Genetically regulated gene expression underlies lipid traits in Hispanic cohorts

Angela Andaleon [1, 2], Lauren S. Mogil [1], Heather E. Wheeler [1, 2, 3, 4*]

1 Department of Biology, Loyola University Chicago, Chicago, IL, USA

2 Program in Bioinformatics, Loyola University Chicago, Chicago, IL, USA

3 Department of Computer Science, Loyola University Chicago, Chicago, IL, USA

4 Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, IL, USA

*** hwheeler1@luc.edu**

Abstract

Plasma lipid levels are risk factors for cardiovascular disease, a leading cause of death worldwide. While many studies have been conducted on lipid genetics, they mainly comprise individuals of European ancestry and thus their transferability to diverse populations is unclear. We performed genome-wide (GWAS) and imputed transcriptome-wide association studies of four lipid traits in the Hispanic Community Health Study/Study of Latinos cohort (n = 11,103), tested the findings for replication in the Hispanic population in the Multi-Ethnic Study of Atherosclerosis (MESA HIS, n = 1,297), and compared the results to the larger, predominantly European ancestry meta-analysis by the Global Lipids Genetic Consortium (GLGC, n = 196,475). GWAS revealed both known and potentially novel SNPs. In the imputed transcriptome-wide association study in multiple tissues and ethnicities, we found 101 significant gene-phenotype associations ($P < 9.65\times10^{-6}$) with 36 unique significant genes, many of which occurred across multiple phenotypes, tissues, and multi-ethnic populations. We found 21 of the 36 significant genes are novel for any lipid association. Of these, 15/21 replicated in MESA HIS, 14/21 replicated in GLGC, and 10/21 associations replicated in both cohorts ($P < 0.05$). We identified genes that associate independently from nearby genes and colocalize with expression quantitative trait loci (eQTLs), indicating a possible mechanism of gene regulation in lipid level variation. We also investigated prediction in multi-ethnic versus European-only models. To fully characterize the genetic architecture of lipid traits in diverse populations, larger studies in non-European ancestry populations are needed.

Author summary

There is an ongoing disparity in genetic studies between populations of European and non-European descent. Notably, Hispanic populations have been both understudied and suffer greater occurrence of health risks such as obesity and high lipid levels. We studied how genetic variation within Hispanic populations associates with lipid traits. We also investigated how the single variant and predicted gene expression associations with lipid traits may work through mechanisms of gene regulation using reference gene

expression data from multiple tissues as well as multiple diverse populations. We found and replicated gene-trait associations, many in multiple tissues and/or populations, that have not been previously seen in much larger European meta-analyses, and proposed potential mechanisms for some of these genes that are independent from known genes. To fully understand genotype-phenotype associations, we must study not just larger, but, more importantly, more diverse cohorts.

Introduction

Lipid levels are a major risk factor for cardiovascular disease, the leading cause of death in the United States [1]. While lipid levels are known to have a highly heritable component, lipid levels as a complex trait are increasingly concerning due to the growing global rate of obesity caused by rapid urbanization and high-fat foods [2]. Hispanic populations have been especially affected by this shift as Hispanic children and adolescents have the highest rate of obesity among ethnicities in the United States [1, 3]. However, like many other genetic trait studies, large lipid meta-analyses, such as the Global Lipids Genetic Consortium (GLGC), acquire information predominantly from Europeans, and these within-European discoveries may not extrapolate to other populations [4–7].

To increase our understanding of the genetic architecture of lipid traits in diverse populations, we chose to study the Hispanic Community Health Study/Study of Latinos (HCHS/SoL) [8]. Phenotypes under investigation include total cholesterol (CHOL), high density lipoproteins (HDL), triglycerides (TRIG), and low density lipoproteins (LDL). Though this cohort has been previously studied in a GWAS for lipid traits, there were no novel loci found that replicated in independent cohorts [9].

Here, we expand upon traditional GWAS by incorporating eQTL data to predict multi-tissue and multi-ethnic transcriptomes for HCHS/SoL and replication cohorts (MESA HIS and GLGC) to further investigate the biological effects of these variants. The quality control pipeline included structurally-aware relationship inference and principal component calculations [10]. We performed a linear mixed model GWAS in each of the four lipid phenotypes [11] and an imputed transcriptome-based association study in PrediXcan [12] for each phenotype using gene expression prediction models

built with data from the Genotype-Tissue Expression Project (GTEx) [13, 14] and MESA [15]. We performed SNP- and gene- level replication in an independent cohort of MESA HIS individuals [16] and using GWAS summary statistics from GLGC [5, 14]. Within these results, we fine-mapped to independent SNPs and genes [17, 18]. We defined "novel" as not explicitly implicated in the GWAS catalog, PubMed, or in the GLGC 2013 paper [5, 19]. We tested whether SNPs in the novel significant genes colocalized with eQTLs, indicating possible mechanisms of action through gene regulation, and investigated if they were associated independently from known genes [14, 20]. All scripts used for analyses are available at `github.com/WheelerLab/px_his_chol`.

Results

Hispanic populations have diverse genetic ancestry between and within self-identified regions

"Hispanic" is a broad cultural term with no genetic meaning, and many people who identify as Hispanic are admixed from three primary continental ancestries (African, Native American, and European) due, in part, to a previous history of colonization and slavery [21]. Caribbean populations, such as Cubans, Dominicans, and Puerto Ricans, tend to be mainly admixed between African and European ancestries, while mainland populations such as Mexicans, Central Americans, and South Americans, tend to be mainly admixed between Native American and European ancestries, with Caribbean and mainland groups also sometimes differing phenotypically [21, 22] (S1 Fig.). We calculated relatedness and genotypic principal components with the software KING, which is robust to population structure [10]. HCHS/SoL participants included in the analyses each reported a self-identified region of ancestry in the Americas (Fig. 1). Not only is the whole HCHS/SoL cohort genetically differentiated and heterogenous, but there is also great diversity within each self-identified region. Most individuals from the same regions tend to cluster in the same direction until the fifth principal component, but have a wide variation in eigenvalues (Fig. 1).

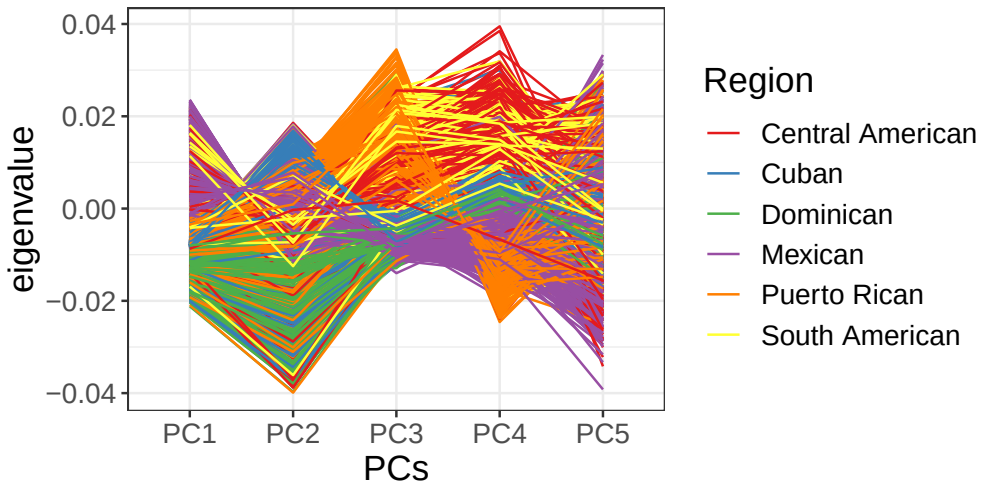

Figure 1. Parallel coordinates of the first 5 genotypic principal components. Hispanic populations are mainly admixed between Native American, West African, and European populations, resulting in a genetically diverse and structured cohort under the umbrella term "Hispanic". Each line represents an individual in HCHS/SoL connected by their eigenvalues and colored by self-identified region. These 5 PCs out of the 20 calculated explain 11.81%, 8.58%, 6.25%, 5.53%, and 5.36% of the variance, respectively.

Joint analysis of GWAS results implicates 13 novel SNPs in lipid traits

We performed GWAS using a linear mixed model implemented in the software GEMMA [11] to investigate individual SNP associations with CHOL, HDL, TRIG, and LDL in HCHS/SoL (n = 11,103). There was no test statistic inflation among the GWAS results (S2 Fig.). We followed the initial GWAS with fine-mapping in GCTA-COJO to find independent signals using linkage disequilibrium calculated with the HCHS/SoL genotypes. GCTA-COJO selects SNPs that are most independent from each other among the loci with multiple SNPs $P < 5\times10^{-8}$, and calculates effect sizes (bJ) and p-values (pJ) using SNPs in high linkage disequilibrium with the selected SNP [17, 18].

In the fine-mapping joint analysis, we found multiple significant, independent SNPs associated with the four phenotypes. These SNPs include 12 with CHOL, 7 with HDL, 0 with TRIG, and 10 with LDL, totalling 29 associations (S1 Table). There were 24 unique independent SNPs across the phenotypes, with some associations replicating in CHOL and an additional lipid trait. 13/24 unique SNPs were novel, defined as not found in the GWAS catalog and having $P > 1\times10^{-5}$ in the GLGC summary statistics (Table 1). 2/13 novel SNPs, rs17041688 and rs117961479, associated with CHOL and had a

MAF < 0.01 in the 1000 Genomes EUR populations, a representative European cohort (Table 1, Figs. 2 and 3) [23–25]. 7/13 novel SNPs are located within introns of genes.

We investigated if SNPs in linkage disequilibrium with the 13 novel SNPs were previously implicated in lipid traits. Of the 2,747 SNPs with $r^2 > 0.6$ in AMR for the SNPs in Table 1, 7 of them were associated with lipid traits in the GWAS Catalog (S2 Table). 2/13 novel SNPs, rs117961479 and rs17766692, are highly correlated with at least one of these 7 lipid associations [26, 27].

Table 1. Novel lipid SNPs from a joint analysis of all GWAS results. In the GWAS of HCHS/SoL, we found 29 independent associations between SNPs and four lipid traits in the joint analysis, including 13 novel SNPs. bJ indicates the effect size and pJ represents the p-value, both from a joint analysis of all the SNPs at the significant loci. The SNPs included in this table have pJ $< 5\times10^{-8}$ and were not previously implicated in the GWAS Catalog or GLGC. Full results for all conditional and joint analyses are in S1 Table.

Chr.	BP	Pheno.	SNP	bJ	pJ	GLGC P	MESA HIS P	MAF	EUR MAF	Intronic gene
2	21199426	CHOL	rs17041688	-0.114	4.8×10^{-9}	0.651	0.205	0.112	0.003	NA
6	108050222	CHOL	rs4945787	-0.098	3.7×10^{-8}	0.223	0.879	0.136	0.188	*SCML4*
7	46171884	CHOL	rs6944635	0.086	1.6×10^{-12}	0.664	0.147	0.431	0.207	NA
7	46171884	LDL	rs6944635	0.082	5.3×10^{-10}	0.689	0.854	0.431	0.207	NA
8	87158577	CHOL	rs11781321	-0.156	2.4×10^{-16}	0.719	0.220	0.160	0.171	*ATP6V0D2*
8	87181395	CHOL	rs144188464	-0.759	1.3×10^{-88}	0.753	0.569	0.392	0.010	NA
8	87204944	CHOL	rs4571715	0.679	2.6×10^{-80}	0.603	0.212	0.423	0.270	NA
14	66769409	HDL	rs6573662	0.188	2.6×10^{-20}	0.806	0.364	0.383	0.118	NA
14	66844743	HDL	rs1457436	-0.506	1.7×10^{-67}	0.693	0.024	0.391	0.101	NA
14	67235722	HDL	rs6573723	0.303	2.2×10^{-34}	0.693	0.500	0.370	0.101	*GPHN*
16	72048632	LDL	rs8062895	-0.081	8.2×10^{-10}	5.8×10^{-5}	0.318	0.442	0.307	*DHODH*
19	11342599	HDL	rs17766692	-0.081	5.8×10^{-12}	8.0×10^{-5}	2.092×10^{-4}	0.189	0.062	*DOCK6*
19	19410750	CHOL	rs117961479	-0.153	1.5×10^{-9}	0.740	0.352	0.060	0.003	*SUGP1*
22	40790565	HDL	rs147122512	-0.052	1.9×10^{-8}	0.328	0.002	0.348	0.157	*SGSM3*

Percent variance explained (PVE) by all genotypes as calculated in GEMMA for each phenotype were similar to those previously observed [28, 29]. These values are CHOL: 0.269 ± 0.029, HDL: 0.180 ± 0.031, TRIG: 0.135 ± 0.030, and LDL: 0.278 ± 0.030.

