nature genetics

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

Ida Surakka[1,2], Momoko Horikoshi[3,4], Reedik Mägi[5], Antti-Pekka Sarin[1,2], Anubha Mahajan[3], Vasiliki Lagou[3,4], Letizia Marullo[6], Teresa Ferreira[3], Benjamin Miraglio[1], Sanna Timonen[1], Johannes Kettunen[1,2], Matti Pirinen[1], Juha Karjalainen[7], Gudmar Thorleifsson[8], Sara Hägg[9–11], Jouke-Jan Hottenga[12], Aaron Isaacs[13,14], Claes Ladenvall[15], Marian Beekman[16,17], Tõnu Esko[5,18–21], Janina S Ried[22], Christopher P Nelson[23,24], Christina Willenborg[25,26], Stefan Gustafsson[9–11], Harm-Jan Westra[7], Matthew Blades[27], Anton J M de Craen[16,28], Eco J de Geus[12], Joris Deelen[16,17], Harald Grallert[29–31], Anders Hamsten[32], Aki S Havulinna[2], Christian Hengstenberg[33,34], Jeanine J Houwing-Duistermaat[35], Elina Hyppönen[36–39], Lennart C Karssen[13], Terho Lehtimäki[40], Valeriya Lyssenko[15,41], Patrik K E Magnusson[9], Evelin Mihailov[5], Martina Müller-Nurasyid[22,34,42,43], John-Patrick Mpindi[1], Nancy L Pedersen[9], Brenda W J H Penninx[44], Markus Perola[1,2,5,45], Tune H Pers[18–20], Annette Peters[29,31,33], Johan Rung[46], Johannes H Smit[44], Valgerdur Steinthorsdottir[8], Martin D Tobin[47], Natalia Tsernikova[5], Elisabeth M van Leeuwen[13], Jorma S Viikari[48,49], Sara M Willems[13], Gonneke Willemsen[12], Heribert Schunkert[33,34], Jeanette Erdmann[25,26], Nilesh J Samani[23,24], Jaakko Kaprio[1,2,50], Lars Lind[51], Christian Gieger[22,29,31], Andres Metspalu[5,52], P Eline Slagboom[16,17], Leif Groop[1,15], Cornelia M van Duijn[13,14], Johan G Eriksson[2,53–55], Antti Jula[2], Veikko Salomaa[2], Dorret I Boomsma[12], Christine Power[36], Olli T Raitakari[56,57], Erik Ingelsson[3,10,11], Marjo-Riitta Järvelin[2,58–61], Unnur Thorsteinsdottir[8,62], Lude Franke[7], Elina Ikonen[63,64], Olli Kallioniemi[1], Vilja Pietiäinen[1], Cecilia M Lindgren[3,20], Kari Stefansson[8,62], Aarno Palotie[1,20,65–67], Mark I McCarthy[3,4,68], Andrew P Morris[3,5,69], Inga Prokopenko[70] & Samuli Ripatti[1,50,71] for the ENGAGE Consortium

Using a genome-wide screen of 9.6 million genetic variants achieved through 1000 Genomes Project imputation in 62,166 samples, we identify association to lipid traits in 93 loci, including 79 previously identified loci with new lead SNPs and 10 new loci, 15 loci with a low-frequency lead SNP and 10 loci with a missense lead SNP, 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 low-density lipoprotein cholesterol and total cholesterol. Altogether, our results highlight the impact of low-frequency variants in complex traits and show that imputation offers a cost-effective alternative to resequencing.

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

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

Stochastic imputation to individuals genotyped using genotyping arrays in samples of sufficient size offers an alternative and cost-effective design to study the associations of low-frequency and rare variants at a genome-wide level. GWA studies of circulating lipids have been highly successful in identifying loci harboring common variants with small effects[2,3]. In previous large-scale GWA studies, 157 loci have been shown to associate with lipid traits[2,3], but the strongest associations have almost exclusively been reported with common variants (minor allele frequency (MAF) > 5%) in European data sets, owing 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 MAFs (≤5%). Although there are almost 40 loci where both FL SNPs and common SNPs implicated in GWA studies reside, it is often not known whether these associations are driven by the same underlying haplotypes and whether the mendelian variants explain the association in population samples.

A full list of affiliations appears at the end of the paper.

Received 25 June 2014; accepted 16 April 2015; published online 11 May 2015; doi:10.1038/ng.3300

We sought to evaluate the impact of common (MAF > 5%), low-frequency (0.5% < MAF ≤ 5%) and rare (MAF ≤ 0.5%) genetic variants on circulating blood lipids in up to 62,166 European samples by imputing variants into the GWA study cohorts using the sequence-based 1000 Genomes Project reference panel[4] (Phase I interim release, June 2011). We aimed to determine (i) what the role of low-frequency variants and the burden of rare variants were in established lipid-associated loci; (ii) whether a dense set of markers from 1000 Genomes Project–based imputation could help to identify additional loci undetected in previous studies focused largely on common variants imputed to less dense reference panels from the HapMap Project; and (iii) how low-frequency and functional lipid variants contribute to the overall trait variance in comparison 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 study cohorts of European ancestry. Within each cohort, we performed sex-stratified inverse-rank normalization of high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglyceride (TG) and total cholesterol (TC) measures, after adjustment of each trait for age, age^2 and study-specific covariates, including principal components to account for population structure. Case-control studies were further subdivided according to original data selection disease status. The GWA genotype scaffold for each cohort was imputed at up to ~37.4 million autosomal variants from the 1000 Genomes Project multi-ancestry reference panel[4] (Phase I interim release, June 2011). Across a subset of studies, ~98% and ~95% of variants present in the reference panel with 1% < MAF ≤ 5% and 0.5% < MAF ≤ 1%, respectively, were well imputed, as defined here by an IMPUTE[5,6] info score of at least 0.4 (**Supplementary Table 1**). However, as expected, imputation of rare variants (MAF ≤ 0.5%) proved more difficult, although ~65% of the rare variants polymorphic in the reference data set 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 Project imputation, with circulating HDL-C, LDL-C, TG and TC levels. Overall, we detected 93 loci with genome-wide significant association (**Supplementary Fig. 1**) to one or more lipid traits ($P < 5 \times 10^{-8}$), of which 10 loci have not been associated with lipids before (**Table 1** and **Supplementary Figs. 2** and **3**). Of the 83 previously established lipid loci, 79 had a new lead SNP for at least one lipid trait in our analysis (**Supplementary Table 2**). In 34 of the 79 loci, the linkage disequilibrium (LD; r^2) between the new lead SNP and the previously identified lead SNP was ≤40% (15 loci had r^2 ≤ 5%), and, in 56 loci, the newly identified variant was not present in the HapMap 2 imputation reference set used in previous studies. In 11 loci, the newly discovered lead SNP had MAF ≤ 5% and an average effect size of 0.18 (in s.d. units) in comparison 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 winner's curse (n = 5,119; **Fig. 1**). These loci included the well-known lipid gene *LPA* for LDL-C (rs186696265: MAF = 0.8%, effect size = 0.26, $P = 4.4 \times 10^{-14}$, 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, $P = 2.2 \times 10^{-92}$, r^2 = 0.9%) and *APOE* for TC (rs7412: MAF = 7.1%, effect size = 0.41, $P = 7.5 \times 10^{-239}$, r^2 = 1.6%), which were already highlighted in the Global Lipids Genetics Consortium fine-mapping analyses[3].

