Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improve the power of transcriptome-wide association studies

Authors

Helian Feng[*1,2], Nicholas Mancuso [3,4], Alexander Gusev[5,6,7], Arunabha Majumdar[8,9], Megan Major[10], Bogdan Pasaniuc[8,9], Peter Kraft[1,2]

Affiliations

1. *Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA;*
2. *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America;*
3. *Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America;*
4. *Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles California, United States of America;*
5. *Department of Medical Oncology, Dana-Farber Cancer Institute & Harvard Medical School, Boston, MA, USA*
6. *Division of Genetics, Brigham & Women's Hospital, Boston, MA, USA*
7. *Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, United States of America;*
8. *Department of Human Genetics, University of California Los Angeles, Los Angeles, California, United States of America;*
9. *Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, California, United States of America*
10. *Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, California*

Abstract

Transcriptome-wide association studies (TWAS) test the association between traits and genetically predicted gene expression levels. The power of a TWAS depends in part on the strength of the correlation between a genetic predictor of gene expression and the causally relevant gene expression values. Consequently, TWAS power can be low when expression quantitative trait locus (eQTL) data used to train the genetic predictors have small sample sizes, or when data from causally relevant tissues are not available. Here, we propose to address these issues by integrating multiple tissues in the TWAS using sparse canonical correlation analysis (sCCA). We show that sCCA-TWAS combined with single-tissue TWAS using an aggregate Cauchy association test (ACAT) outperforms traditional single-tissue TWAS. In empirically motivated simulations, the sCCA+ACAT approach yielded the highest power to detect a gene associated with phenotype, even when expression in the causal tissue was not directly measured, while controlling the Type I error when there is no association between gene expression and phenotype. For example, when gene expression explains 2% of the variability in outcome, and the GWAS sample size is 20,000, the average power difference between the ACAT combined test of sCCA features and single-tissue, versus single-tissue combined with Generalized Berk-Jones (GBJ) method, single-tissue combined with S-MultiXcan or summarizing cross-tissue expression patterns using Principal Component Analysis (PCA) approaches was 5%, 8%, and 38%, respectively. The gain in power is likely due to sCCA cross-tissue features being more likely to be detectably heritable. When applied to publicly available summary statistics from 10 complex traits, the sCCA+ACAT test was able to increase the number of testable genes and identify on average an additional 400 additional gene-trait associations that single-trait TWAS

53 missed. Our results suggest that aggregating eQTL data across multiple tissues using sCCA can
54 improve the sensitivity of TWAS while controlling for the false positive rate.
55

56 ## Author summary

57 Transcriptome-wide association studies (TWAS) can improve the statistical power of genetic
58 association studies by leveraging the relationship between genetically predicted transcript
59 expression levels and an outcome. We propose a new TWAS pipeline that integrates data on the
60 genetic regulation of expression levels across multiple tissues. We generate cross-tissue
61 expression features using sparse canonical correlation analysis and then combine evidence for
62 expression-outcome association across cross- and single-tissue features using the aggregate
63 Cauchy association test. We show that this approach has substantially higher power than
64 traditional single-tissue TWAS methods. Application of these methods to publicly available
65 summary statistics for ten complex traits also identifies associations missed by single-tissue
66 methods.
67

Introduction

Genome-wide association studies (GWASs) have successfully identified thousands of associations between single-nucleotide polymorphisms (SNPs) and complex human phenotypes. Yet, the interpretation of these identified associations remains challenging, and several lines of evidence suggest that many additional associated loci remain to be identified [1, 2]. A recently proposed approach transcriptome-wide association study (TWAS) [3, 4] identifies genetic associations by combining GWAS data with expression quantitative trait locus (eQTL) data. TWAS can be used both to identify new associations and prioritize candidate causal genes in GWAS-identified regions [5]. TWAS integrates gene expression with GWAS data using only genotype expression imputation from a gene expression model built from eQTLs, and then test for the association between imputed gene expression level and a phenotype of interest. The main strength of TWAS is that it can infer the association of imputed gene expression with the phenotype using only GWAS summary statistics data [3, 4]. TWAS can increase the statistical power by combining single-SNP association tests in a biologically motivated fashion and reducing the number of tests performed. The applications of TWAS have led to novel insights into the genetic basis for several phenotype and diseases [6].

Despite the successes of TWAS, the approach has multiple limitations [7]. First, the most relevant tissue for many human diseases and phenotypes remains unclear, and the eQTL data for these relevant tissues are usually challenging to access in large samples. The choice of the most relevant tissue-specific eQTL sample for building gene expression prediction model in TWAS remains largely ad-hoc. Two commonly adopted approaches are: (1) using the largest eQTL sample accessible (usually whole blood [3]), or (2) using the most relevant tissue based on

previous knowledge and experience [6, 8]. Second, the power of TWAS is mainly bounded by the sample size of eQTL data; power of TWAS increases dramatically with the eQTL sample size, approaching an empirical maximum when eQTL sample size is close to 1,000 [3]. However, most available eQTL data sets have a sample size substantially smaller than 1,000. For example, Genotype-Tissue Expression(GTEx) project [9, 10] have generated matched genotype and expression data for 44 human tissues, but with sample size for each tissue varying from only 70 to 361. Researchers do not always know which tissue to use, and sometimes the sample size for the tissue that they prefer to use is too small to have enough power.

Recent work in gene regulation patterns across tissues suggests that local gene expression regulation is often shared across tissues [9-11]. Thus, combining eQTL data across multiple tissues can improve the power of TWAS, by increasing the effective eQTL sample size or increasing the likelihood that the causal tissue (or a close proxy) is included in the eQTL training data. Two previously proposed approaches, UTMOST [12] and S-MultiXcan [13], have shown the advantage of a multi-tissue TWAS approach. However, these two approaches still conduct the TWAS test with single-tissue TWAS weights first, and then combine multiple single-tissue associations into a single powerful metric to quantify. UTMOST uses a generalized Berk-Jones (GBJ) test, which is a set-based method [12]. S-MultiXcan proposes a combined chi-square test that uses principal components from the tissue-specific genetically predicted expression values to integrate univariate S-PrediXcan results [13]. We refer to these two approaches as single-tissue based cross-tissue TWAS approach. We propose to leverage the correlated gene expression pattern across tissues in the eQTL dataset directly to build more stable and representative cross-tissue gene expression features using sparse canonical correlation analysis (sCCA) [14], and thus

114 improve the gene expression prediction model for TWAS. The potential advantage of sCCA is
115 that it can capture any genetic contribution to gene expression that is shared across multiple
116 tissues. Because sCCA maximizes the correlation between a linear combination of tissue-specific
117 expression values and linear combination of cis-genotypes, sCCA features are more likely to be
118 detectably heritable than cross-tissue features constructed using principal components analysis
119 (PCA), which constructs linear combinations to capture total (genetic plus non-genetic)
120 expression variance [14]. In addition, we also propose an omnibus test that combines the single
121 tissue TWAS test results with the sCCA-TWAS test results using the aggregate Cauchy
122 association test (ACAT). ACAT is a computationally efficient P-value combination method for
123 boosting the power in sequencing study, and has proved to be powerful for detecting a sparse
124 signal [15].
125
126 Specifically, we propose a novel four-step pipeline to perform multi-tissue TWAS: 1. generate
127 sparse canonical correlation analysis (sCCA) [14] -based cross-tissue features (sCCA-features)
128 integrating eQTL data across multiple tissues; 2. fit TWAS weights for these sCCA-features as
129 well as single tissue-specific gene expression [3, 4]; 3. perform TWAS with weights built from
130 sCCA-features and singe tissue gene expression [3, 4]; 4. combine the test results of sCCA
131 TWAS results and single tissue TWAS results using the aggregated Cauchy association test
132 (ACAT) [15]. We use extensive simulations to compare this approach with four other cross-
133 tissue approaches, including: 1. performing TWAS on single most relevant tissue, 2. performing
134 TWAS on all single tissues available and combining the test results via Bonferroni or generalized
135 Berk-Jones (GBJ) test [16]; 3. using Principal Components Analysis (PCA) to create cross-tissue
136 features; and 4. the recently proposed S-MultiXcan approach [13].

Through simulations we show that sCCA-features identify a larger number of cis-heritable transcripts than single tissue and PCA-features, and the combined test substantially improves statistical power. Importantly, all approaches successfully control the type I error rate. We also show by simulations that the power of our combined test compares favorably to other approaches despite using incomplete gene expression matrix for all individuals and all tissues thus requiring imputation, as is often the case for multi-tissue gene expression dataset like GTEx [9, 10].

We applied our four-step approach to eQTL data from GTEx and 10 sets of publicly available GWAS summary statistics data. We built sCCA-features on an expression matrix including 134 individuals with data in 22 tissues. The sCCA-TWAS results were then compared with the single-tissue based TWAS results available on TWAS HUB (http://twas-hub.org). sCCA-TWAS was able to increase the number of testable genes by 81% and double the number of identified gene-phenotype associations.

Results

Methods Overview

Our proposed method entails four steps: the feature generating step, weight building step, TWAS step, and tests combining step (Fig. 1). In the feature generating step, we considered three approaches to build TWAS weights using eQTL data from multiple tissues. The first approach builds gene expression prediction weights in each tissue one at a time. The second approach, which we call the PCA-TWAS approach, first performs PCA on the gene expression matrix to create the top PCs (we restricted to top 3 PCs in this work). These PCs are then used as new gene

expression feature to build the gene expression weights and perform TWAS. In the final approach, sCCA-TWAS, we propose to use sCCA to build cross-tissue gene expression features as the weighted average of gene expression across multiple tissues (see Methods). These weighted averages maximize the correlation between the weighted average of gene expressions across tissues and *cis*-genotypes (within 500kb of the gene boundary). In the weight building step, we build TWAS weights for each of these multi-tissue features by regressing the feature on cis-SNPs in the gene's window. In the second step of TWAS, we perform tests for association using these set of weights (for each single-tissue or multi-tissue feature) separately. Finally, in the tests combining step, the single-tissue TWAS tests are combined using a Bonferroni multiple testing adjustment, the Generalized Berk-Jones (GBJ) procedure, or S-MultiXcan [13, 16] (see Methods for more details). We also propose a combined test of single-tissue test and sCCA cross-tissue test by combining the test results with ACAT [15].