Imputed transcriptome-based association study implicates 21 novel genes in lipid traits

We performed imputed transcriptome-based association studies to investigate the associations of predicted gene expression with the four lipid traits while also accounting for relatedness and structure in the two genotyped cohorts, HCHS/SoL (n = 11,103) and MESA HIS (n = 1,297) [11, 12, 16]. We used two main sets of prediction models: GTEx V6 and MESA. GTEx predictor models include 44 individual tissue models

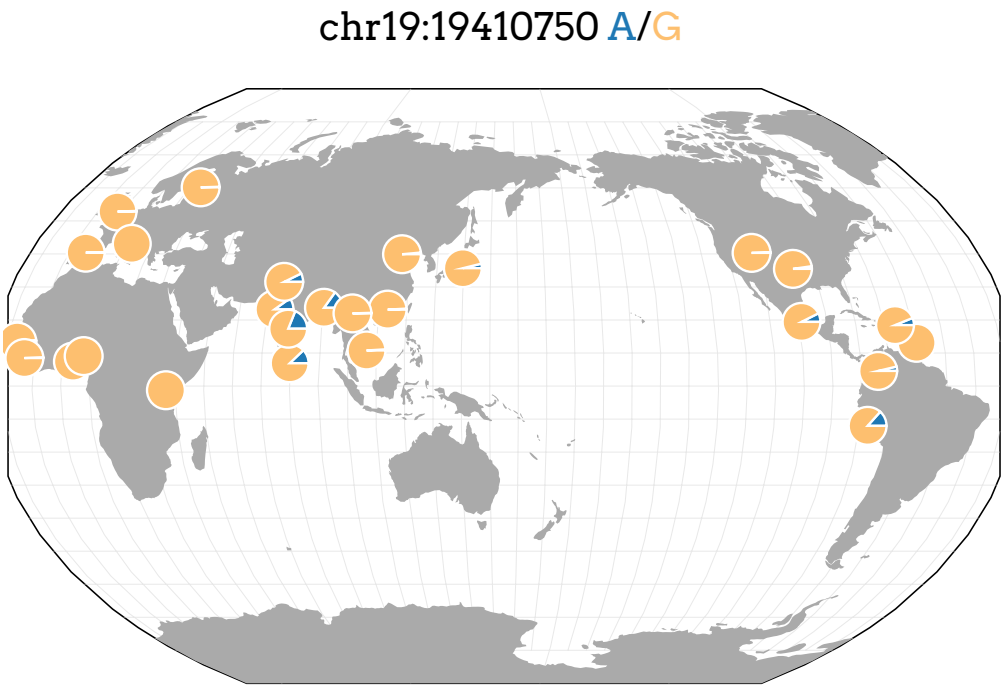

Figure 2. Population allele frequencies of rs117961479. The pie charts represent the frequency of the rs117961479 alleles by 1000 Genomes populations, located where they were collected, with the color corresponding to the nucleotide at the site (blue=A, yellow=G). rs117961479 is significantly associated with CHOL in the GWAS, and it is also a rare variant in Europe, Africa, and East Asia. It is a common variant in HCHS/SoL (MAF = 0.060) and other Hispanic populations in the Americas. This visualization was made using the Geography of Genetic Variants Browser [25].

(predictor population n > 70), and MESA predictor models include monocyte models from 5 multi-ethnic populations (predictor population n > 233) [13–15, 30]. Model abbreviations, keys, and sample sizes are available in S3 Table. We additionally performed replication in MESA HIS and in summary statistics from GLGC (n = 196,475) in all models and traits using S-PrediXcan [5, 14]. We defined discovery significance as $P < 9.65\times10^{-6}$, which is 0.05/(average number of genes in PrediXcan model), and replication significance as $P < 0.05$. There was no test statistic inflation among the PrediXcan results (S3 Fig.).

Across 4 phenotypes, 44 GTEx models, and 5 MESA models, we found 101 significant gene-phenotype associations, including 36 uniquely significant genes. Fifteen of these 36 genes were previously implicated in the GWAS catalog, and many of them replicated in MESA HIS (11/15) and GLGC (13/15). These include well-studied lipid genes such as *APOB*, *PSRC1*, and *CETP* (S4 Table, Fig. 4) [19, 31, 32].

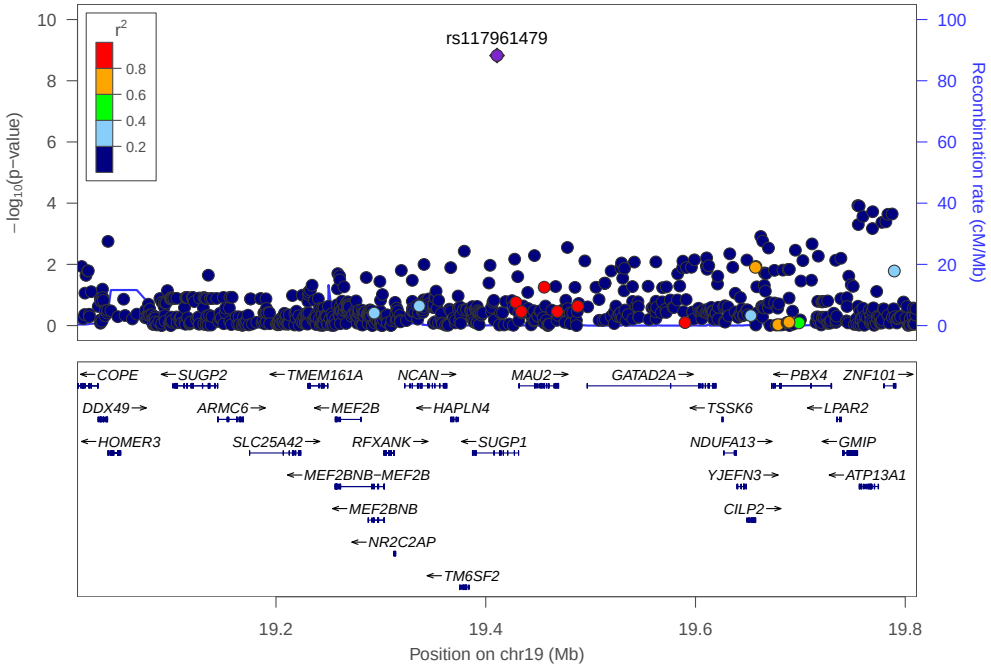

Figure 3. Conditional association of rs117961479 with CHOL in HCHS/SoL. Conditional and joint analysis (COJO) [18] of the HCHS/SoL CHOL GWAS results revealed a genome-wide significant signal for rs117961479 (pJ = 1.5×10^{-9}). Plotted p-values for other SNPs in the region are from a conditional GWAS where the rs117961479 genotype was used as a covariate. The color of each dot represents the SNP's linkage disequilibrium r^2 with the labeled SNP in the 1000 Genomes American populations. Many nearby SNPs within or near *SUGP1* and *MAU2* are in high LD with rs117961479. These conditional results indicate one signal in this region, which was plotted with LocusZoom [24].

We found that 21/36 unique significant genes are novel for any lipid association in the GWAS Catalog across 29 gene-phenotype associations. Most of these genes replicate in CHOL and one of the other three lipid traits. Of these novel genes, 15/21 replicated in MESA HIS, 14/21 replicated in GLGC, and 10/21 genes replicated in both cohorts. Many of the novel genes physically clustered with previously known genes. For example, *ICAM1*, significant in HDL, is located near known lipid genes on chromosome 19 (Fig. 5, Table 2). Complete significant PrediXcan gene association results in HCHS/SoL across all four phenotypes are available in S4 Table.

Stepwise selection reveals independently associated lipid genes

We also performed backward stepwise selection regression to find independent associations for significant PrediXcan genes on a chromosome-phenotype basis in

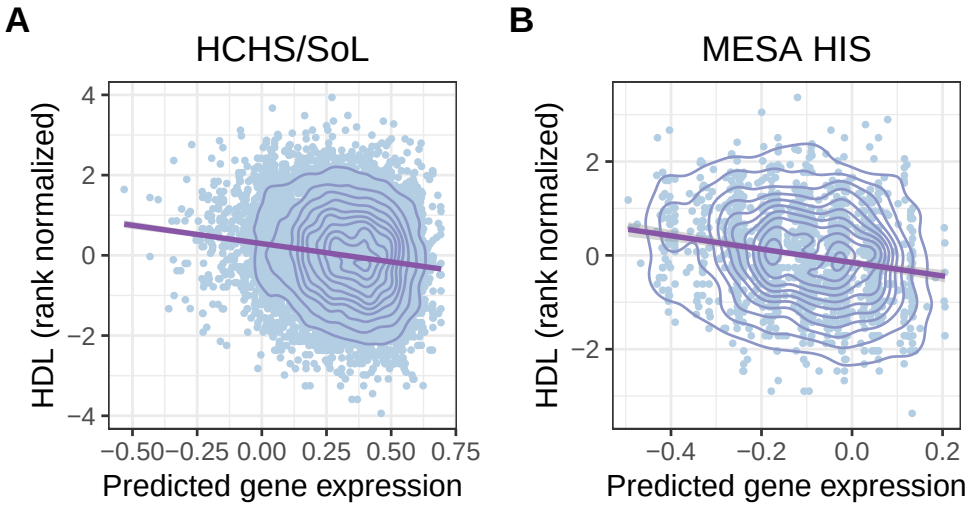

Figure 4. Predicted expression vs. observed phenotype for *CETP*. Using the GTEx Artery-Coronary expression prediction model, increased predicted gene expression of *CETP* is significantly associated with decreased observed HDL in both HCHS/SoL (A) and MESA HIS (B), which is consistent with previous studies of *CETP* [33].

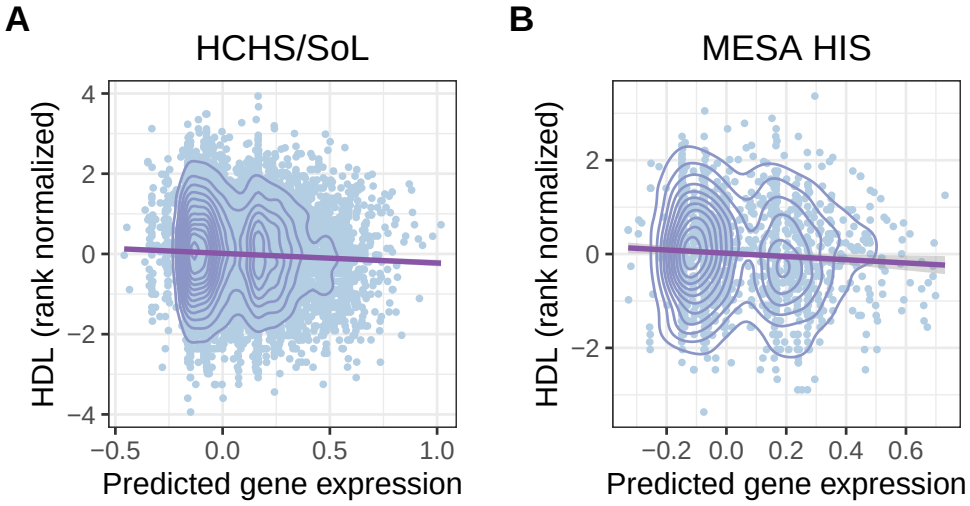

Figure 5. Predicted expression vs. observed phenotype for *ICAM1*. In the GTEx Esophagus Muscularis model, increased predicted gene expression of *ICAM1* is significantly associated with decreased observed HDL in both HCHS/SoL (A) and MESA HIS (B). An observation between expression of *ICAM1* and cholesterol has not been previously observed in humans, but it has been previously observed in mice [34].

Table 2. Novel gene associations with replication in both GLGC and MESA HIS. We performed PrediXcan for four lipid phenotypes in HCHS/SoL using GTEx and MESA monocyte gene expression prediction models [13–15]. The novel genes that replicated in the independent Hispanic (MESA HIS) and majority European (GLGC) cohorts are shown here. In total, there are 10 unique genes across 14 gene-tissue associations. Effect size direction for all genes was the same across all 3 populations. Prediction model abbreviations are defined in S3 Table. Replication results for all gene associations are in S4 Table.

Chr.	Start bp	Pheno.	Gene name	Model	HCHS/SoL P	Beta	GLGC p	MESA HIS p
1	109234945	CHOL	*PRPF38B*	AFA	2.8×10^{-6}	0.140	6.1×10^{-8}	0.009
1	109234945	LDL	*PRPF38B*	AFA	9.9×10^{-7}	0.157	1.3×10^{-10}	0.010
1	109648573	LDL	*C1orf194*	BRNACC	1.1×10^{-6}	0.198	1.0×10^{-67}	0.008
1	110026101	CHOL	*ATXN7L2*	LIVER	1.6×10^{-7}	-0.084	1.4×10^{-36}	0.016
1	110230436	CHOL	*GSTM1*	SNTTRM	1.4×10^{-8}	-4.469	9.5×10^{-143}	0.003
1	110230436	LDL	*GSTM1*	SNTTRM	1.7×10^{-10}	-5.404	2.0×10^{-208}	0.002
11	60618555	LDL	*PTGDR2*	SPLEEN	4.0×10^{-6}	1.229	1.0×10^{-34}	0.031
11	61717293	LDL	*BEST1*	HRTAA	1.6×10^{-6}	-0.391	4.0×10^{-24}	0.020
16	57023397	HDL	*NLRC5*	FIBRBLS	3.1×10^{-18}	-0.107	0	1.2×10^{-4}
16	57023397	HDL	*NLRC5*	HRTAA	1.6×10^{-9}	0.156	0	0.005
16	57392684	HDL	*CCL22*	ESPMSL	1.4×10^{-38}	-9.841	0	5.8×10^{-13}
19	10381511	HDL	*ICAM1*	ESPMSL	1.9×10^{-8}	-0.181	1.0×10^{-10}	0.021
19	11039409	CHOL	*C19orf52*	PNCREAS	1.5×10^{-7}	0.123	1.2×10^{-6}	4.164×10^{-3}
19	11039409	HDL	*C19orf52*	PNCREAS	5.8×10^{-6}	0.080	0.004	9.3×10^{-6}

HCHS/SoL. For each chromosome-phenotype combination, we ran a backward stepwise selection linear regression model that started with a predicted expression term for each significant gene ($P < 9.65\times10^{-6}$). Out of all 101 PrediXcan associations, we found 13/101 remained genome-wide significant ($P < 9.65\times10^{-6}$) after backward stepwise selection, with 7/13 of these comprising novel genes (Table 3). Backward stepwise selection results from all gene associations are in S4 Table.

