nature genetics

Exome chip meta-analysis identifies novel loci and East Asian–specific coding variants that contribute to lipid levels and coronary artery disease

Xiangfeng Lu[1–3], Gina M Peloso[4,5], Dajiang J Liu[6], Ying Wu[7], He Zhang[2,3], Wei Zhou[8], Jun Li[9], Clara Sze-man Tang[10], Rajkumar Dorajoo[11], Huaixing Li[12], Jirong Long[13], Xiuqing Guo[14], Ming Xu[15], Cassandra N Spracklen[7], Yang Chen[16], Xuezhen Liu[9], Yan Zhang[17], Chiea Chuen Khor[11,18,19], Jianjun Liu[11], Liang Sun[12], Laiyuan Wang[1], Yu-Tang Gao[20], Yao Hu[12], Kuai Yu[9], Yiqin Wang[12], Chloe Yu Yan Cheung[21], Feijie Wang[12], Jianfeng Huang[1,22], Qiao Fan[19,23], Qiuyin Cai[13], Shufeng Chen[1], Jinxiu Shi[24], Xueli Yang[1], Wanting Zhao[19], Wayne H-H Sheu[25], Stacey Shawn Cherny[26–28], Meian He[9], Alan B Feranil[29,30], Linda S Adair[31], Penny Gordon-Larsen[31,32], Shufa Du[31,32], Rohit Varma[33], Yii-Der Ida Chen[14], Xiao-Ou Shu[13], Karen Siu Ling Lam[21,34,35], Tien Yin Wong[18,23,36,37], Santhi K Ganesh[2,3], Zengnan Mo[16], Kristian Hveem[38–40], Lars G Fritsche[38,39,41], Jonas Bille Nielsen[2], Hung-fat Tse[21,34,42], Yong Huo[17], Ching-Yu Cheng[19,37,43], Y Eugene Chen[2], Wei Zheng[13], E Shyong Tai[23,36,44], Wei Gao[15], Xu Lin[12], Wei Huang[24], Goncalo Abecasis[41], GLGC Consortium[45], Sekar Kathiresan[4,46], Karen L Mohlke[7], Tangchun Wu[9], Pak Chung Sham[26–28,47], Dongfeng Gu[1,47] & Cristen J Willer[2,3,8,47]

Most genome-wide association studies have been of European individuals, even though most genetic variation in humans is seen only in non-European samples. To search for novel loci associated with blood lipid levels and clarify the mechanism of action at previously identified lipid loci, we used an exome array to examine protein-coding genetic variants in 47,532 East Asian individuals. We identified 255 variants at 41 loci that reached chip-wide significance, including 3 novel loci and 14 East Asian–specific coding variant associations. After a meta-analysis including >300,000 European samples, we identified an additional nine novel loci. Sixteen genes were identified by protein-altering variants in both East Asians and Europeans, and thus are likely to be functional genes. Our data demonstrate that most of the low-frequency or rare coding variants associated with lipids are population specific, and that examining genomic data across diverse ancestries may facilitate the identification of functional genes at associated loci.

Genome-wide association studies (GWASs) have identified over 175 genetic loci that contribute to lipid levels[1–6], which are heritable risk factors for cardiovascular disease, fatty liver disease, age-related macular degeneration, and type 2 diabetes[7–9]. However, most of the published lipid-associated variants are found in non-protein-coding regions of the genome, are without obvious biological significance, and explain only a small fraction of the heritability of lipid levels. The examination of low-frequency and potentially functional variants, which are poorly captured by standard GWAS arrays, has the potential to pinpoint causal variants and genes for follow-up and functional analyses, thereby promoting translation of the findings of genetic studies into new therapeutic targets. For example, low-frequency coding variants in *PCSK9* reduce plasma levels of low-density lipoprotein cholesterol (LDL-C), reduce the risk of coronary artery disease (CAD), and have prompted the development of a new class of therapeutics[10]. Thus, we investigated the effect on lipid levels of the rare and low-frequency variants in the coding portion of the genome in an East Asian population, as East Asians have not been as extensively studied as the European population[11–13].

We carried out a meta-analysis of exome-wide association studies of blood lipid levels (high-density lipoprotein cholesterol (HDL-C), LDL-C, triglycerides (TGs), and total cholesterol (TC)) in a total of 47,532 East Asian samples that were genotyped by exome array. We further integrated the exome array data for plasma lipids in over 300,000 individuals, primarily of European ancestry (84%), from a study conducted by the Global Lipids Genetics Consortium (GLGC)[14]. We aimed to determine whether novel or population-specific variants and genes that influence lipid levels could be identified in a meta-analysis of East Asian and multi-ancestry sample groups. We also aimed to determine whether the protein-altering variants in known

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

Received 16 October 2016; accepted 26 September 2017; published online 30 October 2017; doi:10.1038/ng.3978

lipid loci explained the association signal or were independent evidence of functional genes. Finally, we examined whether exome data implicated the same putatively functional genes at lipid loci in both European and East Asian cohorts.

RESULTS

To improve the coverage for the low-frequency variants in Asian populations and follow up on various GWAS variants, we added approximately 60,000 custom-content variants to the standard exome array. Among 319,272 variants that passed quality control, 204,408 (64.0%) were polymorphic in East Asian individuals, of which about 25% (50,126) were from the custom content. Approximately 76.1% (155,566) of the polymorphic variants were annotated as nonsynonymous or loss-of-function (stop-gain, stop-loss, and splice variants) (**Supplementary Table 1**). By determining the proportion of variants observed in Exome Aggregation Consortium (ExAC) East Asian samples (n = 4,327 individuals) that were successfully genotyped by the array, we estimated that the exome array captured a large fraction of common and low-frequency coding variants (71.15% and 72.59% for variants with minor allele frequency (MAF) > 5% and MAF = 1–5%, respectively). Among rare coding variants identified in ExAC-sequenced individuals, 59.91% (MAF = 0.1–1%) and 19.92% (two or more copies) were captured by the array. Therefore, the array provided good coverage for low-frequency variants and moderate coverage for rare coding variants in East Asians. In addition, we examined 76,000 polymorphic coding variants that were unavailable or monomorphic in ExAC East Asian samples.

Discovery of novel variants associated with lipid levels

Our analysis identified three variants with study-wide significance in three novel loci in East Asians, located at least 1 Mb from previously reported GWAS signals of lipid levels (**Table 1**). These were rs4377290 in *ACVR1C* (TC; $P = 4.69 \times 10^{-8}$), rs7901016 in *MCU* (LDL-C; $P = 5.12 \times 10^{-9}$), and the missense variant rs4883263 (encoding p.Ile342Val) in *CD163* (HDL-C; $P = 5.24 \times 10^{-11}$). Each of these three variants demonstrated evidence for association ($P = 1.80 \times 10^{-3}$ to 6.68×10^{-5}) in over 300,000 GLGC individuals.

Summary of association results

We assessed the association of 110,986 polymorphic variants that had at least 20 minor alleles in 47,532 East Asian samples. Overall, we detected 255 variants (including 51 coding variants) at 41 loci with exome-wide significant association with one or more lipid traits ($P < 4.5 \times 10^{-7}$), of which 3 loci had not been previously reported (**Fig. 1**). Collectively, the overall variance in each lipid trait that could be explained by exome-wide significant variants in East Asian samples was 5.97% for TC, 6.20% for LDL-C, 6.93% for HDL-C, and 6.89% for TG levels, of which 3.22%, 4.77%, 3.35%, and 3.86%, respectively, could be attributed to coding variants (**Fig. 2**). Our results also showed that an additional seven known loci were associated with lipid levels with suggestive significance ($P < 4.46 \times 10^{-6}$, Bonferroni correction of 11,215 variants) (**Supplementary Table 2**), and that, taken together, they increased the percentage of trait variance explained to 6.08–7.20%.

Evaluation of known lipid signals

Among the 38 previously established lipid loci that reached significance, we identified a more significant candidate variant at 14 loci (**Supplementary Table 3** and **Fig. 1**) where the initially reported GWAS index variants showed no significant associations or were independent of our lead variants ($r^2 < 0.02$) (*APOB* and *APOE*), demonstrating allelic heterogeneity between people of East Asian ancestry and those of European ancestry. The lead variants in the remaining 24 loci were the same as or strongly related to ($r^2 > 0.69$) the reported GWAS index variants from previous studies in primarily European samples. Sequential conditional analyses showed that 12 loci with evidence of association had two or more significant signals (**Supplementary Table 4**). For example, we detected a novel missense variant (rs2075260, encoding p.Val2141Ile) at *ACACB* that was largely independent of the originally reported GWAS index variant rs7134594 at *MVK* ($r^2 = 0.01$)[2], and thus represented a previously unreported association. The GWAS index variant rs7134594 could be explained by another missense variant (rs9593, encoding p.Met239Lys) at *MMAB* (conditional $P = 0.73$).