We compared the performance of sCCA based cross-tissue TWAS with single tissue based cross-tissue TWAS approaches (Bonferroni, GBJ, S-MultiXcan) and PCA based cross-tissue TWAS through 2,000 simulations based on GTEx data. We conducted the simulations varying gene expression heritability, genetic correlation in expression across tissues, the proportion of tissues correlated with the causal tissue, the scale of non-centrality parameters in the GWAS z-score distribution (to model GWAS sample size), and whether gene expression from the underlying causal tissue is observed (i.e. not included in model training) or not.

sCCA improves statistical power to detect heritable gene expression

The first step of the TWAS approaches we consider tests the cis-heritability of each gene expression feature; the features that demonstrate significant heritability are only analyzed further. Fig. 2 compares the power of this heritability test for single-tissue, PC and sCCA expression features in the scenario where half of the tissues are correlated with the causal tissue, and the causal tissue is not observed. The relative performance of these features is very similar in the other scenarios (S1 and S2 Figs). The power of detecting heritable genes at a set alpha level increases as the correlation between correlated tissue and causal tissue or the heritability for gene expression in causal tissue increases. On average, the sCCA-features have a consistently higher chance of being heritable: they were 2.78x and 3.72x more likely to be heritable compared to the single tissue-based features and PCA based features.

The power of heritability test for PCA based cross-tissue TWAS is generally low, and the PC that captures the genetic signal best varies across scenarios (S3 Fig). The PCs that explain more of the variance in gene expression are not necessarily more heritable. Sometimes the second or third PC is heritable, but the first PC is not. We also observed that the chance of the PCA based feature to be heritable decreased with as the correlation between the genetic effect of the correlated tissue and the causal tissue increased. This may occur because non-genetic sources of correlation in expression across tissues outweigh genetic sources when the genetic contributions to expression are highly correlated. In this setting, the top PCs often do not capture the genetic effects.

Because sCCA features are constructed by maximizing the correlation between gene expression and genotype, the Type I error rate for the cis-heritability test can be inflated due to overfitting.

In fact, we did observe an inflated Type I error rate for heritability test under null for sCCA (S4 Fig). Considering individual features, the sCCA-feature1 had the highest Type I error rate at 0.43, while PC-feature1 had a slightly inflated type I error rate at 0.06 and the single tissue features maintained the Type I error rate at 0.05 level. But when we account for overall testing of 3 sCCA features, 3 PCS features and 22 single tissue features, the Type I error rate for at least one single tissue being significantly heritable at 0.05 level was 0.65 which is similar to the observed Type I error rate for at least one of the sCCA features being heritable. We note that standard TWAS pipelines typically do not adjust for the number of tissue features tested at the heritability stage. Most importantly, even though the cis-heritability test had an inflated rate of Type I error, the final Type I error rate for the sCCA-TWAS while testing for an association between predicted expression and phenotype was still well controlled (S5 Fig).

sCCA-features increase power of cross-tissue TWAS

Next, we compare the power of various approaches to multi-tissue TWAS to detect gene-trait associations via simulation. We simulated genotype and expression data using linkage disequilibrium (LD) and expression correlation information from GTEx. We set the gene expression in one tissue to be causal for the phenotype and varied the variance explained by genotype for the causal tissue, number of tissues with gene expression correlated with the causal tissue and the corresponding correlation (see Methods for more details). All methods control the Type I error when expression is not associated with the outcome (S5 Fig). In simulations, we varied the correlation between the casual and correlated tissue, the proportion of other tissues correlated with the casual tissue, whether the test results from the causal tissue was observed or not, and the proportion of gene expression variation explained by genotype in the casual tissue

(see Methods for details). In the simulation scenarios that we considered—all of which involved some correlation between the genetic contribution to gene expression in the causal tissue and at least one other tissue—we observed that the relative performance of different methods did not change as a function of the genetic correlation between the casual tissue and the correlated tissues, or the proportion of all tissues correlated with the casual tissue, or whether the causal tissue was analyzed (S1 and S2 Figs, S1-3 Tables).

We considered three sets of methods: (1) single tissue TWAS based approaches, which perform the single tissue based TWAS and either account for multiple testing using Bonferroni or GBJ corrections, or combine the test results using S-MultiXcan ; (2) tests based on cross-tissue features (using PCA or sCCA to build cross-tissue features); and (3) combined test of both single-tissue based methods and cross-tissue feature based methods, using either Bonferroni or ACAT to adjust for multiple testing [15].

First, for the single tissue-based approaches, GBJ and S-MultiXcan had either similar power or GBJ had slightly higher power than S-MultiXcan. For example, when gene expression explains 2% of the variability in outcome and the GWAS sample size is 20,000, the average power of single-tissue test combined with GBJ and single-tissue combined with S-MultiXcan was 0.34, and 0.29, respectively. Second, for the approaches using cross-tissue features, sCCA yielded a substantially higher power than PCA under all scenarios (the average power is 0.26 for sCCA and $<10^{-4}$ for PCA, S1-3 Tables). Third, for approaches to combine sCCA-TWAS and single tissue TWAS test results, combining sCCA-TWAS and single tissue TWAS test results with

ACAT [15] yielded 1.37 times greater power than combining them with Bonferroni (the average power is 0.38 for ACAT and 0.37 for Bonferroni, S1-3 Tables).

Finally, we compared single-tissue, cross-tissue, and combined single- and cross-tissue approaches. For simplicity, we only present comparisons between single-tissue based tests using GBJ to combine evidence across tissues, cross-tissue feature based approach with sCCA-features, and combined test of single-tissue based approach and sCCA-feature with ACAT.

Under the alternative, when gene expression has local genetic effects and gene expression is associated with the trait, the combined test of sCCA-features and single tissue-features using ACAT had the greatest power to detect a gene associated with the outcome, even when expression in the causal tissue was not directly measured (Fig 3). For example, when gene expression explains 2% of the variability in outcome and the GWAS sample size is 20,000, the average power for the ACAT [15] combined test of sCCA features and single-tissue test, sCCA-TWAS and single-tissue tests combined with GBJ was 0.38, 0.23 and 0.34, respectively (S1-3 Tables). The gain in power is likely because sCCA cross-tissue features are more likely to be significantly heritable, and thus increase the number of testable genes. This is particularly relevant for genes with low heritability: for such a gene, sCCA-TWAS has superior power (Fig 3 left panels). On the other hand, for highly heritability genes, single-tissue-based tests have better power than sCCA features. The combined test using both sCCA-TWAS and single-tissue TWAS results thus has superior power in both low- and high-heritability settings. Of note, the gain in power due to combining tests that perform well in different settings can be offset by the potential increased multiple testing burden. Fig. 3 presents power comparisons under the scenario where

273 half of the tissues are correlated with the causal tissue and the causal tissue is not observed
274 (power under other scenarios are reported in S1-3 Tables).
275
276 **sCCA-features provide insight into tissues where gene expression is associated with**
277 **outcome**
278 Although our primary motivation for combining multiple tissues when building expression
279 weights is to increase the power of TWAS, since sCCA performs feature selection on the tissues
280 as well as the cis-SNPs, it has the potential to suggest which tissues may be responsible for an
281 identified TWAS association.
282
283 Fig. 4 shows the sensitivity and specificity for the first sCCA component placing non-zero
284 weight on the causal tissue (if included in the expression panel), or a tissue whose genetic
285 contribution is correlated with that of the causal tissue. The sensitivity of the first sCCA
286 component putting a non-zero weight on a causal or correlated tissue increases with the gene
287 expression h_g^2 and the correlation between the causal tissue and the correlated tissues. Under our
288 simulation assumption, the specificity of the first sCCA component is consistently high, which
289 indicates that when combining gene expression across tissues with sCCA, it is less likely that
290 non-relevant tissues would be included in the top sCCA expression feature. Thus, sCCA can
291 effectively increase sample size while excluding noise. The tissues with non-zero weights in
292 sCCA have a higher probability of being causal.
293
294
295

296 **sCCA performance is stable to missing data imputation in the expression data**

297 The sCCA-TWAS approach requires a complete gene expression matrix: every individual used
298 to train the sCCA features must have expression data from every tissue included in the analysis.
299 However, this is typically not true for multi-tissue gene expression datasets like GTEx [9], where
300 not all donors have samples or expression data from all tissues. A complete case analysis can
301 greatly reduce the sample size available to train sCCA features. On the other hand, imputing
302 missing expression data may induce measurement error or bias. We evaluate the impact of
303 imputing missing expression data via simulation. We simulate complete gene expression and
304 genotype data based on correlations in gene expression observed across GTEx; we then perform
305 single-tissue based TWAS using weights trained in the complete data set. For sCCA and PCA
306 based approaches, we mask the expression data matrix randomly based on the missing proportion
307 pattern for each tissue in GTEx, then impute the missing expression data with MICE [17], using
308 the "predictive mean matching" method. We then perform sCCA-TWAS or PCA-TWAS on the
309 imputed gene expression dataset. sCCA-TWAS applied to imputed expression data still correctly
310 controlled the Type I error rate. Although the power for sCCA-TWAS was lower when using
311 imputed expression data (across all scenario decreased from 0.38 to 0.21), the sCCA-TWAS still
312 provide valuable information when the genetic signal for gene expression is weak.
313
314 **Real-data application**
315
316 **Applying sCCA to GTEx data increased the number of testable genes**
317 We applied the sCCA-TWAS approach based on top 3 sCCA-features to integrate GTEx data
318 (version 6) and GWAS summary statistics data for 10 complex traits using the same *cis*-

heritability filter as TWAS HUB, and compared the results with single tissue based TWAS results on TWAS HUB [18]. The phenotype information is included in Table 1 and the tissue expression dataset information is included in S4 Table. We choose to include top 3 sCCA-features as we observed in the simulation study that the gain in power due to including more features was negligible (S7 Fig). With sCCA cross-tissue features, we increased the number of testable genes to 21,740 compared to 12,027 (all GTEx tissues on TWAS HUB) and 18,954 (all panels on TWAS HUB). Among the genes that we could test using sCCA-TWAS, 10,649 genes were not testable in any of the other single-tissue panels available (that is, they did not pass the filtering criterion for cis-heritability or prediction strength set by TWAS-HUB). At the same time, with sCCA-features that combine expression profiles across multiple tissues, we reduced the multiple testing burden from 84,964 (GTEx tissues) and 157,316 (all panels in TWAS HUB) to 38,620. When the *cis*-genetic regulation is shared across multiple tissues, sCCA-TWAS reduces the redundnacy in expression features tested. Using sCCA-TWAS as opposed to single-tissue TWAS increased the number of testable genes relative to single GTEx tissues by 81% and reduced the multiple testing burden by 55%; realtive to all panels in TWAS HUB we increased the number of testable genes by 56% and reduced the multiple testing burden by 75% [18].