A majority of gene associations have colocalized GWAS and eQTL signals

We further investigated whether the PrediXcan gene associations colocalized with known eQTLs. Colocalization provides additional evidence that the SNPs in a given expression model are functioning via gene regulation to affect lipid traits. We used the software COLOC, which estimates the colocalization probability for a SNP between an eQTL and a GWAS hit [20]. We have previously applied COLOC to gene-level PrediXcan results and observed clustering of significant genes into non-colocalized, colocalized, or unknown signal [14]. Similarly, we defined significance as $P4 > 0.5$ for each gene, corresponding to shared eQTL and GWAS association of variants within the

Table 3. Gene associations with significant results in gene expression backward stepwise selection analyses. We tested for independence of predicted gene expression signals through backward stepwise selection analysis by chromosome-phenotype groups for HCHS/SoL. P values presented are those from the backward stepwise selection model, and this table only includes $P < 9.65\times10^{-6}$. Full results are in S4 Table.

Novel	Chr.	Start bp	Pheno.	Gene name	Tissue/Pop.	Model P
Y	1	109648573	LDL	*C1orf194*	BRNACC	6.6×10^{-6}
N	1	109852192	CHOL	*SORT1*	LIVER	8.7×10^{-8}
N	2	21224301	CHOL	*APOB*	ADPSBQ	4.4×10^{-8}
N	2	21224301	CHOL	*APOB*	SKINNS	2.3×10^{-6}
Y	3	119385766	HDL	*COX17*	OVARY	6.5×10^{-7}
Y	5	35617946	LDL	*SPEF2*	PRSTTE	3.8×10^{-9}
N	6	32407619	CHOL	*HLA-DRA*	MSCLSK	4.2×10^{-8}
N	15	58702768	HDL	*LIPC*	LIVER	2.2×10^{-12}
N	16	56995762	HDL	*CETP*	ARTCRN	3.7×10^{-7}
Y	16	57392684	HDL	*CCL22*	ESPMSL	1.8×10^{-10}
Y	16	71392616	LDL	*CALB2*	BRNHPP	1.2×10^{-6}
Y	19	11039409	CHOL	*C19orf52*	PNCREAS	1.4×10^{-7}
Y	19	12125547	LDL	*ZNF433*	PRSTTE	1.6×10^{-6}

prediction model. We included P3 estimations, which indicate the probability of independent signals from an eQTL association and a GWAS association, with P3 > 0.5 indicating no colocalization. If neither P3 nor P4 is > 0.5, whether or not the GWAS and eQTL signals are colocalized is unknown [14, 20]. We used eQTL data from GTEx V6 and MESA monocytes, the same populations as the PrediXcan predictors [15, 30].

We calculated colocalization probabilities for all 101 gene-phenotype-tissue associations, and found that 60/101 of the associations were significant for colocalization (P4 > 0.5), including 19 unique genes (S4 Table). For 23/101 associations, there was neither evidence for colocalization nor independent signals (P3 < 0.5 and P4 < 0.5, S4 Table). Of the 29 novel gene-phenotype-tissue associations, 6 had evidence for colocalization and 18 had insufficient evidence for independence (Table 4). These colocalization results between the GWAS results and eQTL summary statistics further support possible biological mechanisms of lipid trait variation acting through gene regulation.

Novel lipid genes implicated by multiple analyses

We investigated which genes from the PrediXcan analyses continued to be implicated in lipid trait variation in multiple downstream analyses. The three downstream procedures we performed include:

Table 4. Novel significant gene associations with high probability of colocalization and/or low probability of independent signals. We tested colocalization of signal between the GWAS results and eQTL data for the novel significant genes in the imputed transcriptome-based association study for HCHS/SoL. A P3 > 0.5 indicates independent and distinct GWAS and eQTL signals. A P4 > 0.5 indicates colocalization of signals for a SNP between a GWAS and an eQTL result, suggesting a biological mechanism through gene regulation. If P3 < 0.5 and P4 < 0.5, the relationship between GWAS results and the eQTL results is ambiguous. Results for all colocalization analyses are in S4 Table.

Chr.	Start BP	Pheno.	Gene name	Model	P3	P4
1	109648573	LDL	*C1orf194*	BRNACC	0.090	0.335
1	110230436	CHOL	*GSTM1*	SNTTRM	0.080	0.339
1	110230436	LDL	*GSTM1*	SNTTRM	0.080	0.339
3	119385766	HDL	*COX17*	OVARY	0.004	0.007
5	35617946	LDL	*SPEF2*	PRSTTE	0.020	0.050
11	60618555	LDL	*PTGDR2*	SPLEEN	0.059	0.450
11	61717293	LDL	*BEST1*	HRTAA	0.018	0.966
16	57023397	HDL	*NLRC5*	HRTAA	0.129	0.060
16	57023397	HDL	*NLRC5*	ALL	0.311	0.658
16	57392684	HDL	*CCL22*	ESPMSL	0.033	0.749
16	57917503	HDL	*CNGB1*	BRNHPP	0.105	0.066
16	71392616	LDL	*CALB2*	BRNHPP	0.032	0.671
19	10381511	HDL	*ICAM1*	ESPMSL	0.022	0.802
19	11039409	CHOL	*C19orf52*	BRNCHA	0.063	0.277
19	11039409	HDL	*C19orf52*	PNCREAS	0.091	0.164
19	11039409	CHOL	*C19orf52*	PNCREAS	0.099	0.100
19	11466062	HDL	*DKFZP761J1410*	PTTARY	0.117	0.174
19	11661962	HDL	*ELOF1*	HRTAA	0.117	0.432
19	11685475	HDL	*ACP5*	CLNSGM	0.071	0.170
19	12125547	LDL	*ZNF433*	PRSTTE	0.062	0.376
19	40325098	HDL	*FBL*	BRNHPT	0.062	0.089
19	40854435	HDL	*PLD3*	CAU	0.026	0.951
19	44645710	CHOL	*ZNF234*	BREAST	0.094	0.250
19	44645710	LDL	*ZNF234*	BREAST	0.094	0.250

- Replication of the HCHS/SoL gene association with the same lipid trait, direction of effect, and expression model in GLGC and MESA HIS ($P < 0.05$)
- Independence of gene signals on the same chromosome in the same phenotype in backward stepwise selection modeling in HCHS/SoL ($P < 9.65 \times 10^{-6}$)
- Colocalization between GWAS signals in HCHC/SoL and eQTL signals in GTEx or MESA (COLOC P4 > 0.5)

Four of 101 associations replicated in all three procedures (Table 5) and 56/101 replicated in at least two procedures (S4 Table). Of these 56, 7/101 reached significance in backward stepwise selection and colocalization, 7/101 reached significance in

backward stepwise selection and independent replication, and 50/101 reached significance in independent replication and colocalization.

Table 5. Gene associations with significant results for replication in independent cohorts, colocalization analyses, and backward stepwise selection modeling. We tested the 101 significant gene associations from the PrediXcan analyses for replication, colocalization, and backward stepwise selection modeling. One gene, *CCL22*, of the four that were significant in all of these procedures is novel for association with any lipid trait.

Novel	Chr.	Start BP	Gene name	Pheno.	Model	P	GLGC P	MESA HIS P	P4	Stepwise P
N	1	109852192	*SORT1*	CHOL	LIVER	5.4×10^{-28}	1.0×10^{-123}	1.3×10^{-9}	0.998	8.7×10^{-8}
N	15	58702768	*LIPC*	HDL	LIVER	2.5×10^{-9}	1.5×10^{-150}	3.3×10^{-4}	0.652	2.2×10^{-12}
N	16	56995762	*CETP*	HDL	ARTCRN	1.2×10^{-27}	0	1.6×10^{-13}	0.818	3.7×10^{-7}
Y	16	57392684	*CCL22*	HDL	ESPMSL	1.4×10^{-38}	0	5.8×10^{-13}	0.749	1.8×10^{-10}

Of the 29 novel gene associations, 14/29 associations replicated in MESA HIS and GLGC, 7/29 were significant ($P < 9.65\times10^{-6}$) in the backward stepwise selection procedure, and 6/29 were significant ($P > 0.5$) in the colocalization procedure. One gene of the 29 novel associations, *CCL22* with the ESPMSL model for HDL, was significant in all three procedures (Table 5). Two of 29 reached significance in backward stepwise selection and colocalization, 3/29 reached significance in backward stepwise selection and independent cohort replication, and 3/29 reached significance in independent cohort replication and colocalization. Six of 29 novel significant gene associations replicated in at least two procedures, including *CCL22*, *C1orf94*, *BEST1*, *CALB2*, *ICAM1*, and *TIMM29* (Table 6).

Table 6. Novel gene associations with significant results with significance in at least two procedures. Out of an original 29 novel gene associations, 6/29 were significant in at least two out of three procedures (replication, colocalization, or backward stepwise selection modeling) with one, *CCL22*, reaching significance in all three.

Chr.	Start BP	Gene name	Pheno.	Model	P	GLGC P	MESA HIS P	P3	P4	Stepwise P
1	109648573	*C1orf194*	LDL	BRNACC	1.1×10^{-6}	1.0×10^{-67}	0.008	0.090	0.335	6.6×10^{-6}
11	61717293	*BEST1*	LDL	HRTAA	1.6×10^{-6}	4.0×10^{-24}	0.02	0.018	0.966	NA
16	57392684	*CCL22*	HDL	ESPMSL	1.4×10^{-38}	0	5.8×10^{-13}	0.033	0.749	1.8×10^{-10}
16	71392616	*CALB2*	LDL	BRNHPP	1.8×10^{-6}	3.0×10^{-14}	0.564	0.032	0.671	1.1×10^{-6}
19	10381511	*ICAM1*	HDL	ESPMSL	1.9×10^{-8}	1.0×10^{-10}	0.021	0.022	0.802	0.145
19	11039409	*TIMM29*	CHOL	PNCREAS	1.5×10^{-7}	1.2×10^{-6}	0.004	0.099	0.100	1.3×10^{-7}

Since *CCL22* shares predictor SNPs and is in close proximity with the well-studied lipid gene *CETP*, we added a *CETP* covariate to the association model comparing *CCL22* and HDL. We used the predicted expression from the most significant prediction models for each gene: ESPMSL for *CCL22* and ESPMCS for *CETP* (S4 Table). After conditioning on predicted expression of *CETP*, *CCL22* is still significant (conditioned P $= 2.2\times10^{-8}$; original P $= 1.4\times10^{-38}$). This indicates that multiple genes in the *CETP*

locus may affect HDL.

Multi-ethnic models with diverse sample populations implicate novel genes

Though most of the results were from the 44 predominantly European GTEx tissue models, we also had significant PrediXcan results using the five multi-ethnic MESA monocyte models [15]. We previously found that eQTL replication is best between populations with ancestry similarity balanced with sample size, and further investigated how this translated to gene expression prediction [15]. Here, we used the multi-ethnic MESA predictors from monocytes in the PrediXcan analyses of lipid traits in HCHS/SoL and the replication cohorts, MESA HIS and GLGC.

Overall, there were 21 gene-population-trait associations in the MESA models, including 8 unique genes such as the well studied lipid genes *CETP* and *PSRC1* (Table 7, Fig. 6). Of these 8 genes found using the MESA models, three have not been previously implicated as lipid genes: *NLRC5*, *PLD3*, and *PRPF38B*. *NLRC5* was significant only in ALL for HDL ($P = 5.0\times10^{-6}$), and it replicated in GLGC ($P = 0$) and had a significant colocalization probability ($P4 = 0.658$). *PLD3* was significant only in CAU for HDL ($P = 8.11\times10^{-6}$) and it replicated in MESA HIS ($P = 0.023$), had a significant colocalization probability ($P4 = 0.951$), and had a nominally significant backward stepwise selection P value ($P = 0.038$). *PRPF38B* was significant only in AFA for CHOL and LDL, where it replicated in both GLGC (CHOL $P = 6.1\times10^{-8}$) and MESA HIS (CHOL $P = 0.009$), and it had a nominally significant backward stepwise selection P (CHOL $P = 0.039$). However, none of these 3 genes pass 2 or more of the previously described significance thresholds (Table 6).

