A unified framework identifies new links between plasma lipids and diseases from electronic medical records across large-scale cohorts

Yogasudha Veturi¹, Anastasia Lucas¹, Yuki Bradford¹, Daniel Hui¹, Scott Dudek², Elizabeth Theusch², Anurag Verma¹, Jason E. Miller¹, Iftikhar Kullo³, Hakon Hakonarson⁴, Patrick Sleiman⁴, Daniel Schaid⁵, Charles M. Stein⁶, Digna R. Velez Edwards⁷,⁸,⁹, QiPing Feng⁶, Wei-Qi Wei⁷, Marisa W. Medina², Ronald M. Krauss², Thomas J. Hoffmann¹⁰, Neil Risch¹⁰, Benjamin F. Voight¹¹,¹², Daniel J. Rader¹¹,¹⁳ and Marylyn D. Ritchie¹¹,¹⁵

Plasma lipids are known heritable risk factors for cardiovascular disease, but increasing evidence also supports shared genetics with diseases of other organ systems. We devised a comprehensive three-phase framework to identify new lipid-associated genes and study the relationships among lipids, genotypes, gene expression and hundreds of complex human diseases from the Electronic Medical Records and Genomics (347 traits) and the UK Biobank (549 traits). Aside from 67 new lipid-associated genes with strong replication, we found evidence for pleiotropic SNPs/genes between lipids and diseases across the phenome. These include discordant pleiotropy in the HLA region between lipids and multiple sclerosis and putative causal paths between triglycerides and gout, among several others. Our findings give insights into the genetic basis of the relationship between plasma lipids and diseases on a phenome-wide scale and can provide context for future prevention and treatment strategies.

Plasma lipids, including total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TGs), are heritable risk factors for atherosclerotic cardiovascular disease. Previous meta-analyses, electronic health record (EHR)-based studies and large-scale biobanks have identified hundreds of loci associated with lipids using genome-wide association studies (GWAS). In addition, transcriptome-wide association studies (TWAS) have identified several genes whose cis-expression levels have been implicated in lipid traits as well as a host of other complex traits and diseases. However, one of the challenges has been validating the robustness of the results obtained using different methods across multiple cohorts.

Our primary hypothesis was that we could identify a robust set of lipid-associated genes by integrating tissue-specific gene expression with genotype and examining the extent of their replication across multiple large-scale cohorts that adopted different study designs. We devised an integrative framework that combines TWAS with statistical colocalization and conditional analyses (using tissue-specific weights from the Genotype-Tissue Expression (GTEx) project (v8)). As part of phase I of our study (Fig. 1), we detected several lipid-associated genes that replicated across multiple cohorts for the same trait–tissue pair. These included a meta-analyzed cohort (Global Lipids Genetics Consortium (GLGC)), an EHR-based cohort (Genetic Epidemiology Resource on Adult Health and Aging (GERA)), a meta-analysis multisite cohort (Electronic Medical Records and Genomics (eMERGE³)) and a population-based biobank (UK Biobank (UKB³)).

Plasma lipids have also been known to be associated with diseases pertaining to multiple organ systems, including diseases of the musculoskeletal system (rheumatoid arthritis¹⁰), skin and subcutaneous tissue (psoriasis¹¹), the circulatory system (coronary heart disease¹²) and the nervous system (multiple sclerosis¹³ and Alzheimer’s disease¹⁴). Statistical pleiotropy (statistical association of a genetic variant with multiple traits) can dissect the genetic basis of interrelationships between lipids and diseases. Long established in model organisms, pleiotropy is pervasive among 90% of loci listed in the GWAS Catalog¹⁴,¹⁵. Aside from studies that present a global view of pleiotropy¹⁴,¹⁵, previous studies have identified pleiotropic relationships among lipids and coronary artery disease¹¹, immune-related disorders¹², cardiometabolic traits¹³ and chronic inflammation¹⁴, as well as between coronary artery disease and nervous system disorders¹⁵. However, the underlying genetic mechanisms that link lipid levels to the broad spectrum of diseases in EHRs, also known as electronic medical records, have not been comprehensively investigated in multiple large-scale cohorts using multi-omics data. To understand the genetic interrelationships between plasma lipids and diseases across the

¹Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ²Department of Pediatrics, University of California, San Francisco, Oakland, CA, USA. ³Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA. ⁴Center for Applied Genomics, Children’s Hospital of Philadelphia, Philadelphia, PA, USA. ⁵Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. ⁶Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. ⁷Department of Biomedical Informatics in School of Medicine, Vanderbilt University, Nashville, TN, USA. ⁸Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA. ⁹Institute of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN, USA. ¹⁰Institute for Human Genetics, and Department of Epidemiology & Biostatistics, University of California, San Francisco, San Francisco, CA, USA. ¹¹Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ¹²E-mail: marylyn@pennmedicine.upenn.edu
phenome, we devised a second integrative framework using data on adults of European ancestry from the eMERGE and UKB cohorts as part of phase II of this study (Fig. 1). This additional framework incorporates lipid-guided phenome-wide association studies (PheWAS), gene expression-based phenome-wide association studies (Xpress-PheWAS) and statistical colocalization between EHR and GTEx v8-based gene expression summary statistics. Finally, as part of phase III of this study, we performed two-sample Mendelian randomization (MR)\(^2\) with lipids (from GERA and UKB) as the exposure and EHRs (from UKB and eMERGE) as the outcome (Fig. 1). This overall framework can (1) visualize the complete landscape of pleiotropy between lipids and diseases (including effects that are concordant, that is, have the same direction of effect, and discordant, that is, have the opposite direction of effect) and (2) identify diseases for which lipids could be modifiable exposures. We present a comprehensive overview of the complex interplay among lipids, genetics, gene expression and diseases in the EHRs. The detected genes/variants could be used as targets for functional validation and downstream drug repurposing studies.

**Results**

**Study workflow.** As outlined in the study workflow (Fig. 1), we first performed GWAS on lipid traits in eMERGE and UKB adults of European ancestry and used the summary statistics to conduct TWAS and statistical colocalization on lipid traits (S-PrediXcan\(^7\)) in eMERGE, UKB, GERA and GLGC. We then conducted ‘lipid-guided’ PheWAS, Xpress-PheWAS and statistical colocalization on curated International Classification of Diseases (ICD) Clinical Modification (CM) diagnosis codes from eMERGE and UKB cohorts using SNPs mapping within 1 Mb of lipid-associated genes derived from lipid TWAS. For the lipid-guided PheWAS and Xpress-PheWAS, we had predominantly ICD-9-CM codes (~82%) in the eMERGE network and ICD-10 disease codes (~98%) in the UKB, which we collapsed into three-character parent codes (see Extended Data Fig. 1 for case–control distribution). While the lipid-guided PheWAS and Xpress-PheWAS helped identify potential pleiotropic SNPs/genes between lipids and diseases, MR analyses helped identify diseases for which lipids could be modifiable exposures. Below, we describe results derived from each of the steps delineated above.

**Phase I—discovery and replication of new lipid-associated genes.** Lipid GWAS on adults of European ancestry from eMERGE (n = 31,575) and UKB (n = 377,921) cohorts revealed the breadth of signals across the four lipid traits (Extended Data Figs. 2 and 3). In addition, we used pre-published lipid GWAS summary statistics from GERA (n = 76,627) and GLGC (n = 188,578) for the lipid TWAS. Supplementary Figs. 1 and 2 show the extent of overlap in GWAS and TWAS signals across the four lipid traits in the four cohorts. For lipid TWAS, we integrated summary statistics from each of the four cohorts with tissue-specific weights derived from four tissues most relevant to lipid metabolism (adipose subcutaneous, adipose visceral omentum, liver, whole blood, small intestine)\(\ast\) software = S-PrediXcan\(^7\) and ICD-10 disease codes (~98%) in the UKB, which we collapsed into three-character parent codes (see Extended Data Fig. 1 for case–control distribution).

