ASSOCIATION STUDIES ARTICLE

Integrating DNA sequencing and transcriptomic data for association analyses of low-frequency variants and lipid traits

Tianzhong Yang¹, Chong Wu², Peng Wei³ and Wei Pan^{1,*}

¹Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA,

²Department of Statistics, Florida State University, Tallahassee, FL, USA and ³Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

*To whom correspondence should be addressed at Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455, USA. Tel: +1 (612)626-2705; Email: panxx014@umn.edu

Abstract

Transcriptome-wide association studies (TWAS) integrate genome-wide association studies (GWAS) and transcriptomic data to showcase their improved statistical power of identifying gene–trait associations while, importantly, offering further biological insights. TWAS have thus far focused on common variants as available from GWAS. Compared with common variants, the findings for or even applications to low-frequency variants are limited and their underlying role in regulating gene expression is less clear. To fill this gap, we extend TWAS to integrating whole genome sequencing data with transcriptomic data for low-frequency variants. Using the data from the Framingham Heart Study, we demonstrate that low-frequency variants play an important and universal role in predicting gene expression, which is not completely due to linkage disequilibrium with the nearby common variants. By including low-frequency variants, in addition to common variants, we increase the predictivity of gene expression for 79% of the examined genes. Incorporating this piece of functional genomic information, we perform association testing for five lipid traits in two UK10K whole genome sequencing cohorts, hypothesizing that cis-expression quantitative trait loci, including low-frequency variants, are more likely to be trait-associated. We discover that two genes, *LDLR* and *TTC22*, are genome-wide significantly associated with low-density lipoprotein cholesterol based on 3203 subjects and that the association signals are largely independent of common variants. We further demonstrate that a joint analysis of both common and low-frequency variants identifies association signals that would be missed by testing on either common variants or low-frequency variants alone.

Introduction

One important mechanistic step between genetic variants and complex traits is through regulating gene expression. Two popular integrative gene-based association tests, PrediXcan (1) and TWAS (2), both of which will be called TWAS for simplicity in the sequel, leverage comprehensive regulatory knowledge

and genome-wide association studies (GWAS) data either at the individual level or at the summary level, providing opportunities to evaluate or re-evaluate GWAS focusing on common variants (CVs), with minor allele frequency (MAF) higher than 5%. These methods and the related extensions (3,4) boost statistical power and provide important functional implications that usually lack in standard GWAS. For example, lipid traits,

Received: October 24, 2019. Revised: December 11, 2019. Accepted: December 16, 2019

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

our phenotypes of interest, have estimated heritabilities of 30–70%, but the GWAS-identified CVs account for only 8–16% of the total phenotypic variation (5). Promisingly, novel CVs are found to be associated with lipid traits by leveraging transcriptomic/expression quantitative trait loci (eQTL) data, suggesting an effective way of improving statistical power with a given sample size for lipid traits (2,6). Compared with CVs, the findings for or even applications to low-frequency variants (LFVs, $MAF < 5\%$) are limited, and the underlying role of regulatory LFVs is less studied. Recently, LFVs are found to contribute to part of the remaining unexplained heritability of lipid traits in several large studies (5,7–9) and even the ultrarare variants, for example, singletons and doubletons, have substantial cis-heritability of human gene expression (10). However, due to their much lower MAF, the statistical power for identifying LFVs with moderate-to-weak effect sizes is expected to be too low. As TWAS, integrating eQTL data might be a viable alternative, and it is thus of great interest to extend such methods to LFVs.

The number of LFVs is much larger than CVs in the genomic landscape of complex traits while the regulatory roles of LFVs are less studied. In this article, we extend TWAS to DNA sequencing data to not only boost statistical power to detect associations but also provide insights into the genetic basis of lipid traits through LFVs. Statistically, TWAS is an application of two-stage least squares (2SLS): in Stage 1, it treats genetic variants within or near a gene as instrumental variables to impute or predict the gene expression in a reference eQTL dataset; in Stage 2, it correlates the imputed and so-called genetically regulated expression with a complex trait in a separate main GWAS dataset. If a statistically significant gene–trait association is uncovered in Stage 2, the gene is claimed to be causal to the trait. This causal conclusion holds only when strong modeling assumptions in 2SLS hold. Of note, there are potential pitfalls for such interpretation in a real-data analysis (11). We herein adapt a conservative and maybe more realistic perspective of taking a weighted association analysis: Stage 1 is to construct weights on genetic variants/eQTLs using a reference eQTL dataset, while Stage 2 is to conduct a weighted association analysis by putting higher weights on the eQTLs with a stronger association with gene expression. Herein, we hypothesize that LFVs associated with gene expression are more likely to be associated with complex traits, such as lipid traits. We construct the weights for LFVs using whole-blood gene expression and individual-level whole genome sequencing (WGS) data based on the Framingham Heart Study (FHS) and validate them with the Alzheimer's Disease Neuroimaging Initiative (ADNI) data (12). Then, we utilize these weights to identify lipid-associated genes in two UK10K WGS cohorts (13).

Our first two aims are to determine (1) to what extent LFVs influence gene expression and (2) whether the weights derived from eQTL data are informative in identifying genetic associations with lipid traits. Regarding aim (1), there is growing evidence suggesting a potential role of LFVs on gene expression across tissues by assessing the enrichment of LFVs for individuals with extreme gene expression levels (14). We use a more general and direct approach that is not limited to extreme values of gene expression. Regarding aim (2), it was pointed out that the assumptions underlying TWAS or 2SLS may not hold exactly due to that the constructed weights may not be precise with often small sample sizes of reference eQTL data, likely widespread horizontal pleiotropy, measurement error in expression quantification and cell-type heterogeneity (11,15,16). Under such considerations, we use a powerful adaptive test,

which covers TWAS as a special case and is more robust to estimation errors of the weights or model misspecifications, and examine the use of binary weights, in contrast to continuous weights in typical TWAS, in a real data application (Materials and Methods). Another critical question to be addressed is the relative contributions of LFVs and CVs in the same gene region to gene expression and lipid traits. Since LFVs and CVs in close physical distance may be in linkage disequilibrium (LD), the observed contribution of LFVs could be masked or originated by CVs, while a recent study suggests additive contribution of *de novo* mutations and CVs to the risk of autism spectrum disorders (17). Therefore, we address two more aims: (3) what are the contributions of LFVs (or CVs), in addition to CVs (or LFVs), to predicting gene expression and (4) whether LFVs have a unique contribution to the association signal(s) of any identified gene(s) for lipid traits.