Discussion

Previous genetic lipid studies in HCHS/SoL

HCHS/SoL was previously studied in a GWAS for lipid traits, and there were no novel SNPs found that replicated in independent cohorts [9]. Within the previous study, they discovered eight potentially novel loci, and defined independence as remaining

Table 7. MESA model gene expression association results. With the multi-ethnic MESA monocyte models, we found novel gene associations not found in GTEx models. This table includes genes that were significant in at least one MESA model. Effect sizes in significant (P $< 9.65\times10^{-6}$) models were in the same direction as the ALL effect size shown. Complete results are in S4 Table.

Novel	Chr.	Start BP	Gene name	Pheno.	ALL beta	AFA P	AFHI P	ALL P	CAU P	HIS P	WHLBLD P
Y	1	109234945	*PRPF38B*	CHOL	0.035	2.9×10^{-6}	0.864	0.286	0.192	0.845	NA
Y	1	109234945	*PRPF38B*	LDL	0.044	9.9×10^{-7}	0.887	0.219	0.143	0.933	NA
N	1	109822178	*PSRC1*	CHOL	-0.082	1.2×10^{-11}	6.4×10^{-16}	1.1×10^{-16}	2.0×10^{-16}	2.2×10^{-17}	1.7×10^{-24}
N	1	109822178	*PSRC1*	LDL	-0.099	6.1×10^{-15}	2.2×10^{-19}	1.5×10^{-20}	1.2×10^{-20}	3.4×10^{-21}	1.1×10^{-31}
N	5	74632154	*HMGCR*	LDL	0.382	NA	0.533	0.025	1.7×10^{-6}	0.134	NA
N	11	61535973	*TMEM258*	LDL	-0.155	NA	2.6×10^{-6}	1.7×10^{-4}	1.1×10^{-5}	4.3×10^{-4}	0.019
N	16	56995762	*CETP*	HDL	-0.259	NA	7.3×10^{-30}	1.1×10^{-27}	7.3×10^{-23}	1.8×10^{-31}	NA
Y	16	57023397	*NLRC5*	HDL	-0.190	4.5×10^{-4}	0.288	5.0×10^{-6}	0.101	0.217	0.064
N	19	19756009	*ATP13A1*	CHOL	-0.070	2.6×10^{-5}	5.1×10^{-6}	7.0×10^{-5}	0.041	1.1×10^{-5}	NA
Y	19	40854435	*PLD3*	HDL	0.097	NA	0.236	0.021	8.1×10^{-6}	NA	0.029

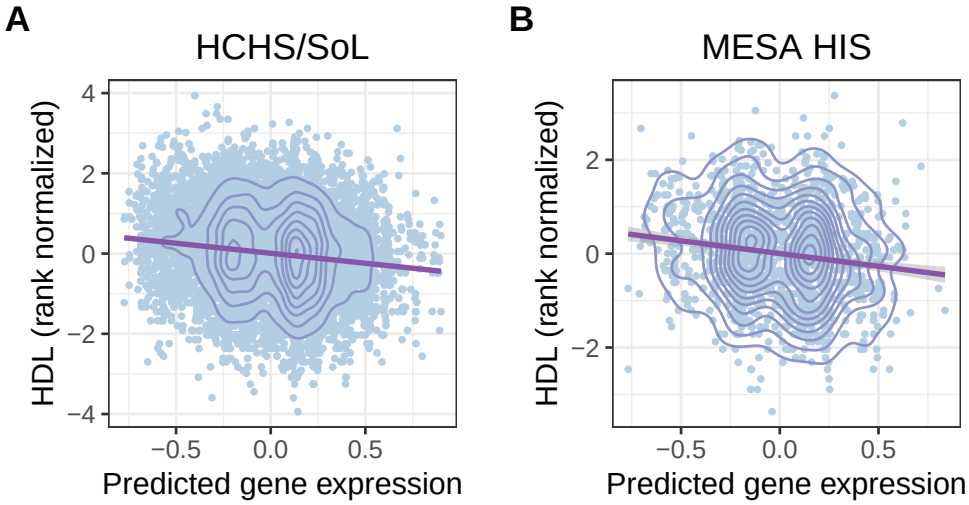

Figure 6. Predicted expression vs. observed phenotype for *CETP*. In the MESA Hispanic model, increased predicted gene expression of *CETP*, a well-studied lipid gene, is significantly associated with decreased observed HDL in both HCHS/SoL (A) and MESA HIS (B).

statistically significant ($P < 5\times10^{-8}$) after conditioning on previously known significant SNPs and the presence of no other known loci within ± 500 kb. Here, we expand upon traditional GWAS, integrating expression quantitative trait loci (eQTL) data from multiple tissues and diverse ethnicities in transcriptome-wide association and colocalization studies to investigate the biological mechanisms underlying lipid trait variation.

Current implications of significant variants

Through conditional and joint analyses, we identified 13 potential novel SNPs associated with lipid traits (Table 1). We found that 2 of these 13 SNPs have linked SNPs ($r^2 > 0.6$) that were previously implicated in lipid traits (S4 Table) [26, 27]. We also found that both the novel SNPs and their linked SNPs are located within both previously implicated and novel lipid genes.

Of the 13 potential novel SNPs, 7 are intronic variants within genes (Table 1). One of these, rs4945787, is an intronic variant within *SCML4*, for which other intronic variants are implicated in coronary artery disease and Alzheimer's disease, and both phenotypes are related to cholesterol levels [35, 36]. rs11781321 is an intronic variant within *ATP6V0D2*, and knockout *Ldlr* mice with lower LDL levels have been observed

to have lower *Atp6v0d2* levels in osteoclast lineage cells [37]. rs6573723 is an intronic variant of *GPHN*, and between high-fat and low-fat fed mice, expression of *GPHN* in liver was nominally significant (P = 0.028) [38]. rs8062895 is an intronic variant of *DHODH*, which has been previously implicated in cholesterol traits [39]. rs17766692 is an intronic variant of *DOCK6* and rs117961479 is an intronic variant of *SUGP1*, which are both well-studied lipid genes [26, 40]. rs147122512 is an intronic variant of *SGSM3*, which is up-regulated in myocardial infarction mice hearts versus control hearts, another phenotype related to cardiovascular disease and cholesterol levels [41]. Of these 7 genes, 3 of them, *DHODH*, *DOCK6*, and *SUGP1*, are previously implicated in lipid levels and the other 4, *SCML4*, *ATP6V0D2*, *GPHN*, and *SGSM3*, are implicated in related phenotypes in humans and/or similar phenotypes in mice.

Current implications of significant genes

In the analyses there were 6 novel significant gene associations with lipid traits that were significant in at least two of the following procedures: replication in two independent cohorts; significance in backward stepwise selection modeling of the gene expression signal; and colocalization of GWAS signal with eQTL signal. These genes include *C1orf194*, *BEST1*, *CCL22*, *CALB2*, *ICAM1*, and *TIMM29* (Table 6).

CCL22 is involved in immunoregulatory and inflammatory processes, and variants within it have been implicated in multiple sclerosis and lupus [42, 43]. In mice, females heterozygous for a *Ccl22* knockout mutation had higher total cholesterol levels than their control counterparts, which is consistent with the findings of higher predicted *CCL22* expression associating with lower LDL levels. Female heterozygotes also have significantly higher body weight than female controls [44].

C1orf194, a currently uncharacterized protein, is expressed most in GTEx brain, lung, and testis tissues. The closest mouse gene at 84% identity is *1700013F07Rik*, whose protein output is also uncharacterized. Currently, *C1orf194* has only been investigated in its potential role in adult spinal ependymoma, and has not been reported in cardiovascular or lipid literature [45].

BEST1 is implicated in ocular diseases in humans, including macular dystrophy and retinal dystrophy [46]. rs17764935, an eQTL for *BEST1* in adipose tissue, was

previously found nominally significant for TRIG for 250 African Americans, which was also replicated in GLGC, but no current literature has implicated *BEST1* in LDL levels [47].

CALB2 produces a calcium-binding protein that is abundant in auditory neurons and may play a role in some cancers [48]. In the GWAS catalog, its only association is for schizophrenia in Ashkenazi Jews [49]. *CALB2* is also extensively studied in the context of mesothelioma, but it has not been studied in context of lipid phenotypes [50, 51].

ICAM1 is an intercellular adhesion molecule, and a major human rhinovirus receptor [52]. In male rats, a high cholesterol diet was found to significantly influence *ICAM-1* molecule levels in the aorta, signifying a correlation between a cholesterol diet and *ICAM-1* expression, and *ICAM-1* molecule levels were negatively correlated with baseline HDL levels, which is consistent with the findings of predicted gene expression levels for *ICAM1* association with lower HDL levels [34] (Fig. 5).

TIMM29, which until recently was known as *C19orf52*, is well expressed across all GTEx tissues. It is associated with transport in the mitochondrial matrix as a subunit of the *TIM22* complex. Proper formation of this complex is essential for survival, and another subunit of the complex, *AGK* functions in mitochondrial lipid metabolism [53]. Since both *AGK* and *TIMM29* are components of the larger complex, their expression may be correlated.

Gene expression prediction in admixed populations

In many studies participants are asked to self-identify under a given label, such as the four race/ethnicity classifications used in MESA: “White, Caucasian”, “Chinese American”, “Black, African-American”, and “Hispanic”. These self-identifications are often not indicative of genetic ancestry, especially in admixed populations such as African-American and Hispanic due to a complex population history [16]. For example, even within self-identified regions in Latin America, populations can be heavily stratified (S1 Fig., Fig. S1). We found that in some MESA models, associations are observed in more diverse models that are not present within European-only models, such as the association of *NLRC5* with HDL using the ALL model that is not seen in

other four models or the association of *PRPF38B* with CHOL and LDL using the AFA model that is not seen in the other four models (Table 7).

A recent study indicates that incorporating the use of local ancestry can improve eQTL mapping and gene expression prediction [54]. Future studies of lipid traits in larger Hispanic populations may include testing local ancestry portions individually in this population to see if less-studied ancestries, such as Native American and West African, are associated with SNPs and genes that have not been previously observed in European-majority cohorts. By combining information from both local ancestry mapping and eQTL studies in diverse populations, we may better characterize and predict gene expression in admixed populations.

There is a dearth of diversity of genetic studies that greatly impacts the ability to apply the results of genetic studies to non-European populations. With many biobank-sized resources only including European data, this gap continues to grow [6]. Without data collection and proper models for non-European populations, there is less potential for accurate implementation of precision medicine. To fully characterize the impact of genetic variation between and within populations and allow all individuals to benefit from biomedical research, we must expand genetic studies to include individuals of diverse ancestries.

Materials and Methods

Ethics statement

This work was approved by the Loyola University Chicago Institutional Review Board (project number 2014). All data were previous collected and de-identified.

Phenotypic and genomic data

The analyses in HCHS/SoL and MESA HIS were performed with whole genome genotypes from the database of Genotypes and Phenotypes (Table 8) [55]. The HCHS/SoL cohort is composed of self-identified Hispanic adults (18-74) recruited from four urban centers in the US: Chicago, IL; Miami, FL; the Bronx, NY; and San Diego, CA, using a previously described household-based sampling technique [8]. Within the

collected lipid phenotypes, we rank-normalized CHOL, HDL, TRIG, and LDL to correct the skewing in the raw data. The majority of participants reported a self-identified region from Cuba (n = 1,987), Dominican Republic (n = 1,132), Mexico (n = 4,056), Puerto Rico (n = 1,984), Central America (n = 1,244), or South America (n = 707). Original genotypes were collected from a standard Illumina Omni2.5M array with 109,571 custom SNPs.

Table 8. Genotyped cohorts

	HCHS	MESA HIS
Accession number	phs000810.v1.p1	phs000209.v13.p3
Pre-QC SNPs	2,536,661	909,622
Pre-QC individuals	12,434	2,174
Post-QC SNPs	2,074,058	809,082
Post-QC individuals	12,235	2,067
Post-imputation SNPs	7,576,834	7,201,805
Ind. with lipid pheno.	11,104	1,297

Within the MESA HIS cohort, self-identified Hispanic individuals were recruited from six urban centers in the US: Baltimore, MD; Chicago, IL; Forsyth County, NC; Los Angeles County, CA; northern Manhattan, NY; and St. Paul, MN. Genotype data were collected from blood samples using an Affymetrix 6.0 SNP array and phenotype data from Exam 1 were used. Many MESA HIS individuals also had similar self-identified regions, and the four lipid phenotypes were also rank-normalized [16].

Summary statistics from the Global Lipids Genetic Consortium were downloaded from `csg.sph.umich.edu/abecasis/public/lipids2013/`. This cohort is >95% European and contained 196,475 individuals. The original study had quantile-normalized phenotypes [5].

Quality control

For quality control, we merged the two HCHS/SoL permission groups from dbGaP and ran all processes in PLINK. All of the genotypes are in genome build GRCh37/hg19. We removed individuals with genotyping rate $< 99\%$. Additionally, we removed SNPs with failed heterozygosity (outside $\pm$ 3 standard deviations from the mean), with failed Hardy-Weinberg equilibrium ($P < 1\times10^{-6}$), and restricted analyses to SNPs on autosomes only. Unlike the typical quality control process for non-admixed populations, we did not remove principal component outliers nor did we remove individuals with >

0.125 relatedness as this would have included most of the sample [56, 57].