**Phase II—phenome-wide.** First performed GWAS on lipid traits in eMERGE and UKB adults (2,447,442 SNPs) and UKB (2,447,442 SNPs) and GTEx v8-based gene expression summary statistics. Finally, as part of phase II of this study, we performed two-sample Mendelian randomization (MR)\(^2\) with lipids (from GERA and UKB) as the exposure and EHRs (from UKB and eMERGE) as the outcome (Fig. 1). This overall framework can (1) visualize the complete landscape of pleiotropy between lipids and diseases (including effects that are concordant, that is, have the same direction of effect, and discordant, that is, have the opposite direction of effect) and (2) identify diseases for which lipids could be modifiable exposures. We present a comprehensive overview of the complex interplay among lipids, genetics, gene expression and diseases in the EHRs. The detected genes/variants could be used as targets for functional validation and downstream drug repurposing studies.

**Phase III—lipid mediation.** First performed GWAS on lipid traits in eMERGE and UKB adults (2,447,442 SNPs) and UKB (2,447,442 SNPs) and GTEx v8-based gene expression summary statistics. Finally, as part of phase II of this study, we performed two-sample Mendelian randomization (MR)\(^2\) with lipids (from GERA and UKB) as the exposure and EHRs (from UKB and eMERGE) as the outcome (Fig. 1). This overall framework can (1) visualize the complete landscape of pleiotropy between lipids and diseases (including effects that are concordant, that is, have the same direction of effect, and discordant, that is, have the opposite direction of effect) and (2) identify diseases for which lipids could be modifiable exposures. We present a comprehensive overview of the complex interplay among lipids, genetics, gene expression and diseases in the EHRs. The detected genes/variants could be used as targets for functional validation and downstream drug repurposing studies.

**Phase I—discovery and replication of new lipid-associated genes.** Lipid GWAS on adults of European ancestry from eMERGE (n = 31,575) and UKB (n = 377,921) cohorts revealed the breadth of signals across the four lipid traits (Extended Data Figs. 2 and 3). In addition, we used pre-published lipid GWAS summary statistics from GERA (n = 76,627) and GLGC (n = 188,578) for the lipid TWAS. Supplementary Figs. 1 and 2 show the extent of overlap in GWAS and TWAS signals across the four lipid traits in the four cohorts. For lipid TWAS, we integrated summary statistics from each of the four cohorts with tissue-specific weights derived from four tissues most relevant to lipid metabolism (adipose subcutaneous, adipose visceral omentum, liver, whole blood, small intestine)\(\ast\) software = S-PrediXcan\(^7\) and ICD-10 disease codes (~98%) in the UKB, which we collapsed into three-character parent codes (see Extended Data Fig. 1 for case–control distribution).
adapted shrinkage to calculate effect sizes on fine-mapped variants obtained using Deterministic Approximation of Posteriors. We devised a new workflow built upon a previous study (Methods) that could prioritize genes in downstream functional analyses to assess causality.

Figure 2 shows lipid–gene associations among autosomal chromosomes for each tissue–cohort combination. We obtained 1,033 Bonferroni-significant genes in total ($P < 5.57 \times 10^{-8}$). These included 79 new genes and 954 previously reported genes (see Methods for details of how genes were classified into new versus previously reported). We subsequently filtered out the linkage disequilibrium (LD)-contaminated genes, a scenario in which gene expression-predictor SNPs (expression quantitative trait loci (eQTL)) and phenotype causal SNPs (GWAS) are different but are in LD. Herein, we only retained genes that had at least one SNP with coloc$^9$ H3 probability $< 0.5$ between GWAS and eQTL datasets for a given lipid–tissue combination. Further, we also conditioned the SNPs at a locus on the top eQTL at that locus (GCTA-COJO$^{30}$) to detect if there are potential secondary independent associations at the locus. Code for identifying LD-contaminated genes and detecting secondary independent associations at a locus is shared on GitHub (https://github.com/RitchieLab/Gene-level-statistical-colocalization/).

In total, 67 new genes that replicated for the same lipid–tissue combination in at least two cohorts also cleared the coloc H3 filter in at least one cohort (Supplementary Table 1). Extended Data Figs. 4 and 5 show across cohorts and tissues (1) the TWAS strength and direction of effects and (2) coloc H4 probabilities, respectively, for the 67 new genes. Figure 2 and Supplementary Tables 1 and 2 also show the extent of replication of genes for a lipid–tissue pair across the four chosen cohorts. We were able to replicate well-known proof-of-concept genes such as CELSR2, SORT1 and PSRC1 on chromosome 1 (PSRC1 replicated in all four cohorts for the same lipid–tissue combination; Fig. 2), as well as ANGPTL3 and PCSK9 (chromosome 1), APOA1 locus (chromosome 11) and PLTP (chromosome 20). We also saw replication in the four cohorts across all five tissues for previously reported genes such as NRBP1 for TGs (chromosome 2), APOAI for HDL-C (chromosome 11), LPL for HDL-C and TGs (chromosome 8) and TMEM258 for TC (chromosome 11). Finally, of the 67 new genes from lipid TWAS with no LD contamination, 41 genes only ‘replicated’ in two cohorts, 18 genes only in three cohorts and 3 genes in all four cohorts with the same direction of effect for a lipid–tissue combination. The four-way replicating genes with coloc $P[H4] > 0.2$ were ZSWIM1 for HDL-C in adipose subcutaneous tissue (chromosome 20) and RP11-136012.2 for HDL-C in liver (chromosome 8). New genes with coloc $P[H4] > 0.5$ included DNAH10OS (dynein axonemal heavy chain 10 opposite strand; chromosome 12), IFI35 (chromosome 17), LILRB1 (chromosome 19), LINC00243 (chromosome 6), RP1-81D8.3 (chromosome 6), RP11-115F16.2 (chromosome 8), RP11-3N5N3.2 (chromosome 2), RAPAP2 (chromosome 1), SDCBP (chromosome 8), XXbac-BPG181B237.3 (also known as LINC01149; chromosome 6) and ZSWIM1 (chromosome 20). LocusZoom plots (Supplementary Figs. 3–14) reveal the strength of lipid and gene expression signal in the region surrounding the shared index SNP for each of these new loci. Finally, among new genes, we also observed evidence of secondary independent associations at the LILRB1 locus (chromosome 19) for HDL-C in whole blood/small intestine (Supplementary Table 1 and Supplementary Fig. 12). Among previously reported genes, we found evidence for secondary signals at APC2, CCDC92 and FADS2, among others (Supplementary Table 2). Although the strength of signal for new genes was lower than for previously reported genes, they replicated in two or more cohorts for the same lipid–tissue combination (Fig. 2), making them targets for further validation.

Phase II—discovery and replication of ICD disease codes. Once we had our list of 1,033 lipid-associated genes, we devised a workflow that only used SNPs mapping within a 1-Mb region of each of these genes and also overlapped MASHR-based prediction models$^{30}$ from PredictDB on GTEx v8 release data$^{1,31,32}$ across 49 available tissues. This resulted in 17,740 and 18,261 SNPs from eMERGE and UKB cohorts, respectively (Methods). We collapsed the ICD codes into three-character parent codes (347 traits in eMERGE and 549 traits in UKB) after excluding nonheritable codes and those with fewer than 200 cases (Methods). We then conducted lipid-guided PheWAS and used the summary statistics to conduct Xpress-PheWAS in both cohorts across 49 tissues available in PredictDB for GTEx v8.

