<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Tang, CS; Zhang, H; Cheung, YY; Xu, M; Ho, JCY; Zhou, W; Cherny, SS; Zhang, Y; Holmen, O; Au, KW; Yu, H; Xu, L; Jia, J; PORSCH, RM; Sun, L; Xu, W; Zheng, H; Wong, LY; Mu, Y; Dou, J; Fong, HY; Wang, S; Hong, X; Dong, L; Liao, Y; Wang, J; Lam, SML; Su, X; Yan, H; Yang, ML; Chen, J; Siu, DCW; Xie, G; Woo, YC; Wu, Y; Tan, KCB; Hveem, K; Cheung, BMY; Zöllner, S; Xu, A; Eugene Chen, Y; Jiang, CQ; Jiang, Y; Lam, TH; Ganesh, SK; Huo, Y; Sham, PC; Lam, KSL; Willer, CJ; Tse, HF; Gao, W</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Nature Communications, 2015, v. 6, p. 10206</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2015</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/226629">http://hdl.handle.net/10722/226629</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese


Blood lipids are important risk factors for coronary artery disease (CAD). Here we perform an exome-wide association study by genotyping 12,685 Chinese, using a custom Illumina HumanExome BeadChip, to identify additional loci influencing lipid levels. Single-variant association analysis on 65,671 single nucleotide polymorphisms reveals 19 loci associated with lipids at exome-wide significance ($P < 2.69 \times 10^{-7}$), including three Asian-specific coding variants in known genes (CETP p.Asp459Gly, PCSK9 p.Arg93Cys and LDLR p.Arg257Trp). Furthermore, missense variants at two novel loci—PNPLA3 p.Ile148Met and PKD1L3 p.Thr429Ser—also influence levels of triglycerides and low-density lipoprotein cholesterol, respectively. Another novel gene, TEAD2, is found to be associated with high-density lipoprotein cholesterol through gene-based association analysis. Most of these newly identified coding variants show suggestive association ($P < 0.05$) with CAD. These findings demonstrate that exome-wide genotyping on samples of non-European ancestry can identify additional population-specific possible causal variants, shedding light on novel lipid biology and CAD.
Dyslipidemia, often defined as elevated levels of total cholesterol (TC), triglyceride (TG) or low-density lipoprotein cholesterol (LDL-C) or decreased level of high-density lipoprotein cholesterol (HDL-C), is an important risk factor for coronary artery disease (CAD) which is a major cause of mortality and morbidity worldwide. The incidence of CAD and the prevalence of its major risk factors, such as dyslipidemia, vary widely according to geographical location and ethnicity. Such differences could be due to environmental factors, like smoking and diet, or underlying population genetic differences. Mapping of the genetic loci affecting blood lipid levels may shed light on disease pathogenesis and provide pharmaceutical targets for prevention and treatment of CAD.

Recently, a multi-ethnic meta-analysis of genome-wide association studies (GWAS) on > 188K individuals, the majority being of European ancestry, discovered a total of 157 loci associated with blood lipid levels that together accounted for about 20–30% of the total genetic contribution. Most of these variants are common, non-coding, and with modest impact on lipid levels, although rare coding variants with larger effects have been identified in a number of genes, including PCSK9 (refs 5,6), LDLR5,7, APOC3 (refs 8,9) and APOB8,6,10. By exploiting the differences in linkage disequilibrium (LD) of the non-European populations (2–9K), the study further narrowed down association signals at five loci, corroborating findings of fine-mapping analyses on Europeans with a much larger sample size.

The vast majority of discovery efforts have been made in samples of European ancestry4,6,11, which contain only a subset of human genetic variation. The degree of overlap between the variants and genes that contribute to lipid heritability in different populations is largely unexplored. During evolutionary history, genetic drift and evolutionary selection have led to different genetic architectures for populations in different geographic areas. For example, variants in APOC3 demonstrate higher frequencies in isolated populations with different ancestry, notably Greeks and Amish. Furthermore, loss-of-function mutations in PCSK9 are more prevalent in individuals with African ancestry. Inter-population differences in allele frequencies may therefore affect the detectability of risk variants. This highlights the importance of association analyses on non-European populations to identify novel loci influencing lipid levels.