Results

LFVs are predictive of gene expression

In the evaluation of our prediction models with LFVs (LFVs-only model), we split the FHS samples into a training dataset with 2000 samples and a test dataset with 443 samples. We predicted the expression level of a gene using LFVs within 0.5 Mb upstream and 0.5 Mb downstream of the gene body region (i.e. from transcription start site to transcription end site) by elastic net regression with 5-fold cross-validation. The initial number of LFVs per gene ranged from 6 to 6594 with mean = 1069 and median = 1047. We found 7145 out of 17 873 genes with $R^2 > 0.01$ in the training dataset, where R^2 was the squared correlation coefficient between the predicted and the observed expression levels. Furthermore, 0.01 is pre-specified as a liberal cutoff point for the predictivity of gene expression. Figure 1 demonstrates that the observed R^2 is much larger than the expected R^2 under the null distribution that none of the LFVs is predictive, suggesting that the selected LFVs are predictive of gene expression levels. As expected, the R^2 in the training dataset tended to be larger than that in the test dataset: 98% of the genes with $R^2 > 0.01$ in the test dataset also had $R^2 > 0.01$ in the training dataset. Next, we used all of the 2443 FHS samples to estimate weights for the downstream-weighted association analyses, and we had a total of 6985 genes passing the cutoff value. We further replicated the results in 798 uncorrelated ANDI samples and showed that our results are consistent for LFVs-only models (Supplementary Material, Fig. S1). The results demonstrate that LFVs play an important and universal role in explaining the gene expression variations. Of note, the elastic net regression method greatly reduced the number of candidate LFVs from thousands to hundreds or less, easing the following association analyses when applied to the UK10K cohorts (Table 1).

Evaluation of the CVs' contribution to gene expression variations

We applied the same method to CVs (CVs-only model) and observed a similar but more divergent trend of the predicted R^2 from the null distribution (Supplementary Material, Fig. S2). We further used all of the FHS samples and got 7937 genes with $R^2 > 0.01$. The initial number of CVs per gene ranged from 25 to 3651 with mean = 387 and median = 356. After variable selection via the elastic net method, the number greatly reduced (Table 1).

Figure 1. Gene expression variations explained by LFVs only. The straight line is the 45° identity line. The left panel shows the observed R^2 in comparison to the expected R^2 under the null hypothesis of no association between gene expression and LFVs; the right panel shows the observed R^2 explained by LFVs in comparison to that explained by CVs in an independent test dataset for the same genes.

Table 1. Summary statistics for different models passing the predictive performance threshold (training $R^2 > 0.01$)

Analysis	Model	No. of genes	MAF category	Min (nSNP)	Max (nSNP)	Median (nSNP)	Mean (nSNP)
LFVs only	$Y^* \sim \text{LFVs}$	6985	LFVs	1	317	18	34
CVs only	$Y^* \sim \text{CVs}$	7939	CVs	1	146	12	15
LFVs only (excluding effects of CVs)	$Y^*_{-CV} \sim \text{LFVs}$	5391	LFVs	1	307	22	37
CVs only (excluding effects of LFVs)	$Y^*_{-LFV} \sim \text{CVs}$	5391	CVs	1	119	13	16
Joint	$Y^* \sim \text{CVs} + \text{LFVs}$	7157	CVs	0	109	7	11
			LFVs	0	44	5	6

Note: nSNP is the number of variants selected in each model. Y^* is the gene expression levels. Y^*_{-CV} and Y^*_{-LFV} are the gene expression levels excluding effects of CVs and LFVs, respectively.

By looking at the LFVs-only and CVs-only models, we found that LFVs contributed to the expression variations of 1539 genes, for which CVs-only models did not reach our proposed predictivity threshold. On the other hand, CVs contributed to the expression variations of 2391 genes, for which LFVs-only models failed the pre-specified cutoff (See right panel of Fig. 1 for an overview). The results suggested that LFVs or CVs might uniquely contribute to the gene expression variations. However, these analyses could not answer the question of the relative contribution of LFVs and CVs to the expression of the remaining 5391 genes with $R^2 > 0.01$ by both the LFVs-only and CVs-only models, which is to be discussed later.

Joint contributions of LFVs and CVs to gene expression variations

As the majority of genes passed our pre-specified predictive cutoff by CVs-only and LFVs-only models, respectively, we want to get a better understanding of whether CVs and LFVs contributed to these genes independently. The fitted models are shown in Table 1, and details are described in Materials and Methods. Simply put, to evaluate the additional contributions of LFVs (or CVs), we regressed out the effects of CVs (or LFVs) and used the residuals as the new outcome to investigate its association with LFVs (or CVs). If LFVs and CVs contributed to gene expression independently, then for either group of variants (i.e. LFVs or CVs), we would have observed no or barely any change in R^2 in the models with and without removing the effects of the other

variant group (i.e. CVs or LFVs) in the test dataset for these 5391 genes. Due to the LD between LFVs and CVs, we saw a clear decreasing trend in R^2 for most of the genes in Figure 2. Examinations of these genes showed that LFVs fully captured the CVs' contribution to the expression of 211 genes (i.e., none of the CVs were selected in the CVs-only models after excluding effects of LFVs), while 2581 genes' expression levels could still be further contributed by CVs (Although the prediction power may decrease, they passed the pre-specified cutoff.). In addition, CVs fully captured LFVs' contribution to the expression of 2397 genes, while LFVs still contributed to the expression of 2500 genes after removing the effects of CVs. In conclusion, although some of the genes whose expression levels may be predicted by CVs only (LFVs only), a majority of them were predictable by both CVs and LFVs. Here, we demonstrate that both LFVs and CVs play important roles in explaining gene expression variations, consistent with findings by others (18,19).

As a more direct approach to validating this conclusion, joint analyses of both CVs and LFVs were performed. Among 8220 genes with both CVs and LFVs available, 7157 genes had $R^2 > 0.01$ in the training dataset. A total of 434 genes consisted of LFVs only, while 267 genes consisted of CVs only. Surprisingly, the rest of the genes were predictable by both CVs and LFVs. The numbers of LFVs and CVs selected per gene are presented in Table 1. Although the total number of variants selected for a joint model was not necessarily larger than a LFVs-only or CVs-only model, an outstanding number of LFVs were selected in addition to CVs across the joint models.

Figure 2. Comparing the contribution to the gene expression variations with (y-axis) and without (x-axis) taking into account the variants from another MAF category in the test dataset. The straight line is the 45° identity line. Left panel: CVs-only models before and after removing the effects of LFVs; right panel: LFVs-only models before and after removing the effects of CVs. Around 97.0% (left) and 97.1% (right) of R^2 decrease after removing the effects of the variants in the other MAF category.

Furthermore, [Figure 3](#) represents the difference of R^2 between the joint models and CVs-only (LFVs-only) models in the test dataset. Clearly, both of the histograms are right-skewed, suggesting that adding variants from the other MAF category improved the predictivity: by including LFVs in addition to CVs, the predictivity for 49% of the examined genes increased; by including CVs in addition to LFVs, the predictivity for 79% of the examined genes increased. Notably, the degree of skewness was largely in agreement with the results above.