Using formal conditional analyses, in the *MAFB* locus, the new low-frequency lead SNP with a large effect (effect size > 0.2, MAF ≤ 5%) explained the association of the previously identified lead SNP in 7 population cohorts (n = 12,834), although the LD between these variants was less than r^2 = 5% (**Fig. 1**). Additionally, there were seven loci with two or more associated lead SNPs over 1 Mb apart that had r^2 < 5%, but in all cases the individual-level formal conditional analyses showed that the associations were completely explained by the known lipid-related SNPs in the regions (*ZCCHC11*, *TMEM48* and *PPAP2B* associations explained by rs11591147 in the *PCSK9* locus, olfactory receptor gene cluster association explained by rs7395581 in the *LRP4-MADD* locus, *CCDC79* association explained by rs73591976 in the *LCAT-RANBP10* locus, and *PSG9* and *IRF2BP1* associations explained by rs7412 in the *APOE* locus).

In 5 of the 79 loci, the lead SNP was a missense variant pointing to either a well-established causal gene (*ANGPTL4*, *APOE*, *PCSK9* or *CILP2*) or to a new candidate gene (*ABCA6*). The *APOE* lead SNP for TC, rs7412 (p.Arg176Cys, MAF = 7.1%, r^2 = 0.7%), has been shown to associate with recessive familial type III hyperlipoproteinemia[7,8], and the *PCSK9* lead SNP for LDL-C, rs11591147 (p.Arg46Leu, MAF = 1.9%, r^2 = 0.9%), has been shown to associate with extreme LDL-C values[9]. In the *ANGPTL4* locus, the lead SNP in our GWA data

Table 1 Newly identified loci associated with HDL-C, LDL-C, TC and/or TG concentrations

					Primary	Secondary	Alleles		Meta-analysis		
Locus	Chr.	Position B37	rsID	Annotation	associated trait	associated trait	(effect/other)	EAF	Effect (SE)	P	n
PROX1	1	214,161,820	rs340839[a]	5′ UTR	TG		A/G	0.47	0.039 (0.006)	4.4×10^{-10}	54,836
CEP68	2	65,284,623	rs2540948	Intronic	TG		C/T	0.35	–0.036 (0.006)	6.6×10^{-9}	59,939
PRKAG3	2	219,699,999	rs78058190	Intergenic	HDL-C		A/G	0.05	–0.141 (0.020)	5.7×10^{-12}	52,934
ADAMTS3	4	73,696,709	rs117087731	Intergenic	TC		T/A	0.01	0.308 (0.051)	2.3×10^{-9}	23,641
MTHFD2L	4	75,084,732	rs182616603	Intronic	TC		T/C	0.01	0.374 (0.044)	1.8×10^{-17}	42,905
MTHFD2L	4	75,084,732	rs182616603	Intronic		LDL-C	T/C	0.01	0.314 (0.045)	2.1×10^{-12}	38,420
GPR85	7	112,722,196	rs2255811	3′ UTR	TG		G/A	0.25	0.041 (0.007)	2.3×10^{-8}	59,962
RMI2	16	11,454,650	rs7188861	Intergenic	HDL-C		A/C	0.20	0.044 (0.008)	6.9×10^{-9}	60,578
TM4SF5	17	4,667,984	rs193042029	Intergenic	TG		G/T	0.01	–0.170 (0.029)	8.1×10^{-9}	50,105
GATA6	18	19,907,770	rs79588679	Intergenic	LDL-C		T/C	0.17	–0.049 (0.009)	3.6×10^{-8}	53,108
ZNF274	19	58,681,861	rs117492019	Intergenic	LDL-C		T/G	0.19	–0.047 (0.008)	1.2×10^{-8}	55,371
ZNF274	19	58,671,267	rs12983728	Intergenic		TC	A/G	0.16	–0.046 (0.008)	4.9×10^{-8}	58,904

[a] Present in the HapMap 2 reference panel.

The table presents the association meta-analysis results for the newly identified loci for the four lipid traits tested. 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 cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol; TG, triglycerides.

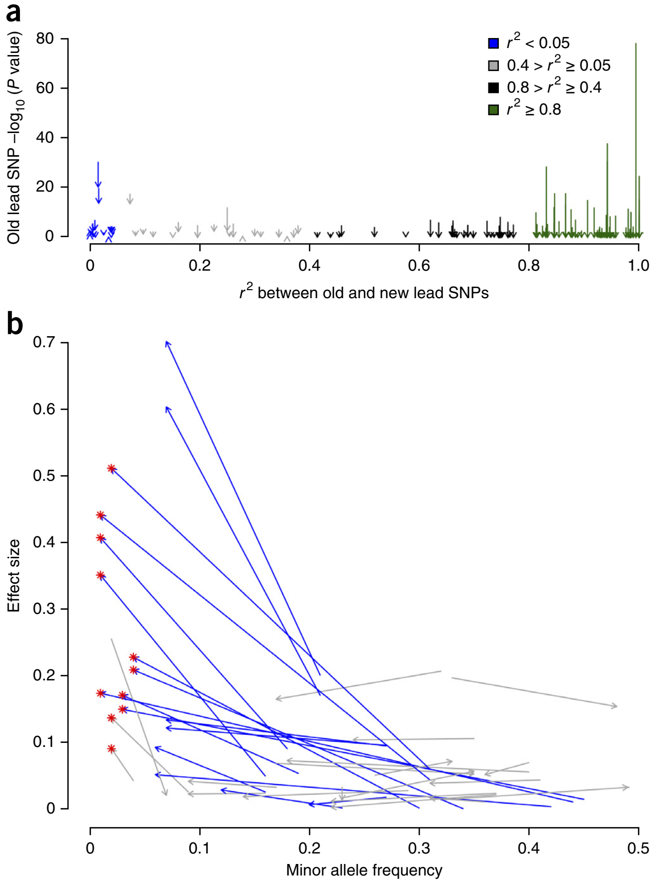

Figure 1 Change in *P* value after analysis conditional on the new lead SNP and comparison of new and previously reported lead SNP effect sizes and allele frequencies per locus. In both plots, each arrow represents one locus and trait where significant association was found in our screening and in one of the previously published large-scale screening studies[2,3]; color is based on the LD between the known and new lead SNPs. (**a**) On the *y* axis are the $-\log_{10}$ (*P* values); arrows start from the *P* value seen in the unconditional analysis in the Finnish subset (n = 12,834) and point to the *P* value in analysis conditional on the new lead SNP. (**b**) Each arrow starts from the effect and MAF for the established lead SNP and points to the corresponding values for the new lead SNP. Red asterisks represent the new low-frequency lead SNPs. Effects have been estimated in the FRCoreExome9702 sample set (n = 5,119), which is independent of the discovery set. For clarity, only results for loci with $r^2 < 0.4$ have been presented.

was a predicted damaging missense variant, rs116843064 (p.Glu40Lys, MAF = 3.0%) with r^2 = 1.8% with the previously associated common lead SNP. The missense variant was associated with TG and HDL-C levels and has previously been associated with extreme TG values[10]. The *CILP2* lead SNP, rs58542926 (in the *TM6SF2* gene encoding a p.Glu167Lys substitution, MAF = 7.8%, r^2 = 98%), was associated with TC levels, risk of myocardial infarction and nonalcoholic fatty liver disease in two papers that appeared while this manuscript was in revision[11,12]. Our new lead SNP in the *ABCA6-ABCA8* locus, rs77542162 (in *ABCA6* encoding a p.Cys1359Arg substitution, MAF = 2.0%, r^2 = 0.6%) associated with LDL-C and TC values ($P = 1.6 \times 10^{-18}$ and 1.9×10^{-13}, respectively).