To systematically evaluate the contribution of coding variants to lipid levels in an Asian population, we perform an exome-wide association analysis using a custom Asian Exomechip on 12,685 individuals selected from 2 Chinese cohorts, the University of Hong Kong Theme-based Research Scheme (HKU-TRS, N = 5,233) and Peking University Health Science Center and the University of Michigan Medical School study of Myocardial Infarction (PUUMA-MI, N = 7,452). First, we address whether known lipid variants identified in Europeans are associated with blood lipid levels in East Asian samples. Then we compare the effect size and power in detecting known variants between HKU-TRS and a Norwegian study of similar sample size (Nord-Trendelag Health Study (HUNT) study, N = 5,643). Next, we seek to identify coding variants in known loci that may point to the functional lipid genes and explore other protein-altering variants for novel lipid-associated genes in this Asian population. We discover a total of three Asian-specific associations involving rare nonsynonymous changes and identify three novel lipid-associated loci not previously reported. Our study demonstrates the importance of ancestry-specific association analysis to discover population-specific associations of rare variants and, most importantly, to provide new insights into novel biological pathways for lipid metabolism.

Results

Using an exome array augmented with custom markers, we successfully genotyped 145,276 polymorphic single nucleotide polymorphisms (SNPs) in 12,685 Chinese individuals (Supplementary Table 1). Single-variant and gene-based association analyses were carried out with respect to the four blood lipid levels (LDL-C, HDL-C, TC and TG). Single-variant score test statistics were presented for the following analyses, unless stated otherwise. Specifically, we assessed association of 65,671 variants that were polymorphic in both Chinese cohorts and had at least 20 minor alleles (MAC ≥ 20). Of these variants, 58% altered protein composition and 26% were Asian-specific variants with minor allele frequency (MAF) < 5% (Supplementary Table 2).

Power and effect size comparison with HUNT-MI. We initially examined the known European-identified GWAS variants4,11 in our combined data set (N = 12,685). Of the 157 index SNPs, 123 variants, including 4 with MAC ≥ 20, were successfully genotyped. We then tested 119 variants for association with their primary lipid traits. The analysis identified 8 variants reaching genome-wide significance (P ≤ 5 × 10^{-8}) and 52 reaching nominal significance (P < 0.05). The large proportions of associated variants were substantially more than expected under the null hypothesis of no association (binomial test P = 3.1 × 10^{-47} and P = 1.5 × 10^{-35}, respectively; Supplementary Table 3). Approximately 83% of the variants (N = 99) had direction of effect concordant to those reported (Supplementary Fig. 1).

To explore if differences in LD may affect power for detecting association of the known GWAS SNPs in Asian samples, we compared the number of significant variants discovered in a Norwegian cohort—HUNT-MI (N = 5,643)—against one of our Chinese cohorts—HKU-TRS (N = 5,233)—of similar sample size. Compared with HUNT-MI, fewer significant associations were detected in HKU-TRS (4 versus 7 for genome-wide significant variants and 32 versus 45 for variants with binomial test P < 0.05; Supplementary Table 4), suggesting a slight decrease in power to detect known European-identified variants in Asian samples. Such differences might be due to LD differences between the index SNP and the true causal variant. This hypothesis was supported by a weaker correlation with the reported effect sizes for Chinese at three of the four lipid traits (HKU-TRS: LDL-C r² = 0.49, HDL-C r² = 0.61, TC r² = 0.13 and TG r² = 0.62) than Norwegian (HUNT-MI: LDL-C r² = 0.46, HDL-C r² = 0.75, TC r² = 0.47 and TG r² = 0.84).