Imputation, relationship inference, and principal component calculation

For imputation, we used the University of Michigan Imputation Server with EAGLE2 phasing [58]. The reference panel was 1000 Genomes Phase 3, using all populations due to the multi-continental nature of the cohort [23]. We then filtered imputation results to SNPs with $R^2 > 0.8$ and minor allele frequency $> 1\%$ within PLINK. Quality control and imputation for MESA HIS was performed by our group previously with the same parameters [15].

Within this cohort, there are potentially confounding amounts of genetic substructure present [59] (Fig. 1). In structured and admixed cohorts, such as Hispanic and African-American populations, software such as KING are robust to complex population structure for relatedness and principal component calculations. [60]. We calculated pedigrees, then inferred relationships across all individuals in the cohort into a relationship matrix, and also calculated the first 20 principal components using KING, and we used the first 5 principal components in the rest of the analyses as performed previously within this cohort (Fig. 1) [10, 61].

Genome-wide association study

We used the software GEMMA for the GWAS portion of the study. We ran separate univariate linear mixed models for each of the four phenotypes and included the relationship matrix as random effects. For the covariates used as fixed effects within the model, we included self-identified region, use of lipid or other cardiovascular medication, and the first 5 of 20 principal components as previously conducted in this cohort [60]. P values presented were calculated using the Wald test [11]. We restricted the analyses to SNPs in PrediXcan models as this was the main focus of the analyses, analyzing in total 1,770,809 SNPs for each phenotype. We performed multi-SNP-based conditional and joint association analysis using GWAS data in GCTA-COJO, and we included the genotype data to use the linkage disequilibrium calculations in the actual genotypes with a collinearity cutoff of 0.9 and a standard genome-wide significance threshold of

5×10^{-8} [17, 18]. To investigate SNPs in linkage disequilibrium with the novel SNPs, we calculated linkage disequilibrium using rAggr in the 1000 Genomes American population [62].

Imputed transcriptome-based association study

We used the software PrediXcan to produce predicted expression from genotype for both HCHS/SoL and MESA HIS. The models we used include 44 tissue models from the GTEx V6 tissues with at least 70 individuals (85% European, 15% African-American) each, with an average of 5,179 genes predicted in each tissue. These include various brain regions, adipose, and liver, and we use their given GTEx abbreviations throughout this work (S3 Table) [12, 14, 30]. The other set of PrediXcan models we used were made from monocyte gene expression in MESA [15]. Each of these five models had a training set of at least 233 individuals, and the African American and Hispanic cohorts are also admixed. These models were retrieved from `predictdb.org`. Individuals used to train the AFHI, ALL, and HIS models are also included in the MESA HIS lipid trait replication cohort [15]. We retrieved information from the models such as SNPs included, SNP weights, and sample sizes using SQLite3 and Python.

After the 49 predicted expression files were made using PrediXcan, we ran them as a pseudo-genotype in a univariate linear mixed model in GEMMA using the command `-notsnp`. This was performed in GEMMA rather than base PrediXcan as base PrediXcan does not have an option to include a relationship matrix, which is important in a heavily related and structured cohort like HCHS/SoL [11]. We again included the KING relationship matrix as random effects and the first five principal components, use of lipid or other cardiovascular medication, and self-identified region as fixed effects. P values presented were calculated using the Wald test.

Since we did not have the GLGC genotypes, we used the software S-PrediXcan, which is an extension of PrediXcan that takes GWAS summary statistics as input [14]. We ran S-PrediXcan on GLGC with all 44 GTEx tissue models and all 5 MESA multi-ethnic models. We considered significance in the discovery population, HCHS/SoL, as $P < 9.65\times10^{-6}$, 0.05/(average number of genes in PrediXcan model), and in the replication populations, MESA HIS and GLGC, as $P < 0.05$ within the

model. We considered gene-level results novel if the gene had no association with any lipid trait in the GWAS Catalog at `ebi.ac.uk/gwas/` or PubMed in humans [19].

Independent association analysis

We performed backward stepwise selection modeling within R to determine lead gene signal of PrediXcan results. We performed separate analyses corresponding to the four lipid phenotypes and each chromosome to determine lead genes within each phenotype-chromosome combination. Within each phenotype-chromosome combination, we ran a backward stepwise selection linear regression model that started with a predicted expression term for each significant gene ($P < 9.65 \times 10^{-6}$) using `step()` in base R.

Colocalization analysis

We performed a colocalization analysis by applying the software COLOC to the lipid GWAS results and eQTL data from GTEx and MESA to determine whether eQTLs within gene prediction models and GWAS hits were shared [20]. A high P4 probability ($P > 0.5$) indicates likely colocalized signals between an eQTL and a GWAS hit, especially in well-predicted genes with a high R^2 value, while a high P3 probability indicates independent signals between an eQTL and a GWAS hit and a high P0, P1, or P2 indicates an unknown association [14]. Analyses were run using scripts from the S-PrediXcan manuscript at `github.com/hakyimlab/summary-gwas-imputation/wiki/Running-Coloc` [14]. We subset the COLOC analyses to only contain SNPs within the predictor models of the significant genes from the PrediXcan analyses.

Data availability statement

HCHS/SoL and MESA HIS phenotype and genotype data are available at dbGaP (phs000810.v1.p1 and phs000209.v13.p3). PrediXcan gene expression models for GTEx and MESA are available at `predictdb.org`, and eQTL data for GTEx and MESA are available at `gtexportal.org/home/datasets` and `github.com/WheelerLab/DivPop`, respectively. Summary statistics of all significant results in this study are available in

the supplemental information, and all major scripts used in analyses are available at `github.com/WheelerLab/px_his_chol`.

Acknowledgements

This work is supported by the NIH National Human Genome Research Institute Academic Research Enhancement Award R15 HG009569 (HEW), start-up funds from Loyola University Chicago (HEW), the Loyola Carbon Undergraduate Research Fellowship (AA), and the Loyola MS Bioinformatics Research Assistant Fellowship (AA). The Hispanic Community Health Study/Study of Latinos is supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Center on Minority Health and Health Disparities, the National Institute of Deafness and Other Communications Disorders, the National Institute of Dental and Craniofacial Research, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Neurological Disorders and Stroke, and the Office of Dietary Supplements. The authors thank the staff and participants of the HCHS/SOL study for their important contributions. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-RR-025005, and UL1-TR-000040. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at `mesa-nhlbi.org`. The following datasets

were downloaded from dbGaP: phs000810.v1.p1 and phs000209.v13.p3. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank Peter Fiorica, Ryan Schubert, and Paul Okoro for their aid in reviewing the manuscript.

Competing interests

The authors have declared that no competing interests exist.

Author contributions

Conceptualization: Angela Andaleon, Heather E. Wheeler

Data curation: Angela Andaleon, Lauren S. Mogil, Heather E. Wheeler

Formal analysis: Angela Andaleon, Lauren S. Mogil, Heather E. Wheeler

Funding acquisition: Angela Andaleon, Heather E. Wheeler

Investigation: Angela Andaleon, Heather E. Wheeler

Methodology: Angela Andaleon, Heather E. Wheeler

Project administration: Heather E. Wheeler

Resources: Angela Andaleon, Lauren S. Mogil, Heather E. Wheeler

Software: Angela Andaleon, Heather E. Wheeler

Supervision: Heather E. Wheeler

Validation: Angela Andaleon, Heather E. Wheeler

Visualization: Angela Andaleon, Heather E. Wheeler

Writing - original draft: Angela Andaleon, Heather E. Wheeler

Writing - review & editing: Angela Andaleon, Lauren S. Mogil, Heather E. Wheeler

Supporting information

Table S1. HCHS/SoL GCTA-COJO results for independent GWAS loci across all lipid phenotypes.
Table S2. List of SNPs with $r^2 > 0.6$ for the 13 novel GCTA-COJO SNPs in Table 1.
Table S3. Abbreviations for all GTEx tissue and MESA population PrediXcan models used.
Table S4. PrediXcan results for all lipid phenotypes, including replication testing in GLGC and MESA HIS, colocalization probabilities, and backward stepwise selection signifiance.

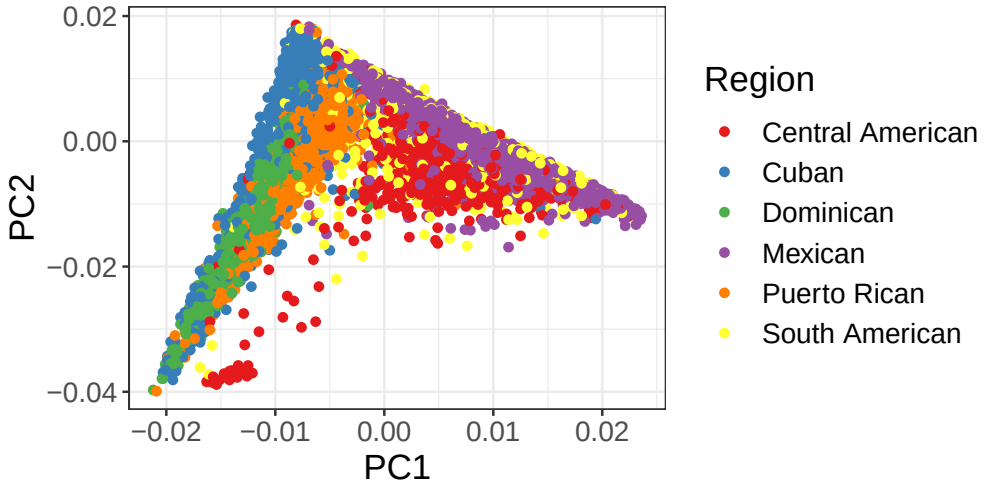

Figure S1. Genotypic principal component analysis in HCHS/SoL with self-identified regions. PC1 vs. PC2 is plotted for each individual colored by their self-identified region. Hispanic populations have diverse continental ancestries due to a previous history of colonization and slavery: African (bottom left), Native American (right), and European (top). Caribbean populations, such as the Cuban, Dominican, and Puerto Rican groups, tend to be mainly admixed between African and European, while mainland populations such as Mexican, Central American, and South American, tend to be mainly admixed between Native American and European.

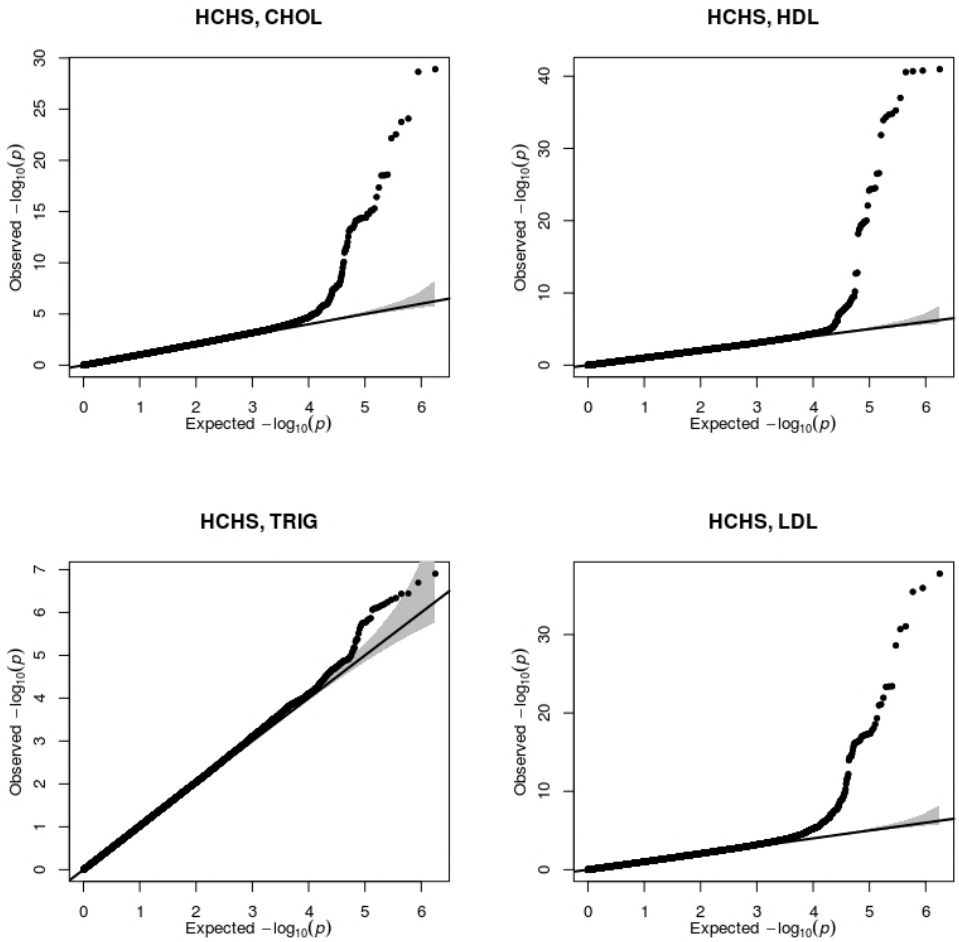

Figure S2. Quantile-quantile plots of the four lipid traits for GWAS results in HCHS/SoL. There is no genome-wide deviation from the significance expectation line in any of the GWAS results.