Application to UK10K WGS cohorts: integration of WGS and transcriptomic data identifies association signals for LFVs

After establishing the relationship between LFVs (and CVs) and gene expression, we incorporated them into association analysis in two UK10K WGS cohorts: the Avon Longitudinal Study of Parents and Children (ALSPAC) and TwinsUK, with five traits, including total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), very low-density lipoprotein (VLDL), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG). We combined the results of the two studies via a fixed-effect meta-analysis model and adopted a Bonferroni-corrected significance threshold at $0.05/22\,081 = 2.3e-06$ because 6985 genes were evaluated for LFVs, 7939 genes for CVs and 7157 genes in the joint models with $R^2 > 0.01$ (Supplementary Material, Fig. S3). In addition to the continuous weights trained above, we defined binary weights, which were one for any selected LFVs and zero otherwise. We used the (weighted) Sum of Powered Score tests (SPU), including SPU (1), SPU (2) and the adaptive Sum of Powered Score (aSPU) test (as detailed in Materials and Methods), where (weighted) SPU (1) with continuous weights is (asymptotically) equivalent to TWAS and SPU (2) is a variance component test equivalent to the sum of squared score (SSU) test (20), and the aSPU test is powerful in adaptively choosing one of the multiple SPU tests, including SPU (1) and SPU (2) (21). [Table 2](#) summarizes the significant genes for five lipid traits with the two types of weights in any of the three types of

tests. The genes identified by binary weights were different from the genes identified by continuous weights. Five genes were significantly associated with LDL-C, among which LDLR was previously identified in a WGS study with 16 324 individuals (9). Two LFVs in LDLR were included, and one (chr19:11231203) with MAF 0.012 was far more significant and had a larger weight than the other in the single variant-based analysis, explaining why the tests with continuous weights were more significant than those with binary weights (Supplementary Material, Table S1).

Three genes, BCL3, MARK4 and PVRL2 are located closely on chromosome 19, and two of them were also found to be associated with TC, which was not surprising because TC and LDL-C were highly correlated (Pearson correlation > 0.85). Additionally, these genes were identified to be significant only by the aSPU tests with binary weights. It indicates that the continuous weights might not be as informative (due to, e.g. inaccurate estimation of the weights) and the signals were sparse in the region, both of which were further confirmed in the single-variant analysis (Supplementary Material, Table S1). In fact, this region is very close to the APOE gene located at chr19:45409038-45412650. APOE was established to be associated with TC and LDL-C in previous GWAS (22–24) and a large sequencing study with more than 60 000 samples (5). Lastly, TTC22 was a novel gene only identified by the aSPU test with continuous weights, and it was missed by either the SPU (1) or SPU (2). The gene resides within 3 Mb upstream of PCSK9 (chr1:55505148–55530526), a well-known gene for lipids and therapeutic targets (25,26), and was previously reported in a sequencing study (5). Interestingly, PCSK9 has long-distance enhancer regions in several cell types, for example, one predicted enhancer region locates at chr1:55271475–55274475 in the HeLa-S3 cells, overlapping with the gene region of TTC22. However, these cell types are not related to blood (27). Additionally, the expression of TTC22 and PCSK9 are in low correlation (0.11), the top signal in TTC22, chr1:55505647, was only marginally associated with PCSK9 (p -value = 0.048), while it was strongly associated with TTC22 (p -value = $4.8e-4$). After adjusting for the expression of TTC22, the p -value of the association between this variant and PCSK9 was

Figure 3. Histograms of the differences of R^2 in an independent test dataset. The vertical solid lines indicate no difference. The dashed curves are the estimated density functions. The left panel shows the difference of R^2 between the joint models and the LFVs-only models; the right panel shows the difference of R^2 between the joint models and the CVs-only models. The top of each histogram is not fully presented to save space.

Table 2. Significant gene–trait associations identified by using weights from LFVs-only models

Gene	R ²	nRV	Chr	TSS	TES	Binary weight			Continuous weight		
						SPU(1)	SPU(2)	aSPU	SPU(1)	SPU(2)	aSPU
LDL-C											
TTC22	0.044	19	1	55 246 751	55 266 941	2.57e − 01	8.44e − 03	2.20e − 05	6.02e − 02	2.90e − 05	≤ 1.00e − 06
LDLR	0.032	2	19	11 200 037	11 244 505	3.37e − 03	2.40e − 05	1.50e − 05	≤ 1.00e − 06	2.00e − 06	2.00e − 06
BCL3	0.016	5	19	45 251 977	45 263 301	4.25e − 03	2.30e − 05	≤ 1.00e − 06	4.62e − 03	2.01e − 03	2.14e − 03
PVRL2	0.240	40	19	45 349 392	45 392 485	6.93e − 01	5.70e − 05	≤ 1.00e − 06	6.39e − 01	4.60e − 01	7.37e − 01
MARK4	0.140	69	19	45 754 515	45 808 541	6.37e − 01	9.40e − 05	≤ 1.00e − 06	4.56e − 01	8.05e − 01	7.75e − 01
TC											
BCL3	0.016	5	19	45 251 977	45 263 301	2.11e − 02	6.00e − 06	≤ 1.00e − 06	5.34e − 03	8.60e − 04	1.62e − 03
MARK4	0.140	69	19	45 754 515	45 808 541	5.79e − 01	3.54e − 03	≤ 1.00e − 06	6.45e − 01	8.52e − 01	9.70e − 01

Note: The significant P-values are marked in boldface. R^2 is the predictive performance in the training dataset. nRV is the number of LFVs in a gene. TSS, transcription start site; TES, transcription end site.

0.05. We further followed Wainberg et al. to check possible false hits due to correlations between the predicted gene expression levels of two genes and/or to their shared variants, but TTC22 did not fall in these two categories (11). Therefore, we claim that there is no strong evidence that PCSK9 drove the signal of TTC22, thus it could be a novel independent signal that warrants further investigation.

The original proposed (aka, unweighted) aSPU test was shown to have superior power over the burden tests (28–30) and variance component tests (20,31) when the number of neutral variants was large (21). Of these four genes, only one or two LFVs were highly significant in the single-variant association analysis, explaining why the aSPU test gave more significant results here. Importantly, we found that using binary weights and the aSPU tests enhanced genetic locus discovery: the quantile–quantile plots of the aSPU tests with binary weights in Supplementary Material, Figure S4 show a very clear (and larger) deviation from the expectation under the null hypothesis (of no associations) for genes with small P-values, reinforcing the necessity of having a robust and powerful test in TWAS (and GWAS as usual).