Refining association signals at known loci. Differences in LD patterns, on the one hand, may affect the detectability of variants tagging the casual SNP; on the other hand, it may help pinpoint the causal variant. We next investigated if variants other than the index SNPs could account for the association of the previously established loci, with the caveat that the majority of the variants we tested are coding. Including the eight loci with significant GWAS SNPs, a total of 16 known loci were associated with at least 1 lipid trait at exome-wide significance (P < 2.69 × 10^{-7}) (Supplementary Figs 2–6; Supplementary Tables 5–8). The association was evaluated after conditioning on nearby index SNP(s), and vice versa, revealing that only signals of four loci (ABCA1 for HDL-C, DOCK7-ANGPTL3, GCKR and MLXIPL for TG) could be completely explained by the GWAS index SNPs (Supplementary Data 1). This list could be further extended to include LIPC if variants reported to show independent associations were considered as well.
One of our non-index lead SNPs, rs13702, at the 3’ untranslated region of LPL, exhibited stronger association with HDL-C ($P = 1.3 \times 10^{-11}$) than the GWAS SNP (rs12678919, $r^2 = 0.47$, $P = 5.7 \times 10^{-15}$). Association of the GWAS SNP was largely attenuated (conditional $P = 4.0 \times 10^{-5}$) once accounted for the effect of rs13702, whereas rs13702 still showed near exome-wide significant association after conditional analysis (conditional $P = 7.2 \times 10^{-7}$). This suggested that the LPL index SNP might act as a proxy for rs13702. Refinement of association signals by population differences in LD could be further demonstrated by our lead SNPs at CELSR2-SORT1 for LDL-C (rs12740374, $P = 4.9 \times 10^{-18}$) and APOA5 for TG (rs651821, $P = 5.0 \times 10^{-10}$), where both SNPs were highlighted as the new lead SNPs in previous fine-mapping analyses involving African and East Asian populations, respectively. Likewise, our lead SNP encoding TM6SF2 p.Glu167Lys (rs5167, $P = 3.8 \times 10^{-12}$), was recently shown to be a causal variant altering TC and TG levels.

New lead SNPs at APOB and DOCK6. Besides refining known signals, some of the new lead SNPs appeared largely independent of the index SNPs ($r^2 < 0.03$), representing novel associations not previously reported (Table 1). At APOB, the new lead SNP corresponded to a novel missense variant (rs13306194, encoding p.Arg532Trp) associated with decreasing LDL-C ($\beta = -0.134$ s.d., $P = 1.2 \times 10^{-12}$) and TC levels ($\beta = -0.129$, $P = 6.0 \times 10^{-12}$), whereas the GWAS index SNP was only weakly associated with both lipid traits ($P > 4 \times 10^{-3}$). Interestingly, although APOB p.Arg532Trp is relatively common among Asians (MAF = 0.13), it has been found to be extremely rare in 1000 Genomes European (MAF = 0.003) and African samples (MAF = 0.004), as well as in HUNT-MI Norwegians (MAF = 7 × 10^{-3}). Similarly, a synonymous DOCK6 variant (rs737337, encoding p.Thr712Thr, MAF = 0.27), known to be associated with HDL-C, was associated with TC level ($P = 7.5 \times 10^{-10}$) in our study and displayed a large difference in allele frequency across populations (MAF = 0.07 in Europeans). Only marginal association has hitherto been reported with TC.

Complex association at APOE and APOA5. For two of the previously implicated loci, Apolipoprotein E (APOE) and APOA5, we observed not only one but multiple independent association signals. In some cases, a different lead SNP was observed for we observed not only one but multiple independent association signals. In some cases, a different lead SNP was observed for

Asian-specific association at CETP, LDLR and PCSK9. In addition, we also identified significant association of three Asian-specific missense variants defined as being polymorphic in Asian but monomorphic in other populations (Table 1). This included a low-frequency missense CETP SNP independently associated with HDL-C (rs2303790, p.Asp459Gly, MAF = 0.027; $\beta = 0.44$, $P = 3.2 \times 10^{-29}$). For LDL-C, associations with strong effect were detected for two probably damaging missense changes—a rare PCSK9 variant encoding p.Arg93Cys (rs151193009, MAF = 0.014; $P = 7.9 \times 10^{-22}$) and a LDLR mutation encoding p.Arg257Trp (rs200990725, MAF = 0.001; $P = 3.0 \times 10^{-8}$). Both variants appeared to be independent of previously reported association signals. The PCSK9 p.Arg464Leu missense variant (rs11591147) was near monomorphic in Chinese and the GWAS index SNP (rs2479409; $r^2 = 4.4 \times 10^{-3}$) was not associated with LDL-C in our study ($P = 0.72$). Similarly, the LDLR p.Arg257Trp was not linked ($r^2 = 1.7 \times 10^{-4}$) with the non-significant index SNP (rs6517220). The two single-variant associations identified in the current study may, by far, represent some of the strongest effect sizes ($\beta = -0.64$ s.d. for PCSK9 p.Arg93Cys and $\beta = 0.91$ for LDLR p.Arg257Trp) for any missense variant known to associate with plasma lipid levels. Interestingly, the association of PCSK9 p.Arg93Cys was more prominent in PUUMA-MI (MAF = 0.017; $\beta = -0.65$, $P = 3.2 \times 10^{-22}$) compared with HKU-TRS (MAF = 0.008; $\beta = -0.61$, $P = 3.7 \times 10^{-8}$). The large discrepancy was mainly a result of the intra-population difference, in which the risk allele frequency of Northern Chinese is at least twice the frequency of Southern Chinese. Such a phenomenon was also observed for other rare variants, which underscores the need for rare variant association analysis in a more fine-scale local population.

