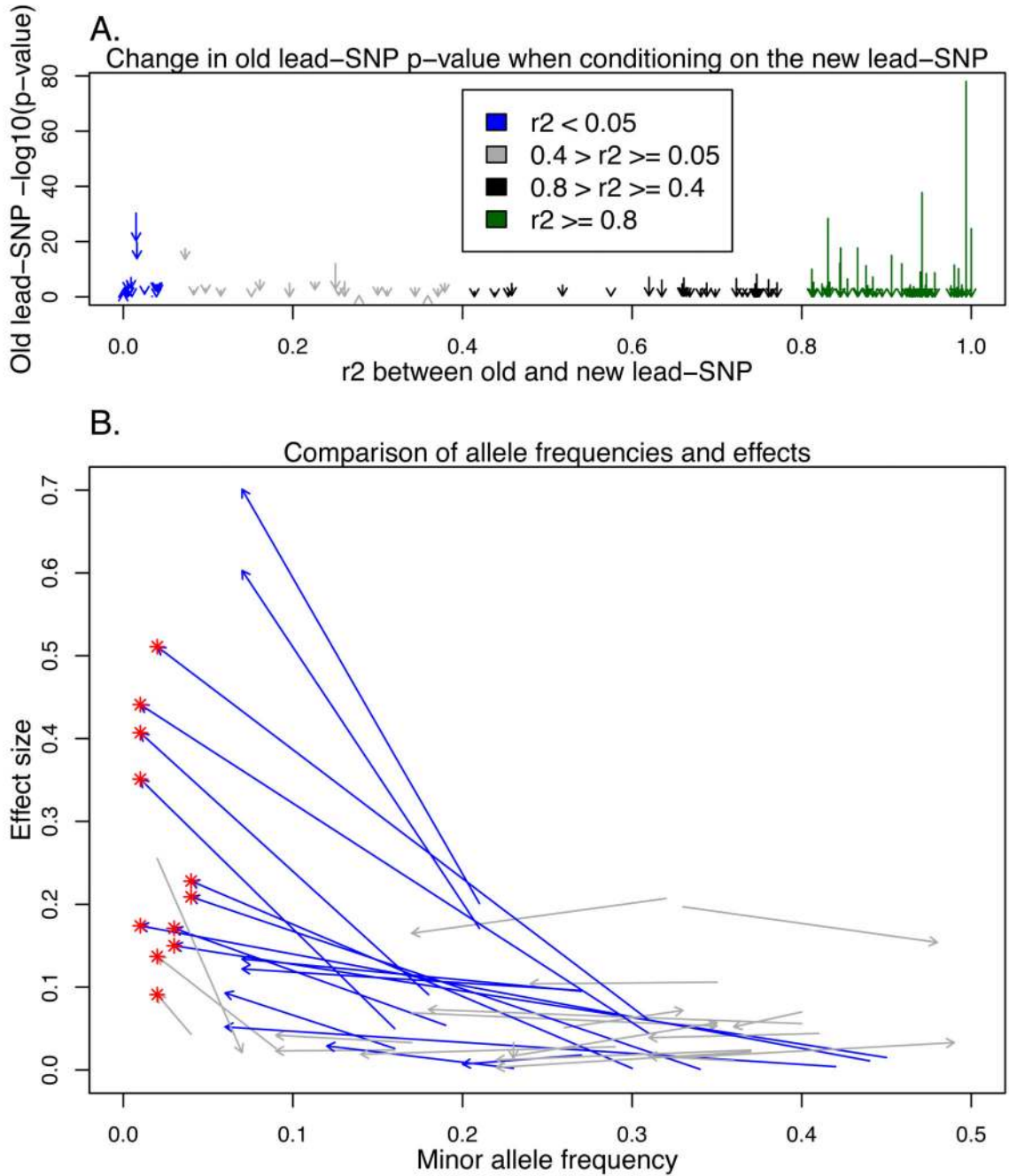
## References

1. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet.* 2012; 13:135–145. [PubMed: 22251874]
2. Teslovich T, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature.* 2010; 466:707–713. [PubMed: 20686565]
3. Willer C, et al. Discovery and Refinement of Loci Associated with Lipid Levels. *Nat. Genet.* 2013;10.1038/ng.2797
4. 1000 Genomes Project Consortium. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012; 491:56–65. [PubMed: 23128226]
5. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 2007; 38:906–913. [PubMed: 17572673]
6. Howie B, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009; 5:e1000529. [PubMed: 19543373]
7. Rall SJ, Weisgraber K, Innerarity T, Mahley R. Identical structural and receptor binding defects in apolipoprotein E2 in hypo-, normo-, and hypercholesterolemic dysbetalipoproteinemia. *J. Clin. Invest.* 1983; 71:1023–1031. [PubMed: 6300187]
8. Rall SJ, Weisgraber K, Innerarity T, Mahley R. Structural basis for receptor binding heterogeneity of apolipoprotein F from type III hyperlipoproteinemic subjects. *Proc. Natl. Acad. Sci. U S A.* 1982; 79:4696–4700. [PubMed: 6289314]



9. Cohen J, Boerwinkle E, Mosley TJ, Hobbs H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* 2006; 354:1264–1272. [PubMed: 16554528]
10. Romeo S, et al. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* 2007; 39:513–516. [PubMed: 17322881]
11. Holmen OL, et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nat Genet.* 2014; 46:345–351. [PubMed: 24633158]
12. Kozlitina J, Smargris E, Stender S, Nordestgaard BG, Zhou HH, Tybjærg-Hansen A, Vogt TF, Hobbs HH, Cohen JC. Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet.* 2014; 46:352–356. [PubMed: 24531328]
13. Dupuis J, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* 2010; 42:105–116. [PubMed: 20081858]
14. Kahn B, Alquier T, Carling D, Hardie D. AMP-activated protein kinase: ancient energy gauge provides clues to modern understanding of metabolism. *Cell. Metab.* 2005; 1:15–25. [PubMed: 16054041]
15. Beer N, et al. The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum. Mol. Genet.* 2009; 18:4081–4088. [PubMed: 19643913]
16. Weisgraber K, Rall SJ, Mahley R. Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *J. Biol. Chem.* 1981; 256:9077–9083. [PubMed: 7263700]