Table 1. Summary of data application results

Trait	GWAS sample size	Number of significant loci	Total number of significant genes in TWAS HUB	Number of signifiacnt genes in GTEx panel	Number of significant genes by sCCA-TWAS	Number of significant genes by ACAT	Reference
Alzheimer's Disease (including proxy)	388324	17	70	34	44	51	Marioni et al. 2018 Nat Comms
Breast Cancer	228951	79	353	162	260	278	Michailidou 2017 Nature
Coronary Artery Disease (CAD)	56422	11	17	11	8	11	Schunkert et al. 2011 Nat Genet

Type 2 Diabetes (T2D) (2012)	48761	5	5	4	2	4	Morris et al. 2012 Nat Genet
Schizophrenia (2018)	65967	38	167	58	138	90	Ruderfer et al. 2018
BMI	457824	255	1592	782	1132	1246	UKBB Loh et al. 2018 Nat Genet
Height	458303	423	5709	2891	4080	5112	UKBB Loh et al. 2018 Nat Genet
Smoking Status	457683	59	233	106	166	164	UKBB Loh et al. 2018 Nat Genet

Chronotype (morning person)	410520	69	202	82	145	140	UKBB Loh et al. 2018 Nat Genet
Tanning	449984	65	382	197	274	325	UKBB Loh et al. 2018 Nat Genet

Real-data application detects novel predicted-expression to phenotype associations

The sCCA-ACAT and sCCA-feature TWAS detected additional associations between predicted gene expression and phenotype for the 10 GWAS traits we considered (Table 1). The single-tissue TWAS tests with GTEx weights identified 4,327 phenotype gene expression associations. In aggregate, sCCA-TWAS identified 4,400 additional associations for 10 phenotypes compared to single tissue GTEx TWAS, and the sCCA-ACAT combined test identified 3,277 additional associations compared to single tissue GTEx TWAS (Figs 5 and 6). The two phenotypes with the largest number of associated genes identified are height and BMI, which are both highly polygenetic. To further contrast the significant associations identified, we considered the overlap between the associations identified with sCCA cross-tissue TWAS and single tissue TWAS for each phenotype. On an average, 18% of the gene-phenotype associations were identified by both single-tissue TWAS and sCCA TWAS, 49% gene-phenotype associations were only identified by sCCA-TWAS, and 34% signals were only detected by single tissue TWAS (Fig 7).

ACAT served as a good combination method for single tissue and sCCA TWAS. Out of the total number of associations identified by either single-tissue TWAS, sCCA-TWAS, or sCCA-ACAT, 85% were significant in the sCCA-ACAT combined test. Among the gene-trait associations that were identified using the sCCA-ACAT approach, 41% were also identified by the single tissue approach but not the sCCA approach; 36% were also identified using the sCCA approach but not the single-tissue approach; 23% were identified using all three approaches; and 1% were identified using only the sCCA-ACAT combined approach. Fig 8 shows the breakdown in the testing performance by phenotype.

Direct comparison of the absolute z-scores from all the single tissue TWAS and sCCA-TWAS shows a correlation of 0.86. The sCCA-TWAS absolute z-score is slightly greater than the median value of single tissue absolute z-score of the same gene from multiple tissues (S6 Fig).

Discussion

We have proposed a novel approach (sCCA-TWAS) to constructing cross-tissue expression features using sparse canonical correlation analysis to boost the power of transcriptome-wide association studies. Through simulations we show that if the genetic component of gene expression in the causal tissue is correlated with the genetic contribution of expression in other tissues, then sCCA-TWAS has greater power than the approaches that use TWAS test statistics based on single-tissue features, including simply applying Bonferroni correction for the number of tissues tested or combining single-tissue tests using a GBJ procedure or S-MultiXcan [13, 16]. We have also proposed to combine sCCA-TWAS tests with single-tissue TWAS tests implementing the aggregate Cauchy association test (sCCA+ACAT). sCCA+ACAT achieves optimal or near-optimal power among the procedures considered both when the causal tissue is genetically correlated with other tissues and when it is not, suggesting that the sCCA+ACAT is a useful method when the genetic architecture of tissue-specific expression and its relationship to outcome is unknown. This increase in power is due in part to the greater number of genes with significantly heritable sCCA features relative to single-tissue features. sCCA-TWAS also greatly improved power relative to another cross-tissue technique using PCA to create cross-tissue features, as the leading principal components often capture non-genetic sources of covariation in gene expression (a general drawback to cross-trait association analysis using PCA[19]). Moreover, the tissue-wise loadings from sCCA factors associated with outcome may provide

some guidance to which tissues are causally related to the outcome (or genetically correlated with the unmeasured causal tissue).

sCCA- and sCCA+ACAT- TWAS can be useful in a situation where eQTL data on germline genetic variation and expression in multiple tissues or cell-types are available on the same set of individuals. sCCA-TWAS cannot be directly applied when eQTL data on different tissues are available on different, non-overlapping samples. When both a multi-tissue reference panel (such as GTEx) and additional large single-tissue reference panels are available, sCCA+ACAT can make use of both the cross-tissue features from the multi-tissue panel and the independent single-tissue panels. Finally, inferring the causal tissue from a set of cross-tissue or single-tissue TWAS results remains an important open question. Although the tissue weights from the sCCA features may provide some clues, further work is needed to develop principled sensitive and specific methods for identifying candidate causal tissues.

Methods

sCCA

Suppose that we have n observations on $p_1 + p_2$ variables, and the variables are naturally partitioned into two groups of p_1 and p_2 variables, respectively. Let $\mathbf{G} \in R(n \times p_1)$ correspond to the first set of variables, and let $\mathbf{X} \in R(n \times p_2)$ correspond to the second set of variables. Assume that the columns of $\mathbf{G}$ and $\mathbf{X}$ have been standardized to have mean zero and standard deviation one. In our setting, $\mathbf{G}$ is a matrix of standardized genotypes with SNPs corresponding

to the columns and $\mathbf{X}$ is a matrix of tissue-specific gene expression values with genes corresponding to the columns.

Standard CCA seeks $\boldsymbol{\omega_1} \in R(p_1)$ and $\boldsymbol{\omega_2} \in R(p_2)$ that maximize correlation between $\mathbf{G\omega_1}$ and $\mathbf{X\omega_2}$ [14], that is:

$$\text{maximize}_{\omega_1,\omega_2}\ \boldsymbol{\omega_1^T G^T X \omega_2} \text{ subject to } \boldsymbol{\omega_1^T G^T G \omega = \omega_2^T G \omega_1 = \omega_2^T X^T X \omega_2 = 1}$$

However, CCA is not appropriate when $p_1, p_2 \approx n$ or $p_1, p_2 >> n$. Witten et al. [14] proposed sparse CCA, a penalized version of CCA, by adding L1 and L2 penalization in the previous optimization problem [14] as:

$$\text{maximize}_{\omega_1,\omega_2}\ \boldsymbol{\omega_1^T G^T X \omega_2} \text{ subject to } \boldsymbol{\omega_1^T G^T G \omega} \leq 1, \boldsymbol{\omega_2^T X^T X \omega_2} \leq 1, \text{and } P_1(\boldsymbol{\omega_1}) \leq c_1, P_2(\boldsymbol{\omega_2}) \leq c_2$$

Using the identity matrix $\mathbf{I}$ as a substitute for $\mathbf{X_1^T X_1}$ and $\mathbf{X_2^T X_2}$ gives what can be termed as "diagonal penalized CCA", and the optimization problem can be re-formulated as:

$$\text{maximize}_{\omega_1,\omega_2}\ \boldsymbol{\omega_1^T G^T X \omega_2} \text{ subject to } ||\boldsymbol{\omega_1}||^2 \leq 1, ||\boldsymbol{\omega_2}||^2 \leq 1, ||\boldsymbol{\omega_1}||_1 \leq c_1, ||\boldsymbol{\omega_2}||_1 \leq c_2$$

For a small c_1 and c_2, this results in $\boldsymbol{\omega_1}$ and $\boldsymbol{\omega_2}$ to be sparse, i.e., many of the elements of $\boldsymbol{\omega_1}$ and $\boldsymbol{\omega_2}$ will be exactly equal to zero. Witten et al. proposed to solve this maximization problem by initializing $\boldsymbol{\omega_2}$ to belong to R^q, and then iteratively maximizing $\boldsymbol{\omega_1^T G^T X \omega_2}$ subject to L_1 and L_2 constraints for $\boldsymbol{\omega_1}$ and $\boldsymbol{\omega_2}$ in turn [14]. $\boldsymbol{\omega_2}$ was initialized to have L_2-norm 1 and was

suggested to use the first right singular vector of $\mathbf{X}$ as the initial value. c_1 and c_2 can be chosen by cross-validation, where c_1 and c_2 are chosen using a grid search to maximize $\mathrm{cor}(\mathbf{G\omega_1}, \mathbf{X\omega_2})$ (across the cross-validation folds). It can be shown that a maximum of $\min(p,q)$ orthogonal $\boldsymbol{\omega_1}$, $\boldsymbol{\omega_2}$ vectors can be generated by repeatedly applying this algorithm to the new correlation matrix $\boldsymbol{G^T X}$ after regressing out the previous canonical component [14].

TWAS

The TWAS pipeline consists of three steps: first, identifying gene expression features that have positive cis-heritability; second, building a linear predictor for each cis-heritable gene feature; and third, constructing the TWAS test statistic combining the prediction weights and summary Z-scores from a trait GWAS.

We computed the p-values for testing *cis*-h_g^2=0 using a likelihood ratio test implemented in GCTA that compares a model with a local random genetic effect to a model without a genetic effect [20]. We included all SNPs that fall within 500 kb of the transcription start and stop sites of a gene. We removed the genes that failed the heritability test from the set of candidate genes, and only the genes with a significant heritability were included in the subsequent prediction model construction.

We then used Elastic Net penalized regression implemented in the R package glmnet [21] to construct linear genetic predictors of gene expression features $\mathbf{W}$ based on all the *cis* SNPs in the eQTL reference panel (500 base-pair window surrounding the transcription start and stop sites). We applied 5-fold cross-validation to choose the elastic net penalty parameters.

We calculated the TWAS test statistic as $Z_{TWAS} = \boldsymbol{wZ}/(\boldsymbol{w\Sigma_{s,s}w}')^{\mathbf{1/2}}$, where $\mathbf{Z}$ is a vector of standardized effect sizes of SNPs for a trait in the *cis* region of a given gene (Wald z-scores), and $\boldsymbol{w} = (w_1\, w_2\, w_3 \ldots\ldots w_j)$ is a vector of prediction weights for the expression feature of the gene being tested, and $\boldsymbol{\Sigma_{s,s}}$ is the LD matrix of the cis SNPs estimated from the 1000 Genomes Project as the LD reference panel. Under null hypothesis that there is no association between the gene expression feature and phenotype, Z_{TWAS} should follow a normal distribution with mean zero and variance one.

sCCA-TWAS

Consider a gene expression array of a certain gene for n individuals and p_2 tissues $\mathbf{X_{nxp_2}}$, and the genotype data $\mathbf{G_{nxp_1}}$ for the same set of individuals at p_1 cis-SNPs. Assume that the columns of $\mathbf{X_{nxp_2}}$, and $\mathbf{G_{nxp_1}}$ have been standardized to have mean zero and variance one.