Figure 3 shows the lipid-guided PheWAS and Xpress-PheWAS results from the eMERGE and UKB cohorts, respectively. In this rotated ‘Hudson’ plot, we see SNP-based signals in the lipid-guided PheWAS plots for eMERGE and UKB (right-hand side of each rotated Hudson plot) and gene-based signals in the Xpress-PheWAS plots (left-hand side of each rotated Hudson plot). We also provide
After mapping the eMERGE ICD-9 codes to ICD-10 using general equivalence mappings, we observed strong replication of PheWAS signals at the Bonferroni threshold ($8.265 \times 10^{-9}$ for eMERGE and $4.987 \times 10^{-9}$ for UKB; Methods) in both cohorts. The replicated diseases/SNP variants spanned diseases of metabolic, nutritional and endocrine, musculoskeletal, circulatory and nervous systems. There were 18 ICD codes that were detected by lipid-guided PheWAS and Xpress-PheWAS in both eMERGE and UKB (Extended Data Fig. 6), of which six also cleared the coloc $P[H3] < 0.5$ filter in Xpress-PheWAS, that is, they had no LD contamination (Supplementary Table 5). These included hypercholesterolemia/disorders of lipoprotein metabolism (chromosome 1), rheumatoid arthritis (chromosome 6 HLA region), pulmonary embolism (chromosome 6 LPA region) and Alzheimer’s disease and senile dementia (chromosome 19). Diseases replicating in the HLA region were all autoimmune diseases (see Supplementary Table 5 for a complete list of detected diseases).

**Phase II—lipid–disease pleiotropy in either cohort.** So far, our lipid-guided PheWAS resulted in SNPs that map not just to the lipid-associated genes (from lipid TWAS) but also to some genes neighboring them. Next, we considered lipid-associated SNPs that were also strictly associated with diseases in either eMERGE or the UKB. Extended Data Fig. 7 shows the number of Bonferroni-significant SNPs overlapping between lipid GWAS (eMERGE, GERA, GLGC and UKB) and lipid-guided PheWAS (eMERGE and UKB). Given its large sample size, the vast majority of lipids

---

**Fig. 3 | Comparison of results between Xpress-PheWAS and lipid-guided PheWAS in eMERGE and UKB.** Plots indicate gene signals from Xpress-PheWAS (left) and SNP signals from lipid-guided PheWAS (right) for eMERGE (left) and UKB (right). In these rotated Hudson plots, the shape of the point indicates the direction of effect. The colors indicate the corresponding disease category. The red lines indicate the Bonferroni-significant thresholds in either cohort for PheWAS (two-sided logistic regression) and Xpress-PheWAS (Methods). For clarity, we truncated the results for the UKB on chromosomes 6 and 2 (indicated by asterisks). All indicated points in the plot map to SNPs that lie within 1Mb of lipid-associated genes from lipid TWAS. See https://ritchie-lab.org/naturegenetics/eMERGE_2020-12-11_scaled.html and https://ritchie-lab.org/naturegenetics/UKB_2020-12-11_scaled.html for interactive versions of this plot.
we have delineated the ICD-9 or ICD-10 codes corresponding to the disease categories, along with the genes associated with each disease category. These genes also have ... pleiotropic associations between lipids and diseases was observed across the genome (Extended Data Fig. 8). We observed the greatest overlap between ICD codes and LDL–C/TC-associated SNPs, specifically for diseases of metabolic, endocrine, circulatory and digestive systems (Extended Data Table 3). In addition, we detected lipid-guided PheWAS associations (P < 4.987 × 10⁻⁸) between Bonferroni-significant lipid SNPs (P < 5.000 × 10⁻⁸) and 73 ICD codes from eMERGE or the UKB. In addition to ICD codes specified above, detected diseases included gonarthrosis, nasal polypl, retinal disorders, benign neoplasms of colon, rectum, anus and anal canal, malignant neoplasms of skin, follicular non-Hodkin’s lymphoma, female genital prolapse, hyperplasia of prostate, cholelithiasis and asthma among others (Supplementary Tables 3 and 5).

Figure 4 shows the Bonferroni-significant genes (rather than SNPs) from Xpress-PheWAS (P < 5.445 × 10⁻¹⁰ for eMERGE and P < 7.262 × 10⁻¹⁰ for UKB) that were associated with ICD codes in either eMERGE or the UKB as well as lipid traits from lipid TWAS (P < 1.390 × 10⁻⁵). In addition, these genes cleared the coloc P[H3] < 0.5 filter for lipid traits and ICD codes in at least one tissue. Similarly to lipid-guided PheWAS, the majority of signal came from the UKB; 125 genes detected from Xpress-PheWAS in the UKB overlapped with previously reported lipid genes without the coloc filter (Extended Data Fig. 9). Again, we found evidence of pleiotropy between lipids and a range of disease categories, most of which were detected from lipid-guided PheWAS: 45 ICD codes were detected from lipid-guided PheWAS and Xpress-PheWAS in the UKB (Extended Data Fig. 6). In addition, with the coloc P[H3] < 0.5 filter, Xpress-PheWAS exclusively detected Bonferroni-significant lipid genes also associated with hematuria, leiomomyoma of uterus and family history of chronic diseases (Supplementary Tables 4 and 5). LocusZoom plots (Supplementary Figs. 15–23) reveal the likely causal variant colocalizing between (1) lipids and gene expression and (2) ICD codes and gene...
expression for a TWAS-significant lipid gene. We also detected putative secondary independent associations at ABO for hemorrhoids, and at new lipid genes LINC00243 for intestinal malabsorption and Xxbac-BPG181B23.7 for hypothyroidism (Supplementary Table 4 and Supplementary Fig. 23).

Phase II—lipid–disease pleiotropy in both cohorts. For lipid-guided PhEWS, we found a smaller subset of ICD codes when we only considered suggestive pleiotropic variants that replicated in both eMERGE and the UKB, on chromosomes 1, 6, 9 and 19 (Fig. 5). These included proof-of-concept SNPs on chromosome 1 that were associated with HDL-C, LDL-C and TC as well as disorders of lipoprotein metabolism and mapped to the previously known CELSR2, SORT1 and PSRC1 lipid genes on chromosome 1. These SNPs had the same direction of effect (protective) for the disease and the lipid traits (Fig. 5), consistent with previous studies. We were also able to replicate SNPs associated with all four lipid traits and Alzheimer’s disease, mapping to the known APOE and TOMM40 genes on chromosome 19, as well as pulmonary embolism, mapping to the previously known ABO gene on chromosome 9 (refs. 35,36). For both pulmonary embolism and Alzheimer’s disease, direction of effect was largely consistent between lipids and disease (discordant). Risk alleles for lipids (positive direction of SNP effect for LDL-C, TC and TGs and negative direction of effect for HDL-C) were also seen to be risk alleles for Alzheimer’s disease and pulmonary embolism and likewise for protective alleles (Fig. 5). We also found SNPs in the HLA region on chromosome 6 that were jointly associated with lipids and autoimmune diseases (seropositive rheumatoid arthritis, multiple sclerosis, hypothyroidism, psoriasis and ulcerative colitis) and insulin-dependent diabetes mellitus. Of these, we saw the opposite direction of effect (risk versus protective) for multiple sclerosis (discordant) and the same direction of effect (concordant) for seropositive rheumatoid arthritis (Fig. 5). In other words, SNPs that led to an increase in lipid levels (risk) were associated with decreased effect among multiple sclerosis cases (protective), whereas SNPs that led to a decrease in lipid levels (protective) led to an increased effect among multiple sclerosis cases (risk). The opposite was true for seropositive rheumatoid arthritis. Finally, SNP rs118039278 mapped to the LPA gene on chromosome 19 and was found to be associated with HDL-C, TGs, TC and LDL-C, as well as angina pectoris, nonrheumatic aortic valve disorders, chronic ischemic heart disease and disorders of lipoprotein metabolism.