Novel loci identified at PKD1L3 for LDL-C and PNPLA3 for TG. Next, we sought to identify novel associations of plausibly functional variants. Along with the 16 known loci, 2 common nonsynonymous variants in genes/loci not previously implicated attained exome-wide significance (Table 1; Supplementary Fig. 2). Regional plots of associations for these two newly identified loci are illustrated in Fig. 1. Here PNPLA3 p.Ile148Met (rs738409), a missense variant known to associate with nonalcoholic fatty liver disease (NAFLD) (β = -0.072, $P = 4.4 \times 10^{-8}$). As NAFLD is a condition tightly linked to obesity and type 2 diabetes mellitus (T2DM), we looked for factors that might confound the observed association. Particularly, the HKU-TRS cohort was overrepresented with T2DM patients relative to the general population. We therefore stratified the sample by T2DM status and tested if this enrichment contributed to the TG-lowering effect. Although stronger association was observed in T2DM cases ($N = 3,366; \beta = -0.114, P = 8.4 \times 10^{-6}$) versus controls ($N = 1,775; \beta = -0.067, P = 0.032$), no significant interaction between the additive effect and T2DM was detected ($P = 0.27$). The effect size observed in the non-T2DM group was very similar to PUUMA-MI ($\beta = -0.058, P = 7.3 \times 10^{-4}$), where no

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms10206 | www.nature.com/naturecommunications
enrichment of T2DM cases was observed. Likewise, association of 
PNPLA3 remained significant (conditional \( P = 1.9 \times 10^{-5} \) in HKU-TRS) after adjusting for body mass index, demonstrating a 
novel association independent of potential confounding factors.

The second strongest novel association mapped to a LD block 
ensembleing the known locus of haptoglobin-related protein (HPR) at 16q22.2. Carriers of the PKD1L3 minor allele (rs7185272, 
encoding p.Thr429Ser) had significantly lower LDL-C (\( P = 5.4 \times 10^{-8} \)) and TC levels (\( P = 2.5 \times 10^{-7} \)). This coding variant is 
not in LD with the GWAS SNP at nearby HPR (rs2000999, \( r^2 = 0.027 \); \( P_{\text{LDL-C}} = 1.1 \times 10^{-3} \); \( P_{\text{TC}} = 1.9 \times 10^{-5} \)), implying an 
independent signal of association (Supplementary Data 1). Another probably damaging, missense variant, rs1559401 encoding 
PKD1L3 p.His571Gln, is indeed in complete LD with PKD1L3 
p.Thr429Ser and had indistinguishable effect sizes on both lipid 
traits (\( P_{\text{LDL-C}} = 5.9 \times 10^{-8} \); \( P_{\text{TC}} = 2.1 \times 10^{-7} \)). We further 
characterized the novel LDL-C lead SNP (rs7185272) by examining its association in other data sets. In silico look up in 
the largest publically available GWAS analyses of GLGC and 
CardioGRAM revealed equally strong associations with LDL-C 
(\( P = 2.8 \times 10^{-7} \)) and TC (\( P = 3.2 \times 10^{-10} \)) and slight though 
significant decrease in CAD risk (\( P = 0.058 \)).

Gene-based test identified TEAD2 as new HDL-associated gene. 
Compared with common or low frequency variants, the statistical 
power of single-variant test to detect association of individual rare 
variant is generally limited. To improve power for identifying 
novel lipid-associated genes, we proceeded to a gene-based 
association test using RAREMETAL, to assess the aggregate 
effect of rare variants across genes. Specifically, we restricted the 
analysis to nonsense variants and missense variants predicted to 
be damaging. Two MAF thresholds (<1% and <5%) were 
employed for the (1) unweighted combined multivariate 
collapsing burden test (CMB) and (2) sequence kernel 
association test (SKAT), whereas only MAF<5% threshold 
was considered for the (3) variable-threshold test (VT).