In gene-based analysis, nine genes (*PCSK9*, *EVI5*, *HMGCR*, *CD36*, *APOA1*, *PCSK7*, *CETP*, *LDLR*, and *PPARA*) reached gene-based significance ($P < 2.8 \times 10^{-6}$) in connection with lipid levels (**Supplementary Fig. 1** and **Supplementary Table 5**). However, our gene-based analyses did not identify any new genes that had not already been highlighted by single-variant tests.

Putative functional coding variants at known loci

The identification of coding variants in known loci has the potential to pinpoint causal genes. We observed that the protein-altering variants were more likely to have strong effect sizes with regard to lipid levels (**Fig. 3** and **Supplementary Table 6**) compared with the non-coding variants that were significantly associated with lipid levels. Ten coding variants in eight genes showed strong effects on lipid levels (β-coefficients ranging from 0.20 to 1.17 s.d.), and eight were low-frequency or rare variants (MAF < 3%). We next sought to quantify what proportion of GWAS loci might be due to a protein-altering variant, and thus implicate a candidate functional gene. We made the reasonably well-supported assumption that a protein-altering variant that is the top signal, explains the signal, or is independent of the original signal is the most likely causal variant for each region[15–17]. Among the 38 known loci for which association evidence attained study-wide significance, 12 loci harbored a protein-altering variant that showed the strongest association with lipid levels, and 4 loci had a protein-altering variant that was not the top signal but could explain the association of the reported index variant (**Supplementary Table 7** and **Fig. 1**). In 8 of these 16 loci (*PCSK9*, *EVI5*, *CD36*, *MMAB*, *ALDH2*, *SLC12A4*, *LDLR*, and *PPARA*), the previously identified lead variants in European populations did not reach exome-wide significance.

Table 1 Genetic variants at novel loci associated with lipid levels in East Asian samples

						East Asian					GLGC		Combined		
Gene	rsID	Position	Alleles	Variants	Trait	AAF	β	S.e.m.	P	N	AAF	P	AAF	P	I^2 (%)
ACVR1C	rs4377290	2:158437683	C/T		TC	0.33	–0.039	0.007	4.69×10^{-8}	46,025	0.46	1.59×10^{-4}	0.44	6.06×10^{-8}	16.2
MCU	rs7901016	10:74637326	C/T		LDL-C	0.27	–0.044	0.008	5.12×10^{-9}	44,985	0.09	1.80×10^{-3}	0.12	2.21×10^{-9}	18.2
CD163	rs4883263	12:7649484	C/T	p.Ile342Val	HDL-C	0.69	–0.047	0.007	5.24×10^{-11}	47,456	0.94	6.68×10^{-5}	0.90	6.30×10^{-13}	2.38

AAF, alternative allele frequency. Positions are reported in human genome build hg19. Alleles are listed as alternative/reference alleles on the forward strand of the reference genome.

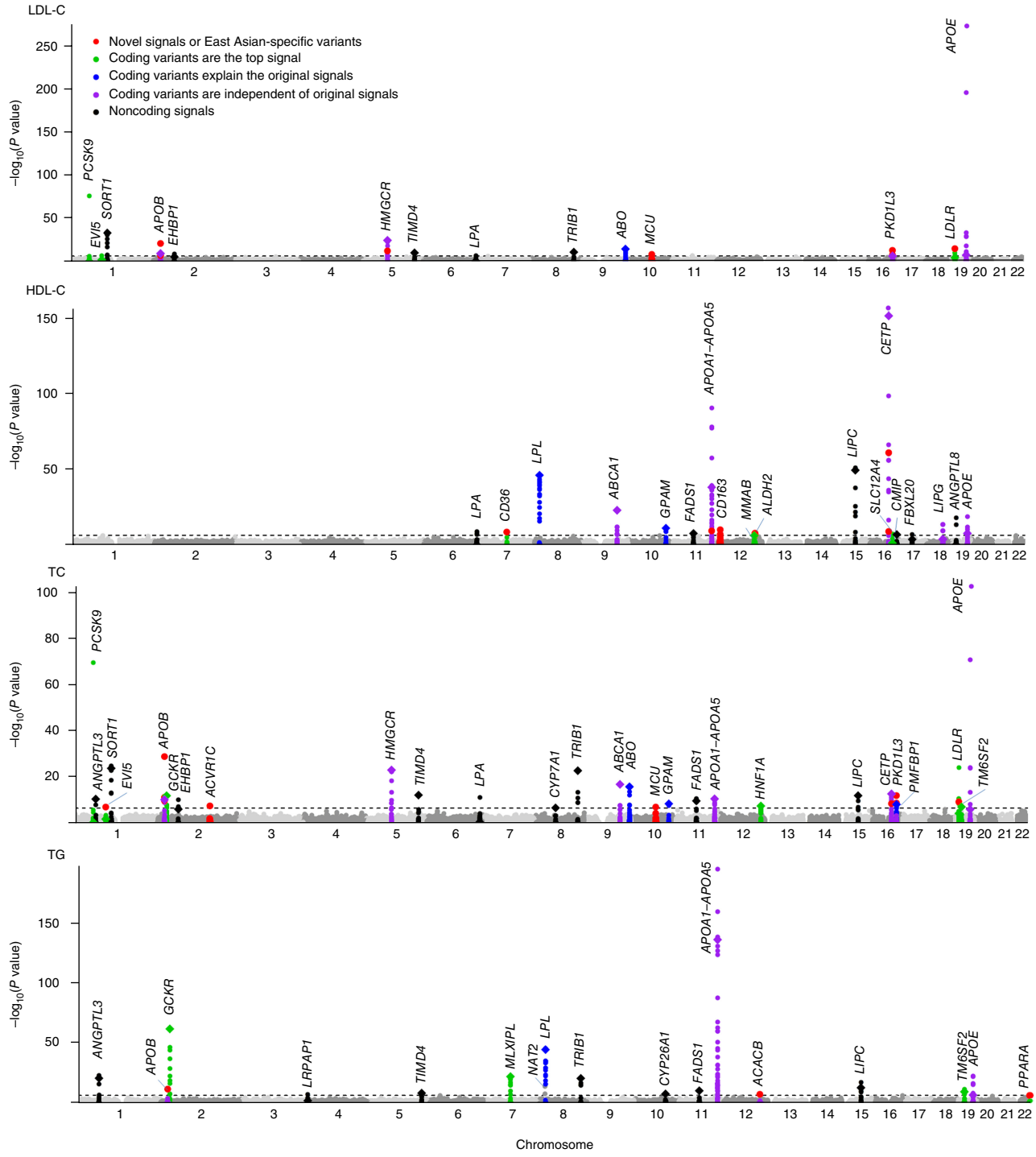

Figure 1 Exome-wide association results for 47,532 East Asians. The Manhattan plots show $-\log_{10}P$ values for variants associated with LDL-C, HDL-C, TC, and TG. Signals with exome-wide significance (horizontal dashed lines; $P < 4.5 \times 10^{-7}$) are highlighted, and the previously reported GWAS lead variants of each region are indicated by diamonds. East Asian–specific variants are defined as variants with conditional *P* values that reached exome-wide significance after conditioning on all independent variants in the corresponding loci identified by GLGC exome-wide association studies.