Next, we only considered suggestive pleiotropic ‘genes’ (as opposed to SNPs) that replicated at the Bonferroni-significance threshold in both eMERGE and the UKB on a smaller subset on chromosomes 1, 6, 9 and 19, in addition to having coloc P[HI] < 0.5 (no LD contamination) in both, lipids and diseases (Fig. 6). Similarly to our lipid-guided PhEWS, we were able to replicate a protective effect of proof-of-concept lipid genes CELSR2, SORT1 and PSRC1 on disorders of lipoprotein metabolism. Finally, we were able to replicate many of the signals found on chromosomes 6, 9 and 19 from lipid-guided PhEWS. These genes were associated with pulmonary embolism (ABO) and Alzheimer’s disease (TOMM40 and APOC1). We were also able to replicate the (tissue-specific) protective/risk effect of genes for these diseases. Finally, we detected a new lipid gene Xxbac-BPG181B23.7 on chromosome 6, which was also associated with ankylosing spondylitis in both cohorts (Fig. 6). This long noncoding gene in the HLA region was also found to be associated with ten other diseases in the UKB only (hypothyroidism, multiple sclerosis, psoriasis, asthma, rheumatoid arthritis, insulin-dependent diabetes mellitus, disorders of lipoprotein metabolism, psoriatic and enteropathic arthropathies, iridocyclitis and intestinal malabsorption; Fig. 4 and Supplementary Figs. 17 and 18).

Phase III—lipids as modifiable exposures for disease. Thus far, we detected SNPs and genes that are suggestive of pleiotropy between lipids and diseases. However, many diseases (especially cardiovascular) are lipid mediated using curated and independent genetic instruments (SNPs) across the genome. To better understand the
Articles NATURE GENETICS

In this study, we implemented a comprehensive integrative framework (Fig. 1) to shed light on the landscape of new and previously reported genetic mechanisms linking lipids to phenome-wide diseases (Fig. 8) in two large cohorts (eMERGE and the UKB). This study was conducted in three phases. In phase I, we developed a framework that integrates TWAS based on fine-mapped eQTLs with statistical colocalization to identify new genes associated with plasma lipids based on the extent of replication in a lipid–tissue pair across four different cohorts: eMERGE, GERA, GLGC and the UKB (Fig. 2). We detected 79 new lipid genes from lipid TWAS (67 of which also cleared colocal $P[H3] < 0.5$ filter in at least one cohort) and 954 previously reported lipid genes, including proof-of-concept genes such as SORT1. Among the replicating new genes with colocal $P[H3] < 0.5$ and colocal $P[H4] > 0.5$, DNASA10OS is a protein-coding gene previously found to be associated with body mass index (BMI) and waist-to-hip ratio (note that DNASA10 is a known lipid gene); ZSWIM1 (phenylacetyl-glutamine) at the PLTP locus is a lymphocyte-expressed gene previously detected as being lipid associated using a powerful gene-based test; RP1039N3.2 on chromosome 2 is a long intergenic noncoding RNA that has recently been implicated in waist-to-hip ratio/BMI. The new lipid-associated genes (Supplementary Table 1) could have evidence for causality and be selected as targets for validation using functional assays.

In the second phase of the study, we conducted lipid-guided PheWAS, Xpress-PheWAS (Figs. 3–6) and statistical colocalization to identify potentially pleiotropic associations between lipids and diseases, while in the third phase we conducted two-sample MR (Fig. 7) to detect diseases that are putatively causally associated with lipids. The study hypothesis was that plasma lipids are likely to have broad effects on complex human diseases across the phenome, given several previous studies that allude to direct links between lipids and diseases of multiple organ systems. To our knowledge, this is the most comprehensive study to date that has been carried out to test this hypothesis, and we used an extensive ensemble of methods that has previously not (1) been applied simultaneously on multiple large cohorts, (2) focused on detecting pleiotropy (discordant and concordant) between plasma lipids and diseases, (3) detected several new and previously reported SNPs and genes role that plasma lipids play as modifiable exposures in diseases across the phenome, we conducted univariable two-sample MR on a chosen subset of diseases. As shown in the workflow in Fig. 1, we ran these analyses in two sets. In the first set, we used the UKB as the exposure (lipid) dataset and eMERGE as the outcome dataset; in the second set, we used GERA as the exposure (lipid) dataset and the UKB as the outcome dataset. After LD clumping SNPs with $P < 5 \times 10^{-8}$ from exposure datasets, we had 183–206 SNPs in set 1 and 53–59 SNPs in set 2 across the selected disease codes (see Methods for protocol). Figure 7 shows the MR estimates and $P$ values for diseases that were significant after Bonferroni correction using at least one of three methods (inverse-variance weighted, Egger and median-based), while Extended Data Fig. 10 sheds light on all the SNP-specific MR effects for a chosen set of diseases from these analyses. Gout was a new disease code that we found to be putatively causally associated with lipids at the Bonferroni threshold in both sets. This association remained even after performing analyses upon excluding SNPs from the HLA region (Supplementary Fig. 24). Other new putative causal associations included disorders of the iris and ciliary body, hyperosmolality and hypernatremia, infective myositis, ingrowing nails and malignant neoplasms of the bronchus and cerebrum. We also found the expected corroboration of lipid-mediated traits such as hypercholesterolemia/disorders of lipoprotein metabolism (Supplementary Table 6), as well as other proof-of-concept diseases such as acute myocardial infarction, primary hypertension, acute ischemic heart disease and atherosclerosis (Fig. 7).

Discussion

In this study, we implemented a comprehensive integrative framework (Fig. 1) to shed light on the landscape of new and previously reported genetic mechanisms linking lipids to phenome-wide diseases (Fig. 8) in two large cohorts (eMERGE and the UKB). This study was conducted in three phases. In phase I, we developed a
as well as putative diseases with causal lipid associations at stringent multiple-comparison thresholds with replication and (4) integrated four types of data (genetic, gene expression, EHR and plasma lipid). While a previous study investigated the overall landscape of genome-wide pleiotropy, it had a high case threshold (>10,000) that resulted in several untested diseases. Also, it did not discuss extent of replication of results or focus on pleiotropy with lipids. Importantly, our analyses are very focused and time effective as we only ran PheWAS on approximately 18,000 fine-mapped eQTLs that mapped to lipid (and neighborhood) genes. Finally, we also present a tool to conduct colocalization integrated with conditional analyses on a chosen set of genes of interest. Our tool can not only identify TWAS-significant genes with no LD contamination but also detect secondary signals at a locus mapping to any gene(s) of interest.

Many signals from the HLA region (as reported in a previous UKB study) correspond to lipid and immune-related pathways (Supplementary Fig. 25). Notably, we saw the opposite direction (Supplementary Table 4). The p53 protein has been previously found to have a new role in regulating lipid metabolism pathways as well as putative diseases with causal lipid associations at stringent multiple-comparison thresholds with replication and (4) integrated four types of data (genetic, gene expression, EHR and plasma lipid). While a previous study investigated the overall landscape of genome-wide pleiotropy, it had a high case threshold (>10,000) that resulted in several untested diseases. Also, it did not discuss extent of replication of results or focus on pleiotropy with lipids. Importantly, our analyses are very focused and time effective as we only ran PheWAS on approximately 18,000 fine-mapped eQTLs that mapped to lipid (and neighborhood) genes. Finally, we also present a tool to conduct colocalization integrated with conditional analyses on a chosen set of genes of interest. Our tool can not only identify TWAS-significant genes with no LD contamination but also detect secondary signals at a locus mapping to any gene(s) of interest.

Fig. 7 | Two-sample univariable Mendelian randomization. Top: exposure dataset is GERA and outcome dataset is the UKB. Bottom: exposure dataset is the UKB and outcome dataset is eMERGE. The diseases are grouped into different categories; direction of triangle corresponds to direction of MR effect. In each panel, the two red horizontal lines correspond to the Bonferroni and false discovery rate (FDR) thresholds. We labeled FDR-significant ICD codes from at least one of two two-sided tests (inverse-variance weighted, Egger and median-based) with an Egger pleiotropy (intercept) \( P > 0.05 \) to have evidence of minimal heterogeneity. Filled circles have confidence intervals (CIs) that do not contain 0, whereas non-filled circles have CIs that contain 0.