Conditional analyses demonstrated independent LFV signals

We further examined these five significant genes in the LFV analysis by conditioning on the CV-regulated gene expression for LDL-C, i.e. the weighted sum of CVs. As shown in Supplementary Material, Table S2, the P-value levels of the genes had only small changes as compared with those in Table 2, suggesting that the association signals were largely independent of the CVs. As a further examination beyond the cis-eQTLs, we performed a single-variant association analysis on the CVs of these genes. We found that none of the CVs within and around LDLR and TTC22 loci was close to being statistically significant (P-values > 0.05), while there were a few highly significant ones residing in BCL3, MARK4 and PVRL2 loci. The leading association signal was chr19:45412079 with a P-value of $8.8e - 43$ and MAF 0.09 in the ALSPAC cohort and a P-value of $1.40e - 19$ and MAF 0.09 in the TwinsUK. After conditioning on this leading CV, the P-values of the three genes became larger than 0.05. However, none of the CV signals were taken into account because they were not selected in the CVs-only models to predict gene expression. Therefore, although the LFV signals may be independent of

Table 3. Significant gene–trait associations identified by using weights from CVs-only models

Gene	R ²	nCV	Chr	TSS	TES	Binary weight			Continuous weight		
						SPU(1)	SPU(2)	aSPU	SPU(1)	SPU(2)	aSPU
LDL-C											
PSRC1	0.012	7	1	109 822 175	109 825 790	9.69e – 02	6.70e – 05	6.00e – 06	≤ 1.00e – 06	≤ 1.00e – 06	≤ 1.00e – 06
CEACAM19	0.034	22	19	45 174 723	45 187 627	2.60e – 03	≤ 1.00e – 06	≤ 1.00e – 06	7.75e – 01	≤ 1.00e – 06	≤ 1.00e – 06
TC											
CEACAM19	0.034	22	19	45 174 723	45 187 627	2.08e – 02	≤ 1.00e – 06	≤ 1.00e – 06	7.75e – 01	1.20e – 04	3.06e – 04
HDL											
MT1DP	0.030	15	16	56 677 598	56 678 853	4.32e – 02	≤ 1.00e – 06	≤ 1.00e – 06	2.64e – 02	1.30e – 03	4.10e – 03
MT1JP	0.049	28	16	56 669 650	56 670 998	1.95e – 01	≤ 1.00e – 06	≤ 1.00e – 06	1.94e – 01	4.00e – 02	1.08e – 01
NLRC5	0.137	20	16	57 050 985	57 117 436	1.15e – 02	≤ 1.00e – 06	≤ 1.00e – 06	5.70e – 02	1.06e – 01	6.29e – 02
POLR2C	0.012	23	16	57 496 550	57 505 921	1.27e – 03	≤ 1.00e – 06	≤ 1.00e – 06	5.00e – 02	1.68e – 01	1.02e – 01

Note: The significant P-values are marked in boldface. nCV is the number of CVs in a gene.

the CV-regulated gene expression/common cis-eQTLs, it is still possible that the signals were driven by their LD with other non-cis-eQTLs or cis-eQTLs excluded due to model misspecification.

Association signals for CVs

Similarly, applying the constructed weights for CVs to the two UK10K cohorts, we found two genes associated with LDL-C (Table 3). CEACAM19 overlaps with APOE on chromosome 19, which is very close to the three genes, MARK4, BCL3 and PVRL2, discovered in the weighted LFV analysis. For the four genes associated with HDL, the genomic regions of MT1DP, MR1JP and NLRC5 overlap with those of CETP and HERPUD1, reported being associated with HDL in the National Human Genome Research Institute (NHGRI) GWAS Catalog (32). None of the genes associated with HDL was found in the LFV analysis.

In summary, some genes discovered in the weighted CV analysis and weighted LFV analysis reside in the same regions, while there was also some evidence for independent signals. In both the CV and LFV association analyses, the aSPU test showed its superior performance as compared with SPU (1), or equivalently, the standard TWAS.

Additional association signals when analyzing CVs and LFVs jointly

By using the weights derived from the joint models, we found that most genes reaching the genome-wide significance were previously identified in CV or LFV association analyses (Table 4). However, for these genes, the included variants were no longer from one MAF category (<5% versus >5%), except for LDLR of which all the variants were low frequency. Additionally, although two CVs were included in TTC22, their P-values from individual-variant analyses were both larger than 0.05, confirming that the signals of TTC22 predominantly came from LFVs. Interestingly, two genes TOMM40 and CCR2 were only observed in this joint analysis, for which more details are provided in Supplementary Material, Table S3. From the single-variant analysis, TOMM40 was in fact mainly driven by one CV (chr19:45422160) with a P-value of 2.78e – 16 and MAF 0.19 in the ALSPAC and a P-value of 4.51e – 4 and MAF 0.17 in the TwinsUK cohort. TOMM40 is located close to the gene regions of BCL3, CEACAM19, MARK4 and PVRL2, whose signals disappeared after adjusting for the leading CV chr19:45412079. TOMM40 was not identified by our CV analysis because its gene expression was not predicted to reach our pre-

specified threshold ($R^2 = 0.001 < 0.01$). In the CEU population (33), chr19:45422160 and chr19:45412079 have a modest correlation ($R^2 = 0.0195$). Conditioning on chr19:45412079, the P-value of the aSPU test increased to 1.50e – 03. On the other hand, CCR2 was driven by a few LFVs and CVs collectively with moderate small P-values. This gene resides in chromosome 3, which was missed in either CV analysis or LFV analysis and was supported by human studies (34,35).

Simulations

We performed a simulation study to evaluate the numerical performance of the aSPU, SPU (1)/TWAS and other tests under either correct or inaccurate weight specification. Briefly, we used the 1000 Genomes Project Phase 3 European ancestry data as a reference to generate DNA sequencing data and the association strengths of LFVs on the outcome were derived from real data. We mimicked various levels of weight misspecification errors. The detailed set-up is described in the Supplementary Material. Supplementary Material, Figure S5 shows that SPU (1) performed the best compared with other SPU tests when the weights were accurate; however, the SPU (1) test lost power as the weight misspecification errors increased, while SPU(γ) tests with higher γ gave better performance. The aSPU test was the overall winner by choosing the most powerful SPU test and thus maintaining high power across a wide range of scenarios.

Discussion

With increasing evidence that LFVs play an important role for complex traits (36), we performed comprehensive analyses on empirical contributions of eQTLs to gene expression and lipid traits. Variants with 1% < MAF < 5% are sometimes considered as low frequency, while variants with MAF < 1% are considered as rare. In this article, we examined variants with 0.2% < MAF < 5% and called them as LFVs for convenience. Interestingly, we found a number of genes whose expression can be well predicted by LFVs only. Further analyses confirmed that both LFVs and CVs contributed to gene expression for most genes. In some cases, CVs were not predictive of gene expression at all, while LFVs were, and vice versa. Furthermore, by up-weighting the cis-eQTLs in applications to lipid traits, we found some association signals driven by LFVs only, such as in genes LDLR and TTC22 (novel) and other signals driven by both LFVs and CVs. Since gene expression serves as an intermediate molecular phenotype

Table 4. Significant gene–trait associations identified by using weights from joint models