In the remaining eight loci (*GCKR*, *MLXIPL*, *HNF1A*, *LPL*, *ABO*, *GPAM*, *PMFBP1*, and *TM6SF2*), the GWAS index variant in each locus (*P* values ranged from 4.86×10^{-8} to 1.26×10^{-62}) was in strong linkage disequilibrium (LD) with the corresponding protein-altering variant ($r^2 > 0.67$) and did not remain significant after the effect of the protein-altering variant was accounted for (conditional *P* values > 0.01), which suggested that the index variant might act as a proxy for the functional protein-altering variant. Together, 42.1% (16/38) of loci seemed to have a protein-altering variant that could account for the original association signal. In addition, we identified 15 protein-altering variants in nine genes (*APOB*, *HMGCR*, *ABCA1*, *APOA1–APOA5*, *ACACB*, *CETP*, *PKD1L3*, *LIPG*, and *APOE*) that were independent of

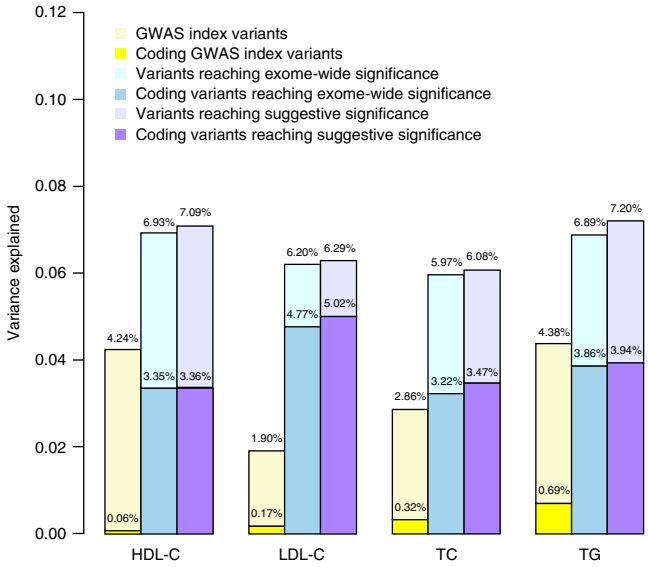

Figure 2 The proportion of total trait variance explained by significant and coding variants. The variance explained by all the variants that reached exome-wide significance ($P < 4.5 \times 10^{-7}$) and by the variants with suggestive significance ($P < 4.46 \times 10^{-6}$) is indicated by light blue and purple bars, respectively. The proportion of variance explained by the corresponding protein-altering variants is represented by dark blue and purple bars. The proportion of variance explained by GWAS index variants is represented by yellow bars.

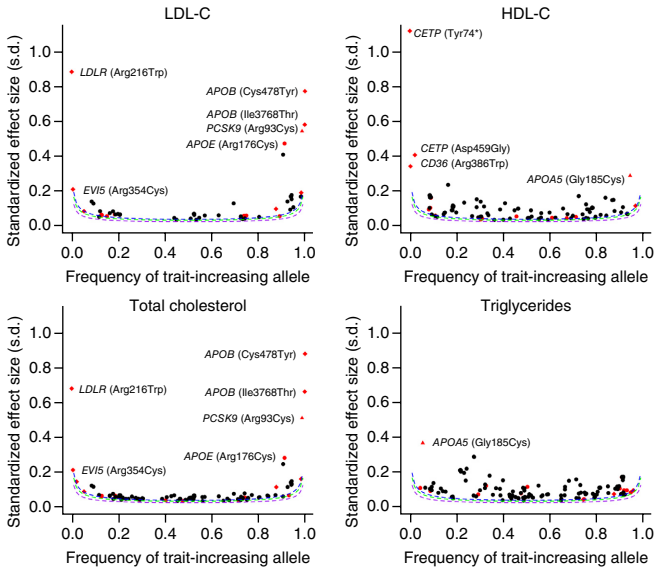

Figure 3 Effect size versus allele frequency for variants associated with blood lipids at exome-wide significance. Protein-altering variants are indicated by red symbols, and noncoding variants by black symbols. East Asian–specific protein-altering variants are labeled by diamonds. The variants indicated by triangles, *PCSK9* (p.Arg93Cys) and *APOA5* (p.Gly185Cys), have extremely rare MAFs in Europeans, although they do not show population-specific association. The protein-altering variants with strong effects on lipid levels (β-coefficient > 0.20 s.d. units) are highlighted. Estimated power curves are shown (dashed lines) for the minimum standardized effect sizes (in s.d. units) that could be identified for a given effect–allele frequency relationship with 10% (purple), 50% (green), and 80% (blue) power, assuming a sample size of 47,532 and an α-level of 4.5×10^{-7}.

the original signal but may highlight functional genes in the region. All of these putative functional variants may point to functional candidate genes—either well-established causal genes (such as the genes that cause Mendelian dyslipidemias (**Supplementary Table 8**)) or potential new candidate genes (*MMAB*, *ACACB*, *SLC12A4*, and *PMFBP1*). In total, the 31 protein-altering variants in the known loci may point to 25 candidate functional lipid genes.

Association with coronary artery disease

To further evaluate whether the novel variants and putative functional variants in known regions identified in our samples also influenced CAD risk, we tested for association in 28,899 Chinese individuals with and without coronary disease (9,661 CAD cases and 18,558 controls) and in the largest publicly available CAD GWAS analysis (CARDIoGRAMplusC4D),which includes ~185,000 CAD cases and controls[18] (**Supplementary Table 9**). For the novel noncoding variant near *MCU* (rs7901016), the C allele associated with lower levels of LDL-C was similarly associated with reduced risk for CAD in Chinese samples (odds ratio (OR) = 0.94; 95% confidence interval (CI) = 0.90–0.98; $P = 2.8 \times 10^{-3}$) and CARDIoGRAMplusC4D samples (OR = 0.94; 95% CI = 0.91–0.98; $P = 4.55 \times 10^{-4}$). Among the 31 putative functional coding variants in the known regions, all 20 non-HDL-C-related variants showed a consistent direction of effect between lipid traits and CAD. Fifteen out of 20 showed nominal significance ($P < 0.05$) in Chinese or CARDIoGRAMplusC4D CAD data, whereas 7 variants in *PCSK9*, *APOB*, *LDLR*, *APOE*, *HNF1A*, and *APOA5* showed significant associations even after multiple testing was accounted for (P values ranged from 5.95×10^{-4} to $8.17 \times 10^{-11} < 0.05/31$). In particular, nearly all of the LDL-associated coding variants demonstrated association with CAD, and the strengths of effect on CAD risk and LDL-C level were strongly correlated ($r^2 = 0.78$; $P = 3.3 \times 10^{-4}$; **Supplementary Fig. 2**).

Novel loci identified in East Asian and GLGC samples

An exome-wide association screen for plasma lipids in >300,000 individuals genotyped by exome array was conducted in parallel by the GLGC[14]. The majority (84%) of the participants were of European ancestry, and only 2.3% were of East Asian ancestry. We further carried out a large-scale trans-ancestry meta-analysis of our East Asian and GLGC samples, being careful to include overlapping samples only once, to seek both novel and population-specific genetic variants for lipid levels.

In the combined GLGC and East Asian samples, nine additional variants that were not significant in the East Asian or GLGC analyses showed significant association ($P < 2.1 \times 10^{-7}$, Bonferroni correction of 242,289 variants analyzed by the GLGC) with at least one lipid trait. All of them were common (MAF > 0.05 in both East Asian and GLGC samples), including four coding variants (**Table 2** and **Supplementary Fig. 3**): *FAM114A2* (p.Gly122Ser; HDL-C; $P = 1.74 \times 10^{-7}$), *MGAT1* (p.Leu435Pro; HDL-C; $P = 9.36 \times 10^{-8}$), *ASCC3* (p.Leu146Phe; LDL-C, $P = 5.84 \times 10^{-8}$; TC, $P = 5.22 \times 10^{-9}$), and *PLCE1* (p.Arg1575Pro; TC; $P = 9.92 \times 10^{-8}$).

Joint analysis of novel signals with additional samples

To strengthen support for the observed associations, we carried out *in silico* replication of significant variants in three additional independent genome-wide data sets comprising a combined total of ~160,000 individuals from the Nord-Trøndelag Health Study[19], GLGC GWAS samples[2], and a Chinese lipids GWAS[20]. We found that the associations of 12 novel variants achieved greater significance

Table 2 Variants at novel loci associated with lipid levels identified from combined East Asian and GLGC samples