In the genome-wide screening, we identified ten loci that have not previously been associated with lipids (near *PROX1*, *CEP68*, *PRKAG3*, *ADAMTS3*, *MTHFD2L*, *GPR85*, *RMI2*, *TM4SF5*, *GATA6* and *ZNF274*), with four having a low-frequency variant (MAF < 5%) as the lead SNP (lowest MAF = 0.7%, rs182616603 in the *MTHFD2L* locus; **Table 1**). All except one of the lead SNPs had not been surveyed in the previous GWA studies based on HapMap 2 imputation. The one lead SNP that was present in the HapMap 2 imputation reference panels was in the 5′ UTR of *PROX1* (rs340839, associated with TG, $P = 4.4 \times 10^{-12}$) and was correlated with a marker previously associated with fasting glucose levels and type 2 diabetes[13] (rs340874, r^2 = 74.7%). The lead SNP in the HDL-C–associated *PRKAG3* locus was located upstream of 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[14].

The role of FL SNPs in the general population

In eight loci (*PCSK9*, *CELSR2-SORT1*, *GCKR*, the human leukocyte antigen (HLA) region, *LPL*, *LIPC*, *CETP* and *APOE*), we tested whether 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 by searching the Online Mendelian Inheritance in Man (OMIM) database and confirmed through analysis of the literature, and SNPs previously reported to affect gene transcription or translation in cellular and/or animal models were taken forward into conditional analyses in 7 population cohorts (n = 12,834; **Supplementary Fig. 4** and **Supplementary Tables 3** and **4**).

The FL SNPs explained the lead SNP association (with $P < 5 \times 10^{-8}$ and conditional $P > 0.01$ for the lead SNP) in four of the eight loci (*CELSR2-SORT1*, *GCKR*, *APOE* and *LIPC*; **Table 2** and **Supplementary Fig. 5**). In the *GCKR* and *APOE* loci, the lead SNPs of our GWA screen were FL SNPs (rs1260326 (p.Pro446Leu)[15] and rs7412 (p.Arg158Cys)[7,8] for *GCKR* and *APOE*, respectively). In the *GCKR* locus, rs1260326 explained the population-level association. Similarly, in the *APOE* locus, the two FL SNPs rs7412 and rs429358 (p.Cys112Arg)[16] defining the APOE ε2, ε3 and ε4 isoforms[17] explained the association (**Supplementary Fig. 5d**,**e**). The *LIPC* association was explained by rs1800588 (−514C/T, MAF = 25.1%)[18] and rs113298164 (p.Thr383Met, MAF = 1.4%)[19] for TC and TG (**Supplementary Fig. 5f**,**g**) but not for HDL-C (**Supplementary Fig. 5h**). All results for the conditional analyses are presented in **Supplementary Table 5**.

Search for new functional candidate SNPs

We then searched for potential candidate causal SNPs in the lipid-associated loci (157 established and 10 newly discovered) 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 (*MLXIPL*, *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 **Supplementary Figure 6** and **Supplementary Table 7**. In the *SOST-DUSP3* and *CILP2* loci, the candidate functional variants explained the genome-wide associations of the lead SNPs in the region in the test set (in both loci, conditional $P > 0.01$). In the *SOST-DUSP3* locus (**Fig. 2a**), a single low-frequency, deleterious missense variant, rs72836561 (p.Arg82Cys, MAF = 2.7%, $P = 1.36 \times 10^{-8}$, effect size = 0.23) in the *CD300LG* gene, explained the whole regional association, implicating *CD300LG* as a likely candidate gene in the locus for TG levels. The same variant has also recently been shown to associate with HDL-C and with fasting serum triacylglycerol levels in exome-wide association studies[20,21].

Table 2 Association results of unconditional analysis and analysis conditional on known mendelian and FL SNPs in loci where the functional SNPs explain the genome-wide association

			Lead SNP in the unconditional analysis									
					Unconditional			Conditional				
Locus	Chr.	Trait	rsID	MAF	Effect (SE)	P	n	Covariate SNP(s) in the model	MAF	Effect (SE)	P	n
CELSR2-SORT1	1	LDL-C	rs646776	0.216	0.159 (0.015)	1.31×10^{-25}	12,739	rs12740374	21.6%	0.001 (0.015)	0.958	12,739
		TC	rs646776	0.216	0.123 (0.015)	4.06×10^{-16}	12,834	rs12740374	21.6%	0.001 (0.015)	0.959	12,834
GCKR	2	TG	rs1260326	0.353	0.128 (0.013)	8.44×10^{-23}	12,815	rs1260326	35.3%	NA	NA	NA
LIPC	15	TC	rs1800588	0.251	0.090 (0.015)	7.23×10^{-10}	12,825	rs113298164 rs1800588	1.4% 25.1%	-2×10^{-6} (0.015)	1.000	11,893
LIPC	15	TG	rs686958	0.252	0.085 (0.015)	6.86×10^{-9}	12,801	rs113298164 rs1800588	1.4% 25.1%	0.022 (0.015)	0.152	11,873
APOE	19	LDL-C	rs7412	0.048	0.648 (0.031)	5.93×10^{-95}	12,730	rs7412 rs429358	4.8% 18.1%	NA	NA	NA
APOE	19	TC	rs7412	0.048	0.456 (0.031)	3.10×10^{-49}	12,827	rs7412 rs429358	4.8% 18.1%	NA	NA	NA
APOE	19	TG	rs483082	0.229	0.089 (0.015)	5.74×10^{-9}	12,799	rs7412 rs429358	4.8% 18.1%	NA	NA	NA

The table shows the 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). If 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 the 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; NA, not applicable.

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%, p.Leu156Pro; for TC, effect size = −0.25 and $P = 2.03 \times 10^{-11}$), and a probably damaging missense variant, rs58542926 (MAF = 6.3%, p.Glu167Lys; for TC, effect size = −0.18 and $P = 6.47 \times 10^{-12}$)—explained the lead SNP association for LDL-C, TC and TG (**Fig. 2b**, **Supplementary Fig. 7** and **Supplementary Table 7** illustrate the result of the conditional analysis for TC).

Biological profiling of the *CD300LG* and *TM6SF2* genes

CD300LG (CD300 molecule–like family member G; also called nepmucin) is a type I cell surface glycoprotein that contains a single immunoglobulin V–like domain[22] and has a role in lymphocyte binding and transmigration[23]. The predicted damaging mutation (encoding p.Arg82Cys) in our TG- and HDL-C–associated locus, rs72836561, affects the immunoglobulin 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 in intracellular vesicles[23,24]. Although CD300 family members have been demonstrated to bind lipids[25], the function of CD300LG in lipid metabolism has not been studied. TM6SF2 (transmembrane 6 superfamily member 2)[26] is a multi-pass membrane protein, in which the predicted deleterious missense mutation (rs1874290064; encoding p.Leu156Pro) maps to the predicted fifth transmembrane domain and the probably damaging missense mutation (rs58542926; encoding p.Glu167Lys) maps to the exposed non-transmembrane domain. TM6SF2 has been shown to localize to the endoplasmic reticulum (ER) compartment/ER-Golgi intermediate compartment (ERGIC) and to influence TG secretion in liver cells[27]. Additionally, the p.Glu167Lys missense substitution was shown to alter serum lipid profiles in humans, and knockdown of *Tm6sf2* in mice was shown to lead to increased liver TG content and decreased very-low-density lipoprotein (VLDL) secretion[11,12].