Gene	R ²	nCV	nRV	Chr	TSS	TES	Binary weights			Continuous weights		
							SPU(1)	SPU(2)	aSPU	SPU(1)	SPU(2)	aSPU
LDL-C												
TTC22	0.028	3	8	1	55 246 751	55 266 941	8.87e − 02	1.81e − 03	1.40e − 05	8.22e − 03	3.00e − 06	3.00e − 06
PSRC1	0.027	5	6	1	109 822 175	109 825 790	1.88e − 01	9.01e − 03	≤ 1.80e − 05	3.84e − 04	≤ 1.00e − 06	≤ 1.00e − 06
LDLR	0.032	0	2	19	11 200 037	11 244 505	3.36e − 03	2.60e − 05	≤ 1.20e − 05	≤ 1.00e − 06	≤ 1.00e − 06	≤ 1.00e − 06
BCL3	0.025	6	3	19	45 251 977	45 263 301	1.85e − 03	6.50e − 05	≤ 1.00e − 06	1.25e − 01	1.07e − 02	1.46e − 02
CEACAM19	0.028	5	6	19	45 174 723	45 187 627	3.10e − 04	2.25e − 03	≤ 1.00e − 06	2.99e − 02	≤ 1.00e − 06	≤ 1.00e − 06
TOMM40	0.024	1	3	19	45 394 476	45 406 946	2.73e − 04	≤ 1.00e − 06	≤ 1.00e − 06	4.14e − 03	8.00e − 06	≤ 1.40e − 05
TC												
BCL3	0.025	0	3	19	45 251 977	45 263 301	9.35e − 03	4.00e − 06	≤ 1.00e − 06	5.44e − 02	5.10e − 03	7.30e − 03
TOMM40	0.024	1	3	19	45 394 476	45 406 946	7.99e − 04	≤ 1.00e − 06	≤ 1.00e − 06	1.77e − 02	3.10e − 04	7.80e − 04
HDL												
CCR2	0.045	8	7	3	46 395 234	46 402 413	5.01e − 01	3.00e − 06	1.30e − 05	4.70e − 01	1.40e − 03	2.80e − 03
NLRC5	0.135	13	12	16	57 050 985	57 117 436	1.44e − 02	≤ 1.00e − 06	≤ 1.00e − 06	9.30e − 02	1.16e − 01	7.79e − 02

Note: The significant P-values are marked in boldface. nCV is the number of CVs included, and nRV is the number of LFVs included in a gene.

between genetic variants and complex traits, it is reassuring that LFVs are important for both gene expression and traits in our findings. With more sequencing data available, our results reinforce the importance of exploring the relationship between complex diseases and LFVs, especially the regulatory LFVs.

Due to the data availability, we focused on gene expression in whole blood, which is probably a relevant tissue for lipid traits. By leveraging sharing of cis-eQTLs across tissues, we may be able to gain higher power for discovering susceptibility loci (37,38). As in TWAS, cis-eQTLs were the main focus of the article, while further analysis on cis-eQTLs with lower frequency and rare trans-eQTLs would be of great interest (39,40). We excluded a large number of very rare genetic variants, such as singletons and doubletons, due to two reasons: (1) we want to ensure the stability and validity of the gene expression prediction models and (2) our association study probably does not have enough power for these variants due to the sample size. However, we acknowledge the potential contributions of ultra-rare variants to gene expression and complex traits (10,41). We checked the signs of the correlations between the observed and predicted gene expression for the significant genes, which were all positive in the training and test data. We used a somewhat arbitrary cutoff of $R^2 > 0.01$ to select the genes to construct the weights for association testing, along the line proposed in (1).

Last but not least, we emphasize the use of the aSPU test and binary weights for the integration of DNA sequencing and transcriptomic data. The aSPU test maintained high power in both simulations and real data applications, due to its adaptiveness to various unknown association patterns and its robustness to underlying model assumptions. In particular, the aSPU test combines the results of the (weighted) Sum test (equivalent to TWAS), SPU (2) (as a variance component test) and other SPU tests. Of course, other (weighted) tests can be possibly applied too (42–44). In addition, continuous weights are often used in TWAS and in general 2SLS. As we relax the strong assumptions on instrumental variables and consider them as weighted association analyses, binary weights are very straightforward to use. The main idea is that we search for and then test only on the cis-eQTLs within and around the gene body region while excluding a large number of non-associated genetic variants, thus gaining statistical power. Binary weighting has shown to be useful both in our real data applications (for both CVs and LFVs) and in some other extensions, such as for methylation QTLs (41,45)

and GxE interactions (Yang et al., manuscript in preparation). On the other hand, while the aSPU test is in general powerful, strictly speaking it is a weighted association test without causal interpretation; in contrast, TWAS (or equivalently, the weighted Sum test) can estimate and test the causal effect, including its size and direction, of genetically regulated components of gene expression on a trait when the stringent modeling assumptions hold for the underlying instrumental variables analysis (i.e. 2SLS). It would be worthwhile to investigate how to combine the advantages of the two types of the approaches to yield more robust, powerful and interpretable methods (46,47).

Materials and Methods

FHS cohort as the expression reference panel

The FHS is a longitudinal multigenerational study aiming at finding risk factors for cardiovascular disease (48,49). The Systems Approach to Biomarker Research in Cardiovascular Disease Initiative (SABRe CVD) gene expression case–control study is a substudy of the FHS, which generated high-throughput gene expression profiles. In addition, as part of the NHLBI Trans-Omics for Precision Medicine (TOPMed) WGS Program, around 4100 subjects in FHS had deep coverage WGS data available.

Gene expression

Gene expression was measured in whole blood samples among the FHS offspring cohort (exam 8) and third-generation cohort (exam 2). RNA expression profiling was conducted using the Affmetrix Human Exon Array ST 1.0. A total of 5626 individuals with 17 873 measured genes whose gene expression data were available, and we further restricted to the individuals with WGS data and without missing values in important covariate variables including gender, age, treatment for lipids, hypertension and diabetes at the corresponding examination. The gene expression data were preprocessed largely following Joehanes et al. (11). We used partial least square with 20 principal components to impute the cell types (whole blood cell, red blood cell, platelet, lymphocyte, monocyte, eosinophil and basophil) based on raw gene expression data and calculated 20 probabilistic estimations of expression residuals (PEER) factors

to capture the hidden confounding factors (50). The imputed cell types and 20 PEER factors were included as additional covariates. Since the subjects in the two cohorts are correlated, we used a mixed-effect model as implemented in an R package GMMAT (version 0.7) to remove the relatedness among subjects (51), whereas the effects of covariates were treated as fixed and effects of genotypes were treated as random. The random effects followed a normal distribution with mean 0 and a covariance matrix proportional to the kinship matrix (52). Residuals of this mixed-effect model were used as the gene expression in the weight construction step described later. A total of 2443 subjects with gene expression and WGS data available and without missing covariates were used to build the prediction models.

Whole genome sequencing

The subjects were sequenced using Illumina HiSeq X with $30 \times$ mean read-depth. We removed genetic variants with Hardy-Weinberg Equilibrium P-value < 0.05 , ambiguous strand mapping, missing call rates exceeding 5%, duplicates. Variants with minor allele counts less than 10 (corresponding to $MAF < 0.2\%$) and on sex chromosomes were excluded. To minimize the impact of high LD among variants and reduce computation burden, we pruned out variants with pairwise R^2 larger than 0.8, leaving 3 298 157 LFVs ($MAF < 5\%$) and 770 828 CVs ($MAF > 5\%$) for analysis, all in hg19 coordinates.