DOE, direction of effect.
SNPs mapping to these genes were not detected from lipid-guided PheWAS, indicating that integrating gene expression information with SNPs likely boosted our signal. This helps us elucidate the underlying genetic architecture of the relationship between HDL-C and neoplasms for TP53. Conversely, lipid-guided PheWAS analyses detected three polymorphisms associated with female genetic prolapse (ICD code: N81) mapping to GDF7 on chromosome 2, of which one SNP (rs9306894), has been previously associated with pelvic organ prolapse in the UKB. Although we also detected GDF7 from Xpress-PheWAS, colocalization analyses filtered out this gene, revealing that PheWAS-based associations found in the literature might not always colocalize with eQTLs.

Finally, the limitations of this study are that TWAS and Xpress-PheWAS cannot distinguish between horizontal pleiotropy and direct gene expression mediation between SNPs and trait. They can also result in false-positive associations due to LD mismatch between GWAS and expression panel, underlying biases in expression panel and sharing of eQTLs with truly causal genes. We addressed this by using a comprehensive ensemble of methods in multiple cohorts with stringent multiple-comparison filters to reduce false-positive associations. Another caveat is that certain disease diagnosis codes are absent in one cohort and not the other, making it difficult to do equivalence mapping between cohorts based on ICD-9 and ICD-10 (for replication). A third caveat is the absence of some genes in prediction models of some tissues. We also restricted our analyses to common variants is the absence of some genes in prediction models of some tissues. A third caveat is the absence of some genes in prediction models of some tissues.

In conclusion, we characterized the landscape of pleiotropy between plasma lipids and diseases from EHR data using a comprehensive suite of methods. Our results provide fresh insights into the genetic relationships underlying lipids and diseases, while our integrative analytical framework can be applied to similarly study pleiotropy for other sets of traits.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00879-y.
30. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).

31. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

32. Gamazon, E. R. et al. A gene-based association mapping for methods targeting use reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

33. Butler, R. The ICD-10 General Equivalence Mappings. Bridging the translation gap from ICD-9. *J. AHIMA* **78**, 84–85 (2007).

34. Xu, L. et al. An association study between genetic polymorphisms related to lipoprotein-associated phospholipase A2 and coronary heart disease. *Exp. Ther. Med.* **5**, 742–750 (2013).

35. Wolpin, B. M. et al. Prospective study of ABO blood type and the risk of pulmonary embolism in two large cohort studies. *Thromb. Haemost.* **104**, 962–971 (2010).

36. Hajizadeh, R., Kavandi, H., Nadiri, M. & Ghaffari, S. The association of ABO blood group with incidence and outcome of acute pulmonary embolism. *Turk Kardiyol. Dern. Ars.* **44**, 397–403 (2016).

37. Zhang, J., Zhao, Z., Guo, X., Guo, B. & Wu, B. Powerful statistical method to detect disease-associated genes using publicly available genome-wide association studies summary data. *Genet. Epidemiol.* **43**, 941–951 (2019).

38. Lumish, H. S., O’Reilly, M. P. & Reilly, M. P. Sex differences in genomic drivers of adipose distribution and related cardiometabolic disorders: opportunities for precision medicine. *Arterioscler. Thromb. Vasc. Biol.* **40**, 45–60 (2020).

39. Reshef, Y. A. et al. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).

40. Cantuti-Castelvetri, L. et al. Defective cholesterol clearance limits remyelination in the aged central nervous system. *Science* **359**, 684–688 (2018).

41. Fard, M. K. et al. BCAS1 expression defines a population of early myelinating oligodendrocytes in multiple sclerosis lesions. *Sci. Transl. Med.* **9**, eaam7816 (2017).

42. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present and future. *Genetics* **193**, 651–669 (2013).

43. Ginn, L., Shi, L., La Montagna, M. & Garofalo, M. LncRNAs in non-small-cell lung cancer. *Noncoding RNA* **6**, 25 (2020).

44. Zhong, R. et al. LINC01149 variant modulates MICA expression that facilitates hepatitis B virus spontaneous recovery but increases hepatocellular carcinoma risk. *Oncogene* **39**, 1944–1956 (2020).

45. Feng, X. & Yang, S. Long noncoding RNA LINC00243 promotes proliferation and glycolysis in non-small-cell lung cancer cells by positively regulating PDK4 through sponging miR-507. *Mol. Cell. Biochem.* **463**, 127–136 (2020).

46. Yu, X., Chen, H., Huang, S. & Zeng, P. Evaluation of the causal effects of blood lipid levels on gout with summary level GWAS data: two-sample Mendelian randomization and mediation analysis. *J. Hum. Genet.* **66**, 465–473 (2021).

47. Marien, E. et al. Non-small-cell lung cancer is characterized by dramatic changes in phospholipid profiles. *Int. J. Cancer* **137**, 1539–1548 (2015).

48. Eggers, L. F. et al. Lipidomes of lung cancer and tumour-free lung tissues reveal distinct molecular signatures for cancer differentiation, age, inflammation and pulmonary emphysema. *Sci. Rep.* **7**, 11087 (2017).

49. Tiwary, S. et al. Metastatic brain tumors disrupt the blood–brain barrier and alter lipid metabolism by inhibiting expression of the endothelial cell fatty acid transporter Mfsd2a. *Sci. Rep.* **8**, 8267 (2018).

50. Sun, H., Zhang, X., Shi, W. & Fang, B. Association of soft tissue infection in the extremity with glucose and lipid metabolism and inflammatory factors. *Exp. Ther. Med.* **17**, 2535–2540 (2019).

51. Gao, S., Cui, X., Wang, X., Burg, M. B. & Dmitrieva, N. I. Cross-sectional positive association of serum lipids and blood pressure with serum sodium within the normal reference range of 135–145 mmol/l. *Arterioscler. Thromb. Vasc. Biol.* **37**, 598–606 (2017).

52. Goldstein, I. et al. p53, a novel regulator of lipid metabolism pathways. *J. Hepatol.* **56**, 656–662 (2012).

53. Mäkinen, N. et al. Exome sequencing of uterine leiomyosarcomas identifies frequent mutations in TP53, ATRX and MED12. *PLoS Genet.* **12**, e1005850 (2016).

54. Parrales, A. & Iwakuma, T. p53 as a regulator of lipid metabolism in cancer. *Int. J. Mol. Sci.* **17**, 2074 (2016).

55. Veturi, Y. & Ritchie, M. D. How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures? *Pac. Symp. Biocomput.* **23**, 228–239 (2018).

56. Olafsdottir, T. et al. Genome-wide association identifies seven loci for pelvic organ prolapse in Iceland and the UK Biobank. *Commun. Biol.* **3**, 129 (2020).

57. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Methods

Ethics. Research conducted in this study complies with all ethical regulations laid out in the Declaration of Helsinki. This study was performed in the eMERGE Network, which is a funded consortium sponsored by the National Human Genome Research Institute (NHGRI) that combined biobiospecimen and EHR data across leading medical institutions. All studies were approved by the institutional review boards of each respective institution. Each participant gave consent for being part of the DNA biobanks. Data from the UKB for this project pertained to application no. 32133.

Datasets. Individual-level data were obtained from the eMERGE network phase III and the UKB. eMERGE network phase III comprises 99,185 genotyped samples across multiple platforms that were imputed to Haplotype Reference Consortium Consortium 1.1 and covered approximately 39 million SNP variants across 78 array genotype batches. The eMERGE sites included in our study were Marshfield Clinic Research Foundation, Vanders University Medical Center, Washington University, University of California, San Francisco, Massachusetts General Hospital, and the University of Alabama at Birmingham. We subsequently retained only those ICD codes that had at least 200 cases.