We apply sCCA (described above) and extract the first three pairs of canonical vectors: $(\boldsymbol{\omega}_1^{(1)}, \boldsymbol{\omega}_2^{(1)})$, $(\boldsymbol{\omega}_1^{(2)}, \boldsymbol{\omega}_2^{(2)})$ and $(\boldsymbol{\omega}_1^{(3)}, \boldsymbol{\omega}_2^{(3)})$. We define three sCCA features as $\mathbf{X}\,\boldsymbol{\omega}_{\mathbf{2}}^{(\mathbf{1})}$, $\mathbf{X}\,\boldsymbol{\omega}_{\mathbf{2}}^{(\mathbf{1})}$ and $\mathbf{X}\,\boldsymbol{\omega}_{\mathbf{2}}^{(\mathbf{3})}$. Then we treat the three sCCA-features as three repeated measure of gene expression across tissue and apply TWAS procedure to them, record the p-value for heritability and z-score of these three features. We account for testing multiple sCCA features per gene via Bonferroni's correction, including only the tests where the sCCA-feature passed the heritability test. We decided to include at most 3 sCCA features, because in simulations, the power gain from including more features appears to be small (Fig. S7).

Single tissue test based cross-tissue TWAS

As a comparison, we also considered single-tissue test based cross-tissue TWAS, where we perform TWAS on the gene expression in each tissue, record the z-scores and p-values for heritability test, respectively. We account for testing multiple tissues for each gene via i) a Bonferroni multiple testing correction or ii) a generalized Berk-Jones (GBJ) test with single-tissue association statistics Z and their covariance matrix $\boldsymbol{\Sigma}$ as inputs [16]. We estimate $\boldsymbol{\Sigma}$ as $\boldsymbol{W}\boldsymbol{\Sigma}_{s,s}\boldsymbol{W}'$, where $\boldsymbol{W}_{qxp}$ is a matrix with the expression weights for each tissue in each row and each SNP [12].

Combined test with sCCA-features and single-tissue features

While sCCA can increase power when sample sizes in individual tissues are small and the genetic contribution to expression is shared across tissues, a single-tissue based approach may be more powerful when the genetic contribution to expression in the causal tissue is uncorrelated with genetic contribution to expression in other tissues. Thus, a combined test for sCCA-features and single-tissue features can have a better average power across a range of scenarios. We therefore consider approaches that combine sCCA and single-tissue expression features, accounting for testing multiple features per gene using a Bonferroni correction, the GBJ test [16], or the ACAT [15]. The GBJ test is a set-based test proposed for GWAS setting, which extended the Berk-Jones (BJ) statistics by accounting for correlation among tests [16]. ACAT is a fast p-value combination method that uses Cauchy distribution to approximate the distribution of a weighted sum of transformed p-values. ACAT has been shown to work well in the context of genetics research, mainly because it does not require the estimation of correlation strauctur among the combined p-values.

494 **PCA based cross-tissue TWAS**

495 We also considered aggregating across tissue signal through Principal Component Analysis
496 (PCA). We first applied PCA to the gene expression matrix $\mathbf{X_{nxq}}$, then used the top 3 principal
497 Components (PCs) as new feature for TWAS. We accounted for testing multiple PCs for each
498 gene by Bonferroni adjustment, including only the tests where the PCs passed the heritability
499 test.

500

501 **S-MultiXcan**

502 Summary-MultiXcan (S-MultiXcan) is another single-tissue based approach for generating
503 multi-tissue gene expression, and draw phenotype associations inference. It utilizes the LD
504 information from a reference panel to integrate univariate S-PrediXcan results. It consists of the
505 following steps: (1) computation of single tissue association test statistics $\hat{\boldsymbol{Z}}$ with S-PrediXcan
506 **[**2**]**; (2) estimation of the correlation in tissue-specific predicted gene expression levels using the
507 LD information from a reference panel (typically GTEx or 1000 Genomes); (3) discarding
508 components of smallest variation from the matrix of correlations in genetically-predicted tissue-
509 specific gene expression levels to avert collinearity and numerical problems (singular value
510 decomposition, analogous to PC analysis in individual-level data). (4) estimation of multi-tissue
511 test statistics from the univariate (single-tissue) results with the help of expression correlation.

512

513 The aggregate S-MultiXcan test statistic is then calculated as $\hat{\boldsymbol{Z}}^{\boldsymbol{T}} Cor(\boldsymbol{X})^{+} \hat{\boldsymbol{Z}} \sim \chi^2_k$, where Cor
514 $(\boldsymbol{X})^{+}$ is the pseudo-inverse of a SVD-regularized version of the correlation matrix of $\boldsymbol{X}$, and k
515 the number of components surviving the SVD pseudo-inverse (the regularized version of the

516 correlation matrix is formed by decomposing the correlation matrix into its principal components
517 and removing those eigenvectors corresponding to the eigenvalues $\frac{\lambda_{max}}{\lambda_i} < 30$).

518

519 **Data simulation settings**

520 We simulated genotype and expression data using linkage disequilibrium (LD) and expression
521 correlation information from the Genotype-Tissue Expression project (GTEx) version 6 [9].
522 GTEx includes data from 449 donors across 44 tissues, with tissue-specific sample sizes ranging
523 from 70 to 361. We removed: (1) individuals with data available for less than 40% of the tissues
524 and (2) tissues where less than 30% of the individuals have data. This results in a 134 (n
525 ndividuals) by 22 (p_2 tissues) ordered expression matrix for each gene. We randomly sampled
526 400 genes in the data set, extracted the cis-SNPs within 500kb around the gene boundary
527 (number of cis-SNPs indicated by p_1) and the gene expression for these 134 individuals and 22
528 tissues, and imputed missing expression values with the column mean. We used this data set to
529 calculate the correlation among gene expression levels across tissues (Σ_X) and the LD structure
530 of cis-SNPs (Σ_G) for each of the 400 genes.

531

532 Individual-level data for a gene expression reference panel data were generated assuming that the
533 gene expression for a particular gene in tissue i is $\mathbf{X_i} = \mathbf{G}\boldsymbol{\beta_i} + \boldsymbol{\epsilon_i}$, where $\mathbf{G}$ is the local genotype
534 matrix, $\boldsymbol{\beta_i}$ is the weight for genotype on gene expression in tissue i, and the residuals $\boldsymbol{\epsilon_i}$ are
535 normally distributed, independent across individual but correlated across tissues. We generated
536 each row in the $n \times p_1$ genotype matrix $\mathbf{G}$ as MVN^{p_1} with mean zero and variance-covariance
537 matrix $\boldsymbol{\Sigma_G}$, the LD matrix calculated from the GTEx genotype data from the gene's cis region.
538 We randomly sampled one tissue to be causal and N_{corr} tissues to be genetically correlated with

the causal tissue. We selected 3% of the cis-SNPs to be causally related to gene expression in the causal tissue and sampled their weights for gene expression, β_{ij}^{causal} from normal distribution with mean zero and variance h_g^2; the remaining β_{ij}^{causal} for j not in the set of causal SNPs were set to 0. To reflect the genetic correlation ρ between the causal tissue and the N_{corr} genetically correlated tissues, the weight for the same SNPs in the correlated tissues were sampled as

$$\boldsymbol{\beta}_{\mathbf{correlated}} \sim \mathrm{MVN}^{\mathrm{Ncorr}\times 1}(\boldsymbol{\beta}_{\mathbf{causal}} \times \rho \times \mathbf{1}_{\mathbf{Ncorr}}, (1-\rho^2) \times h_g^2 \cdot \mathbf{I}_{\mathbf{Ncorr}})$$

This resulted in a p_1 by p_2 weight matrix for genotype on tissue-specific gene. Residual gene expression values were simulated as:

$$\mathbf{e} \sim \mathrm{MVN}^{n\times p}(\mathbf{0}, \mathbf{diag}(\boldsymbol{s_e}) \times \boldsymbol{\Sigma_X} \times \mathbf{diag}(\boldsymbol{s_e}))$$

where $\boldsymbol{s_e} = \sqrt{\mathrm{Var}(\mathbf{X}\boldsymbol{\beta}^{q\times p}) \times (\frac{1}{h^2 g} - 1)}$, so that the variance in gene expression explained by genotype in each tissue is h_g^2. We considered four scenarios, defined by combination of the proportion of tissues genetically correlated with the causal tissue and whether the causal tissue was observed in the analysis: all or half of the tissues were correlated with causal tissue; the causal tissue was or was not observed. We varied h_g^2 from 0.01 to 0.1, and the genetic correlation coefficient ρ between the causal and other tissues from 0.3 and 1.

Given the SNP-expression weights in a tissue and assuming that the trait under study Y has unit variance and the true mean of the trait is related to expression levels in the causal tissue via $E[Y] = r X_{causal}$, the cis-SNP GWAS z-scores for tissue i are distributed as $\mathbf{Z} \sim \mathrm{MVN}(\Sigma_G \times b \times \boldsymbol{\beta}^{\mathbf{i}}, \Sigma_G)$, where $b = \sqrt{N_{gwas} \times r^2}$. For each tissue, we randomly sampled the z-scores from this multivariate normal and set b to 0.00, 6.78, 11.18, 14.36, 17.07, 19.60, 22.13, 24.84, 28.02, 32.42 to achieve the theoretical power of 5%, 10%, ..., 90% at alpha level of 0.05. For example, when

r^2 equals 1% (i.e., variation in gene expression in the target tissue explains 1% of the variability in the trait), the GWAS sample size N_{gwas} ranges from 4,602 to 105,074. We repeated the whole procedure on 400 randomly selected genes. For each gene, we further replicated 5 times for a total of 2000 replicates. For each statistical test procedure (sCCA, PCA, s-MultiXcan, etc.), and for each replicate, there are three possible outcomes: A: the gene is not heritable [i.e., no sCCA feature is significantly heritable, or no PCA, or no single tissue, depending on the procedure]; B: the gene is heritable but not significantly associated with the trait (after accounting for multiple testing across heritable tissues/features); and C: the gene is heritable and significant. We calculate Type I error as B/(B+C) and power as C/2000.