Models for weights construction

We built prediction models in FHS offspring and third-generation cohorts using the elastic net regression ($\alpha = 0.5$) with 5-fold cross-validation (53). To evaluate the performance of the elastic net model, 2000 subjects were used in the cross-validation dataset, while 443 subjects were used as the independent test dataset. After assessing the model performance, we used all of the 2443 subjects to construct weights in the following weighted association analyses.

The prediction model is as follows:

$$Y^* = \sum_{k=1}^p w_k X_k^* + \varepsilon$$

where Y^* is the (residual) gene expression (after adjusting for covariates and subject relatedness). X_k^* is the standardized number of alleles for variant k , p is the number of genetic variants and ε is the error term that could not be explained by the genetic variants, for example, attributable to the environmental and behavioral factors. w_k is the regression coefficient for genetic variant k and the weight for the following association analyses. We constructed the weights by using CVs-only elastic net regression including variants with $MAF > 5\%$ or LFVs-only elastic net regression including variants with $MAF < 5\%$ within 0.5 Mb upstream and 0.5 Mb downstream of the gene body region. Methods other than elastic net, such as Lasso regression or Bayesian sparse linear mixed models, have been used in building weights for CVs (1,2). These other methods were shown to perform similarly to elastic net regression, though elastic net regression might yield more robust and stable results in general (54).

We assessed the model performance by R^2 , calculated as $R^2 = [\text{cor}(Y^*, \hat{Y}^*)]^2$, where $\hat{Y}^*$ is the predicted residualized gene

expression. If the training $R^2 > 0.01$, we used these weights for the following analyses. Suppose q variants are selected, the null distribution of R^2 with sample size n^* follows a beta distribution when the estimated coefficients are ordinary least squares estimators (55): $R^2 \sim \text{Beta}\left(\frac{q-1}{2}, \frac{n^*-q}{2}\right)$. Therefore, the expected R^2 under the null distribution is $\frac{q-1}{n^*-1}$, which serves as an approximation of the expected R^2 in penalized regression.

Joint analyses

LFVs and CVs may both contribute to the variations of gene expression. However, it is not clear whether their contributions are independent or are due to their LD. Therefore, we used the following methods for each gene as follows.

First, we calculated the CV-regulated gene expression denoted as $\hat{Y}_{CV}^*$. If the R^2 was larger than 0.01 in the training dataset, we removed the effects of CVs by deducting the CV-regulated gene expression from the observed gene expression, denoted as Y_{-CV} , that is

$$Y_{-CV} = Y^* - \hat{Y}_{CV}^*$$

Second, we performed the elastic net regression on the Y_{-CV} with LFVs as the predictors. The R^2 of this model was regarded as the additional variation that LFVs explained for gene expression and denoted as R_{-CV}^2 (LFV).

We calculated the additional variation that CVs explained according to the above two steps, except that we calculated Y_{-RV} by deducting the LFV-regulated gene expression $\hat{Y}_{LFV}^*$, which we then regressed with CVs as predictors. The R^2 of this model was regarded as the additional variation that CVs explained for gene expression and denoted as R_{-LFV}^2 (CV).

Since such method is quite conservative for exploring the relative contributions of LFVs and CVs, we used the joint models where both LFVs and CVs were included. We constructed the weights as described above: variants with any MAF within 0.5 Mb upstream and 0.5 Mb downstream of the gene body region were included in building the joint models.

Finding associations in two UK10K WGS cohorts

Two UK10K WGS cohorts—the ALSPAC and TwinsUK—were used for association analyses with the lipid traits (13). In the ALSPAC, a sample of 1976 children was selected for WGS and passed the variant calling procedure. The lipid traits were measured at age of 9 years. In the TwinsUK, 1934 female subjects with age from 17 to 84 years had WGS passing variant calling procedure. In total, 3910 subjects were sequenced using Illumina next-generation sequencing technology at an average $6.7 \times$ mean read-depth. Subjects with missing values in the lipid traits were excluded from the analysis, which leads to a total of 1497 samples in the ALSPAC and 1706 samples in the TwinsUK. A more detailed description of the two cohorts is available somewhere else (7).

Phenotypes

Five lipid traits, including LDL-C, HDL, TG, VLDL and TC, were standardized for each cohort following two steps. In the first step, outliers deviating more than 4 standard deviations from the sample mean for a given trait were excluded from the analysis. In the second step, the inverse normal rank transformation was

performed on the traits. The two steps were performed on the original scale of the LDL-C, HDL, VLDL and TC, but on the log scale of TG. Among the five traits, the correlation between LDL-C and TC and correlation between TC and log(TG) were high (>0.85), while others are relatively small (<0.3). Gender was adjusted in the ALSPAC cohort, while age and age-squared were adjusted in the TwinsUK cohort.

Method for association testing: the aSPU test

To find the association between the genetic variants and lipid traits, we used the following model:

$$Tr = \beta_0 + Z\theta + \sum_{k=1}^p (X_k \hat{w}_k) \beta_k + \varepsilon$$

Tr is the lipid trait, Z is the covariates, $\hat{w}_k$ is the external weights in the above weight construction step after matching the reference panel and X_k is the standardized number of alleles for variant k within and near the gene. $X_k \hat{w}_k$ is the weighted genotype score. The goal is to test $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$. This model prioritizes cis-eQTLs, i.e. a cis-eQTL with larger association strength with gene expression is assumed to have a stronger association with the lipid traits. Since the genetic markers may not be valid instrumental variables and more generally, the assumptions underlying 2SLS (or Mendelian randomization) could be violated and the estimated weights may be biased. In addition, the measurement error of the constructed weights is inevitable, especially for LFVs. Therefore, instead of using TWAS with a stronger assumption on the magnitude of the weights, we used the aSPU test in the context of weighted association analysis, which is shown to be powerful previously (45).

Let the weighted score function for variant k be $U_k = \sum_{i=1}^n X_k^T \hat{w}_k (Tr_i - \mu_i)$, where μ_i is the fitted mean of Tr_i under H_0 and n is the number of subjects included in the association study. The aSPU test is built upon a family of the SPU test, defined as

$$T_{SPU(\gamma)} = \sum_{k=1}^p U_k^{\gamma-1} U_k = \sum_{k=1}^p U_k^{\gamma},$$

where γ is a positive integer. $U_k^{\gamma-1}$ can be considered as a type of estimated weight where the data are used to construct weights (as opposed to external weight $\hat{w}_k$ constructed from an independent reference cis-eQTL dataset) (56). As γ increases, the SPU test puts larger estimated weights on the variants, which are more strongly associated with the trait. Of note, TWAS is a special case of the SPU test when $\gamma = 1$. In addition, weighted SSU test is equivalent to the SPU ($\gamma = 2$) test. Furthermore, when $\gamma = \infty$, $T_{SPU(\gamma)} \propto \left(\sum_{k=1}^p (|U_k|)^{\gamma} \right)^{1/\gamma} \rightarrow \max |U_k| = T_{SPU(\infty)}$. The SPU(∞) chooses the largest weighted score function among all the genetic markers, which is closely related to the UminP test (20). In practice, using $\gamma \in \Gamma = \{1, 2, \dots, 8, \infty\}$ suffices in most scenarios (57) thus we adopt it herein. The estimated weighting improved the robustness of the SPU tests to the use of external weights. In addition, we relaxed our assumptions by performing analyses using binary weights. Basically, we set weights $\hat{w}_k$ to one for all selected variants in the prediction models for gene expression and zero otherwise. Thus, we include the cis-eQTLs for hypothesis testing and allow the aSPU test to decide the adaptive estimated weights for each score function.