17. Ghebranious N, Ivacic L, Mallum J, Dokken C. Detection of ApoE E2, E3 and E4 alleles using MALDI-TOF mass spectrometry and the homogeneous mass-extend technology. *Nucleic Acids Res.* 2005; 33:e149. [PubMed: 16204452]
18. Deeb S, Peng R. The C-514T polymorphism in the human hepatic lipase gene promoter diminishes its activity. *J. Lipid. Res.* 2000; 41:155–158. [PubMed: 10627514]
19. Durstenfeld A, Ben-Zeev O, Reue K, Stahnke G, Doolittle M. Molecular characterization of human hepatic lipase deficiency. In vitro expression of two naturally occurring mutations. *Arterioscler Thromb.* 1994; 14:381–385. [PubMed: 8123642]
20. Liu DJ, et al. Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* 2013 epub ahead of print:Dec 15. 10.1038/ng.2852
21. Albrechtsen A, et al. Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia.* 2013; 56:298–310. [PubMed: 23160641]
22. Takatsu H, Hase K, Ohmae M, Ohshima S, Hashimoto K, Taniura N, Yamamoto A, Ohno H. CD300 antigen like family member G: A novel Ig receptor like protein exclusively expressed on capillary endothelium. *Biochem. Biophys. Res. Commun.* 2006; 348:183–191. [PubMed: 16876123]
23. Umemoto E, et al. Nepmucin, a novel HEV sialomucin, mediates L-selectin-dependent lymphocyte rolling and promotes lymphocyte adhesion under flow. *J. Exp. Med.* 2006; 203:1603–1614. [PubMed: 16754720]
24. Jin S, Umemoto E, Tanaka T, Shimomura Y, Tohya K, Yang BG, Jang MH, Hirata T, Miyasaka M. Nepmucin/CLM-9, an Ig domain-containing sialomucin in vascular endothelial cells, promotes lymphocyte transendothelial migration in vitro. *FEBS Lett.* 2008; 582:3018–3024. [PubMed: 18675811]
25. Cannon JP, O'Driscoll M, Litman GW. Specific lipid recognition is a general feature of CD300 and TREM molecules. *Immunogenetics.* 2012; 64:39–47. [PubMed: 21800138]
26. Carim-Todd L, Escarceller M, Estivill X, Sumoy L. Cloning of the novel gene TM6SF1 reveals conservation of clusters of paralogous genes between human chromosomes 15q24-->q26 and 19p13.3-->p12. *Cytogenet. Cell Genet.* 2000; 90:255–260. [PubMed: 11124529]
27. Mahdessian H, Taxiarchis A, Popov S, Silveira A, Franco-Cereceda A, Hamsten A, Eriksson P, Van't Hooft F. TM6SF2 is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid droplet content. *Proc Natl Sci U S A.* 2014 epub ahead of print June 4<sup>th</sup>.