We further characterized these two genes by using the Gene Network database[28] (see the Online Methods for details) for tissue-specific expression, pathway analysis and prediction of mouse knockout phenotype, based on the Mouse Genome Informatics (MGI) database[29]. We found that CD300LG is coexpressed with genes whose knockout increases circulating VLDL particle levels in mice (prediction $P = 1.4 \times 10^{-9}$), 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 coexpression of genes, included abnormal lipid levels (decreased LDL-C, prediction $P = 8.6 \times 10^{-19}$; decreased VLDL, prediction $P = 2.5 \times 10^{-29}$; decreased TC, prediction $P = 6.3 \times 10^{-24}$) among the most highly significant predictions, in line with recent publications and our association results. All associated MGI-based knockout predictions ($P < 1 \times 10^{-6}$) are shown in **Supplementary Table 8**, and lists of genes with the same and stronger MGI-based predictions can be found in **Supplementary Table 9**.

Both genes were found to be among the most highly expressed genes in tissues important for lipid absorption and/or metabolism on the basis of the analysis using the Gene Network database (**Supplementary Table 10**). *CD300LG* is highly expressed in muscles, plasma and adipose tissue, and *TM6SF2* is highly expressed in liver, plasma and intestines. Furthermore, on the basis of the gene expression network analysis, TM6SF2 likely interacts with proteins involved in intestinal absorption (**Supplementary Table 11**), and it is most highly predicted to function as a lipid transporter ($P = 1.05 \times 10^{-14}$, prediction is based on coexpressed genes; **Supplementary Table 12**).

Contribution of low-frequency variants to lipid variation

We estimated the proportion of the variance in lipid traits explained by variants in the 157 previously established and 10 newly identified loci in an additional cohort of 5,119 individuals from the Finrisk cohort (FRCoreExome9702) not included in our discovery meta-analysis. The lead SNPs from all three GWA screens (Teslovich *et al.*[2], Willer *et al.*[3] and this study), together with the FL SNPs and new candidate functional SNPs, were divided into two groups on the basis of their allele frequency in the FRCoreExome9702 data set. Common SNPs explained 8.2% (TG), 11.9% (HDL-C),

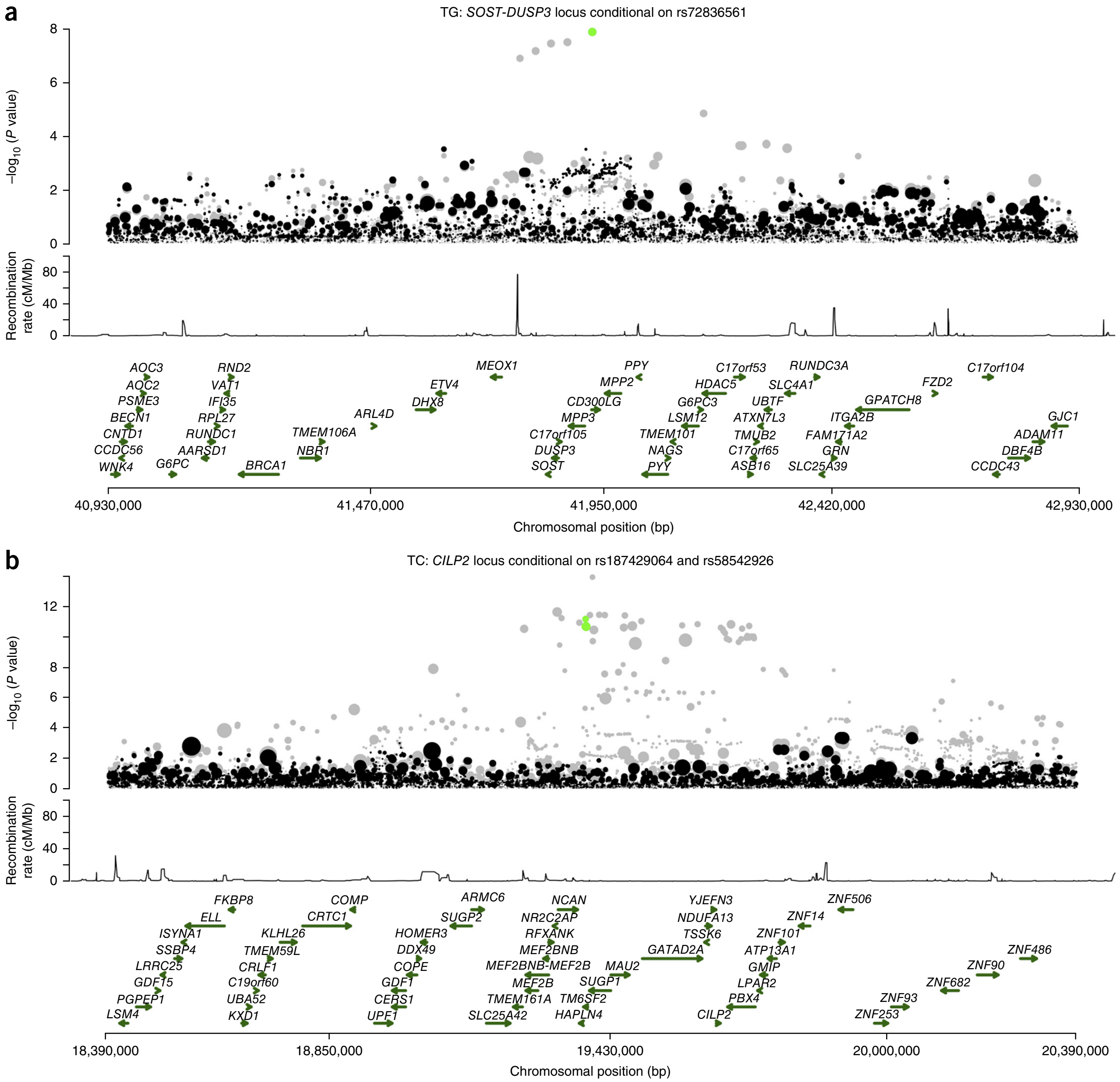

Figure 2 Regional association plots of the conditional analysis in loci where the new candidate functional SNPs explain the genome-wide association. (**a**) Results at the *SOST-DUSP3* locus for TG concentrations in the Finnish subset ($n = 12{,}834$). (**b**) Results at the *CILP2* locus for TC concentrations in the Finnish subset ($n = 12{,}834$). In both plots, the top panel shows the $-\log_{10}$ (*P* value) of each variant as a dot whose size reflects effect size. The middle panel shows the recombination rate in the area, and the bottom panel shows the positions of genes. The *x* axis shows physical position in the genome. In gray are the association results from the unconditional analysis, with green dots representing the new candidate functional SNPs. Black dots are the results from the conditional analysis.

16.3% (LDL-C) and 16.2% (TC) of the variance in lipid levels (**Fig. 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 concentrations, respectively.

We also evaluated the contribution of our SNPs to the additive genetic variance estimated by a linear mixed model (LMM) applied to 10,472 individuals from 6 Finnish GWA cohorts (Online Methods) and to narrow-sense heritability estimates obtained from a large twin study[30]. The narrow-sense heritability estimates from the twin study were 40%, 51%, 51% and 33%, and the LMM estimates derived from the Finnish subset were 26%, 29%, 27% and 19% (for HDL-C, LDL-C, TC and TG, respectively; see the Online Methods for details). We estimate that the SNP set explained 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 in TG, HDL-C, LDL-C and TC concentrations, respectively.