					Combined							GLGC		East Asian	
Gene	rsID	Position	Alleles	Variants	Trait	AAF	β	S.e.m.	*P*	*N*	I^2 (%)	AAF	*P*	AAF	*P*
PDGFC	rs4691380	4:157720124	T/C		HDL-C	0.35	0.014	0.003	1.07×10^{-7}	335,481	0.54	0.36	2.80×10^{-7}	0.31	0.14
FAM114A2	rs2578377	5:153413390	T/C	p.Gly122Ser	HDL-C	0.67	–0.014	0.003	1.74×10^{-7}	335,481	4.25	0.63	2.35×10^{-7}	0.87	0.36
MGAT1	rs634501	5:180218668	G/A	p.Leu435Pro	HDL-C	0.72	–0.015	0.003	9.36×10^{-8}	337,027	1.70	0.76	2.35×10^{-5}	0.52	3.96×10^{-4}
ASCC3	rs9390698	6:101296389	A/G	p.Leu146Phe	LDL-C	0.39	0.014	0.003	5.84×10^{-8}	331,991	0.40	0.41	1.15×10^{-6}	0.26	1.18×10^{-2}
					TC	0.39	0.015	0.003	5.22×10^{-9}	358,251	0.70	0.41	1.89×10^{-7}	0.26	5.05×10^{-3}
C6orf183	rs884366	6:109574095	A/G		HDL-C	0.31	–0.015	0.003	1.45×10^{-8}	327,673	0.04	0.30	4.06×10^{-6}	0.38	1.88×10^{-4}
EEPD1	rs4302748	7:36191699	A/G		LDL-C	0.18	0.018	0.003	2.10×10^{-8}	333,359	4.30	0.20	5.55×10^{-7}	0.09	3.82×10^{-3}
PLCE1	rs2274224	10:96039597	C/G	p.Arg1575Pro	TC	0.44	–0.020	0.004	9.92×10^{-8}	150,798	17.73	0.44	2.80×10^{-7}	0.56	0.41
EIF4B	rs7306523	12:53393964	G/A		LDL-C	0.70	–0.017	0.003	1.38×10^{-7}	313,750	1.10	0.77	1.75×10^{-5}	0.29	1.27×10^{-3}
					TC	0.70	–0.017	0.003	5.36×10^{-8}	338,266	0.00	0.77	1.42×10^{-6}	0.29	1.15×10^{-2}
SLC17A8	rs7965082	12:100800193	T/C		LDL-C	0.52	–0.013	0.002	9.21×10^{-8}	333,359	0.00	0.54	1.89×10^{-6}	0.41	1.31×10^{-2}
					TC	0.52	–0.014	0.002	8.28×10^{-9}	358,251	0.00	0.54	1.47×10^{-6}	0.41	3.86×10^{-4}

AAF, alternative allele frequency. Positions are reported in human genome build hg19. Alleles are listed as alternative/reference alleles on the forward strand of the reference genome.

Table 3 Inter-ancestry allelic heterogeneity at lipid genes

				Protein-altering variants									Lookup in the other sample	
Gene	Study	rsID	Note[a]	Position	Variants	Alleles	Trait	β	S.e.m.	*P*	AAF (%)	Variance explained (%)	AAF (%)	*P*
PCSK9	GLGC	rs11591147	Protein-altering is top	1:55505647	p.Arg46Leu	T/G	LDL-C	–0.475	0.011	0.00	1.48	0.70	0.01	0.26
	Asian	rs151193009	Protein-altering is top	1:55509585	p.Arg93Cys	T/C	LDL-C	–0.542	0.029	7.62×10^{-77}	1.32	0.77	0.01	7.62×10^{-9}
APOB	GLGC	rs1367117	Explaining index	2:21263900	p.Thr98Ile	A/G	LDL-C	0.105	0.003	3.61×10^{-278}	28.44	0.43	13.01	4.26×10^{-10}
	Asian	rs13306194	Protein-altering is top	2:21252534	p.Arg532Trp	A/G	LDL-C	–0.098	0.010	9.53×10^{-22}	12.45	0.20	0.20	8.13×10^{-3}
CD36	GLGC	rs3211938	Protein-altering is top	7:80300449	p.Tyr325*	G/T	HDL-C	0.181	0.021	1.43×10^{-18}	0.47	0.03	0.001	0.87
	Asian	rs148910227	Protein-altering is top	7:80302116	p.Arg386Trp	T/C	HDL-C	0.342	0.058	3.17×10^{-9}	0.31	0.07	0.02	0.01
ABCA1	GLGC	rs146292819	Independent of index	9:107556776	p.Asn1800His	G/T	HDL-C	–0.843	0.059	3.99×10^{-46}	0.05	0.07	0.00	NA
	Asian	rs2230808	Independent of index	9:107562804	p.Lys1587Arg	C/T	HDL-C	0.047	0.007	2.49×10^{-12}	60.97	0.10	72.96	9.78×10^{-19}
CETP	GLGC	rs5880	Independent of index	16:57015091	p.Ala330Pro	C/G	HDL-C	–0.258	0.007	4.08×10^{-321}	4.81	0.60	0.64	6.90×10^{-7}
	Asian	rs2303790	Independent of index	16:57017292	p.Asp459Gly	G/A	HDL-C	0.407	0.025	7.53×10^{-62}	2.23	0.72	0.02	3.16×10^{-5}
PMFBP1	GLGC	rs34832584	Independent of index	16:72162966	p.Thr505Lys	T/G	TC	0.020	0.003	1.62×10^{-8}	15.93	0.01	2.57	0.50
	Asian	rs16973716	Explaining index	16:72156842	p.Lys768Asn	G/T	TC	0.042	0.008	1.75×10^{-7}	28.96	0.07	44.69	2.66×10^{-7}
LIPG	GLGC	rs77960347	Independent of index	18:47109955	p.Asn396Ser	G/A	HDL-C	0.259	0.012	1.62×10^{-98}	1.07	0.14	0.01	0.79
	Asian	rs2000813	Independent of index	18:47093864	p.Thr111Ile	T/C	HDL-C	0.043	0.007	1.04×10^{-9}	31.06	0.08	28.61	1.76×10^{-41}
LDLR	GLGC	rs139043155	Independent of index	19:11217344	p.Asp225Glu	A/T	LDL-C	1.644	0.214	1.53×10^{-14}	0.004	0.02	0.00	NA
	Asian	rs200990725	Protein-altering is top	19:11217315	p.Arg257Trp	T/C	LDL-C	0.882	0.109	6.35×10^{-16}	0.094	0.15	0.001	1.96×10^{-4}
PPARA	GLGC	rs1042311	Protein-altering is top	22:46627780	p.Ala268Val	T/C	TC	0.123	0.018	7.40×10^{-12}	0.50	0.01	0.01	0.23
	Asian	rs1800234	Protein-altering is top	22:46615880	p.Val227Ala	C/T	TG	–0.094	0.018	3.17×10^{-7}	4.21	0.07	0.15	0.12

AAF, alternative allele frequency. Positions are reported in human genome build hg19. Alleles are listed as alternative/reference alleles on the forward strand of the reference genome. [a]Protein-altering is top: protein-altering variants are the most significant variants in the known loci. Explaining index: conditional on the coding variants; adjusted *P* for index variants > 0.01. Independent of index: conditional on the index variants; adjusted *P* for coding variants with exome-wide significance.

Table 4 Loci for which East Asian and GLGC samples identified the same putatively functional protein-altering variant

Gene	rsID	Position	Variant	Alleles	Trait	Study	β	S.e.m.	*P*	AAF	Variance explained (%)	Note[a]
GCKR	rs1260326	2:27730940	p.Leu446Pro	C/T	TG	GLGC	−0.121	0.003	0.00	0.628	0.64	Protein-altering is top
					TG	Asian	−0.114	0.007	1.26×10^{-62}	0.496	0.64	Protein-altering is top
MLXIPL	rs35332062	7:73012042	p.Ala358Val	A/G	TG	GLGC	−0.124	0.004	5.22×10^{-205}	0.117	0.30	Protein-altering is top
					TG	Asian	−0.109	0.011	2.03×10^{-23}	0.109	0.23	Explaining index
LPL	rs328	8:19819724	p.Ser474*	G/C	TG	GLGC	−0.184	0.004	0.00	0.098	0.58	Explaining index
					TG	Asian	−0.169	0.012	1.93×10^{-45}	0.095	0.46	Explaining index
GPAM	rs2792751	10:113940329	p.Ile43Val	C/T	TC	GLGC	−0.028	0.003	7.14×10^{-22}	0.728	0.03	Explaining index
					TC	Asian	−0.043	0.007	5.67×10^{-9}	0.706	0.07	Protein-altering is top
HNF1A	rs1169288	12:121416650	p.Ile27Leu	C/A	TC	GLGC	0.037	0.003	9.99×10^{-40}	0.333	0.06	Protein-altering is top
					TC	Asian	0.038	0.007	4.86×10^{-8}	0.404	0.07	Protein-altering is top
TM6SF2	rs58542926	19:19379549	p.Glu167Lys	T/C	TC	GLGC	−0.129	0.005	7.03×10^{-155}	0.074	0.22	Protein-altering is top
					TC	Asian	−0.066	0.013	4.25×10^{-7}	0.070	0.06	Protein-altering is top
APOE	rs7412	19:45412079	p.Arg176Cys	T/C	LDL-C	GLGC	−0.539	0.006	0.00	0.075	3.80	Independent of index
					LDL-C	Asian	−0.472	0.016	4.87×10^{-197}	0.088	3.49	Protein-altering is top

AAF, alternative allele frequency. Positions are reported in human genome build hg19. Alleles are listed as alternative/reference alleles on the forward strand of the reference genome. [a]Protein-altering is top: protein-altering variants are the most significant variants in the known loci. Explaining index: conditional on the coding variants; adjusted *P* for index variants > 0.01. Independent of index: conditional on the index variants; adjusted *P* for coding variants with exome-wide significance.

than in the discovery study and reached genome-wide significance in the joint analysis (*P* values ranged from 3.0×10^{-8} to 7.6×10^{-15}) (**Supplementary Table 10**).