Three known genes—CEPT (\( P_{\text{HDL-C}} = 9.0 \times 10^{-5} \)), PCSK9 
(\( P_{\text{LDL-C}} = 1.4 \times 10^{-3} \); \( P_{\text{TC}} = 5.1 \times 10^{-27} \)) and APOE 
(\( P_{\text{LDL-C}} = 2.2 \times 10^{-3} \); \( P_{\text{TC}} = 2.0 \times 10^{-13} \))—exhibited strong
association in gene-level tests with HDL-C, LDL-C and/or TC (Table 2). Except APOE, the gene-based $P$ values exceeded the minimal $P$ value among all single-variant tests; however, the associations were largely attenuated ($P > 0.01$) while conditioning on the most significant missense variant, indicating that the signals were largely driven by a single coding SNP.

More importantly, via CMC burden test, we identified a significant gene-based association with HDL-C at a novel gene, TEAD2 ($\beta = 1.11, P = 1.9 \times 10^{-5}$). Though two rare, missense variants (rs142665148, encoding p.Asp12Asn, MAC = 1 and rs13306194, encoding p.Arg532Trp, MAC = 0.08%) predicted to be damaging were included in the TEAD2 gene set, the gene-based association was mainly driven by the near exome-wide significant TEAD2 p.Asp12Asn SNP ($\beta = 1.11, P = 3.8 \times 10^{-7}$). This missense variant was consistently found to increase HDL-C levels in both PUUMA-MI ($\beta = 1.16, P = 9.7 \times 10^{-5}$) and HKU-TRS ($\beta = 1.09, P = 1.2 \times 10^{-5}$). Another singleton missense SNP (TEAD2 p.Ala266Val) also showed a consistent, though non-significant, HDL-C-increasing effect ($\beta = 0.15, P = 0.25$).

Exome-wide significant variants were associated with CAD. We further explored the relationship between the lipid-associated loci and cardiovascular disease in the Chinese population. Specially, we tested whether the newly identified protein-altering variants showing independent associations with lipid traits also influenced CAD risk (Table 3). Among the seven variants tested, five SNPs showed a consistent direction of effect between lipid traits and CAD ($P < 0.05$). In particular, the (or near) Asian-specific variants associated with LDL-C were also strongly associated with CAD. The LDL-decreasing PCSK9 R93C (OR = 0.48, $P = 3.8 \times 10^{-7}$) and APOB R532W (OR = 0.85, $P = 2.9 \times 10^{-4}$) significantly protected against CAD, whereas the risk for CAD was 3.66 times higher for subjects carrying the LDL-increasing LDLR R257W allele ($P = 1.1 \times 10^{-4}$). Compared with the common variants in the same region, these rare variants conferred much larger protective or damaging effects to CAD. In fact, we observed a strong positive correlation between the strengths of effect on plasma lipid levels and CAD risk (Fig. 2), highlighting the clinical relevance of rare variants with large effect in personalized medicine.

**Discussion**

This study is the first Asian exome-wide association analysis on blood lipid levels. By genotyping 12,685 Chinese individuals using a custom exome array, we identified significant association at 19 loci. Three of these loci harbour lipid-associated missense variants, implicating three novel genes—PNPLA3, PKD1L3 and TEAD2—not previously described. For the other 16 known loci, we not only confirmed the association signals of known GWAS hits, but also identified novel independent associations and, above all, revealed 3 Asian-specific associations involving rare non-synonymous changes. This has important implications in study design, such that, by examining the less studied non-European populations, novel associations can be found even with a relatively small sample size. A major finding of our current study is the discovery of three novel lipid-associated loci. The probably damaging, missense variant encoding PNPLA3 p.Ile148Met, a TG lipase, was previously demonstrated to show strong association with NAFLD27–29. NAFLD refers to a wide spectrum of chronic liver disorders, ranging from hepatic steatosis to nonalcoholic steatohepatitis. It is characterized by increased hepatic fat content and hepatic TG level, and is tightly linked with signatures of cardiovascular diseases such as obesity, insulin resistance and T2DM30. The nonsynonymous change was demonstrated to reduce TG-rich very low density lipoprotein efflux, limit substrate-binding to the nearby catalytic residue and thereby impair hydrolyase activity31,32. While PNPLA3 has been...
LDL-cholesterol locus of haptoglobin-related protein (proliferator-activated receptor (PPAR) based burden test. It is involved in the pathway of peroxisome (TEF-4), was identified as a HDL-associated gene through gene-metabolism, according to the Reactome database. TEF-4 also was implicated as a putative sour taste receptor and PKD1L3 patients. Although these studies have shown suggestive lower TG levels in extremely obese individuals and NAFLD the minor allele G (encoding methionine) was associated with association was reported in population-based studies; however, the minor allele G (encoding methionine) was associated with lower TG levels in extremely obese individuals and NAFLD patients. Although these studies have shown suggestive association with lipid variation, our current study was first to demonstrate a genome-wide significant association with TG. Multiple lines of evidence from animal studies suggest that the catalytic inactivation of PNPLA3 can be exacerbated by overexpressing p.Ile148Met under dietary conditions, which expose the liver to high levels of insulin. Moreover, insulin increased PNPLA3 transcription through LXR-mediated activation of SREBP-1c, in line with the increased PNPLA3 mRNA levels observed in obese, insulin-resistant animals. These observations may provide an explanation for the slightly stronger effect of PNPLA3 p.Ile148Met with TG in the T2DM patients observed in HKU-TRS.

