Novel expression-based Phenome-wide association study (exPheWAS) identifies links between lipid-associated genes and multiple disease outcomes in UK Biobank and eMERGE-III

Yogasudha Veturi1, Anastasia Lucas1, Scott Dudek1, Anurag Verma1, Thomas Hoffmann2, Elizabeth Theusch3, Yuki Bradford4, Hakon Hakonarson4, Iftekhar Kullo5, Jason E Miller1, Patrick Sleiman6, Daniel Schaid6, QiPing Feng7, Charles M. Stein8, Wei-Qi Wei9, Digna R. Velez Edwards8,9,10, Marisa Medina1, Ronald M. Krauss3, Neil Risch2, Marylyn D. Ritchie1

1Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
2Epidemiology & Biostatistics, University of California and San Francisco, San Francisco, CA, 94117, USA
3Children's Hospital Oakland Research Institute, Oakland, CA, 94609, USA
4Center for Applied Genomics, Children's Hospital of Philadelphia, PA 19104, USA
5Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN 55905, USA
6Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA
7Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37230, USA
8Department of Biomedical Informatics in School of Medicine, Vanderbilt University, Nashville, TN 37230, USA
9Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37230, USA
10Division of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN 37230, USA

Introduction: Plasma lipids, i.e. high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), and triglycerides (TG) are known heritable factors associated with risk for atherosclerotic cardiovascular disease. Previous meta-analyses and studies based on electronic health records (EHR) and biobanks have identified hundreds of loci associated with plasma concentrations of lipids using genome-wide association studies (GWAS) and transcriptome-wide association studies (TWAS). Increasing evidence also supports shared genetics between cardiovascular disease and other disease categories such as Alzheimer’s disease, multiple sclerosis, psoriasis, gout, and arthritis. In this study, we devised (1) an integrative framework that combines TWAS with probabilistic fine mapping across multiple large-sized cohorts to identify novel, replicating lipid-associated genes with evidence for causality and (2) an innovative expression-based phenome-wide association study (exPheWAS) approach conducted on lipid-associated genes to identify pleiotropic signatures among phenome-wide disease outcomes.

Methods: We first conducted GWAS for each of the four lipid traits using individual-level data from 31,838 European American (EA) adults with 7,666,366 SNPs from electronic Medical Records and Genomics (eMERGE) network Phase III (covariates: age, sex, batch, and the first 6 ancestry-derived PCs) and 377,921 EA adults with 8,284,911 SNPs from the UK Biobank (UKBB) cohort (covariates: age, sex, batch, and the first 20 ancestry-derived PCs). We also used summary-level data from 76,627 Non-Hispanic EA adults with 11,196,892 SNPs from Genetic Epidemiology Resource on Adult Health and Aging (GERA) and 188,578 predominantly EA individuals with 2,447,442 SNPs from Global Lipids Genetics Consortium (GLGC) 2013. We next performed TWAS using S-PrediXcan on summary-statistics obtained from each of the four cohorts. We used tissue-specific weights (~8300 genes) derived from Genotype Tissue Expression (GTEx) Project v7 for adipose subcutaneous, adipose visceral omentum, liver, small-intestine terminal ileum, and whole blood. We then conducted statistical colocalization analyses in each of five tissues using the coloc R package to filter out LD-contaminated genes and subsequently performed probabilistic fine mapping of causal gene sets using the software FOCUS to fine-map associated regions that have strong evidence for causality. We also estimated local SNP heritability for all four traits using HESS. We selected
Bonferroni-significant genes that (a) had FDR-adjusted prediction accuracy p-value<0.05 and (b) replicated for the same trait-tissue pair in at least two cohorts. We then classified the selected genes as either novel or known lipid-associated genes. Subsequently, we identified all variants in a 1 Mb window upstream and downstream from the transcription start and end sites of each selected gene that also overlapped with GTEx v7 prediction model databases across all 48 tissues. There were 66,871 such SNPs from eMERGE-III and 62,813 SNPs from UKBB. We then collapsed the ICD-10 and ICD-9 codes in UKBB and eMERGE-III respectively, to 3-digit categories. After removing injury, poisoning, accident and other non-genetic codes, there remained 646 ICD-10 categories in UKBB and 414 ICD-9 categories in eMERGE-III with at least 200 cases per category. Next, we ran PheWAS (logistic regression) using PLINK2 on (1) 42,930 EA adults and 66,871 SNPs from eMERGE-III (covariates: age, sex, site) and (2) 377,921 EA adults and 62,813 SNPs from UKBB (covariates: age, sex, batch, and the first 10 ancestry-derived PCs). Finally, we performed exPheWAS using S-PrediXcan in eMERGE-III and UK Biobank across all 48 GTEx v7 tissues and examined the extent of replication among the disease categories across both cohorts.

**Results:** We identified 374 Bonferroni-significant genes from *S-PrediXcan* with FDR-significant gene expression prediction accuracy, of which 87 were novel and 287 were known lipid-associated genes. All novel lipid genes replicated, i.e. had the same direction of effect for the same trait-tissue pair in at least two cohorts: 48 of these replicated in two cohorts, 35 replicated in three cohorts while 4 replicated in all four cohorts. 25 novel lipid genes had *FOCUS* posterior inclusion probabilities > 0.5 in at least two cohorts for the same trait-tissue pair. These included genes that replicated in *all four cohorts* for the same trait and tissue (*ZSWIM1* for TG in adipose subcutaneous tissue and *MT3* for HDL in small intestine) as well as genes that replicated in two to three cohorts for the same trait in *all five tissues* (*RP11-109L13.1* and *CYP21A1P* for TG). We used the 374 lipid-associated genes as gene regions for the exPheWAS. 116 of these genes (including 16 HLA genes) associated with 45 ICD-10 categories in UKBB and 11 ICD-9 categories in eMERGE-III at the Bonferroni-significance threshold. These include circulatory-system disorders (angina pectoris, hypertension, phlebitis and thrombophlebitis), endocrine and metabolic disorders (diabetes mellitus, hypothyroidism and hyperthyroidism), digestive disorders (cholelithiasis, ulcerative colitis, diverticulosis), respiratory disorders (asthma), musculoskeletal disorders (ankylosing spondylitis, gout, rheumatoid arthritis, lupus erythematosis), retinal disorders, and even neoplasms (non-Hodgkin’s lymphoma), among others. 30 of the 116 Bonferroni-significant genes (including 9 HLA genes) associated with the same disease categories in the same tissue in both UKBB and eMERGE-III. An example is a novel lipid-associated pseudogene *STK51P* (chr. 6), which was also associated with hypothyroidism when under-expressed in adrenal gland and psoriatic arthropathy when over-expressed in the hippocampus in both cohorts.

**Discussion:** In this study, we uncovered and replicated novel and known lipid-associated genes and also identified their associations with several diseases pertaining to the nervous system, musculoskeletal system, circulatory system, digestive system and endocrine system, among other organ systems. We were also able to elucidate the interplay between tissue, gene-expression, and trait. We demonstrate the performance of the exPheWAS framework in identifying novel lipid-associated genes that have strong evidence for causality as well as in visualizing the landscape of their associations with disease outcomes from across the phenome. This framework can likewise be extended to other traits or clinical laboratory measures of interest and can not only deepen our insights into the underlying disease etiology but also help with drug repurposing in future research. Finally, this exPheWAS can also be scaled to perform the expression-based PheWAS across the entire transcriptome and phenome for a comprehensive analysis of potential genome-phenome-transcriptome relationships.

---

859