Because we focused on adults only, we did not include individuals from Boston Children's Hospital, Cincinnati Children's Hospital Medical Center and the Children's Hospital of Philadelphia. The UKB release 2 has deep genetic and phenotypic data on ~500,000 individuals across the United Kingdom that were genotyped on two genotype arrays across 106 batches and imputed to 96 million variants. We used the 'best-practice' quality-control (QC) pipeline to clean eMERGE phase III imputed genotypic data. We included genetic variants with a genotype call rate of 99.5% or greater, a Hardy-Weinberg equilibrium (HWE) p-value > 10−8, and a minor allele frequency (MAF) > 1%. We also excluded SNPs with HWE p-values < 10−8 and removed individuals with ambiguous sex. There were 7,666,566 SNPs and 47,229 unrelated European American individuals remaining after QC procedures.

For QC in the UKB, we excluded individuals with poor quality genotyping according to a previous publication. We dropped one of a pair of related individuals with p< 0.25 and those with mismatches between self-reported and genetically inferred sex. We also excluded variants with an imputation information score < 0.3 and minor allele frequency < 1%. We only retained European American individuals and estimated identity by descent using PLINK (1.9). We dropped one of a pair of related individuals with p< 0.25. We also excluded SNPs with Hardy-Weinberg equilibrium exact test P < 1 x 10−14 and removed individuals with ambiguous sex. There were 7,666,566 SNPs and 47,229 unrelated European American individuals remaining after QC procedures.

Summary-level data for the four lipid traits were obtained from (1) the GLGC consortium, (2) the UKB where we included all individuals, and (3) the eMERGE study where we only included individuals from the GERA5 cohort, which comprised 76,627 non-Hispanic white individuals with available to us had multiple measurements for each individual, so we decided to use median lipid values for GWAS analyses. For individuals with an even number of measurements, the median lipid value was chosen as whichever of the two central measurements was closest to the mean, and the age at which that measurement was made was used as the associated age in subsequent analyses. For individuals with only one lipid measurement, that measurement was considered the median. For individuals who had only two measurements or identical measurements on different dates, the earliest date on which that value was measured was chosen as the associated age. We filtered on age > 18 years. We removed individuals with phenotypic values greater than three times the standard deviation because they skewed the distribution, and we log transformed TG values to approximate a normal distribution. We sex stratified the phenotypes and regressed them on age, age-squared, batch, statin medication and the first six ancestry-derived PCs. Statin medication was a binary variable corresponding to whether or not an individual received statin medication. We also dropped to an increased number of sex-stratified GWAS analyses. All GWAS analyses were run on PLINK (2.0)24. The genomic control inflation was < 1.1 for each of the four lipid traits with the lowest being 1.04 (for LDL-C). After QC assessment, there only remained European American adults, including n = 31,565 (14,775 males and 16,790 females) for HDL-C, n = 30,509 (14,374 males and 16,135 females) for LDL-C, n = 31,575 (14,747 males and 16,828 females) for TC and n = 31,074 (14,638 males and 16,436 females) for TGs in eMERGE.

For the UKB, similarly to eMERGE, we used medians of first two lipid measurements in our GWAS analyses. We again removed individuals with phenotypic values greater than three times the standard deviation, and we log transformed TG values to approximate a normal distribution. We sex stratified the phenotypes and regressed them on age, age-squared, batch, lipid medication (codes 6153 and 6177) and 20 ancestry-derived PCs. The lipid medication variable was dichotomized by setting all individuals that received cholesterol-lowering medication to one and then setting individuals to zero who did not respond. Similarly to eMERGE, we stacked together the sex-stratified inverse-normalized residuals and used them as the response in all subsequent GWAS analyses. All GWAS analyses were run on PLINK (2.0).

Lipid TWAS framework. We used tissue-specific weights from five tissues in the GTEx Consortium v8: adipose subcutaneous (n = 581), adipose visceral omentum (n = 469), blood (n = 670), small intestine terminal ileum (n = 174) and liver (n = 208) using the MASHR model (http://www.predictdb.org/). S-PrediXcan was used to perform TWAS analyses for the four lipid traits in all four cohorts (eMERGE, UKB, GERA and GLGC). We devised a new workflow by building upon the one proposed by Barbeira et al. Based on the protocol delineated on GitHub (https://github.com/hakymul/MetaXcan/wiki/Tutorial-GETEX-v8-Methods-integration-with-a-Coronary-Artery-Disease-GWAS), we first harmonized our results with GTEx v8, imputed missing GWAS summary statistics in sub-samples using current GTEx v8 summary statistics in a region-based approach based on LD blocks identified by Berisa and Pickrell, and finally merged the imputed results before running S-PrediXcan on each of the five tissues.

New versus previously reported genes. We first obtained Bonferroni-significant GWAS associations in each cohort (across all lipid–tissue combinations). We then investigated the GWAS Catalog and GRASP as well as previous literature for TWAS, colocalization analyses, candidate-gene analyses, gene-based aggregate tests and exome-WAS applied to lipid traits and divided the observed signals into ‘new’ and ‘replicated’ categories. We subsequently focused on the ‘new’ category that ‘replicated’ at least twice, defining ‘replication’ as any gene that cleared the Bonferroni threshold and had the same direction of effect for the same lipid–tissue pair in at least two cohorts.

LD-contaminated gene removal. We removed LD-contaminated associations by performing statistical colocalization on the results obtained in the previous step on a gene-by-gene basis. For running colocalization, we first identified a list of GWAS-significant genes and corresponding lipid traits and tissues across the four cohorts. Next, for genes across all combinations of lipid trait, tissue and GWAS cohort, we identified all the SNPs in the GWAS cohort that were within a 1-Mb region of the transcription and transcription start site of the gene. Of these, we considered as ‘lead SNPs’ all SNPs in the GWAS cohort with P < 0.0001 that were 200–400kb apart from each other.

For primary signals, we collected the SNPs overlapping between GWAS and eQTL datasets in a 200-kb window on either side of each lead SNP. We subsequently estimated the colocal probability of H3 (alternative hypothesis...
that eQTL and GWAS associations correspond to independent signals) and H4 (alternative hypothesis that eQTL and GWAS associations correspond to the same signal) for the lead SNP. We assumed a priori probability that a SNP is associated with (1) lipid phenotype $= 1 \times 10^{-10}$, (2) gene expression $= 1 \times 10^{-10}$ and (3) both GWAS and gene expression $= 1 \times 10^{-4}$ for all coloc analyses. For the given gene, we then selected the lead SNP with the highest $P[H4]$ and $P[H3] < 0.5$.

For secondary signals, we used the following GCTA-COJO (v1.26) protocol in the GWAS and eQTL datasets for each lead SNP to identify potential secondary independent associations at a locus. We ran GCTA-COJO to perform association analysis on all SNPs conditioned on the top-associated eQTL ($P < 0.001$) at that locus using the --cojo-cond option. We used 5,000 randomly chosen European American adults from eMERGE as the reference dataset to calculate pairwise LD in eMERGE, GERA and GLGC; we used 5,000 randomly chosen adults of white British ancestry from the UKB as a reference dataset for the UKB. We then used the combined $P$ from GCTA-COJO in coloc to identify potential secondary signals between the lipid trait and gene expression using the same protocol for primary signals as delineated above.

We repeated this protocol for TWAS–significant genes across all combinations of GWAS cohort, lipid trait and tissue and filtered out genes for which all lead SNPs had $P[H3] < 0.5$, which were termed ‘LD-contaminated genes’.

Egger pleiotropy (intercept) $P > 0.05$ to have evidence of minimal heterogeneity. All analyses were conducted using the MendelianRandomization package (v0.4.3) in R packages.