Data application

We applied sCCA-TWS approach to GTEx and 10 real life-style, polygenic complex traits and diseases (Table 1): whether a morning person [22], smoking status [22], body mass index [22], height [22], hair color [22], schizophrenia [23, 24], type 2 diabetes [25], coronary artery disease [26], breast cancer [27] and Alzheimer's disease [12]. Before applying sCCA to the GTEx data (version?), we removed individuals with data available in less than 40% of the tissues. We also removed tissues where less than 30% of the donors have sample. This resulted in a 134 (n) individual by 22 (p_2) tissue expression matrix for each gene (list of tissues provided in S4 Table). We imputed the missing expression data using the predictive mean method in R package MICE [17]. We performed sCCA on the imputed gene expression and genotype data from GTEx, extracted the top 3 canonical vectors for gene expression for each gene, and built three sCCA-features for each of the gene. Then we adopted the standard TWAS pipeline with the sCCA

584 features, filtering out sCCA-features that failed to converge in GCTA or had a heritability test p-
585 value greater than 0.01. We built linear genetic weights with the rest sCCA-features using Lasso,
586 Elastic Net (eNet), and top eQTL models, and performed TWAS with the model of highest cross
587 validation R^2.
588

589 **Acknowledgements**

590 This work was supported by grants from the U.S. National Institutes of Health:
591 U01 CA194393, R01 HG009120, R01 CA22723, R35 CA197449 and U01
592 HG009088.

593
594
595

References

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal of human genetics. 2017;101(1):5-22. doi: 10.1016/j.ajhg.2017.06.005. PubMed PMID: 28686856; PubMed Central PMCID: PMCPMC5501872.
2. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat Genet. 2018;50(9):1318-26. doi: 10.1038/s41588-018-0193-x. PubMed PMID: 30104760.
3. Alexander G, Arthur K, Huwenbo S, Gaurav B, Wonil C, Brenda WJHP, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics. 2016;48(3). doi: 10.1038/ng.3506.
4. 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. Nat Genet. 2015;47(9):1091-8. Epub 2015/08/11. doi: 10.1038/ng.3367. PubMed PMID: 26258848; PubMed Central PMCID: PMCPMC4552594.
5. Mancuso N, Kichaev G, Shi H, Freund M, Gusev A, Pasaniuc B. Probabilistic fine-mapping of transcriptome-wide association studies. bioRxiv. 2018.
6. Wu L, Cox A, Zheng W. Identification of novel susceptibility loci and genes for breast cancer risk: A transcriptome-wide association study of 229,000 women of European descent. 2018.
7. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles D, Golan D, et al. Transcriptome-wide association studies: opportunities and challenges. bioRxiv. 2018:206961. doi: 10.1101/206961.
8. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. 2018;50(4):621-9. Epub 2018/04/11. doi: 10.1038/s41588-018-0081-4. PubMed PMID: 29632380; PubMed Central PMCID: PMCPMC5896795.
9. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60. Epub 2015/05/09. doi: 10.1126/science.1262110. PubMed PMID: 25954001; PubMed Central PMCID: PMCPMC4547484.
10. Erratum: Genetic effects on gene expression across human tissues. Nature. 2018;553(7689):530. Epub 2017/12/21. doi: 10.1038/nature25160. PubMed PMID: 29258290.
11. Liu X, Finucane HK, Gusev A, Bhatia G, Gazal S, O'Connor L, et al. Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. American journal of human genetics. 2017;100(4):605-16. Epub 2017/03/28. doi: 10.1016/j.ajhg.2017.03.002. PubMed PMID: 28343628; PubMed Central PMCID: PMCPMC5384099.
12. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. bioRxiv. 2018. doi: 10.1101/286013.
13. Barbeira AN, Pividori MD, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection.(Research Article)(Report). PLoS Genetics. 2019;15(1). PubMed PMID: BarbeiraAlvaroN.2019Iptf.

14. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics (Oxford, England). 2009;10(3):515-34. Epub 2009/04/21. doi: 10.1093/biostatistics/kxp008. PubMed PMID: 19377034; PubMed Central PMCID: PMCPMC2697346.
15. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. American journal of human genetics. 2019;104(3):410-21. doi: 10.1016/j.ajhg.2019.01.002. PubMed PMID: 30849328; PubMed Central PMCID: PMCPMC6407498.
16. Sun R, Hui S, Bader G, Lin X, Kraft P. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. bioRxiv. 2018. doi: 10.1101/361436.
17. van Buuren S, Groothuis-Oudshoorn CGM. mice: Multivariate Imputation by Chained Equations in R. Journal of statistical software. 2011;45(3):urn issn 1548--7660. PubMed PMID: vanBuurenStef2011mMIb.
18. Gusev A. TWAS HUB 2017. Available from: http://twas-hub.org/.
19. Aschard H, Vilhjalmsson BJ, Greliche N, Morange PE, Tregouet DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. American journal of human genetics. 2014;94(5):662-76. Epub 2014/04/22. doi: 10.1016/j.ajhg.2014.03.016. PubMed PMID: 24746957; PubMed Central PMCID: PMCPMC4067564.
20. 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. PubMed PMID: 21167468; PubMed Central PMCID: PMCPMC3014363.
21. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software. 2010;33(1):1-22. Epub 2010/09/03. PubMed PMID: 20808728; PubMed Central PMCID: PMCPMC2929880.
22. Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nature Genetics. 2016;48(7):811--6. PubMed PMID: LohPo-Ru2016Faal.
23. Cowell R, Lucas E, Molina J, Meador-Woodruff J, Kleinman J, McCullumsmith R, et al. The Expression of Developmentally-regulated PGC-1alpha-Dependent Genes is Reduced in the Cortex of Patients with Schizophrenia. Neuropsychopharmacology. 2014;39:S587--S. PubMed PMID: CowellR2014TEoD.
24. Ruderfer DM, Ripke S, McQuillin A, Boocock J, Stahl EA, Pavlides JMW, et al. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. Cell. 2018;173(7):1705--8674. PubMed PMID: RuderferDouglasM.2018GDoB.
25. Morris KV. The theory of RNA-mediated gene evolution. Epigenetics. 2015;10(1):1--5. PubMed PMID: MorrisKevinV2015TtoR.
26. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nature Genetics. 49(9):1385--91. PubMed PMID: NelsonChristopherPAabo.
27. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. 2017;551(7678). PubMed PMID: MichailidouKyriaki2017Aai6.

Supporting information

S1 Fig. Proportion of significant (p<0.05) heritability tests for different expression features when cis genetic variation is associated with expression in *all* tissues. ρ denotes the strength of the genetic correlation between expression in the causal tissue and tissues where expression is also associated with cis germline variation ("correlated tissues"). "Non-correlated tissues" are tissues where local germline variation is not associated with gene expression. Here expression in all of the tissues is genetically correlated with the causal tissue, and the causal tissue is not observed (performance in the causal tissue is included as a reference). PC1 is the first principal component of cross-tissue gene expression; sCCA-feature1 is the linear combination of tissue expression values from the first pair of sCCA canonical variables. h^2 denotes the proportion of expression variance in the causal tissue explained by cis genetic variation.

S2 Fig. Proportion of significant (p<0.05) heritability tests for different expression features when cis genetic variation is associated with expression in *some* tissues. ρ denotes the strength of the genetic correlation between expression in the causal tissue and tissues where expression is also associated with cis germline variation ("correlated tissues"). "Non-correlated tissues" are tissues where local germline variation is not associated with gene expression. Here expression in half of the tissues is genetically correlated with the causal tissue, and the causal tissue *is* observed. PC1 is the first principal component of cross-tissue gene expression; sCCA-feature1 is the linear combination of tissue expression values from the first pair of sCCA canonical variables. h^2 denotes the proportion of expression variance in the causal tissue explained by cis genetic variation.

S3 Fig. Proportion of significant (p<0.05) heritability tests for the top three principal components summarizing gene expression across features (*half* of the tissues are correlated with the causal tissue and causal tissue *not* observed). ρ denotes the strength of the genetic correlation between expression in the causal tissue and tissues where expression is also associated with cis germline variation. Half of the tissues are genetically correlated with the causal tissue, which is not observed. h^2 denotes the proportion of expression variance in the causal tissue explained by cis genetic variation.

S4 Fig. Type I error rate for cis-heritability tests. Proportion of simulations where local genetic variation was nominally statistically significantly associated with gene expression, in the scenario where no association was present. sCCA-Feature_1: testing only the leading sCCA expression feature at the α=0.05 level; PCA-feature_1: testing only the lead cross-tissue expression principal component at the α=0.05 level; All_PCA-features and All_sCCA-features: proportion of simulations where at least one of the top three PCA (resp. sCCA) features was significant at the α=0.05 level; All_single_tissue: proportion of simulations where at least one of the 22 single-tissue tests was significant at the α=0.05 level.

S5 Fig. Type I error rate for cross-tissue TWAS methods. Proportion of significant results under null average over all scenarios (Gene expression not associated with phenotype).

S6 Fig. Comparison of the absolute z-score for sCCA-TWAS and single tissue TWAS using weights calculated form GTEx data and GWAS summary statistics from 10 complex traits. The TWAS test statistics using sCCA feature 1 and all single tissue weights from Fusion are

732 plotted on the x-axis and y-axis respectively. The blue line is the fitted regression line and red
733 line is y=x.
734
735 **S7 Fig. Cumulative power for identify heritable gene when include sCCA feature 1 to**
736 **feature 3.** The Y axis indicate the cumulative power of detecting heritable genes when include
737 only sCCA feature 1, sCCA feature 1 and 2, and sCCA feature 1 to 3, average over all scenarios.
738
739 **S1 Table. Summary of simulation power when gene expression in other tissues *not***
740 **correlated with the causal tissue**
741
742
743 **S2 Table. Summary of simulation power when gene expression in *half* of the tissues**
744 **correlated with the causal tissue**
745
746 **S3 Table. Summary of simulation power when gene expression in *all* of the tissues**
747 **correlated with the causal tissue**
748
749 **S4 Table. Summary of GTEx expression data**
750
751