Gene-based association analysis

To complement the single-variant tests for low-frequency variation, we used GRANVIL[31] to test for the association of

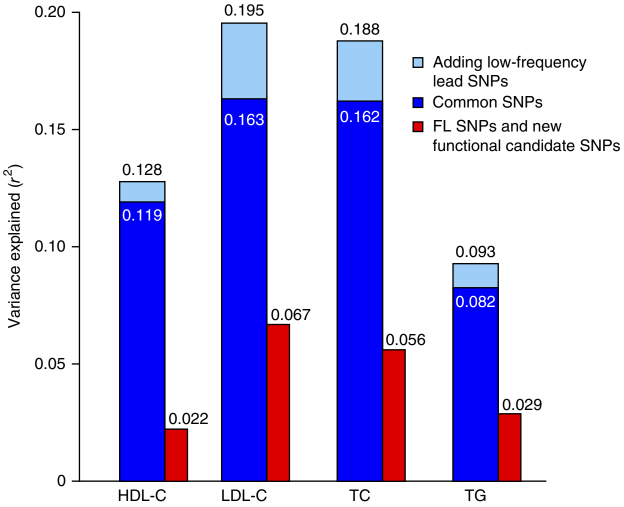

Figure 3 Proportion of total trait variance explained by the lead SNPs and functional SNPs. The proportion of the variance in the trait explained by different SNP sets has been estimated in the independent FRCoreExome9702 sample set (n = 5,119). All lead SNPs from the three association screens (Teslovich *et al.*[2], Willer *et al.*[3] and our screen), together with the known FL SNPs and new candidate functional SNPs, were grouped on the basis of their allele frequency in the FRCoreExome9702 data set into common SNPs (allele frequency > 5%) and low-frequency SNPs (allele frequency ≤ 5%). The variance explained by these two groups is presented with blue bars. The proportion of variance explained by the FL SNPs and candidate functional variants is represented by the red bars.

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 and **Supplementary Table 1**).

We observed genome-wide significant evidence of association ($P < 1.7 \times 10^{-6}$, Bonferroni correction for 30,000 genes) of HDL-C with the mutational load of rare nonsynonymous variants in *LIPC* ($P = 2.1 \times 10^{-7}$, mean MAF = 0.26%; **Supplementary Fig. 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 nonsynonymous variants in *LIPC* remained relatively unchanged (conditional $P = 3.6 \times 10^{-6}$), 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 concentration at genome-wide significance, both mapping to the APO gene cluster: *ZPR1* ($P = 1.5 \times 10^{-11}$, mean MAF = 0.25%) and *APOA5* ($P = 5.0 \times 10^{-8}$, mean MAF = 0.24%). Conditional analyses, adjusting for the lead SNP (rs964184) for the association at the APO gene cluster, reduced the strength of the association of rare variants in both *ZPR1* and *APOA5* with TG concentration but could not fully explain the effect of the mutational load of these genes (**Supplementary Table 13**). As *ZPR1* and *APOA5* map within 2 kb of each other, we further investigated the impact of LD on the association signal in the region with conditional analyses adjusting the association for the mutational load of each gene by that at the other gene (Online Methods). The strength of association of both genes was reduced but not fully attenuated after adjusting for the effect of the other gene (*ZPR1* conditional $P = 1.4 \times 10^{-5}$; *APOA5* conditional $P = 6.3 \times 10^{-4}$), suggesting that the effects of rare variants in these two genes are only partially correlated with each other.

DISCUSSION

Using 1000 Genomes Project–imputed data with a dense SNP set, we were able to impute 9.6 million common and low-frequency SNPs with good quality in 62,166 European samples. With GWA meta-analysis on these data, we identified 10 new loci associated with blood lipid traits and new lead SNPs in 79 previously known lipid-related loci. In 11 previously known loci, the new lead SNP had MAF ≤5%, and, on average, the newly identified low-frequency variants showed an effect size that was 3.6 times greater than that of the corresponding lead SNP in previous meta-analysis studies. Moreover, in four of the ten newly discovered loci, the lead SNPs were low-frequency variants.

Our association results show that low-frequency variants have a much larger contribution to lipid variation in the general population than has previously been shown[2,3]. In several cases, associations that had previously been tagged by common variants are now led by variants with an allele frequency of 0.5–5% and larger effect sizes. The large effect sizes also show in the population variance in lipid traits explained, where low-frequency variants add 3.2% to the variance explained for LDL-C when added to 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.

Although GWA studies have typically identified associations with lipid levels in cohorts with normal population variation, known functional variants—some causing mendelian forms of lipid syndromes and others changing protein structure or disturbing gene transcription—have 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 may generalize to European samples with normal lipid variation. Taken together, the successfully imputed and tested functional SNPs, in combination with the new functional candidate variants, explained 2.2–6.7% of the variation in lipid traits at the population level.

As the FL SNPs explained the population-level association through LD structure in four of the eight loci, we reversed this connection to identify potential candidate genes by finding SNPs with a similar functional profile to the FL SNPs in lipid-associated loci with no previous strong functional candidates. Using this strategy, we identified two loci where missense variants with predicted damaging or deleterious consequences explained the lead SNP associations from the GWA meta-analysis, thus, together with previous evidence, supporting the role of *CD300LG* (TG) and *TM6SF2* (TC, LDL-C and TG) in lipid metabolism, as evidenced by gene network analysis, gene expression correlations, predicted functions in mice and expression patterns across organs, with each type of data suggesting potential links to lipid metabolism. *TM6SF2* was recently listed among genes potentially affecting LDL-C uptake in a small interfering RNA (siRNA) screen focused on cellular lipid phenotypes within previously published blood lipid-associated GWA loci[32]. Additionally, two reports showing strong evidence of a role for one of the two *TM6SF2* missense variants, encoding p.Glu167Lys, on VLDL and TG metabolism were recently published[11,12]. In our data, this variant does not by itself explain the whole regional association; however, the association was explained when this variant was considered together with a second missense variant with lower MAF and larger effect. Overall, our results reinforce the importance of *CD300LG* and *TM6SF2* for blood lipid levels in the general population.

In two established GWA study loci with common lead SNPs, our analyses showed associations of the mutational load of rare variants with lipid traits. The association of HDL-C with rare variants in *LIPC* has previously been reported[33], and we also demonstrate that this signal is independent of the common lead GWA study SNP at this locus. We identified significant association with TG concentration for the accumulation of rare variants in *APOA5*, but conditional analysis on the GWA 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[34], but we report here an impact of this gene on TG concentrations at a population level. Although imputation enables recovery of ~65% of rare variants that are present in the 1000 Genomes Project haplotypes, many will not be represented in the reference panel. Resequencing in large sample sizes will be required to fully elucidate the role of rare variation at these GWA loci on HDL-C and TG levels and to inform functional studies to determine the underlying mechanisms through which these genes influence the regulation of lipids.