**Data visualization.** A modified version of the Hudson R package was used for comparing association results from eMERGE and the UKB (Fig. 3). A modified version of Synthesis-View (http://visualization.ritchielab.org/synthesis_views/plot/) was used to make Extended Data Figs. 4 and 5. The Venn diagrams (Supplementary Figs. 1 and 2 and Extended Data Figs. 6, 7 and 9) were created by UpSetR (v1.4.0) in R. Pleiotropy between lipids and disease categories was visualized using the circlize package v0.4.12.1004 in R (Fig. 4 and Extended Data Fig. 8). LocusZoom was used to generate regional LD plots (Supplementary Figs. 3–23). The SNP-wise MR plots (Extended Data Fig. 10) were made using the two-sample MR package in MR-base (v0.4.3) in R. The overrepresentation analysis (Supplementary Fig. 25) was conducted on WebGestalt (http://www.webgestalt.org/).

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability.** This project corresponds to UKB application ID 32133 and eMERGE Network phase III (dbGaP accession no. phs001584.v1.p1). Lipid GWAS summary statistics for GLGC 2013 (ref. 2) are publicly available for download (http://csg.sph.umich.edu/willer/public/lipids2013/). Lipid GWAS summary statistics for GERA are available via dbGaP (accession no. phs000674.v2.p2). Expression prediction models with LD reference data using MASHR are available on Zenodo (https://zenodo.org/record/352899/files/mashr eql.tar/download=1). GTEx Analysis Release v8 (dbGaP accession no. phs000424.v8.p2) is available for download via the GTEx Portal (https://gtxportal.org/home/datasets/). Summary statistics for lipid GWAS, lipid TWAS, lipid-guided PhEWA and Xpress-PhEWA generated in this study are available on Figshare (https://figshare.com/s/d3b91e1bb6c64c5b2c1a).

**Code availability.** Code for identifying LD-contaminated genes and detecting secondary independent associations at a locus is shared on GitHub (https://github.com/RitchieLab/Gene-level-statistical-colocalization/).

**References.**

58. Verma, S. S. et al. Imputation and quality-control steps for combining multiple genome-wide datasets. *Front. Genet.* 5, 370 (2014).
59. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
60. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295 (2015).
61. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *Hum. Mol. Genet.* 27, 3641–3649 (2018).
62. Bertrais, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* 32, 283–285 (2015).
63. Macarthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D986–D981 (2017).
64. Eicher, J. D. et al. GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and Phenotypes. *Nucleic Acids Res.* 43, 799–804 (2014).
65. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50, 693–698 (2018).
66. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* 46, 1734–1739 (2017).
67. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenotype. *eLife* 7, e34408 (2018).
68. anaustasia-lucas/hudson. A Hudson Plot Package version 0.1.0. *GitHub*. https://rdrr.io/github/anaustasia-lucas/hudson/. Accessed 5 March 2020.
69. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940 (2017).
70. Zuguang Gu. Circlize R package. https://cran.r-project.org/web/packages/circlize/index.html (2019).
71. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337 (2010).
Acknowledgements

Phase III of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG8657 (Group Health Cooperative/University of Washington); U01HG8685 (Brigham and Women’s Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children’s Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children’s Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine). For the UKB, all data for this cohort pertained to project 32133: ‘Integration of multi-organ imaging phenotypes, clinical phenotypes and genomic data’. Y.V., R.M.K., T.J.H., N.R., M.W.M., E.T. and M.D.R. acknowledge funding from the National Institutes of Health (NIH GM115318: Pharmacogenomics of Statin Therapy (POST)). Y.V. and M.D.R. also acknowledge NIH AI077505 (Pharmacogenomics of HIV Therapy). J.E.M. acknowledges NHGRI T32HG009495–01; C.M.S. acknowledges R35GM131770 (Pharmacogenetics to improve Drug Therapy). B.F.V. acknowledges NIH DK101478, NIH HG010067 and a Linda Pechenik Montague Investigator Award for their time on this project.

Author contributions

Y.V. and M.D.R. conceptualized and designed the study. Y.V. conducted all statistical analyses. Y.V. and D.H. conducted phase III analyses. Y.V., A.L. and S.D. performed data visualization. Y.V., Y.B. and A.L. conducted phenotype curation. Y.V., M.D.R. and A.V. performed data acquisition for the UKB. H.H., P.S., I.K., D.S., C.M.S., D.R.V.E., Q.F. and W.-Q.W. performed data acquisition for eMERGE. T.J.H., N.R., R.M.K., M.W.M. and E.T. performed data acquisition for GERA. Y.V. and B.F.V. conceptualized phase III of this study. Y.V. and J.E.M. performed overrepresentation analysis. D.J.R. provided guidance for phases I and II. Y.V. and M.D.R. wrote the manuscript. All authors provided interpretation of the results and critical feedback on the manuscript.

Competing interests

M.D.R. is on the scientific advisory board for Goldfinch Bio and Cipherome. D.J.R. serves on Scientific Advisory Boards for Alnylam, Novartis, Pfizer and Verve and is a founder of Staten Biotechnology. The other co-authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00879-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00879-y.

Correspondence and requests for materials should be addressed to M.D.R.

Peer review information Nature Genetics thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Case-control distribution for ICD codes. Distribution of cases (blue) and controls (yellow) for the collapsed 3-digit ICD codes in eMERGE (top) and UKB (bottom). eMERGE has predominantly ICD-9 codes whereas UKB has predominantly ICD-10 codes.
Extended Data Fig. 2 | Lipid GWAS in eMERGE. Manhattan plots from GWAS (two-sided linear regression) conducted on the four plasma lipid traits (HDL-C, LDL-C, TC, TG) for the eMERGE cohort. In each plot we have chromosomes 1 to 22 on the x-axis and -log(P) value on the y-axis.
Extended Data Fig. 3 | Lipid GWAS in UKB. Manhattan plots from GWAS (two-sided linear regression) conducted on the four plasma lipid traits (HDL-C, LDL-C, TC, TG) for the UKB cohort. In each plot we have chromosomes 1 to 22 on the x-axis and -log(P) value on the y-axis.
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Lipid TWAS P-values for novel lipid genes. Synthesis-view plot indicating $-\log_{10} P$-values for Bonferroni-significant ‘novel’ genes (two-sided gene-based tests: $P < 5.57 \times 10^{-7}$) from lipid TWAS. These genes passed coloc $P[H3] < 0.5$ filter in at least one cohort. The direction of triangle corresponds to the direction of gene-effect from TWAS (left facing-negative and right facing-positive). Colors indicate the five selected tissues from GTEx v8 (adipose subcutaneous, adipose visceral omentum, liver, small intestine terminal ileum, whole blood).
Extended Data Fig. 5 | See next page for caption.
**Extended Data Fig. 5 | Colocalization probabilities of shared causal variant between lipids and gene expression for novel lipid genes.** Synthesis-view plot indicating coloc $P[H4]$ for Bonferroni-significant ‘novel’ genes (two-sided gene-based tests: $P < 5.57 \times 10^{-7}$) obtained from lipid TWAS. These genes passed coloc $P[H3] < 0.5$ filter in at least one cohort. The direction of triangle corresponds to the direction of gene-effect from TWAS (left facing-negative and right facing-positive). Colors indicate the five selected tissues from GTEx v8 (adipose subcutaneous, adipose visceral omentum, liver, small intestine terminal ileum, whole blood). We present coloc results for all regions corresponding to a gene.
Extended Data Fig. 6 | Overlap of detected ICD codes between cohorts. UpSet plot indicating overlap of diseases (ICD codes) with Bonferroni-significant genes between PheWAS and Xpress-PheWAS conducted on eMERGE and UKB, respectively.
Extended Data Fig. 7 | Overlap of significant SNPs between lipid GWAS and lipid-guided PheWAS across cohorts. UpSet plot indicating overlap of GWAS-significant SNPs (Bonferroni threshold) between each of the four plasma lipids (HDL-C, LDL-C, TC, TG) aggregated across the four considered cohorts (eMERGE, GERA, GLGC, UKB) and lipid-guided PheWAS conducted in eMERGE and UKB, respectively.
Extended Data Fig. 8 | Lipid-disease pleiotropy from lipid-guided PheWAS in either eMERGE or UKB. Circos plot indicates Bonferroni-significant SNPs in either cohort (eMERGE or UKB) from lipid-guided PheWAS (two-sided logistic regression). Outer track, the number of SNPs detected in either cohort; inner track, significant ICD codes per disease category. Links, SNPs connecting lipids (in salmon) to diseases (in blue); link thickness, # SNPs; link color, chromosome. Due to large number of SNP associations involved, this plot does not show associations (links) in the HLA region (chromosome 6).
**Extended Data Fig. 9** | Overlap of significant genes between lipid TWAS and Xpress-PheWAS across cohorts. UpSet plot indicating overlap of detected Bonferroni-significant genes between lipid TWAS and Xpress-PheWAS conducted on eMERGE and UKB, respectively. Lipid TWAS genes have been split into two categories: (1) novel; (2) previously reported.
Extended Data Fig. 10 | Effect sizes and confidence intervals from two-sample univariable Mendelian randomization analyses. Mendelian randomization funnel plots depicting MR effect size (using two-sided IVW and Egger approaches) across ICD codes detected as FDR significant (excluding proof-of-concept diseases such as E78 Disorders of lipoprotein metabolism and other lipidemias and I10 Essential primary hypertension; see Fig. 7 for a full list of FDR-significant diseases). Top 5 plots: exposure dataset (lipid), GERA; outcome dataset, UKB. Remaining plots: exposure dataset (lipid), UKB; outcome dataset, eMERGE.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- [ ] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- [ ] A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [ ] The statistical test(s) used AND whether they are one- or two-sided
  *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- [ ] A description of all covariates tested
- [ ] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- [ ] A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- [ ] For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  *Give P values as exact values whenever suitable.*
- [ ] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- [ ] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- [ ] Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used in this manuscript during data collection stage. We used PLINK 1.9 for quality control.