Figures

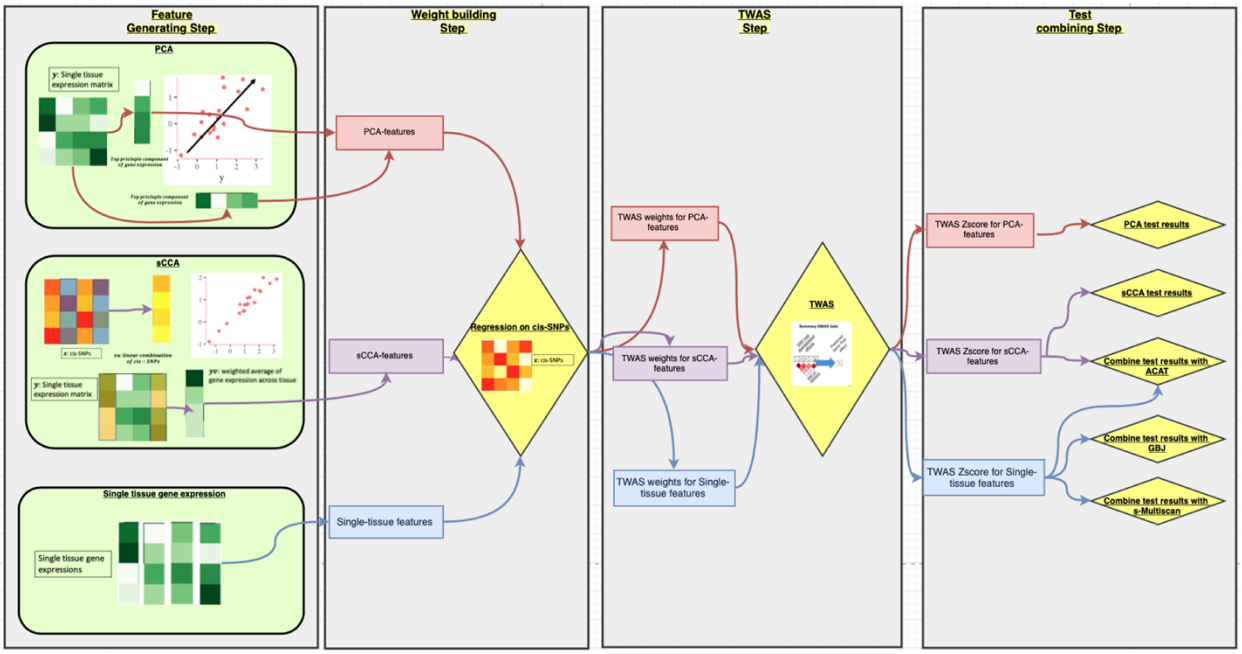


Fig 1. Methods overview. The single tissue based cross-tissue TWAS approach is shown in blue arrows, the PCA based cross-tissue TWAS approach is shown in red arrows, and the sCCA-TWAS approach is shown in purple arrows.

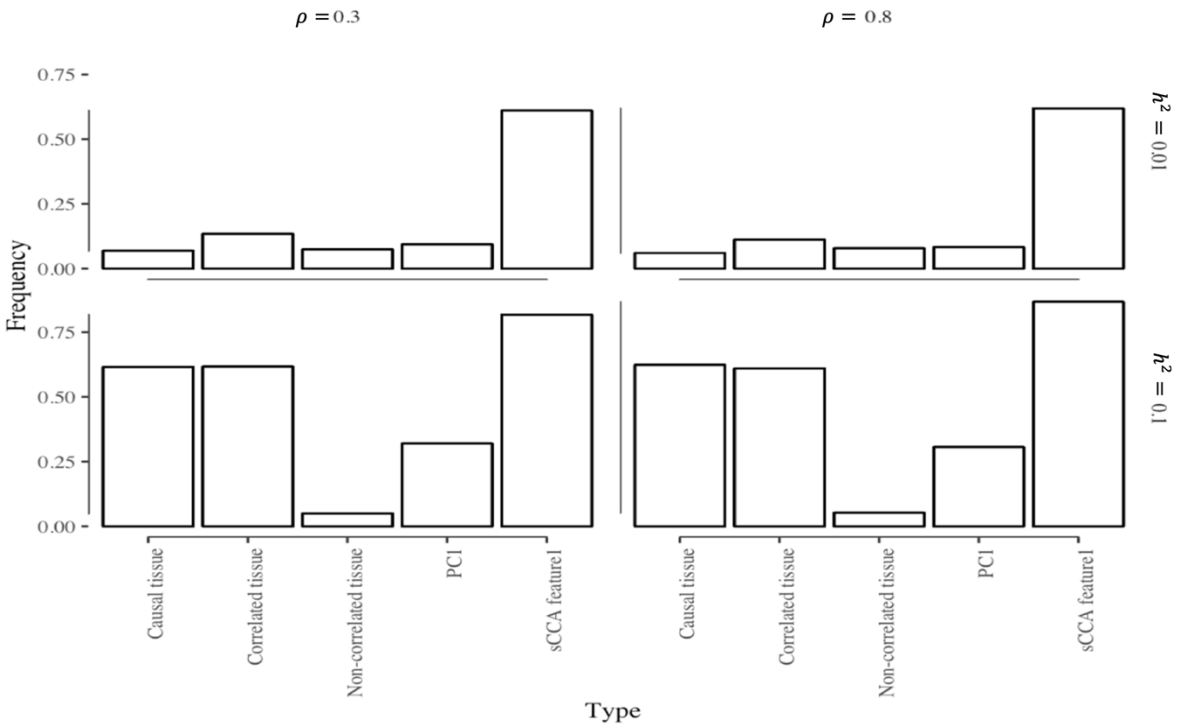
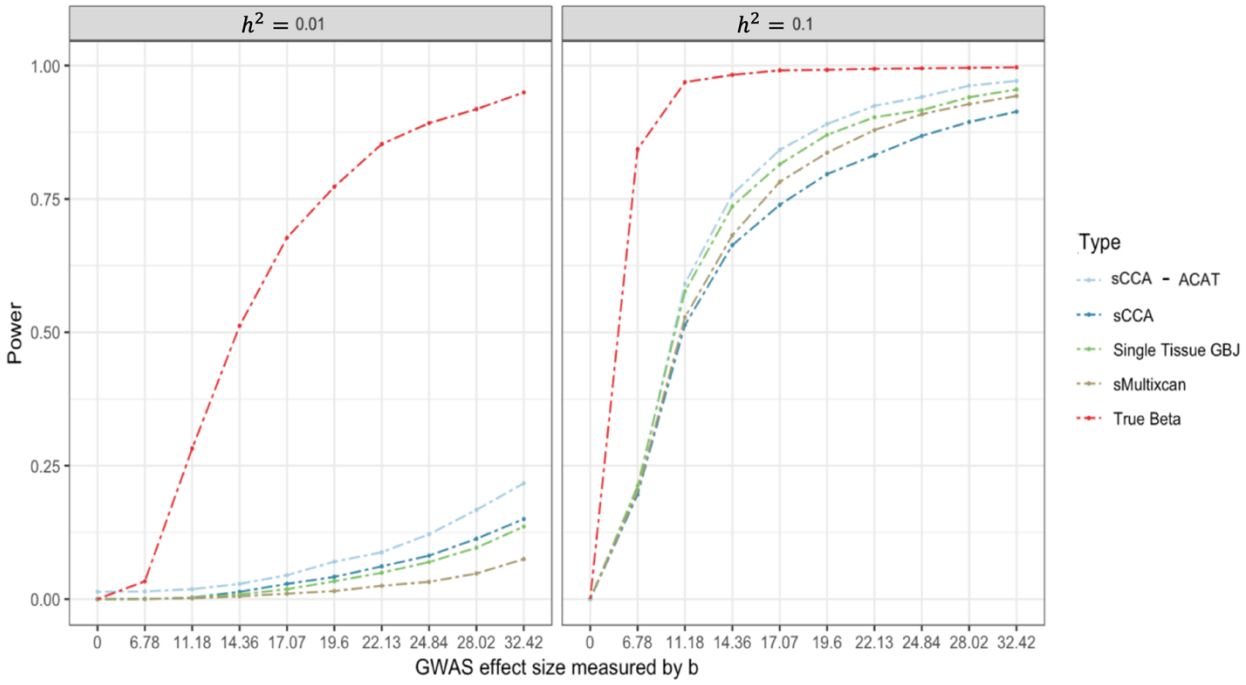


Fig 2. Proportion of significant (p<0.05) heritability tests for different expression features when cis genetic variation is associated with expression in some tissues. Here ρ denotes the strength of the genetic correlation between expression in the causal tissue and another tissue in which the expression is also associated with cis-germline variation ("correlated tissues"). "Non-correlated tissues" are the tissues where local germline variation is not associated with the gene expression. Here expression in half of the tissues is genetically correlated with that in the causal tissue, and the causal tissue is not observed (performance in the causal tissue is included as a reference). PC1 is the first principal component of cross-tissue gene expression; sCCA-feature1 is the linear combination of tissue expression values from the first pair of sCCA canonical variables. h^2 denotes the proportion of expression variance in the causal tissue explained by cis-genetic variation.



773

774 **Fig 3. Power comparison for cross-tissue TWAS methods.** Power (at α=0.05) as a function of
775 GWAS effect size. For each tissue, we randomly sampled the z-scores from this multivariate
776 normal and set b = $\sqrt{N_{gwas} \times r^2}$ to 0.00, 6.78, 11.18, 14.36, 17.07, 19.60, 22.13, 24.84, 28.02,
777 32.42 to achieve theoretical power of 5%, 10%, ..., 90% at alpha level of 0.05. That is, when r^2
778 =1% (when variation in gene expression in the target tissue explains 1% of the variability in the
779 trait), the GWAS sample size N_{gwas} ranges from 4,602 to 105,074. h^2 denotes the proportion of
780 expression variance in the causal tissue explained by cis-genetic variation. sCCA-ACAT:
781 combining 3 sCCA-features and 22 single-tissue tests with ACAT; sCCA: combining top 3
782 sCCA-features tests using a Bonferroni correction; Single Tissue_GBJ: combining 22 single-
783 tissue TWAS statistics using the GBJ test; s-MultiXcan: combining 22 single tissue based test
784 using s-MultiXcan); true weights: a TWAS test using the true (simulated) weights relating SNPs
785 to expression in the causal tissue.