Coding variants point to the same genes across ancestries

We further evaluated whether the variants identified in East Asian samples were also defined as putative functional variants in GLGC samples (**Supplementary Table 11**). We found that East Asian and GLGC samples both pointed to the same nine functional genes, but that different associated variants were present in each ancestry (**Table 3**). The eight coding variants (MAF, 0.004–15.9%) at *PCSK9*, *CD36*, *ABCA1*, *CETP*, *PMFBP1*, *LIPG*, *LDLR*, and *PPARA* identified by GLGC showed lower MAFs (0–2.57%) in the East Asian samples and thus achieved no or only suggestive significance (*CETP*). Conversely, the coding variants at *PCSK9*, *APOB*, *CD36*, *CETP*, *LDLR*, and *PPARA* identified in East Asian samples (MAF, 0.094–12.45%) also had lower MAFs in GLGC samples (0.001–0.20%). In addition, the same putatively functional coding variants and genes at seven loci (*GCKR*, *MLXIPL*, *LPL*, *GPAM*, *HNF1A*, *TM6SF2*, and *APOE*) were identified in both East Asian and GLGC samples, with similar common MAFs (**Table 4**).

East Asian–specific association signals

We next attempted to identify variants that were associated with lipids in East Asian samples only. Among the known lipid loci, we identified 363 independent variants by sequential conditional analysis in GLGC exome-wide association studies (**Supplementary Table 11**). After conditioning on the independent variants in the corresponding loci, we identified 14 independent coding variant associations at 11 loci in East Asian samples with conditional *P* values < 4.5×10^{-7} (**Table 5**, **Figs. 1** and **3**). All 14 East Asian–specific variants were included in the list of putative functional variants that we identified. Eight of these loci (*EVI5*, *APOB*, *HMGCR*, *CD36*, *APOA1*, *CETP*, *LDLR*, and *PPARA*) harbored at least one low-frequency or rare independent coding variant (MAF, 4.21–0.03%). All of these variants either were monomorphic or had a frequency that was at least one order of magnitude lower in Europeans and thus showed only suggestive significance in ~300,000 GLGC individuals.

DISCUSSION

This study represents a large discovery effort to identify coding variation that influences lipid levels in the East Asian population, and it enabled us to systemically evaluate protein-altering variants that identify candidate functional genes. Meta-analyses of East Asian and multi-ancestry samples by exome-chip genotyping array identified 12 novel loci, 5 of which harbored nonsynonymous variants. In the 38 known loci that were replicated, we identified 31 protein-altering variants pointing to 25 functional lipid genes. Moreover, significant association with protein-altering variants identified the same 16 putative functional genes in European and East Asian samples, and 9 of those genes were identified by independent protein-altering variants in the two ancestries.

Among the novel genetic loci identified, several have been implicated in cardiovascular and metabolic phenotypes, which may provide mechanistic insight into the regulation of lipid levels and potential targets for treatment. The significant novel variant associated with both lipids and CAD is located in an intron of *MCU*. *MCU* encodes a mitochondrial inner membrane calcium uniporter that mediates calcium uptake into mitochondria. Mitochondrial calcium has an important role in the regulation of metabolism in the heart[21]. *CD163* encodes a macrophage-specific receptor involved in the clearance and endocytosis

Table 5 East Asian–specific variants associated with blood lipids (conditional $P < 4.5 \times 10^{-7}$)

					East Asian					GLGC		
Gene	Position	rsID	Alleles	Variant	Trait	AAF (%)	β	S.e.m.	P	P_{adj}	AAF (%)	β	S.e.m.	P
EVI5	1:93159927	rs117711462	A/G	p.Arg354Cys	TC	0.69	0.212	0.040	1.41×10^{-7}	2.15×10^{-7}	0.03	0.097	0.080	0.245
APOB	2:21228437	rs376825639	G/A	p.Ile3768Thr	TC	0.15	−0.659	0.097	8.44×10^{-12}	9.96×10^{-12}				
					LDL-C	0.15	−0.579	0.098	3.35×10^{-9}	4.44×10^{-9}				
	2:21252807	–	T/C	p.Cys478Tyr	TC	0.09	−0.876	0.138	2.08×10^{-10}	1.65×10^{-10}				
					LDL-C	0.09	−0.772	0.141	4.19×10^{-8}	3.22×10^{-8}				
	2:21252534	rs13306194	A/G	p.Arg532Trp	TC	12.39	−0.114	0.010	1.45×10^{-29}	2.01×10^{-17}	0.19	−0.084	0.031	6.74×10^{-3}
					LDL-C	12.45	−0.098	0.010	9.53×10^{-22}	2.08×10^{-13}	0.20	−0.085	0.032	8.13×10^{-3}
					TG	12.43	−0.073	0.010	1.38×10^{-12}	4.96×10^{-15}	0.19	−0.133	0.032	2.96×10^{-5}
HMGCR	5:74646765	rs191835914	C/A	p.Tyr311Ser	LDL-C	1.73	−0.190	0.026	2.20×10^{-13}	2.68×10^{-9}	0.04	−0.117	0.067	0.079
CD36	7:80302116	rs148910227	T/C	p.Arg386Trp	HDL-C	0.31	0.342	0.058	3.17×10^{-9}	3.60×10^{-9}	0.02	0.215	0.084	0.010
APOA1	11:116707736	rs12718465	T/C	p.Ala61Thr	HDL-C	3.27	−0.116	0.058	5.50×10^{-10}	1.41×10^{-7}	0.02	0.075	0.099	0.449
ACACB	12:109696838	rs2075260	A/G	p.Val21411Ile	TG	74.34	0.043	0.008	3.95×10^{-8}	7.64×10^{-8}	80.23	0.011	0.003	5.32×10^{-4}
ALDH2	12:112241766	rs671	A/G	p.Glu457Lys	HDL-C	20.43	−0.048	0.008	1.16×10^{-8}	1.85×10^{-8}	0.08	−0.005	0.052	0.928
CETP	16:56997025	rs201790757	G/T	p.Tyr74*	HDL-C	0.03	1.117	0.182	8.97×10^{-10}	4.33×10^{-10}	0.001	0.719	0.352	0.041
	16:57017292	rs2303790	G/A	p.Asp459Gly	HDL-C	2.23	0.407	0.025	7.53×10^{-62}	1.89×10^{-31}	0.02	0.384	0.092	3.16×10^{-5}
PKD1L3	16:71967927	rs17358402	T/C	p.Arg1572His	LDL-C	5.40	0.085	0.015	2.11×10^{-8}	1.86×10^{-9}	24.44	−0.013	0.003	8.47×10^{-5}
					TC	5.41	0.088	0.015	1.96×10^{-9}	1.40×10^{-10}	24.44	−0.009	0.003	3.72×10^{-3}
LDLR	19:11217315	rs200990725	T/C	p.Arg257Trp	TC	0.09	0.677	0.109	5.57×10^{-10}	5.00×10^{-9}	0.001	1.897	0.502	1.57×10^{-4}
					LDL-C	0.09	0.882	0.109	6.35×10^{-16}	6.15×10^{-15}	0.001	1.869	0.502	1.96×10^{-4}
PPARA	22:46615880	rs1800234	C/T	p.Val227Ala	TG	4.21	−0.094	0.018	3.17×10^{-7}	3.36×10^{-7}	0.15	−0.058	0.037	0.118

AAF, alternative allele frequency. Positions are reported in human genome build hg19. Alleles are listed as alternative/reference alleles on the forward strand of the reference genome. P_{adj}, conditioning on the independent variants in the corresponding loci identified by GLGC exome-wide association studies (**Supplementary Table 11**).