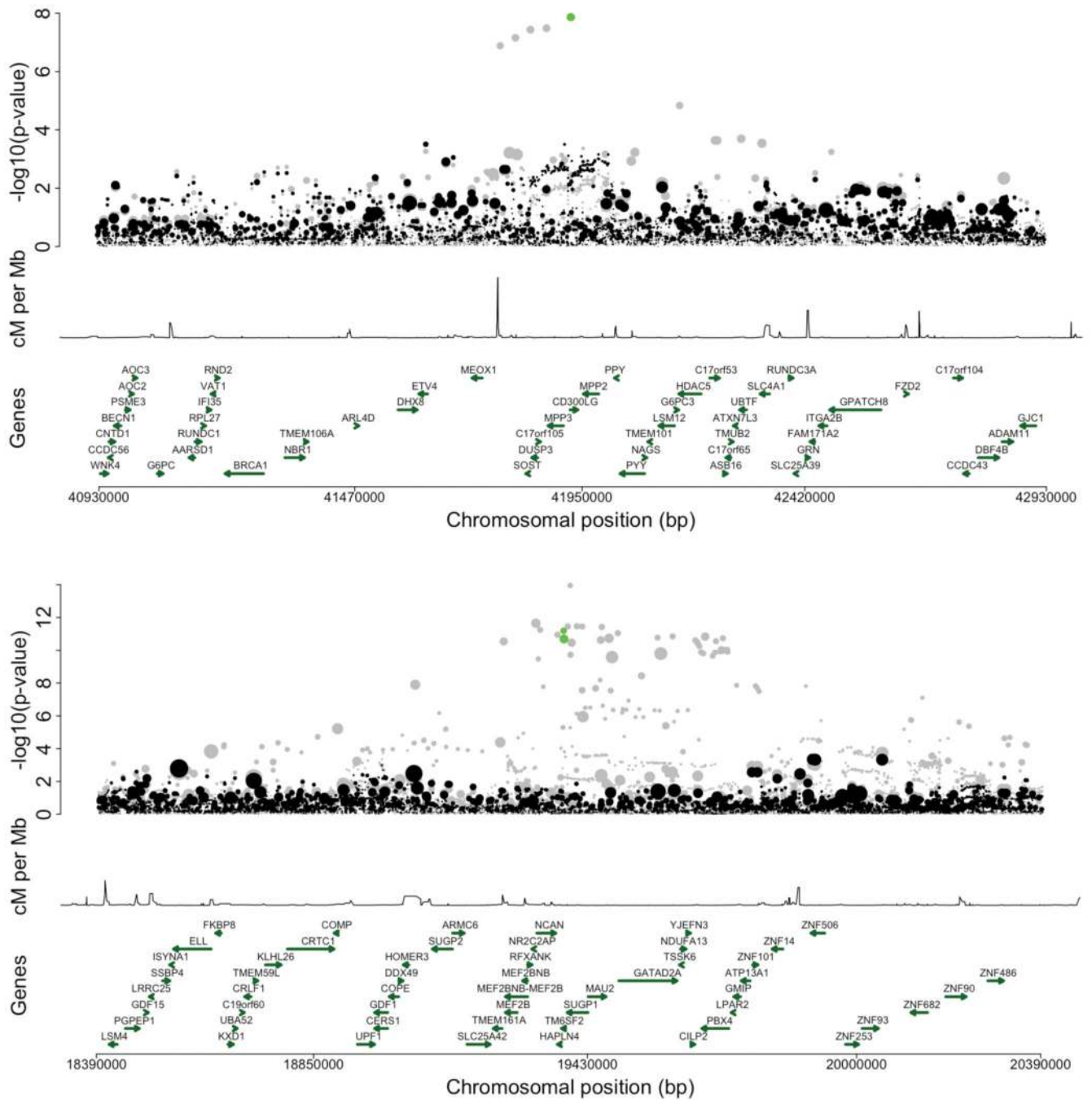
28. Fehrmann SN, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat Genet.* 2015 epub ahead of print Jan 12<sup>th</sup>.
29. Wood AR, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Gen.* 2014 Epub Oct 5. 10.1038/ng.3097
30. van Dongen J, Willemsen G, Chen WM, de Geus EJC, Boomsma DI. Heritability of metabolic syndrome traits in a large population-based sample. *J. Lipid Res.* 2013; 54:2914–2923. [PubMed: 23918046]
31. Mägi R, et al. Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. *Genet Epidemiol.* 2012; 36:785–796. [PubMed: 22951892]
32. Blattman P, Schubert 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]
33. Service SK, et al. Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.* 2014; 10:e1004147. [PubMed: 24497850]
34. Johansen CT, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet.* 2010; 42:682–687.
35. 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]
36. Mägi R, Morris A. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics.* 2010; 11:288.10.1186/1471-2105-11-288 [PubMed: 20509871]
37. Mägi R, Lindgren C, Morris A. Meta-analysis of sex-specific genome-wide association studies. *Genet. Epidemiol.* 2010; 34:846–853. [PubMed: 21104887]
38. Choi Y, Sims G, Murphy S, Miller J, Chan A. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE.* 2012; 7:e46688. [PubMed: 23056405]
39. Ng P, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res.* 2001; 11:863–874. [PubMed: 11337480]
40. Adzhubei I, et al. A method and server for predicting damaging missense mutations. *Nat. Methods.* 2010; 7:248–249. [PubMed: 20354512]
41. Cvejic A, et al. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* 2013; 45:542–545. [PubMed: 23563608]
42. R Development Core Team. R: A language and environment for statistical computing. 2008. <<http://R-project.org>>
43. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 2011; 88:76–82. [PubMed: 21167468]