PKD1L3, encoding an ion channel of the polycystic kidney disease-like family, is located at ~100 kb away from a known LDL-cholesterol locus of haptoglobin-related protein (HPR). Although associated variants in both genes increase levels of LDL-C and TC, we have demonstrated in our study that their associations are highly independent. Besides the probably damaging TC lead SNP, its LD proxy (rs7185272), also the lead SNP for LDL-C, maps to a region with transcriptional factor-binding signals for FOXA1 and theрегулятору генома PCKS9 and TEAD2—showed strong evidence of association with blood lipid levels, alongside with three Asian-specific association signals in known loci (CETP, PSCK9 and LDLR).

Understanding the genetic architecture of how these deleterious rare and regulatory common variants together affect disease risk could provide more insights into the causal mechanisms, and also facilitate the translation of these genetic findings to drug discovery and personalized medicine.

**Methods**

**Subjects and phenotypes.** Our study comprised two main cohorts, HKU-TRS and PUUMA-MI, with a total of 12,685 study participants of Chinese ancestry genotyped and passed quality control. Study protocols were approved by the institutional review boards of all institutions involved in the study and written informed consent was obtained from all study participants. We genotyped 6,048 Southern Chinese subjects recruited from the Chinese CAD Cohort of the Queen Mary Hospital in Hong Kong, Hong Kong Cardiovascular Risk Factor Prevalence Study (CRIPS) and Hong Kong West Diabetes Registry (HKWDR). Detailed descriptions of the corresponding cohorts are provided in Supplementary Note 1.

Fasting venous blood samples were collected for DNA and laboratory-based biochemical analyses. General physical measurements and medical and drug histories of each subject were recorded. Plasma lipids were measured by standard enzymatic methods. LDL-C level was calculated using the Friedewald equation or by direct enzymatic colorimetric test if TG was > 4.5 mmol l⁻¹. Extreme lipid levels (> 4 s.d. from the mean) were set as missing to avoid false positive findings due to outliers. For the quantitative blood lipids analyses, we included only 5,233 subjects who were not taking any lipid-lowering drug or those with their pre-treatment lipid levels available. Clinical characteristics of subjects involved in the lipid association analysis are shown in Supplementary Table 9.

To evaluate if the lipid-associated loci also influence CAD, 2,372 CAD cases and 3,388 non-CAD controls were included in the data set (Supplementary Table 10). CAD cases were defined as having coronary artery revascularization interventions (including percutaneous coronary intervention and/or coronary artery bypass graft surgery); or those who had been diagnosed with MI; or those with stenosis of > 50% in one or more of the major epicardial coronary arteries by angiogram. Subjects who had no documented history or angiogram evidence of CAD were defined as non-CAD controls. Written informed consents were obtained from all subjects.
subjects, and the study protocol was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster. PUUMA-MI. PUUMA-MI is a large-scale project designed to study cardiovascular disease and related traits, including MI and plasma lipid levels, in China. Samples were collected by the Joint Institute of the PUUMA-MI. In total, 8,621 subjects were collected in 4 hospitals: 5,606 from Peking University First Hospital, 1,209 from Peking University Third Hospital, 774 from Beijing Shijinghao Hospital and 482 from Asia Heart Disease Hospital of Wuhan. Fasting plasma lipid levels were measured for Peking University First Hospital-based samples and Beijing Shijinghao cohort after overnight fasting. For samples collected from Peking University Third Hospital, TC and TG were measured by an enzymatic method while HDL-C and LDL-C were measured by a liquid selective detergent method. Details of the cohort description are provided in Supplementary Note 1 and Supplementary Table 11.