In addition to the 93 loci identified, there were 7 loci showing 2 or more association signals that were more than 1 Mb from each other and where the LD between the lead SNPs was 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 also explained the other associations, even over a physical distance of 1 Mb or more or a low level of LD. As these observations were only seen after careful conditional testing of individual-level data, they also highlight how challenging it is to interpret 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, this panel does not cover all low-frequency and rare variants in Europeans. 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 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 sequenced and the resulting data are 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 substantially 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, establishing a similar role for these variants in patient series with extreme phenotypes and in general populations. In addition, we found ten new lipid-associated loci for further investigation, 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 variation in lipid traits. 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 variation in lipid levels in the general population and represent new therapeutic opportunities for treating dyslipidemias and preventing cardiovascular diseases. They also highlight the idea that imputation is a cost-effective approach for assessing association with low-frequency and rare variants without the need for costly resequencing experiments.

URLs. Online Mendelian Inheritance in Man (OMIM) database, http://www.omim.org/; Gene Network database, http://genenetwork.nl/genenetwork/; Mouse Genome Informatics (MGI) database, http://www.informatics.jax.org/; 1000 Genomes Project, http://www.1000genomes.org/; SNPTEST software, http://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html; R: a language and environment for statistical computing, http://r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge CSC (IT Center for Science) and the Technology Centre of the Institute for Molecular Medicine for computational services. The High-Throughput Biomedicine Unit of the Institute for Molecular Medicine Finland, R. Kovanen and A. Uro are acknowledged for technical expertise. M. Jauhiainen is acknowledged for sharing his expertise in the manuscript writing process. This research was supported through funds from the European Union's Seventh Framework Programme (FP7/2007–2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413. I.S. was partly funded by the Helsinki University Doctoral Programme in Biomedicine (DPBM). M.H. was funded by a Manpei Suzuki Diabetes Foundation Grant-in-Aid for young scientists working abroad. A.P.M. and A. Mahajan acknowledge funding from the Wellcome Trust under awards WT098017, WT090532 and WT064890. V. Lagou, L.M., S.H. and I.P. were funded in part through the European Union's Seventh Framework Programme (FP7/2007–2013), ENGAGE project, grant agreement HEALTH-F4-2007- 201413. L.M. was in part sponsored by '5 per mille' contribution assigned to the University of Ferrara, income tax return year 2009, and in part by the ENGAGE Exchange and Mobility Program for ENGAGE training funds. M.D.T. holds a UK Medical Research Council Senior Clinical Fellowship (G0902313). S.T. is supported by the Sigrid Juselius Foundation. J.S.R. and C.G. have received funding from a grant from the RFBR (Russian Foundation for Basic Research)–Helmholtz Joint Research Group (12-04-91322). C.G. received funding from the European Union's Seventh Framework Programme (FP7-Health-F5-2012) under grant agreement 305280 (MIMOmics). M. Perola has been supported by the European Union's Seventh Framework Programme (grant agreements 313010; BBMRI-LPC, 305280; MIMOmics, and 261433; BioSHaRE-EU), Finnish Academy grant 269517, the Yrjö Jahnsson Foundation and the Juho Vainio Foundation. N.J.S. holds a chair funded by the British Heart Foundation (BHF) and is an NIHR Senior Investigator. C.P.N. is funded by the BHF and was preciously funded by the NIHR Leicester Cardiovascular Biomedical Research Unit. V. Salomaa was supported by the Finnish Foundation for Cardiovascular Research and the Finnish Academy (grant 139635). E. Ikonen was supported by the Academy of Finland Centre of Excellence in Biomembrane Research (272130), the Academy of Finland (263841) and the Sigrid Juselius Foundation. V.P. was supported by a University of Helsinki Postdoctoral Researcher grant, the Magnus Ehrnrooth Foundation and the Kymenlaakso Cultural Foundation. O.K., J.-P.M. and V.P. have received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement 258068; EU-FP7-Systems Microscopy Network of Excellence. S.R. was supported by the Academy of Finland (251217 and 255847), the Center of Excellence in Complex Disease Genetics, the European Union's Seventh Framework Programme projects ENGAGE (201413) and BioSHaRE (261433), the Finnish Foundation for Cardiovascular Research, Biocentrum Helsinki and the Sigrid Juselius Foundation. Cohort-specific acknowledgments are provided in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

I.S., M.H., R.M., A.-P.S., A. Mahajan, V. Lagou, L.M., T.F., E. Ikonen, O.K., V.P., C.M.L., U.T., A. Palotie, M.I.M., A.P.M., I.P. and S.R. designed and performed experiments, analyzed data and wrote the manuscript. B.M., S.T., J. Kettunen, M. Pirinen, J. Karjalainen, H.-J.W., J.-P.M., T.H.P. and L.F. performed follow-up experiments and analyzed the data. G.T., S.H., J.-J.H., A.I., C.L., M. Beekman, T.E., J.S.R., C.P.N., C.W. and S.G. analyzed cohort-specific data. H.S., J.E., N.J.S., J. Kaprio, L.L., C.G., A. Metspalu, P.E.S., L.G., C.M.v.D., J.G.E., A.J., V. Salomaa, D.I.B., C.P., O.T.R., E. Ingelsson, M.-R.J. and K.S. designed cohort-specific experiments. M. Blades, A.J.M.d.C., E.J.d.G., J.D., H.G., A.H., A.S.H., C.H., J.J.H.-D., E.H., L.C.K., T.L., V. Lyssenko, P.K.E.M., E.M., M.M.-N., N.L.P., B.W.J.H.P., M. Perola, A. Peters, J.R., J.H.S., V. Steinthorsdottir, M.D.T., N.T., E.M.v.L., J.S.V., S.M.W. and G.W. performed cohort-specific experiments and analyzed cohort-specific data. All authors contributed to the research and reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2011).
2. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
3. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
4. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
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.* **39**, 906–913 (2007).
6. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
7. Rall, S.C., Weisgraber, K.H., Innerarity, T.L. & Mahley, R.W. Identical structural and receptor binding defects in apolipoprotein E2 in hypo-, normo-, and hypercholesterolemic dysbetalipoproteinemia. *J. Clin. Invest.* **71**, 1023–1031 (1983).
8. Rall, S.C., Weisgraber, K.H., Innerarity, T.L. & Mahley, R.W. Structural basis for receptor binding heterogeneity of apolipoprotein F from type III hyperlipoproteinemic subjects. *Proc. Natl. Acad. Sci. USA* **79**, 4696–4700 (1982).
9. Cohen, J.C., Boerwinkle, E., Mosley, T.H. & Hobbs, H.H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
10. Romeo, S. *et al.* Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* **39**, 513–516 (2007).
11. Holmen, O.L. *et al.* Systematic evaluation of coding variation identifies a candidate causal variant in *TM6SF2* influencing total cholesterol and myocardial infarction risk. *Nat. Genet.* **46**, 345–351 (2014).
12. Kozlitina, J. *et al.* Exome-wide association study identifies a *TM6SF2* variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* **46**, 352–356 (2014).
13. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
14. Kahn, B.B., Alquier, T., Carling, D. & Hardie, D.G. AMP-activated protein kinase: ancient energy gauge provides clues to modern understanding of metabolism. *Cell Metab.* **1**, 15–25 (2005).
15. Beer, N.L. *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.* **18**, 4081–4088 (2009).
16. Weisgraber, K.H., Rall, S.C. & Mahley, R.W. Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *J. Biol. Chem.* **256**, 9077–9083 (1981).
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.* **33**, e149 (2005).
18. Deeb, S.S. & Peng, R. The C-514T polymorphism in the human hepatic lipase gene promoter diminishes its activity. *J. Lipid Res.* **41**, 155–158 (2000).
19. Durstenfeld, A., Ben-Zeev, O., Reue, K., Stahnke, G. & Doolittle, M.H. Molecular characterization of human hepatic lipase deficiency. *In vitro* expression of two naturally occurring mutations. *Arterioscler. Thromb.* **14**, 381–385 (1994).
20. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
21. Albrechtsen, A. *et al.* Exome sequencing–driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**, 298–310 (2013).
22. Takatsu, H. *et al.* CD300 antigen like family member G: a novel Ig receptor like protein exclusively expressed on capillary endothelium. *Biochem. Biophys. Res. Commun.* **348**, 183–191 (2006).
23. Umemoto, E. *et al.* Nepmucin, a novel HEV sialomucin, mediates I-selectin–dependent lymphocyte rolling and promotes lymphocyte adhesion under flow. *J. Exp. Med.* **203**, 1603–1614 (2006).
24. Jin, S. *et al.* Nepmucin/CLM-9, an Ig domain–containing sialomucin in vascular endothelial cells, promotes lymphocyte transendothelial migration *in vitro*. *FEBS Lett.* **582**, 3018–3024 (2008).
25. Cannon, J.P., O'Driscoll, M. & Litman, G.W. Specific lipid recognition is a general feature of CD300 and TREM molecules. *Immunogenetics* **64**, 39–47 (2012).
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.* **90**, 255–260 (2000).
27. Mahdessian, H. *et al.* TM6SF2 is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid droplet content. *Proc. Natl. Acad. Sci. USA* **111**, 8913–8918 (2014).
28. Fehrmann, R.S. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).
29. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
30. van Dongen, J., Willemsen, G., Chen, W.M., de Geus, E.J. & Boomsma, D.I. Heritability of metabolic syndrome traits in a large population-based sample. *J. Lipid Res.* **54**, 2914–2923 (2013).
31. Mägi, R. *et al.* Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. *Genet. Epidemiol.* **36**, 785–796 (2012).
32. Blattmann, 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.* **9**, e1003338 (2013).
33. Service, S.K. *et al.* Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.* **10**, e1004147 (2014).
34. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684–687 (2010).