**Figures 1A-B.**

Change in p-value after analysis conditional on the new lead-SNP and comparison of new and previously reported lead-SNP effect sizes and allele frequencies per locus. In both figures, each of the arrows represent one locus and trait, where significant association was found in our screening and in one of the previously published large-scale screening studies<sup>2,3</sup> and the colouring is based on the linkage disequilibrium (LD) between the old and new lead-SNP. The red '\*' represents for the new low-frequency lead-SNPs. In the Figure 1A, on the Y-axis are the  $-\log_{10}$  p-values, arrows starting from the p-value seen in the

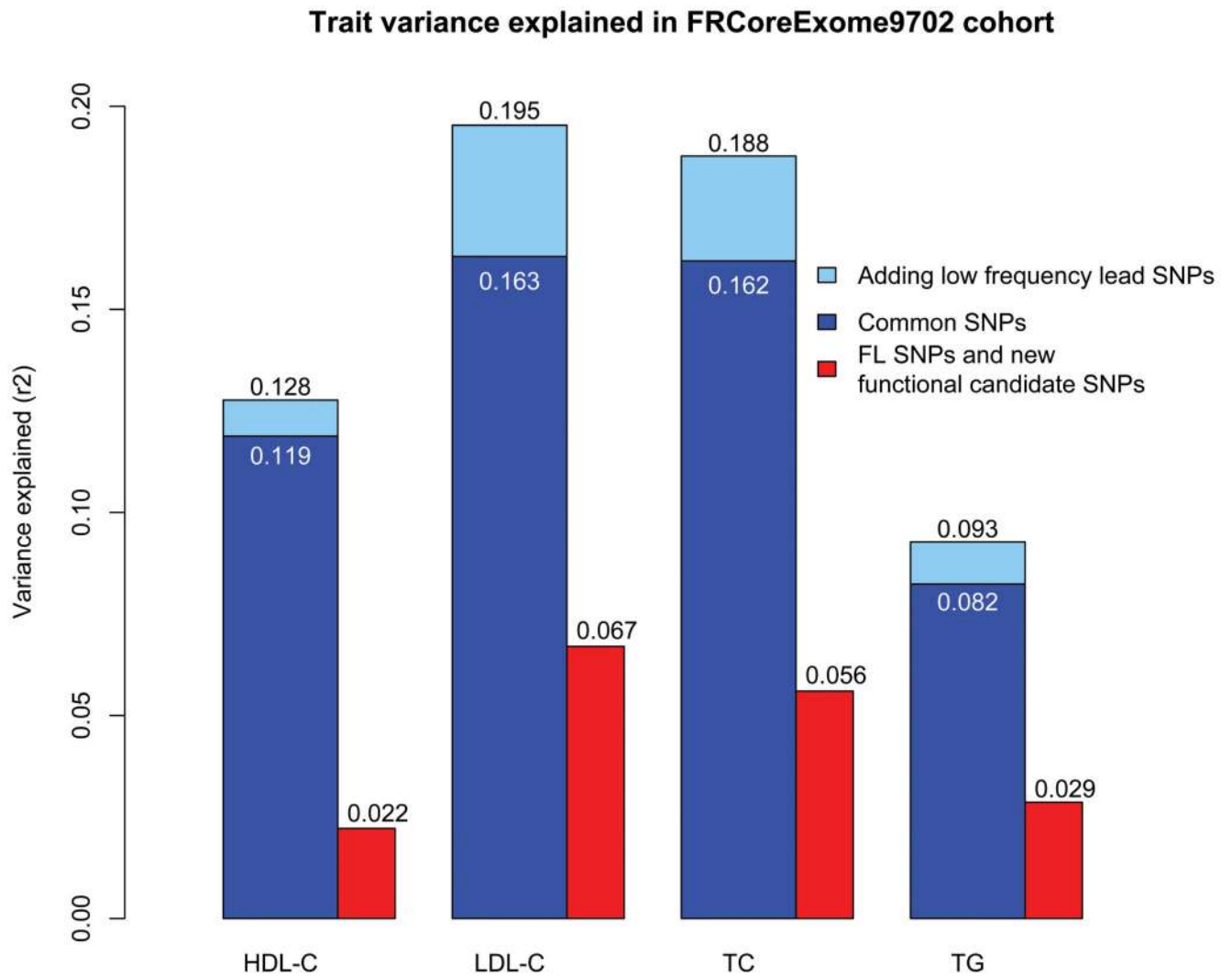
unconditional analysis in Finnish subset ( $N = 12,834$ ) and pointing to the p-value in analysis conditional on the new lead-SNP. In Figure 1B, each arrow starts from established lead-SNP effect and minor allele frequency (MAF) and points to the corresponding values for the new lead-SNP. The effects have been estimated in the FRCORExome9702 sample set ( $N = 5,119$ ), independent of the discovery set. Only results for loci with  $r^2 < 0.4$  have been presented for clarity.