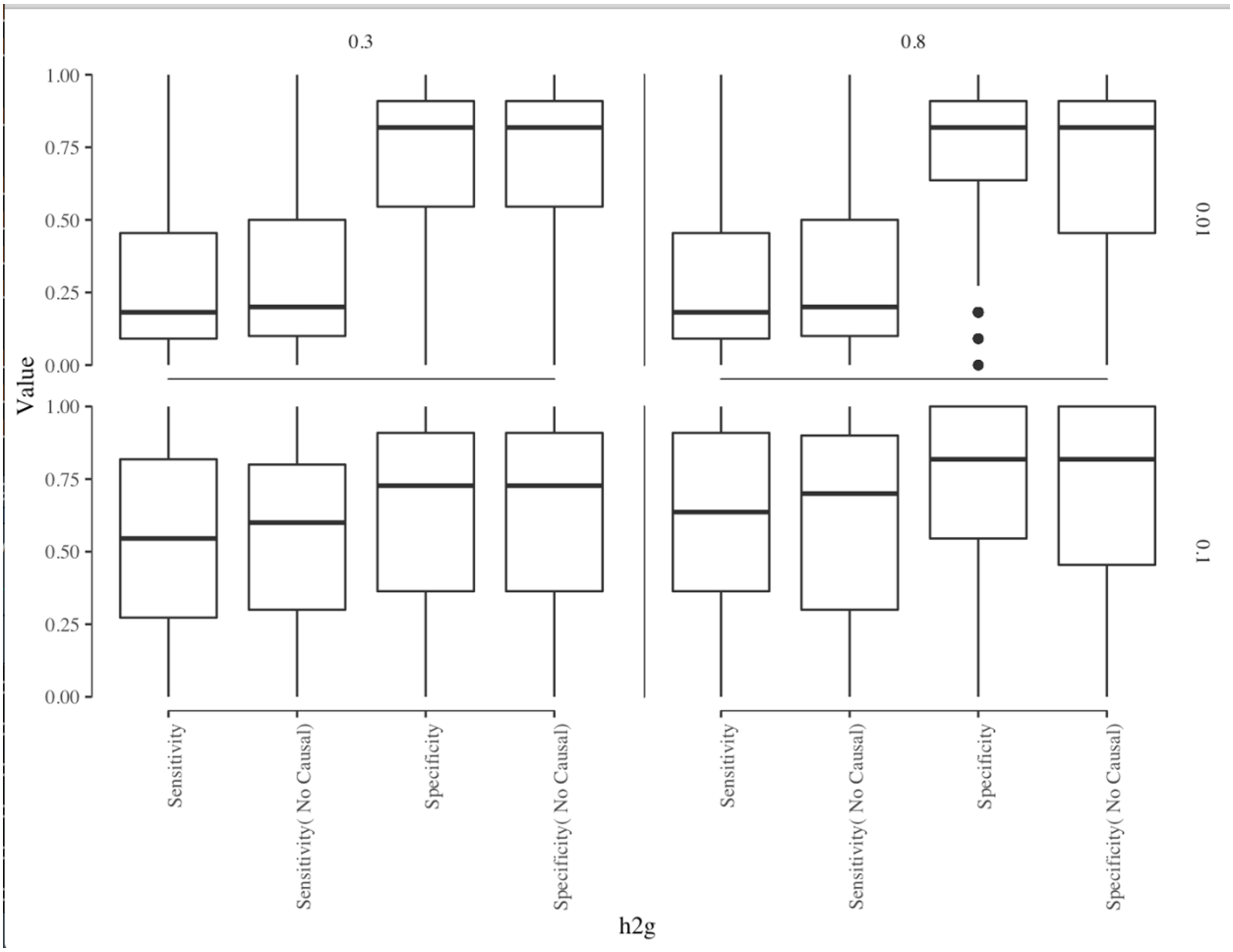
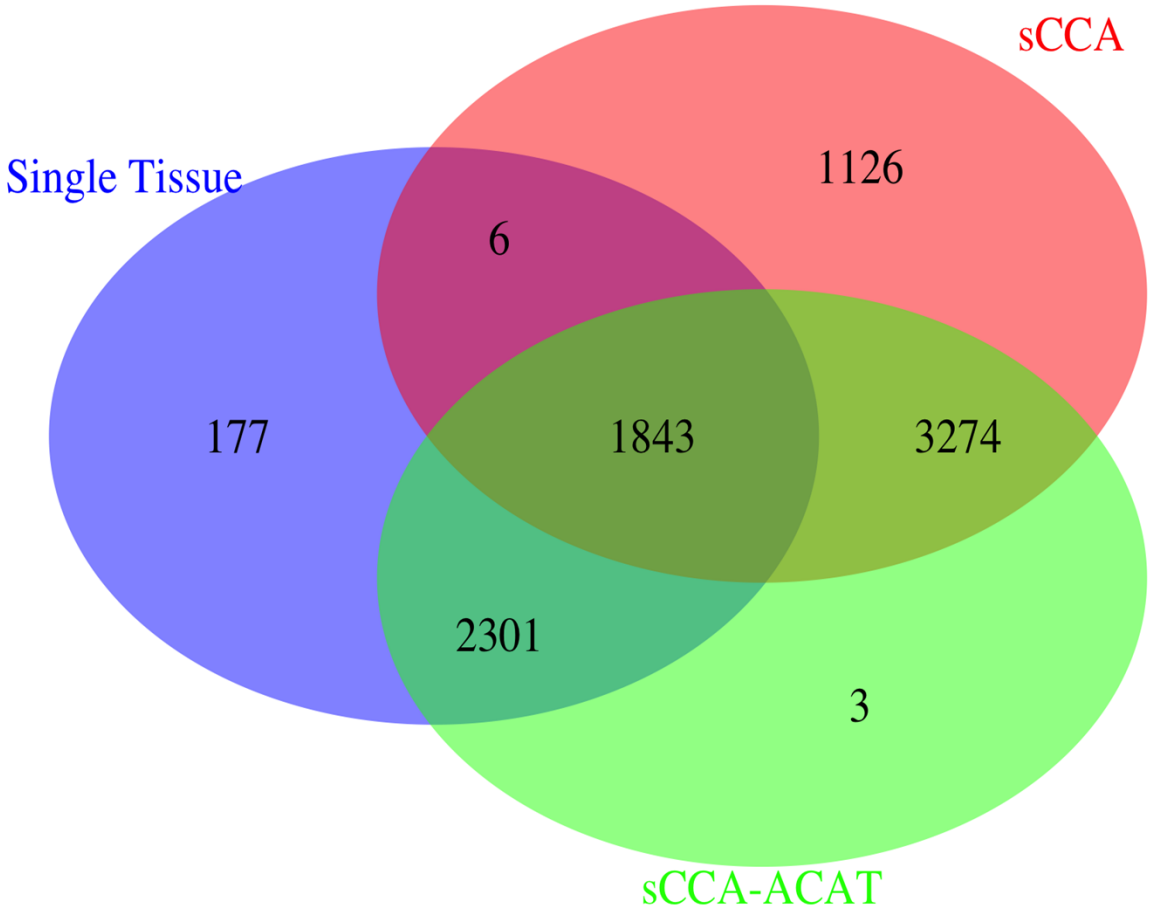


Fig 4. Sensitivity and Specificity of sCCA features.

The box plot of sensitivity and specificity of sCCA putting non-zero weights on the tissue where genotype regulates gene expression. We varied underlying gene expression heritability (h^2) and correlation (ρ) with the causal tissue as: (a) $h^2 = 0.01, \rho = 0.3$; (b) $h^2 = 0.01, \rho = 0.8$; (c) $h^2 = 0.1, \rho = 0.3$; (d) $h^2 = 0.1, \rho = 0.8$.



793
794
795 **Fig 5. Venn Diagram of the significant expression-phenotype associations.** The Venn
796 Diagram of the significant expression-phenotype associations for single tissue test results, sCCA-
797 TWAS test results and ACAT combined results. sCCA-ACAT: combining 3 sCCA-features and
798 22 single-tissue tests with ACAT; sCCA: combining top 3 sCCA-features tests using a
799 Bonferroni correction; Single Tissue: combining 22 single-tissue TWAS statistics using
800 Bonferroni.
801

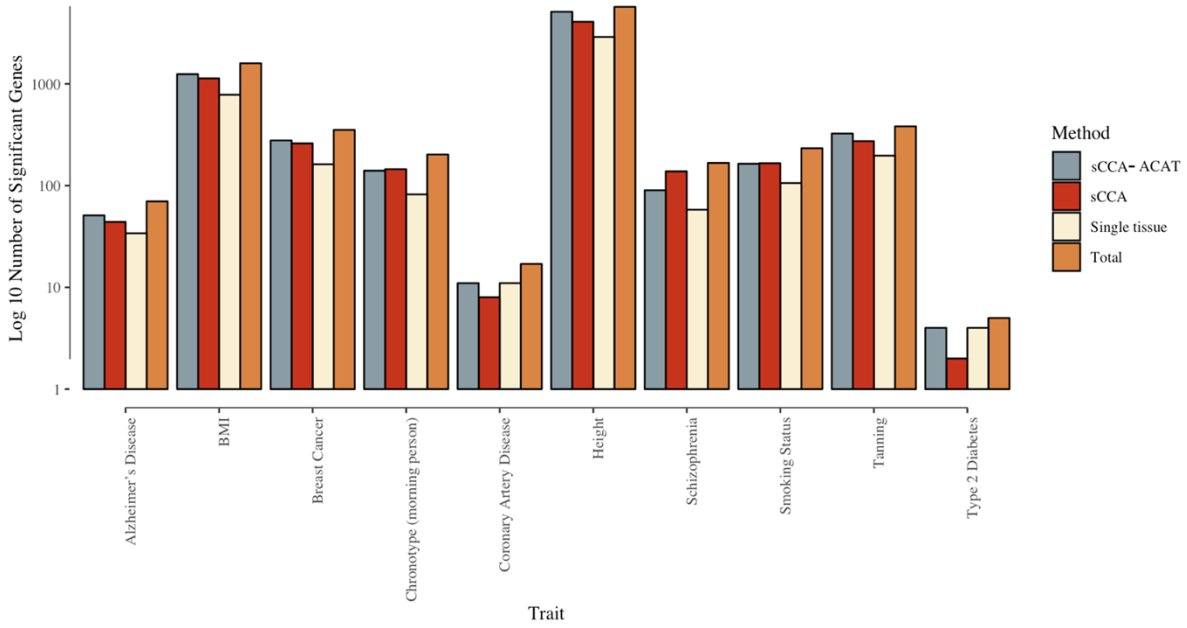


Fig 6. Number of significant genes identified by ACAT combined test, sCCA-TWAS, TWAS using single tissue GTEx data and the total number of significant genes identified by all three methods. Different phenotypes are arranged along the x-axis and the number of significant genes identified by ACAT combined test, sCCA-TWAS, TWAS using single-tissue GTEx data and the total number of significant genes identified by all three methods are shown in the y-axis on log 10 scale. The information about the phenotype are provided in Table 1. sCCA-ACAT: combining 3 sCCA-features and 22 single-tissue tests with ACAT; sCCA: combining top 3 sCCA-features tests using a Bonferroni correction; Single Tissue: combining 22 single-tissue TWAS statistics using Bonferroni.