Since we have two UK10K cohorts with different demographic features, we combined them using the fixed-effect meta-analysis instead of simple pooling. Denote the SPU test statistics in the ALSPAC cohort as $T_{SPU(\gamma)}^A$, and these in the TwinsUK cohort as $T_{SPU(\gamma)}^T$, we combined the two cohorts as follows: $T_{SPU(\gamma)} = T_{SPU(\gamma)}^A + T_{SPU(\gamma)}^T$. Since the power of SPU(γ) test depends on the underlying genetic architecture, which varies from gene to gene, it is better to select γ adaptively. To this end, the aSPU test chooses the most powerful test among the class of SPU (γ) tests, which is defined as $T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)}$, where $P_{SPU(\gamma)}$ is the P-value of the combined SPU (γ) test. The P-value of this aSPU test can be calculated by one layer of Monte Carlo simulations as a simple modification of the original aSPU test P-value calculation procedure (57).

In the application to UK10K WGS datasets, we used a step-wise procedure that gradually increases the number of Monte Carlo simulations (16,57,58) up to a total of one million simulations and stopped until either the gene reached genome-wide significance level or it was obviously not significant. The implementation of Monte Carlo method is even faster with parallel computing: it took 2 h or less for 22 computing cores to complete the genome-wide scans for the UK10K WGS data, i.e. it took at most 44 h for all examined genes by a single core processor.

Data and code availability

The WGS and transcriptomic data of the FHS study are accessible from the National Center for Biotechnology Information dbGap with access numbers phs000974 and phe000002.v7. The study also makes use of data generated by the UK10K Consortium, derived from the TwinsUK and ALSPAC cohorts. Data are available from UK10K Data Access Committee for researchers who meet the criteria for access to confidential data. The core R code for implementation is packaged in R called 'metaRV', available at <https://github.com/ytzhong/metaRV>. The trained weights based on full FHS samples for LFVs-only, CVs-only and joint models are available at <https://github.com/ytzhong/RVweights>.

Supplementary Material

Supplementary Material is available at HMG online

Funding

National Institutes of Health (grants R21AG057038, R01HL116720, R01GM113250, R01GM126002 and R01HL105397).

Conflict of Interest statement

The authors declare no conflict of interest.

Acknowledgements

We thank the Minnesota Supercomputing Institute and Texas Advanced Computing Center for providing computing resources. The FHS is conducted and supported by the NHLBI in collaboration with Boston University (Contract No. N01-HC-25195 and HHSN268201500001). This manuscript was not prepared in collaboration with investigators of the FHS and does not necessarily reflect the opinions or views of the FHS, Boston University or NHLBI. WGS for TOPMed program was supported by

the NHLBI. WGS for 'NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study' (phs000974.v1.p1) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). This study makes use of data generated by the UK10K Consortium, including samples from the TwinsUK and ALSPAC cohorts. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc. and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

- Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J. and Im, H.K. (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.
- Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Peninx, B., Jansen, R., De Geus, E.J., Boomsma, D.I., Wright, F.A. et al. (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.
- Cai, M., Chen, L., Liu, J., Yang, C. Quantifying the impact of genetically regulated expression on complex traits and diseases. (2019) *bioRxiv*. Available from <https://www.biorxiv.org/content/10.1101/546580v1>
- Yang, C., Wan, X., Lin, X., Chen, M., Zhou, X. and Liu, J. (2019) CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*, **35**, 1644–1652.
- Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S. et al. (2015) The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.*, **47**, 589–597.
- Mogil, L.S. Genetically regulated gene expression underlies lipid traits in Hispanic cohorts. (2018) *bioRxiv*. Available from <https://www.biorxiv.org/content/10.1101/507905v1>
- Timpson, N.J., Walter, K., Min, J.L., Tachmazidou, I., Malerba, G., Shin, S.Y., Chen, L., Futema, M., Southam, L., Iotchkova, V. et al. (2014) A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat. Commun.*, **5**, 4871.
- Igartua, C., Mozaffari, S.V., Nicolae, D.L. and Ober, C. (2017) Rare non-coding variants are associated with plasma lipid traits in a founder population. *Sci. Rep.*, **7**, 16415.
- Natarajan, P., Peloso, G.M., Zekavat, S.M., Montasser, M., Ganna, A., Chaffin, M., Khera, A.V., Zhou, W., Bloom, J.M., Engreitz, J.M. et al. (2018) Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.*, **9**, 3391.
- Hernandez, R.D., Uricchio, L.H., Hartman, K., Ye, C., Dahl, A. and Zaitlen, N. (2019) Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.*, **51**, 1349–1355.
- Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermerl, R., Ruusalepp, A., Quertermous, T., Hao, K. et al. (2019) Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.*, **51**, 592–599.
- Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al. (2008) The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J. Magn. Reson.*, **27**, 685–691.
- Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., Lawson, D. et al. (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–89.
- Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T.J., Lee, C.M., Banskota, S., Marigorta, U.M., Bao, G. and Gibson, G. (2016) A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.*, **98**, 299–309.
- Hemani, G., Bowden, J. and Davey, S.G. (2018) Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.*, **27**, R195–R208.
- Xu, Z., Wu, C., Wei, P. and Pan, W. (2017) A powerful framework for integrating eQTL and GWAS summary data. *Genetics*, **207**, 893–902.
- Weiner, D.J., Wigdor, E.M., Ripke, S., Walters, R.K., Kosmicki, J.A., Grove, J., Samocha, K.E., Goldstein, J.I., Okbay, A., Bybjerg-Grauholm, J. et al. (2017) Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.*, **49**, 978–985.
- Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J. et al. (2017) The impact of rare variation on gene expression across tissues. *Nature*, **550**, 239–243.
- Aguet, F., Brown, A.A., Castel, S.E., Davis, J.R., He, Y., Jo, B., Mohammadi, P., Park, Y.S., Parsana, P., Segre, A. et al. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