Data analysis

PLINK 1.9 and 2.0, S-PrediXcan, GCTA-COJO v1.26, Coloc with our own Ii-contamination code in GitHub https://github.com/RitchieLab/Gene-level-statistical-colocalization, PLINK2, MR-base v4.0.3, MR-PRESSO https://github.com/rondolab/MR-PRESSO, MendelianRandomization 0.50 in R, Hudson https://github.com/anastasia-lucas/hudson in R, synthesis-view http://visualization.ritchielab.org/synthesis_views/plot, UpSetR v1.4.0, circize v0.12.1004. In R, locuszoom http://locuszoom.org/genform.php?type=yourdata, webgestalt http://www.webgestalt.org/

For manuscripts utilizing custom algorithms or software that are not central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

This project corresponds to UK Biobank application ID 32133 and eMERGE Network Phase III [dbGaP study accession number phs001584.v1.p1]. Lipid GWAS summary statistics for GLGC 2013 are publicly available for download http://csg.sph.umich.edu/willer/public/lipids2013/. Lipid GWAS summary statistics for GERA are available via dbGaP (accession number phs000674.v2.p2). Expression prediction models with LD reference data using MASHR are available on Zenodo https://zenodo.org/record/3518299/files/mashr_eqtl.tar?download=1. GTEx Analysis Release v8 [dbGaP Accession phs000424.v8.p2] is available for download via the GTEx...
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/rr-reporting-summary-list.pdf](http://nature.com/documents/rr-reporting-summary-list.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | We focused on individuals of European ancestry in this study. Out of 99,185 genotyped samples in the eMERGE consortium, we used best-practice QC pipelines (Verma et al., Front Genet 5, 370 [2014]) after which we were left with 47,229 unrelated European American individuals. For lipid GWAS and TWAS (Phase I), the sample sizes were n=31,565 (14,777 males and 16,790 females) for HDL-C; n=30,509 (14,374 males and 16,135 females) for LDL-C; n=31,575 (14,747 males and 16,828 females) for TC and n=31,074 (14,638 males and 16,436 females) for TG. For lipid-guided PhEWS (Phases II and III), we further removed Columbia University Health Sciences and Mt Sinai due to their low number of case-counts, leaving us with a sample size of 41,981 unrelated individuals of European ancestry (19,556 males and 22,425 females). For UKB, we excluded individuals with poor genotyping quality (Brycroft et al., Nature 562, 203-209 [2018]) and those with mismatches between self-reported and genetically inferred sex, as well as one of a pair of related individuals with pi-hat>0.25. After QC, there were 377,921 individuals of white British ancestry (203,087 females and 174,384 males). This sample size was the same for Phases I, II and III. We have described these procedures in detail in the Methods section. For Phase I, in particular the following were the sample sizes: n=329,480 (153,509 males and 175,971 females) for HDL-C; n=358,482 (165,546 males and 192,936 females) for LDL-C; n=359,096 (165,876 males and 193,220 females) for TC and n=357,709 (164,757 males and 192,912 females) for TG. |
| Data exclusions | We excluded samples that failed genotype or phenotype quality control. To maintain genetic homogeneity, we also focused on samples within the largest ancestral group: i.e. individuals of European ancestry. We have explained above the procedures used for excluding samples that did not meet QC criteria; these have also been explained in the Methods section. |
| Replication | For Phase I, we repeated the analyses on lipid summary statistics from 4 independent datasets; eMERGE, GERA, GLGC and UKB. We used publicly available summary statistics for GERA and GLGC and generated our own for eMERGE and UKB (available on figshare). For Phases II and III, we repeated analyses in two independent datasets, eMERGE and UKB. For Phase I, we present results (genes) that replicate at the Bonferroni threshold in at least two cohorts with the same direction of effect in the same lipid-tissue combination as well as meet the set colocalization criteria (P<0.05). For Phases II and III, we present Bonferroni-significant results that meet our colocalization filter in either dataset. |
| Randomization | The randomization aspect is irrelevant for GERA and GLGC since we only had access to summary statistics. For eMERGE and UKB, we ran Phase I analyses after controlling for age, age-squared, sex, statin medication and first n principal components (n=6 for eMERGE and n=20 for UKB). |
| Blinding | All phenotypes are based on electronic health records and so the investigators were blind with respect to phenotype assignments. All data are de-identified to the research team, as well. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|--------------------------------|---------|
| n/a | Involved in the study |
| ☒ | Antibodies |
| ☒ | Eukaryotic cell lines |
| ☒ | Palaeontology and archaeology |
| ☒ | Animals and other organisms |
| ☒ | Human research participants |
| ☒ | Clinical data |
| ☒ | Dual use research of concern |
| | ChIP seq |
| | Flow cytometry |
| | MRI-based neuroimaging |
### Policy information about studies involving human research participants

**Population characteristics**
We studied common variants from the imputed genotype data from both eMERGE and UK Biobank datasets. For eMERGE, we studied individuals with age larger or equal to 25, and about 50% female and 50% male. The age of the samples in the UK Biobank is larger than 40 and also have about 50% female and 50% male. We started with transcriptome-wide analysis of lipid traits and based on those results and replication with the other lipid TWAS studies, we followed up with Phenome-wide association analyses in the eMERGE and UK Biobank datasets.

**Recruitment**
The participants were recruited by eMERGE consortium and the UK Biobank. All datasets where summary statistics were included were recruited by the individual sites that published the lipid GWAS analyses.

**Ethics oversight**
eMERGE network and UK Biobank provided oversight for their individual studies.

Note that full information on the approval of the study protocol must also be provided in the manuscript.