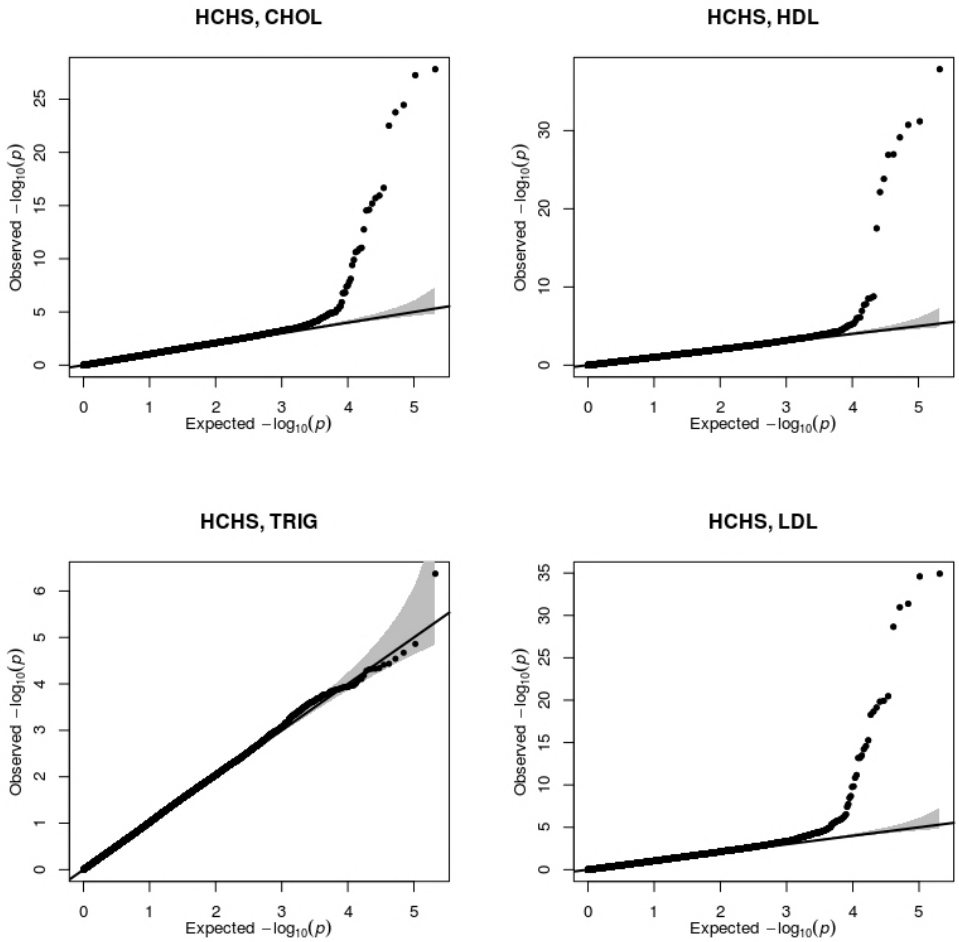

Figure S3. Quantile-quantile plots of the four lipid traits for PrediXcan results in HCHS/SoL. PrediXcan results for all 44 GTEx tissue models and 5 MESA monocyte populations are combined. Each point is a gene-tissue or gene-population association. There is no genome-wide deviation from the significance expectation line in any of the PrediXcan results.

References

1. Writing Group Members, Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, et al. Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. Circulation. 2016;133(4):e38–360. doi:10.1161/CIR.0000000000000350.

2. Wu Y, Waite LL, Jackson AU, Sheu WHH, Buyske S, Absher D, et al. Trans-Ethnic Fine-Mapping of Lipid Loci Identifies Population-Specific Signals and Allelic Heterogeneity That Increases the Trait Variance Explained. PLoS Genetics. 2013;9(3). doi:10.1371/journal.pgen.1003379.

3. Rodriguez CJ, Cai J, Swett K, González HM, Talavera GA, Wruck LM, et al. High Cholesterol Awareness, Treatment, and Control Among Hispanic/Latinos: Results From the Hispanic Community Health Study/Study of Latinos. Journal of the American Heart Association. 2015;4(7):1–10. doi:10.1161/JAHA.115.001867.

4. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161–164. doi:10.1038/538161a.

5. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nature Genetics. 2013;45(11):1274–1285. doi:10.1038/ng.2797.

6. Martin AR, Kanai M, Kamatani Y, Okada Y. Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities. bioRxiv. 2018; p. 1–26.

7. Andaleon A, Mogil LS, Wheeler HE. Gene-based association study for lipid traits in diverse cohorts implicates BACE1 and SIDT2 regulation in triglyceride levels. PeerJ. 2018;6:e4314. doi:10.7717/peerj.4314.

8. LaVange LM, Kalsbeek WD, Sorlie PD, Avilés-Santa LM, Kaplan RC, Barnhart J, et al. Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. Annals of Epidemiology. 2010;20(8):642–649. doi:10.1016/j.annepidem.2010.05.006.

9. Graff M, Emery LS, Justice AE, Parra E, Below JE, Palmer ND, et al. Genetic architecture of lipid traits in the Hispanic community health study/study of

Latinos. Lipids in Health and Disease. 2017;16(1):1–12. doi:10.1186/s12944-017-0591-6.

10. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–2873. doi:10.1093/bioinformatics/btq559.

11. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012;44(7):821–824. doi:10.1038/ng.2310.

12. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015;47(9):1091–1098. doi:10.1038/ng.3367.

13. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Cox NJ, et al. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genetics. 2016;12(11):1–23. doi:10.1371/journal.pgen.1006423.

14. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nature Communications. 2018;9(1):1–20. doi:10.1038/s41467-018-03621-1.

15. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, et al. Genetic architecture of gene expression traits across diverse populations. PLoS Genetics. 2018;14(8). doi:10.1371/journal.pgen.1007586.

16. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, et al. Multi-Ethnic Study of Atherosclerosis: Objectives and Design. American Journal of Epidemiology. 2002;156(9):871–881.

17. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics. 2011;88(1):76–82. doi:10.1016/j.ajhg.2010.11.011.

18. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics

identifies additional variants influencing complex traits. Nature Genetics. 2012;44(4):369–375. doi:10.1038/ng.2213.

19. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research. 2017;45(D1):D896–D901. doi:10.1093/nar/gkw1133.

20. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genetics. 2014;10(5). doi:10.1371/journal.pgen.1004383.

21. Belbin GM, Nieves-Colón MA, Kenny EE, Moreno-Estrada A, Gignoux CR. Genetic diversity in populations across Latin America: implications for population and medical genetic studies. Current Opinion in Genetics and Development. 2018;53:98–104. doi:10.1016/j.gde.2018.07.006.

22. Kramer HJ, Stilp AM, Laurie CC, Reiner AP, Lash J, Daviglus ML, et al. African Ancestry–Specific Alleles and Kidney Disease Risk in Hispanics/Latinos. Journal of the American Society of Nephrology. 2017;28(3):915–922. doi:10.1681/ASN.2016030357.

23. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393.

24. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics. 2011;27(13):2336–2337. doi:10.1093/bioinformatics/btq419.

25. Marcus JH, Novembre J. Visualizing the geography of genetic variants. Bioinformatics. 2017;33(4):594–595. doi:10.1093/bioinformatics/btw643.

26. Spracklen CN, Chen P, Kim YJ, Wang X, Cai H, Li S, et al. Association analyses of East Asian individuals and trans-ancestry analyses with European individuals

reveal new loci associated with cholesterol and triglyceride levels. Human Molecular Genetics. 2017;26(9):1770–1784. doi:10.1093/hmg/ddx062.

27. Below JE, Parra EJ, Gamazon ER, Torres J, Krithika S, Candille S, et al. Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. Scientific Reports. 2016;6(January):1–13. doi:10.1038/srep19429.

28. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature Genetics. 2009;41(1):35–46. doi:10.1038/ng.271.

29. Zhou X. a Unified Framework for Variance Component Estimation With Summary Statistics in Genome-Wide Association Studies. The annals of applied statistics. 2017;11(4):2027–2051. doi:10.1214/17-AOAS1052.

30. Ardlie KG, DeLuca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348(6235):648–660. doi:10.1126/science.1262110.

31. Zhou L, He M, Mo Z, Wu C, Yang H, Yu D, et al. A genome wide association study identifies common variants associated with lipid levels in the Chinese population. PLoS ONE. 2013;8(12):e82420. doi:10.1371/journal.pone.0082420.

32. Thompson JF, Lira ME, Durham LK, Clark RW, Bamberger MJ, Milos PM. Polymorphisms in the CETP gene and association with CETP mass and HDL levels. Atherosclerosis. 2003;167(2):195–204. doi:10.1016/S0021-9150(03)00005-4.

33. Barter PJ, Brewer HB, Chapman MJ, Hennekens CH, Rader DJ, Tall AR. Cholesteryl ester transfer protein: A novel target for raising HDL and inhibiting atherosclerosis. Arteriosclerosis, Thrombosis, and Vascular Biology. 2003;23(2):160–167. doi:10.1161/01.ATV.0000054658.91146.64.

34. Fotis L, Agrogiannis G, Vlachos IS, Pantopoulou A, Margoni A, Kostaki M, et al. Intercellular adhesion molecule (ICAM)-1 and vascular cell adhesion molecule

(VCAM)-1 at the early stages of atherosclerosis in a rat model. In vivo. 2018;26(2):243–50.

35. Herold C, Hooli BV, Mullin K, Liu T, Roehr JT, Mattheisen M, et al. Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. Molecular Psychiatry. 2016;21(11):1608–1612. doi:10.1038/mp.2015.218.

36. Li Y, Wang DW, Chen Y, Chen C, Guo J, Zhang S, et al. Genome-Wide Association and Functional Studies Identify SCML4 and THSD7A as Novel Susceptibility Genes for Coronary Artery Disease. Arteriosclerosis, Thrombosis, and Vascular Biology. 2018;38(4):964–975. doi:10.1161/ATVBAHA.117.310594.

37. Okayasu M, Nakayachi M, Hayashida C, Ito J, Kaneda T, Masuhara M, et al. Low-density lipoprotein receptor deficiency causes impaired osteoclastogenesis and increased bone mass in mice because of defect in osteoclastic cell-cell fusion. Journal of Biological Chemistry. 2012;287(23):19229–19241. doi:10.1074/jbc.M111.323600.

38. Pruis MGM, Lendvai Á, Bloks VW, Zwier MV, Baller JFW, de Bruin A, et al. Maternal western diet primes non-alcoholic fatty liver disease in adult mouse offspring. Acta Physiologica. 2014;210(1):215–227. doi:http://dx.doi.org/10.1016/j.bbrc.2017.07.081.

39. Surakka I, Horikoshi M, Mägi R, Sarin AP, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. Nature Genetics. 2015;47(6):589–597. doi:10.1038/ng.3300.

40. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707–713. doi:10.1038/nature09270.

41. Lee CY, Choi JW, Shin S, Lee J, Seo HH, Lim S, et al. Interaction of small G protein signaling modulator 3 with connexin 43 contributes to myocardial infarction in rat hearts. Biochemical and Biophysical Research Communications. 2017;491(2):429–435. doi:10.1016/j.bbrc.2017.07.081.

42. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nature Genetics. 2013;45(11):1353–1362. doi:10.1038/ng.2770.

43. Langefeld CD, Ainsworth HC, Graham DSC, Kelly JA, Comeau ME, Marion MC, et al. Transancestral mapping and genetic load in systemic lupus erythematosus. Nature Communications. 2017;8(May). doi:10.1038/ncomms16021.

44. Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016;537(7621):508–514. doi:10.1038/nature19356.

45. Olsen TK, Panagopoulos I, Gorunova L, Micci F, Andersen K, Kilen Andersen H, et al. Novel fusion genes and chimeric transcripts in ependymal tumors. Genes, Chromosomes and Cancer. 2016;55(12):944–953. doi:10.1002/gcc.22392.

46. Marmorstein AD, Cross HE, Peachey NS. Progress in Retinal and Eye Research Functional roles of bestrophins in ocular epithelia. Progress in Retinal and Eye Research. 2009;28(3):206–226. doi:10.1016/j.preteyeres.2009.04.004.

47. Sajuthi SP, Sharma NK, Comeau ME, Chou JW, Bowden DW, Freedman BI, et al. Genetic regulation of adipose tissue transcript expression is involved in modulating serum triglyceride and HDL-cholesterol. Gene. 2017;632(August):50–58. doi:10.1016/j.gene.2017.08.019.

48. Parmentier M, Szpirer J, Levan G, Vassart G. The human genes for calbindin 27 and 29 kDa proteins are located on chromosomes 8 and 16, respectively. Cytogenetic and Genome Research. 1989;52(1-2):85–87. doi:10.1159/000132847.

49. Goes FS, Mcgrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics. 2015;168(8):649–659. doi:10.1002/ajmg.b.32349.

50. Kresoja-Rakic J, Sulemani M, Kirschner MB, Ronner M, Reid G, Kao S, et al. Posttranscriptional regulation controls calretinin expression in malignant pleural

mesothelioma. Frontiers in Genetics. 2017;8(May):1–12. doi:10.3389/fgene.2017.00070.

51. Blum W, Pecze L, Rodriguez JW, Steinauer M, Schwaller B. Regulation of calretinin in malignant mesothelioma is mediated by septin 7 binding to the CALB2 promoter. BMC Cancer. 2018;18(1):1–15. doi:10.1186/s12885-018-4385-7.