**Figures 2A-B.**

Regional association plots of the conditional analysis in loci where the new functional candidate SNPs explain the genome-wide association. Figure 2A illustrates the results in *SOST-DUSP3* locus for TG and Figure 2B results in *CILP2* locus for TC in Finnish subset ( $N = 12,834$ ). In these figures, the first panel shows the  $-\log_{10}$  p-value of each variant as a dot whose size reflects the effect size. The second panel shows the recombination rate in the area and the third panel shows the positions of genes. X-axis is the physical position in the genome. In grey are the association results from the unconditional analysis with green dots

representing the new functional candidate SNPs. Black dots are the results from the conditional analysis.



**Figure 3.**

Proportion of total trait variance explained by the lead-SNPs and functional SNPs. The proportion of the trait variance explained by different SNP-sets has been estimated in independent FRCORExome9702 sample set ( $N = 5,119$ ). All lead-SNPs from the three association screens (Teslovich et al.<sup>2</sup>, Willer et al.<sup>3</sup> and our screen) together with the known functional lipid SNPs (FL SNPs) and new functional candidate SNPs were grouped based on their allele frequency in the FRCORExome9702 dataset to common SNPs (allele frequency  $> 5\%$ ) and to low-frequency SNPs (allele frequency  $\leq 5\%$ ). The variance explained by these two groups is presented with blue bars. The proportion of variance explained by the FL SNPs and functional candidates is presented with the red bar.

Table 1

Newly identified loci associated to HDL-C, LDL-C, TC and/or TG

Locus	CHR	Position B37	rsID	Annotation	Primary associated trait	Secondary associated trait	Alleles (effect/ other)	EAF	Effect	SE	Meta-analysis p-value	N
<i>PROX1</i>	1	214161820	rs340839*	UTR5	TG		A/G	0.47	0.039	0.006	4.4×10 <sup>-10</sup>	54836
<i>CEP68</i>	2	65284623	rs2540948	intronic	TG		C/T	0.35	-0.036	0.006	6.6×10 <sup>-9</sup>	59939
<i>PRKAG3</i>	2	219699999	rs78058190	intergenic	HDL-C		A/G	0.05	-0.141	0.020	5.7×10 <sup>-12</sup>	52934
<i>ADAMTS3</i>	4	73696709	rs117087731	intergenic	TC		T/A	0.01	0.308	0.051	2.310 <sup>-9</sup>	23641
<i>MTHFD2L</i>	4	75084732	rs182616603	intronic	TC	LDL-C	T/C	0.01	0.374	0.044	1.8×10 <sup>-17</sup>	42905
<i>GPR85</i>	7	112722196	rs2255811	UTR3	TG		G/A	0.25	0.041	0.007	2.3×10 <sup>-8</sup>	59962
<i>RM12</i>	16	11454650	rs7188861	intergenic	HDL-C		A/C	0.20	0.044	0.008	6.9×10 <sup>-9</sup>	60578
<i>TM4SF5</i>	17	4667984	rs193042029	intergenic	TG		G/T	0.01	-0.170	0.029	8.1×10 <sup>-9</sup>	50105
<i>GATA6</i>	18	19907770	rs79588679	intergenic	LDL-C		T/C	0.17	-0.049	0.009	3.6×10 <sup>-8</sup>	53108
<i>ZNF274</i>	19	58681861	rs117492019	intergenic	LDL-C		T/G	0.19	-0.047	0.008	1.2×10 <sup>-8</sup>	55371
		58671267	rs12983728	intergenic		TC	A/G	0.16	-0.046	0.008	4.9×10 <sup>-8</sup>	58904

\* Present in the HapMap 2 reference panel.

Table presents the association meta-analysis results for the newly identified loci for the four tested lipid traits. Effect sizes are presented in s.d units. CHR: Chromosome, EAF: Effect allele frequency, SE: standard error of the effect, N: number of samples, HDL-C: high density lipoprotein



Table 2

Association results of unconditional analysis and analysis conditional on known Mendelian and functional lipid SNPs in loci where the functional SNPs explain the genome-wide association.