of hemoglobin–haptoglobin complexes by macrophages. Soluble CD163 was recently proposed as a biomarker of the well-known variables of metabolic syndrome, including HDL-C[22]. *ACVR1C* encodes activin-receptor-like kinase 7 (ALK7), one of the type I transforming growth factor–β receptors. ALK7 has recently been shown to have an important role in the maintenance of metabolic homeostasis[23]. ALK7 is expressed at high levels in human adipose tissue and is correlated with body fat and lipids. ALK7 dysfunction may cause increased lipolysis in adipocytes and leads to decreased fat accumulation. *MGAT1* encodes mannosyl (α-1,3)-glycoprotein β-1,2-*N*-acetylglucosaminyltransferase, which is involved in the synthesis of protein-bound and lipid-bound oligosaccharides. It has been found that a variant of *MGAT1* is associated with body weight and obesity[24]. We note that *CD163* and *PDGFC* were associated with lipid levels in an East Asian lipids GWAS meta-analysis published after our manuscript was submitted[25]. To further clarify the possible transcriptional mechanisms underlying the identified loci in association with lipids, we investigated the relationships of the novel variants and proxies with expression quantitative trait loci (eQTLs) by using the Genotype-Tissue Expression (GTEx) eQTL browser. We found significant *cis*-eQTL effects in human tissues at five loci at $P < 4.5 \times 10^{-7}$ (**Supplementary Table 12**). We further predicted putatively regulatory variants in seven novel non-coding regions in 81 cell lines on the basis of deltaSVM scores[26], and found that the variants in *PDGFC*, *C6orf183*, and *MCU* had high regulatory potential with extreme deltaSVM scores greater than 10 in absolute value (**Supplementary Fig. 4**).

Our data allow a more comprehensive understanding of the genetic architecture of lipid susceptibility by revealing novel lipid genes and identifying allelic heterogeneity across populations of different ancestries. We detected multiple independent association signals or new lead variants in known lipid-associated loci that frequently showed no or moderate LD with the corresponding GWAS index variants in European populations. Specifically, we identified 14 East Asian–specific variants that could not be explained by all the independent variants in the corresponding loci identified in GLGC samples. Our study demonstrated the benefits of distinct LD patterns between ancestry groups for the investigation of validated loci. We also found substantial inter-ancestry differences in the identification of rare coding variants across populations, which may have been subjected to natural selection during human evolution or genetic drift. All the low-frequency or rare functional coding variants identified in East Asians (MAF, 0.03–4.21%) appeared to be population specific, and were monomorphic or not present in European individuals who were part of the 1000 Genomes Project; this allelic heterogeneity across populations of different ancestry has been reported in part[6,11]. However, we observed that these rare variants were not monomorphic in more than 300,000 GLGC individuals, but had 15-fold to 160-fold lower frequencies (MAF, 0.001–0.15%) in Europeans than in East Asians (**Supplementary Table 13** and **Supplementary Fig. 5**), with little power to indicate association in Europeans. Similarly, the low-frequency and rare coding variants identified in GLGC samples were extremely rare or monomorphic in East Asian samples (**Supplementary Fig. 6** and **Supplementary Table 11**). Overall, our findings demonstrate that rare and low-frequency coding variants are more likely to be population specific, which underscores the value of discovering ancestry-specific rare variants in diverse populations, particularly for low-frequency variations.

As most GWAS index variants are located in noncoding regions, the identification of associated protein-coding variants may allow scientists to prioritize functional genes and variation. Among the 38 known loci that reached chip-wide significance in our data, coding variants at 16 loci (42.1%) were found to completely account for the original association signal. At an additional nine loci, an independent protein-altering variant indicated a likely functional gene. The coding variants were more likely to have consistent effect sizes across ethnic groups than noncoding variants were. For the GWAS index variants that could not be replicated in East Asian samples, the effect sizes were poorly correlated with those observed in Europeans. In contrast, the effect sizes of the putatively functional coding variants in the same loci were strongly related across ethnic groups (**Supplementary Fig. 7**). Trans-ancestry comparisons provided additional credible evidence to support the identification of the same 16 genes as putative functional genes. The functional genes pointed to by coding variants were either well-known genes or genes with previously unknown roles in lipid metabolism (such as *GPAM* and *PMFBP1*), which may be good candidates for functional assessment. More important, we found that the effects of these putative functional coding variants on levels of LDL-C, TG, and TC were highly correlated with the effect on CAD, but the effects on HDL-C levels were not correlated with CAD. Our findings are in agreement with recent genetic studies showing that both LDL-C and TG levels, but not HDL-C levels, are causally related to CAD risk[27–30].

This large-scale exome-wide association study allowed us to detect a greater number of low-frequency and rare variants than previously identified, 30% of which were not polymorphic in an earlier exome-wide study involving 12,685 Chinese individuals[11]. Nonetheless, the exome array offered moderate coverage for rare variants observed in ExAC East Asian samples. Power calculations indicated that the available sample size provided 80% power to detect variants with an effect size of 0.27 s.d. and MAFs as low as 0.5% at $P < 4.5 \times 10^{-7}$. However, we had considerably less power to evaluate extremely rare variants (MAF < 0.1%). Studies with larger sample sizes and of sequenced samples are therefore needed to fully investigate associations of rare variants with lipid levels.

In conclusion, we identified 12 new loci associated with lipid levels. We also identified coding variants that highlight 25 likely functional genes at previously known loci, including several with previously undiscovered roles in lipids. We also found an abundance of population-specific coding variant associations that underlie lipid traits, highlighting the importance of including individuals of diverse ancestral backgrounds. At the same time, our data demonstrate that the integration of genomic data across diverse ancestral groups may enable researchers to identify functional variants and genes for further functional study.

URLs. Genotype-Tissue Expression (GTEx) Portal, http://www.gtexportal.org/home; Genezoom, http://genome.sph.umich.edu/wiki/Genezoom; ExAC, http://exac.broadinstitute.org; RareMETALS, http://genome.sph.umich.edu/wiki/RareMETALS; RVTESTS, http://genome.sph.umich.edu/wiki/RvTests; RAREMETALWORKER, http://genome.sph.umich.edu/wiki/RAREMETALWORKER.

METHODS

Methods, including statements of data availability and any associated accession codes and 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 thank all the participants of this study for their contributions. X. Lu is supported by the CAMS Innovation Fund for Medical Sciences (grants 2016-I2M-1-009, 2017-I2M-1-004, and 2016-I2M-1-011) and the National Science Foundation of China (grants 81422043, 91439202, 81370002, 81773537, and 81230069). C.J.W. is supported by the National Institutes of Health (grant HL135824). S.K. and C.J.W. are supported by the National Institutes of Health (grant HL127564). P.C.S. was supported by the Hong Kong Research Grants Council (grants TRS T12/705/11 and GRF 17128515). G.M.P. is supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (award K01HL125751). We thank P. Marshall for professional editing. Additional acknowledgments of funding sources for the primary studies are provided in **Supplementary Note 1**.

AUTHOR CONTRIBUTIONS

X. Lu, C.J.W., G.M.P., D.J.L., D.G., and K.L.M. drafted the manuscript. C.J.W., D.G., X. Lu, P.C.S., S.K., K.L.M., and Y.E.C. coordinated the project. X. Lu, D.J.L., G.M.P., and H.Z. served as the central meta-analysis group. X. Lu and J.B.N. carried out eQTL analysis. X. Lu and W. Zhou carried out DeltaSVM analysis. X. Lu, G.M.P., D.J.L., Y. Wu, H.Z., J. Li, C.S.T., R.D., J. Long, X.G., C.N.S., Y.C., Y. Wang, C.Y.Y.C, Q.F., J.S., X.Y., W. Zhao, M.H., and J.B.N. carried out cohort data analysis. W. Zhou, H.L., C.C.K., J. Liu, L.W., F.W., J.S., and W.H. carried out cohort genotyping. H.L., M.X., X. Liu, Y.Z., L.S., Y.G., Y. Hu, K.Y., J.H., Q.C., S.C., A.B.F., L.S.A., P.G.-L., S.D., K.H., and L.G.F. carried out cohort phenotyping. X. Lu, W.H.-H.S., S.S.C., A.B.F., L.S.A., P.G.-L., S.D., R.V., Y.-D.I.C., X.-O.S., K.S.L.L., T.Y.W., S.K.G., Z.M., K.H., L.G.F., H.T., Y. Huo, C.Y.C., Y.E.C., W. Zheng, E.S.T., W.G., X. Lin, W.H., G.A., S.K., K.L.M., T.W., P.C.S., D.G., and C.J.W. were the principal investigators for the cohort.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
2. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
3. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
4. Asselbergs, F.W. *et al.* Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am. J. Hum. Genet.* **91**, 823–838 (2012).
5. Kim, Y.J. *et al.* Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat. Genet.* **43**, 990–995 (2011).
6. Wu, Y. *et al.* Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet.* **9**, e1003379 (2013).
7. Lambert, N.G. *et al.* Risk factors and biomarkers of age-related macular degeneration. *Prog. Retin. Eye Res.* **54**, 64–102 (2016).
8. Herder, C., Kowall, B., Tabak, A.G. & Rathmann, W. The potential of novel biomarkers to improve risk prediction of type 2 diabetes. *Diabetologia* **57**, 16–29 (2014).
9. Wierzbicki, A.S. & Oben, J. Nonalcoholic fatty liver disease and lipids. *Curr. Opin. Lipidol.* **23**, 345–352 (2012).
10. Kathiresan, S. A *PCSK9* missense variant associated with a reduced risk of early-onset myocardial infarction. *N. Engl. J. Med.* **358**, 2299–2300 (2008).
11. Tang, C.S. *et al.* Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese. *Nat. Commun.* **6**, 10206 (2015).
12. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014).
13. Lange, L.A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94**, 233–245 (2014).
14. Liu, D.J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* http://dx.doi.org/10.1038/ng.3977 (2017).
15. 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).
16. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).

17. Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet.* **7**, e1002198 (2011).
18. Nikpay, M. *et al.* A comprehensive 1,000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
19. Krokstad, S. *et al.* Cohort profile: the HUNT Study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
20. Lu, X. *et al.* Genetic susceptibility to lipid levels and lipid change over time and risk of incident hyperlipidemia in Chinese populations. *Circ. Cardiovasc. Genet.* **9**, 37–44 (2016).
21. Williams, G.S., Boyman, L. & Lederer, W.J. Mitochondrial calcium and the regulation of metabolism in the heart. *J. Mol. Cell. Cardiol.* **78**, 35–45 (2015).
22. Parkner, T. *et al.* Soluble CD163: a biomarker linking macrophages and insulin resistance. *Diabetologia* **55**, 1856–1862 (2012).
23. Carlsson, L.M. *et al.* ALK7 expression is specific for adipose tissue, reduced in obesity and correlates to factors implicated in metabolic disease. *Biochem. Biophys. Res. Commun.* **382**, 309–314 (2009).
24. Johansson, A. *et al.* Linkage and genome-wide association analysis of obesity-related phenotypes: association of weight with the *MGAT1* gene. *Obesity (Silver Spring)* **18**, 803–808 (2010).
25. Spracklen, C.N. *et al.* Association analyses of East Asian individuals and transancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Hum. Mol. Genet.* **26**, 1770–1784 (2017).
26. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
27. Do, R. *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45**, 1345–1352 (2013).
28. Rosenson, R.S., Davidson, M.H., Hirsh, B.J., Kathiresan, S. & Gaudet, D. Genetics and causality of triglyceride-rich lipoproteins in atherosclerotic cardiovascular disease. *J. Am. Coll. Cardiol.* **64**, 2525–2540 (2014).
29. Helgadottir, A. *et al.* Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat. Genet.* **48**, 634–639 (2016).
30. Voight, B.F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572–580 (2012).

[1]Department of Epidemiology, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [2]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA. [3]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA. [4]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [5]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA. [6]Department of Public Health Sciences, Institute of Personalized Medicine, Penn State University, University Park, Pennsylvania, USA. [7]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [8]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. [9]MOE Key Lab of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science & Technology, Wuhan, Hubei, China. [10]Department of Surgery, Li KaShing Faculty of Medicine, The University of Hong Kong, Hong Kong, China; Dr. Li Dak-Sum Research Centre, The University of Hong Kong–Karolinska Institutet Collaboration in Regenerative Medicine, Hong Kong, China. [11]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore. [12]Key Laboratory of Nutrition and Metabolism, Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and University of the Chinese Academy of Sciences, Shanghai, China. [13]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [14]Institute for Translational Genomics and Population Sciences, LABioMed at Harbor–UCLA Medical Center, Los Angeles, California, USA. [15]Department of Cardiology, Institute of Vascular Medicine, Peking University Third Hospital, Key Laboratory of Molecular Cardiovascular Sciences, Ministry of Education, Beijing, China. [16]Center for Genomic and Personalized Medicine, Medical Scientific Research Center and Department of Occupational Health and Environmental Health, School of Public Health, Guangxi Medical University, Nanning, Guangxi, China. [17]Department of Cardiology, Peking University First Hospital, Beijing, China. [18]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore, Singapore. [19]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [20]Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China. [21]Department of Medicine, the University of Hong Kong, Hong Kong, China. [22]Community Health Center, The 3rd Affiliated Hospital of Shenzhen University, Shenzhen, China. [23]Duke–National University of Singapore Graduate Medical School, Singapore, Singapore. [24]Department of Genetics, Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, Shanghai, China. [25]Division of Endocrine and Metabolism, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan. [26]Department of Psychiatry, the University of Hong Kong, Hong Kong, China. [27]Centre for Genomic Sciences, Li KaShing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [28]State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong, China. [29]USC–Office of Population Studies Foundation, University of San Carlos, Cebu City, Philippines. [30]Department of Anthropology, Sociology, and History, University of San Carlos, Cebu City, Philippines. [31]Department of Nutrition, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, USA. [32]Carolina Population Center, University of North Carolina, Chapel Hill, North Carolina, USA. [33]USC Eye Institute, Department of Ophthalmology, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA. [34]Research Centre of Heart, Brain, Hormone and Healthy Aging, Li KaShing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [35]State Key Laboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong, China. [36]Saw Swee Hock School of Public Health, National University Health System, National University of Singapore, Singapore, Singapore. [37]Department of Ophthalmology, National University of Singapore, Singapore, Singapore. [38]HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [39]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway. [40]Department of Medicine, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway. [41]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, USA. [42]Hong Kong–Guangdong Joint Laboratory on Stem Cell and Regenerative Medicine, the University of Hong Kong, Hong Kong, China. [43]Ophthalmology and Visual Sciences Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. [44]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore, Singapore. [45]A full list of members and affiliations appears in **Supplementary Note 1**. [46]Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA. [47]These authors jointly supervised this work. Correspondence should be addressed to C.W. (cristen@umich.edu), D.G. (gudongfeng@vip.sina.com) or P.C.S. (pcsham@hku.hk).

ONLINE METHODS

Study cohorts. Twenty-three studies, including both population-based studies and case-control studies of CAD and type 2 diabetes, were genotyped with the Illumina HumanExome array, resulting in a total of 47,532 participants, all of whom were of East Asian ancestry (**Supplementary Table 14**). All participants provided written informed consent, and ethics approval for data generation and analyses was individually obtained for each contributing study. The relevant human genetic data were also approved by the Ministry of Science and Technology of China. For the GLGC exome study, 95 studies contributed association results for exome chip genotypes and plasma lipid levels (**Supplementary Note 1** and **Supplementary Table 15**).

Phenotypes. For most East Asian subjects (86%), TC, HDL-C, and TGs were measured at >8 h of fasting. LDL-C levels were directly measured in 18 studies (88% of total study individuals) and were estimated via the Friedewald formula in the remaining studies, with missing values assigned to individuals with >400 mg/dl TGs. We adjusted the TC values for individuals on lipid-lowering medication by replacing their TC values by TC/0.8 with lipid medication status available. If measured LDL-C was available in a study, the treated LDL-C value was divided by 0.7. No adjustment for individuals using medication was made for HDL-C or TG.

Exome array genotyping and quality control. All study participants were genotyped on the HumanExome Bead-Chip (Illumina), and most samples (83%) also included the custom Asian Vanderbilt content. This custom content was added to the standard Illumina HumanExome Bead-Chip to improve the coverage of low-frequency variants in Asian populations. The variants were selected from 1,077 (581 female Chinese subjects and 496 male Singapore Chinese) whole-exome-sequenced East Asian samples generously provided by W. Zheng and J. Liu[31]. Approximately 29,000 additional common variants were added to the array, including previously identified GWAS variants selected from the GWAS catalog. Genotype calling was done with GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina) in combination with zCall version 2.2 (ref. 32). Within each study, individuals with low genotype completion rates, individuals expressing gender mismatches or a high level of heterozygosity, related individuals, and PCA outliers were excluded from further analysis (**Supplementary Table 16**). In addition, variants that did not meet the 95% or 98% genotyping threshold or that showed deviation from Hardy–Weinberg equilibrium were removed.