52. Greve JM, Davis G, Meyer AM, Forte CP, Yost SC, Marlor CW, et al. The major human rhinovirus receptor is ICAM-1. Cell. 1989;56(5):839–847. doi:10.1016/0092-8674(89)90688-0.

53. Pacheu-Grau D, Callegari S, Emperador S, Thompson K, Aich A, Topol SE, et al. Mutations of the mitochondrial carrier translocase channel subunit TIM22 cause early-onset mitochondrial myopathy. Human Molecular Genetics. 2018;00(00):1–10. doi:10.1093/hmg/ddy305.

54. Zhong Y, Perera M, Gamazon ER. On Using Local Ancestry to Characterize the Genetic Architecture of Human Phenotypes : Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations as a Model. bioRxiv. 2018;doi:10.1101/483107.

55. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nature Genetics. 2007;39(10):1181–1186. doi:10.1038/ng1007-1181.

56. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81(3):559–575. doi:10.1086/519795.

57. Turner S, Armstrong LL, Bradford Y, Carlson CS, Dana C, Crenshaw AT, et al. Quality control procedures for genome wide association studies. Current Proceedings in Human Genetics. 2011;68(1):1–24. doi:10.1002/0471142905.hg0119s68.Quality.

58. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016;48(10):1284–1287. doi:10.1038/ng.3656.

59. Daviglus ML, Talavera GA, Avilés-Santa ML, Allison M, Cai J, Criqui MH, et al. Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. JAMA. 2012;308(17):1775–84. doi:10.1001/jama.2012.14517.

60. Conomos MP, Laurie CA, Stilp AM, Gogarten SM, McHugh CP, Nelson SC, et al. Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. American Journal of Human Genetics. 2016;98(1):165–184. doi:10.1016/j.ajhg.2015.12.001.

61. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genetic Epidemiology. 2015;39(4):276–293. doi:10.1002/gepi.21896.

62. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263–265. doi:10.1093/bioinformatics/bth457.

chr19:19410750 A/G

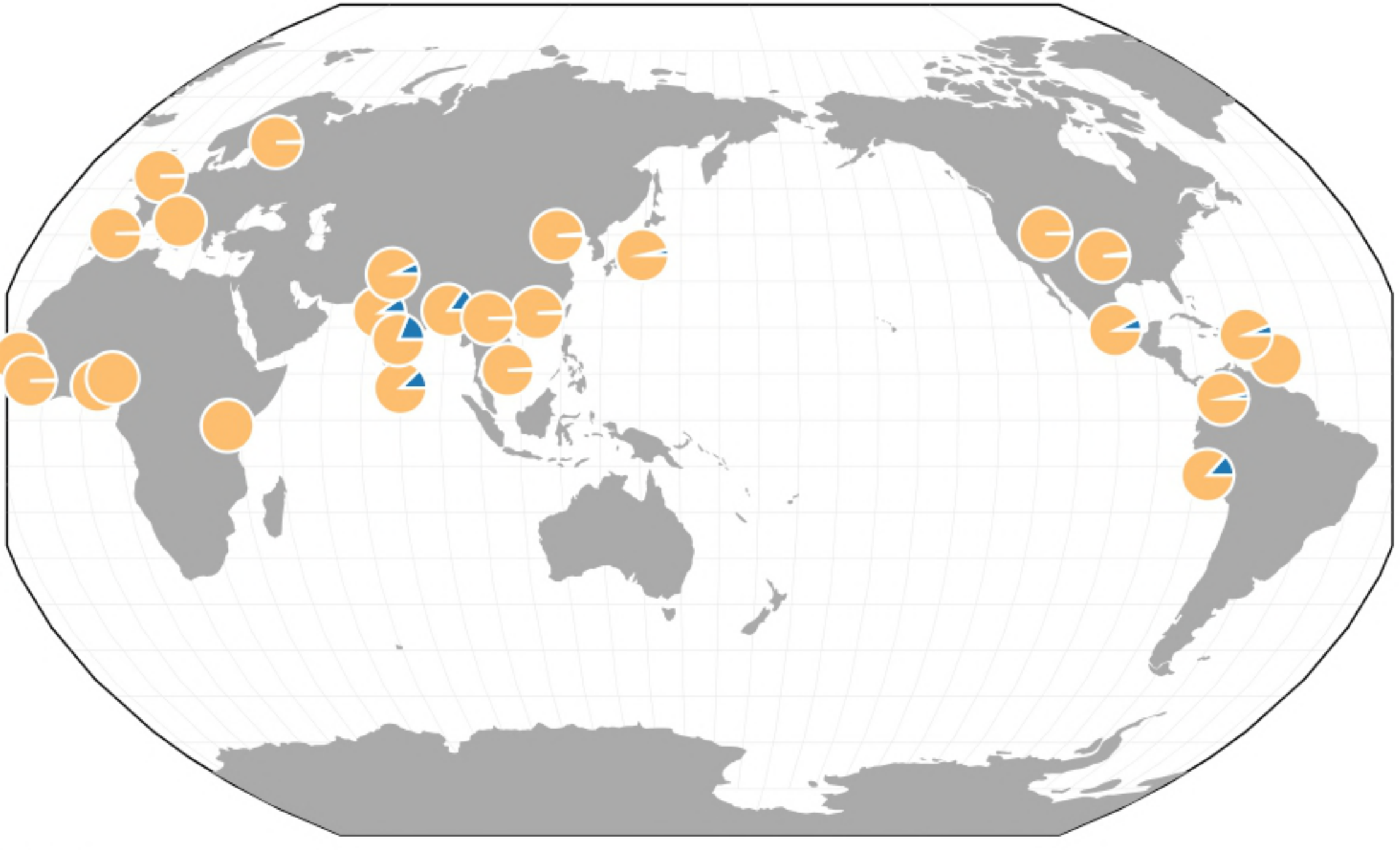

Fig2

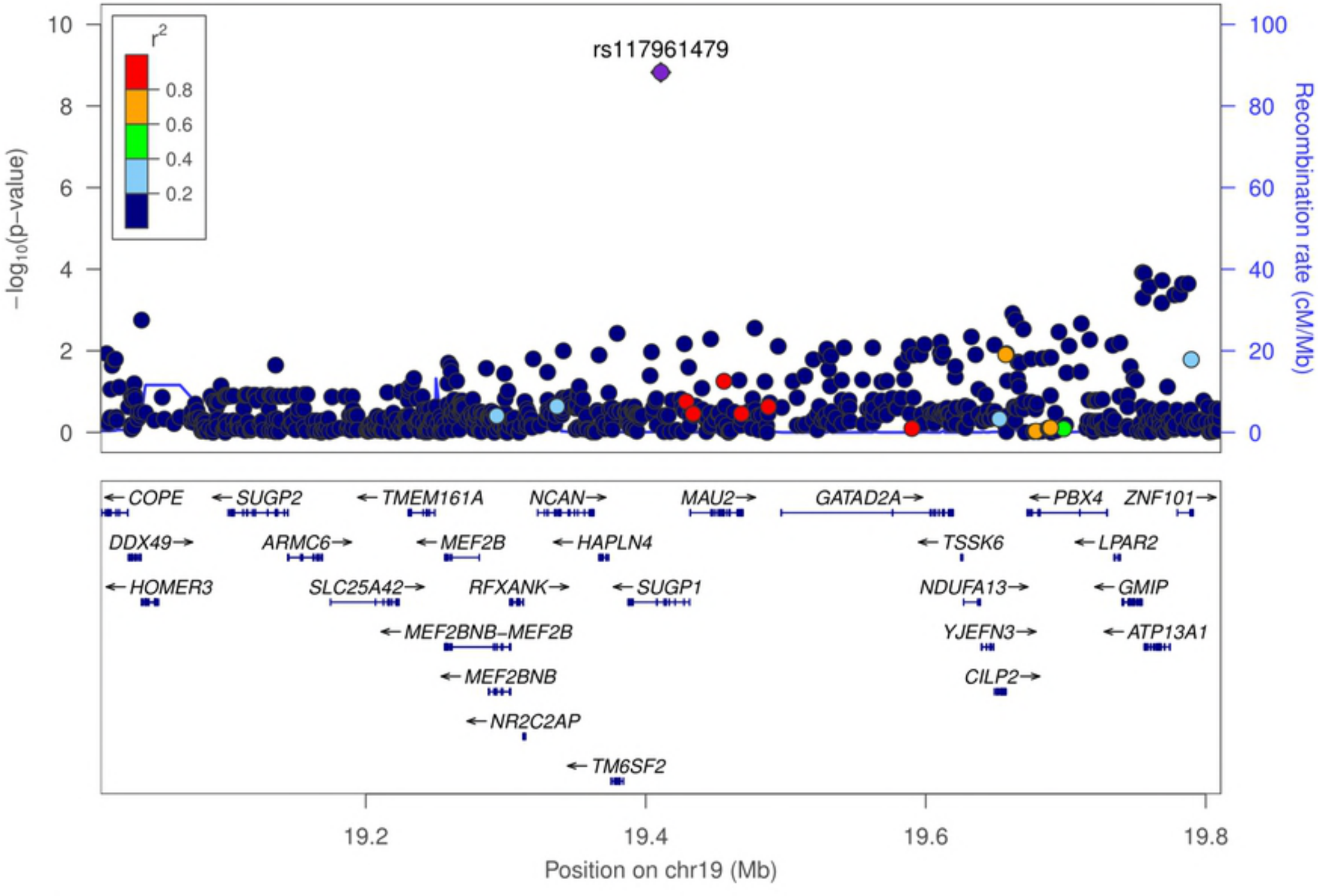
rs117961479
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
COPE
SUGP2
TMEM161A
NCAN
MAU2
GATAD2A
PBX4
ZNF101
DDX49
ARMC6
MEF2B
HAPLN4
TSSK6
LPAR2
HOMER3
SLC25A42
RFXANK
SUGP1
NDUFA13
GMIP
MEF2BNB−MEF2B
YJEFN3
ATP13A1
MEF2BNB
CILP2
NR2C2AP
TM6SF2
19.2
19.4
19.6
19.8
Position on chr19 (Mb)

Fig3

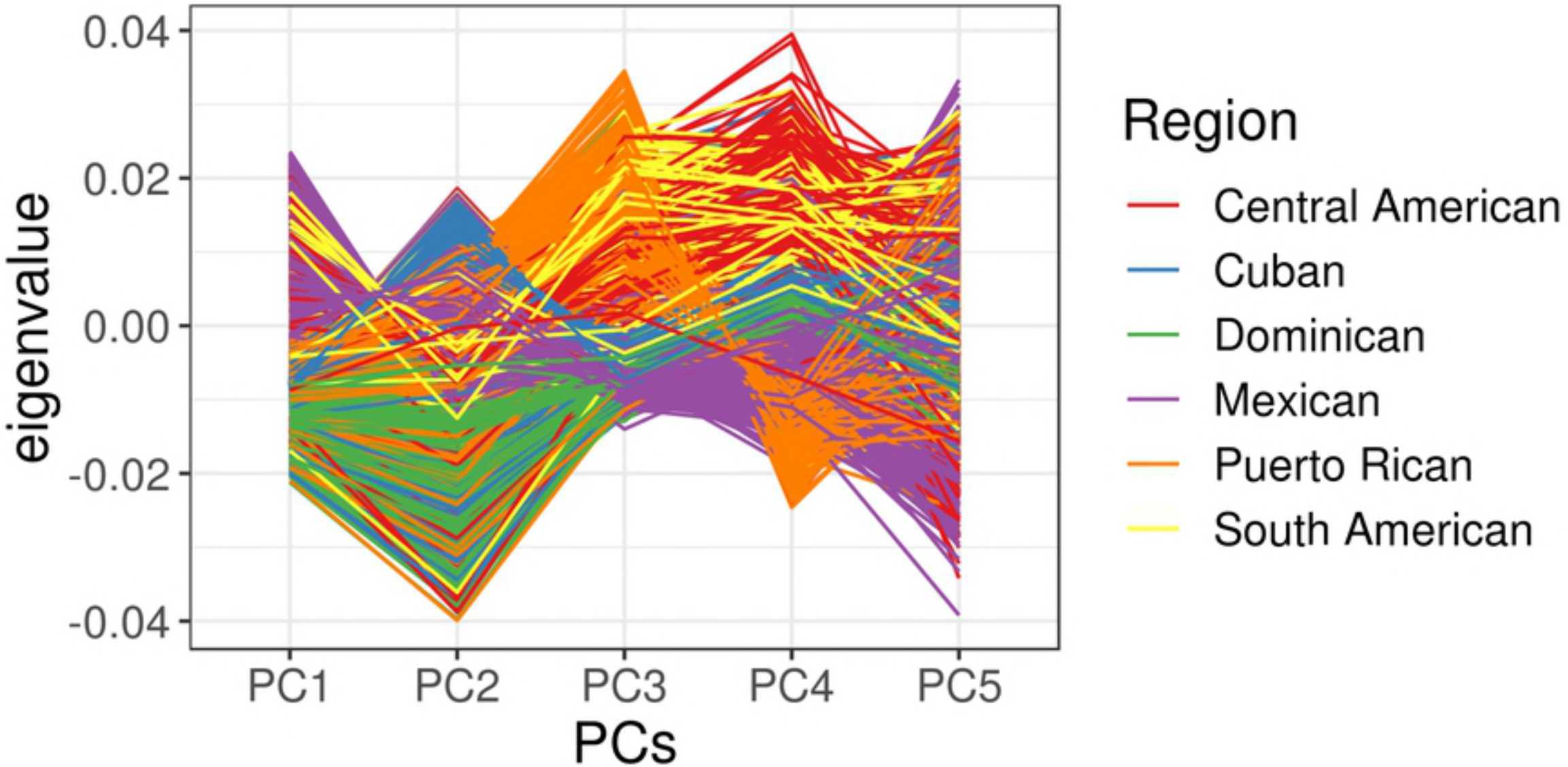
0.04
0.02
0.00
-0.02
-0.04
eigenvalue
PC1
PC2
PC3
PC4
PC5
PCs
Region
Central American
Cuban
Dominican
Mexican
Puerto Rican
South American