Locus	CHR	Trait	rsID	MAF	Lead-SNP in the unconditional analysis			p-value N	
					Effect (SE)	p-value N	Covariate SNPs in the model (MAF)		
<i>CELSR2-SORT1</i>	1	LDL-C	rs646776	0.216	0.159 (0.015)	1.31×10 <sup>-25</sup> 12739	rs12740374 (21.6 %)	0.001 (0.015)	0.958 12739
					0.123 (0.015)	4.06×10 <sup>-16</sup> 12834	rs12740374 (21.6 %)	0.001 (0.015)	0.959 12834
<i>GCKR</i>	2	TG	rs1260326	0.353	0.128 (0.013)	8.44×10 <sup>-23</sup> 12815	rs1260326 (35.3 %)	NA	NA
					0.090 (0.015)	7.23×10 <sup>-10</sup> 12825	rs113298164 (1.4 %) rs1800588 (25.1 %)	-2×10 <sup>-6</sup> (0.015)	1.000 11893
<i>LIPC</i>	15	TC	rs1800588	0.251	0.085 (0.015)	6.86×10 <sup>-9</sup> 12801	rs13298164 (1.4 %) rs1800588 (25.1 %)	0.022 (0.015)	0.152 11873
					0.648 (0.031)	5.93×10 <sup>-95</sup> 12730	rs7412 (4.8 %) rs429358 (18.1 %)	NA	NA
<i>APOE</i>	19	TC	rs7412	0.048	0.456 (0.031)	3.10×10 <sup>-49</sup> 12827	rs7412 (4.8 %) rs429358 (18.1 %)	NA	NA
					0.089 (0.015)	5.74×10 <sup>-9</sup> 12799	rs7412 (4.8 %) rs429358 (18.1 %)	NA	NA

The table shows results for unconditional association analysis and analysis conditional on variants known to cause Mendelian forms of dyslipidemic syndromes and, more broadly, variants with known functional impact on lipids (FL SNPs). In case multiple candidate variants were observed in a locus, they were all included in the same model. Results for the lead-SNP from the unconditional analysis are presented from the meta-analysis of Finnish subset (N = 12,834). Effect sizes are presented in s.d units. CHR: Chromosome, MAF: Minor allele frequency, SE: Standard error of effect estimate, N: Number of samples, HDL-C: High-density lipoprotein cholesterol, LDL-C: Low-density lipoprotein cholesterol, TC: Total cholesterol, TG: Triglycerides.

Table 3

Association results of unconditional analysis and analysis conditional on the functional candidate SNPs in four loci with genome-wide significant p-value and new functional candidate SNPs.

Locus	CHR	Trait	Tested lead-SNP rsID	MAF	Unconditional		Testing the lead-SNP effects					
					Effect (SE)	p-value N	Candidate SNP used as a single covariate in the model (MAF)	Lead-SNP effect when adjusting for single SNP (SE)	p-value N	Candidate SNPs used jointly as covariates in the model (MAF)	Lead-SNP effect when adjusting for multiple SNPs (SE)	p-value N
<i>MLXIPL</i>	7	TG	rs35797675	0.174	-0.119 (0.017)	7.89×10 <sup>-13</sup> 12810	rs35332062 (12%)	-0.044 (0.017)	0.0073 12810	rs35332062 (12%) rs3812316 (12%)	-0.044 (0.017)	0.0074 12810
							rs3812316 (12%)	-0.044 (0.017)	0.0073 12810			
<i>LRP4-MADD</i>	11	HDL-C	rs2596401	0.426	-0.092 (0.013)	3.55×10 <sup>-12</sup> 11894	rs2279238 (27%)	-0.039 0.013	0.0033 11894			
							rs2290148 (27%)	-0.040 0.013	0.0028 11894			
							rs34312154 (20%)	-0.060 0.013	6.0×10 <sup>-6</sup> 11894	rs2279238 (27%) rs2290148 (27%) rs34312154 (20%)		
							rs75352463 (3%)	-0.077 0.013	6.4×10 <sup>-9</sup> 11894	rs75352463 (3%) rs1064608 (36%) rs5896 (23%)		
							rs1064608 (36%)	-0.060 0.013	7.1×10 <sup>-6</sup> 11894			
<i>SOST-DUSP3</i>	17	TG	rs72836561	0.027	0.234 (0.041)	1.36×10 <sup>-8</sup> 12806	rs5896 (23%)	-0.055 0.013	3.0×10 <sup>-5</sup> 11894			
							rs72836561 (3%)	NA	NA			
<i>CILP2</i>	19	LDL-C	rs8100204	0.137	-0.124 (0.019)	1.75×10 <sup>-10</sup> 12723	rs2228603 (7%)	-0.067 (0.019)	5.7×10 <sup>-4</sup> 12723	rs58542926 (6%) rs187429064 (4%)	-0.006 (0.019)	0.744 12723
							rs2228603 (7%)					