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [2]National Institute for Health and Welfare, Helsinki, Finland. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK. [5]Estonian Genome Center, University of Tartu, Tartu, Estonia. [6]Department of Life Sciences and Biotechnology, Genetic Section, University of Ferrara, Ferrara, Italy. [7]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [8]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [9]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [10]Molecular Epidemiology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden. [11]Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [12]Department of Biological Psychology, EMGO Institute for Health and Care Research, VU University and VU University Medical Center, Amsterdam, the Netherlands. [13]Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands. [14]Centre for Medical Systems Biology, Leiden, the Netherlands. [15]Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Centre, Skåne University Hospital, Malmö, Sweden. [16]Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, the Netherlands. [17]Netherlands Consortium for Healthy Ageing, Leiden, the Netherlands. [18]Division of Endocrinology, Children's Hospital, Boston, Massachusetts, USA. [19]Center for Basic and Translational Obesity Research, Children's Hospital, Boston, Massachusetts, USA. [20]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [21]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [22]Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [23]Department of Cardiovascular Sciences, University of Leicester, Leicester, UK. [24]National Institute for Health Research (NIHR) Leicester Cardiovascular Disease Biomedical Research Unit, Glenfield Hospital, Leicester, UK. [25]Institut für Integrative und Experimentelle Genomik, Universität zu Lübeck, Lübeck, Germany. [26]Deutsches Zentrum für Herz-Kreislauf-Forschung (DZHK), partner site Hamburg, Lübeck and Kiel, Germany. [27]Bioinformatics and Biostatistics Analysis Support Hub (BBASH), University of Leicester, Leicester, UK. [28]Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands. [29]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [30]German Center for Diabetes Research (DZD), Neuherberg, Germany. [31]Institute of Epidemiology II, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [32]Cardiovascular Genetics and Genomics Group, Atherosclerosis Research Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. [33]Deutsches Herzzentrum München, Technische Universität München, Munich, Germany. [34]Deutsches Zentrum für Herz-Kreislauf-Forschung (DZHK), partner site Munich Heart Alliance, Munich, Germany. [35]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands. [36]Population, Policy and Practice, University College London Institute of Child Health, London, UK. [37]South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia. [38]School of Population Health, University of South Australia, Adelaide,

South Australia, Australia. [39]Sansom Institute, University of South Australia, Adelaide, South Australia, Australia. [40]Department of Clinical Chemistry, Fimlab Laboratories and School of Medicine, University of Tampere, Tampere, Finland. [41]Steno Diabetes Center, Gentofte, Denmark. [42]Department of Medicine I, University Hospital Großhadern, Ludwig Maximilians Universität, Munich, Germany. [43]Chair of Genetic Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians Universität, Munich, Germany. [44]Department of Psychiatry, VU University Medical Center, Amsterdam, the Netherlands. [45]Diabetes and Obesity Research Program, University of Helsinki, Helsinki, Finland. [46]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [47]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, UK. [48]Department of Medicine, University of Turku, Turku, Finland. [49]Division of Medicine, Turku University Hospital, Turku, Finland. [50]Department of Public Health, University of Helsinki, Helsinki, Finland. [51]Department of Medical Sciences, Uppsala University, Akademiska Sjukhuset, Uppsala, Sweden. [52]Institute of Molecular and Cell Biology of the University of Tartu, Tartu, Estonia. [53]Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland. [54]Folkhälsan Research Centre, Helsinki, Finland. [55]Unit of Primary Health Care, Helsinki University Hospital, Helsinki, Finland. [56]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland. [57]Department of Clinical Physiology and Nuclear Medicine, University of Turku and Turku University Hospital, Turku, Finland. [58]Department of Epidemiology and Biostatistics, Medical Research Council (MRC) Health Protection Agency (HPA), Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. [59]Institute of Health Sciences, University of Oulu, Oulu, Finland. [60]Biocenter Oulu, University of Oulu, Oulu, Finland. [61]Unit of Primary Care, Oulu University Hospital, Oulu, Finland. [62]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [63]Anatomy, Institute of Biomedicine, University of Helsinki, Helsinki, Finland. [64]Minerva Foundation Institute for Medical Research, Helsinki, Finland. [65]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [66]Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA. [67]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA. [68]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, UK. [69]Department of Biostatistics, University of Liverpool, Liverpool, UK. [70]Department of Genomics of Common Disease, School of Public Health, Imperial College London, Hammersmith Hospital, London, UK. [71]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. Correspondence should be addressed to S.R. (samuli.ripatti@fimm.fi).

ONLINE METHODS

Genotype quality control and imputation. Before imputation, all cohorts (see the **Supplementary Note** for cohort information) went through a quality control pipeline with the following criteria: samples with genotype call rate <95%, sex discrepancies, excess heterozygosity or cryptic relatedness were removed. Additionally, ancestry outliers and multidimensional scaling (MDS) outliers were excluded. SNPs with MAF <1%, with call rate <95% (or <99% if the SNP had MAF <5%) or that failed the Hardy-Weinberg equilibrium exact test (precise threshold depending on study) and sex-chromosome SNPs were removed. Genotyping platforms, study-specific quality control criteria and other details are presented in **Supplementary Table 14**. Imputation for the data sets was performed using IMPUTE (v2.0)[5,6] (unless stated otherwise) with the 1000 Genomes Project June 2011 imputation reference panel with 2,188 haplotypes[4].