According to the 2004 American College of Cardiology/American Heart Association guideline, criteria for acute MI are detection of rise and/or fall of cardiac biomarkers (preferably troponin) with at least one value above the 99th percentile of the upper reference limit together with evidence of myocardial ischaemia with at least one of the following: symptoms of ischaemia; ECG changes indicative of new ischaemia (new ST-T changes or new left bundle branch block); development of pathological Q waves in the ECG; imaging evidence of new loss of viable myocardium or new regional wall motion abnormality. Any one of the following criteria meets the diagnosis for prior MI: development of new pathological Q waves with or without symptoms; imaging evidence of a region of loss of viable myocardium that is thinned and fails to contract, in the absence of a non-ischaemic cause. For other individuals, self-reported CAD status was collected as described by the Joint Institute of the PUUMA-MI. In total, 550 MI cases and 5,606 non-CAD controls from Peking University First Hospital, 876 MI cases and 333 non-CAD controls from Peking University Third Hospital, 256 MI cases and 226 non-CAD controls from Wuhan Asia Heart Hospital, and 774 non-CAD controls from Beijing Shijinghao Hospital.

HUNT-MI. A total of 5,643 Norwegians were selected from the second survey of the HUNT, including 2,349 medical-record confirmed MI cases and 2,317. Details of the Norwegian sample, genotyping and phenotypes have been described18.

Genotyping. All subjects were genotyped using the Asian Exomechip, a specially designed exome array with a custom content of 58,317 variants on top of the standard Illumina HumanExome BeadChip (Illumina, CA), which interrogated a total of 302,218 variants. Details of the Asian Exomechip design has been described in a number of studies, including the breast cancer exome-wide association analysis on Chinese17. In brief, the original design of the exome array includes 242,901 markers and 302,218 variants. Details of the Asian Exomechip design has been described in Supplementary Methods. Genotype calling was performed using GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina) independently for both GWAS follow-up or fine mapping studies. Additional information regarding the design is provided in Supplementary Methods. Genotype calling was performed using GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina) independently for both

Data quality controls. HKU-TR5. Manual inspection of genotype clusters was first carried out for >55,000 variants that either (1) showed evidence of bad genotype clustering in exome array genotyping of over 9,000 subjects by collaborators12,16, or (2) had missing rates >0.8 or (3) had high missingness (>1%). A total of 4,350 markers were removed due to poor genotype clustering. Individual-level QC was carried out with regard to duplication, gender mismatch, possible sample contamination and biological relatedness. We further employed a variant-level QC that removed SNPs with >2% missingness or violated Hardy–Weinberg equilibrium (P < 10−4) or SNPs originally designed with the purpose of quality control. Details regarding the quality controls are given in the Supplementary Methods. After all quality control measures, 5,233 samples and 286,795 SNPs, of which 176,149 variants were monomorphic, remained in the data set.

Phenotype transformation. For the HKU-TR5 data set, log transformation was first performed on TG and HDL-C. Phenotypes were then defined as the standardized residuals of blood lipid levels after adjustment on age, gender and two principal components computed from ~22K common SNPs (MAF > 5%). For the PUUMA-MI data, blood lipid levels were transformed to rank-based inverse normal residuals to minimize the impact of outliers and ensure normality. Age, sex, MI status (case or control) and 10 principal components estimated from AIMs were included as covariates in the association tests.

Meta-analysis for lipid traits. We then proceeded with meta-analyses of both single-variant and gene-based association tests using a framework of rvtests coupled with RAREMETAL. First, in each cohort, we carried out single-variant association tests for all markers passing quality controls using rvtests (http://zhanxw.github.io/rvtests/). The computed score test statistics as well as the corresponding variance-covariance matrix, which summarizes the LD between markers, were shared between studies. Overall, 145,276 SNPs were polyomorphic in at least 1 individual in the combined data set.