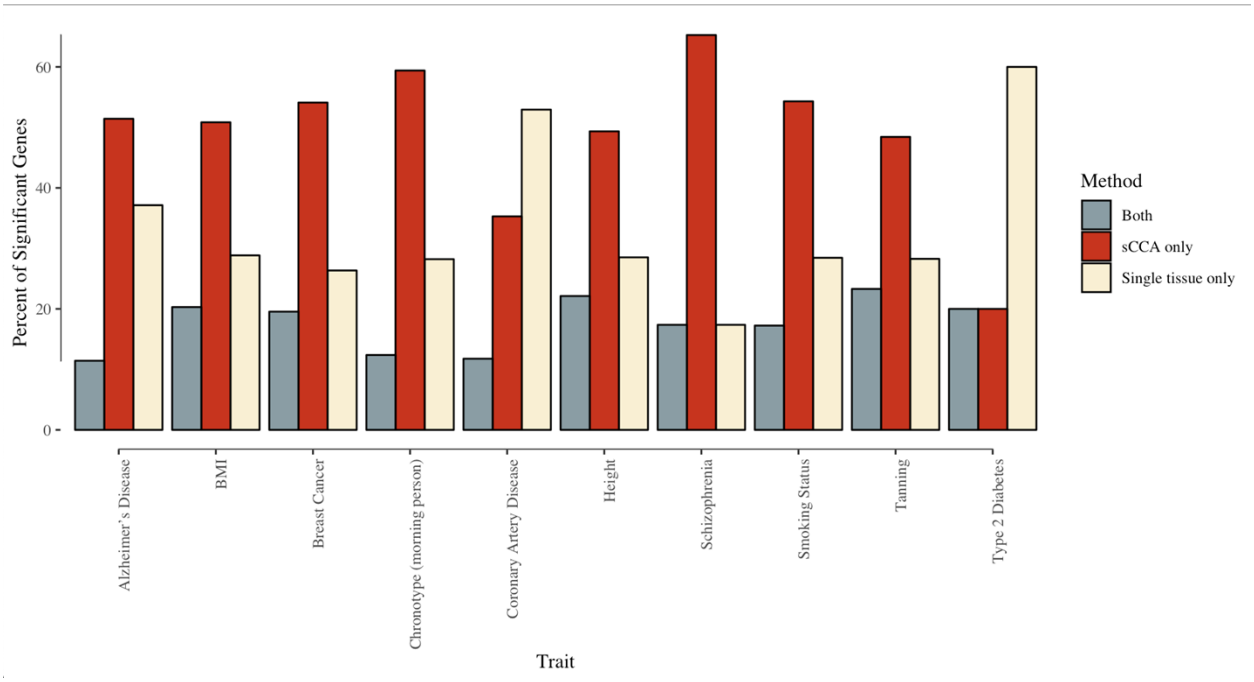


Fig 7. Percentage of significant associations identified by both single tissue TWAS and sCCA TWAS, by only sCCA-TWAS, and by only identified by single tissue TWAS, among all associations identified with sCCA cross-tissue TWAS or single tissue TWAS. Different phenotypes are arranged along the x-axis and the percentage of significant identified by both single tissue TWAS and sCCA-TWAS, by only sCCA-TWAS, and by only identified by single tissue TWAS are shown in the y-axis. The information about the phenotype are provided in Table 1.

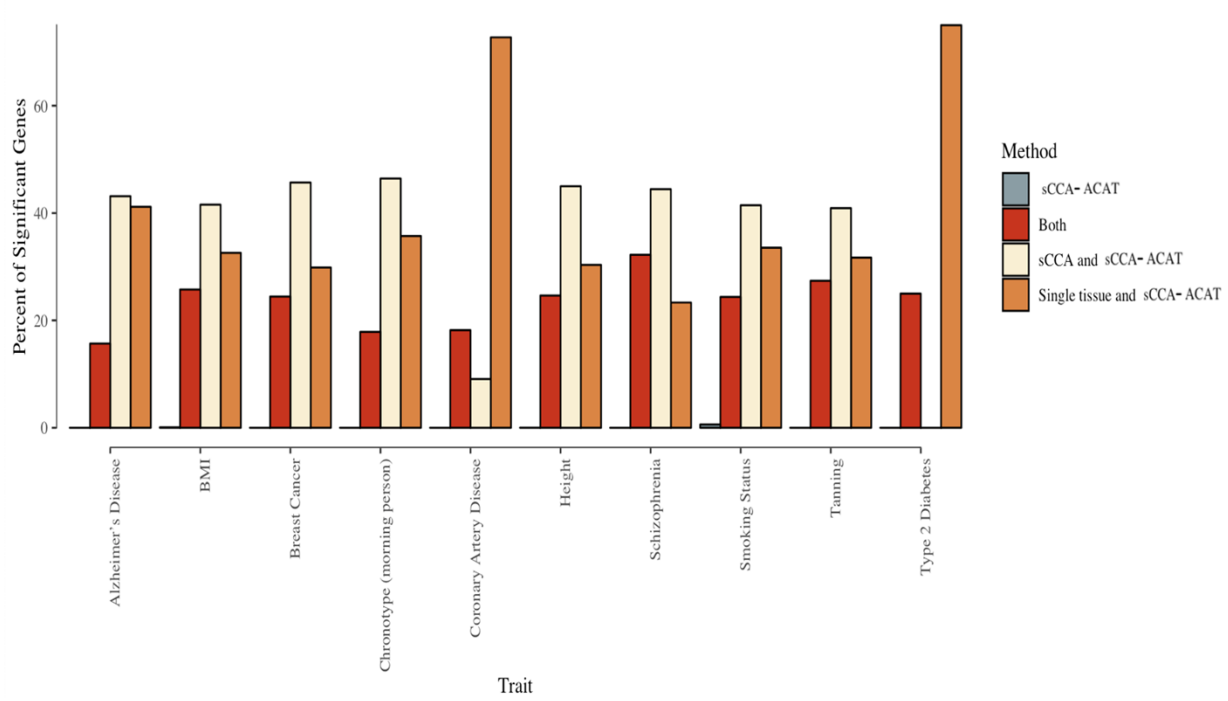


Fig 8. Percent of significant identified by only sCCA-ACAT, by sCCA-ACAT, sCCA-TWAS and single tissue TWAS, by both sCCA-TWAS and sCCA-ACAT, by both single tissue TWAS and sCCA-ACAT among all significant genes. Different phenotypes are arranged along the x-axis and the percentage of significant associations by only ACAT, by ACAT, sCCA-TWAS and single tissue TWAS, by both sCCA-TWAS and ACAT, by both single tissue TWAS and ACAT are shown in the y-axis. The information about the phenotype are provided in Table 1. sCCA-ACAT: combining 3 sCCA-features and 22 single-tissue tests with ACAT; sCCA: combining top 3 sCCA-features tests using a Bonferroni correction; Single Tissue: combining 22 single-tissue TWAS statistics using Bonferroni.

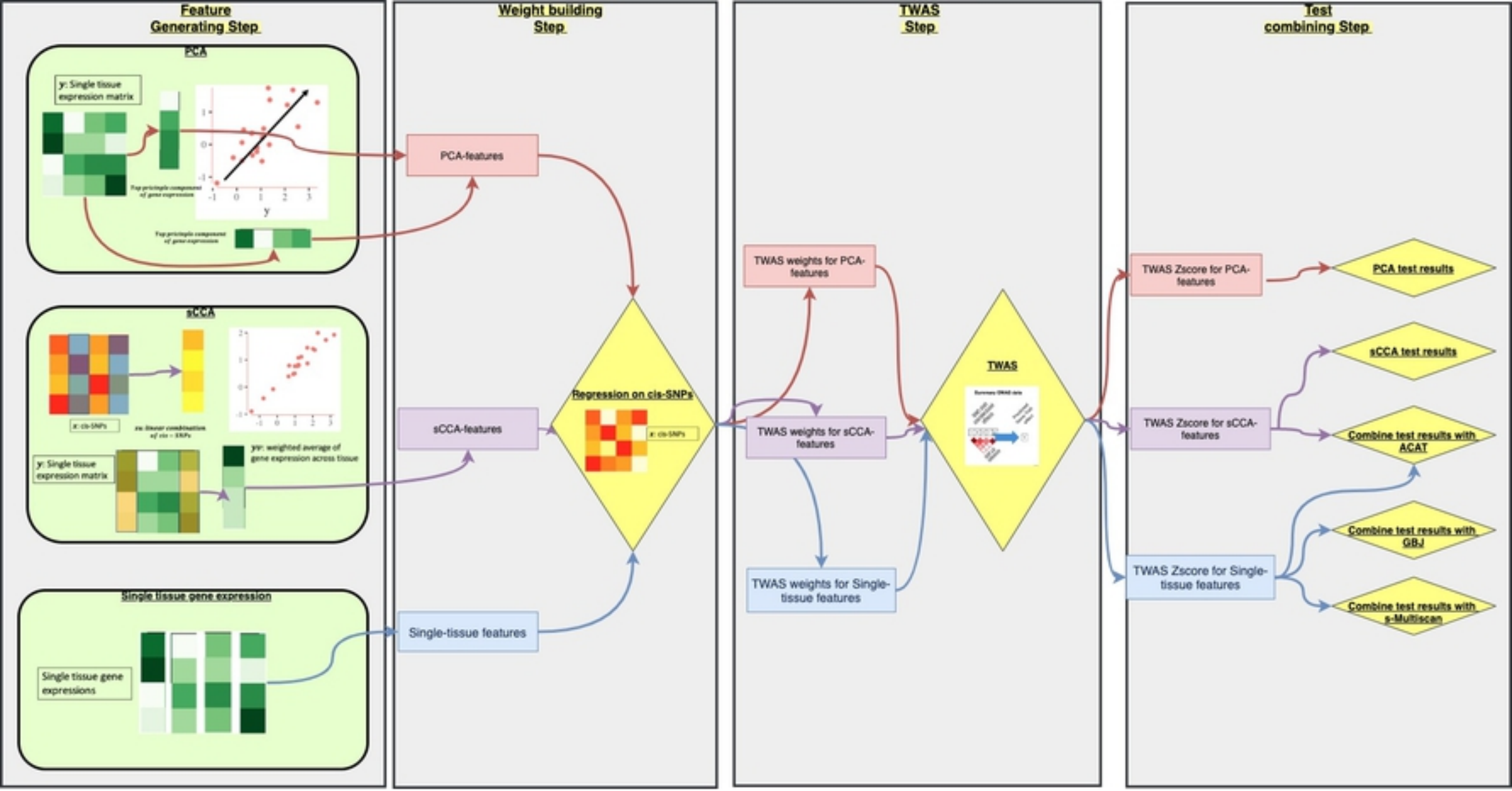

Feature Generating Step
PCA
y: Single tissue expression matrix
sCCA
x: cis-SNPs
y: Single tissue expression matrix
yv: weighted average of gene expression across tissue
Single tissue gene expression
Single tissue gene expressions
Weight building Step
PCA-features
sCCA-features
Single-tissue features
Regression on cis-SNPs
x: cis-SNPs
TWAS Step
TWAS weights for PCA-features
TWAS weights for sCCA-features
TWAS weights for Single-tissue features
TWAS
Test combining Step
TWAS Zscore for PCA-features
TWAS Zscore for sCCA-features
TWAS Zscore for Single-tissue features
PCA test results
sCCA test results
Combine test results with ACAT
Combine test results with GBJ
Combine test results with s-Multiscan

Figure 1