Statistical analyses. For each cohort, HDL-C, LDL-C, TG, and TC measurements were transformed via the inverse normal distribution after adjustment of each trait for age, age squared, and study-specific covariates, including principal components to account for population structure. In studies on diabetes or cardiovascular disease status, cases and controls were analyzed separately.

We performed both single-variant and gene-level association tests. Single-variant analyses in each cohort were carried out with either RAREMETALWORKER or RVTESTS[33], both of which generate single-variant score statistics and their covariance matrix between single-marker statistics. The test statistics, as visualized in a quantile–quantile plot, appeared well-calibrated (**Supplementary Fig. 8**). Gene-based tests were restricted to variants that were predicted to alter the coding sequence of the gene product (defined as missense, stop-gain, stop-loss, or splice-site variants) to enhance the likelihood of identifying causal variants and to reduce the multiple-testing burden. For each trait, we ran four gene-based tests: variable-threshold burden tests with MAF cutoffs of <5% and <1%, and sequence kernel association tests with MAF cutoffs of <5% and <1%. Next, we carried out meta-analyses of single-variant and gene-level association tests with RAREMETALS[34] for HDL-C, LDL-C, TC, and TG. For single variants, we applied a significance threshold of $P < 4.5 \times 10^{-7}$, corresponding to a Bonferroni correction for 110,986 polymorphic variants that had at least 20 minor alleles. For gene-level tests, we used a significance threshold of $P < 2.8 \times 10^{-6}$, corresponding to a Bonferroni correction for 17,614 gene-level tests.

To identify putative functional coding variants that could account for the effects at known lipid loci, we performed reciprocal conditional analyses to control for the effects of known lipid GWAS or coding variants. Loci where the initial lead variant had conditional $P > 0.01$ were considered to be explained by the variants used in the conditional analyses. To dissect East Asian–specific association signals in the reported loci, we also carried out conditional association analysis for variants within 1 Mb of each locus using covariance matrices between single-variant association statistics. Details of the methods can be found in ref. 33. To evaluate two or more independent association signals, we carried out sequential conditional association analyses using the lead variant at each locus as a covariate until results after conditional analysis were no longer significant ($P > 4.5 \times 10^{-7}$). We estimated the LD metric r^2 by using the cohort-combined variants and LD matrices. LD for variants not included on the exome array was estimated from the 1000 Genomes Project cohort of East Asian individuals.

To further assess whether the identified functional coding variants also relate to CAD, we tested their associations with CAD in PUUMA-MI[11], HKU-TRS, HuCAD[35], and two GWAS samples[36] (the Beijing Atherosclerosis Study and the China Atherosclerosis Study (CAS)) involving 9,661 CAD cases and 18,558 controls. The effect estimates and s.e. were meta-analyzed with METAL via the fixed-effect inverse-variance method[37]. We also looked up the CAD association in the largest publicly available CAD GWAS analysis (CARDIoGRAMplusC4D), which consists of ~185,000 CAD cases and controls[18].

***In silico* replication samples.** The *in silico* replication study was conducted with data from additional independent individuals of European ancestry from the Nord-Trøndelag Health Study (HUNT)[19] and GLGC GWAS[2], and Chinese subjects from a Chinese lipids GWAS[20]. HUNT encompasses a population-based cohort of 62,168 individuals with genome-wide genotypes (Illumina Human CoreExome), imputation from the Haplotype Reference Consortium panel, and non-fasting lipid phenotypes. The Chinese lipids GWAS was a meta-analysis of over 13,000 Han Chinese who underwent standardized blood lipid measurements in four independent GWASs. These studies included CAS, the Beijing Atherosclerosis Study, the Genetic Epidemiology Network of Salt Sensitivity study[38], and CAS phase II.

Heritability and estimated proportion of variance explained. We estimated the proportion of variance explained by the set of independently associated variants. Joint effects of variants in a locus were approximated by $\vec{\beta}_{\text{Joint}} = \mathbf{V}_{\text{Meta}}^{-1}\vec{U}_{\text{Meta}}$, where $\vec{U}_{\text{Meta}}$ represents single-variant score statistics and $\mathbf{V}_{\text{Meta}}^{-1}$ is the covariance matrix between them. The covariance between single-variant genetic effects was approximated by the inverse of the variance–covariance matrix of score statistics, that is, $\mathbf{V}_{\text{Meta}}^{-1}$. The phenotype variance explained by independently associated variants in a locus was given by $\hat{\vec{\beta}}_{\text{Joint}}^{T}\,\text{cov}(G)\hat{\vec{\beta}}_{\text{Joint}}$.

Annotation. We used ANNOVAR (version 2012-05-25)[39] to annotate variants as missense, splice, stop-gain/loss, synonymous, or noncoding. Variant identifiers and chromosomal positions are listed with respect to the hg19 genome build.

DeltaSVM analysis. DeltaSVM uses a gapped *k*-mer support vector machine to estimate the effect of a variant in a cell-type-specific manner[26]. DeltaSVM can accurately predict variants associated with DNase I hypersensitivity. Precomputed weights were available from a total of 222 ENCODE samples—99 from the Duke University set, and 123 from the University of Washington set[40]. For the current study, genetic variants were scored for deltaSVM in 81 cell lines from four tissues (blood, blood vessel, heart, and liver). For each of the seven novel noncoding regions, all proxies ($r^2 > 0.8$) were identified on the basis of data from 1000 Genomes.

Data availability. Summary statistics are available for download from the University of Michigan Center for Statistical Genetics (http://csg.sph.umich.edu/abecasis/public/lipids2017EastAsian). Additional supporting data are provided in the supplementary material.

A **Life Sciences Reporting Summary** for this paper is available.

31. Zhang, Y. *et al.* Rare coding variants and breast cancer risk: evaluation of susceptibility loci identified in genome-wide association studies. *Cancer Epidemiol. Biomarkers Prev.* **23**, 622–628 (2014).
32. Goldstein, J.I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* **28**, 2543–2545 (2012).

doi:10.1038/ng.3978

33. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
34. Feng, S., Liu, D., Zhan, X., Wing, M.K. & Abecasis, G.R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828–2829 (2014).
35. Lu, X. *et al.* Coding-sequence variants are associated with blood lipid levels in 14,473 Chinese. *Hum. Mol. Genet.* **25**, 4107–4116 (2016).
36. Lu, X. *et al.* Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nat. Genet.* **44**, 890–894 (2012).
37. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
38. GenSalt Collaborative Research Group. GenSalt: rationale, design, methods and baseline characteristics of study participants. *J. Hum. Hypertens.* **21**, 639–646 (2007).
39. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
40. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

Corresponding Author: Cristen Willer
Date: 9/7/2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   We identified the largest sample size possible of East Asian individuals with exome chip genotyping (N=47,532).

2. Data exclusions

   Describe any data exclusions.

   At the meta-analysis level, no studies were excluded. At the primary study level, exclusions were made using standard quality control criteria: exclusion of samples were made based on sex mismatch, high missing data rate, etc. and exclusion of variants was made based on high missing data rate, Hardy-weinberg disequilibrium, etc.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   All variants that reached study-wide significance were replicated in three additional cohorts and the statistical evidence for all variants became stronger after including replication data (Supplementary Table 10).

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   All East Asian exome chip samples were analyzed together in the discovery phase. Replication was obtained after the discovery analysis was complete using other datasets (HUNT Europeans GWAS, GLGC European GWAS, Chinese GWAS) as listed in Supplementary Table 10.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was not relevant to this study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

Nature Genetics: doi:10.1038/ng.3978

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All software used is publicly available.
Genezoom, http://genome.sph.umich.edu/wiki/Genezoom
RareMETALS, http://genome.sph.umich.edu/wiki/RareMETALS
RVTESTS, http://genome.sph.umich.edu/wiki/RvTests
RAREMETALWORKER, http://genome.sph.umich.edu/wiki/RAREMETALWORKER

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Primary summary statistics were provided by individual study PIs. No companies were involved.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

Nature Genetics: doi:10.1038/ng.3978

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Summaries of research participants for each primary study are given in Supplementary Table 14. Since only summary statistics by variants were shared, and no individual-level genetic or phenotypic data were shared, this study is not considered to be Human Subjects research.

Nature Genetics: doi:10.1038/ng.3978