20. Pan, W. (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.*, **33**, 497–507.
21. Pan, W., Kim, J., Zhang, Y., Shen, X. and Wei, P. (2014) A powerful and adaptive association test for rare variants. *Genetics*, **197**, 1081–1095.
22. Rasmussen-Torvik, L.J., Pacheco, J.A., Wilke, R.A., Thompson, W.K., Ritchie, M.D., Kho, A.N., Muthalagu, A., Hayes, M.G., Armstrong, L.L., Scheftner, D.A. et al. (2012) High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin Transl Sci*, **5**, 394–399.
23. Zhu, Y., Zhang, D., Zhou, D., Li, Z., Li, Z., Fang, L., Yang, M., Shan, Z., Li, H., Chen, J. et al. (2017) Susceptibility loci for metabolic syndrome and metabolic components identified in Han Chinese: a multi-stage genome-wide association study. *J. Cell. Mol. Med.*, **21**, 1106–1116.
24. Chasman, D.I., Giulianini, F., MacFadyen, J., Barratt, B.J., Nyberg, F. and Ridker, P.M. (2012) Genetic determinants of statin-induced low-density lipoprotein cholesterol reduction: the justification for the use of statins in prevention: an intervention trial evaluating rosuvastatin (JUPITER) trial. *Circ. Cardiovasc. Genet.*, **5**, 257–264.
25. Burnett, J.R. and Hooper, A.J. (2018) PCSK9 — a journey to cardiovascular outcomes. *N. Engl. J. Med.*, **379**, 2161–2162.
26. Everett, B.M., Smith, R.J. and Hiatt, W.R. (2015) Reducing LDL with PCSK9 inhibitors—the clinical benefit of lipid drugs. *N. Engl. J. Med.*, **373**, 1588–1591.
27. Gao, T., He, B., Liu, S., Zhu, H., Tan, K. and Qian, J. (2016) EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics*, **32**, 3543–3551.
28. Li, B. and Leal, S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.
29. Morgenthaler, S. and Thilly, W.G. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, **615**, 28–56.
30. Madsen, B.E. and Browning, S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.
31. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.
32. Kilpeläinen, T.O., Bentley, A.R., Noordam, R., Sung, Y.J., Schwander, K., Winkler, T.W., Jakupovic, H., Chasman, D.I., Manning, A., Ntalla, I. et al. (2019) Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity. *Nat. Communications*, **10**, 376.
33. Machiela, M.J. and Chanock, S.J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, **31**, 3555–3557.
34. Sun, R.L., Huang, C.X., Bao, J.L., Jiang, J.Y., Zhang, B., Zhou, S.X., Cai, W.B., Wang, H. and Zhang, Y.L. (2016) CC-chemokine ligand 2 (CCL2) suppresses high density lipoprotein (HDL) internalization and cholesterol efflux via CC-chemokine receptor 2 (CCR2) induction and p42/44 mitogen-activated protein kinase (MAPK) activation in human endothelial cells. *J. Biol. Chem.*, **291**, 19532–19544.
35. Barter, P. (2005) The role of HDL-cholesterol in preventing atherosclerotic disease. *Eur. Heart J.*, **7**, F4–F8.
36. Wainschein, P., Jain, D.P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, et al. (2019) Recovery of trait heritability from whole genome sequence data. *bioRxiv*, 588020. Available from <https://www.biorxiv.org/content/10.1101/588020v1>
37. Barbeira, A.N., Pividori, M.D., Zheng, J., Wheeler, H.E., Nicolae, D.L. and Im, H.K. (2019) Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.*, **15**, e1007889.
38. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M., Yu, Z., Li, B., Gu, J., Muchnik, S. et al. (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.*, **51**, 568–576.
39. Wheeler, H.E., Ploch, S., Barbeira, A.N., Bonazzola, R., Andaleon, A., Fotuhi Siahpirani, A., Saha, A., Battle, A., Roy, S. and Im, H.K. (2019) Imputed gene associations identify replicable trans-acting genes enriched in transcription pathways and complex traits. *Genet. Epidemiol.*, **43**, 22205.
40. Yang, F., Wang, J., Pierce, B.L. and Chen, L.S. (2017) Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res.*, **27**, 1859–1871.
41. Hernandez, R.D., Uricchio, L.H., Hartman, K., Ye, C., Dahl, A. and Zaitlen, N. (2018) Singleton variants dominate the genetic architecture of human gene expression. *bioRxiv*, 219238. Available from <https://www.biorxiv.org/content/10.1101/219238v2>
42. Lee, S., Abecasis, G.R., Boehnke, M. and Lin, X. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.
43. Basu, S. and Pan, W. (2011) Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.*, **35**, 606–619.
44. Zhang, X., Basile, A.O., Pendergrass, S.A. and Ritchie, M.D. (2019) Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics*, **20**, 46.
45. Wu, C. and Pan, W. (2019) Integration of methylation QTL and enhancer–target gene maps with schizophrenia GWAS summary results identifies novel genes. *Bioinformatics*, **35**, 3576–3583.
46. Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A. and Pasaniuc, B. (2019) Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.*, **51**, 675–682.
47. Wu, C., Pan, W. (2019). A powerful fine-mapping method for transcriptome-wide association studies. *Hum. Genet.* doi:10.1007/s00439-019-02098-2
48. Feinleib, M., Kannel, W.B., Garrison, R.J., McNamara, P.M. and Castelli, W.P. (1975) The Framingham Offspring Study. Design and preliminary data. *Prev. Med.*, **4**, 518–525.
49. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J. et al. (2007) The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am. J. Epidemiol.*, **165**, 1328–1335.
50. Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, **7**, 500–507.

51. Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Ceder, J.C., et al. (2016) Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.*, **98**, 653–666.
52. Cao, Y., Maxwell, T.J. and Wei, P. (2015) A family-based joint test for mean and variance heterogeneity for quantitative traits. *Ann. Hum. Genet.*, **79**, 46–56.
53. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat Methodol.*, **67**, 301–320.
54. Fryett, J.J., Inshaw, J., Morris, A.P. and Cordell, H.J. (2018) Comparison of methods for transcriptome imputation through application to two common complex diseases. *Eur. J. Hum. Genet.*, **26**, 1658–1667.
55. Meloun, M., Militky, J. (2011) *Statistical Data Analysis: A Practical Guide*. Woodhead Publishing India.
56. Roeder, K. and Wasserman, L. (2010) Genome-wide significance levels and weighted hypothesis testing. *Stat Sci*, **24**, 398–413.
57. Yang, T., Kim, J., Chong, W., Ma, Y. and Wei, P. (in press) An adaptive test for meta-analysis of rare variant association studies. *Genet. Epidemiol.*
58. Ma, Y. and Wei, P. (2019) FunSPU: a versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. *PLoS Genet.*, **15**, e1008081.
59. Wu, C. and Pan, W. (2018) Integration of enhancer-promoter interactions with GWAS summary results identifies novel schizophrenia-associated genes and pathways. *Genetics*, **209**, 699–709.
60. Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S., Nguyen, Q.T., Demirkale, C.Y., Feolo, M.L., Sharopova, N.R., Sturcke, A. et al. (2017) Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.*, **18**, 16.