Fig1

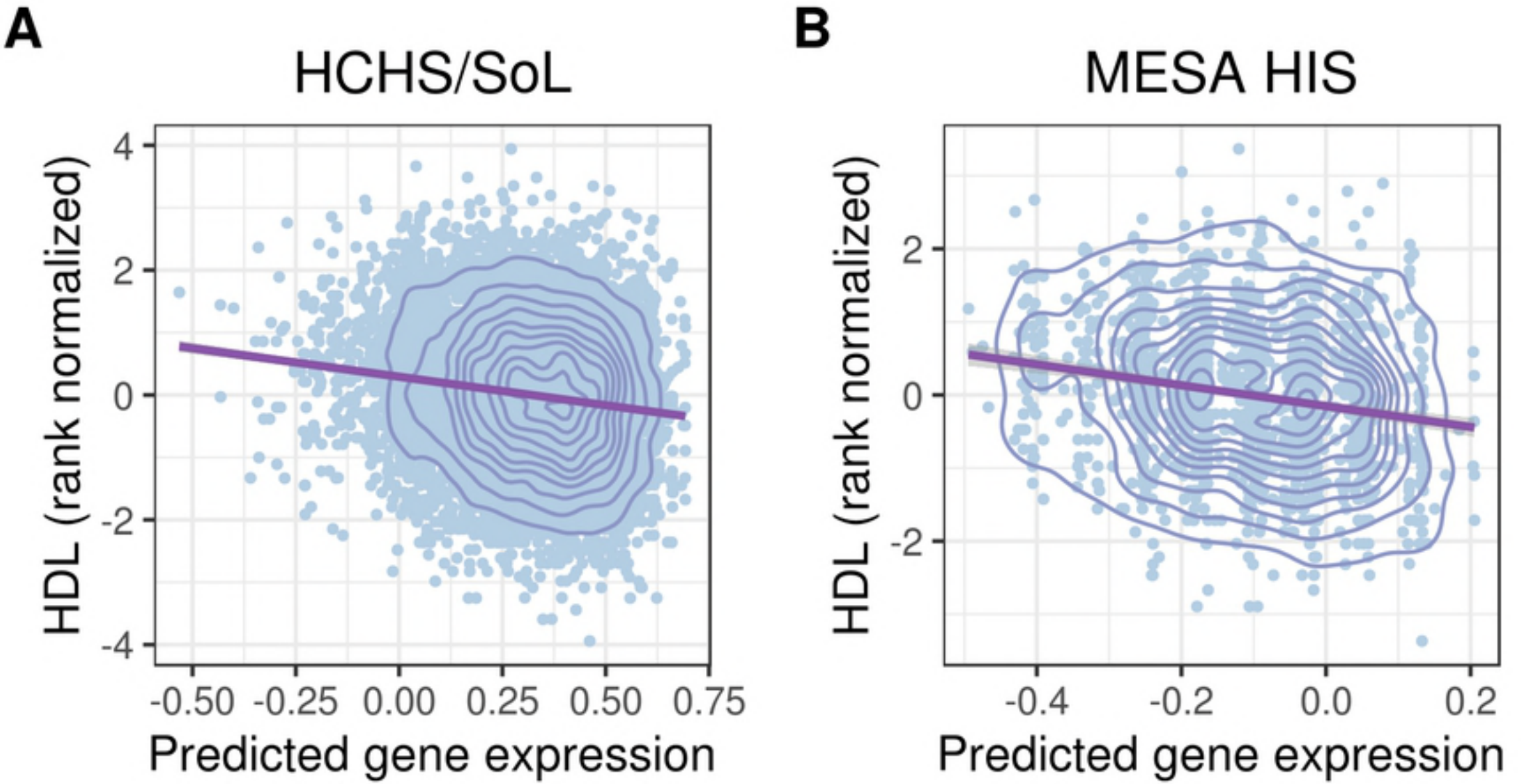
A
HCHS/SoL
HDL (rank normalized)
4
2
0
-2
-4
-0.50
-0.25
0.00
0.25
0.50
0.75
Predicted gene expression
B
MESA HIS
HDL (rank normalized)
2
0
-2
-0.4
-0.2
0.0
0.2
Predicted gene expression

Fig4

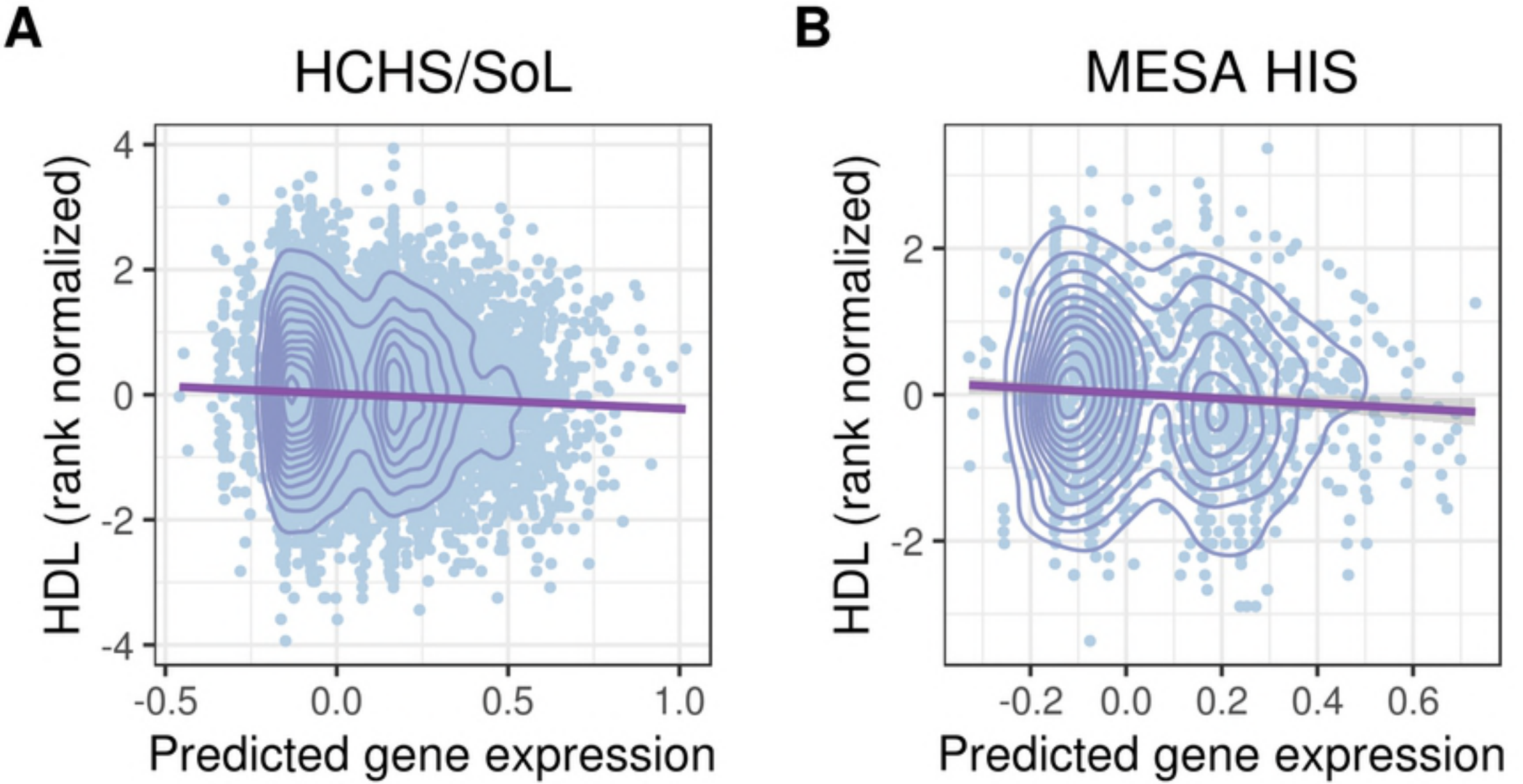
A
HCHS/SoL
HDL (rank normalized)
4
2
0
-2
-4
-0.5
0.0
0.5
1.0
Predicted gene expression
B
MESA HIS
HDL (rank normalized)
2
0
-2
-0.2
0.0
0.2
0.4
0.6
Predicted gene expression

Fig5

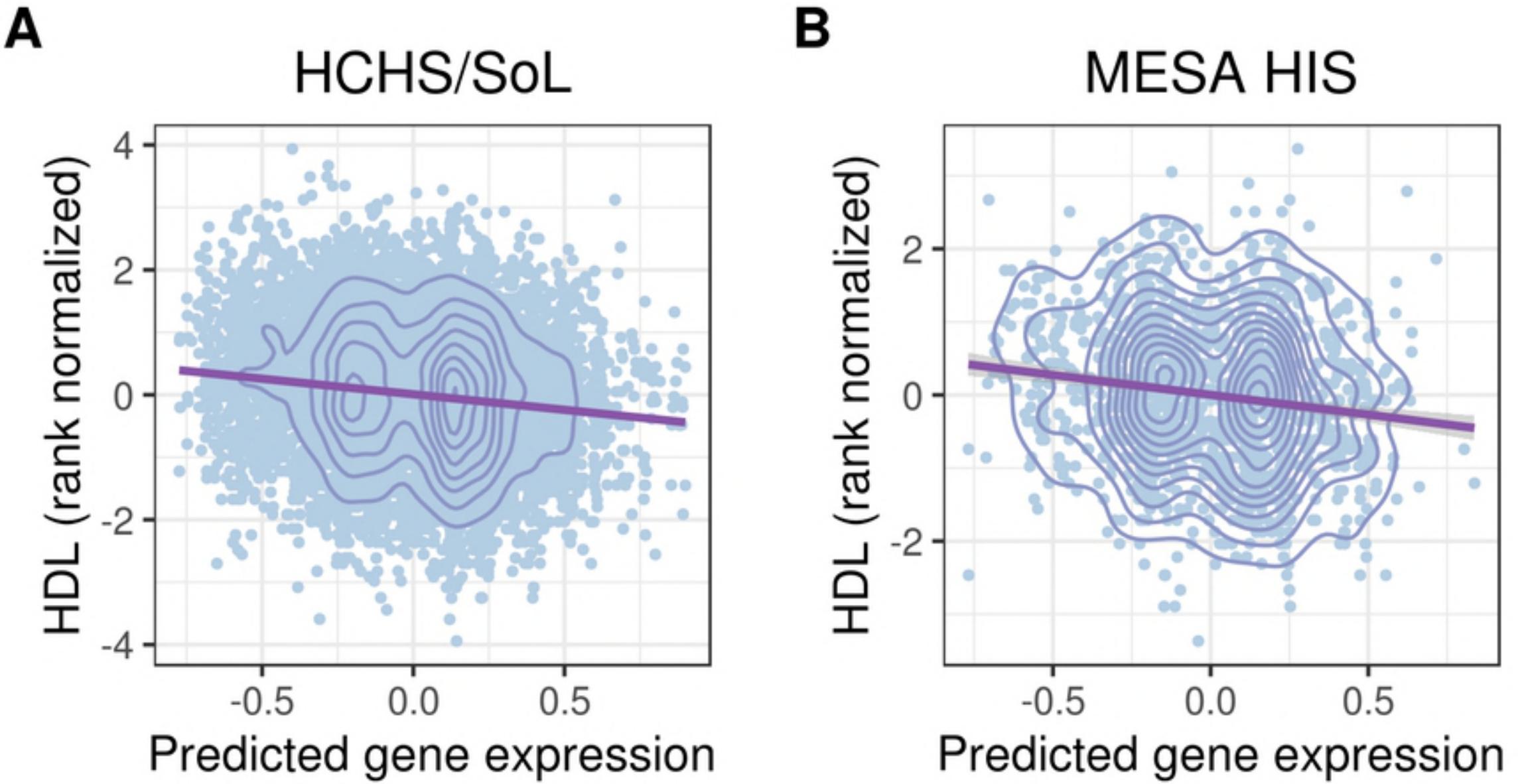
A
HCHS/SoL
HDL (rank normalized)
4
2
0
-2
-4
-0.5
0.0
0.5
Predicted gene expression
B
MESA HIS
HDL (rank normalized)
2
0
-2
-0.5
0.0
0.5
Predicted gene expression

Fig6