Next, we performed single-variant meta-analysis by combining the score test results using RAREMETAL. We restricted the analysis to SNPs that were polyomorphic in both studies and had at least 20 minor alleles in the combined data set (MAC > 20). Exome-wide significance threshold was defined as 2.69 × 10−7, which accounted for multiple testing of 65,671 SNPs passing the filtering criteria (see next section). The test statistics were visualized in a quantile–quantile plot, well-appeared calibrated (Supplementary Fig. 10).

To leverage power for detecting low frequency and rare variants, we performed gene-based meta-analyses, including a CMC24, VT25 and SKAT25, to evaluate their aggregate effects in each gene. We only included loss-of-function (stopgain and frameshift missense variants predicted to be damaging into the gene set). Damaging variants were predicted by Kpgseq23, which combined multiple functional prediction methods (for example, SIFT, PolyPhen2 and CADD) to predict pathogenic variants. Two MAF thresholds (<1% and <5%) were employed for CMC and SKAT, whereas only MAF<5% threshold was considered for VT. Genes with at least 20 copies of rare alleles were considered (n = 8,369 for MAF<1% and 9,025 for MAF<5%). Gene-based significance threshold was defined as P<8.43 × 10−7, which was equivalent to the multiple testing correction of 59,313 effective number of tests (see next section).

Meta-analysis for CAD. Logistic regression was carried out in HKU-TR5 and PUUMA-MI to evaluate if the newly identified lipid-associated SNPs influence CAD risk. The effect estimates and s.e. were meta-analysed using METAL by the fixed-effect inverse-variance method. Considering multiple testing of seven association tests, we defined SNPs with P<7.1 × 10−5 as being associated with CAD.

Identification of independent signals. To delineate SNPs with independent association, we performed stepwise conditional analysis using the known lipid-associated variants within 500 kb of the lead SNPs as covariates by RAREMETAL. For each independent SNPs with conditional P values reaching exome-wide significance (P<2.69 × 10−7) were considered as having independent association and the lead SNPs showing strongest evidence of independent association was later included in the regression model as a covariate. The process was repeated until no variant reached exome-wide significance after adjustment.

Estimating the effective number of independent tests. While the 4 blood lipid traits are highly correlated and there existed LD among common and low-frequency variants, traditional Bonferroni correction for multiple testing of all 262,684 correlated single-variant tests (P<1.90 × 10−10) was likely to be too stringent, leading to an overcorrection of the association. To take into account of the dependencies, we estimated the effective number of independent tests (M⊥) for both phenotypes and genotypes using methods based on eigenvalues of the corresponding correlation matrix69. We estimated the M⊥ to be 3.41 for phenotypes and 54,463 for genotypes, totalling 185,719 independent single-variant tests. Exome-wide significance threshold for single-variant meta-analysis was therefore defined as 2.69 × 10−7 (=0.05/185,719). Similarly, for gene-based meta-analysis, exome-wide significance threshold was defined as P<8.43 × 10−7, equivalent to multiple testing correction of 59,313 (8,369 + 9,025) × 3.41 independent tests for the 8,369 and 9,025 genes with MAF<1% and <5%, respectively.

Asian-specific variants. According to the ancestry of samples (http://www.1000genomes.org), we divided the individuals of 1000 Genomes into three popula-tions: Asian (CHB, CHS and JPT), European (CEU, FIN, GBR, IBS and TSI) and African (ASW, LWK and YRI). For each population, the MAF was estimated separately. Individuals were classified as Asian-specific if they were polymorphic (MAF>0) in the Asian population but monomorphic in both the European and African populations.
Acknowledgements
This work was supported by the Hong Kong Research Grant Council: Theme Based Research Scheme (T12-705/11), General Research Fund (777511M, 776412M, 776513M and 17128515) and Innovation and Technology Support Programme (Tier 3) (ITS/303/12). We thank Joyce MS Fung, KH Lee and TK Leung for their technical support. The PUUMA-MI project is a collaboration between Peking University Health Science Center and The University of Michigan Medical School. Funding for the project was provided by the University of Michigan Medical School and the Peking University Health Sciences Center Joint Institute for Clinical and Translational Research (Principal Investigators: C.J.W. and W.G.). We thank the volunteer study participants for their contribution to research. We also thank the University of Michigan leadership of the Joint Institute for facilitating this research, particularly David Finsky, Amy Huang and Joseph Kolars. C.J.W. is additionally supported by HL094535 and HL109946.

Author contributions

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.