Locus	CHR	Trait	Tested lead-SNP rsID	MAF	Unconditional		Testing the lead-SNP effects				p-value N	
					Effect (SE)	p-value N	Candidate SNP used as a single covariate in the model (MAF)	Lead-SNP effect when adjusting for single SNP (SE)	p-value N	Candidate SNPs used jointly as covariates in the model (MAF)		Lead-SNP effect when adjusting for multiple SNPs (SE)
<i>CILP2</i>	19	TC	rs8100204	0.137	-0.149 (0.019)	1.12×10 <sup>-14</sup> 12819	rs58542926 (6%)	-0.064 (0.019)	9.9×10 <sup>-4</sup> 12723	rs58542926 (6%)	-0.011 (0.019)	0.565 12819
							rs187429064 (4%)	-0.078 (0.019)	5.4×10 <sup>-5</sup> 12723	rs187429064 (4%)	-0.047 (0.026)	0.075 12809
							rs2228603 (7%)	-0.079 (0.019)	3.8×10 <sup>-5</sup> 12819	rs2228603 (7%)	-0.047 (0.026)	0.075 12809
							rs58542926 (6%)	-0.079 (0.019)	4.2×10 <sup>-5</sup> 12819	rs58542926 (6%)	-0.047 (0.026)	0.075 12809
							rs187429064 (4%)	-0.093 (0.019)	1.3×10 <sup>-6</sup> 12819	rs187429064 (4%)	-0.047 (0.026)	0.075 12809
							rs2074300 (14%)	-0.099 (0.019)	3.2×10 <sup>-7</sup> 12819	rs2074300 (14%)	-0.047 (0.026)	0.075 12809
							rs12151060 (14%)	-0.125 (0.019)	1.0×10 <sup>-10</sup> 12819	rs12151060 (14%)	-0.047 (0.026)	0.075 12809
							rs2228603 (7%)	-0.064 (0.026)	0.015 12809	rs2228603 (7%)	-0.047 (0.026)	0.075 12809
							rs58542926 (6%)	-0.048 (0.026)	0.066 12809	rs58542926 (6%)	-0.047 (0.026)	0.075 12809
							rs187429064 (4%)	-0.147 (0.026)	2.5×10 <sup>-8</sup> 12809	rs187429064 (4%)	-0.047 (0.026)	0.075 12809
			rs12151060 (14%)	-0.127 (0.026)	1.5×10 <sup>-6</sup> 12809							

The table shows results for the lead-SNP (column rsID) before and after conditional analysis on the functional candidate variants. In case multiple candidate variants were observed in a locus, multiple linear regression models were fitted to explore the effect of each individual SNP and of all SNPs together. For *CILP2* locus, result of the final model with two SNPs is presented. Results for the lead-SNP in the unconditional analysis are presented from the meta-analysis of Finnish subset ( $N = 12,834$ ). Effect sizes are presented in s.d units. CHR: Chromosome, MAF: Minor allele frequency, SE: Standard error of effect estimate,  $N$ : Number of samples, HDL-C: High-density lipoprotein cholesterol, LDL-C: Low-density lipoprotein cholesterol, TC: Total cholesterol, TG: Triglycerides.