Phenotype measures. All four lipid traits (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 15**. Individuals on lipid-lowering medication were excluded, and measures deviating by more than 5 s.d. were set to missing. All four phenotypes were adjusted for age, age^2 and the first three genetic principal components. Principal components were derived from the GWA data using principal-component analysis of the identity-by-state (IBS) sharing matrix for each study separately[35]. Both the removal of outliers and the adjustments were performed 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 metabolic syndrome and type 2 diabetes case status, respectively. Only men were available in the GerMIFS I and II and ULSAM cohorts. As the NTR cohort has related samples, males and females were analyzed together to account for relatedness.

Single-variant association and meta-analysis methods. A GWA analysis was run in each of the cohorts separately (see **Supplementary Table 14** for software details). Quality control on the association results was performed centrally to have as harmonized a data set as possible. In this procedure, the following SNPs were removed: SNPs with minor allele count <3; SNPs with imputation quality Proper_INFO <0.4; duplicates; and genotyped SNPs with Hardy-Weinberg equilibrium $P < 1 \times 10^{-4}$. The meta-analysis was run using the GWAMA software tool[36,37], 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 quality control steps, the maximum number of SNPs in the analysis was 9,657,952.

SNP associations with $P < 5 \times 10^{-8}$ were considered genome-wide significant, and lead SNPs were required to be at least 1 Mb away from adjacent lead SNPs. In areas with long-range LD, 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 OMIM database for information on 167 loci that had been found to associate with one of the studied traits (HDL-C, LDL-C, TC and TG) in either of the 2 previously published GWA studies[2,3] or in our genome-wide screen. For each of these 167 loci, every gene in a 2-Mb region centered on the published lead variant was looked up in the OMIM database, and the variants associated with lipid-related syndromes or population-extreme lipid values were collected. Of the 167 loci, 38 had OMIM-listed lipid-related SNP variants within the searched window. As our genotype data only included SNP variants, we could not study deletions, insertions and other copy number variations. Each of the OMIM-listed lipid-related SNPs was subsequently mapped to genome build 37 using the dbSNP database for rsID identification. Of the 38 loci, 18 had at least one polymorphic OMIM-listed SNP in the imputed Finnish test set of 7 cohorts (Corogene controls, FTC, GenMets, HBCS, NFBC1966, YFS and PredictCVD; combined n = 12,834). To be sure of the functionality of these SNPs, additional literature searching was performed to find evidence of effects on gene transcription or translation. Of the 18 loci, 8 showed genome-wide significant association in the Finnish meta-analysis and had at least one variant with evidence of 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 analyzed separately with linear regression analysis implemented by SNPTEST software. In each cohort, the imputation quality threshold of Proper_INFO > 0.4 was applied. Each locus was analyzed only for the trait(s) for which it had previously been reported in already published GWA studies. In conditional analysis on SNP(s), the phenotype was first adjusted with the SNP(s), and a linear regression model was then fitted for the remaining residuals. When we performed iterative conditional analyses at a locus, the signal was first conditioned on the most significant variant and then conditioned on the top variant from the initial conditional analysis and so forth. Loci where the initial lead SNP association in the conditional analyses had conditional $P < 0.01$ and no further significant associations (conditional $P < 5 \times 10^{-8}$) were found within the 2-Mb window were considered to be explained.

The results from the seven Finnish cohorts were combined using GWAMA. Because the conditional analyses were run only for the preselected 2-Mb windows, genomic inflation factor (λ) correction could not be applied. However, we did not see substantial inflation in the GWA analysis of all four traits in the seven Finnish cohorts (λ ranged from 0.992.991.029, depending on the trait and cohort).

Search for functional candidate SNPs. To explore suggestive functional variants causing association signals at loci that do not have lipid-related OMIM-listed variants, we selected nine loci: *GALNT2*, *MLXIPL*, *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[2,3] or in our genome-wide screen, as well as in the meta-analysis using 7 Finnish cohorts (n = 12,834). In each of these loci, 2-Mb windows were searched for functional variants that had association $P < 5 \times 10^{-4}$. Candidate SNPs were annotated using the Ensembl database, and functional effects were predicted using the Provean[38], SIFT[39] and PolyPhen[40] databases. If a variant was annotated as a missense mutation predicted to be damaging in at least one of the prediction databases, it was treated as an FL SNP, and formal conditional analysis was performed to investigate whether it explained the association.

Gene network analysis. We used 2,206 principal components that had been derived from a data set of 77,840 samples using Affymetrix microarrays (54,736 human, 17,081 mouse and 6,023 rat). Because gene set enrichment analysis showed that each of these components was enriched for at least one biological pathway, we used these components to develop a gene function prediction algorithm. To do so, we first determined whether each of the components was enriched for a given gene set by performing a t test (contrasting genes known to be part of this pathway with all other genes) and transformed the T statistics into z scores. Subsequently, we calculated 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.*[28]; see Cvejic *et al.*[41] and Wood *et al.*[29] for a short description). We used a permutation strategy to determine the significance of the predictions, controlling the false discovery rate at 5%. On the basis of the MGI mouse knockout database, we predicted that *CD300LG* would increase circulating VLDL-C levels. For *TM6SF2*, the most significantly predicted biological process was intestinal absorption. Only highly significant predictions (permuted $P < 1 \times 10^{-6}$) were taken into account when profiling the two genes.

We text-mined the sample descriptions provided by experimenters who uploaded microarray data to the Gene Expression Omnibus (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* was highly expressed in adipose tissue, heart, muscle and plasma and that *TM6SF2* was highly expressed in ileum and intestinal mucosa.

doi:10.1038/ng.3300

Modeling the proportion of variance explained. To estimate the phenotypic variance explained by different types of SNPs, we ran multiple linear regression models in R using the FRCoreExome9702 data set ($n = 5{,}119$), an independent sample set from the Finrisk cohort. For the models, all lead SNPs (Teslovich *et al.*[2], Willer *et al.*[3] and our study), together with FL SNPs and new functional candidates, were divided into two groups on the basis of the MAF of each variant in the FRCoreExome9702 data set. The tested SNP sets were (i) common (MAF > 5%) lead SNPs and functional SNPs; (ii) low-frequency (MAF ≤ 5%) lead SNPs and functional SNPs added to SNP set (i); and (iii) FL SNPs and the three identified functional candidates.

These SNP sets were used to explain the variation in trait residuals adjusted for sex, age, age^2 and population stratification. 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 explained using the linear mixed-model approach implemented in GCTA[42] (v.1.13). This estimate is a lower bound of the total additive genetic variance, as 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 in 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[31] was used to test for association of each trait with accumulations of minor alleles (mutational load) at successfully imputed rare variants (MAF ≤ 1% and Proper_INFO ≥ 0.4; **Supplementary Table 1**) within genes in a linear regression framework: (i) irrespective of annotation and (ii) restricted to nonsynonymous 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 $P < 1.7 \times 10^{-6}$, corresponding to a Bonferroni correction for 30,000 genes. Conditional analyses were performed to assess the evidence for association of traits with 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 the effects of rare variants in two genes by including the mutational load of one as a covariate in the regression model for trait association with the other.

35. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
36. Mägi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
37. Mägi, R., Lindgren, C.M. & Morris, A.P. Meta-analysis of sex-specific genome-wide association studies. *Genet. Epidemiol.* **34**, 846–853 (2010).
38. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, e46688 (2012).
39. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
40. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
41. Cvejic, A. *et al. SMIM1* underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* **45**, 542–545 (2013).
42. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

doi:10.1038/ng.3300